Pervasive and Mobile Computing 12 (2014) 1–16

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Review

Mining GPS data for mobility patterns: A survey

Miao Lin *, Wen-Jing Hsu

School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 26 November 2012
Received in revised form 17 June 2013
Accepted 26 June 2013
Available online 8 July 2013

Keywords:
Mobility pattern
GPS data
Ubiquitous computing
Mobile computing

ABSTRACT

With the help of various positioning tools, individuals' mobility behaviors are being continuously captured from mobile phones, wireless networking devices and GPS appliances. These mobility data serve as an important foundation for understanding individuals' mobility behaviors. For instance, recent studies show that, despite the dissimilarity in the mobility areas covered by individuals, there is high regularity in the human mobility behaviors, suggesting that most individuals follow a simple and reproducible pattern. This survey paper reviews relevant results on uncovering mobility patterns from GPS datasets. Specially, it covers the results about inferring locations of significance for prediction of future moves, detecting modes of transport, mining trajectory patterns and recognizing location-based activities. The survey provides a general perspective for studies on the issues of individuals' mobility by reviewing the methods and algorithms in detail and comparing the existing results on the same issues. Several new and emergent issues concerning individuals' mobility are proposed for further research.

Contents

* Corresponding author. Tel.: +65 97270920.
E-mail addresses: linm0018@e.ntu.edu.sg (M. Lin), hsu@ntu.edu.sg (W.-J. Hsu).

1574-1192/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.pmcj.2013.06.005

Table 1
A comparison of different types of data repositories for the study of individuals' mobility. Velocity, pause time, inter-contact time (ict for short) are the basic features needed in describing individuals' mobility. Resolution indicates the degree of spatial accuracy of the data. Scale indicates the area covered by an individual's data. GSM cell-tower data contain the ict information, whereas both the wireless network traces and GPS data need manual labels to indicate the ict.

Data type	GSM cell-tower data	Wireless network traces	GPS data
Resolution	km	m	m
Scale	Metropolitan area	Campus/work place	Global
Velocity	No	No	Yes
Pause time	Approximate	Approximate	Exact
Ict	Exact	Labels needed	Labels needed
Path info	Rough	Rough	Exact

1. Introduction

Mining individuals' mobility patterns is an emergent research area. Owing to the prevalence of mobile phone, a wide variety of mobility-related data are now being captured, including, e.g., the usage of WiFi, and communications between individuals via sending messages or making phone calls. The availability of such data on personal devices will inevitably engender the study of individuals' mobility patterns at an unprecedented scale both in terms of the areas covered by the trajectories and also the number of individuals involved in the study. Many interesting results have been obtained [1–4]. For instance, despite great dissimilarities in individuals' mobility areas, ranging from a few square kilometers to more than a few thousands square kilometers, individuals' mobility is found to be highly regular [1]. Most individuals return to a handful of locations and share a rather universal probability density function for places visited [2]. After removing individuals' idiosyncrasies, the rules that govern the individuals in exploring new locations or returning to the previously visited locations are found to be similar [3]. Also, members of a same social group, such as researchers from the same labs, present similar mobility behaviors [4].

The study on mining individuals' mobility patterns have found applications in a wide range of researches, such as personal positioning [5,6], complex system [1,7], computational sociology [4,8], wireless network analysis [9–11], etc. Both individuals and the society as a whole can benefit from the study of individuals' mobility. For instance, understanding individuals' mobility can help recognize regular activities or abnormal events [12–15], predict future movements or destinations [2,5,16,17], identify similarities among individuals [18,19]. The study may also be applied in traffic forecast [20], city planning [21,22], mobile virus control [7], epidemic prevention [23–25], evaluating mobile network protocols [9–11], logistics management [26], energy consumption and carbon footprint [27,28].

Due to the coarse description of individuals' locations provided by the GSM cell-tower data, however, many existing studies of mobility patterns, are incomplete and less than convincing. In GSM cell-tower data, a location is indicated by the cell tower that provides the connecting services. Hence the resolution for the location is constrained by the service area of a cell tower, whose size is in the order of hundreds of square meters to several square kilometers. The paths, often indicated by a series of discontinuous sudden jumps, are hardly observable in fine details between the destinations.

Besides the GSM cell-tower data [29], some other data repositories, such as wireless network traces [30–32] and GPS data [33,34], have recently become available.[1] A comparison among the data repositories in presenting the basic mobility features, is summarized in Table 1. Specifically, GPS data allow to track the movements of individuals in latitude and longitude along with timestamp. GSM cell-tower data are constrained by the service area of cell towers, whereas wireless network is mostly available within even smaller regions, such as campuses or work places, although their accessibility is being improved gradually. By comparison, GPS devices are able to provide positioning data globally. GPS data, however, suffer from three major limitations, (1) GPS signals are usually blocked indoor or underground places, (2) GPS devices may get interferences near tall buildings, and (3) continuously collecting GPS data may consume devices' energy quickly. These limitations have prevented collection of continuous high quality GPS records over long durations.

This survey will focus on the mobility studies based on high resolution positioning data, GPS data in particular, with a target on researches that aim to uncover the patterns in individuals' movements, e.g., [37–39]. The study of individuals' mobility may be broadly classified into two subareas, (1) mining mobility patterns, and (2) constructing mobility models. Typically, the following mobility patterns are of great interest: locations of significance, modes of transport, trajectory patterns and location-based activities. The mobility models of great interest are the ones that emphasize on the statistical patterns of the individuals learned from trajectories [3,9–11], or i.e., the trace-based mobility models. Since there is an excellent survey on this topic [40], we will not include this topic in this paper.

The remainder of this paper is organized as follows. Section 2 to Section 5 focus on issues and results in mining various mobility patterns. Section 6 highlights a few new and emergent research issues. In Section 7, we briefly conclude this review.

[1] A spatial–temporal datasets generator is given in [35]. Since this survey targets on mining of real-world mobility patterns, the issue of the generation of synthetic dataset will not be addressed here. Nevertheless, one of the studies based on synthetic data is included in Section 4.1 [36].

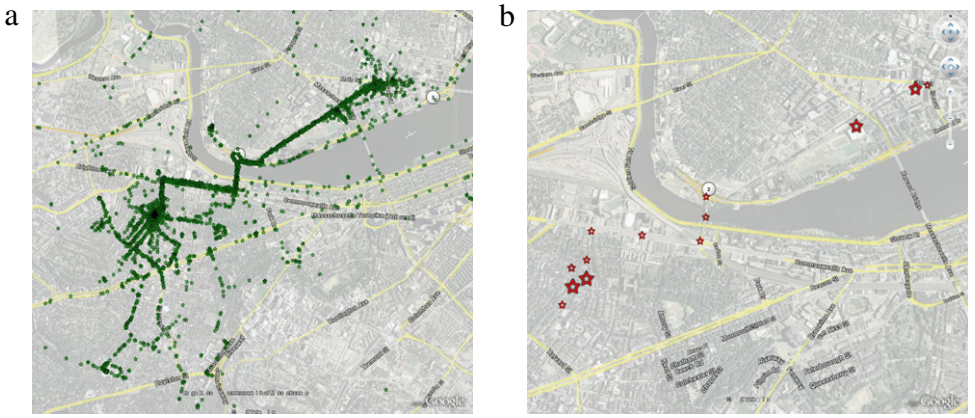

Fig. 1. (a) One individual's GPS traces mapped on Google Earth. (b) The red stars are the detected significant locations for this individual. *Source:* These two figures are from [6].

2. Locations inference and movement prediction

In the context of mobility studies, significant locations are the places where an individual or a group visits frequently or dwells for a long time. These locations may be different from locations that have geographic or social meanings, such as the parks, coffee shops, etc. Fig. 1(b) highlights a few significant locations learned from one individual's trajectory. Some of the locations carry geographic meanings, such as office buildings, but others may be merely the intersection points on the road, where the individuals may stay for a while. Another example may be given by an individual who regularly parks his car in a specific parking lot. Thus a large number of samples may be collected around this area.

In a typical GPS dataset, a trace is a time-stamped sequence of pairs of latitude and longitude. Because of measurement errors, the GPS coordinates of the same location may vary from time to time. Inferring the significant locations from GPS traces is the process of mapping the raw readings to discrete location symbols. In [41], Andrienko et al. summarize four steps for extracting the significant places from mobility data. Firstly, a subset of points indicating certain events, e.g., traffic jam, is extracted from raw data. Secondly, meaningful locations are generated by clustering the points based on the spatial–temporal attributes. The last two steps analyze the aggregated trajectories that fall with the clusters. In the following, we discuss a few recent papers according to the spatial–temporal clustering methods they used, e.g., the density-based clustering algorithm or its variants [16,42,43], or time-based clustering algorithms [44].

Ashbrook et al. [16] propose a two-step method to infer the significant locations, and further construct a Markov model to predict future locations. In the first step, the significant places are inferred by the points where the GPS devices lose the signals from the satellites. These places are clustered into locations using a variant of the K-Means algorithm in the second step. In the clustering procedure, the clusters are initially centered at K chosen points with a given radius, and iteratively move to a denser area, until no further changes arise in the clusters. When given a large radius, the significant locations may correspond to the city level, or the campus-level when given a small radius. A second-order Markov model is constructed based on the transition probabilities between locations. According to the high relative frequency of movement patterns between locations, the future location is predicted. Since the loss of GPS signals serves as the main clue to identify significant locations, mostly buildings are found. Other types of significant locations where the signals are continuously collected, such as certain open-air canteens, may hardly be detected.

Zhou et al. [42] propose a density- and join-based clustering algorithm called DJ-Cluster to infer significant locations. The dense points are those with at least certain number of other points lying within a distance of their neighborhood. The clusters are formed from a set of dense points, which are *density-joinable* when the neighborhood of the dense points share a common point. A two-step preprocessing procedure is introduced to improve the performance of the algorithm, which eliminates a GPS reading if either its speed is greater than zero or if it is within a small distance of the preceding reading. The experimental results indicate great improvements in terms of both recall and precision over those obtained from the K-Means algorithm. Moreover, DJ-Cluster overcomes the drawbacks of DBSCAN[2] [45] by reducing the memory and time requirements. However, some useful information in the original data may be lost during the preprocessing step. For instance, when a stay is incurred at a location and a large amount of readings arising from the adjacent locations are generated, then removing the series of reading within a few meters may conceal the fact that a stay is incurred.

Cao et al. propose a semantics-enhanced clustering algorithm (SEM-CLS) to extract semantically meaningful locations and further rank the locations through a unified probabilistic model [43]. SEM-CLS splits the locations of different semantics in a cluster and merges the locations sharing the same semantics from adjacent clusters, which enhanced the results

[2] In DBSCAN, it clusters the points that have a high density within a given area.

from OPTICS[3] [46] and K-Means. A two-layered graph is constructed to model the trajectory, and each layer represents the relations between the locations and the individuals as well as the relations among the locations. When ranking the significance of the locations, a unified probabilistic model, based on random walks with the Markov Chains built from each layer, is proposed along with an extended model, which is able to incorporate stay durations and distances between locations. Locations of various semantics can be extracted more accurately than those by using a single factor, such as rank-by-visit, rank-by-duration, Hyperlink-Induced Topic Search (HITS) [47], etc. Since the model is constructed based on the trajectory of a group of individuals, the detected locations are semantically meaningful for a group and individuals' bias is removed.

Hariharan et al. [48] construct a probabilistic model to condense, understand and predict individual's location based on a GPS log. A *stay* is a single instance for an individual to have spent sufficiently long duration in one place. A *destination* is generated from merging the stays according to the geographic scale of interest. Based on the time-dependent probability involved with the destinations, such as the probability of an individual starting from a destination at a given time interval, a first-order Hidden Markov Model is constructed to represent the probability for a transition between two destinations at the given time interval. After estimating the parameters through either a Markovian or non-Markovian solution, the relative likelihood of a new location is estimated from this model. Based on the data of two individuals over one year, the experimental results show that a non-Markovian approach is suitable for identifying typical activity patterns, while the Markovian approach is better in revealing atypical behaviors. However, whether the first-order Markov model is able to capture all of the individuals' activity patterns has not been validated.

Krumm et al. [49] propose a *closed-world* model and an *open-world* model to predict individual's future locations. A closed-world model merely considers the probability of where a driver has been to, and the destination refers to a rectangular region of the map. The open-world model enriches the closed-world model by introducing (1) the probability of a certain type of area being a destination, (2) the probability of each destination according to the GPS data, and (3) the driving efficiency and trip time. Both closed-world model and open-world model are trained based on the historical data, but a complete data model, sharing the theory with the open-world model, is trained by the data both before and after the test day. The experimental results show that in both the complete data model and the open-world model, the prediction errors decrease when the trip fraction increases, whereas the prediction errors merely fluctuate a little bit in the simple closed-world model. When associating the locations to a grid map, some large places may be separated and some irrelevant places are included, therefore the probability of each small grid of being a destination may not be directly relevant.

We have reviewed the approaches for the inference of significant locations and various probabilistic models along with efficient algorithms proposed to predict individual's next move with varying levels of success. Yet little has been discussed whether the predicting accuracy is already close to the limit. In the following paragraphs, we review some theoretical results on the predictability of individuals' mobility sequences [1,50,51]. Song et al. [1] define *predictability* as the maximum probability that any algorithm may at best achieve in predicting the next locations based on historical records. Given an individual's mobility sequence, two kinds of information are extracted, (1) the number of distinct locations N, and (2) the entropy rate $\hat{H}_n$ of the sequence which measures its complexity. The first information can be easily calculated from the sequence, while the entropy rate $\hat{H}_n$ can be estimated through the Lempel–Ziv algorithm [52]

$$\hat{H}_n = \left(\frac{1}{n} \sum_{i=2}^{n} \frac{L_i}{\log i} \right)^{-1} \tag{1}$$

where L_i is the length of the shortest substring starting at index i that has not appeared before.

The predictability $\Pi_{\max}$ derived from Fano's Inequality, is of the form,

$$\hat{H}_n = -[\Pi \log_2 \Pi + (1 - \Pi_{\max}) \log_2 (1 - \Pi_{\max})] + (1 - \Pi_{\max}) \log_2 (N - 1). \tag{2}$$

Lin et al. [51] analyze the predictability of individuals' mobility based on 40 individuals' GPS records, each lasted for 16 weeks. Specifically, a location is defined as a grid cell in a grid map. A grid cell is a rectangular region characterized by the *origin* and *spatial scale s*. The origin is a given point of reference. The scale s is a numeric value used to specify the size of the grid cell, where adjacent grid lines differ by $0.001^\circ \times s$ in latitude or longitude. The grid map is a collection of disjoint grid cells that collectively cover the mobility area in a GPS dataset. After specifying the size of each location and the origin of the map, an individual's trajectory is encoded into a location string. The predictability is inferred based on the analysis of the entropy and number of distinct locations discussed before. When the size of a location is about 3 km^2, or that of a village, the predictability is centered around 93%, which concurs with the finding based on GSM cell-tower data with relatively low spatial resolution. When shrinking the size of the locations to roughly that of a large building, the predictability is still around 90%. When varying the scale of the grid and hence the size of locations over a wide range, an invariance between the predictability and the size of the locations is observed. This property suggests that a trade-off between the predicting accuracy and spatial resolution of the locations is unavoidable, which has implications to algorithms

[3] It adopts the theory in DBSCAN, but the difference is that in OPTICS the clusters of points are not generated, whereas all the points are sorted according to the defined radius, within which a high density is achieved.

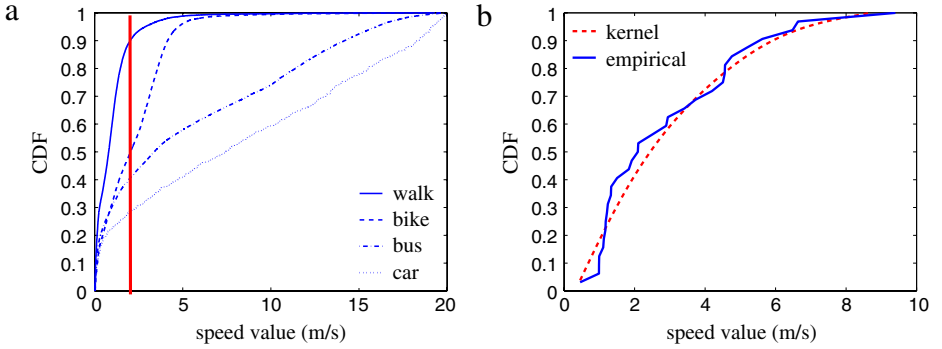

Fig. 2. (a) The cumulative speed distributions are differentiable among different modes of transport, especially the high speed distribution indicated on the right-hand side of the red vertical line. (b) The difference between cumulative speed distributions estimated empirically or from a kernel method. *Source:* These two figures are from [57].

for predicting the mobility of individuals. Moreover, the predictability is found to be independent of the mobility area covered by the individuals. Since the other two results [1,50] on predictability are based on other types of data rather than GPS data, we do not discuss them here.

Significant locations can forward the study of individuals' mobility on the remaining issues. In particular, typical locations, such as bus stop and parking lot, are the indications of changing modes of transport as well as activities. The locations are useful in representing the path patterns graphically.

3. Modes of transport

Mode of transport is another fundamental property of individuals' mobility. Individuals may apply a combination of modes of transport, such as walk, bike, bus and car, etc., in the daily activities. Basically, each mode of transport may present different speed values, but recognizing a mode is not easy, especially when multiple modes are utilized during a trip or when traffic congestion is involved. Besides speed values, some additional features, such as the presence of a car park or bus stop, heading change rate, stop rate, etc., are found to be effective in detecting modes of transport based on supervised methods [53–56], such as decision tree, support vector machine, Markov Models, etc. Various supervised methods are presented by Zheng et al. [53–55] and Reddy et al. [56]. Lin et al. [57] demonstrate that the high speed distributions are apparently differentiable among different modes, thus an unsupervised method is proposed.

Among all the approaches [53–55], the preprocessing step is similar, which divides the GPS log into segments based on criteria such as changing points, uniformity in duration and distance. The changing points are the locations where changes of transportation modes from walking to non-walking modes or vice versa are likely to have occurred. The differences between these approaches lie in the various features selected and the post-processing step applied. Basic features such as the distance covered, mean velocity, etc., are used in [53], while three additional features, namely, heading change rate, stop rate and velocity change rate, are further introduced in [54,55]. Several instance-based classifiers are developed to infer the transportation modes based on the corresponding features through standard training and predicting steps. In the post-processing step, the result is adjusted according to the transition probability between modes [53], and the probability between locations of each mode is used in [54,55]. Experimental results indicate that changing points based segmentation method combined with a Decision Tree algorithm shows the highest inference accuracy. Since manually labeled data are used in both the training phase and the post-processing phase, these methods suffer from laborious labeling as well as tedious training and predicting process.

Reddy et al. [56] detect the modes of transport based on the information from various sensors. Features such as the mean, variance, energy, and the DFT energy coefficients between 1 and 10 Hz, are calculated based on the data obtained from accelerometer, GPS, WiFi, and GSM. In addition to typical instance classifiers, the authors evaluated continuous Hidden Markov Model and Decision Tree combined with a Discrete Hidden Markov Model (DHMM). Among all these classifiers, Decision Tree combined with DHMM achieves the highest accuracy (about 93.6%) to date for detecting the modes of transport.

Different from the supervised methods mentioned before, Lin et al. propose an unsupervised method to infer modes of transport from unlabeled GPS logs [57]. The underlying assumptions are that, (1) the speed distributions from the segments of the same mode are similar, and (2) the speed distributions among different modes are differentiable, especially for the high speed distributions (as shown in Fig. 2(a)). Initially a GPS log is divided into segments that are likely to have been recorded under a single mode of transport. Given a segment S_i and its speed values segment $SV_i = \{v_{i_1}, v_{i_2}, \ldots, v_{i_1+|S_i|}\}$, each speed value v_j is assigned with probability $p_j = v_j^2 / \sum_j v_j^2$, where $i_1 \le j \le i_1 + |S_i|$. The high speed values from each segment are chosen from a weighted bootstrapping technique. Based on the re-sampled high speed values, both empirical method and kernel method are applied to estimate the underlying cumulative speed distribution for each segment. Fig. 2(b) shows that,

the kernel method generates a smooth cumulative distribution in contrast to the empirical method. The difference between the cumulative speed distributions is measured by the two-sample Kolmogorov–Smirnov Test,

$$D_{ij} = \max_x |F_i(x) - F_j(x)| \tag{3}$$

where $F_i(x)$ and $F_j(x)$ are the cumulative distribution estimated based on the high speed value x empirically or by a kernel method from segment i and j, respectively.

The p-value, calculated from the Kolmogorov distribution, serves as a metric to measure the similarity. Based on this metric, the linkage clustering algorithm hierarchically clusters the segments into a few clusters, where each cluster corresponds to the segments generated from one mode of transport. The segments that lack similarity in the high speed distribution are labeled according to the empirical probabilities for using various modes in transferring between specific locations. The unsupervised method achieves a level of accuracy comparable to that of the best known supervised methods. Since there is no need to manually label the data for training, and it can also cater for differences in individual behaviors, the unsupervised method may be more widely applicable than the supervised methods.

4. Trajectory mining

Trajectory mining is the most popular topic in the study of individuals' mobility patterns. A host of studies has been conducted on various mobility datasets, e.g., individual human mobility data [5,6,18,19,44], animal movement data [58,59], hurricane data [58,59], vehicle movement data [14,60–63], and synthetic mobility data [36], etc. We group these studies of trajectory mining into two sub-topics, namely, (1) trajectory clustering, and (2) trajectory patterns extraction. The main difference between the two topics lies in their objectives and results. The first topic aims to cluster or classify the trajectories generated by the same group of objects, e.g., animals of the same kind. The second topic aims to present a variety of methods for modeling common mobility behaviors and for predicting the future move.

4.1. Trajectory clustering

In this context, a cluster is usually a group of trajectories that share common sub-traces or are relatively "close" to each other. For extraction of sub-trajectories, some of the studies are conducted with the assistance of visualization [64,65]. To measure the proximity or closeness of the traces, a distance measure between trajectories is usually defined, then some variants of the density-based clustering methods are introduced in order to cluster the trajectories [36,58–61]. Specifically, both [58,59] apply the idea of DBSCAN, and [36] uses OPTICS, while [60] is based on fuzzy C-means, and [61] applies a sequential pattern mining method.

In DBSCAN, the core points that have at least *MinPts* points within its ϵ distance form the center of the clusters. Similarly, in [58,59], the distance measurement between two line segments is defined as the mean of perpendicular distance $d_\perp$, parallel distance $d_\parallel$ and angle distance d_θ, as shown in Fig. 3(a). Rather than clustering the whole trajectories, Lee et al. consider it to be more meaningful to cluster the common sub-trajectories [58]. Therefore, by using the principle of minimum description length [66], each trajectory is partitioned into a list of line segments, simulating the rapid changes of the directions. By making an analogy to the idea of DBSCAN, the clusters are the group of lines where each line is within ε distance to the core line, which has sufficient number of lines *MinLns* with that distance. The representative trajectory of a cluster is a sequence of line segments that show the general trend of the trajectories within the cluster. Further, motivated by the observations that discriminative features mostly appear in sub-trajectories and regions, Lee et al. classify the trajectory based on the features extracted from region-based clusters and trajectory-based clusters [59]. The target regions are of variable rectangular size and they contain the sub-trajectories of a cluster. These regions are extracted by iteratively partitioning the space following the principle of minimum description length, which is similar to [58] but with a different objective function. The trajectory-based clustering aims to find the homogeneous ε-neighborhood, indicating the trajectories within ε distance form the same clusters. The clusters of homogeneous ε-neighborhood is detected by the trajectory-version DBSCAN [58], and for each cluster the representative trajectory is similarly generated. The classification of the trajectories is based on the features generated from both the region-based and trajectory-based clusters. Both [58,59] effectively detect the discriminative features in each class' trajectory, which demonstrates high classification accuracy in animal movement data as well as hurricane data. However, a major issue with these two methods is that the temporal information is ignored in measuring the distance.

Nanni et al. [36] cluster the trajectories by incorporating the temporal information into OPTICS [46]. Recall in OPTICS the points are reordered by a characteristic distance, namely, the reachability-distance [46]. For clustering the trajectories, the distance between two trajectories τ_1 and τ_2 in time interval T is given by

$$D(\tau_1, \tau_2)|_T = \frac{\int_T d(\tau_1(t), \tau_2(t))}{|T|} dt \tag{4}$$

where $d(x, y)$ is the $2D$ Euclidean distance and $\tau_i(t)$ is the position of the trajectory τ_i at time t.

The problem of clustering the trajectories is formalized to find the best temporal interval within which the minimum mean reachability-distance of all the non-noise points are achieved, namely, TF-OPTICS. The clustering result in each time

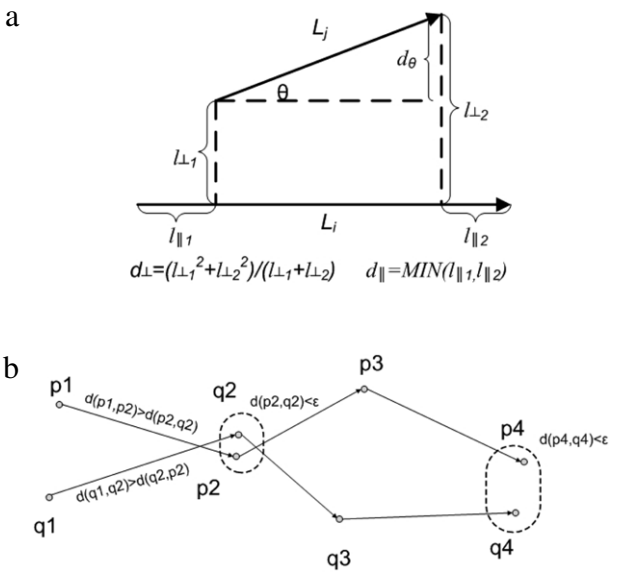

Fig. 3. (a) The three components, namely, perpendicular distance $d_\perp$, parallel distance $d_\parallel$ and angle distance d_θ, in measuring the distance between trajectories in [58,59]. (b) A simple example for the Pairwise Curve-Merging algorithm given two traces P and Q according to tolerance ϵ in [6]. The distance between p_2 and q_2, denoted by dist(p_2, q_2), is less than ϵ, and also dist(p_2, q_2) is less than the previous jump distance on the corresponding traces, e.g., dist(p_1, p_2), dist(q_1, q_2), then either p_2 is replaced by q_2 or vice versa. Fang et al. prove that independent of the choice of replacing either p_2 by q_2 or vice versa, the Hausdorff distance [67] between two traces is always less than ϵ.
Source: Figure (a) is from [58,59], and Figure (b) is from [6].

interval is evaluated by considering both the negative mean reachability-distance, and the case in which long time interval has a lower quality value than that achieved by the subset of the long time interval. Thus, the quality function is given by

$$\mathcal{Q}_2(D, I, \varepsilon') = \frac{-R_c}{\log_{10}(10 + |I|)} \tag{5}$$

where R_c is the mean reachability-distance within the time interval I, D is the trajectory set, and ε' is the density threshold parameter in OPTICS.

The algorithm for searching the maximum time interval is also given. Specifically, an initial time interval is randomly chosen from the subset of the largest time interval. The iteration is to continually find a better time interval that achieves larger $\mathcal{Q}_2(D, I, \varepsilon')$ in the set that extends the previous set by one unit time both backward and forward. The effectiveness of the TF-OPTICS is evaluated by C4C, a synthetic datasets generator [35]. It shows that, (1) the TF-OPTICS is effective since the computation time grows linearly with the number of trajectories, and (2) with respect to the clustering results, TF-OPTICS always shows better clusters than K-means, average-based hierarchical approach, or OPTICS, while it shows low coverage than the other methods. The main problem of this method is that, in the synthetic mobility traces, the mobility traces are normally generated at a regular time. In the real-world movement data, however, it is common to have some traces collected irregularly in time, therefore these trajectories may mislead the clustering results.

Pelekis et al. [60] propose a variant of Fuzzy C-Means (FCM) to cluster the trajectories with the discovered Centroid Trajectory (CenTra). Basically, each trajectory is divided into p sub-trajectories, with each of equal temporal intervals, and each sub-trajectory is mapped to the regular grids with user defined size. By extending the sub-trajectory ε-buffer along the directions that is vertical to its moving direction, the feature of each sub-trajectory is represented as a fuzzy set specifying the extent of the sub-trajectory occupies a list of cells, called region. The distance between two trajectories is measured by, (1) the distance between corresponding regions that contain the two sub-trajectories, and (2) the distance between fuzzy sets representing the features of the sub-trajectories. The clustering of the trajectory is initially to find the Centroid Trajectory that exceeds a region density threshold, and the clusters are generated by evaluating the distance to the Centroid Trajectory. The evaluations of this approach based on the trajectories of GPS-tracked trucks show that, the cluster representative fits the real movements better than an earlier method called TRACLUS [58]. However, the authors did not elaborate on the suitable choice of the number of sub-trajectories, which is a crucial factor in determining the length of the sub-trajectory and also the size of the region.

Lee et al. demonstrate that the sequential patterns are useful for classifying the trajectories on road networks [61]. The sequential patterns are a list of ordered road segments that exceed the minimum support threshold, which is determined by the information gain. The discriminative features are selected according to the *F-score* by comparing the negative set and positive set, and Support Vector Machine is used to classify the trajectories according to the selected features. The effectiveness of the methods is verified by using both synthetic data and real taxi trajectories. However, in both dataset the

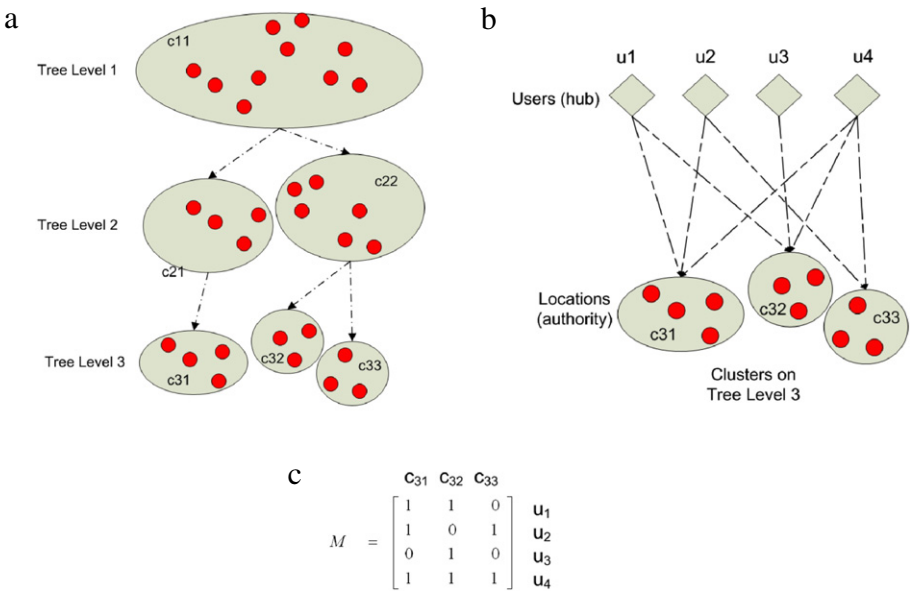

Fig. 4. (a) A framework for demonstrating the arrangement of the stay points, represented by the red nodes, in the tree-based hierarchical graph. In each level of the tree, the stay points are grouped into different clusters, based on which each individual's sequence is constructed according to the different tree level. (b) A sample for demonstrating the connections between individuals and locations at the third level of the tree on the left. (c) The adjacency matrix showing the relation between individuals (hubs) and locations (authorities) according to (b).
Source: Both figures are from [18].

number of classes is small, specifically two, thus the classification effectiveness on more classes is unclear, especially when the trajectories share some common roads.

4.2. Graph-based trajectory mining

There are two types of approaches for describing trajectory patterns from raw traces. In the first type of approaches, the trajectories are described as transitions among places of significance to the given scenarios. Therefore, individuals' mobility is described as a graph, where each node indicates the places the individuals have passed by and the edges are enriched with other features, such as the duration of transitions and the frequencies [5,6,18,19,44]. In the second type of approaches, the trajectories are converted to spatio-temporal items [14,62,63], in which the frequent patterns and periodic patterns are discovered.

Fang et al. [5,6] construct a personalized activity map from raw trajectories and further model individual dynamics. Significant locations are detected based on both dwelling time and visiting frequency. Representative paths are extracted by Pairwise Curve-Merging (PCM) algorithm. At each time, PCM algorithm iteratively compares two traces, e.g., trace P and Q, under a given tolerance ϵ, such that each point p from P is compared with its nearest neighbor q in Q. If p is within the tolerance distance of q, and if p is closer to q than any adjacent point of p in P, then p is replaced by q, where an example is given in Fig. 3(b). The personalized activity map, including significant locations, representative paths, and some additional information such as edge width, speed and number of traversals on the edges, is useful for modeling individual dynamics [5]. Individual's current location will be mapped to the nearest edge on the map. When reaching a crossroad, an outgoing edge is chosen by applying a particle filtering-based inference along with the historical probability of turning at each outgoing edge. However, using particle filtering-based inference to predict the location at a crossroad requires sufficient number of historical trajectories because sparse data may bias the probability of choosing a road.

In [18,19,44], the path patterns are presented in a graph-theoretic manner, where the definitions for the nodes in the graph are different.

A tree-based hierarchical graph (TBHG) is proposed for modeling individual's location histories, based on which some useful information is extracted [18,19]. A location history is a sequence of *stay* points, where an individual stays over a minimum period. The TBHG, consisting of a tree-like hierarchy and a graph associated with each level of the tree, is constructed from the location histories of all individuals. The hierarchy is obtained from all the individual's stay points in a divisive manner. The graph associated with each level of the tree is obtained by sequentially connecting the clusters in a same journey. For instance, a three-level hierarchy is presented in Fig. 4(a), where smaller clusters are formed on the lower level of the tree, and larger clusters are on the higher level of the tree.

Locations of interest and individual travel experiences are inferred based on a HITS-based inference model [19]. In this model, each individual is considered as a hub and the clustered locations at a specific level are taken as authorities, as shown

Table 2
A sample of continuous trajectories and discontinuous trajectories with origins and destinations, and it is from [44]. In the discontinuous trajectories row, a × indicates the loss of current trace. If the trajectories are continuously collected the time-based clustering algorithm can successfully detect the significant locations a, b, and c under a density threshold of 2. In the discontinuous trajectories, however, none of these significant locations are observed since all the locations bear a density less than 2 within a temporal neighborhood. The underlying assumption is that only the locations in the adjacent traces are within the same temporal neighborhood. For instance, in the discontinuous case the location a in trace 1 is not temporally adjacent to the same location in trace 3, thus the requirement of minimum density is not met.

	Trace 1	Trace 2	Trace 3	Trace 4	Trace 5	Trace 6
Continuous	$a \to b$	$b \to c$	$c \to a$	$a \to b$	$b \to c$	$c \to a$
Discontinuous	$a \to b$	×	$c \to a$	×	$b \to c$	×

Table 3
An example of the containment relation. Given two spatial–temporal sequences T and I, each of them is indicated by the transitions between location along with time. In sequence I, the locations b and d are considered as synonymous since they are close to each other. Given the time tolerance τ, T is contained in I if the corresponding travel time satisfies, $|t_1 - t'_1| \leq \tau$ and $|t_2 - (t'_2 + t'_3)| \leq \tau$.

$T : \{a\} \xrightarrow{t_1} \{b\} \xrightarrow{t_2} \{c\}; I : \{a\} \xrightarrow{t'_1} \{b; d\} \xrightarrow{t'_2} \{f\} \xrightarrow{t'_3} \{c\}$

in Fig. 4(b). This method is similar to HITS, in that the good hubs and good authorities are always connected. Let a be the authority scores and h be hub scores, e.g., according to the example in Fig. 4(b), $a = (a_{31}, a_{32}, a_{33})^T$, $h = (h_1, h_2, h_3, h_4)^T$, where each a_{3i} is the score of the cluster c_{3i} and h_i is the score of u_i. The relation between authority scores and hub scores are calculated as $a = M^T \cdot h$, $h = M \cdot a$, where M denotes the adjacency matrix which is derived from the individuals' travel histories, e.g., the adjacency matrix in Fig. 4(c) is derived from Fig. 4(b). This method outperforms some baseline algorithms such as rank-by-count and rank-by-frequency either separately or jointly, in ranking the interesting locations and inferring travel sequences. The similarity between individuals is quantified through a similarity metric considering both the sequence of the trajectory and the hierarchical properties of the locations in the graph [18]. Although individuals sharing similar travel sequences can be effectively measured in [18], a group of colleagues or friends who do not share similar travel sequences may not be differentiable merely through similar sequences.

Chen et al. [44] mine the movement patterns from individual GPS trajectories, and further predict both the destination and route through a pattern matching process. Forward–Backward Matching, a variant of time-based clustering algorithm, is proposed to cluster the origins and destinations from the trajectories. It initially generates the candidate clusters both in a forward and backward manner, and merges the candidate clusters that are at a small distance. Forward–backward Matching avoids the case when the trajectories are not continuously collected. Table 2 illustrates several such instances. The remaining points, except for the origins and destinations, are labeled according to a space-partitioning method. Chen et al. avoid the answer-loss problem [68] in the following method. Firstly, each cell is labeled by the origin–destination pairs of the trajectories that pass it along with the counts. The cell c_i is merged into its adjacent cell c_j, under the case that each origin–destination pair with a path passing through cell c_i has a count less than that in cell c_j. Therefore, the trajectories are simplified into, (1) origin–destination pair, and (2) a cell sequence along with the pairs of origin–destination with a path passing through it and the corresponding counts. The proposed algorithm is superior to the 1st- and 2nd-order Markov model with higher accuracy in 1-step prediction, and resulting in smaller Hausdorff distance [69] in continuous route prediction. Considering that mere geographic information, such as the location at each time step, is applied in the proposed algorithm, incorporating additional information, such as time, date, modes of transport, etc., may increase the performance of the algorithm.

4.3. Trajectory mining as spatial–temporal item

By describing the trajectories as a list of regions-of-interest or rectangular regions, the path patterns are considered as the frequent spatial–temporal item sets in [14,62,63]. Note that, [62] is the only study considering the temporal information in mining path patterns.

The objective of the study in [62] is to find the trajectory patterns that exceed the minimum support. A trajectory pattern is defined as the trajectories sharing both similar sequence of locations and the travel times. A spatial–temporal sequence is a location sequence with travel times. The *containment* relation among the sequences is defined based on the neighborhood function and the time tolerance, shown in Table 3. The neighborhood function is used to cluster the locations into meaningful areas, i.e., static or dynamic Regions-of-Interest. The static Regions-of-Interest are generated before the mining process, while the dynamic Regions-of-Interest are generated along with the process of mining the sequences. The experimental results on real and synthetic data indicate that the running time of both algorithms are almost linear to the length of input sequences. The main problem in defining the trajectory pattern in [62] is that, the transition time is not well defined, which involves both the travel time between two locations as well as the dwelling time at the destination.

The mined pattern or the so-called *T-Pattern* is useful in predicting the next location [70]. Firstly, the extracted T-patterns are ordered in the *T-Pattern tree*, where each node contains the entries of region ID, support, and the pointers to the children. The region in this case is referring to the regular grid of high visiting frequency. One node points to another node only when they are observed sequentially in the trajectories, and the connection is enriched by the minimum and maximum travel time between the corresponding regions in the nodes. When predicting the next location, the given trajectory history is matched with the tree according to the specified spatial and temporal tolerance. By tuning the spatial tolerance, indicating the maximum distance of the current trajectory to the matched T-pattern tree, the balance between the predicting rate and accuracy is adjustable. The authors suggest that the major contribution is that the prediction based on T-pattern tree allows the end users to specifically choose between predicting rate and the accuracy by tuning the spatial tolerance. A similar study about balancing between the predictability and the accuracy of the predictions is also reported in [51], which is discussed in Section 2. However, when the spatial tolerance is small, e.g., 10 m, the predicting rate is less than 40%, and the corresponding predicting accuracy is around 60%. When the predicting rate is low, they did not propose any method to handle the case. Some possible solutions regarding the spatial and temporal relevance of the locations can be applied in these cases [71]. For instance, when the current prediction is not in the existing prediction cases, the choice can be made according to the visitation frequency of each location, the distance to the current location, as well as combining the information of these two.

Cao et al. [14,63] aim to mine frequent patterns and periodic patterns from a spatio-temporal sequence extracted from a GPS trajectory. A series of rectangular regions are used to represent the trajectory into a spatio-temporal sequence. Frequent singular patterns are discovered by using a heuristic called *growing* [14]. A segment with median length is first chosen. In the filtering and verification steps, the chosen segment is merged with other segments that share similar characteristics. This process is repeated until either all the segments have been assigned to a region or no region is found for the segments—in the latter case the segments are outliers. To mine the frequent substrings, a stack is employed in a substring tree, which is a rooted directed tree with the root linking to multiple substring sub-trees. According to the experimental results on verified bus trajectories, the proposed method is more effective than the grid-based method, since some of the sequential patterns within a grid cannot be captured in the grid-based method. Also, the substring tree method is more efficient, in terms of execution time, than the level-wise method and grid-based method.

The objective in [63] is to discover all the valid, frequent and nonredundant patterns with respect to a minimum support. A *valid pattern* is a sequence of valid regions, with each representing a dense cluster following the definition of DBSCAN [45]. A pattern is *redundant* if it can be implied by the other patterns. The frequent 1-patterns are initially generated from the dense clusters. An example of valid pattern and redundant pattern is given in Fig. 5(b) and (c). Three methods are proposed to find longer patterns, which are a variant of level-wise Apriori-TID algorithm [72] (STPMine1), a faster top-down approach (STPMine2), and a simplified version of STPMine2 (STPMine2-V2). In STPMine1, the candidates for the frequent patterns with length k are generated from those of length $k-1$. The ones with support exceeding a threshold and an additionally valid pattern are chosen to be the frequent patterns of length k. In STPMine2, the trajectory is converted to a time series sequence, where periodic patterns are mined according to a modified algorithm in [73]. STPMine2-V2 mines the patterns approximately in an efficient manner. The experimental results show that both STPMine1 and STPMine2 identify the same longest patterns, but STPMine2-V2 is able to find similar longest patterns. Both STPMine2 and STPMine2-V2 are more efficient than STPMine1, since the former two scan the data less number of times than the latter. However, in all the cases, the length of periodic patterns has to be specified, and the frequent periodic patterns of different lengths cannot be automatically identified.

In summary, the key point in the study of trajectory mining is to find certain representative locations[4] in describing the raw trajectory. By converting the raw trajectory into a list of representative points, the topics such as trajectory clustering and trajectory pattern extracting can be conducted in a similar manner to the study of clustering, frequent item mining or graph studies.

5. Location-based activity recognition

Inferring location-based activities is a procedure of learning behaviors from movement or positioning related data. The data can be, for instance, the reading from the GPS devices, the road that the individuals currently on, the information from a digital map, etc. The behaviors can be the destination for the trip, the purpose of this trip, or the kind of activities that will take place at the destination, etc. A few location-based activities can be recognized through GPS data, such as working, visiting friends, on or off a car, dining, shopping, etc. Generally, the individual's mobility is modeled by types of conditional models, where a few examples are shown in Figs. 6 and 7.

Do et al. infer individual's next place and the dwelling time at the place by using a factorized conditional model according to the extracted features [17]. The factorized conditional models can greatly reduce the size of parameter space than that from the full conditional model given in Fig. 6. Specifically, the location information they applied is from the source of GPS sensors and WiFi data. The individuals' trace is converted to a sequence of meaningful locations indicated by grid cells, which are obtained by merging the locations recorded inside. The tasks of predicting the next location and the duration of

[4] In the study of trajectory mining, not all the representative locations in describing the trajectories share the idea of significant locations that we have discussed in Section 2 since in some cases all the points in the raw trace are retained.

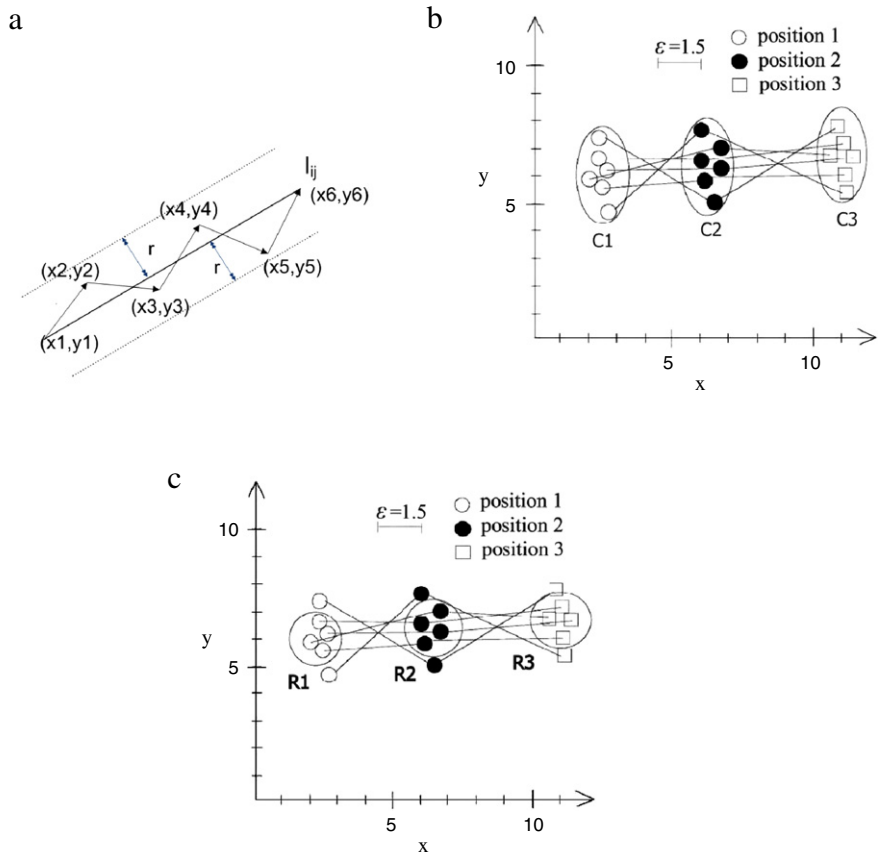

Fig. 5. (a) A sample of simplified trajectory pattern in [14,63]. The trajectory for (x_i, y_i), where $1 \leq i \leq 6$, is simplified by its representative line l_{ij} according to the distance threshold r. The path patterns of two trajectories are considered to be similar if the corresponding representative lines satisfy two conditions, namely, (1) the angle between two lines is less than a threshold, (2) the distance between the two lines is less than a threshold. (b) and (c) A sample of valid and redundant patterns defined in [63], respectively. In Figure (b), each point falls to the corresponding cluster indicated by the ellipse, and the valid pattern is $P = C1 - C2 - C3$, where each Ci indicates a cluster. Figure (c) indicates a pattern $P' = R1 - R2 - R3$, which can be implied from P, thus it is a redundant pattern.
Source: Figure (a) is from [14], and Figures (b) and (c) are from [63].

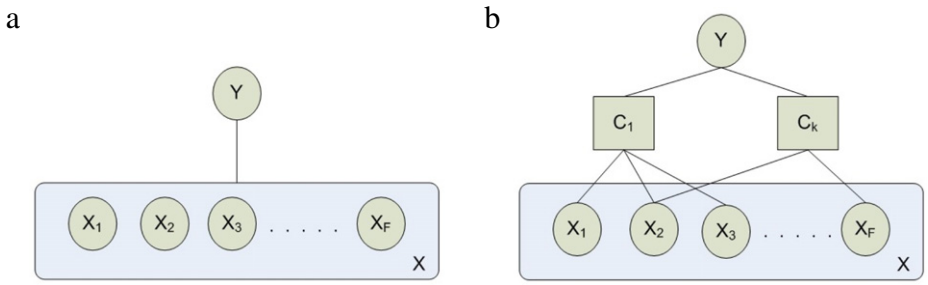

Fig. 6. (a) The general full conditional model. (b) The factorized conditional model given in [17]. The number of parameters of the full conditional model increases exponentially with the increasing size of the feature space. Therefore, when the feature space is large, the required training dataset becomes very large. However, by handling the features in a group rather than individually in the factorized conditional models the parameter space may be greatly reduced, which is especially useful in modeling individual's mobility.

stay are reduced to finding the conditional relations between the feature space **X** to the behavior space **Y** in the factorized conditional model

$$P(Y|\mathbf{X}) = \frac{\prod_{k=1}^{K} P_k(Y|\mathbf{C}_k)^{w_k}}{Z(\mathbf{X})} \tag{6}$$

where $\mathbf{C}_k \subset \mathbf{X}$ indicates a subset of features, w_k is the weight for each distribution P_k, and $Z(\mathbf{X})$ is the normalizing constant.

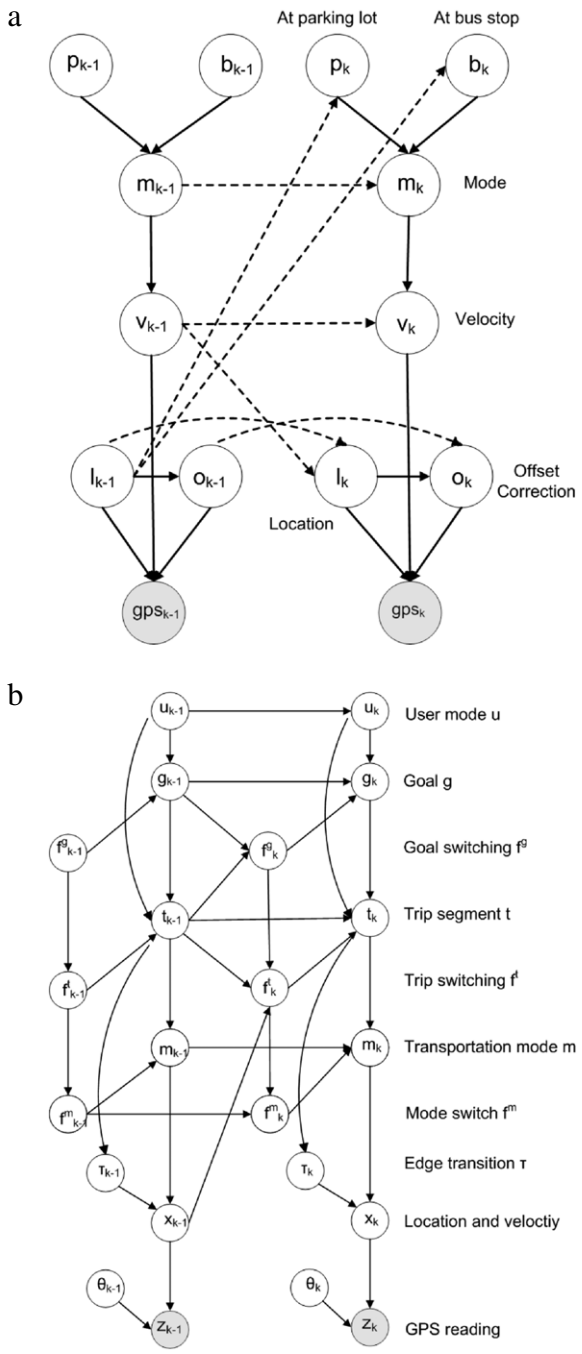

Fig. 7. Two typical dynamic Bayesian networks [12,74], including the observable nodes, hidden nodes and the dependencies between them, are presented in modeling individual's mobility. In the dynamic Bayesian network, each node represents the variable at a given time, and the arrows indicate the dependencies between different variables. In both cases, only the GPS reading are the observable nodes, and the remaining ones have to be learned from the mode. The main difference between (a) and (b) is that, the model in (b) is in a fine-grained level, where the switching between modes, trips and goals are considered. In (a), the solid lines indicate the intra-temporal causal links and the dashed lines indicate the inter-temporal links.

The weight of each distribution is learnt based on the training data collected from all the individuals, whereas the parameters for each distribution is exclusively learnt according to the individual's data. The experimental results show that the accuracy in predicting the next location by using a combination of the subset of the features in the factorized conditional model is more than 10% higher than that based on any single feature or the combination of a few, and a similar result is reported for the prediction for the duration of stay. However, one issue regarding applying the factorized conditional model is the choice of appropriate subset of features. In the original paper, the authors obtained the combination by a greedy

search process, which always chooses the features that achieve the highest predicting accuracy. However, in general, the best overall results may not be obtained by choosing the list of the features that show the best performance individually.

In the Bayes inference, the posterior probability density for the variable of concern, e.g. individual's state at current time x_k, conditioned on the data so far $z_1, \ldots, z_k$ can be recursively estimated,

$$p(x_k|z_{1:k}) \propto p(z_k|x_k) \int p(x_k|x_{k-1}) p(x_{k-1}|z_{1:k-1}) dx_{k-1} \tag{7}$$

where the state x_k may contain different factors. In the following we review a few approaches in details.

Patterson et al. [12] construct a dynamic Bayes net model to infer high level behaviors, such as modes of transport and transportation routes, from noisy GPS data. The dependencies, observable nodes and hidden nodes are presented in Fig. 7(a).

The state x_k, defined as $x_k = (e, d, v)$, indicates that the individual is on edge e, at a distance d to the start vertex of e and is traveling at a speed v on the edge. The current location is $l_k = (e, d)$, and o_k is the expected sensor error. The transportation mode m_k is chosen from $\{BUS, FOOT, CAR\}$, and the velocity v_k from each mode is generated according to the Gaussian distribution, where the parameters for the Gaussian distributions of three different modes are learned by a Gaussian mixture model. Since the mode of *FOOT* is unlikely to generate a high speed value, the velocity from this mode is generated by the Gaussian distribution with smallest mean. The mode of *CAR*, however, is still possible to generate a low speed value, therefore the speed from *CAR* is assumed to be evenly chosen from the three Gaussian distributions, which is an intuitive choice but it has not been validated.

The posterior distribution of the state x_k is represented in a particle filter method, where $S_k = \{(x_k^i, w_k^i)|i = 1, \ldots, n\}$, and the weight $w_k^i = p(z_k|x_k^i)$ is assigned according to the likelihood of the observation. The parameters, such as the edge transitions, and mode transitions, etc., for this model are learned by iteratively maximizing the posterior probability in the EM algorithm.

Liao et al. model an individual's daily movements by a three-level hierarchical activity model [74], where the dependencies and variables are shown in Fig. 7(b). In this model, the lowest level estimates the individual's location and velocity from GPS readings. The middle level represents the modes of transport and segments of a trip. The highest level represents the individual's next locations. The posterior distribution of the parameters based on the observations $z_{1:k}$ is separated into continuous and discrete parts,

$$p(x_k, m_k, f_k^t, f_k^m, \theta_k, \tau_k|z_{1:k}) = p(x_k|m_k, f_k^t, f_k^m, \theta_k, \tau_k, z_{1:k}) p(m_k, f_k^t, f_k^m, \theta_k, \tau_k|z_{1:k}) \tag{8}$$

where the continuous parts are the locations and velocity in the state x_k, the discrete parts are transportation mode m_k, edge transition τ_k, mode switching f_k^m and trip switching f_k^t.

Similar to [12], Rao–Blackwellized particle filters [75] are used for inference. In detail, the discrete states are sampled by a particle filter and the continuous states are sampled by a Kalman filter conditional on the discrete samples. The next locations, transfer locations, and transition matrices are obtained from the model given the structure and parameters estimated from the EM algorithm. Moreover, the errors are detected through two trackers. One tracker uses the learned model to predict individual's ordinary routine, and the second tracker uses a flat model accounting for general physical constraints but it is not adjusted to the individual's routine to detect something unexpected.

A three-level hierarchical model is applied in [76]. However, the inference algorithms are different from [74]. In [76], a discriminative relational Markov network (RMN) is used to identify significant locations, meanwhile a generative dynamic Bayesian network is used to learn transportation routines, infer goals and errors.

In [77], the hierarchically structured conditional random fields are used to extract an individual's activities and significant places [13], which are a simplified version compared to those in [74,76]. The lowest level is the associations between each GPS reading and the relevant patch of street map, in which three feature functions are defined to indicate their relationships. The middle level consists of the activity objects, each of which is featured by temporal information, average speed, geographic information, and the compatibility between types of activities. The top level includes place objects, which are featured by the frequency of activities occurred, and cliques that count the number of different homes and work places. A conditional random field [77] is constructed for labeling the activities and places based on the features extracted.

All these hierarchical probabilistic models are not naturally inheritable. Given additional training data, the structure of the model must be learned again, as the parameters of the old model become inapplicable with the new data.

There are several other studies [78–80] that target on recognizing individuals' daily activity patterns from the GSM-based localization, blue tooth communications [78,79], and even the activity survey data [80]. These studies apply formal approaches such as, Latent Dirichlet Allocation [81], or Hidden Markov Models [82], etc. The interested readers should consult these papers directly.

6. Issues for further research

Besides the topics discussed in the previous section, many issues concerning individuals' mobility are of great interest. This section discusses a few new and emergent issues for further research.

6.1. Further study on predictability

Although many findings of predictability have been presented on GPS data [51], major questions remain unanswered. In both [1,51], individuals' mobility records are generated at a fixed sampling rate. The fixed sampling rate incurs two problems. (1) If multiple locations are visited within a sampling period, only one location will be captured, which leads to incomplete modeling of individuals' mobility. (2) The sampling rate determines the length of sequences, the entropy it contains, etc. Thus, from a theoretical perspective, what would be the optimal rate to perfectly preserve the information about individuals' mobility without much redundancy?

Besides sampling at a fixed rate, there are other alternatives for generating mobility sequences [83], such as movement-based encoding, and a hybrid of time- and movement-based encoding. In movement-based encoding, a transition between locations is encoded when there is a change of location, while the hybrid of time- and movement-based encoding retains the benefits of the two methods. For a given sampling rate, if there is a change of locations in between two periodic samples, the transitions are encoded as the movement-based encoding method, meanwhile the transitions are encoded periodically as the time-based method. Intuitively, the movement-based encoding method may overestimate individuals' mobility, since the entropy contained in the string may be relatively large. The time-based encoding method may miss some mobility information, such as the transitions between locations within a sampling period. Therefore, a mixture of time- and movement-based encoding method may be a good choice, provided an appropriate sampling rate is chosen.

Also, the predictability Π of individual's mobility should be a function of both the spatial scale λ and the temporal scale τ, respectively representing the size of each location and the sampling rate in capturing the mobility string,

$$\Pi = f(\lambda, \tau). \tag{9}$$

The issue of determining how the predictability varies with respect to both λ and τ is clearly fundamental to the mobility research and it deserves a thorough study.

6.2. Modeling the process of mobility

Recall that when estimating the entropy from a location string, the Lempel–Ziv data compression algorithm is applied under the assumption that the process of human mobility is a stationary and ergodic process [1]. It is reasonable to assume that the human mobility process has these general statistical properties. However, it is unclear whether individuals' mobility has the properties of a Markov process. The empirical study based on real human trajectory is necessary to validate the natural behavior of human mobility, such as the dependency between locations, the probability of transition between locations, etc.

6.3. Mobility data compression and prediction

Feder and Merhav [84] show that each lossless data compression algorithm may be transformed to an effective algorithm for predicting next symbols. Although the Lempel–Ziv algorithm is applied to estimate the entropy rate in the location string in [1], no corresponding predicting algorithm is further mentioned or applied. The existing study shows that predictors based on Lempel–Ziv data compression algorithm are effective in different domains: text, music and proteins [85]. However, it is unclear whether the Lempel–Ziv algorithm is an appropriate choice for predicting individuals' next move. Besides, some alternatives have been proposed, such as prediction by partial match [86], context tree weighting method [87], probabilistic suffix tree [88], etc. The fundamental theory underlying these predictors is similar, which uses Markov model to model the dependency in the sequence and the order of the Markov model is gradually increased along with the length of the input sequence. The main differences among the approaches lie in the methods for counting frequency, estimating the probability of unobserved events, and determining the variable length etc. Thus, one relevant question is: which data compression algorithms is more suitable for modeling individuals' mobility sequences? How do we tailor the algorithms based on the special characteristics of individual's mobility behaviors?

6.4. Constructing mobility models

Among various mobility models, the trace-based mobility models are the most realistic models since the objective of these models is to fit the statistical properties observed from real trajectories. However, most trace-based mobility models [9–11] are constructed based on the trajectories under the walk mode. In reality, however, each individual may apply a combination of modes of transport when traveling around, which may have different statistical properties from the trajectory under the walk mode. The mobility models should exhibit both regularity and randomness. The high regularities as recently reported in [1,2] include: high frequency of visiting a few locations, stable transitions among the locations, etc. Meanwhile, human mobility is seldom deterministic. For instance, the path followed by an individual from one location to another may be slightly different each time, and the locations that are seldom visited should also be considered. Thus, random models such as Brownian motion, Lévy flight, etc., may be appropriate models for modeling the randomness in human mobility.

7. Conclusions

We have reviewed many recent results on mining mobility patterns from GPS data, and have covered issues such as locations of significance, modes of transport, trajectory patterns and location-based activities. We also have suggested a few outstanding issues for further study. Although this survey aims to include as many topics concerning individuals' mobility as possible, some issues are necessarily left out; the mobility studies based on the other types of data, such as GSM cell-tower data and wireless network logs, are also not included. Nevertheless, we hope that this survey may serve as a quick guide for those who are interested in making contributions to the emerging area of individuals' mobility.

References

[1] C. Song, Z. Qu, N. Blumm, A.-L. Barabái, Limits of predictability in human mobility, Science 327 (2010) 1018–1021.
[2] M.C. González, C.A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453 (2008) 779–782.
[3] C. Song, T. Koren, P. Wang, A.-L. Barabái, Modelling the scaling properties of human mobility, Nature Physics 6 (2010) 818–823.
[4] N. Eagle, A.S. Pentland, Reality mining: sensing complex social systems, Personal and Ubiquitous Computing 10 (2006) 255–268.
[5] H. Fang, W.-J. Hsu, L. Rudolph, Cognitive personal positioning based on activity map and adaptive particle filter, in: MSWiM, 2009, pp. 405–412.
[6] H. Fang, W.-J. Hsu, L. Rudolph, Mining user position log for construction of personalized activity map, in: ADMA, 2009, pp. 444–452.
[7] P. Wang, M.C. González, C.A. Hidalgo, A.-L. Barabái, Understanding the spreading patterns of mobile phone viruses, Science 324 (5930) (2009) 1071–1076.
[8] M.A. Bayir, M. Demirbas, N. Eagle, Mobility profiler: a framework for discovering mobility profiles of cell phone users, Pervasive and Mobile Computing 6 (4) (2010) 435–454.
[9] I. Rhee, M. Shin, S. Hong, K. Lee, S. Chong, On the Levy-walk nature of human mobility in: The 27th Conf. on Computer Communications. IEEE, 2008, pp. 924–932.
[10] K. Lee, S. Hong, S.J. Kim, I. Rhee, S. Chong, Slaw: a mobility model for human walks, in: The 28th Conf. on Computer Communications. IEEE, 2009, pp. 855–863.
[11] S. Hong, K. Lee, I. Rhee, Step: a spatio-temporal mobility model for humans walks, in: MASS, 2010, pp. 630–635.
[12] D.J. Patterson, L. Liao, D. Fox, H. Kautz, Inferring high-level behavior from low-level sensors, in: UbiComp, 2003, pp. 73–89.
[13] L. Liao, D. Fox, H. Kautz, Extracting places and activities from GPS traces using hierarchical conditional random fields, International Journal of Robotics Research 26 (2007) 119–134.
[14] H. Cao, N. Mamoulis, D.W. Cheung, Mining frequent spatio-temporal sequential patterns, in: ICDM, 2005, pp. 82–89.
[15] L. Liao, D. Fox, H. Kautz, Location-based activity recognition using relational Markov networks, in: Proc. of the 19th Int. Joint Conf. on Artificial Intelligence, 2005, pp. 773–778.
[16] D. Ashbrook, T. Starner, Using GPS to learn significant locations and predict movement across multiple users, Personal and Ubiquitous Computing 7 (2003) 275–286.
[17] T.M.T. Do, D. Gatica-Perez, Contextual conditional models for smartphone-based human mobility prediction, in: UbiComp, 2012, pp. 163–172.
[18] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W.-Y. Ma, Mining user similarity based on location history, in: GIS, 2008, pp. 1–10.
[19] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, in: WWW, 2009, pp. 791–800.
[20] G. Krings, F. Calabrese, C. Ratti, V.D. Blondel, Urban gravity: a model for inter-city telecommunication flows, Journal of Statistical Mechanics: Theory and Experiment 7 (2009) L07003.
[21] H.A. Makse, S. Havlin, H. Stanley, Modelling urban growth patterns, Nature 377 (1995) 608–612.
[22] M.W. Horner, M.E. O'Kelly, Embedding economies of scale concepts for hub network design, Journal of Transport Geography 9 (4) (2001) 255–265.
[23] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A. Vespignani, Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions, PLoS Medicine 4 (2007) e13.
[24] J. Kleinberg, The wireless epidemic, Nature 449 (2007) 287–288.
[25] L. Hufnagel, D. Brockmann, T. Geisel, Forecast and control of epidemics in a globalized world, Proceedings of the National Academy of Sciences of the United States of America 101 (42) (2004) 15124–15129.
[26] Transportation & logistics 2030, Tech. Rep., PricewaterhouseCoopers.
[27] A. Schäfer, Global mobility, energy use and climate constraints, in: International Energy Workshop, 2011.
[28] A. Schäfer, H.D. Jacoby, J.B. Henwood, I.A. Waitz, The other climate threat: transportation, American Scientist 97 (6) (2009) 476.
[29] MIT media lab, 2012. http://realitycommons.media.mit.edu/index.html.
[30] D. Kotz, T. Henderson, Crawdad: a community resource for archiving wireless data at dartmouth, IEEE Pervasive Computing 4 (2005) 12–14.
[31] M. McNett, G.M. Voelker, Access and mobility of wireless pda users, Mobile Computing and Communication Reviews 9 (2005) 40–55.
[32] T. Henderson, D. Kotz, I. Abyzov, The changing usage of a mature campus-wide wireless network, in: MobiCom, 2004, pp. 187–201.
[33] Y. Zheng, Geolife, 2010. http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx.
[34] I. Rhee, M. Shin, S. Hong, K. Lee, S. Kim, S. Chong, CRAWDAD data set ncsu/mobilitymodels, v. 2009-07-23, 2009. http://crawdad.cs.dartmouth.edu/ncsu/mobilitymodels.
[35] F. Giannotti, A. Mazzoni, S. Puntoni, C. Renso, Synthetic generation of cellular network positioning data, in: GIS, 2005, pp. 12–20.
[36] M. Nanni, D. Pedreschi, Time-focused clustering of trajectories of moving objects, Journal of Intelligent Information Systems 27 (3) (2006) 267–289.
[37] F. Giannotti, D. Pedreschi, Mobility, Data Mining and Privacy, Springer, 2008.
[38] Y. Zheng, X. Zhou, Computing with Spatial Trajectories, Springer, 2011.
[39] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti, Unveiling the complexity of human mobility by querying and mining massive trajectory data, The VLDB Journal 20 (5) (2011) 695–719.
[40] N. Aschenbruck, A. Munjal, T. Camp, Trace-based mobility modeling for multi-hop wireless networks, Computer Communications 34 (6) (2011) 704–714.
[41] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, S. Wrobel, From movement tracks through events to places: extracting and characterizing significant places from mobility data, in: IEEE VAST, 2011, pp. 161–170.
[42] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, L. Terveen, Discovering personal gazetteers: an interactive clustering approach, in: GIS, 2004, pp. 266–273.
[43] X. Cao, G. Cong, C.S. Jensen, Mining significant semantic locations from GPS data, Proceedings of the VLDB Endowment 3 (2010) 1009–1020.
[44] L. Chen, M. Lv, G. Chen, A system for destination and future route prediction based on trajectory mining, Pervasive and Mobile Computing 6 (2010) 657–676.
[45] M. Ester, H.Peter Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: KDD, 1996, pp. 226–231.
[46] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, in: SIGMOD, 1999, pp. 49–60.
[47] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.
[48] R. Hariharan, K. Toyama, Project lachesis: parsing and modeling location histories, in: GIS, 2004, pp. 106–124.
[49] J. Krumm, E. Horvitz, Predestination: inferring destinations from partial trajectories, in: Ubicomp, 2006, pp. 243–260.
[50] B.S. Jensen, J.E. Larsen, K. Jensen, J. Larsen, L.K. Hansen, Estimating human predictability from mobile sensor data, in: MLSP, 2010, pp. 196–201.

[51] M. Lin, W.-J. Hsu, Z.Q. Lee, Predictability of individuals' mobility with high-resolution positioning data, in: Ubicomp, 2012, pp. 381–390.
[52] J. Ziv, A. Lempel, Compression of individual sequences via variable-rate coding, IEEE Transactions on Information Theory 24 (5) (1978) 530–536.
[53] Y. Zheng, L. Liu, L.-H. Wang, X. Xie, Learning transportation mode from raw GPS data for geographic applications on the web, in: WWW, 2008, pp. 247–256.
[54] Y. Zheng, Y. Chen, Q. Li, X. Xie, W.-Y. Ma, Understanding transportation modes based on GPS data for web applications, ACM Transactions on the Web 4 (1) (2010) 1–36.
[55] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on GPS data, in: UbiComp, 2008, pp. 312–321.
[56] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, M. Srivastava, Using mobile phones to determine transportation modes, ACM Transactions on Sensor Networks 6 (2) (2010) 1–27.
[57] M. Lin, W.-J. Hsu, Z.Q. Lee, Detecting modes of transport from unlabeled positioning sensor data, Journal of Location Based Services (2013).
[58] J.-G. Lee, J. Han, Trajectory clustering: a partition-and-group framework, in: SIGMOD, 2007, pp. 593–604.
[59] J.-G. Lee, J. Han, X. Li, H. Gonzalez, Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering, Proceedings of the VLDB Endowment 1 (1) (2008) 1081–1094.
[60] N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos, Y. Theodoridis, Clustering trajectories of moving objects in an uncertain world, in: ICDM, 2009, pp. 417–427.
[61] J.-G. Lee, J. Han, X. Li, H. Cheng, Mining discriminative patterns for classifying trajectories on road networks, IEEE Transactions on Knowledge and Data Engineering 23 (5) (2011) 713–726.
[62] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, in: KDD, 2007, pp. 330–339.
[63] H. Cao, N. Mamoulis, D.W. Cheung, Discovery of periodic patterns in spatiotemporal sequences, IEEE Transactions on Knowledge and Data Engineering 19 (4) (2007) 453–467.
[64] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, G. Andrienko, Visually driven analysis of movement data by progressive clustering, Information Visualization 7 (3) (2008) 225–239.
[65] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, F. Giannotti, Interactive visual clustering of large collections of trajectories, in: IEEE VAST, 2009, pp. 3–10.
[66] P.D. Grnwald, I.J. Myung, M.A. Pitt, Advances in Minimum Description Length: Theory and Applications, The MIT Press, 2005.
[67] F. Aurenhammer, T.U. Graz, R. Klein, F. Hagen, P.I. Vi, Voronoi diagrams, in: Handbook of Computational Geometry, 2000, pp. 201–290.
[68] C.S. Jensen, D. Lin, B. Chin, O.R. Zhang, Effective density queries on continuously moving objects, in: ICDE, 2006, p. 71.
[69] D. Huttenlocher, G. Klanderman, W. Rucklidge, Comparing images using the hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9) (1993) 850–863.
[70] A. Monreale, F. Pinelli, R. Trasarti, F. Giannotti, Wherenext: a location predictor on trajectory pattern mining, in: KDD, 2009, pp. 637–646.
[71] M. Lin, Z.Q. Lee, W.-J. Hsu, Uncovering temporal and spatial localities in individuals' mobility, in: MDM, 2013.
[72] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: VLDB, 1994, pp. 487–499.
[73] J. Han, G. Dong, Y. Yin, Efficient mining of partial periodic patterns in time series database, in: ICDE, 1999, pp. 106–115.
[74] L. Liao, D.J. Patterson, D. Fox, H. Kautz, Learning and inferring transportation routines, Artificial Intelligence 171 (2007) 311–331.
[75] A. Doucet, N. de Freitas, K.P. Murphy, S.J. Russell, Rao-blackwellised particle filtering for dynamic Bayesian networks, in: UAI, 2000, pp. 176–183.
[76] L. Liao, D.J. Patterson, D. Fox, H. Kautz, Building personal maps from GPS data, Annals of the New York Academy of Sciences (2006) 249–265.
[77] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: ICML, 2001, pp. 282–289.
[78] N. Eagle, A.S. Pentland, Reality mining: sensing complex social systems, Personal and Ubiquitous Computing 10 (4) (2006) 255–268.
[79] L. Ferrari, M. Mamei, Classification and prediction of whereabouts patterns from the reality mining dataset, Pervasive and Mobile Computing 9 (4) (2012) 516–527.
[80] S. Jiang, J. Ferreira, M.C. González, Clustering daily patterns of human activities in the city, Data Mining and Knowledge Discovery 25 (2012) 478–510.
[81] D.M. Blei, A.Y. Ng, M.I. Jordan, J. Lafferty, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[82] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.
[83] A. Bhattacharya, S.K. Das, Lezi-update: an information-theoretic approach to track mobile users in pcs networks, in: MobiCom, 1999, pp. 1–12.
[84] M. Feder, N. Merhav, Relations between entropy and error probability, IEEE Transactions on Information Theory 40 (1) (1994) 259–266.
[85] R. Begleiter, R. El-Yaniv, G. Yona, On prediction using variable order Markov models, Journal of Artificial Intelligence Research 22 (2004) 385–421.
[86] J. Cleary, I. Witten, Data compression using adaptive coding and partial string matching, IEEE Transactions on Communications 32 (4) (1984) 396–402.
[87] F. Willems, Y. Shtarkov, T. Tjalkens, The context-tree weighting method: basic properties, IEEE Transactions on Information Theory 41 (3) (1995) 653–664.
[88] D. Ron, Y. Singer, N. Tishby, The power of amnesia: learning probabilistic automata with variable memory length, Machine Learning (1996) 117–149.