Research Article

A Spatial-Contextual Indoor Trajectory Prediction Approach via Hidden Markov Models

Peng Wang , Jing Yang , and Jianpei Zhang

College of Computer Science and Technology, Harbin Engineering University, Harbin 150000, China

Correspondence should be addressed to Jing Yang; yangjing@hrbeu.edu.cn

Received 29 June 2021; Revised 17 November 2021; Accepted 23 November 2021; Published 11 January 2022

Academic Editor: Chakchai So-In

Copyright © 2022 Peng Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Unlike outdoor trajectory prediction that has been studied many years, predicting the movement of a large number of users in indoor space like shopping mall has just been a hot and challenging issue due to the ubiquitous emerging of mobile devices and free Wi-Fi services in shopping centers in recent years. Aimed at solving the indoor trajectory prediction problem, in this paper, a hybrid method based on Hidden Markov approach is proposed. The proposed approach clusters Wi-Fi access points according to their similarities first; then, a frequent subtrajectory based HMM which captures the moving patterns of users has been investigated. In addition, we assume that a customer's visiting history has certain patterns; thus, we integrate trajectory prediction with shop category prediction into a unified framework which further improves the predicting ability. Comprehensive performance evaluation using a large-scale real dataset collected between September 2012 and October 2013 from over 120,000 anonymized, opt-in consumers in a large shopping center in Sydney was conducted; the experimental results show that the proposed method outperforms the traditional HMM and perform well enough to be usable in practice.

1. Introduction

During the past decade, there have been a large amount of researches focusing on trajectory prediction [1, 2]. While most of the researches to date concentrated on outdoor scenario with GPS or GPS-like positioning only, researches show that human beings spend around 87% of times in indoor environments such as shopping malls, office buildings, airports, conference facilities, and private homes [3, 4]; thus, the research of trajectory prediction falls short in another important setting, namely, indoor scenario. The main reason is that there are no reliable and accurate indoor positioning systems and the outdoor positioning technologies such as GPS cannot identify the indoor location of a user accurately either.

With the prevalent of mobile devices which support Wi-Fi-enabled connectivity and the increasing number of indoor Wi-Fi-enabled venues, breakthrough in indoor trajectory prediction has been made in recent years. Shopping malls are providing free Wi-Fi connections to attract and retain users. For example, the giant supermarket chain Tesco has carried out free Wi-Fi services in hundreds of its stores in the UK [5]. Wi-Fi is almost a must have for shopping malls, and because of the

ubiquitous Wi-Fi services, it is becoming easier to track shoppers' foot-path and physical movements by capturing the Wi-Fi signals emitted by mobile devices and collecting the MAC address while the shoppers move around in the shopping mall.

In this paper, we aim to propose an effective method to predict a customer's next visiting place in the indoor space based on his history locations where each location is represented by a specific Wi-Fi access point. There are many potential ways that shoppers and retailers can benefit from the application of indoor trajectory prediction. For example, by being aware of the movement of customers in advance, vendors can quickly target possible shoppers and push advertisements through online advertisement system with the objective to attract shoppers' attention and make it possible to boost the sales even before customers physically approach the store. Moreover, through the application, the pedestrian flow in the shopping mall can be predicted making it possible to avoid the traffic block and maximize the effect of shopping coupon promotion which delivers the coupons in a specific region at a certain time. Last but not least, through the indoor prediction system, managers can rearrange the workers in the indoor space more efficiently.

In all, the contributions of the paper are as follows:

- (1) We investigate which are the appropriate hidden states of Hidden Markov model in indoor trajectory prediction problem
- (2) A novel clustering algorithm based on the similarity of Wi-Fi access points has been studied
- (3) From spatial perspective, we employ frequent sub-trajectory into indoor trajectory prediction problem which improves the prediction ability of our approach
- (4) From contextual perspective, we investigate and integrate the unique contextual information of shop category in indoor scenario into our trajectory prediction problem which improves the prediction ability
- (5) Extensive experiments have been carried out on a real-world dataset to evaluate our method. The results show notable improvement of indoor trajectory prediction in accuracy by adopting our approach

The reminder of this paper is organized as follows. We first present literature review on the research area in Section 2 before describing the preliminaries about indoor trajectory and related issues of Hidden Markov mode in Section 3. We present the data processing technique in Section 4; then, we build the HMM-associated indoor trajectory prediction models and present the fusion framework which integrates trajectory prediction and shop category prediction in Section 5. In Section 6 we evaluate the proposed method and analyze experiment results. Finally, we conclude the paper in Section 7.

2. Related Work

There have been extensive researches in localization and mobility prediction utilizing wireless network and mobile devices [6–8]. However, unlike the outdoor positioning system (GPS), indoor positioning systems such as Bluetooth, Infrared, RFID, or Wi-Fi have only been mature in recent years and started to emerge in commercial markets [9]. While much more researches focusing on trajectory prediction in the context of outdoor positioning systems have been published [10], there is still not sufficient related work on the same application in indoor spaces.

Based on the data collected, there are two types of trajectory predictions: continuous where the positioning system generates continuous coordinate values and discrete where the positioning system provides discrete coordinate values such as the IDs of positions [11]. The accuracy of continuous prediction can be evaluated by the distance between the predicted result and the corresponding actual position while the accuracy of discrete prediction which returns the predicted IDs of positions can be evaluated as the probabilities that the results are actual IDs. Our work falls into the second category where each indoor trajectory point is represented by a Wi-Fi access point with a unique ID.

The methods proposed for trajectory prediction can be divided into two main categories: (1) mine frequent patterns

of trajectories and (2) establish the model of moving objects to achieve the prediction.

In order to extract the trajectory patterns for prediction, a number of research works have been conducted [12–14]. Apart from the work based on geographical characteristics of trajectories, semantic trajectories have also been mined and used for the prediction [15]. While the aforementioned work adopts the trajectory patterns for prediction directly, our work is distinguished from them in that we treat the patterns as hidden states that can not be observed.

Though there are many pattern mining-based prediction methods, they do not have a mechanism to allow continuous updating of movement patterns. So various machine learning techniques have been developed to predict users' location given their historical movements, such as dynamic Bayesian networks [16, 17]. However, during our background research, the most frequently used methods are Markov Chain and Hidden Markov model and their variations.

In [18], the authors proposed the Markov transition probability which is based on a cell-based organization of a target space and the Markov Chain model. In the recent work [2], the authors present a model based on Markov chain for predicting the next location of a student in campus, they consider the notion of time in the prediction algorithm that coined as trajectory prediction algorithm (TPA).

When considering the variations of the Markov model, the authors proposed a mixed Markov Chain model (MMM) which is an intermediate model between an individual model and a universal one where it clusters individuals into a set of latent groups based on their mobility behaviors and transition traces, each group has its specific Markov Chain model that needs to be learned. The results show that the prediction accuracy is higher than that of Markov-chain model and HMMs [11]. However, it will consume much time and storage space for learning and storing the generated models. In order to solve the problem, the authors extended a mobility model called Mobility Markov Chain (MMC) to incorporate the n previous visited locations [19].

When referring to HMM-based trajectory prediction, it can be classified into two categories: parameter learning and structure learning [20]. For parameter learning, [21] developed the mixed hidden Markov models (MHMM) and [22] proposed a HMM-based trajectory prediction algorithm where the main parameters can be adapted autonomously. For structure learning, [23] introduces an algorithm for modeling and recognizing temporal structure of visual activities based on HMM; in [24], the authors presented models for pedestrian behavior analysis by building geometric and probabilistic models sequentially.

Existing works aiming at predicting indoor movements employ the method of pattern mining and variance of Markov Chain [1, 10, 17, 19, 25, 26]; however, the unique feature of semantic information in indoor environment such as shop categories has not been studied for prediction due to the limitation of their data source. What is more, their methods are only tested on small data sets, which only contain a small number of users or trajectory points, and in [19], the datasets used were collected in a controlled environment: data was gathered from participants who are aware of the

TABLE 1: Notations used in the paper.

Symbol	Description
AP_i	Wi-Fi access point in the shopping mall
$\text{sim}(AP_i, AP_j)$	Similarity between access points according to their shop categories
$P_i = \{AP_i, t\}$	Trajectory point is a combination of AP_i and association time t
δ	Similarity threshold between access points
$T_i = \{P_1, P_2, \dots, P_n\}$	A sequence of time-ordered indoor trajectory points
$D = \{T_1, T_2, \dots, T_n\}$	Indoor trajectory database
$H = \{H_1, H_2, \dots, H_n\}$	Hidden states in the HMM
$O = \{O_1, O_2, \dots, O_n\}$	Observation states in the HMM
$\Pi = \{\pi_i\}$	Initial probability of choosing state i
$A = \{h_{ij}\}$	Probability of transforming from hidden state i to hidden state j
$B = \{b_{ik}\}$	Probability of transforming from hidden state i to observation state k
$\lambda = \{H, O, \Pi, A, B\}$	Indoor Hidden Markov model
$\Omega = \{H_c, O_c, \Pi_c, A_c, B_c\}$	The Hidden Markov model regarding shop category
$C_i = \{AP_i, AP_j, \dots, AP_k\}$	A group of trajectory points
Δ	Threshold of time duration between two consecutive trajectory points

experiments. Different from the existing works, in this article, we used a large-scale dataset collected between September 2012 and October 2013 from over 120,000 anonymized users (907,084 associations are detected forming 261,369 indoor trajectories) which is operated by a large mall with 67 Wi-Fi access points (APs) across 90,000 square meters. Additional information about the physical environment is provided by the owner of the mall including floor plans of the stores, the shop categories, and the location of Wi-Fi access point. Such data provides a unique opportunity for us to analyze the interaction between users' physical movement and semantic movement in indoor spaces and our method takes the frequent subtrajectory and shop categories the customer visits into consideration which have not been investigated before.

3. Problem Statement and Preliminaries

In order to better depict the method proposed, we first describe indoor trajectory and its related properties. Table 1 lists all the notations used in the paper; specific meanings have been given when they emerge in the paper.

3.1. Basic Concepts. In this section, we will introduce some basic concepts first.

Definition 1. An AP denotes a Wi-Fi access point in the shopping mall, each AP has a unique id and it covers several shops, and each shop belongs to a specified shop category. We denote an AP as $AP_i = \{i, \text{sub}_{\text{category}_i}\}$, where $\text{sub}_{\text{category}_i}$ is a subset of all the shop categories.

As the example illustrated in Table 2, Wi-Fi access point AP_{13} covers six shops and their shop categories belong to

TABLE 2: Example of Wi-Fi access point.

AP	Shop name	Shop category
AP_{13}	Caffee Contessa	Cafe
	Ignazia	Women's Fashion
	Maggie	Women's Fashion
	Oxford	Unisex Fashion
	Peep Toe	Women's Footwear
	Peter Sheppard	General Footwear (MM)

Women's Fashion, Unisex Fashion, Women's Footwear, and General Footwear; thus, we denote $AP_{13} = \{13, [\text{Cafe}, \text{Women's Fashion}, \text{Unisex Fashion}, \text{Women's Footwear}, \text{General Footwear}]\}$. where $\|AP_i\|$ denotes the number of AP_i 's shop categories.

Definition 2. The similarity between two APs is defined as follows:

$$\text{sim}(AP_i, AP_j) = \frac{\|AP_i\| \cap \|AP_j\|}{\|AP_i\| \cup \|AP_j\|}, \quad (1)$$

Definition 3. An indoor trajectory T is composed of a sequence of time ordered APs. Where $T = \{P_1, P_2, \dots, P_k\}$ and $P_i = \{(AP_m, t_i)\}$, $i \in [1, k]$, P_i represents the indoor trajectory point AP_m when associating at timestamp t_i in the indoor space.

Definition 4. Indoor trajectory base D stores a large number of customers' logging information in the indoor space, $D =$

$\{T_1, T_2, \dots, T_n\}$ represents the trajectory set, and $\|D\| = n$ is the total number of trajectories.

3.2. Hidden Markov Models. In this section, we briefly introduce the basic concept of *Hidden Markov model* [27], *indoor trajectory HMM*, and *shop category HMM*.

The Hidden Markov model is a statistical model used to describe the Markov process with unknown parameters. It is a two-tiered random process in which the upper layer is composed of Markov Chain that describes the transition between hidden states, and the bottom layer is a random model that depicts the relationship between observation symbols (observation states) and hidden states. Based on HMM, our problem can be formalized as follows.

Definition 5. In indoor trajectory prediction based on HMM, assume an indoor trajectory is depicted as $T = \{P_1, P_2, \dots, P_n\}$, where n is a distinct timestamp, and $\lambda = \{H, O, \Pi, A, B\}$ is the HMM-based indoor trajectory prediction model, where H stands for hidden states and O corresponds for observation states in the Hidden Markov model, Π is the initial state distribution matrix, A is the hidden states transition matrix, and B is the transition matrix; the goal is to compute the position of $AP_{(n+1)}$ in the $(n+1)^{th}$ timestamp based on the previous n^{th} timestamps.

Similar to the classic Hidden Markov model, our indoor trajectory HMM is depicted as follows:

Definition 6. The indoor trajectory HMM is a five-tuple $\lambda = \{H, O, \Pi, A, B\}$; it then follows:

H is a set of indoor trajectory hidden states. The hidden states are denoted by $H = \{H_1, H_2, \dots, H_n\}$, which meet the Markov Property, where n indicates the number of hidden states.

O is a set of observation symbols; in our case, it denotes a set of Wi-Fi access points which is expressed by $O = \{O_1, O_2, \dots, O_m\}$, where m corresponds to the total number of observable states.

$\Pi = \{\pi_i\} (1 \leq i \leq n)$ denotes the initial probability of choosing state i ($\sum \pi_i = 1$). For example, when $t = 1$, the possibilities of choosing hidden states H_1 , H_2 , and H_3 are $P_1(H_1) = \pi_1$, $P_1(H_2) = \pi_2$, and $P_1(H_3) = \pi_3$, respectively; then, initial state probability matrix is $\Pi = \{\pi_1, \pi_2, \pi_3\}$.

$A = \{h_{ij}\}$, is the indoor trajectory hidden state transition probability matrix, where $h_{ij} = P\{H(t+1) = H_j | H(t) = H_i\}$, $1 \leq i, j \leq n$, indicates that at timestamp $t+1$, the probability of choosing state H_j is h_{ij} , in the condition of state H_i at timestamp t .

$B = \{b_{ik}\}$ is the confusion matrix (transition probability matrix); it describes the transition probability between the hidden states and observation states in the HMM, where $b_{ik} = P\{O(t) = o_k | H(t) = H_i\} (1 \leq i \leq n, 1 \leq k \leq m)$ represents the probability of transformation from hidden state H_i to observation state o_k .

After defining the parameters in the indoor trajectory HMM, we will introduce how to compute the three basic

parameters, namely, initial probability matrix Π , transition probability matrix A , and confusion matrix B , respectively. Overall speaking, we calculate the parameters of HMM by a statistical method.

The calculation of initial probability matrix Π is as follows:

$$P(H_i) = \frac{\text{Num}(H_i)}{\sum_{j=1}^N (\text{Num}(H_j))}, \quad (2)$$

where N denotes the number of hidden states and $\text{Num}(H_i)$ denotes the frequency of hidden state H_i appearing in all hidden state sequences corresponding to observation states.

Then, the computation of state transition probability matrix A is as follows:

$$h_{ij} = P\{H(t+1) = H_j | H(t) = H_i\} = \frac{\text{Num}(H_i, H_j)}{\text{Num}(H_i)}, \quad (3)$$

where $\text{Num}(H_i, H_j)$ denotes the count of hidden state H_j appearing after hidden state H_i in the training set and $\text{Num}(H_i)$ represents the count of hidden state H_i emerging in the training set.

Finally and similar to the calculation of probability matrix A , the calculation of confusion matrix B is as follows:

$$b_{ik} = P\{O(t) = o_k | H(t) = H_i\} = \frac{\text{Num}(H_i, O_k)}{\text{Num}(H_i)}, \quad (4)$$

where $\text{Num}(H_i, O_k)$ denotes the count of observed state O_k occurring along with the hidden state H_i and $\text{Num}(H_i)$ represents the count of hidden state H_i occurring in the training set.

In shop category HMM, in real world scenarios, when a customer goes into the shopping mall, it is highly possible that his visiting sequences have certain patterns; as the example shown in Figure 1, after buying clothes, he would like to eat something; after that, he may go to the cinema to have a rest; thus, the customers' behaviors are highly predictable. In order to validate to what extent the shop categories a customer visits can be accurately predicted, we design a shop category-based Hidden Markov model to test our assumptions.

Similar to the prediction of indoor trajectory, the prediction of shop category based on HMM is depicted as follows:

Definition 7. In shop category prediction based on HMM, assume that a customer has visited n shops in a shopping mall that belong to m categories $\{\text{category}_1, \text{category}_2, \dots, \text{category}_m\}$; our goal is to predict what is the next shop category the customer is going to visit based on the Hidden Markov model.

FIGURE 1: An example of a user's movement in a shopping mall.

Since the details of shop category HMM are similar to that of indoor trajectory HMM, we omit the detailed descriptions here.

4. Indoor Trajectory Data Preprocessing

Before constructing the HMM, we have to preprocess the data which includes grouping access points and extracting hidden states.

4.1. Group Analysis for APs. In order to obtain the hidden states, we first propose the similarity-based grouping method which groups the access points according to their similarities between each other. Similarity parameter δ plays a critical role in the algorithm which differs fundamentally from traditional clustering methods because it does not need to iteratively visit each Wi-Fi access point. Algorithm 1 depicts the procedures of the grouping method. The basic idea of our grouping algorithm is given as follows: (1) initialize and specify the ID of groups (line 1); (2) visit the access points in TLink, and check whether a point has been visited; if so, omit it; else, specify groupNum to its group ID and mark it as “visited” (lines 2-8); and (3) traverse the j^{th} access point in Tlink, where $j = i + 1$; if the similarity between i^{th} and j^{th} points is no less than δ , then we assign the group ID of the j^{th} AP to groupNum and mark it as “visited” (lines 9-17).

4.2. Extracting Subtrajectory Patterns. In this section, we describe how to extract the frequent subtrajectory patterns from the training data set.

The subtrajectory in this study means consecutive trajectory points; the length of subtrajectory ranges from 2 to n , where n is the number of trajectory points in the given trajectory. As the example shown in Table 3, given a trajectory sequence $A \rightarrow B \rightarrow C \rightarrow D \rightarrow B \rightarrow C$, the set of subtrajectory with length two is $\{(A \rightarrow B), (B \rightarrow C), (C \rightarrow D), (D \rightarrow B), (B \rightarrow C)\}$; then, we sort the subtrajectory patterns in a descending order and choose the top- k patterns as hidden states.

After computing the top frequent subtrajectory patterns, an additional procedure is to preprocess the raw trajectories into frequent subtrajectory based trajectories. First, we select

input: A linklist *Tlink* storing the access points and similarity threshold δ
output: A set of groups $C = \{C_1, C_2, \dots, C_n\}$ satisfying the condition with respect to δ

```

1: groupNum = 1;
2: for i = 0 ; i < Tlink.count ; i ++ do
3:   if Tlink[i].visit == TRUE then
4:     CONTINUE;
5:   else
6:     TLink[i].id = groupNum;
7:     TLink[i].visit = TRUE;
8:     Create a new group  $C_{groupNum}$ 
9:      $C_{groupNum}.append(TLink[i]);$ 
10:  end if
11:  for j = i + 1 to Tlink.count do
12:    if Tlink[j].visit == TRUE then
13:      CONTINUE;
14:    end if
15:    if  $sim(Tlink[i], Tlink[j]) \geq \delta$  then
16:      Tlink[j].id = groupNum;
17:      Tlink[j].visit = TRUE;
18:       $C_{groupNum}.append(Tlink[j]);$ 
19:    end if
20:  end for
21:  groupNum ++;
22: end for
23: Return  $C = \{C_1, C_2, \dots, C_n\}$ 

```

ALGORITHM 1: Wi-Fi access point grouping algorithm.

TABLE 3: Example of extracting frequent patterns.

Pattern	Frequency
(B→C)	2
(A→B)	1
(C→D)	1
(D→B)	1

frequent subtrajectory patterns according to the number of them; after that, we merge the trajectory points into a super-node. The procedure is illustrated in Figure 2.

FIGURE 2: Converting raw trajectories into subtrajectory-based trajectories.

5. Indoor Trajectory Prediction via HMM

In this section, we start by giving the working mechanism of the proposed method; then, we will present the details of each step.

5.1. Working Mechanism. The working mechanism of our proposed method is illustrated in Figure 3. At its most basic level, our solution contains two essential phases, namely, (1) model training: which retrieves hidden states corresponding to the historical trajectories and build the indoor trajectory Hidden Markov model, and (2) prediction: given a sequence of observation trajectory points, solve the decoding problem in HMM, and calculate the possibilities of candidate access points to be visited by the user next. Hidden states contain much information that can not be observed directly, but the information is highly valuable. We are interested in finding the most possible hidden state sequences $H(t) = \{H_i | i \in [1, 2, 3, \dots, n]\}$ with t timestamps corresponding to the given trajectory sequence $O(t) = \{o_i | i \in [1, 2, 3, \dots, m]\}$. It can be answered by solving the decoding problem in indoor trajectory HMM.

Definition 8. In decoding problem, it is given that

$$P(O | \lambda) = \text{Max}_{H_1, H_2, \dots, H_n} \{P(O | H_i, \lambda) * P(H_i | \lambda)\}, \quad (5)$$

where O represent a chain of trajectory points and $H_i, i \in [1, 2, \dots, N]$ is the corresponding hidden state sequence of O . We can solve this problem using the Viterbi algorithm by calculating the maximum probabilities and storing the hidden states that result in the max.

5.2. Viterbi Algorithm. In this section we will briefly introduce the Viterbi algorithm. First, we will introduce four symbols used here:

- (1) $\delta_t(i)$ represents the maximum probability that produces $O(t) = \{o_i | i \in [1, 2, 3, \dots, m]\}$ and along the

path $q_1, q_2, \dots, q_t$, when observed state is o_i at time-stamp t

- (2) $\varphi_t(i)$ represents a status value, which saves the last optimal state leading to the current state
- (3) P^* is the output probability
- (4) q_T^* saves the final choice of optimal hidden state at timestamp t

The pseudocode of the Viterbi algorithm is shown in Algorithm 2, it first initializes the end cache and initializes the variable which is used to track the index of the max node to zero (lines 2-5); then, it calculates over all remaining observation sequences and states the partial max and stores away the index that delivers to it (lines 7-12); then, we find the Viterbi path with the maximum transition probability by calculating the inverse function $\delta_T(i)$ (lines 14-15); and finally, it will return the hidden state sequence corresponding to the given observation sequence (line 17). Thus, when given a sequence of trajectory points, we can obtain the most probable hidden state sequences through the Viterbi algorithm.

5.3. Shop Category Augmented Trajectory Prediction. In this section, we will present the details of shop category prediction and fuse it with the indoor trajectory HMM to help improve the prediction accuracy.

Given a trajectory sequence $T = \{P_1, P_2, \dots, P_n\}$, the number of shop category sequences is

$$N = \prod_{i=1}^n |\text{sub}_{\text{category}_i}|, \quad (6)$$

where $|\text{sub}_{\text{category}_i}|$ is the number of shop categories that the i^{th} trajectory point covers. As the example shown in Table 4, given a trajectory sequence $\{P_1, P_2, P_3, P_4\}$, each Wi-Fi access point will cover several shop categories; in our example, P_1 covers four shop categories, namely, a, b, c , and d ; P_2 covers three shop categories, namely, a, d , and e ; and P_3 covers three shop categories, namely, a, b , and c . First, we generate the possible shop category sequences according to the $(n-1)$ trajectory points; in the example, there are a total of $4 * 3 * 3 = 36$ possible shop category sequences. Then the generated shop sequences are used to do the prediction, based on shop category HMM, we can get the possible shop categories to be visited next. In our example, if a or d (the shop categories that P_4 covers) is in the set of predicted shop categories, we say the shop category to be visited next is accurately predicted.

In order to improve the precision of trajectory prediction, we integrate the shop category prediction with trajectory prediction. Each candidate' shop category weight is computed according to Algorithm 3: candidate list $L2$ stores possible shop categories to be visited next; then for each candidate access point AP_i , if one of its surrounding Shopcategory $_i$ is in $L2$, we assign β to its shop category weight, where $\beta = 1/\text{length}(L2)$; finally, we apply Formula

FIGURE 3: Working mechanism.

input: Observations of length n , indoor trajectory HMM
output: best path that generates the observations

```

1: // initialization
2: for each state in hidden states do
3:    $\delta_t(i) = \pi_i b_i(o_1)$ 
4:    $\varphi_1(i) = 0$ 
5: end for
6: //recursion
7: for each time step  $t$  from 2 to  $n$  do
8:   for each state in hidden states do
9:      $\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(o_t)$ 
10:     $\varphi_t(j) = \operatorname{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}]$ 
11:   end for
12: end for
13: //termination
14:  $P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$ 
15:  $q_T^* = \operatorname{argmax}_{1 \leq i \leq N} [\delta_T(i)]$ 
16: // backtracking
17: Return  $q_t^* = \varphi_{t+1}(q_{t+1}^*), t = T-1, T-2, \dots, 1$ 

```

ALGORITHM 2: Viterbi algorithm.

TABLE 4: Shop category example.

Trajectory sequence	P_1	P_2	P_3	P_4
	a	a	a	a
	b	d	b	d
Shop category the AP covers	c	e	c	
	d			

(7) to compute the unified score $\operatorname{Pro}(\text{AP}_i)$ of each access point AP_i to be visited; α is the weight we signed to shop category.

$$\operatorname{ProNext}(\text{AP}_i) = (1 - \alpha)\operatorname{Pro}(\text{AP}_i) + \alpha W_{\text{category}}[i]. \quad (7)$$

5.4. Indoor Trajectory Prediction Algorithm. In this section, we propose the details of the algorithm for predicting the next location in indoor scenarios.

As shown in Algorithm 4, we aim to predict the position of the $(n+1)^{\text{th}}$ time stamp beyond the previous n trajectory points; the details of the algorithm are described as follows: (1) initialize the parameters (line 1); (2) check whether the

```

input: Candidate AP list  $L1$ , Predicted shop category list  $L2$ 
output: Each candidate  $AP_i$ 's shop category weight
1:  $\beta = 1/\text{len}(L2)$ ,  $W_{\text{category}} = []$ 
2: for  $AP_i$  in  $L1$  do
3:   for  $Shopcategory$  in  $Sub_{\text{category}_i}$  do
4:     if  $Shopcategory$  in  $L2$  then
5:        $W_{\text{category}}[i] += \beta$ 
6:     end if
7:   end for
8: end for
9: Return  $W_{\text{category}}$ 

```

ALGORITHM 3: Computing shop category weight.

```

input: A set of indoor trajectory sequences  $T = \{P_1, P_2, \dots, P_n\}$ 
output: The predicted trajectory point  $P_{n+1}$  at  $(n+1)^{th}$  timestamp.
1: Initialize parameters;
2: if  $T.\text{length} \leq 0$  then
3:   Return FALSE;
4: end if
5:  $RT(AP_1, AP_2, \dots, AP_n) = \text{Viterbi}(T)$ ;
6: for  $i = 0$  to  $m\_DimA$  do
7:   Compute  $\text{ProNext}(AP_i)$ 
8:   Rank( $AP_i$ )
9: end for
10: Return  $P_{n+1}$ .

```

ALGORITHM 4: Trajectory prediction.

length of the trajectory meets the requirement (lines 4-5); (3) use the Viterbi algorithm to compute the most probable hidden states corresponding to the observation sequence (line 5); (4) traverse all the next possible places to be visited, calculate the possibilities, and rank the possible APs in a decreasing order (lines 6-9); and (5) return the next possible access point (line 10).

6. Experimental Evaluation

In this section, we would like to evaluate and compare our algorithm against baseline methods. All algorithms were implemented using Python 3.5. The dataset was split into two sets, namely, training dataset and testing dataset, in terms of check-in time rather than using a random partition method. The intuition is that in practice we can only use the past check-in data to predict the future check-in events. The training set was used to learn the predicting models of the proposed method, and the testing set was used to evaluate the accuracy of the prediction model.

6.1. Experiment Setup

6.1.1. Dataset. The trajectory dataset we used was obtained from a large inner city shopping mall with 67 Wi-Fi access points (APs) across 90,000 square meters between September 2012 and October 2013. The mall contains over 200 stores that belong to 34 shop categories (e.g., Women/Men's

TABLE 5: Statics of the dataset.

Entry	Value
Number of APs	67
Number of associations	907,084
Number of shops	200
Number of shop categories	34
Number of trajectories	261,369

Fashion and General Footwear). Such data provides a unique opportunity for us to analyze the interaction between users' physical movement and semantic movement in indoor spaces. The statistics of the data set are shown in Table 5.

Figure 4 depicts the trend about the number of trajectories w.r.t their length. We omit the trajectories whose lengths are smaller than those of the three which are too short to convey meaningful visiting patterns. As can be seen from the figure, more than 90% of people tend to access three to nine Wi-Fi access points when they are in the shopping mall.

Based on the floor plan, the proximal areas of APs are classified into three main categories: food court which consists of 11 APs, retail which is made up of 46 APs, and navigational where 10 APs belong to this main category. Table 6 shows the distribution of visiting time per category. Seven percent of users' visiting time was spent in navigational areas, 23% was spent in the food court, and 70% was spent

FIGURE 4: Number of trajectories w.r.t the length.

TABLE 6: Context in relation to visiting time, as a proportion of all association time spent at a given category of AP (and per AP within category), as well as average time.

Context	% of association time	Av time (hr)
Food court	23% (2.06% per AP)	1.39
Retail	70% (1.52% per AP)	2.29
Navigational	7% (0.68% per AP)	1.00
Total	100% (1.49% per AP)	2.77

in a retail context. The largest average duration per AP was measured in the food court. In addition, from the average of visiting time per user visit, we observed that indoor users tend to spend more time in a retail context than in other physical contexts.

6.1.2. Parameters. Table 7 summarizes all of the parameters used in this work. A parameter was set to the default value in the experiments where any other parameters were being varied.

6.1.3. Performance Metrics. In general, our trajectory predicting technique computes a score for each candidate item (i.e., AP in this paper) regarding a target user and returns APs with the top- k highest scores as prediction results to the target user. To evaluate the quality of our method, we apply three metrics:

- (1) $p@1$: it measures the percentage of times in which we found the correct next location.
- (2) $p@5$: it measures the percentage of times in which the correct next location was given in the top-five most probable locations.
- (3) Mean reciprocal rank: it measures the average of the reciprocal ranking positions for the correct next location.

$$\text{MRR} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\text{ran } k_i}. \quad (8)$$

TABLE 7: Parameters that we used in our experiments.

Parameter	Values	Default
Similarity threshold (δ)	0.3, 0.4, 0.5, 0.6, 0.7	0.5
Number of patterns (n)	1, 2, 3, ..., 20	15
Weight of shop category (α)	0.9, 0.8, ..., 0.1	0.8

FIGURE 5: Performance on whole dataset.

To examine the effectiveness of our proposed method, we have implemented four comparing algorithms: (1) a naive prediction algorithm based on HMM (Naive), (2) prediction algorithm that considers AP similarities (SM), (3) prediction algorithm that takes frequent subtrajectory pattern into consideration and finally (FS), and (4) prediction algorithm which fuses subtrajectory pattern and shop categories into a unified model (FSS).

6.2. Experimental Results. This section analyzes the extensive experimental results. We first investigate the effect of similarity parameter δ on the prediction accuracy; then, we will check the trend of accuracy when varying the number of frequent subtrajectory patterns; finally, we will compare the fusion method with the aforementioned approaches and discuss some important findings about trajectory prediction in indoor scenarios.

6.2.1. Effect Analysis of the Similarity Threshold. In this section, we will investigate the performance of grouping APs into a cluster as hidden states according to AP similarities and check the trend of predicting accuracy under different similarity thresholds. The experiments evaluate the effect of parameter δ varying from 0.3 to 0.7 with the step of 0.1 each time to obtain a new HMM. Note that we only report the results of δ ranging from 0.3 to 0.7 because when δ is larger than 0.5; it will have little impact to the final results corresponding to Figure 5. Then, we check the performance of different HMM and analysis the accuracy trend.

As shown in Figure 5, with the increasing of similarity threshold, the accuracy of $p@5$ increases and remains quite stable when the similarity threshold is larger than 0.5. The reason may be that with the increment of similarity

FIGURE 6: Prediction precision evaluation: optimization techniques when varying the similarity threshold.

FIGURE 7: Prediction precision evaluation: optimization techniques when varying the number of patterns.

threshold, Wi-Fi access points are less likely to be clustered into a group; thus, the hidden states in the model increase also, which strengthens the model's predicting ability, so at the early stage, the prediction accuracy increases. However, after a certain threshold, the newly generated hidden states will not affect the model's predicting ability too much. Thus, throughout the experiment, we find that the model's predicting ability will not always improve with the increasing of hidden states.

TABLE 8: Prediction of shop sequence.

Length	Number	Number of shop sequence	Hit time
4	1000	24870	651
5	1000	91656	660
6	1000	370672	402
7	1000	919024	723

FIGURE 8: Prediction precision evaluation: optimization techniques when varying α .

The trend is more obvious as depicted in Figure 6; we illustrate the prediction accuracy for different lengths of trajectories here. The length of trajectories ranges from five to nine; we can conclude from the figure that with the increasing of similarity threshold, the accuracy of $p@5$ increases steadily, and another important finding is that as the length of the trajectories increases, the prediction accuracy also increases. The reason may be that as the trajectory length increases, when computing the most probable hidden state sequence, more historical trajectory points will be taken into consideration leading to a better prediction result. This is true in a real-world scenario, because we will know better about a user when his visiting sequence becomes longer.

6.2.2. Effect Analysis of the Number of Frequent Subtrajectories. In this section, we will investigate the performance of frequent subtrajectory-based HMM. We incrementally add top frequent subtrajectory patterns into the hidden states of our model. The number of frequent subtrajectories ranges from 5 to 21 covering 14% to 61% of total Wi-Fi access points, respectively. The results are depicted in Figure 7.

As can be seen from the figure, the trend of predicting accuracy for $p@1$, $p@5$, and MRR do not increase stably when we increase the number of frequent subtrajectories. But the best result occurs when we employ the top 15 frequent patterns that cover 44% Wi-Fi access points.

6.2.3. Accuracy on Shop Category Prediction. In this section, we will check to what extent can we accurately predicate the shop categories going to be visited next.

TABLE 9: Accuracy comparison.

Methods	$p@1$	$p@5$	MRR
Naive	0.05313	0.27189	0.17212
SM	0.07234	0.30077	0.19238
FS	0.12409	0.37591	0.25054
FSS	0.13509	0.40414	0.26204

Table 8 shows the prediction results. We investigated trajectories whose length ranges from 4 to 7, the number of generated shop sequences ranges from 24870 to 919024, and the prediction accuracy varies from 40% to 70%. Thus, we can surmise that human movement exhibits certain patterns and adopting the shop category will increase the accuracy of trajectory prediction.

6.2.4. Effect Analysis of Fusion Methods. In this section, we will check whether combining the access point and shop category into a unified model will obtain a better a result or not. Figure 8 shows the results obtained when we vary the weight parameter of shop category α in which the optimal value of α is around 0.8. We can conclude from the results that introducing shop category can help increase the location prediction accuracy, with additional information about the customers; we can predict their whereabouts more precisely. Table 9 compares the results obtained through the four different methods. As can be seen from the table, a naive method performs worst mainly because it does not consider the features in the indoor settings. Our frequent subtrajectory-based method outperforms the naive and similarity-based methods greatly for $p@1$, $p@5$, and MRR. Especially for $p@1$, the accuracy of frequent subtrajectory-

based method nearly doubles the result of the similarity-based method indicating the physical movement pattern plays a crucial rule in indoor trajectory prediction. Further, when we integrate shop category into trajectory prediction, the prediction accuracy improves for $p@1$, $p@5$, and MRR, indicating that the customers' movements have certain semantic intentions in real life which can be used to help increase the accuracy of indoor trajectory prediction.

7. Conclusion

Trajectory prediction in indoor space is a challenging and hard topic, unlike in the outdoor scenario where travelers walk through the foot path or road; in the indoor setting, even if in a small region, there will be multiple choices for the next location to go to. In order to accurately predict the mobility of customers, we employed the HMMs to precisely depict transition between trajectory points. The proposed approach first mines hidden states of indoor trajectory points; in order to obtain the hidden states, we put forward two methods here: the first method groups the Wi-Fi access points according to their similarities as hidden states, and the second method considers the frequent subtrajectories as hidden states; then, we use HMM to model and calculate relative parameters. After that, the most likely hidden sequences corresponding to the given trajectory are calculated through the Viterbi algorithm and the ranking list for the next probable access point is generated. In order to enhance the prediction ability, we fuse the results of trajectory prediction and shop category prediction into a unified framework. Finally, we have conducted extensive experiments to evaluate the prediction accuracy of our proposed method using a large-scale real dataset collected from a shopping mall and compared it with baseline methods; results show that our newly proposed approach outperforms the baselines significantly in terms of prediction accuracy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant Nos. 61672179, 61370083, and 61402126; the Natural Science Foundation of Heilongjiang Province under grant No. F2015030; the Youths Science Foundation of Heilongjiang Province of China under grant No. QC2016083; the Fundamental Research Funds for the Central Universities under Grant No. HEUCFM180601; and the Heilongjiang Postdoctoral Science Foundation under grant No. LBH-Z14071.

References

- [1] B.-K. Ang, D. Dahlmeier, Z. Lin, J. Huang, M.-L. Seeto, and H. Shi, "Indoor next location prediction with Wi-Fi," in *The Fourth International Conference on Digital Information Processing and Communications (ICDIPC2014)*, pp. 107–113, Kuala Lumpur, Malaysia, 2014.
- [2] B. Wang, Y. Hu, G. Shou, and Z. Guo, "Trajectory prediction in campus based on Markov chains," in *International Conference on Big Data Computing and Communications*, pp. 145–154, Springer, 2016.
- [3] P. L. Jenkins, T. J. Phillips, E. J. Mulberg, and S. P. Hui, "Activity patterns of californians: use of and proximity to indoor pollutant sources," *Atmospheric Environment. Part A. General Topics*, vol. 26, no. 12, pp. 2141–2148, 1992.
- [4] N. E. Klepeis, W. C. Nelson, W. R. Ott et al., "The national human activity pattern survey (nhaps): a resource for assessing exposure to environmental pollutants," *Journal of Exposure Science and Environmental Epidemiology*, vol. 11, no. 3, p. 231, 2001.
- [5] S. Shearman, "Tesco provides free Wi-Fi in larger UK & Ireland stores," 2014, <http://www.essentialretail.com/in-store-ops/article/546b3eeb912b1-tesco-provides-free-wi-fi-in-larger-uk-ireland-stores>.
- [6] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 3, pp. 55–66, 2001.
- [7] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless atm networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 922–936, 1998.
- [8] P. S. Prasad and P. Agrawal, "Mobility prediction for wireless network resource management," in *System Theory, 2009. SSST 2009. 41st Southeastern Symposium on (2009)*, pp. 98–102, Tullahoma, TN, USA, 2009.
- [9] C. S. Jensen, H. Lu, and B. Yang, "Indoor-a new data management frontier," *IEEE Data Engineering Bulletin*, vol. 33, no. 2, pp. 12–17, 2010.
- [10] J. Huang, D. Dahlmeier, Z. Lin, B.-K. Ang, M.-L. Seeto, and H. Shi, "Wi-fi based indoor next location prediction using mixed state-weighted Markov-chain model," *International Journal of Machine Learning and Computing*, vol. 4, no. 6, p. 505, 2014.
- [11] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," *ACMProceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp. 25–33, Chicago, IL, USA, 2011.
- [12] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung, "Mining, indexing, and querying historical spatiotemporal data," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 236–245, Seattle, Washington, USA, 2004.
- [13] M. Morzy, "Mining frequent trajectories of moving objects for location prediction," in *Machine Learning and Data Mining in Pattern Recognition*, pp. 667–680, Springer, 2007.
- [14] Q. Zhao and S. S. Bhowmick, *Sequential pattern mining: a survey*, ITechnical Report CAIS Nayang Technological University Singapore, 2003.

- [15] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Mining user similarity from semantic trajectories," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 19–26, San Jose, CA, USA, 2010.
- [16] L. Nguyen, H.-T. Cheng, P. Wu, S. Buthpitiya, and Y. Zhang, "Pnlum: system for prediction of next location for users with mobility," in *Nokia Workshop at the Tenth International Conference on Pervasive Computing*, Newcastle, UK, June 2012.
- [17] J. Petzold, A. Pietzowski, F. Bagci, W. Trumler, and T. Ungerer, "Prediction of indoor movements using Bayesian networks," in *International Symposium on Location-and Context-Awareness*, pp. 211–222, Springer, 2005.
- [18] Y. Ishikawa, Y. Tsukamoto, and H. Kitagawa, "Extracting mobility statistics from indexed spatio-temporal datasets," *STDBM*, pp. 9–16, 2004.
- [19] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility Markov chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, p. 3, Bern, Switzerland, 2012.
- [20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [21] R. M. Altman, "Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 201–210, 2007.
- [22] S. Qiao, D. Shen, X. Wang, N. Han, and W. Zhu, "A self-adaptive parameter selection trajectory prediction approach via hidden Markov models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 284–296, 2015.
- [23] S. Gong, M. Walter, and A. Psarrou, "Recognition of temporal structures: learning prior and propagating observation augmented densities via hidden Markov states," in *The Proceedings of the Seventh IEEE International Conference on (1999)*, pp. 157–162, Kerkyra, Greece, 1999.
- [24] D. Makris and T. Ellis, "Spatial and probabilistic modelling of pedestrian behaviour," *BMVC*, pp. 1–10, 2002.
- [25] O. O. Barzaiq, S. W. Loke, and H. Lu, "On trajectory prediction in indoor retail environments for mobile advertising using selected self-histories," in *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 304–307, Beijing, China, 2015.
- [26] L. D. Lam, A. Tang, and J. Grundy, "Predicting indoor spatial movement using data mining and movement patterns," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 223–230, Jeju, Korea (South), 2017.
- [27] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.