Context-Aware Semantic Annotation of Mobility Records

HUANDONG WANG and YONG LI, Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University
JUNJIE LIN, University of California
HANCHENG CAO, Department of Computer Science, Stanford University
DEPENG JIN, Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University

The wide adoption of mobile devices has provided us with a massive volume of human mobility records. However, a large portion of these records is unlabeled, i.e., only have GPS coordinates without semantic information (e.g., Point of Interest (POI)). To make those unlabeled records associate with more information for further applications, it is of great importance to annotate the original data with POIs information based on the external context. Nevertheless, semantic annotation of mobility records is challenging due to three aspects: the complex relationship among multiple domains of context, the sparsity of mobility records, and difficulties in balancing personal preference and crowd preference. To address these challenges, we propose CAP, a context-aware personalized semantic annotation model, where we use a Bayesian mixture model to model the complex relationship among five domains of context—location, time, POI category, personal preference, and crowd preference. We evaluate our model on two real-world datasets, and demonstrate that our proposed method significantly outperforms the state-of-the-art algorithms by over 11.8%.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Information systems** → **Spatial-temporal systems**; • **Networks** → *Network services;*

Additional Key Words and Phrases: Mobility trajectories, semantic annotation, point of interest, graphical model

ACM Reference format:
Huandong Wang, Yong Li, Junjie Lin, Hancheng Cao, and Depeng Jin. 2021. Context-Aware Semantic Annotation of Mobility Records. *ACM Trans. Knowl. Discov. Data.* 16, 3, Article 47 (October 2021), 20 pages.
https://doi.org/10.1145/3477048

This work was supported in part by the National Key Research and Development Program of China under grant 2018YFB1800804, the National Nature Science Foundation of China under U1936217, 61971267, 61972223, 61941117, 61861136003, and 6217012092, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, the China Postdoctoral Science Foundation under 2021M691830, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

Authors' addresses: H. Wang, Y. Li, and D. Jin, Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; emails: {wanghuandong, liyong07, jindp}@tsinghua.edu.cn; J. Lin, University of California, San Diego 92093; email: jjlin@ucsd.edu; H. Cao, Department of Computer Science, Stanford University, CA 94305; email: hanchcao@stanford.edu.

1556-4681/2021/10-ART47 $15.00
https://doi.org/10.1145/3477048

1 INTRODUCTION

With the popularization of mobile devices and the booming of **location-based social networks** (**LBSNs**), a massive volume of human mobility records becomes available to service providers. Among them, records labeled with semantic information (e.g., **Point of Interest** (**POI**)) are of the greatest importance because they indicate where users actually visit (e.g., a particular restaurant), which acts as a powerful context for accurate user behavior understanding and analysis [2, 4]. For example, using mobility records with semantic information, service providers can make more accurate location prediction [5, 21], better understand human mobility patterns [1, 36], know the user preference better and recommend corresponding products or POIs for them [24], and better facilitate other applications (e.g., healthcare [14]). Also, local businesses and governments can use the information to better understand region demands and support their decision making process [3, 12, 16].

Although labeled records with semantic information are useful, such data are limited in number since they can only be obtained when LBSN users actively post their status. Under most circumstances, however, only unlabeled GPS records are available to service providers when users use related applications. Moreover, these GPS records are often noisy and not perfectly located on certain POIs. Thus, it is important and challenging to annotate unlabeled records with correct POIs to make them as informative as labeled records.

To achieve this goal, a number of methods have been proposed in the literature. Existing studies consider the distance between the GPS coordinates and the location of POIs [35]. Gong et al. [7] further take the temporal context and POI category into consideration. Wu et al. [34] explore the history of a user's mobility records, yet they model each user independently without considering the crowd preference. Zhang et al. [48] further leverage user grouping to annotate mobility records with POIs. These studies are still far from satisfactory since they did not properly address the following three challenges:

- **Complexity of Context.** Semantic annotations are based on external context, including the distance between the user and POIs, the popularity of POIs at different times of the day, user's history, and other users' history. It is challenging to consider all these factors simultaneously and combine them in one model of annotation.
- **Sparsity of Records.** It is challenging to model POIs and users before annotation due to the skew and sparsity of mobility records. Since a small number of POIs contribute to the majority of mobility records, it is challenging to model POIs that are rarely visited. Furthermore, it is difficult to model the mobility pattern of users who leave only a few records.
- **Personal and Crowd Preference.** To achieve personalized annotation, we should balance personal preference and crowd preference. When considering personal preference, most existing works ignore crowd preference, which is important information especially when the data are highly sparse.

To tackle the three challenges, in this article, we propose CAP, a **context-aware personalized (CAP)** semantic annotation model based on the complex context and user preference. More specifically, we propose a Bayesian mixture model by considering five domains of context simultaneously—location, time, POI category, personal preference, and crowd preference. In order to alleviate data skew, we assume POIs in the same category share the same temporal pattern, which is captured by a Gaussian mixture distribution [7]. To overcome the problem of data sparsity, we use the semi-supervised learning method by adding unlabeled records to the training set. We model user preference with the multinomial-Dirichlet distribution. By tactfully choosing

Table 1. Notations and Description

Notation	Description
$\mathcal{U}, \mathcal{C}, \mathcal{P}$	The set of all users, POI categories, and POIs, respectively.
$\mathcal{P}_{\boldsymbol{l}}$	The set of candidate POIs at location $\boldsymbol{l}$.
$\mathcal{D}_L^u, \mathcal{D}_U^u$	The set of labeled and unlabeled records for user u, respectively.
$\boldsymbol{d}_i, u_i, \boldsymbol{l}_i, t_i, p_i$	The ith record and its user, location, timestamp and POI.
r_i, c_i, z_i	The region, POI category and temporal Gaussian component of the ith record.
$\boldsymbol{\mu}_r, \Sigma_r$	The mean and covariance matrix corresponding to region r.
$\nu_{c,z}, \sigma_{c,z}$	The mean and standard deviation of temporal Gaussian component z for category c.
$\boldsymbol{\theta}$	Parameters of multinomial region distribution.
$\boldsymbol{\psi}_c$	Parameters of multinomial temporal Gaussian component distribution for category c.
$\boldsymbol{\Phi}_{r,u}$	Parameters of multinomial category distribution for region r, user u.
$b \cdot \boldsymbol{\Phi}_r$	Parameters of Dirichlet prior for $\boldsymbol{\Phi}_{r,u}$ in region r, where b is a hyper-parameter to adjust the influence of $\boldsymbol{\Phi}_r$.
$\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}_0, \rho_0$	Hyper-parameters of NIW prior for $(\boldsymbol{\mu}_r, \Sigma_r)$.
$\nu_0, \lambda_0, \epsilon_0, \tau_0$	Hyper-parameters of NIG prior for $(\nu_{c,z}, \sigma_{c,z})$.
α, β, γ	Hyper-parameters of Dirichlet prior for $\boldsymbol{\theta}, \boldsymbol{\Phi_r}, \boldsymbol{\psi_c}$, respectively.

the Dirichlet prior, we successfully balance user preference and common behavior patterns. Our contributions can be summarized as follows:

- We propose a novel Bayesian mixture model, which is the first work to simultaneously model five domains of context in human mobility. By taking location, time, POI category, personal preference, and crowd preference into consideration, we present by far the most systematic approach, which achieves considerable performance gain in the task of semantic annotation.
- We use a semi-supervised learning method to alleviate data sparsity by using both labeled and unlabeled mobility records. Specifically, we first model crowd preference using labeled records and further capture personal preference using unlabeled records.
- We evaluate our proposed model with real-world datasets of two cities through extensive experiments. The results show that our method significantly outperforms the state-of-the-art algorithms on all datasets by over 11.8%.

2 SYSTEM MODEL AND OVERVIEW

2.1 Mathematical Setup and Problem Description

We denote a human mobility record as a 4-tuple $\boldsymbol{d}_i = \{u_i, \boldsymbol{l}_i, t_i, p_i\}$, where u_i represents the user ID, $\boldsymbol{l}_i$ represents the geographic coordinates, and t_i represents the timestamp. If the record is labeled, p_i is the associated POI. For unlabeled records without POI, p_i is null. Given any user $u \in \mathcal{U}$, we define the set of all the labeled and unlabeled records as $\mathcal{D}_L^u$ and $\mathcal{D}_U^u$, respectively. The set of all labeled records of all users is denoted as $\mathcal{D}_L = \bigcup_{u \in \mathcal{U}} \mathcal{D}_L^u$.

We further denote a POI as a 2-tuple $p = \{\boldsymbol{l}, c\}$, where $\boldsymbol{l}$ represents the geographic coordinates and c represents the category. The set of POIs and POI categories are denoted as $\mathcal{P}$ and $\mathcal{C}$, respectively. Given a location $\boldsymbol{l}$, we further define a set containing candidate POIs within a distance δ

from location $\boldsymbol{l}$ as $\mathcal{P}_{\boldsymbol{l}}$, i.e., $\mathcal{P}_{\boldsymbol{l}} = \{p | p \in \mathcal{P} \wedge dist(p.\boldsymbol{l}, \boldsymbol{l}) < \delta\}$. Overall, we summarize major notations used in this article in Table 1. Then, the problem of this article can be defined as follows:

Semantic Annotation Problem:

Given: POIs $\mathcal{P}$, labeled records $\mathcal{D}_L$, a target user u with the corresponding unlabeled records $\mathcal{D}_U^u$.

Problem: Find the associated POI $p_i \in \mathcal{P}$ for each unlabeled record $\boldsymbol{d}_i \in \mathcal{D}_U^u$.

2.2 System Overview

Figure 1 shows the framework of our POI annotation system that models users' mobility behavior based on human mobility records. Our system takes labeled and unlabeled mobility records of all users as input. The goal is to annotate each unlabeled mobility record with the correct POI.

When annotating an unlabeled record with POI, our system takes the following context into consideration: the distance between the user's location and candidate POIs, POI category obtained from map services, the popularity of POIs at different times of the day, crowd preference in different functional regions, and personal preference inferred based on user history.

As we can observe from Figure 1, for better modeling users' mobility in terms of the spatial dimension, our system first automatically divides the city into functional regions, where most users follow similar crowd preference to POIs of different categories [45]. The motivation of using such automatically divided regions rather than existing static regions (e.g., administrative regions) is that users' behavior may be not consistent with existing static regions. Furthermore, POIs and users are evolving over time. Thus, the boundaries of functional regions are ambiguous, and we should use regions automatically divided by our system.

At the same time, for better modeling users' mobility in terms of the temporal dimension, based on timestamps and POI categories from the mobility records of all users, our system seeks to estimate the temporal patterns of visitations to different POIs. However, due to data skew, i.e., some POIs are popular while others are rarely visited, it is hard to model the temporal visitation pattern of each POI accurately. Thus, we utilize the fact that POIs of the same category share the same temporal visitation pattern, e.g., restaurants are popular during lunchtime and dinnertime. Then, they share the same parameters describing the temporal visitation pattern in our model, and our system only needs to estimate the temporal patterns of visitations to POIs of each category.

Our system also takes personal preference into consideration. By utilizing the crowd preference of automatically divided functional regions as prior, our model further learns the personal preference of each user based on his/her historical labeled mobility records.

Finally, our system combines the contexts of location, time, POI category, crowd preference, and personal preference in a cohesive manner to estimate the associated POI of each unlabeled mobility record.

3 SEMANTIC ANNOTATION MODEL

As introduced in the system overview, we design our system based on the following three observations: First, different areas of a city show distinct functions. In different functional regions (e.g., financial district, airport, uptown), people tend to visit POIs of corresponding categories. Second, different categories of POI have unique temporal visiting patterns, i.e., their popularity varies at different times of the day. Third, although users usually follow crowd preference when visiting an area, they also demonstrate personal preference.

3.1 Bayesian Mixture Model

Given user $u \in \mathcal{U}$, we propose a Bayesian mixture model to describe the mobility patterns based on labeled records of all users $\mathcal{D}_L$ and unlabeled records $\mathcal{D}_U^u$. The ith record is denoted as $\boldsymbol{d}_i \in$

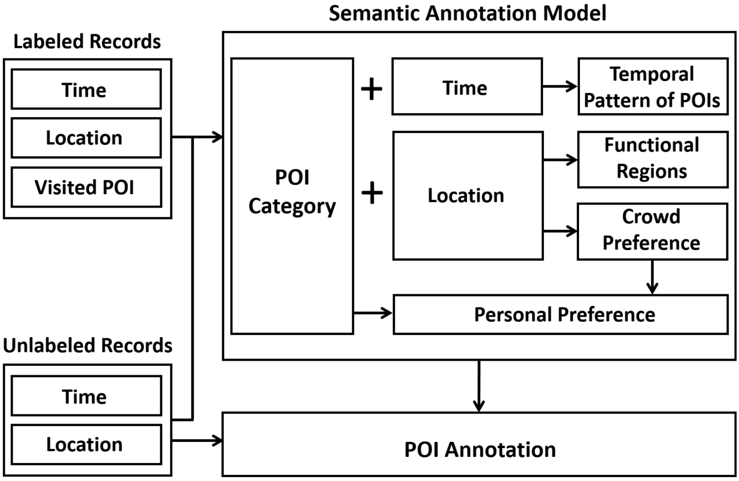


Fig. 1. The framework of our system.

$\mathcal{D}_L \bigcup \mathcal{D}_U^u$. Our model jointly utilizes spatial context, temporal context, and user preference, and can be decomposed into three components: location, time, and POI category modeling.

3.1.1 Location. It has been found that different areas of a city have distinct functions in a number of existing studies [41, 42]. Thus, our model divides a city into functional regions based on users' mobility records. Given record $\boldsymbol{d}_i$ with location $\boldsymbol{l}_i$ characterized by geographic coordinates, we introduce a latent discrete random variable r_i to indicate which region it belongs to. The set of all regions is denoted as R. We use multinomial distribution Multi$(\boldsymbol{\theta})$ to model r_i, where $\boldsymbol{\theta}$ is a $|R|$-sized vector and $p(r_i = r) = \theta_r$ for each $r \in R$. In order to model the randomness of human behavior and the inherent GPS error, for records belonging to each $r \in R$, we use bivariate Gaussian distribution to model their location by following existing approaches [24, 45, 46]. In addition, in order to avoid too much parameters, the location l_i is modeled to be independent with the category c_i conditioned on the region r_i. Therefore, given record $\boldsymbol{d}_i$ belonging to region r, the probability density of its geographic coordinate $\boldsymbol{l}_i$ can be represented as follows:

$$p(\boldsymbol{l_i}|r) = \mathcal{N}(\boldsymbol{l_i}|\boldsymbol{\mu}_r, \Sigma_r) = \frac{\exp\left(\frac{-(\boldsymbol{l_i}-\boldsymbol{\mu}_r)^T\Sigma_r^{-1}(\boldsymbol{l_i}-\boldsymbol{\mu}_r)}{2}\right)}{2\pi\sqrt{|\Sigma_r|}}, \tag{1}$$

where $\boldsymbol{\mu}_r$ and Σ_r are the mean vector and covariance matrix of region r, respectively.

3.1.2 Time. To alleviate data skew, i.e., some POIs are popular while others are rarely visited, we assume POIs of the same category share the same temporal visitation pattern. For example, restaurants are popular during lunchtime and dinnertime, whereas nightclubs are popular in the evening. We model the temporal pattern as Gaussian mixture distribution with H temporal Gaussian components [49]. Suppose that record $\boldsymbol{d}_i$ belongs to category c, the probability density of its time t_i can be written as

$$p(t_i|c)=\sum_{z=1}^{H}\boldsymbol{\psi}_{c,z}\mathcal{N}(t_i|\nu_{c,z},\sigma_{c,z})=\sum_{z=1}^{H}\boldsymbol{\psi}_{c,z}\frac{\exp\left(\frac{-(t_i-\nu_{c,z})^2}{2\sigma_{c,z}^2}\right)}{\sqrt{2\pi\sigma_{c,z}^2}}, \tag{2}$$

where $\mathcal{N}(t_i|\nu_{c,z},\sigma_{c,z})$ is a Gaussian component, z is a latent discrete random variable indicating which component it belongs to, and $\nu_{c,z}$ and $\sigma_{c,z}$ are mean and standard deviation of component

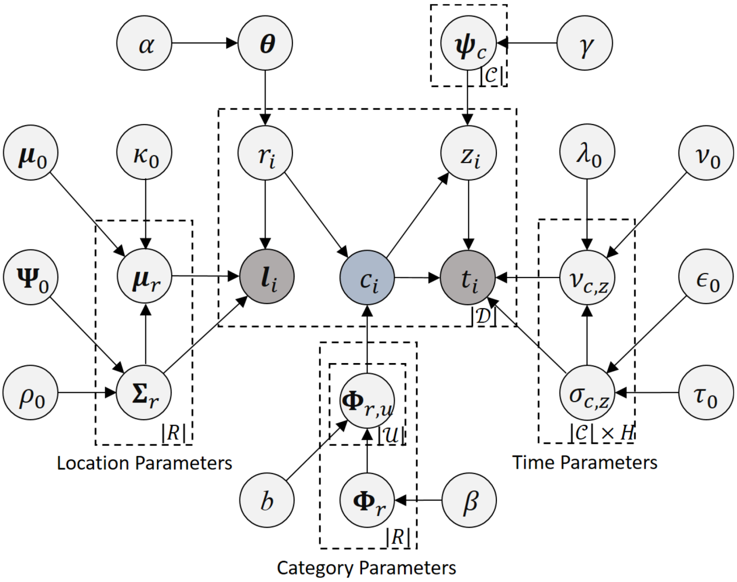


Fig. 2. Graphical model of CAP.

z for category c. Based on observations from existing studies [11, 20], most POIs have one or two peak hours in a day, we set $H = 2$ in our model.

3.1.3 POI Category. To leverage personal preference, we assume that different users tend to visit different categories of POI in a certain region. For example, a user may visit office buildings and fast food restaurants in one region frequently if he works there, while another user who resides there may always visit an apartment instead. Given record $\boldsymbol{d}_i$, we use variable c_i to indicate its category. Given region r and user u, we use multinomial distribution Multi($\boldsymbol{\Phi}_{r,u}$) to model the category, where $\boldsymbol{\Phi}_{r,u}$ is a $|C|$-sized vector and $p(c_i = c|r,u) = \Phi_{r,u,c}$ for each $c \in C$. Note that the correlation between c_i and $\boldsymbol{l}_i$ are modeled by combining $p(\boldsymbol{l}_i|r)$ and $p(c_i|r,u)$ with the joint correlated variable r_i. In addition, when the record $\boldsymbol{d}_i$ is labeled, c_i is an observable variable, but it becomes a latent variable when $\boldsymbol{d}_i$ is unlabeled.

However, the labeled mobility records of users are usually sparse [9], which leads to overfitting if we directly train $\boldsymbol{\Phi}_{r,u}$ for each region and each user using such limited data. To overcome this problem, we further consider crowd preference by first replacing $\boldsymbol{\Phi}_{r,u}$ with a universal vector $\boldsymbol{\Phi}_r$ for all users, i.e., supposing every user has the same interest in a region. We use labeled records of all users $\mathcal{D}_L$ to train the model and learn $\boldsymbol{\Phi}_r$ for each region. After that, for each user $u \in \mathcal{U}$, we add the corresponding unlabeled records D_U^u to the training set and use $b \cdot \boldsymbol{\Phi}_r$ as a prior of $\boldsymbol{\Phi}_{r,u}$, i.e., $\boldsymbol{\Phi}_{r,u} \sim$ Dirichlet$(b \cdot \boldsymbol{\Phi}_r)$ where b is a hyper-parameter to adjust the influence of $\boldsymbol{\Phi}_r$.

Further, we employ common conjugate prior distributions as prior distribution for parameters in our model to simplify the parameter estimation. For $\boldsymbol{\theta}$, $\boldsymbol{\psi}_c$ and $\boldsymbol{\Phi}_{r,u}$, which are the parameters of multinomial distribution, we employ Dirichlet distribution as their prior distribution. For $\boldsymbol{\mu}_r$ and Σ_r, which are the parameters of bivariant Gaussian distribution, we employ **Normal-inverse-Wishart** (**NIW**) distribution as their prior. For $\nu_{c,z}$ and $\sigma_{c,z}$, which are the parameters of Gaussian distribution, we use **Normal-inverse-gamma** (**NIG**) distribution.

In summary, we divide a city into several regions and apply a Gaussian distribution for each region. For each POI category, we use the Gaussian mixture distribution to model its temporal pattern. Also, we adopt a multinomial distribution to model user preference in each region. Figure 2 plots the graphical model and the generative progress is shown in Algorithm 1.

3.2 Parameter Estimation

We use collapsed Gibbs sampling to approximate parameters in the model. For a record $\boldsymbol{d}_i$ belonging to region r, category c, temporal Gaussian component z, and user u, its collapsed probability density can be represented as follows:

$$\begin{aligned} & p(\boldsymbol{d}_i|\boldsymbol{d}_{-i}, r, c, z, u) \\ = & \underbrace{\iint p(\boldsymbol{l}_i|\boldsymbol{\mu}_r, \Sigma_r)p(\boldsymbol{\mu}_r, \Sigma_r|\boldsymbol{d}_{-i})d\boldsymbol{\mu}_r d\Sigma_r}_{(1)} \underbrace{\int p(r|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{d}_{-i})d\boldsymbol{\theta}}_{(2)} \underbrace{\int p(c|\boldsymbol{\Phi}_{r,u})p(\boldsymbol{\Phi}_{r,u}|\boldsymbol{d}_{-i})d\boldsymbol{\Phi}_{r,u}}_{(3)} \\ & \underbrace{\iint p(t_i|\nu_{c,z}, \sigma_{c,z})p(\nu_{c,z}, \sigma_{c,z}|\boldsymbol{d}_{-i})d\nu_{c,z}d\sigma_{c,z}}_{(4)} \underbrace{\int p(z|\boldsymbol{\psi}_c)p(\boldsymbol{\psi}_c|\boldsymbol{d}_{-i})d\boldsymbol{\psi}_c}_{(5)}, \end{aligned} \tag{3}$$

where, we omit the hyper-parameters $\{\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}_0, \rho_0, \nu_0, \lambda_0, \epsilon_0, \tau_0, \alpha, \beta, \gamma\}$ for simplicity.

For term (1) in Equation (3), since we use the conjugate prior for parameters $\boldsymbol{\mu}_r$ and Σ_r, $(\boldsymbol{\mu}_r, \Sigma_r|\boldsymbol{d}_{-i})$ follows NIW distribution with parameters $(\boldsymbol{\mu}_r, \kappa_r, \boldsymbol{\Psi}_r, \rho_r)$, which can be calculated as follows:

$$\begin{cases} \kappa_r = \kappa_0 + n_r, \quad \boldsymbol{\mu}_r = \dfrac{\kappa_0\boldsymbol{\mu}_0 + n_r\bar{\boldsymbol{l}}_r}{\kappa_r}, \quad \rho_r = \rho_0 + n_r, \\ \boldsymbol{\Psi}_r = \boldsymbol{\Psi}_0 + \sum\limits_{r_j=r, j\neq i} (\boldsymbol{l}_j - \bar{\boldsymbol{l}}_r)(\boldsymbol{l}_j - \bar{\boldsymbol{l}}_r)^T + \dfrac{\kappa_0 n_r}{\kappa_r}(\bar{\boldsymbol{l}}_r - \boldsymbol{\mu}_0)(\bar{\boldsymbol{l}}_r - \boldsymbol{\mu}_0)^T; \end{cases} \tag{4}$$

ALGORITHM 1: Generative Process

```
Draw $\boldsymbol{\theta} \sim \text{Dirichlet}(\cdot|\alpha)$
for each region $r$ do
    Draw spatial distribution
    $(\boldsymbol{\mu}_r, \Sigma_r) \sim \text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}_0, \rho_0)$
    Draw $\boldsymbol{\Phi}_r \sim \text{Dirichlet}(\cdot|\beta)$
    for each user $u$ do
        Draw $\boldsymbol{\Phi}_{r,u} \sim \text{Dirichlet}(\cdot|b \cdot \boldsymbol{\Phi}_r)$
    end
end
for each category $c$ do
    Draw $\boldsymbol{\psi}_c \sim \text{Dirichlet}(\cdot|\gamma)$
    for each temporal Gaussian component $z$ do
        Draw temporal distribution
        $(\nu_{cz}, \sigma_{cz}) \sim \text{NIG}(\nu_0, \lambda_0, \epsilon_0, \tau_0)$
    end
end
for each record $d$ do
    Draw a region $r \sim \text{multi}(\boldsymbol{\theta})$
    Draw a location $\boldsymbol{l} \sim \mathcal{N}(\boldsymbol{\mu}_r, \Sigma_r)$
    Draw a category $c \sim \text{multi}(\boldsymbol{\Phi}_{r,u})$
    Draw a temporal Gaussian component $z \sim \text{multi}(\boldsymbol{\psi}_c)$
    Draw a time $t \sim \mathcal{N}(\nu_{c,z}, \sigma_{c,z})$
end
```

where n_r is the number of records belonging to region r and $\bar{\boldsymbol{l}}_r$ is the mean vector of these records, i.e., $\bar{\boldsymbol{l}}_r = \frac{1}{n_r} \sum_{r_j=r, j\neq i} \boldsymbol{l}_j$. Therefore, the first term of the collapsed probability density can be calculated as

$$\iint p(\boldsymbol{l}_i|\boldsymbol{\mu}_r,\Sigma_r)p(\boldsymbol{\mu}_r,\Sigma_r|\boldsymbol{d}_{-i})d\boldsymbol{\mu}_r d\Sigma_r = t_{\rho_r-1}\left(\boldsymbol{l}_i|\boldsymbol{\mu}_r,\frac{\Psi_r(\kappa_r+1)}{\kappa_r(\rho_r-1)}\right),$$

where $t_{\rho_r-1}(\cdot)$ is the probability density function of bivariate t-distribution.

Similarly, for term (4), $(\nu_{c,z},\sigma_{c,z}|\boldsymbol{d}_{-i})$ follows NIG distribution with parameters $(\nu_{c,z},\lambda_{c,z},\epsilon_{c,z},\tau_{c,z})$, which can be calculated as follows:

$$\begin{cases} \lambda_{c,z}=\lambda_0+n_{c,z},\ \nu_{c,z}=\dfrac{\lambda_0\nu_0+n_{c,z}\bar{t}_{c,z}}{\lambda_{c,z}},\ \tau_{c,z}=\tau_0+\dfrac{n_{c,z}}{2}, \\ \epsilon_{c,z}=\epsilon_0+\displaystyle\sum_{c_j=c,z_j=z,j\neq i}\frac{(t_j-\bar{t}_{c,z})^2}{2}+\frac{\lambda_0 n_{c,z}}{2\lambda_{c,z}}(\bar{t}_{c,z}-\nu_0)^2, \end{cases} \tag{5}$$

where $n_{c,z}$ is the number of records belonging to category c and temporal Gaussian component z; $\bar{t}_{c,z}$ is the mean value of these records, i.e., $\bar{t}_{c,z} = \frac{1}{n_{c,z}}\sum_{c_j=c,z_j=z,j\neq i} t_j$. Therefore, term (4) can be calculated as

$$\iint p(t_i|\nu_{c,z},\sigma_{c,z})p(\nu_{c,z},\sigma_{c,z}|\boldsymbol{d}_{-i})d\nu_{c,z}d\sigma_{c,z}=t_{2\tau_{c,z}}\left(t_i|\nu_{c,z},\frac{\epsilon_{c,z}(\lambda_{c,z}+1)}{\lambda_{c,z}\tau_{c,z}}\right),,$$

where $t_{\tau_{c,z}}(\cdot)$ is the probability density function of t-distribution.

For term (2), $\boldsymbol{\theta}|\boldsymbol{d}_{-i}$ follows Dirichlet distribution and we have

$$\int p(r|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{d}_{-i})d\boldsymbol{\theta} = \theta_r = \frac{n_r+\alpha}{\sum_{r'=1}^{|R|}(n_{r'}+\alpha)}. \tag{6}$$

For term (5), $\boldsymbol{\psi}_c|\boldsymbol{d}_{-i}$ follows Dirichlet distribution and we have

$$\int p(z|\boldsymbol{\psi}_c)p(\boldsymbol{\psi}_c|\boldsymbol{d}_{-i})d\boldsymbol{\psi}_c = \psi_{c,z} = \frac{n_{c,z}+\gamma}{\sum_{z'=1}^{2}(n_{c,z'}+\gamma)}. \tag{7}$$

As for term (3), we first replace $\boldsymbol{\Phi}_{r,u}$ with $\boldsymbol{\Phi}_r$. $\boldsymbol{\Phi}_r|\boldsymbol{d}_{-i}$ follows Dirichlet distribution and we have

$$\int p(c|\boldsymbol{\Phi}_{r,u})p(\boldsymbol{\Phi}_{r,u}|\boldsymbol{d}_{-i})d\boldsymbol{\Phi}_{r,u} = \Phi_{r,c} = \frac{n_{r,c}+\beta}{\sum_{c'=1}^{|C|}(n_{r,c'}+\beta)}, \tag{8}$$

where $n_{r,c}$ is the number of records belonging to region r and category c.

After that, we consider each user's personal preference and $\boldsymbol{\Phi}_{r,u}|\boldsymbol{d}_{-i}$ follows Dirichlet distribution with prior $b\cdot\boldsymbol{\Phi}_r$ and we have

$$\int p(c|\boldsymbol{\Phi}_{r,u})p(\boldsymbol{\Phi}_{r,u}|\boldsymbol{d}_{-i})d\boldsymbol{\Phi}_{r,u}=\Phi_{r,u,c}=\frac{n_{r,u,c}+b\Phi_{r,c}}{\sum\limits_{c'=1}^{|C|}(n_{r,u,c'}+b\Phi_{r,c'})}, \tag{9}$$

where $n_{r,u,c}$ is the number of records belonging to region r, user u, and category c.

Then, for every record $\boldsymbol{d}_i$, we can simultaneously sample region r_i, category c_i, and temporal Gaussian component z_i according to the following posterior probability:

$$\begin{aligned} p(r_i=r,c_i=c,z_i=z|\boldsymbol{d}_{-i}) &\propto t_{\rho_r-1}\left(\boldsymbol{l}_i|\boldsymbol{\mu}_r,\frac{\Psi_r(\kappa_r+1)}{\kappa_r(\rho_r-1)}\right) \\ &\cdot\theta_r\cdot\Phi_{r,u,c}\cdot t_{2\tau_{c,z}}\left(t_i|\nu_{c,z},\frac{\epsilon_{c,z}(\lambda_{c,z}+1)}{\lambda_{c,z}\tau_{c,z}}\right)\cdot\psi_{c,z}. \end{aligned} \tag{10}$$

However, if we directly sample r_i, c_i, z_i based on Equation (10), the computational complexity will be $O(M|R||C||\mathcal{D}|)$, which is tough. To reduce the computation cost, we find that when $\boldsymbol{d}_i$ is labeled, i.e., category c_i is fixed, r_i and z_i will become independent. Therefore, if we sample r_i

and z_i separately, the computational complexity can be reduced to $O(M|R||\mathcal{D}|)$. Their posterior probability distributions are as follows:

$$\begin{aligned} p(r_i = r|\boldsymbol{d}_{-i}) &\propto t_{\rho_r - 1}\left(\boldsymbol{l}_i|\boldsymbol{\mu}_r, \frac{\boldsymbol{\Psi}_r(\kappa_r + 1)}{\kappa_r(\rho_r - 1)}\right)\theta_r\Phi_{r,c}, \\ p(z_i = z|\boldsymbol{d}_{-i}) &\propto t_{2\tau_{c,z}}\left(t_i|\nu_{c,z}, \frac{\epsilon_{c,z}(\lambda_{c,z} + 1)}{\lambda_{c,z}\tau_{c,z}}\right)\psi_{c,z}. \end{aligned} \tag{11}$$

The detailed process of collapsed Gibbs sampling is shown in Algorithm 2. It takes the number of iterations M, region number $|R|$, category number $|C|$, human mobility records $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U^u$, and hyper-parameters as the input. First, it randomly initializes r_i, z_i for all records and c_i for unlabeled records. Then, it iterates sampling them for each labeled records $\boldsymbol{d}_i \in \mathcal{D}_L$. The statistics are updated after each time of sampling. After M iterations, we use crowd preference $\boldsymbol{\Phi}_r$ as a prior of user preference $\boldsymbol{\Phi}_{r,u}$ and further learn $\boldsymbol{\Phi}_{r,u}$ using unlabeled records of user u. Similarly, it iterates sampling r_i, z_i, c_i for each unlabeled records $\boldsymbol{d}_i \in \mathcal{D}_U^u$ and updates the statistics after each time of sampling. After M iterations, this algorithm outputs the statistics $\{\boldsymbol{\mu}_r,\kappa_r,\boldsymbol{\Psi}_r,\rho_r,\nu_{c,z},\lambda_{c,z},\epsilon_{c,z},\tau_{c,z},\theta_r,\psi_{c,z},\Phi_{r,u,c}\}$ as the final results.

The computational complexity of collapsed Gibbs sampling shown in Algorithm 2 is $O(M(|R||C||\mathcal{D}_U^u| + |R||\mathcal{D}_L|))$. Since the number of all labeled records is much more than unlabeled records of user u, i.e., $|\mathcal{D}_L| \gg |\mathcal{D}_U^u|$, the complexity can be approximated as $O(M|R||\mathcal{D}|)$. Thus, the computation cost roughly grows linearly with the number of iterations, number of regions, and number of records, which is feasible in practice.

3.3 POI Annotation

Given a record $\boldsymbol{d}_i$ with geographical location $\boldsymbol{l}_i$, time t_i, user u_i, and the set of candidate POI $\mathcal{P}_{\boldsymbol{l}_i}$, we aim at finding the real POI $p_i \in \mathcal{P}_{\boldsymbol{l}_i}$ the user visits. To achieve this goal, we calculate the probability of user visiting each candidate POI, i.e., $p(p_i = p|\boldsymbol{l}_i, t_i)$ and choose the POI with maximum probability. We first calculate the probability that record $\boldsymbol{d}_i$ appears in region r as $p(r_i = r) \propto \theta_r\mathcal{N}(\boldsymbol{l_i}|\boldsymbol{\mu}_r, \Sigma_r)$. Then, the probability of each candidate POI $p \in \mathcal{P}_{\boldsymbol{l}_i}$, whose category is c, is given by

$$\begin{aligned} p(p_i = p|\boldsymbol{l}_i, t_i) \propto &\left(\sum_{r=1}^{|R|} p(r_i = r)\Phi_{r,u,c}\right), \\ &\left(\sum_{z=1}^{2} \boldsymbol{\psi}_{c,z}\mathcal{N}(t_i|\nu_{c,z}, \sigma_{c,z})\right)\exp\left(-\frac{dist(p.\boldsymbol{l}, \boldsymbol{l}_i)^2}{\zeta^2}\right), \end{aligned} \tag{12}$$

where ζ is a hyper-parameter and the last term models distance decay.

To sum up, for semantic annotation of mobility records, we develop a Bayesian mixture model based on five domains of context—location, time, POI category, personal preference, and crowd preference. We further adopt collapsed Gibbs sampling to approximate parameters using both labeled and unlabeled records. Finally, we annotate each record with POI that has maximum probability based on Equation (12).

4 PERFORMANCE EVALUATION

4.1 Datasets

Our experiments are conducted on two human mobility datasets provided by WeChat, the most popular mobile instant messenger in China. They were collected between April 17–May 17, 2018, and cover the entire metropolitan area of Beijing and Guangzhou, which are two of the largest cities in China. The datasets contain both labeled and unlabeled human mobility records: labeled records are posted by users when they actively share POIs with friends and unlabeled records are collected by WeChat when users use other location services. The labeled records are check-ins containing anonymized user ID, GPS coordinates, time, POI, and category, while the unlabeled records do not contain POI and category. Since a visit may generate more than one unlabeled record, we combine them through stopping point detection methods [51]. On both datasets, we focus on users with more than two check-ins as our target users. Figures 3(a) and (b) show the distribution of the number of labeled and unlabeled records for each user. On the Beijing dataset, 80% of the users have less than three labeled records and 63 unlabeled records; on the Guangzhou dataset, 80% of the users have less than four labeled records and 79 unlabeled records.

Our POI dataset is also provided by WeChat. We only consider POI with at least one check-in record during April 17–May 17, 2018. In Guangzhou and Beijing, the average number of candidate POIs for each annotation is 19.4 and 8.1, respectively. For each POI, we obtain two levels of categories. The lower level has 16 coarse-grained categories, while the higher level has 54 fine-grained categories. For example, the category "restaurant" in the lower level is subdivided into seven categories in the higher level, including "Chinese restaurant", "western restaurant", and so on. We use

ALGORITHM 2: Collapsed Gibbs Sampling

```
Input:$M, |R|, |C|, \mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U^u, \boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}_0, \rho_0, \nu_0,$
$\lambda_0, \epsilon_0, \tau_0, \alpha, \beta, \gamma, b$.
Output:$\boldsymbol{\mu}_r,\kappa_r,\boldsymbol{\Psi}_r,\rho_r,\nu_{c,z},\lambda_{c,z},\epsilon_{c,z},\tau_{c,z},\theta_r,\psi_{c,z},\Phi_{r,u,c}$.
Initialize: Randomly initialize $r_i, c_i, z_i$.
for $iter \in \{1, \dots, M\}$ do
    for $\boldsymbol{d}_i \in \mathcal{D}_L$ do
        Remove $\boldsymbol{d}_i$ from $r_i, z_i$;
        Update $\boldsymbol{\mu}_{r_i}, \kappa_{r_i}, \boldsymbol{\Psi}_{r_i}, \rho_{r_i}, \nu_{c_i,z_i}, \lambda_{c_i,z_i}, \epsilon_{c_i,z_i},$
        $\tau_{c_i,z_i}, \theta_{r_i}, \psi_{c_i,z_i}, \Phi_{r_i,c_i}$ based on (4)~(8);
        for $r \in \{1, \dots, |R|\}, z \in \{1, 2\}$ do
            Calculate $p(r_i = r, z_i = z|\boldsymbol{d}_{-i})$ based on (11);
        end
        with probability $p(r_i = r, z_i = z|\boldsymbol{d}_{-i})$do
            $r_i \leftarrow r, z_i \leftarrow z$;
        end
    end
end
for $iter \in \{1, \dots, M\}$ do
    for $\boldsymbol{d}_i \in \mathcal{D}_U^u$ do
        Remove $\boldsymbol{d}_i$ from $r_i, c_i, z_i$;
        Update $\boldsymbol{\mu}_{r_i}, \kappa_{r_i}, \boldsymbol{\Psi}_{r_i}, \rho_{r_i}, \nu_{c_i,z_i}, \lambda_{c_i,z_i}, \epsilon_{c_i,z_i},$
        $\tau_{c_i,z_i}, \theta_{r_i}, \psi_{c_i,z_i}, \Phi_{r_i,u,c_i}$ based on (4)~(9);
        for$r \in \{1, \dots, |R|\}, c \in \{1, \dots, |C|\}, z \in \{1, 2\}$ do
            Calculate $p(r_i = r, c_i = c, z_i = z|\boldsymbol{d}_{-i})$ based on (10);
        end
        with probability$p(r_i = r, c_i = c, z_i = z|\boldsymbol{d}_{-i})$do
            $r_i \leftarrow r, c_i \leftarrow c, z_i \leftarrow z$;
        end
    end
end
```

Table 2. Statistics of Two Datasets

Dataset	Guangzhou	Beijing
# users	8,493	4,465
# target users	427	390
# labeled records	12,953	6,397
# unlabeled records	59,172	20,752
# POIs	5,773	3,506

both levels of the POI category to evaluate our model. Table 2 summarizes the statistics of two datasets.

4.2 Baselines and Metrics

We denote our proposed model as **CAP**, and compare it with several state-of-the-art POI annotation algorithms as follows:

4.2.1 Distance-Based Method (DIST). This method only considers the geographic distance between user and candidates POIs. It simply annotates the geographically closest POI to each record.

4.2.2 Bayesian Activity Inference Model (Bayes). Gong et al. [7] propose this method to infer the trip purpose of taxi passengers by taking both location and time into consideration. It formulates the visit probability function of a candidate POI $p \in \mathcal{P}_l$, based on Bayes' rules. The function is given by

$$p(p|\boldsymbol{l}, t) \propto A_p \cdot dist(p.\boldsymbol{l}, \boldsymbol{l})^{-\lambda} \cdot pop(p|t). \tag{13}$$

A_p is the attractiveness of p, which is not available in our datasets, so we omit this term. $pop(p|t)$ is the popularity of p at time t.

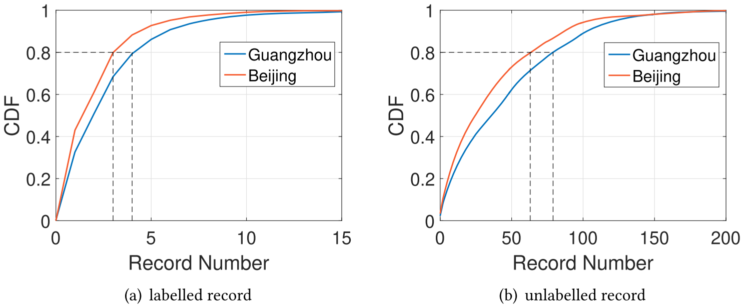
4.2.3 Visit Probability Propagating Model (VPPM). Wang et al. [32] model the probability of POIs visited by passengers as the following parametric function:

$$p(p|\boldsymbol{l}) = \frac{\beta_1}{\beta_2} dist(p.\boldsymbol{l}, \boldsymbol{l}) \cdot \exp(1 - dist(p.\boldsymbol{l}, \boldsymbol{l})/\beta_2). \tag{14}$$

The probability distribution achieves the maximum value with the distance of β_2 and has an exponential heavy tail, which is used to propagate the visit probability of passengers from the drop-off point to their target POIs.

4.2.4 Spatial-Temporal Regularity-Based Model (STR). Wu and Li [34] propose a Markov random field model by utilizing spatial and temporal regularity of human mobility. Based on the observation that a user tends to visit similar POIs (e.g., POIs with the same category) when observed at close locations or similar time of the day, the POIs visited by a user can be inferred through minimizing pairwise potentials between spatial-temporal related records. **Spatial-temporal Regularity** (**STR**) captures the personal preference of each user independently by using all the location records of each user, i.e., it does not take crowd preference into consideration.

4.2.5 Semi-Supervised Version of STR (STR+). The basic version of STR only utilizes time and location of human mobility records, ignoring the ground truth labels that are available in labeled records. The **semi-supervised version of STR** (**STR+**) further leverages those ground truth labels by setting the node potential of labeled records with a fixed value [34].



(a) labelled record

(b) unlabelled record

Fig. 3. Distribution of numbers of labeled and unlabeled records.

4.2.6 *Context-Aware Annotation Model (CA).* This model is a simplified version of our method, which does not consider personal preference and only use labeled records. In this model, we approximate parameters in the Bayesian mixture model using only labeled records and replace personal preference vector $\mathbf{\Phi}_{r,u}$ with a universal vector $\mathbf{\Phi}_r$ for all users.

4.2.7 *Supervised Version of CAP (CA+).* Compared with **Context-aware Annotation** (**CA**), this model further considers personal preference without using unlabeled records, which is a **supervised version of CAP** (**CA+**). It utilizes labeled records to approximate $\mathbf{\Phi}_{r,u}$ for each user in the Bayesian mixture model.

In order to measure the correctness of the annotation result, we use a widely-used metric: accuracy@k, which is defined as $\frac{N^k}{N}$, where N^k is the number of records whose associated POI is correctly predicted in the top-k results, and N is the total number of records. Note that STR only has accuracy@1 due to its model.

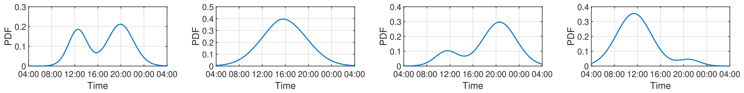
We also compare accuracy@1 of each method with DIST to calculate the relative improvement. Let N^1_{DIST} be the number of records being correctly annotated using DIST, we define the relative improvement as $\frac{N^1 - N^1_{DIST}}{N^1_{DIST}}$.

4.3 Parameter Settings

On both datasets, we use the last check-in of target users as the test set. All the other records are used as the training set. We set the parameters of our model as follows to achieve best performance: $\delta = 500\ m$, $\zeta = 100\ m$, $M = 20$, $\kappa_0 = 1$, $\rho_0 = 10$, $\lambda_0 = 1$, $\tau_0 = 2$, $\alpha = 10000$, $\beta = 10$, $\gamma = 10$, $b = 3$. $\boldsymbol{\mu}_0$, and $\mathbf{\Psi}_0$ are mean vector and covariance matrix of locations of all records. ν_0 and ϵ_0 are the mean and standard deviation of timestamps of all records. $|R|$ is set to be 225 and 125 for Guangzhou and Beijing, respectively. We further discuss these parameters in Parameter Study. As for the baselines, parameters are tuned to achieve the best performance regarding the accuracy@1 metric.

4.4 Experimental Results

4.4.1 *Case Study.* To evaluate the performance of our model, we first visualize the spatial and temporal pattern generated by our model based on the training set. We conduct a case study on the Beijing dataset. Figure 4 shows the temporal patterns of four most representative categories among 54 fine-grained categories. As we can observe, each category has a unique temporal pattern: fast food restaurants are popular during lunchtime and suppertime, which coincides with our prior knowledge. Most people visit shopping malls from 10AM to 10PM, corresponding with the malls' opening time. Obviously, places of entertainment are more popular in the evening, and people



(a) fast food restaurant (b) shopping mall (c) entertainment (d) hospital

Fig. 4. Temporal pattern of four categories.

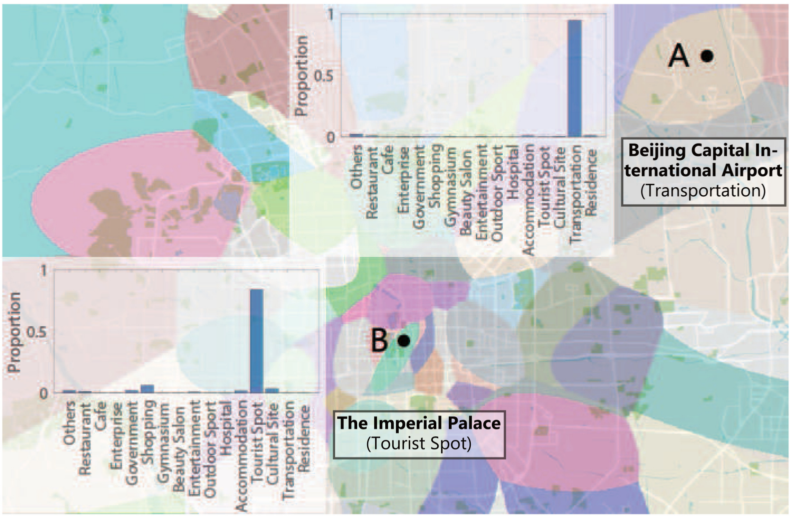


Fig. 5. Generated regions where A represents Beijing Capital International Airport Region and B is the Imperial Palace Region.

tend to see a doctor in the daytime. Our model learns those temporal patterns from the training set without any prior knowledge, which demonstrates its effectiveness.

Figure 5 visualizes the functional regions that our model generates. Basically, the function of these regions is consistent with our prior knowledge. To further demonstrate the effectiveness of our model, we select two typical regions and plot their category distribution in Figure 5. Region A represents Beijing Capital International Airport Region, and 94% mobility records in this region are related to transportation; Region B represents the Imperial Palace Region, and 84% mobility records there are related to tourist spots. Note that for brevity and readability, here, we present results for coarse-grained POI categories. Comparing with administrative regions that most map services use, the functional regions that our model learns and generates are more informative. This is because functional regions in a city are generally not consistent with administrative regions since their boundaries are ambiguous and evolving over time. By revealing distinct functions of different areas, we can better infer the POIs that users visit.

In summary, both the temporal and spatial observations demonstrate the strong ability of our model to capture human mobility patterns.

4.4.2 Performance Comparison. Table 3 shows the annotation performance in terms of accuracy@k with relative improvement. For methods other than DIST, we use both levels of the POI category and denote them as (16) and (54), respectively. As we can observe, our method outperforms all baselines on both datasets. Specifically, on the Guangzhou dataset, CAP surpasses DIST by 16.7%, with 49.6% relative improvement, and it performs better than VPPM, Bayes, STR, and

Table 3. Performance on Two Datasets

	Guangzhou				Beijing			
	Accuracy@k			Relative Improvement	Accuracy@k			Relative Improvement
	k=1	k=2	k=3		k=1	k=2	k=3	
DIST	33.7%	46.1%	55.7%	–	42.6%	58.7%	68.7%	–
VPPM	34.0%	46.4%	55.7%	0.9%	42.6%	58.7%	68.7%	0.0%
Bayes(16)	35.4%	49.2%	55.5%	5.0%	43.3%	61.0%	70.5%	1.6%
Bayes(54)	34.4%	48.0%	55.3%	2.1%	43.6%	57.9%	68.7%	2.3%
STR(16)	34.7%	–	–	3.0%	44.9%	–	–	5.4%
STR(54)	34.4%	–	–	2.1%	44.6%	–	–	4.7%
STR+(16)	37.9%	–	–	12.5%	47.4%	–	–	11.3%
STR+(54)	38.6%	–	–	14.5%	46.7%	–	–	9.6%
CA(16)	43.3%	57.8%	64.9%	28.5%	48.2%	64.9%	72.6%	13.1%
CA(54)	44.0%	57.6%	65.1%	30.6%	47.2%	64.4%	70.3%	10.8%
CA+(16)	47.1%	62.8%	69.1%	39.8%	49.0%	67.7%	73.1%	15.0%
CA+(54)	49.4%	64.9%	69.1%	46.6%	50.0%	56.7%	74.1%	17.4%
CAP(16)	48.5%	63.0%	69.8%	43.9%	51.8%	69.0%	74.6%	21.6%
CAP(54)	**50.4%**	**65.3%**	**70.0%**	**49.6%**	**52.1%**	**69.7%**	**76.4%**	**22.3%**

STR+ by 16.4%, 15.0%, 16.4%, and 11.8%, respectively. Similarly, on the Beijing dataset, CAP outperforms DIST, VPPM, Bayes, STR, and STR+ by 9.5%, 9.5%, 8.5%, 9.0%, and 3.9%, respectively.

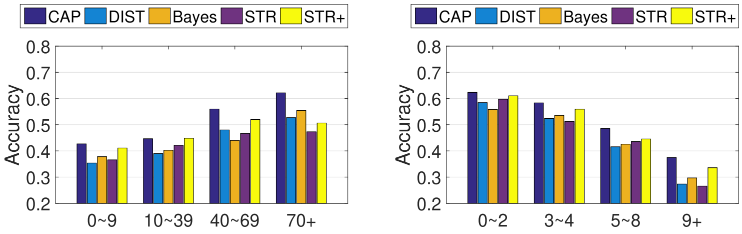
Note that our method achieves a very large improvement on the Guangzhou dataset. We believe that this performance gain is due to the fact that Guangzhou has a higher POI density than Beijing, i.e., there are more candidate POIs for each annotation, which makes the POI annotation problem more challenging. In addition, we can observe that VPPM outperforms DIST on the Guangzhou dataset, while they achieve similar performance on the Beijing dataset, which also indicates that Guangzhou has a higher POI density than Beijing. Since with a larger distance to POIs, the probability distribution (14) degrades into being inversely proportional to the distance, similar to the DIST algorithm. Thus, they achieve similar performance on the Beijing dataset.

Further, in Figure 6(b), we rank the records based on the number of candidate POIs and divide them into four groups. We can observe that as the number of candidate POI increases, the advantage of our model is more evident. For records with less than three candidate POIs, our method surpasses the baseline with the highest accuracy by only 1.3%, while for records with more than nine candidate POIs, the performance gain of our method increases to 3.9%.

We also rank the users based on the number of mobility records and divide them into four groups. As shown in Figure 6(a), our method achieves greater performance gain for users with more mobility records. For users with less than 10 records, the performance gain of our method is 1.6%, while for users with 70+ records, it comes to 6.8%.

In addition, as we can observe from Table 3, the performances are different when using different levels of category. While the coarse-grained categories cannot fully leverage information of all categories, it alleviates data skew and avoids overfitting. In comparison, the fine-grained POI categories illustrate mobility patterns more detailedly, while suffering from difficulties in training with limited data.

4.4.3 Parameter Study. In our model, region number $|R|$ is an important hyper-parameter. For each dataset and each level of categories, we plot the accuracy as a function of $|R|$ in Figure 7. Generally, a larger region number leads to better performance, since it can better capture user preference in different areas of a city. However, due to data sparsity, an excessively large region



(a) record number of user

(b) number of candidate POI

Fig. 6. Performance vs. data quality.

number may reduce the number of records in each region and therefore cause overfitting. According to Figure 7, We set $|R|$ as 225 and 125 for Guangzhou and Beijing, respectively.

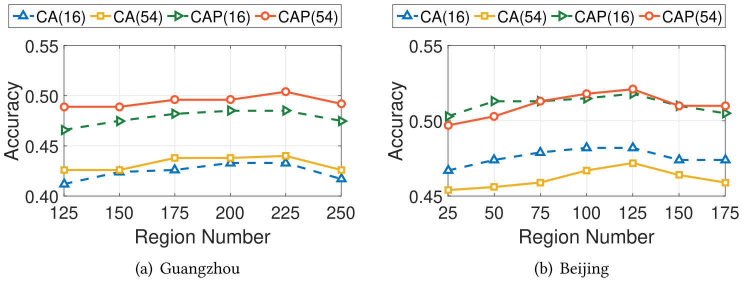
Next, we examine how the parameters of NIW and NIG prior influence the performance. Figures 8(a) and (b) show the impact of parameters κ_0 and ρ_0 in the prior of NIG distribution. We can observe that they achieve the best performance near our default settings. Figures 8(c) and (d) show the impact of parameters λ_0 and τ_0 in the prior of NIW distribution. Compared with κ_0 and ρ_0, they show less impact on the performance. Figures 8(e), (f), and (g) plot the impact of parameter α, β, γ in the **Dirichlet process** (**DP**). Since all these curves demonstrate a moderate variance, we conclude that our model is not sensitive to the hyper-parameters, indicating the robustness of our method. Figure 8(h) plots the impact of parameter b, which is used to adjust the influence of crowd preference, where $b = 0$ means only personal preference is considered in the model. By choosing a proper b value, we can balance personal preference and crowd preference so as to achieve the best performance.

4.4.4 Computational Time. Finally, we evaluate the performance of our proposed model in terms of computational time. Without loss of generality, we only show the performance of our proposed CAP algorithm compared with STR+, DIST, Bayes, and VPPM algorithms on the Beijing dataset. In addition, the computational time of all algorithms is evaluated on a 16-Core 2.10 GHz Linux server. We first show the average computational time for POI annotation of each user in Figure 9(a), where the region number is set to be 125. We can observe that DIST, Bayes, and VPPM have short computational time, since they annotate mobility records directly based on probability statistics of POI, where there are not many parameters to be estimated. Different from them, CAP and STR+ algorithms require relatively large computational time to estimate their parameters. Specifically, we can observe that the computational time of CAP is a little larger than that of STR+. Compared with the performance improvement, the additional computational time is acceptable. Further, we show the computational time as the function of the region number in Figure 9(b). We can observe an obvious linear relationship between the computational time and region number, which is consistent with our theoretical analysis in Section 3.2 and indicates that our proposed algorithm is feasible in practice.

5 RELATED WORK

5.1 Semantic Annotation of Mobility Records

Researchers have studied the semantic annotation problem using various methods [7, 8, 13, 34, 35, 37, 38]. Guc et al. [8] introduce a conceptual annotation model based on the notion of episodes. Gong et al.[7] formulate the probability that people visit certain POI based on Bayes' rule, in order to infer the trip purpose of taxi passengers. Yan et al. [37, 38] propose a **hidden Markov model** (**HMM**) to infer the POI based on the previous place the user visits. Lian et al. [13] leverage the



(a) Guangzhou

(b) Beijing

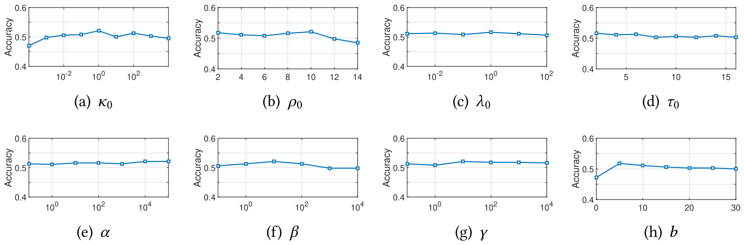
Fig. 7. Performance vs. region number.

learning to rank method by considering four features—user history, distance, review, and web popularity. Wu et al. [35] use **Kernel Density Estimation** (**KDE**) methods to annotate mobility data with dynamic semantics based on external contextual data. To achieve personalized annotation, Wu et al. [34] propose a Markov random field model by utilizing spatial and temporal regularity of human mobility. Zhang et al. [48] further leverage user grouping to annotate mobility records with POIs. Different from these methods, our work is the first to consider location, time, category, user preference, and similarity among users simultaneously using the probabilistic graphical model.

5.2 Mobility Modeling

Mobility modeling is a classical task in urban computing [1, 4, 5, 21, 36]. A number of approaches model users' mobility behavior by using the Markov-based model. Lu et al. [18] use the Markov model, where each state corresponds to a location. Mathew et al. [19] use a variant of the Markov model, i.e., HMM, where each hidden state corresponds to a probabilistic distribution over locations. Zhang et al. [47] further enhance HMM by using a user grouping scheme. In addition, some approaches use the DP to model user mobility. Jeong et al. [10] use DP to cluster users based on their mobility patterns. McInerney et al. [20] use DP to share parameters describing the periodicity of mobility among the users. Wang et al. [25, 27] use DP to model users' app usage and mobility simultaneously. Wang et al. [31] develop a probabilistic model to infer the purposes of taxi passengers. Zheng et al. [49] develop a Bayesian-based model to cluster users based on their characteristics behavior patterns. Further, a number of studies consider context information, such as social media or social networks, in mobility modeling. Specifically, Zhang et al. [46] leverage the information from the geo-tagged tweet. Wang et al. [23] use the information of social networks. Huai et al. [9] propose a sequential model by extending the Bayesian hidden Markov model for modeling the personalized contextual information of mobile users. Wang et al. [26] integrate the neural attention mechanism into the extended Markov model to predict future movements of users. Wang et al. [29] investigate the mobile user profiling problem based on POI check-in data, and they propose an adversarial substructured learning method to solve this problem. Liu et al. [17] predict the bus travel demand based on mobility patterns obtained from the location traces of taxicabs and the mobility records in bus transactions, which is used for bus routing optimization. Wang et al. [30] propose a collective embedding framework to learn the community structure based on periodic spatial-temporal mobility graphs of humans. Fu et al. [6] focus on identifying and quantifying the urban forms of residential communities, and they propose a collective learning approach based on individual-level human mobility data to solve this problem. Yao et al. [39] focus on utilizing the "co-occurrence" of origin–destination zones in mobility trajectories to learn their embeddings.

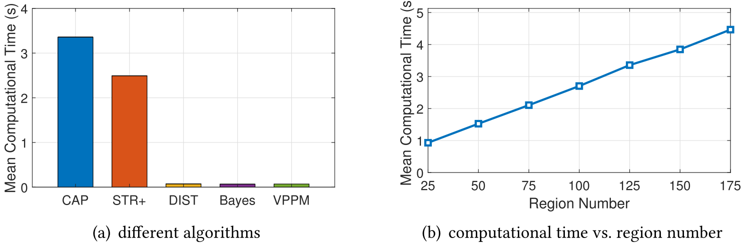
Wang et al. [28] consider the continuity of human mobility in terms of temporal dimension by utilizing von Mises distribution in the Bayesian mixture model. However, this model mainly



(a) κ_0 (b) ρ_0 (c) λ_0 (d) τ_0

(e) α (f) β (g) γ (h) b

Fig. 8. Performance vs. parameters.



(a) different algorithms (b) computational time vs. region number

Fig. 9. Performance in terms of computational time.

considers the spatial and temporal mobility patterns of users, and factors including the POI category and the temporal patterns of different POI categories are not considered, which is different from our proposed CAP algorithm. Different with the goal of predicting users' future movement in these approaches, we seek to annotate unlabeled mobility records with correct POIs, i.e., extracting the semantic information of these mobility records.

5.3 Location Recommendation

Location context plays an important role in location-based services. Ye et al. [40] consider the social and geographical characteristics of users and locations in their location recommendation system. Zheng et al. [50] recommend locations to users based on their accurate locations and comments. Sun et al. [22] focus on inferring users' intent from app usage behavior and other contextual information to provide personalized recommendations. Yuan et al. [43] recommend POIs to users by using a collaborative filtering method and incorporating temporal information. Liu et al. [15] propose a geographical probabilistic factor analysis framework for POI recommendations, which strategically takes multiple factors, which influence the user check-in decision process, into consideration. Wang et al. [33] recommend spatial items by modeling and fusing the sequential influence, cyclic patterns, and personal interests. A nonparametric Bayesian approach is further developed in [44] to capture spatio-temporal contextual information and users' topical interests and intentions for accurate recommendation and retrieval. These approaches seek to find out new locations or POIs that users will visit in the future. Different from them, our goal is to annotate unlabeled historical user trajectories with POIs.

6 CONCLUSIONS

In this article, we propose a context-aware personalized semantic annotation model. We use a Bayesian mixture model to capture human mobility patterns by considering five domains of information simultaneously—location, time, POI category, personal preference, and crowd preference. Extensive results on two real-world datasets demonstrate that our method significantly outperforms the state-of-the-art algorithms by over 11.8%. In the future, we will take periodic mobility patterns into consideration so as to develop better models for semantic annotation of mobility records. Also, we will further model temporal patterns of POI in a similar way as user preference, i.e., modeling each POI separately using category-level distribution as prior. Since the results of CAP can facilitate further applications, including POI recommendation, population estimation, and public security, we will study related problems based on our proposed method in future work.

REFERENCES

[1] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: A comparison across pois, websites, and smartphone apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–24.

[2] Hancheng Cao, Jie Feng, Yong Li, and Vassilis Kostakos. 2018. Uniqueness in the city: Urban morphology and location privacy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–20.

[3] Hancheng Cao, Jagan Sankaranarayanan, Jie Feng, Yong Li, and Hanan Samet. 2019. Understanding metropolitan crowd mobility via mobile cellular accessing data. *ACM Transactions on Spatial Algorithms and Systems* 5, 2 (2019), 1–18.

[4] Hancheng Cao, Fengli Xu, Jagan Sankaranarayanan, Yong Li, and Hanan Samet. 2020. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. *IEEE Transactions on Mobile Computing* 19, 5 (2020), 1096–1108.

[5] Zhilong Chen, Hancheng Cao, Huangdong Wang, Fengli Xu, Vassilis Kostakos, and Yong Li. 2020. Will you come back/check-in again? Understanding characteristics leading to urban revisitation and re-check-in. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[6] Yanjie Fu, Guannan Liu, Yong Ge, Pengyang Wang, Hengshu Zhu, Chunxiao Li, and Hui Xiong. 2018. Representing urban forms: A collective learning model with heterogeneous human mobility data. *IEEE Transactions on Knowledge and Data Engineering* 31, 3 (2018), 535–548.

[7] Li Gong, Xi Liu, Lun Wu, and Yu Liu. 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science* 43, 2 (2016), 103–114.

[8] Baris Guc, Michael May, Yucel Saygin, and Christine Kopp. 2008. Semantic annotation of GPS trajectories. *11th AGILE International Conference on Geographic Information Science* 38, 6 (2008), 1–9.

[9] Baoxing Huai, Enhong Chen, Hengshu Zhu, Hui Xiong, Tengfei Bao, Qi Liu, and Jilei Tian. 2014. Toward personalized context recognition for mobile users: A semisupervised bayesian HMM approach. *ACM Transsactions on Knowledge Discovery from Data* 9, 2 (2014), 1–29.

[10] J. Jeong, M. Leconte, and A. Proutiere. 2016. Cluster-aided mobility predictions. In *Proceedings of the 35th Annual IEEE International Conference on Computer Communications.*

[11] Levente Juhąsz and Hartwig Hochmair. 2020. Studying spatial and temporal visitation patterns of points of interest using safegraph data in florida. *GI Forum* 2020 (2020), 119–136.

[12] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[13] Defu Lian and Xing Xie. 2011. Learning location naming from user check-in histories. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.*

[14] Zongyu Lin, Shiqing Lyu, Hancheng Cao, Fengli Xu, Yuqiong Wei, Hanan Samet, and Yong Li. 2020. HealthWalks: Sensing fine-grained individual health condition via mobility data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–26.

[15] Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. 2013. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[16] Yanchi Liu, Chuanren Liu, Xinjiang Lu, Mingfei Teng, Hengshu Zhu, and Hui Xiong. 2017. Point-of-interest demand modeling with human mobility patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[17] Yanchi Liu, Chuanren Liu, Nicholas Jing Yuan, Lian Duan, Yanjie Fu, Hui Xiong, Songhua Xu, and Junjie Wu. 2014. Exploiting heterogeneous human mobility patterns for intelligent bus routing. In *Proceedings of the IEEE International Conference on Data Mining.*

[18] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. *Scientific Reports* 3, 1 (2013), 1–9.

[19] Wesley Mathew, Ruben Raposo, and Bruno Martins. 2012. Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*.

[20] James McInerney, Jiangchuan Zheng, Alex Rogers, and Nicholas R. Jennings. 2013. Modelling heterogeneous location habits in human populations for location prediction under data sparsity. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[21] Hongzhi Shi, Hancheng Cao, Xiangxin Zhou, Yong Li, Chao Zhang, Vassilis Kostakos, Funing Sun, and Fanchao Meng. 2019. Semantics-aware hidden markov model for human mobility. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 774–782.

[22] Yu Sun, Nicholas Jing Yuan, Yingzi Wang, Xing Xie, Kieran McDonald, and Rui Zhang. 2016. Contextual intent tracking for personal assistants. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[23] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[24] Hao Wang, Yanmei Fu, Qinyong Wang, Hongzhi Yin, Changying Du, and Hui Xiong. 2017. A location-sentiment-aware recommender system for both home-town and out-of-town users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[25] Huandong Wang, Yong Li, Mu Du, Zhenhui Li, and Depeng Jin. 2021. App2Vec: Context-aware application usage prediction. *ACM Transactions on Knowledge Discovery from Data* 15, 6 (2021), 1–21.

[26] Huandong Wang, Yong Li, Depeng Jin, and Zhu Han. 2021. Attentional markov model for human mobility prediction. *IEEE Journal on Selected Areas in Communications* 39, 7 (2021), 2213–2225.

[27] Huandong Wang, Yong Li, Sihan Zeng, Gang Wang, Pengyu Zhang, Pan Hui, and Depeng Jin. 2019. Modeling spatio-temporal app usage for a large user population. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–23.

[28] Huandong Wang, Sihan Zeng, Yong Li, Pengyu Zhang, and Depeng Jin. 2020. Human mobility prediction using sparse trajectory data. *IEEE Transactions on Vehicular Technology* 69, 9 (2020), 10155–10166.

[29] Pengyang Wang, Yanjie Fu, Hui Xiong, and Xiaolin Li. 2019. Adversarial substructured representation learning for mobile user profiling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[30] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Xiaolin Li, and Dan Lin. 2018. Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology* 9, 6 (2018), 1–28.

[31] Pengfei Wang, Guannan Liu, Yanjie Fu, Yuanchun Zhou, and Jianhui Li. 2018. Spotting trip purposes from taxi trajectories: A general probabilistic model. *ACM Transactions on Intelligent Systems and Technology* 9, 3 (2018), 1–26.

[32] Pengyang Wang, Jiawei Zhang, Guannan Liu, Yanjie Fu, and Charu Aggarwal. 2018. Ensemble-spotting: Ranking urban vibrancy via POI embedding with multi-view spatial graphs. In *Proceedings of the SIAM International Conference on Data Mining*.

[33] Weiqing Wang, Hongzhi Yin, Xingzhong Du, Quoc Viet Hung Nguyen, and Xiaofang Zhou. 2018. TPM: A temporal personalized model for spatial item recommendation. *ACM Transactions on Intelligent Systems and Technology* 9, 6 (2018), 1–25.

[34] Fei Wu and Zhenhui Li. 2016. Where did you go: Personalized annotation of mobility records. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*.

[35] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. 2015. Semantic annotation of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*.

[36] Fengli Xu, Tong Xia, Hancheng Cao, Yong Li, Funing Sun, and Fanchao Meng. 2018. Detecting popular temporal modes in population-scale unlabelled trajectory data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.

[37] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2011. SeMiTri: A framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology*.

[38] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2013. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology* 4, 3 (2013), 1–38.

[39] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. 2018. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.

[40] Mao Ye, Peifeng Yin, and Wang-Chien Lee. 2010. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

[41] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 186–194.

[42] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2014. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2014), 712–725.

[43] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[44] Quan Yuan, Gao Cong, Kaiqi Zhao, Zongyang Ma, and Aixin Sun. 2015. Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users. *ACM Transactions on Information Systems* 33, 1 (2015), 1–33.

[45] Quan Yuan, Wei Zhang, Chao Zhang, Xinhe Geng, Gao Cong, and Jiawei Han. 2017. PRED: Periodic region detection for mobility modeling of social media users. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*.

[46] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. 2017. TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[47] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han. 2016. GMove: Group-level mobility modeling using geo-tagged social media. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[48] Yong Zhang, Hua Wei, Xuelian Lin, Fei Wu, Zhenhui Li, Kaiheng Chen, Yuandong Wang, and Jie Xu. 2018. Context-aware location annotation on mobility records through user grouping. In *Advances in Knowledge Discovery and Data Mining*. D. Phung, V. Tseng, G. Webb, B. Ho, M. Ganji, and L. Rashidi (Eds.), Lecture Notes in Computer Science, Vol. 10939, Springer, Cham.

[49] Jiangchuan Zheng and Lionel M. Ni. 2012. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*.

[50] Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th International Conference on World Wide Web.*

[51] Yu Zheng and Xing Xie. 2011. Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology* 2, 1 (2011), 1–29.

Received March 2020; revised March 2021; accepted July 2021