Tensor Factorization Based POI Category Inference

Yunyu He[1], Hongwei Peng[1], Yuanyuan Jin[1],
Jiangtao Wang[1](⋆✉), and Patrick C. K. Hung[2]

[1]National Trusted Embedded Software Engineering Technology Research Center (NTESEC), East China Normal University, Shanghai, China
[2]Faculty of Business and Information Technology, University of Ontario Institute of Technology(UOIT),Canada
yyhe@stu.ecnu.edu.cn,penghongwei_phw@163.com,yyj@stu.ecnu.edu.cn
jtwang@sei.ecnu.edu.cn,patrick.hung@uoit.ca,

Abstract. Trajectory data is an important kind of data with different aspects of the user information like demographics, user behavior and activities. Therefore, it is significant and essential to infer point-of-interests (POI) categories from trajectory data for user modeling and user preferences mining in many location-based services (LBS). Recent researches focus more on recommendation and prediction of next POI, which are based on the check-in data. Check-in data is only a partial aspect of the user's behavior which collected by a certain LBS, while trajectory data describes the user from all around, which can help modeling user's interest preferences in a great degree. However, due to a deviation between the GPS-coordinate and the actually visited location, it is significant to infer the ultimate POI categories people accessed from trajectory data instead of mapping location coordinates to POIs directly. In this paper, we propose a collaborative inferring framework to analyze the actually visited POI categories from users' historical trajectory data. Through modeling relationships among the user, time and POI category, the tensor decomposition method can effectively complement the missing data and provides accurate predictions when user trajectory data is absent. Extensive experiments have been conducted with various state-of-the-art baseline on real-world trajectory data, and experiment results have demonstrated the promising performance in this framework.

1 Introduction

With the increasing popularity of GPS equipped mobile devices and vehicles, geographical records have become prevalent on the web. These geographical records not only reflect the user's mobility patterns, but also provide some help for modeling the user's interest preferences through the point-of-interests (POI) user visited. Besides, it is significant to mine users' preferences [17] [9] and establishing user profile from user's geo-spatial data for various location-based

⋆ ✉is the corresponding author

services (LBS), such as personalized advertising services, urban planning and location-based recommendation services. Some recent researches focus on recommendation POI [10] [18], which also consider the user's interest preferences. But these all based on the check-in data that only relates to a certain LBS, such as Facebook, Twitter, Foursquare, etc. Only when using these platforms to share their experience and check-in information associated with a POI, the location data will be collected by these platforms. They can only model the user's interest preferences and make recommendation based on the certain platform, but not suitable for modeling user profile. Compared with the check-in data, the trajectory data can track the user's behavior better and record the user's access information more completely. It could be concluded that the trajectory data has a greater value to estimate the POIs user has accessed, which is the key step of modeling user profile and can provide better help for follow-up work.

However, the number of POIs is substantially larger than the number of POI categories in the whole city map. Compared with POIs, the POI categories can perform the common characteristics of users' interest offering more help for user profile. Thus, the POI category are chosen in this work to represent user's interest preference, which can better reflect similar behavior patterns among users.

To this end, the crucial problem solved in this paper is how to infer the POI categories visited by each user from his or her trajectory data. By our best knowledge, it is a challenging research task due to following reasons:

First of all, there is usually a large deviation between the GPS-coordinate and the actual visited POI category. In our daily life, the location acquisition device is not always close to the user within a small range. For example, if a user wants to eat in a restaurant next to a supermarket parking lot he/she chooses to park the car, it is obvious that the supermarket parking lot is not really the POI category user has visited.

What's more, the POI categories user accessed may suffer from data sparsity for representing human mobility among a certain period, given that few people are willing to along with this device all the time or to use the device every day.

To cope with these challenges, some related methods that can be used to made some exploration. Recent studies have also found that human mobility follows a high degree of regularity over time [5]. Based on these observations, a collaborative method called tensor factorization is adopted, which can learn a universal model for solving this problem from users' trajectory data, instead of learning a separate model for each user isolated. Although, the tensor factorization considering the global factor collaboratively, it will ignore the specific geographical factors for each stopping point. Yi et al. [19] defined *negative-unlabeled* (NU) learning problem which can be used to take advantage of the specific geographical situations for each stopping point. However, this work conduct a matrix decomposition losing a part of latent factors of the tensor and reducing accuracy to some extent.

Considering the advantages and disadvantages of tensor factorization and negative-unlabeled learning problem, we propose a novel collaborative framework to solve the POI categories inferring problem. First, hidden relationships among

users, time slots and POI categories are integrated into a three-dimensional tensor $\mathcal{X}$ and conduct a tensor factorization method with Tikhonov Regularization. This tensor factorization method helps to overcome sparsity problem of data, meanwhile it complements the missing data. So that, the framework can make accurate predictions when user trajectory data is absent. Then, negative-unlabeled constraints are adopted to make use of the specific geographical situations for each staying point by normalizing the probability of the POI categories within the candidate set. At last, an efficient alternating minimization algorithm is employed to combine these two method and solve the problem.

To summarize, the major contributions of this paper are:

1. We develop a collaborative method under negative-unlabeled constraints to model relationships among user, time and POI category, which overcomes the data sparsity problem and infers POI categories accurately.
2. The proposed collaborative inferring framework can complement the missing data and make accurate predictions when user trajectory data is absent.
3. Based on two real-word dataset collected, the extensive experiments has been conducted to validate the effectiveness of the proposed approach. The results show that our approach vitally outperforms baseline models with a significant margin in several scenarios.

The rest of this paper is structured as following: Section 2 introduces related work. Data preprocessing and problem definition is presented in Section 3. The collaborative inferring framework under negative-unlabeled constraints will be proposed in Section 4. Experimental results based on a large-scale dataset are presented in Section 5. Finally, the conclusion of the paper will be drawn in Section 6 with a brief discussion of limitations and directions of future research.

2 Related Work

In order to introduce the related work of this paper in a more orderly way, this section is divided into two parts: tensor factorization and user profile on trajectory data.

2.1 Tensor Factorization

Comparing with many other collaborate methods, tensor has great advantages. As a generalization of matrices, tensor can model correlations among more than two dimensions while matrix factorization methods can only apply to two-order data. In recent years, tensor factorization has been widely applied in a variety of fields. Narita et al. [14] and Ge et al. [4] focus on how to utilize auxiliary information improve the quality of tensor decomposition. Zhong et al. [22] extracted feature from heterogeneous data set based on tensor factorization to infer user profiles. Yi et al. [19] focus on the similar problem with this paper which learned mobile users' location categories from highly inaccurate mobility data and proposed NUTF. NUTF treat this problem as a NU learning problem which

assigned the non-zero probabilities with any non-negative values that sum up to one, then optimized objective function by a low-rank tensor factorization. But there is a trade off between accuracy and efficiency of the algorithm. The NUTF decomposed the tensor by unfolding it into a two-dimension matrix and adopted a randomness matrix decomposition [6], which could reduce the complexity but loss a part of latent factors of the tensor as well as accuracy to some extent.

2.2 User Profile on Trajectory Data

Since former researches seldom focused the same task, we review three widely studied researches contributing to user profile on trajectory data :

User's important locations detection These researches identify important locations (e.g., home or working place) to understand users' behavior at these locations. Cao et al. [1] proposed a framework which can extract staying points from each users' GPS data and then clustering them to find significant semantic locations. An unsupervised collaborative approach is applied in Liu et al. [12] to identify home and working locations of individuals from geo-spatial trajectory data which also define the user-location signatures to describe users' behavior at the location.

Interesting regions discovery Zheng et al. [21] means the culturally important places and then models multiple individuals' location histories to mine interesting locations. Van Canh et al. [16] discovered regional communities by exploiting methods based on spatial latent Dirichlet allocation(SLDA). Yuan et al. [20] discovers regions of different functions using both human mobility and POIs located in a region. These works try to learn semantics of regions so that it can do help for user preference mining, as they are meaningful for finding user communities rather than modeling individuals' profile.

POI prediction Feng et al. [3] predicts which POI the user will visit at next time through historical check-in records. Li et al. [10] recommends POIs in location-based social networks (LBSN) by uncovering potential check-in information primarily based on Matrix Factorization. Like most recommendation systems, they are all based on actual check-in dataset collected from a certain LBSN. From this perspective, the user preference based on above studies can only describe the interest user perform on the certain LBSN rather than a comprehensive profile of the person.

3 Preliminaries

We first introduce the negative-unlabeled constraints, and then formally define the POI category inferring problem for trajectory data.

3.1 Problem Formulation

As shown in Figure 1, the brown circle draws the certain range out and all the POI categories in the circle (i.e. theater, hotel, restaurant and mansion) will be

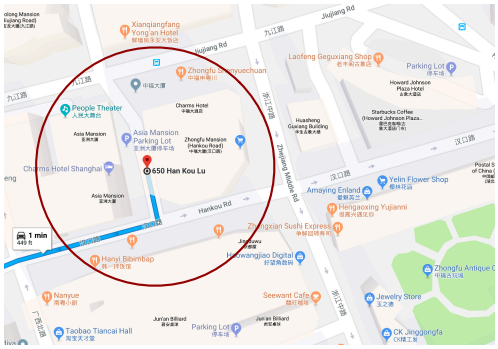


Fig. 1: The blue line illustrates the trajectory of vehicle, the red mark means where the car is parked.

defined as the candidate set of the GPS coordinate. Indeed, user's true visited POI must be within a certain range of the coordinate updates and the POI categories out of the range is impossible to be visited.

The collection vehicle trajectory data is in the form: <vehicle id, time stamp, location coordinates, vehicle state[1]>. Our ultimate goal is to infer the probability of POI category user visited after getting off his or her car. To achieve the goal, we first preprocess the raw data by several key steps:

1. Extract users' significant visiting points and dwelling time by filtering vehicle state.
2. Normalize time into time slots with the consideration of the proposition that human mobility would follow a high degree of regularity over time.
3. Match GPS coordinates with POI information. For the staying point in each time slot, the possible POI categories considered can be within an uncertain range of the update GPS coordinates. More details will be explained later in the experiments described in Section 5.

Let $I_{ut:}$ be the indicator of POI possible categories the user n visited during the time slot t. In this paper, we formally define the POI category inferring problem as follows:

DEFINITION *(POI category inferring):* Given a targeting user n, a observation time slot t, a candidate set $I_{ut:}$ of POI category, the POI category inferring problem is to predict the value of probability $\mathcal{X}_{utc} \in [0, 1]$. Specifically, the value $\mathcal{X}_{utc}$ more close to one means that the user n has greater chance to visit the category c within time slot t.

The Table 1 lists the notations and their meanings used in this paper.

3.2 Negative-Unlabeled Constraints

Based on the precondition mentioned above, possible POI categories must be within an candidate set of the staying point extracted from the vehicle trajectory

[1] The state flag illustrates whether the vehicle is running or stopping.

Table 1: Notation and Description

Notation	Description
U, T, C	users set, time lots set, category set
u, t, c	user id, time lot id, POI category
$\mathcal{X}$, $\mathcal{Y}$	POI probability tecnsor
$\mathcal{X}_{utc}$	the probability of user u visiting POI category c at time slot t
I_{utc}	indicates whether POI category c is in the candidate set of user u and time t
R	the number of latent factors

data. The possible POI categories contribute a candidate set. In addition, there can be only one actual category of visited POI for a user at one time. Therefore, the probabilities of POI categories in the candidate set sum up to one.

If we treat this problem as a two-class learning problem, we can easily label the POI categories out of range as negative class, leaving the categories in the candidate set unlabeled.

Under this scenario, we turn to negative-unlabeled learning, which is a counterpart concept to positive-unlabeled learning (PU), to solve this problem. PU [11] also called learning from positive and unlabeled examples, which aims to build a binary classifier to classify the test set containing positive and unlabeled examples into two classes. The research works [2] [7] propose semi-supervised PU learning methods to take advantage ofthe unlabeled data, which contains the instances belonging to the predefined class rather than the labeled categories. On the opposite, negative-unlabeled learning problem samples all the labeled examples from the negative class while the unlabeled examples come from both negative and positive classes. Accordingly, our collaborative inferring model is developed under the negative-unlabeled constraints.

4 Collaborative Inferring Framework

4.1 Tensor Factorization Using Tikhonov Regularization

To make use of the collaborative capabilities among users, times and categories and to complete missing data at the same time, we propose to estimate the possibility of different POI categories based on tensor models.

As shown in Figure 2, we use the tensor $\mathcal{X}$ to represents the observed data. The three-ways of $\mathcal{X} \in \mathbb{R}^{U,T,C}$ represent users, time slot and POI category respectively, where each element $\mathcal{X}_{utc} \in [0, 1]$ represent the probability of user u visiting POI category c during time slot t. U, T, C respectively denoting the number of user, time slot and POI category.

In order to capture the common behavioral characteristics of user, each day is segmented into several time bins, which results in $\mathcal{X}$ being sparse and low-rank. Similar to matrix factorization, tensor factorization can decompose a tensor into the sum of several rank-one tensors that can best approximates the given

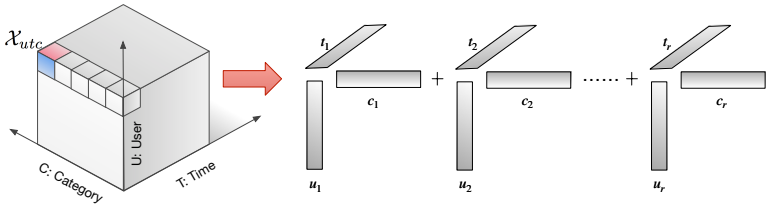


Fig. 2: Tensor decomposition model.

tensor. This paper will build our model on a CANDECOMP/PARAFAC (CP) decomposition model and can be represented as follow:

$$\mathcal{X} \approx \sum_{r=1}^{R} u_r \circ t_r \circ c_r \tag{1}$$

Where u_r, t_r, c_r are latent vectors of size $U \times 1$, $T \times 1$, and $C \times 1$ respectively, $R \leqslant min\{U, T, C\}$ is the number of latent factors as the rank of a tensor, and the symbol "$\circ$" stands for the outer product. More specifically, $\mathcal{X}$ is approximately equal to the sum of R tensors.

$$\underset{\mathcal{X},\mathcal{Y}\in\mathbb{R}^{U,T,C}}{minimize} \|\mathcal{X} - \mathcal{Y}\|_F^2 \tag{2}$$

The vectors u_r, t_r, c_r have be collected in latent factor matrices U, T, C for user, time and category, i.e. $U = [u_1, u_2 \cdots u_R]$, which are of sizes $U \times R$, $T \times R$, and $C \times R$, respectively. With these definitions, it can also be represented in matrix form:

$$\begin{aligned} \mathcal{X} \approx [\![\mathrm{U,T,C}]\!] &= \sum_{r=1}^{R} u_r \circ t_r \circ c_r \\ \mathcal{X}_{(1)} &= U(T \odot C)^\intercal \\ \mathcal{X}_{(2)} &= T(C \odot U)^\intercal \\ \mathcal{X}_{(3)} &= C(U \odot T)^\intercal \end{aligned} \tag{3}$$

The symbol "$\odot$" denotes the Khatri-Rao product [2] and the $\mathcal{X}_{(i)}$ means the model - i unfolding of tensor $\mathcal{X}$.

Further more, to avoid overfitting and to provide a unique solution, Tikhonov regularization terms is added with the regularization parameter $\lambda_U, \lambda_T, \lambda_C > 0$ to the objective function. Thus, the goal of tensor decompose problem can be represented by the following optimization problem:

[2] Khatri-Rao product of matrices A and B with k columns, given by $A \odot B = [a_1 \otimes b_1 \; a_2 \otimes b_2 \cdots a_k \otimes b_k]$, where $\otimes$ denotes Kronecker product.

$$min \left\| \mathcal{X} - [\![\mathrm{U,T,C}]\!] \right\|_F^2 + \frac{\lambda_U}{2} \left\| U \right\|_F^2 + \frac{\lambda_T}{2} \left\| T \right\|_F^2 + \frac{\lambda_C}{2} \left\| C \right\|_F^2 \tag{4}$$

Where the symbol "−" denotes the element-wise subtraction (which computes a tensor with each element equals $\mathcal{X}_{utc} - \mathcal{Y}_{utc}$) and "$\|\cdot\|_F$" indicates the Frobenius Norm of Tensor (similar to matrix) which is defined as: $\|\mathcal{X}\|_F = \sqrt{\sum_{u=1}^{U} \sum_{t=1}^{T} \sum_{c=1}^{C} \mathcal{X}_{utc}{}^2}$.

To solve the above optimization problem, we chose the alternating least square (ALS) algorithm, which is commonly used for CP decomposition. It works by iteratively optimizing one parameter while leaving the others fixed (i.e. fixes T and C to update U) on the base of Equation 5, 6 and 7, until meeting the convergence condition.

$$U = \mathcal{X}_{(1)}(T \odot C) \left[(T \odot C)^\intercal (T \odot C) + \lambda_U I_R \right]^\dagger \tag{5}$$

$$T = \mathcal{X}_{(2)}(U \odot C) \left[(U \odot C)^\intercal (U \odot C) + \lambda_T I_R \right]^\dagger \tag{6}$$

$$C = \mathcal{X}_{(3)}(U \odot T) \left[(U \odot T)^\intercal (U \odot T) + \lambda_C I_R \right]^\dagger \tag{7}$$

Where I_R is the unit matrix of size $R \times R$ and $[\cdot]^\dagger$ means generalized inverse matrix.

In this part, tensor decomposition method model the hidden relationships among users, time slots and POI categories, and generate collaborative latent factors. It helps to relieve sparsity problem of data and complement the missing data, which provides the Collaborative Inferring Framework the capability to make accurate predictions when user trajectory data is absent.

4.2 Collaborative Inferring Framework under Negative-Unlabeled Constraints

Algorithm 1 Projection vector under NU constraints

Input: $\mathcal{X}, \mathcal{I}$
Output: $\mathcal{Y}$

1: **for** $\forall\ u,\ t$ **do**
2: initial $v \in \mathcal{X}_{ut:}$
3: sort v in the desending order
4: $j = max\left\{c \in [\mathcal{I}_{ut:}] \mid v_c + \frac{1}{c}(1 - \sum_{i=1}^{c} v_i) > 0\right\}$
5: $\rho = \frac{1}{j}\left(1 - \sum_{i=1}^{j} v_i\right)$
6: $\mathcal{Y}_{ut:} \leftarrow s$ s.t. $s_i = max\left\{v_i + \rho, 0\right\}, i \in [\mathcal{I}_{ut:}]$
7: **end for**

For the POI category inferring problem, it is not sufficient to utilize only collaborative latent factors obtained from the tensor decomposition, because it only take the global factor into consider. To take advantage of the specific geographical factors for each staying point, the tensor is required to satisfy the

negative-unlabeled constraints. For each user u and time slots t, there can be only one actual category of visited POI meanwhile several possible POI categories. Therefore, the possible POI categories contributes a candidate set and the probabilities of them should be sum up to one. Meanwhile, it is impossible to visite the POI categories out of candidate set of every staying point for user u and time slots t. Similar to [19] , the NU constraints under our problem definition can be given as following:

$$\begin{cases} \mathcal{Y}_{utc} \geq 0, \forall u, t, c \\ \mathcal{Y}_{utc} = 0, \forall u, t, and\ c \notin I_{ut:} \\ {\mathcal{Y}_{ut:}}^{\intercal} I_{ut:} = 1, \forall u, t \end{cases} \tag{8}$$

The category latent vector is projected onto the probability simplex for each user u and time slot t to achieve the goal. Only the categories among candidate set $I_{ut:}$ can be calculated and the value of them are sum up to one, while the other categories not included in the candidate set $I_{ut:}$ are assigned to zero. As described in Algorithm 1., the project algorithm can be efficiently computed in $O(|I_{ut:}| \times \log |I_{ut:}|))$ time and perform better than other normalization method i.e. *softmax function* in the case of similar vector values. So far, we have considered the specific geographic information around each staying point, which improve the accuracy of inferring the POI categories.

Combining tensor decomposition and NU constraints , the final model in this paper can be expressed as follows:

$$\underset{\mathcal{X},\mathcal{Y} \in \mathbb{R}^{U,T,C}}{minimize} \|\mathcal{X} - \mathcal{Y}\|_F^2 \tag{9}$$

$$\begin{aligned} s.t. \quad & min \left\| \mathcal{X} - \hat{\mathcal{X}} \right\|_F^2 + \frac{\lambda_U}{2} \|U\|_F^2 + \frac{\lambda_T}{2} \|T\|_F^2 + \frac{\lambda_C}{2} \|C\|_F^2 \\ & where \quad \hat{\mathcal{X}} = [\![\mathrm{U,T,C}]\!] \\ & \begin{cases} \mathcal{Y}_{utc} \geq 0, \forall u, t, c \\ \mathcal{Y}_{utc} = 0, \forall u, t, and\ c \notin I_{ut:} \\ {\mathcal{Y}_{ut:}}^{\top} I_{ut:} = 1, \forall u, t \end{cases} \end{aligned} \tag{10}$$

In order to solve the collaborative inferring framework efficiently, we employ an alternating minimization scheme that iteratively updates one of $\mathcal{X}$ and $\mathcal{Y}$ and minimize their difference.

As shown in Algorithm 2, the learning strategy is summarized through alternating least square (ALS). First, to initialize the output tensor $\mathcal{Y}$ and the number of iterations. And then the tensor decomposition with tikhonov regularization procedure is described from lines 2 to 8. It is important to note that after updating all the three latent matrix, it is demanded to update the three unfolding of tensor $\mathcal{X}$ in each round of decomposition iteration. After the tensor decomposition procedure reaches the convergence condition, lines 9-10 expound projection procedure to meet the negative- unlabeled constraints by using $\mathcal{X}$ generated in last procedure as input.

Algorithm 2 Solving CFNU via ALS

Input: $\mathcal{X}, \lambda_U, \lambda_T, \lambda_C$
Output: $\mathcal{Y}$
Initialize: $\mathcal{Y} \leftarrow \mathcal{X}$, iter = 0
while Not Converged and $iter \leq I_{max}$ **do**
  repeat
    $\mathcal{X} \leftarrow \mathcal{Y}$
    Update $\tilde{U}, \tilde{T}, \tilde{C}$ according to Equation (5), (6), (7)
    Update $\mathcal{X}$ by update each unfolding with $\tilde{U}, \tilde{T}, \tilde{C}$ according to Equation (3)
  until Converged
  Compute $\mathcal{Y}$ according to Algorithm 1
end while
return $\mathcal{Y}$

5 Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed framework for inferring the actual POI category that user visited. Concretely, the experiments aim at answering following questions:

1. How effective is the proposed method compared with alternative state-of-the-art methods on inferring POI categories from trajectory data?
2. How do the parameters contribute to the inferring accuracy? That is to say, it is needed to give special care for tuning the approach, or is there a wider choice of parameters leading to high robustness?
3. How is the data complementing ability? Can the framework effectively predict POI categories when user GPS-coordinate is absent?

5.1 Datasert

Two real-world datasets are evaluated: electric vehicle trajectory data [3] and DianPing check-in data. All the methods are run on the same machine with an Intel Core 2.90GHz CPU and 16GB RAM of a single-thread for fair comparison.

The trajectory data set is sampled from fifty electric vehicles in Shanghai between 1 June 2015 and 31 December 2015. When the the vehicle is used, various types of information such as local time, GPS coordinates, vehicle states, travelled distance, running speed, etc. are uploaded every 30 to 50 seconds. In this study, we conduct three pretreatment steps on the raw data in order to apply the framework: (1)We first extract meaningful staying points from the raw trajectory data by estimating the dwell time. The dwelling time users spend on the staying points can be calculated easily based on local time, GPS coordinates and vehicle states. In this work, the threshold of dwelling time is half an hour. (2)Then, each day is split into 7 time bins as following: 0am-7am, 7am-10am, 10am-13pm, 13pm-16pm, 16pm-19pm, 19pm-22pm, 22pm-24pm. The non-uniform scheme is

[3] provided by Shanghai EV data platform: http://www.shevdc.org

chosen since there is little activity during the early morning. (3)In the end, we convert the GPS coordinates associated with POI categories and build the candidate set via Baidu map API [4]. Specifically, 28 POI categories of Baidu map hierarchy are chosen, including parking lot, hospital, school, governmental agency, tourist attraction, etc. Finally, each staying point can be represented by a tuple <user id, time bins id, category candidate set>.

DianPing data set, which contains over three hundred thousand check-in records from 2756 users and 22212 POIs corresponding to 66 categories over the whole 2014, is also evaluated. Each check-in includes a user ID, a time stamp, a POI ID, and the category of the POI.

5.2 Experimental Setup and Metrics

To investigate the quality of the proposed framework, we adopt the Accuracy@k as evaluation metric. Since there is only one true category for every user during a time slot, precision and recall are essentially equal in this situation.

Accuracy@k represents the percentage of correct category emerging in the top-k predictions and is calculated as:

$$Accuracy@k = \frac{\sum_{u=1}^{|U|}\sum_{t=1}^{|T|}\left|S_{u,t}\bigcap\tilde{S}_{u,t}\right|}{\#staying\ points} \tag{11}$$

Where $S_{u,t}$ is the visited categories set observed by the user u at time slot t and $\tilde{S}_{u,t}$ is the predicted value set about user u and time slot t. Besides, $\#staying\ points$ is the total number of staying points and k means the number of predicted values. To achieve the best performance, different k values are picked due to different feature of the two data sets for the parameters. Since experimental results are insensitive to regularization parameters in Tikhonov regularization (Equation 4), we set $\lambda_U, \lambda_T, \lambda_C = 0.01$.

5.3 Baseline

To the best of our knowledge, there is seldom model directly predicting POI category from human trajectory data. Our collaborative method is then compared with the following baselines.

- *Negative-Unlabeled Tensor Factorization (NUTF)*: This baseline computes user's location categories by unfolding the inferring tensor to a matrix and adopting random SVD algorithm.
- *Non-negative matrix factorization (NMF)*: This baseline [8] is a matrix decomposition method under the condition that all the elements in the matrix are non-negative constraints. It is used to solve Collaborative filtering problems in recommend system [13].
- *Singular Value Decomposition (SVD)*: We adopt the regularized SVD [15], which is a collaborative filtering algorithm predicting users' preferences for items, to obtain a prediction for POI category.

[4] http://lbsyun.baidu.com/

5.4 Evaluating over Electric Vehicle Trajectory Data

Table 2: The performance on trajectory data for accuracy

method	p=40%			p=60%			p=80%			avg
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	improvement
CFNU	0.3134	0.5993	0.7875	0.3133	0.6242	0.7875	0.3133	0.6242	0.7874	N/A
NUTF	0.2395	0.4954	0.5754	0.2762	0.5432	0.7190	0.2967	0.5817	0.6190	18.51%
NMF	0.2057	0.5309	0.6519	0.2385	0.4697	0.6270	0.1581	0.4640	0.5635	38.71%
SVD	0.2312	0.4713	0.5305	0.2779	0.5478	0.7247	0.2968	0.5787	0.7231	18.76%

Table 3: The performance on Check-in data for accuracy

method	p=40%			p=60%			p=80%			avg
	k=1	k=5	k=10	k=1	k=5	k=10	k=1	k=5	k=10	improvement
CFNU	0.1342	0.5241	0.8202	0.1319	0.5226	0.8182	0.1330	0.5249	0.8124	N/A
NUTF	0.1244	0.4269	0.6135	0.1124	0.3315	0.5230	0.1068	0.3691	0.5298	35.11%
NMF	0.1135	0.3439	0.5266	0.0793	0.2642	0.4003	0.0520	0.2076	0.3474	93.06%
SVD	0.1634	0.4979	0.7004	0.0741	0.2714	0.4752	0.0522	0.2602	0.4534	64.78%

Since the ground truth (the actual POI category visited by electric vehicle user) cannot be obtained from the trajectory data, we make a rule to pick ground truth as the verification set which can evaluate prediction performance. If the categories in candidate set of a staying point are all the same, this category is considered as the true single user visited, which is also called ground truth. For both the verification set and the remaining data set, the noisy data is added by generate categories randomly according to the same user id and time bins id. The ratio of created noise categories to total categories is represented by p. Our experiment also evaluate the influence of p. And then all the three data mentioned above are integrated as the training set to our framework.

The model performance in terms of accuracy and improvement are shown in Table 2. For this dataset a fraction of noise data is selected randomly. Then the possibility of each POI category is estimated . The results are reported at the fraction of 40%, 60% and 80%. Based on the results, we can observe that NMF perform worse than other methods in general. It is interesting that SVD and NUTF have similar performance on this dataset. The random svd method is also used in the NUTF, and NUTF can not effectively improve the accuracy with a small number of users, they have almost the same accuracy on this datasets. Our approach CFNU (short for Collaborative Inferring Framework under Negative-Unlabeled Constraints) outperforms baseline methods and achieves an average improvement of 18.51% relative to NUTF when k equals to 1, 3 and 5. These results likely due to the lack of exploitation of the potential relevance between time, user and category.

On the other hand, with the increase of fraction p, NMF shows a significant decrease while our method still maintains the high accuracy. A possible reason

is that not only the user's preference but also the time latent factors have been learned in the training, so that the possibility on the target POI category can be accurately predicted. This consequence also answers the question 2 raised at beginning of this section, and the parameter p has little influence on the accuracy of the proposed model.

5.5 Evaluating over Check-In data

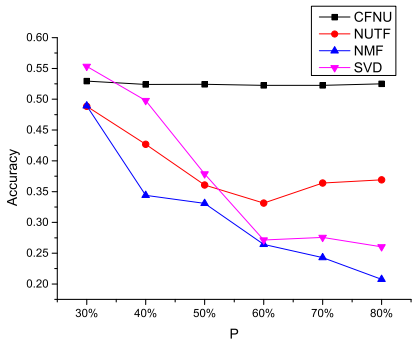
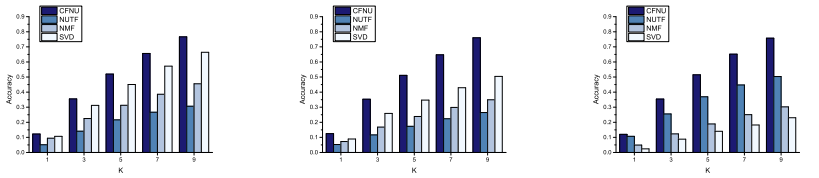


Fig. 3: The performance comparison of basic methods at different p .

To further illustrate the effectiveness of our approach, DianPing check-in data set is evaluated. DianPing check-in data is different from general trajectory data, it records the users visited POIs through the web service, that can directly obtain corresponding POI catrgories. Hence, to simulate the real situation defined to solve in the section 3, we randomly sample 10% of all the check-in records as the validation set and add the noise data in the same way as trajectory data. In addition, the 80:20 among validation set split is chosen to divide training set and testing set for testing complement and prediction capability. More specifically, to evaluate the accuracy of inferring result, only 80 percent of the validation set is selected, the remaining data (without ground truth) and the created noise data are used as the input of the framework. For evaluating complement and prediction capability, the other 20% of the validation set is divided as the test data without putting into the framework. The result verifies that proposed framework can predict POI category accurately even in the deficiency of check in data.

Firstly, the fraction of noise data is set as $p = 40\%$, 60% and 80% and Table 3 presents the evaluation results for inferring POI categories on check-in data. It is shown in the table that the proposed model consistently outperforms NUTF, NMF and SVD across k and gives an average improvement of over 30% in terms of accuracy over other alternative methods, demonstrating the effectiveness of our method.

Effects of noise proportion: To simulate the real situations, the noise percentage p is set from 30% to 80% with the step of 10%. The accuracy of these models under this scenario are shown in Figure 3. When the p is at a low percentage, the function of this area is relatively simple, which gives the NUTF, NMF and SVD methods an opportunity to outperform the proposed model. But with the increase of p, the POIs in the region will be more diversified and these methods show a huge downward trend, while our model consistently performs the high accuracy. This result is encouraging since it demonstrates that the proposed framework for POI categories inferring can achieve high robustness in a realistic scenario where varies categories are rounding the staying points. It also proves that the parameter p has little influence on the accuracy, and gives evidence of the high robustness since there is a wider choice of parameter p.



(a) accuracy@p=40% (b) accuracy@p=60% (c) accuracy@p=80%

Fig. 4: The performance of accuracy at p= $40\%, 60\%, 80\%$.

Performance of complement and prediction: To test the complementing and predictive capability of the model without check-in data, the predictions on the test data (20% of the validation set without putting into the framework) are evaluated. Figure 4 shows the predicted result with varying number of k. SVD achieves a relatively high accuracy when the noise proportion is low, but drops rapidly as the noise ratio increases. However, the proposed method has better performance than others in this scenario regardless of the p, which indicates proposed method can predict POI categories people visiting in the future accurately. This result provides the answer to question 3 , and the proposed collaborative inferring method can complement the missing data meanwhile make effectively predictions when user check-in data is absent.

From these results, under various scenarios, our proposed collaborative inferring method consistently performs best among all and outperforms state-of-the-art methods with a significant improvement. The effectiveness, robustness and superiority of the proposed model for the POI category inferring problem is empirically confirmed.

6 Conclusion

Inferring the user's visited POI category is indispensable for modeling the user's interest preferences and establishing user profile. It plays an important role in

the location-based service because of the comprehensive, accurate, detailed user portraits for every individual, and can help the system to provide better personalized service. The problem differs with many current POI researches with several new characteristics, which requires the deployment of new models. Our proposed collaborative inferring method exploits the collaborative capabilities among users, time slots and categories. Meanwhile, it effectively alleviates the problem of sparsity and improve the inferring accuracy. Through complementing the missing data using tensor decomposition with Tikhonov regularization, this framework can provide accurate predictions when user trajectory data is absent. Extensive experiments with two real-world data sets have validated the effectiveness of our collaborative inferring model. In our future work, contextual information can be explored to improve the prediction accuracy of current framework further.

Acknowledgments. This work is supported by NSFC grants (No. 61532021), Shanghai Knowledge Service Platform Project (No. ZF1213), SHEITC and Shanghai Agriculture Applied Technology Development Program (No. G20160201). Jiangtao Wang is the corresponding author.

References

1. Cao, X., Cong, G., Jensen, C.S.: Mining significant semantic locations from gps data. Proceedings of the VLDB Endowment 3(1-2), 1009–1020 (2010)
2. Fei, G., Liu, B.: Social media text classification under negative covariate shift. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2347–2356 (2015)
3. Feng, S., Li, X., Zeng, Y., Cong, G., Chee, Y.M., Yuan, Q.: Personalized ranking metric embedding for next new poi recommendation. In: IJCAI. pp. 2069–2075 (2015)
4. Ge, H., Caverlee, J., Zhang, N., Squicciarini, A.: Uncovering the spatio-temporal dynamics of memes in the presence of incomplete information. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 1493–1502. ACM (2016)
5. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. nature 453(7196), 779 (2008)
6. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review 53(2), 217–288 (2011)
7. Hu, H., Sha, C., Wang, X., Zhou, A.: A unified framework for semi-supervised pu learning. World Wide Web 17(4), 493–510 (2014)
8. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
9. Lee, R.K.W., Hoang, T.A., Lim, E.P.: On analyzing user topic-specific platform preferences across multiple social media sites. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1351–1359. International World Wide Web Conferences Steering Committee (2017)

10. Li, H., Ge, Y., Hong, R., Zhu, H.: Point-of-interest recommendations: Learning potential check-ins from friends. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 975–984. ACM (2016)
11. Li, X.L., Liu, B.: Learning from positive and unlabeled examples with different data distributions. In: European Conference on Machine Learning. pp. 218–229. Springer (2005)
12. Liu, R., Buccapatnam, S., Gifford, W.M., Sheopuri, A.: An unsupervised collaborative approach to identifying home and work locations. In: Mobile Data Management (MDM), 2016 17th IEEE International Conference on. vol. 1, pp. 310–317. IEEE (2016)
13. Luo, X., Zhou, M., Xia, Y., Zhu, Q.: An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. IEEE Transactions on Industrial Informatics 10(2), 1273–1284 (2014)
14. Narita, A., Hayashi, K., Tomioka, R., Kashima, H.: Tensor factorization using auxiliary information. Data Mining and Knowledge Discovery 25(2), 298–324 (2012)
15. Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of KDD cup and workshop. vol. 2007, pp. 5–8 (2007)
16. Van Canh, T., Gertz, M.: A spatial lda model for discovering regional communities. In: Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. pp. 162–168. IEEE (2013)
17. Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., McAuley, J.: Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1103–1112. International World Wide Web Conferences Steering Committee (2017)
18. Ye, M., Yin, P., Lee, W.C., Lee, D.L.: Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 325–334. ACM (2011)
19. Yi, J., Lei, Q., Gifford, W., Liu, J.: Negative-unlabeled tensor factorization for location category inference from inaccurate mobility data. arXiv preprint arXiv:1702.06362 (2017)
20. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 186–194. ACM (2012)
21. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on World wide web. pp. 791–800. ACM (2009)
22. Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X.: You are where you go: Inferring demographic attributes from location check-ins. In: Proceedings of the eighth ACM international conference on web search and data mining. pp. 295–304. ACM (2015)