Accepted Manuscript

Privacy-preserving and sparsity-aware location-based prediction method for collaborative recommender systems

Shunmei Meng, Lianyong Qi, Qianmu Li, Wenmin Lin, Xiaolong Xu, Shaohua Wan

PII: S0167-739X(18)31805-3
DOI: https://doi.org/10.1016/j.future.2019.02.016
Reference: FUTURE 4769

To appear in: *Future Generation Computer Systems*

Received date : 28 July 2018
Revised date : 19 December 2018
Accepted date : 11 February 2019

Please cite this article as: S. Meng, L. Qi, Q. Li et al., Privacy-preserving and sparsity-aware location-based prediction method for collaborative recommender systems, *Future Generation Computer Systems* (2019), https://doi.org/10.1016/j.future.2019.02.016

Privacy-preserving and Sparsity-aware Location-based Prediction Method for Collaborative Recommender Systems

Shunmei Meng[1,2], Lianyong Qi[3], Qianmu Li[1], Wenmin Lin[4], Xiaolong Xu[5], Shaohua Wan[6*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]State Key Laboratory for Novel Software Technology, Nanjing University, China
[3]School of Information Science and Engineering, Qufu Normal University, China
[4] Department of Computer Science, Hangzhou Dianzi University, China
[5] School of Computer and Software, Nanjing University of Information Science and Technology, China
[6]School of Information and Safety Engineering, Zhongnan University of Economics and Law, China

{mengshunmei@njust.edu.cn, lianyongqi@gmail.com, qianmu@njust.edu.cn, linwenmin@hdu.edu.cn, xlxu@nuist.edu.cn, shaohua.wan@ieee.org}

Abstract. With the rapid growth of public cloud offerings, how to design effective prediction models that provide appropriate recommendations for potential users has become more and more important. In dynamic cloud environment, both of user behaviors and service performance are sensitive to contextual information, such as geographic location information. In addition, the increasing number of attacks and security threats also brought the problem that how to protect critical information assets such as sensitive data, cloud resources and communication in a more effective and secure manner. In view of these challenges, we propose a privacy-preserving and sparsity-aware location-based prediction method for collaborative recommender systems. Specifically, our method is designed as a three-phase process. Firstly, two privacy-preserving mechanisms, i.e., a randomized data obfuscation technique and a region aggregation strategy are presented to protect the private information of users and deal with the data sparsity problem. Then a location-aware latent factor model based on tensor factorization is applied to explore the spatial similarity relationships between services. Finally, predictions are made based on both global and spatial nearest neighbors. Experiments are designed and conducted to validate the effectiveness of our proposal. The experimental results show that our method achieves decent prediction accuracy on the premise of privacy preservation.

Keywords: Location-aware recommendation, Privacy-preserving, Data sparsity, Tensor factorization

1. Introduction

Recommendation has been a hot research topic with the rapid growth of cloud services [1-2]. Great efforts have been done both in industry and academia to develop effective prediction models for recommender systems, which mainly aim at exploiting

* Corresponding author

available information to provide users with satisfying recommendations [3-4]. With the popularity of mobile applications and devices, most cloud services could be invoked everywhere [5]. Because of the dynamics of cloud environment, most cloud services become region-sensitive. Actually, user preferences, quality of service (QoS) and the popularity of services are all varying with the change of users' geographic location. Location information plays an increasingly important role in both users' behaviors and service performance, especially in dynamic cloud environment and real-world applications.

Although there have been some researches focusing on studying location influence to recommendation models [6-8]. Most of them merely focused on the location influence on user preferences. Few work paid attention to the location influence on QoS performance of services. Compared with traditional internet services, QoS of cloud services is more sensitive to location due to the dynamics of their environment. Both of QoS of cloud services and user behaviors are usually changing over geographic location. Thus it is still a fundamental task for recommender systems to provide the most beneficial suggestions to potential users with the consideration of location information. Moreover, data sparsity is always a serious threat that deteriorates the performance of recommendation methods [9-10], where users may only use a small number of services and provide limited QoS records. Under a data-sparsity scenario, existing collaborative recommendation models fail to capture the similarity relationships between users or services effectively. Factorization technique has been a successful prediction model used in recommender systems and proved to be an effective way to address the data sparsity problem [11-12].

In addition, the ever-increasing number of attacks and security threats also bring the privacy preservation problem, which has been an important issue emerged to be addressed in complex cyber environment [13-14]. To make effective recommendations, user sensitive information, such as observed QoS values, activity patterns, location information, social relationships, etc., are collected by recommender systems, which puts users at risk. The behavior data and location information of users may be abused or even resold to unauthorized parties for profits. In location-aware recommendation models, both location information and observed QoS data could disclose the private information of users. Thus effective privacy-preserving mechanisms should be integrated into recommendation models to protect the private information of users in a more effective and secure manner [15-17].

Based on these observations, in this paper, we propose a novel privacy-preserving and sparsity-aware location-based prediction model based on tensor factorization. The proposed model aims to achieve a tradeoff between prediction accuracy and privacy preservation. The main contributions of our proposal are summarized as follows:

- A privacy-preserving location-based collaborative recommendation algorithm is proposed to achieve a tradeoff between prediction accuracy and privacy preservation. Firstly, a user-service-location model and a security model used in our method are defined.
- Then two privacy-preserving mechanisms are proposed: 1) A random perturbation technique is employed to protect the observed QoS data of users; 2) A

region aggregation strategy is presented to preserve the specific location of users and deal with the data sparsity problem.

- Moreover, a location-aware tensor factorization model is employed to mine the similarity relationships between services over location adaptively so as to provide location-aware predictions.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Then the problem statement of our work is presented in Section 3. Based on the analysis in Section 3, a privacy-preserving and sparsity-aware location-based prediction algorithm is proposed in Section 4. Section 5 empirically studies the empirical performance and efficiency of our method. Finally, Section 6 concludes this paper and provides some future work.

2. Related Work

In this section, we review the related work of recommendation models from four aspects: location-social influence, spatial-temporal influence, factorization techniques and privacy-preserving recommendation models.

Location-Social Influence. Many researches have demonstrated that there is a strong correlation between user's locations and his/her POIs (Points of Interest) as well as social connections. Recent location-aware recommendation studies mainly focuses on exploiting the geographical influence or social influence to improve prediction accuracy [18-25]. Lian et al. [18] propose a collaborative location recommendation framework to exploit the relations between users, activities and locations, so as to provide location-aware recommendations. The research [19] presents a location-aware probabilistic generative model that leverages location-based ratings to model user profiles and provide location-recommendations. Chen et al. [20] employ the location information and QoS values to cluster users and services to provide personalized service recommendations. Jiang et al. [21] propose a personalized travel sequence recommendation approach by learning topical package model from big multi-source media, travelogues and community-contributed photos. The references [22-24] focus on analyzing the location influence on users' check-in behaviors, and combining user preferences, location influence or social influence into a geo-social recommendation model. The authors in [25] present an instance-region neighborhood matrix factorization model where two levels of geographical characteristics are integrated into the learning of latent factors of users and locations to predict users' preferences on locations.

Spatial-Temporal Influence. Recently, to obtain more accurate recommendations for users, many researches not only consider location influence but also temporal influence [26-31]. Zhang et al. [26] present a personalized trip recommendation approach based on not only the temporal-spatial constraints but also user specific preferences on POIs. The reference [27] proposes a spatial-temporal topic model to infer user preferences, the spatial and temporal patterns of topics embedded in users' check-in behaviors, and the correlation between sentimental tags and rating scores from users' check-in and rating behaviors. Yuan et al. [28] study users' mobility behaviors from users, location information, temporal information, and activities, and propose a nonparametric bayesian model for context-aware applications. Wang et al. [29] propose a spatial-temporal QoS prediction method where the temporal QoS prediction is formulated as a generic regression problem and a zero-mean Laplace prior distribution assumption is made on the residuals of QoS prediction. The authors in [30] present a spatial-temporal latent ranking approach based on a ranking-based pairwise tensor factorization framework to model the interactions among users, POIs, and time information. Yang et al. [31] design a spatial-temporal activity preference model and apply a context-aware fusion framework to integrate the spatial and temporal preference models for preference inference.

Factorization Techniques. Factorization Techniques including both matrix factorization and tensor factorization have been successfully utilized in prediction models since the Netflix Prize [3, 32-37].The reference [3] proposes a healthcare recommendation model which presents a topic model based on hybrid matrix factorization methods to mine user preference distribution and doctor feature distribution. He et al. [33] design a novel learning algorithm based on matrix factorization technique for online recommendation, which aims to mine user preferences from implicit feedback. Zhang et al. [34] propose a temporal QoS-aware recommendation approach based a non-negative tensor factorization technique to deal with the triadic relations among users, services and time. The authors in [35] proposes two distributed approaches based on high-order and large-scale tensor factorization to make a trade-off between convergence speed and prediction accuracy. Shi et al. [36] factorize user-item rating matrix and other contextual movie similarity matrices to integrate contextual information into the recommendation models. The literature [37] designs a mashup service recommendation model by combining the implicit API correlations regularization into probabilistic matrix factorization model to enhance the recommendation diversity.

Privacy-preserving Recommendation Algorithms. In recommendation models, the requirement to collect users' QoS data and other sensitive information probably puts users at risk. To enable effective recommendation from shared data under privacy protection, there have been many works on privacy-preserving recommender systems. Existing works on privacy preserving recommendation models can be divided into two categories, i.e., cryptography based recommendation approaches and data perturbation based recommendation approaches.

Cryptography based recommendation approaches usually adopt homomorphic encryption to encrypt user private information [38-41]. Qi et al. [38] present a privacy-preserving distributed service recommendation method based on Locality-Sensitive Hashing strategy to achieve a tradeoff among prediction accuracy, privacy preservation and efficiency. The reference [39] proposes a privacy-preserving collaborative QoS prediction framework which combines Yao's garbled circuit and additively homomorphic encryption by additively secret sharing to address non-linear computations in QoS prediction. Kaur et al. [40] present a privacy-preserving collaborative filtering scheme on arbitrary distributed data based on multi-party random masking and polynomial aggregation techniques. Li et al. [41] propose an efficient privacy-preserving collaborative filtering algorithm for online recommendations, where an unsynchronized secure multi-party computation protocol is presented.

Data perturbation based recommendation approaches generally inject noise on user data to protect user privacy [17, 42-48]. Zhu et al. [17] design a similarity-maintaining privacy preservation strategy to obfuscate the QoS data from the users' perspective and propose a location-aware low-rank matrix factorization method to improve the robustness of recommendation models. The research [42] proposes a simple yet effective privacy-preserving framework by applying data obfuscation techniques, and introduces two privacy-preserving QoS prediction approach under the privacy-preserving framework. Boutet et al. [43] firstly design an obfuscation mechanism revealing only the least sensitive information and then propose a randomization-based dissemination algorithm ensuring differential privacy. Polatidis et al. [44] propose a multi-level privacy-preserving method for collaborative filtering systems by perturbing each rating based on multiple levels and different ranges of random values for each level. The authors in [45] propose a hybrid privacy-preserving protocol for matrix factorization by combining partially homomorphic encryption with Yao's garbled circuits. Casino et al. [46] propose a novel privacy-preserving collaborative filtering method based on micro-aggregation, which guarantees k-anonymity and makes a tradeoff between the privacy of users' preferences and recommendation accuracy.

Different from previous research work, in our work, we consider location influence into recommendation models by distinguishing region-sensitive QoS metrics from region-insensitive QoS metrics. A randomized data obfuscation technique and a region aggregation algorithm are used to preserve the observed QoS data and location information of users respectively. Besides, in most existing recommendation works, factorization techniques are usually used to predict the rating of users for services directly. While in our proposal, we use a high-order tensor factorization technique to mine the similarity between services.

3. Problem Statement

In this section, we first present a motivation scenario of our proposal and formulate our problem. Then a tensor decomposition model used in our proposal is introduced.

3.1 A Motivation Scenario

In this section, we will present a recommendation scenario to show the research problem of our work. Tom is a software engineer working in China, and he wants to rent some cloud virtual machines (VM). However, there are large-scale candidate services that can satisfy his functional requirement. Then the problem that he faces is how to find an optimized service that is most suitable for him in nonfunctional requirements (QoS). The QoS metrics of VM services contain price, stability, speed and security. Here, we assume that Tom concerns more about stability and security.

Now, there are two candidate VM resources A and B, which both meet the functional requirements of Tom. And Tom has not used both of them before. The overall ratings for the two candidate services are almost the same. However, in dynamic cloud environment, the QoS performance (such price and response time) of the same service invoked in different locations maybe different. For example, the rating of A and B in China are respectively {4.7, [4, 4.6, 4.8, 4.6]} and {4.5, [4.8, 4, 4.4, 4.5]}, where 4.7 is the overall rating and [4, 4.6, 4.8, 4.6] is the rating vector for individual QoS. Then Tom may choose service A in China. While in the U.S., the rating of A and B are respectively {4.4, [4, 4.2, 4.4, 4.5]} and {4.6, [4.8, 4.5, 4.7, 4.5]}. Then Tom may choose resource B when he travels to the U.S. for business. But if we make predictions based on all ratings without considering location influence, the prediction for Tom in both countries maybe the same, which is obviously unreasonable.

From this example, we can find that the QoS performance of cloud services may vary over geographic location. And the correlation between users and ratings over mobile location may be weakened. Most existing recommendation methods make predictions for the target user (the user needs to be recommended) based on the collected ratings without considering the location influence to QoS performance of cloud services. Besides, both of the observed QoS data and location information could disclose the private information of users, such as their habits and affiliations. Thus, the security and privacy problem is also an important issue to be addressed in location-aware prediction problem. In addition, privacy-preserving mechanisms may lead to less accurate prediction performance, as the real information of users (such as the observed QoS data and location information) is blurred. Thus we should make a trade-off between prediction accuracy and privacy preservation.

To address the privacy-preserving location-aware prediction problem, we propose a privacy-preserving and location-aware collaborative recommendation model to obtain an optimized recommendation.

3.2 Problem Formulation

Some important concepts and definitions are presented in this section. To mine the triadic relations among users, cloud services, and location features, we first introduce the user-service-location model used in our proposal. Then a security model for privacy preservation is presented.

(1) User-Service-Location Model:

In the user-service-location model, given a user set U and a service set S, the number of users and services are respectively N and M. Each service in S is associated with an H-dimensional QoS vector $\vec{Q} = [q_1, q_2, ..., q_H]$, which indicates the features of non-functional properties of the services. The rating of user i on service j at location l_{ij} is denoted as $\{r_{ij}, \boldsymbol{RQ}_{ij}, l_{ij}\}$, where r_{ij} ($r_{ij} \geq 0$) denotes the overall rating of user i on service j, $\boldsymbol{RQ}_{ij} = [rq^1_{ij}, rq^2_{ij}, ..., rq^H_{ij}]$ is the rating vector for individual QoS metric, l_{ij} is a location tag indicating the specific location where user i invoked service j.

Region Division: To analyze the spatial influence to recommendation performance in dynamic cloud environment, the invoked location of services is divided into G regions, i.e., $\{R_1, R_2, ..., R_G\}$. We assume that services invoked in the same region are likely to have similar location-aware QoS performance.

Region-sensitive QoS metrics and Region-insensitive QoS metrics: To mining the location features of QoS, we distinguish region-sensitive QoS metrics from region-insensitive QoS metrics. Region-sensitive QoS metrics are the metrics that have clear location features, which are dynamic features relative to location information (such as response time). And region-insensitive QoS metrics are regular features which are usually evolving at a rather slow speed (such as reputation).

In our work, we distinguish region-sensitive QoS metrics from region-insensitive QoS metrics by measuring the fluctuation of the rating for each individual QoS metric over different regions, which is defined in equation (1).

$$Flu(q_h) = \frac{1}{G \cdot |S|} \sum_{s=1}^{|S|} \sum_{g=1}^{G} (avg(R_g, rq_h^s) - avg(rq_h^s))^2 , \quad (1)$$

Where $avg(R_g, rq_h^s)$ is the average rating of metric q_h in region R_g of service s, and $avg(rq_h^s)$ is the overall average rating of q_h in all regions of service s. A fluctuation threshold δ is given to determine whether a QoS metric is region-sensitive or region-insensitive. If $Flu(q_h) > \delta$, then q_h can be seen as a region-sensitive QoS metric, otherwise, q_h is a region-insensitive QoS metric. So the QoS metric vector $\boldsymbol{Q} = [q_1, q_2, ..., q_H]$ can be divided into two parts, i.e., region-insensitive QoS vector $\boldsymbol{IQ} = [iq_1, iq_2, ..., iq_a]\ (0 \le a \le H)$ and region-sensitive QoS vector $\boldsymbol{SQ} = [sq_1, sq_2, ..., sq_b] \quad (0 \le b \le H, a + b = H)$.

Global Nearest Neighbors and Spatial Nearest Neighbors: In our model, the neighborhood model is item-based. Traditional item-based CF algorithms usually make predictions based on the ratings of "neighbor" items selected from the whole collected data without considering location influence. Different from previous work, we define two kinds of nearest neighbors in our proposal, i.e., global nearest neighbors and spatial nearest neighbors, which are described in the following.

Definition 1. (Global Nearest Neighbors) For a candidate service s, its global nearest neighbors are the services that have the most similar QoS ratings with service s at all regions.

Definition 2. (Spatial Nearest Neighbors) The spatial nearest neighbors of service s in region R_g are the services that have the most similar QoS performance with service s at the aggregated region of R_g. (The definition of the aggregated region is presented in Section 4.1)

(2) Security Model:

With the ever-increasing number of attacks and security threats in cloud environments, the privacy and security problem has been an important issue emerged to be addressed in recommender systems. In most existing location-aware recommendation models, few works consider the security problem in their prediction models. The requirement to collect users' QoS record data and other sensitive information probably puts users at risk. Fig. 1 shows the security model for privacy preservation of our proposal. As shown in Fig. 1, both observed QoS data and location information could disclose the private information of users. In complex cyber environment, the private information of service users may be abused by the potential privacy attackers such as insincere recommender systems and other unauthorized parties. Thus effective privacy-preserving mechanisms should be integrated into recommendation models to protect the private information of users in a more secure manner. More detail analysis about security and privacy problem in location-aware recommendation model is described in the following.

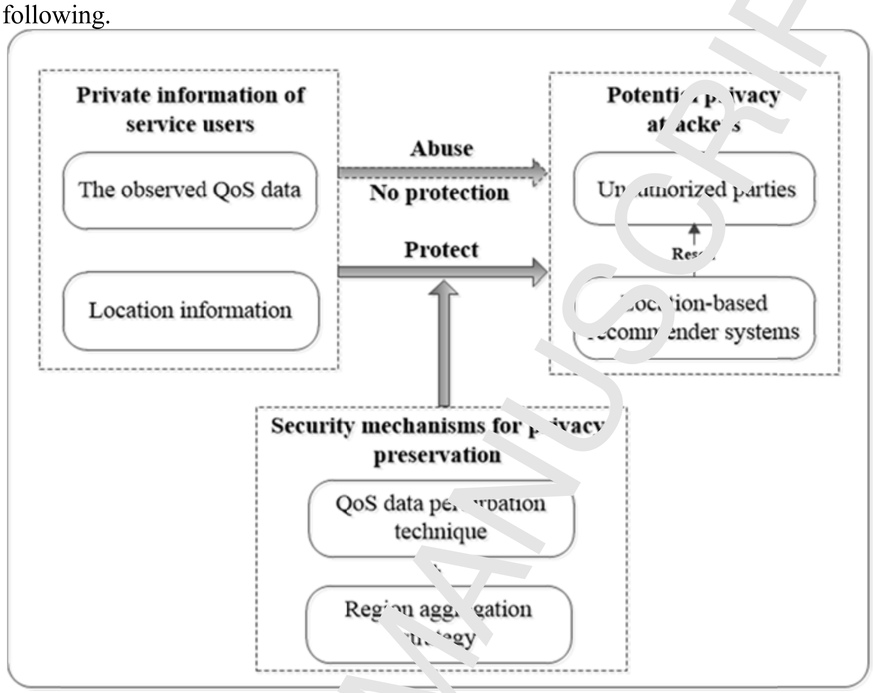


Fig. 1 Security model for privacy preservation.

1) Risk of disclosing the observed QoS data: The observed QoS data records contain both the subjective and objective feedbacks of users on services, which reflect not only user preferences but also the contextual information of users. Thus more accurate predictions could be made by mining the relationships between users and services based on history QoS data records of users. However, the observed QoS data may be abused by insincere recommender systems or even resold to unauthorized parties for profits. The real QoS information may disclose the private information of users such as access manners, specific location information, or even user habits and user identity. In our proposal, to protect the private information of users form potential privacy attackers, we apply a randomized data obfuscation technique to disguise the observed QoS data. The randomized obfuscated QoS data will reduce the association among the QoS data, users and services, and then it will be difficult for attackers to infer user habits or the specific identity of users so as to protect the privacy information of users.

2) Risk of disclosing location information of users: With the popularity of mobile applications and devices, specific location information could be obtained easily by today's network technology. In complex cyber environment, specific location information has become important personal private information of users. According to the location information, not only the specific geographic location of users (such as city and region) could be obtained, but also more specific identity information of users could be inferred by integrating service category information. Thus it is also important to protect users' location information. Most existing location-aware recommendation approaches utilize the specific location information, or even the specific latitude and longitude information of users to make personal recommendations. Once the specific location information of users is obtained by the attackers, it will put users at risk. Thus it is important to blur the specific location information of users to make users in a more secure environment. To address this problem, a region aggregation strategy is proposed in our work to expand and fuzzy the target region so as to blur the specific location of users.

Based on the data obfuscation technique and the region aggregation strategy, the private information of users could be protected, but the prediction accuracy of the location-aware recommendation model will be affected. Thus in privacy-aware recommendation model, it is important to get a trade-off between prediction accuracy and privacy-preservation.

3.3 CP Decomposition Model

To analyze the latent factors among services, neighbors and location information, high-order decomposition techniques are necessary. CANDECOMP/ PARAFAC (CP) decomposition model [49] has been proved to be one of the most successful approaches of high-order decomposition for its uniqueness and related interpretability of the components. In our work, we will apply a location-aware LFM (latent factor model) model based on CP decomposition to mine the triadic relations among services, neighbors and location features. In CP decomposition, an N-dimensional tensor $\boldsymbol{X} \in \Re^{I_1 \times I_2 \times \ldots \times I_N}$ can be decomposed into a sum of rank-one tensors, which can be written as:

$$\boldsymbol{X} \approx \sum_{r=1}^{R_X} x_r^{(1)} \circ x_r^{(2)} \circ \cdots \circ x_r^{(N)}, \quad (2)$$

where R_X is the rank of tensor $\boldsymbol{X}$ and vector $x_r^n \in \Re^{I_n \times R_X}$ ($r = 1,\ldots,R_X$ and $n = 1,\ldots,N$). More details of CP decomposition can be found in reference [49].

4. Privacy preserving and Sparsity-aware Location-based Prediction Method

In this paper, to make effective predictions with privacy protection, a privacy-preserving and sparsity-aware location-based prediction method is proposed. Our method is designed as a three-phase process. In phase 1, two privacy-preserving mechanisms are proposed. Firstly, to protect the observed QoS data of users, a simple but effective data perturbation technique is applied. To further blur the geographic location of users, regions are aggregated, which also addresses the data sparsity problem. In phase 2, similarity between services is first calculated based on the obfuscated QoS data firstly, and then a location-aware LFM model based on CP decomposition is applied to predict the missing similarities. In phase 3, predictions are made based on the history ratings of both the global neighbors and the spatial nearest neighbors. The three phases are described in detail in the following.

4.1 Privacy-preserving Mechanisms

1) QoS Data Perturbation:

For privacy preservation, we use a randomized data obfuscation technique to disguise the observed QoS values. The basic idea of the randomized data obfuscation is to add a noise to the real QoS data. By the empirical analysis, the literature [42] and [50] have proved that some approximate computations (such as scalar product) on the aggregated data of the disguised QoS data can be done with decent or even better accuracy. The randomized perturbation on the observed QoS values can be performed with the following equation:

$$r'^{h}_{ij} = r^{h}_{ij} + \varepsilon_{ij}, \tag{3}$$

where r^h_{ij} is the real QoS data, and r'^h_{ij} is the disguised QoS data of r^h_{ij}, $\varepsilon_{ij} \in [-\beta, \beta]$ is a random value generated based on the uniform distribution in $[-\beta, \beta]$, and β is the random range. If the range is too large, the accuracy of the prediction will be very low. While if the range is too small, the perturbed QoS value could still disclose the private information of users. Thus, to make a trade-off between user privacy and the prediction accuracy, the range of the random value β is critical and should be well chosen, which will be discussed in detail in the experiment.

As the observed QoS data are obfuscated randomly, it is impossible to infer the real QoS values based on the obfuscated QoS values. Thus the observed QoS values are protected.

2) Region Aggregation:

To make more effective recommendations, location information of users will be utilized in location-aware recommendation model. Users also have privacy concerns about their geographic location. Since the location preferences or social relationship of users are private. The disclosure of such information may lead to security threats. Besides, as the QoS performance of cloud services is usually spatially dynamic, user preferences and ratings for the services also vary over location. Then the neighbors (similar services) of the same service may also vary with different invoked location. To make more accurate recommendations, we should make predictions based on the history ratings at the target region. However, the data sparsity problem is always a shortcoming in recommender systems, especially in location-aware recommendation models. The related rating dataset of the target user at the target region may be very spare, since users may only use a small number of items and provide limited QoS records in a certain region.

To solve the privacy preservation problem and data sparsity problem, we employ an aggregation strategy to aggregate the similar regions for the target region, so as to expand the target region and blur the specific location of users. The algorithm of the region aggregation strategy is presented in Algorithm 1, which is an improved version of the threshold-based clustering algorithm proposed in [51].

Algorithm 1 Region Aggregation Algorithm

Input: Regions $\{R_1, R_2,\ldots, R_G\}$, and region similarity threshold θ

Output: New region division: $\{FR_1, FR_2,\ldots, FR_K\}$

1 $K \leftarrow 1$, $FR_1 \leftarrow R_1$
2 **for** $i \leftarrow 2$ to G
3 **for** $j \leftarrow 1$ to K
4 $f \leftarrow 0$, $max \leftarrow 0$
5 **if** $\eta(R_i, FR_j) \geq \theta$ & $\eta(R_i, FR_j) \geq max$
6 **then** $max \leftarrow \eta(R_i, FR_j)$, $f = j$
7 **end if**
8 **end for**
9 **if** $f \neq 0$
10 **then** $FR_f \leftarrow FR_f \cup R_i$
11 **do** merged ← False
12 **for** $j \leftarrow 1$ to K
13 **if** $f \neq j$ & $\eta(FR_f, FR_j) \geq \theta$
14 **then** $FR_f \leftarrow FR_f \cup FR_j$
15 $FR_j \leftarrow FR_K$, $K \leftarrow K - 1$
16 merged ← True
17 **break;**
18 **end if**
19 **end for**
20 **while** (merged)
21 **else** $K \leftarrow K + 1$
22 $FR_K = R_i$
23 **end if**
24 **end for**
25 return $\{FR_1, FR_2,\ldots, FR_K\}$;

In Algorithm 1, a region is randomly selected and assigned as a region cluster itself (Line 1). Then for every unassigned region, calculate its similarity with existing clusters. If the similarity is no less than the region similarity threshold θ, then aggregate the region to the cluster with the maximum similarity (Line 3-10). If no such cluster could be found after checking all existing clusters, then assign the region as the seed for a new cluster (Line 21-23). After the step above, if the similarity of the new region cluster with another existing cluster is no less than threshold θ, then merge the two clusters together, and recompute the cluster similarities (Line 11-20). From Algorithm 1, it can be found that each original region R_g only belongs to an aggregated gion FR_k ($1 \leq k \leq K$, $1 \leq K \leq G$).

As presented in Algorithm 1, we define a spatial similar coefficient $\eta(R_i, R_j) \in [0,1]$, which is denoted as the spatial closeness of the region-sensitive QoS metrics for candidate services between region R_i and R_j. The larger $\eta(R_i, R_j)$ is, the closer the spatial features of candidate services between R_i and R_j is. The spatial similarity $\eta(R_i, R_j)$ is determined based on Pearson Correlation Coefficient (PCC), which is defined in equation (4). Given a region similarity threshold θ, if $\eta(R_i, R_j) \geq \theta$, then region R_i and R_j can be considered to be similar.

$$\eta(R_i,R_j)=\frac{\sum_{s\in S(R_i)\cap S(R_j)}(\overline{\boldsymbol{RS}}_{is}-\overline{\boldsymbol{RS}}_i)\bullet(\overline{\boldsymbol{RS}}_{js}-\overline{\boldsymbol{RS}}_j)}{\sqrt{\sum_{s\in S(R_i)\cap S(R_j)}\left\|\overline{\boldsymbol{RS}}_{is}-\overline{\boldsymbol{RS}}_i\right\|^2}\cdot\sqrt{\sum_{s\in S(R_i)\cap S(R_j)}\left\|\overline{\boldsymbol{RS}}_{js}-\overline{\boldsymbol{RS}}_j\right\|}}\,, \tag{4}$$

where $S(R_i)\cap S(R_j)$ is the set of coinvoked services by users at region R_i and R_j, $\overline{\boldsymbol{RS}}_{is}$ is the average spatial QoS-rating vector of service s at region R_i, $\overline{\boldsymbol{RS}}_i$ is the average spatial QoS-rating vector of all candidate services in $S(R_i)\cap S(R_j)$ at R_i. Here, $\overline{\boldsymbol{RS}}_{is}$ and $\overline{\boldsymbol{RS}}_i$ are calculated based on obfuscated QoS data. Note that the QoS data appearing below are all obfuscated data.

Once the regions are aggregated, the specific region is blurred. The location information used in our recommender model is no longer the accurate locations, but only represented by a region number. And the sub-regions in the aggregated region clusters are discontinuous which would make it more difficult to get the real locations of users. And it will be meaningless for attackers to get the aggregated region information. Thus the private location information of users can be protected.

Based on the above two privacy-preservation mechanisms, i.e., the data obfuscation technique and the region aggregation strategy, the private information of users could be protected. To achieve decent prediction accuracy on the premise of privacy preservation, appropriate values for random range β and region similarity threshold θ should be set, which will be analyzed in the experiment.

4.2 Nearest Neighbors Determination

1) **Similarity Computation:**

As presented in Section 3, there are two kinds of nearest neighbors, i.e., global nearest neighbors and spatial nearest neighbors, which can be calculated as follows.

The global nearest neighbors of a service s can be determined by equation (5).

$$sim_{sv}^{GNN}=\frac{\sum_{u\in U(s)\cap U(v)}\left(\boldsymbol{RQ}_{us}-\overline{\boldsymbol{RQ}}_s\right)\bullet\left(\boldsymbol{RQ}_{uv}-\overline{\boldsymbol{RQ}}_v\right)}{\sqrt{\sum_{u\in U(u)\cap U(v)}\left\|\boldsymbol{RQ}_{us}-\overline{\boldsymbol{RQ}}_s\right\|^2\times\left\|\boldsymbol{RQ}_{uv}-\overline{\boldsymbol{RQ}}_v\right\|^2}}\,, \tag{5}$$

where $U(s)\cap U(v)$ is the set of users that rated both service s and service v in all regions. Here, we give a preset similarity threshold δ_{sim}, then the services that have similarity with service s no less than δ_{sim} can be considered as global nearest neighbors of service s.

The spatial nearest neighbors of service s at aggregated region FR_k can be determined by equation (6).

$$sim_{sv}^{SNN}(FR_k)=\frac{\sum_{u\in U_{sv}(FR_k)}\left(\boldsymbol{RQ}_{us}-\overline{\boldsymbol{RQ}}_s\right)\bullet\left(\boldsymbol{RQ}_{uv}-\overline{\boldsymbol{RQ}}_v\right)}{\sqrt{\sum_{u\in U_{sv}(FR_k)}\left\|\boldsymbol{RQ}_{us}-\overline{\boldsymbol{RQ}}_s\right\|^2\times\left\|\boldsymbol{RQ}_{uv}-\overline{\boldsymbol{RQ}}_v\right\|^2}}\,, \tag{6}$$

where $U_{sv}(FR_k)=\{u \mid u \in U(s) \cap U(v) \,\&\, l_{us} \in FR_k \,\&\, l_{uv} \in FR_k\}$. Then the spatial nearest neighbors of service s in R_g are the services in FR_k that have similarity with service s no less than δ_{sim} .

2) **Similarity prediction based on tensor factorization:**

Though the data sparsity problem can be relieved by aggregating similar regions. The rating data in the aggregated region is still sparse. It's still hard to determine similarity relationships between users without enough knowledge of history service experience. Thus it is difficult to find the spatial nearest neighbors for the target user *u*. To solve the data sparsity problem further, a spatial-aware LFM model based on CP decomposition is applied to predict the spatial similarity between services.

The triadic relations among services, neighbors and location features can be formulated as a three-dimensional similarity tensor $\boldsymbol{Sim} \in \Re^{M \times M \times K}$. The element in tensor ***Sim*** is denoted as sim_{ijk}^{SNN} , which represents the spatial similarity of service *i* and service *j* at the aggregated region FR_k . Based on the CP decomposition model, the tensor $\boldsymbol{Sim} \in \Re^{M \times M \times K}$ can be expressed as the inner-product of three *R*-dimensional vectors:

$$\boldsymbol{Sim} \approx [\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{L}] \equiv \sum_{r=1}^{R} s_r \circ v_r \circ l_r \;,1 \tag{7}$$

where *R* is actually the rank of tensor ***Sim***, which is defined as the smallest number of rank-one tensors. $\boldsymbol{S}=[s_1, s_2, ..., s_R]$, $\boldsymbol{V}=[v_1, v_2, ..., v_R]$ and $\boldsymbol{L}=[T_1, T_2, ..., T_R]$. s_r , v_r and l_r represent the latent factor vectors associated with service, neighbor and location, respectively. Actually, s_r and v_r are both user vectors, the tensor ***Sim*** should be symmetric and $sim_{ijk}^{SNN} = sim_{jik}^{SNN}$. So s_r and v_r should be theoretically equivalent. Then equation (7) can be rewritten as follows:

$$\boldsymbol{Sim} \approx \sum_{r=1}^{R} s_r \circ s_r \circ t_r \;1 \quad . \tag{8}$$

As shown in equation (8), compared with the traditional user-item matrix factorization model, we consider not only the latent factors between services, but also the relation with the "geographical trend" reflected in location information. Then the miss spatial similarity can be predicted by equation (9).

$$\hat{sim}_{ijk}^{SNN} = \mu + b_{lk} + \sum_{r=1}^{R} s_{ir} \circ s_{jr} \circ t_{kr} \tag{9}$$

In equation (9), the observed spatial similarity can be broken into two components: biases [52] and service-neighbor-location interaction. The bias component contains the overall average similarity μ and location bias b_{lk} . To learn the involved parameter b_{lk} and the involved vectors, i.e., s_{ir} , s_{jr} and l_{kr} , we minimize the regularized squared error function:

$$\min_{b,s,l} \sum_{(i,j,k) \in Train} \left\| sim_{ijk}^{SNN} - \hat{sim}_{ijk}^{SNN} \right\|^2 + \lambda W \;, \tag{10}$$

where *Train* is the set of the (*i*, *j*, *k*) pairs for sim_{ijk}^{SNN} , which is known as the training set. sim_{ijk}^{SNN} is obtained by equation (6). $W = b_{lk}^2 + \|s_{ir}\|^2 + \|s_{jr}\|^2 + \|l_{kr}\|^2$, which is applied to regularize the learned parameters to avoid overfitting [53] and the constant λ controls the extent of regularization.

In this paper, we adopt stochastic gradient descent to solve equation (10) by looping through all similarity values in the training set. For each given training case, the associated prediction error is denoted as e_{ijk} :

$$e_{ijk} \overset{def}{=} sim_{ijk}^{SNN} - \hat{sim}_{ijk}^{SNN} = sim_{ijk}^{SNN} - \mu - b_{lk} - \sum_{r=1} s_{ir} \cdot s_{jr} \cdot l_{kr} \tag{11}$$

Modifying the parameters by a magnitude proportional to γ (learning rate) in the opposite direction of the gradient, yielding the following recurrence equations:

$$\begin{aligned} b_{lk} &\leftarrow b_{lk} + \gamma(e_{ijk} - \lambda \cdot b_{lk}) \\ \boldsymbol{S}_i &\leftarrow \boldsymbol{S}_i + \gamma(2e_{ijk} \cdot (\boldsymbol{S}_j \circ \boldsymbol{L}_k) - \lambda \cdot \boldsymbol{S}_i) \\ \boldsymbol{L}_k &\leftarrow \boldsymbol{L}_k + \gamma(e_{ijk} \cdot (\boldsymbol{S}_i \circ \boldsymbol{S}_j) - \lambda \cdot \boldsymbol{L}_k) \end{aligned} \tag{12}$$

By iterative learning based on equation (12), the spatial similarity between service *i* and service *j* at aggregated region FR_k can be predicted and obtained. And then the spatial nearest neighbors can be determined based on the spatial similarity and the preset similarity threshold δ_{sim} .

4.3 Rating Prediction

Once the global nearest neighbors (denoted as S_{GNN}) and spatial nearest neighbors (denoted as S_{SNN}) are determined, then the prediction of target user *i* on candidate service *j* at the aggregated region FR_k (denoted as r_{ijk}) is defined in equation (13). The prediction consists of two parts, i.e., prediction based on the global nearest neighbors and prediction based on the spatial nearest neighbors, which are combined by the SAW (Simple Additive Weighting) technique.

$$r_{ijk} = \alpha \cdot \left(\bar{r}_j + \frac{\sum_{s \in S_{GNN}(j)} (r_{is} - \bar{r}_s) \cdot sim_{is}^{GNN}}{\sum_{s \in S_{GNN}(j)} \left| sim_{is}^{GNN} \right|} \right) + (1-\alpha) \cdot \left(\bar{r}_j^k + \frac{\sum_{s' \in S_{SNN}(j,FR_k)} (r_{is'} - \bar{r}_{s'}) \cdot sim_{is'}^{SNN}(FR_k)}{\sum_{s' \in S_{SNN}(j,FR_k)} \left| sim_{is'}^{SNN}(FR_k) \right|} \right) \tag{13}$$

Where α is the weight coefficient, $\bar{r}_j$ is the average rating of service *i*, $\bar{r}_j^k$ is the average rating of service *j* at the aggregated region FR_k , $S_{GNN}(j)$ and $S_{SNN}(j, FR_k)$ respectively represent service *j's* global and spatial nearest neighbor set where the services have been used by user *i*, $\bar{r}_s$ is the overall average rating of services in $S_{GNN}(j)$, and $\bar{r}_{s'}$ is the overall average rating of services in $S_{SNN}(j, FR_k)$.

In our proposal, there are two schemes to determine α , which can be a fixed value or an empirical value. For the fixed scheme, α can be set as a/H , and $1-\alpha = b/H$. Then if the number of region-sensitive QoS metrics is more than the number of region-insensitive QoS metrics, i.e., $a > b$, then the prediction based on the spatial part makes up the larger part of the total. For the empirical scheme, α can be empirically well chosen in the experiment, which will be discussed in detail in Section 5.2.

5 Experiments

In this section, experiments are designed to evaluate the efficiency of our proposal. We first present the experiment settings and then demonstrate the experimental results with detailed analysis.

5.1 Experimental Settings

1) Experimental Setup and Dataset

We implement our method in Java programming language and run it on a cluster server consisting of 17 nodes. Each node has an Intel(R) Xeon(R) CPU E5-2650 (2.6Ghz/30M Cache) processor and 64 GB RAM. In the experiment, we employ two real datasets to evaluate the efficiency of our privacy-preserving location-based prediction algorithm. The two datasets are described as follows:

WSDREAM-Dataset-1 [54]: This dataset is a publicly available QoS dataset of real-world Web services and contains the QoS performance (throughput and response time) of 5825 services from 339 users with location information. This dataset provides two kinds of geographic location information: latitude & longitude, and regions. In our experiment, we only use the region information of users to represent the location of service invocations.

TRIPADVISOR-Dataset: The second dataset is a real-world dataset collected from a well-known travel review site (www. tripadvisor.com), where many travelers give ratings and comments to various travel services. We collect ratings for hotels from 15 regions, which contain the overall ratings and individual QoS ratings (totally six QoS metrics). After cleaning, there are about 1681722 records left, with 76177 users and 6547 hotels. All the ratings for hotels range from 1 to 5, with 5 as the excellent. To analyze the location influence to QoS, we preprocess the rating dataset by aggregating the hotels that have both the same stars and similar tags. After preprocessing, there are about 160 kinds of hotels, and every kind of hotels can appear in different regions.

In our experiment, we use the five-fold cross validation method, and the dataset was split into 80% training data and 20% test data.

2) Comparative Approaches:

To evaluate the effectiveness of our proposal, we compare our method with four alternative approaches:

IPCC [55]: This method is an item-based collaborative filtering method using Pearson Correlation Coefficient, which is widely used in e-commerce scenarios.

LFM [53]: This method is a recommendation model based on matrix factorization, which is proposed to exploit the latent factors of the original data. Both IPCC and LFM have no data obfuscation.

P-UIPCC and P-PMF [42]: These two methods are two QoS prediction methods based on a generic privacy-preserving framework with data obfuscation techniques. In P-UIPCC, the similarity between users is integrated with the similarity between services to make predictions. P-PMF is collaborative recommendation method based on probabilistic matrix factorization model and data obfuscation technique.

3) **Performance Metrics:**

Four widely used performance metrics are applied to evaluate the statistical accuracy of recommendation approaches: mean absolute error (MAE) [56], area under the ROC curve (AUC) [57], precision and recall [58].

MAE is a statistical accuracy metrics used to measure the prediction accuracy, which is defined as the average absolute deviation between the predicted rating and the real rating. Lower MAE presents more accurate predictions. Besides, we apply three classical Information Retrieval (IR) metrics, i.e., AUC, precision and recall, to evaluate Top-N recommendation performance. AUC is the area under the ROC curve, and larger AUC value indicates higher prediction accuracy. The equation of precision and recall are presented as follows:

$$Precision@N = \frac{\sum_{u \in U} |T_u \cap R_u|}{\sum_{u \in U} |R_u|} @N$$

$$Recall@N = \frac{\sum_{u \in U} |T_u \cap R_u|}{\sum_{u \in U} |T_u|} @N ,$$

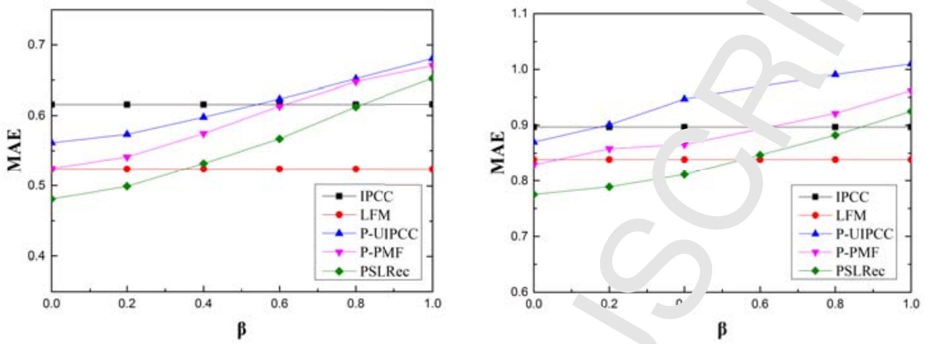
where T_u is the recommendation set of services for user u in train set, R_u is the recommendation set of services for user u in test set.

5.2 Experimental Results

1) **Prediction Effectiveness:**

In our experiment, to evaluate the prediction accuracy of our proposal, we compare our method (denoted as PSLRec) with the four methods. The comparison results are shown in Figs. 2-4.

Comparison in MAE: Fig. 2 (a) and Fig. 2(b) respectively present the result of prediction performance on the two datasets of all approaches in MAE with the change of the random range β. From Fig. 2, we can see that our approach degrades in prediction accuracy (i.e., MAE increases) when β becomes larger, as the observed QoS data is better disguised. However, when β is small (e.g., $\beta \leq 0.8$ in Fig. 2 (a), and $\beta \leq 0.8$ in Fig. 2(b)), our method performs better than IPCC (baseline method with no data obfuscation). Similarly, our method also performs better than LFM when $\beta \leq 0.2$ in Fig. 2 (a), and $\beta \leq 0.4$ in Fig. 2(b). Thus a tradeoff can be made between the prediction accuracy and privacy preservation by setting appropriate β. Additionally, we also observe that our method consistently outperforms P-UIPCC and P-PMF with the same random range, which depicts the effectiveness of considering location influence into recommendation models.



(a) WSDREAM-Dataset-1 (b)TRIPADVISOR-Dataset

Fig. 2 Performance comparison in MAE with the change of random range β.

Comparison in AUC: Fig. 3 provides the prediction performance of different approaches on AUC with the change of β, which is conducted on WSDREAM dataset. We observe that our method degrades in AUC when β becomes larger. When $\beta \leq 0.2$, the prediction accuracy of our proposal is better than LFM, and when $\beta \leq 0.2$, our method also performs better than IPCC. It suggests that an appropriate value for β should be selected to make a good balance between the prediction accuracy and user privacy. However, we find that the advantage of our method over P-UIPCC and P-PMF decreases with the increase of β, which indicates that data obfuscation have worse effect to prediction accuracy with the increase of β.

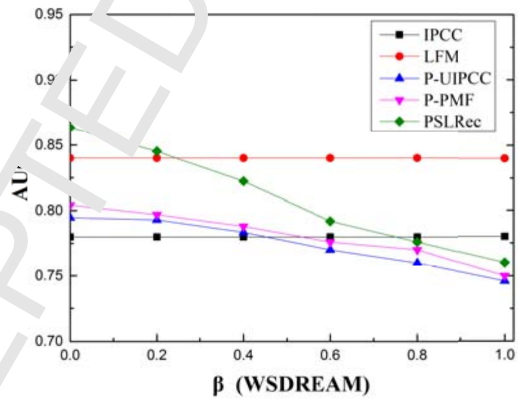
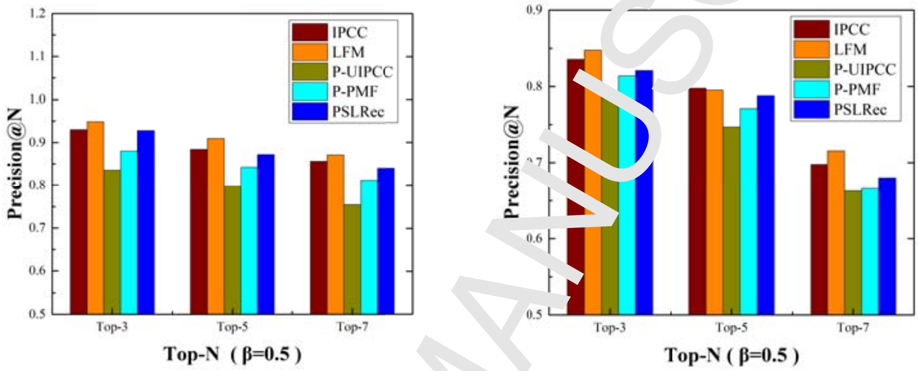


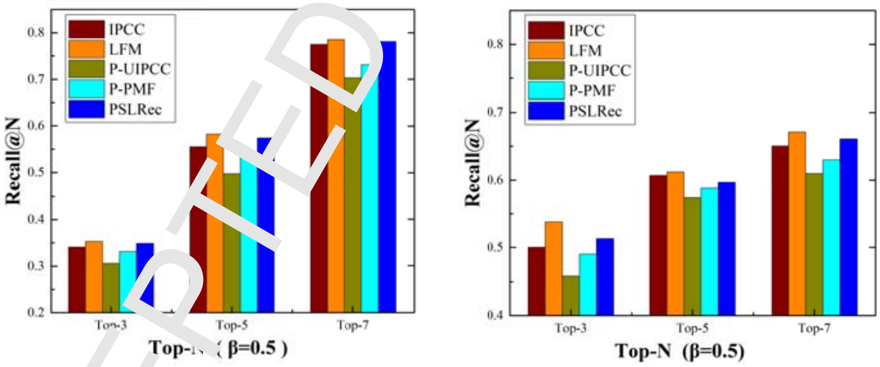
Fig. 3 Performance comparison in AUC with the change of random range β.

Comparison in Precision and Recall: Fig. 4 and Fig. 5 depict the performance of Top-N (N=3, 5, 7) recommendations of all approaches, where β is set as 0.5. And the QoS data of users could be well disguised when $\beta = 0.5$. Fig. 4 shows the precision@N performance, and Fig. 5 shows the recall@N performance. In both datasets, the precision of our method degrades with the increase of N, while the recall upgrades. From Fig. 4 and Fig. 5, we can observe that our method, i.e., PSLRec, outperforms both P-UIPCC and P-PMF in both precision and recall. For example, in case of the WSDREAM dataset, the precision@7 and recall@7 of PSLRec are respectively 5.2% and 7.6% higher than P-PMF. Besides, we can find that when β=0.5, the prediction accuracy of our proposal is close to IPCC and a little lower than LFM. And on the TRIPADVISOR dataset, the precision@7 and recall@N of PSLRec are only 2.5% and 2.3% lower than LFM. The results present that our proposal could achieve decent accuracy on the premise of privacy preservation.



(a) WSDREAM-Dataset-1 (b)TRIPADVISOR-Dataset

Fig. 4 Performance comparison in Top-N prediction accuracy on precision@N.



(a) WSDREAM-Dataset-1 (b)TRIPADVISOR-Dataset

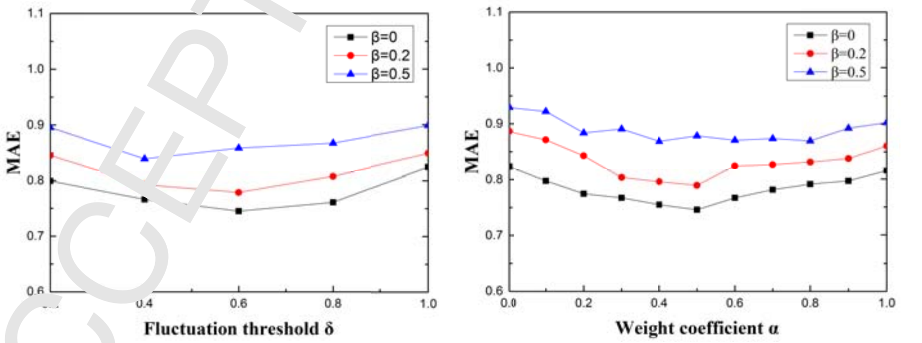
Fig. 5 Performance comparison in Top-N prediction accuracy on recall@N.

2) **Impact of Parameter Settings:**

In this section, we discuss the impact of parameter settings to prediction accuracy under the TRIPADVISOR-Dataset. Here, we mainly analyze the impact of the fluctuation threshold δ and the weight coefficient α associated with random range β.

Impact of δ ***:*** The fluctuation threshold δ is used to determine whether a QoS metric is a region-sensitive QoS metric or not. Different δ may bring different division of region-sensitive QoS and region-insensitive QoS. To evaluate the effect of δ to prediction performance, we vary δ from 0.2 to 1 with the step of 0.2. The region similarity threshold θ and the weight coefficient α are set as 0.4 and 0.5, respectively. Fig. 6 (a) shows how δ effect the final prediction performance. Here, we get MAE of our method in case of $\beta = 0$ (no data obfuscation), $\beta = 0.2$ and $\beta = 0.5$. The results show that the MAE curve changes similarly when $\beta = 0$ and $\beta = 0.2$, and MAE reaches the optimal point when δ is about 0.6 in both cases. And MAE reaches the optimal point when δ is about 0.4 in case of $\beta = 0.5$. The trend of MAE curve in case of $\beta = 0.5$ is inconsistent with the MAE curves of other two cases. It implies that larger random range may have more uncertain effects.

Impact of α ***:*** The weight coefficient α determines whether prediction based on global neighbors or prediction based on spatial neighbors plays the larger part in the final prediction. In our proposal, we give two schemes to decide α. As presented in Section 4.3, α can be a fixed value a/H or an empirical value. Here, we study the impact of α to the final performance by varying α from 0 to 1 with a step of 0.1. δ and θ are set as 0.6 and 0.4, respectively. α =0 and α =1 are two special cases (That α =0 considers only prediction based on the global part, and α =1 considers only prediction based on the spatial part). Actually, we can get that a/H=0.5 (a=3 and H=6) when δ=0.6 and θ =0.4. Thus the two schemes for α adopted in our proposal are coincident. Through this study, we can observe that the distinguishing of region-sensitive QoS metrics from region-insensitive QoS metrics is necessary and meaningful. In addition, from Fig. 6 (b), we can see that MAE reaches the optimal point when α falls around 0.5 in case of $\beta = 0$ and $\beta = 0.2$. It can be found that the optimal experimental value of α and a/H are approximately equal. However, when β is set as 0.5, the trend of MAE curve is completely inconsistent with the MAE curves of other two cases, and there's no determined optimal point. It implies that the impact of different parameter settings to prediction accuracy may be weakened with larger β. Thus to get decent prediction accuracy on the premise of privacy preservation, β and other critical parameters such as δ and α should be well chosen.



(a) θ =0.4, α =0.5 (b) δ =0.6, θ =0.4

Fig. 6 Parameters impact on prediction performance (MAE). (a) Impact of the fluctuation threshold δ. (b) Impact of the weight coefficient α.

3) Security and Privacy Analysis:

This section presents the security and privacy analysis of our approach. As demonstrated in Section 4.1, to overcome the security issues in location-aware recommendation models, two privacy-preserving mechanisms are proposed to protect the observed QoS data and specific location information of users in a more secure manner. However, more secure privacy mechanisms may lead to less accurate prediction performance. Thus it is important to make a trade-off between prediction accuracy and privacy preservation. In the following, we analyze the impact of the random range β (in data obfuscation technique) and the region similarity threshold θ (in region aggregation strategy) to privacy preservation and prediction accuracy.

From Fig. 2 and Fig. 3, we can see that our approach (lower MAE and AUC indicates worse prediction performance) degrades in prediction accuracy when β becomes larger (higher β indicates that the observed QoS data is better disguised). Thus an appropriate value for β should be set. On the WSDREAM dataset, our approach performs better than IPCC and LFM when $\beta \leq 0.8$ and $\beta \leq 0.3$, respectively. On the TRIPADVISOR dataset, our approach performs better than IPCC and LFM when $\beta \leq 0.8$ and $\beta \leq 0.4$, respectively. Moreover, we could find that our approach consistently outperforms or no worse than P-UIPCC and P-PMF with the same random range. Then we could set an appropriate β (For example, here, β cloud be set as 0.3 on WSDREAM dataset and 0.6 on TRIPADVISOR dataset, respectively) to get a good balance between privacy preservation and prediction accuracy.

Based on the region aggregation algorithm, the location information used in our prediction model is no longer the specific locations, but only the fuzzy location information (a region number) which makes it difficult for the attackers to get the real location information of users. As described in Algorithm 1, to blur the specific location information, an appropriate value for the region similarity threshold θ should be chosen. We can see that if the region similarity threshold θ is small, the number of aggregated region clusters is also small, which weakens location features but could better hide user's specific location information. While if θ is large, the location influence to recommendations is emphasized but the location information of users may be disclosed. Thus an appropriate θ should also be set. For example, based on the above experimental analysis on TRIPADVISOR dataset, when θ is set as 0.4 in this case, we could get acceptable prediction accuracy with the specific location information blurred.

The above security analysis depicts that a good balance could be made between the prediction accuracy and the goal of privacy preservation based on our proposal.

6 Conclusions

In this paper, a privacy-preserving and sparsity-aware location-based prediction model is proposed for collaborative recommender systems. It aims to provide the most beneficial products for users with the consideration of privacy preservation. First, location influence is considered into the classical neighborhood-based collaborative prediction model by distinguishing region-sensitive QoS metrics from region-insensitive QoS metrics. Accordingly, global nearest neighbors and spatial nearest neighbors are defined. To protect user privacy, a data obfuscation technique is utilized to disguise the observed QoS data of users. In addition, a region aggregation algorithm is presented to deal with the data sparsity problem and blur the specific location information of users. After data obfuscation and region aggregation, a location-aware LFM model based on tensor factorization is applied to mine the spatial similarity relationships between services and identify the spatial nearest neighbors. Then, the final predictions can be made by combining the predictions based on the global nearest neighbors and spatial nearest neighbors. Finally, experiments based on real-world datasets are conducted to demonstrate the effectiveness of our proposal. In our future work, we will do further study collaborative recommendation models based on multi-model integration and consider more privacy-preservation mechanisms.

Acknowledgements

This paper is partially supported by the National Science Youth Foundation of China under Grant No. 61702264 and No. 61702277, the Open Research Project of State Key Laboratory of Novel Software Technology (Nanjing University) under Grant No. KFKT2017B07.

References

[1] Liu J, Jiang Y, Li Z, et al. Domain-Sensitive Recommendation with User-Item Subgroup Analysis. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(4): 939-950.

[2] ur Rehman M H, Ahmed E, Yaqoob I, et al. Big Data Analytics in Industrial IoT Using a Concentric Computing Model. IEEE Communications Magazine, 2018, 56(2): 37-43.

[3] Meng S, Wang H, Qi L, et al. Spatial-Temporal Aware Intelligent Service Recommendation Method Based on Distributed Tensor factorization for Big Data Applications. IEEE Access 2018, 6, 59462-59474.

[4] Meng S, Dou W, Zhang X, et al. KASR: A Keyword-Aware Service Recommendation method on MapReduce for big data applications. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(12): 3221-3231.

[5] Mehmood Y, Haider N, Imran M, et al. M2m Communications in 5g: State-of-the-art Architecture, Recent Advances, and Research Challenges. IEEE Communications Magazine, 2017, 55(9): 194-201.

[6] Lian D, Zheng K, Ge Y, et al. GeoMF++: Scalable Location Recommendation via Joint Geographical Modeling and Matrix Factorization. ACM Transactions on Information Systems (TOIS), 2018, 36(3): 33.

[7] Bagci H, Karagoz P. Context-aware Location Recommendation by Using a Random Walk-based Approach. Knowledge and Information Systems, 2016, 47(2): 241-260.

[8] Xu X, Xue Y, Qi L, et al. An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles. Future Generation Computer Systems, 2019.

[9] Wen J, Zhou Z, Liu Z, et al. Sharp Sufficient Conditions For Stable Recovery of Block Sparse Signals by Block Orthogonal Matching Pursuit. Applied and Computational Harmonic Analysis, 2018, to be published.

[10] Hu Y, Peng Q, Hu X, et al. Time aware and Data Sparsity Tolerant Web Service RecomMendation based on Improved Collaborative filtering. IEEE Transactions on Services Computing, 2015, 8(5): 782-794.
[11] Wang X, Zhong Y, Zhang L, et al. Spatial Group Sparsity Regularized Nonnegative Matrix Factorization for Hyperspectral Unmixing. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(11): 6287-6304.
[12] Gao Z, Zhang H, Xu G P, et al. Multi-view Discriminative and Structured Dictionary Learning With Group Sparsity For Human Action Recognition. Signal Processing, 2015, 112: 83-97.
[13] Peng M, Zeng G, Sun Z, et al. Personalized App Recommendation based on App Permissions. World Wide Web, 2018, 21(1): 89-104.
[14] Qi L, Meng S, Zhang X, et al. An Exception Handling Approach for Privacy-Preserving Service Recommendation Failure in a Cloud Environment. Sensors, 2018, 18(7): 2037.
[15] Wu Z, Li G, Liu Q, et al. Covering the sensitive subjects to protect personal privacy in personalized recommendation. IEEE Transactions on Services Computing, 2018, 11(3): 493-506.
[16] Ding S, Qu S, Xi Y, et al. A long video caption generation algorithm for big video data retrieval. Future Generation Computer Systems, 2019, 93, 583-595.
[17] Zhu X, Jing X Y, Wu D, et al. Similarity-Maintaining Privacy Preservation and Location-aware Low-rank Matrix Factorization for QoS Prediction Based Web Service Recommendation. IEEE Transactions on Services Computing, 2018, to be published, DOI: 10.1109/TSC.2018.2839741.
[18] Lian D, Ge Y, Zhang F, et al. Scalable Content-Aware Collaborative Filtering for Location Recommendation. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1122-1135.
[19] Yin H, Cui B, Chen L, et al. Modeling Location-based User Rating Profiles for Personalized Recommendation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2015, 9(3): 19.
[20] Chen X, Zheng Z, Yu Q, et al. Web Service Recommendation via Exploiting Location and QoS Information. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(7): 1913-1924.
[21] Jiang S, Qian X, Mei T, et al. Personalized Travel Sequence Recommendation on Multi-Source Big Social Media. IEEE Transactions on Big Data, 2016, 2(1): 43-56.
[22] Yin H, Cui B, Zhou X, et al. Joint Modeling of User Check-in Behaviors for Real-time Point-of-Interest Recommendation. ACM Transactions on Information Systems (TOIS), 2016, 35(2): 11.
[23] Yin H, Zhou X, Cui B, et al. Adapting to User Interest Drift for Poi Recommendation. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2566-2581.
[24] He X, Liao L, Zhang H, et al. Neural collaborative filtering. Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 173-182.
[25] Liu Y, Wei W, Sun A, et al. Exploiting Geographical Neighborhood Characteristics For Location Recommendation. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014: 739-748.
[26] Zhang C, Liang H, Wang K. Trip Recommendation Meets Real-world Constraints: POI Availability, Diversity, and Traveling Time Uncertainty. ACM Transactions on Information Systems (TOIS), 2016, 35(1): 5.
[27] He T, Yin H, Chen Z, et al. A Spatial-Temporal Topic Model for the Semantic Annotation of POIs in LBSNs. ACM Transactions on Intelligent Systems and Technology (TIST), 2016, 8(1): 12.

[28] Yuan Q, Cong G, Zhao K, et al. Who, Where, When, and What: A Nonparametric Bayesian Approach to Context-Aware Recommendation and Search for Twitter Users. ACM Transactions on Information Systems (TOIS), 2015, 33(1): 2.

[29] Wang X, Zhu J, Zheng Z, et al. A Spatial-Temporal QoS Prediction Approach for Time-aware Web Service Recommendation. ACM Transactions on the Web (TWEB), 2016, 10(1): 7.

[30] Zhao S, Zhao T, Yang H, et al. STELLAR: Spatial-Temporal Latent Ranking for Successive Point-of-Interest Recommendation. AAAI. 2016: 315-322.

[31] Yang D, Zhang D, Zheng V W, et al. Modeling User Activity Preference By Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(1): 129-142.

[32] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer, 2009 (8): 30-37.

[33] He X, Zhang H, Kan M Y, et al. Fast Matrix Factorization for Online Recommendation With Implicit Feedback. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016: 549-558.

[34] Zhang W, Sun H, Liu X, et al. Temporal QoS-aware web service recommendation via non-negative tensor factorization. Proceedings of the 23rd international conference on World wide web, 2014: 585-596.

[35] Shin K, Sael L, Kang U. Fully Scalable Methods For Distributed Tensor Factorization. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 100-113.

[36] Shi Y, Larson M, Hanjalic A. Mining contextual movie similarity with matrix factorization for context-aware recommendation. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(1): 16.

[37] Yao L, Wang X, Sheng Q Z, et al. Mashup Recommendation by Regularizing Matrix Factorization with API Co-Invocations. IEEE Transactions on Services Computing, 2018, to be published, DOI: 10.1109/TSC.2018.2803171.

[38] Qi L, Xiang H, Dou W, et al. Privacy-Preserving Distributed Service Recommendation based on Locality-Sensitive Hashing. 2017 IEEE International Conference on Web Services (ICWS), 2017: 49-56.

[39] Li L, Liu A, Li Q, et al. Privacy-preserving Collaborative Web Services QoS Prediction via Yao's Garbled Circuits and Homomorphic Encryption. Journal of Web Engineering, 2016, 15(3-4): 203-225.

[40] Kaur H, Kumar N, Batra S. An Efficient Multi-party Scheme For Privacy Preserving Collaborative Filtering For Healthcare Recommender System. Future Generation Computer Systems, 2018, to be published.

[41] Li D, Chen C, Lv Q, et al. An Algorithm For Efficient Privacy-preserving Item-based Collaborative Filtering. Future Generation Computer Systems, 2016, 55: 311-320.

[42] Zhu J, He P, Zheng Z, et al. A Privacy-preserving QoS Prediction Framework for Web Service Recommendation. 2015 IEEE International Conference on Web Services (ICWS), 2015: 241-248.

[43] Boutet A, Frey D, Guerraoui R, et al. Privacy-preserving Distributed Collaborative Filtering. Computing, 2016, 98(8): 827-846.

[44] Polatidis N, Georgiadis C K, Pimenidis E, et al. Privacy-preserving Collaborative Recommendations Based on Random Perturbations. Expert Systems with Applications, 2017, 71: 18-25.

[45] Nikolaenko V, Ioannidis S, Weinsberg U, et al. Privacy-preserving Matrix Factorization. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. ACM, 2013: 801-812.

[46] Casino F, Domingo-Ferrer J, Patsakis C, et al. A K-anonymous Approach to Privacy Preserving Collaborative F5iltering. Journal of Computer and System Sciences, 2015, 81(6): 1000-1011.

[47] Gao Z, Wang D, Wan S, et al. Cognitive-inspired class-statistic matching with triple-constrain for camera free 3D object retrieval. Future Generation Computer Systems, 2019, 94, 641-653.
[48] Wan S, Zhao Y, Wang T, et al. Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things. Future Generation Computer Systems, 2019, 91, 382-391.
[49] Kolda T G, Bader B W. Tensor decompositions and applications. SIAM review, 2009, 51(3): 455-500.
[50] Polat H, Du W. Privacy-preserving Collaborative Filtering Using Randomized Perturbation Techniques. Third IEEE International Conference on Data Mining (ICDM), 2003: 625-628.
[51] Bhatia S K. Adaptive K-Means Clustering. FLAIRS conference. 2004: 695-699.
[52] Koren Y. Collaborative Filtering with Temporal Dynamics. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 447-456.
[53] Wen J, Wang J, Zhang Q. Nearly Optimal Bounds For Orthogonal Least Squares. IEEE Trans. Signal Process, 2017, 65(20): 5347-5356.
[54] Zheng Z, Zhang Y, Lyu M R. Investigating QoS of real-world web services. IEEE transactions on services computing, 2014, 7(1): 32-39.
[55] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003 (1): 76-80.
[56] Chai T, Draxler R R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?–Arguments Against Avoiding RMSE in the Literature. Geoscientific Model Development, 2014, 7(3): 1247-1250.
[57] Bradley A P. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern recognition, 1997, 30(7): 1145-1159.
[58] Carullo G, Castiglione A, De Santis A, et al. A Triadic Closure and Homophily-based Recommendation System for Online Social Networks. World Wide Web, 2015, 18(6): 1579-1601.

Shunmei Meng received her PhD degree in Department of Computer Science and Technology from Nanjing University, China, in 2016. Now, she is an assistant professor of School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China. She has published papers in international journals and international conferences such as TPDS, ICWS, and ICSOC. Her research interests include recommender systems, service computing, and cloud computing.

Lianyong Qi received his PhD degree in Department of Computer Science and Technologyfrom Nanjing University, China, in 2011. Now, he is an associate professor of the School of Information Science and Engineering, Chinese Academy of Education Big Data, Qufu Normal University, China. He has already published more than 30 papers including JSAC, TCC, TBD, FGCS, JCSS, CCPE. His research interests include recommender systems and services computing.

Qianmu Li received the BSc and PhD degrees from Nanjing University of Science and Technology, China, in 2001 and 2005, respectively. He is a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include information security, computing system management, and data mining. He received the China Network and Information Security Outstanding Talent Award and multiple Education Ministry Science and Technology Awards.

Wenmin Lin received her PhD degree in Department of Computer Science and Technology from Nanjing University, China, in 2015. Now, she is an assistant professor of Department of Computer Science, Hangzhou Dianzi University, China. She has published papers in international journals and international conferences. Her research interests Block chain, and service computing.

Xiaolong Xu received his B.Sc. and M.Sc. degrees in computer science from Nanjing University of Information Science and Technology, in 2010 and 2013, respectively, and the Ph.D. degree from the Nanjing University, China, in 2016. He worked as a research scholar at Michigan State University, USA, from Apr. 2017 to May 2018. He is currently an assistant professor with the school of computer and software, Nanjing University of Information Science and Technology. He has published more than 40 peer review papers in international journals and conferences. His research interests include mobile computing, edge computing, IoT, cloud computing and big data.

Shaohua Wan received Ph.D. degree from School of Computer, Wuhan University in 2010. From 2015, he worked as a postdoc at State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology. From 2016 to 2017, he worked as a visiting scholar at Department of Electrical and Computer Engineering at Technical University of Munich, Germany. At present, he is an associate professor and master advisor at school of Information and Safety Engineering, Zhongnan University of Economics and Law. His research interests include massive data computing for Internet of Things and edge computing.

1. Shunmei Meng

2. Lianyong Qi

3. Qianmu Li

4. Wenmin Lin

5. Xiaolong Xu

6. Shaohua Wan

Highlight

- Location influence is considered into neighborhood-based collaborative recommendation model by distinguishing region-sensitive QoS metrics from region-insensitive QoS metrics.
- A privacy-aware region aggregation is proposed to deal with the data sparsity problem and protect user privacy.
- A location-aware latent factor model based on tensor factorization is proposed to identify the spatial nearest neighbors.
- Extensive experiments are conducted on real-world dataset.