GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations

Jia-Dong Zhang
Department of Computer Science
City University of Hong Kong
jzhang26-c@my.cityu.edu.hk

Chi-Yin Chow
Department of Computer Science
City University of Hong Kong
chiychow@cityu.edu.hk

ABSTRACT

Recommending users with their preferred points-of-interest (POIs), e.g., museums and restaurants, has become an important feature for location-based social networks (LBSNs), which benefits people to explore new places and businesses to discover potential customers. However, because users only check in a few POIs in an LBSN, the user-POI check-in interaction is highly sparse, which renders a big challenge for POI recommendations. To tackle this challenge, in this study we propose a new POI recommendation approach called GeoSoCa through exploiting *geographical correlations*, *social correlations* and *categorical correlations* among users and POIs. The geographical, social and categorical correlations can be learned from the historical check-in data of users on POIs and utilized to predict the relevance score of a user to an unvisited POI so as to make recommendations for users. First, in GeoSoCa we propose a kernel estimation method with an adaptive bandwidth to determine a personalized check-in distribution of POIs for each user that naturally models the geographical correlations between POIs. Then, GeoSoCa aggregates the check-in frequency or rating of a user's friends on a POI and models the social check-in frequency or rating as a power-law distribution to employ the social correlations between users. Further, GeoSoCa applies the bias of a user on a POI category to weigh the popularity of a POI in the corresponding category and models the weighed popularity as a power-law distribution to leverage the categorical correlations between POIs. Finally, we conduct a comprehensive performance evaluation for GeoSoCa using two large-scale real-world check-in data sets collected from Foursquare and Yelp. Experimental results show that GeoSoCa achieves significantly superior recommendation quality compared to other state-of-the-art POI recommendation techniques.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Information Filtering

Keywords

Location-based social networks; point-of-interest recommendations; geographical correlations; social correlations; categorical correlations; popularity

SIGIR'15, August 09 - 13, 2015, Santiago, Chile

DOI: http://dx.doi.org/10.1145/2766462.2767711.

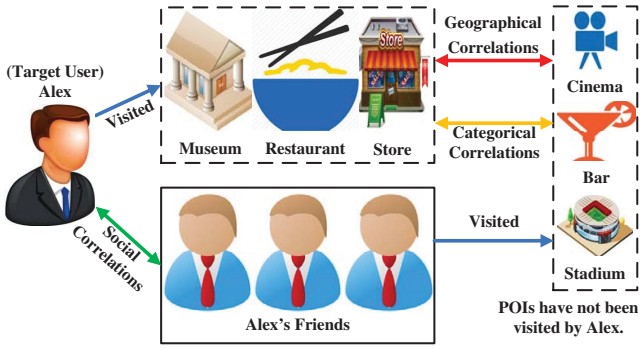


Figure 1: Three important types of correlations

1. INTRODUCTION

As an increasingly popular application of location-based services, location-based social networks (LBSNs), such as Foursquare and Yelp, have attracted millions of users to check in their preferred points-of-interest (POIs), e.g., museums, restaurants and stores, and share their experiences of visiting these POIs with friends. For example, as of December 2014, Foursquare had over 6 billion check-ins with millions more every day contributed by more than 55 million users worldwide. These historical check-in data incorporate rich information about users and POIs, and thus bring new opportunities to mine the user visiting preferences for personalized POI recommendations that not only help users explore new POIs but also benefit for businesses to discover potential customers.

In the problem of personalized POI recommendations, the key tasks are to estimate the preference or relevance scores of a user to her unvisited POIs and return the POIs with the top-k highest preference scores for the user. Most existing POI recommendation methods (e.g., [3, 5, 6, 10, 11, 14, 25, 26, 36]) apply the traditional memory-based or model-based collaborative filtering techniques with the user-POI check-in matrix to compute the preference score between a user and an unvisited POI. However, the user-POI check-in matrix is highly sparse because users have only visited a very small proportion of POIs in an LBSN. As a result, these methods usually suffer from low recommendation quality.

In this paper, we predict the preference scores of any target user to her unvisited POIs based on **three important types of correlations among the user and unvisited POIs** that can be derived from the historical check-in data of users on POIs, as depicted in Figure 1. (1) **Geographical correlations.** The first law of geog-

raphy [17] states that: "Everything is related to everything else, but near things are more related than distant things." For example, in reality a person usually visits a POI, e.g., museums, and then travel to its nearby POIs, e.g., restaurants and stores. That is, the close POIs have the stronger geographical correlations than the POIs that are far from each other. Thus, we can estimate a user's relevance score for an unvisited POI based on the geographical correlations between *the user's visited POIs* and *unvisited POI*. (2) **Social correlations.** In the real world, a person may prefer POIs highly recommended by her friends. For instance, they often visit POIs, e.g., museums, restaurants or stores together. In other words, social friends are more likely to share common interests on POIs than strangers. Likewise, in LBSNs, users establish social links with each other to share their experiences of visiting POIs. Hence, we can compute the relevance score of a user for an unvisited POI in terms of the social correlations between *the user* with *her friends* who have visited the POI. (3) **Categorical correlations.** LBSNs often predefine a universal set of categories and attach each POI to a certain subset of categories. The category of a POI reflects its usual business activities and nature. In reality people have different biases on the categories of POIs: a foodie often visits restaurants to taste a variety of food, and a tourism enthusiast usually travels on tourism attractions all over the world. Accordingly, we can deduce the relevance score of a user to an unvisited POI based on the categorical correlations in the categories of *the user's visited POIs* and *the unvisited POI*.

Therefore, we are motivated to propose a new probabilistic approach for point-of-interest recommendations through exploiting and integrating **Geo**graphical, **So**cial and **Ca**tegorical correlations, called **GeoSoCa**. (1) **Geographical correlation modeling.** With the visited POIs of a user, we estimate a personalized check-in distribution over the geographical latitude and longitude coordinates for the user; the personal check-in distribution naturally models the geographical correlations between *the user's visited POIs* and *her unvisited POIs* and is able to compute the geographical relevance score of the user to any unvisited POI. (2) **Social correlation modeling.** Given an unvisited POI of a user, we first aggregate the check-in frequency or rating on the POI from her social friends; then the social check-in frequency (or rating) is transformed into a social relevance score of the user to the unvisited POI, based on the social check-in frequency (or rating) distribution that is estimated from the historical check-in data of all users. (3) **Categorical correlation modeling.** At first, we derive the bias of a user to a certain category according to her visited POIs; then the bias is used to weigh the popularity of an unvisited POI in the corresponding category. Further, the weighed popularity for the user to the unvisited POI is also mapped into a categorical relevance score based on the popularity distribution estimated from the historical check-in data of all users. The categorical correlation also takes into account the POI popularity from all users which indicates the quality of the POI and benefits for POI recommendations.

The main contributions of this study can be summarized:

- To model the geographical correlations, we extend the kernel density estimation by applying an adaptive bandwidth that is learned from the underlying check-in data. Our adaptive kernel estimation method can improve the predictive ability of the estimated check-in distribution for a user, in comparison to the kernel density estimation with a fixed bandwidth [30, 31, 33, 34] and the common distance distribution for all users [9, 13, 15, 22, 23, 28, 29]. (Section 3.2)

- To model the social correlations, we develop a method to estimate the social check-in frequency or rating by a user's friends to a POI as a power-law distribution that is learned from the historical check-in data of all users. Our method is distinct from the current works [2, 4, 20, 21, 23, 27, 30, 31, 33, 34, 35] that derive the similarities between users in terms of their social links and then integrate them into the traditional collaborative filtering techniques. (Section 3.3)

- To model the categorical correlations, we devise a method to combine the category bias of a user and the popularity of a POI into a relevance score between the user and POI in terms of the estimated popularity distribution from the historical check-in data of all users. Our method is different from the existing works [1, 8, 15, 18, 27, 35] that separately utilize the category and/or popularity information of POIs. (Section 3.4)

- To the best of our knowledge, our proposed GeoSoCa is the first study to integrate the geographical, social, categorical and popularity information for POI recommendations. (Section 3.5)

- We conduct extensive experiments to evaluate the recommendation accuracy of GeoSoCa using two large-scale real-world data sets collected from Foursquare and Yelp. Experimental results show that GeoSoCa significantly outperforms other state-of-the-art POI recommendation techniques. (Section 4)

The rest of this paper is organized as follows. We review the related work on POI recommendations in Section 2. Our proposed methods to model geographical, social and categorical correlations are presented in Section 3, followed by experimental evaluation in Section 4. Finally, Section 5 concludes this paper.

2. RELATED WORK

This section reviews existing POI recommendation techniques on how they employ the geographical, social, categorical, and popularity information.

POI recommendations using geographical information. Based on the fact that the geographical proximity significantly affects the check-in behaviors of users on the POIs, the geographical information has been intensively used in POI recommendations. One way is to simply consider the current locations of users to filter out the POIs that are far from the users [1, 4, 20, 14]. Another way is to apply the geographical latent factor or topic models to derive latent features of regions or POIs [7, 8, 12, 16, 25, 26]. The more sophisticated way is to estimate the geographical correlations of check-in POIs as a common distance distribution for all users, e.g., a multi-center Gaussian distribution [2], a power-law distribution [9, 13, 15, 22, 23, 28, 29], or a personalized non-parametric distribution for each user [30, 32, 33]. In particular, the recent works [31, 34] employed the kernel density estimation method with the fixed bandwidth to model the geographical check-in distribution of POIs for each user over the latitude and longitude coordinates. Further, in this paper we develop an adaptive kernel estimation method to enhance the ability of the obtained check-in distributions to predict the relevance score between a user and an unvisited POI.

POI recommendations using social information. Social links between users have also been widely utilized to improve the quality of location-based recommender systems, since the social friends are more likely to share common interests on POIs than strangers. Most current works derive the similarities between users from social links and put them into the traditional memory-based or model-

based collaborative filtering techniques. For example, some literatures [4, 20, 23, 27, 30, 31, 33, 34] seamlessly integrated the similarities of users into the user-based collaborative filtering techniques, while others [2, 21, 35] employed the user similarities as the regularization terms or weights of latent factor models. In this paper, we contrive a new method to exploit the social correlations between users by aggregating the check-in frequency or rating of friends to POIs and transforming them into relevance scores, based on the estimated social check-in frequency or rating distribution from the historical check-in data of all users.

POI recommendations using categorical information. The categories of POIs visited by a user implicitly indicate the activities of the user in the POIs. For instance, a person checking in a cinema means that she is watching a movie there. Thus, the category information of POIs is useful for modeling the specific preference of a user. However, there are only a few studies that utilize the category information for POI recommendations. Rahimi and Wang [18] simply identified the preference of a user to a POI with the bias of the user to the category of the POI. Bao et al. [1] calculated the category biases of users to compute the similarity of the users for the user-based collaborative filtering method. Besides *the category biases of users*, Ying et al. [27] also derived *the category weights of POIs* from the tags annotated on the POIs and then estimated the relevance scores between users and POIs based on the inner product of the category biases and weights. Liu et al. [15] clustered POIs into groups based on their categories, built a user-category transition matrix instead of user-POI check-in matrix from the historical check-in data of users, and applied the matrix factorization technique to discover the next top-k categories that a user would like to check in. Zhao et al. [35] clustered users into communities and represented each community as a weighted category vector, in which each dimension is the check-in counts of a particular POI category by users in the community; to apply the user-based collaborative filtering method, they further computed the similarity of users according to the category vectors of the communities of the users. Hu et al. [8] leveraged the matrix factorization technique to associate each category with a latent vector and deuced the relevance score of a user to a POI based on the latent vectors of the categories of the POI.

POI recommendations using popularity information. The popularity of a POI reflects the quality of products or services provided by the POI. For example, a restaurant receiving a lot of visits from customers indicates that the restaurant provides high-quality foods for its customers. Thus, the popularity of POIs is also helpful for POI recommendations. Most existing works regard the popularity of a POI as the universal prior preference of users to the POI. Specifically, the study [27] employed the prior preference of users to unvisited POIs as the weight of the edges between the users and POIs in the complete bipartite graph, while other studies utilize *the prior preference* to adjust *the posterior preference* that are derived from the geographical information [8, 13, 14, 28]. However, in these studies the prior preference is not personalized for users and thus in practice the benefit from the popularity is considerably limited. On the other hand, the current works [1, 8, 15, 18, 27, 35] **separately** modeled the effect of the category and popularity of POIs and may not take full advantage of them for POI recommendations. In this paper, we devise a new method to combine the category bias of a user and the popularity of a POI into a relevance score between the user and POI so as to personalize the effect of the popularity of the POI on the user.

3. MODELING CORRELATIONS FOR POI RECOMMENDATIONS

Table 1: Key Notations in the Paper

Symbol	Meaning
U	Set of all users in an LBSN
u	Some user: $u \in U$
L	Set of POIs in an LBSN
l	Some POI with a pair of geographical latitude and longitude coordinates (x,y): $l \in L$
C	Set of categories of POIs in an LBSN
c	Some category: $c \in C$
$\mathbf{R}_{\|U\|\times\|L\|}$	Check-in matrix: $\mathbf{R}_{u,l}$ is the check-in frequency or rating of user u on POI l
$\mathbf{S}_{\|U\|\times\|U\|}$	Social link matrix: if $u,u' \in U$ have a social link, $\mathbf{S}_{u,u'}=1$; otherwise, $\mathbf{S}_{u,u'}=0$.
$\mathbf{B}_{\|U\|\times\|C\|}$	Categorical bias matrix: $\mathbf{B}_{u,c}$ is the frequency of user u visiting category c
$\mathbf{P}_{\|C\|\times\|L\|}$	Popularity matrix: $\mathbf{P}_{c,l}$ is the popularity of POI l in category c

In this section, we introduce the research problem with the required data structures (Section 3.1), model the geographical, social, and categorical correlations (Sections 3.2, to 3.4), and integrate these three correlations to recommend POIs to users (Section 3.5).

3.1 Problem Statement

Here we define the data structures and the research problem in this paper. These data structures can be extracted from the rich information in an LBSN, incorporating the historical check-in data of users on POIs, social links between users, the categories of POIs, and the geographical latitude and longitude coordinates of POIs. Table 1 lists the key symbols used in this paper.

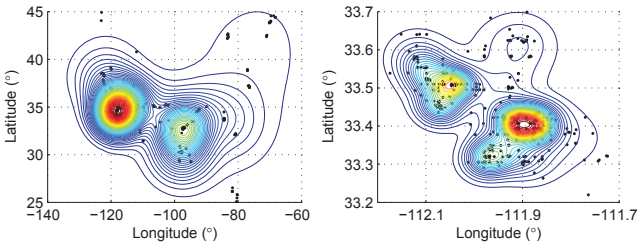
DEFINITION 1 (**CHECK-IN MATRIX**). Given the historical check-in data of users on POIs from an LBSN, we can easily build a check-in matrix $\mathbf{R}_{|U|\times|L|}$, in which each entry $\mathbf{R}_{u,l}$ represents the check-in frequency (e.g., Foursquare) or rating (e.g., Yelp) of user $u \in U$ on location $l \in L$, and U and L are the sets of users and POIs in the LBSN, respectively. Note that most entries in $\mathbf{R}$ are zero, since users have only visited a very small proportion of POIs in the LBSN.

DEFINITION 2 (**SOCIAL LINK MATRIX**). Given the social links between users from an LBSN, it is easy to construct a social link matrix $\mathbf{S}_{|U|\times|U|}$, in which if there exists a social link between two different users $u,u' \in U$, $\mathbf{S}_{u,u'}=1$; otherwise, $\mathbf{S}_{u,u'}=0$.

DEFINITION 3 (**CATEGORICAL BIAS MATRIX**). Given the historical check-in data of users on POIs and the categories of POIs from an LBSN, we construct a categorical bias matrix $\mathbf{B}_{|U|\times|C|}$, in which each entry $\mathbf{B}_{u,c}$ represents the frequency of user u visiting the POIs that belong to category $c \in C$, and C is the universal set of categories of POIs that is often predefined in the LBSN. Note that a POI could belong to multiple categories.

DEFINITION 4 (**POPULARITY MATRIX**). We build a popularity matrix $\mathbf{P}_{|C|\times|L|}$, in which each entry $\mathbf{P}_{c,l}$ represents the popularity of POI l in category c, i.e., the check-in frequency or overall rating of all users on l. Note that most entries of $\mathbf{P}$ are zero, because a POI only belongs to a certain subset of the universal category set C.

DEFINITION 5 (**GEOGRAPHICAL COORDINATES**). A POI $l \in L$ is associated with a pair of geographical latitude and longitude coordinates (x,y).



(a) A user in Foursquare (b) A user in Yelp

Figure 2: Geographical check-in distribution (contour line) of random users from the real-world data

Problem definition on POI recommendations. Given the geographical coordinates of POIs, check-in matrix $\mathbf{R}_{|U|\times|L|}$, social link matrix $\mathbf{S}_{|U|\times|U|}$, categorical bias matrix $\mathbf{B}_{|U|\times|C|}$, and popularity matrix $\mathbf{P}_{|C|\times|L|}$, the goal is to predict the preference score $s(u,l)$ of a user u to an unvisited POI l (i.e., $\mathbf{R}_{u,l}=0$), and then return the top-k POIs with the highest preference score $s(u,l)$ for the user u.

3.2 Geographical Correlations

Distinctly from non-spatial items, such as books and music in conventional recommendation systems, in LBSNs users are required to physically interact with POIs to consume their offered products or services, e.g., eating food at restaurants or watching movies at cinemas. Thus, the geographical proximity of POIs plays a significant role in the check-in behaviors of users. In other words, the close POIs have the stronger geographical correlations than the far POIs. Therefore, we can exploit the geographical correlations between *a user's visited POIs* and *her unvisited POI* to estimate the relevance score of the user on the unvisited POI. To model the geographical correlations between POIs, we estimate a personalized check-in distribution over the geographical coordinates for each user based on her own visited POIs.

Due to the fact that the check-in distributions of users are different from one another, e.g., indoorsy persons like visiting venues around their living areas while outdoorsy persons prefer exploring new interesting places by traveling around the world. Accordingly, the current works [30, 31, 33, 34] learn the distribution form from the check-in POIs of a user based on a nonparametric estimation method, i.e., the kernel density estimation with a fixed bandwidth [19]. Nonetheless, the fixed bandwidth does not reflect the facts in the user check-in data: dense urban areas will have high check-in density and sparsely-populated rural areas will have low check-in density. To this end, in this paper we adapt the kernel bandwidth to each check-in data point and the adaptive bandwidth itself is also learned from the underlying check-in data. In general, the adaptive kernel estimation method includes three steps: *pilot estimation*, *local bandwidth determination*, and *adaptive kernel estimation for geographical relevance score*.

Step 1: Pilot estimation. First, we find a pilot estimation based on the kernel density estimation with a fixed bandwidth [19]. Let $L_u=\{l_1,l_2,\cdots,l_n\}$ be the set of check-in POIs of user u, in which each POI l_i is associated with a pair of latitude and longitude (x_i,y_i). Specifically, we use the frequency or rating of user u on POI l_i, i.e., $\mathbf{R}_{u,l_i}$ (DEFINITION 1), as the weight of l_i because a higher frequency or rating at a POI indicates that it is more important to the user. The pilot estimation $\tilde{f}_{Geo}(l|u)$ of the check-in distribution of user u on an unvisited POI l is given by

$$\tilde{f}_{Geo}(l|u)=\frac{1}{N}\sum_{i=1}^{n}\left(\mathbf{R}_{u,l_i}\cdot K_H(l-l_i)\right) \tag{1}$$

together with

$$N=\sum_{i=1}^{n}\mathbf{R}_{u,l_i} \tag{2}$$

and

$$K_H(l-l_i)=\frac{1}{2\pi H_1H_2}\exp\left(-\frac{(x-x_i)^2}{2H_1^2}-\frac{(y-y_i)^2}{2H_2^2}\right), \tag{3}$$

where $K_H(l-l_i)$ is the normal kernel function with the fixed bandwidth H consisting of two global bandwidths (H_1,H_2) for the latitude and longitude, given by

$$H_1=1.06n^{-\frac{1}{5}}\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(\mathbf{R}_{u,l_i}\cdot x_i-\frac{1}{N}\sum_{j=1}^{n}\mathbf{R}_{u,l_j}\cdot x_j\right)^2} \tag{4}$$

and

$$H_2=1.06n^{-\frac{1}{5}}\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(\mathbf{R}_{u,l_i}\cdot y_i-\frac{1}{N}\sum_{j=1}^{n}\mathbf{R}_{u,l_j}\cdot y_j\right)^2}, \tag{5}$$

where (H_1,H_2) are computed from the standard deviation of the latitude and longitude values in the check-in data of user u, respectively [19].

Step 2: Local bandwidth determination. Further, instead of directly using the pilot estimation $\tilde{f}_{Geo}(l|u)$ in Equation (1) to predict the relevance score of user u to POI l, we utilize the pilot estimation to determine the local bandwidth h_i for each check-in POI l_i, given by

$$h_i=\left(g^{-1}\cdot\tilde{f}_{Geo}(l_i|u)\right)^{-\alpha}, \tag{6}$$

where α is the sensitivity parameter with $0\le\alpha\le1$, i.e., the larger the parameter α, the more sensitive the local bandwidth h_i will be to the pilot estimation $\tilde{f}_{Geo}(l_i|u)$, and g is the geometric mean:

$$g=\sqrt[n]{\prod_{i=1}^{n}\tilde{f}_{Geo}(l_i|u)} \tag{7}$$

which imposes the constraint that the geometric mean of the h_i ($i=1,2,\cdots,n$) is equal to one.

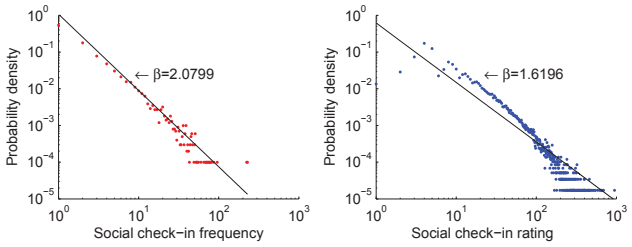
Step 3: Adaptive kernel estimation for geographical relevance score. Finally, with the global bandwidth $H=(H_1,H_2)$ in Equations (4) and (5) and the adaptive local bandwidth h_i in Equation (6), the adaptive kernel estimation $f_{Geo}(l|u)$ of the check-in distribution of user u on an unvisited POI l is computed through

$$f_{Geo}(l|u)=\frac{1}{N}\sum_{i=1}^{n}\left(\mathbf{R}_{u,l_i}\cdot K_{Hh_i}(l-l_i)\right), \tag{8}$$

where

$$K_{Hh_i}(l-l_i)=\frac{1}{2\pi H_1H_2h_i^2}\exp\left(-\frac{(x-x_i)^2}{2H_1^2h_i^2}-\frac{(y-y_i)^2}{2H_2^2h_i^2}\right). \tag{9}$$

It is important to note that: When a check-in POI l_i lies in the area with a higher check-in density, it has a larger pilot estimation $\tilde{f}_{Geo}(l_i|u)$ in Equation (1) or a smaller local bandwidth h_i in Equation (6) that will produce a peak adaptive kernel estimation $f_{Geo}(l|u)$ around l_i in Equation (8). In contrast, when a check-in POI l_i lies in the area with a lower check-in density, it has a smaller pilot estimation $\tilde{f}_{Geo}(l_i|u)$ in Equation (1) or a larger local bandwidth h_i in Equation (6) that will generate a smooth adaptive kernel estimation $f_{Geo}(l|u)$ around l_i in Equation (8). Therefore, our adaptive kernel estimation method can improve the predictive



(a) Foursquare (b) Yelp

Figure 3: Social check-in frequency or rating distribution in the real-world data

ability of the estimated check-in distribution for the geographical relevance score of user u on an unvisited POI l, i.e., $f_{Geo}(l|u)$.

Example. Figure 2 depicts the contour line of the geographical check-in distribution estimated through Equation (8) with $\alpha = 0.5$ for a user who is randomly selected from two publicly available real-world data sets: Foursquare [6] and Yelp [24], respectively. The estimated geographical distribution is able to capture the geographical preference of a specific user on POIs: the user in Foursquare visits POIs over the world, with concentration on two areas, while the user in Yelp checks in POIs only in Arizona of USA, with focus on three areas.

3.3 Social Correlations

In reality, friends often go to some places like movie theaters or restaurants together or a person may travel on spots highly recommended by her friends. In other words, the social correlations between users greatly affect the check-in behaviors of users to POIs. Likewise, in LBSNs users create social links with each other to indicate the social friendships among them. Accordingly, we can deduce the relevance score of a user and an unvisited POI through leveraging the social correlations between *the user* with *her friends* who have visited the POI. The process consists of three steps: *social aggregation*, *distribution estimation of social frequency or rating*, and *social relevance score computation*.

Step 1: Social aggregation. Formally, given a user u and an unvisited POI l, we aggregate the check-in frequency or rating $x_{u,l}$ of the user u's friends (i.e., u' with $\mathbf{S}_{u,u'} = 1$) on the POI l, given by

$$x_{u,l} = \sum_{u' \in U} \mathbf{S}_{u,u'} \cdot \mathbf{R}_{u',l}, \tag{10}$$

where $\mathbf{R}_{u',l}$ is the check-in frequency or rating of user u' on POI l (DEFINITION 1) and $\mathbf{S}_{u,u'}$ indicates whether there exists a social link between users u and u' (DEFINITION 2).

One can naively regard the social check-in frequency or rating $x_{u,l}$ as the relevance score between user u and POI l or simply divide $x_{u,l}$ by the number of friends of u as in the traditional collaborative filtering techniques. More sophisticatedly, in this study we transform the social check-in frequency (or rating) into a normalized relevance score based on the social check-in frequency (or rating) distribution that is learned from the historical check-in data of all users.

Step 2: Distribution estimation of social frequency or rating. In real-world data sets, the social check-in frequency or rating random variable x follows a power-law distribution, the probability density function of which is defined by

$$f_{So}(x) = (\beta - 1)(1+x)^{-\beta}, x \geq 0, \beta > 1, \tag{11}$$

where β is estimated by the check-in matrix $\mathbf{R}$ and social link matrix $\mathbf{S}$:

$$\beta = 1 + |U||L| \left[\sum_{u' \in U} \sum_{l' \in L} \ln \left(1 + \sum_{u'' \in U} \mathbf{S}_{u',u''} \cdot \mathbf{R}_{u'',l'} \right) \right]^{-1}, \tag{12}$$

in which $\sum_{u'' \in U} \mathbf{S}_{u',u''} \cdot \mathbf{R}_{u'',l'}$ is the social check-in frequency or rating of the friends u'' of user u' on POI l'.

To observe the real distribution of the social check-in frequency or rating, we conducted analysis on the two publicly available real-world data sets: Foursquare [6] and Yelp [24]. Figure 3 shows that the social check-in frequency or rating (i.e., the dots) in the two real-world data sets fits a certain power-law distribution very well (i.e., the line), estimated through Equations (11) and (12). Thus, modeling the social check-in frequency or rating as a power-law distribution is reasonable and effective.

Step 3: Social relevance score computation. The estimated probability density function f_{So} in Equation (11) is monotonically decreasing with respect to the social check-in frequency (or rating) x, but the social relevance score should be monotonically increasing with regard to the social check-in frequency (or rating) because friends share more common interests on POIs. Thus, we define the social relevance score of $x_{u,l}$ in Equation (10) based on the cumulative distribution function of f_{So}, given by

$$F_{So}(x_{u,l}) = \int_0^{x_{u,l}} f_{So}(z)dz = 1 - (1+x_{u,l})^{1-\beta}, \tag{13}$$

where F_{So} is an increasing function respecting the social check-in frequency or rating $x_{u,l}$ because of $1 - \beta < 0$. Moreover, based on the cumulative distribution function F_{So} in Equation (13), the social check-in frequency (or rating) $x_{u,l}$ is transformed into a social relevance score that reflects the relative position of $x_{u,l}$ in all social check-in frequencies (or ratings) of users on POIs.

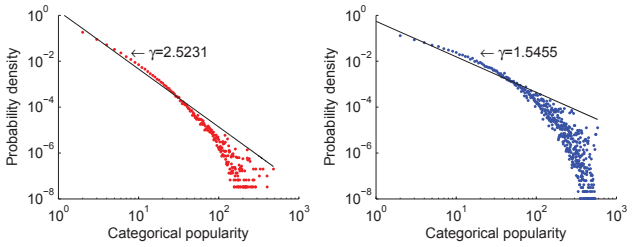
3.4 Categorical Correlations

In LBSNs, each POI is attached to a few categories. The category of a POI has a strong indication about what activities happen in the POI and what products or services are provided by the POI. For instance, a person visiting a restaurant means that she may have a meal there and a Chinese restaurant indicates that Chinese food will be provided for customers. In practice, people have shown distinct biases on the categories of POIs, e.g., a foodie likes visiting restaurants to taste various food and a tourism enthusiast prefers traveling all over the world to view tourism attractions. Hence, we also can derive the relevance score of a user to an unvisited POI through exploiting the categorical correlations between *the visited POIs* and *the unvisited POI* of the user.

In addition, the popularity of a POI reflects the quality of products or services offered by the POI, e.g., a popular restaurant usually indicates that it supplies high-quality foods. Therefore, it is helpful to utilize the popularity for POI recommendations. Specifically, we develop a new method to combine the category bias of a user and the popularity of a POI into a relevance score between the user and POI through three steps: *weighing popularity by categorical bias*, *distribution estimation of categorical popularity*, and *categorical relevance score computation*.

Step 1: Weighing popularity by categorical bias. At first, we take the bias of a user u to a certain category c as $\mathbf{B}_{u,c}$, i.e., the frequency of user u visiting the POIs that belong to category c (DEFINITION 3). Then, the bias $\mathbf{B}_{u,c}$ is used to weigh the popularity of an unvisited POI l in category c, i.e., $\mathbf{P}_{c,l}$ (DEFINITION 4). Correspondingly, we obtain the categorical popularity $y_{u,l}$ for user u on POI l:

$$y_{u,l} = \sum_{c \in C} \mathbf{B}_{u,c} \cdot \mathbf{P}_{c,l}, \tag{14}$$



(a) Foursquare (b) Yelp

Figure 4: Categorical popularity distribution in the real-world data

where C is the predefined universal set of categories in an LBSN. A larger value of $y_{u,l}$ indicates that the category of POI l is more satisfied with the bias of user u and the POI l is more popular to the general public.

One may naively consider the categorical popularity $y_{u,l}$ as the relevance score between user u and POI l or simply normalize the categorical bias $\mathbf{B}_{u,c}$ in advance. Nevertheless, in this research the categorical popularity of a user to an unvisited POI is sophisticatedly mapped into a relevance score based on the distribution of the categorical popularity that is learned from the historical check-in data of all users.

Step 2: Distribution estimation of categorical popularity. As the distribution of the social check-in frequency or rating, we apply the similar process to build the distribution of the categorical popularity. Formally, we assume the probability density function of the categorical popularity random variable y, defined by

$$f_{Ca}(y) = (\gamma - 1)(1+y)^{-\gamma}, y \geq 0, \gamma > 1, \tag{15}$$

in which γ can be learned from the categorical bias matrix $\mathbf{B}$ and popularity matrix $\mathbf{P}$:

$$\gamma = 1 + |U||L| \left[\sum_{u' \in U} \sum_{l' \in L} \ln \left(1 + \sum_{c \in C} \mathbf{B}_{u',c} \cdot \mathbf{P}_{c,l'} \right) \right]^{-1}, \tag{16}$$

where $\sum_{c \in C} \mathbf{B}_{u',c} \cdot \mathbf{P}_{c,l'}$ is the categorical popularity of user u' on POI l'.

As depicted in Figure 4, we have also observed that the categorical popularity (i.e., the dots) in the two real-world data sets approaches to the power-law distribution (i.e., the line) that is estimated in terms of Equations (15) and (16). Thus, these results have validated that the assumption of the power-law distribution is in accordance with reality.

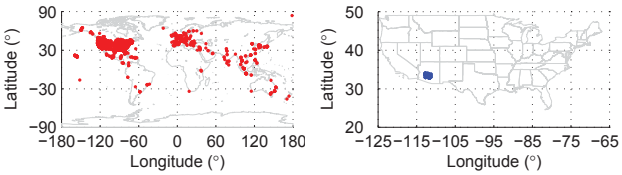
Step 3: Categorical relevance score computation. Similarly, the estimated probability density function f_{Ca} in Equation (15) is monotonically decreasing regarding the categorical popularity y; however, the categorical relevance score is monotonically increasing respecting the categorical popularity, since people prefer the popular POIs that also meet their categorical biases. To this end, we employ the cumulative distribution function of f_{Ca} to obtain the categorical relevance score of $y_{u,l}$ in Equation (14), given by

$$F_{Ca}(y_{u,l}) = \int_0^{y_{u,l}} f_{Ca}(z)dz = 1 - (1+y_{u,l})^{1-\gamma}, \tag{17}$$

where due to $1 - \gamma < 0$, F_{Ca} is an increasing function with respect to the categorical popularity $y_{u,l}$. Importantly, the categorical $y_{u,l}$ is also normalized into a categorical relevance score, i.e., the relative position of $y_{u,l}$ compared to other categorical popularities of users on POIs.

Table 2: Statistics of the two data sets

	Foursquare	Yelp
Number of users	4,163	70,817
Number of POIs	121,142	15,579
Number of categories	35	591
Number of check-ins or ratings	483,813	335,022
Number of social links	32,512	303,032
User-POI matrix density	2.83×10^{-4}	1.46×10^{-4}



(a) Foursquare (b) Yelp

Figure 5: Distribution of POIs in the two data sets

3.5 POI Recommendations

Finally, we integrate the geographical, social and categorical relevance scores, given by Equations (8), (13) and (17), into a unified preference score $s(u,l)$ for user u to unvisited POI l based on the product rule:

$$s(u,l) = f_{Geo}(l|u) \cdot F_{So}(x_{u,l}) \cdot F_{Ca}(y_{u,l}). \tag{18}$$

The top-k POIs l with the highest score $s(u,l)$ are recommended to user u. It is worth mentioning that the product rule has been widely used to fuse different factors for POI recommendations in the previous works [2, 13, 30, 33] and has shown high robustness.

4. EXPERIMENTS

In this section, we conduct intensive experiments to evaluate the recommendation quality of GeoSoCa, compared to state-of-the-art POI recommendation techniques. We present experimental settings in Section 4.1 and analyze experimental results in Section 4.2.

4.1 Experimental Settings

Two real data sets. We use the two publicly available large-scale real check-in data sets that have been studied in Section 3 and were crawled from Foursquare [6] and Yelp [24]. Note that the Foursquare data set provides the check-in frequencies of users to POIs and the Yelp data set offers the ratings of users on POIs. Table 2 shows the statistics of the data sets and Figure 5 depicts the distribution of POIs over the world from Foursquare while the greater Phoenix, Arizona, USA from Yelp. The half of check-in data with earlier check-in times are used as the training set and the other half of check-in data are used as the testing set, because in practice we can only utilize the past check-in data to predict the future check-in events.

Evaluated techniques. Our proposed method, i.e., GeoSoCa, is compared with the state-of-the-art POI recommendation techniques including:

- USG: This method is a unified location recommendation framework that integrates User preferences, Social and Geographical information [23].
- CoRe: This method fuses social collaborative filtering and geographical check-in distribution using kernel density estimation with a fixed bandwidth [31].

- DRW: This method is a Dynamic Random Walk model that combines social, category and popularity information [27].
- LCARS: This method builds a Location-Content-Aware Recommender System based on the well-known topic model (i.e., latent Dirichlet allocation) to infer personal interest and local preference [25, 26].
- NCPD: This method applies matrix factorization to incorporate the influence of Neighborhood, Category, Popularity and Geographical distance of POIs [8].

Performance metrics. To evaluate the quality of POI recommendations, it is important to find out how many POIs actually visited by a user **in the testing set** are discovered by the recommendation techniques. For this purpose, we employ two standard metrics:

$$\text{Precision} = \frac{\text{No. of discovered POIs}}{\text{No. of POIs recommended for the user: } k},$$

$$\text{Recall} = \frac{\text{No. of discovered POIs}}{\text{No. of POIs actually visited by the user}}.$$

Parameter settings. We examine the precision and recall with respect to various numbers of POIs recommended for users (top-k from 2 to 50) and numbers of POIs visited by users **in the training set** (n from 2 to 50). Further, we investigate the recommendation accuracy of the three components in GeoSoCa and the effect of sensitivity parameter α, by default $\alpha = 0.5$, in Equation (6). Note that β and γ are not free parameters and are learned from check-in data according to Equations (12) and (16), respectively.

4.2 Experimental Results

We compare our GeoSoCa with the state-of-the-art POI recommendation techniques in Section 4.2.1, study the recommendation quality of the three components in GeoSoCa in Section 4.2.2, and investigate the effect of sensitivity parameter α in Section 4.2.3.

4.2.1 Method Comparison

Figures 6 and 7 depict the recommendation accuracy of our GeoSoCa compared to the state-of-the-art POI recommendation techniques with respect to the number k of POIs recommended for users and the number n of POIs visited by users in the training set. The trends in the precision and recall of all evaluated methods are intuitive. For example, as k increases, the precision gets lower and the recall becomes higher, because recommending more POIs for users can discover more POIs that the users would like to check in, but some recommended POIs are less possible to be visited by the users. With the increase of n, the precision and recall gradually raise, since all methods can learn better recommendation models through using more check-in data.

The absolute accuracy of POI recommendation techniques is usually not high, because the density of user-POI check-in matrix is pretty low as shown in Table 2, but POI recommendation techniques will perform better as more check-in data are collected. This phenomenon has been repeatedly observed in previous works (e.g., [23, 31]). Instead, we concentrate on contrasting the relative accuracy of the evaluated POI recommendation techniques.

USG. This method [23] linearly integrates user preference from user-based collaborative filtering, social influence from social collaborative filtering, and geographical influence from a common distance distribution for all users, but it does not take into account the category and popularity information of POIs. Moreover, it is not advisable to apply the universal linear weights for user preference, social influence, and geographical influence, since some users are affected by social friends more and other users may rely on the geographical influence more. According to Figures 6 and 7, USG subsequently gives the third worst recommendation result on the Foursquare data set and the worst recommendation accuracy on the Yelp data set that has two time lower density than the Foursquare data set, as depicted in Table 2.

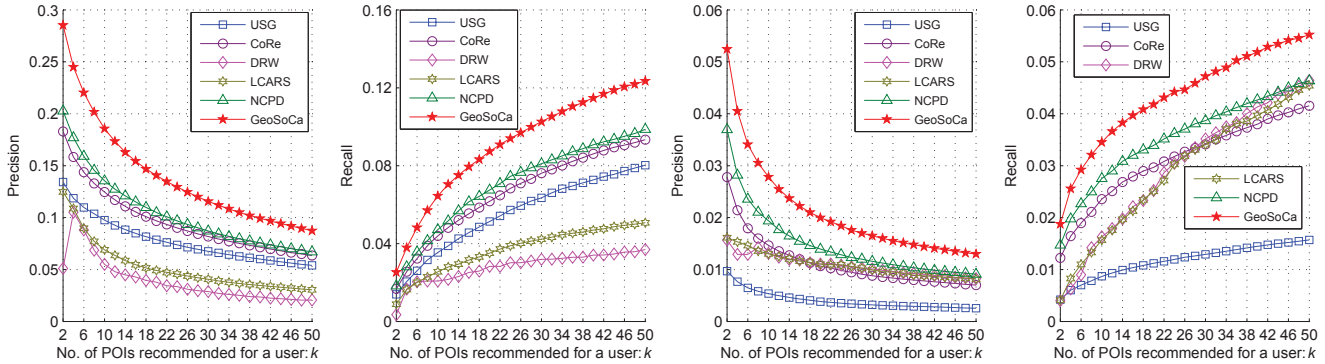
CoRe. This method [31] employs the social influence in the same way as USG, but it models a personalized geographical check-in distribution for each user and combines the social and geographical influences by a more robust product rule rather than using the linear sum rule. CoRe accordingly outperforms USG to some extent as depicted in Figures 6 and 7. However, CoRe still misses the category and popularity information of POIs and estimates the geographical check-in distribution based on the kernel density estimation method with a fixed bandwidth. As a result, in general it only generates the third best recommendation precision and recall on the two data sets.

DRW. This method [27] adopts a dynamic random walk model to fuse the social links between users and the category and popularity information of POIs, but it does not consider the unique characteristic of POI recommendations for LBSNs, i.e., the influence of geographical information of POIs on the check-in behaviors of users. Consequently, DRW reports the lowest recommendation accuracy on the Foursquare data set, as shown in Figures 6(a), 6(b), 7(a), and 7(b). Interestingly, on the Yelp data set with lower density in Figures 6(c), 6(d), 7(c), and 7(d), DRW is superior to USG through taking full advantage of the popularity of POIs to deal with the sparsity of data.

LCARS. This method [25, 26] exploits the well-known topic model, i.e., latent Dirichlet allocation, to infer the personal interest of users and local preference (i.e., local specialty) of regions (e.g., a city). The personal interest or local preference is represented as a mixture of topics, in which each topic is a distribution over POIs and is learned from the check-in data and categories of POIs. Nonetheless, LCARS ignores the geographical and social characteristics in the check-in behaviors of users on POIs, so it also suffers from the low recommendation accuracy as in DRW.

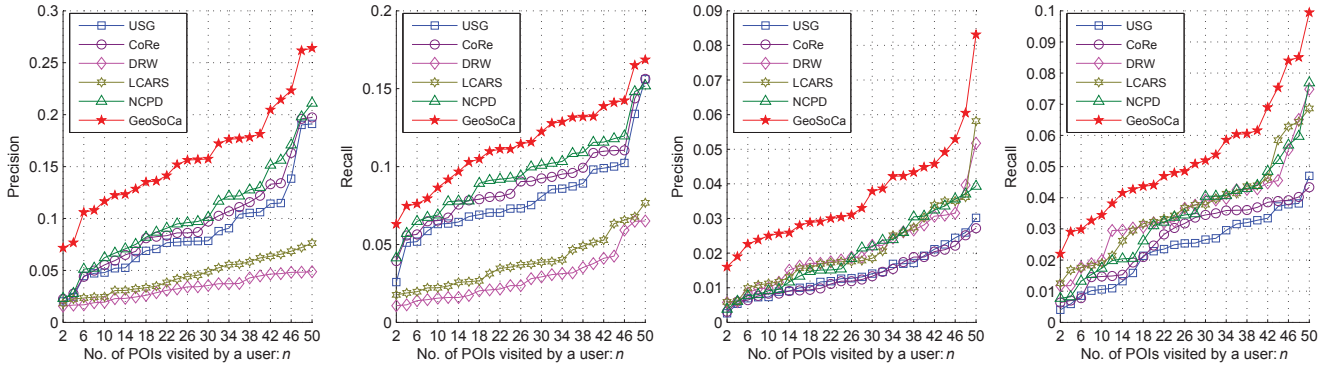
NCPD. This method [8] utilizes matrix factorization to derive a latent factor vector for each user, POI and category, a geographical bias for each user, and a popularity bias for each POI. Then, NCPD computes the preference score of a user to a POI based on (1) the latent factor vectors of the user, the categories that the POI belongs to, and the neighborhoods of the POI, (2) the user's geographical bias, and (3) the POI's popularity bias. Accordingly, NCPD shows the second best recommendation result as depicted in Figures 6 and 7. However, its improvement is very limited in comparison to CoRe, because NCPD simply regards the geographical and popularity influence as a bias, instead of modeling them as a geographical or popularity distribution.

GeoSoCa. Our proposed GeoSoCa always exhibits the best recommendation quality in terms of both precision and recall. In particular, it achieves the significant improvement compared to the second best recommendation technique NCPD. The reason is threefold: (1) GeoSoCa models the geographical check-in distribution using the adaptive kernel estimation method, in which the bandwidth is adaptive to each check-in data point rather than using a fixed bandwidth. This adaptive method can capture the real geographical check-in patterns of users on POIs, e.g., high check-in density in dense urban areas and low check-in density in sparsely-populated rural areas. (2) GeoSoCa exploits the social check-in frequency or rating of friends based on the power-law distribution which is learned from the historical check-in data of users. This method can effectively transform the social check-in frequency or



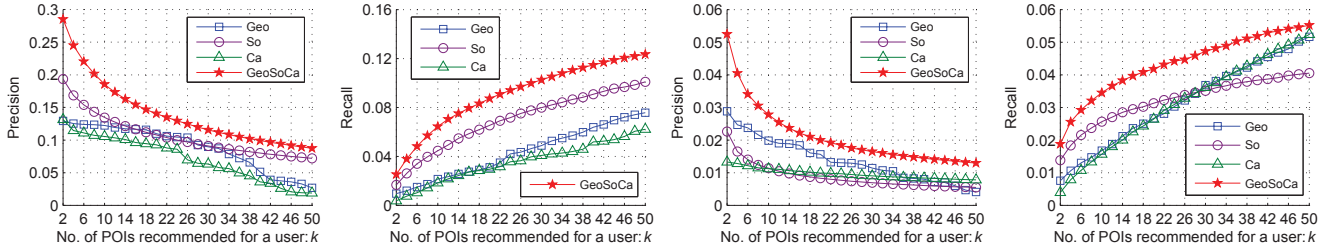
(a) Foursquare (b) Foursquare (c) Yelp (d) Yelp

Figure 6: Recommendation accuracy with respect to top-k values



(a) Foursquare (b) Foursquare (c) Yelp (d) Yelp

Figure 7: Recommendation accuracy with respect to given-n values



(a) Foursquare (b) Foursquare (c) Yelp (d) Yelp

Figure 8: Recommendation accuracy of the three components of GeoSoCa

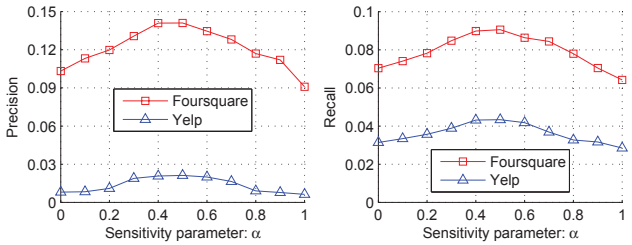
rating into a reasonable relevance score and is superior to the traditional social collaborative filtering. (3) GeoSoCa takes full advantage of the category and popularity information through seamlessly combining the categorical bias of users and the popularity of POIs into a reasonable relevance score based on the estimated power-law distribution from check-in data of users.

4.2.2 *Study on Three Components in GeoSoCa*

Here we study the three components of GeoSoCa including geographical, social and categorical correlations, written as Geo, So and Ca, respectively. Figure 8 depicts the recommendation accuracy of the three components based on Equations (8), (13) and (17), respectively. We have two observations: (1) All the three components play a key role in GeoSoCa for POI recommendations and they are competitive to one another. For example, So outperforms Geo and Ca on the Foursquare data set, but reversely on the Yelp data set. And the performance of Geo and Ca is similar on the two data sets. (2) The integration of the three components is helpful for enhancing the recommendation quality, since GeoSoCa is significantly superior to each component, i.e., Geo, So and Ca. The behind reason is that in practice people are affected by the geographical, social and categorical correlations in varying degrees; it is unable to model the check-in behaviors of all users by considering only one type of correlations. Thus, POI recommendations should take full advantage of various types of useful information implied the check-in behaviors of users on POIs.

4.2.3 *Effect of Sensitivity Parameter α*

Figure 9 depicts the effect of the sensitivity parameter α in Equation (6) on the precision and recall of GeoSoCa in the two real-world data sets; note that β and γ are not free parameters and are learned from the check-in data. We have the following three obser-



(a) GeoSoCa - Precision (b) GeoSoCa - Recall

Figure 9: Effect of sensitive parameter α on recommendation accuracy of GeoSoCa

vations: (1) The optimal value of α usually lies between 0.4 and 0.6 that generates the highest recommendation accuracy. (2) As the value of α varies from 0.4 to 0, the local bandwidth gets less sensitive to the pilot estimation in terms of Equation (6), i.e., the local bandwidth is less related to the check-in data. Especially, when $\alpha = 0$, the adaptive local bandwidth is degraded to the fixed bandwidth that has nothing to do with the check-in data. As a result, the precision and recall decrease. (3) In contrast, when the value of α becomes higher from 0.6 to 1, the local bandwidth gets more sensitive to the pilot estimation in Equation (6), i.e., the local bandwidth is prone to over-fitting the check-in data. Thus, the recommendation quality deteriorates as well.

5. CONCLUSION AND FUTURE WORK

This paper proposes a new POI recommendation approach GeoSoCa by exploiting three types of correlations that are derived from the historical check-in data of users on POIs. (1) GeoSoCa uses the kernel estimation with an adaptive bandwidth to model the geographical correlations between POIs as a personalized geographical check-in distribution of POIs for each user. (2) GeoSoCa models the social check-in frequency or rating as a power-law distribution to utilize the social correlations between users. (3) GeoSoCa models the popularity of POI categories as a power-law distribution to employ the categorical correlations between POIs. The experimental results on the two large-scale real-world data sets collected from Foursquare and Yelp show that GeoSoCa significantly improves the recommendation accuracy of the current state-of-the-art POI recommendation approaches. As a part of our future work, we plan to extend our GeoSoCa by integrating the sentiment or opinion extracted from the textual reviews of users commenting POIs to improve the recommendation quality of GeoSoCa.

6. ACKNOWLEDGEMENTS

The authors were supported by Guangdong Natural Science Foundation of China under Grant S2013010012363.

7. REFERENCES

[1] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *ACM SIGSPATIAL*, 2012.

[2] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, 2012.

[3] C. Cheng, H. Yang, M. R. Lyu, and I. King. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, 2013.

[4] G. Ference, M. Ye, and W.-C. Lee. Location recommendation for out-of-town users in location-based social networks. In *ACM CIKM*, 2013.

[5] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *ACM RecSys*, 2013.

[6] H. Gao, J. Tang, X. Hu, and H. Liu. Content-aware point of interest recommendation on location-based social networks. In *AAAI*, 2015.

[7] B. Hu and M. Ester. Spatial topic modeling in online social media for location recommendation. In *ACM RecSys*, 2013.

[8] L. Hu, A. Sun, and Y. Liu. Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction. In *ACM SIGIR*, 2014.

[9] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *ACM WSDM*, 2013.

[10] K. W.-T. Leung, D. L. Lee, and W.-C. Lee. CLR: A collaborative location recommendation framework based on co-clustering. In *ACM SIGIR*, 2011.

[11] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. LARS: A location-aware recommender system. In *IEEE ICDE*, 2012.

[12] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *ACM KDD*, 2014.

[13] B. Liu, Y. Fu, Z. Yao, and H. Xiong. Learning geographical preferences for point-of-interest recommendation. In *ACM KDD*, 2013.

[14] B. Liu and H. Xiong. Point-of-interest recommendation in location based social networks with topic and location awareness. In *SDM*, 2013.

[15] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *ACM CIKM*, 2013.

[16] Y. Liu, W. Wei, A. Sun, and C. Miao. Exploiting geographical neighborhood characteristics for location recommendation. In *ACM CIKM*, 2014.

[17] H. J. Miller. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004.

[18] S. Rahimi and X. Wang. Location recommendation based on periodicity of human activities and location categories. In *PAKDD*, 2013.

[19] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, 1986.

[20] H. Wang, M. Terrovitis, and N. Mamoulis. Location recommendation in location-based social networks using user check-in data. In *ACM SIGSPATIAL*, 2013.

[21] D. Yang, D. Zhang, Z. Yu, and Z. Wang. A sentiment-enhanced personalized location recommendation system. In *ACM HT*, 2013.

[22] Z. Yao, B. Liu, Y. Fu, H. Xiong, and Y. Ge. User preference learning with multiple information fusion for restaurant recommendation. In *SDM*, 2014.

[23] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *ACM SIGIR*, 2011.

[24] Yelp. Challenge Data Set [accessed 25-april-2014].

http://www.yelp.com/dataset_challenge, 2014.

[25] H. Yin, B. Cui, Y. Sun, Z. Hu, and L. Chen. LCARS: A spatial item recommender system. *ACM TOIS*, 32(3):11:1–11:37, 2014.

[26] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen. LCARS: A location-content-aware recommender system. In *ACM KDD*, 2013.

[27] J. J.-C. Ying, W.-N. Kuo, V. S. Tseng, and E. H.-C. Lu. Mining user check-in behavior with a random walk for urban point-of-interest recommendations. *ACM TIST*, 5(3):40:1–40:26, 2014.

[28] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Time-aware point-of-interest recommendation. In *ACM SIGIR*, 2013.

[29] Q. Yuan, G. Cong, and A. Sun. Graph-based point-of-interest recommendation with geographical and temporal influences. In *ACM CIKM*, 2014.

[30] J.-D. Zhang and C.-Y. Chow. iGSLR: Personalized geo-social location recommendation - a kernel density estimation approach. In *ACM SIGSPATIAL*, 2013.

[31] J.-D. Zhang and C.-Y. Chow. CoRe: Exploiting the personalized influence of two-dimensional geographic coordinates for location recommendations. *Information Sciences*, 293:163–181, 2015.

[32] J.-D. Zhang and C.-Y. Chow. TICRec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations. *IEEE TSC, accepted to appear*, 2015.

[33] J.-D. Zhang, C.-Y. Chow, and Y. Li. iGeoRec: A personalized and efficient geographical location recommendation framework. *IEEE TSC, accepted to appear*, 2014.

[34] J.-D. Zhang, C.-Y. Chow, and Y. Li. LORE: Exploiting sequential influence for location recommendations. In *ACM SIGSPATIAL*, 2014.

[35] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. *ACM TIST*, 5(3):50:1–50:26, 2014.

[36] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Towards mobile intelligence: Learning from gps history data for collaborative recommendation. *Artificial Intelligence*, 184-185:17–37, 2012.