Point-of-Interest Recommendations in Location-Based Social Networks

Jia-Dong Zhang, Chi-Yin Chow
Department of Computer Science, City University of Hong Kong, Hong Kong

Abstract

Location-based social networks (LBSNs), e.g., Foursquare, Gowalla and Yelp, bridge the physical world with the virtual online world. LBSNs have accumulated plenty of community-contributed data such as social links between users, check-ins of users on points-of-interest (POIs), geographical information and categories of POIs, which reflect the preferences of users to POIs. Recommending users with their preferred POIs benefits people to explore new places and businesses to discover potential customers. This paper aims to recommend personalized POIs for users based on their preferences that are learned from the community-contributed data. To this end, this paper models the social, categorical, geographical, sequential, and temporal influences on the visiting preferences of users to POIs.

1 Introduction

With the rapid pervasiveness of mobile devices embedded with wireless communication and location acquisition abilities, location-based social networks (LBSNs) such as Foursquare, Gowalla, Brightkite, Yelp, and Facebook places, have become some of the most popular Internet applications and attracted millions of users. The LBSNs bridge the physical world with the virtual online world. In an LBSN (Figure 1), users can establish social links with each other to share their experiences of visiting some **interesting locations**, also known as **points-of-interest (POIs)**, e.g., restaurants, stores, and museums, through performing check-ins to these POIs in the LBSNs via their handheld device. In LBSNs, there are plenty of community-contributed data including *social links between users*, *check-ins of users on POIs*, *geographical information and categories of POIs*. These rich data are the reflection of human behaviors in reality and bring new opportunities to model the decision making process of users visiting POIs. In the LBSNs, it is crucial to recommend personalized POIs to users based on their preferences learned from the community-contributed data, which benefits for users to know new POIs and discover a city while for businesses to delivery advertisements to targeted users and improve business profits.

In LBSNs, **there are five major characteristics that affect the visiting preferences or check-in behaviors of users to POIs**. **(1) Social influence.** In the real world, people interact with each other. For example, friends often go to some places like movie theaters or restaurants together, or a person may travel on spots highly recommended by her friends. Thus, a person's preference on POIs can be influenced by her close friends or a group of friends that are likely to share some common interests. Accordingly, in LBSNs, users establish social links and form communities to share their experiences of visiting POIs. **(2) Categorical influence.** The category of a POI reflects its usual business activities and nature. For instance, a person checking in a restaurant indicates that she may have a meal and checking in a cinema means that she is watching a movie there. In practice, people have shown 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. **(3) Geographical influence.** Spatial POIs are totally different from other non-spatial items, e.g., books, music and movies in conventional

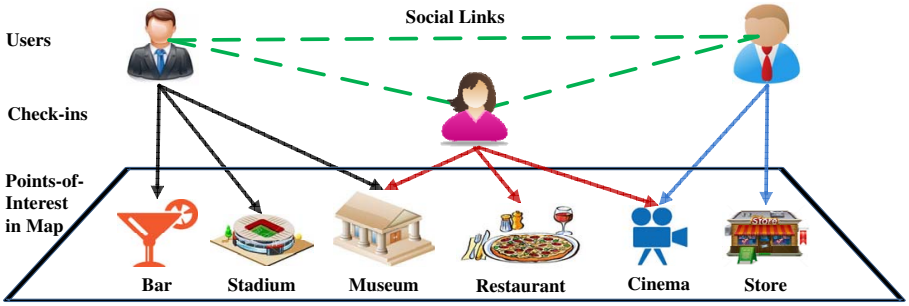


Figure 1: A location-based social network

recommender systems, because physical interactions are required for users to visit POIs. The geographical information (i.e., latitude and longitude coordinates) of POIs significantly affects users' check-in behaviors. For instance, people tend to visit POIs close to their homes or offices, and also may be interested in exploring the nearby places of their visited POIs. **(4) Sequential influence.** In reality, human movements exhibit sequential patterns. For instance, cinemas or bars may be usually visited after restaurants as users would like to relax after dinner, and checking in stadiums first and then restaurants is better than the reverse way because it is not healthy to exercise right after a meal. Thus, the influence of sequential patterns is also important for users' check-in behaviors. **(5) Temporal influence.** Time is a very important factor influencing human activities at different times on weekdays and weekends. For example, users often visit restaurants at noon on weekdays and bars at midnight on weekends. These weekday and weekend patterns reflect the temporal check-in preferences of users to POIs, which can be used to make time-aware POI recommendations by suggesting properly visiting time.

This paper aims to exploit the social, categorical, geographical, sequential, and temporal influences to recommend personalized POIs for users, in which 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.

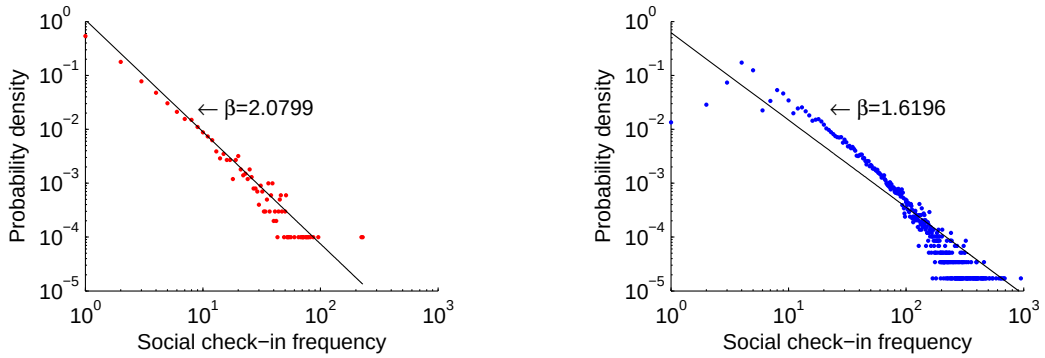
2 Modeling Social Influence

In reality, the social links between users greatly affect the check-in behaviors of users to POIs. Existing works simply employ the social links of users to derive the similarities between users and integrate them into the traditional collaborative filtering techniques. Nevertheless, the traditional collaborative filtering techniques often suffer from the data sparsity problem in the user-POI check-in matrix, since users only visit a very small proportion of POIs in an LBSN. Thus, it is much better to devise a new and sophisticated approach to exploit the social links between users for POI recommendations. In our recent study [5], we 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 frequency aggregation*, *distribution estimation of social frequency*, and *social relevance score computation*.

Step 1: Social frequency 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

$$x_{u,l} = \sum\nolimits_{u' \in U} \mathbf{S}_{u,u'} \cdot \mathbf{R}_{u',l}, \tag{1}$$

where $\mathbf{R}_{u',l}$ is the frequency or rating of user u' visiting POI l and $\mathbf{S}_{u,u'}$ indicates whether there exists a social link between users u and u'. One can naively regard the social check-in frequency $x_{u,l}$ 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, but more sophisticatedly in this study we transform the social check-in frequency into a



(a) Foursquare (b) Yelp

Figure 2: Social check-in frequency distribution in the real-world data

normalized relevance score based on the social check-in frequency distribution that is learned from the historical check-in data of all users.

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

$$f_{So}(x) = (\beta - 1)(1 + x)^{-\beta}, x \geq 0, \beta > 1, \tag{2}$$

where β is estimated by the check-in or rating matrix $\mathbf{R}$ and social link matrix $\mathbf{S}$ based on maximum likelihood estimation:

$$\beta = 1 + |U||L| \left[\sum_{u' \in U} \sum_{l' \in L} \ln\left(1 + \sum_{u'' \in U} \mathbf{S}_{u',u''} \cdot \mathbf{R}_{u'',l'}\right)\right]^{-1}, \tag{3}$$

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

To observe the real distribution of the social check-in frequency, we conducted analysis on the two publicly available real-world data sets with social links between users and categories of POIs: Foursquare [1] and Yelp [2]. Figure 2 shows that the social check-in frequency (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 (2) and (3). Thus, modeling the social check-in frequency as a power-law distribution is reasonable and effective.

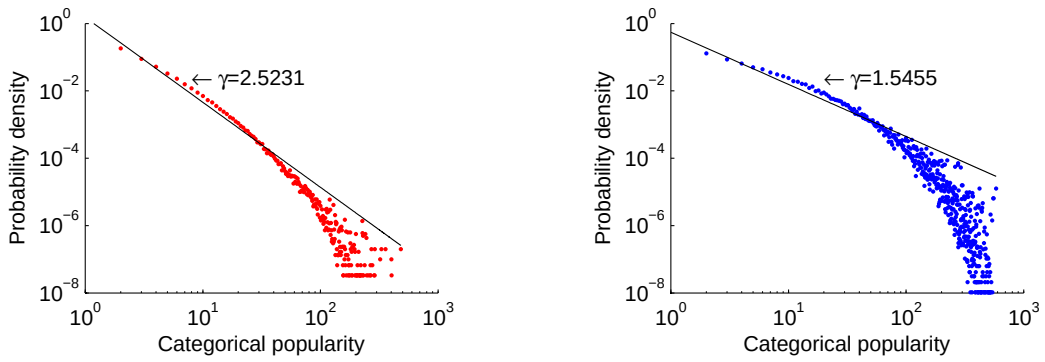
Step 3: Social relevance score computation. The estimated probability density function f_{So} in Equation (2) is monotonically decreasing with respect to the social check-in frequency x, but the social relevance score should be monotonically increasing with regard to the social check-in frequency because friends share more common interests on POIs. Thus, we define the social relevance score of $x_{u,l}$ in Equation (1) based on the cumulative distribution function of f_{So}, given by

$$F_{So}(x_{u,l}) = \int_0^{x_{u,l}} f_{So}(z)dz = 1 - (1 + x_{u,l})^{1-\beta}, \tag{4}$$

where F_{So} is an increasing function on the social check-in frequency $x_{u,l}$ because of $1-\beta < 0$. Moreover, based on the cumulative distribution function F_{So} in Equation (4), the social check-in frequency $x_{u,l}$ is transformed into a social relevance score that reflects the relative position of $x_{u,l}$ in all social check-in frequencies.

3 Modeling Categorical Influence

In practice, the category of a POI has a strong indication about what activities happen in the POI and people have shown distinct biases on the categories of POIs. Hence, we also can derive the relevance score of a user



(a) Foursquare (b) Yelp

Figure 3: Categorical popularity distribution in the real-world data

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 approach [5] 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. Then, the bias $\mathbf{B}_{u,c}$ is used to weigh the popularity or overall rating of an unvisited POI l in category c, i.e., $\mathbf{P}_{c,l}$. Correspondingly, we obtain the categorical popularity $y_{u,l}$ for user u on POI l as follows:

$$y_{u,l} = \sum\nolimits_{c \in C} \mathbf{B}_{u,c} \cdot \mathbf{P}_{c,l}, \tag{5}$$

where 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.

Step 2: Distribution estimation of categorical popularity. As the distribution of the social check-in frequency, 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

$$f_{Ca}(y) = (\gamma - 1)(1 + y)^{-\gamma}, y \geq 0, \gamma > 1, \tag{6}$$

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

$$\gamma = 1 + |U||L| \left[\sum\nolimits_{u' \in U} \sum\nolimits_{l' \in L} \ln \left(1 + \sum\nolimits_{c \in C} \mathbf{B}_{u',c} \cdot \mathbf{P}_{c,l'} \right) \right]^{-1}, \tag{7}$$

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 3, we have also observed that the categorical popularity (i.e., the dots) in the two real-world data sets [1, 2] approaches to the power-law distribution (i.e., the line) that is estimated in terms of Equations (6) and (7). In addition, when the categorical popularity is higher than 200, the deviation of the

estimated power-law distribution becomes larger. Fortunately, the categorical popularity has a considerably low probability with the value that is higher than 200. 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 (6) 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 (5), given by

$$F_{Ca}(y_{u,l}) = \int_0^{y_{u,l}} f_{Ca}(z)dz = 1 - (1 + y_{u,l})^{1-\gamma}, \tag{8}$$

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.

4 Modeling Personalized Geographical Influence

The geographical information of POIs plays a significant influence on users' check-in behaviors and has been intensively exploited to make POI recommendations for users. Current works usually model the geographical influence as a universal distance distribution for all users. However, the geographical influence on users' check-in behaviors is unique. For instance, indoorsy persons like visiting POIs around their living areas while outdoorsy persons prefer traveling around the world to explore new POIs. Therefore, we argue that the influence of geographical information on individual users' check-in behaviors should be personalized when recommending POIs for users. In our previous studies, we model the geographical influence for each user as an individual *one-dimensional distance distribution* [3, 9] or *two-dimensional check-in distribution* [4].

This paper presents the approach that models the geographical influence as two-dimensional check-in distributions over latitude and longitude coordinates, which are more reasonable and intuitive than one-dimensional distance distributions. The reason is twofold. (1) The probability of a user visiting a location is not simply monotonous respecting their distance, because the visiting probability is not only affected by the distance but also the location's intrinsic characteristics. For example, in reality the check-in locations of a user are usually distributed in several areas. (2) It is hard to compute a visiting probability for a location based on a distance distribution, since it needs to find a reference location to derive a reasonable distance for the location in the first place. Conversely, it is considerably intuitive to employ a two-dimensional check-in distribution to compute a visiting probability for any location with latitude and longitude.

Hence, we utilize the personalized two-dimensional geographical influence for POI recommendations. Specifically, we estimate a personalized two-dimensional check-in probability density for each user, based on the kernel density estimation (KDE) that does not have any assumption on the form of the underlying distribution. Let $L_u = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\}$ be the set of locations of POIs visited by the user u, the two-dimensional check-in density f using L_u is given by:

$$f(\mathbf{l}) = \frac{1}{n\sigma^2} \sum_{i=1}^{n} K\left(\frac{\mathbf{l} - \mathbf{l}_i}{\sigma}\right), \tag{9}$$

where each location $\mathbf{l}_i = (lat_i, lon_i)^T$ is a two-dimensional column vector with the latitude (lat_i) and longitude (lon_i), $K(\cdot)$ is the kernel function and σ is a smoothing parameter, called the bandwidth. In our paper [4], we apply the widely used standard two-dimensional normal kernel:

$$K(\mathbf{x}) = \frac{1}{2\pi} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x}). \tag{10}$$

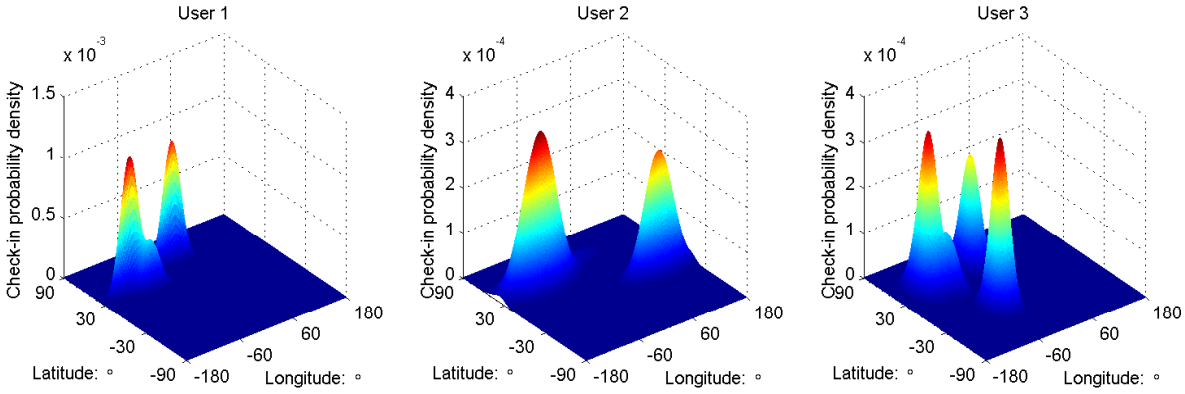


Figure 4: Personal check-in probability density over two-dimensional geographic coordinates

Figure 4 depicts the individual check-in probability density of three users randomly chosen from Foursquare based on Equation (9). We have the following two findings. (1) The geographical influence of locations on these three users' check-in behaviors is unique since their check-in probability densities are distinct from each other. (2) These check-in probability densities are usually multimodal rather than unimodal or monotonous.

5 Modeling Sequential Influence

Human movement exhibits spatiotemporal sequential influence. The sequential influence may associate with the time of a day (e.g., *people usually visit museums or libraries at daytime, go to restaurants for dinner in the evening, and then relax in cinemas or bars at night*), the geographical proximity of POIs (e.g., *tourists often orderly visit London Eye, Big Ben, Downing Street, Horse Guards, and Trafalgar Square*), the place nature and human preference (e.g., *checking in stadiums first and then restaurants is better than the reverse way because it is not healthy to exercise right after a meal*). To utilize the sequential influence for POI recommendations, current methods apply the first-order Markov chain by assuming that the next possibly visiting POI of a user only relies on her latest visited POI. Nevertheless, in reality the next POI may depend on her all visited POIs. Hence, in our previous research [6, 8], we propose a new POI recommendation approach with sequential influence based on additive Markov chain (AMC) that considers the effect of all visited POIs on the next visiting POI. In the AMC, the sequential patterns are represented as a location-location transition graph (L^2TG).

Definition 1 (L^2TG): A location-location transition graph (L^2TG) $G = (L, E)$ consists of a set of nodes L and a set of edges $E \subseteq L \times L$. Each node $l_i \in L$ represents a POI associated with an outgoing count of l_i as a transition predecessor to other POIs denoted by $OCount(l_i)$. And each edge $(l_i, l_j) \in E$ represents a transition $l_i \rightarrow l_j$ associated with a transition count denoted by $TCount(l_i, l_j)$.

In Definition 1, L^2TG is associated with *transition counts* and *outgoing counts* instead of *transition probabilities* so that L^2TG can be incrementally updated in an online fashion. In terms of *transition counts* and *outgoing counts* associated with L^2TG, *transition probabilities* can be determined based on Definition 2.

Definition 2 (Transition probability): If the outgoing count of l_i is non-zero, i.e., $OCount(l_i) > 0$, the transition probability of $l_i \rightarrow l_j$, denoted $TP(l_i \rightarrow l_j)$, is calculated by

$$TP(l_i \rightarrow l_j) = TCount(l_i, l_j)/OCount(l_i). \tag{11}$$

Otherwise, $TP(l_i \rightarrow l_j) = 1$ for $l_j = l_i$ and $TP(l_i \rightarrow l_j) = 0$ for $l_j \neq l_i$.

By Definition 2, if the outgoing count of l_i is non-zero, the transition probability of $l_i \rightarrow l_j$ is defined as the proportion of $TCount(l_i, l_j)$ to $OCount(l_i)$ in Equation (11), which is essentially the relative frequency definition of probability. On the other hand, if $OCount(l_i) = 0$ that means all users do not check in any other POIs after l_i; accordingly we define the transition probability of l_i to itself is one for simplicity.

Therefore, given a POI sequence $S_u = \langle l_1, l_2, \ldots, l_n \rangle$, our AMC defines the sequential probability of visiting a new POI l_{n+1} by

$$p(l_{n+1}|S_u) \propto \sum_{i=1}^{n} 2^{-\alpha \cdot (n-i)} \cdot TP(l_i \rightarrow l_{n+1}), \tag{12}$$

where $2^{-\alpha \cdot (n-i)}$ represents the sequence decay weight with the decay rate parameter $\alpha \geq 0$ and the larger α is, the higher is the decay rate. More importantly, the transition probability $TP(l_i \rightarrow l_{n+1})$ of l_i to l_{n+1} is weighed through leaning towards recently visited POIs, since the POIs with recent check-in timestamps usually have stronger influence on a newly possibly visiting POI than the POIs with old timestamps.

6 Modeling Temporal Influence for Time-aware POI Recommendations

Heretofore, all aforementioned modeling approaches cannot suggest appropriate time for users to visit a recommended POI, because they do not consider the influence of the temporal context when users visiting POIs on their check-in behaviors. In reality, time is a very important factor influencing human activities at different times on weekdays and weekends. To suggest properly visiting time when recommending POIs for users, existing methods split a day into time slots, e.g., 24 hours, and apply collaborative filtering recommendation techniques to infer users' preferences on POIs at each time slot separately. Unfortunately, these methods generally suffer from two major limitations due to discretization: *time information loss* and *lack of temporal influence correlations between different times.* Thus, we propose a probabilistic framework to model continuous temporal influence for time-aware POI recommendations in our recent study [7].

In the problem of time-aware POI recommendations, it is required to not only recommend interesting POIs to users based on their preferences but also suggest proper time for users to visit recommended POIs. That is, we need to predict the probability $p(l|u, T)$ of user u visiting POI $l \in L$ at time interval T. In terms of probability theory,

$$p(l|u, T) = \frac{p(l|u)p(T|u, l)}{p(T|u)} \propto p(l|u)p(T|u, l) = p(l|u) \int_{t \in T} f(t|u, l)dt, \forall l \in L, \tag{13}$$

where $p(l|u)$ is the prior probability of user u visiting POI l that is independent of time interval T and can be derived using any non-time-aware methods, and $f(t|u, l)$ is the time probability density conditioned on user u and POI l that is essential to utilize the temporal influence. We also estimate the time probability density based on KDE, given by

$$f(t|u, l) \propto \sum_{t_i \in S_{u,l}} W_{u,l}(t_i) \frac{1}{\sigma} K\left(\frac{t \ominus t_i}{\sigma}\right), \tag{14}$$

where $t \ominus t_i$ is their time difference, $S_{u,l}$ is the time sample for estimating $f(t|u, l)$ and $W_{u,l}(t_i)$ is the weight of the sample point t_i.

Note that usually u has not checked in POI l yet, so we need to obtain the time sample $S_{u,l}$ based on the two important kinds of temporal influence correlations: (1) *The check-in behaviors of different users to the same POI at different times* may be correlated. For example, a group of friends may visit a POI at different times, because they have the common interest in the POI, but with different available time. (2) *The check-in behaviors of the same user to different POIs at different times* may be correlated as well. For instance, the POIs belonging to the same category may be visited by a user at different times, because she could visit the POIs for different purposes (e.g., a user visits a restaurant for a breakfast, lunch or dinner). Thus, we can derive the time sample $S_{u,l}$ of user u to POI l by combining the check-in samples: (i) $D_{u',l}$ of another user u' visiting l (i.e., $u, u' \in U \wedge u \neq u'$)

and (ii) $D_{u,l'}$ of u visiting another POI l' (i.e., $l, l' \in L \wedge l \neq l'$). Formally,

$$S_{u,l} = \left(\bigcup_{u' \in U} \{t_i | t_i \in D_{u',l}\}\right) \bigcup \left(\bigcup_{l' \in L} \{t_j | t_j \in D_{u,l'}\}\right), \tag{15}$$

Further, we consider the cosine similarity between users or POIs using their check-in data as the sample weight of the corresponding time sample points, since the higher the similarity is, the smaller is the time difference of users visiting POIs, i.e.,

$$\forall t_i \in D_{u',l}, W_{u,l}(t_i) = sim(u, u'); \forall t_j \in D_{u,l'}, W_{u,l}(t_j) = sim(l, l'). \tag{16}$$

7 Conclusions and Future Research Directions

To recommend personalized POIs for users, this paper proposes the approaches for modeling the social, categorical, geographical, sequential, and temporal influences on the visiting preferences of users to POIs. These approaches complement each other and can be integrated together to improve the quality of recommendation results. For example, the work [5] employs the robust product rule to combine the social, categorical, and geographical influence, while the literature [10] develops a gravity model to fuse the social influence with spatiotemporal sequential influence. Hence, one research direction is to devise new methods to integrate all influences. Our recent study [10] mines the user opinions on POIs from textual comments to derive the user preferences and obtains better POI recommendations for users. Thus, another research direction is to explore new opinion mining methods for understanding the specific preferences of users on different aspects of POIs.

References

[1] H. Gao, J. Tang, X. Hu, and H. Liu. Content-aware point of interest recommendation on location-based social networks. In *AAAI*, pages 1721–1727, 2015.

[2] Yelp. Challenge Data Set. `http://www.yelp.com/dataset_challenge`, 2014.

[3] J.-D. Zhang and C.-Y. Chow. iGSLR: Personalized geo-social location recommendation - a kernel density estimation approach. In *ACM SIGSPATIAL*, pages 334–343, 2013.

[4] 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.

[5] J.-D. Zhang and C.-Y. Chow. GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *ACM SIGIR*, pages 443–452, 2015.

[6] J.-D. Zhang and C.-Y. Chow. Spatiotemporal sequential influence modeling for location recommendations: A gravity-based approach. *ACM TIST*, 7(1):11:1–11:25, 2015.

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

[8] J.-D. Zhang, C.-Y. Chow, and Y. Li. LORE: Exploiting sequential influence for location recommendations. In *ACM SIGSPATIAL*, pages 103–112, 2014.

[9] J.-D. Zhang, C.-Y. Chow, and Y. Li. iGeoRec: A personalized and efficient geographical location recommendation framework. *IEEE TSC*, 8(5):701–714, 2015.

[10] J.-D. Zhang, C.-Y. Chow, and Y. Zheng. ORec: An opinion-based point-of-interest recommendation framework. In *ACM CIKM*, pages 1641–1650, 2015.