International Journal of Pervasive Computing and Communications

A study on individual mobility patterns based on individuals' familiarity to visited areas

Jungkyu Han Hayato Yamana

A study on individual mobility patterns based on individuals' familiarity to visited areas

Jungkyu Han and Hayato Yamana
Department of Computer Science and Communication Engineering, Waseda University, Tokyo, Japan

Received 4 August 2015
Revised 6 October 2015
Accepted 1 February 2016

Abstract

Purpose – The purpose of this paper is to clarify the correlations between amount of individual's knowledge of a specific area and his/her visit pattern to point of interest (POI, interested places) located in the area.

Design/methodology/approach – This paper proposes a visit-frequency-based familiarity estimation method that estimates individuals' knowledge of areas in a quantitative manner. Based on the familiarity degree, individuals' visit logs to POIs are divided into a set of groups followed by analyzing the differences among the groups from various points of view, such as user preference, POI categories/popularity, visit time/date and subsequent visits.

Findings – Existence of statistically significant correlations between individuals' familiarity to areas and their visit patterns is observed by our analysis using 1.4-million POI visit logs collected from a popular location-based social network (LBSN), Foursquare. There exist different skewness of the visit time and visited POI distribution/popularity with regard to the familiarity. For instance, users go to unfamiliar areas on weekends and visit POIs for cultural experiences, such as museums. A notable point is that the correlations can be detected even in the areas in home city, which have not been known so far.

Originality/value – This is the first in-depth work that studies both estimation of individuals' familiarity and correlations between the familiarity and individuals' mobility patterns by analyzing massive LBSN data. The methodologies used and the findings of this work can be applicable not only to human mobility analysis for sociology, but also to POI recommendation system design.

Keywords Recommendation, Mobility, Familiarity, LBSN, POI

Paper type Research paper

1. Introduction

Human mobility analysis plays a great role in many applications, such as personalized point of interest (POI, interested places or venues) recommendations, and location-aware advertisements. Thanks to the pervasiveness of mobile handheld devices, applications can recommend a proper suggestion in a timely manner. However, for convenience, applications must minimize their user interaction because they are used while moving. Thus, capturing users' precise requirements becomes indispensable. Here, users' requirements can be extracted by analyzing their mobility patterns in a given situation.

As a huge number of location-based social network (LBSN) users share information on visited venues (e.g. geo-coordinates, venue names), researchers have studied human mobility patterns using LBSN data. One research group analyzed the mobility patterns from various points of view. The results revealed that user preferences and activity patterns can be retrieved from visited venues (Cheng *et al.*, 2011; Cho *et al.*, 2011; Joseph *et al.*, 2012). LBSN data analysis is a useful tool for business analysis (Qu and Zhang,

International Journal of Pervasive Computing and Communications
Vol. 12 No. 1, 2016
pp. 23-48

1742-7371
DOI 10.1108/IJPCC-01-2016-0010

2013; Georgiev *et al.*, 2014), city planning (Cranshaw *et al.*, 2012) and even for cultural-difference investigation (Silva *et al.*, 2014). Another group is designing POI recommendations by incorporating spatiotemporal factors such as time and distance into traditional preference models inferred from LBSN data (Ye *et al.*, 2011; Cheng *et al.*, 2012).

Although the factors considered by previous studies have improved the precision of mobility pattern analysis and the expressive power of models, we should consider more local and ephemeral factors, such as users' personal contexts, to improve the analysis. Familiarity, that is, how much an individual knows about a visited area, is an interesting context because it may bias the influence of the existing factors. Baltrunas *et al.* (2012) investigated the effect of familiarity on recommendations through the study of context-aware recommendations for sightseeing sites. Wang *et al.* (2015) proposed a familiarity-aware POI recommendation. However, they treated familiarity as one of many contexts (Baltrunas *et al.*, 2012) and applied binary familiarity concept to city-sized areas (Baltrunas *et al.*, 2012; Wang *et al.*, 2015). In addition, they did not analyze how the familiarity influences actual user mobility patterns in detail. What does happen to user mobility patterns if there is an unfamiliar area in their home city? Does unfamiliarity really exist in this case?

In this paper, we study user mobility-pattern variations caused by familiarity by using data from a popular LBSN, Foursquare. We study how user preferences and venue popularities influence users' visits, depending on their familiarity with a given area. Then, we analyze the time and category distributions of their visits in familiar/unfamiliar areas to investigate when and why users go to unfamiliar areas. Our contribution is two-fold:

(1) We propose a visit-frequency-based method to quantify the familiarity concept from LBSN data. The method allows us to investigate the familiarity concept from large-scale LBSN data.

(2) Our study findings contribute to both analyzing human visit pattern for sociology and designing POI recommendation systems which provide more user satisfactory results. This paper not only reports mobility-pattern variations with regard to the familiarity but also points out how the variations are captured in "topic model" which is one of the popular methods used in POI recommendation.

In Section 2, we describe the previous studies related to our study. In Section 3, we explain the research questions and define features such as familiarity. We describe the LBSN data we used in Section 4. In Section 5, our analysis results are shown, followed by discussion in Section 6. We conclude our study in Section 7.

2. Related work

In this section, we introduce previous research on both human mobility analysis and POI recommendations that adopt visiting patterns, and then we examine studies closely related to the familiarity concept in detail.

2.1 Human mobility analysis

Cheng *et al.* (2011) investigated check-ins (visit logs) of LBSN users and analyzed their mobility pattern and its relationship with time, distance and the user's demographic. Cho *et al.* (2011) showed that people have multiple activity areas, such as their home and

workplace. In addition, individual's current location is predictable by Multi-center Gaussian models with time. Joseph *et al.* (2012) showed that users' latent preferences to venues can be inferred from their check-ins.

Other researchers analyzed LBSN data to study economic/cultural sociology, such as business analysis, city planning and cultural diversity studies. Qu and Zhang (2013) studied the demographics and movement patterns of customers between different types of shops. Georgiev *et al.* (2014) studied the London Olympic game's impact on local retailers. Cranshaw *et al.* (2012) showed that the border lines of administrative districts and the districts derived from residents' check-ins are different. Silva *et al.* (2014) studied food consumption pattern differences between cities or countries, to investigate the cultural differences of people's eating patterns.

2.2 POI Recommendations

User mobility patterns have also been studied to devise personalized POI recommendations. Many of the previous POI recommendation studies proposed POI recommendation combining mobility pattern and user preference. The mobility pattern is represented by density distribution of users' activity over geographic areas. The activity density is inferred from geolocations of individual user's check-ins by using kernel density estimation (Zhang and Chow, 2013; Kurashima *et al.*, 2013) or simply penalizing the POIs located far from the user's nearest visited area (Ye *et al.*, 2011). A user's preference is inferred from the user's check-ins and opinions about the visited places by using memory-based (Ye *et al.*, 2011; Zhang and Chow, 2013) or model-based collaborative filtering such as matrix factorization (Cheng *et al.*, 2012; Gao *et al.*, 2015a, 2015b), topic model (Kurashima *et al.*, 2013) and combination of two different models (Liu and Xiong, 2013, Liu *et al.*, 2013a).

Gao *et al.* (2013) and Yuan *et al.* (2013) exploited the fact that a suitable visit time is different for each venue to improve the matrix factorization or memory-based collaborative filtering. Cheng *et al.* (2013) and Liu *et al.* (2013b) studied a way to utilize patterns of subsequent visits to recommend the next venue when his/her current venue was given. Yang *et al.* (2013) and Zhang *et al.* (2015) used the users' sentiments or opinions inferred from comments related to the users' check-ins to obtain more user satisfaction from the recommendations.

Some approaches focused on mitigating the cold-start problem: directly available data to infer user preference are not enough to achieve accurate recommendation. Gao *et al.* (2015a, 2015b) proposed a method exploiting neighbors' visit patterns to recommend POIs to the users who left insufficient visit logs. Bao *et al.* (2012), Yin *et al.* (2013) and Hu *et al.* (2014) studied the method to address the cold-start problem which arises when users visit a non-home city. Bao *et al.* (2012) used the POIs that are popular among local experts whose preferences are similar to the user who requests recommendation. Yin *et al.* (2013) inferred user preferences and popularity distributions over POIs in their visited city by topic model and combined them to generate recommendation. Social relationships and geographical proximity are also used to address data-sparsity problem. Zhang *et al.* (2015) exploited the fact that users tend to go to the POIs where their friends visited. Hu *et al.* (2014) improved recommendation quality by using the fact that users tend to give similar ratings to geographically closely located POIs.

2.3 Studies related to familiarity

Bao *et al.* (2012) recommended venues frequently visited by local experts whose preferences are similar to the user who wants a recommendation in a non-home city. Lian *et al.* (2014) assumed that the reason why a venue was unvisited by a user was either disliked or simply unknown by the user. They gave different weights to each type of un-visit. To distinguish the two different types, they used the distance from the activity area, that is, how far the unvisited venue is from the areas where the user frequently visits. For example, if an unvisited venue is near or in his/her activity area, we conclude that he/she may dislike the venue. If the venue is far from the activity area, he/she may simply not know about the venue.

Yin *et al.* (2013) tried to find a proper blending ratio between the influence of the area where the user visits and his/her visiting preferences to build an area-aware POI recommendation. Gao *et al.* (2015a, 2015b) used venues visited by neighbors to complement insufficient visit logs of the recommendation-requested-user under the condition that they have similar activity areas and preferences. Baltrunas *et al.* (2012) considered the familiarity as part of user context to build a context-aware recommendation. Wang *et al.* (2015) studied a familiarity-aware POI recommendation to provide better recommendation to users who visit both home and other cities.

Our research focuses on mobility pattern differences with regard to the degree of the user's familiarity. In contrast, Bao *et al.*'s (2012) method recommends constant venues that are popular among local experts, regardless of the user's familiarity. Therefore, from an analysis perspective, Bao *et al.* (2012) analyzed the mobility pattern of local experts who are familiar with the city. Lian *et al.* (2014) adopted the familiarity concept; however, the study did not focus on the user's mobility-pattern variation, but focused on how to infer the user preferences to venues. Thus, they are different from our study.

Yin *et al.* (2013) tried to blend the visited areas' influences and the users' visit patterns. The blending factor varies for each user, but from single user's point of view, the factor remains static to all areas. This means that the model cares about the visited areas' influence, but cannot care about the user's familiarity with the areas. Gao *et al.* (2015a, 2015b) may implicitly include the familiarity concept in the model to some extent, but they did not explicitly exploit the familiarity concept. Baltrunas *et al.* (2012) and Wang *et al.*'s (2015) works are strongly related to our study. However, in Baltrunas *et al.*'s method, the degree of familiarity is explicitly given by the user, while our study focuses on inferring the familiarity automatically from data. In addition, both of the methods used the simple familiarity to a city-sized area, for example, familiar home city and unfamiliar cities far from home city. Thus, their study did not give a detailed analysis about the mobility-pattern variation with regard to gradual changes of the familiarity degree.

3. Questions and definitions

In this section, we explain our investigation and definitions.

3.1 Research questions

We investigate how individual user's mobility pattern in an area varies with regard to the user's familiarity with the area. We try to answer the four research questions below. In this paper, we define "user u's familiarity with area a" as the degree of how many

times user u has visited area a. We assume that the number of visits indicates the user's knowledge level about the area:

RQ1. How are users' preferences and venues' popularity affected by the users' familiarity with areas?

RQ2. When and why do people go to familiar/unfamiliar areas?

RQ3. How does individual's subsequent visit pattern differ between familiar and unfamiliar areas?

RQ4. Can the familiarity-caused visit pattern differences be captured by adopting "topic model" which is frequently used for POI recommendations?

Both of an individual's preference to a venue and the venue's popularity compose a major part of the individual's willingness to visit the venue. As most POI recommendation models (Ye *et al.*, 2011; Cheng *et al.*, 2012; Zhang and Chow, 2013, Kurashima *et al.*, 2013; Gao *et al.*, 2013; Yuan *et al.*, 2013) are based on these two factors, it is worth investigating *RQ1*. It is interesting to know why and when people go to a familiar/unfamiliar area because it reflects both their activity pattern and the cultures in the areas. A subsequent visit pattern gives a hint for understanding the actual movements in familiar/unfamiliar areas. Because of the aforementioned reasons, we set *RQ2* and *RQ3*. We select *RQ4* because we want to examine how the familiarity-related visit patterns are captured by "topic model" to find more specific directions toward familiarity-aware POI recommendations.

3.2 Definitions

We define six concepts indispensable to our investigation: check-in, area, activity area, familiarity, preference and popularity.

3.2.1 Check-in. Check-in indicates a user visit log. A check-in is shown as a tuple of *<user, venue, category, tags, location, timestamp>*, where user represents "user ID"; venue shows the visited "venue ID"; category represents the venue's classification listed in Section 4.2; tags is a set of the short free-keywords that describe the venue's characteristics; location is a geographic coordinate of the venue; and timestamp shows the time of check-ins. We will use small letters u, v and c as instances and capital letters U, V and C as the set of all users, venues and categories, respectively.

Here, a single physical venue is mapped to a single venue ID. Thus, different venue IDs are assigned to the venues that have the same name, but are located at different locations. For example "64 ice-cream" in 1st Street and "64 ice-cream" in 2nd Street have different IDs. Besides, we assume that each venue has only one category. For example, the venue named "OO Hamburger" is classified into "Burger Joint" category only. As for tags, a venue may have plural tags that are generated by users voluntarily. For instance, "OO Hamburger"'s tags can be "Hamburger" and "Delicious".

3.2.2 Area and activity area (A.A.). An area is a geographic area where we assume the user's familiarity is the same, that is, a unit of geographic area. As shown in Figure 1, we divide a city into 2×2 km areas that do not intersect with each other. We do not use an entire city as a single area because we want to observe the continuous movement-pattern-change, even in the same city. We use a small letter a and its capital A as an instance and the set of all areas, respectively.

User u's A.A. is the area where a substantial number of u's check-ins occur and are concentrated. The concentration of check-ins is important to distinguish whether the area is a real A.A. or a normal area whose check-ins came from the outskirts of adjacent A.A.s. Most users have more than one A.A., for example their home area and workplace (Cho *et al.*, 2011). To find A.A.s from check-in data, we define an A.A. as the area which has at least n check-ins concentrated in a given range in the area. We set n as 5 and a 0.5-km-radius circle as the range. We also define the "Primary A.A. of user u" as the A.A. with the largest number of u's check-ins.

3.2.3 Familiarity. Familiarity, $f_{u,a}$, represents the degree of how many times user u visits (and therefore knows more about) area a. We define $f_{u,a}$ as the ratio of the number of check-ins made in a by u to the largest number of check-ins made by u in a single area. Therefore, $f_{u,a}$ is defined for each (user, area) pair. Here, the value of $f_{u,a}$ becomes 1 for area a where her/his largest check-ins are left by u, and becomes 0 for area a where no check-ins are left by u.

Regardless of the definition of $f_{u,a}$, we must consider the situation that it is possible that the number of check-ins in an area underrepresents the user's actual visits to the area, because check-in data are voluntarily generated by users. Therefore, we adopt the expected number of user's check-ins in a given area a instead of the actual number of user's check-ins in a when the expected number is larger than the actual number [Eq. (1)].

To calculate the expected number of check-ins in a given area a, we exploit the probability of check-ins made by the user. Figure 2 plots the check-in probability of NYC Foursquare users with regard to the distance between the checked-in venue and the user's nearest A.A. (please refer to Section 4 for details on the data). If we know the

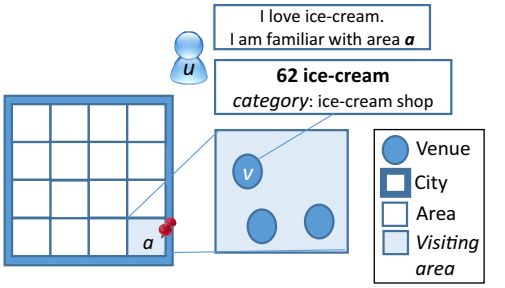


Figure 1.
Areas in a city

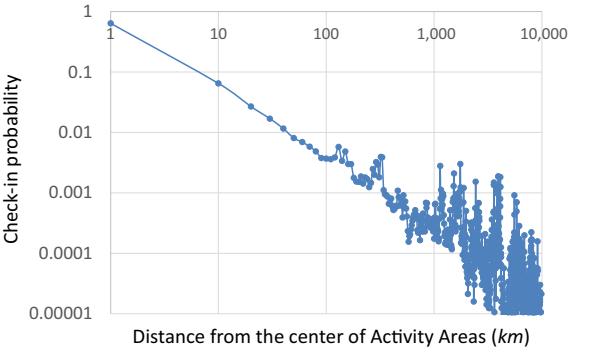


Figure 2.
Check-in probability with regard to distance (New York City)

check-in probability of the distance, we can calculate the expected number of check-ins of user u for area a by multiplying the check-in probability of a and the number of check-ins of u in u's primary A.A. [equation (2)].

It is well-known that the decrement probability follows the power-law distribution (Cheng *et al.*, 2011; Ye *et al.*, 2011) [equation (3)]. As equation (2) decreases by the power-law as the distance increases, we use a logarithm when we calculate the ratio in equation (1) to prevent $f_{u,a}$ from an extremely steep decrease compared to the distance increase. If the logged value is less than 0, we use 0.

$$f_{u,a} = \max\left(\frac{\log(n'_{u,a})}{\max_{a' \in A}(\log(n_{u,a'}))}, \frac{\log(n_{u,a})}{\max_{a' \in A}(\log(n_{u,a'}))}\right) \tag{1}$$

$$n'_{u,a} = \left(\max_{a' \in A} n_{u,a'}\right) \cdot p_{chk}(d_{u,a}) \tag{2}$$

$$p_{chk}(d_{u,a}) = z_1 \cdot d_{u,a}^{\,z_2} \tag{3}$$

Notations: $n'_{u,a}$: the estimated number of check-ins made in area a by user u; $n_{u,a}$: the actual number of check-ins made in area a by user u; $d_{u,a}$: km distance between area a and the nearest A.A. of user u (use 1 km when the distance is less than 1 km); and $z1$, $z2$: constants ($z_1 = 1.0$, $z_2 = -1.1$).

3.2.4 Preference and popularity. We define preference and popularity to find an answer to *RQ1*.

Preference, $pref_{u,c}$, represents the importance degree of category c to user u. $pref_{u,c}$ is defined for each (user, category) pair. A larger value of $pref_{u,c}$ indicates that category c is more important to user u. We adopt both category and check-in count-based preferences, because a category-based preference is more interpretable than a venue-based preference when we compare the same user's preferences for multiple areas with different familiarities. We use a simple count-based algorithm to eliminate any unexpected latent factor's effect that could be introduced from more complex models. We followed Bao *et al.*'s (2012) term frequency–inverse document frequency (*tf-idf*) style preference [equation (5)], but used a normalized version [equation (4)] as $pref_{u,c}$ to limit the range to [0,1].

$$pref_{u,c} = \frac{r_{pref_{u,c}} - \min_{c'' \in C} r_{pref_{u,c''}}}{\max_{c' \in C} r_{pref_{u,c'}} - \min_{c'' \in C} r_{pref_{u,c''}}} \tag{4}$$

$$r_{pref_{u,c}} = \frac{s_{u,c} + 1}{\left(\sum_{c' \in C} s_{u,c'}\right) + 1} \times log\left(\frac{|U|}{i_{c,*}}\right) \tag{5}$$

Notations: $pref_{u,c}$: u's preference to category c; $s_{u,c}$: the number of visits of u to venues in category c; and $i_{c,*}$: the number of distinct users who visited at least one venue in category c.

Popularity, $pop_{v,a}$, indicates how many distinct users visited venue v located in area a [equation (6)]. Popularity is calculated for each identical venue but is normalized to the

area in which they are located. Therefore, the most-popular venue in a given area has value 1 and the least-popular venue in the area has value 0. We use a logarithm to prevent an abnormally popular venue from over lowering the popularity of the other venues.

$$pop_{v,a} = \frac{\log(i_{v,*}) - \min_{v'' \in located(a)} (\log(i_{v'',*}))}{\max_{v' \in located(a)} (\log(i_{v',*})) - \min_{v'' \in located(a)} (\log(i_{v'',*}))} \quad (6)$$

Notations: located(a): the set of venues located in area a and $i_{v,*}$: the number of distinct users who visited venue v at least once.

4. Data

In this section, we explain the LBSN data and the venue categories used in our analysis.

4.1 LBSN data

We used Foursquare[1] check-in data for our analysis. We gathered publicly available check-in data via Twitter[2] using a Twitter API[3] during a 1.5-year period (2013.7-2015.1). We selected three cities for analysis: New York City (NYC), Tokyo and Los Angeles (LA). We selected large cities because we expected that an unfamiliar area would exist in the city, even for users who live in or near the city. In addition, these cities have more check-ins in the data set than other cities. We only used the check-ins of local users whose home was in the city because we wanted to find a user's mobility-pattern difference between areas in their home city. We assumed that a user is a local user of the city if half of their activity areas are in the city. We applied the density based spatial clustering of applications with noise (DBSCAN) (Sander *et al.*, 1996) (eps:1, mitPts:5) clustering algorithm to the check-in data of each user to find the user's activity areas. Table I shows the statistics of the data we used.

In Table I, we also list the category classification error and the location error of the check-in data. The category classification error indicates how many venues in the check-in data have misclassified categories. If a human-judged category of a venue is different from the category given by Foursquare, then the venue has a misclassified category. For example, if a Burger Joint was classified into Bar by Foursquare, then we say the category was misclassified because the Burger Joint should be classified into Fast-food. The location error indicates the geographical distance between the Foursquare-given geolocation for the venue and the actual geolocation retrieved from map information.

Table I. Check-in data statistics

City	Range [from, to:(lat, long)], width/height	Home users	Venues	Check-ins	Error (Category, Location)
NYC	(40.49, −74.27), (40.92, −73.67) 50.87/47.75 km	4,199	53,718	710,010	1%, 20 m
Tokyo	(35.59,139.50),(35.86,139.93) 38.97/29.96 km	2,754	58,973	331,704	2%, 20 m
LA	(33.68, −118.67),(34.34, −118.04) 58.42/73.21 km	1,790	34,982	393,497	1%, 20 m

We randomly sampled 100 venues for each of the cities and then manually investigated the error by using Web searching and an online map service. In result, the sampled data sets have 2 per cent of the category classification error at the most, and less than 20 m of the location error.

4.2 Categories

Our adopted categories are shown in Table II. It is based on the Foursquare hierarchical categories[4]. We directly used three depth-1 Foursquare categories: Arts & Entertainment, Shop & Services and Outdoors & Recreation. In addition, we adopt four more categories by modifying two depth-1 Foursquare categories: "Food" and "Nightlife spot". The categories mainly contain food and drinking places. Silva *et al.* (2014) showed that peoples' eating styles well-represent their preferences.

The additional categories are constructed by re-grouping the sub-categories in "Food" and "Nightlife spot" into the four categories: "Slow food", "Fast food", "Hard (Alcoholic) drink" and "Soft (Non-Alcoholic) drink". The sub-categories in "Food" or "Nightlife spot" that cannot be a sub-category in any of the four categories are discarded. For instance, "Night Market" in "Nightlife spot" was discarded because it is not related to eat or drink. Our only difference from Silva *et al.* (2014) is that we divided the "Drink" category into "Hard drink" and "Soft drink" because we think that there is a visit-pattern difference between pubs and coffee shops.

We omit the other Foursquare categories because users do not have much freedom of choice when they visit those venues. For example, sub-categories of "Professional & Other Categories" are office, medical center, etc. In most cases, the visited venues in the sub-categories are not a matter of preference. The distribution of check-ins and venues over the seven categories is shown in Table III.

Category	Sub-category
Arts & Entertainment	Movie theater, music venue, stadium, museum, performing arts venue
Shop & Services	Food & drink shop, clothing store, department store, gym/fitness center
Outdoors & Recreation	Park, plaza, athletics & sports, beach
Slow food	American/Japanese/Mexican restaurant
Fast food	Café, pizza place, burger joint, ramen/noodle house
Hard drink	Bar, pub, sake bar
Soft drink	Coffee shop, juice bar, tea room

Table II. Seven categories and examples of sub-categories

	NYC		Tokyo		LA	
Category	Check-in	Venue	Check-in	Venue	Check-in	Venue
Arts & Entertainment	11	8	9	4	11	7
Shop & Services	26	34	24	21	31	37
Outdoors & Recreation	7	8	8	5	6	7
Slow food	22	21	24	36	22	22
Fast food	13	17	19	19	13	17
Hard drink	15	7	8	11	8	4
Soft drink	5	4	5	3	7	4
Total	100	100	100	100	100	100

Table III. Distribution of check-ins/venues over categories

5. Analysis result

In this section, we describe our analysis to answer the four research questions introduced in Subsection 3.1 and discuss the results of the analysis. In the analysis, we use only the areas that have more than 120 distinct venues to avoid the polarized popularity caused by an extremely small number of venues in the area.

5.1 Familiarity versus Preference/Popularity

RQ1. How are users' preferences and venues' popularity affected by the users' familiarity with areas?

Here, we investigate how both of user u's preference to category c ($pref_{u,c}$) of visited venue v and v's popularity in area a ($pop_{v,a}$) change with regard to u's familiarity to a ($f_{u,a}$) where v is located. We group the check-ins of all users into one of the 11 groups, according to the value of $f_{u,a}$. Group 0 has familiarity with $f_{u,a}$ range [0.0, 0.1), Group 0.1 has [0.1, 0.2) … and Group 1 has [1.0, 1.0]. We will refer to Group x as *familiarity x* for simplicity.

5.1.1 Movement distance. Before we investigate the preference and popularity's effects from the familiarity's perspective, we list the statistics in Figure 3 to grasp a rough picture of the visited areas' distribution, and the distance from the nearest A.A, with regard to the familiarity. In Figure 3(a), represents the accumulated percentage of the number of visited areas, and Figure 3(b) represents the mean of the distances between the users' visited areas and the users' nearest activity areas, over the 11 groups. From the statistics, we can observe that most areas, except for the areas with familiarity 0, are located within 6 km from the users' A.A.; the familiarity increases linearly as the distance decreases. The number of visited areas starts to decrease at about 3 km from the users' A.A. (at familiarity 0.5) as the distance increases.

5.1.2 Validity test. We calculate the mean of the all users' preferences to each category and the mean of the visited venues' popularity to each category, over the 11 groups, that is, 11 familiarity-degrees. We carried out two tests for each category. The first test is a statistical significance test ($0.01 > p$) of the difference of the means between each of familiarity pairs (0, 0.5) and (0.5, 1). If at least one pair passed the first test, then we test for correlation (Pearson, 1895) between familiarity and the preference/popularity to the ranges that passed the first test.

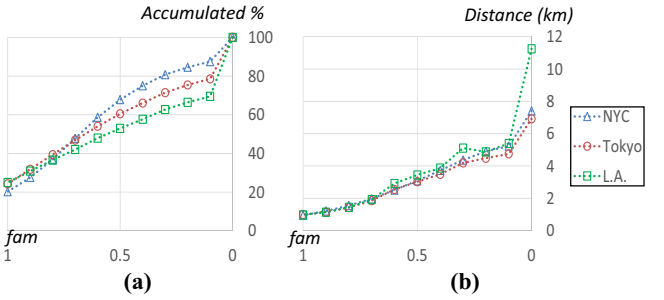


Figure 3.
Familiarity versus number of areas, distance

Notes: (a) # of Areas; (b) distance

If the coefficient is greater than 0.7 or less than -0.7, we can say that there is a strong positive or negative correlation. We can also say that a category is familiarity related if either the preference or the popularity of the category passed the first test and the absolute value of the coefficient is over 0.7. We plot the final results of five familiarity-related categories in Figure 4 and list the top-five sub-categories of the five categories in Table IV. For comparison, if a category of a single city is familiarity related, we also plot the values of the other cities in the same category.

5.1.3 Slow food. The preference has a strong positive correlation with the familiarity, while the popularity has a strong negative correlation. This implies that users tend to visit restaurants they like in familiar areas, while they visit popular restaurants in unfamiliar areas, regardless of their preference.

Slow food usually requires more cost and more time than fast food. Therefore, users put more weight on their preference when they consider slow-food venues. In addition, the result can be explained by the concept of "Knowledge and Objective". As users frequently visit familiar areas, they have more chances to find restaurants they like or to develop opinions that may contradict popular opinions. However, in unfamiliar areas, popular venues have a higher chance of being recognized by the users. In addition, in most cases, users do not have a strong opinion about the venues in unfamiliar areas. From an objective point of view, visiting an unfamiliar area specifically to eat some food is rare; in most cases, food is not the primary goal for visiting unfamiliar areas. Therefore, a preference for food is less important in an unfamiliar area than in a familiar area like workplace, where dining can be a primary objective.

Sub-category distributions in the same city are similar between familiar and unfamiliar areas (underlined sub-categories in Table IV. Slow food, indicate the sub-categories that appears in both of familiar and unfamiliar areas). As the top-1

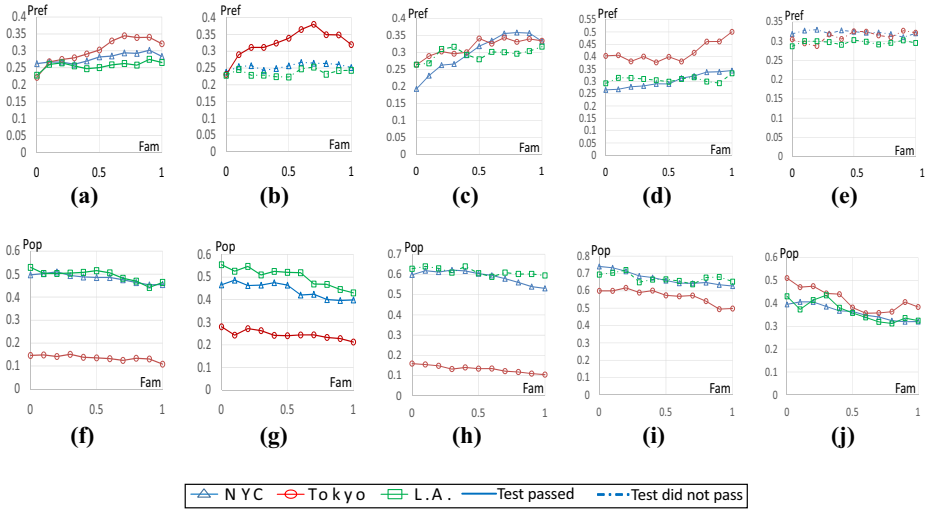


Figure 4.
Preference to/popularity of visited venues with regard to familiarity

Notes: (a) Slow food-pref; (b) fast food-pref; (c) hard drink-pref; (d) arts & Ent.-pref; (e) shop & serv.-pref; (f) slow food-pop; (g) fast food-pop; (h) hard drink-pop; (i) arts & Ent.-pop; (j) shop & Serv.-pop

Table IV. Top five sub-category distributions of five categories in familiar/unfamiliar areas (100 per cent in total, for each fam)

City	Fam	Slow food (%)	Fast food (%)	Hard drink (%)	Arts & Entertainment (%)	Shop & Services (%)
NYC	[0.9, 1]	American restaurant (16)	Burger joint (11)	Bar (43)	**Music venue (24)**	**Gym/fitness center. (42)**
		Italian restaurant (11)	Pizza place (11)	Pub (10)	**Movie theater (23)**	Food & drink shop (13)
		Mexican restaurant (10)	Cafe (10)	Cocktail bar (8)	Performing arts (19)	Clothing store (7)
		Dinner (4)	Bakery (9)	Sports bar (8)	Stadium (9)	Salon/barbershop (4)
		New America restaurant (4)	Sandwich place (9)	Gay bar (8)	General entertainment (7)	Electronics store (3)
	[0, 0.1]	American restaurant (15)	Burger joint (17)	Bar (45)	**Stadium (24)**	Gym/fitness c. (22)
		Italian restaurant (9)	Pizza place (15)	Beer garden (12)	**Museum (17)**	Food & drink shop (12)
		Mexican restaurant (7)	Bakery (9)	Cocktail bar (8)	Performing arts (13)	**Mall (9)**
		Chinese restaurant (6)	Café (8)	Pub (7)	Music venue (13)	Clothing store (4)
		BBQ joint. (5)	Food truck (7)	Sports bar (7)	Movie theater (12)	Department store (8)
Tokyo	[0.9, 1]	Japanese restaurant (34)	**Ramen/noodle (51)**	Sake bar (37)	**Movie theater (26)**	Mall (13)
		Chinese restaurant (11)	Café (23)	Bar (36)	**Music venue (19)**	Convenience store (10)
		Italian restaurant (8)	Fast food restaurant (8)	Pub (8)	Arcade (9)	**Gym/fitness center (10)**
		Indian restaurant (7)	Bakery (3)	Karaoke bar (5)	Theme park (8)	Food & drink shop (8)
		Dinner (6)	Burger joint (2)	Wine bar (4)	Art gallery (7)	Electronics store (8)
	[0, 0.1]	Japanese restaurant (35)	**Ramen/noodle (52)**	Sake bar (44)	Music venue (20)	**Mall (25)**
		Chinese restaurant (8)	Café (25)	Bar (21)	**Stadium (13)**	Electronics store (10)
		BBQ joint. (7)	Fast food restaurant (6)	Pub (10)	**Museum (12)**	Department store (7)
		Italian restaurant (7)	Bakery (3)	Karaoke bar (6)	Concert hall (11)	Food & drink shop (6)
		Restaurant (6)	Dessert shop (2)	Beer garden (6)	Movie theater (10)	Record shop (6)
L.A.	[0.9, 1]	American restaurant (17)	Burger joint (16)	Bar (30)	**Movie theater (26)**	**Gym/fitness center (33)**
		Mexican restaurant (13)	Café (13)	Gay bar (14)	**Music venue (19)**	Food & drink shop (15)
		Sushi restaurant (8)	Pizza place (9)	Gastro pub (9)	General entertainment (12)	Mall (8)
		Italian restaurant (7)	Sandwich place (7)	Sports bar (7)	Stadium (11)	Clothing store (5)
		Dinner (6)	Fast food restaurant (7)	Cocktail bar (6)	Performing arts (8)	Department store (3)
	[0, 0.1]	American restaurant (18)	Burger joint (18)	Bar (26)	Movie theater (23)	**Mall (21)**
		Mexican restaurant (9)	Café (13)	Gastro pub (16)	Music venue (14)	Food & drink shop (11)
		Italian restaurant (8)	Pizza place (8)	Cocktail bar (9)	Performing arts (13)	Gym/fitness center (11)
		Sushi restaurant (6)	Breakfast spot (7)	Pub (7)	General entertainment (12)	Clothing store (8)
		Dinner (5)	Sandwich place (6)	Gay bar (7)	Concert hall (10)	Department store (5)

sub-category represents the country where the city is located, everyone visits the sub-category regardless of the familiarity degree. However, in unfamiliar areas, the proportion of the other sub-categories decreases, and thus has a more diverse distribution than in a familiar area.

5.1.4 Fast food. The popularity increases in unfamiliar areas. However, there is no strong correlation between the preference and the familiarity, except in Tokyo.

The reason why the preference does not show a strong correlation to the familiarity is caused from the purpose of fast food. People put a higher priority on saving time than on their preference when they visit fast-food venues. However, we can find a strong preference change between familiarity 0.0 and 0.7 in Tokyo. This is interesting because, in "Fast food" in Tokyo, the top five sub-category-distributions between familiar areas and unfamiliar areas are similar to each other (Table IV).

We think that the phenomenon is caused by the Ramen/Noodle sub-category, the most popular fast food in Japan. Ramen/Noodle has over 50 per cent of the fast-food check-in in both familiar and unfamiliar areas. Unlike other fast foods, there are many Ramen fans in Japan. The fans eat Ramen mostly for lunch and sometimes for dinner (thus, the location is mostly near an activity area). However, many ordinary people also consume Ramen in unfamiliar areas because Ramen is considered as fast food. We think that this is the reason why the preference for fast food in Japan has a positive correlation to the familiarity.

5.1.5 Hard drink. NYC and Tokyo's hard-drink patterns are similar to the pattern of slow food. We can explain this by using the reasoning similar to that of slow food. However, LA could not pass the statistical tests. We suspect that the socializing characteristic of Hard drink contributes to LA's results. Bars or pubs are frequently visited with friends, which makes it difficult to follow individuals' preferences. Thus, it is possible that the pattern becomes different due to the characteristics of L.A. residents. However, we do not have any social information, such as friend links; therefore, we will leave the issue for future work.

5.1.6 Arts & Entertainment. For users in NYC and Tokyo, preference has a strong positive correlation to the familiarity, while popularity has a strong negative correlation, as with slow food. However, L.A. again could not pass the statistical tests.

In NYC and Tokyo, movie theater and music venue are major sub-categories in familiar areas, while stadium and museum are emerging sub-categories in unfamiliar areas. The movie theater and music venue's check-in percentage drops in unfamiliar areas. Normally, we do not visit stadiums and museums more frequently than movie theaters and music venues. As our preference is a visit-frequency-based metric, it is natural that the preference is lower in unfamiliar areas. The results can also explain why people go to unfamiliar areas. A stadium or museum provides unique services that cannot be experienced at any other type of venue. For example, we must go to the Louvre to meet the real Mona Lisa. In contrast, a newly released movie is available at many movie theaters. In L.A., movie theater, music venue and general entertainment have about 50 per cent of the check-ins in both familiar and unfamiliar areas.

5.1.7 Shop & Service. There is no correlation between the preference and the familiarity. However, people still tend to go to more-popular venues when they visit unfamiliar areas.

We think that the phenomenon is caused by the fact that users visit Shop & Service venues only when they need something from the venues. For example, users visit a clothing store when they need clothes. Therefore, despite the many check-ins related to the category, users' preferences vary over the sub-categories. However, there are

categories that appear frequently in familiar or unfamiliar areas. Gym/fitness center is a popular sub-category in familiar areas, while Mall emerged in unfamiliar areas. As gym/fitness centers are related to private exercise, they appear in areas near activity area. A mall is a large enclosed shopping area. Therefore, going to a mall becomes a good reason for users to go to unfamiliar areas.

5.1.8 Others. We discuss neither "Outdoors & Recreation" nor "Soft Drink", because we did not find any statistical significance in the categories.

5.2 Time and reasons to visit

RQ2. When and why do people go to familiar/unfamiliar areas?

People go to unfamiliar areas after work on weekdays, or enjoy a weekend lifestyle more in unfamiliar areas. Going to venues of Arts & Entertainment becomes an important motivation for people to visit unfamiliar areas. Especially, going to the venues that provide unique services, such as museums or stadiums, may become a strong motivation (Subsection 5.1.). As we know, food-related venues are popular venues to visit, regardless of familiarity; however, they are more important in familiar areas. To confirm the above discussion, we analyzed the temporal pattern and category distribution of the check-ins.

5.2 1 Temporal pattern. To answer *RQ2*, we first investigate how many check-ins are made on weekdays and weekends. We grouped check-ins into two groups, weekday (Mon. to Fri.) and weekend (Sat. to Sun.), according to the timestamp of the check-in. We further sorted each of the groups into the 11 sub-groups, according to the user's familiarity with the area. In Figure 5, we plot the ratio of the number of check-ins made on weekends to the number of check-ins made on weekdays to each of the 11 groups. The red line indicates the weekends to weekdays ratio calculated from all check-ins. (Familiarity is not considered.) The ratio increases as familiarity decreases. This implies that people tend to go to unfamiliar areas on weekends.

In Figure 6, we plot the distribution of check-ins over the hours of the day. A solid line represents the check-in distribution in familiar areas (range of the familiarity: [0.9, 1.0]), and the dashed line represents that of unfamiliar areas ([0.0, 0.1]). On weekdays, check-ins in unfamiliar areas are more concentrated in the evening (17: 00-20:00). On weekends, the overall shape of the unfamiliar area distribution is similar to the familiar area. However, the unfamiliar area has a higher percentage of check-ins at the peak period. Regardless of weekdays and weekends, there are fewer check-ins in the morning in unfamiliar areas.

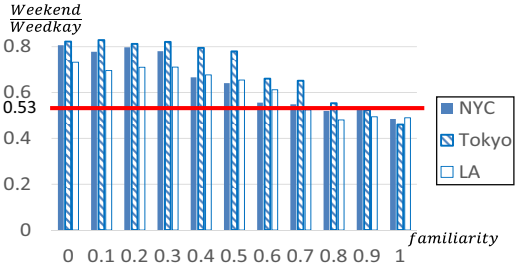
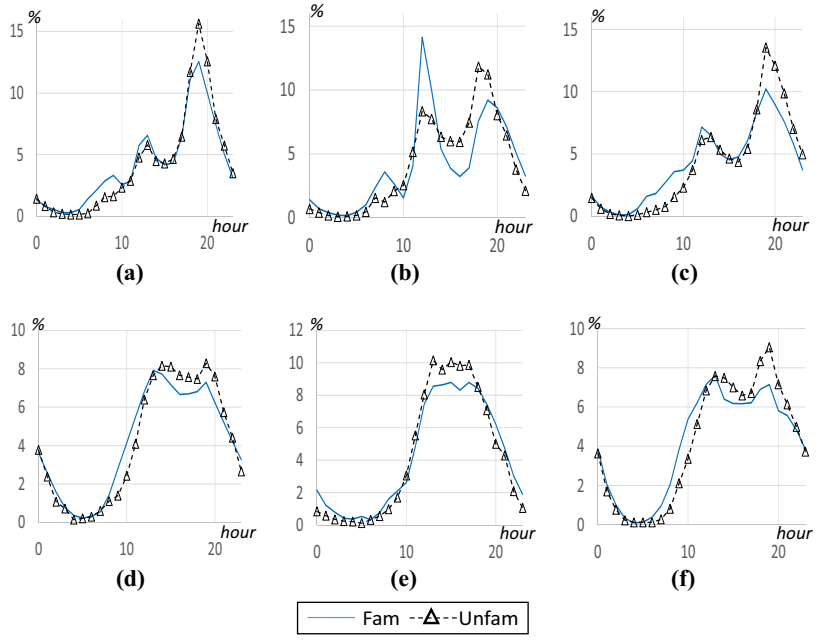


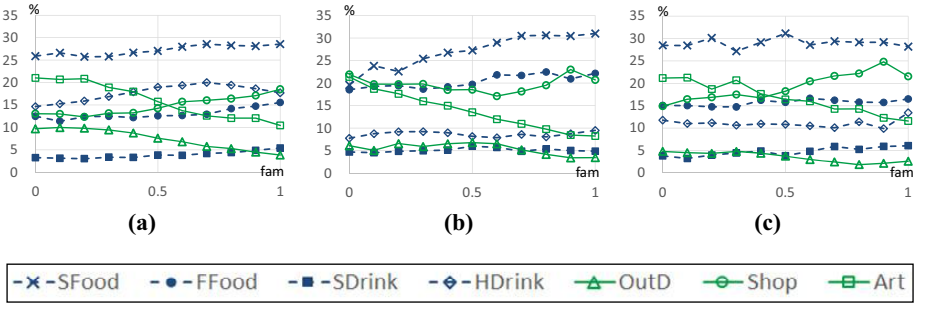
Figure 5.
Weekends to weekdays check-in ratio



Notes: (a) NYC-weekdays; (b) Tokyo-weekdays; (c) L.A.-weekdays; (d) NYC-weekends; (e) Tokyo-weekends; (f) L.A.-weekends

Figure 6. Check-in distributions over time period

5.2 2 Category distribution. We show the category distributions over the familiarity degrees in Figure 7. In all three cities, the percentage of Art & Entertainment in unfamiliar areas is almost twice than the percentage in familiar areas. The percentages of food-related categories (dashed lines) decreased slightly in unfamiliar areas, but they still have over 50 per cent of the total check-ins. We can also find city-specific phenomena. Shop & Service-related check-ins decreased in unfamiliar areas in NYC and LA. However, shopping is still an important activity in unfamiliar areas in Tokyo.



Notes: (a) NYC; (b) Tokyo; (c) L.A

Figure 7. Category distributions over areas of different familiarities

Instead, "Slow food" is decreased in unfamiliar areas in Tokyo, while the changes are small in the other cities.

5.3 Familiarity versus visit pattern

RQ3. How does individual's subsequent visit pattern differ between familiar and unfamiliar areas?

In this subsection, we investigate the categories, time and distance of two subsequent check-ins to determine the users' visit-pattern differences between familiar and unfamiliar areas.

We extracted all of the check-in pairs that were made subsequently in the same area and the same day by the same users. Then, we grouped the check-in pairs into the 11 groups, according to the familiarity. For each group, we counted the number of the category pairs appearing in the check-in pairs in the group. We treated the two different temporal orders of the subsequent categories (e.g. category A to B and B to A) as the same category pair. For each group, we also calculated the timestamp difference between the check-ins in the check-in pairs, and then calculated the mean of the difference for each of the category pairs.

Figures 8(a)-(c) show the distribution of each pair, and Figures 8 (d)-(f) show the mean time difference between visits of each pair to unfamiliar areas (range of the familiarity: [0.0, 0.1]), middle areas ([0.6, 0.5]) and familiar areas ([0.9, 1.0]) of the three cities. Due to space limitations, we omitted the 10 lowest category pairs. Regardless of city and familiarity, (Slow food, Slow food), (Fast food, Fast food), (Shop & Service, Shop & Service) and (Hard drink, Hard drink) show high occupation. This is natural because we eat lunch and dinner. We also frequently go to another shop or bar when we go shopping or drinking. The time between visits of the shop pair is shorter than the food-related pairs because shopping does not require time for digestion.

In unfamiliar areas, "Arts & Entertainment"-related pairs become important. The (Shop & Service, Shop & Service) pair has the greatest occupation difference between familiar and unfamiliar areas. This fact supports that people go to unfamiliar areas to visit "Arts & Entertainment". Shopping is the most important non-food-related activity in familiar areas, which was discussed in the previous subsections.

The time between visits normally increases as the familiarity increases, regardless of the category. We think that this is due to users' more-concentrated activity in unfamiliar areas (Figure 6). Despite minor differences between the cities, the mean and median distance between two consecutive check-ins is roughly less than 350 m and 250 m in all the ranges of familiarity.

5.4 Familiarity represented in topic model

RQ4. Can the familiarity-caused visit-pattern differences be captured by adopting "topic model" which is frequently used for POI recommendation?

In the previous subsections, we analyzed the familiarity effects as statistics of whole users or areas. In this subsection, we examine how the familiarity-affected visit patterns are captured by topic model to find more specific directions toward familiarity-aware POI recommendation which is one of the most probable and important applications.

We adopt topic model because of two reasons. First, there exist many POI recommendation algorithms (Kurashima *et al.*, 2013; Yin *et al.*, 2013; Liu and Xiong,

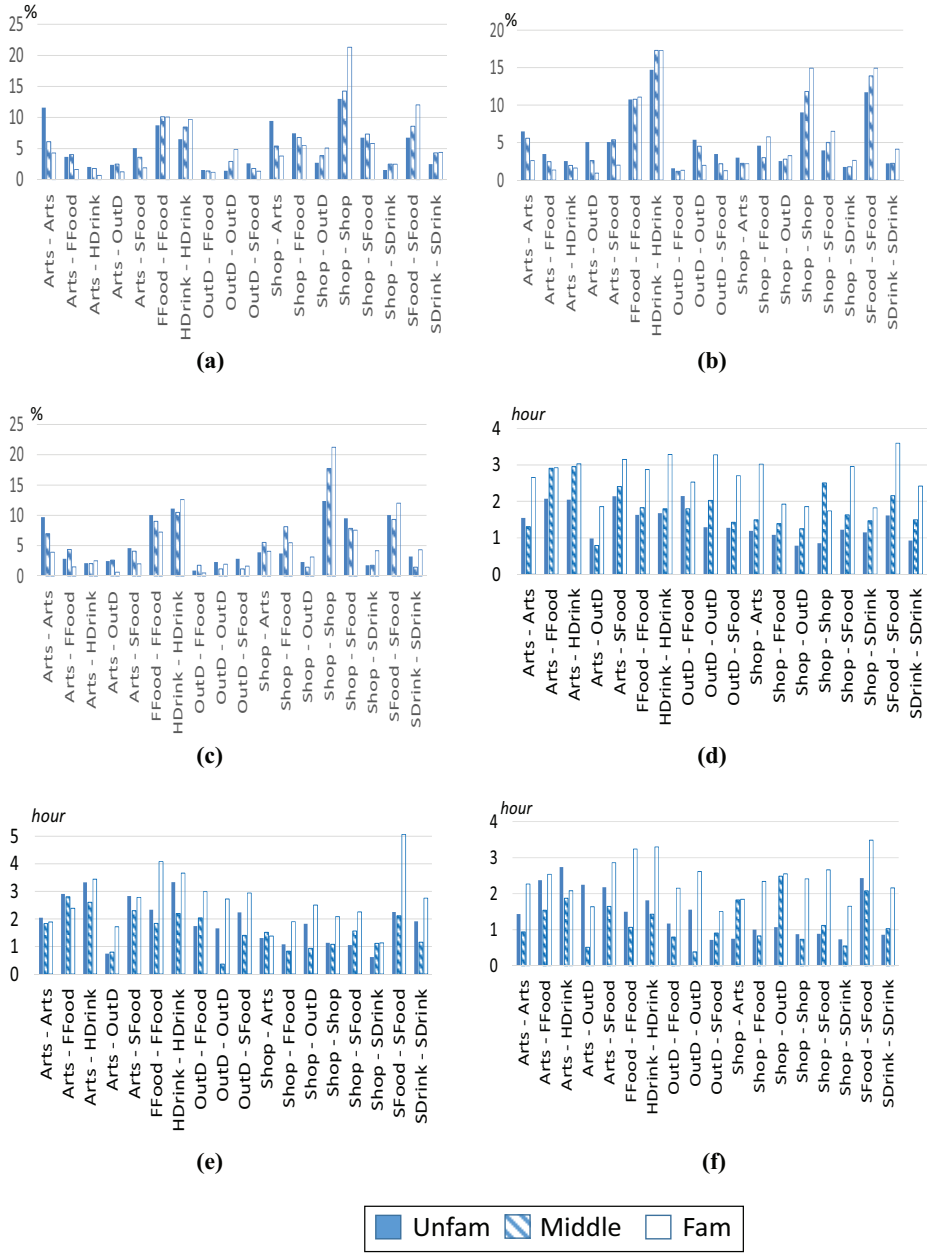


Figure 8.
Category pairs distribution of subsequent check-ins, and time difference between subsequent check-ins

Notes: (a) NYC-distribution; (b) Tokyo-distribution; (c) L.A.-distribution; (d) NYC-time; (e) Tokyo-time; (f) L.A.-time

2013; Wang *et al.*, 2015) using topic model. Wang *et al.* (2015) studied the familiarity-aware POI recommendation. Therefore, using topic model for the analysis is realistic. Second, in topic model, topics are represented by words, that is, a set of tags and categories, which is easily interpretable by human approach compared to other popular approaches such as matrix factorization (Cheng *et al.*, 2012).

5.4.1 Objective. In this subsection, we try to find the answers to the following two questions to clarify that the visit-pattern variations affected by familiarity differences have some effect on better personalized POI recommendations. In the questions, the term "target user" indicates a user who requested the recommendation.

SQ1. To complement a target user's visit pattern in an area, can we adopt the other users' visit patterns under the condition where the target user and the other users have the same familiarity degree in the area?

SQ2. To complement a target user's visit pattern in an area, can we adopt the target user's whole visit patterns in all the areas regardless of his/her familiarity degrees?

In less-familiar areas, only a small number of check-ins of the target user are available to recommend POIs. Thus, both *SQ1* and *SQ2* are important questions to complement the target user's insufficient visit-pattern information. Wang *et al.*'s (2015) study proved that the answer to *SQ1* is "yes" in cities far from home city, that is, in strongly unfamiliar areas. However, we do not know whether the answer is still valid or not to the areas in home city, that is, in the areas consisting of weakly unfamiliar areas and familiar areas.

5.4.2 Approach. Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003), a popular implementation of topic model, is used. A user is represented by a "bag of tags and categories" each of which is related to the venues they visited. Please remember 3.2.1. Check-in for tags and keywords. For simplicity, we call both tags and categories as words. In LDA, "Corpus" represents a set of all users U, and "Vocabulary" indicates the set of words in the corpus. LDA classifies words in vocabulary into pre-defined topics (i.e. latent topic) by learning relation-strength between each word-topic pair. As LDA learns word-topic-relation-strength in the form of probability (Resnik and Hardisty, 2010) and users are represented by the bag of words, we can infer the probability distribution over topics for any given user in the Corpus by accumulating word-topic-relation-strength of each word related to the venues that the given user visited. As venues are also related to words, we can calculate venue-topic-relation-strength by accumulating word-topic-relation-strength of each word related to each of the venues. In this subsection, we denote the probability distribution over topics as preference distribution or popularity distribution for term simplicity. We used 20 topics as the number of pre-defined topics to balance topic's resolution and easy human interpretability.

To answer *SQ1*, we extracted three grouped users whose familiarity to area ($f_{u,a}$) is in the ranges of [0.0, 0.2), [0.4, 0.6) and [0.8, 1.0] followed by examining the difference of popularity distribution among the three groups. We examine the difference among the three groups, area by area. If we can see clear difference among the three groups, we can say that the answer to *SQ1* is "yes".

A set of popularity distributions in an area is calculated by using whole check-ins left by users in the three groups in the area. As we know venue-topic-relation-strength, we accumulate venue-topic-relation-strength of the venue appearing in each of the

check-ins to make accumulated probability distribution over topics, which we call "popularity distribution".

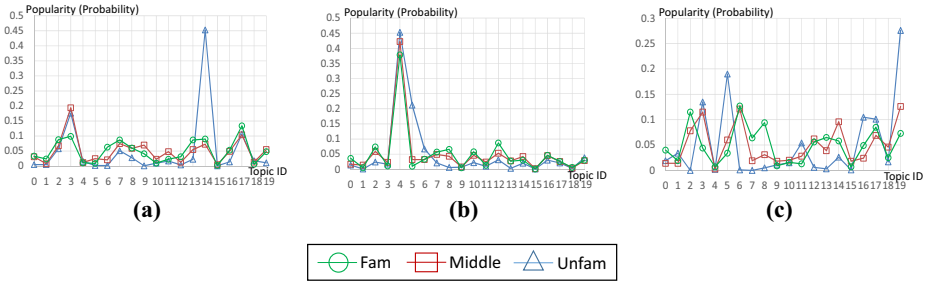
To answer *SQ2*, we extracted five grouped areas whose familiarity of the target user is in the ranges of [0.0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6,0.8) and [0.8, 1.0]. Here, the target user's check-ins in each of the five grouped areas show his/her preference distributions over his/her familiarity. To calculate the preference distribution, we adopt similar ways to the popularity distribution calculation. Single user's check-ins are classified into one of the five groups according to the user's familiarity degree to the areas where the check-ins were made. Then, the accumulated probability distribution over topics is calculated for each of the five check-in groups of single user. If we can detect some correlation between the preference distributions in single user, we can say that the answer to *SQ2* is "yes".

5.4.2 Results. From the result shown later, we can say "yes" to RQ1. However, we are required to deal with the visit pattern of the public in a more detailed layer than the topic model, for instance in individual POI layer. We can partially say "yes" to RQ2, because we have found weak correlation between the preference distributions in single user.

5.4.3 Familiarity versus Individual areas. In Figure 9, we plotted the popularity distributions of three areas in Tokyo: Fam indicates the users whose familiarities are in [0.8, 1.0], Middle and Unfam indicate the users whose familiarities are in [0.4, 0.6) and in [0.0, 0.2), respectively. We also plotted top-5 related words to each topic in Table V. Figure 9(a) represents Shibuya, a famous fashion/music area for young people, Figure 9(b) represents a famous electronics store and hobby goods area Akihabara and Figure 9(c) represents Roppongi which is famous for IT ventures, museums and popular clubs. Shibuya and Akihabara have small number of sharp peaks because the two cities have clear characteristics. Both Topic ID 3 in Shibuya and Topic ID 4 in Akihabara indicate music venue and electronics store. The topics can match to the human perception of Shibuya and Akihabara. Roppongi has more complex distributions because the area has more complicated mixture of venues from different purpose, which results in having various visitors.

We can also confirm that the results found in previous subsections indicate that unfamiliar users go for Arts & Entertainment-related venues, while food-related venues are more important to familiar users. The results are also valid in individual areas.

The difference between the popularity distributions in the same area can be detected in all the three areas. In addition, the popularity distributions indicate the publics'



Notes: (a) Shibuya; (b) akihabara; (c) roppongi

Figure 9. Popularity distributions over topics in the three famous areas in Tokyo, Japan

Table V.
20 Topics of Tokyo

Topic ID (Interpretation)	0 (Supermarket)	1 (Gym/Fitness)	2 (Slow food 1)	3 (Music Venue)	4 (Electronics store)
Words	Convenience store	Gym	Japanese restaurant	Rock club	Electronics store
	Supermarket	Gym/Fitness Center	Sake bar	Record shop	Camera
	Mall	Tipness	BBQ Joint	Music venue	Photobooth
	Bridge	Pool	Sushi restaurant	Live	Yodobashi
	Grocery store	Flower shop	Italian restaurant	Music	Shopping

Topic ID (Interpretation)	5 (Baseball/Concert)	6 (Coffee shop)	7 (Noodle)	8 (Slow food 2)	7 (Movie theater)
Words	Tokyo	Coffee shop	Ramen/Noodle Ramen	Japanese restaurant	Theater
	Concert hall	Coffee	Tsuke-men	Chinese restaurant	Movie
	stadium	Starbucks	Udon	Indian restaurant	Multiplex
	Baseball stadium	cafe	Bath house	Diner	Cinema
	Field	Shop		Italian restaurant	Movie theater

Topic ID (Interpretation)	10 (Hobby)	11 (Shopping mall)	12 (Historic Site)	13 (Hard drink)	14 (Apple store)
Words	Arcade	Mall	Shrine	Bar bar	Electronics store
	Hobby shop	Department store	edo	beer	Department store
	photobooth	Theater	Bridge	Pub	Mac
	Hot spring	Museum	Shogun	Sake bar	iphone
	Animate	Halls	Dokugawa	Wine bar	ipod

Topic ID (Interpretation)	15 (Theme park)	16 (Fast food 1)	17 (Fast food 2)	18 (Outdoor Activity)	19 (Books/Museum)
Words	Theme park	Café	Café	Park	Bookstore
	Tokyo	Fast food restaurant	Bakery	Parks	Gallery
	Disney	Coffee Shop	Italian restaurant	Outdoors	Art gallery
	Resort	Restaurant	French restaurant	Pond	Art
	Disney sea	Diner	Dessert shop	Jogging	Art museum

common visit pattern to the specific areas. Therefore, as Lian *et al.* (2014) and Wang *et al.* (2015) proposed, it is reasonable to pay attention to the visit pattern of the users who visited the same area. However, as Table VI shows, the popularity-distribution difference between familiar users and middle users is relatively small compared to other pairs, one of which is unfamiliar user group. The two popularity distributions are strongly positively correlated. This means that distinguishing familiarity-oriented visit-pattern difference at the topic level is difficult if the familiarity difference is not huge.

Therefore, it is better to capture the difference in more specific layers than in topic layer, such as individual POIs. Figure 10 shows the difference of top 10 popular venues among familiar users and among moderately familiar users related to the specific topics. We can see that popular venues are significantly different between the familiar and moderately familiar users, even if the two user groups' popularity difference over the topic is small. Therefore, modeling the visit-pattern variation in more detailed layer is required to capture the familiarity's influence more effectively.

5.4.4 Familiarity versus individual user preference. Figure 11 shows an example of four users' preference distributions with three different familiarity degrees; Fam, Middle and Unfam. As shown in Figure 11, we can see that the preference distributions of the three different familiarity degrees are not similar to each other, even if these three are made by the same user.

Table VII shows the average Pearson's coefficient between the two preference distributions of the same users who left their check-ins in NYC, Tokyo or LA. Table VII indicates that the user preference distributions of the same user have weak positive correlation between each other, regardless of the familiarity degree. This supports our first impression from Figure 10. However, this also implies that users have some

Pair	Shibuya	Akihabara	Roppongi
(Unfam, Middle)	0.49	0.90	0.50
(Middle, Fam)	**0.75**	**0.98**	**0.62**
(Fam, Unfam)	0.52	0.85	0.03

Table VI. Pearson's coefficients between different user groups

(a) **(b)** **(c)**

Notes: Red = popular amongst "Fam" users; Green = popular amongst "Middle" users; Yellow = popular amongst Both; (a) shibuya: hard drink; (b) akihabara: electronics Store; (c) roppongi: coffee shop

Figure 10. Top 10 popular venues in familiar users and moderately familiar users related to a specific topic

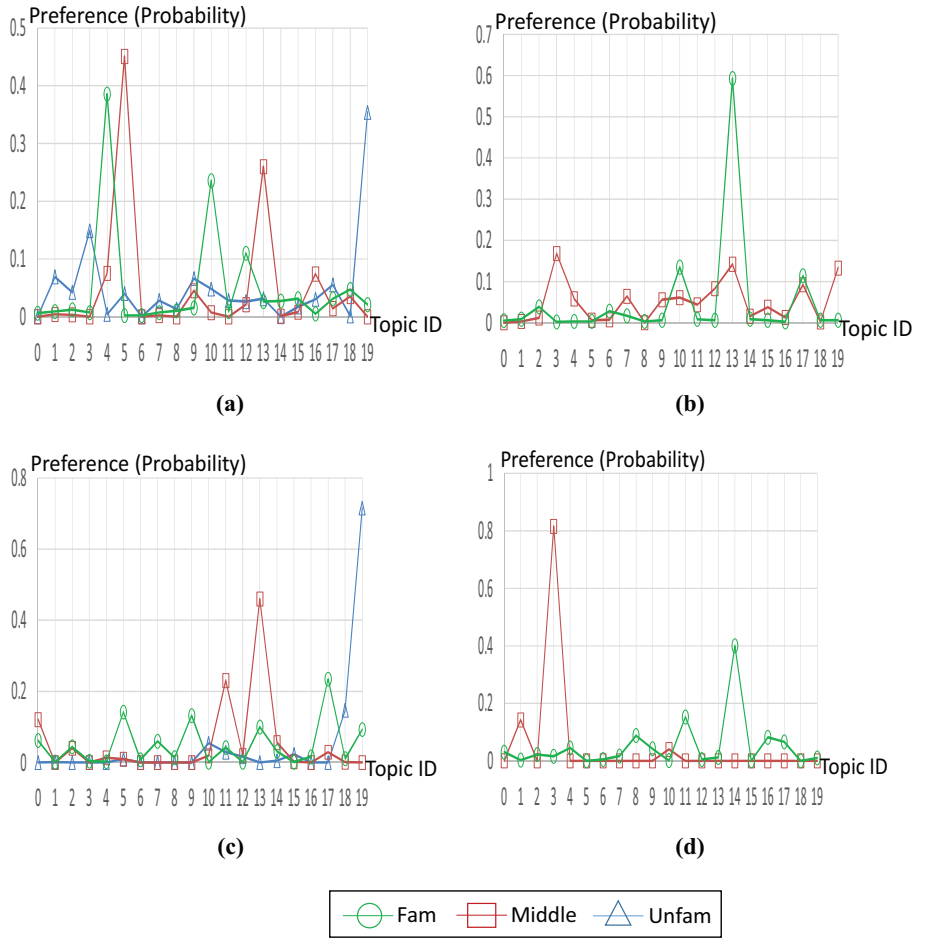


Figure 11.
Preference distributions over topics of NYC and Tokyo individuals

Notes: (a) User 1 in NYC (53 check-ins in total); (b) User 2 in NYC (143 check-ins in total); (c) User 3 in Tokyo (36 check-ins in total); (d) User 4 in Tokyo (134 check-ins in total)

consistent preference distribution which is not varied with regard to the familiarity degree. As there is a weak correlation, to some extent, we can infer the user preference distribution observed in a given familiarity degree by using the same users' preference distributions observed in other familiarity degrees.

The coefficient differences between "subsequent familiarity degrees (No 1-4)" and "the given familiarity degree and the all other familiarity degrees (No.5-9)" are not significant. Therefore, we think that it is better to infer the preference with a given familiarity degree by using all available check-ins left by users, because the approach can mitigate data sparseness problem while maintaining similar error scale to the approach that uses the preference distribution of subsequent familiarity degrees only.

6. Discussion

In this section, we briefly discuss possible ways of integrating familiarity concept with recommendation systems.

From our analysis, it is obvious that user preferences over categories (or venues) vary with regard to the familiarity degrees. Therefore, the familiarity concept can be integrated with existing recommendation algorithms. One possible way of integration is to put the familiarity-related weights on user preferences. We are able to infer the common preference tendency of users over categories (or venues) for a given familiarity degree from the users' check-ins with the given familiarity degree. When a user requests recommendation in an area with a given familiarity degree, recommendation systems are able to provide recommendations to the user according to the mixed preference, that is, the mixture of the user's preference and the public's preference tendency with the given familiarity degree. We expect that we can learn the users' preferences, the public's preference tendencies for each familiarity degree and the blending factors, simultaneously, by adopting machine-learning algorithms such as multi-faceted topic models (Yin *et al.*, 2013; Eisenstein *et al.*, 2011; Mei *et al.*, 2007) or context-aware matrix factorization (Baltrunas *et al.*, 2012).

The tendency that users visit less-popular venues in familiar areas compared to those in unfamiliar areas is an interesting issue. Such visit tendency forces familiarity-aware recommendation systems to choose less-popular venues from the candidate venues in the same category. However, we have a problem, that is, the systems have many recommendation candidates when the systems turn their eyes to less-popular venues, because the popularity distribution follows long-tail distribution. The facts prevent most of the existing recommendation algorithms from achieving accurate recommendation because the algorithms depend on popularity concept, somehow, to filter out irrelevant candidates. Therefore, we need an additional criterion to address the problem. Fortunately, the problem mostly occurs in familiar areas where many of users' check-ins are available. We might borrow the idea of combining "topic model" and "matrix factorization" (Liu *et al.*, 2013a). We adopt "topic model" primarily to guess users' higher level (category level) preferences in unfamiliar areas. In contrast, matrix factorization which focuses more on lower level (individual venue level) preferences gets more weight in familiar areas.

We believe that time and subsequent visit-pattern characteristics with regard to the familiarity can be used as features that can be integrated with related algorithms to enrich the time-aware recommendations (Gao *et al.*, 2013; Yuan *et al.*, 2013) or next-place recommendations (Cheng *et al.*, 2013; Liu *et al.*, 2013b).

Table VII. Pearson's coefficients between different familiarity ranges of the same user

No.	Pair ("−": set subtraction)	NYC	Tokyo	LA
1	[0.0, 0.2), [0.2, 0.4)	0.11	0.13	0.12
2	[0.2, 0.4), [0.4, 0.6)	0.12	0.14	0.15
3	[0.4, 0.6), [0.6, 0.8)	0.12	0.16	0.17
4	[0.6, 0.8), [0.8, 1.0]	0.28	0.25	0.24
5	([0.0,1.0] – [0.0, 0.2)), [0.0, 0.2)	0.08	0.14	0.11
6	([0.0,1.0] – [0.2, 0.4)), [0.2, 0.4)	0.09	0.16	0.15
7	([0.0,1.0] – [0.4, 0.6)), [0.4, 0.6)	0.12	0.19	0.18
8	([0.0,1.0] – [0.6, 0.8)), [0.6, 0.8)	0.27	0.26	0.24
9	([0.0,1.0] – [0.8, 1.0]), [0.8, 1.0]	0.25	0.24	0.23

7. Conclusion

In this paper, we investigated how a user's movement pattern varies over areas with different familiarity by analyzing over 1.4-million check-in data. People tend to go to unfamiliar areas on weekends to enjoy cultural experiences (visit movie theaters, museums, stadiums, etc.), while they put more weight on shopping in familiar areas. Food-related venues are always important despite their relative-importance changes with regard to the familiarity.

User's preference and venue's popularity are two important factors that are influenced by the degree of familiarity with an area. However, the preference-varying pattern is different according to venue's category, while the popularity-varying pattern is relatively constant over the categories. The variation of the preference is closely related to users' perception about the category. For instance, different from gorgeous restaurants, users think that they should visit the fast-food venues when they have to save time. In this case, they think that their preference is a negotiable factor.

The results have some limitations. Our results do not represent all people in the cities. As we gathered the LBSN data via Twitter, the user demographic is biased by the user distribution of Twitter. Despite the limitations, we can detect familiarity-related visit-pattern changes even in the same city (relatively short range). We believe that familiarity is a useful concept for both mobility analysis and POI recommendations.

Our future work includes the inference of the unique geographic factors that vary with regard to the familiarity by using data from diverse LBSNs. Further, it should include integration of the familiarity-related mobility-pattern variations with the POI recommendation.

Notes

1. https://foursquare.com
2. https://twitter.com
3. https://dev.twitter.com/rest/public
4. https://developer.foursquare.com/categorytree

References

Baltrunas, L., Ludwig, B., Peer, S. and Ricci, F. (2012), "Context relevance assessment and exploitation in mobile recommender systems", *Personal and Ubiquitous Computing*, Vol. 16 No. 5, pp. 507-526.

Bao, J., Zheng, Y. and Mokbel, M.F. (2012), "Location-based and preference-aware recommendation using sparse geo-social networking data", *Proceeding of the 20th Int'l Conference on Advances in Geographic Information Systems, ACM, New York, NY*, pp. 199-208.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Cheng, C., Yang, H., King, I. and Lyu, M.R. (2012), "Fused matrix factorization with geographical and social influence in location-based social networks", *Proceeding of 26th Int'l Conference on Artificial Intelligence*.

Cheng, C., Yang, H., Lyu, M.R. and King, I. (2013), "Where you like to go next: successive point-of-interest recommendation", *Proceeding of the 23th Int'l joint Conference on Artificial Intelligence, AAAI Press*, pp. 2605-2611.

Cheng, Z., Caverlee, J., Lee, K. and Sui, D.Z. (2011), "Exploring millions of footprints in location sharing services", *Proceeding of the 5th Int'l Conference on Weblogs and Social Media, AAAI Press*.

Cho, E., Myers, S.A. and Leskovec, J. (2011), "Friendship and mobility: user movement in location-based social networks", *Proceeding of the 17th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*, pp. 1082-1090.

Cranshaw, J., Schwartz, R., Hong, J. and Sadeh, N. (2012), "The livehoods project: utilizing social media to understand the dynamics of a city", *Proceeding of the 6th Int'l Conference on Weblogs and Social Media, AAAI Press*.

Eisenstein, J., Ahmed, A. and Xing, E.P. (2011), "Sparse additive generative models of text", *Proceeding of the 28th Int'l Conference on Machine Learning*.

Gao, H., Tang, J., Hu, X. and Liu, H. (2013), "Exploring temporal effects for location recommendation on location-based social networks", *Proceeding of the 7th ACM Conference on Recommender Systems, ACM, New York, NY*, pp. 93-100.

Gao, H., Tang, J., Hu, X. and Liu, H. (2015b), "Content-aware point of interest recommendation on location-based social networks", *Proceeding of the 29th AAAI Conference on Artificial Intelligence*.

Gao, H., Tang, J. and Liu, H. (2015a), "Addressing the cold-start problem in location recommendation using geo-social correlations", *Data Mining and Knowledge Discovery*, Vol. 29 No. 2, pp. 299-323.

Georgiev, P., Noulas, A. and Mascolo, C. (2014), "Where business thrive: predicting the impact of the olympic games on local retailers through location-based service data", *Proceeding of the 8th Int'l Conference on Weblogs and Social Media, AAAI Press*.

Hu, L., Sun, A. and Liu, Y. (2014), "Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction", *Proceeding of the 37th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 345-354.

Joseph, K., Tan, C.H. and Carley, K.M. (2012), "Beyond 'local', 'categories', and 'friends': clustering foursquare users with latent 'topics'", *Proceeding of the 2012 ACM Conference on Ubiquitous Computing, ACM, New York, NY*, pp. 919-926.

Kurashima, T., Iwata, T., Hoshide, T., Takaya, N. and Fujimura, K. (2013), "Geo topic model: joint modeling of user's activity area and interests for location recommendation", *Proceeding of the 6th ACM Int'l Conference on Web Search and Data Mining, ACM, New York, NY*, pp. 375-384.

Lian, D., Zhao, C., Xie, X., Sun, G., Chen, E. and Rui, Y. (2014), "GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation", *Proceeding of the 20th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York*, pp. 831-840.

Liu, B. and Xiong, H. (2013), "Point-of-interest recommendation in location based social networks with topic and location awareness", *Proceeding of the 2013 SIAM Int'l Conference on Data Mining*.

Liu, B., Fu, Y., Yao, Z. and Xiong, H. (2013a), "Learning geographical preferences for point-of-interest recommendation", *Proceeding of the 19th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*, pp. 1043-1051.

Liu, X., Liu, Y., Aberer, K. and Miao, C. (2013b), "Personalized point-of-interest recommendation by mining user's preference transition", *Proceedingof the 22nd ACM Int'l Conference on Information & Knowledge Management, ACM, New York, NY*, pp. 733-738.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C.X. (2007), "Topic sentiment mixture: modeling facets and opinions in weblogs", *Proceeding of the 16th Int'l Conference on World Wide Web, ACM, New York, NY*, pp. 171-180.

Pearson, K. (1895), "Notes on regression and inheritance in the case of two parents", *Proceeding of the Royal Society of London*, Vol. 58, pp. 240-242.

Qu, Y. and Zhang, J. (2013), "Trade area analysis using user generated mobile location data", *Proceeding of the 22nd Int'l Conference on World Wide Web, 2013*, *Republic and Canton of Geneva*, *Switzerland*, pp. 1053-1964.

Resnik, P. and Hardisty, E. (2010), "Gibbs sampling for the uninitiated", Technical Report, UMIACS, available at: www.umiacs.umd.edu/~resnik/pubs/LAMP-TR-153.pdf (accessed 22 January 2016).

Sander, J., Ester, M., Kriegel, H.P. and Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proceeding of the 2nd ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*.

Silva, T.H., Melo, P.O.S.V., Almeida, J., Musolesi, M. and Loureiro, A. (2014), "You are what you eat (and Drink): identifying cultural boundaries by analyzing food & drink habits in foursquare", *Proceeding of the 6th Int'l Conference on Weblogs and Social Media*, *AAAI Press*.

Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S. and Zhou, X. (2015), "Geo-SAGE: a geographical sparse additive generative model for spatial item recommendation", *Proceeding of the 21th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*, pp. 1255-1264.

Yang, D., Zhang, D., Yu, Z. and Wang, Z. (2013), "A sentiment-enhanced personalized location recommendation system", *Proceeding of 24th ACM Conference on Hypertext and Social Media, ACM, New York, NY*, pp. 119-128.

Ye, M., Yin, P., Lee, W. and Lee, D. (2011), "Exploiting geographical influence for collaborative point-of-interest recommendation", *Proceeding of the 34th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 325-334.

Yin, H., Sun, Y., Cui, B., Hu, Z. and Chen, L. (2013), "LCARS: a location-content-aware recommender system", *Proceeding of the 19th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, ACM, New York, NY*, pp. 221-229.

Yuan, Q., Cong, G., Ma, Z., Sun, A. and Magnenat-Thalmann, N. (2013), "Time-aware point-of-interest recommendation", *Proceeding of the 36th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 363-372.

Zhang, J. and Chow, C. (2013), "iGSLR: personalized geo-social location recommendation – a kernel density estimation approach", *Proceeding of the 21st ACM SIGSPATIAL Int'l Conference on Advances in Geographic Information Systems, ACM, New York, NY*, pp. 334-343.

Zhang, J.D., Chow, C.Y. and Zheng, Y. (2015), "ORec: an opinion-based point-of-interest recommendation framework", *Proceeding of the 24nd ACM Int'l Conference on Information & Knowledge Management, ACM, New York, NY*, pp. 1641-1650.

Further reading

Zhang, J.D. and Chow, C.Y. (2015), "GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations", *Proceeding of the 38th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 443-452.

Corresponding author

Jungkyu Han can be contacted at: han.jungkyu@akane.waseda.jp