# Geo Topic Model: Joint Modeling of User's Activity Area and Interests for Location Recommendation

Takeshi Kurashima*, Tomoharu Iwata+, Takahide Hoshide*, Noriko Takaya* and Ko Fujimura*
*NTT Service Evolution Labs., NTT Corporation
+NTT Communication Science Labs., NTT Corporation
*1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
+2-4 Hikaridai, Seika-Cho, Soraku-gun, Kyoto, 619-0237 Japan
{kurashima.takeshi,iwata.tomoharu,hoshide.takahide,takaya.noriko,fujimura.ko}@lab.ntt.co.jp

## ABSTRACT

This paper proposes a method that analyzes the location log data of multiple users to recommend locations to be visited. The method uses our new topic model, called *Geo Topic Model*, that can jointly estimate both the user's interests and activity area hosting the user's home, office and other personal places. By explicitly modeling geographical features of locations and users, the user's interests in other features of locations, which we call latent topics, can be inferred effectively. The topic interests estimated by our model 1) lead to high accuracy in predicting visit behavior as driven by personal interests, 2) make possible the generation of recommendations when the user is in an unfamiliar area (e.g. sightseeing), and 3) enable the recommender system to suggest an interpretable representation of the user profile that can be customized by the user. Experiments are conducted using real location logs of landmark and restaurant visits to evaluate the recommendation performance of the proposed method in terms of the accuracy of predicting visit selections. We also show that our model can estimate latent features of locations such as art, nature and atmosphere as latent topics, and describe each user's preference based on them.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

## General Terms

Algorithms

## Keywords

Topic model, Location recommendation

## 1. INTRODUCTION

In recent years, social media has become ubiquitous and important for social networking and user-generated content sharing. In the tourism and food domains, more and more people are sharing their experiences on photo sharing sites (e.g. Flickr [10] and Google Picasa [14]), social networking services (e.g. foursquare [11] and Facebook [9]), travel communities (e.g. TripAdvisor [27]), and restaurant review sites (e.g. Tabelog [26], yelp [29] and ZAGAT.com [31]). Mining and modeling the visit behavior contained in this tremendous volume of location log data is an important task, because it will provide people with efficient ways to develop the most satisfying plan of action.

In this paper, we propose a method for recommending new locations that are potentially interesting to users based on a new topic model, called *Geo Topic Model*. We define a location as a uniquely identified specific site, including landmark and restaurant. Our concept is based on understanding the processes that underlie a user's decision about where and what to visit in the future; the location that the user will visit depends on 1) user's daily-life range and 2) user's topic interest.

We have two criteria for designing a method to recommend locations. First, recommending locations that match the user's interests is important. For example, if the user is interested in the arts, we would expect the recommender system to emphasize locations associated with art such as art museums and picture galleries. Our approach is motivated by the success of using topic models to analyze human behavior log data [7, 19, 18]. A topic model is a hierarchical Bayesian model, in which a user (document) is modeled as a mixture of topics, and a topic is modeled as a probability distribution over locations (words). However, when we use existing topic models to analyze visit behavior based on location log data, we are unable to discover the users' interests in the features (latent topics) of locations such as "art" and "nature". In the tourism and food domains, the user's choice of items is largely influenced by her/his geographical coordinates, and locations in the user's immediate neighborhood are likely to be chosen. The estimated topic, therefore, groups nearby locations, and most of the topics are used to describe the user's spatial area of activity. Thus existing topic models tend to recommend easily accessible but uninteresting locations to the user.

Second, augmenting location recommendations with feedback is important. When a user is not satisfied with the first suggestions, she/he would like to review the estimated user profile in order to understand why and how the recommendations were made, and customize it to receive better recommendations. Providing an interpretable representation of the user profile is very useful for understanding the original user information and reformulating the profile, resulting in improved recommendation performance. Existing collaborative filtering methods can make use of the travel histories of a group of similar users (i.e., user-based collaborative filtering) or a set of similar locations (i.e., location-based collaborative filtering) to generate location recommendations. The main


*WSDM'13,* February 4–8, 2013, Rome, Italy.

disadvantage of the method is that the learnt latent space is not easy to interpret. The user of a recommender system based on existing collaborative filtering methods can not easily understand why and how the recommendations were made.

With these criteria in mind, we develop a machine learning algorithm for recommending geographical locations. Our key idea is to model the user's interests separately from location accessibility. By modeling user's spatial area of activity based on geographical features of visited locations (i.e. latitude and longitude coordinates), the user's interests in other latent features (latent topics) of locations can be inferred effectively. Furthermore, we use annotation data of each location to construct a meaningful representation of latent features (latent topics). Thus our proposed model can generate an interpretable representation of each user profile; it can be presented alongside location recommendations and customized by the user if necessary.

Our proposed model can yield impressive results when recommending interesting locations in two scenarios: 1) recommending locations in the vicinity of the user's daily life areas and 2) recommending locations when the user is in less familiar areas. Recommendations for the user's daily life generated by our model are easy to access from the user's home, office, and other personal places, and suit the user's interests. The latent topics estimated from our model lead to high accuracy in predicting the visit behavior as driven by personal interests. We also introduce a recommendation method customized for a user in the mobile environment who wants to find locations to suit her immediate surroundings; the recommended locations are easy to access from the user's current position, and match the user's interests. Even when the user visits a new place (e.g. sightseeing and business trips), this method can generate recommendations from the user's topic interests learned from the characteristics found in the user's location log.

We demonstrate the effectiveness of the proposed model using real-life location logs of landmark and restaurant visits. We show the utility of our model by generating recommendations in two areas: daily life and those yet to be visited. We also analyze the content of estimated latent topics using annotation data of locations and find that our model can lucidly describe each user's interests in terms of latent features of locations such as art, nature, construction, popularity and atmosphere.

In summary, the major contributions of this paper include:

- Geo Topic Model: learns the user's activity area and personal interests in the latent features of locations simultaneously;
- A recommendation method that presents easy-to-access and interesting locations for 1) a user in daily life and 2) a traveler in an unfamiliar territory;
- A method of discovering latent features (latent topics) of landmarks/restaurants from the actual logs of visited landmarks/restaurants and an interpretable description of the user's interest based on latent features (latent topics).

This paper is organized as follows: The next section describes related works. Section 3 defines our research problem and presents our proposed model and recommendation methods. Section 4 uses actual location logs to confirm the effectiveness of the proposed model, and Section 5 concludes this paper.

# 2. RELATED WORK

## 2.1 Behavior Modeling for Recommendation

Collaborative filtering, the traditional approach to recommendation, makes use of the location logs of a group of similar users (i.e., user-based collaborative filtering) or a set of similar locations (i.e., location-based collaborative filtering) to generate location recommendations. Most successful collaborative filtering methods are based on latent factor models; one technique, matrix factorization, offers good performance [25]. Matrix factorization allows us to represent users and items in a shared low-dimensional space, and predict whether the user will like an item from the inner product of their latent representations. Among latent factor models, matrix factorization performs well for behavior prediction, but its main disadvantage is that the learnt latent low-dimensional space is not easy to interpret.

Analyzing human behaviors using probabilistic topic models yields a more interpretable representation of users and locations than latent factor models. Topic modeling algorithms are used to discover a set of latent features of locations (topics) from the location histories of multiple users (a collection of documents), where a latent feature (topic) is a probabilistic distribution over locations (words). Topic models also provide an interpretable low-dimensional representation of the user's interests, where interests are modeled as a mixture of location features (topics). We can discover the latent feature of locations based on what types of locations are contained, and understand the user's preference profile based on the latent feature of the locations that he or she likes. We adopt a topic modeling approach to realize our goal of obtaining knowledge that offers a deeper understanding of the decision-making process (i.e. answer the question, "how do we decide where to visit next?").

Probabilistic Latent Semantic Analysis (PLSA) [16, 17] and Latent Dirichlet Allocation (LDA) [4] are two representative topic models; they have been successfully used in a wide variety of applications such as information retrieval, language modeling, and modeling the interests underlying human behavior. For example, Das et al. use a linear model to combine recommendations from different algorithms including PLSA for generating personalized recommendations from Google News [7]. We proposed the Topic Tracking Model for analyzing purchase behavior; it captures interests and temporal changes in item trends [18]. The discovery of patterns in the navigational behavior of Web users based on PLSA was tackled by [19].

In the tourism domain, we adopted the topic model for estimating user's interests based on geo-tag based histories on Flickr [20]. The method proposed in [20] has, unfortunately, difficulty in recommending locations that suit the user's preference because most of the latent topics describe the user's immediate area of activity (i.e. the estimated topic of existing topic models groups nearby locations). The topic-based model proposed in [30] also infers an area (a set of nearby locations) that could have similar functions (e.g. residential, commercial and entertainment area in a city). This is one of the major problems when analyzing location histories, and the challenge tackled by this study is to estimate the user's interests by considering the extent of the user's geographical range.

## 2.2 GPS Trajectory Mining

Data mining from GPS trajectories gathered by mobile devices is somewhat related to our work because its main goal is to predict where a person may be going. GPS point data is assumed to contain only the information of location (latitude and longitude) and timestamp; a GPS trajectory is a time series of points. Ashbrook et al. apply a Markov model to GPS data in an attempt to model traveler behavior [2]. Their assumption is that the location that the tourist will move to next depends on the current and recently visited locations. Zheng et al. applied a graph mining method to a GPS dataset generated by 107 users in order to extract major sightseeing spots and classical travel routes between them [34, 32]. Several papers have mined travel routes and itineraries linking popular landmarks from the GPS trajectories of multiple users [20, 1, 5]. Sequential pattern mining from GPS trajectories was also performed by Monreale et al. [23]. Zheng et al. proposed hierarchical-graph-based measurement for estimating the similarity between multiple users' movements [33]. Leung et al. proposed CLM graph extraction from GPS trajectory data with a User-Activity-Location tripartite data structure [21]. Their model assumes that the temporal sequence of visited locations represents the user's activity. This is because most users visit several locations in order to perform a specific activity (e.g. go to office and its nearest station for work). The methods that depend on the sequential patterns found in GPS trajectories are designed for capturing the current temporal-spatial properties in user activities. Different from these methods, we want to predict the next location assuming that the user's actions are being driven by her/his personal interests (e.g. vacation spot, place to eat on the weekend).

## 2.3 Location-Based Recommender System

There are some related studies on mobile city guide applications that can personalize the user's recommendations. COMPASS is a tourist guide system that makes recommendations after considering user profile, schedule, shopping list, time since last visit and context such as weather and traffic conditions [28]. MobyRec is a context-aware mobile recommender system that allows preferences to be specified for restaurants etc. [24]. However, these systems require the user's stated query or initially specified interests. Ours generates recommendations using only location log data; the user does not have to explicitly enter her/his profile, interests or queries.

Based on extensive field work in Tokyo, Bellotti et al. designed and developed a location recommender system that infers user activity (e.g. "eating" and "shopping") and preference [3]. Their work is similar to our work in that they attempt to generate recommendations based on the user's implicit (inferred) interests. Their system, however, combines the results from a variety of existing recommendation models in an ad hoc fashion depending on the user's context or pre-defined rules. No new recommendation model is proposed in [3]. Again, this paper proposes a new probabilistic model specialized for recommending locations.

# 3. PROPOSED MODEL

In this section, we first describe the proposed Geo Topic Model, and the recommendation method based on the model for predicting visit behavior in daily life. We then describe how the recommendation method supports a mobile user who visits a new place.

## 3.1 Geo Topic Model

Suppose that we have a set of $N$ users $\boldsymbol{U} = \{u\}_{u=1}^{N}$ and a set of $I$ locations $\boldsymbol{I} = \{i\}_{i=1}^{I}$. Each location is associated with geotag $r_i$, which represents the latitude and longitude coordinates of the location in a shared bi-axial space. Each user has a visited-location log denoted by $\boldsymbol{x}_u = \{x_{u1}, ..., x_{uM_u}\}$, where $x_{um}$ represents the $m$th visited location of user $u$, and $x_{um} \in \boldsymbol{I}$. Our research problem is simple, and is described as follows; given a user's location log $\boldsymbol{x}_u$, the task is to recommend a set of locations that the user will want to visit next. The notations used are summarized in Table 1.

We make the following two assumptions about visit behavior in daily life:

1. The location that the user visits next largely depends on the user's activity areas, i.e. areas that encompass her home, office, and other personal places. For example, a user whose office is located in Manhattan is much more likely to visit Grand Central Station than Newark Pennsylvania Station because Grand Central Station is much closer.
2. The choice of what to do next is also determined by the user's interests. For example, a user who is interested in art is likely to visit the Metropolitan Museum of Art while a user who is interested in sports is likely to visit Yankee Stadium.

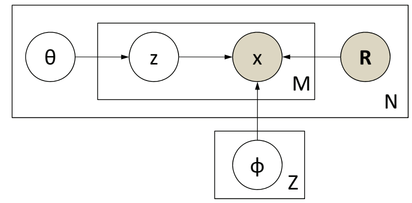



**Figure 1: Graphic model representation of Geo Topic Model. Shaded and hollow nodes indicate observed and latent variables, respectively.**

**Table 1: Notation**

| Symbol | Description |
|---|---|
| $\boldsymbol{U}$ | set of users |
| $u$ | user, $u \in \boldsymbol{U}$ |
| $\boldsymbol{I}$ | set of locations |
| $i$ | location, $i \in \boldsymbol{I}$ |
| $z$ | latent topic |
| $M_u$ | number of locations in the history of user $u$ |
| $x_{um}$ | the $m$th visited location of user $u$, $x_{um} \in \boldsymbol{I}$ |
| $\boldsymbol{x}_u$ | location log of user $u$ denoted by $\boldsymbol{x}_u = \{x_{u1}, ..., x_{uM_u}\}$ |
| $r_i$ | geotag of location $i$ represented by latitude and longitude coordinates in bi-axial space |
| $\boldsymbol{R}_u$ | the set of geotags of locations of user $u$ denoted by $\boldsymbol{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$ |
| $\theta_{uz}$ | probability that topic $z$ is chosen by user $u$ |
| $\boldsymbol{\theta}_u$ | topic proportions of user $u$ denoted by $\boldsymbol{\theta}_u = \{\theta_{uz}\}_{z=1}^{Z}$ |
| $\phi_{zi}$ | probability that location $i$ is chosen for topic $z$ |
| $\boldsymbol{\phi}_z$ | location proportions of topic $z$ denoted by $\boldsymbol{\phi}_z = \{\phi_{zi}\}_{i=1}^{I}$ |
| $N$ | number of users |
| $I$ | number of locations |
| $Z$ | number of latent topics |
| $\beta$ | bandwidth parameter for the width of the activity area |

Our model incorporates these assumptions by assuming the following generative process of visit locations. Each user has user-dependent topic proportions $\boldsymbol{\theta}_u = \{\theta_{uz}\}_{z=1}^{Z}$ that represent the interests of the user, where $\sum_z \theta_{uz} = 1$, $\theta_{uz} \geq 0$, and $Z$ is the number of latent topics. For each visit, topic $z$ is chosen according to $\boldsymbol{\theta}_u$, and then location $x$ is generated depending on both topic-specific parameters $\boldsymbol{\phi}_z = \{\phi_{zi}\}_{i=1}^{I}$ and the set of geotags of previously visited locations: $\boldsymbol{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$. Here, $\phi_{zi}$ represents how likely location $i$ is to be visited given topic $z$. The proposed model is illustrated in Figure 1, where shaded and hollow nodes indicate observed and latent variables, respectively.

In our model, the probability that user $u$ with geotag history $\boldsymbol{R}_u$ visits location $i$ is calculated by the following equation:

$$P(i|u, \boldsymbol{R}_u, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{z=1}^{Z} P(z|u, \boldsymbol{\Theta}) P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi}), \qquad (1)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_u\}_{u=1}^N$, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_z\}_{z=1}^Z$, and $P(z|u, \boldsymbol{\Theta}) = \theta_{uz}$. $P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi})$ represents the probability that location $i$ is chosen from topic $z$ after consideration of the user's geotags $\boldsymbol{R}_u$. We assume that a location that is close to the locations visited by the user is more likely to be visited. Probability $P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi})$ can be written as follows:

$$P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi}) = \frac{1}{C} \exp(\phi_{zi}) \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_i - r\|^2), \quad (2)$$

where $\|\cdot\|$ represents the Euclidean norm in the geographical space, $C = \sum_{i' \in \boldsymbol{I}} \exp(\phi_{zi'}) \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_{i'} - r\|^2)$ is the normalization constant, and $\beta$ is a parameter describing the width of each activity area. As bandwidth parameter $\beta$ is decreased, the width of the activity area increases. The first factor, $\exp(\phi_{zi})$, implies the contribution of interest in choosing a location. Parameter $\phi_{zi}$ represents the probability that location $i$ is chosen for topic $z$; existing topic models such as PLSA also have this parameter. If latent topic $z$ represents "art", location $i$ associated with art, such as a museum or a theater, is likely to be chosen as the topic. The second factor $\sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_i - r\|^2)$ implies the contribution of the user's movement potential. Location $i$ that is near the visit locations in the user's visit log $\boldsymbol{x}_u$ is likely to be chosen. If the user often visits restaurants on the city's main street, other restaurants on the street are likely to be chosen. By using the geographical features of visited locations to explicitly model the user's spatial area of activity, we can effectively estimate other latent features (latent topics) of locations and the user's interest in them (i.e. user-dependent topic proportions).

The number of parameters of the proposed model is $N(Z-1) + ZI$, where there are many fewer latent topics than users and locations ($Z \ll N$ and $Z \ll I$). The proposed method has the same order of computational cost as PLSA because the number of parameters in PLSA is $N(Z-1) + Z(I-1)$. Topic models, including ours, can establish a low-dimensional representation of the location of a user (expressed as user $*$ location matrix), so we can represent $N * I$ probabilities with fewer parameters.

## 3.2 Parameter Estimation

We use maximum likelihood to estimate the parameters in the proposed model. The unknown parameters are a set of topic probabilities $\boldsymbol{\Theta}$ and a set of location parameters $\boldsymbol{\Phi}$. All unknown parameters are represented by $\boldsymbol{\Psi} = \{\boldsymbol{\Theta}, \boldsymbol{\Phi}\}$.

Parameter log likelihood is, given a set of user location logs $X = \{\boldsymbol{x}_u\}_{u=1}^N$, derived as follows:

$$L(\boldsymbol{\Psi}|\boldsymbol{X}) = \sum_{u=1}^{N} \sum_{m=1}^{M_u} \log \sum_{z=1}^{Z} \theta_{uz} P(x_{um}|z, \boldsymbol{R}_u, \boldsymbol{\Phi}). \quad (3)$$

The log likelihood can be maximized with the EM algorithm [8]. The conditional expectation of the complete-data log likelihood is represented as follows:

$$\begin{aligned} & Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}) \\ = & \sum_{u=1}^{N} \sum_{m=1}^{M_u} \sum_{z=1}^{Z} P(z|u, m; \hat{\boldsymbol{\Psi}}) \log \theta_{uz} P(x_{um}|z, \boldsymbol{R}_u, \boldsymbol{\Phi}), \end{aligned} \quad (4)$$

where $\hat{\boldsymbol{\Psi}}$ represents the current estimate, and $P(z|u, m; \hat{\boldsymbol{\Psi}})$ represents the topic posterior probability of the $m$th location of user $u$ given the current estimate. In E-step, we use Bayes rule to compute the topic posterior probability:

$$P(z|u, m; \hat{\boldsymbol{\Psi}}) = \frac{P(z|u, \boldsymbol{\Theta}) P(x_{um}|z, \boldsymbol{R}_u, \boldsymbol{\Phi})}{\sum_{z'=1}^{Z} P(z'|u, \boldsymbol{\Theta}) P(x_{um}|z', \boldsymbol{R}_u, \boldsymbol{\Phi})}, \quad (5)$$

where $P(x_{nm}|z, \boldsymbol{R}_u, \boldsymbol{\Phi})$ is calculated by (2). In M-step, the next estimate of topic proportion $\theta_{uz}$ is given by:

$$\hat{\theta}_{uz} = \frac{\sum_{m=1}^{M_u} P(z|u, m; \hat{\boldsymbol{\Psi}})}{\sum_{z'=1}^{Z} \sum_{m=1}^{M_u} P(z'|u, m; \hat{\boldsymbol{\Psi}})}. \quad (6)$$

The next estimates of $\boldsymbol{\phi}_z$ can not be solved in closed form. Therefore, we estimate them by maximizing $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ through the use of a gradient-based numerical optimization method such as the quasi-Newton method [22]. The gradient of $\boldsymbol{\phi}_z$ w.r.t. $\boldsymbol{\phi}_z$ is as follows:

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\phi}_z} = & \sum_{u=1}^{N} \sum_{m=1}^{M_u} P(z|u, m; \hat{\boldsymbol{\Psi}}) \\ & - \sum_{u=1}^{N} \sum_{m=1}^{M_u} P(z|u, m; \hat{\boldsymbol{\Psi}}) P(x_{um}|z, \boldsymbol{R}_u, \boldsymbol{\Phi}). \end{aligned} \quad (7)$$

By iterating the E-step and the M-step until convergence, we obtain a local optimum solution for $\boldsymbol{\Psi}$. The number of topics $Z$ and width parameter $\beta$ can be estimated by using cross validation.

## 3.3 Recommendation to Suit Current Position

In the previous sections, we described Geo Topic Model, a probabilistic model that predicts locations to visit. The recommendations generated from our model have two features; 1) easy to access from the user's home, office, and other familiar locations 2) and match user's interests. This type of information need occurs often in daily life. In this section, we show another recommendation method customized for a user in the mobile environment. The prevalence of location-aware devices enables the user to search for local information using his current position (i.e. GPS point) as an implicit search query. A mobile user wants to search for locations that suit her/his current position when she/he is in an unfamiliar territory. This type of information need occurs in several situations such as sightseeing, traveling, and business trips. The required properties of these recommendations are 1) offer immediate access from the user's current position 2) and match user's interests. The first property is addressed by using the current position as a substitute for the area of activity. The second property is addressed by using parameters $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ in the same way as used for generating daily life recommendations.

Suppose that the user's current position is $c$, as identified by latitude and longitude coordinates. Given the user's visit log $\boldsymbol{x}_u$ and $c$, the task is to recommend a set of personalized locations that the user will want to visit next. The probability that user $u$ at position $c$ visits location $i$ is given by the following equation:

$$P(i|u, c, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{z=1}^{Z} P(z|u, \boldsymbol{\Theta}) P(i|z, c, \boldsymbol{\Phi}), \quad (8)$$

where the probability that location $i$ is visited given topic $z$ and current position $c$ is as follows:

$$P(i|z, c, \boldsymbol{\Phi}) = \frac{\exp(\phi_{zi}) \exp(-\frac{\beta}{2}\|r_i - c\|^2)}{\sum_{i' \in \boldsymbol{I}} \exp(\phi_{zi'}) \exp(-\frac{\beta}{2}\|r_{i'} - c\|^2)}, \quad (9)$$

where we assume that the probability depends on the Euclidean distance from the current position (latitude and longitude coordinates in bi-axial space).

An important characteristic of our model, described in the previous section, is that it can extract the latent features (topics) of locations beyond just their geographical positions. Once user-dependent topic proportions $\boldsymbol{\theta}_u$ (and a set of location parameters $\boldsymbol{\Phi}$) are estimated from the location log formed in her/his own hometown, recommendations based on the user's topic interests can be generated anywhere since spatially-disbursed locations are grouped in each latent topic of our model. Furthermore, our method satisfies a key requirement of mobile recommender systems; it must generate real-time recommendations according to the current position. This is because the probability $P(i|u, c, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ can be calculated with low computational cost by using (8) and (9) without parameter estimation. User-dependent topic proportions $\boldsymbol{\theta}_u$ do not strongly vary with her/his position.

# 4. EXPERIMENTS

This section evaluates the performance of the proposed method by conducting quantitative experiments on a location log of visits to restaurants and landmarks. Two quantitative experiments were conducted based on two scenarios; one is recommending locations to a user who wants to know interesting locations in his daily-life range (Section 4.1), the other is recommending locations to a user who wants to know interesting locations while traveling (Section 4.2). Finally, we analyze the content of estimated latent topics (position and types of locations grouped in each topic of our model) using the annotations attached to each location, and compare them to the output of an existing topic model (Section 4.3).

## 4.1 Experiment 1: Behavior Prediction Accuracy in Daily Life

### 4.1.1 Dataset

Our experiment used restaurant reviews from Tabelog and a Flickr-sourced geotag collection. Tabelog, which provides restaurant information, has successfully collected a large number of customer reviews. According to our research, at least 2,500,000 customer reviews of about 680,000 restaurants in Japan, uploaded by over 260,000 users, are available on Tabelog. A set of reviews of each user can be taken as the location log of restaurants the user visited. We generated four types of location log datasets from Tabelog data; location logs for Tokyo, Osaka, eastern Japan (including Tokyo, Kanagawa and Saitama prefectures) and western Japan (including Osaka, Kyoto and Hyogo prefectures). For each region, we first crawled the restaurant data including geotags (latitude and longitude), and reviews of crawled restaurants were then downloaded. A set of reviews of each user was then translated into a location log. We eliminated users traveling through or to these destinations using the static location information annotated by each user in order to evaluate prediction accuracy for local inhabitants.

We also collected large-scale-sets of geotagged photographs on Flickr using the site's public API. The geotagged photographs of each user can be assumed to be the user's personal location log. We generated three types of location log datasets from Flickr data; location logs for New York City, San Francisco and Los Angeles. For each city, geotagged photos taken within 50km from the city center were crawled. Given a collection of photos with geolocations, we automatically extracted often-photographed landmarks in the city because we want to recommend popular locations. We define a landmark as a uniquely identified specific location within the city; such as a sightseeing spot, a store, a building, a bridge, an outlet, and so on. We used *mean-shift* clustering to extract often-photographed landmarks from a set of geotags on Flickr since the previous work by Crandall et al. showed that the mean-shift procedure was effective for landmark extraction from spatial data [6]. Based on the results of landmark detection, each location in a location log was converted into one of the landmarks found by the mean-shift procedure. This yielded a set of location logs of visit landmarks from the Flickr-sourced geotag collection. Details of the number of users (location logs) and the number of locations (landmarks) to be recommended are shown in Table 2.

**Table 2: Location log data. Users who visited fewer than five locations were omitted from each. Restaurants reviewed by fewer than one hundred, forty, one hundred and fifty users were omitted from the data for Tokyo, Osaka, eastern Japan and western Japan, respectively. Each dataset had about the same number of recommended restaurants. Landmarks visited by fewer than ten users were also omitted from Flickr-sourced data.**

| Region | Logs | Locations | Ave. log length |
|---|---|---|---|
| Tokyo | 6,257 | 910 | 14.872 |
| Osaka | 2,304 | 763 | 15.995 |
| Eastern Japan | 7,326 | 1,067 | 14.871 |
| Western Japan | 3,531 | 956 | 15.245 |
| New York City | 3,363 | 967 | 9.778 |
| San Francisco | 4,433 | 1,030 | 12.231 |
| Los Angeles | 2,848 | 943 | 11.686 |

### 4.1.2 Model Comparison

We evaluate the validity of our model in terms of its accuracy in predicting behavior in daily life. This is because, according to a previous survey, prediction accuracy is by far the most commonly used measure for evaluating the performance of recommender systems [15]. We compared the Geo Topic Model against the following four probabilistic models:

**Multinomial model:** recommends locations based on popularity. Popularity can be calculated by using the multinomial probability distribution over locations $P(i)$.

**Kernel model:** recommends locations based on user's activity area. New locations close to visited locations are likely to be recommended.

$$P_K(i|u) = \frac{\sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_i - r\|^2)}{\sum_{i' \in \boldsymbol{I}} \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_{i'} - r\|^2)}. \quad (10)$$

**Topic model:** recommends locations based on user interest as estimated by the PLSA model.

$$P_T(i|u) = \sum_{z \in \boldsymbol{Z}} P(z|u)P(i|z). \quad (11)$$

**Kernel-Topic model:** recommends locations based on user's activity area and interests. This model is the linear combination of the Kernel model and the topic model (PLSA) as follows:

$$P_{KT}(i|u) = \alpha P_K(i|u) + (1-\alpha)P_T(i|u), \quad (12)$$

where $\alpha$ is a parameter that balances the influence of Kernel and Topic models.

We used data excluding the last visited locations of all users as training data to infer the model, and predicted the last visited locations of all users as test data. Thus, the number of training sets equals the number of location logs (users) in Table 2, the number of test sets also equals the number of location logs (users) in Table 2. Since the model predicts the next location likely to be visited by the user, we employed the 5-best (accuracy) prediction of visited locations as the evaluation metric. In other words, we calculated the percentage of visited locations contained in the set of 5-highest items over all test examples. The results of 5-best (accuracy) prediction of visited location are shown in Figure 2. In

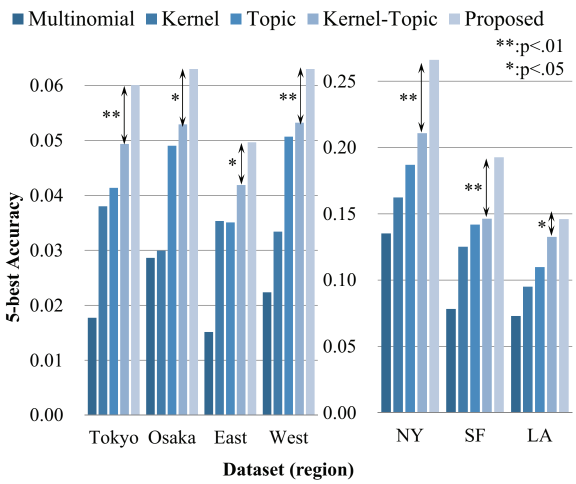



**Figure 2: Comparisons of 5-best (accuracy) predictions of daily-life behavior. X-axis plots the datasets.**

**Table 3: Best parameter values as identified in the first experiment. $\beta$ was chosen from $10$, $10^2$, $10^3$, $10^4$, $10^5$, $10^6$, $10^7$ and $10^8$. $Z$ was set in the range of $3$ to $50$.**

| Model | Region | $\beta$ | $Z$ | $\alpha$ |
|---|---|---|---|---|
| Kernel-Topic | Tokyo | $10^5$ | 22 | 0.2 |
| | Osaka | $10^6$ | 13 | 0.1 |
| | Eastern Japan | $10^5$ | 15 | 0.3 |
| | Western Japan | $10^6$ | 19 | 0.1 |
| | New York City | $10^2$ | 5 | 0.9 |
| | San Francisco | $10^3$ | 20 | 0.5 |
| | Los Angeles | $10^2$ | 16 | 0.7 |
| Proposed | Tokyo | $10^4$ | 8 | - |
| | Osaka | $10^4$ | 7 | - |
| | Eastern Japan | $10^3$ | 9 | - |
| | Western Japan | $10^3$ | 10 | - |
| | New York City | $10^3$ | 9 | - |
| | San Francisco | $10^2$ | 7 | - |
| | Los Angeles | $10^3$ | 6 | - |

this figure, the X-axis plots datasets and the Y-axis plots the 5-best (accuracy) prediction. We used the sign test to determine the statistical significance between the proposed model and the baselines. To sum up, for both data sets, the proposed model yields better 5-best (accuracy) prediction than Multinomial, Kernel, Topic model or Kernel-Topic model, and the differences are significant (two-sided sign test: $p < 0.05$). This shows that the proposed method can appropriately predict the user's visit behavior in regular/daily life because it recommends locations by considering both the user's activity area and interests. Furthermore, unlike the Kernel-topic model, our model does not need not to estimate the value of $\alpha$ because it can automatically learn the influence of location and interest in parameter estimation (i.e. automatically decide which to emphasize).

Note that best parameter values of all models, except Multinomial model, were determined using 5-fold cross validation; they are summarized in Table 3. Parameters that influence the performance of our method are the number of topics $Z$ and bandwidth parameter $\beta$. In our model, as bandwidth parameter $\beta$ is decreased, the width of the activity area increases. For example, restaurants in eastern Japan (western Japan) are distributed more widely than

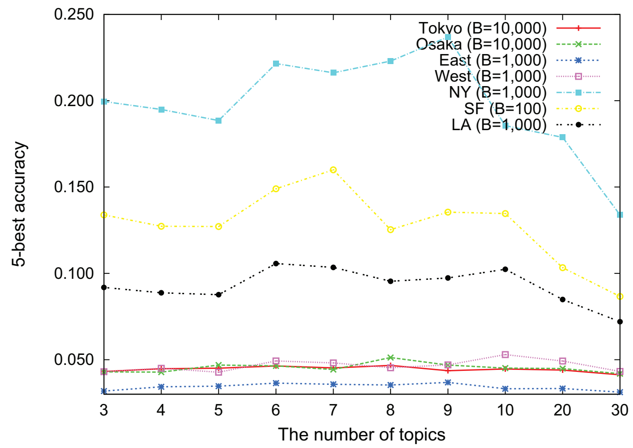



**Figure 3: 5-best (accuracy) predictions of daily-life behavior of the proposed model with different numbers of topics. The X-axis plots $Z$. The Y-axis plots the averaged 5-best (accuracy) prediction across 5 runs.**

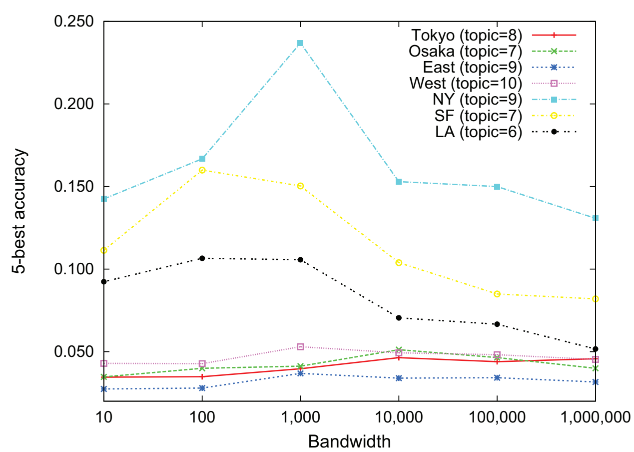



**Figure 4: 5-best (accuracy) predictions of daily-life behavior of the proposed model with different bandwidth parameters. The X-axis plots $\beta$. The Y-axis plots the averaged 5-best (accuracy) prediction across 5 runs.**

those in Tokyo (Osaka), so the estimated $\beta$ of eastern Japan (western Japan) is smaller than that of Tokyo (Osaka). Furthermore, best parameters of $Z$ of eastern Japan (western Japan) are larger than those of Tokyo (Osaka) since eastern Japan (western Japan) has more and a wider variety of restaurants than Tokyo (Osaka).

Figure 3 shows the results of 5-fold cross validation of the 5-best (accuracy) predictions of the proposed model with different numbers of topics (for each dataset, bandwidth parameter $\beta$ is set to the best parameter value). When the number of topics is small, the 5-best prediction accuracy is low. The scores of both datasets are maximized at about 10, and after that, it gradually decreases with further increases in the number of topics. The results of 5-fold cross validation of 5-best (accuracy) predictions of the proposed model at different values of $\beta$ are shown in Figure 4 (for each dataset, the number of topics, $Z$, is set to the best parameter value). These results show that setting the appropriate number of topics and best bandwidth parameter $\beta$ can yield better recommendation performance.

**Table 4: Location log data used in the second experiment. Users who visited fewer than five restaurants and Restaurants reviewed by fewer than one hundred users were omitted from each dataset.**

| Home - Destination | Restaurants (home/destination) | Logs | Ave. length |
|---|---|---|---|
| East - West | 1424 (1103/321) | 1,558 | 24.335 |
| West - East | 1424 (321/1103) | 820 | 19.162 |

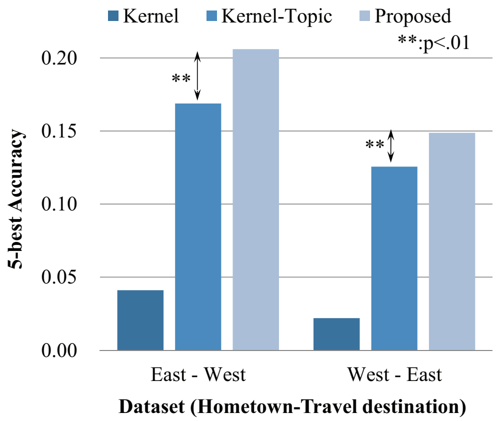



**Figure 5: Comparisons of 5-best (accuracy) predictions of travel behavior. The X-axis plots datasets and the Y-axis plots the 5-best prediction accuracy averaged across all runs.**

## 4.2 Experiment 2: Travel Behavior Prediction Accuracy

Our method was tested by recommending locations to suit the user's position while traveling (see part 3.3). In this section, we show that our method can predict visit locations while traveling with higher accuracy than is possible with the baselines. The dataset used in our experiments was generated from crawled Tabelog data. Details of the data collection and generation procedure are as follows. We first crawled users living in eastern Japan (Tokyo, Kanagawa, and Saitama prefecture). For each user, the location logs of her/his hometown and travel destination were then downloaded. Travel destinations were western Japan (including Osaka, Kyoto and Hyogo prefecture). Only users who visited restaurants in both western and eastern Japan were chosen from the users living in eastern Japan (i.e. each location log had to contain restaurants in both eastern Japan and western Japan). The dataset holds travel behavior log data because there is a great distance (about 500 kilometers) between these regions. In the evaluation, we predicted the visited restaurants in western Japan while traveling based on the location log of her/his hometown (eastern Japan). The converse dataset consists of location logs of users living in western Japan, and was used to predict travel destinations in eastern Japan from location logs for western Japan. Tokyo station and Osaka station were set as the user's current position while traveling east and west, respectively, since they are the start points of most travelers. Details of the number of users (location logs) and the number of restaurants to be recommended are shown in Table 4.

We compared three probabilistic models that can incorporate the user's current position information: Kernel model, Kernel-topic model and the proposed model. Kernel model recommends locations in the order of distance from the current position. Kernel-topic model recommends locations based on user's current position and interests estimated by existing topic model (PLSA). As is the case in the first experiment, we employed the 5-best (accuracy) predictions of restaurant visit as the evaluation metric. We tested prediction accuracy by leave-one-out cross validation. That is to say, one data was used for testing, and the others were used for training. This was repeated using each of the location log entries as a test set. So the number of tests equals the number of location log entries as shown in Table 4. Figure 5 shows the 5-best (accuracy) predictions averaged across all runs. We also used the sign test to determine the statistical significance of the differences between the average 5-best (accuracy) predictions of the proposed and the baselines. For each dataset, the result of the sign test was $p < 0.01$ (two-sided sign test). The results show that the proposed model is significantly better than the baselines in terms of the prediction accuracy of travel behavior because it generates recommendations by using user interests that do not specifically depend on user's activity area.

**Table 5: Best parameter values in the second experiment. $\beta$ was chosen from $10$, $10^2$, $10^3$, $10^4$, $10^5$, $10^6$, $10^7$ and $10^8$. $Z$ was set from 5 to 50 with interval of 5. $\alpha$ was chosen from 0.1 to 0.9 with interval of 0.1**

| Model | Hometown - Travel destination | $\beta$ | $Z$ | $\alpha$ |
|---|---|---|---|---|
| Kernel-Topic | East - West | $10^2$ | 5 | 0.9 |
| Kernel-Topic | West - East | $10^2$ | 10 | 0.7 |
| Proposed | East - West | $10^3$ | 10 | - |
| Proposed | West - East | $10^3$ | 10 | - |

All comparative models were run using the best parameters estimated using 5-fold cross validation, see Table 5. As to the proposed model, we used the best combination of $Z$ and $\beta$ from among all possible combinations. As to the Kernel-topic model, we used the best combination of $Z$, $\beta$ and $\alpha$ from among all possible combinations. The two datasets used in this experiments were generated by combining two types of location logs; the location logs of eastern and western Japan used in the first experiment. Thus the best parameters, $Z$ and $\beta$, of the combined dataset in the second experiment are almost the same as those of the location logs of eastern and western Japan in the first experiment.

## 4.3 Experiment 3: Topic Extraction

An important characteristic of probabilistic latent topic models is that they provide an interpretable low-dimensional representation of the users or locations. In this section, we qualitatively discuss the differences between the data-driven topics identified by Geo Topic Model and PLSA model.

Maps in Figure 6 shows the high-probability locations given the topics extracted by the proposed model and PLSA from the Flickr-sourced location log data used in the first experiment (New York City and San Francisco). The top 30 landmarks for each topic and each region are mapped on the map interface based on their geo-tags; the color of the mapped icon represents the topic. As shown, compared with our model, the PLSA model tends to group nearby locations together due to their positions. On the other hand, in the proposed model, each topic contains landmarks without any undue weighting from their positions. This means that the topics of the proposed model represent location-independent features of landmarks.

So what types of locations are grouped in each latent topic of our model? In order to answer this question, we further analyzed the estimated topics using the annotations (tags) associated with each landmark on the Flickr site. These annotations indicate the location names, the photographic subjects and the content of activity and are attached by the Flickr users. We extracted representative tags of each topic.

Details of the representative tag extraction procedure are as follows. Suppose that we have tag $g$ and a set of tags $G$. For each topic $z$, we calculated the $lift$ value of tag $g$ in order to extract representative tags specific to each topic. Given topic $z$, the $lift$ value of tag $g$ is calculated as follows:

$$Lift_z(g) = \frac{P(g|z)}{\sum_{z' \in \boldsymbol{Z}} P(g|z')P(z')}. \tag{13}$$

$Lift$, one of the measures for discovering interesting rules or patterns, is the ratio of $P(g|z)$ over the average for the population as a whole [12]. $Lift$ takes a low value for noisy tags that appear in all topics such as "nyc" and "sf" since it considers the global occurrence probability of $g$. $P(g|z)$ is the probability that tag $g$ is chosen from topic $z$, and calculated as follows:

$$P(g|z) = \frac{1}{C} \sum_{i \in \boldsymbol{I}} \exp(\phi_{zi}) P(g|i), \tag{14}$$

where $C = \sum_{i' \in \boldsymbol{I}} \exp(\phi_{zi'})$ is the normalization constant, and $P(g|i)$ is the probability that tag $g$ is chosen from landmark $i$, and calculated as follows:

$$P(g|i) = \frac{H(g,i)}{H(i)}. \tag{15}$$

$H(i)$ is the number of users who visited landmark $i$, and $H(g,i)$ is the number of users who annotated landmark $i$ with tag $g$.

High-value tags of the topics of the proposed model and PLSA are shown in the lower part of each map in Figure 6. Topics from 1 to 10 are ranked in descending order of $P(z) = \sum_{u \in \boldsymbol{U}} P(z|u)P(u)$ as estimated by each model. For both cities, the most probable topics (topic 1) as estimated by our model group famous sightseeing spots including Central Park, Guggenheim Museum, Golden Gate Bridge and Palace of Fine Arts Theatre. These location, contained in the most probable topic (i.e. topic 1), are interesting to sightseers. Other topics from 2 to 10 in our model tend to group a set of landmarks in terms of their features (e.g. nature, sea, construction, art, entertainment, sports, great views, and driving). On the other hand, each topic in PLSA is likely to contain location names, and does not contain tags that clearly identify location features.

As shown in Figure 6, one of the important characteristics of the recommender system based on our model is that it provides interpretable representations of latent topics. Each latent topic is transformed into an easy-to-understand representation such as popularity, art and nature. Furthermore, for each user $u$, estimated topics are ranked by user interest $\boldsymbol{\theta_u}$ as estimated and provided together with location recommendations. This offers user interaction for increasing recommendation satisfaction. For example, if a user is not satisfied with the initial location recommendations, the user first looks at the list of topics and understands the implicit estimation of profile information. Next, the user explicitly indicates which topics are interesting; At which point, our system reformulates the location recommendations based on $\boldsymbol{\phi_z}$, the location proportions of specified topic $z$.

Topic extraction is also performed using the location logs of restaurant visits. High-value tags of the topics of the proposed model and PLSA are shown in Table 6 and Table 7, respectively. We used Tabelog-based location log data from Tokyo, and the number of topics of the proposed model and PLSA were set to 10. As shown in Table 7, the high-probability tags of each topic are, in the PLSA model, pretty much the same. For example, topics 2, 4, 7 and 8 have almost all the same tags. On the other hand, in our model, restaurants appear to be grouped according to their features such as taste of food, atmosphere, and cooking style. For example, restaurants grouped in topic 3 are preferred by users who like sweet-tasting foods, while restaurants in topics 1 and 9 are preferred by users who like spicy or salty foods. The restaurants with a nice atmosphere in topics 4, 5 and 7 tend to be chosen by users on social dates (e.g. French restaurants and wine bars). Topics 2 and 6 tends to contain Japanese restaurants that serve tempura, sushi, soba and yakitori, while topics 8 and 10 tend to contain the restaurants that serve a wider variety of international cuisines such as American, Vietnamese and Spanish foods. The results show that ours can learn and extract latent features of locations from other type of location log data (i.e. location log data of restaurant visits).

# 5. CONCLUSION

This paper proposed Geo Topic Model to estimate both the user's interest and the user's spatial area of activity, simultaneously. By using the geographical features of visited locations to explicitly model the user's spatial area of activity, we can effectively estimate the user's interests in the latent features of locations beyond just their geographical positions.

The recommendation method based on Geo Topic Model meets the two criteria of a truly effective recommender system; one is recommending locations that match user's interests, and the other is augmenting location recommendations with the estimated user profile that can be customized as desired. We conducted quantitative experiments using Tabelog-based and Flickr-based real location log data to compare Geo Topic Model against other probabilistic models in terms of prediction accuracy of behavior. The results show that the latent topics estimated by our model lead to high accuracy in predicting the next location to be visited as driven by personal interest. We also showed a method for expressing latent topics in everyday language; we analyzed the content of estimated latent topics using annotation data of each location, and found that our model discovered latent topics related to art, nature, construction, great views, atmosphere and so on from actual logs of visited locations. This enables the recommender system to explain the estimated user profile in an easy-to-understand way since the user's interest can be represented as a probability distribution over the topics.

As the first and most important step, we showed the effectiveness of our proposed model in terms of the restaurant and landmark recommendation tasks which are the most popular services. A future direction is to incorporate more information such as user features to improve the performance and meet more practical information needs. Our future work also includes an evaluation in the field. The offline evaluations using behavior log data conducted in our experiments are an important starting point for exploring the validity and effectiveness of our model, but are only one approach to evaluating models and recommendation methods. We will let real users in realistic mobile settings evaluate the real strengths of our approach.

# 6. REFERENCES

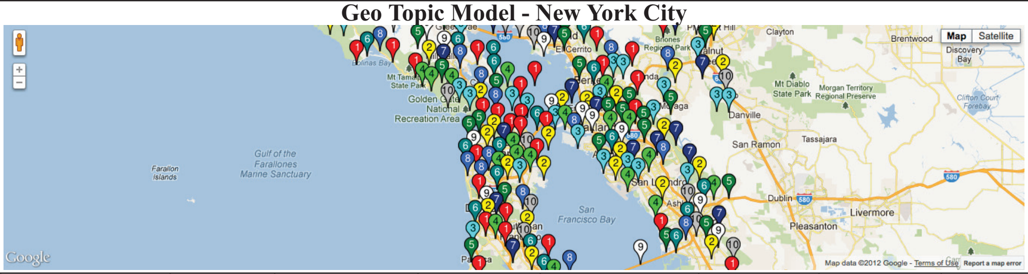

[1] Y. Arase, X. Xie, T. Hara, and S. Nishio. Mining people's trips from large scale geo-tagged photos. In *MM '10*, pages 133–142, 2010.

[2] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[3] V. Bellotti, B. Begole, E. H. Chi, N. Ducheneaut, J. Fang, E. Isaacs, T. King, M. W. Newman, K. Partridge, B. Price,

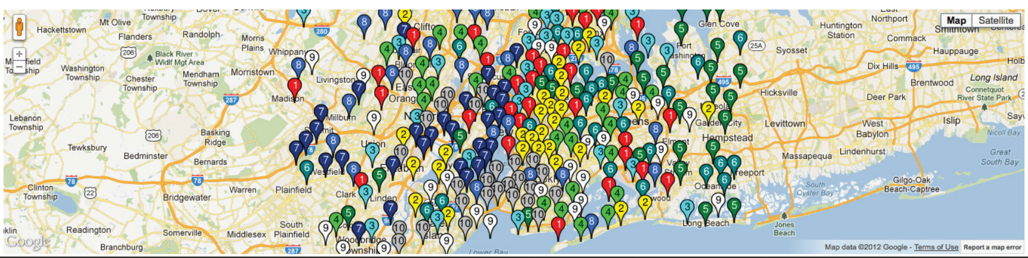
| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| centralpark | ferry | construction | art | yankeestadium | cityscape | garden | livemusic | beach | spring |
| guggenheim | bird | architecture | bar | coneyisland | liberty | flowers | graffiti | zoo | bike |
| met | sunset | chelsea | streetart | baseball | batterypark | prospectpark | music | vacation | bicycle |
| fifthavenue | lake | show | soho | summer | panorama | green | party | animals | color |
| rockefeller | boat | urban | dumbo | mets | skyline | nature | waterfront | garden | library |

**Geo Topic Model - New York City**



| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| goldengate | woods | ship | store | landscape | graffiti | concert | sausalito | bike | coast |
| palaceoffinearts | goldengatepark | alcatraz | hotel | marin | art | live | urban | bicycle | sailing |
| downtown | trees | sunrise | house | sunrise | color | fireworks | cafe | bus | summer |
| presidio | muirwoods | sea | dinner | sea | mural | show | bar | ship | bakerbeach |
| chinatown | park | bridge | marketstreet | coast | haightashbury | livemusic | restaurant | birds | bike |

**Geo Topic Model - San Francisco**



| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| met | parkslope | greenwichvillage | newark | coneyisland | yankees | jerseycity | libertyisland | harlem | jerseycity |
| guggenheim | prospectpark | soho | ice | neon | mets | airport | ellisisland | upperwestside | bridge |
| columbuscircle | astoria | eastvillage | jersey | beach | bronx | statenisland | statueofliberty | library | statenisland |
| centralpark | williamsburg | jfk | nature | island | queens | ferry | statenisland | broadway | astoria |
| fifthavenue | eastvillage | lowereastside | tree | parade | eastriver | newark | batterypark | centralpark | trees |

**PLSA - New York City**

| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| prison | palaceoffinearts | alameda | giants | goldengatepark | airport | muirwoods | oceanbeach | marinheadlands | mission |
| alcatraz | presidio | oakland | embarcadero | castro | sfo | sausalito | pacificocean | coast | mural |
| island | marina | eastbay | soma | museum | berkley | marincounty | museum | marin | hill |
| pier39 | fishermanswharf | berkeley | baseball | oceanbeach | baybridge | marin | soma | marincounty | house |
| presidio | pier39 | train | baybridge | garden | baseball | marinheadlands | alameda | pacificocean | bart |

**PLSA - San Francisco**

**Figure 6: The top five representative tags for topics estimated by our model and PLSA. We use the location logs in New York City and San Francisco. Bandwidth parameter $\beta$ of our model is set to the best parameter as estimated in the first experiment. The number of topics is 10. For each topic, $z$, tags are ordered according to $Lift_z(g)$. An interface is implemented using Google Maps API [13].**

Table 6: The top five representative tags for topics estimated by our model. We use the location logs of Tokyo. The number of topics is 10, and bandwidth parameter $\beta$ of proposed model is set to the best parameter as estimated in the first experiment.

| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| ramen | soba | cafe | Italian food | French dish | broiled meat | bar | hamburger | Indian curry | seafood |
| cold noodle | broiled eel | cake | French dish | broiled eel | beef giblets | shouchu | sandwich | Indian food | Spanish cuisine |
| dumpling | wheat noodle | chocolate | wine | tempura | yakitori | sake | steak | curry | Vietnamese food |
| dan dan noodle | beef giblets | bread | pasta | sushi | chicken | pub | pizza | Thai dish | western dish |
| Chinese noodle | tempura | bagel | pizza | soba | kebab | wine | western dish | Thai curry | Thai dish |

Table 7: The top five representative tags for topics estimated by PLSA. We use the location logs of Tokyo. The number of topics is 10.

| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 | topic 6 | topic 7 | topic 8 | topic 9 | topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| cafe | ramen | French dish | ramen | broiled meat | cafe | ramen | ramen | soba | French dish |
| cake | cold noodle | Italian food | western dish | Italian food | cake | Italian food | cold noodle | wheat noodle | western dish |
| bread | Indian curry | sushi | cold noodle | French dish | Italian food | French dish | wheat noodle | curry | cafe |
| chocolate | Indian food | wine | curry | Chinese food | hamburger | cold noodle | dumpling | western dish | chicken |
| bagel | curry | Chinese food | dumpling | cafe | French dish | pizza | Indian curry | cafe | hamburger |


P. Rasmussen, M. Roberts, D. J. Schiano, and A. Walendowski. Activity-based serendipitous recommendations with the magitti mobile leisure guide. In *CHI '08*, pages 1157–1166, 2008.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] M. D. Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *HT '10*, pages 35–44, 2010.

[6] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09*, pages 761–770, 2009.

[7] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *WWW '07*, pages 271–280, 2007.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical*, 39(1):1–38, 1977.

[9] Facebook. http://www.facebook.com/.

[10] Flickr. http://www.Flickr.com/.

[11] foursquare. http://foursquare.com/.

[12] L. Geng and H. J. Hamilton. Interestingness measures for data mining: a survey. *ACM Computing Surveys*, 38(3), 2006.

[13] Google Maps. http://maps.google.com/.

[14] Google Picasa. http://picasa.google.com/.

[15] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transaction on Information systems*, 22(1):5–53, 2004.

[16] T. Hofmann. Probabilistic latent semantic analysis. In *UAI '99*, pages 289–296, 1999.

[17] T. Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. In *SIGIR '03*, pages 259–266, 2003.

[18] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09*, pages 1427–1432, 2009.

[19] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *KDD '04*, pages 197–205, 2004.

[20] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *CIKM '10*, pages 579–588, 2010.

[21] K. W. T. Leung, D. L. Lee, and W. C. Lee. CLR: a collaborative location recommendation framework based on co-clustering. In *SIGIR '11*, pages 305–314, 2011.

[22] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[23] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. WhereNext: a location predictor on trajectory pattern mining. In *KDD '09*, pages 637–646, 2009.

[24] F. Ricci and Q. N. Nguyen. Acquiring and revising preferences in a critique-based mobile recommender system. *IEEE Intelligent Systems*, 22(3):22–29, 2007.

[25] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS '08*, pages 1257–1264, 2008.

[26] Tabelog. http://tabelog.com/.

[27] TripAdvisor. http://www.tripadvisor.com/.

[28] M. van Setten, S. Pokraev, and J. Koolwaaij. Context-aware recommendations in the mobile tourist application COMPASS. In *AH '04*, pages 235–244, 2004.

[29] yelp. http://www.yelp.com/.

[30] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *KDD '12*, pages 186–194, 2012.

[31] ZAGAT.com. http://www.zagat.com/.

[32] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *WWW '10*, pages 1029–1038, 2010.

[33] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Y. Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, 5(1), 2011.

[34] Y. Zheng, L. Zhang, X. Xie, and W. Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *WWW '09*, pages 791–800, 2009.