Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness

Bin Liu* Hui Xiong†

Abstract

The wide spread use of location based social networks (LBSNs) has enabled the opportunities for better location based services through Point-of-Interest (POI) recommendation. Indeed, the problem of POI recommendation is to provide personalized recommendations of places of interest. Unlike traditional recommendation tasks, POI recommendation is personalized, location-aware, and context depended. In light of this difference, this paper proposes a *topic and location* aware POI recommender system by exploiting associated textual and context information. Specifically, we first exploit an aggregated latent Dirichlet allocation (LDA) model to learn the interest topics of users and to infer the interest POIs by mining textual information associated with POIs. Then, a *Topic and Location-aware* probabilistic matrix factorization (TL-PMF) method is proposed for POI recommendation. A unique perspective of TL-PMF is to consider both the extent to which a user interest matches the POI in terms of topic distribution and the word-of-mouth opinions of the POIs. Finally, experiments on real-world LBSNs data show that the proposed recommendation method outperforms state-of-the-art probabilistic latent factor models with a significant margin. Also, we have studied the impact of personalized interest topics and word-of-mouth opinions on POI recommendations.

Keywords: Location-Based Social Networks, Recommender systems, Topic modeling

1 Introduction

Recent years have witnessed the increased development of location-based social networking (LBSN) services, such as Foursquare, Facebook Places and Google Latitude. LBSNs allow users to explore Places-of-Interests (POIs) for better services through sharing check-in experiences and opinions on the POIs they have checked in. For example, in Foursquare, users can (1) categorize a POI to help describe what type of places this POI is; (2) tag a POI to let people know what they can expect from it; (3) share their experiences of check-ins with others; (4) know how many people have visited a specific POI and how much time they spent there.

Indeed, the task of Point-of-Interest (POI) recommendation is to provide personalized recommendations of places of interest. It plays an important role in providing better location based services in location based social networks. Both LBSN users and POI owners are expected to have effective POI recommendations. For owners, they could have more targeted customers. Also, for users, they could identify the most relevant POIs and have better user experiences.

Unlike traditional recommendation tasks, POI recommendation is personalized, location-aware, and context depended. This can be illustrated by the following scenario. Bob lives in the New York City, usually he has a coffee in the morning at a Starbucks near his home, then has his lunch at an Italian restaurant near his office. Also, he prefers to hang out with his friends at a certain bar before he returns home. At weekends, he sometimes go to the Central Park with his family. Now, if Bob would spend the holiday in Florida, then what kind of POIs Bob would be interested in for his trip? This POI recommendation will certainly be personalized, location-aware, and context depended.

The development of POI recommender systems is much more complex than the development of traditional recommender systems. The reasons are as follows. First, for POI recommendations, the users' interest can vary dramatically at different time and locations. For instance, what POIs should we recommend to a resident in the New York City when he travels to Florida? Second, the LBSN user behaviors are intrinsically spatio-temporally correlated. The heterogeneous nature of spatio-temporal data is a big challenge for recommendation. Third, a POI is usually associated with categories and tags to describe the POI. However, unlike traditional recommendation (i.e. article recommendation [16]), the textual information associated with POIs is usually incomplete and ambiguous. Finally, even two POIs with similar or even the same semantic topics can be ranked differently if they are in two different regions.

*†Management Science and Information Systems Department, Rutgers University, NJ, USA. Email: {benbin.liu, hxiong}@rutgers.edu.

In light of the above challenges, we propose a topic and location aware method for POI recommendation. The proposed method allows to effectively exploit the textual information associated with POIs to better profile users and POIs, as well as to take into account of context aware information. Then, we develop a *Topic and Location-aware* probabilistic matrix factorization (TL-PMF) method for POI recommendation based on the learned user and POI topic distribution, and simultaneously incorporating location information. A unique perspective of this proposed method is to consider both the extent to which a user interest matches the POI in terms of topic distribution and the word-of-mouth opinions of the POI.

Finally, experimental results on real-world LBSNs data show that the proposed POI recommendation method outperforms state-of-the-art probabilistic latent factor models with a significant margin in terms of both prediction and Top-N recommendation.

2 Problem Formulation

Here, we consider POI recommendations in LBSNs. Intuitively, a user chooses a POI at a given time by matching her/his personal preferences with the service content of that POI. A user would have her/his own taste for the choice of POIs, and the personal preference can be represented by an interest topic distribution. However, even two POIs with similar or the same semantic terms can be rated differently if they are located differently. For example, a certain kind of outdoor recreation is very popular in warm and sunny California can be much less popular in a chilly northeastern area. Therefore, to provide better personalized recommendations of POIs, we need to consider both the extent to which a user's interest matches a POI in terms of topics as well as the word-of-mouth opinions of the POI.

Typically, there is textual and location-aware information associated with a POI as shown above in Table 1, which can be mined to improve location services. LBSNs such as Foursquare allow users to (1) categorize a POI; (2) tag a POI; (3) record how many different people have visited a POI and the total number of visits to this POI. As a result, the category and tag words provide semantic information about this POI. Meanwhile, the check-in numbers provide important local popularity information of that POI, which represents the word-of-mouth opinion of the POI.

From an example of POI and its associated information in Table 1, we can know detailed semantic and location information of this POI. The textual information, the categories and tags, provides meaningful semantics which can be presented in terms of topics. The last two numbers, the total number of people associated with and total number of visits to the POI, indicate

Name: Columbia Heights Coffee
Address: 3416 11th Street Northwest, Washington, DC 20010
Categories: Coffee Shop, General Entertainment, Sandwich Place
Tags: lounge chairs, tea, closes early, hipsters, coffee, outdoor seating, sandwiches, bagles, pastries, free wifi, neighborhood
Total people: 630, Total check-ins: 2,056.

Table 1: A POI and its associated information.

the word-of-mouth opinion of the POI. The larger these numbers, the more popular this POI is in this area.

Formally, we are given the historical check-in records $R_{M\times N}$ of M LBSN users $U = \{u_1, u_2, ..., u_M\}$ and N POIs $C = \{c_1, c_2, ..., c_N\}$ with r_{ij} as the number of times user u_i checked in POI c_j. r_{ij} is similar to the rating score of user u_i for item c_j in general recommendation setting. Also, for each POI, we have additional profile information such as location information, regional information in terms of city and state names, textual information in terms of categorical and tag words, and the regional popularity score P_j of POI c_j in terms of how many people associated with and how many times people visited this POI. Categories and tags are words that are assigned to describe the POI. So we have a document d_{c_j} for each POI c_j.

We build the location aware recommender system by exploring the textual and context information associated with the POIs. We argue that the rating r_{ij} of user u_i for a POI c_j is determined by two factors: (1) The extent to which the POI's interest matches a user's personalized interest in terms of topic, and (2) The regional level word-of-mouth opinion for a POI in terms of popularity score. We profile users and PoIs by mining the textual information through topic modeling.

We will use following mathematical notations in this paper. $U = \{u_1, u_2, ..., u_M\}$: a set of M users. $C = \{c_1, c_2, ..., c_N\}$: a set of N POIs. $R_{M\times N}$ with r_{ij} being the number of times user u_i checked in POI c_j. d_{c_j}: the textual items, both the tags and categories, associated with POI c_j. d_{u_i}: the items associated with POIs that user u_i visited. P_{c_j}: popularity score of POI c_j derived from the "total people" and "total check-ins". $W = \{w_1, w_2, ..., w_V\}$: unique V words set of all the associated textual information.

3 User and POI Profiling

In this section, we profile users and POIs in terms of interest distribution by performing topic models on the associated textual information.

3.1 Topic Distillation

The goal of topic distillation is to learn the interest of a user in terms of topic distribution based on the textual information of the POIs the user have checked in. Also, we need to infer the topic of interest a POI can provide. Unlike previous studies on collaborative filtering which only rely on other user's ratings to infer a given user's

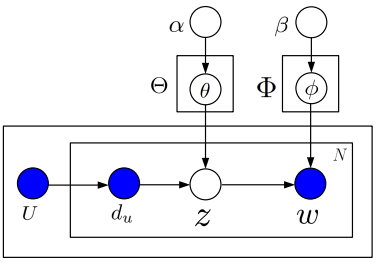

Figure 1: The aggregated LDA model.

rating on a specific item, we propose to profile user and POI through topic distillation. The Latent Dirichlet Allocation (LDA) model [3] is an popular technique to identify latent topic information from a large document collection. In LDA, each document is represented as a probability distribution over topics and each topic is represented as a probability distribution over a number of words. The model has two latent variables that can be inferred from the data: (1) the document-topic distributions Θ, and (2) the topic-word distributions Φ. Then information can be obtained about which topics users are typically interested in as well as textual representation of POIs in terms of these topics.

To distill the topics in which LBSN users are interested by applying LDA, we propose to aggregate all the documents of the POIs user u_i have checked in into a user document d_{u_i}. We combine all the terms, both the tags and categories associated with a POI, into a POI document d_{c_j}. One reason for aggregation is that the terms associated with a single POI are usually short, incomplete and ambiguous. The aggregation process can better learn a user's interest in terms of topic. Thus, the topics of d_{u_i} can represent user u_i's interest topics.

In this way, we build an aggregated LDA model as shown in Figure 1. Each document essentially corresponds to a LBSN user. As a result, the topic distribution of document d_{u_i} represents the interests of u_i. Each user u is associated with a multinomial distribution over topics, represented by θ. Each interest topic is associated with a multinomial distribution over textual terms, represented by ϕ. The generation process of the area aware user interest topic is as following:

1. For each topic $z \in \{1, ..., K\}$, draw a multinomial distribution over terms, $\phi_z \sim Dir(\beta)$.
2. For the document d_{u_i} given a user u_i
   (a) Draw a topic distribution, $\theta_{d_{u_i}} \sim Dir(\alpha)$
   (b) For each word $w_{d,n}$ in document d_{u_i}:
      i. Draw a topic $z_{d,n} \sim Mult(\theta_{d_{u_i}})$
      ii. Draw a word $w_{d,n} \sim Mult(\phi_{z_{d,n}})$

Then, we have: (1) Matrix $\Theta_{M\times K}$, where M is the number of users and K is the number of topics. θ_{ij} represents the probability that user i is interested in topic t_j. (2) Matrix $\Phi_{K\times V}$ where K is the number of topics and V is the number of unique terms in the dataset. Vector $\phi_{i\cdot}$ is the probability distribution of topic i over the V terms.

We further infer the topic distribution π_j of POI c_j based on the learned user topic term distribution $\Phi_{K\times V}$. Therefore, we can compute the topic similarity.

3.2 Model Parameter Learning

For the aggregated LDA model, we have two sets of unknown parameters of interest: the user level document-topic distributions Θ, and the topic-word distributions Φ. There is also the latent variable z corresponding to the assignments of individual words to topics. We also need to infer the topic distribution π_j for each POI through the learned model as well as the POI document d_{c_j}.

Given the two hyperparameters α and β, the complete likelihood of the model of the M user documents as shown in Figure 1 is:

$$p(W, Z, \Theta, \Phi|\alpha, \beta) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\phi_{z_{m,n}})p(z_{m,n}|\theta_m)\cdot p(\theta_m|\alpha)\cdot p(\Phi|\beta) \tag{3.1}$$

Note that it is computational intractable to directly estimate Θ and Φ in the likelihood of the LDA model as shown in Equation (3.1). During parameter estimation, we only need to keep track of $\Phi_{K\times V}$(word by topic) matrix, and $\Theta_{M\times K}$ (user by topic) matrix. From these matrices, we can estimate the topic-word distributions and user-topic distributions using Gibbs sampling [9]. First we need to sample the conditional distribution of the latent variable z as follows.

$$p(z_i = k|\mathbf{w}_i = w_i, \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{k,-i} + \beta}{n^{(\cdot)}_{k,-i} + V\beta} \cdot \frac{n^{(k)}_{d_i,-i} + \alpha}{n^{(\cdot)}_{d_i,-i} + K\alpha}$$

where the counts $n^{(\cdot)}_{\cdot,-i}$ indicate term i is excluded from the corresponding document or topic.

With the sampling results, we can estimate ϕ and θ using $\phi_{kw} = \frac{n^{(w)}_k + \beta}{\sum_{w=1}^{V} n^{(w)}_k + V\beta}$ and $\theta_{ik} = \frac{n^{(k)}_i + \alpha}{\sum_{k=1}^{K} n^{(k)}_i + K\alpha}$ where $n^{(w)}_k$ is the frequency of word assigned for topic k and $n^{(k)}_i$ the topic observation counts for document d_{u_i} of user u_i. V is the number of the unique words and K is the number of topics. α and β are two priors and here we set symmetrical priors.

Next, we infer the topic distribution $p(\pi_j|d_{c_j}, \mathcal{M})$ of a POI with document d_{c_j} given the trained model $\mathcal{M} : \{\Theta, \Phi\}$ and hyperparameters α and β. Similar to the parameter estimation for the aggregated LDA model, we use the Gibbs sampling method to derive the topic distribution for each POI [10]. The full conditional distribution of the Gibbs sampling is

$$p(z_{d_{c_j}} = k|\mathbf{w}_i = w_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \mathcal{M}) \propto \phi_{k,w_i}(n^{(k)}_{d_{c_j},-i} + \alpha)$$

Then, the topic distribution for POI d_{c_j} is: $\pi_{jk} = \frac{n_j^{(k)}+\alpha}{\sum_{k=1}^{K} n_j^{(k)}+K\alpha}$, where $n_j^{(k)}$ is the topic observation count for POI document d_{c_j}.

3.3 Interest Matching Score

After deriving the interests of both users and POIs in terms of topic distribution, we can compute the extent to which a POI's interest matches a user's personalized interest by a matching score. The matching score between user u_j and POI c_j is defined as the similarity in terms of user interest topic distribution θ_i and POI topic distribution π_j. We use the symmetric Jensen-Shannon divergence between user u_i and POI c_j is:

$$D_{JS}(u_i, c_j) = \frac{1}{2}D(\theta_i \parallel M) + \frac{1}{2}D(\pi_j \parallel M)$$

where $M = \frac{1}{2}(\theta_i + \pi_j)$ and $D(\cdot \parallel \cdot)$ is the Kullback-Leibler distance. Then we define the matching score as $S(u_i, c_j) = 1 - D_{JS}(u_i, c_j)$.

4 A Topic and Location Aware Probabilistic Matrix Factorization (TL-PMF) Model

Since the POI recommendation is personalized, location-aware, and context depended, we introduce a *Topic and Location-aware* probabilistic matrix factorization (TL-PMF) method for POI recommendation by considering both the extent to which a user interest matches the POI in terms of topic distribution and the word-of-mouth opinions of the POI.

4.1 The Topic and Location-Aware POI Recommendation in LBSNs

In addition to the POI textual information and word-of-mouth opinions, we have the LBSN user's historical check-in record matrix R with r_{ij} being the number of times user u_i has checked in POI c_j. This also applies when r_{ij} is binary variable ($r_{ij} = 1$ meaning u_i interested in POI c_j and $r_{ij} = 0$ meaning not). We see r_{ij} as the rating of a user u_i for POI c_j.

For POI recommendation in LNSNs, we need to consider both (1) the extent to which the POI interest topic matches a user's personalized interest in terms of topics, and (2) the regional level word-of-mouth opinion for a POI in terms of popularity scores in a region. The rating r_{ij} of a user u_i for POI c_j is determined by user factors and POI factors. On the one hand, the rating should reflect the matching between the POI topic and the user interest topic. The rating is higher if two topic distributions match better. On the other hand, the rating should reflect the word-of-mouth opinion index P_j of the local area.

We define the Topic and Location influence index of user u_i for POI c_j as

$$TL_{ij} = \gamma S(u_i, c_j) + (1-\gamma)P_j \tag{4.2}$$

Here, $S(u_i, c_j)$ is a marching score between user u_j and POI c_j in terms of user interest topic distribution θ_i and POI topic distribution π_j. The second term P_j is a regional level popularity factor for POI c_j as a word-of-mouth opinion on the POI. γ is a factor to balance these two factors. Then TL_{ij} considers both interest topic match between user and POI, and location aware word-of-mouth opinions for a POI.

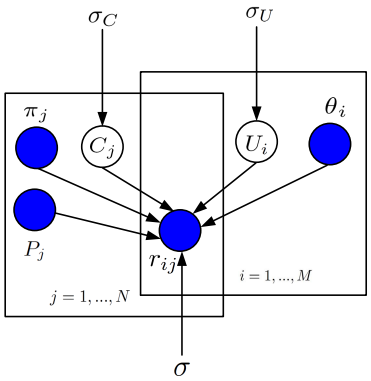

Figure 2: The TL-PMF model.

4.2 The TL-PMF Model

To leverage the influence interest topic and location aware word-of-mouth opinions for POI recommendation, we propose a *Topic and Location-aware* probabilistic matrix factorization (TL-PMF) model. The graphical representation of TL-PMF is shown in Figure 2. Let r_{ij} be the rating of user u_i for POI c_j, U_i and C_j are the user and POI latent feature space vector respectively. The distribution over the observed ratings as well as the textual information is

$$p(R|U, C, TL, \sigma^2) = \prod_{i=1}^{M}\prod_{j=1}^{N}\left[\mathcal{N}(r_{ij}|f(U_i, C_j, TL_{ij}), \sigma^2)\right]^{I_{ij}} \tag{4.3}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. I_{ij} is the indicator function. Function $f(U_i, C_j, TL_{ij})$ is to approximate the rating of user u_i for POI c_j.

Consider the influence interest topics and location aware word-of-mouth opinions on user u_i's preference for POI c_j, we define

$$f(U_i, C_j, TL_{ij}) = TL_{ij} \cdot U_i^T C_j \tag{4.4}$$

where U_i and C_j are $D-$dimensional latent factors for user u_i and POI c_j respectively, TL_{ij} is the topic and location index of user u_i for POI c_j. Here we use a weighted product of user latent factors and POI factors by incorporating topic and location index to improve PMF model. TL_{ij} is derived from the aggregated topic model and the popularity score as shown in Section 3.

We set zero mean Gaussian prior to user and POI latent space [15]: $p(U|\sigma_U^2) = \prod_{i=1}^{M}\mathcal{N}(U_i|0, \sigma_U^2\mathbf{I})$ and $p(C|\sigma_C^2) = \prod_{j=1}^{N}\mathcal{N}(C_j|0, \sigma_C^2\mathbf{I})$. Then, the posterior distribution of Equation (4.3) becomes

$$p(U, C|R, \sigma^2, TL, \sigma_U^2, \sigma_C^2) \propto p(R|U, C, \sigma^2, TL, \sigma_U^2, \sigma_C^2) p(U|\sigma_U^2) p(C|\sigma_C^2)$$
$$= \prod_{i=1}^{M} \prod_{j=1}^{N} \left[\mathcal{N}(r_{ij}|f(U_i, C_j, TL_{ij}), \sigma^2)\right]^{I_{ij}}$$
$$\times \prod_{i=1}^{M} \mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I}) \times \prod_{j=1}^{N} \mathcal{N}(C_j|0, \sigma_C^2 \mathbf{I})$$

We need to estimate parameters in terms of maximizing likelihood. The log posterior distribution is:

$$\mathcal{L}(U, C|R, \sigma^2, TL, \sigma_U^2, \sigma_C^2)) =$$
$$- \frac{1}{2\sigma^2} \sum_{i=1}^{M} \sum_{j=1}^{N} I_{ij} (r_{ij} - f(U_i, C_j, TL_{ij}))^2 - \frac{1}{2\sigma_U^2} \sum_{i=1}^{M} U_i^T U_i$$
$$- \frac{1}{2\sigma_C^2} \sum_{j=1}^{N} C_j^T C_j - \frac{1}{2} \left[\left(\sum_{i=1}^{M} \sum_{j=1}^{N} I_{ij} \right) \ln\sigma^2 + MD \ln\sigma_U^2 + ND \ln\sigma_C^2 \right]$$

where D is the dimension of the latent factors. Maximizing the log posterior equals to minimizing the following function

$$E = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} I_{ij} \left(r_{ij} - TL_{ij} \cdot U_i^T C_j\right)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{M} ||U_i||_F^2 + \frac{\lambda_C}{2} \sum_{j=1}^{N} ||C_j||_F^2 \quad (4.5)$$

where $\lambda_U = \sigma^2/\sigma_U^2$, $\lambda_C = \sigma^2/\sigma_C^2$, and $|| \cdot ||_F^2$ is the Frobenius norm. Performing a gradient descent method on U and C can lead to a local minimum solution to Equation (4.5) using: $\frac{\partial E}{\partial U_i} = -\sum_{j=1}^{N} I_{ij} \left(r_{ij} - TL_{ij} \cdot U_i^T C_j\right) \cdot TL_{ij} C_j + \lambda_U U_i$ and $\frac{\partial E}{\partial C_j} = -\sum_{i=1}^{M} I_{ij} \left(r_{ij} - TL_{ij} \cdot U_i^T C_j\right) \cdot TL_{ij} U_i + \lambda_C C_j$

4.3 Prediction and Recommendation

After the user interest topic and parameters U, C are learned, the TL-PMF model prediction of the rating of a user for a given POI is estimated as $\mathbb{E}(r_{ij}|u_i, c_j) = TL_{ij} \cdot U_i^T C_j$ where γ can adjust the weight of matching score and the local popularity score.

Since recommendation in LBSNs is highly location sensitive, the recommendation list should be close to the user's current region and thus it advisable to recommend POIs near the user's physical location. Our TL-PMF model provides global predicted preference scores global. In real practice, we need take into consideration of location information to make reasonable personalized POI recommendations. Given a user's current location L_{u_i}, one possible way to make recommendations is to recommend N POIs corresponding to top N prediction scores within a certain range $\text{Range}_{L_{u_i}}$.

5 Experimental Results

In this section, we provide an empirical evaluation of the performances of the proposed model. All the experiments were performed on a large real-world LBSN dataset collected from Foursquare, one of the largest and most popular LBSN community.

The Wonderland Ballroom, Washington, DC, 2010-07-24, 04:25:41
Black Squirrel, Washington, DC, 2010-07-24, 16:42:28
Columbia Heights Coffee, Washington, DC, 2010-07-25, 01:19:02
The Wonderland Ballroom, Washington, DC, 2010-07-25, 02:08:44
Commonwealth Gastropub, Washington, Dc, 2010-07-25, 07:45:51
Washington National Airport, Arlington, Va, 2010-07-26, 19:20:47
.
Fornelletto, Atlantic City, NJ, 2010-08-10, 18:45:42
.
Panera Bread, Knoxville, TN, 2010-08-26, 17:10:08
.
Lou Malnati's Pizzeria, Chicago, IL, 2010-10-19, 00:26:25
.

Table 2: A check-in trace of a user.

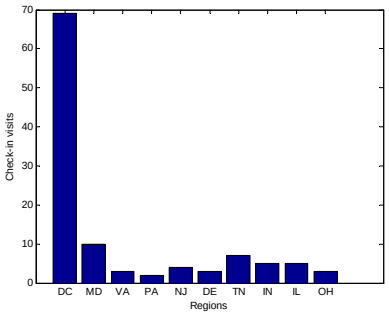

Figure 3: An example of check-in reports for a user

5.1 The Experimental Data

The dataset is formulated as follows [4]: Foursquare users usually report their check-ins of POIs via Twitter. When a LNSN user posted a Tweet, which indicates a check-in of a POI, we consider it as the user has checked in physically. Also from Foursuqre, we have detailed information of each POI with its location in terms of latitude and longitude, region, the associated categories, tags, the total number of people, and the total number of check-ins. With both the LBSN's tweet check-in reports, in which latitude and longitude are available, and the LBSN check-in profiles have latitude and longitude values, we can match these two sources of information to obtain LBSN users' check-in profiles with additional information for the POIs.

Table 2 shows an example of the check-in trace for a user, who had reported her/his visit to different POIs at different states in USA. Figure 3 shows the check-in report times for the user in different regions. A user would usually have her/his home address, which corresponds to the highest frequent report region, and may visit POIs at different regions.

Table 3: Data Description

user	POIs	rating	avg # rates	sparsity
35,025	49,779	1,080,824	30.85	99.94%

As a lot of users may have checked in or reported few check-ins, we exclude those users with less than 6 check-in records. As the number of words associated with each POI vary dramatically, we select the POIs with the minimum 10 tags. We finalized a dataset as

topic	terms
1	airport terminal travel airlines delta gate tsa high mile gogo gogoinflight united southwest wifi baggage continental airplane handler airways
2	san technology apple office bar diego home iphone gym shop computer ipod store video mac coffee center ipad restaurants
3	sea seattle tac bar seatac wifi coffee beer free waterfront airport food fish limo restaurant limousine square hour colman
4	train station transit new bus subway rail metro public food nj transportation line amtrak york penn city jersey express
5	bar food beer coffee wine restaurant free bbq music burgers trivia patio pool delivery italian chicken american outdoor bon
6	theater movie movies theatre food gallery photo booth photobooth popcorn mall cinema pizza douchebag cineplex shopping imax art store
7	college university frat gas food library school pizza student coffee center state boys bar building campus store station gym
8	marketing design media social web office music corporate advertising food coffee search seo agency development restaurant internet digital toronto
9	attorney law injury accident lawyer personal lawyers attorneys city atlanta firm bar beer restaurant bankruptcy office food sports oc
10	mall store food mobile accessories shopping american wireless apple cell court phone department macy coffee photobooth body women shoes

Table 4: Some selected topics (identified by aggregated LDA when $K = 30$)

shown in Table 3. Here, we use implicit rating, namely the number of checks-in for a POI as the rating for the POI. This is different from the rating in movie recommendation, in which the rating is usually in a range from 1 to 5. So, we need to transfer the discrete rating to a value between [0,1] by using $f(x) = (x - 1)/(K - 1)$ with K is the maximum rating value [15]. We can see that the rating matrix is very spare with 99.94% missing ratings.

5.2 Evaluation Metrics

We adopt the *Root Mean Squared Error* (RMSE) to measure the prediction error. RMSE is defined as $\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i,j}(\hat{r}_{ij} - r_{ij})^2}$ where r_{ij} denotes the rating of POI j by user i, $\hat{r}_{ij}$ denotes the corresponding rating predicted by the model, and N denotes the total number of the tested rating. The smaller the value of RMSE, the more precise a recommendation.

Ranking a recommendation list is often more important than the rating prediction. So we also evaluate the algorithms in terms of ranking. We present each user with N POIs sorted by their predicted rating and evaluate based on which of these POIs were actually visited by the user. However, a direct use of top N based metric like *recall@N* and *precision@N* would incur underlying biases as this metric depends heavily on the percentage of relevant items that each user has rated [11]. In our dataset, a user has rated only a very small percentage (about 0.06%). We adopt the relative rank evaluation method as described in [12]

First we select the $|\mathcal{T}|$ highest rating set $\mathcal{T}$ from the test dataset. For each POI $c_j \in \mathcal{T}$ for user u_i, we add another $|\mathcal{C}|$ randomly selected POIs $\mathcal{C}$, and predict the rating for $\{c_j, \mathcal{C}\}$. Then, we sort the $|\mathcal{C}| + 1$ predicted rating scores in a descending order. In this way, we can find the relative place of these interesting POIs in the total order of the recommendation list for a given user. We can obtain a cumulative distribution of the relative ranking based on the selected rating set $\mathcal{T}$.

5.3 Implementation Details

We divided the data into training (80%) and testing (20%) data. We compared TL-PMF with PMF. We did not use other matrix factorization methods like SVD based methods as the benchmark because it has been shown that PMF outperforms SVD approaches [15].

For TL-PMF, we further set different parameter γ in the topic and location index $TL_{ij} = \gamma S(u_i, c_j) + (1 - \gamma)P_j$ to test how local popularity factor influence user's preference choice. When $\gamma = 1$, it means that the recommendation is made by only including user interest topic and is denoted as TL-PMF$_{\mathbf{T}}$; $\gamma = 0$ means that the rating mainly relies on local word-of-mouth popularity information and is denoted as TL-PMF$_{\mathbf{L}}$, and $0 < \gamma < 1$, denoted as TL-PMF$_{\mathbf{TL}}$, means that the recommendation is made by combining both user interest topic and local popularity opinion.

We normalize the local rating score for POI j in area to $[0, 1]$ range by the following equation. $\hat{P}_j = \frac{1}{2}\left\{\frac{\text{totalPeo}_j - 1}{\max_j\{\text{totalPeo}_j\} - 1} + \frac{\text{totalCk}_j - 1}{\max_j\{\text{totalCk}_j\} - 1}\right\}$ where $\max_j\{\text{totalPeo}_j\}$ and $\max_j\{\text{totalCk}_j\}$ are the maximum total people value and total check-in value in the area respectively.

We set $\lambda_U = 0.01$ and $\lambda_C = 0.01$ for PMF, TL-PMF$_{\mathbf{T}}$, TL-PMF$_{\mathbf{TL}}$ and TL-PMF$_{\mathbf{L}}$. We set $\alpha = 50/K$ and $\beta = 0.1$ in the aggregated LDA model.

Table 4 shows some of the user interest topics learned from the aggregated LDA when $K = 30$. These topics include transportation, technology, recreation, restaurant, school, company, shopping and so on.

5.4 Performance Comparisons

Here, we compare the performances of different approaches in terms of RMSE and Top-N metrics.

5.4.1 Performance comparison I: RMSE

With RMSE, we compare TL-PMF and PMF at different settings. We first set the number of topics $K = 30$ and $K = 50$ to learn user topic interest, and thus get the topic and location index TL_{ij}. We do not directly use $\mathbb{E}(r_{ij}|u_i, c_j) = g(TL_{ij} \cdot U_i^T C_j)$ for prediction but pass the results through a logistic function $g(x) = \frac{1}{1+\exp(-x)}$ to bound the prediction score to range $[0, 1]$. Then the prediction becomes: $\mathbb{E}(r_{ij}|u_i, c_j) = g(TL_{ij} \cdot U_i^T C_j)$. In each topic number case, we perform TL-PMF with different user and POI factor dimensions ($D = 10$ and $D = 30$). Also, we compare the effect of local popularity P_j in recommendation.

	D=10		D=30	
Model	30 topics	50 topics	30 topics	50 topics
PMF	0.2488		0.2470	
TL-PMF$_{\mathbf{T}}$	0.2362	0.2345	0.2388	0.2380
TL-PMF$_{\mathbf{TL}}$	0.2305	0.2301	0.2324	0.2319

Table 5: A prediction comparison of TL-PMF with PMF in terms of RMSE with two different factor dimensions in two different topic number settings (Note: PMF does not involve topics.).

As shown in Table 5, no matter whether incorporating only topic model or both topic model and local popularity rating, TL-PMF outperforms PMF. For example, when topic number $K = 30$ and factor dimension $D = 10$, comparing to PMF, TL-PMF$_{\mathbf{T}}$ improves RMSE by 5.1%, and TL-PMF$_{\mathbf{TL}}$ with $\gamma = 0.5$ improves RMSE by 7.3%. We can see that TL-PMF$_{\mathbf{T}}$ improves recommendation performances by incorporating user's personal interest learned by topic model. TL-PMF$_{\mathbf{TL}}$ further improves recommendation by balancing both a user's personal interest and the word-of-mouth opinion.

To further investigate the effect of word-of-mouth opinions on recommendation performances, we perform another experiment by adjusting γ, which controls the weight of personal and word-of-mouth opinion factors. Table 6 shows that RMSE varies according to the different rating determination factor parameter γ. The change of RMSE with λ is shown in Figure 4. In the figure, we can see that word-of-mouth opinions not always compensate personalized interests to improve recommendation performances. When we depend too much on local popularity score, happening when γ approaches to 0, the recommendation performance starts decreasing, and even can be worse than PMF without additional information. Another problem with too much weight to local popularity score is the slow convergence of the algorithm. Note that the RMSE value 0.398 (corresponding to $\gamma = 0$) is the result after 5000 iterations. One explanation is that the personal interest is not always consistent with word-of-mouth opinions.

5.4.2 Performance comparison II: Top N

Since POI recommendation in LBSNs is highly location sensitive, the recommendation list should be close to the user's current region. Figure 3 shows an example of check-in reports for a user in different regions. A user would visit POIs at different regions. Therefore, we measure the Top N performance by considering the recommendation list within a certain range of the target user's current location.

We use the relative ranking measure as introduced in Section 5.2. We select the highest $|\mathcal{T}|$ rating $\mathcal{T}$ set from the test data as the probe POIs. Then, for each probe POI and the corresponding user, we randomly select $|\mathcal{C}| = 500$ POIs $\mathcal{C}$ within a certain range Range$_{L_{c_j}}$

γ	0	0.1	0.2	0.5	0.7	1
RMSE	0.398	0.2418	0.2318	0.2305	0.2320	0.2362

Table 6: A comparison of TL-PMF$_{\mathbf{TL}}$ with different γ values in topic and location index $TL_{ij} = \gamma S(u_i, c_j) + (1-\gamma)P_j$. Here, $K = 30$ and $D = 10$.

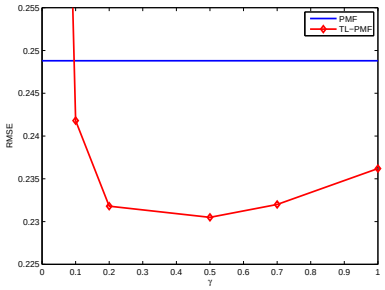

Figure 4: The RMSE values of TL-PMF$_{\mathbf{TL}}$ with different γ values in topic and location index $TL_{ij} = \gamma S(u_i, c_j) + (1-\gamma)P_j$ (red line) vs. PMF (blue line).

of the probe POI location L_{c_j}, In this way, we can find the relative rank of these probe POIs in the total order of the recommendation list for a given user.

We compare the top N performances of TL-PMF$_{\mathbf{TL}}$, TL-PMF$_{\mathbf{T}}$ and PMF using the relative ranking measure. Figure 5 shows the cumulative distribution of the percentile relative rank for TL-PMF$_{\mathbf{TL}}$, TL-PMF$_{\mathbf{T}}$ and PMF. Note that the straight line connecting the bottom-left and top-right corners is for random prediction. As can be seen, our TL-PMF models, both the TL-PMF$_{\mathbf{T}}$ model with just topic model and the TL-PMF$_{\mathbf{TL}}$ model with topic model as well as regional level word-of-mouth opinion, outperforms PMF (dot blue) significantly. Indeed, for the case when x-axis value is equal to 0.1 or 10%, which corresponds to recommend top-50 POI recommendation: probabilistically, the numbers of POIs will match user interest are $50 \times 22.38\% \approx 11$ with PMF, $50 \times 91.33\% \approx 45$ with TL-PMF$_{\mathbf{T}}$ model, and $50 \times 96.52\% \approx 48$ with TL-PMF$_{\mathbf{TL}}$ model respectively. In the experiment, we set the location range Range$_{L_{c_j}}$ as a state level. We can potential expect to make more relevant POI recommendations by narrowing the location range value.

5.4.3 Summary

In summary, the proposed models can outperform the baseline method dramatically in terms of both RMSE and Top N metrics. We have observed that both personalized user interest topic as well as location dependent word-of-mouth opinion can be incorporated into the proposed flexible framework to improve recommendation performance.

5.5 Topic Analysis in LBSNs

Here, we analyze the topic characteristics of POIs across different geographical regions. We have shown

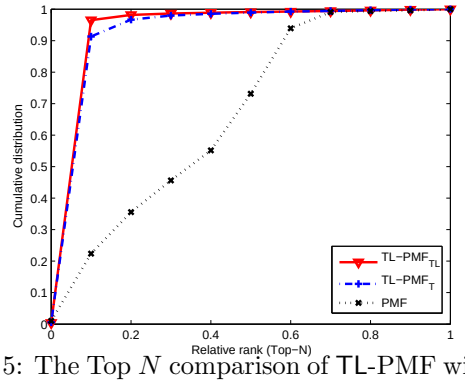

Figure 5: The Top N comparison of TL-PMF with PMF in terms of relative rank. X-axis stands for the relative rank in percentage of the probe POIs.

that the user generated textual tags, which are aimed to better describe what type of places a POI is, help to improve POI recommendation. We are further interested in studying whether different areas would present different topics, and what is the effect of topic difference on recommendation. To this end, we select eight areas from all the POI dataset: California (CA), Arizona (AZ), Texas (TX), Florida (FL), Chicago area (IL), Washington DC (DC), Boston area (MA) and New York area (NY), and form a region level POI data set. These areas cover different regions and are representative of regional differences.

We aggregate all the POIs in an area into a region-level document and have eight region-level documents. For each region, we infer the region document-topic distribution π based on the topics we learned by setting $K = 30$. For each region pair $\{R_i, R_j\}$ within the selected regions, we can compute the correlation of the topic distribution Corr_{ij} by using $\text{Corr}_{ij} = \frac{\sum_{k=1}^{K}(\pi_{ik}-\bar{\pi}_i)(\pi_{jk}-\bar{\pi}_j)}{\sqrt{\sum_{k=1}^{K}(\pi_{ik}-\bar{\pi}_i)^2}\sqrt{\sum_{k=1}^{K}(\pi_{jk}-\bar{\pi}_j)^2}}$ where $\bar{\pi}_i$ and $\bar{\pi}_j$ are the average topic probability for regions R_i and R_j respectively. Then, we have the region-level topic correlation in Table 7 and its visualization in Figure 6.

In Table 7 and Figure 6, we can see that the region difference poses different topics because the correlation between the region level topics are almost near 0 or negative. In the selected regions, the California and Florida areas share the highest correlation. Florida shares high correlation with both Arizona and Texas, but Arizona and Texas do not have high correlation (the correlation between Arizona and Texas is -0.0637).

Through the topic analysis of both user interest topic and regional level topic comparison, we revealed that (1) Most POIs of LBSNs are dominated by a few topics, which are common life topics, as shown in Table 4; (2) Topics differ in different regions even in contiguous regions. This implies that we should take into consideration of both personalized user interests as

	AZ	TX	FL	IL	DC	MA	NY
CA	0.1163	0.0836	0.2974	0.0572	0.0302	0.0943	0.0937
AZ		-0.0637	0.0577	-0.1096	-0.0061	-0.0330	-0.0876
TX			0.1294	-0.0588	-0.0372	-0.0152	-0.0495
FL				0.0202	-0.0529	-0.0247	-0.0071
IL					-0.0200	-0.0504	-0.0598
DC						0.0159	0.0198
MA							-0.0424

Table 7: The regional level topic correlation when the topic number $K = 30$.

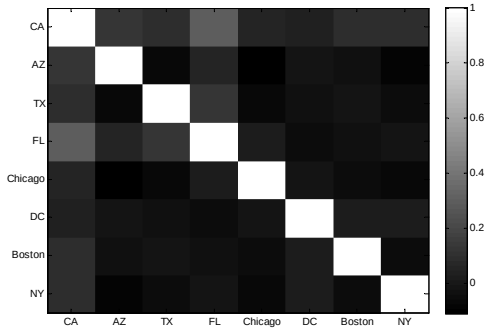

Figure 6: The correlation of topic distributions between different selected areas.

well as the regional level word-of-mouth opinions for POI recommendations in LBSNs.

6 Related Work

Related work can be grouped into two categories: works study place of interest recommendation from the application perspective, and works study how to exploit textual information to improve recommendation from the methodology perspective.

With increasing popularity of LBSNs, applying POI recommendation to provide better location based service has caught a lot of attentions from both academia and industry. Previous studies on POI recommendations mainly relied on user trajectory data. For example, various works [21, 2, 20, 8, 13, 7] applied collaborative filtering based method to recommend locations and travel packages based on user trajectory data. By considering the geographical influence due to the spatial clustering phenomenon in LBSN users, Ye et al [18] explored user preference, social influence and geographical influence for recommending POIs in LBSNs.

More recent work began to explore textual information to better understand patterns in LBSN and to improve LBSN services. For instance, [5] applied topic models to identify daily location-driven routines by mining text from mobile phone data. [17] presented a work on semantic annotation for LBSNs to annotate places with category tags by exploring explicit patterns of individual places and implicit relatedness among similar places. [19] proposed a latent geographical topic analysis method to explore both location and associated text of locations and found this can help to discover meaningful geographical topics. Finally, [6] analyzed Twitter

posts and performed LDA on the data to extract urban patterns, such as hotspots and crowd behaviors.

There are works to explore textual information for recommendation. A straightforward way is to combine collaborative filtering with topic models. By mining the textual information associated with each item, we could combine probabilistic matrix factorization (PMF) [15] and topic models [16]. The fLDA model in [1] follows this line, but they associated the rating by regularizing both user and item factors simultaneously through user features and words associated with each item. In addition to exploring topic models for item recommendation, there are also studies which use topic models to learn social-media user interests to recommend new friends with similar interests [14].

Unlike the tasks of recommending movies and scientific papers [16], the problem of POI recommendation in LBSN services is location-aware, personalized, and context depended. In addition, the textual terms associated with POIs are usually incomplete and ambiguous. This study explores both associated textual and context information to address these challenges.

7 Conclusion

In this paper, we studied the POI recommendation problem in LBSNs by exploiting textual information as well as regional word-of-mouth opinions. There are several advantages of the proposed recommendation method. First, the textual terms associated with POIs are usually incomplete and ambiguous. To meet this challenge, the proposed method exploits location dependent word-of-mouth opinions in addition to users' personalized interests learnt from the insufficient POI textual information. Second, the location-aware aggregated LDA recommendation approach allows to profile user interests by performing topic modeling of the users' historical textual information. This provides a way to match the user interests to the POI topic, and thus alleviate the cold start problem in recommendation. Third, the proposed recommendation method can strike a balance between the use of individual information and the use of location-aware word-of-mouth opinions. This helps to avoid the excessive use of personalized information, and thus reducing the possibility of overfitting. Last but not least, the proposed method is flexible and could be extended to incorporate other types of context-aware information to enhance POI recommendation.

Acknowledgement. This research was partially supported by National Science Foundation (NSF) via grant numbers CCF-1018151, IIS-1256016, and DUE-1241315. Also, it was supported in part by Natural Science Foundation of China (70890082, 71028002). Finally, the author gratefully acknowledges the support of K. C. Wong Education Foundation, Hong Kong.

References

[1] D. Agarwal and B.-C. Chen. flda: matrix factorization through latent dirichlet allocation. WSDM'10, pages 91–100.

[2] B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In *4th Workshop on Social Network Systems*, 2011.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[4] Z. Cheng, J. Caverlee, K. Y. Kamath, and K. Lee. Toward traffic-driven location-based web search. CIKM'11, pages 805–814, 2011.

[5] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27, Jan. 2011.

[6] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli. Extracting urban patterns from location-based social networks. In *3rd ACM SIGSPATIAL Workshop on Location-Based Social Networks*, pages 9–16, 2011.

[7] Y. Ge, Q. Liu, H. Xiong, A. Tuzhilin, and J. Chen. Cost-aware travel tour recommendation. In *KDD*, pages 983–991, 2011.

[8] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani. An energy-efficient mobile recommender system. In *KDD*, KDD '10, pages 899–908, 2010.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[10] G. Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, 2009.

[11] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, Jan. 2004.

[12] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. KDD '08, pages 426–434, 2008.

[13] Q. Liu, Y. Ge, Z. Li, E. Chen, and H. Xiong. Personalized travel package recommendation. In *ICDM*, pages 407–416, 2011.

[14] M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *WWW 2011*, pages 101–102, 2011.

[15] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, volume 20, 2008.

[16] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD 2011*, pages 448–456, 2011.

[17] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *KDD 2011*, pages 520–528, 2011.

[18] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. SIGIR '11, pages 325–334.

[19] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. WWW '11, pages 247–256.

[20] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. WWW'10, pages 1029–1038.

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