CoSoLoRec: Joint Factor Model with Content, Social, Location for Heterogeneous Point-of-Interest Recommendation

Hao Guo[1], Xin Li[3], Ming He[1], Xiangyu Zhao[1], Guiquan Liu[1](✉), and Guandong Xu[2]

[1] University of Science and Technology of China, Hefei, China
[2] University of Technology Sydney, Sydney, Australia
[3] IFLYTEK Research, Hefei, China
{guoh916,zxy1105}@mail.ustc.edu.cn xinli2@iflytek.com gqliu@ustc.edu.cn
Guandong.Xu@uts.edu.au mheustc@gmail.com

Abstract. The pervasive use of Location-based-Social-Networks calls for more precise Point-of-Interest recommendation. The probability of a user's visit to a target place is influenced by multiple factors. Though there are several fusion models in such fields, heterogeneous information are not considered comprehensively. To this end, we propose a novel probabilistic latent factor model by jointly considering the social correlation, geographical influence and users preference. To be specific, a variant of Latent Dirichlet Allocation is leveraged to extract the topics of both user and POI from reviews which is denoted as explicit interest. Then, Probabilistic Latent Factor Model is introduced to depict the implicit interest. Moreover, Kernel Density Estimation and friend-based Collaborative Filtering are leveraged to model users geographic allocation and social correlation respectively. Thus, we propose CoSoLoRec, a fusion framework, to ameliorate the recommendation. Experiments on two real-word datasets show the superiority of our approach over the state-of-the-art methods.

Keywords: Location-based Social Network, Point-of-Interest Recommendation, Topic Model, Probabilistic Latent Factor Model, Heterogeneous Information

1 Introduction

In recent years, with rapidly development of Location-based Social Networks (LBSNs), boundary between the physical world and virtual networks is broken. As an interlink between these two worlds, Point-of-Interest (POI) refers to a place, such as restaurant that users may find useful or tend to visit and plays an essential role in LBSN thereby leading to an application - POI recommendation. This application can not only benefit merchants by increasing their revenue through virtual marketing but also benefit customers accrelating their decision-making by filtering out uninteresting places thus makes them satisfied.

Traditional recommender systems can be seamlessly applied by treating POI as an ordinary item, however there are several characteristics of POI recommendation that make it different from conventional recommendations thus if well considered, the performance would be improved in a significant margin.

- Tobler's Law of Geographical Influence. As Tobler [1] indicates, the aggregation of a user's check-ins depicts that users' check-in probability is inversely

proportional to the geographical distance. Geographical influence can be denoted as a physical metric between the user and the POI.
- Homophily of Social Correlation. Homophily is one of the most important theories in sociology and also works in social network [2]. The depiction of homophily suggests that people tend to trust and have similar favorite as their friends psychologically, thus making social correlation a psychological metric between the user and the POI.
- Heterogeneous Information. A LBSN contains heterogeneous information, such as geographical location, social network, rating data and text reviews. Undoubtedly, utilizing heterogeneous data indubitably results in more precise user profiling and more personalized recommendations.

Despite the successes and improvements of the existing studies, the heterogeneous information is not comprehensively considered in one model. Generally speaking, users' behavior in choosing POIs can be influenced by multiple factors. So our recommendation will show more accurate and efficient if we consider more factors which will influence users' bahaviors. However, under some circumstance in reality, one or several pieces of information are not available, so real life applications demand for robust modeling. To this end, we propose a novel probabilistic latent factor model by considering the geographical location, social correlation and textual reviews simultaneously. Specifically, our model consists of four-fold measurements, i.e., physical distance, psychological distance, explicit interest and implicit interest. Firstly, physical distance denotes the distance between the user and the POI in real world. We apply Kernel Density Estimation (KDE) to estimate the visiting probability of a user at a target POI based on his or her historical visiting records in adherence to Tobler's Law. Secondly, for calculating psychological distance, friend-based Collaborative Filtering (CF) is simultaneously utilized to predict the visiting probability under the assumption called "the phenomenon of homophily" which means a user's visiting behavior is influenced by his or her friends. Thirdly, in order to leverage a user's explicit interest, we aggregate all the reviews written by him or her as a document and apply Latent Dirichlet Allocation (LDA) on it to derive the topic distribution of such user afterwards. The matching on the corresponding topics reveals preference for a user to a POI. Finally, under the belief that there are several factors that implicitly affect a user's decision-making, we apply the Latent Factor Model to depict the implicit interest thereby augmenting it to the explicit interest to form the entire interest in the model.

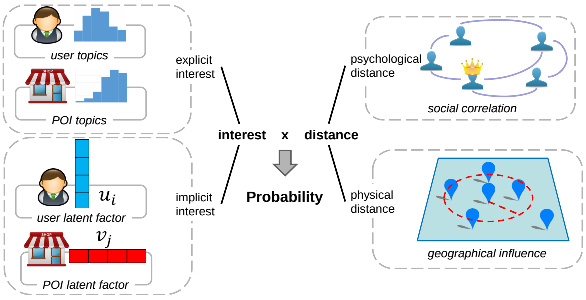


Fig. 1. The Architecture Framework of CoSoLoRec model

In summary, we have made several contributions in this paper:

- We propose a novel probabilistic latent factor model for heterogeneous Point-of-Interest recommendation by simultaneously considering the geographical location, social correlation and the user preference.
- Our model can keep its robustness due to its modularization. It means every heterogeneous information is embedded into the model as a module thus removing anyone of them won't affect the praticality but only decline the performance somewhat.
- We conduct extensive experiments on two real-world large scale datasets to evaluate both the efficiency and the effectiveness of our model. The results demonstrate that our approach outperforms other state-of-the-art methods.

The rest of the paper is organized as follows. We review the recent studies in Section 2. Problem is defined in Section 3.1 along with the model formation in Section 3.2 and the model description in Section 4. The experiment is demonstrated in Section 5. We conclude our work in Section 6.

2 Related Work

In recent years, POI recommendation has grown in popularity with the increasing demand of LBSNs. In reality, geographical information plays critical roles in influencing user's behaviors[5][4] since physical interactions are required in LBSNs which differs from other non-spatial situations totally. To exploit influence of geographical information in improving the quality of location recommendations, Some techniques model the distance between two locations visited by a user as a distribution for all users. In [4], Ye *et al.* employed power-law distribution (PD) to model user's checkin behaviors using naive bayesian method. Instead of making PD assumption, Cheng *et al.*[5] proposed the Multi-center Gaussian Model (MG-M) to capture features of user's checkin behaviors. However, the geographical influence on individual user's visiting behaviors should be personalized rather than appearing as a common distribution. So Zhang *et al.*[17] used kernel density estimation (KDE) to model geographical influence as personalized distance distribution for each user. In our work, we adopt KDE to model geographical factor since its superior in personalized modeling.

According to homophily of social correlation, friends tend to share common interests which will make recommendation more accurate and efficient. H. Ma proposed social trust concept[12] summarized as friend-based Collaborative Filtering (CF) to explain social influence. In [13], H.Tong proposed Random Walk with Restart(RWR) to capture social relations. In this paper, we adopt friend-based CF since its lower computation cost and more accuracy.

Exploring text information can also better understand patterns in LBSNs and improve LBSNs services. Ye *et al.*[24] explored explicit patterns of individual places and implicit revelance among similar places through semantic annotation. Liu *et al.*[9] proposed TL-PMF model to consider both the extent to which a user interest matches the POI in terms of topic distribution and the word-of-mouth opinion of the POI. Kurashima *et al.*[19] proposed Geo-Topic Model which jointly estimates user's interests and activity areas. Yin[20] proposed LCA-LDA model by giving consideration to both personal interest and local preference. Zheng *et al.*[22] proposed a cross-region collaborative filtering method based on hidden topics about check-in records to recommend new POIs. Zhang *et al.*[23] distinguished the user preferences on the content of POIs from the POIs themselves and combined the predicted rating on content and location of POI.

Previous studies mainly focused on just one or two aspects which affects user's visiting behaviors. However, overall consideration on the joint effects of heterogeneous information can show great superiority in POI recommendation. Ye *et al*[4] incorporated social and geographical influences into user-based CF framework by using linear interpolation. In [5], Cheng *et al.* considered geographical influence, check-in patterns, frequency and social networks in POI recommendation. Hu and Ester[6] considered spatial and textual aspects of posts published by mobile users and predicted user's willing locations. However, this work is more similar to a location prediction problem rather than POI recommendation. B.Liu[10] proposed the Geo-BNMF model to embrace geographical influence, popularity, text information and latent factor model. Based on this, They also proposed Geo-PFM[11] to combine geographical influence, latent factor model with latent region. Lian *et al.*[21] proposed a weighted matrix factorization method incorporating the modeling of the spatial clustering phenomenon. In this paper, we jointly fused heterogeneous information with latent factor model to describe users' behaviors and make our recommendation more efficient and accurate.

3 Fusion Model with Heterogeneous Information

In this section, we define the problem of POI recommendation and introduce a fused probabilistic latent factor framework for heterogeneous information.

3.1 Problem Definition

The major task for POI recommendation is to recommend POIs which a user has not visited using heterogeneous information in LBSNs. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_i, \ldots, u_m\}$ be the set of users and $\mathcal{V} = \{v_1, v_2, \ldots, v_j, \ldots, v_n\}$ be the set of POIs. Each user u_i visited some POIs L_i historically and rated on these POIs.

User's visiting behavior may be not only influenced by geographical distance between his destination and visited POIs but also influenced by ratings made by user's friends F_i. We regard these factors as geographical influence and social influence which can also be regarded as physical distance and psychological distance respectively. Text information such as reviews may also reflects explicit interest of user u_i with user-topic distribution $\boldsymbol{\theta_i}$ in topic model.

The problem under investigation is essentially how to effectively and accurately estimate user's probability in visiting new POIs by employing information containing above three aspects. To attack this problem, in Section 3.2, We formulate a fused probabilistic latent factor model by incorporating these factors.

As for POIs, each POI v_j has its own location l_j labeled as vector $< lon_j, lat_j >$ in representing latitude and longitude respectively. Also each POI has its own textual profiles with its topic distribution $\boldsymbol{\pi_j}$ in topic model. For convenience, we term i and j as user u_i and location l_j respectively.

3.2 Fused Probabilistic Latent Factor Model

To make CoSoLoRec model concrete, we assume the follow factor representation: (1) each user i is associated with his or her interest $\eta\,(i, j)$ with respect to POI j. (2) each user has an intended visiting probability $p_f\,(i, j)$ with respect to POI j on the basis of friend-based CF. (3) geographical influence impels user

Table 1. Mathematical Notations

Symbol	Size	Meaning
$\mathcal{U}$	m	Set of all users in one LBSN
$\mathcal{V}$	n	Set of all POIs in one LBSN
L_i	$\lvert L_i \rvert$	Set of locations visited by a user
X_i	$\binom{\lvert L_i \rvert}{2}$	sample of distances between L_i
$\mathcal{F}_i$	$\lvert \mathcal{F}_i \rvert$	friends of user i
F	$m \times n$	a predicted $m \times n$ data matrix
$\boldsymbol{u_i}$	d	preference of user i
$\boldsymbol{v_j}$	d	affinity of POI j
$\boldsymbol{\theta_i}$	K	user i's review topic distribution
$\boldsymbol{\pi_j}$	K	POI j's review topic distribution

i to estimate the probability he or she will visits POI j denoted as $p_l\left(i,j\right)$. We integrally consider user's interest, physical distance and psychological distance. Finally we got a joint model with these three factors:

$$p\left(i,j\right) \propto \eta\left(i,j\right)\left(\left(1-\lambda\right)p_l\left(i,j\right)+\lambda p_f\left(i,j\right)\right) \tag{1}$$

The recommend process of user i for POI j can be represented in a generative way. For user's preference, $\eta\left(i,j\right)$ can be represented as a linear combination of latent factor $\mathbf{u_i^T v_j}$ and function of user's and POI's observable properties which can be expressed as topic distribution of user i and POI j named as $\boldsymbol{\theta_i}, \boldsymbol{\pi_j}$. We denote these two parts as implicit interests and explicit interests of a user respectively. We use $\eta_1\left(i,j\right)$ and $\eta_2\left(i,j\right)$ to notate them. Also, user's rating $y\left(i,j\right)$ can be influenced by his visiting probability. Here we adopt Possion distribution to describe this relation. So fused probabilistic latent factor model can be expressed as follows:

1. Draw a user interest
   (a) Generator user latent factor $u_{iw} \sim Gamma(\alpha_U, \beta_U)$
   (b) Generator item latent factor $v_{jw} \sim Gamma(\alpha_V, \beta_V)$
   (c) user's explicit interest $\eta_1\left(i,j\right)=\boldsymbol{\theta_i^T}\boldsymbol{\pi_j}$, implicit interest $\eta_2\left(i,j\right)=\mathbf{u_i^T v_j}$
   (d) user's interest $\eta\left(i,j\right)=\eta_1\left(i,j\right)+\eta_2\left(i,j\right)$
2. $y\left(i,j\right) \sim P\left(p\left(i,j\right)\right)$ where
   $p\left(i,j\right)=\left(\eta_1\left(i,j\right)+\eta_2\left(i,j\right)\right)\left(\left(1-\lambda\right)p_l\left(i,j\right)+\lambda p_f\left(i,j\right)\right)$

4 Model Specification

In this section, we introduce detailed model specifications and present our fusion model called **CoSoLoRec**.

4.1 Geographical Influence

We aim to exploit geographical Influence by measuring the distance from a user's visited POIs to an unvisited POI. Thus we employ Kernel Density Estimation (KDE) to model the geographical influence of POIs on users' visiting behaviors.

Like MGM, KDE is also a widely-adopted method to estimate geographical influence. What's more, it shows superior to other methods which model geographical influence in considering visited POIs. We can evaluate general influence of all POIs using the following method:

$$p_l\left(i,j\right)=P\left(\bigcup_{t=1}^{\lvert L_i \rvert}\left(c_t \rightarrow c_0\right)\right)=1-P\left(\bigcap_{t=1}^{\lvert L_i \rvert}\overline{c_t \rightarrow c_0}\right)=1-\prod_{t=1}^{\lvert L_i \rvert}\left(1-P\left(c_t \rightarrow c_0\right)\right) \tag{2}$$

From the equation (2), we can discover that before fetching geographical influence of locations $p_l\left(i,j\right)$, our task is to learn the probability that event $c_t \to c_0$ occurs. Here we use the same algorithm proposed in [17] to learn it.

$$P\left(c_t \to c_0\right) = \frac{1}{|X_i|}\sum_{x\in X_i} K\left(\frac{z_t - x}{\delta}\right) = \frac{1}{\sqrt{2\pi}|X_i|}\sum_{x\in X_i} e^{-\frac{(z_t - x)^2}{2\delta^2}} \tag{3}$$

Here, z_t is the distance between user visiting POI c_0 and each of user's historical POIs, which can be used to derive the probability of c_0. $K(\cdot)$ represents kernal function. Also δ is a smoothing parameter which is called the bandwidth. We use optimal bandwidth[3] $\delta \approx 1.06\hat{\delta}|X_i|^{-1/5}$. However, the computational complexity grows rapidly with the increment of L_i. So we use efficient approximation algorithm[17] to measure $p_l\left(i,j\right)$.

Eventually, by combining equation (2) and (3), we can exploit *geographical influence of locations* by using $p_l\left(i,j\right)$. However, only the geographical information is not sufficient. Thus social correlation is introduced.

4.2 Friend-based Collaborative Filtering

With the exponential growth of online social network, social relationship plays an important role in influencing users' behaviors. Friends usually have similar behaviors due to the phenomenon that sociologists call homophily[4].

Aiming to predict the probability of user i to a POI j, we adopt the user-based collaborative filtering(CF) by regarding all of i's friends as neighbors. In order to determine the probability in interval $[0,1]$, We devise the calculation as:

$$p_f\left(i,j\right) = \frac{\sum_{i'\in\mathcal{F}_i} sim\left(i,i'\right) r_{i'j}}{\sum_{i'\in\mathcal{F}_i} sim\left(i,i'\right)} \cdot \frac{1}{r_{max}} \tag{4}$$

Here, r_{max} can ensure $p_f\left(i,j\right)$ is normalized. $sim\left(i,j\right)$ refers to similarity between user i and user j. In our study, we choose cosine similarity.

4.3 Probabilistic Latent Factor Model

Probabilistic Factor Model (PFM)[14] is a generative probabilistic model. The notations involved in PFM are defined in Table 1. Here, $\hat{f_{ij}}$ is assumed to follow Possion Distribution, the mean is $\hat{y_{ij}}$. Also, $\boldsymbol{u_i}$ and $\boldsymbol{v_j}$ are given certain distributions as priors. Here, $\boldsymbol{u_i} = \left(u_{i1}, u_{i2}, ..., u_{iw}, ..., u_{id}\right)$ and $\boldsymbol{v_j} = \left(v_{j1}, v_{j2}, ..., v_{jw}, ..., v_{jd}\right)$.

Therefore, the process of Probabilistic Factor Model is as follows:

1. for all w, generate $u_{iw} \sim p\left(u_{iw}|\mathit{\Phi}_{u_{iw}}\right)$
2. for all w, generate $v_{jw} \sim p\left(v_{jw}|\mathit{\Phi}_{v_{jw}}\right)$
3. generate $\hat{y_{ij}}$ from user i to location j with equation $\hat{f_{ij}} = \sum\limits_{w=1}^{d} u_{iw}v_{jw} = \boldsymbol{u_i v_j}$
4. generate $\hat{y_{ij}} \sim P\left(\hat{f_{ij}}\right)$

Here, $\mathit{\Phi}_{u_{iw}}$ and $\mathit{\Phi}_{v_{jw}}$ are hyperparameter lists respect to u_{iw} and v_{jw}. We assume latent factors are non-negative in real situations. So u_{iw} and v_{jw} are given Gamma distributions as empirical priors[14][15]. The gamma distribution of U and V can be represented as functions:

$$p\left(U|\alpha_U,\beta_U\right) = \prod_{i=1}^{m}\prod_{w=1}^{d} \frac{{u_{iw}}^{\alpha_U - 1}\exp\left(-u_{iw}/\beta_U\right)}{{\beta_U}^{\alpha_U}\mathit{\Gamma}\left(\alpha_U\right)} \tag{5}$$

$$p\left(V|\alpha_V,\beta_V\right) = \prod_{j=1}^{n}\prod_{w=1}^{d} \frac{{v_{jw}}^{\alpha_V - 1}\exp\left(-v_{jw}/\beta_V\right)}{{\beta_V}^{\alpha_V}\mathit{\Gamma}\left(\alpha_V\right)} \tag{6}$$

where $u_{iw}, v_{jw}, \alpha_U, \beta_U, \alpha_V, \beta_V$>0, $\mathit{\Gamma}\left(\boldsymbol{.}\right)$ is the Gamma function. Given user latent factor $\boldsymbol{u_i}$ and item latent factor $\boldsymbol{v_j}$, the The possion distribution of y_{ij} given f_{ij} is

$$P\left(\hat{y_{ij}}|\hat{f_{ij}}\right) = \left(\boldsymbol{u_i v_j}\right)^{\hat{y_{ij}}} \frac{\exp\left(-\boldsymbol{u_i v_j}\right)}{\hat{y_{ij}}!} \tag{7}$$

Expressed in matrix of the above equation with $F = UV^T$. With the method of maximum a posterior(MAP), posterior distribution of Y can be modeled as

$$p\left(U, V|Y, \alpha_U, \beta_U, \alpha_V, \beta_V\right) \propto P\left(Y|F\right) p\left(U|\alpha_U, \beta_U\right) p\left(V|\alpha_V, \beta_V\right) \tag{8}$$

Finally, we can infer parameters with stochastic gradient descent (SGD) method.

4.4 Textual Analysis

In order to extract users' explicit interest, we use an aggregated LDA model. The model has two latent variables with corresponding super parameters as priors: (1) document-topic distributions $\mathit{\Theta}$. (2) topic-word distributions $\mathit{\Phi}$. In order to learn users' interests, we aggregate all the reviews written by each user into a document. Thus, user and document are interchangable in reflecting user's interest.

In this way, we build an aggregated Topic Model. Each user in LBSN is associated with topics following a multinomial distribution, denoted as $\boldsymbol{\theta}$. Also, each topic is associated with textual items according to a multinomial distribution. As we have sampled $\mathit{\Theta}$ and $\mathit{\Phi}$ of Topic Model of users. Obviously, the dimension for document-topic distribution of both Topic Model should be the same. Using gibbs sampling, the topic distribution for POI j and document-topic distribution θ of users is:

$$\pi_{js} = \frac{n_j^{(s)} + \alpha}{\sum\limits_{s=1}^{K} n_j^{(s)} + K\alpha} \qquad \theta_{is} = \frac{n_i^{(s)} + \alpha}{\sum\limits_{s=1}^{K} n_i^{(s)} + K\alpha} \tag{9}$$

Here, $n_j^{(s)}$ is the topic observation count for POI j. $n_i^{(s)}$ is the topic observation counts for user i(document d). V and K are the number of unique words and topics. α and β are hyperparameters in corresponding to topic model.

4.5 Learning and Inference

Parameter Estimation Let us denote all parameters as $\mathit{\Lambda} = \{U, V\}$ and let $\mathit{\Omega} = \{\alpha_U, \beta_U, \alpha_V, \beta_V\}$ be the hyperparameters. Hyperparameters are all apriori given. Given the observed data collection $\mathcal{D} = \{p\left(i, j\right)\}^{I_{ij}}$ where $p\left(i, j\right)$ is the user visiting probability, and $I_{ij} = 1$ when user i visited POI j, and $I_{ij} = 0$ otherwise.

To estimate the parameters $\mathit{\Lambda}$, we use maximum likelihood estimation (MLE) method and sampling algorithm to learn all the parameters. So the postprior probability can be expressed as the following:

$$P\left(\mathit{\Lambda}|\mathcal{D}, \mathit{\Omega}\right) \propto \prod_{\mathcal{D}} P\left(y\left(i, j\right)|U, V, \mathit{\Omega}\right)^{I_{ij}} \times P\left(U|\alpha_U, \beta_U\right) P\left(V|\alpha_V, \beta_V\right) \tag{10}$$

For simplicity, we use logarithmic form of posterior distribution instead. We express this as follows:

$$\begin{aligned} \mathcal{L}\left(U, V; \mathcal{D}\right) = \sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}\left(y\left(i, j\right)\log p\left(i, j\right) - p\left(i, j\right)\right) + \sum_{i=1}^{M}\sum_{w=1}^{d}\left(\left(\alpha_U - 1\right)\log u_{iw} - \frac{u_{iw}}{\beta_U}\right) \\ + \sum_{j=1}^{N}\sum_{w=1}^{d}\left(\left(\alpha_V - 1\right)\log v_{jw} - \frac{v_{jw}}{\beta_V}\right) \end{aligned} \tag{11}$$

Here, $p\left(i,j\right)=\left(\boldsymbol{u_i^T}\boldsymbol{v_j}+\boldsymbol{\theta_i^T}\boldsymbol{\pi_j}\right)\left(\left(1-\lambda\right)p_l\left(i,j\right)+\lambda p_f\left(i,j\right)\right)$ In order to approximate actual value of $\boldsymbol{u_i}$ and $\boldsymbol{v_j}$, we use stochastic gradient descent (SGD) method to optimize them and update parameters iteratively using all training samples, ξ_i and ξ_j are learning rates.

$$u_{iw}\leftarrow u_{iw}+\xi_i\times\frac{\partial\mathcal{L}}{\partial u_{iw}}\qquad v_{jw}\leftarrow v_{jw}+\xi_j\times\frac{\partial\mathcal{L}}{\partial v_{jw}}\tag{12}$$

4.6 Recommendation

After we learn the parameters Λ, CoSoLoRec model predicts the ratings of a user for a given POI using $\mathbb{E}\left(y\left(i,j\right)\right)=\left(\mathbf{u_i^T}\mathbf{v_j}+\boldsymbol{\theta_i^T}\boldsymbol{\pi_j}\right)\left(\left(1-\lambda\right)p_l\left(i,j\right)+\lambda p_f\left(i,j\right)\right)$. We adjust hyperparameters in training process and adjust parameter λ to balance physical and psychological influence in making decisions. Our model learns the latent factors by SGD effectively. Therefore we make POI recommendation via our trained model.

5 Experiment

In this section, we conduct several experiments based on CoSoLoRec model and baselines to evaluate the performance of our proposed approach empirically. All experiments are conducted on two real-world datasets in LBSNs, collected from Yelp and Foursquare.

5.1 Dataset

Yelp dataset. Yelp is a famous website which provides reviews and ratings for restaurants and other business places[16]. To make the experimental results more convinced, we manually choose two cities named “Phoenix” and “Las Vegas” which consist of 80% of original datasets to evaluate our approaches. We filter out users who have more than 300 friends to avoid spam generated by brushed from robots and less than 20 friends to make our datasets dense. With the same reason, we filter out users who have more than 300 reviews and less than 20 reviews. Finally, we obtain a dataset consists of 3059 users, 26446 business along with 180755 review records.

Foursquare dataset. Besides the Yelp dataset, we also evaluate our models on Foursquare (4sq). The dataset includes POIs distributed in the United States. With the same reason in handling Yelp dataset, we filter out users with more than 500 friends and less than 18 friends. Similarly, we filter our ratings given by more than 500 users and less than 20 users. We finalize a dataset of 6895 users for 13208 POIs with 166989 ratings. Table 2 indicates the data statistics for Yelp and Foursquare.

Table 2. Data Description

	Yelp	Foursquare
Number of users	366715	571700
Number of locations	61184	8318919
Review items	1569265	5550203
User-location matrix density	6.99×10^{-5}	1.17×10^{-6}
Number of Cities	10	50

To unify ratings with probability to visit a POI, we normalize the discrete rating using $f\left(x\right)=\frac{x}{max\{x\}}$, where $max\{x\}$ represents the largest rating value[7].

5.2 Evaluation Metrics

We present each user with N POIs sorted by the predicted probability and evaluated based on which of these POIs were actually visited by users.

rPrecision To unify evaluation of a universal baseline and baseline with region-based attached, we introduce relative precision for evaluation. We assume C as the candidate POIs. The precision in a top-K list of a random recommendation system is $\frac{|S_{visited}|}{|C|}$. Then, the relative precision[10][25] is defined as: $rPrecision@N = \frac{|S_{N,rec} \cap S_{visited}| \cdot |C|}{|S_{visited}| \cdot N}$

RMSE We also use Root Mean Square Error(RMSE) for evaluation. $RMSE = \sqrt{\frac{1}{N} \sum_{(u,i) \in E} \left(r_{ui} - \hat{r_{ui}}\right)^2}$. where E is the test dataset. r_{ui} and $\hat{r_{ui}}$ represent the observed and predicted performance used for user i on business j respectively. Smaller value of RMSE implies better performance in our recommendation.

5.3 Baseline Comparison

In this section, we introduce baselines and parameters involved in experiments and evaluate our model with baselines in a number of experiments.

Comparative Approaches In this section, in order to show the effectiveness of our CoSoLoRec model, we compare our model with the following baselines:

①**Probabilistic Matrix Factorization (PMF)**. PMF [7] is a recommendation method widely used for different recommendation tasks.

②**Non-negative Matrix Factorization (NMF)**. NMF [8] is a method used in recommender system which constrains the factors to be non-negative.

③**Bayesian Non-Negative Matrix Factorization (BNMF)**. BNMF [18] is an unsupervised learning method using Markov chain Monte Carlo sampling method based on Non-negative Matrix Factorization.

④**Geographical-Topical Bayesian Nonnegative Matrix Factorization (GT-BNMF)**[9]. This is a new method combining geographical information, textual information with other aspects based on BNMF.

⑤**Geographical Probabilistic Factor Model (Geo-PFM)**[11]. This is a fusion method using latent factor model considering geographical information with no textual information and social correlation.

To further understand the benefits using different forms of implicit interest and distinction between implicit interest and explicit interest, we implement three modifications of CoSoLoRec model.

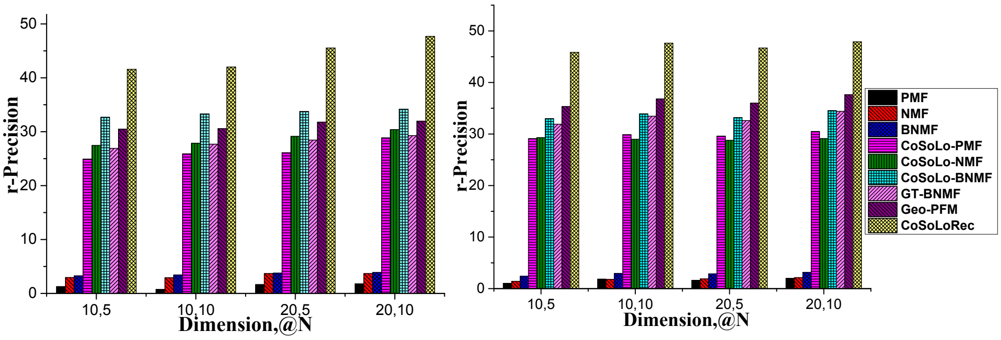
⑥**CoSoLo-PMF** is a modification using PMF to depict implicit interest.

⑦**CoSoLo-NMF** is a modification using NMF to depict implicit interest.

⑧**CoSoLo-BNMF** is a modification using BNMF to depict implicit interest.

We divide data as training set and test set on the ratio of 7:3 with the review time order. For CoSoLoRec model, we set $\alpha_U = 5$ and $\beta_U = 0.2$ as hyper parameters of prior of U. Also, we set $\alpha_V = 20$ and $\beta_V = 0.2$ as hyper parameters of prior of V. We initialize $\lambda = 0.5$ for comparing our model with baselines. For PMF, we set hyper parameters $\sigma_U = 0.05$ and $\sigma_V = 0.05$ as priors of U and V. Also, we set $\sigma = 0.2$ as prior of $y\left(i, j\right)$. For textual aspects, we initialize the topic number $K = 30$, $\alpha = 40/K$, $\beta = 0.3$. We set the number of regions $|R| = 50$ for foursquare which is the number of regions according to all the states in USA(expect Alaska). For yelp, we set $|R| = 2$ for data from two cities.

Expermental Results To see how the models actually outperform a universal baseline and baseline with region-based attached, we use relative performance. Figure 2(a) and 2(b) indicates CoSoLoRec model outperforms all the basslines



(a) rPrecision@N on Yelp

(b) rPrecision@N on Foursquare

Fig. 2. r-Precision with different latent factor dimension and @N

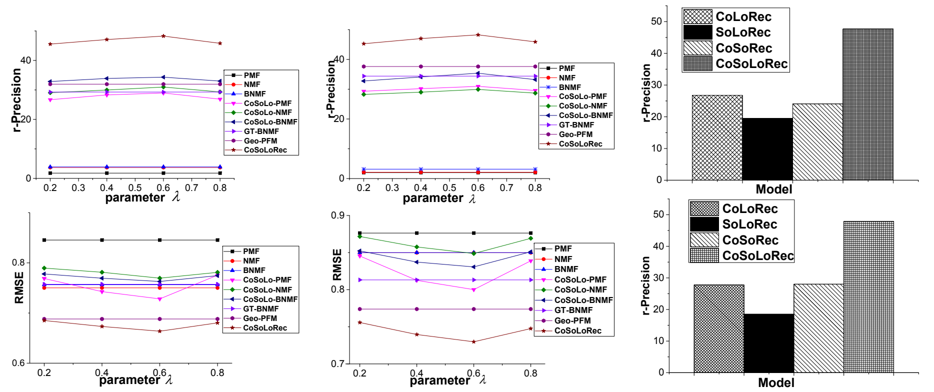
including classical baseline models (PMF,NMF,BNMF) as well as recent proposed model (GT-BNMF,Geo-PFM) in both datasets. Furthermore, CoSoLo-PMF, CoSoLo-NMF and CoSoLo-BNMF show almost equivalent performance in precision. This phenomenon indicates heterogeneous information can reflect user's interests accurately. GT-BNMF model considers heterogeneous information such as geographical information, popularity and user's interests. However, our model considers further more about these factors. Geo-PFM model use Possion factor model which can guarantees a rigorous probabilistic generative process. However, our model performs better since heterogeneous In order to measure the difference between estimated rating values and real rating values in test datasets, RMSE is introduced. We conduct experiments on different latent factor dimensions. From table 3, we can conclude that our model achieves less mean square error than baselines with different data divisions. So it is obvious that our model is more accurate in recommendation.

Table 3. Performance Comparison in different dimensions

	D	Metrics	PMF	NMF	BNMF	⑥	⑦	⑧	GT-BNMF	Geo-PFM	CoSoLoRec
Yelp	10	RMSE	0.8225	0.7644	0.766	0.7639	0.7824	0.7769	0.7241	0.7076	**0.6692**
		Improve	18.64%	12.45%	12.64%	12.40%	14.47%	13.86%	7.58%	5.43%	
	20	RMSE	0.8455	0.7502	0.7564	0.7365	0.7716	0.7672	0.7573	0.6881	**0.6693**
		Improve	20.84%	10.78%	11.52%	9.12%	13.26%	12.76%	11.62%	2.73%	
4sq	10	RMSE	0.8792	0.8515	0.8624	0.8335	0.8612	0.8454	0.8282	0.7815	**0.7476**
		Improve	14.97%	12.20%	13.31%	10.31%	13.19%	11.57%	9.73%	4.34%	
	20	RMSE	0.8763	0.8498	0.85	0.8019	0.8509	0.8334	0.8132	0.7739	**0.7319**
		Improve	16.48%	13.87%	13.89%	8.73%	13.99%	12.18%	10.00%	5.43%	

5.4 Parameters Sensitivity

As mentioned, both geographical influence and social influence play important roles in estimating user's interest on unvisited POIs. So in the following part, we respectively set λ as 0.2, 0.4, 0.6, 0.8 to detect importance of geographical and social factors. From Figure 3(a) and 3(b), we can find that rPrecision rises first and falls later while RMSE shows a reverse trend. Since PMF,NMF,BNMF,GT-BNMF and Geo-PFM do not consider relation between geographical distance and psychological distance, results show no change with different parameter λ. We can find both geographical and social influence play comparative roles. Our model outperforms baselines in every value of λ while three modifications show almost equivalent performance with Geo-PMF and GT-BNMF.



(a) r-Pre and RMSE on Yelp (b) r-Pre and RMSE on 4sq (c) Robust for Yelp and 4sq

Fig. 3. Experimantal Results in Parameter Sensitivity and Model Robust

5.5 Impact of Geographical, Social and Textual information

In some situations, it is ideal that all aspects are covered in our model. So it is necessary to evaluate our model if not all the data is present. In this part, we choose three models which are based on CoSoLoRec model: 1)**CoLoRec:** social correlation removed. 2)**CoSoRec:** geographical factor removed. 3)**SoLoRec:** textual information removed. All the experiments are conducted in $K = 20, N = 10$.

Figure 3(c) shows results comparing the above three models with CoSoLoRec model. We conclude that above three models perform worse than CoSoLoRec model since user preference can not be completely described if one of factors in CoSoLoRec removed. However, relevant indicators show not much decline comparing with CoSoLoRec model. In particular, CoLoRec and CoSoRec models show not much decline compared to SoLoRec model which shows users' text information contributes greater than geographical factor and social correlation.

6 Conclusion

In this paper, we proposed CoSoLoRec model fusing heterogeneous information like geographical factor, social correlation and text information. We incorporate user preference with geographical factor realized by KDE and social correlation realized by friend-based CF. Further more, we devide user interest into explicit interest implemented by Topic Model and implicit interest implemented by probabilistic latent factor model. Experimental result conducted with Yelp and foursquare dataset demonstrated that CoSoLoRec model is superior to all other approaches evaluated, such as PMF, NMF, GT-BNMF and Geo-PFM and three different forms of CoSoLoRec model. Also we can conclude text information is more important than geographical factor and social correlation. Our model performs better in any different combinations between geographical and social influence.

References

1. Tobler, Waldo R.:"A computer movie simulating urban growth in the Detroit region." Economic geography (1970): 234-240.
2. McPherson, Miller, Lynn Smith-Lovin, and James M. Cook.:"Birds of a feather: Homophily in social networks." Annual review of sociology (2001): 415-444.
3. Dehnad, Khosrow.:"Density estimation for statistics and data analysis." Technometrics 29.4 (1987): 495-495.

4. Ye, Mao, et al.:"Exploiting geographical influence for collaborative point-of-interest recommendation." Proceedings of ACM SIGIR. ACM, 2011.
5. Cheng, C,Yang, H.,et al.:"Fused matrix factorization with geographical and social influence in location-based social networks." AAAI. 2012
6. Hu, Bo, and Martin Ester.:"Spatial topic modeling in online social media for location recommendation." Proceedings of the 7th ACM Recsys. ACM, 2013.
7. Mnih, Andriy, and Ruslan Salakhutdinov.:"Probabilistic matrix factorization." NIPS. 2007.
8. Lee, Daniel D., and H. Sebastian Seung.:"Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788-791.
9. Liu, Bin, and Hui Xiong.:"Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness." SDM. Vol. 13. 2013.
10. Liu, Bin, et al.:"Learning geographical preferences for point-of-interest recommendation." Proceedings of ACM SIGKDD. ACM, 2013.
11. Liu, Bin, et al.:"A general geographical probabilistic factor model for point of interest recommendation." TKDE 2015: 1167-1179.
12. Ma, Hao, Michael R. Lyu, and Irwin King.:"Learning to recommend with trust and distrust relationships." Proceedings of ACM Recsys. ACM, 2009.
13. Tong, Hanghang, Christos Faloutsos, and Jia-Yu Pan.:"Fast random walk with restart and its applications." Proc. IEEE ICDM, IEEE Computer Society, 2006.
14. Ma, Hao, et al.:"Probabilistic factor models for web site recommendation." Proceedings of ACM SIGIR. ACM, 2011.
15. Chen, Ye, et al.:"Factor modeling for advertisement targeting." NIPS. 2009.
16. Anderson, Michael, et al.:"Learning from the crowd: Regression discontinuity estimates of the effects of an online review database*." The Economic Journal 122.563 (2012): 957-989.
17. Zhang, Juyong, Chi-Yin Chow, et al. "iGeoRec: A personalized and efficient geographical location recommendation framework." Services Computing, IEEE Trans on, IEEE, 2015.
18. Schmidt, Mikkel N and Winther, Ole and Hansen, Lars Kai. "Bayesian non-negative matrix factorization" Springer, 2009
19. Kurashima, Takeshi and Iwata, Tomoharu and Hoshide, Takahide and Takaya, Noriko and Fujimura, Ko. "Geo topic model: joint modeling of user's activity area and interests for location recommendation" Proc. ACM WSDM, ACM, 2013
20. Yin, Hongzhi and Sun, Yizhou and Cui, Bin and Hu, Zhiting and Chen, Ling. "Lcars: a location-content-aware recommender system" Proc. ACM SIGKDD, ACM, 2013
21. Lian, Defu and Zhao, Cong and Xie, Xing and Sun, Guangzhong and Chen, Enhong and Rui, Yong. "Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation" Proc. ACM SIGKDD, ACM, 2014
22. Zheng, Ning and Jin, Xiaoming and Li, Lianghao. "Cross-region collaborative filtering for new point-of-interest recommendation" Proc. WWW companion, 2013
23. Zhang, Chenyi and Wang, Ke. "POI recommendation through cross-region collaborative filtering" KIS, Springer, 2016
24. Ye, Mao and Shou, Dong and Lee, Wang-Chien, et al. "On the semantic annotation of places in location-based social networks" Proc. ACM SIGKDD, ACM, 2011
25. Yin, Peifeng and Luo, Ping and Lee, Wang-Chien and Wang, Min. "App recommendation: a contest between satisfaction and temptation" ACM WSDM, ACM, 2013
26. Zhou, Dequan and Wang, Bin and Rahimi, Seyyed Mohammadreza and Wang, Xin "A study of recommending locations on location-based social network by collaborative filtering" Advances in Artificial Intelligence, Springer, 2012
27. Noulas, Anastasios and Scellato, Salvatore and Lathia, Neal and Mascolo, Cecilia "A random walk around the city: New venue recommendation in location-based social networks" PASSAT and SocialCom, IEEE, 2012