Neurocomputing 373 (2020) 56–69

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Where to go: An effective point-of-interest recommendation framework for heterogeneous social networks

Xi Xiong[a], Shaojie Qiao[b], Nan Han[c,*], Fei Xiong[d], Zhan Bu[e], Rong-Hua Li[f], Kun Yue[g], Guan Yuan[h]

[a] School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China
[b] School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China
[c] School of Management, Chengdu University of Information Technology, Chengdu 610103, China
[d] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[e] College of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210023, China
[f] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[g] School of Information Science and Engineering, Yunnan University, Kunming 650500, China
[h] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

ARTICLE INFO

Article history:
Received 21 March 2019
Revised 23 July 2019
Accepted 14 September 2019
Available online 25 September 2019

Communicated by Dr. Weike Pan

Keywords:
Location-based social networks
POI recommendation
Heterogeneous networks
Probabilistic graphical model

ABSTRACT

Point-of-Interest (POI) recommendation is one of the most essential tasks in LBSNs to help users discover new interesting locations, especially when users travel out of town or to unfamiliar areas. Current studies on POI recommendation in LBSNs mainly focus on modeling multiple factors extracted from users' profiles and checking-in records. Data sparsity and incompleteness of user-POI interaction matrix are very common problems in POI recommendation, especially for the out-of-town scenario. Another challenge is that most information in the LBSNs is unreliable due to users' different backgrounds or preferences. Because of the close relationship between users, information from trustable friends on Communication-Based Social Networks (CBSNs) is more valuable than that in LBSNs, which can give a preferable suggestion instead of trustless reviews in LBSNs. In this study, we propose a latent probabilistic generative model called HI-LDA (Heterogeneous Information based LDA), which can accurately capture users' words on CBSNs by taking into full consideration the information on LBSNs including geographical effect as well as the abundant information including social relationship, users' interactive behaviors and comment content. In particular, the parameters of the HI-LDA model can be inferred by the Gibbs sampling method in an effective fashion. Beyond these proposed techniques, we introduce an POI recommendation framework integrating geographical clustering approach considering the locations and popularity of POIs simultaneously. Extensive experiments were conducted to evaluate the performance of the proposed framework on two real heterogeneous LBSN-CBSN networks. The experimental results demonstrate the superiority of HI-LDA on effective and efficient POI recommendation in both home-town and out-of-town scenarios, when compared with the state-of-the-art baseline approaches.

1. Introduction

Point-of-Interest (POI) recommendation is one of the most essential tasks in Location-Based Social Networks (LBSNs) to help users discover new interesting places based on the pervasive application of mobile information technology and location services [1]. The most popular LBSNs, e.g., Foursquare[1] and Yelp,[2] can help users find favorite places, i.e., hotels and restaurants, according to their historical records and preferences [2]. POI recommendation can benefit advertising agencies in an effective way of launching advertisements to the potential consumers, and improve user viscosity (i.e., improve the possibility of consumption) to LBSN service providers [3]. In general, users often post subjective reviews and share experience with their friends in these LBSN applications. The LBSN applications become more necessary when a user travels to an unfamiliar area, where he has little knowledge of local environment.

One of the most important problems for POI recommendation in LBSNs is *the unreliable data in LBSNs*, especially for travelers.

* Corresponding author.
E-mail addresses: sjqiao@cuit.edu.cn (S. Qiao), hannan@cuit.edu.cn (N. Han).

[1] https://foursquare.com/
[2] https://www.yelp.com

https://doi.org/10.1016/j.neucom.2019.09.060
0925-2312/

Table 1
Comparison of the percentage that a visitor's friends in Foursquare and Facebook visits the restaurant in the following month.

City	Restaurant	Fs[a]	Fb[b]
San Francisco (US)	Neighbor Bakehouse	9.4%	15.2%
	Pushkin	8.9%	15.7%
	Kuma Sushi + Sake	8.4%	10.6%
	Game of Burger	8.8%	13.0%
Istanbul (Turkey)	Match Cafe	8.7%	13.1%
	Beygah	8.1%	10.7%
	Free State Kitchen	10.7%	16.5%
Liverpool (UK)	Baltic Bakehouse	9.2%	14.6%
	Leaf	10.0%	13.2%

[a] Fs is short for Foursquare.
[b] Fb is short for Facebook.

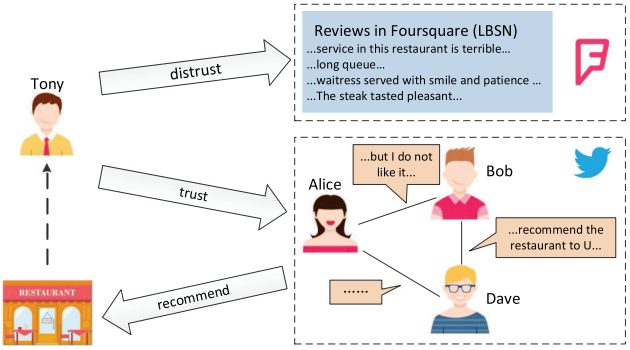


Fig. 1. Example of users' different attitudes towards words in LBSNs and CBSNs.

When a user asks for a recommendation in LBSNs, most of users in the LBSNs cannot give accurate suggestions due to their different backgrounds or preferences. For instance, a person in favor of spicy food travels to a seaside city will probably be recommended of seafood via LBSNs, which may be not suitable for him.

By means of our designed crawler, we acquired from Foursquare the reviews and the check-in data of the top-3 popular restaurants in San Francisco, Istanbul and Liverpool, respectively, during May to June 2017. Some visitors made a review or a check-in after a visit to a restaurant. For each popular restaurant, 200 active visitors that posted reviews in May were selected. Only a small fraction of their friends checked in the same POI in the next month. The third column of Table 1 illustrates the percentage of his friends that checked in the restaurant during June after his visits in May. We find that the textual reviews from visitors in LBSNs have limited influence on the following decisions of their friends. The distrust between individuals in LBSNs degrades the accuracy of POI recommendation.

Then we extracted the comments about the restaurants from the textual content in Facebook, which indicates that the individuals have visited the places. Similar to the third column of Table 1, the last column represents the percentage of his friends in Facebook that visited the restaurant in June after his visits in May. We observe that the percentage becomes larger when users are influenced by the recommendation from Facebook than from Foursquare.

From the above comparison, we find that an appropriate way for alleviating user distrust, especially in the out-of-town recommendation scenario, is to exploit and integrate the information from CBSNs (Communication Based Social Networks) [4] such as Facebook and Twitter. CBSNs are platforms for communication and promoting friendship, which characterizes the interaction mechanisms or behaviors including posting, reposting and replying in order to facilitate the communication among users. All these behaviors generate subjective comments on a POI to show its explicit preference. Users are also able to share interesting places or favorite food of daily life with their friends by their behaviors in CBSNs. Due to the close relationship between users, the information in CBSNs is more trustable than that in LBSNs. As illustrated in Fig. 1, friends' suggestion from Twitter (a CBSN), can help the user Tony make a decision. The information from his trustable and close friends on Twitter, named Alice, Bob and Dave, can give a preferable suggestion instead of trustless reviews on Foursquare (a LBSN). It is worthwhile to notice that review and comment are two different concepts, that is, a user express his feeling by reviews in LBSNs after visiting POIs and discuss with his friends on POIs in CBSNs which generates relevant comments.

Users often have multiple accounts in LBSNs and CBSNs which can be modeled by heterogeneous networks. A heterogeneous network consists of only one type of vertices, but there are more than one kind of relationships between these vertices. Due to its flexibility in modeling data heterogeneity, heterogeneous networks have been adopted in recommendation systems to characterize rich auxiliary data [5]. The difference of this work from the others lies in that we take into full consideration the characteristics of heterogeneous networks, while existing works only focus on analyzing a single type of networks. However, it is a difficult task to collect these data across two different websites because the corresponding users in the two networks can hardly be aligned, i.e., it is difficult to match the user accounts which are created by an individual on different websites. We observe that some users provide their Twitter and Facebook accounts on their Foursquare homepages. Connections between the accounts of Foursquare, Twitter and Facebook are constructed as the anchor links across these social networks. In this study, we generate heterogeneous networks such as the Foursquare-Twitter network and the Foursquare-Facebook network to suggestion POIs. Moreover, social relationships and interactive behaviors in CBSNs can help greatly improve the recommendation accuracy which were often ignored by the previous works as presented in [6,7].

When a user visits an unfamiliar area, his interest may drift across different geographical areas, i.e., he might change his preference when he comes to other areas which he isn't familiar with. Interest drift may reduce the quality of recommendation by using existing approaches, especially for out-of-town users [8]. As a matter of fact, users from the same social community often have similar interests in the same places, even taking into consideration the effect of interest drift. We can find the fact that the knowledge and experience from friends within the same communities will provide valuable information [9].

Based on the aforementioned discussion, we propose a latent probabilistic generative model called **H**eterogeneous **I**nformation based **LDA** (HI-LDA), which can accurately capture users' words on CBSNs by taking into full consideration the information on LBSNs including geographical effect as well as the abundant information on LBSNs including social relationship, users' interactive behaviors and comment content, in order to adapt to untrusted relationship and improve the effectiveness of POI recommendation, especially for out-of-town users. Fig. 2 shows the basic idea of the proposed HI-LDA model. Inspired by recent works about user interest modeling [7], HI-LDA adopts users' relationship in CBSNs as well as latent sentiments [10,11] in comments from CBSNs to characterize users' interests to overcome the untrusted problem in LBSNs. As shown in Fig. 2, a user has different accounts in Twitter and Foursquare, which are indicated in blue color (Twitter account) and red color (Foursquare account),

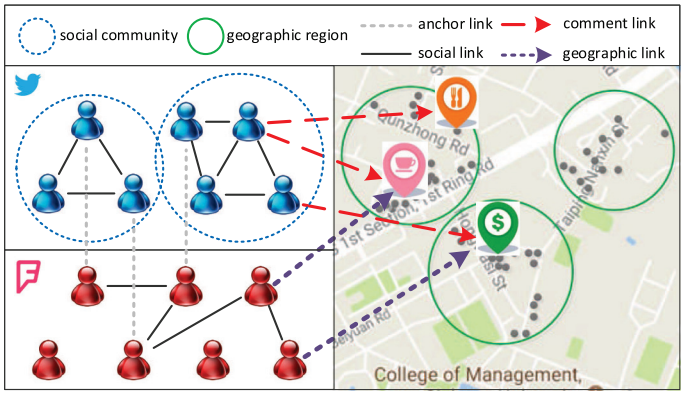


Fig. 2. Graphical representation of the HI-LDA model.

respectively, and connected by an anchor link which is used to show the correlation between the LBSN and the CBSN. We only show parts of anchor links for the convenience of understanding. Specifically, we first partition the geographical area into multiple regions and classify the users in the CBSN into multiple communities, and then infer individual user's sentiment distribution w.r.t. a geographical region and a social community according to the textual content of comments. The red directed lines indicate that the users in this LBSN make comments on the geographical locations by multiple categories of behaviors, e.g., posting and reposting. We can see that the HI-LDA model integrates the characteristics of heterogeneous networks (described by anchor links), social relationship between users (represented by social links) and comment words on CBSNs (represented by comment links). To the best of our knowledge, this is the first work to consider the distrust problem for POI recommendation over a heterogeneous network.

In order to demonstrate the applicability of HI-LDA, we investigate how it supports the following two recommendation scenarios in a unified way: (1) home-town recommendation that assumes that the target user is located in his home town, i.e., to meet users' information need in their daily life, and (2) out-of-town recommendation that aims to meet users' requirement when they travel out of town, especially in unknown areas. It is worthwhile to mention that both of the recommendation scenarios should be personalized and location-related, i.e., to recommend different POIs for the same target user at different time and different locations.

In this study, we make the following original contributions:

- We investigate the phenomenon of user's distrust in LBSNs and argue that the capability of handling the problem of user's distrust is essential for POI recommendation, especially for out-of-town recommendation. We make a thorough analysis of the hybrid effect of social communities, geographical locations and comment words corresponding to posts in CBSNs (i.e., their comments on different POIs posted by their friends or others), which is based on the following assumptions: (1) users from the same social communities might be interested in similar locations based on their frequent online communication, (2) a better POI recommendation for a user in specific areas can be inferred from the comment words of his friends in the same community.
- We propose a latent probabilistic generative model called HI-LDA, which can accurately capture comment words on CBSNs by taking into account social communities, interactive behaviors and the geographical information. In addition, we infer the parameters of the HI-LDA model effectively by the Gibbs sampling method.
- Via the HI-LDA model, we introduce a framework for POI recommendation including a geographical clustering algorithm based on locations and popularity.
- We conduct extensive experiments to evaluate the performance of the proposed HI-LDA model in two heterogeneous networks by the metric of hit-rate accuracy, and the results demonstrate that our approach outperforms the state-of-the-art baseline methods in POI recommendation in an effective and efficient fashion.

The remainder of this paper is organized as follows: Section 2 surveys the related works. Section 3 introduces some significant definitions and formulate the problem. Section 4 illustrates the application framework of POI recommendation. Section 5 introduces a geographical clustering approach via DBSCAN on popularity. Section 6 details the proposed HI-LDA model, and then presents the inference algorithm. The experimental results are presented in Section 7. Lastly, we conclude this study in Section 8.

2. Related works

In order to improve the quality of POI recommendation, the recent works mainly focus on integrating important subjective and objective factors including the effect of posted content, social relationship between individuals and characteristics of heterogeneous geo-social networks.

Effect of textual content. Currently, researchers have intensively studied the textual information of spatial items to alleviate the data sparsity problem of user-item matrices in LBSNs. Hu et al. [12] introduced a spatial topic model for POI recommendation by taking into account the spatial as well as textual information of users' posts. Yin et al. [7,13] integrated the textual information of the given POIs with local preferences and personal interests of users, to handle the data sparsity problem for out-of-town recommendation (the distance between user location and target area is larger than a threshold). Liu et al. [14] explored the effect of textual information with an integrated topic model and matrix factorization method. Wang et al. [6] introduced a latent probabilistic generative model to integrate geographical information and user reviews with location-aware and sentiment-aware individual interests. However, most of these approaches do not take into consideration the social connections and communication between users. Yang et al. [15] proposed a novel sentiment learning method by integrating photos posted by social users with comments from their friends. This model can discriminate the comments that are closely relevant to the sentiment retrieved from photos from the other irrelevant ones. But the sentiment derived from photos is inaccurate and ambiguous, which cannot be applied to POI recommendation. However, the textual contents derive from LBSNs in regard to these exiting works, which may cause the problem of data sparsity, especially for out-of-town areas. Although this problem can be coped with by integrating users' behavior information such as check-in records, the inherent data quality has a significant impact on the effectiveness of POI recommendation.

Geo-social effect. Users are socially clustered into multiple communities and POIs are geographically clustered into multiple regions. Both of them are ubiquitous in a geo-social network and evolve simultaneously. Many recent studies showed that there is a strong correlation between user check-in activities and geographical distance as well as social connections, so most of current POI recommendation approaches mainly focus on leveraging the geographical and social influence to improve recommendation accuracy. Han et al. [16] proposed an easy-to-compute metric and a community similarity degree to calculate the interest similarity among users in communities. Yuan et al. [17] proposed a frame-

work considering the influence of social relationships, which can distinguish each user's trusted and susceptible friends, respectively. Ference et al. [18] proposed a collaborative recommendation approach to recommend POI for out-of-town users in LBSNs based on the current location. The recommendation result depends on users' preferences, social influence as well as geographical proximity. Yin [19] proposed a latent probabilistic generative model to mimic user interest drift by integrating the factors of region-dependent personal interests and crowd's preferences. It alleviates the data sparsity for out-of-town regions by considering the social and geographical correlation. Yin [20] proposed a model to incorporate network proximities, spatio-temporal co-occurrences and semantic similarity into the generative process of communities. However, the aforementioned works do not consider the effect of user's behavior. For instance, the forwarded texts in Twitter always contain more useful information than the replying words as concluded in [21]. In this study, we propose a latent probabilistic generative model called HI-LDA, which can accurately capture comment words on CBSNs by taking into account social the interactive behaviors of users.

Recommendation in heterogeneous networks. Users often have multiple accounts in different social networks simultaneously. Before users join into another network, they may have been using other networks for a long time. Zhang et al. [22] introduced an approach to address the problem of recommending different categories of items simultaneously for a new LBSN across partially aligned social networks. The recommendation services in LBSNs can be viewed as the problem of link prediction. Sajadmanesh et al. [23] proposed an effective general meta-path-based approach to address the missing target problem. In the model, multiple kinds of social factors may impact a user's decision of joining the target network, which also influence the generation of a new anchor link. Min et al. [24] investigated the problem of cross-platform multimedia recommendation. In order to help users to enjoy different recommendation platforms in an effective manner, they proposed a cross-platform multi-modal topic model, which can distinguish two types of topics and align multiple modalities of different platforms. Wang et al. [25] explored the effects of various diffusion patterns on the information diffusion process and inferred multi-aspect diffusion networks with multi-pattern cascades. However, most of the above methods do not recommend a POI by taking into full consideration of the geographical as well as social effects. In this study, we make a thorough analysis of the hybrid effect of social relationship, geographical locations and comment words corresponding to posts in CBSNs. Qiao et al. [26] proposed a hidden Markov model based location prediction algorithm, which can self-adaptively select important parameters. Furthermore, Qiao et al. [27] proposed a three-in-one location suggestion model, which predicts possible location of moving objects with relative uncertainty. However, the above three works suggest locations based on geographical information without considering the social relationship and textual information derived from CBSNs. For POI recommendation, the prediction accuracy cannot be guaranteed.

3. Problem statement

In this section, we first formalize the POI recommendation problem and then introduce important definitions and notations.

Problem statement. Given the dataset of a LBSN-CBSN network, which contains POI relevant comments and all relevant users' profiles, our goal is to suggest a list of POIs that a target user u may be interested in. Given a distance threshold d_0, the problem takes into consideration two scenarios: the out-of-town recommendation, i.e., the distance between user location and target area is larger than d_0, and the home-town recommendation when distance is smaller than d_0.

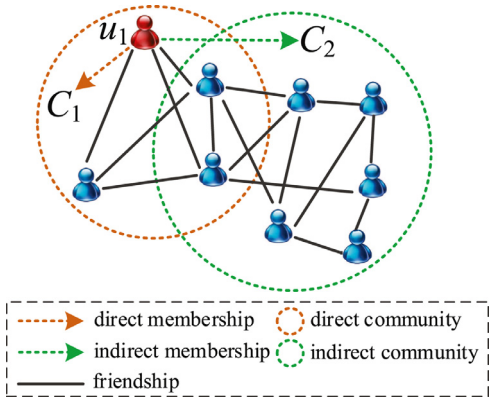


Fig. 3. Direct membership and indirect membership.

To facilitate understanding, Table 2 describe the important notations used in this study.

For ease of proposing our method, it is essential to present some definitions beforehand.

Definition 1 (POI)**.** A POI is a uniquely identified site (e.g., a restaurant or a park). In the proposed model, a POI has two attributes, i.e., identifier and location represented by longitude and latitude coordinates, which is denoted by the notations v and l_v, respectively.

Definition 2 (User's home location)**.** Given a user u, we denote the userâs home location where the user lives in. However, it is hard to directly obtain a userâs home location. Thus, we use the method developed by Scellato et al. [28], such that we regard the spatial item where a user frequently presents reviews as her home location.

Definition 3 (Community)**.** Social communities are groups of connected users with a high density, while the connections between groups are sparse [21].

Definition 4 (Community membership)**.** In a social network, each user might have multiple roles and belong to different communities. The community membership is used to represent the affiliation degree w.r.t. a user to a community, which represents a user's tendency to communicate with others in this community. A collection of communities $\mathcal{C}$ follows a multinomial distribution with user u as the parameter, i.e., $\psi_u = \{\psi_{u,c} : c \in \mathcal{C}\}$: where $\psi_{u,c}$ denotes the probability of user u belonging to community c.

The community membership w.r.t. a user involves two categories: direct membership and indirect membership. As shown in Fig. 3, user u_1 is not involved in the community C_2, but two of u_1's friends are C_2's members. Therefore, user u_1 has an indirect membership of community C_2.

Definition 5 (Interaction behaviors)**.** Social media feature some interaction mechanisms called behaviors to facilitate the communication, including (a) posting a new comment on a POI, (b) forwarding the post from friends and add some comments, and (c) replying to the post of friends. All these behaviors can generate subjective comments on a POI, which show its explicit features. The notation $\mathcal{B}_c$ is used to represent the collection of behaviors within community c, each of which can be denoted by b. A collection of behaviors $\mathcal{B}_c$ follows a multinomial distribution with community c as the parameter, i.e., $\sigma_c = \{\sigma_{c,b} : b \in \mathcal{B}_c\}$, where $\sigma_{c,b}$ denotes the probability of behavior b happening in community c.

Table 2
Notations in the proposed model.

Symbol	Description
$\mathcal{U}, u$	Set of users and a specific user u
R, N_r	Number of regions and number of POIs w.r.t. region r
C, N_c	Number of communities and number of behavior types w.r.t. community c
M, N_m	Number of comments on a POI w.r.t. POI v and behavior b, number of words w.r.t. a comment
K	Number of sentiments
$\mathcal{W}_{b,v}$	Collection of words from the reviews in the LBSN and comments in the CBSN w.r.t. behavior b and POI v
ψ	Parameter of the multinomial distributions over community c w.r.t. u
σ	Parameter of the multinomial distributions over behavior b w.r.t. community c
ϑ	Parameter of the multinomial distributions over region r w.r.t. u
σ	Parameter of the multinomial distributions over behavior b w.r.t. c
μ, Σ	Mean value and variance of the geographical Gaussian distributions w.r.t. region r
θ	Parameter of the multinomial distributions over the latent variable α w.r.t. a comment $(u, b, v, \mathcal{W}_{b,v})$
φ	Parameter of the multinomial distributions over the latent variable β w.r.t. a topic z
$\alpha, \beta, \gamma, \tau, \eta, \pi$	Hyper-parameters of the Dirichlet distribution corresponding to $\theta, \varphi, \vartheta, \phi, \psi, \sigma$, respectively

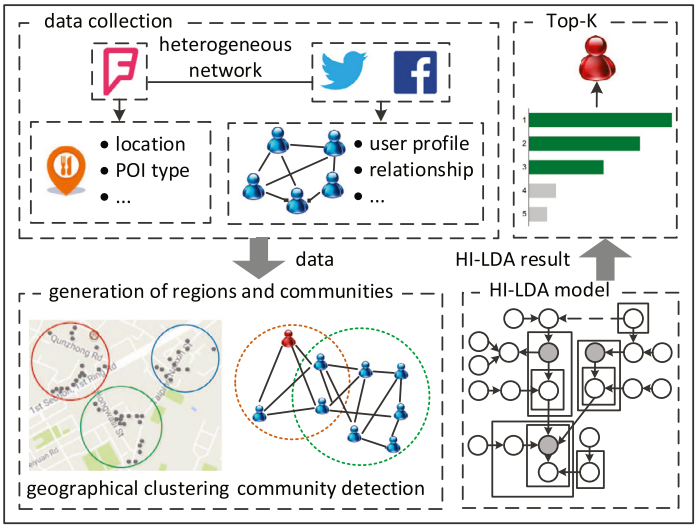


Fig. 4. A framework of POI recommendation by the HI-LDA model.

Definition 6 (Comments on a POI)**.** Users in a community make comments on a POI by interaction mechanism. Each word in a textual comment belongs to one kind of sentiment. The notation $W_{b,v}$ is used to denote the collection of words marking on a POI v by behavior b.

Definition 7 (Sentiment)**.** Given a collection of words $\mathcal{W}$, a sentiment z is defined as a multinomial distribution over $\mathcal{W}$, i.e., $\phi_z = \{\phi_{z,w} : w \in \mathcal{W}\}$, where $\phi_{z,w}$ denotes the probability of a sentiment z generating word w.

In this study, we focus on make effective and efficient POI recommendation for both home-town and out-of-town scenarios. And, the problem of POI recommendation can be described as follows.

4. An effective framework of POI recommendation via HI-LDA

In order to achieve our goal presented in Section 3, we illustrate the framework of POI recommendation in Fig. 4.

Basically, the framework of POI recommendation contains four essential phases:

(1) Data is collected from LBSN and CBSN portals and then build heterogeneous networks.
(2) Cluster POIs based on their locations and popularity, and assign each noise point to the nearest cluster. Then, a series of geographical clusters called regions are generated. Moreover, users are grouped into different communities by community detection algorithms [29], while each user is associated with multiple communities.
(3) The probability of user choosing each unvisited POI is calculated based on the HI-LDA model.
(4) Recommended the top-K POIs with the highest probabilities to a given user.

5. Geographical clustering approach via DBSCAN on popularity

A POI is an interesting place represented by a point with longitude and latitude position on the map. Given a set of POIs in home-town or out-of-town, POIs with many nearby neighbors can be grouped together by the clustering algorithms such as DBSCAN [30]. The results of DBSCAN include some noise points, which are located far from any cluster in the spatial distance. However, some noise points may be attractive to a specific group of people, which are viewed as potential POIs.

In this section, we present a comprehensive description of the geographical clustering approach based on the DBSCAN algorithm called popularity based DBSCAN (P-DBSCAN). Different from DBSCAN, our proposed approach reports each noise point as a member of the nearest cluster under some conditions.

Four important parameters are involved in the algorithm. The ϵ-neighborhood and the minimum number of points (*MinPts*) have the same meaning as given in DBSCAN, while the noise distance threshold δ and the popularity threshold $\mathcal{T}_{pop}$ are two new parameters in the proposed algorithm. The appropriate values of these

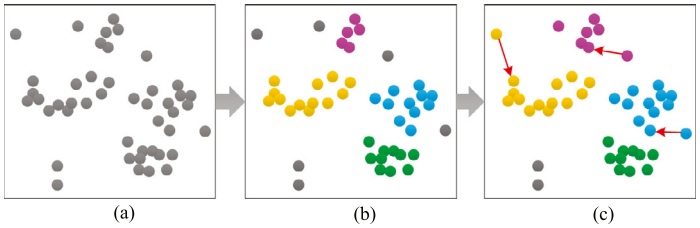


Fig. 5. Working mechanism of location clustering by DBSCAN on popularity.

four parameters are specified by taking into consideration the distribution of points.

Definition 8 (ϵ-neighborhood)**.** The ϵ-neighborhood of a given point p, represented by $N_\epsilon(p)$, is defined by $N_\epsilon(p)=\{q \in D | dist(p, q) \leq \epsilon\}$, where ϵ is the radius of area D.

Definition 9 (Geographical region)**.** Given $\mathcal{D}$ is a geographical database, a cluster C w.r.t. ϵ and *MinPts* is a non-empty subset of $\mathcal{D}$ satisfying the following conditions:

(1) $\forall$ p, q: if $p \in C$ and q is density-reachable from p w.r.t. ϵ and *MinPts*, then $q \in C$.
(2) $\forall$ p, $q \in C$: p is density-connected to q w.r.t. ϵ and *MinPts*.

Definition 10 (Noise distance threshold)**.** It is denoted by δ and used for assigning the noise points to the nearest cluster within the threshold δ. The clustering result of DBSCAN contains core areas and noise points. Some noise points that can be included into the clusters are called marginal areas. The noise distance threshold is used to identify the marginal areas, thus we have $\delta > \epsilon$.

Definition 11 (Popularity threshold)**.** The popularity indicates the popularity of a point and can be measured by the number of reviews or the rating score. The popularity threshold is viewed as a constraint to identify the marginal areas.

The detail of this geographical clustering algorithm is given in Algorithm 1.

Algorithm 1: Geographical clustering based on popularity.

Input: A geographical database $\mathcal{D}$, the cluster radius ϵ, and the minimum number of points *MinPts*.
Output: A set of regions $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$, where n is the number of regions.

1 n=0;
2 Initialize a set *NoiseSet*;
3 $\mathcal{R}$, *NoiseSet* = *DBSCAN*($\mathcal{D}, \epsilon$, *MinPts*);
4 **foreach** $i \in$ NoiseSet *AND* $i.pop > \mathcal{T}_{pop}$ **do**
5 $R_i = GetNearestCluster(i, \mathcal{R}, \delta)$;
6 $UpdateCluster(R_i, i)$;
7 **end**
8 **return** $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$;

Algorithm 1 first uses the DBSCAN algorithm to cluster points by longitude and latitude position of points (lines 1–3). For each point with a popularity larger than $\mathcal{T}_{pop}$ in the set of noise points (line 4), it finds the nearest region within the distance δ (line 5), and add the point to the region (line 6). Lastly, it returns all the regions (line 8).

Fig. 5 demonstrates the working mechanism of our proposed P-DBSCAN algorithm. Fig. 5(a) is the initial state with all points painted as gray. Then points are clustered into several regions by DBSCAN as in Fig. 5(b). Fig. 5(c) shows that the noise points are finally added to the nearest region within the noise distance threshold.

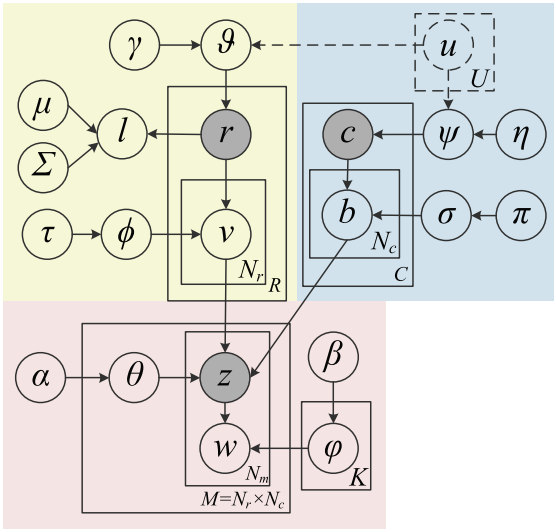


Fig. 6. Graphical representation of the proposed model.

Similar to DBSCAN, this approach does not require to specify the number of regions, and can find arbitrarily shaped regions. DBSCAN cannot handle datasets with large differences in densities, since the *MinPts* – ϵ combination cannot then be chosen appropriately for all clusters. In our algorithm, it does works well when specifying a relatively small ϵ value, and assigning the noise points to their nearest regions.

6. Heterogeneous information based LDA model (HI-LDA model)

In this section, we propose our **H**eterogeneous **I**nformation based **LDA** model (HI-LDA model). Firstly, Section 6.1 introduces the model structure of HI-LDA. Then, Section 6.2 introduces the probabilistic generative process of observed variables, e.g., interactive behaviors and comment words on POIs, which are generated according to their distributions. Moreover, we infer the parameters in Sections 6.3 and 6.4. Finally, we analyze the computational complexity of the inference process in Section 6.6.

6.1. Model structure

In order to infer the distribution of POIs from geo-social networking data, we propose a probabilistic generative model called Heterogeneous Information based LDA model (HI-LDA model), which jointly models users' interactive behaviors by taking into full consideration geo-partitioned areas and the content of posts as well. Basically, the HI-LDA model takes into account the following factors: geographical distance, semantic similarity and social homogeneity.

Fig. 6 shows the probabilistic generative process of HI-LDA, and the notations of the model are given in Table 2.

In general, the generative process of the proposed model consists of three components:

Community-behavior component. We assume that the number of the occurrence of each community and the number of the occurrence of each behavior within a community follow the multinomial distribution with the parameter of ψ and σ, respectively, which is shown in the blue area of Fig. 6.

Region-POI component. A popular region or POI always invokes more comments. The number of the occurrence of each region and the number of the occurrence of each POI within a region follow the multinomial distribution with the parameters ϑ and ϕ, respectively. Based on the literature [19], the geographical location of each POI v follows a geographical Gaussian distribution [14] in region r, which is characterized by μ and Σ. This component is shown in the yellow area of Fig. 6.

Sentiment-word component. It describes the generative process of different sentiments and the relevant words w.r.t. a specific POI by one kind of interactive behavior, which is similar to the LDA model [31]. The red area of Fig. 6 illustrates this component.

6.2. Generative process

The probabilistic generative process is given in Algorithm 2. When user u decides to generate a word, he first selects a geographical region r, then chooses a POI v with the location l_v which follows the geographical Gaussian distribution. u belongs to several communities depending on interests or special relationship. He chooses a community r from these communities, including direct and indirect ones, then communicates with one of the users in the community r by behavior b. After the above process, a sentiment distribution w.r.t. the combination (v, b) is generated. With the given topic z, words $\mathcal{W}_{v,b}$ are generated from the distribution φ_z.

Algorithm 2: Generative process in HI-LDA.

1 **foreach** *user* $u \in \mathcal{U}$ **do**
2 Sample $\psi_u \sim Dirichlet(\cdot|\eta)$;
3 Sample $\vartheta_u \sim Dirichlet(\cdot|\gamma)$;
4 **foreach** *community* $c \in \mathcal{C}$ **do**
5 Sample $\sigma_{u,c} \sim Dirichlet(\cdot|\pi)$;
6 **end**
7 **end**
8 **foreach** *region* $r \in \mathcal{R}$ **do**
9 Sample $\phi_r \sim Dirichlet(\cdot|\tau)$;
10 **foreach** *community* $c \in \mathcal{C}$ **do**
11 Sample $\theta_{c,r} \sim Dirichlet(\cdot|\alpha)$;
12 **end**
13 **end**
14 **foreach** *sentiment type* $z \in \mathcal{Z}$ **do**
15 Sample $\varphi_z \sim Dirichlet(\cdot|\beta)$;
16 **end**
17 **foreach** *user* $u \in \mathcal{U}$ **do**
18 **foreach** *comment* $(u, v, b, \mathcal{W}_{v,b}) \in \mathcal{M}_{v,b}$ **do**
19 Sample the index of a community $c \sim Multi(\psi_u)$;
20 Sample the index of a behavior $b \sim Multi(\sigma_c)$;
21 Sample the index of a region $r \sim Multi(\vartheta_u)$;
22 Sample the index of a POI $v \sim Multi(\phi_r)$;
23 Sample the coordinate of a location $l_v \sim N(\mu_r, \Sigma_r)$;
24 Sample a sentiment type $z \sim Multi(\theta_{v,b})$;
25 **foreach** *word index* $w \in \mathcal{W}_{v,b}$ **do**
26 Sample word $w \sim Multi(\varphi_z)$;
27 **end**
28 **end**
29 **end**

The probabilistic generative process is given in Algorithm 2. When making a POI recommendation for u, it first samples the distribution parameter of ψ_u, ϑ_u, $\sigma_{u,c}$, ϕ_r, $\theta_{c,r}$, φ_z corresponding to the hyper-parameters (lines 1–16). Then, it samples each variable according to its distribution (lines 17–29).

6.3. Model inference

Based on the generative process, the joint probability of the observed (c, b, r, v, l, z, w) and hidden variables $(\psi, \sigma, \vartheta, \phi, \theta, \varphi)$ are represented in Eq. (1).

$$
\begin{aligned}
&p(\psi, \sigma, \vartheta, \phi, \theta, \varphi, c, b, r, v, l, z, w|\alpha, \beta, \gamma, \tau, \eta, \pi, \mu, \Sigma, \mathcal{U}) \\
&= \prod_{u\in\mathcal{U}} \underbrace{p(b|\sigma, c)p(c|\psi)p(\psi|\eta, u)p(\sigma|\pi)}_{\text{community-behavior}} \cdot \\
&\prod_{u\in\mathcal{U}} \underbrace{p(l|r, \mu, \Sigma)p(v|r, \phi)p(r|\vartheta)p(\vartheta|\gamma, u)p(\phi|\tau)}_{\text{region-POI}} \cdot \\
&\underbrace{p(z|\theta, v, b)p(\theta|\alpha)p(w|\varphi, z)p(\varphi|\beta)}_{\text{sentiment-word}} \\
&= \prod_{u\in\mathcal{U}} \underbrace{p(b|\sigma_c)p(c|\psi)p(\psi|\eta_u)p(\sigma|\pi)}_{\text{community-behavior}} \cdot \\
&\prod_{u\in\mathcal{U}} \underbrace{p(l|\mu_r, \Sigma_r)p(v|\phi_r)p(r|\vartheta)p(\vartheta|\gamma_u)p(\phi|\tau)}_{\text{region-POI}} \cdot \\
&\underbrace{p(z|\theta_{v,b})p(\theta|\alpha)p(w|\varphi_z)p(\varphi|\beta)}_{\text{sentiment-word}}
\end{aligned} \tag{1}
$$

As shown in Eq. (1), the joint probability is partitioned into three segments corresponding to three components introduced in Section 6.1. Given the hyper-parameters α, β, γ, η, τ, π and the observed values c, b, r, v, l, z, w, our goal is to infer the latent variables ψ, σ, ϑ, ϕ, θ and φ, as well as the parameters μ and Σ. As it is a very difficult task to accurately calculating these posteriors, the Markov Chain Monte Carlo (MCMC) method is adopted in this study to retrieve samples, as given in [32]. The MCMC method enables the process to converge to a target stationary distribution by constructing a Markov chain, then samples are drawn from this Markov chain. Each sampled variable can be assigned a value in each state of the chain.

As a widely used MCMC algorithm, Gibbs sampling can iteratively draw the values of latent variables from a Markov chain, whose stationary distributions agree with the posteriors. By employing the Gibbs sampling approach, we sample latent regions and sentiments from the comments by fixing all other parameters, e.g., μ and Σ. The posterior distribution of Gibbs sampling is stationary. Therefore, the samples can be used to estimate the distribution.

For each comment $(u, b, v, \mathcal{W}_{b,v})$, behavior b and POI v belong to a specific community c and a specific region r, respectively. We first update the values of community c and region r according to the values of other variables, i.e., sample c and r in terms of the following posterior probability:

$$
p(c|c_{\neg}, b, u) \propto \frac{n^{\neg}_{u,c} + \eta}{\sum_{u'}(n^{\neg}_{u',c} + \eta)} \cdot \frac{n^{\neg}_{b,c} + \pi}{\sum_{b'}(n^{\neg}_{b',c} + \pi)} \tag{2}
$$

$$
p(r|r_{\neg}, v, l_v, u) \propto \frac{n^{\neg}_{u,r} + \gamma}{\sum_{u'}(n^{\neg}_{u',r} + \gamma)} \cdot \frac{n^{\neg}_{v,r} + \tau}{\sum_{v'}(n^{\neg}_{v',r} + \tau)} \cdot p(l_v|\mu_r, \Sigma_r) \tag{3}
$$

where $c_{\neg}$ represents all communities except the current one; $r_{\neg}$ represents all regions except the current one; $n_{u,c}$ and $n_{u,r}$ are the

numbers of occurrence that community c and region r are sampled w.r.t. user u, respectively; $n_{b,c}$ is the number of occurrence that behavior b is sampled from community c; $n_{v,r}$ is the number of occurrence that location v is drawn from region r; the number $n^{\neg}$ denotes the number of items, e.g., comments, words and regions, excluding the current item.

By Eqs. (2) and (3), POI v may belong to another region and behavior b may belong to another community. Then we update the mean value and the variance of region r, i.e., μ_r and Σ_r by the following equations:

$$\mu_r = E(r) = \frac{1}{|S_r|}\sum_{v\in S_r} l_v \tag{4}$$

$$\Sigma_r = D(r) = \frac{1}{|S_r|-1}\sum_{v\in S_r}((l_v-\mu_r)^T(l_v-\mu_r)) \tag{5}$$

where S_r is the set of locations assigned to region r.

A comment is generated by behavior b on POI v, and can be represented by a tuple (v, b). Then we can sample the sentiment within comment (v, b) according to the following posterior probability:

$$p(z|z_{\neg}, v, b, \mathcal{W}_{v,b}) \propto \frac{n^{\neg}_{v,b,z}+\alpha}{\sum_{(v',b')}(n^{\neg}_{v',b',z}+\alpha)} \cdot \prod_{w\in\mathcal{W}_{v,b}} \frac{n^{\neg}_{w,z}+\beta}{\sum_{w'}(n^{\neg}_{w',z}+\beta)} \tag{6}$$

where $z_{\neg}$ represents sentiments for all comment words w.r.t. a POI-community combination except the current one; $n_{v,b,z}$ is the number of occurrence that sentiment z is sampled from POI v by behavior b; $n_{w,z}$ is the number of times that word w is generated from sentiment z.

The hyper-parameters are specified to: α=50/K, γ=50/R, η=50/C, $\beta = \tau = \pi$=0.01 by experimental studies, following the literature [7]. The geographical region r is initialized by our proposed geographical clustering approach in Section 5, and then randomly initialize the sentiment z assigned to each word. During each iteration, the assignments for communities, regions and sentiments are updated by Eqs. (2), (3) and (6), respectively. After each iteration, the geographical Gaussian distribution parameters are updated by Eqs. (4) and (5). The iteration is repeated until convergence. Lastly, the posterior samples can be used to estimate the parameters by examining the numbers of c, r and z assignments to the comments.

6.4. Inference framework

After an adequate number of sampling iterations, the approximated posteriors can be used to estimate the parameters by examining the numbers of c, r and z assignments to users' comments in CBSNs. The inference framework is presented in Algorithm 3.

Algorithm 3 includes the following steps:

(1) Randomly initialize the community and region assignments for each comment (lines 1–3);

(2) In each iterative operation, update the community, region and sentiment assignments for each comment by Eq. (2), Eq. (3) and Eq. (6), respectively (lines 7–11).

(3) The iteration is repeated until convergence (lines 6–21). In addition, a burn-in process [33] is introduced in the first hundreds of iterations to remove unreliable sampling results (line 12). The sample lag, i.e., the interval between samples after burn-in, is also introduced to sample only periodically thereafter to avoid correlations between samples.

Algorithm 3: The inference process of HI-LDA.

Input: number of iteration I, number of burn-in I_b, sample lag I_s
Output: estimated parameters $\hat{\theta}, \hat{\varphi}, \hat{\vartheta}, \hat{\psi}, \hat{\phi}, \hat{\sigma}$

1 **foreach** *comment* $(u, v, b, \mathcal{W}_{v,b}) \in \mathcal{M}_{v,b}$ **do**
2 Sample community, region and sentiment randomly;
3 **end**
4 $\theta_{sum} \leftarrow 0$, $\varphi_{sum} \leftarrow 0$, $\vartheta_{sum} \leftarrow 0$, $\psi_{sum} \leftarrow 0$, $\phi_{sum} \leftarrow 0$ and $\sigma_{sum} \leftarrow 0$;
5 *count*=0;
6 **foreach** *iter* $\in [1, IterSum]$ **do**
7 **foreach** *comment* $(u, v, b, \mathcal{W}_{v,b}) \in \mathcal{M}_{v,b}$ **do**
8 Sample a community according to Eq. 2;
9 Sample a region according to Eq. 3;
10 Sample a sentiment according to Eq. 6;
11 **end**
12 **if** *(iter* $> I_b$*) and (iter mod* I_s*==0)* **then**
13 count←count+1;
14 $\theta_{sum} \leftarrow \theta_{sum} + \frac{n_{v,b,z}+\alpha}{\sum_{v'}\sum_{b'}(n_{v',b',z}+\alpha)}$;
15 $\varphi_{sum} \leftarrow \varphi_{sum} + \frac{n_{w,z}+\beta}{\sum_{w'}(n_{w',z}+\beta)}$;
16 $\psi_{sum} \leftarrow \psi_{sum} + \frac{n_{c,z}+\eta}{\sum_{c'}(n_{c',z}+\eta)}$;
17 $\vartheta_{sum} \leftarrow \vartheta_{sum} + \frac{n_{u,r}+\gamma}{\sum_{u'}(n_{u',r}+\gamma)}$;
18 $\phi_{sum} \leftarrow \phi_{sum} + \frac{n_{v,z}+\tau}{\sum_{v'}(n_{v',z}+\tau)}$;
19 $\sigma_{sum} \leftarrow \sigma_{sum} + \frac{n_{b,c}+\pi}{\sum_{b'}(n_{b',c}+\pi)}$;
20 **end**
21 **end**
22 $\hat{\theta} \leftarrow \theta_{sum}/count$;$\hat{\varphi} \leftarrow \varphi_{sum}/count$;$\hat{\vartheta} \leftarrow \vartheta_{sum}/count$;$\hat{\psi} \leftarrow \psi_{sum}/count$;$\hat{\phi} \leftarrow \phi_{sum}/count$;$\hat{\sigma} \leftarrow \sigma_{sum}/count$;;
23 **return** $\hat{\theta}, \hat{\varphi}, \hat{\vartheta}, \hat{\psi}, \hat{\phi}, \hat{\sigma}$;

6.5. POI recommendation based on HI-LDA

Once we have estimated the model parameter set $\hat{\Psi} = \{\hat{\theta}, \hat{\varphi}, \hat{\vartheta}, \hat{\psi}, \hat{\phi}, \hat{\sigma}\}$, given a target user u at a specific location, we calculate the probability of user u selecting each unvisited POI v by the following equation:

$$P(v|u) = \sum_r P(v|r)P(r|u) \tag{7}$$

where $P(r|u)$ is calculated as follows based on Bayes rule:

$$P(r|u) = \sum_c P(c|u)P(r|c) = \sum_c P(c|u)\frac{P(r)P(c|r)}{\sum_{r'}P(r')P(c|r')} \propto P(r)\sum_c P(c|u)P(c|r) \tag{8}$$

where $P(r)$ is calculated by:

$$P(r) = \sum_u P(r|u)P(u) \tag{9}$$

The probability that community c generate words about r is the total probability that the set of words $\mathcal{W}_{b,v}$ appears, thus $P(c|r)$ is computed by:

$$P(c|r) \propto P(cr) = \sum_{v\in r}\sum_b P(v|r)P(l_v|r)P(b|c)P(\mathcal{W}_{b,v}) \tag{10}$$

where $\mathcal{W}_{b,v}$ denotes the collection of words generated by the specific behavior b targeting to POI v, which is calculated by the following equation:

$$P(\mathcal{W}_{b,v}) = \sum_z P(z|b,v)\left(\prod_{w\in\mathcal{W}_{b,v}} P(w|z)\right)^{\frac{1}{\mathcal{W}_{b,v}}} \tag{11}$$

Based on Eqs. (8)–(11), the original Eq. (7) can be reformulated as in Eq. (12).

$$P(v|u) \propto \sum_{u'} P(r|u')P(u') \sum_{r} P(v|r) \sum_{c} P(c|u)$$
$$\sum_{v}\sum_{b} P(v|r)P(l_v|r)P(b|c) \sum_{z} P(z|b,v) \left(\prod_{w \in \mathcal{W}_{b,v}} P(w|z) \right)^{\frac{1}{\mathcal{W}_{b,v}}}$$
$$= \sum_{u'} P(u') \hat{\vartheta}_{u',r} \sum_{r} \hat{\phi}_{r,v} \sum_{c} \hat{\psi}_{c,u} \sum_{v} \hat{\phi}_{r,v} P(l_v|\hat{\mu}_r, \hat{\Sigma}_r)$$
$$\sum_{b} \hat{\sigma}_{b,c} \sum_{z} \hat{\theta}_{b,v} \left(\prod_{w \in \mathcal{W}_{b,v}} \hat{\varphi}_z \right)^{\frac{1}{\mathcal{W}_{b,v}}} \quad (12)$$

The value of $P(v|u)$ is the probability of POI v that will be recommended to user u. The remaining work of POI recommendation is to sort all the POIs based on the $P(v|u)$ value and recommend the top-K POIs with the largest probabilities to user u.

6.6. Computational complexity

The computational complexity of the inference framework is analyzed in this section. Suppose that the whole process executes for I iterations, in each of which, all comments are scanned. Given the notations introduced in Table 2, for each comment, $\mathcal{O}(C+R+K)$ operations are required to sample latent communities, latent regions and latent sentiments. The series of comments from a user is indicated by $\mathcal{M}_u$, so the whole computational complexity is $\mathcal{O}(I(C+R+K)\sum_u |\mathcal{M}_u|)$.

7. Experimental results

In this section, the experimental results of the proposed HI-LDA model are represented in order to evaluate its effectiveness and efficiency by comparing with the state-of-the-art works.

7.1. Experimental setup

7.1.1. Datasets

Popular LBSN and CBSN datasets are used to conduct experiments in this study. Foursquare is a famous LBSN, while Twitter and Facebook are popular CBSNs. Users exist in all or part of these datasets. The topologies of these datasets vary by their follower or friend relationships.

Foursquare dataset. Foursquare is a popular location-based social network and can offer many location-based services, e.g., POI check-ins and posting online reviews for a POI. The dataset was collected from 75,140 users living in the areas of San Francisco, California in USA, and contains the check-in information, including identities, names, check-in time and locations in terms of latitude and longitude.

Twitter dataset. As a popular microblog portal, Twitter has attract a huge number of users from the Internet, and has been viewed as a platform for communication and promoting friendship. Users like to share interesting places or favorite food of daily life with their friends. The data collected from Twitter includes information about 28,553 users and their related 186,589 comments.

Facebook dataset. Facebook is another popular CBSN portal with more than 2.19 billion monthly active users during the first quarter of 2018. It provides several communication channels for users, including user messages, chatting and status updates, et al. The numbers of collected users and relevant comments on POIs are 52,772 and 378,117, respectively.

Since there is no dataset which can meet the requirements of heterogeneous networks, we need to collect new data from these three networks and fuse these data together to obtain the geographical relationship of POIs and the social relationship of users. Data are obtained in San Francisco during May 2017 by our designed crawler. As shown in Fig. 7, if users provide their Twitter and Facebook accounts on their Foursquare homepages, connections between the accounts of Foursquare, Twitter and Facebook are constructed as the anchor links across these social networks. After removing abundant data, we create two heterogeneous networks with anchor links: the Foursquare-Twitter network and the Foursquare-Facebook network. These datasets contain sensitive and privacy information such as comment words and locations. We use some privacy protection methods to preprocess these data. For example, all POIs' names and users' names are converted into digital numbers, and the longitudes and latitudes of POIs are changed into relative coordinates.

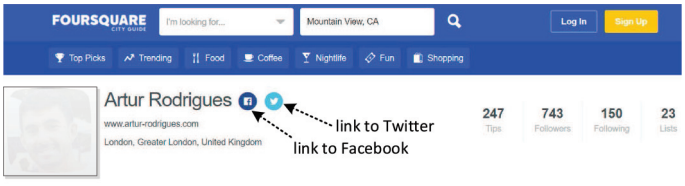


Fig. 7. Crawl on the three websites.

Table 3
Description of the datasets .

	Foursquare	Twitter	Facebook
Number of POIs	45,322	–	–
Number of regions	1561	–	–
Number of comments	634,088	186,589	378,117
Number of users	75,140	28,553	52,772

Table 4
Description of the heterogeneous networks.

	Fs*-Twitter	Fs-Facebook
Number of POIs	36,771	31,891
Number of comments	35,510	25,299
Number of users	18,279	8,601

* Fs is short for Foursquare.

The statistics of the datasets and the heterogeneous networks are shown in Tables 3 and 4.

7.1.2. Baseline approaches

We compare our proposed HI-LDA model with the following state-of-the-art POI recommendation approaches.

UPS-CF [18] is a collaborative recommendation approach to recommend POI for out-of-town users in LBSNs based on the current location. The recommendation result depends on users' preferences, social influence as well as geographical proximity.

ST-LDA [19] is a latent probabilistic generative model to mimic user interest drift by integrating the factors of region-dependent personal interests and crowd's preferences. It alleviates the data sparsity for out-of-town regions by considering the social and geographical correlation.

JIM [13] is a joint probabilistic generative model which adapts users' check-in behaviors to temporal effect, content effect, geographical influence and word-of-mouth effect, especially suitable for out-of-town users. Compared with our HI-LDA, JIM does not take into account the crowd sentiments, and only considers the temporal effect.

UCGT [20] infers users' social communities by incorporating their spatio-temporal and semantic information. However, the

Table 5
Features of different recommendation methods.

	Geo.	Temp.	Pref.	Soc.	Behav.	Tex.
UPS-CF	•		•	•	•	
ST-LDA	•	•			•	•
JIM	•	•			•	•
UCGT	•	•		•	•	•
LSARS	•		•		•	•
HI-LDA	•		•	•	•	•

geographical effect is not involved in this model, which is not viewed as a rational recommendation approach.

LSARS [6] is a latent probabilistic generative model that mimic the decision-making process of users' check-in activities both in home-town and out-of-town scenarios by adapting to user interest drift and crowd sentiments, which can learn location-aware and sentiment-aware interests of individuals from the profiles of POIs and the reviews of users. However, the geographical influence is not considered when generating the review words in LSARS.

The features of these recommendation methods are compared in Table 5, whose headers refer to several factors including geographical influence (denoted by Geo.), temporal effect (denoted by Temp.), users' preferences (denoted by Pref.), social relationship (denoted by Soc.), users' behaviors (denoted by Behav.) as well as textual content (denoted by Tex.).

In order to further validate the improvement obtained by HI-LDA, we designed four variant models based on HI-LDA by distinguishing users' individual interests in different communities, discriminating comment sentiments w.r.t. different behaviors and different locations, and exploiting the effect of a region's size.

HI-LDA-V1 assumes that users' interactive behaviors are community-independent, i.e., the distribution of sentiments is similar across communities.

HI-LDA-V2 is the second variant version of HI-LDA where users' comment sentiments are activity-independent, i.e., users have the same sentimental distribution by different behaviors.

HI-LDA-V3 does not consider the influence of locations on comment sentiments, i.e., all the POIs in the a region have the same sentimental distribution.

HI-LDA-V4 adopts larger administrative divisions like cities instead of a small regions, and then infer city-dependent personal interests and make a recommendation.

HI-LDA-V5 only exploits the reviews in the LBSN, instead of the mixture that is composed of the reviews in the LBSN and the comments in the CBSN.

HI-LDA-V6 only consider the LBSN including the reviews and the friend relationships, etc. the CBSN is removed in this variant.

7.1.3. Evaluation metrics

HI-LDA can be applied to both home-town and out-of-town recommendation, so we evaluate the recommendation performance under these two scenarios, separately. To determine whether the recommendation happens in home-town or out-of-town, we measure the distance between the visiting place and a user's home, which is extracted from his profile. If the distance is greater than d_0, the user is supposed to be located in an out-of-town region. Following the previous study [34], we specify the threshold distance d_0 to 100km in this study.

In order to evaluate the overall recommendation efficiency and effectiveness, we utilize *Accuracy@k* according to the methodological framework [12] by 10-fold cross validation. Specifically, we divide the dataset into subsets based on the clusters after the clustering of locations. A subset with a comparable number of locations involves only adjacent clusters. For each comment $(u, v, b, \mathcal{W}_{v,b})$ in the testing set, 1) the ranking scores are computed for all POIs including the POI v; 2) a ranked list is obtained by ordering all the POIs, where p denotes the position of POI v in the list. The best result is v exceeding all unvisited POIs; 3) a top-k recommendation list consists of the first k POIs from the list. If $p \leq k$, a hit happens (i.e., the actual visited POI v is recommended to user u); otherwise, a miss case happens. *Accuracy@k* is defined as below:

$$Accuracy@k = \frac{|hit@k|}{|D_{test}|} \tag{13}$$

where $|hit@k|$ and $|D_{test}|$ represent the number of hits in the testing set and the number of all the test cases, respectively.

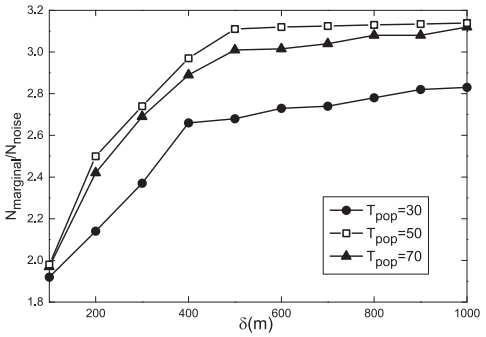


Fig. 8. Results of distance-based clustering method.

7.1.4. Parameter selection

The average number of reviews can indicate the popularity of an area, and we use the algorithm of P-DBSCAN in Section 5 to cluster the POIs in terms of their popularities. In order to analyze the effect of the proposed Algorithm 1 and specify the values of parameters, we use $N_{cluster}/N_{noise}$, i.e., the ratio of the average number of reviews in non-noisy area to that in noisy area, to measure the effectiveness of the clustering algorithm. Based on the Foursquare-Facebook network, we can obtain the appropriate value of ϵ by experiments, i.e., ϵ=60 m. Then we compare the ratios under different noise distance thresholds ranging from 100 m to 1000 m. Each curve specifies a threshold of the number of reviews. The results are given in Fig. 8.

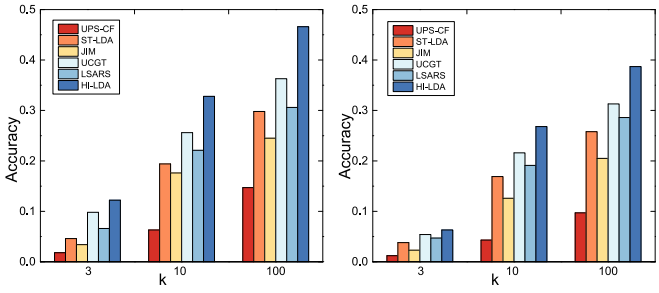
According to the results in Fig. 8, we can obtain the following observations:

- The popularity within the marginal regions grows with δ. It increases more rapidly with a threshold smaller than 400 m, which implies that some popular points are located outside the core regions and reside within a distance less than 400 m. The points far away from these regions are often small restaurants, such as noodle restaurants and cafes, which are convenient to nearby residents.
- Most popular points can be involved when the threshold of the review number $\mathcal{T}_{pop}$ is larger than 30, and most of ordinary points are viewed as marginal or noise points, which can degrade the effectiveness of clustering. Thus, we compare these three curves with $\mathcal{T}_{pop}$ larger than 30, and find that the appropriate value for $\mathcal{T}_{pop}$ is 50. This can be explained by the reason that a large $\mathcal{T}_{pop}$ excludes some relatively popular points, while a small $\mathcal{T}_{pop}$ might include a great number of unpopular points.

We have the same results in the Foursquare-Twitter network. Based on the aforementioned discussion, we specify δ=500 and $\mathcal{T}_{pop}$=50 with ϵ=60 in experiments.

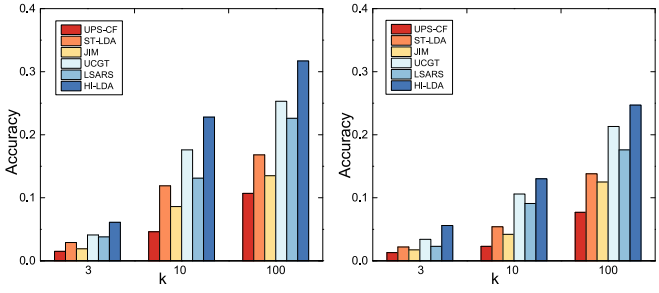
7.2. Recommendation effectiveness

In this section, we compare the performance of different POI recommendation approaches with well-tuned parameters.



(a) Home-town recommendation (b) Out-of-town recommendation

Fig. 9. Top-k recommendation performance on the Foursquare-Facebook network.



(a) Home-town recommendation (b) Out-of-town recommendation

Fig. 10. Top-k recommendation performance on the Foursquare-Twitter network.

Figs. 9 and 10 demonstrate the performance of POI recommendation with different k values in the Foursquare-Facebook network and the Foursquare-Twitter network, respectively. A larger k value is often ignored in top-k recommendation.

According to Figs. 9 and 10, we observe that HI-LDA outperforms the other baseline methods on both networks, and the gaps range from 0.9% to 36.9%, which indicates that the recommendation effectiveness can be significantly improved, by taking into full consideration the factors of geographical influence, users' profiles, social relationship, users' behaviors and textual effect simultaneously. And then, we come to the following conclusions:

(1) Due to the data sparsity, the CF-based method of UPS-CF performs worse than text-based methods, i.e., ST-LDA, JIM, UCGT, LSARS and HI-LDA, which shows that text information, such as reviews, profiles of users, descriptions of POIs in LBSNs and comments on POIs in CBSNs, are valuable information to alleviate the problem of data sparsity, especially for the out-of-town scenario. Specifically, the textual content of HI-LDA comes from another social network with users having frequent interactions, which provide a large amount of useful textual information. We can conclude that someone's friends or friends of a friend may visit the POIs, and the comments from them can show their preferences which are probably similar to the given user.
(2) Different from other methods only contains the interactive behavior of check-in, there are several categories of interactive behaviors in HI-LDA, including post, repost and reply, which give a more comprehensive description of a user. When a user travels out of town, recommendation information from his friends may help him make a good decision.
(3) HI-LDA and UCGT perform better than ST-LDA and JIM, which shows that the social influence is capable of significantly improving the accuracy of recommendation both in home-town and out-of-town scenarios, and we find that social relationship has a greater impact on users' decision-making than temporal effect in POI recommendation.
(4) We find that HI-LDA has a better recommendation performance in the Foursquare-Facebook network than the Foursquare-Twitter network. This is because these two CBSNs have different features such as user behaviors. Users often exchange messages, post status updates and share photos with their close friends by Facebook, and follow celebrities and pay attention to their status updates in Twitter. In addition, users tend to trust the friends in Facebook instead of the friends in Twitter.

7.3. Impact of different factors

In this section, we explore the effect of each factor in the HI-LDA model, including the effect of communities (F1), the effect of different behaviors over sentiments (F2), the effect of locations over sentiments (F3), the effect of region scope (F4), the effect of comments in CBSNs (F5) and the effect of multi-source data (F6). HI-LDA is compared with its six variant versions introduced in Section 7.1.2, and the comparison results are illustrated in Fig. 11.

As shown in Fig. 11, HI-LDA outperforms its six variant versions with the gap ranging from 1.5% to 37.0% for home-town recommendation, and ranging from 2.6% to 35.4% for out-of-town recommendation. The results indicate the benefit brought by each factor, respectively. For example, the performance gap between HI-LDA and HI-LDA-V2 validates the benefit of distinguishing the multiple categories of user behaviors in a specific region.

Each factor owns different contribution to improving the effectiveness of recommendation. Similarly, the same factor has different contributions to home-town and out-of-town recommendation. In terms of the importance of the six factors in the home-town recommendation scenario, they can be ranked as $F6 > F5 > F1 > F2 > F3 > F4$, while in the out-of-town recommendation scenario, they can be ranked as follows: $F6 > F5 > F3 > F4 > F1 > F2$. This can be explained by the reason that the two scenarios have different characteristics:

(1) Users in the their home towns are more likely to rely on their own experience, while the suggestion from close friends play an important role when they travel to out-of-town regions without any experience.
(2) The geographical effect is apparent in out-of-town regions because many POIs nearby are attractive to users and it is unnecessary to visit a POI far away except scenic spots and historical sites.

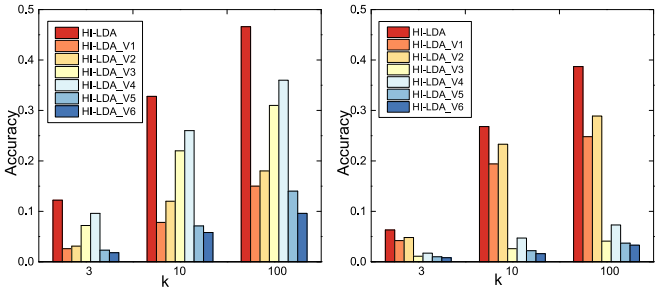
7.4. Analysis of parameter sensitivity

It is essential to choose appropriate parameters in HI-LDA to achieve better performance. In this section, we will analyze the sensitivity of parameters in HI-LDA.

Based on the aforementioned discussion, the hyper-parameters α, β, γ, τ, η and π are fixed as α=50/K, γ=50/R, η=50/C, $\beta = \tau = \pi$=0.01 based on experimental results. However, the performance of HI-LDA is highly sensitive to the number of regions, communities and topics. Thus, we evaluate the performance of HI-LDA by changing the number of regions, communities and topics.

Tables 6, 7, 8, 9 show the results of POI recommendation in the Foursquare-Facebook network.

Based on the results, we can see that the recommendation accuracy of HI-LDA initially increases with the number of topics, and then it drops slowly when the topics is more than 50. Similar observation can be obtained by increasing the number of regions, and we find that the recommendation accuracy of HI-LDA grows with the number of regions, and then it becomes stable when the regions is more than 60. A larger C value, which implies more communities w.r.t. a user, results in a higher recommendation accuracy. It is because that K, R and C are the parameters which affect



(a) Home-town recommendation (b) Out-of-town recommendation

Fig. 11. Impact of different factors on the Foursquare-Facebook network.

Table 6
Home-town recommendation accuracy (Foursquare-Facebook, C=50).

	$K=10$	$K=20$	$K=30$	$K=40$	$K=50$	$K=60$
$R=20$	0.242	0.263	0.295	0.304	0.301	0.301
$R=30$	0.246	0.272	0.313	0.337	0.333	0.333
$R=40$	0.249	0.283	0.337	0.374	0.368	0.368
$R=50$	0.265	0.306	0.364	0.386	0.381	0.381
$R=60$	0.274	0.313	0.382	0.401	0.395	0.398
$R=70$	0.274	0.313	0.382	0.402	0.395	0.398

Table 7
Out-of-town recommendation accuracy (Foursquare-Facebook, C=40).

	$K=10$	$K=20$	$K=30$	$K=40$	$K=50$	$K=60$
$R=20$	0.307	0.328	0.361	0.406	0.401	0.401
$R=30$	0.321	0.343	0.373	0.432	0.428	0.428
$R=40$	0.326	0.355	0.386	0.448	0.437	0.437
$R=50$	0.336	0.368	0.406	0.468	0.457	0.457
$R=60$	0.339	0.385	0.416	0.479	0.472	0.472
$R=70$	0.339	0.421	0.462	0.478	0.474	0.474

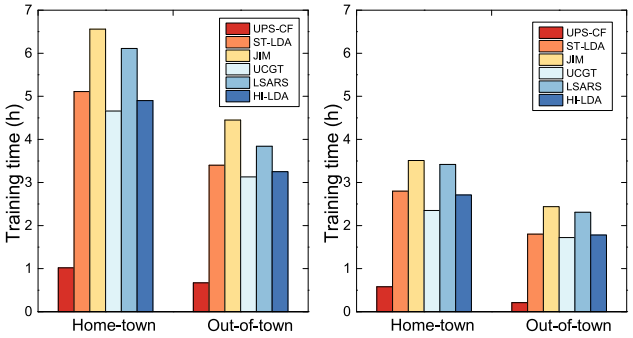
Table 8
Home-town recommendation accuracy (Foursquare-Facebook, $R=60$).

	$K=10$	$K=20$	$K=30$	$K=40$	$K=50$	$K=60$
$C=20$	0.246	0.272	0.343	0.364	0.356	0.356
$C=30$	0.258	0.283	0.356	0.374	0.368	0.368
$C=40$	0.264	0.295	0.367	0.386	0.381	0.381
$C=50$	0.275	0.314	0.390	0.405	0.398	0.398
$C=60$	0.280	0.315	0.388	0.407	0.403	0.403

Table 9
Out-of-town recommendation accuracy (Foursquare-Facebook, R=60).

	$K=10$	$K=20$	$K=30$	$K=40$	$K=50$	$K=60$
$C=20$	0.336	0.371	0.404	0.442	0.438	0.438
$C=30$	0.345	0.384	0.406	0.467	0.463	0.463
$C=40$	0.360	0.397	0.427	0.485	0.483	0.483
$C=50$	0.366	0.405	0.433	0.491	0.488	0.488
$C=60$	0.366	0.405	0.433	0.491	0.488	0.488

the complexity of the model. When K, R and C are too small, i.e., smaller than the real values, the model has limited capability of describing the data, e.g., the inherent relationship among the communities, regions and textual sentiments. In this case, the accuracy of recommendation is reduced. When each of these parameters exceeds a threshold, the proposed model is comprehensive enough to describe the data, and we do not need to increase K, R and C for improving the recommendation accuracy. Finally, the optimal parameter combination for POI recommendation is $K = 40$, $R = 60$, and C is specified to 50 for the home-town scenario, and 40 for the out-of-town scenario.



(a) The Foursquare-Facebook network (b) The Foursquare-Twitter network

Fig. 12. Comparison of training time among algorithms.

7.5. Recommendation efficiency

In this section, we evaluate the efficiency of our proposed model by comparing it with the baseline methods in the Foursquare-Facebook and Foursquare-Twitter heterogeneous networks. All the experiments are conducted in the environment with Java (JDK 1.7), Windows 10, and run on a PC with a CPU of Core i7 (7500U) and 8GB RAM. The comparison results of the average training time for these methods are given in Fig. 12 with the parameter K=10.

As shown in Fig. 12, most models have similar training time except the UPS-CF method. This is because the computational complexity of UPS-CF is $\mathcal{O}(n^2)$. However, all of the other models mainly contain two costly operations: iterative sampling and textual analysis, and the computational complexity of each model is $\mathcal{O}(n^3)$. The cost of training time in the Foursquare-Twitter network is about half of that in the Foursquare-Facebook network. This can be explained by the reason that the Foursquare-Facebook network has more users, comments and POIs, which greatly increases the training time.

8. Conclusion

POI Recommendation is a very challenging and difficult problem in LBSNs. In order to effectively suggest POIs in LBSNs, we model the factors including social communities and geographical regions, and textual comments in CBSNs. To the best of our knowledge, most of social networks especially CBSNs are categorized into a number of social communities. And, users in one of the communities probably visit the same locations because of their frequent online communication. A preferable POI recommendation for a user in specific regions can be inferred from comment words of his friends in the same community. Based on the above consider-

ations, we proposed a latent probabilistic generative model called HI-LDA, which can accurately capture comment words in CBSNs by taking into account social communities, interactive behaviors and the geographical information. The important parameters of the HI-LDA model are obtained by the Gibbs sampling approach. Lastly, we introduced a framework for POI recommendation including a geographical clustering algorithm based on the locations and popularity of POIs. We conducted extensive experiments to evaluate the performance of the proposed HI-LDA model on two heterogeneous social networks, and the results demonstrate that our approach outperforms the state-of-the-art baseline methods in effectiveness and efficiency of POI recommendation.

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. .

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61772091 61802035 and 61962006 the Youth Foundation for Humanities and Social Sciences of Ministry of Education of China under Grant No. 17YJCZH202; the Sichuan Science and Technology Program under Grant Nos. 2018JY0448, 2019YFG0106, 2019YFS0067 and 2018GZ0253; the Natural Science Foundation of Guangxi under Grant No. 2018GXNSFDA138005; the Innovative Research Team Construction Plan in Universities of Sichuan Province under Grant No. 18TD0027; the Scientific Research Foundation for Advanced Talents of Chengdu University of Information Technology under Grant Nos. KYTZ201715 and KYTZ201750; the Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology under Grant No. J201701; Guangdong Key Laboratory Project under Grant No. 2017B030314073.

References

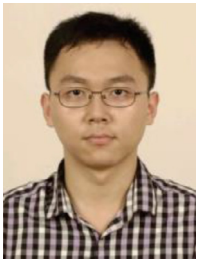
[1] S. Qiao, N. Han, J. Zhou, R.-H. Li, C. Jin, L.A. Gutierrez, SocialMix: a familiarity-based and preference-aware location suggestion approach, Eng. Appl. Artif. Intell. 68 (2018) 192–204.
[2] R. Baral, T. Li, Exploiting the roles of aspects in personalized POI recommender systems, Data Min. Knowl. Discov. 32 (4) (2018) 1–24.
[3] S. Zhao, I. King, M.R. Lyu, A survey of point-of-interest recommendation in location-based social networks, arXiv:1607.00647 (2016).
[4] Z. Li, F. Xiong, X. Wang, H. Chen, X. Xiong, Topological influence-aware recommendation on social networks, Complexity 2019 (2019) 6325654.1–12.
[5] C. Shi, Y. Li, J. Zhang, Y. Sun, P.S. Yu, A survey of heterogeneous information network analysis, IEEE Trans. Knowl. Data Eng. 29 (1) (2017) 17–37.
[6] H. Wang, Y. Fu, Q. Wang, H. Yin, C. Du, H. Xiong, A location-sentiment-aware recommender system for both home-town and out-of–town users, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1135–1143.
[7] H. Yin, Y. Sun, B. Cui, Z. Hu, L. Chen, Lcars: a location-content-aware recommender system, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 221–229.
[8] Q. Zhang, D. Wu, G. Zhang, J. Lu, Fuzzy user-interest drift detection based recommender systems, in: IEEE International Conference on Fuzzy Systems, 2016, pp. 1274–1281.
[9] X. Zhang, J. Cheng, T. Yuan, B. Niu, H. Lu, Toprec:domain-specific recommendation through community topic mining in social network, in: International Conference on World Wide Web, 2013, pp. 1501–1510.
[10] X. Xiong, S. Qiao, Y. Li, H. Zhang, P. Huang, N. Han, R.-H. Li, ADPDF: a hybrid attribute discrimination method for psychometric data with fuzziness, IEEE Trans. Syst., Man, Cybern. 49 (1) (2019) 265–278.
[11] X. Xiong, S. Qiao, Y. Li, F. Xiong, L. He, N. Han, Affective impression: sentiment-awareness POI suggestion via embedding in heterogeneous lbsns, IEEE Trans. Affect. Comput. (2019). 1–1
[12] B. Hu, M. Ester, Spatial topic modeling in online social media for location recommendation, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, 2013, pp. 25–32.
[13] H. Yin, X. Zhou, Y. Shao, H. Wang, S. Sadiq, Joint modeling of user check-in behaviors for point-of-interest recommendation, ACM Trans. Inf. Syst. 35 (2) (2016).
[14] B. Liu, H. Xiong, Point-of-interest recommendation in location based social networks with topic and location awareness, in: Proceedings of the 2013 SIAM International Conference on Data Mining, 2013.
[15] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, J. Tang, How do your friends on social media disclose your emotions? in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 306–312.
[16] X. Han, L. Wang, R. Farahbakhsh, n. Cuevas, R. Cuevas, N. Crespi, L. He, Csd: a multi-user similarity metric for community recommendation in online social networks, Expert Syst. Appl. 53 (2016) 14–26.
[17] T. Yuan, J. Cheng, X. Zhang, Q. Liu, H. Lu, How friends affect user behaviors? an exploration of social relation analysis for recommendation, Knowl. Based Syst. 88 (C) (2015) 70–84.
[18] G. Ference, M. Ye, W.C. Lee, Location recommendation for out-of-town users in location-based social networks, in: ACM International Conference on Information and Knowledge Management, 2013, pp. 721–726.
[19] H. Yin, X. Zhou, B. Cui, H. Wang, K. Zheng, Q.V.H. Nguyen, Adapting to user interest drift for POI recommendation, IEEE Trans. Knowl. Data Eng. 28 (10) (2016) 2566–2581.
[20] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, Q.V.H. Nguyen, S. Sadiq, Discovering Interpretable Geo-social Communities for User Behavior Prediction, in: IEEE International Conference on Data Engineering, 2016, pp. 942–953.
[21] X. Xiong, Y. Li, S. Qiao, N. Han, Y. Wu, J. Peng, B. Li, An emotional contagion model for heterogeneous social media with multiple behaviors, Physica A 490 (2018) 185–202.
[22] J. Zhang, X. Kong, P.S. Yu, Transferring heterogeneous links across location-based social networks, in: ACM International Conference on Web Search and Data Mining, 2014, pp. 303–312.
[23] S. Sajadmanesh, H.R. Rabiee, A. Khodadadi, Predicting anchor links between heterogeneous social networks, in: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2016, pp. 158–163.
[24] W. Min, B.K. Bao, C. Xu, M.S. Hossain, Cross-platform multi-modal topic modeling for personalized inter-platform recommendation, IEEE Trans. Multimedia 17 (10) (2015) 1787–1801.
[25] S. Wang, X. Hu, P.S. Yu, Z. Li, Mmrate:inferring multi-aspect diffusion networks with multi-pattern cascades, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1246–1255.
[26] S. Qiao, D. Shen, X. Wang, N. Han, W. Zhu, A self-adaptive parameter selection trajectory prediction approach via hidden markov models, IEEE Trans. Intell. Transp. Syst. 16 (1) (2015) 284–296.
[27] S. Qiao, N. Han, W. Zhu, L.A. Gutierrez, Traplan: an effective three-in-one trajectory-prediction model in transportation networks, IEEE Trans. Intell. Transp. Syst. 16 (3) (2015) 1188–1198.
[28] S. Scellato, A. Noulas, R. Lambiotte, C. Mascolo, Socio-spatial properties of online location-based social networks, Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[29] X. Ma, D. Di, W. Quan, Community detection in multi-layer networks using joint nonnegative matrix factorization, IEEE Trans. Knowl. Data Eng. PP (99) (2019). 1–1
[30] J. Sander, M. Ester, H. Kriegel, X. Xu, Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications, Data Min. Knowl. Discov. 2 (2) (1998) 169–194.
[31] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2012) 993–1022.
[32] A.D. Bolton, N.A. Heard, Malware family discovery using reversible jump MCMC sampling of regimes, J. Am. Stat. Assoc. (4) (2018).
[33] Y. Papanikolaou, J.R. Foulds, T.N. Rubin, G. Tsoumakas, Dense distributions from sparse samples: improved gibbs sampling parameter estimators for LDA, Statistics 18 (62) (2017) 1–58.
[34] W. Wang, H. Yin, L. Chen, Y. Sun, S. Sadiq, X. Zhou, Geo-SAGE: A geographical sparse additive generative model for spatial item recommendation, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1255–1264.

Xi Xiong received the B.S. and M.S. degrees from the Beijing Institute of Technology and the Ph.D. degree in information security from Sichuan University, Chengdu, in 2013. He is currently a lecturer with the School of Cybersecurity, Chengdu University of Information Technology, Chengdu, China. He has published over 20 papers in the most prestigious journals and conferences. His research interests include web mining, social computing and machine learning.

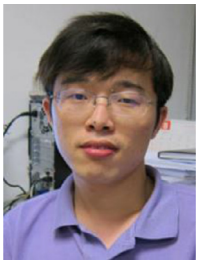
Shaojie Qiao received the B.S. and Ph.D. degrees from Sichuan University, Chengdu, China, in 2004 and 2009, respectively. From 2007 to 2008, He worked as a visiting scholar in the School of Computing at the National University of Singapore. He is currently a Professor with the School of Software Engineering, Chengdu University of Information Technology, Chengdu, China. He has led several research projects in the areas of databases and data mining. He has authored more than 40 high quality papers, and coauthored more than 90 papers. His research interests include location-based social networks and trajectory data mining.

Nan Han received the M.S. and Ph.D. degrees from Chengdu University of Traditional Chinese Medicine, Chengdu, China. She is currently an associate professor with the School of Management, Chengdu University of Information Technology. Her research interests include trajectory prediction and data mining.

Fei Xiong received the B.E. degree and the Ph.D. degree in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2007 and 2013. He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. From 2011 to 2012, he was a visiting scholar at Carnegie Mellon University. He has published over 60 papers in refereed journals and conference proceedings. He was a recipient of National Natural Science Foundations of China and several other research grants. His current research interests include the areas of web mining, complex networks and complex systems.



Zhan Bu received his Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2014. He is currently an Associate Professor at Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China. His recent research interests include complex network, data mining and game theory.



Rong-Hua Li received the Ph.D. degree from the Chinese University of Hong Kong in 2013. He is currently an associate professor at Beijing Institute of Technology, China. His research interests include algorithmic aspects of social network analysis, graph data management and mining, as well as sequence data management and mining.

Kun Yue received his M.S. degree in computer science from Fudan University in 2004, and received his Ph.D. degree in computer science from Yunnan University in 2009. He is currently a professor of School of Information Science and Engineering at Yunnan University. His research interests include massive data analysis, artificial intelligence, and knowledge engineering. He is the director of Chinese Association for Artificial Intelligence (CAAI), vice Chairman designate of CAAI Uncertainty in Artificial Intelligence Society.

Guan Yuan received the M. S. and Ph.D. degrees in Computer Science and Technology from China University of Mining and Technology in 2009 and 2012, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, China University of Mining and Technology. His research areas include spatial-temporal data mining and social networks.