Knowl Inf Syst (2017) 53:153–177
DOI 10.1007/s10115-017-1036-2

REGULAR PAPER

Spatial and semantical label inference for social media

A cross-network data fusion approach

Yuchi Ma[1] · **Ning Yang**[1] · **Lei Zhang**[1] · **Philip S. Yu**[2,3]

Received: 5 February 2016 / Revised: 21 February 2017 / Accepted: 21 February 2017 /
Published online: 3 March 2017

Abstract Exploring the spatial and semantical knowledge from messages in social media offers us an opportunity to get a deeper understanding about the mobility and activity of users, which can be leveraged to improve the service quality of online applications like recommender systems. In this paper, we investigate the problem of the spatial and semantical label inference, where the challenges come from three aspects: diverse heterogeneous information, uncertainty of individual mobility, and large-scale sparse data. We address the challenges by exploring two types of data fusion, the fusion of heterogeneous social networks and the fusion of heterogeneous features. We build a 4-dimensional tensor, called spatial–temporal semantical tensor (STST), to model the individual mobility and activity by fusing two heterogeneous social networks, a social media network and a location-based social network (LBSN). To address the challenge arising from diverse heterogeneous information and the uncertainty of individual mobility, we construct three types of heterogeneous features and fuse them with STST by exploring their interdependency relationships. Particularly, a

Y. Ma and N. Yang: These authors contributed equally to this study and share first authorship.

This work is supported by National Science Foundation of China through Grant 61173099, the Basic Research Program of Sichuan Province with Grant 2014JY0220, and NSF through Grants CNS-1115234, DBI-0960443, OISE-1129076, IIS-1526499, CNS-1626432, and NSFC 61672313.

✉ Yuchi Ma
scu.Richard.Ma@gmail.com

✉ Ning Yang
yangning@scu.edu.cn

Lei Zhang
Leizhang@scu.edu.cn

Philip S. Yu
psyu@uic.edu

1 College of Computer Science, Sichuan University, Chengdu, China

2 Department of Computer Science, University of Illinois at Chicago, Chicago, USA

3 Institute for Data Science, Tsinghua University, Beijing, China

spatial tendency feature is constructed to constrain the inference of individual mobility and reduce the uncertainty. To deal with large-scale sparse data, we propose a parallel contextual tensor factorization (PCTF) to concurrently factorize STST. Finally, we integrate these components into an inference framework, called spatial and semantical label inference SSLI. The results of extensive experiments conducted on real datasets and synthetic datasets verify the effectiveness and efficiency of SSLI.

Keywords Heterogeneous Social Networks · Social Media · Tensor Decomposition · Data Fusion

1 Introduction

Social media, such as Twitter and Weibo, are platforms for users to share news or their stories with their friends, which have become part of our daily life. Exploring the spatial and semantical knowledge from the messages in social media is an important task. For example, if we know the location and activity type of a user through his/her messages, we can timely make a contextual and personalized recommendation to him/her. So far, existing researches can be grouped into two categories, factorization-based methods [7,12,21,28,29] and probability graph-based methods [17,18,24]. Some of these works only focus on the prediction of locations, but ignore the activities. Some other works propose methods for prediction of both the locations and the activities, but fail to take into consideration the interrelation among users, locations, activities, and time, such as social relationship and geographical features. Besides, the messages in social media usually have no spatial and semantical labels. As a result, in this paper, we aim at the problem of inferring hidden spatial and semantical labels of the messages in social media, which is not easy due to the following challenges:

- *Diverse heterogeneous information* There are diverse heterogeneous information that can be collected, e.g., social links among individuals, check-in information on each individual, geographical information on regions, including POIs in each region, and neighboring relationship between regions. It is a challenge on how to utilize this information to construct meaningful features and relationships to help improve the spatial and semantical label prediction.
- *Uncertainty of individual mobility* Gonzalez et al. [6] find that an individual's mobility usually rotates around at a few previously visited locations. However, individual mobility is affected by many factors, such as social circles. As a result, it is difficult to make an accurate inference for an individual only based on his/her own messages in social media.
- *Large-scale sparse data* The volume of social media data considered here is huge, which makes the traditional methods impractical. Besides, few users have messages with spatial and semantical labels. The proportion of these labeled messages is very low compared with their messages without labels, which makes it difficult to make an accurate inference.

In this paper, we address the challenges by exploring two types of data fusion, the fusion of heterogeneous features and the fusion of heterogeneous social networks. At first, to model the mobility and activity of users, we fuse two types of heterogeneous social networks, a social media network and a location-based social network (LBSN), to build a 4-dimensional tensor, called spatial–temporal temantical tensor (STST).

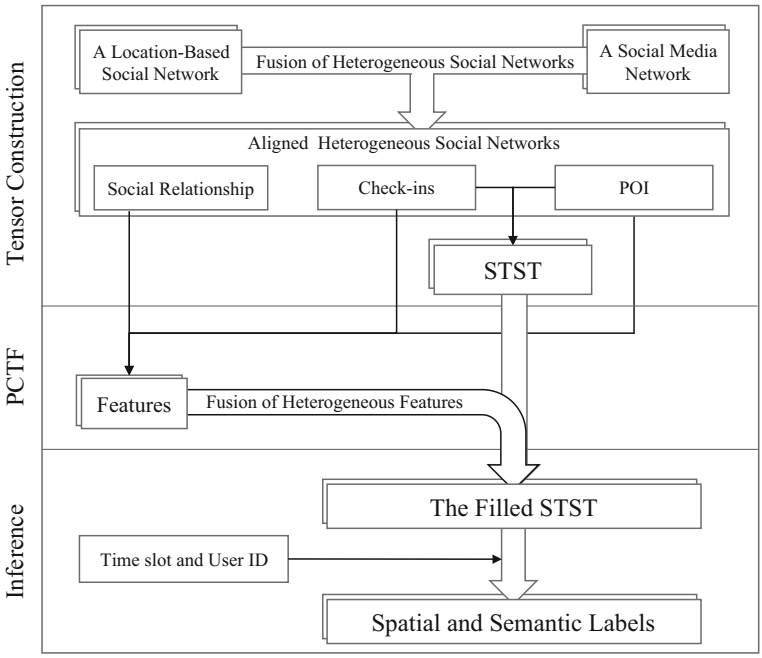


Fig. 1 Schematic of the proposed Spatial and semantical label inference (SSLI) framework

To address the challenge of diverse heterogeneous information, we construct meaningful features from three types of heterogeneous information, social relationship feature, geographical feature, and spatial tendency feature and fuse them with STST by exploring their interdependency relationships, where the spatial tendency feature is constructed to constrain the inference of individual mobility and reduce the uncertainty. To deal with the large-scale sparse data, we propose a parallel factorizing algorithm, called parallel contextual tensor factorization (PCTF), to reconstruct the missing entries of STST in a parallel fashion.

As shown in Fig. 1, we integrate our ideas into a unified framework, called spatial and semantical label inference (SSLI), which consists of three stages. At the first stage, we construct the STST based on check-ins and the information of point of interest (POI) from an LBSN and a social media network. At the second stage, we utilize PCTF to factorize and fill the STST. Particularly, during the factorization, PCTF explores three types of relationships by identifying additional types of information not captured in the STST model to construct three types of features and fuses them with STST. The three types of features include social relationship feature, geographical feature, and spatial tendency feature. At the third stage, we make spatial and semantical label inference based on the filled STST. Our main contributions can be summarized as follows:

1. We propose a novel inference framework called spatial and semantical label inference (SSLI). SSLI is able to make an inference with better accuracy through a data fusion approach as compared to existing approaches.
2. To fulfill SSLI, we model the mobility and activity of individuals as a 4-dimensional tensor, called spatial–temporal semantical tensor (STST), by fusing two heterogeneous social networks, a social media network and a location-based social network (LBSN).

We also propose a parallel contextual tensor factorization (PCTF) algorithm which can reconstruct a sparse STST with a divide-and-conquer strategy.
3. With the diverse heterogeneous information that can be collected, we devise a novel way to utilize this information to construct meaningful features and relationships to help improve the spatial and semantical label prediction. This is achieved via constructing three types of novel features and fusing them with STST by exploring their interdependency relationships. Particularly, a spatial tendency feature is constructed to constrain the inference of individual mobility to reduce the uncertainty.
4. We compare SSLI with four baseline methods. The results verify the effectiveness and efficiency of SSLI.

The rest of this paper is organized as follows. We give the notations and preliminaries in Sect. 2. We construct STST in Sect. 3 and describe the fusion of heterogeneous features in Sect. 4. We propose PCTF and the inference model of SSLI in Sects. 5 and 6 respectively. We present the experimental results and analysis in Sect. 7. Finally, we discuss related works in Sect. 8 and conclude in Sect. 9.

2 Notations and preliminaries

2.1 Notations

In this paper, a scalar is denoted by a capital letter (e.g., $N \in \mathbb{R}$) and a vector is denoted by a italic boldface lowercase letter (e.g., $\boldsymbol{v} \in \mathbb{R}^{N\times 1}$). A matrix is denoted by an italic boldface capital letter (e.g., $\boldsymbol{X} \in \mathbb{R}^{N\times M}$), and a tensor is denoted by a calligraphy capital letter (e.g., $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times L\times M\times K}$), where the elements in $\boldsymbol{X}$ and $\boldsymbol{\mathcal{A}}$ are denoted by x_{nm}, a_{nlmk} respectively. A set is denoted by a bold capital letter (e.g., $\mathbf{G}$).

2.2 Preliminaries

Definition 1 (*Heterogeneous social network*) A social network is heterogeneous if it contains multiple kinds of nodes and links. Heterogeneous social networks can be represented as $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \bigcup_i \mathbf{V}^i$ is the union of different node sets and $\mathbf{E} = \bigcup_i \mathbf{E}^i$ is the union of heterogeneous link sets.

Definition 2 $(Venue)$ Venues are certified locations that allow users checked in on a location-based social network (LBSN) like Foursquare. When a user checks in at a venue, the category of the venue can be considered as the activity of the user. (e.g., a user is shopping while checking in at a supermarket).

Note that we use the terms activity and activity type interchangeably in this paper.

Definition 3 (Tip) In a location-based social network (LBSN), users can post comments with time stamps and location labels when they check-in at certificated venues, where the comments are also called tips. The set of tips is denoted by $\mathbf{W}$, where a tip, denoted by w, consists of three parts, the user $w.u$, the venue $w.v$, and the time slot of the check-in time stamp $w.t$.

3 Spatial–temporal semantical tensor

In this section, we model user mobility and activity as a 4-dimensional tensor, called spatial–temporal semantical tensor (STST), by fusing two heterogeneous social networks, a social media network and a location-based social network (LBSN).

3.1 Fusion of heterogeneous social networks

In social media networks like Twitter, most of messages have no spatial and semantical labels, while in LBSNs, users often share venues and post comments for the venues. The tips on LBSNs contain not only text and time stamps, but also the information of venues including categories and locations. It gives us an opportunity to infer spatial and semantical labels by applying the fusion of heterogeneous social networks, which aligns users between the social media network and the LBSN.

Since network alignment is an independent subject and not the focus of this paper, we employ the method proposed by Zhang et al. [27], called multi-network link identifier (MLI), to align the users across a social media network and an LBSN. MLI can be classified as a transfer learning-based method according to the data fusion category summarized by Zheng [30]. MLI is proposed to solve the multi-network link prediction problem based on the heterogeneous topological features that are extracted from the selected "social meta paths" in aligned heterogeneous social networks. MLI introduces a multi-PU link prediction framework to predict the anchor links among social networks, where the features are defined by meta-path. The definition of aligned heterogeneous social networks is given as follow:

Definition 4 (*Aligned heterogeneous social networks*) [27]

If two different heterogeneous social networks share some common users, then the two networks are called aligned networks. Multiple aligned heterogeneous social networks are formulated as $\mathbf{A}_{\mathrm{G}} = ((\mathbf{G}^1, \mathbf{G}^2, ..., \mathbf{G}^n), (\mathbf{A}^{1,2}, \mathbf{A}^{1,3}, ..., \mathbf{A}^{1,n}, \mathbf{A}^{2,3}, ..., \mathbf{A}^{(n-1),n}))$. $\mathbf{G}^i, i \in \{1, 2, ..., n\}$ is a heterogeneous social network. $\mathbf{A}^{i,j} \neq \varnothing, i, j \in \{1, 2, ..., n\}$, is the set of anchor links between $\mathbf{G}^i$ and $\mathbf{G}^j$, where **an anchor link** is an undirected link between $\mathbf{G}^i$ and $\mathbf{G}^j$ iff $(u_i \in U^i) \wedge (v_j \in U^j)$ (u_i and v_j are the accounts of the same user in $\mathbf{G}^i$ and $\mathbf{G}^j$, respectively).

As Fig. 2 shows the social media network account *User B* has no spatial and semantical information. After aligning the LBSN account *User B'* with the account *User B*, we can transfer his/her spatial and semantical information from the LBSN to the social media network, which makes it possible to model the mobility and activity of *User B*.

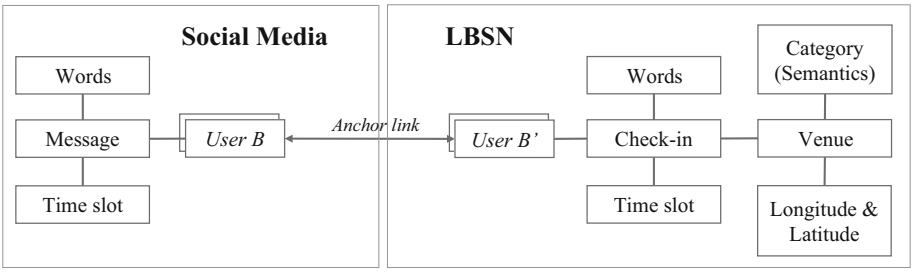


Fig. 2 Aligned heterogeneous social networks

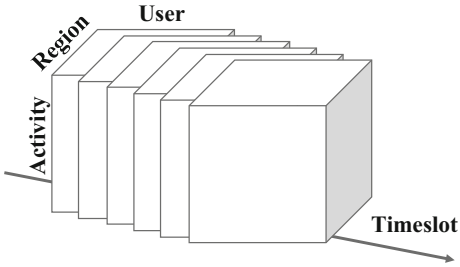


Fig. 3 An illustration of STST

3.2 Model of the mobility and activity

After aligning users, we model the mobility and activity of users as a 4-dimensional tensor, which is defined as follow:

Definition 5 (*Spatial–temporal semantical tensor* (*STST*))
An STST is a tensor $\mathcal{A} \in \mathbb{R}^{N \times L \times M \times K}$ of four dimensions, respectively, representing users, regions, activities, and time slots, where N, L, M, K are the numbers of users, activities, regions, time slots, respectively. An entry $a_{nlmk} \in \mathcal{A}$ stores the number of tips posted by user n on a specific activity l that happens in a specific region m at a specific time slot k, where a time slot represents 1 h of a day.

STST models users from three aspects, namely the spatiality, the temporality, and the semantics, where the semantics is represented by the category of venue which can be taken as the user-activity. As Fig. 3 shows the STST can be understood as a series of user tensors with three dimensions along the user dimension. An entry in a user tensor records how often the user does a specific activity in a specific region at a specific time slot.

4 Fusion of heterogeneous features

Although using tensor factorization to identify latent features have been applied in many applications [16,22,31], here we attempt to take advantage of the diverse heterogeneous information that can be collected, e.g., social links among individuals, check-in information of each individual, geographical information on regions, including POIs in each region, and neighboring relationship between regions to help address the data sparsity and noise issues. Particularly, in order to reduce the uncertainty of individual mobility, a spatial tendency feature is constructed.

4.1 Social relationship feature

With the help of user alignment, we are able to use not only the explicit social links in the social media network, but also the implicit social links across the social media network and the LBSN, to extract social relationship features for users. Figure 4 shows an example of the extraction of social relationship feature from aligned users. As shown in Fig. 4, we can figure out users' both one-step friendships and multi-step friendships across two heterogeneous social networks. We represent the social relationships of users by a social relationship feature matrix, which is defined as follow:

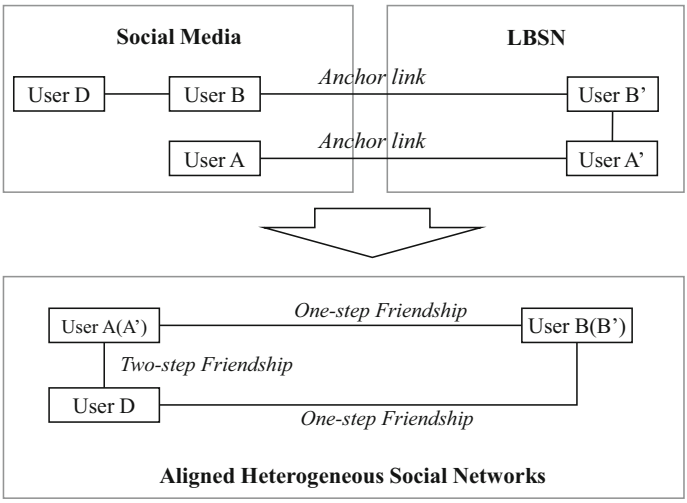


Fig. 4 Extracting social relationship feature from aligned heterogeneous social networks

Definition 6 (*Social relationship feature matrix*)
For a set of users in aligned heterogeneous social networks $\mathbf{A}_{\mathrm{G}}$, their social relationship feature matrix, denoted by $\boldsymbol{U}_{\mathrm{S}} \in \mathbb{R}^{N \times N}$, is a matrix in which an entry $u_s^{(n_1 n_2)}$ is the weighted distance between user n_1 and user n_2:

$$u_s^{(n_1 n_2)} = \begin{cases} 1 & SDis(n_1, n_2) = 0 \\ 0.1 & SDis(n_1, n_2) = 1 \\ 0.01 & SDis(n_1, n_2) = 2 \\ 0 & SDis(n_1, n_2) > 2 \end{cases}$$

where N is the number of individuals and $SDis(n_1, n_2)$ counts the steps between user n_1 and user n_2.

For example, as shown in Fig. 4, $SDis$(User A, User A) $= 0$, so $u_s(UserAUserA) = 1$; $SDis$(User A, User D) $= 2$ after aligning users, so $u_s(UserAUserD) = 0.01$.

Intuitively, the social relationships of individuals are affected by their social preference and the features of their latent friends. Inspired by this idea, we can approximate $\boldsymbol{U}_{\mathrm{S}} \in \mathbb{R}^{N \times N}$ by

$$\boldsymbol{U}_{\mathrm{S}} \approx \boldsymbol{U} \times \boldsymbol{U}_{\mathrm{L}}^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{U}$ is the user latent feature matrix, $\boldsymbol{U}_{\mathrm{L}} \in \mathbb{R}^{N \times D}$ is the social preference latent feature matrix, and D is the number of latent features.

4.2 Spatial tendency feature

In order to reduce the uncertainty of individual mobility, we construct the spatial tendency feature. Intuitively, an individual's mobility usually rotates around at a few previously visited locations [6]. The individual's mobility is also affected by his/her own social circle (e.g., a user may check-in at a restaurant nearby his/her friends' home for meeting). Inspired by these intuitions, we define the spatial tendency feature matrix as follow:

Definition 7 (*Spatial tendency feature matrix*)
For a set of individuals and a set of regions, their spatial tendency feature matrix, denoted by $\boldsymbol{S}_{\mathrm{T}} \in \mathbb{R}^{N \times M}$, is defined as

$$\boldsymbol{S}_{\mathrm{T}} = \boldsymbol{U}_{\mathrm{S}} \times \boldsymbol{U}_{\mathrm{R}}, \tag{2}$$

where $\boldsymbol{U}_{\mathrm{S}}$ is the social relationship feature matrix defined in Definition 6, and $\boldsymbol{U}_{\mathrm{R}}$ is the User-Region matrix in which an entry $u_r^{(nm)} = \sum_{l=1}^{L} \sum_{k=1}^{K} a_{nlmk}, a_{nlmk} \in \boldsymbol{\mathcal{A}}$, stores the number of check-ins of a user n at a region m.

4.3 Geographical features

Regions are divided in terms of administrative divisions. The spatial label of a tip can be represented by the corresponding region index. We build eight geographical features for a region, which are described as follows:

Popularity: For a region r_m, its popularity, denoted by $R_1^{(r_m)}$, is defined as the total check-in number in region r_m.

Repeat business: For a region r_m, its repeat business, denoted by $R_2^{(r_m)}$, is defined as the average check-in number over people who check-in at r_m. **Richness**: For a region r_m, its richness, denoted by $R_3^{(r_m)}$, is defined as the number of venue types in region r_m.

Density: For a region r_m, its density, denoted by $R_4^{(r_m)}$, is defined as the number of venues in region r_m.

The above four geographical features reflect the prosperity of a region from the four aspects.

Region entropy: For a region r_m, its region entropy, denoted by $R_5^{(r_m)}$, is defined as the heterogeneity of the venue types in region r_m:

$$R_5^{(r_m)} = -\sum_{\rho} \frac{M_\rho^{(r_m)}}{R_4^{(r_m)}} \times log \frac{M_\rho^{(r_m)}}{R_4^{(r_m)}},$$

where ρ is the type of venue, $R_4^{(r_m)}$ is the density of region m, $M_\rho^{(r_m)}$ is the number of venues of type ρ in region r_m.

This feature assesses the influence of the spatial heterogeneity of a region [10]. For example, the region entropy of a business center might be higher than the region entropy of a residential area, since the venues in the business center might be more diversified.

Inward flow: For a region r_m, its inward flow, denoted by $R_6^{(r_m)}$, is defined as the total inward flow that users transfer from outside to region r_m within time threshold T_{h}:

$$R_6^{(r_m)} = |\{(c_1, c_2) \mid c_1.u = c_2.u, c_1.v \notin r_m, c_2.v \in r_m, 0 < c_1.t - c_2.t < T_{\mathrm{h}}\}|.$$

This feature reflects whether a region is attractive to users [10]. Here the idea is that regions with high inward flow are mostly the last places where the users check-in at, which means these places can really retain consumers.

Outward flow: For a region r_m, its outward flow, denoted by $R_7^{(r_m)}$, is defined as the total outward flow that users transfer from region r_m to outside within time threshold T_{h}:

$$R_7^{(r_m)} = |\{(c_1, c_2) \mid c_1.u = c_2.u, c_1.v \in r_m, c_2.v \notin r_m, 0 < c_1.t - c_2.t < T_{\mathrm{h}}\}|.$$

A region with high outward flow suggests that the region is a springboard. For example, when a region contains a station or airport, the region might have a high outward flow.

Transition pair: For a region r_m, its transition pair, denoted by $R_8^{(r_m)}$, is defined as the total number of transition pairs that users transfer between venues in region r_m within time threshold T_h:

$$R_8^{(r_m)} = |\{(c_1, c_2) \mid c_1.u = c_2.u, c_1.v \in r_m, c_2.v \in r_m, 0 < c_1.t - c_2.t < T_\mathrm{h}\}|.$$

This feature also accesses the prosperity of a region. A region with massive transition pairs indicates that the region contains abundant and various venues.

Based on the eight geographical features, we can represent a set of regions by a geographical feature matrix, which is defined as follow:

Definition 8 (*Geographical feature matrix*)
For a set of regions, $\{r_1, ..., r_m, ..., r_M\}$, their geographical feature matrix, denoted by $\boldsymbol{R}_\mathrm{F} \in \mathbb{R}^{M\times 8}$, is defined as

$$\boldsymbol{R}_\mathrm{F} = \begin{Bmatrix} R_1^{(r_1)} & R_2^{(r_1)} & R_3^{(r_1)} & R_4^{(r_1)} & R_5^{(r_1)} & R_6^{(r_1)} & R_7^{(r_1)} & R_8^{(r_1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ R_1^{(r_m)} & R_2^{(r_m)} & R_3^{(r_m)} & R_4^{(r_m)} & R_5^{(r_m)} & R_6^{(r_m)} & R_7^{(r_m)} & R_8^{(r_m)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ R_1^{(r_M)} & R_2^{(r_M)} & R_3^{(r_M)} & R_4^{(r_M)} & R_5^{(r_M)} & R_6^{(r_M)} & R_7^{(r_M)} & R_8^{(r_M)} \end{Bmatrix}$$

where M is the total number of regions.

Intuitively, the geographical features of regions are affected by the regions and the features themselves. Inspired by this idea, we can factorize $\boldsymbol{R}_\mathrm{F} \in \mathbb{R}^{M\times 8}$ as follow:

$$\boldsymbol{R}_\mathrm{F} \approx \boldsymbol{R} \times \boldsymbol{R}_\mathrm{L}, \tag{3}$$

where $\boldsymbol{R} \in \mathbb{R}^{M\times D}$ is a region latent feature matrix, $\boldsymbol{R}_\mathrm{L} \in \mathbb{R}^{D\times 8}$ is a latent feature matrix of geographical features, and D is the number of latent features.

5 Parallel contextual tensor factorization

We describe the parallel contextual tensor factorization (PCTF) in this section, and summarize the notations in Table 1. PCTF fuses the contexts defined in last section to a tensor factorization and factorizes the STST into a tensor product of a low-rank core identity tensor and four latent feature matrices, which carry the information about users, activities, regions, and time, respectively. To make the factorization scalable, PCTF takes a divide-and-conquer strategy. PCTF first divides the tensor into sub-tensors, and then factorizes each sub-tensor in parallel by invoking the contextual tensor factorization (CTF). At last, PCTF generates the complete filled STST by integrating the filled sub-tensors. Figure 5 gives an illustration of PCTF. The definition of CTF is given as follow:

Definition 9 (*Contextual tensor factorization* (CTF))
The factorization of a given STST, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times L\times M\times K}$, is defined as

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{U} \times_2 \boldsymbol{C} \times_3 \boldsymbol{R} \times_4 \boldsymbol{T},$$

Table 1 Notations

	Dimension	Description	Fixed
$\mathcal{A}$	$N \times L \times M \times K$	The given STST	Yes
$\mathcal{I}$	$D \times D \times D \times D$	Identity tensor with four dimensions	Yes
$\boldsymbol{U}_{\mathrm{S}}$	$N \times N$	Social relationship feature matrix	Yes
$\boldsymbol{R}_{\mathrm{F}}$	$M \times D_r$	Geographical feature matrix, where $D_r = 8$	Yes
$\boldsymbol{S}_{\mathrm{T}}$	$N \times M$	Spatial tendency feature matrix	Yes
$\boldsymbol{U}_{\mathrm{R}}$	$N \times M$	The User-Region matrix	Yes
$\boldsymbol{U}$	$N \times D$	User latent feature matrix	No
$\boldsymbol{C}$	$L \times D$	Activity latent feature matrix	No
$\boldsymbol{R}$	$M \times D$	Region latent feature matrix	No
$\boldsymbol{T}$	$K \times D$	Time slot latent feature matrix	No
$\boldsymbol{U}_{\mathrm{L}}$	$N \times D$	Social relationship latent feature matrix	No
$\boldsymbol{R}_{\mathrm{L}}$	$D \times D_r$	Latent feature matrix of geographical features, where $D_r = 8$	No
$\overline{\boldsymbol{U}}_{\mathrm{R}}$	$N \times M$	The inferred User-Region matrix	No

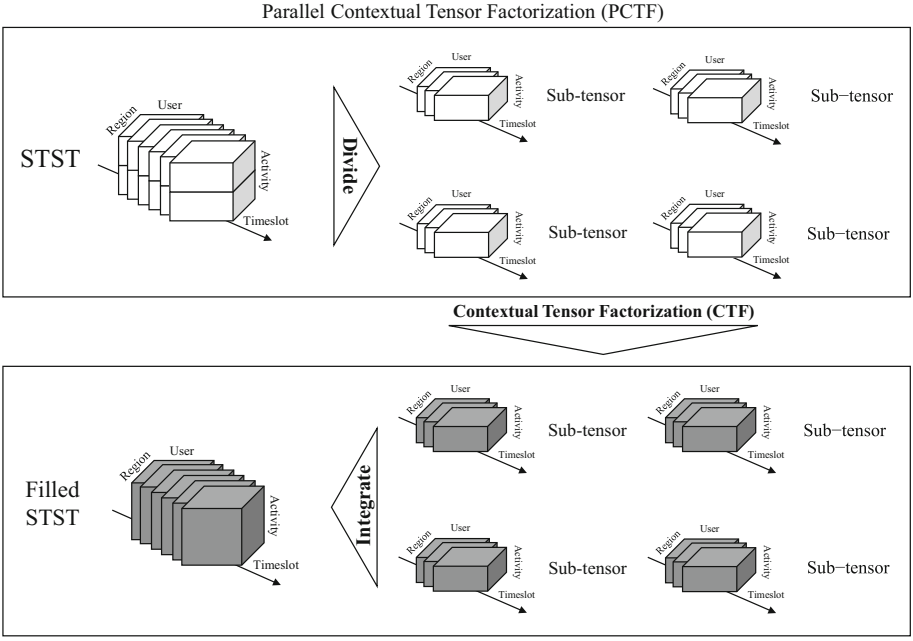


Fig. 5 The framework of parallel contextual tensor factorization (PCTF)

where the symbol $\times_i$ stands for the tensor-matrix multiplication along the ith dimension of the tensor; $\boldsymbol{U}$, $\boldsymbol{C}$, $\boldsymbol{R}$, $\boldsymbol{T}$ are the latent feature matrices of users, activities, regions, and time slots, which are calculated by solving the following optimization problem:

$$\underset{\boldsymbol{U},\boldsymbol{C},\boldsymbol{R},\boldsymbol{T},\boldsymbol{U}_{\mathrm{L}},\boldsymbol{R}_{\mathrm{L}}}{\operatorname{argmin}} \Delta\left(\boldsymbol{U},\boldsymbol{C},\boldsymbol{R},\boldsymbol{T},\boldsymbol{U}_{\mathrm{L}},\boldsymbol{R}_{\mathrm{L}}\right). \tag{4}$$

Δ is the cost function of the tensor $\boldsymbol{\mathcal{A}}$, which is defined as

$$\Delta\left(\boldsymbol{U},\boldsymbol{C},\boldsymbol{R},\boldsymbol{T},\boldsymbol{U}_{\mathrm{L}},\boldsymbol{R}_{\mathrm{L}}\right) = F_0 + \lambda_1 F_1 + \lambda_2 F_2 + \lambda_3 F_3 + \lambda_4 F_4, \tag{5}$$

where

1. $F_0 = \frac{1}{2}\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{U} \times_2 \boldsymbol{C} \times_3 \boldsymbol{R} \times_4 \boldsymbol{T}\|_{\mathrm{F}}^2$ is used to control the error of decomposing, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm;
2. $F_1 = \frac{1}{2}\|\boldsymbol{U}_{\mathrm{S}} - \boldsymbol{U} \times \boldsymbol{U}_{\mathrm{L}}^{\mathrm{T}}\|_{\mathrm{F}}^2$ is used to control the error of the constraint of social relationship feature (see Definition 6), and $\boldsymbol{U}$ is shared with F_0;
3. $F_2 = \frac{1}{2}\|\overline{\boldsymbol{U}}_{\mathrm{R}} - \boldsymbol{S}_{\mathrm{T}}\|_{\mathrm{F}}^2$ is used to control the error of the constraint of spatial tendency feature, $\overline{\boldsymbol{U}}_{\mathrm{R}}$ is the sub-matrix of the inferred User-Region matrix in which an entry $\overline{u}_r^{(nm)} = \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{d=1}^{D} u_{nd}c_{ld}r_{md}t_{kd}$, and $\boldsymbol{S}_{\mathrm{T}}$ is the sub-matrix of the Spatial Tendency Feature Matrix (see Definition 7);
4. $F_3 = \frac{1}{2}\|\boldsymbol{R}_{\mathrm{F}} - \boldsymbol{R} \times \boldsymbol{R}_{\mathrm{L}}\|_{\mathrm{F}}^2$ is used to control the error of the constraint of geographical features (see Definition 8), and $\boldsymbol{R}$ is shared with F_0;
5. $F_4 = \frac{1}{2}(\|\boldsymbol{U}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2 + \|\boldsymbol{R}\|_{\mathrm{F}}^2 + \|\boldsymbol{T}\|_{\mathrm{F}}^2 + \|\boldsymbol{U}_{\mathrm{L}}\|_{\mathrm{F}}^2 + \|\boldsymbol{R}_{\mathrm{L}}\|_{\mathrm{F}}^2)$ are regularization penalties;
6. λ_1, λ_2, and λ_3 are the weights for the contexts, and λ_4 is the weight for regularization penalties.

Algorithm 1 *PCTF (**P**arallel **C**ontextual **T**ensor **F**actorization))*

INPUT:
$\boldsymbol{\mathcal{A}}$: The original tensor;
$\boldsymbol{U}_{\mathrm{S}}$: Social Relationship Feature Matrix;
$\boldsymbol{R}_{\mathrm{F}}$: Geographical Feature Matrix;
OUTPUT:
$\overline{\boldsymbol{\mathcal{A}}}$: The filled sub-tensor;
1: Divided $\boldsymbol{\mathcal{A}}$ into sub-tensors;
2: **parallel foreach** (sub-tenor $\boldsymbol{\mathcal{A}}^{(\gamma)}$) **{**
3: Call **CTF** for each sub-tenor $\boldsymbol{\mathcal{A}}^{(\gamma)}$;
4: **}**
5: Joint all the filled sub-tensors to $\overline{\mathcal{A}}$
6: Output $\overline{\boldsymbol{\mathcal{A}}}$;

Algorithm 1 gives the outline of PCTF, and Algorithm 2 gives the outline of CTF which is invoked by PCTF for sub-tensors. Note that as there is no closed-form solution to Eq. (4), CTF searches a local minimum solution for a sub-tensor based on gradient descent.

6 SSLI inference model

We take the filled spatial–temporal semantical tensor (STST) as the input for the spatial and semantical label inference. Given a message in a social media network, an inferred activity

Algorithm 2 *CTF (Contextual Tensor Factorization)*

INPUT:
$\mathcal{A}$: The original tensor;
$\boldsymbol{U}_{\mathrm{S}}$: Social Relationship Feature Matrix;
$\boldsymbol{R}_{\mathrm{F}}$: Geography and Mobility Feature Matrix;
OUTPUT:
$\overline{\mathcal{A}}$: The filled sub-tensor;
Init $\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{R}, \boldsymbol{T}, \boldsymbol{U}_{\mathrm{L}}, \boldsymbol{R}_{\mathrm{L}}$;
$Last\Delta = MaxDouble$; $\Delta = MaxDouble/2$;
while $(Last\Delta - \Delta > \epsilon)$ **{**
  $Last\Delta = \Delta$;
  Set η according to **backtracking line search** [19];
  $\boldsymbol{U} = \boldsymbol{U} - \eta \frac{\partial \Delta}{\partial \boldsymbol{U}}$;
  $\boldsymbol{C} = \boldsymbol{C} - \eta \frac{\partial \Delta}{\partial \boldsymbol{C}}$;
  $\boldsymbol{R} = \boldsymbol{R} - \eta \frac{\partial \Delta}{\partial \boldsymbol{R}}$;
  $\boldsymbol{T} = \boldsymbol{T} - \eta \frac{\partial \Delta}{\partial \boldsymbol{T}}$;
  $\boldsymbol{U}_{\mathrm{L}} = \boldsymbol{U}_{\mathrm{L}} - \eta \frac{\partial \Delta}{\partial \boldsymbol{U}_{\mathrm{L}}}$;
  $\boldsymbol{R}_{\mathrm{L}} = \boldsymbol{R}_{\mathrm{L}} - \eta \frac{\partial \Delta}{\partial \boldsymbol{R}_{\mathrm{L}}}$;
  $\overline{\mathcal{A}} = \mathcal{I} \times_1 \boldsymbol{U} \times_2 \boldsymbol{C} \times_3 \boldsymbol{R} \times_4 \boldsymbol{T}$;
  Compute Δ according to Eq. 5;
}
Output $\overline{\mathcal{A}}$;

weight vector $\overline{\boldsymbol{v}}_a \in \mathbb{R}^{L\times 1}$ and an inferred region weight vector $\overline{\boldsymbol{v}}_r \in \mathbb{R}^{M\times 1}$ are expected as the output, where an entry $\overline{\boldsymbol{v}}_a^{(l)}$ in $\overline{\boldsymbol{v}}_a$ stores the probability that $s.activity = l$; an entry $\overline{\boldsymbol{v}}_r^{(m)}$ in $\overline{\boldsymbol{v}}_r$ stores the probability that $s.region = m$.

Given a message s in social media, its corresponding user ID is n and the time slot is k, the semantical label inference can be made by Eq. (6):

$$\overline{\boldsymbol{v}}_a = \boldsymbol{I}\left(\overline{\mathcal{A}}_{n**k}\overline{\mathcal{A}}_{n**k}^{\mathrm{T}}\right)\boldsymbol{X}, \tag{6}$$

where $\boldsymbol{I} \in \mathbb{R}^{L\times L}$ is an identity matrix, $\boldsymbol{X} \in \mathbb{R}^{L\times 1}$ is filled with 1, $\overline{\mathcal{A}}_{n**k} \in \mathbb{R}^{L\times M}$ is a matrix which is a slice of $\overline{\mathcal{A}}$, and the symbol '$*$' denotes the corresponding dimensions between the tensor and the matrix.

The spatial label inference can be made by Eq. (7):

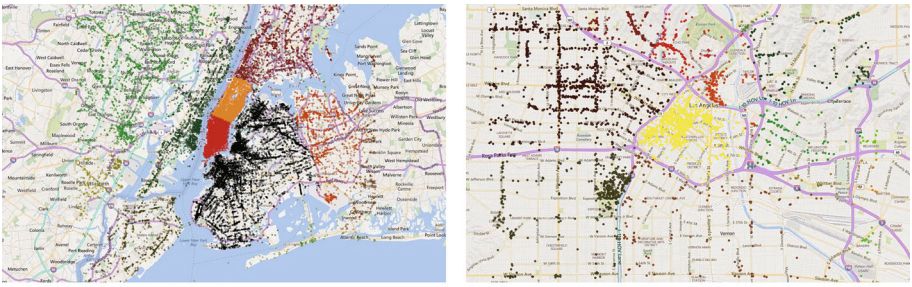
$$\overline{\boldsymbol{v}}_r = \boldsymbol{I}\left(\overline{\mathcal{A}}_{n**k}^{\mathrm{T}}\overline{\mathcal{A}}_{n**k}\right)\boldsymbol{X}, \tag{7}$$

where $\boldsymbol{I} \in \mathbb{R}^{M\times M}$ is an identity matrix, $\boldsymbol{X} \in \mathbb{R}^{M\times 1}$ is filled with 1, and $\overline{\mathcal{A}}_{n**k} \in \mathbb{R}^{L\times M}$ is a matrix which is a slice of $\overline{\mathcal{A}}$.

Because L and K are both constants, the time complexity of semantical and spatial label inference are both constants.

7 Experiment

In this section, we first verify the effectiveness of three contexts defined in Sect. 4 and geographical features defined in Sect. 4.2 by investigating their interdependency relationships. To verify the effectiveness and efficiency of our proposed method, we then compare SSLI with the baseline methods. All experiments are executed on a Windows 7 PC with an Intel



(a) **(b)**

Fig. 6 Region segmentation in **a** NYC and **b** LA

Xeon CPU of 3.4GHz and 16 GB RAM, and all programs are implemented in C#. The source code is available on [13].

7.1 Datasets

We conduct our experiments on two different real-world heterogeneous social network datasets, Foursquare and Twitter. Due to the sparsity of check-in data, we joint two Foursquare datasets which are respectively collected by Zhang et al. [11,25,26] and Bao et al. [1].

Social media network: twitter dataset [11,25,26] contains 5,223 users, 9,490,707 tweets, and 164,920 social links.

LBSN: foursquare dataset part 1 [11,25,26] contains 5,392 users, 48,756 tips, 38,921 venues, and 76,972 social links.

LBSN: foursquare dataset part 2 [1] contains 221,128 tips generated by 49,062 users at 46,065 venues in New York City (NYC) and 104,478 tips generated by 31,544 users at 31,864 venues in Los Angeles (LA).

There are totally 14 categories of venues in the two foursquare datasets: Government, health and beauty, legal and finance, estate and construction, entertainment and arts, education, food and dining, home and family, professional and services, shopping, sports and recreation, travel, nightlife, outdoors. We find 68,936 venues in New York City (NYC) and Los Angeles (LA) from two Foursquare datasets. We crawl 2678 check-ins and tweets with spatial and semantical labels which are posted by the users from 174 anchor users who have accounts on the two heterogeneous social networks.

We use the foursquare dataset as the training set and the twitter dataset as the test set. To construct the time slot dimension of spatial–temporal semantical tensor (STST), we map the time of check-ins into one of 24 time slots corresponding to the 24 h in a day. To build the region dimension, we divide regions by administrative divisions artificially. For example, as shown in Fig. 6, each point represents a venue, and each color represents a region. New York City is divided into 14 regions, and Los Angeles is divided into 14 regions as well. Note that Manhattan is divided into 3 regions, since the number of venues in Manhattan is much larger than those in other regions.

Additionally, to verify the efficiency of the proposed method, we also generate a very large synthetic dataset consisting of 50 billion entries.

7.2 Inference metric

For a message in a social media network, its ground-truth activity weight vector and region weight vector, denoted by $\boldsymbol{v}_a$ and $\boldsymbol{v}_r$, are defined as

$$\boldsymbol{v}_a = \left\{ v_a^{(1)}, ..., v_a^{(l)}, ...v_a^{(L)} \right\},$$
$$\boldsymbol{v}_r = \left\{ v_r^{(1)}, ..., v_r^{(m)}, ...v_r^{(M)} \right\},$$

where $v_a^{(l)} = 1$ if l is the ground-truth activity label, and 0 otherwise; $v_r^{(m)} = 1$ if m is the ground-truth region label, and 0 otherwise.

To evaluate the performance of the inference, we propose an accuracy, which is defined as

$$Acc = 1 - \frac{\sum |\boldsymbol{v} - \overline{\boldsymbol{v}}|}{2S},$$

where S is the number of samples.

7.3 Baseline methods

In order to demonstrate the effectiveness of our framework, we compare our approach with the following baseline methods:

1. $\mathbf{TF}$, $\mathbf{TF} + \mathbf{F}_1$, $\mathbf{TF} + \mathbf{F}_2$, $\mathbf{TF} + \mathbf{F}_3$, $\mathbf{TF} + \mathbf{F}_1 + \mathbf{F}_2$, $\mathbf{TF} + \mathbf{F}_1 + \mathbf{F}_3$, $\mathbf{TF} + \mathbf{F}_2 + \mathbf{F}_3$ In order to verify the effectiveness of the contexts defined in Sect. 4, we will compare PCTF ($\mathbf{TF} + \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3$) with the different combinations of $\mathbf{TF}$ and the three contexts, where $\mathbf{TF}$ is the tensor factorization part of the cost function, i.e., $F_0 + \lambda_4 F_4$ in Eq. (5).
2. **Matrix factorization (MF)** MF is proposed by B. Webb [23] to solve the movie recommender problem in Netflix Price. MF assumes the latent features of objects are described by vectors, and different types of objects have factors with the same size. When predicting the rating of users to objects, the estimated ratings can be expressed as the product of the latent features of the given users and the given objects. The general expression of matrix factorization is:

$$\boldsymbol{R} = \boldsymbol{P}\,\boldsymbol{Q}^{\mathrm{T}}, \tag{8}$$

   where $\boldsymbol{P} \in \mathbb{R}^{N \times D}$ is the latent feature matrix of objects. $\boldsymbol{Q} \in \mathbb{R}^{M \times D}$ is the latent feature matrix of the objects. N and M are the number of the users and objects, respectively. D is the number of latent features.
   To infer the spatial and semantical labels of tips, we apply two factorizations: user-activity matrix factorization $\boldsymbol{R}_{\mathrm{A}} = \boldsymbol{P}_{\mathrm{A}}\,\boldsymbol{Q}_{\mathrm{A}}^{\mathrm{T}}$ and user-region matrix factorization $\boldsymbol{R}_{\mathrm{R}} = \boldsymbol{P}_{\mathrm{R}}\,\boldsymbol{Q}_{\mathrm{R}}^{\mathrm{T}}$, where $\boldsymbol{R}_{\mathrm{A}}$ and $\boldsymbol{R}_{\mathrm{R}}$ are filled according to users' historical check-in data. Therefore, for a message, the inferred spatial and semantical label vectors are the corresponding user's row vectors in $\boldsymbol{P}_{\mathrm{A}}\,\boldsymbol{Q}_{\mathrm{A}}^{\mathrm{T}}$ and $\boldsymbol{P}_{\mathrm{R}}\,\boldsymbol{Q}_{\mathrm{R}}^{\mathrm{T}}$.
3. **Biased Matrix Factorization (Biased MF)** Biased MF is proposed by Paterek [14], which is an extension of Basic Matrix Factorization. Biased MF adds biased rates to the objects of either type. The formula of inference is given as follow:

$$\hat{r}_{nm} = b_n + b_m + \textstyle\sum_{d=1}^{D} p_{nd} q_{md}, \tag{9}$$

   where $\hat{r}_{nm}$ is an estimate of rate that the user n gives to the object m, D is the number of latent features. The p_{ud} and q_{md} are the entries of the latent feature matrices of users and

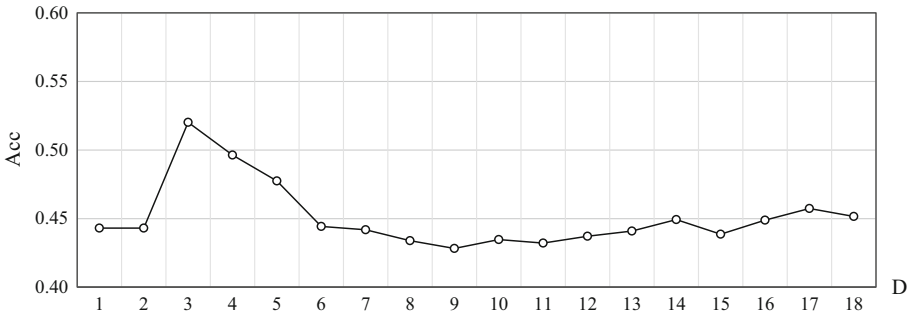


Fig. 7 Tuning the number of latent features D

objects. b_u and b_m are the biases of user u and object m, respectively.

To infer spatial and semantical labels, we do the same thing as what we do in MF.

4. **Who+Where+When+What (W4)** W4 is a probabilistic model, which discovers users' mobility behaviors from spatial, temporal, and semantical aspects [24]. After learning the model on training set, W4 can directly infer a given message's spatial and semantical labels.
5. **Text-Mining-based method (TMBM)** TMBM is a part of W4. Given collections of words with spatial and semantical labels, TMBM can learn the relationships between the representation of words and the spatial and semantical labels. Given the text of a tip, we can infer its spatial and semantical labels based on the learned relationships.

7.4 Parameter setting

The parameters, D, λ_1, λ_2, λ_3 and λ_4 are all real numbers, which cannot be determined in polynomial time. As a result, to determine the parameters, we use a heuristic method that tuning them one by one in fixed step size. First, we determine the number of latent features D by setting $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = 0$ and $\lambda_4 = 0$ on training data. According to the result shown in Fig. 7, we set the number of latent features $D = 3$.

Next, we determine the parameters, λ_1, λ_2, λ_3 and λ_4, by tuning them one by one on the training set. According to the average Acc of spatial and semantical label inference shown in Fig. 8, we set λ_1, λ_2, λ_3 and λ_4 to 0.3, 0.5, 0.2 and 0.1, respectively.

7.5 Effectiveness verification of contexts

Now, we verify the effectiveness of three contexts, social relationship ($\mathbf{F}_1$), spatial tendency ($\mathbf{F}_2$), and geographical features ($\mathbf{F}_3$), which are defined in Sect. 4.

During the experiment, if one context is not selected, its weight parameter is set to 0. For example, the setting of parameters for factorization combination "$\mathbf{TF} + \mathbf{F}_1 + \mathbf{F}_2$" will be $\lambda_1 = 0.3$, $\lambda_2 = 0.5$, $\lambda_3 = 0$ and $\lambda_4 = 0.1$.

As Fig. 9 shows when all contexts are considered, SSLI performs best. One can note that not all the contexts are equally important. When only considering one context, we find that the spatial tendency ($\mathbf{F}_2$) is the least important context. However, when considering two contexts, we find that the combination dropping $\mathbf{F}_2$ performs worst. According to the Definition 7, $\mathbf{F}_2$ involves two aspects, users and regions, which are connected to the other two contexts. And, there is no direct correlation between $\mathbf{F}_1$ and $\mathbf{F}_3$. As a result, in addition

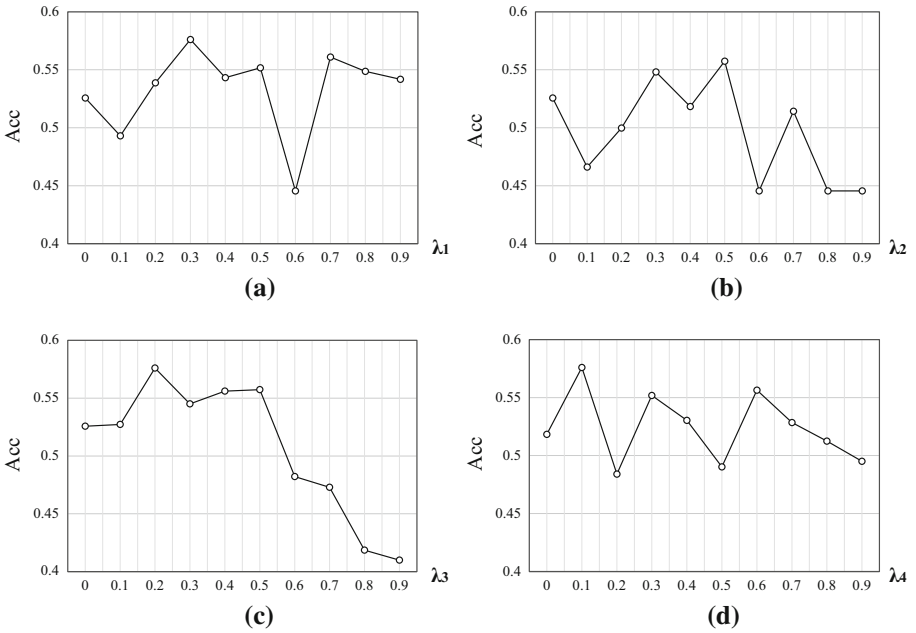


Fig. 8 Tuning the weights for the contexts in PCTF **a** λ_1, **b** λ_2, **c** λ_3, **d** λ_4

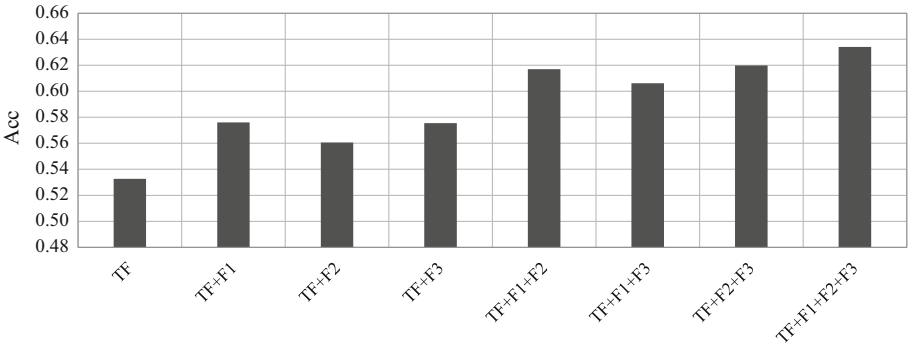


Fig. 9 Context effectiveness verification

to **TF**, $\mathbf{TF}+\mathbf{F}_2$ performs worst, $\mathbf{TF}+\mathbf{F}_1+\mathbf{F}_2$ and $\mathbf{TF}+\mathbf{F}_2+\mathbf{F}_3$ outperform $\mathbf{TF}+\mathbf{F}_1+\mathbf{F}_3$, and $\mathbf{TF}+\mathbf{F}_1+\mathbf{F}_2+\mathbf{F}_3$ performs best.

7.6 Effectiveness verification of geographical features

Now, we verify the effectiveness of the eight geographical features by running SSLI with different settings where one includes all eight features, and the others each drops one of the eight features. Note that in all settings, the contexts of social relationship and spatial tendency are considered .

As Fig. 10 shows when all the geographical features are included, SSLI performs best. One can note that not all the features are equally important. For example, the geographical feature "Richness" is less important than the other features. This is because "Richness" represents the

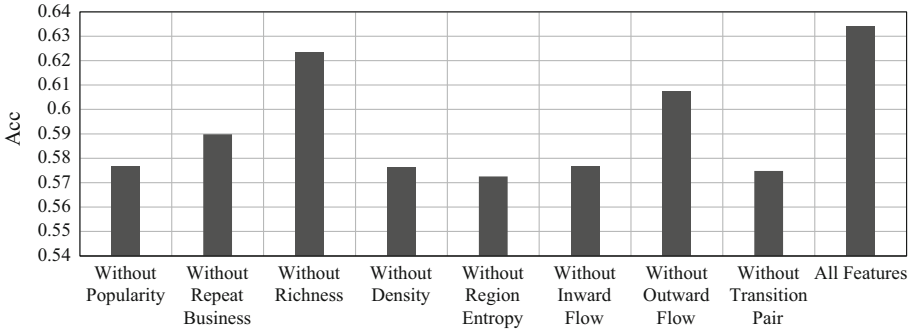


Fig. 10 Geographical feature effectiveness verification

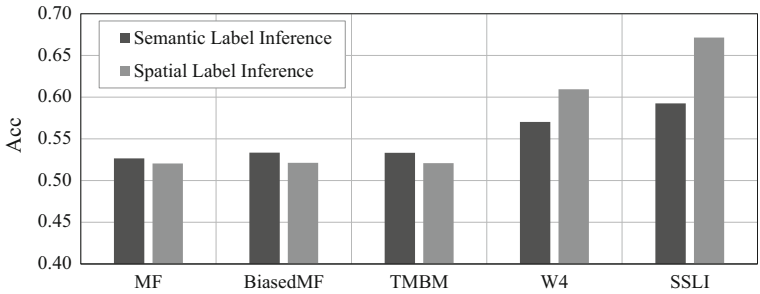


Fig. 11 The accuracy comparison between SSLI and baseline methods

number of venue types in a region which cannot accurately distinguish wide-range regions, which likely contain all types of venues. The geographical feature "outward flow" is also less important than the other features. This is because "outward flow" is not identifying enough as well. Many factors may lead to the increasing of "outward flow" of a region, e.g., the station and the convention center in the region.

7.7 Accuracy

We run SSLI and the baseline methods 100 times as they are all sensitive to initialization. Figure 11 shows the average accuracy of each method. We can give some analyses as follows:

1. Biased MF outperforms Basic MF, as Biased MF is originally proposed to improve basic MF by adding biased rates to the objects of both spatial and semantical labels. SSLI, however, still outperforms Biased MF, which is because SSLI can take into consideration not only the profiles of the users, but also the post time of tweets.
2. SSLI performs far better than TMBM, which is because the representations of posts in a social media network are quite different from those in LBSN. We also note that TMBM performs better for semantical label inference than for spatial label inference. This is because in many cases, the words of a post in a social media network partly reflect the activities of users, but the regions are less mentioned than the activities.
3. SSLI outperforms W4, which is because SSLI can take into consideration not only the three aspects of individuals, but also the practical experiences in real world.

Next, we do the Friedman test [2] to show the superiority of our method. We run SSLI and baseline methods 50 times, where we randomly exchange 5% check-ins of training set

Table 2 The performance rank comparison of SSLI and baseline methods

	MF	BiasedMF	TMBM	W4	SSLI
Overall average rank	4.08	3.94	3.92	2.04	1.02

Table 3 The rank differences between SSLI and four baseline methods

	MF	BiasedMF	TMBM	W4
Rank differences	3.06	2.92	2.90	1.02

and test set at each time. Let r_j^i be the rank of the jth method on the ith test. The Friedman test compares the average ranks of our methods and the baseline methods, where the average rank of the jth method is $R_j = \frac{1}{N}\sum_i r_j^i$. The null-hypothesis of Friedman test states all the algorithms are equivalent, and so their ranks R_j should be equal [4]. Iman and Davenport [15] show that the original Friedman statistic χ_F^2 is undesirably conservative and propose a better statistic

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{10}$$

where $\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$. The rejection region of null-hypothesis at 95% confidence level is $F_F > \mu_{0.95} = 2.4177$, where 0.95 is the confidence level. Based on Eq. (10) and the results shown in Table 2, we get $\chi_F^2 = 7583.6320$ and $F_F = 167.6608 > 2.4177$. Thus, the test statistics fall into the rejection region, which indicates that the alternative hypothesis, the average accuracies of our method and the baseline methods are not equivalent, is accepted.

Since the alternative hypothesis of Friedman test is accepted, we further proceed with a post hoc test, Bonferroni–Dunn test [5], to make a further comparison between SSLI and each baseline method. The performances of a pair of compared methods are significantly different if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}, \tag{11}$$

where the critical value q_α is based on the Studentized range statistic divided by $\sqrt{2}$ [4]. By comparing the performance of SSLI with those of the baseline methods, we calculate the average rank differences between SSLI and baseline methods. As shown in Table 3, all the differences are greater than the critical difference $CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 0.7899$, where $k = 5$, $N = 50$, and $q_\alpha = 2.498$, at 95% confidence level. As a result, SSLI is superior to all the baseline methods at 95% confidence level.

In summary, SSLI is an effective method for inferring the spatial and semantical labels for messages in social media, and the inference results of SSLI are better than all the baseline methods.

7.8 Efficiency

First, we investigate how the cost of CTF changes with the number of iterations on the real dataset, where the cost is evaluated according to Eq. (4). The real dataset contains

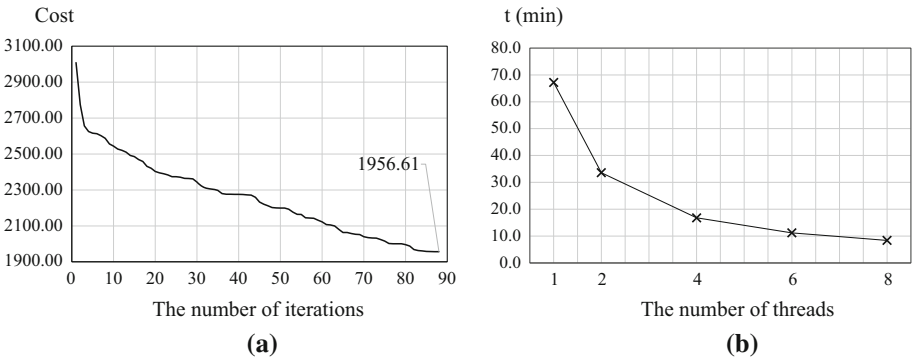


Fig. 12 **a** The overall cost of PCTF with the number of iterations; **b** The running time of PCTF with the number of threads

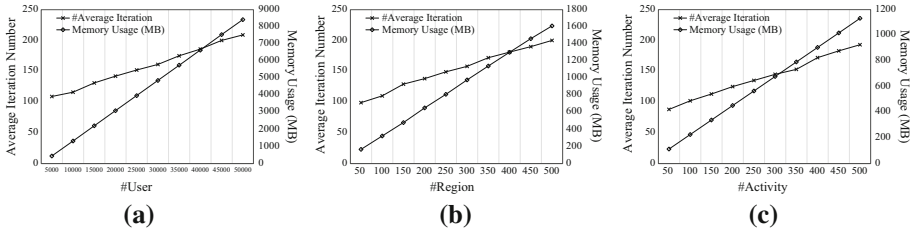


Fig. 13 The iteration number and memory usage of PCTF on tensors with different number of users, regions, and activities. **a** 10 activities, 10 regions. **b** 1000 users, 10 activities. **c** 1000 users, 10 regions

174 users, 28 regions, 14 activities, and 24 timeslots. As shown in Fig. 12a, the cost of CTF rapidly decreases with the number of iterations and finally converges at about 1956.6. Then, we verify the efficiency of PCTF on different number of threads. We first divide the original tensor into 24 sub-tensors equally and use C threads to run PCTF by setting $C = 1$, $C = 2$, $C = 4$, $C = 6$, $C = 8$, respectively. Figure 12b indicates how the overall running time of PCTF decreases with the increase of the thread number. When doubling the threads, the sub-tensors needed to be factorized for each thread will be halved. As a result, compared with one thread strategy, the parallel strategy can cut down the running time to $1/C$.

Then, we investigate how the iteration number and the memory usage of CTF changes with the number of users, regions, and activities, on a synthetic dataset. We generate a synthetic tensor contains 50,000,000,000 entries consisting of 50,000 users $\times$ 500 regions $\times$ 500 activities $\times$ 4 timeslotes, where 0.2% entries of the tensor are randomly selected and filled according to uniform distribution.

We use eight threads to run PCTF on randomly generated tensors with different number of users, regions, and activities. As shown in Fig. 13, the average iteration number and the memory usage of CTF increases linearly with the number of users, regions and activities.

To verify the running time of PCTF, we also use eight threads to run PCTF on the randomly generated STSTs with different number of users, regions, and activities and check the running time of each parallel factorization. Then, we infer the spatial and semantical labels for 100 synthetic tweets on each of the filled synthetic STST. As shown in Fig. 14,

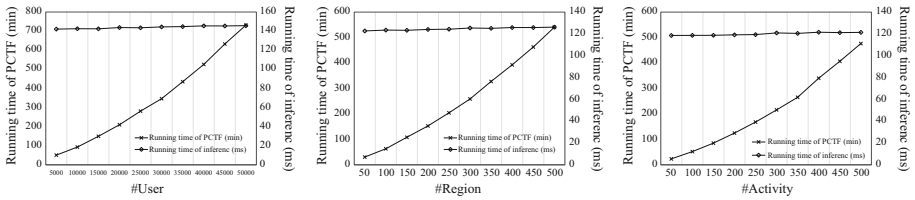


Fig. 14 The running time of PCTF and inference on tensors with different number of users, regions, and activities. **a** 10 activities, 10 regions. **b** 1000 users, 10 activities. **c** 1000 users, 10 regions

Table 4 The average time consumption and memory usage of PCTF

#Users × *#Regions* × *#Activities*	Memory usage (MB)	Running time (min)
1,000,000	623.87	69.33
2,000,000	1399.02	162.35
3,000,000	2172.91	273.63
4,000,000	2946.20	419.51
5,000,000	3719.83	582.74

the curves marked by crosses denote the running time of PCTF, and the curves marked by rhombuses denote the running time of inference. The results show the running time of PCTF grows with a polynomial time complexity of about order 1.5 as the number of users, regions, and activities increases, while the running time of inference keeps constant with the increase of the scale of training set. Note that the influence of the number of users, regions, and activities made on the running time and memory usage of PCTF are different as shown in Figs. 13 and 14. Table 4 summarizes the average time consumption and memory usage of PCTF when fixing the product of the number of users, regions, and activities, respectively.

We test the limitation under the current experimental environment. Here we assume the default parameter setting is 1000 users, 10 regions, and 10 activities over four timeslots. We explore the limit on each parameter separately, while keeping the other parameters with the default setting. We found the maximum number of users that can be supported is up to 754,400, while the maximum numbers of regions and activities that can be supported are up to 4100 and 5800, respectively.

8 Related work

There exist four classes of methods for label inference which are data fusion-based methods, text-mining-based methods, probabilistic model-based methods, and tensor decomposition-based methods.

Data fusion-based methods

Zhang et al. [27] propose a framework, called multi-network link identifier (MLI), to solve the multi-network link prediction problem by aligning the users between a social media network and an LBSN. Zheng et al. [31] propose a context-aware tensor decomposition to infer the fine-grained noise situation in New York City. However, their objective problems

are different from ours, and their methods involve only one type of data fusion, while we develop two types of data fusion, i.e., the fusion of heterogeneous social networks and the fusion of heterogeneous features.

Text-mining-based methods
Traditional methods mine the spatial and semantical labels from words. Wang et al. [20] learn the relationship between locations and words based on latent dirichlet allocation (LDA). By using an LDA-based model, Hao et al. [8] extract locations from travelogues. However, in social media, the words of messages are mostly independent with spatial and semantical labels, which makes the traditional text-mining methods unsuitable for these kind of messages.

Probabilistic model-based methods
Gonzlez et al. [6] find that an individual's mobility usually rotates around at a few previously visited locations, (e.g., home or office) and the mobility of an individual can be modeled as a stochastic process centered at several specific points. Song et al. [17,18] focus on the predictability of human mobility, and report that there is 93% human mobility, which is contributed by the high regularity of human behavior. Cho et al. [3] observe that the mobility of each user is centered at two regions, "work" and "home". They model each region as a Gaussian distribution over latitude and longitude. The probability that a user stays at the two regions is modeled as a function of time. They propose a generative model, periodic mobility model (PMM), to predict the location of a user. PMM takes users and timestamps as input; it generates a region, and the region further generates a geographical location. However, none of these works take users' activities into consideration.

Some probabilistic models infer both the region and the topic of a tweet according to users' regularities and motilities. Hong et al. [9] present a Bayesian network which depicts the dependency among region, user, topic, and geographical location. Based on the network, Hong et al. learn the geographical topics for tweets. In addition, Yuan et al. [24] further study temporal aspect to model both of the users' spatial and semantical topics, which offers the first solution to jointly model individuals from the spatial, temporal, and topical aspects. However, these models assume that "an individual's mobility usually centers at different personal geographical regions, e.g., home region and work region [3] and users tend to visit places within these regions [24]". In real-world social networks, such as Foursquare, the geographical location labels of check-ins are very randomness, especially on weekends. As social animals, the mobility and activity of users are largely affected by the social circles. However, none of these works take the social relationship among users into consideration.

Tensor decomposition-based methods
Zheng et al. [28,29] propose a tensor decomposition-based method to recommend locations and activities for users by using the location data based on GPS and users' comments. They also model the mobility and activity of users by a tensor and propose a context-aware CP decomposition to address the sparse data problem in mobile information retrieval. Similarly, a series of context-aware tensor decomposition-based methods are proposed [16,22,31] to solve problems in Urban Computing. However, the volume of social media data considered here is huge, which makes the traditional serial methods unpractical.

9 Conclusion

In this paper, we propose an inference framework, called spatial and semantical label inference (SSLI), to infer the spatial and semantical labels for the messages in social media by exploring two types of fusion, namely the fusion of heterogeneous social networks and the fusion of heterogeneous features.

We first model the mobility and activity of users as a spatial–temporal semantical tensor (STST) by fusing two heterogeneous social networks, a social media network and an LBSN. Then, we construct three types of heterogeneous features including social relationship feature, geographical features, and spatial tendency feature and fuse them with STST by exploring their interdependency relationships. Particularly, the spatial tendency feature is constructed to constrain the inference of individual mobility and reduce the uncertainty of mobility. We propose a factorizing algorithm, called parallel contextual tensor factorization (PCTF), to fill the missing entries of STST in a parallel fashion. The spatial and semantical labels can be inferred by retrieving the filled STST. We conduct the experiments on real datasets from two different domains, Twitter and Foursquare. The results verify the effectiveness and efficiency of SSLI.

The matrix and tensor factorization with constraints can be considered as a fusion of heterogeneous features. By using heterogeneous features to constrain latent features which are produced by the factorization, we are able to integrate priori knowledge about those heterogeneous features so as to achieve a higher accuracy of filling in the missing entries of the original tensor. So far, our method demands at least one of the source network contains abundant check-in data. In the future, we are going to study a pointed social network fusion method, so as to apply our method on those networks which have only little location information, such as Facebook and LinkedIn. Besides, we are going to apply the factorization-based feature fusion to broader domains, such as information diffusion and explicable recommendation.

References

1. Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: GIS '12 Proceedings of the 20th international conference on advances in Geographic Information systems pp 199–208
2. Casella G, Berger RL (2002) Statistical inference, vol 2. Duxbury Pacific Grove, Pacific Grove, CA
3. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining pp 1082–1090
4. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
5. Dunn OJ (1980) Multiple comparisons among means. J Am Stat Assoc 56:52–64
6. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453:479–482
7. Guerzhoy M, Hertzmann A (2014) Learning latent factor models of travel data for travel prediction and analysis. In: Canadian conference on artificial intelligence
8. Hao Q, Cai R, Wang C, Xiao R, Yang JM, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: WWW '10 Proceedings of the 19th international conference on World Wide Web pp 401–410
9. Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: WWW '12 Proceedings of the 21th international conference on World Wide Web pp 769–778
10. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C (2013) Geo-spotting: Mining online location-based services for optimal retail store placement. In: KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining pp 793–801

11. Kong X, Zhang J, Philip YS (2013) Inferring anchor links across multiple heterogeneous social networks. In: CIKM '13 Proceedings of the 22nd ACM international conference on information and Knowledge Management pp 1289–1294
12. Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y (2014) Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation.In: SigKDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
13. Ma Y (2016) Source code. http://pan.baidu.com/s/1qY2GHWS
14. Paterek A (2007) Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of KDD cup and workshop pp 5–8
15. Ronald IL, Davenport JM (1980) Approximations of the critical region of the fbietkan statistic. Commun Stat Theory Methods 9(6):571–595
16. Shang J, Zheng Y, Tong W, Chang E, Yu Y (2014) Inferring gas consumption and pollution emission of vehicles throughout a city. In: KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
17. Song C, Koren T, Barabasi AL (2010) Modelling the scaling properties of human mobility. Nat Phys 6:818–823
18. Song C, Qu Z, Blumm N, Barabasi A (2010) Limits of predictability in human mobility. Science 327:1018–1021
19. Stephen-Boyd lV Convex optimiation
20. Wang C, Wang J, Xie X, Ma WY (2007) Mining geographic knowledge using location aware topic model. In: GIR '07 Proceedings of the 4th ACM workshop on Geographical information retrieval pp 65–70
21. Wang Y, Yuan NJ, Lian D, Xu L, Xie X, Chen E, Rui Y (2014) Regularity and conformity: Location prediction using heterogeneous mobility data. In: SigKDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
22. Wang Y, Zheng Y, Xue Y (2014) Travel time estimation of a path using sparse trajectories. In: KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
23. Webb B (2006) Netflix update: Try this at home. http://sifter.org/simon/journal/20061211.html
24. Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Who, where, when and what: discover spatio-temporal topics for twitter users. In: KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 605–613
25. Zhang J, Kong X, Philip YS (2013) Predicting social links for new users across aligned heterogeneous social networks. In: ICDM '13 Proceedings of the 13th International conference on data Mining pp 1289–1294
26. Zhang J, Kong X, Philip YS (2014) Transferring heterogeneous links across location-based social networks. In: WSDM '14 Proceedings of the 7th ACM international conference on Web search and data Mining pp 303–312
27. Zhang J, Philip YS, Zhou Z (2014) Meta-path based multi-network collective link prediction. In: KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining pp 1286–1295
28. Zheng VW, Cao B, Zheng Y, Xie X, Yang Q (2010) Collaborative filtering meets mobile recommendation: a user-centered approach. In: AAAI 10
29. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data. In: WWW '10 Proceedings of the 19th international conference on World Wide Web
30. Zheng Y (2015) Methodologies for cross-domain data fusion: an overview. IEEE Trans. Big Data 1(1):16–34
31. Zheng Y, Liu T, Wang Y, Zhu Y, Liu Y, Chang E (2014) Diagnosing new york city's noises with ubiquitous data. In: Ubicomp '14 Proceedings of the 2014 ACM international joint conference on Pervasive and Ubiquitous Computing pp 247–256

Yuchi Ma was born in 1990. He is currently a doctoral student in Sichuan University. His recent research interests include social network and recommender system.

Ning Yang was born in 1974. He received the Ph.D. degree in data mining from School of Computer Science, Sichuan University, in 2010. He is a member of IEEE. He is a lecturer at Sichuan University. His current research interests include database and data mining.

Lei Zhang was born in 1980. She received the Ph.D. degree in computer software and theory from School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), in 2008. She is a professor at Sichuan University. Her research interests include theories and applications of machine intelligence, control theories and applications.

Philip S. Yu received the PhD degree in electrical engineering from Stanford University. He is a distinguished professor in computer science at the University of Illinois at Chicago and is also the Wexler Chair in information technology. His research interests include big data and data mining. He is a fellow of the ACM and the IEEE.