Predicting Activity Attendance in Event-Based Social Networks: Content, Context and Social Influence

Rong Du †, **Zhiwen Yu** †, **Tao Mei** ‡, **Zhitao Wang** †, **Zhu Wang** †, **Bin Guo** †
† Northwestern Polytechnical University, Xi'an 710129, Shaanxi, China
‡ Microsoft Research, Beijing 100080, China
zhiwenyu@nwpu.edu.cn, tmei@microsoft.com

ABSTRACT

The newly emerging event-based social networks (EBSNs) connect online and offline social interactions, offering a great opportunity to understand behaviors in the cyber-physical space. While existing efforts have mainly focused on investigating user behaviors in traditional social network services (SNS), this paper aims to exploit individual behaviors in EBSNs, which remains an unsolved problem. In particular, our method predicts activity attendance by discovering a set of factors that connect the physical and cyber spaces and influence individual's attendance of activities in EBSNs. These factors, including content preference, context (spatial and temporal) and social influence, are extracted using different models and techniques. We further propose a novel Singular Value Decomposition with Multi-Factor Neighborhood (SVD-MFN) algorithm to predict activity attendance by integrating the discovered heterogeneous factors into a single framework, in which these factors are fused through a neighborhood set. Experiments based on real-world data from Douban Events demonstrate that the proposed SVD-MFN algorithm outperforms the state-of-the-art prediction methods.

Author Keywords

Activity prediction; event-based social networks; content preference; context; social influence.

ACM Classification Keywords

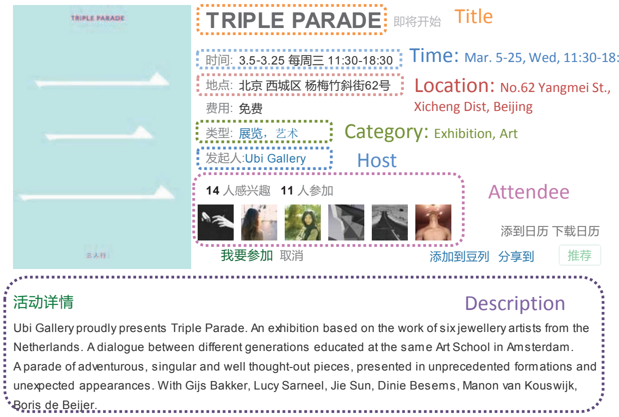
H.3.5 Online Information Services: Web-based services

General Terms

Algorithms, Experimentation, Performance

INTRODUCTION

With the proliferation of event and activity-based applications such as Facebook Events[1], Google+ Events[2], Meetup[3], and Douban Events[4], it has become possible for users to expand their online interactions to offline activities. People can propose and attend a variety of offline social activities through these online services, which can further promote face-to-face social interactions. This kind of social media is called Event-Based Social Networks (EBSNs) [20]. Although there have been studies on EBSNs, very few attempts have been paid on behavior prediction. Understanding the collective dynamics of user participation in events is crucial to better understand social networks in both the physical and cyber worlds and to provide critical insights that help personalized event recommendation and targeted advertising. In contrast to traditional social network services (SNS), user behavior in EBSNs is predominantly driven by offline activities and highly influenced by a set of unique factors, such as spatio-temporal constraints and special social relationships (host and attendee). Therefore, these properties of social activities play an important role in behavior prediction in EBSNs, making it different from the online behavior prediction in SNS. Figure 1 shows the main elements of activities in Douban Events, a popular EBSNs in China, described as follows.


TRIPLE PARADE 即将开始

时间: 3.5-3.25 每周三 11:30-18:30
地点: 北京 西城区 杨梅竹斜街62号
费用: 免费
类型: 展览，艺术
发起人:Ubi Gallery

14 人感兴趣 11 人参加

活动详情

Ubi Gallery proudly presents Triple Parade. An exhibition based on the work of six jewellery artists from the Netherlands. A dialogue between different generations educated at the same Art School in Amsterdam. A parade of adventurous, singular and well thought-out pieces, presented in unprecedented formations and unexpected appearances. With Gijs Bakker, Lucy Sarneel, Jie Sun, Dinie Besems, Manon van Kouswijk, Boris de Beijer.
Title
Time: Mar. 5-25, Wed, 11:30-18:30
Location: No.62 Yangmei St., Xicheng Dist, Beijing
Category: Exhibition, Art
Host
Attendee
添到日历 下载日历
我要参加 取消
添加到豆列 分享到
推荐
Description

Figure 1. An example of an activity on Douban Events with five key elements: location, time, attendees, host, and content.

- *Location*: where the activity will be held. Usually, an activity is held at a convenient and popular place.
- *Time*: when the activity will start and end. In fact, the distribution of time highly depends on the content of an activity.

[1] www.facebook.com/events
[2] plus.google.com/events
[3] www.meetup.com
[4] beijing.douban.com

UbiComp'14, September 13 - 17 2014, Seattle, WA, USA.
Copyright 2014 ACM 978-1-4503-2968-2/14/09...$15.00.
http://dx.doi.org/10.1145/2632048.2632063

- *Attendees*: users who click the "I want to attend" button to show their intention to attend the event.
- *Host*: the user or site who will hold the event and be responsible for offline activity organization.
- *Content*: the detailed information of an activity, which includes three aspects: category, title, and description. Categories could be music, film, sport, party, travel, exhibition, drama, and so on.

Based on these elements, we are able to predict a user's activity attendance using different methods. However, how to comprehensively combine the heterogeneous information in a systematic way still remains an open issue. In this paper, we formulate and parameterize the above elements into a multi-factor model, which mainly consists of three different factors: 1) **content**. One key factor that determines whether a user will attend a certain activity is its content. We use content preference to represent a user's preference for different activities. 2) **context**. The context in our method refers to spatial context and temporal context. Specifically, users may have different time and location preferences while attending different offline activities. For example, if a user is a student, the possibility for him/her to attend an activity held during a weekday would be low, even if he/she likes the activity. Similarly, if a user is far away from the activity location, her willingness to attend would be low. 3) **social influence**. In EBSNs, users follow each other based on common interests. One's willingness to attend a certain activity could be impacted by his or her social relationships with the host and other attendees. However, due to the time and location constraints of activities, the host usually plays a much more significant role in user's attendance decision than ordinary followers. The reason is that, on one hand, the host can recommend activities to its followers and, on the other , if a user often attends activities hosted by a influential host, the possibility for attending other events organized by this host will also be high.

Based on these factors and the multi-factor model, we propose a novel method, named Singular Vector Decomposition (SVD) with Multi-Factor Neighborhood (SVD-MFN), to predict activity attendance in EBSNs. The method is based on SVD and is capable of integrating different features into a multi-factor model as the neighborhood, which fully takes advantage of both the SVD and multi-factor model. The main contributions can be summarized as follows:

- We propose to investigate the prediction of activity attendance in the emerging EBSNs, which is, to the best of our knowledge, one of the first attempts in the area of ubiquitous computing.
- We have discovered and modeled three key factors that influence individual behavior, e.g., content preference, spatio-temporal contexts, and social influence.
- We have developed a novel algorithm (i.e., SVD-MFN) to optimally integrate multiple heterogeneous factors we mentioned above into a single framework that outperforms the state-of-art methods.

The remainder of this paper is organized as follows. We first review related work. Then we formally define the problem and present the system framework. Feature extraction and fusion are then described, followed by an elaboration of the proposed SVD-MFN algorithm. We next present experimental results. We conclude our work and discuss possible future directions in the final section.

RELATED WORK

We briefly review related work, which can be classified into three categories.

The first category is research on understanding relationship between online and offline social interactions in EBSNs. Liu *et al.* proposed the concept of EBSNs and focused on community detection using both online and offline social links [20]. Han *et al.* sought to gain insights into user behavior for attending offline events based on Douban Events [12]. By studying the events in Douban, they present results linked to the event properties, user behavior of participants of an event, and social influence on an event, which help us better understand what affects user behavior during events. Xu *et al.* conducted a quantitative analysis that revealed the relationship between online following behavior and characteristics of real-world events [27]. The difference between our work and this line of research is that we move one step further to leverage the unique characteristics of EBSNs to predict whether a user will attend an activity or not.

The second category includes research into activity or event prediction. There are some approaches towards activity recommendation. For example, in the Pittsburgh area, a cultural event recommender was built around trust relations [16]. The recommender system for academic events focused more on social network analysis (SNA) combined with collaborative filtering (CF) [14]. Cornelis *et al.* developed a hybrid event recommendation approach where both CF and content-based algorithms were employed [7]. Daly *et al.* established an interesting event management service that considers the location of events when making recommendations [9]. They found that event attendees are sometimes from nearby locations and proposed a location-based method to recommend local events to targeted users. Minkov *et al.* presented a collaborative approach for event recommendation which outperforms approaches that only consider content information [22]. Zhuang and Sang *et al.* explored using temporal and spatial context for entity recommendation [32, 25]. However, the existing research only considers one or two aspects of events in EBSNs, and none of them have developed a comprehensive yet systematic method to combine different and heterogeneous information.

The third category focuses on location-based social networks, which also contain both online and offline social interactions [26, 30]. Although adjacent check-ins may indicate implicit social interactions and social ties [6], the check-in data is usually too sporadic to represent human behavior [23]. Crandall *et al.* examined the geographical features to infer social ties [8]. Similarly, Zhuang *et al.* formalized the problem of predicting geo-graphic coincidences in ephemeral social networks, and used a factor graph model to predict the possibility of two users will meet in future [31]. However, neither of these prediction methods is event-driven. Additionally, while the above research mainly used spatial information for prediction, the "offine" (social events) features considered in this paper contain multiple specific characteristics of the activity, such as content preference, spatial and temporal context, and social influence.

PROBLEM STATEMENT AND SYSTEM OVERVIEW

In this section, we first formulate the activity attendance prediction problem, and then present the proposed system framework.

Problem Statement

In EBSNs, there is a list of activities $\{a_1, a_2, \cdots, a_m\}$ and a list of users $\{u_1, u_2, \cdots, u_n\}$. Let A_u denote all the activities that a specific user u has ever attended. User preference is usually extracted from historical behavior. We use $F_u(a)$ to represent the preference of user u for the activity a, which consists of the following five parts. $CP_u(a)$ is the content preference, $DP_u(a)$ is the distance preference in the spatial context, $WP_u\,(a)$ and $SP_u(a)$ are the day of the week and the hour of the day preferences in the temporal context, and $SP_u(a)$ represents social influence, which means the relationship of u and a's host. These factor pairs can be concluded as three macro aspects: content, context and social influence. We can also get an activity neighbor set $NS_u(a)$ for each user-activity pair selected from A_u by combining the above three aspects, which means the nearest activities selected from user's attendance history. The activity attendance prediction task of a target user in EBSNs can be described as: given a user a attendance history A_u and the set of attendees U_a of an upcoming activity a, we are predicting whether the target user u will attend a or not. The prediction result is represented as r_{ua}, which is a binary value, i.e., 1 means "attend" and 0 "ignore." In this paper, we take both $F_u(a)$ and $NS_u(a)$ into account for effective activity attendance prediction. The key problems are listed as follows:

- How to evaluate contribution of different factors including $CP_u(a)$, $DP_u(a)$, $WP_u(a)$, $HP_u(a)$, and $SP_u(a)$? How to organize and fuse these factors in a single model, and then obtain $F_u(a)$?
- How to extract the neighbor set $NS_u(a)$ using the factor model?
- How to combine $F_u(a)$ and $NS_u(a)$ in the prediction of r_{ua}?

System Overview

Figure 2 shows an overview of our framework, which consists of three components: the attendance matrix construction component (left), the neighborhood discovery with multi-factor component (right), and the prediction component (bottom).

Attendance Matrix Construction. On one hand, given the target user u and the current attendees of an upcoming activity a, we can construct a binary attendance matrix for activity a. On the other hand, the historical attendance matrix can be built for all the users. Afterwards, based on matrix updating, we can get a global attendance matrix, which is one input for the prediction component.

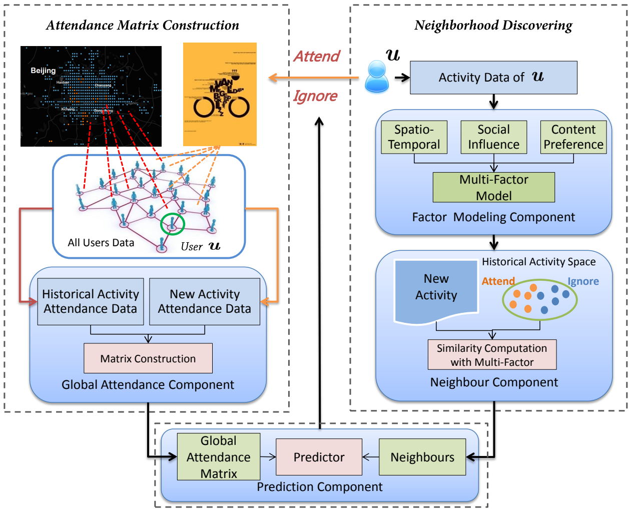


Figure 2. The proposed framework for activity prediction in EBSNs.

Neighborhood Discovery with Multi-Factor. The second component is based on the target user's historical attendance, which is the key component in our framework. We first extract features for each activity from three aspects: content preference, spatio-temporal context, and social influence. Then, considering the different influence of these three aspects, we propose a multi-factor(MF) model using decision tree to evaluate their contributions. Based on this model, we can compute the similarity between each pair of events from the target user's perspective. Furthermore, we can discover the neighbor activities for the target user from the attendance history.

Prediction Component. To combine the previous parts into our system, we propose the SVD-MFN predictor in the prediction component. We will give the details in the SVD-MFN algorithm section.

FEATURE MODELING

In this section, we first elaborate the features used in our multi-factor model from three macro aspects: content preference, spatio-temporal context, and social influence. The foundation of this model is a user study of the factors that influence activity attendance, which shows that users pay the most attention to these three aspects when they choose activities. Then, we present how different features are fused together.

Feature Extraction

Content

The content of an activity plays a major role in determining the likelihood of a user participating in an event [22]. Therefore, the calculation of content similarity between the user's historical activities and the upcoming activity is a key step. A proper similarity measure can greatly influence the performance of our system. In the case of Douban Events, there are three elements for characterizing the activity content: category, title and description, as shown in Figure 1. We put these elements together as a whole text, and then define the content similarity between a pair of activities as their text similarity. Numerous studies have attempted to resolve the text similarity problem. One approach is to expand and enrich the keyword in the text with a search engine [4]. Another approach uses an external lexical database, such as WordNet, to mine the relationships among words [17]. Although there are a lot of words and their semantic relationships in the lexical database, the application scope of dictionary-based similarity computation is quite limited. The Latent Dirichlet Allocation (LDA) is a probabilistic topic model that can solve all the above problems [3].

LDA based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. We begin by removing stop words (i.e., punctuation) and short words (i.e., “of” and “and”), and then format the remaining text as the input of LDA. Afterwards, the formatted text is mapped to the subject space by using the Gibbs sampling. And Then the Jensen-Shannon (JS) distance is often used to compute text similarity [19]. In the end, in order to get the best number of topics in LDA, we use a clustering method [4]. The process of generating a text with n words based on LDA can be described by a marginal distribution:

$$P(d) = \int\limits_{\theta} \left(\prod_{i=1}^{n} \sum_{T^{(i)}} P\left(W^{(i)}|T^{(i)},\beta\right) P\left(T^{(i)}|\beta\right) \right) P(\theta|\alpha)\, d\theta \tag{1}$$

where $P(\theta|\alpha)$ is derived from Dirichlet distribution parameterized by α, and $P\left(W^{(i)}|T^{(i)},\beta\right)$ is the probability of word $W^{(i)}$ under topic $T^{(i)}$ parameterized by β. The topic-word distribution $P\left(W^{(i)}|T^{(i)},\beta\right)$ in Eq. (1) is an important factor for the implementation of LDA. We use the Gibbs sampling method to extract topics from the corpus [10], and then adopt the result of sampling as the input of text similarity computation. A standard function to measure the divergence between two distributions p and q is the Kullback Leibler (KL) divergence [19]:

$$\mathrm{D}_{KL}(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \tag{2}$$

The KL divergence is asymmetric but convenient to be applied in the JS divergence which has been proved symmetrized [19] and its value ranges from 0 to 1. The equation is:

$$JS(p,q) = \frac{1}{2}\left[D_{KL}\left(p, \frac{p+q}{2}\right) + D_{KL}\left(q, \frac{p+q}{2}\right) \right] \tag{3}$$

As an example, consider the three sample activities which have been preprocessed:

a_1: “ drama British Shakespeare classic Macbeth. ”

a_2: “ British Shakespeare King Lears. ”

Applying LDA with topic number Z = 2 would yield topics to:

T_1: “ British Shakespeare ”

T_2: “ drama classic ”

Obviously, a_1 would have a 50% membership in both topics, since it contains words from both topics to an equal degree, and activity a_2 would have a 100% membership in T_1 and T_2, respectively. We could then represent each activity as a vector of their topic memberships:

a_1 = [0.5 , 0.5]

a_2 = [1.0 , 0.0]

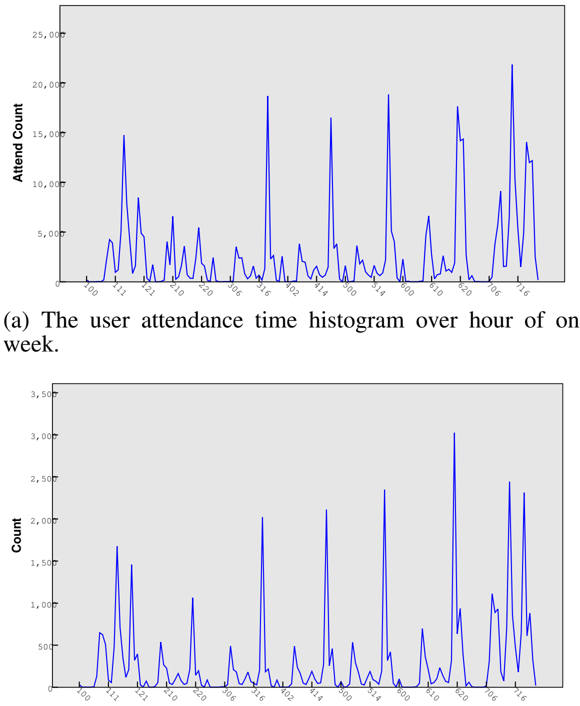
where the first element in each vector corresponds to their distributions in topic T_1 and the second element to distributions in T_2, represented as $\theta^{(a_1)}$ and $\theta^{(a_2)}$. So the JS divergence between a_1 and a_2 can be computed by taking $\theta^{(a_1)}$ and $\theta^{(a_2)}$ into Eq. (3), which equals to 0.31. Consequently, we could get the content similarity between a_1 and a_2 with JS divergence as:

$$Sim(a_1,a_2) = 1 - JS\left(\theta^{(a_1)}, \theta^{(a_2)}\right) \tag{4}$$

which equals to 0.69 in the example.

Based on the content similarity of two activities, we can compute the user’s content preference/interest. Specifically, we adopt the interest drift with forgetting mechanism in our work [5], and the key idea is explained below:

On one hand, people’s interest wane as time goes by like memory. For example, an activity that was attended recently by a user should have a higher impact on the prediction of future behavior than an event that happened a long time ago. On the other hand, the forgetting speed slows down as the accumulated interests become more stable. Based on these two principles, we have constructed two interest models for different purposes by incorporating the forgetting mechanism, using Short Term Interest Model (STIM) to represent the user’s recent interests and Long Term Interest Model (LTIM) to denote accumulated stable interests. The forgetting function is implemented to simulate the attenuation of the user’s interests based on the following equation: $I(a,a_i) = exp\{-\frac{ln2\times(t_a-t_{a_i})}{hl}\}$, where the forgetting coefficient $I(a,a_i)$ denotes the degree that the original interest have declined, t_{a_i} means the start date on an activity that the user has attended, t_a means the date of the future activity to be predicted, and hl denotes the half-life (in days) controlling the speed of forgetting. The larger hl is, the slower interests fade. When $t_a - t_{a_i} = hl$, $I(a,a_i)$ descends to $1/2$. For the short interest model, hl is a stable constant, and we set it as 90. For the long interest model, the half-life is not a constant any more. The user’s interests usually become more stable over time. We use the $I_l(a,a_i) = exp\{-\frac{ln2\times(t_a-t_{a_i})}{hl_0+d_{acc}\times s}\}$ forgetting formula to cal-



(a) The user attendance time histogram over hour of one week.

(b) The activity start time histogram over hour of one week

Figure 3. Time histogram over hour of one week

culate coefficient $I_l(a, a_i)$, where hl_0 represents an initial half-life value, and d_{acc} denotes how many days the original LTIM has evolved. Constant s reflects the impact of d_{acc} on the forgetting speed and we set it as 0.5 by experience. By involving factor $d_{acc} \times s$, the user's interests fall more slowly than original formula. To sum up, we use the following forgetting formula to calculate the coefficient:

$$I(a, a_i) = \left\{ \begin{array}{l} e^{-\frac{ln2 \times (t_a - t_{a_i})}{hl_0}}, (t_a - t_{a_i}) \leq d_{acc} \\ e^{-\frac{ln2 \times (t_a - t_{a_i})}{hl_0 + d_{acc} \times s}}, (t_a - t_{a_i}) > d_{acc} \end{array} \right. \quad (5)$$

Consequently, for target user u, the content similarity between future activity a and past event a_i is $CS_u(a, a_i) = I(a, a_i) \times Sim(a, a_i)$, where the $Sim(a, a_i)$ means the content similarity we computed in Eq. (4). Considering both the content similarity between two activities and the attenuation degree of the user's interests, we define the content preference of user u to activity a as follows:

$$CP_u(a) = \frac{\sum_{a_i \in A_u} I(a, a_i) \times Sim(a, a_i)}{\sum_{a_i \in A_u} Sim(a, a_i)} \quad (6)$$

Spatial and Temporal Context

1) *Temporal Context*

The temporal context is important for activity prediction. On one hand, as Figure 3(a) shows, human behavior shows strong daily and weekly periodical patterns. On the other hand, the start time of activities is also periodic, as shown in Figure 3(b).

a) *Day of Week Factor*

A user's daily life is usually weekly periodic, so we introduce the day of the week factor $WP_u(a)$ as the first temporal user preference as follows:

$$WS_u(a, a_i) = \left\{ \begin{array}{l} 1, wd(t_a) = wd(t_{a_i}) \\ 0, wd(t_a) \neq wd(t_{a_i}) \end{array} \right. \quad (7)$$

where $wd(t_a)$ represents which day of the week the user attended activity a, and $wd(t_a) \in \{1, 2, 3, 4, 5, 6, 7\}$ corresponds to the day of the week (i.e., Monday, Tuesday, Wednesday, ..., Sunday). Thereby, u's day of week preference can be defined as:

$$WP_u(a) = \frac{\sum_{a_i \in A_u} WS_u(a, a_i) \times Sim(a, a_i)}{\sum_{a_i \in A_u} Sim(a, a_i)} \quad (8)$$

b) *Hour of Day Factor*

In Figure 3(a), we found that a user's activity attendance in one day is also periodic, which can be explained as follows. Most users on Douban are either students or white collars. If an activity is held during study or work time, then these users are not likely to participate even though they are interested in the event. In contrast to the day of the week factor, we employ the Gauss formula instead of the binary formula to express the similarity in the hour level: $HS_u(a, a_i) = exp\{-\frac{(t_a - t_{a_i})^2}{2}\}$, thus we can express the hour of the day factor as:

$$HP_u(a) = \frac{\sum_{a_i \in A_u} HS_u(a, a_i) \times Sim(a, a_i)}{\sum_{a_i \in A_u} Sim(a, a_i)} \quad (9)$$

2) *Spatial Context*

We noticed that the likelihood of attending an activity decreases as the distance between the user's home location and the activity's location increases, which is not surprising and has been proven by numerous researchers [28]. Moreover, users have individual preference for locations. For example, if the transportation to a place is convenient, activities around it will be more popular. Figure 4 illustrates the location distribution of all activities in our dataset. We observed that there were more activities in the Haidian, Chaoyang and Dongcheng districts in Beijing, China, which are represented by the bigger and warmer circles in the figure. Additionally, similar activities tend to be located in the same areas. For example, most education activities are held in Haidian district, where there are many colleges [27].

To derive the spatial similarity between two activities, we calculate the user's location preference based on her attendance history, as the user's real home location is difficult to obtain. Similar to the hour of day factor, we adopt the Gauss formula to calculate the distance similarity: $DS_u(a, a_i) = exp\{-\frac{Distance(a, a_i)^2}{2}\}$, and the user's spatial factor can be defined as :

$$DP_u(a) = \frac{\sum_{a_i \in A_u} DS_u(a, a_i) \times Sim(a, a_i)}{\sum_{a_i \in A_u} Sim(a, a_i)} \quad (10)$$

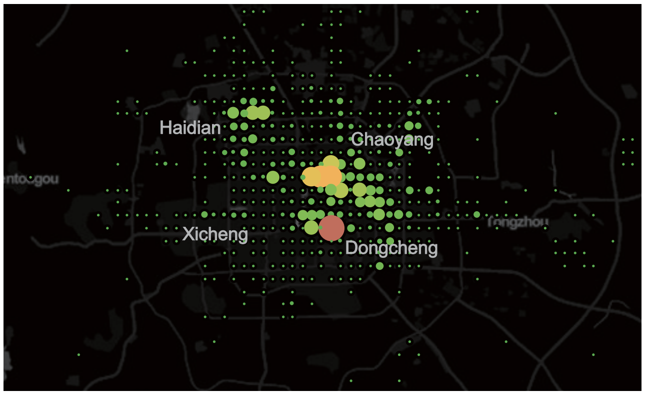


Figure 4. Distribution of activity locations in Beijing, China.

Social Influence

Social friendship is beneficial for event prediction and recommendation [29], which is a key factor motivating users to participate in social events [1]. We define two types of social relationships between the user and the host. The first type is the *following relationship*. In Douban Events, the activity host is allowed to send invitations to her followers. Therefore, if user u follows a host, she is likely to be more willing to attend the activities organized by this host once she is invited. The second type is *preferring relationship*. For example, a user may have attended a lot of sports events organized by one host who runs a gym, because the user is interested in this host or she is just a member of the gym. Therefore, she is more likely to attend future sports activities organized by this gym host whether she follows the host or not. Based on the above description, we define the social similarity between an upcoming event a and a past event a_i as

$$SS_u\left(a, a_i\right) = S_1\left(u, H\left(a\right)\right) * \delta + S_2\left(H\left(a\right), H\left(a_i\right)\right) * \left(1-\delta\right) \tag{11}$$

where $H\left(a\right)$ means the host of a, S_1 means the *following relationship* and S_2 means the *preferring relationship*. If u follows $H\left(a\right)$, $S_1\left(u, H\left(a\right)\right)$ =1, otherwise $S_1\left(u, H\left(a\right)\right) = 0$. The same goes for *preferring relationship*: for two activities a and a_i organized by the same host, if $H\left(a\right) = H\left(a_i\right)$, then $S_2\left(H\left(a\right), H\left(a_i\right)\right) = 1$, otherwise it equals 0. We set the parameter δ as 0.5, which means these two types of relationships have the same weights. Consequently, the user's social influence is defined as:

$$SP_u\left(a\right) = \frac{\sum_{a_i \in A_u} SS_u\left(a, a_i\right) \times Sim\left(a, a_i\right)}{\sum_{a_i \in A_u} Sim\left(a, a_i\right)} \tag{12}$$

Now, we have extracted similarity features from three different aspects, which need be fused together to enable effective activity prediction.

Feature Fusion

As different features have different impacts on user preferences, the challenge is to evaluate their significance and then fuse them together. In EBSNs, a user's rating for an activity is binary (i.e., 1 or 0), which means she attends or ignores an event. Therefore, we can transform activity attendance prediction into a classification issue. Then classification algorithms can be used to deal with our problem. Based on the result of classification, we first get the contribution weights of different features and then combine the features linearly. The total similarity between an upcoming activity a and a past event a_i for the target user u is:

$$\begin{aligned} Sim_u\left(a, a_i\right) &= \alpha * CS_u\left(a, a_i\right) + \beta * WS_u\left(a, a_i\right) \\ &+ \gamma * HS_u\left(a, a_i\right) + \delta * DS_u\left(a, a_i\right) + \varepsilon * SS_u\left(a, a_i\right) \end{aligned} \tag{13}$$

where $\alpha, \beta, \gamma, \delta$ and ε represent the weights of different features, respectively. In detail, α means the whole content preference weight; β and γ are the day of week and hour of day weights in the temporal context; δ is for the spatial context and ε denotes the weight of the social influence feature and their value ranges from 0 to 1.

SVD-MFN ALGORITHM

To effectively leverage different features for activity attendance prediction, we proposed a novel Singular Value Decomposition with Multi-Factor Neighborhood (SVD-MFN) algorithm, which is based on SVD. Thereby, in this section we first introduce SVD and then explain the algorithm in detail.

Matrix factorization: SVD

Singular value decomposition (SVD) is a well-known matrix factorization technique that addresses the problems of synonymy, polysemy, sparsity, and scalability for large datasets [15]. It has been proved in different areas, such as advertising [21] and e-commerce [18]. The basic idea of matrix decomposition in SVD is applicable to our system, because it delivers good results and is easy to tune and customize. In its basic form, every user u and activity a is associated with vectors as $p_u, q_a \in \mathrm{R}^d$. The vectors p_u and q_a are generally referred to as d -dimensional latent user and activity factors, respectively. Based on these definitions, user u's attendance at activity a is predicted via the following equation: $r_{ua} = {p_u}^T q_a$.We use a logistic function to transform the preference score into the interval $(0, 1)$, and set 0.5 as the threshold, i.e., if it is bigger than 0.5 the user will attend, otherwise the user will ignore the event. In this work, parameters are generally learned by solving the following regularized least squares problem:

$$\min_{p_*, q_*} \sum_{(u,a) \in R} \left(r_{ua} - {p_u}^T q_a\right)^2 + \lambda\left(\sum_u \|p_u\|^2 + \sum_a \|q_a\|^2\right) \tag{14}$$

Here, the constant λ is a parameter determining the extent of regularization, which is set as 0.01. As we use the gradient descent learning algorithm, the training time grows linearly with the value of $|R|$. Therefore, the running time of each user-activity pair is $O\left(1\right)$.

SVD-MFN: Integrating Multi-Factor Neighborhood into SVD

Traditional neighborhood methods focus on computing the relationships between activities or, alternatively, users. While they are effective at detecting localized relationships and performing predictions on a few similar neighbors, they may fail to work when there is no or few observed ratings within the neighborhood of limited size. In contrast, the latent factor model in SVD is effective at capturing global information

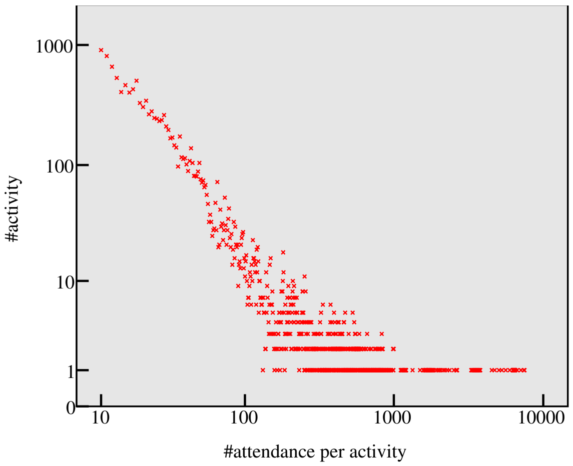


Figure 5. Distribution of the number of attendees.

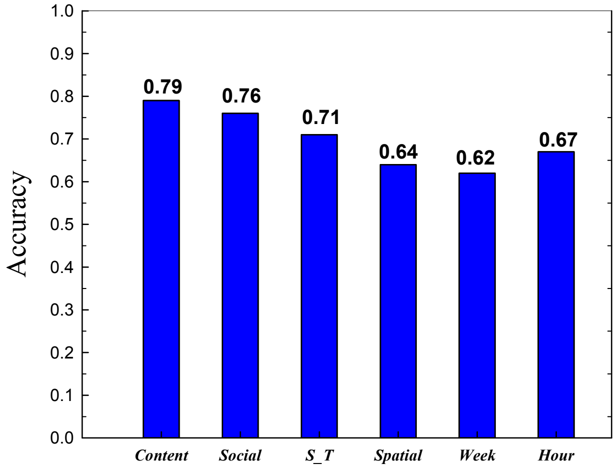


Figure 6. Prediction performance based on Decision Tree.

and has much better generalization capability due to its ability to represent users/activities more comprehensively. The neighborhood based prediction model can be combined with the matrix factorization model as follows:

$$\widetilde{r_{ua}} = {p_u}^T q_a + |N(u,a;k)|^{-\frac{1}{2}} \sum_{i\in N(u,a;k)} w_{ai}(r_{ui} - \overline{r_u}) \tag{15}$$

Here $\overline{r_u}$ means the average rate for user u and $N(u,a;k)$ consists of all the activities that are selected as the k-nearest neighbors of activity a attended by u. The parameter k refers to the number of neighbors. Specifically, we treat w_{ai} as free parameters which are learned together along with the matrix factorization model parameters. During computation, we only need to store and update the parameters for k-nearest neighbors of each activity instead of all the activity pairs. We set $N(u,a;k)$ as the k-nearest neighbors determined by the similarity measure $Sim_u\left(a, a_i\right)$, which integrates all the extracted factors.

EXPERIMENTAL RESULTS

In this section, we evaluate the proposed framework and method based on a real EBSNs data set. We first introduce the EBSNs dataset and then present the parameter learning experiments for different procedures in the framework. Afterwards, we evaluate the weights of different features in our Multi-Factor model based on a decision tree. Finally, we compare the SVD-MFN method with another three existing approaches and the results demonstrate that our method performance better.

Dataset

Our experiments were conducted on a dataset collected from Douban Events. We used APIs offered by Douban to crawl all valid activities within a specified time interval. To make the data set manageable, we only selected users from Beijing who attended more than three activities in 2013. Afterwards, we crawled the activity attendance history of these users from Feb. 2012 to Oct. 2013. In total, we got three kinds of data including:

1) User-activity data. i.e., the user's attendance history.

2) Activity attributes data. For each activity, we could get its start time, address, longitude, latitude, category, title and description.

3) Social data. To compute social influence, we also collect the relationship between users and activity hosts.

The statistics of the dataset are shown in Table 1. In order to explore the characteristics of people's activity attendances, we first take a statistic analysis about the distribution of attendee number, which is shown in Figure 5. The result belongs to power-law or long-tail distribution, which indicates that most activities were attended by a small number of users, while few activities attracted a large number of users. The average number of attendees of each activity is 31.98 in our dataset.

Based on the obtained activity attributes data, we constructed the feature model and implement our prediction method. Specifically, in the experiments, we split our dataset into two parts. The first part (February 2012 to May 2013) was used for training and the second part (June 2013 to October 2013) for testing. Meanwhile, the collected user-activity pairs were regarded as positive samples. However, to train the binary classifier unbiasedly, there should be an equal number of positive and negative samples. As there were a total of 481,325 positive samples in our data set, we randomly chose 481,325 user-activity pairs from the active users and the corresponding collection of activities that they did not attend as negative samples.

Feature evaluation

In the multi-factor model, we considered user preference from three different aspects: content, spatio-temporal contex-

Time Interval	2012-01-01 to 2013-10-01
Number of Users	15,050
Number of Activities	45,561
Number of Hosts	6,570
Number of Following Relationship with Hosts	313,479
Number of User-Activity Pairs	481,325

Table 1. Statistics of the Douban Events dataset

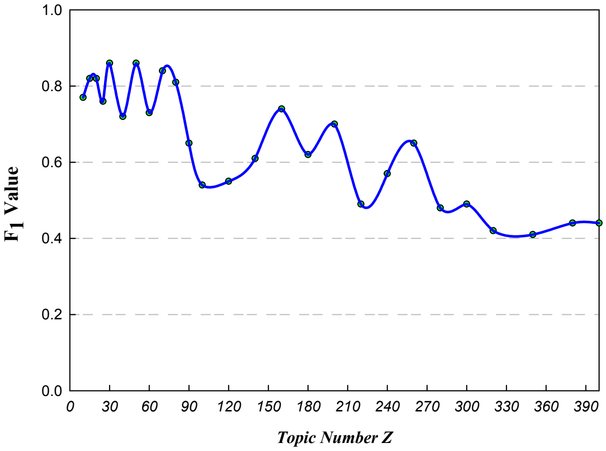


Figure 7. The F_1 Values under Different Topic Number.

t, and social influence. To evaluate their contribution to activity attendance prediction, we adopted a decision tree to predict attendance using each category of features separately. We ran 10-fold cross validation to perform prediction and adopt the accuracy as the overall performance measure. All the tests were based on WEKA [11], and the results are shown in Figure 6.

According to the results, we found that different features had distinct performances in the prediction. Specifically, the content preference achieved the best performance with an accuracy of 0.79. Social influence was in the second place with an accuracy of 0.76. Spatial and temporal context also contributed to the prediction, although its performance was not as good as the other two factors, which is shown as S_T in the figure.

Parameter learning

Parameter learning in LDA

During the extraction of content preference, we need to determine the optimal values for parameters in LDA, which are α, β and the number of topics Z. The optimal values of α and β depend on Z and the size of the vocabulary in the document collection, and they are typically set at $\alpha = 50/Z$ and $\beta = 0.01$ [19]. Z can affect the interpretability of the extracted topics and accordingly influence the performance of calculating the similarity between two pieces of text. If Z has a small value, the obtained topics will be too general to describe the result, while a large Z will result in very narrow topics.

To select the best value of topic number Z, we used a text classification method. All activities in Douban Events are categorized into 10 categories: music, film, salon, sport, commonwealth, party, travel, exhibition, drama and others. Based on the JS distance, we first calculated the content similarity between each pair of activities. Afterwards, we used k-means++ to classify all the activities into 10 clusters. Then we adopted the F_1 measure for performance evaluation [24], which is commonly used in text classification. The best topic number Z should correspond to the highest F_1 value. The result shown in Figure 7. Based on the Figure, we set the topic number as 50 where F_1 achieves the best performance which equals to 0.86.

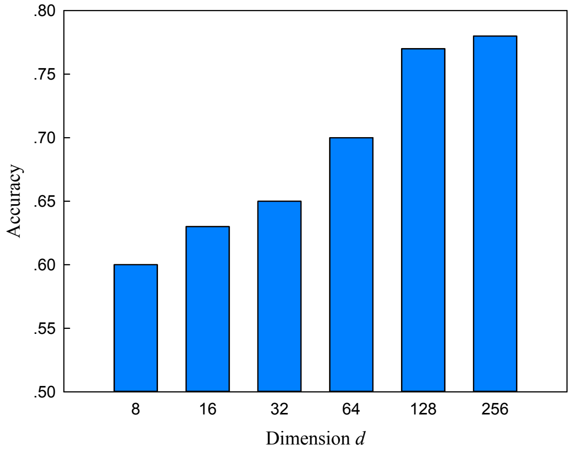


Figure 8. Performance under Different Dimensions d.

Parameter learning in SVD

The number of dimension d is the parameter we considered most closely during our experiments. We used accuracy to evaluate its influence on the performance of SVD.

As SVD is a dimensionality reduction methodology, the number of dimensions has a great impact on prediction accuracy, as shown in Figure 8. In other words, the number of dimensions is critical for the effectiveness of the low dimensional representation. To avoid over-fitting, we conducted experiments with diverse values of d ranging from 8 to 256. According to Figure 8, while the number of dimensions was low, the performance of SVD improved continuously as d increases. However, there was little performance improvement once the number of dimensions surpassed 128, where the algorithm gradually achieved optimal performance. However, the time consumption increased sharply when d was larger than 128. To balance the accuracy and time complexity, d was fixed at 128 in the following experiments, and the accuracy here was 0.77.

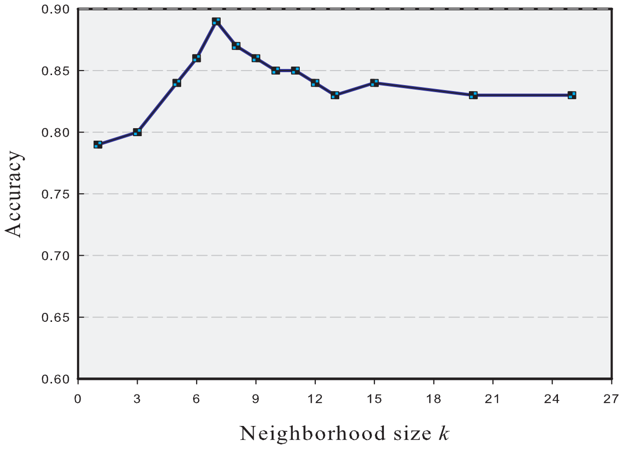


Figure 9. Performance under Different Neighborhood Size k.

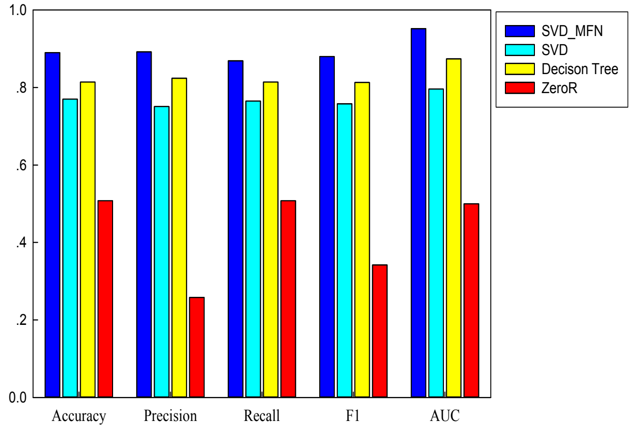


Figure 10. Performance Comparison with Other Approaches

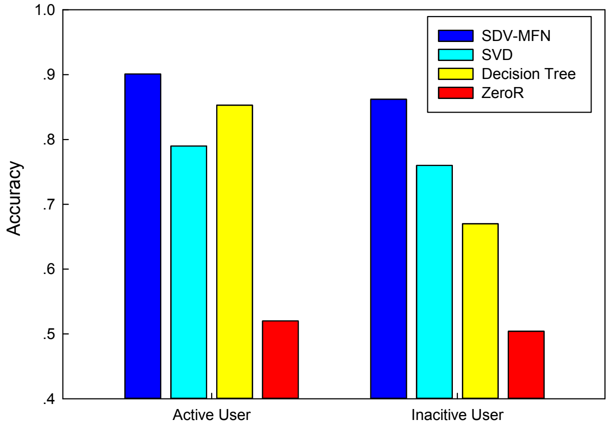


Figure 11. Performance for Different Type of Users

Parameter learning in SVD-MFN

In SVD-MFN, the main parameter that needs to be learned is the number of neighbors, which has also been proven to have a significant impact on prediction accuracy [13]. To examine the sensitivity of the neighborhood size, we conducted an experiment where the number of neighbors varied from 1 to 25 and then calculated the corresponding prediction accuracy. Figure 9 shows the experimental results.

Accordingly, the size of the neighborhood does affect the prediction accuracy. Specifically, the prediction accuracy increased as the number of neighbors increased at the beginning, and then after achieving its extreme, it started to decrease. This might be due to the fact that too many neighbors results in too much noise. In our experiment, accuracy reached its peak (0.89) when the neighborhood size was set as 7.

Comparison with other approaches

To examine the performance of SVD-MFN in attendance prediction, we compared it with three approaches, i.e.,decision tree, SVD and ZeroR. For the decision tree, we use all features in our factor model. For SVD, we set the number of dimensions at 128 and keep the other parameters the same as those in SVD-MFN. For ZeroR, it is the simplest classification method and often used as a benchmark for other classification methods.

We compared these three approaches based on different measures, including *accuracy*, *precision*, *recall*, *F-value*, and *AUC*. According to the results shown in Figure 10, we found that our approach achieved better performance than SVD, decision tree and ZeroR for all the measures. Specifically, in the case of accuracy, SVD-MFN reached 0.89, which was 0.08 higher than decision tree, 0.12 higher than SVD, and 0.39 higher than ZeroR. Similarly, the F-value of SVD-MFN was also the highest (0.88), which was 0.122 higher than SVD, 0.067 higher than decision tree, and 0.538 higher than ZeroR. In the case of AUC, the performance of SVD-MFN is as high as 0.952. SVD has a bad performance since it only used the user-activity matrix. The performance of decision tree was better than SVD, as it also considered activity characteristics. For more insights into our algorithm's performance, we also computed the value of Cohen's Kappa [2]. Cohen's kappa was first introduced as a measure of agreement between observers of psychological behavior. It was only later found that it can also be used as a meter for classifiers accuracy by evaluating the degree of agreement between the classifier and reality. The Cohen's Kappa in SVD-MFN is 0.758, which is a pretty high value and proves the good agreement of our algorithm.

Performance test for different types of users

In EBSNs, users often behave differently in terms of active level. Some users attend activities much more frequently than others. Due to this phenomenon, each user can be defined as either active or inactive. In our dataset, the average number of attendance count per user was 31.98. To gain insights into the performance of our algorithm in different type of users, we split our test dataset into two subsets: active users(attendance count $\geq$ 32) and inactive users(attendance count $<$ 32). The results are illustrated in Figure 11. We observed that compared with other approaches, SVD-MFN achieved high accuracy for both active and inactive users, which proves its consistent. The reason for its outstanding performance is SVD-MFN leveraged the merits of both SVD and multi-factor model, and experimental result verified its advantages over other approaches.

CONCLUSION AND FUTURE WORK

In this paper, we have focused on the challenging issues of activity attendance prediction in emerging EBSNs. We have explored the modeling of EBSNs users by utilizing content preference, spatio-temporal context, and social influence features and built a multi-factor model that incorporates the weights of different features. Based on this model, we obtained neighbor activities for the current user-activity pair. In particular, we proposed SVD-MFN to combine SVD with neighbor activities as well as multiple features to address the activity attendance prediction issue. Our experimental results showed that our method outperformed existing activity prediction methods. Our work is crucial for better understanding emerging EBSNs and providing critical insights that will help in personalized event recommendation and targeted advertising that can increase customer satisfaction and trust in EBSNs services.

Although the experimental results suggest that the proposed method is effective at activity attendance prediction, the results are only based on a particular social networking service, i.e., Douban Events. Therefore, in the future it will be necessary expand the SVD-MFN algorithm to other social media. Furthermore, based on the proposed prediction algorithm, we will develop a system to recommend events to users and attendees to hosts. We also want to expand our work to advertising by combining influence maximization. For example, as a host, who should he/she invites to the event in order to get the best advertising effectiveness.

ACKNOWLEDGEMENTS

This work was partially supported by the National Basic Research Program of China (2012CB316400), the National Natural Science Foundation of China (61222209, 61373119, 61332005), the Specialized Research Fund for the Doctoral Program of Higher Education (20126102110043), and Microsoft Corporation.

REFERENCES

1. L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Knowledge Discovery and Data Mining*, pages 44–54, 2006.
2. A. Ben-David. Comparison of classification accuracy using cohens weighted kappa. *Expert Systems with Applications*, 34(2):825–832, 2008.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Advances in neural information processing systems*, 1:601–608, 2002.
4. D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *World Wide Web Conference Series*, pages 757–766, 2007.
5. Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen. Model bloggers' interests based on forgetting mechanism. In *World Wide Web Conference Series*, pages 1129–1130, 2008.
6. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Knowledge Discovery and Data Mining*, pages 1082–1090. ACM, 2011.
7. C. Cornelis, X. Guo, J. Lu, and G. Zhang. A Fuzzy Relational Approach to Event Recommendation. In *Indian International Conference on Artificial Intelligence*, pages 2231–2242, 2005.
8. D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
9. E. M. Daly and W. Geyer. Effective event discovery: using location and social information for scoping event recommendations. In *Proceedings of ACM Conference on Recommender Systems*, pages 277–280. ACM, 2011.
10. T. L. Griffiths. Finding scientific topics. *Proceedings of The National Academy of Sciences*, 101:5228–5235, 2004.
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *Sigkdd Explorations*, 11:10–18, 2009.
12. J. Han, J. Niu, A. Chin, W. Wang, C. Tong, and X. Wang. How online social network affects offline events: A case study on douban. In *Ubiquitous Intelligence & Computing and International Conference on Autonomic & Trusted Computing*, pages 752–757, 2012.
13. J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Research and Development in Information Retrieval*, pages 230–237, 1999.
14. R. Klamma, P. M. Cuong, and Y. Cao. You never walk alone: Recommending academic events based on social network analysis. In *Complex Sciences*, pages 657–670. 2009.
15. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
16. D. H. Lee. PITTCULT: trust-based cultural event recommender. In *Conference on Recommender Systems*, pages 311–314, 2008.
17. Y. Li, D. Mclean, Z. A. Bandar, J. D. O'Shea, and K. A. Crockett. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18:1138–1150, 2006.
18. Y.-M. Li, C.-T. Wu, and C.-Y. Lai. A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. *Decision Support Systems*, 55(3):740–752, 2013.
19. J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.
20. X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *Knowledge Discovery and Data Mining*, pages 1032–1040, 2012.
21. A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *Knowledge Discovery and Data Mining*, pages 1032–1040, 2012.
22. E. Minkov, B. Charrow, J. Ledlie, S. J. Teller, and T. Jaakkola. Collaborative future event recommendation. In *International Conference on Information and Knowledge Management*, pages 819–828, 2010.
23. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *International Conference on Weblogs and Social Media*, pages 70–573, 2011.
24. T. Peng, W. Zuo, and F. He. Svm based adaptive learning method for text classification from positive and unlabeled documents. *Knowledge and Information Systems*, 16(3):281–301, 2008.
25. J. Sang, T. Mei, J.-T. Sun, C. Xu, and S. Li. Probabilistic sequential pois recommendation via check-in data. In *Proceedings of International Conference on Advances in Geographic Information Systems*, pages 402–405, 2012.
26. Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu. Discovering and profiling overlapping communities in location-based social networks. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(4):499–509, April 2014.
27. B. Xu, A. Chin, and D. Cosley. On how event size and interactivity affect social networks. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 865–870, 2013.
28. D. Yang, D. Zhang, Z. Yu, and Z. Yu. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 479–488, 2013.
29. M. Ye, X. Liu, and W.-C. Lee. Exploring social influence for recommendation: a generative model approach. In *Proceedings of ACM International Conference on Research and Development in Information Retrieval*, pages 671–680, 2012.
30. Z. Yu, Y. Yang, X. Zhou, Y. Zheng, and X. Xing. Investigating how user's activities in both virtual and physical world impact each other leveraging lbsn data. *International Journal of Distributed Sensor Networks*, 2014.
31. H. Zhuang, A. Chin, S. Wu, W. Wang, X. Wang, and J. Tang. Inferring geographic coincidence in ephemeral social networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 613–628. 2012.
32. J. Zhuang, T. Mei, S. C. Hoi, Y.-Q. Xu, and S. Li. When recommendation meets mobile: contextual and personalized recommendation on the go. In *Proceedings of the ACM International Conference on Ubiquitous Computing*, pages 153–162, 2011.