A Combined Collaborative Filtering Model for Social Influence Prediction in Event-Based Social Networks

Xiao Li, Xiang Cheng, Sen Su(✉), Shuchen Li, and Jianyu Yang

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
{lixiao,chengxiang,susen,lsc,jyyang}@bupt.edu.cn

Abstract. Event-based social networks (EBSNs) provide convenient online platforms for users to organize, attend and share social events. Understanding users' social influences in social networks can benefit many applications, such as social recommendation and social marketing. In this paper, we focus on the problem of predicting users' social influences on upcoming events in EBSNs. We formulate this prediction problem as the estimation of unobserved entries of the constructed user-event social influence matrix, where each entry represents the influence value of a user on an event. In particular, we define a user's social influence on a given event as the proportion of the user's friends who are influenced by him/her to attend the event. To solve this problem, we present a combined collaborative filtering model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model, by incorporating event-based neighborhood method into matrix factorization. Due to the fact that the constructed social influence matrix is very sparse and the overlap values in the matrix are few, it is challenging to find reliable similar event neighbors using the widely adopted similarity measures (e.g., Pearson correlation and Cosine similarity). To address this challenge, we propose an additional information based neighborhood discovery (AID) method by considering three event-specific features in EBSNs. The parameters of our MF-EN model are determined by minimizing the associated regularized squared error function through stochastic gradient descent. We conduct a comprehensive performance evaluation on real-world datasets collected from DoubanEvent. Experimental results demonstrate the superiority of the proposed model compared to several alternatives.

1 Introduction

In the past few years, event-based social networks (EBSNs), such as Plancast[1] and DoubanEvent[2], have proliferated to be the online platforms for users to organize, attend and share social events to be held in offline physical venues [18]. EBSNs link online and offline social worlds, providing not only typical

[1] http://www.plancast.com.

[2] http://www.douban.com.

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 142–159, 2016.
DOI: 10.1007/978-3-319-32055-7_13

online social networking services, but also face to face offline communication by attending events. Previous studies [6,26,29] have shown that users could influence others to attend events in EBSNs, especially for close social ties. For instance, when an organizer publishes an event, users can express their willingness to join the event by RSVP ("yes" or "maybe")[3] and broadcast posts about their participating information to their friends (i.e., followers), who hesitate in making decisions. When a user's friends see his/her participating post, they might want to attend the event together with the user.

Social influence aims to study the behavioral change of a person because of the perceived relationships with other people in social networks. Since it has a wide range of applications, such as social recommendation [27] and social marketing [12], considerable works have been conducted to the influence analysis or prediction in social networks (e.g., Twitter and Facebook) [2,24,25]. However, as a newly emerged online social service, EBSN has its unique characteristics, such as event location and event organizer, which leads to the social influence analysis or prediction approaches used in conventional social networks might be ineffective in EBSNs. Nevertheless, social influences of EBSN users can provide valuable insights. For an upcoming event (i.e., the event which has not been held but has been published in the EBSNs), the event organizer hopes to maximize the attendees. This goal makes him/her desire to target the influencers on this event. These influencers are able to let many friends to attend this event by sharing the event. In this case, the event organizer needs to know users' influences to their friends. Therefore, understanding users' influences is a key issue in EBSNs.

In this paper, we focus on the social influence prediction problem in EBSNs. We formulate this prediction problem as the estimation of unobserved entries of the constructed social influence matrix S, where each entry (u, e) represents the social influence of user u on event e. Notice that, we focus on predicting users' influences on upcoming events which could provide valuable information for event organizers. In particular, similar to the definition of item-level social influence in conventional social networks [7], we define user u's influence on event e as the proportion of u's friends who are influenced by u to attend e. Different from the structure-level influence [21] and the topic-level influence [17,23], the predicted event-level influence can be used in two angles. On one hand, given an upcoming social event, we could find out the influencers to attract more friends for attending the event. On the other hand, given a user, we could recommend events for him/her to share, which can improve the interactions between the user and their friends. Matrix factorization is a straightforward approach to solve this prediction problem. By using users' observed influences, we could predict their influences on the upcoming events which have already some RSVPs ("yes" or "maybe"). However, as matrix factorization does not detect associations among the closely related items (i.e., users or events), the prediction performance of this approach might be poor. To improve the prediction accuracy, a potential

[3] The RSVP ("yes" or "maybe") indicates that a user wants to attend or is interested in an event. We assume that a user will attend the events which he/she has expressed RSVP ("yes") to.

approach is to integrate neighborhood method with matrix factorization [13]. However, since the social influence matrix S is very sparse and the overlap values in S are few, it is hard to find reliable similar neighbors using the widely adopted similarity measures (e.g., Pearson correlation and Cosine similarity). Therefore, how to discover reliable event neighbors in S is a challenging problem.

In EBSNs, event content, event location and event organizer are the major components of an event which affect users' decisions in attending the event. Therefore, if two events are similar on these three aspects, we can consider these two events as similar events. To this end, we propose an additional information based neighborhood discovery (AID) method to identify event neighborhood. To find the neighborhood of a targeted event e, we first capture three event-specific features (i.e., event content, event location and event organizer) and compute the similarities between e and other events on each feature. Then, we take the most similar events on each feature as a neighborhood of e. Such that, we obtain three neighborhood sets corresponding to the three features. Finally, a neighborhood aggregation strategy is proposed to derive the final neighborhood. In particular, in such strategy, we pick up the events contained in at least two neighborhood sets to make up the final neighborhood of e.

Based on AID, we present a combined collaborative filtering model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model, to predict users' social influences on upcoming events in EBSNs. The model incorporates event-based neighborhood method into matrix factorization and thus can take advantages of both matrix factorization and neighborhood method. Model parameters are determined by minimizing the associated regularized squared error function through stochastic gradient descent. In summary, the major contributions of our work are listed as follows:

- We present a novel combined collaborative filtering model, namely, Matrix Factorization with Event-User Neighborhood (MF-EN) model, which incorporates event-based neighborhood method into matrix factorization, for social influence prediction in EBSNs. To the best of our knowledge, this is the first attempt to define and solve the event-level social influence prediction problem in EBSNs.
- To find reliable similar neighbors, we propose an additional information based neighborhood discovery (AID) method by considering event-specific features (i.e., event content, event location and event organizer) in EBSNs.
- We evaluate the performance of our prediction model on real-world datasets collected from DoubanEvent, which is the biggest event-based social network in China. Experimental results demonstrate the superiority of our MF-EN model compared to several alternatives.

The remainder of this paper is organized as follows. In Sect. 2, we give the definition of the social influence prediction problem, and show our model framework. In Sect. 3, the AID method is discussed in detail. We present our MF-EN model for social influence prediction in Sect. 4, followed by experimental evaluation in Sect. 5. We review the related work in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Preliminaries

2.1 Event-Based Social Network

Users can establish, join and share events held offline in physical venues in event-based social networks (EBSNs). Users, events and organizers are three essential types of entities in EBSNs. As shown in Fig. 1, users in an EBSN denoted as U_1, U_2, U_3 and U_4 are interconnected via social links to form an online network. They are the participants of social events. Events denoted as E_1, E_2 and E_3 contain textual content information and locations (i.e., L_1 and L_2) where they are held offline. Organizers denoted as O_1 and O_2 are a special kind of users who establish as well as attend social events. They are the owners of social events and an organizer may hold more than one event.

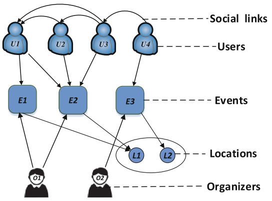

Fig. 1. The description of EBSNs

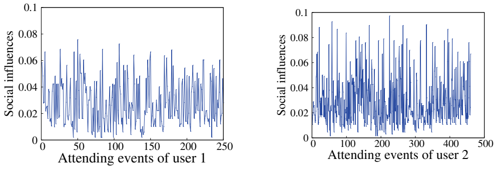

Fig. 2. Influence varies with users and events in EBSNs

2.2 Social Influence in EBSNs

Users may influence others to attend events in EBSNs. Let us consider the following scenario. Bob discovers a drama in DoubanEvent. Since he is not quite sure whether it is worth watching, Bob hesitates in making a decision to watch it until his friend Alice broadcasts her participation information on this drama. Given that Alice has broadcasted several wonderful social events before, Bob is quite confident about her taste in dramas and finally attends the event together with Alice.

We randomly select two active users in DoubanEvent, which is the largest event-based social network in China. For each user, we calculate the proportion of their influenced friends on each of his/her attending events and plot them in Fig. 2. From Fig. 2, we can observe that: different users have different social influences to their friends; users' social influences are different on different events.

2.3 Problem Definition

In EBSNs, there is a set of users $U = \{u_1, u_2, ..., u_M\}$, a set of events $E = \{e_1, e_2, ..., e_N\}$ and a set of organizers $O = \{o_1, o_2, ..., o_C\}$. For each user $u \in U$, it has a user profile (e.g., user id and username), friends (i.e., followers) collection

$F(u)$ and a set of past and upcoming events $HE_u \subseteq E$ that user u has expressed RSVP ("yes"). For each event $e \in E$, it has an event id, content information, an event organizer $o(e)$, a set of users who have expressed RSVP ("yes") (denoted as $HU_e \subseteq U$) and a physical location $l_e = \{lon_e, lat_e\}$ in terms of longitude and latitude where event e is held. For each organizer $o \in O$, it has a set of events E_o organized by him/her.

Xu et al. [26] have validated that mutual influences have effects on the event participation between friends in EBSNs by using statistics analysis. Similar to the definition of item-level social influence in conventional social networks [7], we define user u's social influence on event e as s_{ue}, which equals to $p_{ue} \Big/ |F(u)|$ (i.e., the proportion of u's influenced friends on event e), where p_{ue} denotes the number of u's friends who are influenced by u to attend event e. The social influence prediction problem can be formally defined as estimating the unknown s_{ue} according to the observed influence values. Notice that, we focus on predicting the social influence of a user on an upcoming event, on which some users have expressed RSVP ("yes"), and this user has already expressed RSVP ("yes") on some past or upcoming events before.

2.4 Model Framework

In this paper, focusing on the social influence prediction problem in EBSNs, we present a combined collaborative filtering predicting model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model, which incorporates neighborhood method into matrix factorization to improve the prediction accuracy. As shown in Fig. 3, the model is composed of three major components: social influence matrix construction, additional information based neighborhood discovery and MF-EN predicting model.

Social Influence Matrix Construction. Using each user's attending events, we construct the user-event social influence matrix S. Obviously, there are $M \times N$ entries in S, and each entry is denoted as s_{ue}. Recall that s_{ue} is user u's social influence on event e. In practice, only some elements of S can be observed and

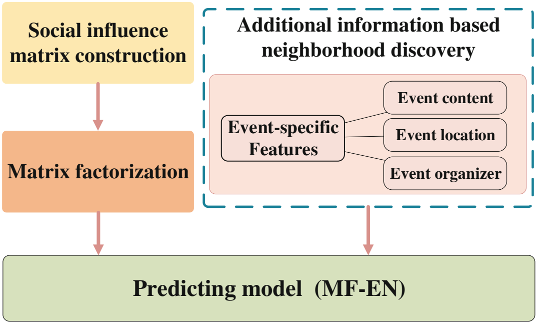

Fig. 3. The framework of the MF-EN model

the unobserved elements (of upcoming events) are represented by $\hat{s}_{ue}$. The social influence prediction problem is to predict $\hat{s}_{ue}$ by the observed s_{ue}.

Additional Information Based Neighborhood Discovery. We propose an additional information based neighborhood discovery (AID) method which takes event-specific features (i.e., event content, event location and event organizer) into consideration for neighborhood discovery.

MF-EN Model. Our proposed MF-EN model incorporates event-based neighborhood method into matrix factorization. Model parameters are learned by solving the associated regularized squared error function through stochastic gradient descent.

3 Additional Information Based Neighborhood Discovery

In EBSNs, since lots of new events are published every day and each user can only attend a small proportion of events, the constructed social influence matrix S is very sparse, and the overlap values in S are few. Therefore, it is challenging to find reliable similar neighbors by using the widely adopted similarity measures such as Pearson correlation and Cosine similarity. To address this challenge, by leveraging three event-specific features in EBSNs, we propose an additional information based neighborhood discovery (AID) method for event neighborhood discovery.

Event content, event location and event organizer are the major components of an event which affect users' decisions in attending the event in EBSNs. Therefore, we consider two events are similar events if they are similar on these three aspects. To this end, our AID method captures three event-specific features (i.e., event content, event location and event organizer) to perform event neighborhood discovery. Given a targeted $\hat{s}_{ue}$ in social influence matrix S, for each event e' in HE_u, which is a set of past and upcoming events user u has expressed RSVP ("yes"), we first compute the similarities of event content, event location and event organizer between e and e', denoted as $ES_c\left(e, e'\right)$, $ES_l\left(e, e'\right)$ and $ES_o\left(e, e'\right)$, respectively. Then, we take the most similar events on each feature as a neighborhood of e. Such that, we derive three neighborhood sets corresponding to the three features. Finally, we utilize a neighborhood aggregation strategy to find the final neighborhood. In particular, in this strategy, we take events contained in at least two neighborhood sets as the final neighborhood of e. We discuss how to determine the size of the neighborhood on each feature in Sect. 5. In the following, we will introduce how to compute the similarities between events on each event-specific feature in detail.

3.1 Event Content

Event content is the key component of an event, which plays a major role in determining users' decisions in attending the event. In order to derive the content similarity between events, we put the event content including event title, description and category as a document and obtain the event distribution on latent topics by employing the Latent Dirichlet Allocation (LDA) model [3], which is an unsupervised machine learning technique to identify latent topics from a large document.

Based on the assumption that documents are mixtures of topics, LDA models document d as a probability topic distribution, denoted as θ_d, and each topic z is represented as a probability distribution over terms in the vocabulary, denoted as ϕ_z. Like previous study [16], we first format content text of event e to document d_e, by removing the stop words from each corpus. Then each event has a corresponding document, which is taken as the input of the LDA model. Finally, we can obtain all events' document-topic distributions Θ and topic-word distributions Φ, where the topic distribution represents the varieties of an event. We suppose there are K latent topics. The generative process of LDA is as follows:

1. For each topic $z \in \{1, 2, ...K\}$, draw $\phi_z \sim Dirichlet\,(\beta)$, which is a multinomial distribution over terms.
2. For each event document d_e
   (a) Draw a topic distribution $\theta_{d_e} \sim Dirichlet(\alpha)$.
   (b) For each word $w_{d_e,n}$ in document d_e,
      (i) Draw a topic $z_{d_e,n} \sim Mult(\theta_{d_e})$,
      (ii) Draw a word $w_{d_e,n} \sim Mult(\phi_{z_{d_e,n}})$.

Given the hyperparameters α and β, the joint distribution of an event document is specified as

$$p\left(w_{d_e}, z_{d_e}, \theta_{d_e}, \Phi|\alpha, \beta\right) = \prod_{n=1}^{N_e} p\left(w_{d_e,n}|\phi_{z_{d_e,n}}\right) p\left(z_{d_e,n}|\theta_{d_e}\right) p\left(\theta_{d_e}|\alpha\right) p\left(\Phi|\beta\right), \quad (1)$$

where N_e is the number of words of event e.

The complete likelihood of N event documents is derived based on the assumption that all the documents are independent of each other:

$$p(W, Z, \Theta, \Phi|\alpha, \beta) = \prod_{e=1}^{N} p(w_{d_e}, z_{d_e}, \theta_{d_e}, \Phi|\alpha, \beta). \quad (2)$$

The model has two unknown parameters to be inferred: the document-topic distributions Θ, and the topic-word distributions Φ. We utilize Gibbs sampling [11] to estimate these parameters, which is a special sort of Markov-chain Monte Carlo (MCMC) simulation.

Given the topic distributions θ_{d_e} and $\theta_{d_{e'}}$ of events e and e', we can use the Jensen-Shannon divergence [11] to compute the content similarity between them. The Jensen-Shannon divergence is defined as follows:

$$D_{JS}(\theta_{d_e}, \theta_{d_{e'}}) = \frac{1}{2}[D_{KL}(\theta_{d_e}, \frac{\theta_{d_e} + \theta_{d_{e'}}}{2}) + D_{KL}(\theta_{d_{e'}}, \frac{\theta_{d_e} + \theta_{d_{e'}}}{2})], \quad (3)$$

where $D_{KL}(\cdot)$ is the Kullback-Leibler divergence. In particular, the value Jenson-Shannon divergence ranges from 0 to 1 and increases when the distinction between θ_{d_e} and $\theta_{d_{e'}}$ becomes larger.

Finally, given two events e and e', we define the content similarity $ES_c\,(e,e')$ $(0 \leq ES_c\,(e,e') \leq 1)$ between them as:

$$ES_c\,(e,e') = 1 - D_{JS}(\theta_{d_e}, \theta_{d_{e'}}). \tag{4}$$

3.2 Event Location

In EBSNs, event location specifies the physical place where the event is held. Compared with conventional social networks, offline social interaction is a unique characteristic of EBSNs. Therefore, event location is an important factor affecting whether a user attends an event. To better understand the impact of event location on users' behaviours, we perform a data analysis on a real-world dataset crawled from DoubanEvent. We calculate the distances between all pairs of events with different locations which each user has attended and plot the probability density function of the distance in Fig. 4 in log-log scale. As shown, we can observe that the distance probability distribution approximately follows a power law, which means that most of the event pairs which a user has attended are within a short distance. To this end, we take event location as an important factor, and consider two events are similar if their locations are close.

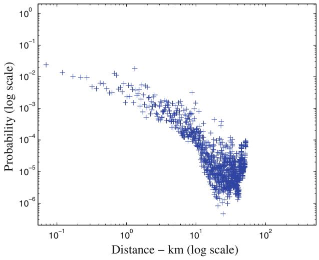

Fig. 4. Distance probability distribution

Let $dis(l_e, l_{e'})$ denote the Euclidean distance between event e and e'. We use the Gauss formula which describes the exponential decay with the distance to measure the location similarity $ES_l\,(e,e')$ $(0 < ES_l\,(e,e') \leq 1)$ between event e and e', which is defined as:

$$ES_l\,(e,e') = \exp\{-\frac{dis(l_e, l_{e'})^2}{2}\}. \tag{5}$$

3.3 Event Organizer

In EBSNs, organizers are the owners of events. Whether a user attends an event is also affected by the event organizer [8]. For example, if the organizer is a user's favorite singer, the user would be interested in attending the singer's concerts. Therefore, we consider, two events are similar if they are held by the same organizer. In particular, we define the organizer similarity $ES_o\left(e,e'\right)$ between event e and e' as a binary value function,

$$ES_o\left(e,e'\right)=\begin{cases}1 & o(e)=o(e')\\ 0 & \text{others}\end{cases}, \tag{6}$$

where $o(e)$ denotes event e's organizer and $o(e)=o(e')$ means that event e and e' are held by the same organizer.

4 MF-EN Predicting Model

To improve the accuracy of social influence prediction in EBSNs, we present a combined collaborative filtering model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model. In this section, we first introduce the basis of our model (i.e., matrix factorization), then show our MF-EN model.

4.1 Matrix Factorization

Matrix factorization is the basis of our combined model. It can be seen as the baseline approach to solve our social influence prediction problem. For the user-event social influence matrix S, matrix factorization maps both users and events into a joint latent factor space of dimensionality d, such that the influence matrix is modeled as inner products in that space [14]. User u is associated with a vector $p_u\in\mathbb{R}^d$ and event e is associated with a vector $q_e\in\mathbb{R}^d$. Both vectors p_u and q_e are referred to as d-dimensional latent factors. The social influence $\hat{s}_{ue}$ of user u on the upcoming event e can be predicted according to the following equation,

$$\hat{s}_{ue}=p_u^T q_e. \tag{7}$$

Parameters p_u and q_e are generally learned by solving the following regularized least squares problem:

$$\min_{p_*,q_*}\sum_{(u,e)\in\kappa}\left(s_{ue}-p_u^T q_e\right)^2+\lambda\left(\parallel p_u\|^2+\parallel q_e\|^2\right), \tag{8}$$

where κ is the collection of the whole observed values of the social influence matrix and the constant λ is a parameter determining the extent of regularization.

4.2 MF-EN: Matrix Factorization with Event Neighborhood Model

Similar to matrix factorization, neighborhood method [4,5] can also be used to predict the unobserved values of the social influence matrix. These two types of methods have their own advantages and disadvantages [13,22]. Neighborhood method is effective at detecting localized relationships, which focus on computing the relationships between similar neighbors. They always ignore the vast majority of values and predict values only dependent on a few significant neighborhood. While matrix factorization is effective at estimating global information and poor at detecting strong associations on closely related neighbors.

In this paper, we incorporate both event-based neighborhood method into matrix factorization by a combined collaborative filtering model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model, which takes advantages of both matrix factorization and neighborhood method to enrich each other. Let $\bar{s}_u$ denote the average social influence values of user u. The model is shown as follows:

$$\hat{s}_{ue} = p_u^T q_e + |N(u,e)|^{-\frac{1}{2}} \sum_{f \in N(u,e)} w_{ef}(s_{uf} - \bar{s}_u), \tag{9}$$

where $N(u,e)$ is the neighbor set of event e among HE_u, which is found by the AID method. Moreover, w_{ef} is the parameter which denotes the influence weights of event f to e.

Equation 9 provides a 2-tier model for social influence prediction: The first tier $p_u^T q_e$ considers the global interaction between users and events; The second tier contributes the fine grained adjustments that the event neighborhood plays roles in. All parameters in Eq. 9 can be determined by minimizing the associated regularized squared error function through stochastic gradient descent:

$$\min_{p_*,q_*,w_*} \sum_{(u,e)\in\kappa} \left(s_{ue} - p_u^T q_e - |N(u,e)|^{-\frac{1}{2}} \sum_{f \in N(u,e)} w_{ef}(s_{uf} - \bar{s}_u) \right)^2 + \lambda_1 \left(\|p_u\|^2 + \|q_e\|^2 + \sum_{f \in N(u,e)} w_{ef}^2 \right), \tag{10}$$

where λ_1 determines the extent of regularization.

5 Experiments

In this section, we evaluate the proposed model based on real-world EBSN datasets. We first describe the experimental setup including the datasets, evaluation metrics and comparison methods. Then, we evaluate the prediction accuracy of our proposed model.

5.1 Experimental Setup

Datasets. The datasets used in our experiments are collected from the website of DoubanEvent. We get the following data: (1) event information, including event id, content information (category, title and textual description), organizer id, physical location (longitude, latitude) and the set of attendees who are recorded in order of time when they express RSVP ("yes"); (2) user information, including user id, username, city and followers IDs. To make data sufficient for evaluation, we remove users who have attended fewer than 5 events (about 5 % of the total users) and events whose participants are fewer than 8 (about 3 % of the total events). After preprocessing, we get 11123 users, 29342 events, 153408 friend links and 356052 user-event pairs. The sparsity of the resulting dataset is 99.9 %.

To capture the social influence in EBSNs, similar to previous social influence studies [10,24], we consider that a user is influenced by a friend when he/she attends an event after that friend's attending. In particular, we assume that all social influences are independent from each other, thus when a user gets multiple broadcastings of an event information before he expresses RSVP ("yes") on it, we simplify the case as that the user is influenced by the latest one.

In our experiments, we randomly sample different number of users and select events attended by these users to form datasets with different sizes, including 1000 users dataset, 5000 users dataset and 11123 users dataset. In particular, we use the 5000 users dataset for parameters setting. We randomly select 50 %, 70 % and 90 % of the observed entries in social influence matrix S of different sizes of datasets (i.e., 1000 users dataset, 5000 users dataset and 11123 users dataset) for training, and the rest for testing.

Evaluation Metrics. To evaluate the accuracy of our proposed method, we adopt two popular evaluation metrics, namely, *Root Mean Square Error* (*RMSE*) and *Mean Absolute Error* (*MAE*), which are defined as:

$$RMSE = \sqrt{\frac{1}{|S'|}\sum_{(u,e)\in S'}(s_{ue}-\hat{s}_{ue})^2}, \qquad MAE = \frac{1}{|S'|}\sum_{(u,e)\in S'}|s_{ue}-\hat{s}_{ue}|, \tag{11}$$

where $|S'|$ denotes the size of the testing set S'. The smaller *RMSE* or *MAE* value indicates better accuracy.

Comparison Methods. We compare our MF-EN model with the following 7 methods.

- **Logistic Regression (LR):** If we regard the event-specific features as variables, and the values of social influences as the response, then the social influence prediction on upcoming events can be formulated as the regression problem. Thus, we use the LR model to combine the event-specific features linearly and learn the regression coefficients of these features from the training data.
- **Event Influence Mean (EM):** This method uses the mean influence value of the corresponding event to predict the unobserved values:

$$\hat{s}_{\cdot,e} = \sum\nolimits_{u\in HU_e} s_{ue} \Big/ |HU_e|. \tag{12}$$

- **User Influence Mean (UM):** Similar to EM, this method uses the mean influence value of the corresponding user to predict the unobserved values:

$$\hat{s}_{u,\cdot} = \sum\nolimits_{e \in HE_u} s_{ue} \Big/ |HE_u|. \tag{13}$$

- **Classical Event-Based Neighborhood Method (P-EN):** The unobserved values are predicted based on the values of events' neighbors discovered by using Pearson correlation:

$$\hat{s}_{ue} = |N(u,e)|^{-\frac{1}{2}} \sum\nolimits_{f \in N(u,e)} w_{ef}(s_{uf} - \bar{s}_u). \tag{14}$$

  Model parameters are determined by Eq. 15. Actually, we have also evaluated Cosine similarity as the similarity measure which gives poorer results compared with Pearson correlation.

$$\min_{w_*} \sum_{(u,e) \in \kappa} (s_{ue} - A_{ue})^2 + \lambda_2 \cdot \sum_{f \in N(u,e)} w_{ef}^2. \tag{15}$$

- **Event-Based Neighborhood Method Using Additional Information (AI-EN):** Different from P-EN, events' neighbors are discovered by using our AID method.
- **Matrix Factorization (MF):** In this method, we predict the unobserved values using matrix factorization (i.e., Eq. 7) and parameters are determined by Eq. 8.
- **HF-NMF:** If we consider the event content as the content of the web post, and the proportion of friends who are influenced by the user to attend the event as the user's social influence on the post, then the hybrid factor non-negative matrix factorization (HF-NMF) approach proposed in [7] can be used to predict the user's social influence on an upcoming event.

5.2 Experimental Results

5.2.1 Parameters Setting

Parameters of LDA. In order to achieve the content similarity between events, we select the optimal LDA parameters: α, β and the number of topics K. According to the setting in [15], we set $\alpha = 50\big/K$ and β=0.01. We evaluate *RMSE* and *MAE* under different values of K from 10 to 100. The results plotted in Fig. 5 show that the performance of LDA increases with the growth of K and there is little performance improvement after $K = 70$. However, the time consumption increases sharply when K is larger than 70. To balance the accuracy and computation complexity, we fix the value of K in LDA to 70 in our experiments.

Parameters of MF. We set $\lambda = 0.01$ for matrix factorization. How to set the dimension number d is important for the prediction performance. If d is too small, we cannot discriminate users and events in the latent space. If d is too large, the computation complexity will be greatly increased. Thus, we evaluate *RMSE* and *MAE* of MF by varying the number of latent dimensions d from

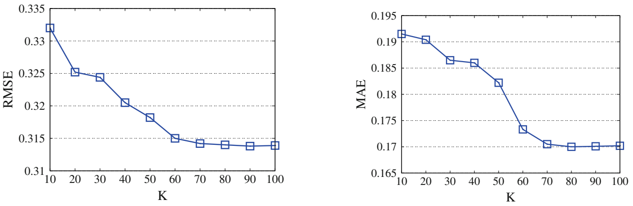

Fig. 5. Prediction performance of LDA with different topic numbers

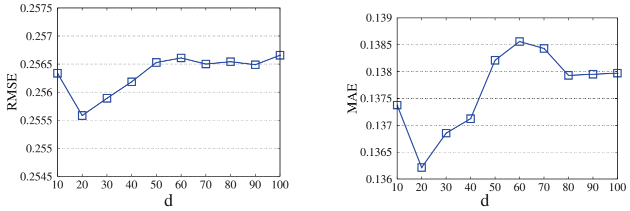

Fig. 6. Prediction performance of MF with different latent dimensions

10 to 100. The performance results are plotted in Fig. 6. According to Fig. 6, we set the dimension number as 20, where *RMSE* and *MAE* achieve the best performance.

Parameters of Neighborhood Size. In our AID method, the size of the event neighborhood set on each event-specific feature determines the final event neighborhood size. For simplicity, the size of the event neighborhood on each event-specific feature is set to be equal, which is denoted as k. Recall that, for a given event e, its neighborhood on the feature of event organizer is consisted of the events which have the same organizer with e. Therefore, the size of the neighborhood on the event organizer feature is not restricted by k (i.e., it might be smaller or larger than k). The prediction performance of our proposed methods under different neighborhood size k is shown in Fig. 7. We can observe that the size affects the prediction performance and the optimal size of these methods is different. To obtain the best performance, we set k of AI-EN and MF-EN to 100 and 70, respectively.

5.2.2 Advantages of the AID Method

In this section, we study the performance of the AID method. We compare the performance of neighborhood methods whose neighbors are found by our proposed AID approach (i.e., AI-EN) with methods whose neighbors are directly obtained by using the Pearson correlation (i.e., P-EN).

Our proposed AID method considers three event-specific features for event neighborhood discovery. In order to demonstrate the advantages of the combination of these features in the event neighborhood discovery, we also compare the performance of AI-EN with methods whose neighbors are discovered only

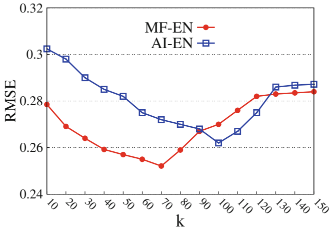

Fig. 7. Performance with different k

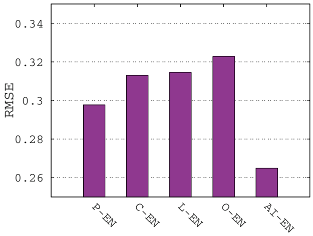

Fig. 8. Performance of neighborhood methods

by a single feature (e.g., event content). The event-based neighborhood methods using the single feature of event content (C), event location (L) and event organizer (O) for event neighborhood discovery are denoted as C-EN, L-EN and O-EN, respectively.

In Fig. 8, we plot the performance of these event-based neighborhood method. We can observe that AI-EN which discovers the event neighborhood by aggregating the neighborhood sets on the event-specific features can obtain the best performance. The reason lies in that: (1) The social influence matrix S is sparse, and similar neighbors found by the Pearson correlation are unreliable; (2) Only using the single feature of EBSNs (e.g. event content) cannot find reliable event neighborhood.

5.2.3 Performance Comparison

The prediction errors measured by RMSE and MAE of the comparison methods on different datasets are shown in Table 1 with best results highlighted in boldface. We make 4 observations from the results.

First, the proposed MF-EN model, which incorporates event-based neighborhood method into matrix factorization by considering multiple event-specific features in EBSNs, achieves the best performance measured by both RMSE and MAE in our experiments. Notice that, similar to the work [7], the values of users' event-level social influences are very small in our datasets (usually less than 0.2). Therefore, a small decrease of the RMSE (or MAE) can give significant performance improvements.

Second, the comparisons between MF-EN and MF reveal that the performance improves when incorporating event-based neighborhood method into matrix factorization. For example, MF-EN is better than MF and EN. This is because the memory-based and model-based collaborative filtering approaches have their own advantages and can enrich each other.

Third, in most of the experiments, HF-NMF performs better than MF, while MF-EN perform better than HF-NMF. This is because HF-NMF incorporates the event content information into matrix factorization. However, since MF-EN integrates event-based neighborhood method with matrix factorization by

Table 1. RMSE and MAE of all methods

Training %	Metrics	LR	EM	UM	AI-EN	MF	HF-NMF	MF-EN
1000 users								
50 %	*RMSE*	0.352436	0.338952	0.325979	0.267345	0.261651	0.258213	**0.255415**
	MAE	0.215672	0.19437	0.18908	0.135672	0.135824	0.135162	**0.133524**
70 %	*RMSE*	0.348563	0.336231	0.316227	0.255742	0.2495	0.249485	**0.246755**
	MAE	0.210435	0.190721	0.170242	0.132275	0.129875	0.129716	**0.127657**
90 %	*RMSE*	0.334436	0.326821	0.309297	0.251824	0.245768	0.244702	**0.241334**
	MAE	0.201534	0.189877	0.169782	0.131835	0.127337	0.126837	**0.124007**
5000 users								
50 %	*RMSE*	0.360245	0.341215	0.332983	0.276524	0.272651	0.266536	**0.259040**
	MAE	0.222875	0.198547	0.197218	0.137282	0.137211	0.135921	**0.133142**
70 %	*RMSE*	0.351895	0.324997	0.320721	0.262412	0.255582	0.255272	**0.252169**
	MAE	0.218854	0.189451	0.189882	0.137410	0.136214	0.135415	**0.132086**
90 %	*RMSE*	0.346537	0.310115	0.318723	0.258927	0.253286	0.252438	**0.249027**
	MAE	0.206981	0.172589	0.170071	0.135446	0.134275	0.133727	**0.129224**
11123 users								
50 %	*RMSE*	0.369857	0.349987	0.349927	0.278562	0.275315	0.274423	**0.270140**
	MAE	0.232471	0.205817	0.203802	0.138741	0.138634	0.137912	**0.135934**
70 %	*RMSE*	0.357635	0.329752	0.343552	0.264921	0.260089	0.259355	**0.253261**
	MAE	0.226934	0.190168	0.198021	0.137486	0.136987	0.136581	**0.134632**
90 %	*RMSE*	0.351537	0.318853	0.338571	0.261374	0.254758	0.254527	**0.251426**
	MAE	0.214325	0.178954	0.191872	0.135961	0.135364	0.135036	**0.131236**

considering some unique characteristics of EBSNs, such as event location and event organizer, they obtain better results than HF-NMF.

Last, the percentage of the training data has a significant impact on the prediction performance. In particular, the more training data, the lower prediction errors (i.e., measured by RMSE and MAE) the method can achieve. The reason lies in that the performance of matrix factorization is poor in the case where there is very little training data. Moreover, the prediction performance of EM is worse than all the collaborative filtering approaches.

6 Related Work

Event-Based Social Network. Liu et al. [18] firstly introduce and define the EBSN, a new type of social network which connects the online and offline social worlds. Some works study the event recommendation problem [19,31] in EBSNs. Qiao et al. [19] present a Bayesian latent factor model for event recommendation by incorporating heterogenous social relationships, geographical features and implicit ratings. Yu et al. [28] study the problem of identifying the most influential and preferable set of invitees by extending the credit distribution model. To predict the event attendance, Du et al. [8] propose a singular value decomposition with multi-factor event-based neighborhood algorithm. Formulating the group-oriented event participation problem as a novel discriminant framework, Xu et al. [26] exploit the impact of dynamic mutual influence for the social event participation prediction. Different from these works, we attempt to quantify the social influences of users in EBSNs.

Social Influence Analysis. Influence is a potential factor which affects users' behaviors. Considerable works have been conducted to qualitatively validate the existence of influence [2,20]. Anagnostopoulos et al. [2] apply the statistical test (i.e., shuffle test) to identify whether influence is a source of social correlation using the time factor in a social network system. Chin et al. [6] investigate the user behaviour on attending offline events and find that social influences exist in EBSNs and users choose to attend an event partly because their friends will attend this event. Recently, some works have been proposed to quantify the social influence in different social networks [1,10,25]. Zhang et al. [30] argue that the influence is continuously dynamic and infer the continuous dynamic social influence for temporal behavior prediction. Goyal et al. [10] propose both static and time-dependent model to capture influence probabilities from a log of past propagations. They consider user u influences user v on an action if they are friends and u performs this action before v. Zhang et al. [29] propose a unified metric to quantify the mutual user influence between social relation and geographical distance in location-based social networks (LBSNs). They evaluate the social influence of each user-pair in the participant set from a random walk perspective. Different from this work, we attempt to estimate each user's social influence to their friends on an event whose potential participants are unknown. There are also some studies focusing on the social influence in a more fine-grained level. Tang et al. [17] analyze the topic-level social influence using the probabilistic model, in which they state, individuals' influences to others could vary greatly across different topics. Weng et al. [25] measure the topic-sensitive influences of users in Twitter by taking both the topical similarity between users and the link structure into account. Embar et al. [9] present online, multi-dimensional approach for topic-specific social influence analysis. Cui et al. [7] consider a user's social influences are different on different posts, thus they define the item-level social influence and propose a hybrid factor non-negative matrix factorization approach (HF-NMF) to solve the influence prediction problem in conventional social network (e.g., Facebook and Twitter). Inspired by their work, we focus on predicting the event-level social influence in EBSNs. Since the method proposed in [7] does not consider the unique characteristics of EBSNs (e.g., event location), it cannot satisfactorily solve our event-level social influence prediction problem. To the best of our knowledge, this is the first attempt to quantify the event-level social influence in EBSNs.

7 Conclusion

In this paper, we study the problem of predicting users' social influences on upcoming events in event-based social networks (EBSNs). In particular, we define a user's social influence on a given event as the proportion of his/her friends who are influenced by the user to attend the event. To solve this problem, we present a combined collaborative filtering model, namely, Matrix Factorization with Event Neighborhood (MF-EN) model, which takes advantages of both matrix factorization and neighborhood method. In the MF-EN model, to find reliable similar neighbors, we propose an additional information based neighborhood discovery (AID) method, which takes three event-specific features (i.e., event content, event location and event organizer) into consideration. We conduct extensive experiments on real-world datasets collected from DoubanEvent. The experimental results demonstrate that our proposed model outperforms several alternatives. In our future work, we plan to incorporate the user-specific features into our model.

Acknowledgement. The work was supported by National Natural Science Foundation of China under Grant 61502047.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM, pp. 207–218 (2008)
2. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In: SIGKDD, pp. 7–15 (2008)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Cai, Y., Lau, R.Y., Liao, S.S., Li, C., Leung, H.F., Ma, L.C.: Object typicality for effective web of things recommendations. Decis. Support Syst. **63**, 52–63 (2014)
5. Cai, Y., Leung, H.F., Li, Q., Min, H., Tang, J., Li, J.: Typicality-based collaborative filtering recommendation. IEEE Trans. Knowl. Data Eng. **26**(3), 766–779 (2014)
6. Chin, A., Tian, J., Han, J., Niu, J.: A study of offline events and its influence on online social connections in douban. In: GreenCom and iThings/CPSCom, pp. 1021–1028 (2013)
7. Cui, P., Wang, F., Liu, S., Ou, M., Yang, S., Sun, L.: Who should share what?: item-level social influence prediction for users and posts ranking. In: SIGIR, pp. 185–194 (2011)
8. Du, R., Yu, Z., Mei, T., Wang, Z., Wang, Z., Guo, B.: Predicting activity attendance in event-based social networks: content, context and social influence. In: Ubicomp, pp. 425–434 (2014)
9. Embar, V.R., Bhattacharya, I., Pandit, V., Vaculín, R.: Online topic-based social influence analysis for the wimbledon championships. In: KDD, pp. 1759–1768 (2015)
10. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: WSDM, pp. 241–250 (2010)
11. Heinrich, G.: Parameter estimation for text analysis. Technical report (2005)
12. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: SIGKDD, pp. 137–146 (2003)
13. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: KDD, pp. 426–434 (2008)
14. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **8**, 30–37 (2009)

15. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991)
16. Liu, B., Xiong, H.: Point-of-interest recommendation in location based social networks with topic and location awareness. In: SDM, vol. 13, pp. 396–404 (2013)
17. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining topic-level influence in heterogeneous networks. In: CIKM, pp. 199–208 (2010)
18. Liu, X., He, Q., Tian, Y., Lee, W.C., McPherson, J., Han, J.: Event-based social networks: linking the online and offline social worlds. In: KDD, pp. 1032–1040 (2012)
19. Qiao, Z., Zhang, P., Cao, Y., Zhou, C., Guo, L., Fang, B.: Combining heterogenous social and geographical information for event recommendation. In: AAAI, pp. 145–151 (2014)
20. Singla, P., Richardson, M.: Yes, there is a correlation: -from social networks to personal behavior on the web. In: WWW, pp. 655–664 (2008)
21. Strogatz, S.H.: Exploring complex networks. Nature **410**(6825), 268–276 (2001)
22. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4 (2009)
23. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: SIGKDD, pp. 807–816 (2009)
24. Wen, Y.T., Lei, P.R., Peng, W.C., Zhou, X.F.: Exploring social influence on location-based social networks. In: ICDM, pp. 1043–1048 (2014)
25. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: WSDM, pp. 261–270 (2010)
26. Xu, T., Zhong, H., Zhu, H., Xiong, H., Chen, E., Liu, G.: Exploring the impact of dynamic mutual influence on social event participation. In: SDM, pp. 262–270 (2015)
27. Ye, M., Liu, X., Lee, W.C.: Exploring social influence for recommendation: a generative model approach. In: SIGIR, pp. 671–680 (2012)
28. Yu, Z., Du, R., Guo, B., Xu, H., Gu, T., Wang, Z., Zhang, D.: Who should i invite for my party?: combining user preference and influence maximization for social events. In: Ubicomp, pp. 879–883(2015)
29. Zhang, C., Shou, L., Chen, K., Chen, G., Bei, Y.: Evaluating geo-social influence in location-based social networks. In: CIKM, pp. 1442–1451 (2012)
30. Zhang, J., Wang, C., Wang, J., Yu, J.X.: Inferring continuous dynamic social influence and personal preference for temporal behavior prediction. PVLDB **8**(3), 269–280 (2014)
31. Zhang, W., Wang, J.: A collective bayesian poisson factorization model for cold-start local event recommendation. In: KDD, pp. 1455–1464 (2015)