Exploring Weather Data to Predict Activity Attendance in Event-based Social Network: From the Organizer's View

JIFENG ZHANG, Hunan University, China
WENJUN JIANG *, Hunan University, China
JINRUI ZHANG, Hunan University, China
JIE WU, Temple University, USA
GUOJUN WANG, Guangzhou University, China

Event-based social networks (EBSNs) connect online and offline lives. They allow online users with similar interests to get together in real life. Attendance prediction for activities in EBSNs has attracted a lot of attention and several factors have been studied. However, the prediction accuracy is not very good for some special activities, such as outdoor activities. Moreover, a very important factor, the weather, hasn't been well exploited. In this work, we strive to understand how the weather factor impacts activity attendance, and we explore it to improve attendance prediction from the organizer's view. First, we classify activities into two categories: the outdoor and the indoor activities. We study the different ways that weather factors may impact these two kinds of activities. We also introduce a new factor of event duration. By integrating the above factors with user interest and user-event distance, we build a model of attendance prediction with the weather named **GBT-W**, based on the Gradient Boosting Tree. Furthermore, we develop a platform to help event organizers estimate the possible number of activity attendance with different settings (e.g., different weather, location, etc), so as to effectively plan their events. We conduct extensive experiments and the results show that our method has a better prediction performance on both the outdoor and the indoor activities, which validates the reasonability of considering weather and duration.

Additional Key Words and Phrases: Attendance prediction, event-based social networks, weather factors, event classification

ACM Reference Format:
Jifeng Zhang, Wenjun Jiang, Jinrui Zhang, Jie Wu, and Guojun Wang. 2020. Exploring Weather Data to Predict Activity Attendance in Event-based Social Network: From the Organizer's View. 1, 1 (February 2020), 26 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Wenjun Jiang is the corresponding author

Authors' addresses: Jifeng Zhang, Hunan University, Changsha, China, jifengzhang@hnu.edu.cn; Wenjun Jiang, Hunan University, Changsha, China, jiangwenjun@hnu.edu.cn; Jinrui Zhang, Hunan University, Changsha, China, JRZhang@hnu.edu.com; Jie Wu, Temple University, Department of Computer and Information Sciences, Philadelphia, PA, USA, jiewu@temple.edu; Guojun Wang, Guangzhou University, Guangzhou, China, csgjwang@gzhu.edu.cn.

XXXX-XXXX/2020/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

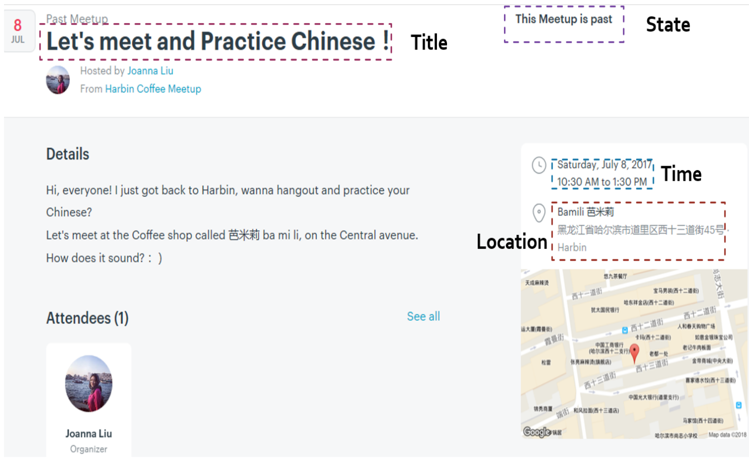


Fig. 1. Example of meetup.com about a real event.

1 INTRODUCTION

Event-based social networks (EBSNs) are becoming more and more popular in recent years. Examples include "douban.com [1], " "meetup. com [2], " "Plancast.com [3]" and so on. These social networks connect the offline world and the online world [14]. In EBSNs, the organizers can create a group with some particular hobby and regularly organize off-line activities (also called events) [17, 45]. Given a created social event, users may express their intent to join by online check-in [18, 44]. People can join one or more groups according to their own preferences, and regularly participate in some activities according to their own needs [30, 49]. Users in EBSNs can communicate online, as well as comment and upload photos like other social networks, which in turn enables similar people to communicate face-to-face in real life [51]. Moreover, online discussion and evaluation of the user's activities will also facilitate the development of future activities [19]. Many scholars have done a lot of researches on EBSNs. In this paper, we study attendance prediction for activities in EBSNs, from the organizers' view. In particular, we deeply study and exploit the effects of weather, which are usually being neglected in existing work.

Organizers have a lot of things to think about when preparing to hold an event. For example, organizers need to consider the location, the time, the content, and the description of the event. In general, the more the organizer plans and prepares for the event, the higher the quality and the popularity of the event. As a result, users who participate in the event will have a better experience [20, 53].

Many factors may impact users' attendance in an activity. Fig. 1 shows some key elements of an activity on the Meetup.com site.

- Location: the place where the activity will be held, which is usually a place where there is a large population.
- Time: the start time and the end time of the activity. The duration of the activity often depends the type of activity.
- Attendants: users who show their wishes to join the activity.
- Organizer: the person who organizes the activities and is responsible for the offline events.
- State: status of the activity.

Based on the elements of the activity, combined with user information, such as hobbies, activity history and home address, we can roughly predict whether a particular user will participate in this activity. However, many factors can affect the presence of users. Apart from the activity itself, there are several external factors that can impact user behaviors [13]. Among them, the weather is an important one. As we may know, the weather is inextricably linked to people's mood, and the mood will inevitably affect their behaviors [52]. However, we study several EBSNs and find that organizers usually do not show the weather conditions associated with the event when presenting their activity plans (as shown in Fig. 1). In literature, many researches on attendance prediction focus more on various attributes of the activity itself and the profiles of users for personalized event recommendation from the user's point of view. They also neglect external factors, including the weather, which leads to a poor prediction for specific activities (e.g., the outdoor activities) [34].

Our motivation. Based on the above analysis, we try to study the effects of weather factors and explore them to improve attendance predictions from the perspective of the organizers. Event organizers may need to estimate the number of attendances in two typical scenarios: one is for initiating an event (e.g. announce the event); the other is adjusting the event (e.g., one or two days before the event or even several hours before the event). We try to provide a comprehensive prediction of different attendance numbers in different settings (including different weather conditions). Based on this, the organizer can plan or adjust the event more flexibly.

Our contribution. In this paper, we classify activities as outdoor or indoor, and we study the different effects of weather factors on these two kinds of activities. We also identify several other factors including user interests, user-event distance, and duration for comprehensive prediction. The main contributions of this paper can be summarized as follows.

- We deeply study the impact of weather on user attendance. We distinguish the direct and indirect impacts of fine-grained weather factors such as temperature, humidity, and wind force.
- We identify several features and build a model of the attendance prediction with weather based on a gradient descent decision tree (called $\boldsymbol{GBT\text{-}W}$) from the organizer's view. It comprehensively integrates the internal factors of users and events, as well as the external factors of weather and distance.
- We collect the data in real EBSNs and gather the corresponding weather data to test our method extensively. The results validate the effectiveness of our work.
- Based on the prediction model, we develop a simple event hosting platform for organizers. It can recommend top-k choices (e.g., the time, the location, and the weather) for an event.

In Section 2, we review the related work. In Section 3, we formally define the problem and introduce our dataset. We identify all possible features/factors in Section 4 and describe the proposed attendance prediction model in Section 5. Next, we conduct experiments and analysis in Section 6. We build a simple event recommendation application for organizers and describe how to apply our work in practice in Section 7. Finally, we sum up our work and discuss some future directions in Section 8.

2 RELATED WORK

This section gives a brief introduction to the related works, focusing on three main research areas: EBSNs, activity recommendation, and attendance prediction.

Research on EBSNs. Many researches qualitatively study the characteristics of EBSNs through data analysis. Liu et al. [26] presented a comprehensive introduction to EBSN in five aspects, which has drawn increasing attention to EBSNs. Han et al. [15] analyzed the behavior of online users' participation in offline activities using the Douban dataset, which checks the relevance of several factors and whether the user participates in the activity. Xu et al. [39] conducted intensive analysis to study the characteristics and user behaviors in real activities, which focuses more on the interaction between social networks and activities, including the preservation or change of social networks and the creativity and diversity of activities. She et al. [32] analyzed the deficiencies of existing EBSNs in the activity arrangement, and proposed a more intelligent EBSN platform that provides personalized event planning for each participant. Zhang et al. [48] study the group dynamics in EBSNs. The above articles introduce the three main bodies of organizers, activities, and users in EBSNs, which leads to related research on activity recommendation and attendance prediction in this field.

Activity recommendation. Liu et al. [26] were the first to propose the issue of activity recommendation and illustrate the main challenge of the cold-start problem. Macedo et al. [28] considered the related background of a user's participation in an activity, including temporal, spatial, content, and membership, assuming that these are all positive effects. Aliannejadi et al. [6] utilized user tags to match location keywords and used many machine learning methods to make personalized recommendations for the location of the event. Wang et al. [37] focused more on the social influence of event organizers and team members in the event recommendation system. Liu et al. [27] analyzed the existing method mismatch with EBSNs, and proposed a Bayesian latent factor model, which can jointly formulate some types of data for friend recommendation to better participate in offline events and enhance user experience. Tu et al. [35] improved the effectiveness of the activity recommendation by finding suitable friends for the user and appropriately referring to the opinions of the friends. However, most of the current research is based on user-centered activities, and there are few studies from the perspective of the organizer. Our paper aims to address this issue.

Attendance prediction. Attendance prediction is similar to activity recommendation but the perspective is different. Liu et al. [25] described the relationship between the success of an event and the group, classifying social groups according to the four characteristics: group, members, activities, and structure. Jiang et al. [17] analyzed event participants for a single organizer based on semantic information, geographical information, and social network information. Feng et al. [10] studied the selection of influential organizers for influence maximization. Zhang et al. [50] discussed the probability of users' participation in one certain activity in the future by considering the time, space, and content of activities. Du et al. [9] divided the impact factors into three aspects: activity content preference, time space, and social influence. Zhang et al. [46] paid more attention to the attributes of user participation in the event, such as preferences and the influence of the user. Lu et al. [34] combined multiple factors such as time, space, background, activity content, users themselves, and mutual influence. Wu et al. [38] proposed a model to explore the dynamic nature of personal presence over time based on LSTM.

Existing works on attendance prediction usually exploit multiple factors of activities and users, such as spatio-temporal factors [9], the context of event [34, 46], social influence [10], and so on. In the prediction algorithm, existing research has used Singular Value Decomposition (SVD) [9], Logistic Regression [46], Naive Bayes [46], Classification and Regression Tree (CART) [34]. In this paper, we try to incorporate two new factors: the internal factor of event duration and the external factor of weather conditions. We divide

Table 1. Notations

Notation	Explanation
C_e	the category of event e
H_u	the set of all the events that a user u has attended
D_I	comfort index
T	the average temperature of one day
F	the relative humidity of one day
L_e^u	the duration of event e for which user u attends
V_e	the weekend feature of event e
I_e^u	the interest degree of user u on event e
D_e^u	the distance feature between user u and event e
P_e^u	the duration feature of user u attending event e

events into outdoor events and indoor events to deeply study the effects of weather on attendance.

3 PROBLEM STATEMENT

In this section, we first formulate the activity attendance prediction problem, and then present our solution framework. The notations are shown in Table 1.

3.1 The activity attendance prediction problem

In EBSNs, the organizers hold a series of activities/events, $\{e_1, e_2, e_3, \ldots, e_m\}$, each of which has a set of participants, $\{u_1, u_2, u_3, \ldots, u_n\}$. The activities can be classified into many categories $\{c_1, c_2, c_3, \ldots, c_k\}$. Given a set of users and a future event e, the tasks of activity attendance prediction are to (1) predict whether a user will attend or not by exploiting all possible features including the weather impact, user interests, and duration; (2) provide the possible attendance numbers according to different settings for organizers, including different dates, locations, weather conditions and so on. The main challenges are listed as follows:

- There are many kinds of factors. Among them, the weather factor has not been deeply studied in previous research. We want to deeply study the different roles that weather conditions play in activity attendance for different types of activities.
- Since there are more users who do not participate in the event than those who do, the selection of negative samples is especially important for the prediction model. Thus, we need to carefully select the negative samples.

3.2 Solution Overview and Data set

Solution overview. Keeping the attendance prediction for organizers in mind, we focus on exploiting the weather factor and several other factors to build a comprehensive prediction model. Our solution has three main steps: (1) Feature extraction; (2) Sample filtering; and (3) Model building.

We first study several key factors affecting attendance, including weather conditions, interests, home-event distance, and duration. We analyze the different ways in which these factors affect different types of activities. The framework of our solution is shown in Fig. 2. It has three steps.

- Step 1: process the activity feature and measure the effects of weather.
- Step 2: select the samples for building model.

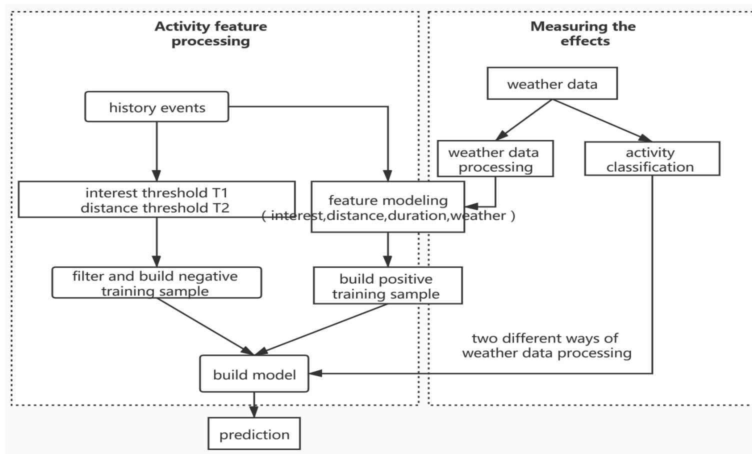


Fig. 2. System framework.

- Step 3: build the prediction model with the above processed feature vectors based on Gradient-Boosting trees ($\boldsymbol{GBT}$).

$\boldsymbol{GBT}$ algorithm has been validated to be very effective in dealing with classification problems. To serve event organizers, we also use our model to create a simple application to help event holders better plan an event.

Dataset. The dataset we use is collected from a popular EBSN: *Meetup.com*. It is collected through the Meetup API from the two cities of London and New York from January 2018 to January 2019, including some information about public events such as participants, duration, topic, and time of events. There are a total of 257,249 users and 146,636 events. There are 63,320 events held in London and 83,316 events held in New York. Among them, the number of outdoor events in London and New York are 6,750 and 9,702, respectively. The details of the dataset are shown in Table 2. There are 32 categories in our dataset, as shown in Table 3. We can see that the outdoor activities (e.g., *outdoor/adventure* and *sports*) occupy a large portion in daily activities. Besides that, we also collect the corresponding weather data from *tianqi*.2345.*com* [4], including the temperature, the weather type (e.g., rainy or sunny), the wind force and the humidity, in London and New York. We will use these data for feature modeling and attendance prediction.

4 FEATURE MODELING

In this section, we first analyze the factors that affect the attendance with our dataset. Then we classify the activities, and take into account the different effects on various activities.

4.1 Factors Analysis

We exploit all possible factors affecting attendance through the collected dataset, including user interests, user-event distance, duration, and the weather factor.

4.1.1 User Interest. Users may have many different hobbies, and various hobbies may encourage them to participate in more activities. We count the number of activity categories that users in the dataset participated in, as shown in Fig. 3. We can see that over 40% of users participate in more than one category of events. Most people prefer to enjoy many categories of events. So, the willingness of users to participate in the event is strongly related to whether the user is interested in the topic of the event.

Table 2. Statistics of dataset

Indicators	London	New York
Number of events	63,320	83,316
Number of users	130,558	126,691
Number of outdoor events	6,750	9,072
Number of indoor events	56,570	74,244
Number of outdoor events in the first quarter	1,697	2,584
Number of indoor events in the first quarter	15,368	21,839
Number of outdoor events in the second quarter	2,038	2,738
Number of indoor events in the second quarter	14,269	20,283
Number of outdoor events in the third quarter	1,706	2,301
Number of indoor events in the third quarter	13,694	16,798
Number of outdoor events in the forth quarter	1,310	1,450
Number of indoor events in the forth quarter	13,238	15,322
Number of User-event pairs	639,789	575,686

Table 3. Activity categories in Meetup

socializing	singles	language/ethnic identity
tech	outdoors/adventure	sports/recreation
fitness	new age/spirituality	dancing
movies/film	games	religion/beliefs
LGBT	writing	fashion/beauty
pets/animals	parents/family	book clubs
paranormal	sci-fi/fantasy	photography
supportmovements/politics	community/environment	hobbies/crafts
music	fine arts/culture	education/learning
career/business	health/wellbeing	movements/politics
new age/spirituality	food/drink	

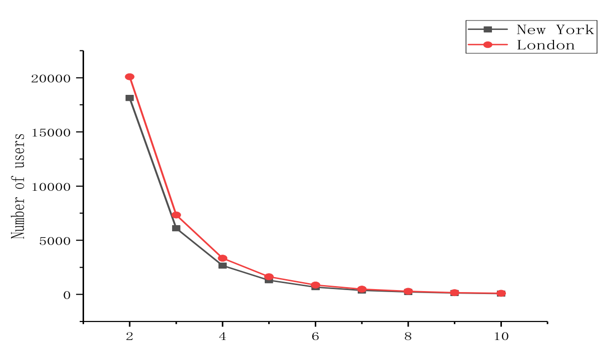
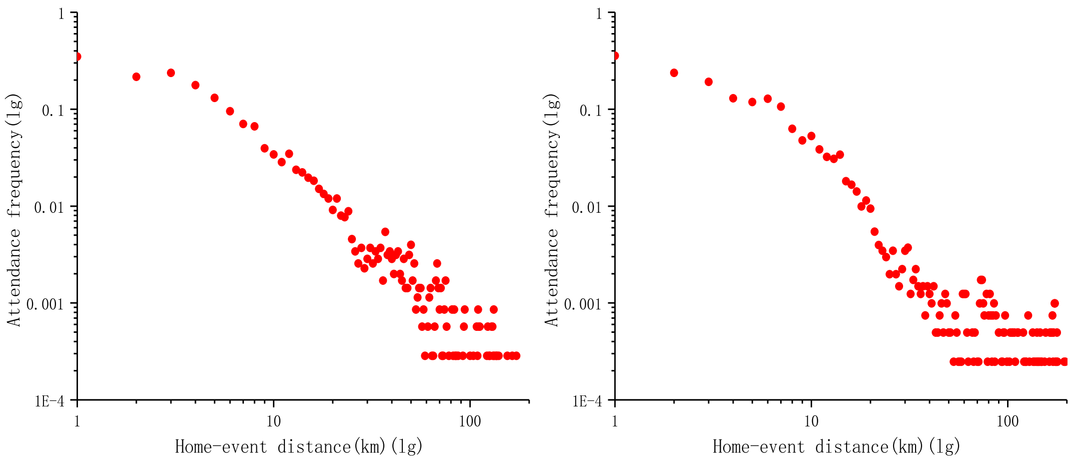


Fig. 3. The number of users that attend more than one category of events.

We use "user interest" to represent user's preference for a certain type of activity. The interest degree of user u on event e, I_e^u is calculated by the following formula:

$$\begin{aligned} I_e^u &= \frac{N_{C_e}^u}{N_{H_u}}. \\ N_{C_e}^u &= |\{e_i \in C_e \cap e_i \in H_u)\}| \end{aligned} \tag{1}$$



(a) New York (b) London

Fig. 4. Distribution of home-event distance vs. the probability of event participation.

In Eq. (1), $N^u_{C_e}$ is the number of events that user u participates in and belongs to C_e; N_{H_u} is the number of activities that u has attended.

4.1.2 User-Event Distance. In real life, if a user is interested in an event, he may attend the activity. However, if his location is too far from the event location, he may abandon his attending plan [33]. The location of the event is especially important in social networks [21]. Therefore, we define the user-event distance to account for this phenomenon. We analyze our dataset from the aspect of the attendance probability with respect to the home-event distance in Fig. 4.

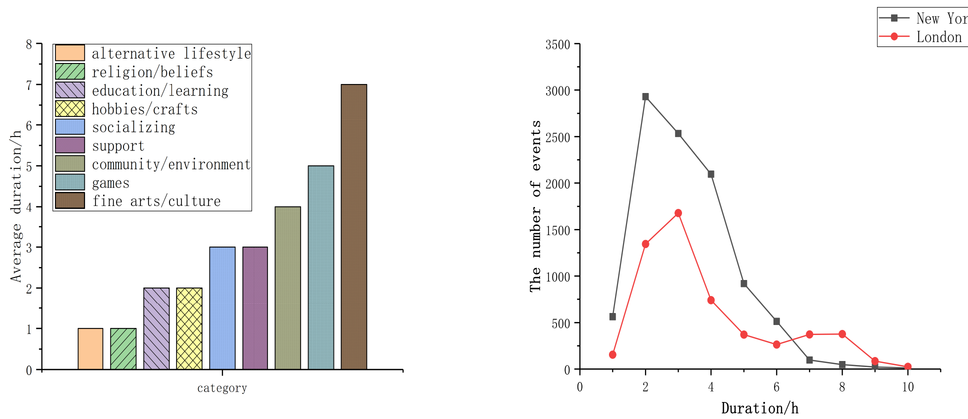
It indicates a power-law distribution: in other words, most users attend nearby events and are less likely to attend events that are far away [34]. Based on this analysis, we define the feature of distance D^u_e for users as:

$$D^u_e = z * d(u,e)^b, \tag{2}$$

where $d(u,e)$ is the distance between user u's home location and event e's location, z and b are coefficients that can be learned via curves fitting. In our paper, z=0.382, b=-0.843 in New York and z=0.390, b= -0.887 in London.

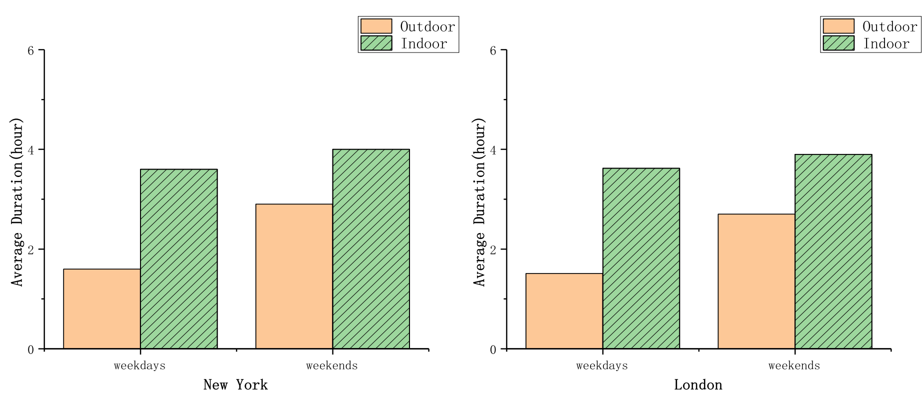
4.1.3 Duration. Duration is an important feature of the activity [7]. We often participate in activities during our free time. Therefore, duration may be a feature that users will pay attention to. Especially for outdoor activities, weather conditions and the event duration are very important to users' willingness to attend. For example, an event of writing communication may last for an hour, a movie watching activity lasts two to three hours, and a concert may last longer.

We analyze the dataset for the duration (shown as in Fig. 5(a)). It indicates that the average duration of different types of activities is quite different. Among them, "game" or "culture" activities usually have a longer duration, which take more than 7 hours, "Religious" activities or "learning" activities have a relatively shorter duration, usually 1-2 hours. Then, we deeply study the category *socializing*. We find that the duration of different events is also quite different. As shown in Fig. 5(b), 90% events take 2-5 hours. Few of them last



(a) The average duration in some event categories. (b) The number of events for different durations in the category *socializing*.

Fig. 5. The duration statistic in our data set.



(a) (b)

Fig. 6. Average duration of events in weekdays and weekends.

shorter or longer. Therefore, we use a method to determine the duration of one kind of activity. The duration of user u on event e, P_e^u is calculated as follows.

$$P_e^u = \frac{L_e^u * N_{C_e}^u}{\sum_{e_i \in C_e} L_{e_i}^u} \tag{3}$$

In Eq. 3, C_e is the category of event e, and L_e^u is defined as the duration of u on the event e. We first calculate the average duration of the activities that the user u attends and that belong to category C_e, and then compute the ratio between it and duration of the activity e.

Meanwhile, the existence of the weekend will also affect people's willingness to attend the event. Users may prefer to participate in activities that allow them to enjoy their vacation. Fig. 6 shows that in both two cities, the average duration of the events are longer on the

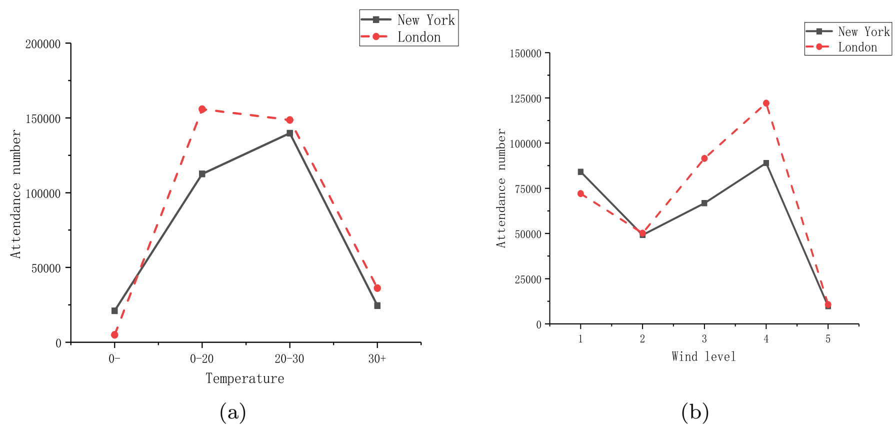


Fig. 7. The number of attendance with respect to different weather conditions

Table 4. The sample records of weather in our dataset.

Time	Temperature	Weather	Wind Force	Humidity
2018-03-01	-1°C/-4°	Cloudy Snow	East Wind 4	46%
2018-03-02	1°C/-1°	Overcast Snow	East-northeast Wind 4	32%
2018-03-03	-1°C/-1°	Moderate Snow	East Wind 5	53%
2018-03-04	3°C/3°	Sleet Rainy	East Wind 3	26%

weekend. Therefore, we define V_e as the weekend feature of event e. If the event e is held at the weekend, we set V_e as 1, otherwise we set V_e as 0.

4.1.4 Weather Factor. Bad weather has a great impact on urban traffic [36]. Meanwhile, many activities can be affected by the weather [16]. We collect the weather data corresponding to our activity data, so as to investigate whether the weather has an impact on user attendance, and if so, how it affects the user's willingness to attend. Table 4 shows a sample of weather data in our dataset. Its main attributes include the date, the temperature, the weather, the wind force, and the relative humidity. First, we count the number of user attendance in different weather conditions. As shown in Fig. 7, for different weather conditions there are significant differences in the number of attendees. Therefore, we can say that the weather does have an impact on the willingness of users to attend an event.

Next, we will explore how weather conditions affect people's attendance willingness. Before this, we need to quantify the weather data to facilitate the calculation in our model. In this article, we propose two methods: frequency based indexing and human comfort based indexing, as follows.

Frequency based indexing. Before indexing, we first process the temperature attribute. According to Jiwei et al. [23] , while mood state is not sensitive to temperature, it is significantly sensitive to temperature change. That is, a dramatic change in temperature will make people feel uncomfortable. Therefore, we first calculate the temperature changes. Next, we define the index rules as follows:

(1) Count the occurrence times of each case according to attributes, and sort it in descending order according to the number of times;

(2) Assign 0 to the most frequent occurrences and 1 to the second most frequent occurrence, so on and so forth.

Taking Table 4 for example, if the value of a temperature appears the most times (e.g., -1 in Table 4), then the temperature will be assigned the index 0. Similarly, for the weather type, the most frequent type will be assigned index 0, the second most frequent type with index 1, so on and so forth. Finally, a record in the data may be represented with "2018-03-01 0 1 0 5" after processing. The temperature, the weather, the wind direction and the humidity we process are denoted as W_1, W_2, W_3, and W_4 respectively.

Human comfort based indexing. We use the human comfort index D_I to measure people's feelings about the weather conditions. It reflects the body's comfort with the environment in a certain temperature, humidity condition, and wind speed, with different combinations of temperature and relative humidity. Through investigation, we select an existing method that is frequently used by the meteorological administration [31]. The index of human comfort index D_I is calculated as follows:

$$D_I = 1.8T - 0.55(1.8T - 26)(1 - F) - 3.2\sqrt{W_s} + 32, \quad (4)$$

where T is the average temperature, F is the relative humidity, and W_s is the wind speed. Generally, the range of D_I is 0-90. When the value is too high or too low, humans will feel uncomfortable. For example, when the temperature is 24°C, the relative humidity is 50%, and the average wind speed is 0.4 m/s, people will feel comfortable. In this case, the value of D_I is 69. When the temperature is 10°C, the relative humidity is 95% and the average wind speed is 4 m/s, people will feel slightly cold and a little uncomfortable. In this case, the value of D_I is 43. Fig. 8 displays the change of D_I over time in details. We can observe the weather characteristics of the two cities very well. The range of D_I in New York is larger due to there being four distinct seasons, while that range in London is more concentrated due to there being less weather change over time.

In this subsection, we quantify the weather data in two ways. These two processing methods will be applied to two types of activities. The frequency based indexing treats each attribute as an independent feature. We will study the impact of each attribute on outdoor activities. The human comfort based indexing combines multiple attributes together. We will study its effect on indoor activities. We will deeply study the effects of the weather factor on different activities in the following part.

4.2 The impacts of weather on different events

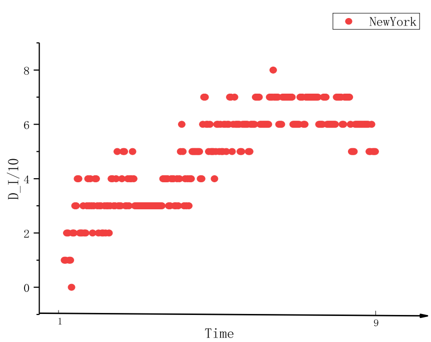
The impact of weather on different kinds of activities may be different. For outdoor activities, a bad weather may directly affect the activities; for indoor activities, the bad weather may indirectly affect people's willingness to participate in activities, and make it harder to arrive at the location of events [41]. Therefore, in order to better differentiate the effects of weather factors on different activities, we first classify the activities into indoor events and outdoor events. The common outdoor events and indoor events are listed in Table 5. Meanwhile, we classify the impacts of weather factors on activities, which are divided into direct and indirect impacts.

Direct Impact. For outdoor activities, weather factors have a direct impact. Therefore, in the treatment of outdoor activities, we use the weather factors directly as a feature vector.

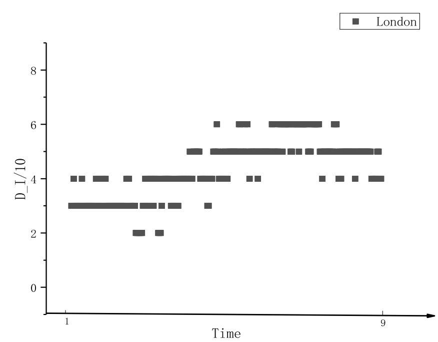
Indirect Impact. The impact of weather data on indoor activities cannot be neglected. For example, bad weather makes it more difficult for users to travel [22]. Even if the distance

Table 5. Common activity classification

Categories of outdoor events	Categories of indoor events
Outdoor/adventure	Film
Environmental Protection	Fitness
Outdoor Photography	Culture/Writing
Car/Bicycle	Fashion/Clothing
Sports	Women



(a)



(b)

Fig. 8. D_I of everday in New York and London.

from the user to the event venue is not very far, users may feel unwilling to go to the destination and give up the participation [8]. This is because the uncomfortable weather has increased our subjective distance to the event destination, exceeding the actual user-event distance. This is a dimension of psychological distance, which is called spatial distance [42]. It is related to the actual distance as well as other factors [5]. The attendance probability of indoor events is indirectly influenced by the changed distance factor.

Next, we define the way that human comfortableness affects the subjective distance. When the human comfortableness index is too high or too low, people will be comfortable and they will not be very satisfied with the weather conditions. In this case, their subjective distance will be longer than the actual distance, and a longer distance will reduce the possibility of participation. The subjective distance D_s^u is calculated as follows.

$$D_s^u = \ln(|\lfloor D_I/10 \rfloor - 6| + \mathrm{E})^{\beta} * D_e^u, \tag{5}$$

where D_e^u is the actual distance, that is, the user-event distance defined in Section 4.1.2. E is a natural constant, i.e., $\mathrm{E} \approx 2.71828$ When D_I is too low or too high, people will be uncomfortable. We use $D_I/10 - 6$ to reflect the impact of comfortableness. β represents the extent to which the weather affects the distance. We will discuss the effects of β in details in the experiment.

5 ATTENDANCE PREDICTION: THE GBT-W MODEL

In this section, we establish the prediction model ***GBT-W*** for activity attendance from the view of organizers. We model attendance prediction as a binary classification problem. We first select positive and negative samples, then we propose the prediction model based on the Gradient Boosting Tree.

5.1 Sample Selection

When establishing a classification model, we need a certain number of positive and negative samples [24]. We can easily use the users who take part in activities as positive samples (denoted as U_{pos}). The main difficulty lies in the selection of negative samples (U_{neg}). In a large group of users, the number of users who do not participate in an activity is much larger than the number of those who do. Therefore, in order to avoid data imbalance, it is very important to select proper negative samples.

We can solve the data imbalance by random sampling that selects the negative samples with the same quantity as that of positive samples. However, random sampling neglects the fact that there are many different types of users who do not participate in an event [12]. Some people may not want to go for special reasons. Most people are actually unaware of such activities. If we select these people to build negative samples, it will have a great side impact on the prediction model. Therefore, we must pick out the right users who can accurately influence the attendance prediction, filtering out those who are unrelated.

We first filter out unrelated users by user interest and user-event distance. Users have a low probability of participating in activities that are very far away. Meanwhile, if someone doesn't like a certain type of activity, then his relevance to the activity will be very low. Based on the two phenomena, we introduce two thresholds to filter samples [40]. They are the threshold of the interest (denoted as T_1) and the threshold of distance (denoted as T_2). The two thresholds are used to select users with appropriate interest or user-event distance.

Next, we construct the set of negative samples (denoted as U_{neg}). We first generate the initial sample set from all users who do not participate in event e, and whose interest is larger than T_1 or whose distance is smaller than T_2, as follows.

$$U_{ca} = \{u \notin U_e : (I_e^u > T_1) \cup (D_e^u < T_2)\} \tag{6}$$

$$U_{neg} = Sample(U_{ca}), \tag{7}$$

where $u \notin U_e$ represents users who do not participate in event e. U_{neg} is generated by random sampling of U_{ca}. Then we mix U_{pos} and U_{neg} as the training sets and build model ***GBT-W*** based on the gradient descent decision tree [47].

5.2 Attendance Prediction Model

With the generated positive and negative samples, we can build the attendance prediction model ***GBT-W***.

Preliminary: Gradient-Boosting trees (***GBT***) is used to predict attendance. ***GBT*** algorithm has been validated to be very effective in dealing with classification problems, and achieves a higher accuracy than other methods [43]. ***GBT*** uses a weak classifier to train on the basis of the residual after each classification [11]. The ***GBT*** gets the results of multiple iterations and takes the weighted average value. Moreover, ***GBT*** uses another iterator to enhance the performance [29].

Prediction method. In order to better integrate the weather factor to improve the attendance prediction, we treat the indoor and outdoor activities separately. For the outdoor events, we use the processed weather data to fuse the feature. For the indoor events, we use the temperature and relative humidity to calculate the human comfortableness. The details are shown in Algorithm 1. We use the extracted feature vectors in combination with the Gradient-Boosting trees to make prediction. For each candidate user, we can use our algorithm to predict whether he will attend the event e. We sum up all the number of users whose prediction result is 1 (i.e., attendance) as the total attendance number of the event e.

ALGORITHM 1: *GBT-W*: Attendance prediction with weather based on the ***GBT***

Input: a set of test users U, an event e, multiple different possible weather data of e
Output: Prediction Result R (an array of binary values, each element is for a user) for different weather conditions
1 Let R_u^e be the prediction result (a binary value) of a user u, indicating whether he will attend the event e, 1 for yes and 0 for no.
2 **for** *each weather condition of* e **do**
3   **for** *each* u *in* U **do**
4   Calculate the interest degree I_e^u by Eq.1.
5   Calculate feature of distance D_e^u by Eq.2.
6   Calculate feature of duration P_e^u by Eq.3.
7   Qualify weather data (temperature, weather type, wind force and humidity): W_1, W_2, W_3, W_4.
8   **if** e *is outdoor event* **then**
9   F_e^u=I_e^u+D_e^u+P_e^u+V_e+W_1+W_2+W_3+W_4;
10   **end**
11   **if** e *is an indoor event* **then**
12   Calculate human comfort index D_I by Eq.4.
13   Calculate actual distance D_s^u by Eq.5.
14             $D_e^u = D_s^u$.
15   F_e^u=I_e^u+D_e^u+P_e^u+V_e.
16   **end**
17   Precision result $R_u^e = MODEL_{GBT}(F_e^u)$.
18   **end**
19   return R.
20 **end**

Organizers often do not get accurate weather conditions when planning an event. Therefore, in Algorithm 1, we take multiple different weather conditions as the input, and our algorithm can predict the user's attendance at events under different weather conditions. This allows event organizers to adjust their event plans in time according to the weather changes.

In Algorithm 1, we treat indoor and outdoor activities separately. For the outdoor events (lines 8-10), we use the processed weather data to fuse the feature. For the indoor events (lines 11-16), we only use the temperature and relative humidity to calculate the human comfortableness. F_e^u is the fusion feature about all aspects of processed features. For outdoor events (line 9), we combine the features of user interest I_e^u, distance feature D_e^u, duration feature P_e^u, weekend feature V_e, weather feature W_1, W_2, W_3, and W_4 to obtain the fusion feature. For indoor events (line 15), we combine the features of user interest I_e^u, distance feature D_e^u, duration feature P_e^u, and weekend feature V_e to obtain the fusion feature. When we get all the feature vectors for prediction, we sort them by the starting time of activities. We use the earliest 80% to train the model and use the next 20% as the testing set.

We also build some other models for comparison in experiments. We build the model ***GBT-1*** by removing feature P_e^u in F_e^u, the model ***GBT-2*** by not distinguishing between outdoor and indoor activities, the model ***GBT-3*** by not distinguishing activities, the model ***GBT-4*** by removing feature I_e^u in F_e^u, and the model ***GBT-5*** by removing feature D_e^u in F_e^u.

For a future event e, event organizers may not get the exact weather conditions when planning events. Our model can provide multiple predictions with multiple weather conditions for the organizers, so that they can conduct more comprehensive planning. Moreover, as the time gets closer to the event date, organizers can use our model to adjust their plans, so as to make the event more successful and attract more people to participate.

6 EXPERIMENTS

In this section, we evaluate the proposed ***GBT-W*** model with the crawled real EBSNs data set as described in Section 3.2. We test the performance of our model from several aspects and compare it with other existing methods. Our experimental environment is Hadoop 2.6.0, Spark 1.6.0, Python 3.6.4. We use five-fold cross-validation and our experimental classifier training takes about 42 minutes.

6.1 Experimental settings

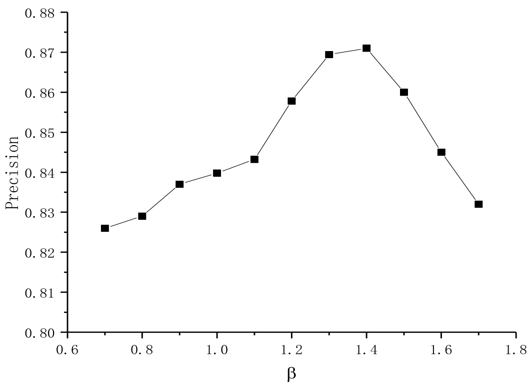
Evaluation Metrics. We compare these methods based on two common metrics of precision and recall.

$$\begin{aligned} &Precision = \frac{N_{TP}}{N_{TP}+N_{FP}}, \\ &Recall = \frac{N_{TP}}{N_{TP}+N_{FN}}. \end{aligned} \tag{8}$$

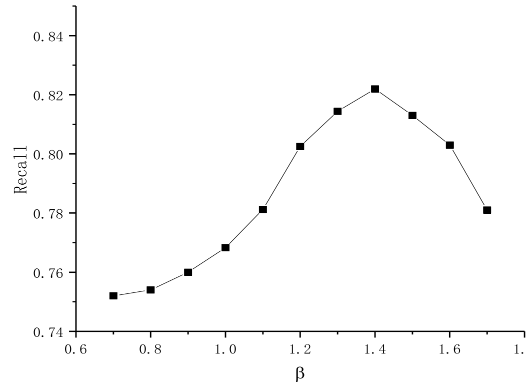
in which N_{TP} is the number of attendances whose both actual and the predicted results are positive (i.e., attendance). N_{FP} represents the number of users whose actual result is absence but whose prediction result is attendance. N_{FN} is the number of users whose actual result is attendance and the prediction result is absence.

Baseline models. It is worth noting that, it is difficult to conduct fair comparison with other attendance prediction models which exploit other factors. However, their prediction models are generally based on typical classifiers including Singular Value Decomposition[9], Decision Tree [34], etc. Therefore, we compare with different classifiers in our experiments. To examine the effectiveness of our method, we compare it with the following models:

- ***GBT-1***: It models all features apart from duration. It is used to validate the impact of duration on event attendance.
- ***GBT-2***: It does not distinguish between outdoor activities and indoor activities, and the weather data is processed in the same way as outdoor activities. It is used to test the impact of different weather factor treatments on event attendance.
- ***GBT-3***: It does not use the thresholds mentioned in Section 5 to select the negative samples. It selects all users that haven't attended the event as the negative samples. It is used to check the effects of different negative sample selection methods on event attendance.
- ***GBT-4***: It models all features apart from user interest. It is used to check the impact of user interest on event attendance.
- ***GBT-5***: It models all features apart from user-event distance. It is used to validate the impact of the distance on event attendance.



(a) Precision



(b) Recall

Fig. 9. Prediction Performance with different β in New York.

- ***SVD-MFN***: It combines user preferences, distance factors, and puts feature vectors into a matrix to predict individual attendance by using Singular Value Decomposition (SVD).
- ***DT-W***: It models all features with the decision tree algorithm.
- ***RT-W***: It models all features with the random forest algorithm.
- ***GBT-W***: It is our proposed attendance prediction model. Our ***GBT-W*** model combines user interest, user-event distance, weather data, and event duration to conduct the prediction with decision tree based on gradient boosting.

Among the above baselines, ***GBT-1***, ***GBT-2***, ***GBT-3***, ***GBT-4***, ***GBT-5*** are used to test the effects of different components or features of our work. ***SVD-MFN***, ***DT-W***, ***RT-W*** are used to test classifiers.

6.2 The effects of Parameter β

In this section, we study the impact of parameter β. When our model predicts the attendance of users in indoor activities, we study the changes in human psychology caused by weather factors. We propose to use the human comfort index to measure people's feelings about weather conditions, and to differentiate actual user-event distance and subjective distance of user psychology. The parameter β shown in Eq. 5 indicates how the subjective distance in our psychology is affected by the actual distance and the comfort index. The prediction performance with different β values in two cities are shown in Fig. 9 and Fig. 10, respectively.

GBT-W achieves the best results when β is 1.4 in New York and 0.8 in London. When its value is too large or too small, the performance of the model drops quickly. We analyze the reason and find that if it is too large, our model will exaggerate the influence of weather factors on people's psychology, and the subjective distance we obtain will be much larger than the actual distance. If β is too small, the proportion of the influence of weather factors becomes extremely small, which also reduces the performance. Meanwhile, we can see that the best β of the two cities are different. The value of β in New York is larger than that in London. It indicates that users in New York may be affected more by bad weather, resulting in lower willingness to participate in the event. The possible reason may be that the weather in London is more erratic, indirectly causing users in London to endure bad weather. In

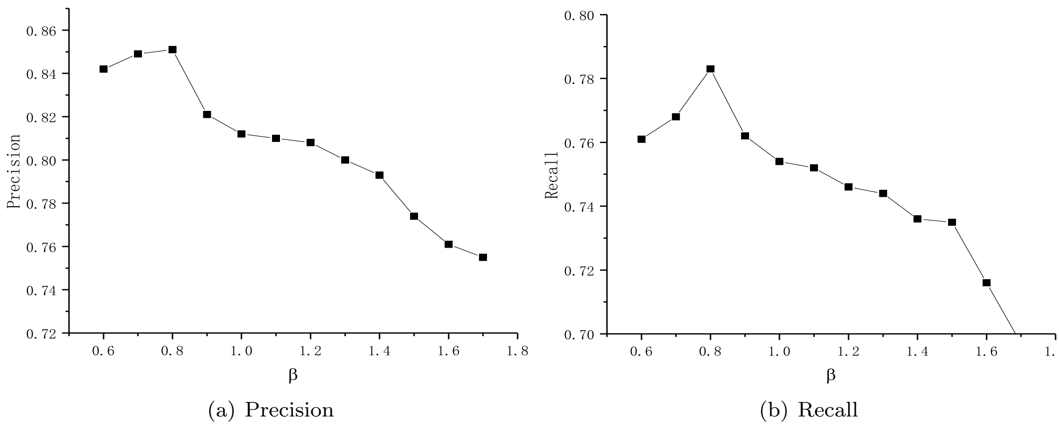


Fig. 10. Prediction Performance with different β in London.

contrast, in New York, where the four seasons are distinct, people have little tolerance for bad weather.

6.3 Performance analysis

We conduct comparison with baseline models in terms of precision and recall.

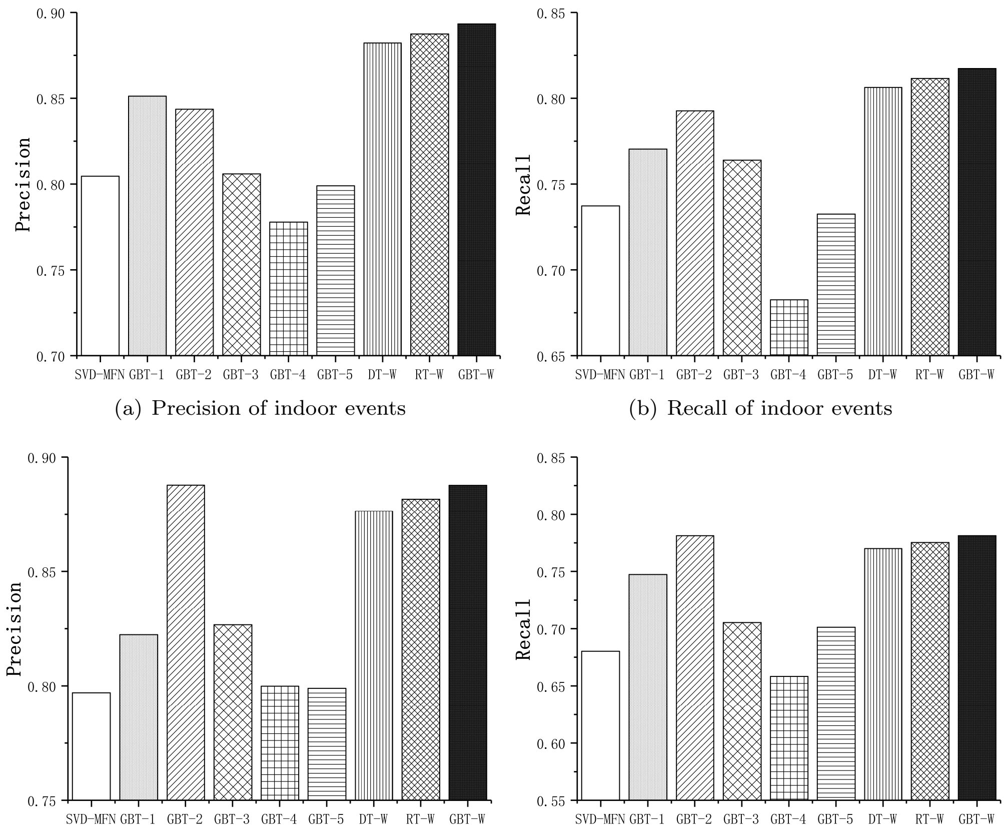
Overall performance. The results of New York and London are shown in Fig. 11 and Fig. 12, respectively. It shows that our ***GBT-W*** model has good performance, whose precision and recall reach over 80% in both New York and London. Moreover, ***GBT-W*** can even reach more than 85% in both cities, whose precision and recall are 8.9% and 9.1% higher than that of ***SVD-MFN*** respectively in New York; the improvements are 14.8% and 13.6% in London. It validates that considering the duration and weather factors in the prediction helps improve the accuracy.

Meanwhile, we observe that the performance of ***GBT-4*** (neglect interest) and ***GBT-5*** (neglect distance) are much worse, which indicates that the interest and the distance have very important impact on attendance prediction. ***DT-W*** and ***RT-W*** have a little worse performance, whose precision is 2% and 1.3% lower than taht of ***GBT-W***. It indicates that the ***GBT*** algorithm is more suitable for our prediction model.

Moreover, the overall performance in London is not as good as that in New York. There is a possible reason: The weather in London is more uncertain, which may influence the performance of the model and increase the uncertainty. Meanwhile, the uncertainty of weather in London may make users care more about the duration of the event, especially those who participate in outdoor activities. Therefore, the performance differences between ***GBT-1*** and ***GBT-W*** in London are bigger than that in New York, since ***GBT-1*** neglects the duration.

In the following part, we will compare ***GBT-1***, ***GBT-2***, ***GBT-3*** with the integrated ***GBT-W*** model, to check the effects of our model components.

The effects of considering duration. Our research integrates one factor : activity duration. The model ***GBT-1*** builds the prediction model without the factor of duration. Fig. 11 and Fig. 12 show that, in both New York and London, ***GBT-W*** has a better performance than that of ***GBT-1***, which is 6.6% and 7.5% higher in precision and recall respectively on average. It indicates the necessity of considering the duration. Moreover, the



(a) Precision of indoor events (b) Recall of indoor events

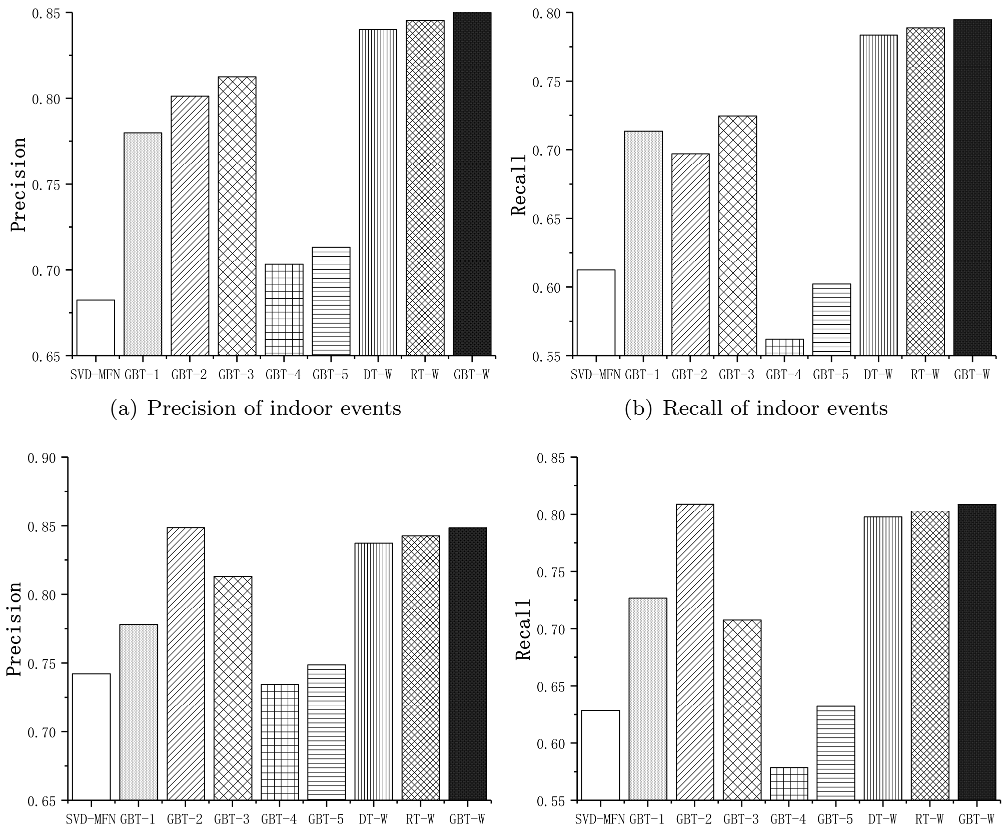
(c) Precision of outdoor events (d) Recall of outdoor events

Fig. 11. Overall Performance in New York.

differences between the two models in the outdoor events are bigger than the changes in the indoor events, which is 6% and 3% higher in outdoor and indoor events respectively on average. It may be because compared with indoor activities, users pay more attention to the duration when they participate in outdoor activities.

The effects of processing weather data. ***GBT-2*** treats the weather data in all events the same as for outdoor events, neglecting the difference between indoor and outdoor events. From the results in Fig. 11 and Fig. 12, we can see that for outdoor events in both cities, the performances of ***GBT-2*** and ***GBT-W*** are the same. This is because they process weather data in outdoor events the same way. As for the indoor activities, ***GBT-W*** performs better than ***GBT-2***, whose average performance is 5% and 6.1% higher in precision and recall, respectively. It indicates the necessity of differentiating the impacts of weather on different activities (i.e., the outdoor and the indoor activities).

The effects of generating negative samples. As mentioned before, we select negative samples by the two thresholds of user interest and user-event distance. The model ***GBT-3*** doesn't utilize the thresholds to select negative samples. Fig. 11 and Fig. 12 show that ***GBT-W*** has much better performance than ***GBT-3*** in the indoor activities. It reaches 5.5% and 5.32% higher than ***GBT-3*** in precision and recall respectively in London, and 6.3% and 7.4% in New York. Meanwhile, for outdoor events, the improvements of ***GBT-W***



(a) Precision of indoor events

(b) Recall of indoor events

(c) Precision of outdoor events

(d) Recall of outdoor events

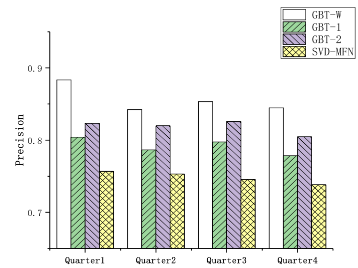
Fig. 12. Overall Performance in London.

over ***GBT-3*** are 5.1% and 3.4% in precision and recall in London, and 6.1% and 6.5% in New York, respectively. It indicates that utilizing the two thresholds helps improve the prediction accuracy. Moreover, the improvement of outdoor events is a bit lower than that of indoor events. It may be because there are fewer records of outdoor events than indoor events.

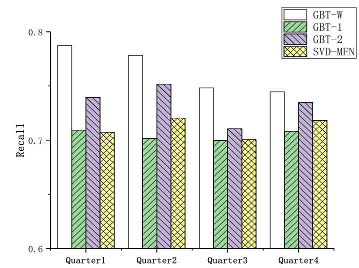
6.4 Deep analysis by quarter

Generally, the weather is different in different quarters. In order to test the performance of our prediction model in different quarters, we divide the experimental data into four quarters according to the event starting time, and sort by the time. Then we select the earliest 80% in these four quarters to train the model and use the next 20% as the testing set. The results are shown in Fig. 13 and Fig. 14 for New York and London, respectively.

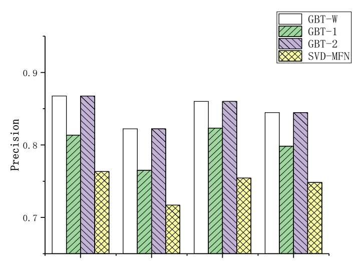
Performance in different quarters. Overall, the prediction performance in New York is better than that of London, which is 3.2% higher than London. The possible reason is that the climate characteristics of the two cities are different. Compared with London, New York has more regular climate change and four distinct seasons. This feature can improve prediction performance.



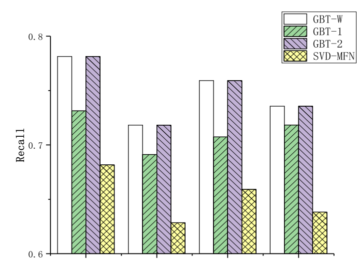
(a) Precision of indoor events



(b) Recall of indoor events

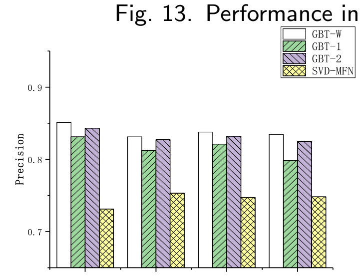


(c) Precision of outdoor events



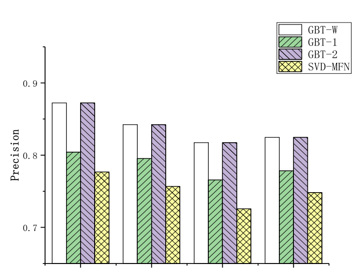
(d) Recall of outdoor events

Fig. 13. Performance in different quarters in New York.

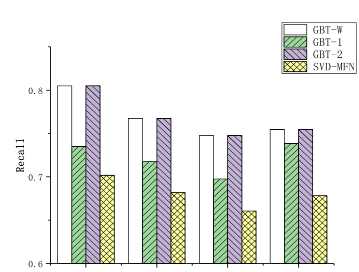


(a) Precision of indoor events

(b) Recall of indoor events



(c) Precision of outdoor events



(d) Recall of outdoor events

Fig. 14. Performance in different quarters in London.

We also find that in New York, the performance of outdoor activities decreases first and then rises with the change of the quarters. One possible reason for this is the four distinct seasons in New York, where temperatures gradually increase over time, making users may pay more attention to the effects of low or high temperatures on their participation in the first and the third quarters. Meanwhile, there is no significant trend for indoor activities in New York like there is for outdoor activities. The reason is that the temperature in the second or the third quarter is relatively comfortable and thus it does not have much impact on the user's travel to indoor events.

Performance in London (as shown in Fig. 13) is very different from that in New York (as shown in Fig. 14). In the four quarters, the performance of outdoor activities has gradually decreased significantly. This may be due to London's climate characteristics. Unlike the distinct seasons in New York, London is like spring all the year round; but London has more rainy weather. Excessive rainy weather and climate change can have a large impact on users' participation in outdoor activities, and this random situation leads to a decline in predictive performance. However, the performance of indoor activities has not declined very quickly, which may be due to the fact that the changing weather has less impact on users' participation in indoor activities.

6.5 Performance for different kinds of users.

The frequencies with which users use social networks are quite different. Therefore, users can be divided into active users and inactive users [9]. In our dataset, the average number of activity attendances for users is 5. The details of active users and inactive users are shown in Fig. 15. In New York, about 6% of people participate in activities more than 10 times, and the number of participants between 5 and 10 is 20.4 percent. 73.6% of users participate in fewer than 5 events. Meanwhile, in London, about 2.5% of people participate in activities more than 10 times, and the number of participants between 5 and 10 is 13.2%. 84.3% of users participate in fewer than 5 events.

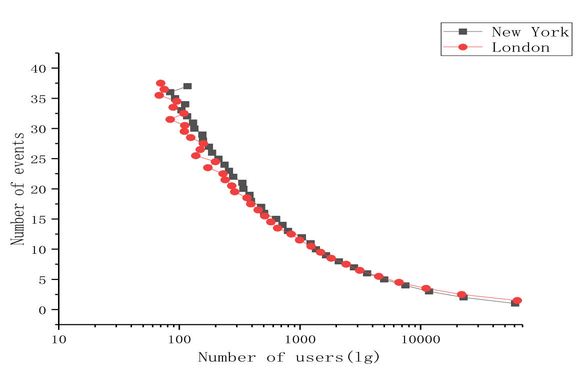
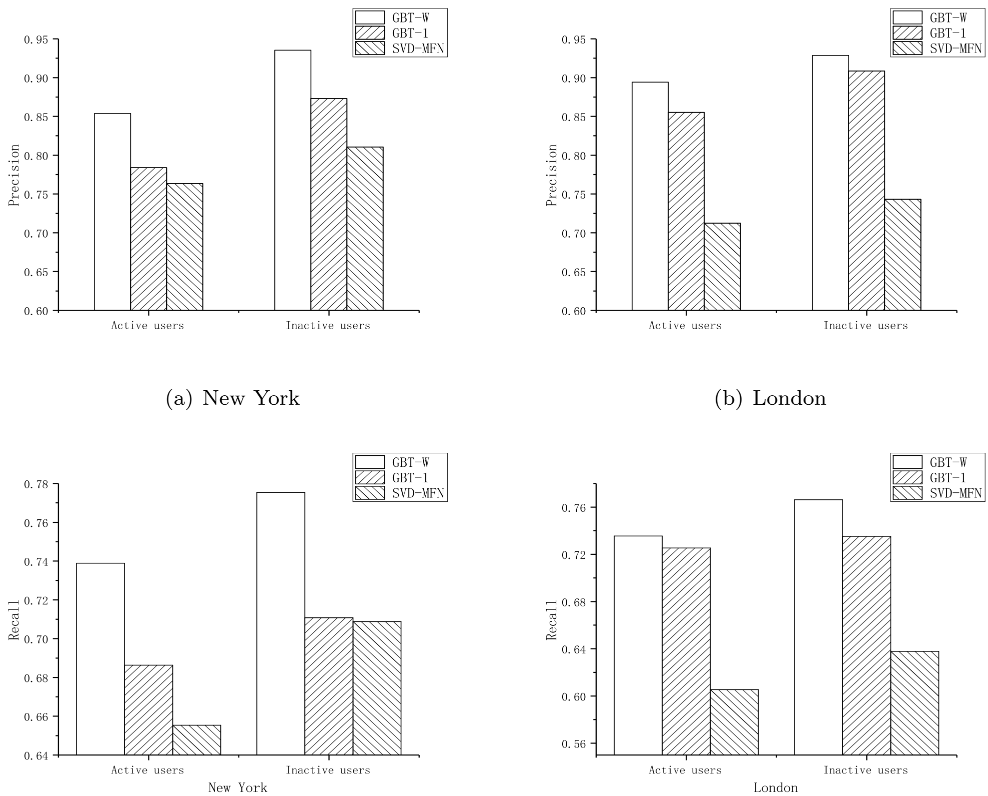


Fig. 15. Statistic of the Number of Users Participating in Activities.

Based on the above statistics, we take 5 as a threshold, and we divide users in our dataset into inactive users and active users. The prediction results of the two users sets are shown in Fig. 16. The proposed ***GBT-W*** model achieves higher precision for both active users and inactive users. It validates the effectiveness of considering the weather factor. Moreover, it is a bit strange that the performance of inactive users is better than that of active users. We analyze the reason and find that it may be because the number of inactive users is more than that of active users.



(a) New York (b) London

(c) New York (d) London

Fig. 16. Performance for different kinds of users in New York and London.

6.6 Summary of experiments

We conducted intensive experiments to validate the effectiveness of the proposed attendance prediction model and test the effects of several key features. Our main findings are summarized as follows:

1. Considering the weather factors in attendance prediction can improve the prediction performance, however the impact of the weather on attendance varies in different cities and for different activity classifications.

2. It is necessary to consider the duration of the event and the selection of negative samples in attendance prediction, which can improve the performance of the prediction.

3. Our model has good prediction performance for both active and inactive users.

4. Different from existing attendance prediction models, our work in this paper focus on deeply study the impact of weather on event attendance of different activities, and exploit it for enhancing attendance prediction from the view of organizers. Intensive data analyses and experiments validate the effects of weather, as well as the effectiveness of our prediction model.

7 APPLICATION OF OUR MODEL FOR EVENT ORGANIZER

In this section, we discuss how our model can help organizers to hold a popular event. Event organizers only need to propose the type of activities they are preparing, as well as several options for the event time and the event venue. Our attendance prediction model can integrate possible factors (e.g. the weather) to predict the number of attendances and recommend top-k settings for the organizers. We implement a demo to illustrate the application, as shown in Fig. 17.

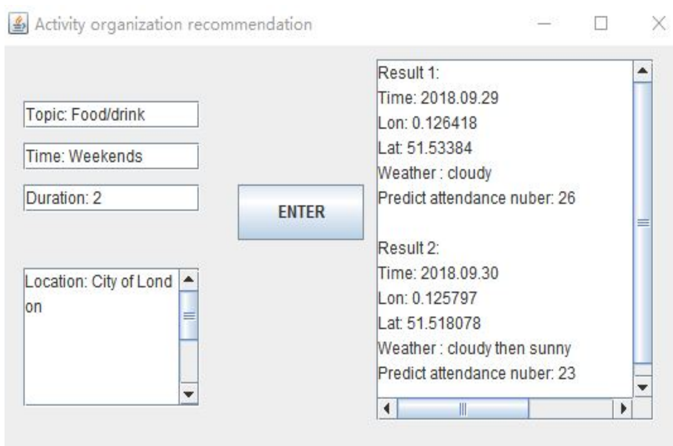


Fig. 17. The demo of activity attendance prediction and recommendation.

It can be seen from Fig. 17 that our method can recommend several candidate settings according to the choices of the organizer, and provide the number of possible attendees to the organizers. This process is somewhat different from traditional activity recommendation, which recommends personalized activities based on the target participant's information. Our work is able to provide the organizers with top-k settings for preparing activities and provide possible attendance numbers with respect to different settings.

Discussion. It is worth discussing more deeply that: if there is no accurate weather information, then how can our model work? First of all, our model can provide multiple choices according to different settings (including different weather). Second, we crawl more data to check the creation time and starting time of the events in our event data for London and New York. We calculate the proportion of the interval days with [0,5], [6,10], [11,15], [16,20], [21,25], [26,30], and more than 30, as shown in Fig. 18.

It indicates that about 30% of events are held within half a month after the events were created, about 38% are held within 20 days, about 45% within 25 days, and about 50% within 30 days. In this case, organizers can check the weather forecast to estimate the weather conditions of the event day when publishing the event. For other cases in which an event is published many days or even weeks before the event day, the organizer can refer to several possible choices by our prediction and make a full consideration. Moreover, he can also adjust some details when the date is closer. Furthermore, thanks to the development and new research in weather forecast, people can expect more accurate weather information for longer time before the event. Therefore, we would say that weather information may limit a little of our prediction model.

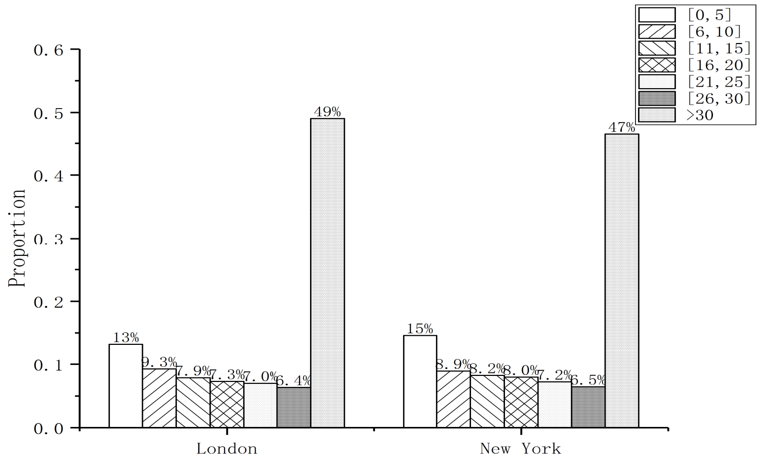


Fig. 18. Time interval (days) between event being created and event day.

8 CONCLUSION AND FUTURE WORK

In this paper, we study the effects of the weather factor in activity attendance prediction for organizers in EBSNs. We build an attendance prediction model by extracting and combining the internal and external factors of events, including users' interest, user-event distance, the duration of the event, and the weather factor. We extensively study the impact of weather on different types of indoor and outdoor activities. Experimental results in our collected real dataset validate the effectiveness of our method. We also develop a demo to show how our work can help the organizers hold activities.

In future work, we will improve our model to update the prediction according to the possible evolution/update of weather or other factors. We are also interested in studying more factors that impact the attendance of activities. We will further consider the internal connections among those factors, such as the air pressure in the environment and the traffic factors. We would also like to apply our method into more EBSNs.

ACKNOWLEDGEMENTS

This research was supported by NSFC grant 61632009, Guangdong Provincial NSF Grant 2017A030308006, Open project of Zhejiang Lab 2019KE0AB02, and in part by NSF grants CNS 1824440 and CNS 1828363.

REFERENCES

[1] https://www.douban.com/
[2] https://www.meetup.com/
[3] http://www.plancast.co.uk/
[4] http://tianqi.2345.com/
[5] A, S., W, P., N, P.: The psychological distance of climate change. Risk Analysis **32**(6), 957–972 (2012)
[6] Aliannejadi, M., Crestani, F.: Personalized context-aware point of interest recommendation. ACM Trans. Inf. Syst. **36**(4), 45:1–45:28 (Oct 2018). https://doi.org/10.1145/3231933
[7] Chen, J., Jiang, W.: Context-Aware Personalized POI Sequence Recommendation, pp. 197–210 (11 2019). https://doi.org/10.1007/978-981-15-1301-5.16
[8] Dong, Y., Jiang, W.: Brand purchase prediction based on time-evolving user behaviors in e-commerce. Concurrency and Computation: Practice and Experience **31**(1) (2019), https://doi.org/10.1002/cpe.4882
[9] 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: ACM International Joint Conference on

Pervasive and Ubiquitous Computing. pp. 425–434 (2014)
[10] Feng, K., Cong, G., Bhowmick, S.S., Ma, S.: In search of influential event organizers in online social networks pp. 63–74 (2014)
[11] Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis **38**(4), 367–378 (2002)
[12] Fu, T., Wang, C., Cheng, N.: Deep learning based joint optimization of renewable energy storage and routing in vehicular energy network. IEEE Internet of Things Journal pp. 1–1 (2020). https://doi.org/10.1109/JIOT.2020.2966660
[13] Georgiev, P., Noulas, A., Mascolo, C.: The call of the crowd: Event participation in location-based social services. Computer Science (2014)
[14] Guo, L., Zhang, D., Wang, Y., Wu, H., Cui, B., Tan, K.L.: Co 2: Inferring personal interests from raw footprints by connecting the offline world with the online world. ACM Transactions on Information Systems **36**, 1–29 (03 2018). https://doi.org/10.1145/3182164
[15] Han, J., Niu, J., Chin, A., Wang, W., Tong, C., Wang, X.: How online social network affects offline events: A case study on douban. In: International Conference on Ubiquitous Intelligence and Computing and International Conference on Autonomic and Trusted Computing. pp. 752–757 (2012)
[16] Jeuring, J.H.G., Peters, K.B.M.: The influence of the weather on tourist experiences: analysing travel blog narratives. Journal of Vacation Marketing **19**(3), 209–219 (2013)
[17] Jiang, J.Y., Li, C.T.: Analyzing social event participants for a single organizer. In: Tenth international AAAI conference on web and social media (2016)
[18] Jiang, W., Wu, J., Li, F., Wang, G., Zheng, H.: Trust evaluation in online social networks using generalized flow. IEEE Transactions on Computers (TC) **65(3)**, 952–963 (2016)
[19] Jiang, W., Wu, J., Wang, G., Zheng, H.: Forming opinions via trusted friends: Time-evolving rating prediction using fluid dynamics. IEEE Transactions on Computers (TC) **65(4)**, 1211–1224 (2016)
[20] Jiang, W., Wang, G., Bhuiyan, M.Z.A., Wu, J.: Understanding graph-based trust evaluation in online social networks: Methodologies and challenges. ACM Comput. Surv. **49**(1), 10:1–10:35 (May 2016). https://doi.org/10.1145/2906151, http://doi.acm.org/10.1145/2906151
[21] Kanza, Y., Kravi, E., Safra, E., Sagiv, Y.: Location-based distance measures for geosocial similarity. ACM Trans. Web **11**(3), 17:1–17:32 (Jul 2017). https://doi.org/10.1145/3054951, http://doi.acm.org/10.1145/3054951
[22] Kumari, A., Singh, A.K., Patil, N.: Travel recommendation system using geotagged photos. In: Proceedings of the 7th International Conference on Computer and Communication Technology. pp. 22–26. ICCCT-2017, ACM, New York, NY, USA (2017), http://doi.acm.org/10.1145/3154979.3154995
[23] Li, J., Wang, X., Hovy, E.: What a nasty day: Exploring mood-weather relationship from twitter. In: ACM International Conference on Conference on Information and Knowledge Management. pp. 1309–1318 (2014)
[24] Liu, P., Ding, Y., Fu, T.: Optimal throwboxes assignment for big data multicast in vdtns. Wireless Networks pp. 1–11 (2019)
[25] Liu, X., Suel, T.: What makes a group fail: Modeling social group behavior in event-based social networks. In: IEEE International Conference on Big Data. pp. 951–956 (2017)
[26] Liu, X., Tian, Y., Tian, Y., Lee, W.C., Mcpherson, J., Han, J.: Event-based social networks: linking the online and offline social worlds. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1032–1040 (2012)
[27] Lu, Y., Qiao, Z., Zhou, C., Hu, Y., Guo, L.: Location-aware friend recommendation in event-based social networks: A bayesian latent factor approach. In: ACM International on Conference on Information and Knowledge Management. pp. 1957–1960 (2016)
[28] Macedo, A.Q., Marinho, L.B., Santos, R.L.T.: Context-aware event recommendation in event-based social networks. In: ACM Conference on Recommender Systems. pp. 123–130 (2015)
[29] Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., Barman, S.: A statistical approach to predict flight delay using gradient boosted decision tree. In: International Conference on Computational Intelligence in Data Science. pp. 1–5 (2018)
[30] Ogundele, T.J., Chow, C.Y., Zhang, J.D.: Eventrec: Personalized event recommendations for smart event-based social networks. In: IEEE International Conference on Smart Computing. pp. 1–8 (2017)
[31] Qin, W.J.: Analysis on weather comfort for traveling in guangxi. Journal of Guangxi Meteorology (2003)
[32] She, J., Tong, Y., Chen, L.: Utility-aware social event-participant planning pp. 1629–1643 (2015)
[33] Shi, Y., Jiang, W.: Point-of-interest recommendations: capturing the geographical influence from local trajectories. In: 2017 IEEE International Symposium on Parallel and Distributed Processing with

Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC). pp. 1122–1129. IEEE (2017)

[34] Shuo Zhang, J., Lv, Q.: Understanding event organization at scale in event-based social networks. ACM Transactions on Intelligent Systems and Technology **10**, 1–23 (01 2019). https://doi.org/10.1145/3243227

[35] Tu, W., Cheung, D.W., Mamoulis, N., Yang, M., Lu, Z.: Activity recommendation with partners. ACM Trans. Web **12**(1), 4:1–4:29 (Sep 2017), http://doi.acm.org/10.1145/3121407

[36] Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P.S., Li, Z., Huang, Z.: Computing urban traffic congestions by incorporating sparse gps probe data and social media data. ACM Trans. Inf. Syst. **35**(4), 40:1–40:30 (Jul 2017). https://doi.org/10.1145/3057281

[37] Wang, Z., Zhang, Y., Li, Y., Wang, Q., Xia, F.: Exploiting social influence for context-aware event recommendation in event-based social networks. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications. pp. 1–9 (2017)

[38] Wu, X., Dong, Y., Shi, B., Swami, A., Chawla, N.V.: Who will attend this event together? event attendance prediction via deep lstm networks. In: Proceedings of the 2018 SIAM International Conference on Data Mining. pp. 180–188. SIAM (2018)

[39] Xu, B., Chin, A., Dan, C.: On how event size and interactivity affect social networks. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems. pp. 865–870 (2013)

[40] Xu, Y., Ren, J., Wang, G., Zhang, C., Yang, J., Zhang, Y.: A blockchain-based nonrepudiation network computing service scheme for industrial iot. IEEE Transactions on Industrial Informatics **15**(6), 3632–3641 (June 2019). https://doi.org/10.1109/TII.2019.2897133

[41] Xu, Y., Ren, J., Zhang, Y., Zhang, C., Shen, B., Zhang, Y.: Blockchain empowered arbitrable data auditing scheme for network storage as a service. IEEE Transactions on Services Computing pp. 1–1 (2019). https://doi.org/10.1109/TSC.2019.2953033

[42] Y, T., N, L.: Construal-level theory of psychological distance. Psychological Review **117**(2), 440–463 (2010)

[43] Ye, J., Chow, J.H., Chen, J., Zheng, Z.: Stochastic gradient boosted distributed decision trees. In: ACM Conference on Information and Knowledge Management. pp. 2061–2064 (2009)

[44] Yin, H., Cui, B., Chen, L., Hu, Z., Zhang, C.: Modeling location-based user rating profiles for personalized recommendation. ACM Trans. Knowl. Discov. Data **9**(3), 19:1–19:41 (Apr 2015). https://doi.org/10.1145/2663356

[45] Yin, H., Cui, B., Sun, Y., Hu, Z., Chen, L.: Lcars: A spatial item recommender system. ACM Trans. Inf. Syst. **32**(3), 11:1–11:37 (Jul 2014). https://doi.org/10.1145/2629461

[46] 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: ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 879–883 (2015)

[47] Zhang, C., Liu, C., Zhang, X., Almpanidis, G.: An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications **82**(C), 128–150 (2017)

[48] Zhang, S., Alanezi, K., Gartrell, M., Han, R., Lv, Q., Mishra, S.: Understanding group event scheduling via the outwithfriendz mobile application. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(4), 175 (2018)

[49] Zhang, S., Lv, Q.: Hybrid egu-based group event participation prediction in event-based social networks. Knowledge-Based Systems **143**, 19–29 (2018)

[50] Zhang, X., Zhao, J., Cao, G.: Who will attend? – predicting event attendance in event-based social network. In: IEEE International Conference on Mobile Data Management. pp. 74–83 (2015)

[51] Zhang, Y., Wu, H., Panangadan, A., Prasanna, V.K.: Integration of heterogeneous web services for event-based social networks. In: IEEE International Conference on Information Reuse and Integration. pp. 57–63 (2015)

[52] Zhou, X., Li, K., Xiao, G., Zhou, Y., Li, K.: Top k favorite probabilistic products queries. IEEE Transactions on Knowledge and Data Engineering **28**(10), 2808–2821 (2016)

[53] Zhou, X., Li, K., Yang, Z., Li, K.: Finding optimal skyline product combinations under price promotion. IEEE Transactions on Knowledge and Data Engineering **31**(1), 138–151 (2018)