2015 16th IEEE International Conference on Mobile Data Management

Who Will Attend? – Predicting Event Attendance in Event-Based Social Network

Xiaomei Zhang, Jing Zhao and Guohong Cao
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA, 16802
Email: {xqz5057, juz139, gcao}@cse.psu.edu

***Abstract*—Human mobility prediction has received considerable attention because it helps addressing many practical problems in mobile networks. Most existing techniques focus on regular mobility prediction by studying the periodic mobility pattern of users. However, they fail to detect users' irregular mobility patterns, like attending a sporadic event. We address this problem by proposing techniques to predict event attendance based on the following basic idea: if a user is interested in events related to a topic, he may also attend future events related to this topic. In our solution, to learn how users are likely to attend the future events, three sets of features are identified by analyzing users' past activities, including semantic, temporal, and spatial features. Then, the supervised learning models are trained to predict event attendance based on the extracted features. To evaluate the performance of the proposed techniques, we collect a dataset based on Meetup that contains semantic descriptions of all events organized over a period of two years. Evaluation results show that the supervised classifiers built by all features outperform those built by individual features, and semantic features are more effective than temporal features and spatial features for predicting event attendance.**

I. Introduction

Understanding and predicting human mobility and activities can help design effective protocols for data dissemination in Delay Tolerant Networks [1][2][3] and mobile social networks [4][5], and can assist resource management in wireless networks. There have been lots of research efforts on characterizing and predicting human mobility [6][7][8][9], and most of them are based on a common observation that human movements exhibit a high degree of recurrence [10][11]; for example, people visit regular places repeatedly during their daily activities.

Although existing techniques can be used to predict users' regular activities based on their social routines, they fail to predict users' activities in many cases. First, existing techniques cannot be applied to predict users' irregular movements. For example, when a sporadic event happens, such as an art festival, it is hard to infer if a user will attend this event based on the user's periodic routines. Second, existing research can only identify periodic visits to landmarks, but fail to capture periodic routines when location changes frequently. For example, a class held every weekday morning may take place at different buildings. The weekend party of a group of students may be held at different places each week.

This work was supported in part by Network Science CTA under grant W911NF-09-2-0053.

In this paper, we address these problems by focusing on irregular mobility prediction; i.e., predicting users' attendance at a future event at specific time and location based on their common interests. For instance, an event of "English group" is formed when people having difficulties in speaking English gather together to practice English, and an event of "football seminar" is formed when a famous football player gives a talk on campus. The *event attendance problem* can be formalized as a problem of predicting if a user will attend a future event by mining past activities. The basic idea is as follows: if a user is interested in events related to a topic, this user may also attend future events related to this topic. For example, if a student usually attends art-related events (e.g., art class on weekends), it is more likely the student will go to the art festival.

It is a challenge to predict event attendance due to two issues; one is the dearth of an event-based dataset, and the other is the difficulty of identifying the topic relevance between events. Thanks to the popularity of online social networks [12], such as Meetup (www.meetup.com), Plancast (www.plancast.com), and Eventbrite (www.eventbrite.com), where people in the neighborhood can create, organize and sign on social events online, we are able to collect event-based datasets. We have collected a dataset based on Meetup that contains $149,089$ users and $132,739$ events organized over a period of two years. Each event in the dataset has a semantic description. The topic relevance between events is characterized by the similarity between the semantic descriptions of events, i.e., semantic similarity between events. To calculate the semantic similarity between events, we design a semantic analysis approach based on categories. In this approach, the semantic similarity between events is calculated as the sum of their similarities at different categories.

To predict event attendance, we identify three sets of features including *semantic*, *temporal*, and *spatial* features. Semantic features characterize how frequently users attended similar events in the past, and the semantic similarity between events is used to identify similar events a user attended in the past. Temporal features measure users' temporal preference when attending events, and spatial features capture users' location preference when attending events. Based on the Meetup dataset, we first evaluate the effectiveness of each feature in predicting event attendance. Next, we train

978-1-4799-9972-9/15 $31.00 © 2015 IEEE
DOI 10.1109/MDM.2015.23

three supervised classifiers based on the defined features: logistic regression [13], J48 decision tree [14] and Naïve Bayes [15], and evaluate the performance of these classifiers. The evaluation results demonstrate that the classifiers built by all features significantly outperform those built by individual feature, which indicates that event attendance can be affected by multiple factors. We also observe that semantic features are more important than temporal features and spatial features for predicting event attendance.

The rest of the paper is organized as follows. Section II presents the preliminaries. In Section III, we present the approach for predicting event attendance. Section IV presents the evaluation results of the predicting approach. The last section concludes the paper.

II. Preliminaries

In this section, we first formulate the event attendance problem, and then introduce the Meetup dataset and our semantic analysis approach.

A. Problem Formulation

The event-attendance problem is formulated as follows:

Event-Attendance Problem: Given a user u and a future event e, the goal is to predict if u will attend e with the following information:

1) The information (description, time, location) of the past events that user u has attended.
2) The information (description, time, location) of event e.
3) The home location of user u

To solve this problem, we extract features (denoted as $F_{u,e}$) that characterize the relationship between user u and event e. A binary variable $a_{u,e}$ is used to represent if u will attend e. Then, the extracted features $F_{u,e}$ are used to predict the value of $a_{u,e}$.

$$F_{u,e} \Longrightarrow a_{u,e}$$

Predicting users' attendance at events requires a semantic representation for each event so that the semantic similarity between events can be computed. To semantically represent an event, a keyword list is extracted from the description of the event: $\{k_1, k_2, ..., k_M\}$, where M is the number of keywords. Each keyword k_i is associated with a weight w_i ($0 \leq w_i \leq 1$) which represents the importance of this keyword. The weight assignment depends on specific circumstances and will be discussed later. An event can be semantically represented by a set of keyword pairs:

$$\mathcal{K} = \{(k_1, w_1), (k_2, w_2), ..., (k_M, w_M)\}$$

B. The Dataset

The dataset is collected based on an online social network called Meetup, which provides people amongst neighborhoods an opportunity to create, organize and sign on "meetups" according to their common interests. One such "meetup" is an event, and therefore Meetup is referred to as an event-based social network [12]. An upcoming event is posted after it is planned by the hosting group. People reply "yes" if they will attend this upcoming event, and we simply consider these people as attendants of the event. The previous events are denoted as past events. From the event website, we can also find who attended the past events, the rating and the comments about this event. Since Meetup's creation in 2001, it has attracted more than 13 million users over the world and over $374,000$ events are hosted every month.

In our paper, we use the events in Meetup that were held in Pennsylvania (US) during 2011 and 2012. The resulting dataset contains $149,089$ users having $132,739$ events organized over two years. To semantically represent an event, keyword pairs are extracted for each event (as discussed in Section II-A). The keywords are extracted from the topics of the hosting group and the event description (the title of event). A weight of 0.8 is assigned to the keywords extracted from the topics of the hosting group, and a weight of 1 is assigned to the keywords extracted from the event title. The weight for the keywords extracted from the event title has higher weight because it contains unique information related to the event.

TABLE I
Categories

Arts/Culture	Career/Business	Cars/Motorcycles
Dancing	Education/Learning	Community/Environment
Fitness	Food/Drink	Fashion/Beauty
Games	Government/Politics	Lesbian/Gay/Bisexual/Transgender
Hobbies/Crafts	Health/Wellbeing	Language/Ethnics
Lifestyle	Movies/Film	Literature/Writing
Music	Spirituality	Outdoors/Adventure
Paranormal	Parents/Family	Pets/Animals
Photography	Religion/Beliefs	Fiction/Fantasy
Singles	Socializing	Sports/Recreation
Support	Technology	Women

C. Semantic Analysis

Since one event can be semantically represented by a set of keyword pairs, we compute the semantic similarity between two events as the similarity between two sets of keyword pairs. The popular methods to compute the similarity include counting the common keywords or computing the Jaccard similarity coefficient [16] between the two sets. These methods are implemented by matching the identical keywords in the two sets. However, they fail to identify the synonymies due to the lack of semantic analysis. For example, the synonymies "speech" and "talk" are considered to be different, even though they may have the same meaning.

To address this problem, we apply a semantic analysis approach to compute the semantic similarity between events based on taxonomy of categories. The similarity at each category is computed first. Then, the overall similarity is computed by summing the similarities in all categories. Here, the category information is pre-defined and we simply adopt the category information collected in Meetup that is used to characterize groups and events. There are 33 categories which are listed in Table I.

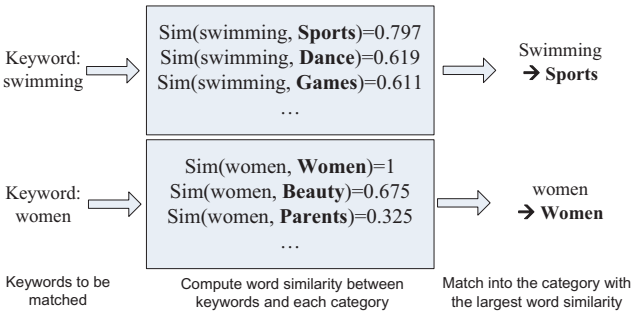


Fig. 1. Matching keywords "swimming" and "women" into categories.

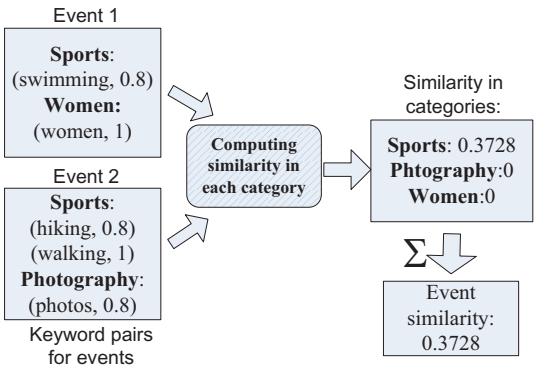


Fig. 2. Computing event similarity by summing similarities in all categories.

In the rest of the paper, for simplicity, the semantic similarity between two events is referred to as "*event similarity*", and the semantic similarity between two words is referred to as "*word similarity*".

1) Matching Keywords into Categories: To compute event similarity, we first match the keywords of an event into categories, i.e., one keyword should belong to one category.

To match a keyword to a category, the word similarity between the keyword and each category is computed, and this keyword belongs to the category that has the largest word similarity. Word similarity quantifies how two words are similar in semantics. For example, Pedersen *et. al* [17] present some popular word similarity measures based on WordNet [18][19], which is a large lexical database of English not only providing the word definitions, but also recording various semantic relations between words. In this paper, we adopt one of the most well-known measures–*Lin* [20] measure to compute the word similarity. *Lin* measure computes the similarity based on the word definitions. We also test other similarity measures, which only shows minor difference from the *Lin* measure.

Figure 1 shows an example of matching the keywords of an event into categories. This event is about a swimming class for women, and it has two keywords: "swimming" and "women". Each keyword is matched to the category that has the largest word similarity. For example, "swimming" has word similarity of 0.797 with "Sports", which is the largest among the word similarities with all categories. Therefore, "swimming" belongs to the category "Sports". Similarly, women has the largest word similarity with "Women", and thus belongs to category "Women".

2) Computing Event Similarity: After matching keywords into categories, the set of keyword pairs is split into multiple subsets, and each subset includes the keyword pairs in that category. The subset of keyword pairs in category i is denoted as $c_i = \{(k_{i,1}, w_{i,1}), (k_{i,2}, w_{i,2}), ...\}$, where $k_{i,u}$ is the u-th keyword and $w_{i,u}$ is the weight of $k_{i,u}$. To compute the similarity between two events e_1, e_2, we first calculate the event similarity in each category and then sum these similarities. Before calculating the event similarity in category i, we first find the two subsets of keywords pairs in category i for e_1 and e_2, denoted as c_i^1, c_i^2. Then the event similarity in category i is the similarity between c_i^1 and c_i^2, denoted as $SimC(c_i^1, c_i^2)$. It is calculated as the largest pairwise similarity between keyword pairs in c_i^1 and c_i^2, where the similarity between two keyword pairs is the word similarity between two keywords multiplied by their weights. The event similarity in category i is represented as follows:

$$\begin{aligned} &SimC(c_i^1, c_i^2)) \\ &\quad = \max_{u,v}(KeywordPairSim((k_{i,u}^1, w_{i,u}^1), (k_{i,v}^2, w_{i,v}^2))) \\ &\quad = \max_{u,v}(WordSim(k_{i,u}^1, k_{i,v}^2) * w_{i,u}^1 * w_{i,v}^2) \end{aligned} \tag{1}$$

where $KeywordPairSim(*,*)$ represents the similarity between two keyword pairs, and $WordSim(*,*)$ represents the word similarity between two words. $k_{i,u}^1$ represents the u-th keyword in c_i^1, $k_{i,v}^2$ represents the v-th keyword in c_i^2, and $w_{i,u}^1$, $w_{i,v}^2$ are the weights of them. Then the *semantic similarity* between two events is defined as the summation of similarities in all categories:

$$Sim(e_1, e_2) = \sum_{i=1}^{n} SimC(c_i^1, c_i^2) \tag{2}$$

Figure 2 shows an example on how to compute the event similarity. One event has keyword pairs (swimming, 0.8) and (women, 1) and the other has (hiking, 0.8), (walking, 1) and (photos, 0.8). In the "Women" category and the "Photography" category, the keywords only appear in one event (either Event 1 or Event 2), and the event similarity is 0 in these categories. In "Sports" category, both events have keywords, and we need to compute the event similarity using Equation (1). Here, the keyword pairs (swimming, 0.8) and (hiking, 0.8) have similarity $0.367 * 0.8 * 0.8 = 0.235$, where 0.367 is the word similarity between keywords "swimming" and "hiking", calculated based on the *Lin* measure. The two 0.8s are the weights of the keywords. Similarly, (swimming, 0.8) and (walking, 1) have similarity $0.466 * 0.8 * 1 = 0.3728$. The keyword pairs (swimming, 0.8) and (walking, 1) have the largest similarity 0.3728, and thus the event similarity in category "Sports" is 0.3728. Since the event similarities in

other categories are 0, the semantic similarity between these two events is 0.3728.

III. Predicting Event Attendance

To find out if a user will attend a future event, three sets of features are defined, including *semantic*, *temporal*, and *spatial* features, which are defined by studying the history behavior of the user. More specifically, semantic feature characterizes the user's interest in future events by studying the user's attendance at past events. Temporal feature exploits temporal information on user activities and measures the user's temporal preference when attending events. Spatial feature is extracted to measure the user's location preference when attending events. In the rest of the paper, user and future event are denoted as u and e respectively. Next, we introduce these three features in detail, and then present a prediction approach based on these extracted features.

A. Semantic Feature

With semantic features, we aim to capture the user's interest in future event by studying the user's attendance at past events. For example, if a student frequently attended football training in the past, it is most likely that he will appear at a football game on campus or attend a seminar from a famous football player visiting his campus. To quantify user u's interest in future event e, two semantic features *total attendance* and *percentage of attendance* are defined.

1) Total attendance: This feature measures the number of similar-topic events (*or referred to as* similar events) the user has attended in the past. Two events are considered to be similar if their semantic similarity is larger than a *similarity threshold* α. Given event e, we count the number of similar events that u has attended during a time period T in the past. Here, both α and T are constants and can be set flexibly. These two parameters will also be used to define the following features. Formally, total attendance is represented as:

$$n_{u,e} = |\{e_i \in E_u : t_e - T < t_{e_i} < t_e \wedge Sim(e, e_i) > \alpha\}| \tag{3}$$

where E_u indicates the set of past events that u has attended, and t_e is the time of event e.

2) Percentage of attendance: Percentage of attendance is another semantic feature that can be used to demonstrate the user's preference for similar topic. It is computed as the percentage of events that the user has attended among all similar events. This feature is effective when the events of a specific topic are only held infrequently but the user attends most of them. In this case, total attendance may not identify the user's interest in this topic due to the low frequency. Therefore, we use percentage of attendance to quantify the user's interest in this kind of events. Specifically, with event e, we find all similar events held in a past time period T and compute the percentage of events that u has attended. Formally, the total number of similar events in a time period T is:

$$n_e = |\{e_i \in E : t_e - T < t_{e_i} < t_e \wedge Sim(e, e_i) > \alpha\}| \tag{4}$$

where E indicates the set of all past events. In Equation (3), we compute the number of events that user u has attended, i.e., $n_{u,e}$. The *percentage of attendance* for user u and event e is computed as:

$$p_{u,e} = \frac{n_{u,e}}{n_e} \tag{5}$$

B. Temporal Feature

Users usually have temporal preference when attending events. For example, a user may prefer to go to gym after work in the afternoon, and another user may prefer to watch football game on Saturday. To characterize the temporal preference of a user u on a future event e, three temporal features are defined: *recent attendance*, *weekly attendance*, and *daily attendance*.

Before defining the temporal features, we first define a metric to quantify the *temporal relation* between two events: $T(e_1, e_2)$. For example, if e_1 and e_2 are both held on Saturday, they have a large temporal relation on the weekly pattern. The temporal relation is defined according to different temporal features and will be discussed when defining specific features.

The temporal features in terms of u and e are calculated by analyzing similar events that u has attended in the past time period T. For each of these events e_i, we find the temporal relation $T(e_i, e)$ between e_i and e. Then, the temporal feature is calculated by averaging the temporal relations of e with all these past events.

$$t_{u,e} = \frac{\sum_{e_i \in E_u^T, Sim(e_i,e)>\alpha} T(e_i, e) * Sim(e_i, e)}{\sum_{e_i \in E_u^T, Sim(e_i,e)>\alpha} Sim(e_i, e)} \tag{6}$$

Here, $e_i \in E_u^T$ indicates the past events that user u has attended in the past time period T. $T(e_i, e)$ is the temporal relation between the past event e_i and the future event e, which is set in accordance with various temporal features. When computing the average of temporal relation, we use $Sim(e_i, e)$, the event similarity between e_i and e, as the weight of the past event e_i. Then, the past event with a larger event similarity with e has more contributions in computing the temporal feature.

1) Recent Attendance: Human behaviors can be best characterized by the most recent activities. For example, if a user frequently takes physical training in the gym recently, it is likely that he will attend another physical training event. Based on this, we first study the feature *recent attendance*, where the temporal relation is simply related to the time interval between e_i and e. We use the reciprocal of the time interval as the temporal relation between the events:

$$T^r(e_i, e) = \frac{1}{t_e - t_{e_i}}$$

Then, *recent attendance* is represented as:

$$t^r_{u,e} = \frac{\sum_{e_i \in E_u^T, Sim(e_i,e)>\alpha} T^r(e_i, e) * Sim(e_i, e)}{\sum_{e_i \in E_u^T, Sim(e_i,e)>\alpha} Sim(e_i, e)} \tag{7}$$

2) Weekly Attendance: People usually have similar behaviors on the same day of a week, as implied by the example that the user prefers to attend football events on Saturday. Here, the temporal relation $T^w(e_i, e)$ indicates if the two events happen on the same day of the week.

$$T^w(e_i, e) = \begin{cases} 0 & dow(t_{e_i}) \neq dow(t_e) \\ 1 & dow(t_{e_i}) = dow(t_e) \end{cases}$$

where $dow(t) \in [1, 2, ..., 7]$ returns a value corresponding to a specific day in a week (Monday, Tuesday,..., Sunday) of time t. Formally, *weekly attendance* is represented as:

$$t^w_{u,e} = \frac{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} T^w(e_i, e) * Sim(e_i, e)}{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} Sim(e_i, e)} \quad (8)$$

3) Daily Attendance: This feature characterizes another human routine, i.e., having similar behaviors at similar time of a day. For example, a person normally attends an after-work exercise class at 5 pm every day. However, his schedule may not be exactly the same (i.e., it may be a little bit early or late when attending the same activity), and hence there may be some deviation. A Gaussian function is applied to measure how two time points (in hours) at a day are similar:

$$s(t_1, t_2) = e^{-\frac{(t_1 - t_2)^2}{2}}$$

Based on this formula, the temporal relation for *daily attendance* is calculated as

$$T^d(e_i, e) = s(t_{e_i}, t_e) = e^{-\frac{(t_{e_i} - t_e)^2}{2}}$$

The *daily attendance* is computed as:

$$t^d_{u,e} = \frac{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} T^d(e_i, e) * Sim(e_i, e)}{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} Sim(e_i, e)} \quad (9)$$

C. Spatial Feature

Spatial features can be extracted to capture the spatial preference of users when attending events. We characterize two spatial features: *home distance* and *location preference*.

1) Home Distance: Home distance measures the distance between the home location of user u and the location of event e:

$$h_{u,e} = \text{Dist}(l^h_u, l_e)$$

where $\text{Dist}(*, *)$ is the distance between two locations, l^h_u is the home location of user u and l_e is the location of event e.

2) Location Preference: This feature characterizes the location preference when a user attends events. For example, a student may only attend local football games, and if the game is held at another city far away, he may not attend. We use similar formula as Equation 6 to characterize how event e is spatially related to past events. Here, the spatial relation is simply the distance between the past event e_i and the event e, i.e., $\text{Dist}(l_{e_i}, l_e)$. The *location preference* can be represented as follows:

$$l_{u,e} = \frac{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} Dist(l_{e_i}, l_e) * Sim(e_i, e)}{\sum_{e_i \in E^T_u, Sim(e_i,e)>\alpha} Sim(e_i, e)} \quad (10)$$

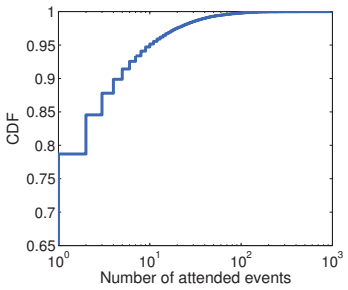


Fig. 3. The Cumulative Distribution Function (CDF) of the number of events attended per user from 2011 to 2012.

D. Predicting Approach

We adopt a supervised learning approach to learn how the extracted features affect users' decisions on event attendance. This process is also referred to as supervised binary classification, considering that 'attend or not' is a binary classification. Specifically, the Meetup data will be split into training data and testing data. The supervised classifier learns how features affect event attendance from the training data, and the learned classifier is evaluated on the testing data. There are many supervised classifiers in the literature [21], and we use three classifiers in this paper, including logistic regression [13], J48 decision tree [14] and Naïve Bayes [15]. One challenge of the supervised learning is how to set the parameters including the *past time period* T and the *similarity threshold* α. In the next section, we will experimentally test the effect of the parameters on the performances and determine the appropriate values for the parameters.

IV. Performance Evaluations

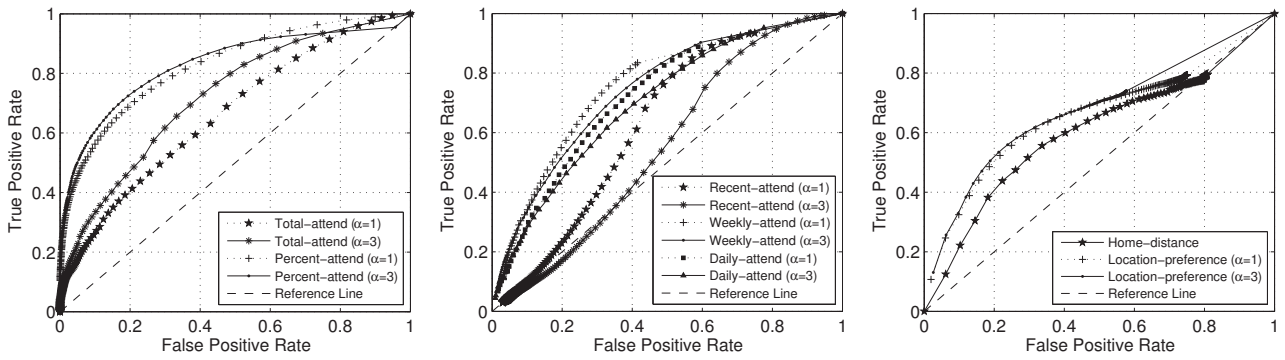
Based on the dataset collected from Meetup, we evaluate the proposed solutions in this section, including the effectiveness of the extracted features and the influence of the parameters on performances. This section starts with the discussion on data selection, evaluation strategy and experiment setting, and then presents the evaluation results.

A. Data Selection

To evaluate the effectiveness of the extracted features and train the supervised classifiers, we use data collected from Meetup. For each user-event pair $\{u, e\}$ in the dataset, $F_{u,e}$ represents the extracted features and $a_{u,e}$ represents if u attends e in Meetup. We include all the information associated with one user-event pair to a data instance:

$$[F_{u,e}, a_{u,e}]$$

Data instances are selected from all users and all events held during 2011-2012 in Pennsylvania. Amongst all these users, only small portion of them actively participate events as shown in Figure 3, which plots the cumulative distribution function (CDF) of the number of events attended per user during the two years. The CDF shows that only about 20% of users have attended events, and only about 1% users actively attend



(a) Semantic Features (b) Temporal Features (c) Spatial Features

Fig. 4. ROC curves for the decision stumps built by individual features: (a) decision stumps built by semantic features, *total attendance* and *percentage of attendance*; (b) decision stumps built by temporal features, *recent attendance*, *weekly attendance* and *daily attendance*; (c) decision stumps built by spatial features, *home distance* and *location preference*. A classifier is better if its ROC curve is closer to the upper-left corner. The reference line in the diagonal represents the ROC curve for the random classifier.

more than one event every month on average. These 1% users are referred to as *active users*. Since inactive users do not attend events frequently, it is not necessary to include them for evaluation. Thus, we only consider the active users when selecting data instances. A data instance is a positive instance if $a_{u,e} = 1$, and it is a negative instance if $a_{u,e} = 0$. To train the binary classifier unbiasedly, there should be equal numbers of positive instances and negative instances. In our dataset, there are $60,221$ positive instances including all the active users and their attended events. To have the same number of negative instances, we randomly choose $60,221$ negative instances from the active users and the events they have not attended.

B. Evaluation Strategy

The *Receiver-Operating-Characteristics (ROC)* [22] curve is commonly used to illustrate the performance of a binary classifier system. It plots the fraction of true positives out of the total actual positives (true positive rate) vs. the fraction of false positives out of the total actual negatives (false positive rate), at various threshold settings. The ROC curve is a monotonic non-decreasing function of true positive rate over the false positive rate. A random classifier only results in a curve $y = x$ in the diagonal, and a classifier is better if it is closer to the upper-left corner. Based on this fact, the *area under the ROC curve (AUC)* is an important metric to evaluate the overall performance of a binary classifier [23]. It is claimed in [24] that AUC is a statistically consistent and discriminating metric. Besides AUC, another metric used in this paper is the prediction *accuracy* which is used to evaluate the performance of various supervised classifiers.

C. Experiment Setting

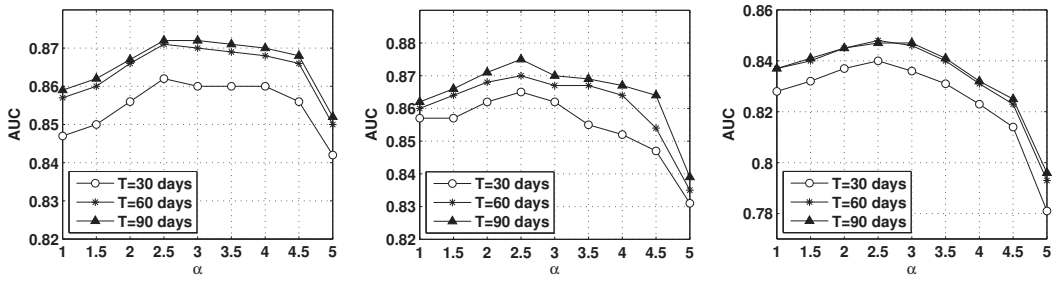
All the experiments are performed using the WEKA software [21]. The supervised classifiers used in our experiments include logistic regression, J48 decision tree and Naïve Bayes. It is easy to inspect the inner structures of these simple classifiers, so that we can learn the role of each feature in prediction. Other advanced classifiers (like random forest, SVM) are also tested using WEKA, but they only show minor improvement over these simple classifiers. For the dataset, the first half is used for training and second half is used for performance evaluation. The WordNet database (which is used to compute word similarity) used in our experiment is the most updated version for Windows (WordNet 2.1).

TABLE II
AUC FOR INDIVIDUAL FEATURES USING DECISION STUMPS

Features	**AUC** ($\alpha = 1$)	**AUC** ($\alpha = 3$)
Random	0.5	0.5
Total Attendance	0.6	0.65
Percentage of Attendance	0.74	0.768
Recent Attendance	0.57	0.632
Weekly Attendance	0.679	0.709
Daily Attendance	0.627	0.67
Home Distance	0.607	0.607
Location Preference	0.653	0.66

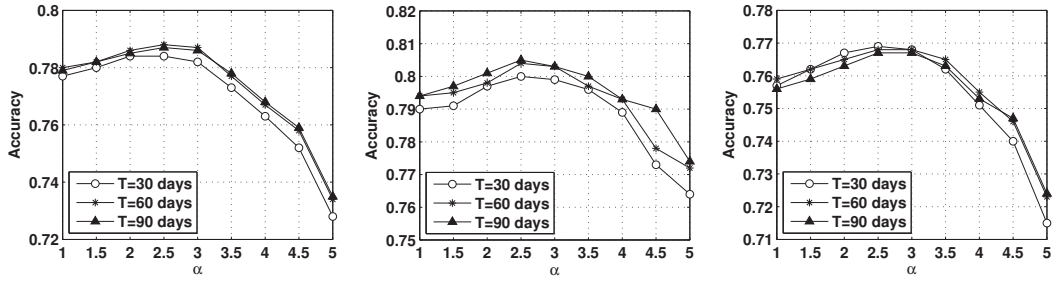
D. Individual Features

We evaluate the prediction accuracy power of each individual feature using decision stump, which is a one-level decision tree [25]. Based on one individual feature, a predicting score is computed for each user-event pair, and a higher score usually means a higher attending probability. For the semantic feature or temporal feature, a higher feature value implies a higher attending probability, so the score is simply set to be the feature value. For the spatial feature, a smaller feature value implies a higher attending probability, so the score is set to be the negative of the feature value. By setting a decision threshold, the instance with score higher than the threshold is predicted as positive (will attend), otherwise it is predicted as negative (not attend). As the decision threshold varies, we get different true positives and false positives, which will generate



(a) Logistic Regression (b) J48 Decision Tree (c) Naïve Bayes

Fig. 5. The impact of the parameters T and α on the prediction performance in terms of AUC



(a) Logistic Regression (b) J48 Decision Tree (c) Naïve Bayes

Fig. 6. The impact of the parameters T and α on the prediction performance in terms of prediction accuracy

the ROC curve. The ROC curves for the decision stumps built by individual features are shown in Figure 4. The reference line in the figure represents the ROC curve generated by the random classifier. To be more general, we test the features computed using different parameters: *past time period* T and *similarity threshold* α. Since the impact of T on the ROC curve is not easy to observe, we only present the ROC curves under different α, with $T = 90$ days.

For semantic features, as shown in Figure 4 (a), the decision stumps built by *total attendance* and *percentage of attendance* have ROC curves higher than the reference line. This result illustrates the effectiveness of both semantic features in predicting event attendance. In addition, we also observe a superiority of the *percentage of attendance* over the *total attendance*. This indicates that the percentage of similar events a user has attended in the past time period is more helpful on predicting whether the user will attend the future event.

Figure 4 (b) shows the ROC curves for the decision stumps built by three temporal features. As can be seen, the decision stumps built by temporal features have ROC curves higher than the reference line, and decision stump built by *weekly attendance* has the highest ROC curve. Therefore, exploiting temporal features about the attendance at similar events, especially the *weekly attendance*, provides significant assistance in predicting event attendance.

For the spatial features, as shown in Figure 4 (c), the ROC curves for the decision stumps built by *home distance* and *location preference* are higher than the reference line, and the decision stump built by *location preference* is the highest. Therefore, location is also an important factor that affects event attendance.

By comparing the predicting power of the features under different similarity threshold α ($\alpha = 1$ and $\alpha = 3$), we find that $\alpha = 3$ has an overall superiority over $\alpha = 1$, which implies that the similarity threshold may have impact on the predicting power.

Finally, we compare the predicting power of all individual features by computing the AUCs corresponding to the ROC curves in Figure 4. As shown in Table II, the results are consistent with what we have observed from the ROC curves, where *percentage of attendance* outperforms other semantic features, *weekly attendance* outperforms other temporal features, and *location preference* outperforms other spatial features. Amongst all these features, *percentage of attendance* has the most predicting power, with AUC ($\alpha = 1$) of 0.74 and AUC ($\alpha = 3$) of 0.768. These results suggest that semantic features are the most important features for predicting event attendance.

E. Supervised Learning Evaluation

In this part, the predicting power of all features are combined in supervised learning models. We first evaluate the effects of the two parameters T and α on prediction. After setting the appropriate parameters, we build supervised classifiers using all features and compare it with those using only one feature.

1) Determining Parameters T and α: The two parameters, *past time period* T and *similarity threshold* α are used in

Section III to define features. The two parameters affect all features except *home distance*, and therefore impact the performance of the supervised classifiers. We train three supervised classifiers, logistic regression, decision tree and Naïve Bayes, using the features defined with different T and α. The performance of these classifiers are shown in Figure 5 and Figure 6. The figures show the prediction performance in terms of AUC and accuracy. As can be seen from Figure 5, as α varies from 1 to 5, AUC reaches the maximum value when $\alpha = 2.5$ for all classifiers. In addition, AUC has a large improvement when T increases from 30 days to 60 days, but the improvement is much smaller when T increases from 60 days to 90 days. We also increase T to be longer than 90 days, but could not find any noticeable change for AUC. These results suggest that the history data in a time period of 90 days is enough to infer user's event attendance. For prediction accuracy, as shown in Figure 6, the highest prediction accuracy is achieved when $\alpha = 2.5$. With the increase of T, the prediction accuracy does not have too much improvement. Thus, α is set to be 2.5 and T is set to be 90 days in the rest of the paper.

2) Comparing With Learning models Built with One Feature: By including all features in the supervised learning models, we should achieve better predicting power than individual features. In this subsection, we compare the performance of the supervised classifiers built with all features and those built with individual features.

The evaluation results of three classifiers are presented in Table III. As can be seen, the classifiers built with all features can increase AUC by 0.02 - 0.25, and increase the prediction accuracy by $0.03-0.20$ compared to classifiers with individual features. These results demonstrate that the predicting power can be increased significantly by combining all features.

The performance of different classifier is different. Logistic regression and Naïve Bayes have worse performance than J48 decision tree, and Naïve Bayes is the worst. The low performance of logistic regression and Naïve Bayes is due to the following two reasons. First, the feature *home distance* does not increase the predicting performance in the two classifiers, since the classifiers built with *home distance* have AUCs even smaller than the random baseline of 0.5. Another reason is related to the intrinsic mechanisms in the classifiers. Logistic regression is based on a linear model, which is not sufficient to characterize the effect of each feature. Naïve Bayes assumes independence of multiple features, but the features may mutually correlate to some extent.

To identify the effect of each feature in the classification, we study the inner structures of the two classifiers: logistic regression and J48 decision tree. Naïve Bayes is not considered here because it has the worst performance. The coefficients of the features (absolute value) in the logistic regression classifier are shown in Table IV, and the top three levels of the J48 decision tree are shown in Figure 7. All results manifest the contribution of the feature *percentage of attendance*, as it has the largest coefficient in logistic regression and dominates in the top three levels of the J48 decision tree. Thus, when

TABLE IV
THE COEFFICIENTS OF FEATURES IN LOGISTIC REGRESSION.

Features	Coefficients
Percentage of attendance	6.0789
Weekly attendance	1.95
Daily attendance	1.05
Total attendance	0.0118
Recent attendance	0.0118
Home distance	0
Location preference	0

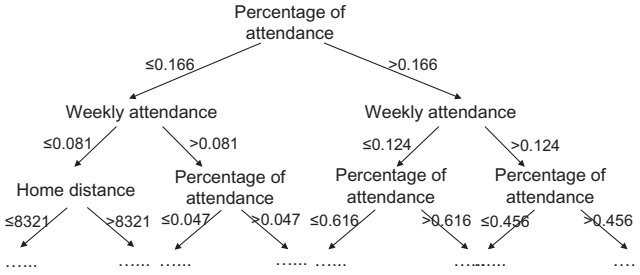


Fig. 7. The top three levels of the J48 decision tree

predicting whether a user will attend an event, the most important task is to find all similar events in the history and the percentage of attendance. In addition to *percentage of attendance*, the temporal features *weekly attendance* and *daily attendance* also play important roles in predicting, since they also belong to the top three coefficients in logistic regression. This further demonstrates that users' behaviors conform with the temporal routines to some extent. The spatial features, *location preference* and *home distance*, do not play important roles in the J48 decision tree and even have 0 coefficients in the logistic regression, thus the spatial features are not as important as the semantic features and temporal features in predicting event attendance.

3) Studying the Importance of Semantic Information: Since an important part of our prediction approach is the consideration of semantic information, we next run experiments to study the importance of the semantic information. In the first experiment, we compare between classifiers built with three sets of features: semantic features, temporal features, and spatial features, to see if the semantic features have better predicting power. The results are shown in Figure 8. As can be seen, the classifier built with semantic features has the best performance in terms of AUC and prediction accuracy. Therefore, semantic features are more important than temporal features and spatial features for predicting event attendance.

In the second experiment, we study the impact of semantic information on individual features. As presented in Section III, besides semantic features, temporal features and spatial features also use semantic information. For example, the temporal feature is calculated by averaging the temporal relations of the future event with all past events. When computing the

TABLE III
COMPARING CLASSIFIERS BUILT WITH ALL FEATURES AND THOSE BUILT WITH INDIVIDUAL FEATURE

Features	Logistic Regression		J48		Naïve Bayes	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Random Baseline	0.5	0.5	0.5	0.5	0.5	0.5
Total Attendance	0.718	0.639	0.702	0.656	0.633	0.602
Percentage of Attendance	0.852	0.76	0.818	0.763	0.834	0.736
Recent Attendance	0.389	0.503	0.658	0.628	0.618	0.504
Weekly Attendance	0.752	0.68	0.746	0.703	0.752	0.658
Daily Attendance	0.715	0.656	0.7	0.66	0.716	0.654
Home Distance	0.395	0.503	0.699	0.637	0.397	0.503
Location Preference	0.671	0.561	0.685	0.658	0.67	0.56
All Features	**0.872**	**0.786**	**0.875**	**0.805**	**0.847**	**0.767**

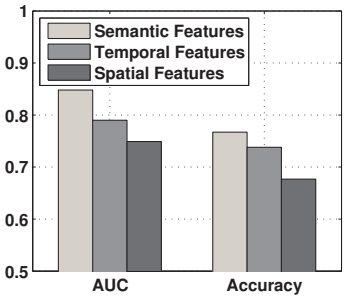


Fig. 8. Comparing the performance of individual features using the J48 decision tree.

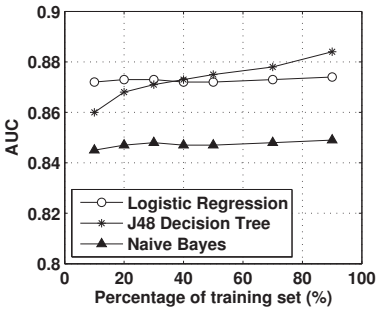
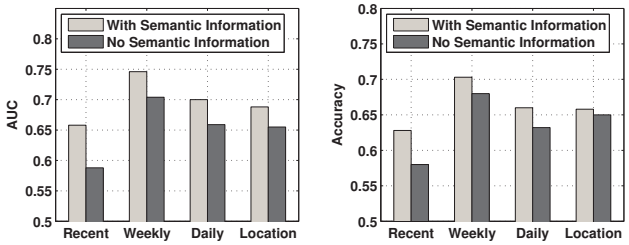


Fig. 10. The effect of the length of the training set on the performance of different classifiers.



(a) AUC (b) Accuracy

Fig. 9. Semantic information can be used to increase the predicting power of various features such as *recent attendance*, *weekly attendance*, *daily attendance* and *location preference* (evaluated using the J48 decision tree).

average of the temporal relations, the event similarity between the past event and the future event is used as the weight of the past event. The spatial feature *location preference* is also calculated using the same method. We run an experiment to see how the semantic information affects the predicting power of these features. Since the spatial feature *home distance* does not use semantic information, it is not tested here. We compare the predicting power of the features calculated using semantic information and the features calculated without semantic information. The results are shown in Figure 9. As can be seen, the features with semantic information can improve AUC by about 0.05, and improve the prediction accuracy by about 0.03. These results further demonstrate the importance of considering semantic information.

4) The Length of the Training Set: In the previous experiments, the first half of the dataset is used for training and the second half is used for evaluation. We further run an experiment to see if the length of the training set (as a percentage of the whole dataset) affects the performance of the supervised classifier. The evaluation results of all three supervised classifiers are shown in Figure 10. As can be seen, with the length of the training set increases, logistic regression is stable, J48 decision tree shows some improvement in term of AUC. Naïve Bayes has a slight increase in AUC, but it has the worst AUC as the training set varies. For the other two classifiers, J48 decision tree is preferred when the percentage of training set is larger than 40%, while logistic regression is preferred if the length of training set is smaller than 40%.

V. RELATED WORK

Characterizing and predicting human mobility have attracted lots of research efforts. Some earlier prediction techniques are based on characterizing user's history spatial trajectories. Gao *et al.* [26] characterized user mobility behaviors at a fine-grained level based on the Hidden Markov Model formulation of user mobility. Yuan *et al.* [27] employed semi-markov process model to describe user mobility as transitions between landmarks. These methods aim to capture the short-term mobility trajectory, but fail to capture the long-term human

mobility. Later prediction approaches are more focused on the analysis of the spatial-temporal patterns of user movements, i.e., the periodic mobility patterns, so as to characterize the long-term human mobility. For example, researchers [9][10][11][28] have studied human periodic mobility pattern by analyzing the daily and weekly movement of users. Sadilek *et al.* [29] further predict human mobility in a longer term with a scale of months or even years. Even though these techniques can capture human mobility most of time, they fail to detect users' irregular movements, which is the focus of this paper.

In addition to location based services, the newly emerging online event-based social networks (EBSN) provide a new venue to analyze human mobility and social behaviors. Liu *et al.* [12] investigated the network properties of EBSN such as the degree distribution and community structures. Researchers in [30] [31] have analyzed how offline human activities at events affect the social networking behaviors. This paper continues the research on EBSN and focuses on predicting users' irregular mobility by analyzing their past activities.

Even though our event attendance prediction utilizes some similar techniques with the recommendation system [32][33] (like finding the interest of a user by analyzing the user's past behaviors), we further consider the temporal and spatial factors that may influence users' decision on event attendance. Moreover, learning the behavior of users on event attendance can help to characterize users' mobility in a more comprehensive way, which is helpful for the potential design of the smartphone applications and networking strategies.

VI. Conclusions

In this paper, we proposed techniques to predict event attendance by mining users' past activities. We identified three sets of features including semantic, temporal, and spatial features. Semantic features characterize how frequently users attended similar events in the past, and the semantic similarity between events is used to identify similar events a user attended in the past. Three supervised learning models are trained to learn how the features affect event attendance. To evaluate the performance of the proposed techniques, we collect a dataset based on Meetup that contains semantic descriptions of all events organized over a period of two years. Evaluation results show that the supervised classifiers built by all features outperform those built by individual features, and semantic features are more effective than temporal features and spatial features for predicting event attendance.

References

[1] K. Fall, "A delay-tolerant network architecture for challenged internets," in *ACM SIGCOMM*, 2003.

[2] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, 2011.

[3] W. Gao and G. Cao, "User-centric data dissemination in disruption tolerant networks," in *IEEE INFOCOM*, 2011.

[4] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network," in *IEEE INFOCOM*, 2009.

[5] X. Zhang and G. Cao, "Efficient data forwarding in mobile social networks with diverse connectivity characteristics," in *IEEE ICDCS*, 2014.

[6] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *ACM SIGKDD*, 2009.

[7] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, p. 3, 2011.

[8] T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," in *ACM Ubicomp*, 2012.

[9] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in *IEEE ICDM*, 2012.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[11] E. Nazerfard, P. Rashidi, and D. J. Cook, "Using association rule mining to discover temporal relations of daily activities," in *Toward Useful Services for Elderly and People with Disabilities*. Springer, 2011, pp. 49–56.

[12] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in *ACM SIGKDD*, 2012.

[13] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

[14] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[15] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI workshop on empirical methods in artificial intelligence*, 2001.

[16] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, 1971.

[17] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet:: Similarity: measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL*. Association for Computational Linguistics, 2004.

[18] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] C. Fellbaum, "Wordnet," in *Theory and applications of ontology: computer applications*. Springer, 2010, pp. 231–243.

[20] D. Lin, "An information-theoretic definition of similarity," in *ICML*, 1998.

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[22] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[23] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[24] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[25] W. Iba and P. Langley, "Induction of one-level decision trees," in *ICML*, 1992.

[26] W. Gao and G. Cao, "Fine-grained mobility characterization: steady and transient state behaviors," in *ACM MobiHoc*, 2010.

[27] Q. Yuan, I. Cardei, and J. Wu, "Predict and relay: an efficient routing in disruption-tolerant networks," in *ACM MobiHoc*, 2009.

[28] X. Zhang and G. Cao, "Transient community detection and its application to data forwarding in delay tolerant networks," in *IEEE ICNP*, 2013.

[29] A. Sadilek and J. Krumm, "Far out: Predicting long-term human mobility." in *AAAI*, 2012.

[30] S. Counts and J. Geraci, "Incorporating physical co-presence at events into digital social networking," in *ACM CHI extended abstracts on Human Factors in Computing Systems*, 2005.

[31] B. Xu, A. Chin, and D. Cosley, "On how event size and interactivity affect social networks," in *ACM CHI extended abstracts on Human Factors in Computing Systems*, 2013.

[32] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *ACM conference on Digital libraries*, 2000.

[33] P. Melville and V. Sindhwani, "Recommender systems," in *Encyclopedia of machine learning*. Springer, 2010, pp. 829–838.