Front. Comput. Sci., 2013, 7(2)
DOI 10.1007/s11704-013-3902-8

RESEARCH ARTICLE

Co-occurrence prediction in a large location-based social network

Rong-Hua LI (✉)[1], Jianquan LIU[2,3], Jeffrey Xu YU[1], Hanxiong CHEN[2], Hiroyuki KITAGAWA[2]

1 The Chinese University of Hong Kong, Hong Kong, China
2 University of Tsukuba, Ibaraki 305-8577, Japan
3 Cloud System Research Laboratories, NEC Corporation, Tokyo 108-8001, Japan

Abstract Location-based social network (LBSN) is at the forefront of emerging trends in social network services (SNS) since the users in LBSN are allowed to "check-in" the places (locations) when they visit them. The accurate geographical and temporal information of these check-in actions are provided by the end-user GPS-enabled mobile devices, and recorded by the LBSN system. In this paper, we analyze and mine a big LBSN data, Gowalla, collected by us. First, we investigate the relationship between the spatio-temporal co-occurrences and social ties, and the results show that the co-occurrences are strongly correlative with the social ties. Second, we present a study of predicting two users whether or not they will meet (co-occur) at a place in a given future time, by exploring their check-in habits. In particular, we first introduce two new concepts, *bag-of-location* and *bag-of-time-lag*, to characterize user's check-in habits. Based on such bag representations, we define a similarity metric called habits similarity to measure the similarity between two users' check-in habits. Then we propose a machine learning formula for predicting co-occurrence based on the social ties and habits similarities. Finally, we conduct extensive experiments on our dataset, and the results demonstrate the effectiveness of the proposed method.

Keywords location-based social networks, Gowalla, co-occurrence

RReceived January 5, 2013; accepted February 22, 2013

E-mail: rhli@se.cuhk.edu.hk

1 Introduction

Currently, big data analysis and processing has become one of challenging tasks in both research and industry community. Big data broadly denotes a large, complex, and dynamic dataset [1]. Various application domains, such as biology, astronomy, social science, computer science, have collected more and more data due to the rapid development of the data collection technology.

Among these application domains, the big online social network data generated by users has attracted much attention in the research community. In recent years, social network services (SNS) such as Facebook and Twitter have become increasingly popular. These traditional SNS platforms allow the users to share thoughts, activities, photos, and other information with friends. Unlike the traditional SNS, recently, location-based social networks (LBSN) rapidly attract a considerable number of people to share their locations through GPS-equipped smart-phones. In LBSN, the users are allowed to check-in locations (spots) when they visit there, and to share their check-in's with friends via their mobile devices. Meanwhile, their accurate geographical locations are provided by GPS-enabled functionality, and the temporal information is recorded by the LBSN systems. This emerging form of applications open the door for researchers to study spatio-temporal features of user's online activities.

However, so far there have not been much research on spatio-temporal features of user's online check-in activities and on how these features affect the interactions between users in LBSN. One interesting feature is the spatio-temporal

co-occurrences [2–4]. This feature describes an event that two users simultaneously appear at the same location. In LBSN, the co-occurrence between two users means that they visited the same location in a short temporal range. In this paper, the co-occurrence is always short for *spatio-temporal co-occurrence* if not specifically stated. Another interesting feature is the spatio-temporal characteristics of the user's check-in habits. This feature can be used to identify the users by their special habits, to group the users by their habit similarities, and to infer social tie between two users, and so forth. Based on the analysis of these spatio-temporal features of users in LBSN, in this paper, we are interested in the problem of predicting the co-occurrence between two users.

Some previous studies have focused on the relationships between the co-occurrences and the social ties. As observed in certain off-line and online social systems [2, 5], a small number of co-occurrences can result in a high likelihood of a social tie. Specifically, in an off-line system, this phenomenon has been used to study city life [5]. In an online context, the authors [2] study this issue on the photo-sharing site Flickr (www.flickr.com/), and they report that a small number of contemporaneous photo-taken activities can lead to a high likelihood of a social link. To show this phenomenon whether or not exists in LBSN, we first crawl a very large LBSN dataset from a former notable LBSN platform, Gowalla. The crawled Gowalla dataset includes 317 815 users, 1 778 487 social ties, 1 777 090 spots, and 16 929 121 check-ins. Based on it, we examine the correlations between co-occurrences and social ties in Gowalla. However, as our observation in Gowalla, this phenomenon is not very significant, but the co-occurrences still exhibit a strong correlation with the social ties. Based on this analysis, we present a machine learning method for co-occurrence prediction by using the social-tie features and the users' check-in habits similarities. In particular, we first propose a *bag-of-location* and a *bag-of-time-lag* representation to capture the spatial and temporal features of user's check-in habits respectively. Based on the bag representations, we define a similarity metric to measure the similarities of users' check-in habits. Then, with the features of social ties and check-in similarities, we use a logistic regression classifier to predict the co-occurrence between two given users. Finally, we conduct extensive experiments on our dataset, and the results show the predictive accuracy is greater than 90%.

The rest of this paper is organized as follows: Section 2 describes the collected Gowalla dataset. Section 3 shows the relationship between the co-occurrences and social ties. The bag representations and the machine learning formula, as well as the experimental results are presented in Section 4. We briefly survey the related work in Section 5, and we conclude this work in Section 6.

2 The big Gowalla data

In this section, we introduce our Gowalla dataset, and present a certain basic analysis on it as well. Gowalla was a notable location-based social networking platform, which was acquired by Facebook recently. Unlike traditional online social network platforms, the location-based social networking services, such as Gowalla and Foursquare (http://foursquare.com/), present a new way for users to share locations with their friends.

In recent years, with the development of mobile device, GPS functionality is widely embedded, which can provide accurate geographical location of user. This evolution is rapidly enriching the structure of social network. The Gowalla integrates this kind of location information into the traditional social network by exploiting the new concept "check-in". In the Gowalla, all the users are mobile device holders, and all the spots are the sites (places) that the users ever visited. A check-in with a timestamp means a user visited a spot at a certain time. By integrating the spots and check-ins, the location-based social network (LBSN) presents a novel graph structure as shown in Fig. 1. Note that in the traditional online social networks such as Twitter and Facebook, the network only contains one type of nodes (users). For instance, in Fig. 1, the traditional social networks usually only contain the users (left nodes), keeping their social ties. In LBSN, the users still have their friends, and can freely check-in some spots when they visited them, or they can check-in the same spot when they visited it at different time (i.e., s_2 in Fig. 1).

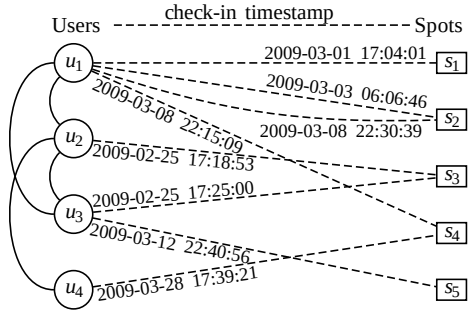

Fig. 1 Example: graph structure of Gowalla

For deep analysis, we implemented a crawler to gather a representative dataset from Gowalla using its open API. We launched the crawler for six months from October 01, 2010.

All the crawled data is summarized in Table 1. In particular, we have collected 317 815 online users which includes 1 778 487 social ties. For the locations, we have crawled 1 777 090 spots, which contains 16 929 121 check-ins in total. Figure 2 visualizes the spots (locations) from the crawled data. Interestingly, from Fig. 2, we can see that the shape of world map is clear, which implies that the collected dataset covers most locations that the users visited in Gowalla. The result further suggests that the collected dataset can represent the whole Gowalla data. Otherwise, the shape could be a portion of the world map if the dataset contains bias data.

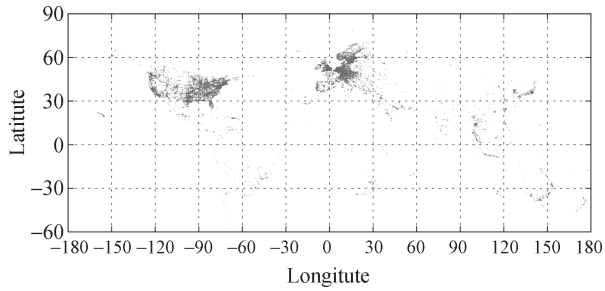

Fig. 2 World check-in map of crawled data

Table 1 Summary of crawled Gowalla dataset

Items	Crawled size
Users	317 815
Social ties	1 778 487
Spots	1 777 090
Check-ins	16 929 121

In our collected dataset, a check-in record is a tuple ⟨*userid, latitude, longitude, timestamp*⟩. Here *latitude* and *longitude* denotes the latitude and longitude of the location where the user visited, and *timestamp* denotes the time stamp of the check-in activity. Each user in the crawled dataset has a check-in list and a check-in list contains a location sequence and a time-stamp sequence. As shown in Fig. 3, the length of the check-in lists (the number of check-ins) of the users exhibit a power law distribution. In other words, the length of the location sequence and time-stamp sequence of users are extremely diverse. This result indicates that the sequence matching based methods, which require the length of the sequences are equivalent, for measuring similarities such as [6] cannot work well in our dataset. To overcome this issue, we resort to the bag-of-words model [7], which is a well-known model to characterize the word features of the documents in information retrieval area. One of the important properties of the bag-of-words model is that it can yield a feature vector for each document with same length. Thus, it can easily define the similarity measure between two feature representations. We shall describe our feature representation method in the following sections. Below, we first present a analysis of the relationship between the co-occurrences and the social ties, and then we propose a machine learning formula for predicting the co-occurrence between two users based on their social ties and check-in activities.

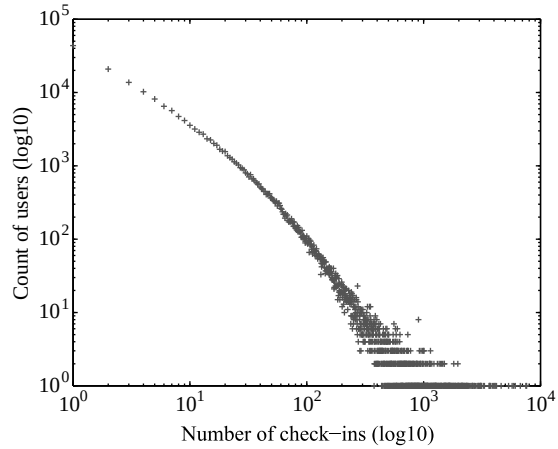

Fig. 3 The log-log plot of user's check-ins distribution

3 Co-occurrences and social ties

Previous study on an online photo-share website, Flickr, shows that a small number of co-occurrences can yield a high empirical likelihood of the social links [2]. Here, we are interested in studying this issue in Gowalla. We choose to study the active users whose number of check-in's greater than 50, resulting in 8% of all users in our dataset (25 516 active users). The reason of using active users for analysis is that most users in Gowalla are inactive, and the median user has only 23 check-in's, thus it has lower probability that two inactive users contain co-occurrences.

We follow the techniques in [2] to study this issue. First, we divide the surface of the earth into grid-like cells, and the side lengths of the cell span z degrees of latitude and longitude. In our analysis, z takes the values of 1.0, 0.1, and 0.01 respectively. Second, for any two users u_i and u_j, we count the number of co-occurrences between them at various time ranges, 1 d, 5 d, 10 d, 15 d, 20 d, 25 d, and 30 d, respectively. Third, we calculate the fraction of social ties $Fra = N_c/N_f$ and the probability of social ties $Pro = N_c/N_u$ w.r.t. a specified number of co-occurrences. Here N_c denotes the number of social ties that involves at least a specified number of co-occurrences, N_f denotes the number of social ties, and N_u denotes the number of user-pairs that involves at least a specified number of co-occurrences. Our results of Fra and Pro as a function of the number of co-occurrences are shown in

Fig. 4(a)–(c) and Fig. 4(d)–(f) respectively.

From Fig. 4(a)–(c), we find that the fraction of social ties decreases as the increasing number of co-occurrences and the decreasing temporal range in all 1.0, 0.1, and 0.01 cells. There are 50%, 40%, and 22.9% of all social ties having one co-occurrence in a 1.0, 0.1, and 0.01 cell at one day, respectively. These results show that the friends exhibit co-occurrences in Gowalla, indicating that the social ties correlate to the co-occurrences. On the other hand, as shown in Fig. 4(d)–(f), the probability of social ties typically increases as the number of co-occurrences increases and temporal range decreases. Surprisingly, the results show a lower probability of social links w.r.t. the number of co-occurrences than the observations in [2]. The largest probability is only 0.22, which appears when the temporal range is one day, the spatial threshold is 0.01, and the number of co-occurrences is 14. In the Flickr dataset, however, the largest probability is greater than 0.95 reported in [2]. One potential explanation could be that two strangers in Gowalla have a significant opportunity to check-in the same cell at the same temporal range. However, the probability of the event that two strangers take photos in the same place at the same temporal range, and then upload their photos into Flickr is relatively low.

4 Predicting co-occurrence

In this section, we are interested in studying the problem of predicting two users in Gowalla whether or not they will contain a co-occurrence from their social ties and historical check-in activities. In the following, we first describe a method to represent the spatial and temporal features of the check-in activities. Then, we present a machine learning formula to predict the co-occurrence phenomenon between any two given users based on such spatial and temporal features.

4.1 Feature representation and similarity metric

As discussed in the previous sections, each user's check-in histories can be represented by a location sequence and a time-stamp sequence. The time-stamp sequence is a discrete time-point sequence, and it could be missing some important temporal features of user's check-in activities as described in a simple example below. To better characterize the temporal features, we turn to use the time-lag sequence, where each element of the sequence is a time lag between two consecutive check-in activities. The time-lag sequence can be deemed as a "first-order" representation of the time-stamp sequence, and it can capture some important temporal characteristics of user's check-in activities as illustrated in the following example.

Consider two users u_1 and u_2, and let the time-stamp sequence of u_1 and u_2 be $(1, 3, 5)$ and $(6, 8, 10)$ respectively. Obviously, the time-stamp sequence of u_1 and u_2 are quite different. However, the time-lag sequences are equivalent, which equal to $(2, 2)$. In fact, these two users exhibit similar temporal characteristics of their check-in activities as both of them have a check-in activity at every other day. Hence, in this case, the time-lag sequence is more powerful than the

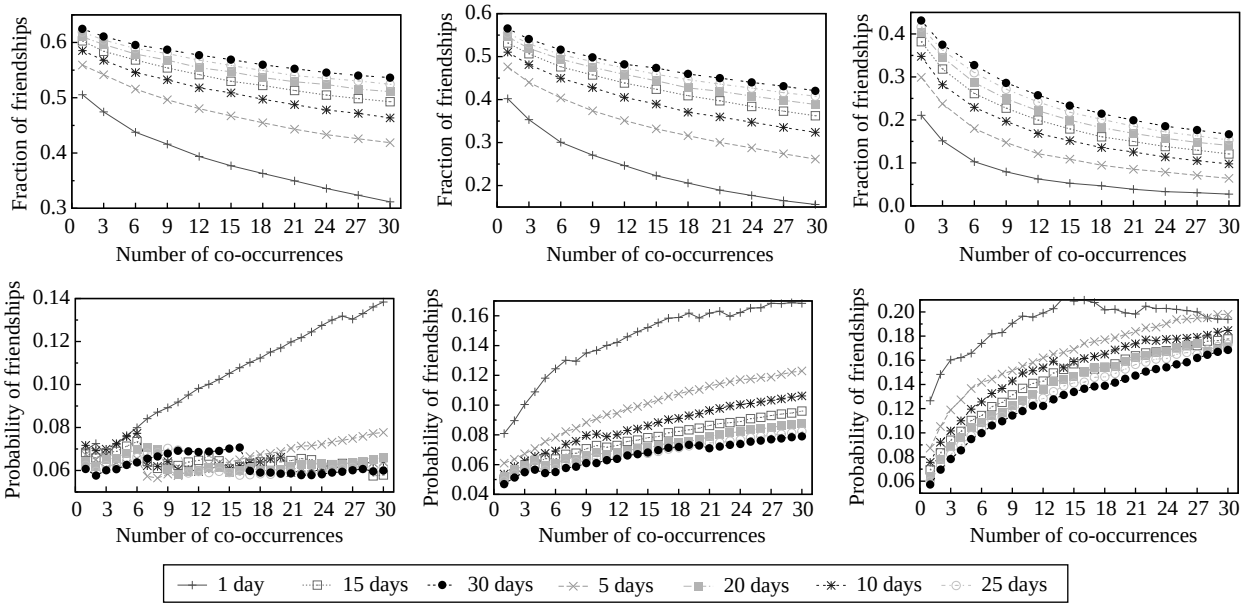

Fig. 4 (a)–(c) Fraction of social ties (friendships) with spatial threshold (a) $z = 1.0$; (b) $z = 0.1$; (c) $z = 0.01$, and (d)–(f) probability of social ties with spatial threshold (d) $z = 1.0$; (e) $z = 0.1$; (f) $z = 0.01$, as a function of number of co-occurrences.

time-stamp sequence for characterizing the temporal features.

To simplify our analysis, inspired by the bag-of-words model [7], we similarly represent the spatial and temporal features of user's check-in activities. We refer to our bag representations as the bag-of-location and the bag-of-time-lag, and the detail descriptions are given as follows.

- **Bag-of-location representation** for each user u_i, we represent the spatial features of u_i by a bag-of-location. For convenience, firstly, we divide the surface of the earth into grid-like cells, and the side lengths of the cell span z degrees of latitude and longitude. Secondly, we map each location that have been visited by users into a cell. And we denote each location by a cell label c_j, where $j = 1, 2, \ldots, m$, and m is the number of cells that have been visited by users. Then, for a given user u_i, the location sequence is given by $C_i = (c_{j_1}, c_{j_2}, \ldots, c_{j_{n_i}})$, where c_{j_k}, for $k = 1, 2, \ldots, n_i$, denotes a cell visited by user u_i, and n_i denotes the number of check-in's of user u_i. Finally, the bag-of-location representation C_i of user u_i is the histogram of the location sequence C_i.
- **Bag-of-time-lag representation** for a given user in our collected dataset, we design a bag-of-time-lag representation to characterize the temporal features of user's check-in records, where the time lag denotes the time difference between two consecutive check-in activities. Specifically, for each user u_i, we extract the time-stamp sequence $\mathcal{T}_i = t_1, t_2, \ldots, t_{n_i}$ from his check-in's. For convenience, we use *day* as the time unit in this work. And then, we sort the time-stamp sequence resulting in $\tilde{\mathcal{T}}_i = \tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{n_i}$ and calculate the time lag as $\Delta t_j = \tilde{t}_{j+1} - \tilde{t}_j$, for $1 \leqslant j \leqslant n_i - 1$. Then, the time-lag sequence of u_i is given by a vector $\Delta\mathcal{T}_i = (\Delta t_1, \Delta t_2, \cdots, \Delta t_{n_i-1})$. Finally, the bag-of-time-lag representation T_i of user u_i is the histogram of the time-lag sequence $\Delta\mathcal{T}_i$.
- **Check-in habits similarity** based on the bag-of-word representation, we are able to easily measure the similarity between two users from their check-in activities. For convenience, we refer to such similarity as the check-in habits similarity, because it reflects the similarity of users' check-in habits. The habits similarity between user u_1 and user u_2 is defined by

$$S_{ij} = \lambda \cos(C_i, C_j) + (1 - \lambda) \cos(T_i, T_j), \qquad (1)$$

where C_i denotes bag-of-location representation of user u_i, T_i the bag-of-time-lag representation of user u_i, $\cos(a, b)$ the cosine distance between the vectors a and b, and $\lambda \in [0, 1]$ s a parameter used to tradeoff the spatial and temporal features. The reason of using the cosine distance is twofold. First, the cosine metric is very easy to compute. Second, it can be applied to the kernel-based classifiers. In addition, it is worth mentioning that if $\lambda = 0$, the habits similarity only captures temporal features, while if $\lambda = 1$, the habits similarity only captures spatial feature. If $0 < \lambda < 1$, then the habits similarity clearly captures both spatial and temporal features.

4.2 Problem definition and methodology

In this subsection, we first formulate the co-occurrence prediction problem. Then, we propose a machine learning method for solving this problem.

4.2.1 Problem definition

Given a social graph G with adjacent matrix A, two users u_i and u_j, and their check-in list up to time t, the goal of the co-occurrence prediction problem is to predict whether or not the contemporaneous event will happen between two users u_i and u_j in the temporal interval $(t, t + \Delta t)$.

4.2.2 Learning methodology

We make use of a logistic regression classifier to predict whether or not two users include a co-occurrence in temporal range Δt, because it is very simple and easy to implement. Moreover, in the experiments, we find that such logistic regression based method can achieve very promising results. We begin with defining two features for the logistic regression classifier. The first is the social-ties feature, which is denoted by a variable A_{ij}. In particular, for two users u_i and u_j, if they are friends, then $A_{ij} = 1$, otherwise $A_{ij} = 0$. The reason why we choose the social-tie as a feature for prediction is that our previous empirical observations have shown the strong correlation between the social ties and the co-occurrences (see Fig. 4). The second feature we chosen is the similarity between two users u_i and u_j, and here we use the similarity measure S_{ij} defined in Eq. (1). This is based on a mild assumption as described as follows. If two users whose check-in habits are similar, then they have high likelihood to include a co-occurrence. After having features, we formulate the logistic regression classifier in the following.

$$P(B_{ij}|A_{ij}, S_{ij}) = \frac{1}{1 + e^{-(a_0 + a_1 A_{ij} + a_2 S_{ij})}}, \qquad (2)$$

where a_0, a_1, a_2 are the coefficients that we estimate on the training data, and B_{ij} is a binary variable, which is $B_{ij} = 1$

if two users u_i and u_j have at least one co-occurrence in temporal range Δt, $B_{ij} = 0$ otherwise, and S_{ij} denotes the habits similarity between user u_i and user u_j.

4.3 Results

We consider the active users whose number of check-ins are greater than 50 to learning the model parameters, as the feature representations of these users are more meaningful than those of inactive users. We divide our dataset into two dataset by check-in's time stamp, namely dataset I and dataset II. Specifically, we generate dataset I by extracting all check-in records from the active users in our collected dataset whose time stamp are less than or equal to t. And then we generate dataset II by extracting the check-in records from our collected dataset whose time stamp belongs to $(t, t + \Delta t)$. Then we generate the training sets as follows. Firstly, from dataset I, we randomly generate 100 000 user-pairs and extract the feature vectors in terms of the bag-of-word model. Then, we compute the habit similarities using Eq. (1). Here, we set the spatial threshold z to 0.1, and similar results can be observed for other z values. For the tradeoff parameter λ (in Eq. (1)), we set it to 0, 0.5, and 1, denoting temporal feature, spatio-temporal feature, and spatial feature respectively. Secondly, we label each user-pair to 1 if they have at least one co-occurrence, 0 otherwise. We also use the same method to generate the test sets with size 100 000 from dataset II.

We conduct the experiments on four various temporal ranges (Δt): 1 d, 10 d, 20 d, and 30 d. In all of our experiments, we estimate logistic regression coefficients over 10-fold cross validation. And we use the predictive accuracy as the evaluation metric. The results are shown in Fig. 5.

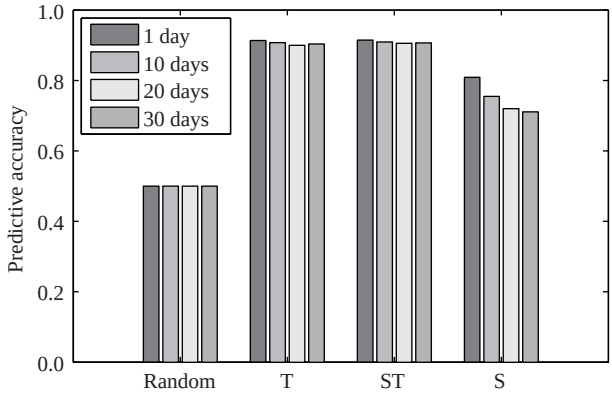

Fig. 5 Accuracy of predicting co-occurrence of different classifiers in various temporal intervals (Δt): 1 d, 10 d, 20 d and 30 d

In Fig. 5, Random denotes the random predictor. T, ST, and S denote the logistic regression classifiers in which the habits similarity S_{ij} is based on the temporal feature ($\lambda = 0$), spatio-temporal feature ($\lambda = 0.5$), and spatial feature ($\lambda = 1$) respectively. From Fig. 5, we can observe that classifier ST slightly outperforms the classifier T and significantly outperforms the classifier S over all the temporal intervals. This result indicate that the spatio-temporal features are more effective to capture the habits similarity between two users than other features, and the temporal features are better than the spatial features to characterize habits similarity. It is worth noting that the predictive accuracies of the classifiers ST and T are very promising, which are greater than 0.9 in all the temporal ranges. In contrast, the classifier S exhibits relatively poor performance whose predictive accuracy is about 0.75 over all the temporal ranges. Although the classifier S perform poorly, it still substantially outperforms the random predictor. These results suggest that the proposed spatio-temporal features indeed capture the important features of users' habits similarities, thus resulting in a high prediction accuracy.

4.3.1 Effect of the parameter λ

Here we study how the parameter λ affects the performance of the classifiers. We set the temporal range to 1 day, and similar results can be observed for other temporal ranges. The results of predictive accuracies over different λ are shown in Fig. 6. From Fig. 6, we can see that the predictive accuracy is robust w.r.t. the parameter λ when $\lambda \leqslant 0.7$. If $\lambda > 0.8$, we can see that the predictive accuracy decreases as increasing λ. The reason is because if λ increases, then the weight of the spatial features increases, thus resulting in that the habits similarities are dominated by the spatial feature. Since temporal features are more effective than the spatial features to capture the habits similarities, the large weight of the spatial feature reduces the predictive accuracy.

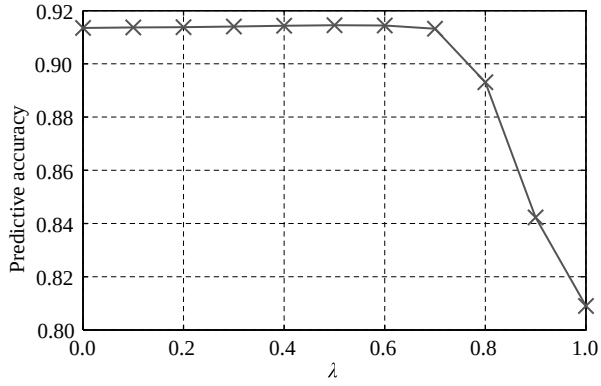

Fig. 6 The effect of parameter λ

4.3.2 Comparison between social ties and habits similarities

To better understand the relationships of the learned model to the features (social ties and habits similarities), we compare the ROC curves of three logistic regression (LR) classifiers with various features. The three classifiers include LR with social ties (Eq. (2) with $a_2 = 0$), LR with habits similarities (Eq. (2) with $a_1 = 0$), and LR with social ties and habits similarities (Eq. (2)). As shown in the previous experiment, the spatio-temporal feature is the best one to characterize the habits similarities, thereby we choose the spatio-temporal feature and set $\lambda = 0.5$ to measure the habits similarities. For the temporal range, we set it to 1 day, and similar results can be obtained in other temporal ranges. Figure 7 depicts the ROC curves of these three classifiers. From Fig. 7, we can see that the classifier with combinational features (both social ties and habits similarities) outperforms the classifier with a signal feature for co-occurrence prediction. The performance of LR with habits similarities is significantly better than the performance of LR with social ties. The results imply that the habits similarities are the crucial features in the proposed learning model for predicting co-occurrence.

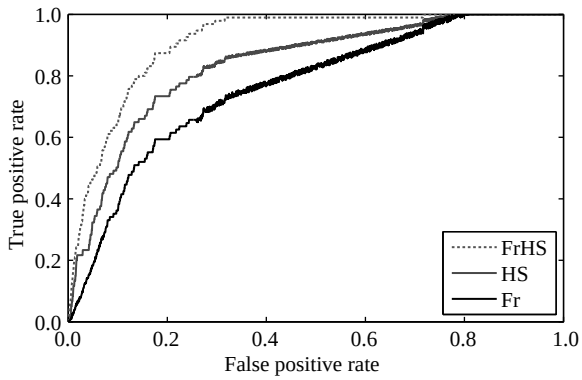

Fig. 7 ROC curve of logistic regression classifier with various features for predicting co-occurrence. Fr: social ties (friendships), HS: habits similarities, FrHS: social ties (friendships) and habits similarities

5 Related work

First, our work is related to the problem of analysis and mining of location-based social network (LBSN). In [8], the authors propose a location recommendation algorithm for LBSN by using social and geographical properties of users and locations. From the analysis point of view, Li et al. [9] analyze a LBSN over the users profiles, activities, mobility characteristics and social graphs. And they compare several LBSNs to study user's location-sharing behavior [10]. Unlike their work, recently, Scellato et al. [11] propose a graph-analysis-based method in LBSN to study how geographic distance affects the network structure. Subsequently, Scellato et al. in [12] study the spatial properties of the users in LBSN, and in [13] they propose to use the spatial properties to predict social ties. Cho et al. [14] investigate the relationship between the social ties and human mobility from the LBSN data, and they find that social ties can explain about 10% to 30% human mobility. More recently, Brown et al. in [14] study the relationship between the spatial properties of users and the community structures. Their results indicate that community structure could arise from both social spatial factors. Most of existing work mainly focus on the spatial features of users in LBSN. In this paper, we study both spatial and temporal features of user's check-in activities.

Second, our work is also related to link prediction in social network [15]. Most link prediction algorithms are based on graph-based similarity measures [15–18]. However, the graph-based similarity measures ignore node's content and only consider the structure information of the network, thus cannot be applied to the case with the link structure of the network totally unknown. This problem is the so-called cold start link prediction problem [19]. In [19], the authors proposed a bootstrap probabilistic graph framework and use node's group information as an initial measure to solve this problem. Using content information of nodes for link prediction is well studied in the literature [19–21]. However, to the best of our knowledge, relatively few studies have focused on applying node's spatio-temporal features to predict social links. In [3,4], the authors use spatio-temporal co-occurrence events to construct a social network, but their work mainly focuses on mining these events. In [22], the authors present a set of spatio-temporal features, namely co-location, to predict social links. But their features are only based on simple statistics of user's location history, thus they cannot fully characterize the complicated spatio-temporal features of user's check-in activities. The most related work is [2], in which the authors observed that a small number of spatio-temporal co-occurrences will lead to high likelihood of a social link in an online photo share site Flickr. However, unlike their observations, our work is based on an LBSN, and our results show a very low probability of a social tie w.r.t. the co-occurrences. In addition, their work aims to model the observed phenomenon, while in our work we mainly focus on how to represent and extract the spatio-temporal features of user's check-in activities as well as use these features for co-occurrence prediction task.

Finally, our work is also related to the spatio-temporal data

mining. The spatio-temporal data mining is well studied in the literature [23–27]. In these work, the authors focus on mining the spatio-temporal patterns from the spatio-temporal database. These patterns involve periodic pattern of the move object [23, 27], frequent spatio-temporal sequential pattern [24], complex sequential pattern [25], and co-occurrence pattern [26]. In addition, there is some other work towards to mining topic related spatio-temporal patterns. For example, in [28], the authors propose a probabilistic model to discover the spatiotemporal theme pattern on Weblogs. In [29, 30], the authors propose a topic model to mining the geographic routines of human from mobile phone data. However, all these work focuses on mining the spatio-temporal patterns. Instead, we exploit spatio-temporal features for iimproving predictions in LBSN.

6 Conclusion

In this paper, we present a study of exploring spatio-temporal features of user's activities to predict co-occurrence in a location-based social network, Gowalla. We first propose bag-of-location and bag-of-time-lag representation to characterize user's check-in activities. Based on these features, we define a similarity measure to represent user's check-in habit similarities. Then, we make use of a logistic regression classifier with social-ties feature and habit similarities to predict co-occurrence between any user-pairs precisely. There are several future directions that deserve further investigation. First, it may be interesting to explore other machine learning approaches (such as SVM and other kernel machines) that might yield better performance for predicting co-occurrence. Second, our feature representation methods can not only be applied to mine location-based social network, but it can also be used in other domains, such as spatio-temporal data mining and activity recognition. Finally, our current algorithm works on the sub-dataset which only includes active users. It is also interesting to develop an efficient learning algorithms that can handle the entire dataset. A promising direction is to use the hashing-based learning algorithms to handle such big data, such as the algorithm is developed in [31, 32].

Acknowledgements The work was supported by grant of the Research Grants Council of the Hong Kong SAR, China (418512).

References

1. Hey T, Tansley S, Tolle K M. The fourth paradigm: data-intensive scientific discovery. Microsoft Research, 2009
2. Crandall D J, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences, 2010, 107(52): 22436–22441
3. Lauw H W, Lim E P, Pang H, Tan T T. Social network discovery by mining spatio-temporal events. Computational & Mathematical Organization Theory, 2005, 11(2): 97–118
4. Lauw H W, Lim E P, Pang H, Tan T T. Stevent: spatio-temporal event model for social network discovery. ACM Transactions on Information Systems (TOIS), 2010, 28(3): 15:1–15:32
5. Milgram S. The experience of living in cities. Science, 1970, 167(3924): 1461–1468
6. Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma W Y. Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2008, 34–44
7. Christopher D, Manning P R, Sch ü tze H. Introduction to Information Retrieval. Cambridge University Press, 2008
8. Ye M, Yin P, Lee W C. Location recommendation for location-based social networks. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2010, 458–461
9. Li N, Chen G. Analysis of a location-based social network. In: Proccedings of the 2009 International Conference on Computational Science and Engineering. 2009, 263–270
10. Li N, Chen G. Sharing location in online social networks. IEEE Network, 2010, 24(5): 20–25
11. Scellato S, Mascolo C, Musolesi M, Latora V. Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd Conference on Online Social Networks. 2010, 8–17
12. Scellato S, Noulas A, Lambiotte R, Mascolo C. Socio-spatial properties of online location-based social networks. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. 2011, 1–8
13. Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011, 1046–1054
14. Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011, 1082–1090
15. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019–1031
16. Ito T, Shimbo M, Kudo T, Matsumoto Y. Application of kernels to link analysis. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005, 586–592
17. Kunegis J, Lommatzsch A. Learning spectral graph transformations for link prediction. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009, 561–568
18. Li R H, Yu J X, Liu J. Link prediction: the power of maximal entropy random walk. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011, 1147–1156
19. Leroy V, Cambazoglu B B, Bonchi F. Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010, 393–402

20. Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining. 2011, 635–644
21. Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010, 243–252
22. Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N. Bridging the gap between physical location and online social networks. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing. 2010, 119–128
23. Mamoulis N, Cao H, Kollios G, Hadjieleftheriou M, Tao Y, Cheung D W. Mining, indexing, and querying historical spatiotemporal data. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004, 236–245
24. Cao H, Mamoulis N, Cheung D W. Mining frequent spatio-temporal sequential patterns. In: Proccedings of the 5th IEEE International Conference on Data Mining. 2005, 8–16
25. Verhein F. Mining complex spatio-temporal sequence patterns. In: Proceedings of the 2009 SIAM International Conference on Data Mining. 2009, 605–617
26. Celik M, Shekhar S, Rogers J P, Shine J A. Mixed-drove spatiotemporal co-occurrence pattern mining. IEEE Transactions on, Knowledge and Data Engineering, 2008, 20(10): 1322–1335
27. Li Z, Ding B, Han J, Kays R, Nye P. Mining periodic behaviors for moving objects. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010, 1099–1108
28. Mei Q, Liu C, Su H, Zhai C. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of the 15th International Conference on World Wide Web. 2006, 533–542
29. Farrahi K, Gatica-Perez D. Probabilistic mining of socio-geographic routines from mobile phone data. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(4): 746–755
30. Farrahi K, Gatica-Perez D. Discovering routines from large-scale human locations using probabilistic topic models. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1): 3:1–3:27
31. Li P, König A C. Theory and applications of *b*-bit minwise hashing. Communications of the ACM, 2011, 54(8): 101–109
32. Li P, Shrivastava A, Moore J L, König A C. Hashing algorithms for large-scale learning. In: Proceedings of the 25th Annual Conference on Neural Information Processing Systems. 2011

Rong-Hua LI is pursuing his PhD in Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, China. His research interests include social network analysis and mining, complex network theory, uncertain graphs mining, Monte-Carlo algorithms, and machine learning.

Jianquan LIU received the BE from Shantou University, China, ME and PhD from the University of Tsukuba, Japan, in 2005, 2009, and 2012, respectively. He was a Development Engineer in Tencent Inc. from 2005 to 2006. He is currently a researcher at the Cloud System Research Laboratories of NEC Corporation, working on the topics of large-scale data processing and cloud computing. His research interests include high-dimensional similarity search, social network analysis, web data mining and information retrieval, cloud storage and computing, and multimedia databases. He is a member of ACM and the Database Society of Japan (DBSJ).

Jeffrey Xu YU received the BE, ME, and PhD in Computer Science from the University of Tsukuba, Japan, in 1985, 1987, and 1990, respectively. He held teaching positions in the Institute of Information Sciences and Electronics, University of Tsukuba, Japan, and the Department of Computer Science, The Australian National University. Currently, he is a professor in the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong. His current main research interest includes graph database, graph mining, keyword search in relational databases, and social network analysis. He is a senior member of IEEE, a member of IEEE Computer Society, and a member of ACM.

Hanxiong CHEN received the BS from Zhongshan University, Guangdong, China, in 1985, the MS and the PhD in Computer Science, from the University of Tsukuba, Japan, in 1990 and 1993, respectively. He is currently an assistant professor at Faculty of Engineering, Information and Systems, University of Tsukuba. His research interests include data engineering, knowledge discovery, data mining and information retrieval. He is a member of ACM, IEEE-CS and IPSJ.

Hiroyuki KITAGAWA received the BS in Physics and the MS and PhD in Computer Science, all from the University of Tokyo, in 1978, 1980, and 1987, respectively. He is currently a full professor at Faculty of Engineering, Information and Systems, University of Tsukuba. His research interests include data integration, data mining, information retrieval, stream processing, data-intensive computing, XML, and scientific databases. He is an editorial board member of IEEE TKDE and WWWJ, a Fellow of IPSJ and IEICE, Vice Chairperson of the Database Society of Japan, and a member of ACM, IEEE Computer Society, and JSSST.