Inferring Social Ties from Multi-view Spatiotemporal Co-occurrence

Caixu Xu(✉) and Ruirui Bai

Soochow University, Suzhou, China
csxucaixu@gmail.com

Abstract. Recently, social ties inferring in spatiotemporal data has attracted widespread attentions. Previous studies, which focused on either co-occurrence or context, do not fully exploit the information of spatiotemporal data. In order to better use the spatiotemporal information, in this paper we introduce two novel co-occurrence feature, namely, topic co-occurrence feature and context co-occurrence feature. The former feature is extracted by the topic model on carefully constructed bag-of-words. The latter feature is extracted by natural language processing tools on carefully constructed context sequence, which considers context, co-occurrence and mobility periodicity simultaneously. These two novel co-occurrence feature are both based on time and space perspectives. Then we infer social ties from these multi-view co-occurrence feature (including baseline co-occurrence, topic and context co-occurrence). The experiments demonstrate that the two novel co-occurrence feature contribute to the social tie inferring significantly.

Keywords: Social ties · Spatiotemporal co-occurrence · Topic co-occurrence
Context co-occurrence

1 Introduction

In recent years, spatiotemporal data has attracted interest from more and more people. Spatiotemporal data usually include time and space information, where time dimension information is represented by check-in time and space dimension represented by check-in location. Specifically, because of the fashionable usage of mobile devices, users can easily share spatiotemporal information with their friends. This phenomenon inspires companies to use spatiotemporal data to mine user behavior patterns and offer customized services to them. Therefore, it is meaningful to mine the social tie between people hidden in this spatiotemporal data. These social tie offer an opportunity to understand users' requirements, such as friend recommendations or targeted advertisements for Internet companies [1].

Intuitively, users with higher social tie would have a greater chance to appear together at the same location, such as colleagues meeting in workdays or friends spending time together at a coffee shop. The methods inferring relationship have been widely studied [2–6] through co-occurrence feature and current context feature. Different from these works, we infer social tie from ***Multi-View Co-occurrence*** (*MVC*). As shown in Fig. 1, we apply the strong explanatory co-occurrence feature as baseline. At

L. H. U and H. Xie (Eds.): APWeb-WAIM 2018, LNCS 11268, pp. 378–392, 2018.
https://doi.org/10.1007/978-3-030-01298-4_31

the same time, we introduce one novel feature named topic co-occurrence feature. Additionally, we further combine context and co-occurrence information as context co-occurrence feature. The two novel features can both capture people's periodic mobility.

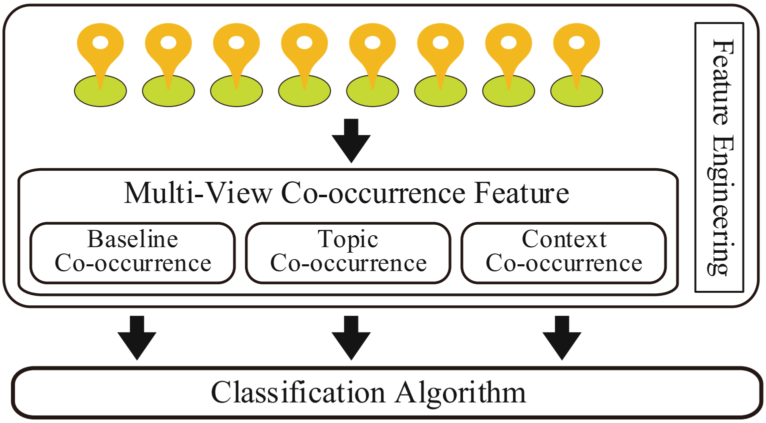

Fig. 1. The overview of Multi-view Co-occurrence

In summary, the main contributions to this paper are as follows:

(1) We carefully build spatiotemporal bag-of-words from both temporal and spatial aspect, and the user is regarded as word. Then we use Latent Dirichlet Allocation algorithm to extract topic feature representing user co-occurrence.
(2) We carefully construct context sequence from two different aspects. The context sequence includes co-occurrence, context and time periodicity information simultaneously. Then we present a novel method to extract context co-occurrence feature based on the context sequence. Our method carefully transfers the user-pair relationship in spatiotemporal data to word-pair relationship in sequence.
(3) The two novel co-occurrence feature contribute to the social ties inferring significantly. In the subset of Brightkite, the topic feature leads 9.1% improvement than the baseline in AUC indicator, context co-occurrence feature leads 9.3% improvement.

The remainders of this paper are organized as follows. Section 2 discusses related work. Section 3 describes our methodology in detail. Section 4 reports our experiments. Finally, we make conclusions in Sect. 5.

2 Related Work

We categorize the related works into three groups based on their focus: trajectory based methods [7–9], context based method [6] and co-occurrence based methods [2–5]. We compare our method with prior works in Table 1.

Table 1. Comparison between MVC and prior works

Characteristics	RWCFR [6]	EBM [3]	TAI [4]	SCI [5]	Multi-View Co-occurrence		
					Baseline Co-occurrence	Topic Co-occurrence	Context Co-occurrence
Location Diversity		√		√	√		
location context	√						√
temporal context							√
location co-occurrence		√	√	√	√	√	√
temporal co-occurrence		√	√	√	√	√	√
mobility periodicity				√	√	√	√

Trajectory based methods relaxed the concept of co-occurrence and use similarity in trajectory to measure likelihood of friendship between two people [4]. Chen et al. [7] applied frequent sequential pattern mining technology to extract the sequence of places that a user frequently visits, then use them to model his mobility profile combined with semantics of spatiotemporal information. [8, 9] focused on measuring user similarity using trajectory patterns, and [8] provide a tool named MinUS which integrates the technologies of trajectory pattern mining on discovering user similarity.

In context based method, the context includes social context, personal preferences context, location context and temporal context. Bagci et al. [6] proposed a random walk based context-aware friend recommendation algorithm (RWCFR). Depending on the location-based social network, they build a graph according to the current context (i.e. social relations, personal preference and location) for user. The method demonstrate that the context can describe the users' social tie. However, spatiotemporal data usually includes co-occurrence information, making full the use of spatiotemporal co-occurrence can further enhance the prediction accuracy.

Co-occurrence based methods had been shown to improve accuracy of social relationship estimation than trajectory based methods because of the co-occurrence feature [4]. Grandall et al. [2] demonstrated that the co-occurrence feature contributes to inferring social ties based on the experiments with a dataset of 38 million geo-tagged photos from Flickr. They also had shown that the probability of a social tie increases as the number of co-occurrence times increases and the temporal range decreases. Pham et al. [3] proposed an entropy-based model (EBM) that estimates the strength of social connections by analyzing people's co-occurrences in space and time through diversity and weighted frequency. Zhou et al. [4] proposed a Theme-Aware social strength Inference (TAI) approach that mines theme (also called the unit for co-occurrence) from co-occurrence behaviors, and then leverages the theme to measure the social strength of two persons. Njoo et al. [5] proposed a unified framework called SCI framework (Social Connection Inference framework). The SCI framework quantified three key co-occurrence features (i.e. diversity, stability and duration), and then aggregate co-occurrence features using machine learning algorithms to predict the social ties.

In summary, [2–5] illustrated the importance of co-occurrence feature which is also considered in our proposed two novel co-occurrence features. [6] shows that location context contribute to social tie prediction, therefore the context co-occurrence introduce location context. The check-in time sequence can reflect social tie between users to some context, and the characteristic is included in the context co-occurrence feature which is not involved in [6]. Different from [3–6], the context co-occurrence has novelty, which is not a traditional fusion. Peoples' mobility periodicity is also an import characteristic [10], and our proposed two novel features both take it into account. Generally, the characteristic of each view in MVC is shown in Table 1.

3 Methodology

In this section, we first describe how to generate baseline co-occurrence feature from co-occurrence times and location diversity. Moreover, we present the method to generate topic co-occurrence feature from location and time aspect. Finally, we describe how to generate context co-occurrence feature based on two carefully constructed context sequence.

3.1 Baseline Co-occurrence Feature

Times Co-occurrence Feature. The number of co-occurrence is powerful signal to infer social tie, which lead us to choose it as one of baseline feature. Intuitively, the more the times of co-occurrence between two users, the stronger the strength of social tie. More formally, co-occurrence set $\psi_{x,y}^{z} \in \psi_{x,y}$ quantifies the meeting frequency between users u_x and u_y in the location l_z during time threshold Δt. The parameter Δt can be set to different granularity (1 h, 2 h or 24 h). The $\psi_{x,y} = \{\psi_{x,y}^{z_1}, \psi_{x,y}^{z_2}, \ldots, \psi_{x,y}^{z_m}\}$ is the meeting frequency set for all meeting locations between users u_x and u_y. The $|\psi_{x,y}|$ is co-occurrence times, which is the number of two users appearing together.

Diversity Co-occurrence Feature. We also consider location diversity as the baseline feature, which is considered by [3, 5]. Variation in the meeting places between users is useful for reducing the possibilities of coincidences. For example, the co-occurrence number of user u_1 and u_2 is equal that of user u_1 and u_3, however u_1 meets u_2 several times in the same location, u_1 meeting u_3 a few times in several locations. The meeting occasions in the former are more likely to happen by chance than those in the latter. The reason is that the possibility of meeting in more diversified locations is lower than the possibly of meeting in the same location. Therefore, the location diversity feature is determined by Eq. 1.

$$Diversity(u_x, u_y) = -\sum \psi_{x,y}^{z} \log(\psi_{x,y}^{z}) \tag{1}$$

3.2 Topic Co-occurrence Feature

The topic feature was used in the paper [11, 12] in other domain, and we transfer it to apply in spatiotemporal data domain. The topic feature is mainly from two aspects.

Location Topic Co-occurrence Feature. There are lots of check-in location information in spatiotemporal data. In a certain location (e.g., a specific longitude and latitude), all users in the same location can be represented as a document and each user as a word. After removing less frequent location, we form vocabulary words from location-based spatiotemporal data. In order to mine co-occurrence feature between two users, we choose the Latent Dirichlet Allocation algorithm [13] to mine topic co-occurrence feature. We use a sparse matrix $x_{W \times M}$ to represent the bag-of-word representation of all locations, where there are $1 \leq m \leq M$ locations and $1 \leq w \leq W$ user. LDA allocates a set of thematic topic labels, $z = \{z^k_{w,m}\}$, to explain non-zero elements in the location-user co-occurrence matrix $x_{W \times M} = \{x_{w,m}\}$, where $1 \leq w \leq W$ denotes the word index in the vocabulary, $1 \leq m \leq M$ denotes the document index, and $1 \leq k \leq K$ denotes the topic index. Usually, the number of topics K is provided by us. The nonzero element $x_{w,m} \neq 0$ denotes the number of user check-in mth location. The objective of LDA inference algorithms is to infer posterior probability from the full joint probability $p(x, z, \theta, \phi)$, where z is the topic labeling configuration, $\theta_{K \times M}$ and $\phi_{K \times W}$ are two non-negative matrices of multinomial parameters for document-topic and topic-word distributions, satisfying $\sum_k \theta_m(k) = 1$ and $\sum_w \phi_w(k) = 1$. Both multinomial matrices are generated by two Dirichlet distributions with hyperparameters α and β. For simplicity, we consider the smoothed LDA with fixed symmetric hyperparameters. We use a coordinate descent (CD) method called belief propagation (BP) [14] to maximize the posterior probability of LDA,

$$p(\theta, \phi | x, \alpha, \beta) = \frac{p(x, \theta, \phi | \alpha, \beta)}{p(x | \alpha, \beta)} \propto p(x, \theta, \phi | \alpha, \beta). \tag{2}$$

The output of LDA contains two matrices $\{\theta, \phi\}$. The $\phi_{K \times M}$ can is the location topic co-occurrence feature of each user, which is useful for us.

Time Topic Co-occurrence Feature. The process of time topic co-occurrence feature is similar to that of location topic co-occurrence feature. In this situation, we see all user that check in the same day as a document, and see each user as a word. Then we use LDA to generate two matrices $\{\theta, \phi\}$. The $\phi_{K \times M}$ can is time topic co-occurrence feature of each user. The time granularity is set as day because people check-in usually present the characteristic of periodicity [10]. For example, the middle class check-in every morning and night in the company.

3.3 Context Co-occurrence Feature

In this section, we first give some important notations definition. Then, we carefully construct context sequence from two aspects (location-time and time-location) to represent spatiotemporal co-occurrence and context information. Specifically, we propose a new method to extract context co-occurrence feature based on two context sequence respectively.

Notation Definition. In Table 2 we list the notations of parameters that we use. We denote $u \in U$ as the user, and $c \in C$ as the check-in data. Each c reflects the appearance of a user u at a specific location l at a specific time t with the form of $\{u, t, l\}$. Each user has many check-ins c information. The C_t= $\{c_1,c_2,\ldots,c_n\}$ represent all check-ins during the same time period. $C_{t\&l}$ is a sequence of elements in C_t that ranked according location shortest distance principle. $C_t^{Sequence} = \{C_t^1, C_t^2, \ldots, C_t^N\}$ is a sequence that elements ranked according time order. C_l= $\{c_1,c_2,\ldots,c_m\}$ represents all check-ins in the same location. $C_{l\&t}$ is a sequence of elements in C_l that ranked according time order. $C_l^{Sequence} = \{C_l^1, C_l^2, \ldots, C_l^M\}$ is a sequence that elements rank according location shortest distance principle. The θ is a context sequence which only consists of user id. The parameters relationship is that $\sum_{i=1}^{M} \left|C_l^i\right| = \sum_{j=1}^{N} \left|C_t^j\right| = |C| = |\theta|$ and the θ usually includes many repetitive user id.

Table 2. Notation of parameters

Variable	Notation
$u \in U$	u is a user id; U is all different user id set $\{u_1, u_2, u_3, \ldots\}$
c	$\{u, t, l\}$, a user check in at specific location l at specific time t
C_t	$\{c_1, c_2, \ldots, c_n\}$, all c at the same time period
$C_{t\&l}$	Elements in C_t rank according shortest distance principle
$C_t^{Sequence}$	$\{C_t^1, C_t^2, \ldots, C_t^N$ $\}$, the elements C_t rank according time order
C_l	$\{c_1, c_2, \ldots, c_m\}$, all c at the same location l
$C_{l\&t}$	Elements in C_l rank according time order
$C_l^{Sequence}$	$\{C_l^1,C_l^2, \ldots, C_l^M\}$, the elements C_l rank according shortest distance principle
θ	A sequence that capture spatiotemporal context co-occurrence information

Location-Time Context Co-occurrence Feature. We first generate context co-occurrence sequence, and the generation process of location-time context sequence is given in Algorithm 1. *SortLocationByDistance* function produce sequence $C_l^{Sequence}$. The elements in $C_l^{Sequence}$ are ranked according distance shortest principle: if there is no location before, the first location is chosen randomly; otherwise, the location closest to the former location is assigned as the current location; the location closest to the *(M − 1)th* location is assigned as the *Mth* location; and so on. *SortTime* function uses quick sort algorithm to rank according time order because time is one-dimensional information. The returned value of the Algorithm 1 is the location-time context sequence θ consisted of user id, shown in Fig. 2 (B). Note that the same ellipse color represents the same location in Fig. 2 (B) and this context sequence capture strong location co-occurrence, meanwhile including shortest location context and time context.

The context co-occurrence feature is not simple fusion between context and co-occurrence, different from traditional approaches. We artfully use the toolkit word2vec to extract context co-occurrence feature through context sequence. The tool takes as its

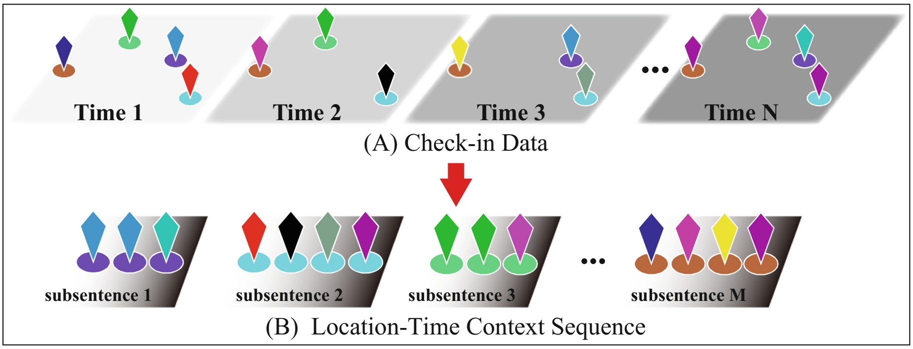

Fig. 2. The rhombuses with different colors denote different users, the ellipses with different colors denote different locations, and the time change is denoted by the shade of the background. (A) shows that at a fixed time period, different users check-in at different locations. (B) shows location-time context sequence. Note that the color of rhombuses, the color of ellipses and the change of background shade.

input a large corpus (corpus also can be seen as a sequence consist of words) and produces a dimensional space, with each unique word being assigned a vector in the space [15]. These word vectors are positioned in the vector space that words that share common contexts in the corpus are located in close proximity to one another in the space [15, 16]. The word vectors is context co-occurrence feature that we need. The feature includes spatiotemporal co-occurrence and context information.

More formally, given a context sequence $\theta = \{u_1, u_2, u_3, u_4, \ldots, u_T\}$ with spatiotemporal semantics representation, our objective is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}[\sum_{j=-k}^{k} logp(u_{t+j}|u_t)], \tag{3}$$

where T is number of elements in θ and k is the size of the window. The inner summation goes from $-k$ to k to compute the log probability of correctly predicting the user u_{t+j} given the user in the middle u_t. The outer summation goes over all users in the context sequence. The values of the two ends of the window are filled by the boundary value. Every user u is associated with two learnable parameter vectors, w_u and v_u. They are the "input" and "output" vectors of u respectively which can be learned [16]. The probability of predicting the user u_i given the user u_j is defined as

$$p(u_i|u_j) = \frac{\exp(w_{u_i}^{\mathrm{T}} v_{u_j})}{\sum_{l=1}^{U} \exp(w_l^{\mathrm{T}} v_{u_j})}, \tag{4}$$

where U is different users in the context sequence θ. The optimization approach is using stochastic gradient descent and the gradient is computed using backpropagation rule [16]. Each user's context semantic feature v_u (also called word vector in the

Natural Language Processing domain) can be learned. The word vector v_u captures word context and word co-occurrence information, which is extremely useful for us.

The toolkit word2vec is usually used to find synonyms in document (a sequence consist of words), and we innovatively apply it to finding user-pair relationship. Context feature captures sequence representation that context sequence has. Therefore, context co-occurrence feature also includes spatiotemporal co-occurrence and context. The feature with co-occurrence and context is different from above literatures [2–6], and not simply merges context and co-occurrence together.

Day-Location Context Co-occurrence Feature. The generation process of time-location context sequence is given in Algorithm 2. *SortTimeByGranularity* function can produce $C_t^{Sequence}$ quickly. The time parameter τ can be accurate to different value. The elements in C_t rank according distance shortest principle. The returned value of the Algorithm 2 is the time-location context sequence θ, shown in Fig. 3 (B). Note the distance between different ellipses in Fig. 3 (B). The time-location context sequence capture time co-occurrence, meanwhile including time order context and shortest location context.

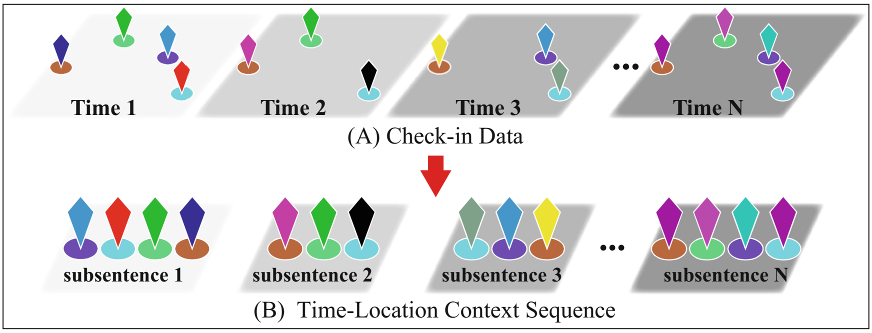

Fig. 3. (A) shows that at a fixed time period, different users check-in at different locations. (B) shows time-location context sequence. Note that the color of rhombuses, the color of ellipses and the change of background shade.

The time parameter τ is set to day. As shown in Fig. 3 (A), every chunk is one day i.e. time 1 is the first day; time 2 is the second day and so on. The elements in the time chunk C_t rank according distance shortest principle. Day-location is special time-location context sequence, because people's periodic movement is based on day [10] such as people commute on working day, people check in home at night. The periodicity of people's movement is a unit of day. Therefore, this context sequence captures people's mobility periodicity, including day co-occurrence, time order context and shortest location context. We do not adopt other parameter τ because the check-in data in location-based social networks is very sparse in time. Of course, if the strong time co-occurrence is needed, and the smaller parameter τ can be assigned.

After achieving the time-location context sequence $\theta = \{u_1, u_2, u_3, u_4, \ldots, u_T\}$ through above process, we use word2vec toolkit to extract context co-occurrence feature. The generation of feature is similar to that of the location-time context feature.

4 Experiments

In this section, we first describe datasets. Second, we describe how to generate three types of co-occurrence feature in detail. Moreover, we describe classifier algorithm learning. Finally, we evaluate our performance.

4.1 Datasets

Our experiments are based on a subset of the real dataset, Brightkite and Gowalla [17]. The original dataset is a global dataset, and the people in the global dataset have a heterogeneous nature. For the isomorphism of the data, we choose the data in the eastern US. The check-in data is handled to a triplet < *u, t, l* > , where the *l* is represented by longitude and latitude. The user-pair data is handled to a triplet < u_1, u_2, *label* > where the *label* indicates whether two users exist relationship.

The original dataset did not provide negative examples [5, 17] (all labels are true). As shown in Fig. 4, we use non-connected graphs to construct negative examples. In the figure, u1 and u2 are friends, u2 and u3 are friends. We see {u1, u2, u3} as a connected graph considering the transitivity of relationship. After constructing negative examples, the user-pair data is divided into training and testing dataset. The overview is shown in Table 3.

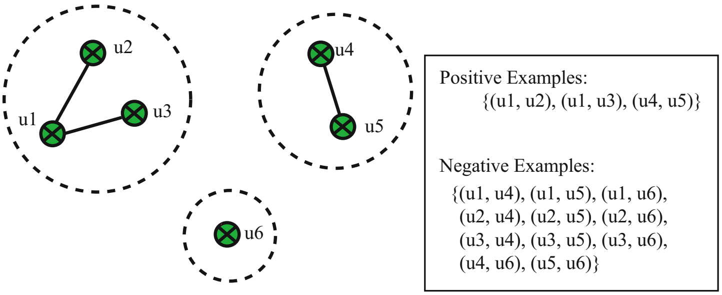
u2
u1
u3
u4
u5
u6
Positive Examples:
{(u1, u2), (u1, u3), (u4, u5)}
Negative Examples:
{(u1, u4), (u1, u5), (u1, u6),
(u2, u4), (u2, u5), (u2, u6),
(u3, u4), (u3, u5), (u3, u6),
(u4, u6), (u5, u6)}

Fig. 4. Negative sample construction

4.2 Multi-view Co-occurrence Feature Generation

Baseline Co-occurrence Feature. There are mainly two types of features, which are times co-occurrence and diversity co-occurrence. In our experiment, the parameter Δt of the feature is set to 1 h (i.e. 3600 s), 2 h (i.e. 7200 s) and 24 h (i.e. one day) respectively. According to these three types of granularity, three types of times co-occurrence sets in different location are generated respectively. Then, we can achieve

Table 3. Experiment datasets

Datasets	Brightkite	Gowalla
Checkins	837,161	732,205
Nodes (Users ID)	5,966	10,585
Train Data (user-pair)	291,543	87,567
Test Data (user-pair)	26,674	5,155

times co-occurrence and diversity co-occurrence based on specific granularity. The combination of the two type of features serves as input to the classifier (i.e. baseline-3600, baseline-7200, baseline-day). Based on the combination of the above three granularities, we call it baseline-merge feature that is used as the object of comparison with the other two novel features.

Topic Co-occurrence Feature. There are mainly two types of features, which are location and time topic co-occurrence feature. For both two type of features, we assigned $K = 100$ dimensional topic features for each location and each time period respectively. The time period is set as day (i.e. 24 h) to the time topic feature. We call these two types of features Topic-Day and Topic-Location respectively.

Context Co-occurrence Feature. Location-time sequence with spatiotemporal co-occurrence and context information can be generated through Algorithm 1. To achieve the time-location sequences through Algorithm 2, the parameter τ is assigned to day (i.e. 24 h). Then, two types of context sequences θ with co-occurrence and context information can be achieved. The word2vec provides an implementation of the skip-gram architecture which is in accord with our objective function [16], so we choose the skip-gram architecture. The context co-occurrence feature size is set as 200 and the window of max skip length between users is set as 10 (the parameter k in formula (3) is 10). The learning rate is set as 0.01 and other parameters are default. After the toolkit learning from context sequences, the context co-occurrence features with spatiotemporal information (the vector parameter v_u called word vector in NLP domain) can be learned. Each user is mapped to two types of 200 dimensional context co-occurrence features. The two context co-occurrence features represent co-occurrence and context information in spatiotemporal data. In our experiment, we call these two types of feature context-location-time and context-day-location.

4.3 Classification Algorithm

The multiple classifiers can be trained through three different views with different degree of co-occurrence information. In our experiment, we choose the XGBoost classifier to make prediction. It is a supervised learning method that uses a tree boosting technique. For a given datasets with n examples and m features $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in [0, 1])$, a tree ensemble model uses K additive functions to predict the output, as follows:

$$\bar{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{5}$$

where F is the space of the regression trees [18]. The output $\bar{y}_i$ is the relative probability of a user-pair relationship strength.

The user is represented with the co-occurrence feature and then the user-pair features combination as the input. Considering relationship is bidirectional in each view feature, a relatively larger vector position value is placed ahead of the corresponding position value. There are some primary parameters: the booster parameter is set as gbtree; the max depth is 3 which avoid overfitting; the boosting learning rate is 0.1; the objective function is binary logistic; the early stopping is 10; and other parameters are set as default. Multiple classifiers are used to predict social ties in the test data, and multiple types of result can be achieved. Generally, the higher the prediction result, the higher strength the two users' social ties.

4.4 Performance Evaluation

The classifier will output a list of top U user-pair that have the higher social tie. We use recall and prediction metrics on top U user-pair to evaluate the prediction results. Generally, increasing U will increase recall but decrease precision. Fix a certain U, the higher recall and precision correspond to the better prediction performance. The definition of recall@U is

$$\mathrm{R@U} = \frac{\text{The number of true user-pair in top U}}{\text{The total number of true user-pair}}. \tag{6}$$

Similarly, the definition of precision@U is

$$\mathrm{P@U} = \frac{\text{The number of true user-pair in top U}}{\mathrm{U}} \tag{7}$$

We also use the area under the ROC curve (AUC) [19] evaluated on the test data, which is the standard scientific accuracy indicator. Generally, we use AUC, R@U and P@U to evaluate the overall predictive performance.

Baseline Co-occurrence Feature. We compare different granularities on baseline co-occurrence in Tables 4 and 5. Coarse-grained time co-occurrence can achieve better result because of the sparseness of the dataset in the time dimension. The baseline-day achieve the best result compared to the baseline-3600 and baseline-7200, which largely depend on user movement periodicity. For example, people commute on working day, and people check in home at night [10]. Meanwhile, the baseline-day is also coarse-grained time granularity. The baseline-day contribute to the baseline-merge significantly. In all baseline features, the baseline-merge achieve the best precision and recall in top 500, because baseline-3600 and baseline-7200 contribute to short time co-occurrence and baseline-day contribute to periodic co-occurrence, which satisfy complementary principle. However, the baseline-merge does not increase in U at

13000, mainly because the sparsity of the dataset in the time dimension makes fine-grained time co-occurrence without any effect and only the baseline-day is in effect.

Table 4. Performance on Brightkite subset

	AUC	Precision@U	Recall@U	Precision@U	Recall@U
		1000 (Top 3.8%)	1000 (Top 3.8%)	13000 (Top 50%)	13000 (Top 50%)
Baseline-3600	0.552	0.789	0.131	0.526	0.540
Baseline-7200	0.566	0.828	0.164	0.533	0.554
Baseline-day	0.670	0.907	0.300	0.583	0.676
Baseline-merge	0.671	0.913	0.317	0.583	0.676
Topic-day	0.762	0.849	0.189	**0.633**	**0.825**
Topic-location	0.631	0.826	0.163	0.57	0.641
Context-day-location	**0.764**	**0.925**	**0.359**	0.622	0.789
Context-location-time	0.653	0.859	0.202	0.581	0.668

Topic Co-occurrence Feature. In the subset of Brightkite dataset, the topic-day features are better than topic-location, because the sparseness of the location is stronger than that of the time, and more difficult to infer the social ties. However, the conclusion is exactly the opposite in Gowalla because sparseness of the time is slightly stronger than that of the location. The sparseness of the time in Gowalla subset leads that topic co-occurrence feature is not good as the baseline-merge. The baseline-merge consider time co-occurrence from different granularities, which is more overall than the topic-day in the Gowalla subset. In the Brightkite subset, the topic-day better portray coarse-grained co-occurrence and takes second place in the AUC indicator. As shown in Figs. 5(a) and 6(a), with the U increasing, the topic-day achieved the best performance on the precision and recall compared to other views.

Context Co-occurrence Feature. In two datasets, the context co-occurrence feature achieve the best performance on AUC because it capture both context and co-occurrence information, other views only capturing co-occurrence information. The context-day-location feature achieve the better performance on AUC than context-location-time in Brightkite subset, because of the sparseness of location dimension information. Due to the sparseness of time in Gowalla subset, the context-location-time feature achieve the better performance on AUC than context-day-location. The context co-occurrence feature usually works better than corresponding topic co-occurrence feature. As shown in Figs. 5 and 6, the solid line of the corresponding color is above the dotted line in most of the time, because the context co-occurrence feature captures both co-occurrence and context information compared to the topic co-occurrence feature. From Figs. 5 and 6, we can also conclude that the context co-occurrence feature overall exceeds the baseline.

In summary, the two novel feature we proposed for extracting co-occurrence have their own advantages over the baseline. We emphasize the context co-occurrence feature because it captures the spatiotemporal context, co-occurrence and periodic mobility simultaneously. In general, it is better than the baseline and topic feature in AUC on the current two dataset subset.

Table 5. Performance on Gowalla subset

	AUC	Precision@U	Recall@U	Precision@U	Recall@U
		500 (Top 10%)	500 (Top 10%)	2500 (Top 50%)	2500 (Top 50%)
Baseline-3600	0.562	0.719	0.198	0.528	0.532
Baseline-7200	0.573	0.748	0.218	0.535	0.545
Baseline-day	0.736	0.939	0.445	0.621	0.719
Baseline-Merge	0.738	**0.944**	**0.454**	0.621	0.719
Topic-day	0.705	0.811	0.27	0.616	0.709
Topic-location	0.727	0.858	0.321	0.628	0.737
Context-day-location	0.761	0.827	0.287	0.649	0.788
Context-location-time	**0.782**	0.879	0.348	**0.656**	**0.805**

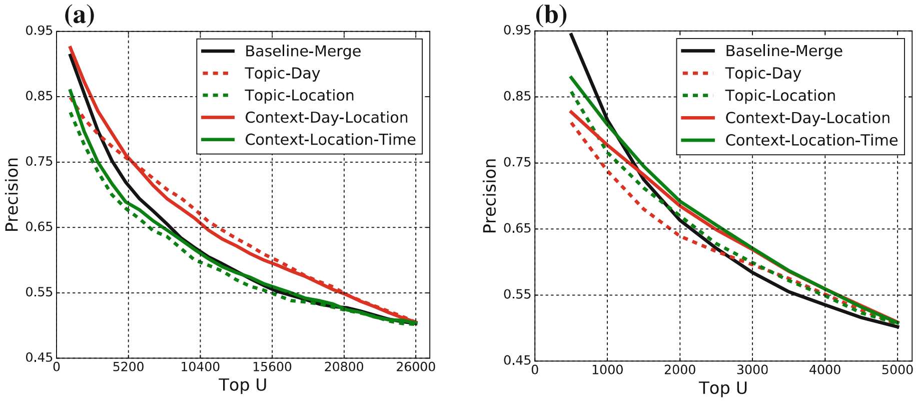

Fig. 5. (a) Precision on Brightkite subset. (b) Precision on Gowalla subset

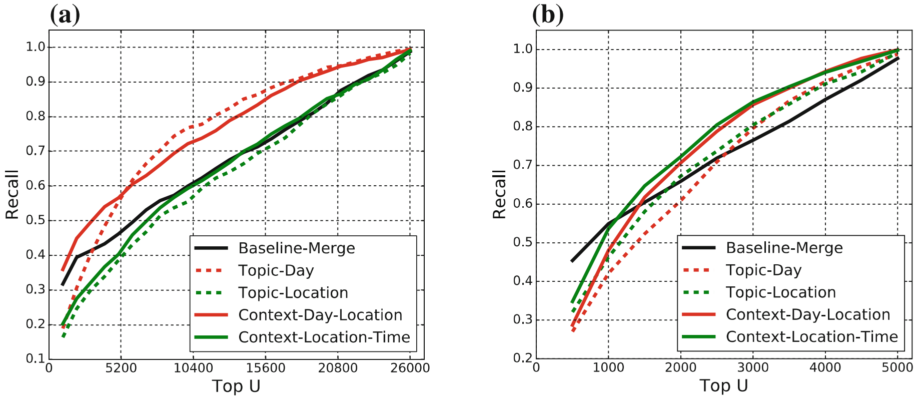

Fig. 6. (a) Recall on Brightkite subset. (b) Recall on Gowalla subset

5 Conclusion

In this paper, we infer social ties from multiple spatiotemporal co-occurrence, where the topic and the context co-occurrence features are presented. The two proposed co-occurrence feature are both from space and time aspects. The latter represents spatiotemporal context, co-occurrence and peoples' periodicity mobility simultaneously. The experiment results demonstrate that our two novel feature contribute to social ties inferring significantly.

References

1. Machanavajjhala, A., Korolova, A., Sarma, A.: Personalized social recommendations: accurate or private. Proc. VLDB Endow. **4**, 440–450 (2011)
2. Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring social ties from geographic coincidences. In: Proceedings of the National Academy of Sciences of America, pp. 22436–22441 (2010)
3. Pham, H., Shahabi, C., Liu, Y.: EBM: An entropy-based model to infer social strength from spatiotemporal data. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 265–276 (2013)
4. Zhou, N., Zhang, X., Wang, S.: Theme-aware social strength inference from spatiotemporal data. In: Li, F., Li, G., Hwang, S.-w, Yao, B., Zhang, Z. (eds.) WAIM 2014. LNCS, vol. 8485, pp. 498–509. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08010-9_56
5. Njoo, G.S., Kao, M.-C., Hsu, K.-W., Peng, W.-C.: Exploring check-in data to infer social ties in location based social networks. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10234, pp. 460–471. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7_36
6. Bagci, H., Karagoz, P.: Context-aware friend recomendation for location based social networks using random walk. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 531–536 (2016)
7. Chen, X., Pang J. and Xue R.: Constructing and comparing user mobility profiles for location-based services. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 261–266 (2013)
8. Chen, X., Kordy, P., Lu, R., Pang, J.: MinUS: mining user similarity with trajectory patterns. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8726, pp. 436–439. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44845-8_29
9. Chen, X., Lu, R., Ma, X., Pang, J.: Measuring user similarity with trajectory patterns: principles and new metrics. In: Chen, L., Jia, Y., Sellis, T., Liu, G. (eds.) APWeb 2014. LNCS, vol. 8709, pp. 437–448. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11116-2_38
10. 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, pp. 1082–1090 (2011)
11. Huang, Y., Zhu, F., Yuan, M., et al.: Telco churn prediction with big data [C]. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 607–618. ACM (2015)

12. Liu, G., Nguyen, T.T., Zhao, G., et al.: Repeat buyer prediction for e-commerce [C]. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155–164. ACM (2016)
13. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation [J]. J. Mach. Learn. Res., pp. 993–1022 (2003)
14. Zeng, J., Cheung, W.K., Liu, J.: Learning topic models by belief propagation[J]. IEEE Trans. Pattern Anal. Mach. Intell., pp. 121–1134 (2013)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013)
16. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting Similarities among Languages for Machine Translation (2013)
17. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection (2014)
18. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System (2016)
19. Bradley, A.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. In: Pattern Recognition. (1997)