IEICE TRANS. INF. & SYST., VOL.E98–D, NO.8 AUGUST 2015

PAPER

Hybrid Markov Location Prediction Algorithm Based on Dynamic Social Ties

Wen LI[†a)], **Shi-xiong XIA**[†], **Feng LIU**[††], ***Nonmembers***, ***and*** **Lei ZHANG**[†], ***Member***

SUMMARY Much research which has shown the usage of social ties could improve the location predictive performance, but as the strength of social ties is varying constantly with time, using the movement data of user's close friends at different times could obtain a better predictive performance. A hybrid Markov location prediction algorithm based on dynamic social ties is presented. The time is divided by the absolute time (week) to mine the long-term changing trend of users' social ties, and then the movements of each week are projected to the workdays and weekends to find the changes of the social circle in different time slices. The segmented friends' movements are compared to the history of the user with our modified cross-sample entropy to discover the individuals who have the relatively high similarity with the user in different time intervals. Finally, the user's historical movement data and his friends' movements at different times which are assigned with the similarity weights are combined to build the hybrid Markov model. The experiments based on a real location-based social network dataset show the hybrid Markov location prediction algorithm could improve 15% predictive accuracy compared with the location prediction algorithms that consider the global strength of social ties.

key words: *location prediction, dynamic social ties, hybrid Markov model, cross-sample entropy*

1. Introduction

The last few years have witnessed a considerable increase in the number of mobile devices along with the use of wireless communication and a rapid development of location-based social networks, which makes a large amount of individuals' movement data be collected. For example, high-resolution GPS mobility data, coarse-grained mobile phone data, social and geographical context obtained from online location-based social networks, etc. However, an increasing number of people are dissatisfied with the application correlated with their current locations. In order to be more proactive, applications must not just sense the user's current context, but also be able to predict the user's future context, which can make preparations for users' future activities and provide more natural and customized services for location-based applications [1]. In location prediction problem, given an object's recent movements, the location of this object at the future time is estimated. Clearly, location prediction is useful in inferring the crowd of a region, estimating the traffic status and so on.

Recently, various predicting technology are used for the location prediction, for example, Decision Tree [2], Bayesian network [3], [4], Markov Model [5], Neural Network [4], State predictor and Blending predictors [4], [5]. Most existing prediction techniques only take the user's own movement history as input to the location predictors. However, inspired by the social phenomenon that the social friends tend to have the similar behaviors, the researchers have started to consider the movements of friends or people with correlated mobility patterns for improving the prediction accuracy [6]–[13]. People may be more likely to visit the places that their friends or similar people visited in the past and human movements are usually affected by their social context, so the interdependence of human movement and social ties could help us to know an individual's potential interests and hints about when and where a particular user would like to go.

Although the studies on the correlations between individuals' historical movement and social behaviors have made great success, there are still several issues with existing location prediction approaches.

(1) In the real world, social network structure changes over time. Like the edges of the social network, the strength of social ties also dynamically changes, for example leveling up, increasing, decreasing, adding a new friend or friendship broken up. Eagle et al. [15] used mobile similarity to infer the social structures of users. The observations showed that spatial and temporal context is likely to be an important indicator of particular types of relationship and there is an evolution of relationships over time. Therefore, if we could find the most similar people or intimate friends at different times, then the interdependence of human mobility and social ties could be reflected more realistic, which reduces the deviation caused by the variation of social ties.

(2) Individuals' historical movement analysis is inevitable, so we integrate social ties into the Order-k Markov model to consider the movement patterns of the individual and his/her friends. Previous studies applied the cosine similarity [8] or the mutual information [7] of two users' location vectors to measure the mobile similarity, which only considers the distribution of different individuals' history of visited locations and neglects the transition information between the location pairs. Hence, this paper proposes a modified cross-sample entropy to quantify the correlation between the movements of different users. In this paper, we combine the user's own historical movements and his/her

Manuscript received August 29, 2014.
Manuscript revised April 2, 2015.
Manuscript publicized May 14, 2015.

[†]The authors are with School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China.
[††]The author is with China National Coal Association, Beijing 100713, China.
a) E-mail: liwen7881687@126.com

DOI: 10.1587/transinf.2014EDP7296

friends' movement data which have the most similar mobility patterns at different time intervals. Specifically, our contributions in this study include:

- we present how to measure the similarity of the behavior sequences of two social friends by a modified cross-sample entropy and we demonstrate it is possible to improve the prediction accuracy by the movement history of his friends.

- we propose a hybrid prediction model to combine the individual's visited history and his friends' movement history with different weights at different time slices, which could reflect the variation of the strength of the social ties with the changes of time.

- we evaluate our algorithm using a online location-based social network (LBSN) dataset. The results show that our hybrid Markov location prediction model based on dynamic social ties (HM-DST) could provide a better predictive accuracy than the location predictors without considering the dynamic social ties.

The remainder of this paper is organized as follows. We first give a brief review of some related work in Sect. 2, then introduce the dataset used in this paper in Sect. 3. Section 4 proposes the HM-DST model. The experimental results are presented in Sect. 5. Finally, we conclude the study in Sect. 6.

2. Related Work

In recent years, the popularity of location-based and preference-aware recommender systems provide users more opportunities to leverage similar friends' experiences to extend individual knowledge and retrieve the information matching their tastes with minimal efforts [11]. Therefore, we could obtain a better predictive performance by virtue of the movement data of their friends or similar users who share similar interests, locations and travel sequences. Moreover, some researchers incorporate geography into the prediction model and apply movement patterns of the group to individual location predictions. Calabrese et al. combined collective movement patterns, time of day, land use and points of interests to build a probabilistic model which can predict an individual's next location within an hour correctly 60% of the time [12]. Xiong introduced the collective behavioral patterns (CBP) and then proposed CBP-based Bayesian model to learn the correlations with time-shifting from the mobility data of crowds [13]. De Domenico used the concept of mutual information to quantify the correlation between two mobility traces and considered the movement of the people who have correlated mobility patterns to improve the prediction accuracy [7]. Cho developed a model of human mobility that combines periodic short range movements with long-distance travel that is influenced by social network ties. The experiments showed that social relationships can explain about 10% to 30% of all human movement, while periodic behavior explains 50% to 70% [6]. Gao proposed a social-historical model that integrates the social and historical effects to explore user's check-in behaviors on the location-based social networks (LBSNs) [8]. Above studies have verified that the social and historical ties could improve the location predictive performance. Further, for new users, the training data is typically insufficient and unavailable, which lead to a poor predictive performance, so Mclnerney presented a framework to enhance prediction using information about the mobility habits of existing users [14]. Hence, the movement patterns could not only increase the prediction accuracy but also decrease the proportion of prediction failures when they visit a new location.

Table 1 Statistics information of the experimental dataset

Number of users	253
Number of check-ins	441182
Number of social friendships	16757
Number of unique locations	80421
Locations visited at least five times	11802
Average check-ins per user	1743
Average number of friends per user	66
Average duration per user (day)	702
Average check-ins per user per day	3.12

3. Dataset Description

We consider a dataset of online location-based social network -Brightkite- to capture human mobility. Brightkite allows users to share their locations with their friends and the friendship relationship is mutual. Users logged in Brightkite could make check-ins at the geographic locations, where users can see who is nearby and who has been there before. We study a dataset collected between Apr. 2008 and Oct. 2010 by Gao [16], which contains a friends list and the list of all the check-ins. Each check-in is represented as a tuple < userId, check-in time, latitude, longitude, location id > and the friend list records which two users have the social relationship. The whole dataset contains 58,228 users, 214,078 social ties and 4,491,143 check-ins. However, the number of the users is very large and the check-ins frequency reflect the user's degree of the activity, so the records of the users who just make small check-ins and spend several days could not show the complete daily life. Therefore, we select the more active users to do the research. In our experiments, 253 users who have at least 1000 check-ins and over 100 locations and spend more than 400 days on Brightkite are selected. We obtain 80,421 unique geographical check-in locations from the whole selected dataset as the location vocabulary and 360 average distinct locations per user. Table 1 lists the summary statistics.

The location-based social networks provide location-based specific data, as one can distinguish between a check-in to the office on the 2nd floor and a check-in to a coffee shop on the 1st floor of the same building [6]. These fine-grained and precise locations make individuals' activity intentions and macro mobility regularity get less attention than the specific movement places. For instance, when a user went into a business district, he was likely to visit several separate places. However, these places did not form a

fixed pattern and might be varied at different times. Therefore, the clustering of check-in locations at different spatial scales could help us learn more about individuals' movement regularity.

In this paper a grid-based clustering method is used to generate the grid-based locations, which contains two parameters, namely the *origin* = $(long_o, lat_o)$ and spatial scale s [17]. The origin is a given reference point to divide the grid map and we set as $(0, 0)$. The parameter s is a numeric value used to specify the size of the grid cell. Suppose the grid cell of a given map has a scale s, then the increment of adjacent grid points differs by $0.001^\circ s$ in longitude or latitude. Every check-in location corresponds to a grid cell. When the scale is 0, the location of each check-in is the latitude and longitude recorded in the original dataset. With the increase of s, more and more original check-in locations are clustered into a single grid cell. The location with larger scales is the central point of the corresponding grid cell. When the scale is 10, the size of each grid cell is approximately $0.9km^2$, or that of a large commercial district.

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be the set of users and $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$ be the set of locations where N and M are the numbers of users and locations respectively. Each check-in action is represented as a tuple $<u_i, l_j, t_k> \in C$, indicating user $u_i \in \mathcal{U}$ checks in at location $l_j \in \mathcal{L}$ at time t_k, where C is the observed check-in set. Let $\mathcal{F}(u)$ denote u's social friends. Let $x_u(t) \in \mathcal{L}$ denote the geographic location of user u at time t, $\mathcal{H}_u(t_1, t_n) = <x_u(t_1), x_u(t_2), \ldots, x_u(t_n)>$, where $x_u(t_i) \in \mathcal{L}$ and $t_1 \le t_i \le t_n$ be the observed historical location sequence of u between time t_1 and t_n, $\mathcal{S}_u(t_1, t_n) = \{\mathcal{H}_{u_f}(t_1, t_n) | u_f \in \mathcal{F}(u)\}$ be the observed location sequences of u's friends or potential similar users between time t_1 and t_n.

4. Hybrid Markov Location Prediction Model Based on Dynamic Social Ties

4.1 The Modified Cross-Sample Entropy

Due to individual's sociality, his/her mobility patterns could be inferred from the movement data of his/her friends. If two individuals usually visit the same places or follow the similar paths, they may have the similar life patterns and each person's movements could have a positive effect on his friends' behaviors. Existing methods for measuring the similarity only consider the individuals who visit the same places or have a meeting at the same places and the similar time, which neglects the similar mobility patterns among the visit histories of the user and his friends. If two friends usually go through the similar location sequence, their future movements may produce more intersection. Hence, we define two users' mobility similarity using a modified cross-sample entropy (MCS) which do not consider the difference between the different embedding dimensions. The modified cross-sample entropy is an effective technique for analyzing the degree of synchrony between two related time series [18], [19]. Greater value of MCS means the existence of some similar patterns in the two time series. For the individuals who have similar life style and preference, the modified cross-sample entropy between their location sequences should be relatively high. We present algorithm 1 to describe our modified cross-sample entropy's detailed calculation steps, where user u_1's location sequence between time t_1 and t_{n_1} is $\mathcal{W} = \mathcal{H}_{u_1}(t_1, t_{n_1}) = <x_{u_1}(t_1), x_{u_1}(t_2), \ldots, x_{u_1}(t_{n_1})>$, u_2's location sequence between time t_1 and t_{n_2} is $\mathcal{V} = \mathcal{H}_{u_2}(t_1, t_{n_2}) = <x_{u_2}(t_1), x_{u_2}(t_2), \ldots, x_{u_2}(t_{n_2})>$, the embedding dimension is m (the length of vectors to be compared) and the similarity tolerance is r (the tolerance for accepting matches, which limits the maximum distance between the corresponding locations).

Algorithm 1: The calculation steps of the modified cross-sample entropy

(1) $n_1 - m + 1$ vectors $\mathcal{W}_i^m = <x_{u_1}(t_i), x_{u_1}(t_{i+1}), \ldots, x_{u_1}(t_{i+m-1})>$ and $n_2 - m + 1$ vectors $\mathcal{V}_j^m = <x_{u_2}(t_j), x_{u_2}(t_{j+1}), \ldots, x_{u_2}(t_{j+m-1})>$ are extracted from location sequence $\mathcal{W}$ and $\mathcal{V}$ respectively.

(2) The distance between vector $\mathcal{W}_i^m$ and its neighbors $\mathcal{V}_j^m$ is defined as $d[\mathcal{W}_i^m, \mathcal{V}_j^m] = max\{gd[x_{u_1}(t_{i+k}), x_{u_2}(t_{j+k})] \mid 0 \le k \le m-1, 1 \le j \le n_2 - m + 1, i \ne j, |i-j| \ge m\}$, where $gd[x_{u_1}(t_{i+k}), x_{u_2}(t_{j+k})]$ is the ground distance between their corresponding representative point of locations $x_{u_1}(t_{i+k})$ and $x_{u_2}(t_{j+k})$.

Suppose that $x_1 = (Lon_1, Lat_1)$ and $x_2 = (Lon_2, Lat_2)$ are two GPS locations and Lon and Lat are the longitude and latitude in degrees. The ground distance between the locations x_1 and x_2 is defined as [20]:

$A = sin(Lat_1) * sin(Lat_2) * cos(Lon_1 - Lon_2) + cos(Lat_1) * cos(lat_2)$

$gd[x_1, x_2] = R * Arccos(A)$

Where R is the average radius of the earth and the value of R is 6371.004 km.

(3) Let $MCS_i(m, r) = (\text{number of } 1 \le j \le n_2 - m + 1 \text{ such that } d[\mathcal{W}_i^m, \mathcal{V}_j^m] \le r, i \ne j, |i-j| \ge m)/(n_2 - m + 1)$ be the ratio of the number of $d[\mathcal{W}_i^m, \mathcal{V}_j^m] \le r$ to the whole number of vectors $n_2 - m + 1$, in which $|i - j| \ge m$ is the Theiler window [21] and removes the contribution of the close location vectors as a result of temporal correlation.

(4) Then define $MCS_{<\mathcal{W},\mathcal{V}>}(m, r) = \frac{\sum_{i=1}^{n_1-m+1} MCS_i(m,r)}{n_1-m+1}$.

4.2 The Dynamic Changes of Social Ties

Due to the fact that users are more likely to visit places that their friends visited in the past, we could obtain a better predictive performance by virtue of the movement data of their friends or similar users who share similar interests, locations and travel sequences. The strength of social ties between the user and his friends changes with time and the changes usually reflect in two aspects:

(1) The relationships vary with the absolute time. The social interactions between two friends usually experience several stages, for example building, enhancing, weakening or disappearing. The relationships between two individu-

als are in different stages at different times. For example, when the user u meets an individual u_1 a few days, their relationship is relatively close, but as more new friends join in u's social circle, u might reduce or strengthen the interactions between u_1. Therefore, using the movement data of the user's social friends who has closest interactions during different time periods could obtain better predictive performance. This paper uses week as the division scale of the absolute time.

(2) The relationships vary with different time slices. Individuals' social interactions not only contain different stages but also vary in different mappings of the relative time, for example the time could be projected down to each day of the week, each hour of the day, the workdays and weekends, etc. Nathan presented the social connections are concentrated on colleagues' traditional behaviors at work. While on weekends the individuals may do activities with their family or the friends who have similar interests [11]. Hence, the mapping of the relative time could distinguish the similarities between the user and his friends according to the user's movement features and preferences. This paper will project the time to workdays and weekends.

As the influences of social ties on user's movement do not always show up as the simultaneously visiting at the same location, the user may pay attention to the locations that their friends recommended and visited in the past. The similar behaviors between the user and his friends not only occur at the same time period, when calculating the strength of the social ties between user u and his friend u_1 at a certain time period between t_1 and t_n, we compare the movement data of u_1 at that time interval with the movements of u before time t_n.

4.3 Hybrid Markov Model Based on Dynamic Social Ties

Order-k Markov model is a very popular model to predict the individual's next location. Among those, Markov model is easy-implemented and effective method. In this paper, Markov model is used to be a base model and We will discuss the performance of Order-k Markov in the future. Markov model could be used to describe dynamic changes in user's movement behaviors, which considers that the current motion state depends only on the previous motion state. Markov model is a probabilistic automation in which states represent individual's historical locations and the transition matrix describes the probabilities of particular transitions between states. Given a series of historical visits, the locations $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$ correspond to the states $\mathcal{EL} = \{E_1, E_2, \ldots, E_M\}$ of Markov model and the movements between locations could be used to calculate the transition matrix between states.

We propose a hybrid Markov location prediction model based on dynamic social ties (HM-DST) to integrate user's personal and social movement data. Let $\mathcal{H}_u(t_1, t_n) = <x_u(t_1), x_u(t_2), \ldots, x_u(t_n)>$ be user u's observed historical location sequence between time t_1 and t_n and $\mathcal{S}_u(t_1, t_n) = \{\mathcal{H}_{u_f}(t_1, t_n) | u_f \in \mathcal{F}(u)\}$ be the observed location sequences of u's friends between time t_1 and t_n. We present algorithm 2 to describe the processes of building the model HM-DST.

Algorithm 2: hybrid Markov location prediction model based on dynamic social ties

(1) In order to reflect the dynamic changes of the social ties between the user u and his/her friends $\mathcal{F}(u)$, the historical location sequences of u and $\mathcal{F}(u)$ are segmented into several location subsequences based on the time division. For the time interval $[t_1, t_n]$, it could be described as a concatenation of a series of time slices. That is, $[t_1, t_n] = \bigcup_{k=1}^{T}(t^s_{k,workdays}, t^e_{k,workdays}) \cup (t^s_{k,weekends}, t^e_{k,weekends})$, where $T = \lceil (t_n - t_1)/7 \rceil$ is the total week number between t_1 and t_n, $t^s_{k,workdays}$ and $t^e_{k,workdays}$ are the start time and end time of the working days on the k-th week respectively, $t^s_{k,weekends}$ and $t^e_{k,weekends}$ have the similar definitions. Therefore, user u's location sequence $H_u(t_1, t_n)$ could be expressed as $\mathcal{H}_u(t_1, t_n) = \bigcup_{k=1}^{T} \mathcal{H}_u(t^s_{k,workdays}, t^e_{k,workdays}) \cup \mathcal{H}_u(t^s_{k,weekends}, t^e_{k,weekends})$.

(2) For evaluating the influence of friends' movements on user u's mobility behaviors in different time intervals, we apply the modified cross-sample entropy to quantify the similarities between friends' movements during different time periods and u's historical movements. The modified cross-sample entropy between the location subsequences of the user and his/her friends is larger and the mobility similarities between the user and his/her friends is higher, and the effect on the user's movements is stronger. Suppose that friend u_f's location subsequence on the workdays of the k-th week is $\mathcal{W}_f(k, workdays) = \mathcal{H}_{u_f}(t^s_{k,workdays}, t^e_{k,workdays})$ and the historical location subsequence of user u on the workdays before k-th week is $\mathcal{V}(k, workdays) = \bigcup_{j=1}^{k} \mathcal{H}_u(t^s_{j,workdays}, t^e_{j,workdays})$, the standardized movement similarity between u_f and u during time $t^s_{k,workdays}$ and $t^e_{k,workdays}$ is defined as $Sim_{<u_f,u>}(k, workdays) = \frac{MCS_{<\mathcal{W}_f(k,workdays),\mathcal{V}(k,workdays)>}(m,r)}{\sum_{j=1}^{|\mathcal{F}(u)|} MCS_{<\mathcal{W}_j(k,workdays),\mathcal{V}(k,workdays)>}(m,r)}$.

Because we assume the current motion state depends only on the previous motion state, the embedding dimension m is set as 2 and the value of r is determined according to the selection of spatial scale s.

(3) In order to reflect the different importance of personal motion and friends' movements for building the location prediction model, we use an parameter $\eta \in [0, 1]$ to control the weights between historical and social ties. A location sequences cluster composed by user u's personal historical location sequence $\mathcal{H}_u(t_1, t_n)$ and friends' location subsequences set $\{\mathcal{H}_{u_f}(t^s_{k,w}, t^e_{k,w}) | u_f \in \mathcal{F}(u), 1 \le k \le T, w \in \{workdays, weekends\}\}$ are used as input to build a hybrid Markov location prediction model based on dynamic social ties (HM-DST). In order to distinguish the contributions of different location subsequences of u's friends to u's potential mobility behaviors, we assign a weight $(1 - \eta) \times Sim_{<u_f,u>}(k, w)$ to the location subsequence $\mathcal{H}_{u_f}(t^s_{k,w}, t^e_{k,w})$. Therefore, the state space of HM-DST is constitute of the states $\{E_1, E_2, \ldots, E_M\}$ which direct to the locations $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$ and the transition probability between state E_i and E_j can be calculated by

$$P_{ij}^{HM-DST} = \frac{\eta \times N_u(i,j) + (1-\eta)\sum_{u_f \in \mathcal{F}(u)} WN_{u_f}(i,j)}{\eta \sum_{q=1}^{M} N_u(i,q) + (1-\eta)\sum_{u_f \in \mathcal{F}(u)} \sum_{q=1}^{M} WN_{u_f}(i,q)}$$

$WN_{u_f}(i,j) = \sum_{w \in \{workdays, weekends\}} \sum_{k=1}^{T} Sim_{<u_f,u>}(k,w) \times N_{u_f,<k,w>}(i,j)$

$N_u(i,j)$ = (number of moving from the location l_i to l_j in u's location sequence $\mathcal{H}_u(t_1,t_n)$)

$WN_{u_f}(i,j)$ = (weighted number of moving from the location l_i to l_j in u_f's location subsequences set)

$N_{u_f,<k,w>}(i,j)$ = (number of moving from the location l_i to l_j in u_f's location subsequence $\mathcal{H}_{u_f}(t^s_{k,w}, t^e_{k,w})$)

Consequently, given the hybrid Markov location prediction model based on dynamic social ties and user u's last location $x_u(t_n) = l_i$, the predictive probability of the next check-in at location l_j is defined as

$P_u^{HM-DST}(x_u(t_{n+1}) = l_j | x_u(t_n = l_i)) = P_{ij}^{HM-DST}$

The next check-in location is predicted to be the location l that has the maximum transition probability from location l_i.

5. Experiments

5.1 Evaluation Metrics

Let $\mathcal{X} = <x_u(t_1), x_u(t_2), \ldots, x_u(t_n)>$ denote an user u's observed location sequence, and $\mathcal{X}^* = <x_u^*(t_1), x_u^*(t_2), \ldots, x_u^*(t_n)>$ be the predicted location sequence by our hybrid Markov location prediction model based on dynamic social ties. We divide the location sequence of each user into 10 parts, and each part has approximately equal visit time. Let the time stamp at the end of each part be $\{T1, T2, \ldots, T10\}$. We predict the locations at each part for the user, with his historical visits before that time as observed context. In addition, we put at least 10% of visits in the training set, so the comparisons of prediction results start from $T2$.

To measure the predictive performance of different mobility models, we use the prediction accuracy (PA) as the evaluation metric. The prediction accuracy for user u at part $T(i)$ is defined as $PA(T(i)) = \sum_{T(i-1)<t_i \le T(i)} I(x_u(t_i) = x_u^*(t_i))$/(check-in times between $T(i-1)$ and $T(i)$), where $I()$ is the Boolean indicator function that returns 1 if its argument evaluate to true and returns 0 otherwise.

5.2 Baseline Model

To evaluate the effectivity of our HM-DST model and other social prediction models, we choose four baseline models with detailed descriptions below: $\mathcal{X} = <x_u(t_1), x_u(t_2), \ldots, x_u(t_n)>$ is the set of visiting history and $x_u(t_n)$ and $x_u(t_{n+1})$ are individual's last visit location and next visit location.

1) Markov Model (Markov)

The Markov model considers the latest visited place as context, and searches for the most frequent patterns to predict the next location. The probability of the next visit $x_u(t_{n+1})$ at location l with Markov model is defined as:

$$P_u^{Markov} = P_u^{Markov}(x_u(t_{n+1}) = l | x_u(t_n) = l_k) = \frac{|x_u(t_r)|x_u(t_r) \in \mathcal{X}, x_u(t_r) = l, x_u(t_{r-1}) = l_k|}{|x_u(t_r)|x_u(t_r) \in \mathcal{X}, x_u(t_{r-1}) = l_k|}$$

2) Markov model with cosine similarity (Markov-Cos)

The Markov model with cosine similarity considers not only the user u's personal movement data, but also the movements of friends with correlated mobility patterns (characterized by high cosine similarity). The parameter η is used to control the weight between historical and social ties. The probability of the next visit $x_u(t_{n+1})$ at location l with Markov-Cos is defined as:

$$P_u^{Markov-Cos}(x_u(t_{n+1}) = l | x_u(t_n) = l_k) = \eta P_u^{Markov} + (1-\eta) \sum_{u_f \in \mathcal{F}(u)} Sim_{Cos}(u, u_f) P_{u_f}^{Markov}$$

where $Sim_{Cos}(u, u_f)$ is the cosine similarity between u and u_f. For each user, let $v \in R^M$ be his visits vector with each element $v(k)$ equal to the number of visits at location $l_k \in \mathcal{L}$, where $M = |\mathcal{L}|$ is the vocabulary size. The cosine similarity of two users u and u_f is defined as:

$$Sim_{Cos}(u, u_f) = \frac{v.v_f}{|v|_2 \times |v_f|_2}$$

where $|.|_2$ is the 2-norm of a vector.

3) Markov model with mutual information similarity (Markov-MI)

The Markov model with mutual information similarity also considers the movements of friends and quantify the correlation of two mobility trace by mutual information. The probability of the next visit $x_u(t_{n+1})$ at location l with Markov-MI is defined as:

$$P_u^{Markov-MI}(x_u(t_{n+1}) = l | x_u(t_n) = l_k) = \eta P_u^{Markov} + (1-\eta) \sum_{u_f \in F(u)} Sim_{MI}(u, u_f) P_{u_f}^{Markov}$$

where $Sim_{MI}(u, u_f)$ is the mutual information similarity between u and u_f. Suppose that X and Y are the visits history of user u and u_f, random samples x drawn from $\mathcal{X}$ and y drawn from $\mathcal{Y}$ correspond to the geographic coordinates. The Probability Density Functions (PDF) of x $P_X(x)$ represents the fraction of the times visited by user u in a particular position x. $P_Y(y)$ is the PDF of y which measures the fraction of the times visited by the user u_f in a particular position y and $P_{XY}(x,y)$ is the joint probability. The mutual information $I(\mathcal{X}, \mathcal{Y})$ is defined as:

$$I(\mathcal{X}, \mathcal{Y}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{\mathcal{X}\mathcal{Y}}(x,y) log \frac{P_{\mathcal{X}\mathcal{Y}}(x,y)}{P_{\mathcal{X}}(x) P_{\mathcal{Y}}(y)}$$

4) Markov model with the modified cross-sample entropy similarity (Markov-MCS)

The Markov model with our modified cross-sample entropy similarity also consider the movements of friends and quantify the correlation of two mobility trace by the modified cross-sample entropy. The probability of the next visit $x_u(t_{n+1})$ at location l with Markov-MCS is defined as:

$$P_u^{Markov-MCS}(x_u(t_{n+1}) = l|x_u(t_n) = l_k)$$
$$= \eta P_u^{Markov} + (1-\eta) \sum_{u_f \in \mathcal{F}(u)} Sim_{MCS}(u, u_f) P_{u_f}^{Markov}$$

where $Sim_{MCS}(u, u_f) = \frac{MCS_{<\mathcal{H}_u(t_1,t_n),\mathcal{H}_{u_f}(t_1,t_n)>}(m,r)}{\sum_{u_q \in \mathcal{F}(u)} MCS_{<\mathcal{H}_u(t_1,t_n),\mathcal{H}_{u_q}(t_1,t_n)>}(m,r)}$

5.3 Performance Evaluation

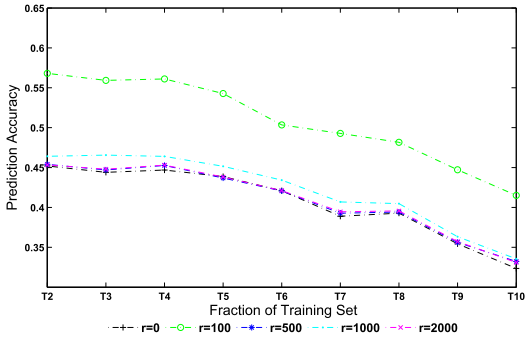
5.3.1 Prediction Performance vs Spatial Scale and Similarity Tolerance

We first study the effects of the prediction accuracy with different spatial scales s and the similarity tolerance r. When s is large, many original GPS locations may be included in a single grid cell. With the increase of s, the prediction locations will become larger and a lower number of locations could be obtained. So the prediction accuracy would be improved. Moreover, the location sequences which are not similar on the small spatial scale may become the same patterns on larger s. As seen in Fig. 1, when $r = 0$, the prediction accuracy of s $= 20$ obtains a 11% and 4% improvements than that of $s = 0$ and $s = 10$.

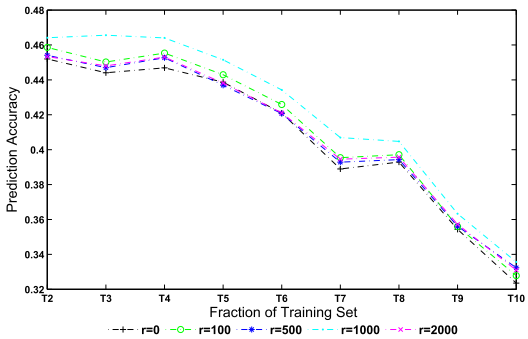
We also investigate the effects of the similarity tolerance r. When r is large, some location sequences which have the long distances would be treated as the similar location sequences. So the movements of the unfamiliar friends may produce a great impact to the user, which reduces the predictive performance. From the Fig. 1, we can see that the prediction accuracy increases at first and then decrease. When $s = 10$ and 20, The figures reach a peak at about $100 * s$. That is because each grid cell is almost $100m$. When the spatial scale is s and the maximum distance of two location sequences is smaller than the length of the grid cell, the two location sequence are similar and the movement history of his friend would make a positive effects to the user's location prediction. However, if the spatial scale is 0, the prediction accuracy will obtain the best performance when the r is about $100m$. the location sequences of the user and his friend will be matched when the latitudes and the longitudes of the two location sequences are exactly the same, which could neglect the randomness of the movements. Hence, in the next experiments, we set $r = 100 * s$ when $s > 0$ and $r = 100$ when $s = 0$.

5.3.2 Comparing with Personal Mobility Model and Other Social Models

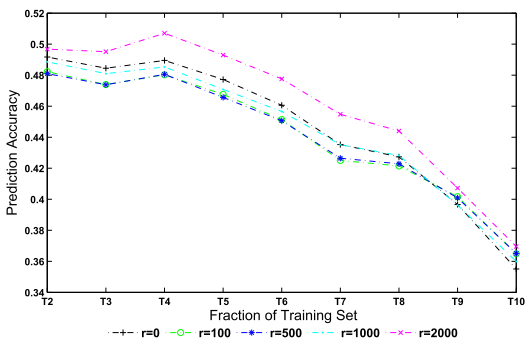
We compare the prediction results of hybrid Markov location prediction model based on dynamic social ties (HM-DST) with Markov model and some social Markov models, for example, Markov-Cos which measures the strength of social ties by cosine similarity, Markov-MI which uses mutual information to quantify the similarity, Markov-MCS which applies our modified cross sample entropy but not considers the dynamic changes. Figure 1 shows the prediction accuracy of Markov, Markov-Cos, Markov-MI,



(a) s=0



(b) s=10

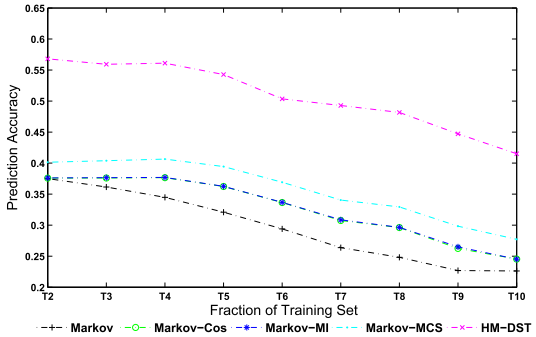


(c) s=20

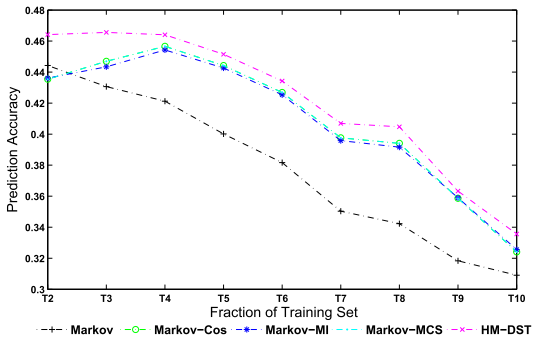
Fig. 1 The prediction performance changes with different spatial scales s and similarity tolerances r

Markov-MCS and HM-DST changes with the fraction of training set when $\eta = 0.5$ and the spatial scale s is set as 0, 10 and 20.

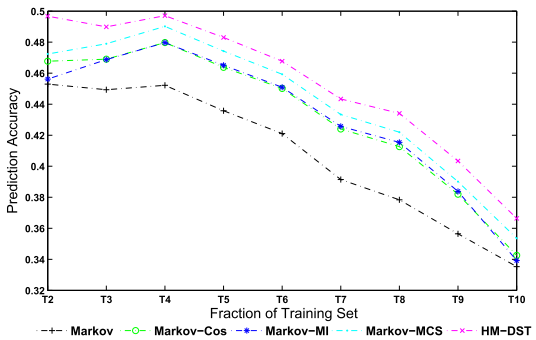
From Fig. 2 (a), we can see HM-DST model performs the best with an average accuracy of 50% and Markov model performs the worst. The social Markov model can obtain more than 3% prediction accuracy than Markov model which doesn't consider the social ties. Markov-Cos model and Markov-MI model get similar predictive performance. Markov-MCS provides 2% improvement over Markov-Cos model and Markov-MI model. That is because the modified cross-sample entropy can be more effective in measuring two users' mobility similarity. HM-DST model consid-



(a) s=0 r=100

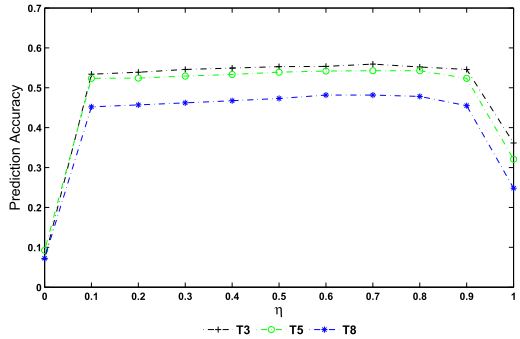


(b) s=10 r=1000

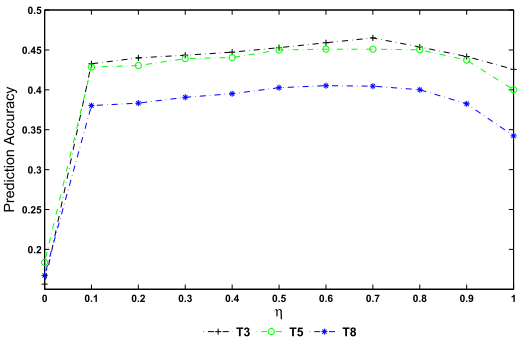


(c) s=20 r=2000

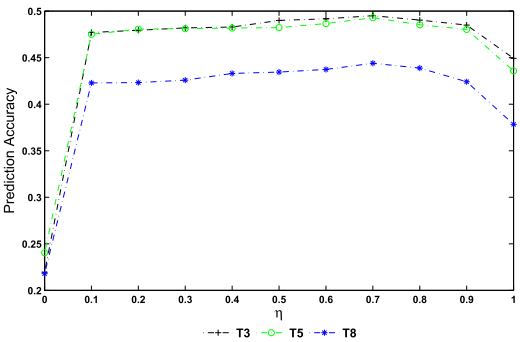
Fig. 2 The prediction accuracy comparison of HM-DST, some social models and Markov model with $s = 0$, $s = 10$ and $s = 20$



(a) s=0 r=100



(b) s=10 r=1000



(c) s=20 r=2000

Fig. 3 The performance changes with η when $s = 0$, $s = 10$ and $s = 20$

ers the dynamic social ties, which obtain a 15% improvement than Markov-MCS model, a 17% improvement over Markov-Cos model and Markov-MI model, and a 20% increase than Markov model. The usages of dynamic social ties and the modified cross-sample entropy could increasingly improve the prediction accuracy. From Fig. 2 (b) and (c), we can see the similar changing trend with (a), but the improvements of predictive performance when $s = 10$ and 20 are smaller than $s = 0$. When increasing the spatial scale, the transitions between different check-in locations might be treated as the same mobility patterns, which could enlarge the movement similarities between some dissimilar friends and block the improvement of the predictive performance when considering the friends' movements. Overall, HM-DST model could provide considerable improvement than Markov model and other social Markov models.

5.3.3 Adjusting the Weight between Historical and Social ties

To investigate the contribution of social ties and historical ties in affecting user's behaviors, we increase the parameter η from 0 to 1 with an increment step of 0.1 and observe the prediction performance at each η. We only show the prediction accuracy at parts $T3$, $T5$ and $T8$ in Fig. 3, since the similar performance can be observed at other parts. When $\eta = 0$, HM-DST model just considers social ties. Its perfor-

mance is always worst, suggesting that considering social information only is not enough to capture the individual's behaviors. By increasing η, the performance first increases to the peak and then decrease, but the variation between 0.1 and 0.9 is small. The prediction accuracy remains stable when η is between 0.5 to 0.8. When $\eta = 1$, HM-DST model converts to the Markov model, which only considers the historical movements. Its performance is not the best and have a 20% decrease than the predictive accuracy of HM-DST model when $s = 0$, which indicates the social ties are also important and beneficial to the location prediction.

6. Conclusions

In this paper, a hybrid Markov location prediction algorithm based on dynamic social ties is presented. Firstly, according to the time division, the movements of the user and his/her friends are segmented to find the changes of the social circle in different time slices. Secondly, the movement similarities between user's historical movements and friends' segmented movements during different time periods are quantified by a modified cross-sample entropy. Finally, the user's historical movement data and his friends' movements at different times which are assigned with the similarity weights are combined to build an HM-DST. The experiments based on Brightkite dataset show the HM-DST model could provide a 15% improvement over the location predictors without considering the dynamic social ties when using check-in locations.

Acknowledgments

This work was supported by the Postgraduate Innovation Project of Jiangsu (CXZZ13_0933) and the Natural Science Foundation of Jiangsu Province, China (BK201302).

References

[1] I.E. Burbey, "Predicting Future Locations and Arrival Times of Individuals," Virginia: Virginia Polytechnic Institute and State University, 2011.
[2] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "WhereNext: a location predictor on trajectory pattern mining," Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.637–646, 2009.
[3] J. McInerney, S. Stein, A. Rogers, and N.R. Jennings, "Breaking the habit: Measuring and predicting departures from routine in individual human mobility," Pervasive and Mobile Computing, vol.9, no.6, pp.808–822, 2013.
[4] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, and P. Thiran, "Where to go from here? Mobility prediction from instantaneous information," Pervasive and Mobile Computing, vol.9, no.6, pp.784–797, 2013.
[5] T.M.T. Do and D. Gatica-Perez, "Where and what: Using smartphones to predict next locations and applications in daily life," Pervasive and Mobile Computing, vol.12, pp.79–91, 2014.
[6] E. Cho, S.A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1082–1090, 2011.
[7] M. De Domenico, A. Lima, and M. Musolesi, "Interdependence and predictability of human mobility and social interactions," Pervasive and Mobile Computing, vol.9, no.6, pp.798–807, 2013.
[8] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," ICWSM, 2012.
[9] D. Kelly, B. Smyth, and B. Caulfield, "Uncovering measurements of social and demographic behavior from smartphone location data," Human-Machine Systems, IEEE Transactions on, vol.43, no.2, pp.188–198, 2013.
[10] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.L. Barabasi, "Human mobility, social ties, and link prediction," Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1100–1108, 2011.
[11] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," ACM Transactions on the Web (TWEB), vol.5, no.1, 2011.
[12] F. Calabrese, G.D. Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, pp.312–317, 2010.
[13] H. Xiong, D. Zhang, D. Zhang, and V. Gauthier, "Predicting mobile phone user locations by exploiting collective behavioral patterns," Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on, pp.164–171, 2012.
[14] J. McInerney, A. Rogers, and N.R. Jennings, "Improving location prediction services for new users with probabilistic latent semantic analysis," Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp.906–910, 2012.
[15] N. Eagle, A.S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," Proceedings of the National Academy of Sciences, vol.106, no.36, pp.15274–15278, 2009.
[16] Brightkite, http://snap.Stanford.Edu/data/loc-brightkite.html.
[17] M. Lin, W.-J. Hsu, and Z.Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp.381–390, 2012.
[18] W. Shi and P. Shang, "Cross-sample entropy statistic as a measure of synchronism and cross-correlation of stock markets," Nonlinear Dynamics, vol.71, no.3, pp.539–554, 2013.
[19] H.-B. Xie, J.-Y. Guo, and Y.-P. Zheng, "Using the modified sample entropy to detect determinism," Physics Letters A, vol.374, no.38, pp.3926–3931, 2010.
[20] http://en.wikipedia.org//wiki//Geographical_distance.
[21] J. Theiler and P.E. Rapp, "Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram," Electroencephalography and clinical Neurophysiology, vol.98, no.3, pp.213–222, 1996.

Wen Li is currently a PhD candidate in the School of Computer Science and Technology, China University of Mining and Technology. The main areas of interests are trajectory data mining and pattern recognition.

Shixiong Xia is a Professor at the School of Computer Science and Technology, China University of Mining and Technology. He has published more than 60 research papers in journals and international conferences. His research interests are wireless sensor networks and intelligent information processing et al.

Feng Liu is an Adjunct Professor at the School of Computer Science and Technology, China University of Mining and Technology. He is also the Secretary General of China National Coal Association. His main research interest is trajectory data mining.

Lei Zhang received his PhD degree in the Department of Computer Application Technology, Nan jing University of Aeronautics and Astronautics in 2006. He is now an associate professor in the School of Computer Science and Technology, China University of Mining and Technology. His current research interests include mobile computing and data mining.