MINING USER MOVEMENT SIMILARITY BASED ON MASSIVE GPS TRAJECTORY DATA WITH TEMPORAL EFFECTS

Hua Yuan
School of Management and Economics
University of Electronic Science and Technology of China
No. 2006, Xiyuan Ave. West Hi-Tech Zone, Chengdu, China
yuanhua@uestc.edu.cn

Lu Feng
School of Management and Economics
University of Electronic Science and Technology of China
No. 2006, Xiyuan Ave. West Hi-Tech Zone, Chengdu, China
fenglu1101@gmail.com

Yu Qian*
School of Management and Economics
University of Electronic Science and Technology of China
No. 2006, Xiyuan Ave. West Hi-Tech Zone, Chengdu, China
qiany@uestc.edu.cn

ABSTRACT

The pervasive use of mobile devices and location-based services has supported the generation of large spatio-temporal datasets reflecting user movement behavior. However, studies on such type of data have depended heavily on geographic overlapping, and information about the time of day of travel visits has been overlooked. In this paper, we proposed an efficient method for mining user movement similarity based on users' travel histories as recorded by GPS trajectories. Our approach also allowed consideration of related temporal effects. To that end, first we introduced a partition method to divide the trajectories into a set of line segments that allowed us to explore the correlation between users and their visited territories. Significantly, we proposed a characteristic point mapping method to transform the sparse GPS trajectories into a set of transactional data. Based on this data, we conducted a series of data mining procedures for efficient discovery of the users' movement information. Second, we proposed a novel, low-rank matrix factorization-based method to cluster users' movements into groups based on their similarity, including temporal characteristics. The experimental results demonstrated that the proposed method can be used to mine the popular roaming routes of users or similar movements efficiently while including temporal patterns. This approach can prove valuable for the development of location-based social network recommendations and human mobility prediction.

Keywords: GPS trajectory; User movement; Matrix factorization; Temporal effect

1. Introduction

As the popularity of smart phones and other GPS-enabled mobile devices continues to grow, location acquisition technologies have become increasingly pervasive, leading to the collection of large spatio-temporal datasets about user movement behavior [Lin et al. 2014; Koutsiouris et al. 2016]. Relatively easy access to large amounts of spatio-temporal data, specifically GPS trajectories, provides an opportunity to discover valuable geographic information concerning individual mobility. In turn, deeper understanding of user movement behavior provides enormous business opportunities with regard to geographic navigation and location-based recommendations [Zheng 2011] by recognizing numerous traffic activities from both the pedestrian side and the transportation side [Liao et al. 2005; Ghourchain 2016].

Recently, user movement similarity has become particularly significant for location-based social network recommendations [Li et al. 2008; Cho et al. 2011] and human mobility prediction [Do et al. 2015], Consequently,

* Corresponding author

research that measures user movement similarity based on travel trajectories has attracted considerable attention [Lv et al. 2013; Chen et al. 2014]. However, for current GPS-enabled applications, collecting information about the similarity of users' trajectories is often difficult and inefficient for the following three reasons:

(1) First, when users turn the GPS-enabled devices on and off casually, the recorded GPS data are often non-uniform, sparse, or lost, and the data collected may be inconsistent with the end-points. Therefore, any two trajectories might not be identical even if they recorded the same path. For instance, see the sample trajectories of TR_1 and TR_2 in Figure 1. They have been used to record the movements of user1 and user2 on the same path. However, the GPS data points in TR_1 and TR_2 are rarely identical.

(2) Second, large amounts of GPS point data are recorded in a trajectory (especially in the case of high recording frequency), but only a few of the data are key to exhibiting interesting geographic information about a user's travel [Zheng 2015]. For example, in Figure 1, the "meaningless roaming" data points on TR_1 may be motion noise involved in user1's movement.

(3) Third, human geographical movement has always exhibited significant temporal characteristics that are strongly related to the locations [Ye 2011]. For example, the intentions of different users to visit the same place may not be the same, so the time of day of their visits may also vary widely.

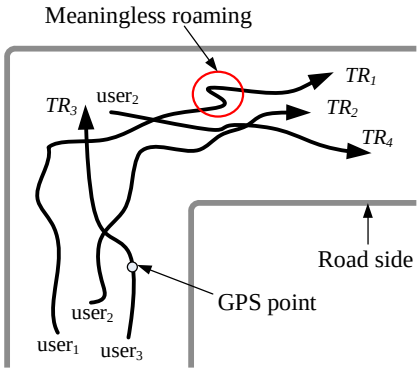
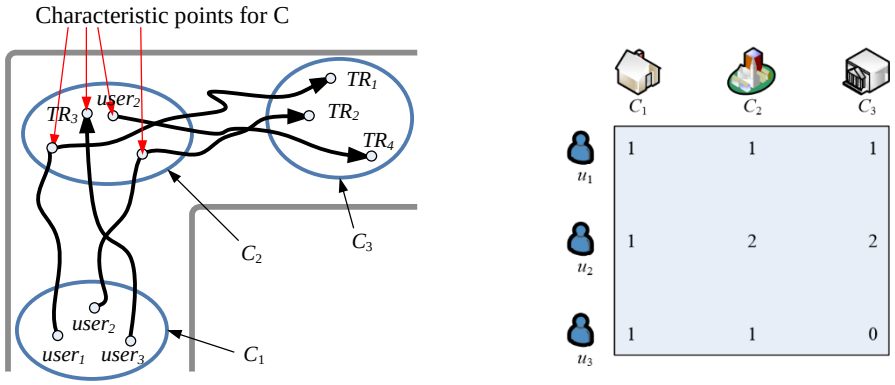


Figure 1: Sample of GPS trajectories

To mine the movement similarity of users efficiently, in this paper, first we introduced an efficient GPS trajectory partition method [Yuan et al. 2014] to trim the sequential GPS data into line segments, taking the key end points as the characteristic points for clustering (See Figure 2(a)). Based on these points, we could map all the trajectories onto a series of abstract trip routes. Next, by taking each of the clusters as a fixed territory (location), a user's trajectory reflected his visits to a certain series of geographic locations. Such a "user-visiting-location" relationship may be represented as a user-location matrix (see Figure 2(b)). Finally, taking temporal effects into consideration, we proposed a novel low-rank matrix factorization based method to solve the problem of mining users' similar movements.



(a) Trajectories partition and endpoints clustering. (b) User-Location matrix.

Figure 2: Method for constructing user-location correlation

The remainder of this paper is organized as follows. Section 2 summarizes significant related work. Section 3 details the novel methods for trajectory partition and fixed territory clustering. Section 4 presents the procedure for mining users' movement similarity with temporal effects. Section 5 shows the experimental results, while Section 6 presents conclusions based on our work.

2. Related Work

2.1. Trajectory information mining

Generally, GPS trajectory data are recorded with very high frequency that provides notably fine-grained information about a movement. However, this capability is also the main reason for data sparsity. Therefore, earlier research in this area attempted to mine path information in a manner that would avoid the data sparsity problem at least partially. Jeung et al. [2008] considered the idea of convoys in trajectory databases, and formalized the concept of a convoy query with density based notions. Han et al. [2012] argued that density and Euclidean distance were no longer effective measures of the utility of spatial clustering of mobile object trajectories. Instead, they proposed a road network-aware approach for fast and effective clustering of spatial trajectories. Sung et al. [2012] presented a clustering method to extract motion patterns from historical data, and used the patterns to generate interception paths.

Trajectory pattern mining, introduced in [Giannotti et al. 2007], has been another important area of research. A trajectory pattern represents a set of individual trajectories that share the property of visiting the same sequence of places with similar travel times. Following this approach, some important efforts were devoted to mining travel sequences [Monreale et al. 2009]. Since trajectory data contains many points with only a small amount of useful information, partitioning the raw trajectory into segments (sub-trajectories), and finding a representative line for each segment, are two feasible methods that help extract significant information, such as characteristic points. Lee et al. [2008] proposed a TraClass method for trajectory data that generated a hierarchy of features by partitioning trajectories and exploring region-based and trajectory-based clustering. Cao et al. [2005] defined the pattern elements as spatial regions around frequent line segments, and the patterns were detected using a substring tree structure. In [Lee et al. 2007], the researchers proposed a partition-and-group framework for clustering trajectories that partitioned a trajectory into a set of line segments. Then they grouped similar line segments together into a cluster. However, this approach suffered from the difficulty of defining the similarity metric for line segments [Sung et al. 2012], which led to complexity in both the mathematical analysis and numerical calculation. To address this problem, Yuan et al. [2014] presented a novel trajectory partition method. The primary advantage of this method was the low computational cost for finding characteristic points from massive trajectories.

2.2. User Movement Similarity Mining

With continuing increased available of personal mobile information, researchers focused extensively on individual location history as represented by GPS trajectories. This range of work included detecting individual locations [Ashbrook and Starner 2003; Hariharan and Toyama 2004], recognizing user-specific activities at each location [Patterson et al. 2003; Pao et al. 2012] to analyze location correlations [Zheng and Xie 2010], and predicting users' movements among these locations [Ashbrook and Starner 2003]. Ultimately, all these research efforts aimed to provide better travel recommendations, a goal that promoted interest in mining data about similar users based on their movements.

To measure user similarity in geographic environments, most of the work in the literature adopted the basic idea of analyzing the movement regularities of mobile users. Zheng et al. [2011] proposed a system for measuring user similarity that first extracted stay points from trajectories and then organized them as a hierarchical framework. Similarity between users was calculated based on the retrieved moving sequences. Lu et al. [2011] proposed a method named LBSAlignment to measure the movement similarity of two mobile users by analyzing the longest common sequence. Thakur et al. [2010] modeled users' visiting histories for various locations with a user-location matrix. The eigen vectors of the matrix were used to measure the similarity of motion of these mobile users. Lin and Su [2008] proposed a simple way to compare spatial shapes of moving object trajectories by introducing a new distance function based on "one way distance" (OWD). Yuan and Raubal [2014] contributed to this research area by developing the Spatio-temporal Edit Distance measure, an extended algorithm to determine the similarity between user trajectories based on detailed call records.

Since data sparsity is a main challenge for these methods, some recent work tried to make use of more geo-social information to measure user similarity at the semantic level [Botzenhardt et al. 2016]. Along this line, Lee and Chung [2011] proposed a method to calculate mobile user similarity using the semantics of the locations they visited, in which, the location semantics were constructed by leveraging social network services. However, Wu et al. [2015] modeled semantic trajectories based on road networks and proposed a Constrained Time-based Common Parts (CTCP) approach to measure the similarity. Ying et al. [2010] proposed a method of MSTP-Similarity to evaluate similarity among users based on their maximal semantic trajectory patterns. In their work, the semantics of the trajectories were

transformed by using a geographic information database. In [Lv et al. 2013], they addressed the problem of mining users' long-term activity similarity based on their trajectories. To that end, they first noted the routine activity from users' daily trajectories. Then user similarity was calculated hierarchically based on the extracted routine activities.

As we have seen, to measure the movement similarity of users, all the above approaches relied too much on individuals' geographic overlapping. Moreover, the temporal information was ignored.

3. Trajectory Partitioning and Characteristic Points Clustering

3.1. Trajectory Partitioning

Let g_j denote the $j\text{-}th$ GPS point. Then a series of time-ordered GPS points is given as

$$TR = \{g_1, g_2, \ldots, g_j, g_{j+1}, \ldots, g_n\} \quad (1)$$

which represents a trajectory *TR* consisting of *n* GPS points, implying the movement of an object from g_1 to g_n via GPS points $g_2,\ldots,g_{n-1}$. As suggested in [Lee et al. 2007; Yuan et al. 2014], partitioning a trajectory TR into a set of stationary sub-trajectories (SSTs) is a feasible way to discover characteristic points. The two end-points of each SST can be collected as two characteristic points of TR because the movement of an object in an SST is relatively stable. In other words, an SST can be as a direct path (i.e., a line segment) between its two end-points.

Given a data set of N trajectories, using the method presented in [Yuan et al. 2014], we can partition each $TR_i = \{g_{1_i}, \ldots, g_{n_i}\}$ ($i = 1,\ldots,N$) into m_i SSTs as shown in Table 1. *SST*(*x*, *y*) means the SST has two GPS data points, *x* and *y*, as its end-points. The adjacent two SSTs may have one common end-point.

Table 1: Partition TR_i into SSTs

TR_i	SST
1	$SST\left(g_{1_1}, g_{c_{11}}\right), \ldots, SST\left(g_{c_{1(m_1-1)}}, g_{n_1}\right)$
$\ldots$ i	$SST\left(g_{1_i}, g_{c_{i1}}\right), \ldots, SST\left(g_{c_{i(m_i-1)}}, g_{n_i}\right)$
$\ldots$ N	$SST\left(g_{1_N}, g_{c_{N1}}\right), \ldots, SST\left(g_{c_{N(m_N-1)}}, g_{n_N}\right)$

As a result, we can obtain further a set of characteristic points as $CP(TR_i)$ from each trajectory of TR_i (Table 2).

Table 2: Characteristic points of trajectories.

TR_i	$CP(TR_i)$
1	$g_{1_1} - g_{c_{11}} - \cdots - g_{c_{1(m_1-1)}} - g_{n_1}$
$\ldots$ i	$g_{1_i} - g_{c_{i1}} - \cdots - g_{c_{i(m_i-1)}} - g_{n_i}$
$\ldots$ N	$g_{1_N} - g_{c_{N1}} - \cdots - g_{c_{N(m_N-1)}} - g_{n_N}$

As we can see, it would be simpler and easier to extract key information from a group of SSTs than from the trajectories directly. The significant effect of an SST is that it removes the noise points while retaining the key GPS points for identifying characteristic points, providing great convenience in dealing with the geographical relationship between GPS points.

3.2. Characteristic Points Clustering

The information contained in each end-point is still very limited, because different users would generate very different characteristic points even if they walked on the same path.

In the following Algorithm 1, we introduce the general *k*-means method to cluster the characteristic points into *l* clusters, i.e., $C_1,\ldots,C_l$, based on their geographic closeness.

Algorithm 1 Cluster the Characteristic Points

1: **Input**: **TR** = $\bigcup_{i=1}^{N} TR_i$;
2: **Output**: Cluster results **C**;
3: **for** i = 1 to N **do**
4: Partition TR_i into SSTs;
5: Collect end-points of SSTs to generate the set of $CP(TR_i)$;
6: **end for**
7: Cluster the points in $\bigcup_{i=1}^{N} CP(TR_i)$ into l clusters: $C_1,...,C_l$;
8: **return C** = $\bigcup_{k=1}^{l} c_k$.

Gathering the geographically adjacent characteristic points together, the clustered point can be assumed to represent a latent location or POI (point of interest). Algorithm 1 shows the general framework of mining the clusters of characteristic points in a given GPS trajectory data set. First, each of the GPS trajectories is partitioned into segments to obtain its characteristic points (Lines 3-6). Then, all the characteristic points are clustered into l groups $C_1,...,C_l$ (Line 7).

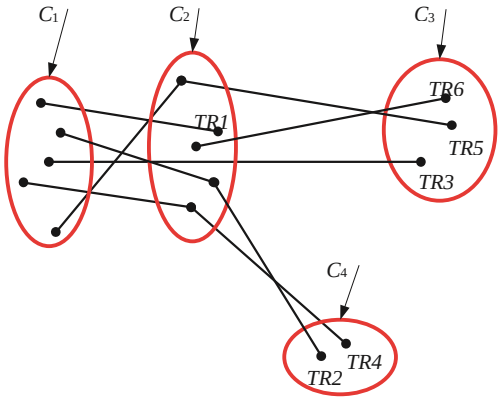


Figure 3: Characteristic points clustering

Figure 3 shows the *characteristic points* of six trajectories ($TR_1,...,TR_6$) clustered into four groups: C_1, C_2, C_3 and C_4.

4. Mining User Movement Similarity with Temporal Effect

4.1. User-location Matrix

In order to study the movement similarity of users, we need to make clear the location-visiting similarity at each movement (represented by a GPS trajectory) of all the users.

We use m to denote the number of observed users, and l to denote the number of clustered locations generated by Algorithm 1. Let $u = \{u_1,u_2,...,u_m\}$ be the set of users, and $c = \{c_1,c_2,...,c_l\}$ be the set of visited locations, respectively. $X \in \mathbb{R}_+^{m \times l}$ is a user-location matrix with each element X_{ij} representing the number of visits made by user ui at location c_j. The following Algorithm 2 shows the generation of X_{ij}.

4.2. Research Problem

Let $U \in \mathbb{R}_+^{m \times d}$ be the user check-in preferences for all the locations, and let $C \in \mathbb{R}_+^{l \times d}$ be the location characteristics, where $d \ll \min(m, n)$ denotes the number of latent preference factors of users.

Algorithm 2 Generate the User-location Matrix

Input: **CP** = $\bigcup_{i=1}^{N} CP(TR_i)$; $c = \bigcup_{k=1}^{l} c_k$.
Output: $\boldsymbol{X}$;
$\boldsymbol{X} = 0$;
for $i = 1$ to N **do**
    Seek out the provider index, i.e., u_s, of TR_i;
    for $j = 1$ to $|CP(TR_i)|$ **do**
        Obtain the j-th characteristic point g_{cj} in $CP(TR_i)$;
        if $g_{cj} \in c_k$ **then**
            $X_{sk} = X_{sk} + 1$;
        end if
    end for
end for
return $\boldsymbol{X}$.

With the notations introduced above, we then define the problem of mining user movement similarity as a co-cluster problem using a constrained nonnegative matrix factorization:

$$min_{U,C\geq 0}\|X - UC^T\|_F^2, \tag{2}$$

where the operational symbol $\|\cdot\|$ denotes the Frobenuis norm of a matrix. By solving the above optimization problem, the result can be used to approximate the check-in preference of u_i on an unvisited location c_j. Basically, this result can be useful for studying the movement similarity of users based on their location preference.

4.3. Temporal Effect Based Regularization

Human geographical movement exhibits significant temporal patterns, and is highly relevant to the location property. Thus, investigating the temporal features embedded in daily patterns provides an opportunity for us to improve our understanding of human mobile behavior.

To model the correlation of users' movements with temporal effects, we constructed a user-user similarity graph in which the graph nodes represent users, and the edges represent the affinity between the movements of users with temporal effect. The adjacency matrix $W^u \in \mathbb{R}_+^{m\times m}$ of the graph is defined as in [Hu et al. 2013]:

$$W_{ij}^U = \begin{cases} 1 & if\ u_j \in \mathcal{N}(u_i)\ or\ u_i \in \mathcal{N}(u_j) \\ 0 & otherwise \end{cases} \tag{3}$$

where $\mathcal{N}(u_i)$ denotes the k-nearest neighbors (KNN) of node u_i.

In order to adopt the KNN method, we proposed the following measurement to evaluate the intensity of users' movement similarity with the constraint of temporal effect. First, in a time interval of [0, T], a user's visiting history for all the locations can be distributed as a "time-location" matrix, i.e., X_{TL}. Therefore, given two users $u_i, u_j \in u$, if they have visited the same locations at the same (or similar) times with high frequency, then their affinity will be high. Based on this assumption, we used the following method to measure the similarity between two users u_i and u_j:

$$sim(u_i, u_j) = e^{-\frac{\left\|x_{TL}^i - x_{TL}^j\right\|_F^2}{\left\|x_{TL}^i\right\|_F^2 \left\|x_{TL}^j\right\|_F^2}} \tag{4}$$

The key idea is that, if two users are close in the graph, their movement affinity is also close. This result can be achieved by minimizing the following loss function:

$$\min tr(U^T L^u U) \tag{5}$$

where tr(·) denotes the trace of a matrix. $L^u = D^u - W^u$ is the Laplacian matrix of the user-user similarity graph and the diagonal matrix D^u is the degree matrix of W^u, i.e., $d_i^u = \sum_j W_{ij}^u$.

By approximating the visiting activities for each temporal state $t \in [0, T]$ and minimizing their aggregation, the time-dependent user visiting preferences can be obtained by solving the following optimization problem:

$$min_{U,C\geq 0}\|X - UC^T\|_F^2 + \lambda tr(U^T L^u U) + \alpha\|U\|_F^2 + \beta\|C\|_F^2, \quad (6)$$

where λ is a parameter to control the temporal regulation.

However, the optimization problem in relation (6) is not convex with respect to all the two variables of $\boldsymbol{U}$ and $\boldsymbol{C}$, i.e., there is no closed-form solution for the problem. Next, we introduced an iterative update algorithm to solve the optimization problem based on the work in [Ding et al. 2006].

4.4. Learning Algorithm

Let Λ_u and Λ_c be the Lagrange multiplier for constraint $\boldsymbol{U} \geq 0$ and $\boldsymbol{C} \geq 0$ respectively. Then the Lagrange function L is defined as follows:

$$L = tr[(X - UC^T)^T(X - UC^T)] + \lambda tr(U^T L^u U) + \alpha tr(U^T U) + \beta tr(C^T C) - tr(\Lambda_u U^T) - tr(\Lambda_c C^T) \quad (7)$$

By setting the derivatives $\nabla_U L = 0$ and $\nabla_C L = 0$, we get the following:

$$\Lambda_u = -2XC + 2UC^T C + 2\lambda(D^u - W^u)U + 2\alpha U \quad (8)$$

$$\Lambda_c = -2XU^T + 2CU^T U + 2\beta C \quad (9)$$

The Karush-Kuhn-Tucker complementary condition [Boyd and Vandenberghe 2004] for the non-negativity constraint of Λ_u and Λ_c gives $\Lambda_u(i,j)U(i,j) = 0$ and $\Lambda_u(i,j)C(i,j) = 0$. Thus, we can obtain the updating rules for $\boldsymbol{U}$ and $\boldsymbol{C}$ as follows:

$$U(i,j) \leftarrow U(i,j)\sqrt{\frac{[XC+\lambda W^u U](i,j)}{[2UC^T C+\lambda D^u U+\alpha U](i,j)}}; \quad (10)$$

$$C(i,j) \leftarrow C(i,j)\sqrt{\frac{[X^T U](i,j)}{[CU^T U+\beta C](i,j)}}. \quad (11)$$

In summary, we present the following Algorithm 3 for solving the optimization problem of (6).

Algorithm 3 Explore users having similar movements with temporal effect.

1: **Input**: user-location matrix $\boldsymbol{X}$; $\cup_{i=1}^{n} X_{TL}^{i}$; parameter λ, α, β;
2: **Output**: user cluster results U;
3: Construct matrix $\boldsymbol{L}^u$ in relation (5);
4: Initialize $\boldsymbol{U},\boldsymbol{C} \geq 0$;
5: **while** Not convergent **do**
6: Update $\boldsymbol{U}(i,j)$ according to relation (10);
7: Update $\boldsymbol{C}(i,j)$ according to relation (11);
8: **end while**
9: **return** $\boldsymbol{U}$.

In the above computational algorithm, initialization for the matrices and Laplacian matrix can be inferred from lines 3-4. The two matrices are updated with their respective updating rules (lines 5-8), and the iterative process would be stopped if the value of these matrices converged, or if the number of iterations exceeded a given threshold.

5. Experiments

5.1. Experiment Setup

In this section, we evaluated the performance of our proposed framework for mining users' movement similarity from massive GPS trajectories. In particular, we evaluated ability to: (1) mine characteristic points from GPS data, and (2) identify users' movement similarity with temporal effects. Before we delve into the details of the experiments, first we will describe the dataset used in this paper.

5.1.1. Data Set

We tested our method using the BJ data set, a real GPS trajectory dataset collected by the Geolife project conducted by Microsoft Research Asia in Beijing1. This data set was generated by 182 users in a period of over three years (from April, 2007, to August, 2012). It contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. All the trajectories are logged in a dense representation, e.g., every 5 seconds or every 20 meters per point. In the following work, 706 trajectories were chosen randomly for the experiments.

[1] http://research.microsoft.com/en-us/projects/urbancomputing/.

Note that the original GPS data adopted the WGS-84 coordinate system. In our experiments, all the GPS point values were translated into Beijing-54 coordinates to facilitate the computation of Euclidean distance. Since the city of Beijing is a relatively small area compared to the size of the Earth's surface, we regarded the translation error for each GPS data point to be negligible.

5.1.2. GPS Data Preprocessing

Obviously, in a massive data environment, a fast and accurate SST partitioning method is important for studying GPS trajectories. As mentioned previously in this paper, Yuan et al. [2014] presented a novel trajectory partition method to obtain the stationary sub-trajectory (SST) from massive trajectory data. The primary advantage of the method was the low computational cost, and the fact that it could be used to partition any number of trajectories with different forms of start movement (start position, speed, direction, and so on). In our research, we conducted a set of experiments to exhibit the advantages of Yuan's work (denoted by Yuan) on GPS data preprocessing. This step was very important because the common trajectories collected by devices are non-uniform and sparse.

The experimental results of the efficiency of this method are shown in Table 3, which exhibits the following:

(1) The efficiency of the selected method was about 10 times faster than that of the method presented by Lee et al. (denoted by Lee).
(2) The computation became faster with the increasing value of d_0.
(3) The curves in Figure 4 were almost straight lines, providing experimental evidence that both methods have computation complexity of $O(n)$, where n is the number of GPS points in a trajectory.

Table 3: Time consumption for the trajectory partition methods (d_0=1).

TR length (n)	Yuan	Lee	Ratio
10	0.0013495	0.0049614	3.67
100	0.005776	0.0637	11.02
300	0.013291	0.17376	13.07
600	0.027756	0.38121	13.73
1000	0.063317	0.61054	9.64

Changing the value of the parameter d_0, the comparative results of algorithm efficiency are shown in Figure 4. As we can see, when the value of d_0 changed to a bigger value, the Yuan's method still had a very high computational efficiency. Moreover, in the experiments, the average number of iterations for partitioning the trajectories in the BJ data set was 3. The minimum number of iterations was about 1 when the GPS trajectory was generated in a direct path (very low position disturbance), whereas the maximum was about 8 for some irregular trajectories (high position disturbance) that contained more than 20,000 GPS points.

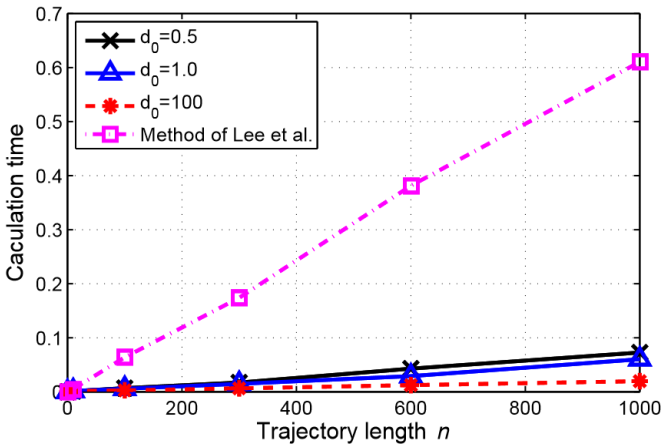


Figure 4: Comparison of Efficiency

In addition, Lee et al. [2007] presented two measurements of preciseness and conciseness to evaluate the optimal partitioning of a trajectory. Preciseness means that the difference between a trajectory and a set of its trajectory partitions should be as small as possible. Conciseness means that the number of trajectory partitions should be as small as possible. Obviously, the method presented by Yuan et al. [2014] has a very high performance on conciseness,

especially when d0 is big. Thus, our other experiment in data preprocessing concerned the preciseness of the trajectory partition method. In Figure 5, the black dashed line represents the base line of raw trajectory, which keeps the information of all the GPS points in a trajectory. The pink line illustrates the number of characteristic points generated by the method presented by Lee et al. [2007], while the blue line shows the number of characteristic points generated by the approach provided by Yuan et al. [2014]. The red line represents the number of characteristic points found by these two methods commonly. The experimental results showed clearly:

(1) The preciseness of both methods decreased with the increase of trajectory length, *n*. The method presented by Lee et al. [2007] had an advantage in preciseness because it kept more GPS points as characteristic points.
(2) With respect to the base line, the preciseness of the method presented by Yuan et al. [2014] was acceptable. If we wanted more conciseness, then a relative smaller value of d_0 would be needed.

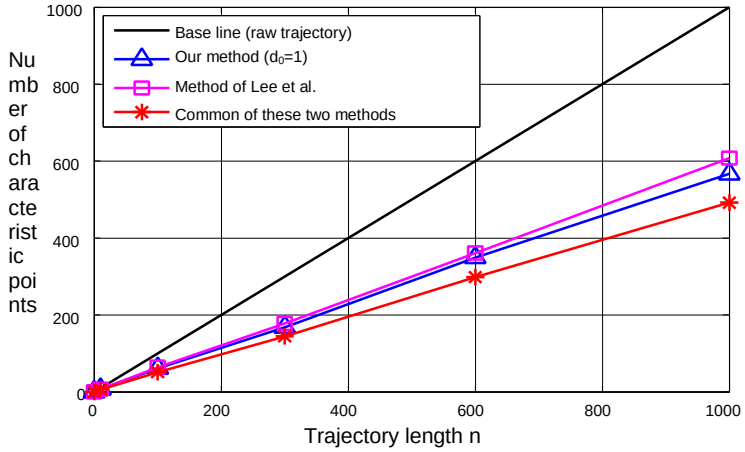


Figure 5: Comparison of Preciseness

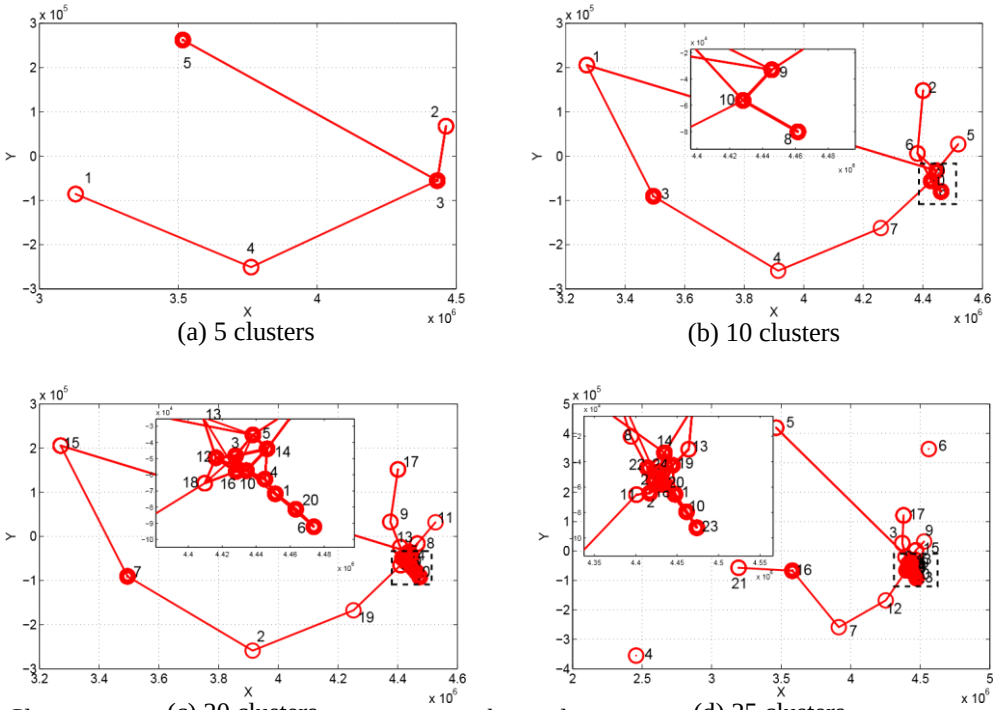
In summary, these results demonstrated that the method presented by Yuan et al [2014] had an advantage in computing speed, which became more significant when we were faced with a large amount of GPS data. In contrast, the method presented by Lee et al. [2007] kept more preciseness (information) about the raw trajectory.

5.2. Characteristic Points Clustering

In this experiment, we first clustered the characteristic points into groups (d_0=100), and then we generated a spatial map of users' typical motions with the trajectory coverage relation between groups, i.e., fixed territories.

5.2.1. Clustering Characteristic Points into Groups

In the clustering process, an important challenge was to set an appropriate number of clusters. For example, using too few clusters would result in a large number of merged network nodes. Some characteristic points in these trajectories were presented by a single node because of their relatively shorter trajectory length. Therefore, much useful geographic information would be concealed (See Figures 6(a) and 6(b)). In contrast, use of a large number of clusters would result in lack of discrimination between clusters. For instance, some clusters in Figure 6(d) are crowded, so it is difficult to distinguish them from each other.



(a) 5 clusters (b) 10 clusters

(c) 20 clusters (d) 25 clusters

Figure 6: Cluster number and geographical information

Since clustering is the task of grouping a set of objects so that objects in the same group are more similar to each other than to objects in other groups, we introduced the following exploratory method to cluster the end-points into groups:

(1) First, we clustered the characteristic points into 2 to *k* groups respectively.
(2) Then, we calculated the average distance between clusters.
(3) Finally, we chose the clustering results that had both bigger cluster distances and larger cluster numbers as the most suitable for the studied trajectories.

Considering the instability of clustering methods, we used the averaged results after repeating the calculation 10 times. Figure 7 reflects the impact of the cluster number on clustering results for some sample trajectories in the BJ data set.

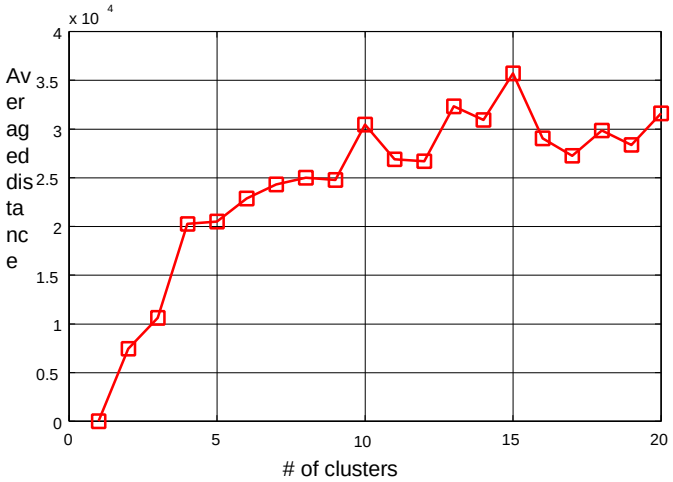


Figure 7: Impact of the cluster number on clustering results

5.2.2. Mapping Trajectory onto Locations

Following on the steps described above, we now had the original trajectories provided by users, along with some clusters of characteristic points that could be deemed as latent locations. To study the users' movements, we needed to explore the correlations between each of the user-generated trajectories and the locations.

Based on the clustering results, if a characteristic point g_{ij} of a trajectory TR_i belonged to a cluster c_j, then we said that the location represented by c_j had been covered by TR_i. Obviously, we can map a trajectory onto a series of sequential locations. Assuming a user's movement from one location to another is a path (or sub-path), we could construct a path network for all the trajectories. Figure 8 demonstrates a local path network generated by the sample trajectories. It is an abstract map of the typical movements for a group of users.

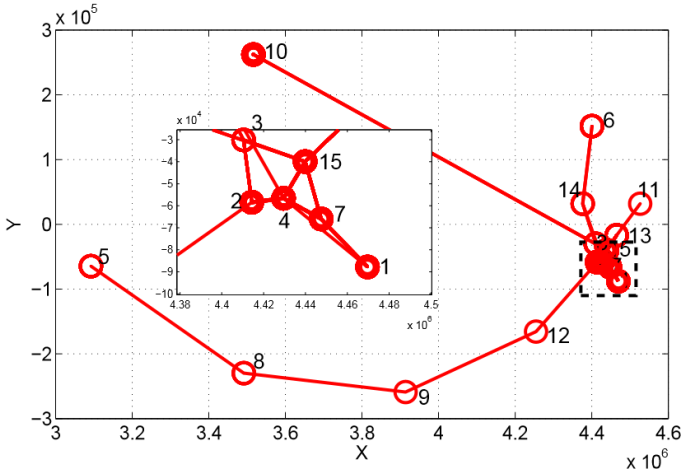


Figure 8: Path networks for the characteristic points.

Importantly, following this approach, we could establish two key relationships: "user u_i-providing-trajectory TR_i" and "trajectory TR_i-covering-location c_j." As such, we obtained the final information about "u_i-visiting-location c_j" which contained valuable information about the correlation of travel movements among fixed territories. In turn, this information could can be transformed easily into a "user-location" matrix. This step was necessary for studying the users' movement similarity and detecting frequent paths in an area for better personalized recommendations.

5.3. Mining User Movement Similarity with Temporal Effects

In this portion of our research, we compared the performance of our method (NMF) with the classic K-means model. To this end, five groups of users with similar movement were annotated manually as ground truth. For each group, we defined an area (a subset of the 100 locations) that the group members visited frequently at almost same time, so that users in the same group would have the same or similar characteristics in terms of their daily movements.

Purity was selected to assess the performance of the proposed methods. Following [Hu et al. 2013], purity was calculated as the weighted sum of individual cluster purity values, as shown:

$$purity = \frac{1}{n}\sum_{i=1}^{k} max_j \left|C_j \cap l_j\right| \tag{12}$$

where $\{l_1,...,l_j,...\}$ is the ground truth, k is the number of clusters and n is the total number of points. Purity measures the extent to which each cluster contains data points from one class, so the higher the purity, the better the clustering result.

For general experimental purposes, a total of 950 users (150 labeled manually) were selected to compare the similarity of their trajectories at 10, 20, and 30 days, respectively. The parameters of λ, α, and β in matrix factorization were set empirically as $\lambda = \alpha = \beta = 1$. In addition, parameter $K = 20$ for the K-nearest neighbor method was defined as in equation (3) and the number of clusters for the k-means method was set at 10. The experiments for each method were repeated 5 times, with the results as follows:

Table 4: Comparison of purity.

Experiment	NMF			k-means		
	Day=10	Day=20	Day=30	Day=10	Day=20	Day=30
1	0.3485	0.4862	0.6810	0.5076	0.2936	0.4310
2	0.3309	0.3910	0.4488	0.4245	0.3609	0.3622
3	0.4159	0.4504	0.5439	0.5664	0.3741	0.3596
4	0.3223	0.5338	0.6260	0.4050	0.2556	0.3496
5	0.2800	0.4621	0.5424	0.2880	0.4318	0.4492

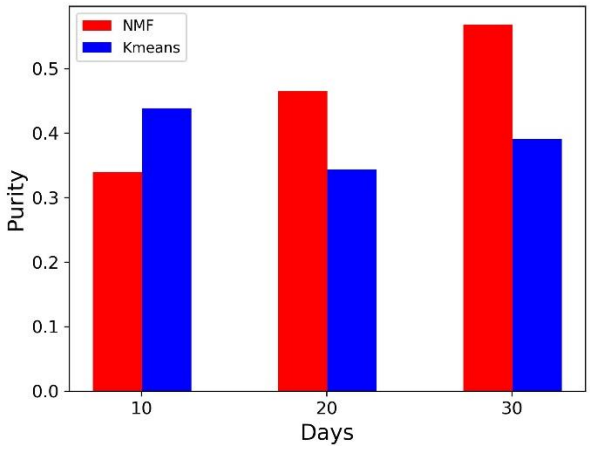


Figure 9: The averaged purity

The average performance of NMF and k-means is shown in Figure 9. When the GPS data covered only 10 days of users' movements, we can see that k-means demonstrated relatively better efficiency. This finding may have resulted because the similarity relationship of users (visiting locations at each time interval) in NMF had not been established effectively. However, as the time (number of days) in the experiment increased, NMF gradually showed a greater advantage for mining the data of users' movement similarity with consideration of temporal components.

6. Conclusion

In this work, we attempted to mine users' movement similarity by leveraging their GPS trajectory data. However, the user-location-time relationship was very sparse in raw GPS data. To address this problem, we proposed a novel method by combining together GPS data sequence partitioning and matrix factorization to achieve the objectives.

First, we introduced an efficient trajectory partitioning method to trim the sequential GPS data into line segments. Next, taking these end-points as the characteristic points, we used a clustering algorithm to deduce the fixed territories (locations) information covered by the trajectories. This method made use of SST trajectory partitioning and clustering of characteristic points to eliminate the data sparsity caused by the raw GPS data, so that the correlation between trajectories could be studied efficiently. More importantly, we constructed a "user-location" matrix (the elements in the matrix implied "which user accessed what location") to represent the users' preferences in terms of visited locations.

Second, we leveraged the timestamp information associated with each GPS point to extend the initial "user-location" relationship to the "which user at what time visiting what location." As a result, the spatio-temporal profile of each user could be built. Next, we put the work of mining users' movement similarity into an equivalent task of clustering the users into groups based on their location visiting history and temporal characteristics. The clustering process was realized by a new proposed matrix factorization model. The experimental results showed that users with similar movement could be discriminated on the basis of our method, and thus demonstrated the effectiveness of our approach.

The methods proposed in this work are efficient for mining information hidden in the trajectory data, especially the frequent paths, fixed territories, and movement intentions. This information can prove valuable for the

development of location-based social network recommendations and human mobility prediction. In turn, this approach can provide businesses with opportunities for geographic information services.

Acknowledgment

The authors would like to thank the supports of the National Natural Science Foundation of China (Nos.: 71671027/71572029/71490723/71271044).

REFERENCES

Ashbrook D., and T. Starner, "Using gps to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*,7(5): 275–286, 2003.

Botzenhardt Achim, Ye Li, and Alexander Maedche, "The Roles of Form and Function in Utilitarian Mobile Data Service Design," *Journal of Electronic Commerce Research*, Vol. 17, No.3, 220-238, 2016.

Boyd S., and L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, USA, 2004.

Cao H., N. Mamoulis, and D. W. Cheung, "Mining frequent spatio-temporal sequential patterns," in: Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM'05, p. 82–89, 2005.

Chen X., P. Kordy, R. Lu, and J. Pang, "MinUS: Mining User Similarity with Trajectory Patterns", in: Calders T., Esposito F., Hüllermeier E., Meo R. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD, p. 436-439, 2014.

Cho E., S. A. Myers, and J. Leskovec, "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, KDD'11, p. 1082–1090, 2011.

Do T. M. T., O. Dousse, M. Miettinen, and D. Gatica-Perez, "A probabilistic kernel method for human mobility prediction with smartphones," *Pervasive and Mobile Computing*, p. 13 – 28, 2015.

Ding C., T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'06, p. 126–135, 2006.

Giannotti F., M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'07, p. 330–339, 2007.

Ghourchian N., "Location-based activity recognition with hierarchical dirichlet process," in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, p. 3990–3991, 2016.

Han B., L. Liu, and E. Omiecinski, "Neat: Road network aware trajectory clustering," in: Proceedings of the 2012 IEEE 32nd International Conference on Distributed Computing Systems, ICDCS'12, IEEE Computer Society, p. 142–151, 2012.

Hariharan R., and K. Toyama, "Project lachesis: Parsing and modeling location histories," in: M. J. Egenhofer, C. Freksa, H. J. Miller (Eds.), GIScience, Vol. 3234 of Lecture Notes in Computer Science, Springer, p. 106–124, 2004.

Hu X., J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in: Proceedings of the 22nd International Conference on World Wide Web, WWW'13, p. 607–618, 2013.

Jeung H., M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen, "Discovery of convoys in trajectory databases," in Proceedings of the VLDB Endowment, 1(1): 1068–1080, 2008.

Koutsiouris Vasilios, Adam Vrechopoulos, and Georgios Doukidis, "Classifying, Profiling and Predicting User Behavior in the Context of Location Based Services." *Journal of Electronic Commerce Research*, Vol. 17, No.4: 340-357, 2016.

Lee J.-G., J. Han, X. Li, and H. Gonzalez, "Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering," Proceedings of the VLDB Endowment, 1(1): 1081–1094, 2008.

Lee J.-G., J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, p. 593–604, 2007.

Lee M.-J., and C.-W. Chung, "A user similarity calculation based on the location for social network services," in: Proceedings of 16th International Conference on Database Systems for Advanced Applications, DASFAA Part I, p. 38–52, 2011.

Li Q., Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems, GIS'08, ACM, p. 34:1–34:10, 2008.

Lin B., and J. Su, "One way distance: For shape based similarity search of moving object trajectories," *GeoInformatica*, 12(2):117–142, 2008.

Lin M. and W. J. Hsu, “Mining gps data for mobility patterns: A survey,” Pervasive and Mobile Computing 12, p. 1-16, 2014.

Liao L., D. Fox, and H. Kautz, “Location-based activity recognition using relational markov networks,” in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI’05, p. 773–778, 2005.

Lu E. H. C., V. S. Tseng, and P. S. Yu, “Mining cluster-based temporal mobile sequential patterns in location-based service environments,” *IEEE Transactions on Knowledge and Data Engineering*, 23 (6):914–927, 2011.

Lv M., L. Chen, and G. Chen, “Mining user similarity based on routine activities,” *Information Sciences*, 236: 17-32, 2013.

Monreale A., F. Pinelli, R. Trasarti, and F. Giannotti, “Where next: a location predictor on trajectory pattern mining,” in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD’09, p. 637–646, 2009.

Patterson D. J., L. Liao, D. Fox, and H. A. Kautz, “Inferring high-level behavior from low-level sensors,” in: A. K. Dey, A. Schmidt, J. F. McCarthy (Eds.), Ubicomp, Vol. 2864 of Lecture Notes in Computer Science, Springer, p. 73–89, 2003.

Pao H.-K., J. Fadlil, H.-Y. Lin, and K.-T. Chen, “Mining frequent trajectory pattern based on vague space partition,” *Knowledge-Based Systems*, 34:81–90, 2012.

Sung C., D. Feldman, and D. Rus, “Trajectory clustering for motion prediction,” in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, p. 1547–1552, 2012.

Thakur G. S., A. Helmy, and W.-J. Hsu, “Similarity analysis and modeling in mobile societies: The missing link,” in: Proceedings of the 5th ACM Workshop on Challenged Networks, CHANTS’10, p. 13–20, 2010.

Wu X., Y. Zhu, S. Xiong, Y. Peng, and Z. Peng, “A new similarity measure between semantic trajectories based on road networks,” in: Web Technologies and Applications: 17th Asia-Pacific Web Conference, APWeb 2015, p. 522–535, 2015.

Ye M., K. Janowicz, C. Mulligann, and W. C. Lee, “What you are is when you are: The temporal dimension of feature types in location-based social networks,” in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS’11, p. 102–111, 2011.

Ying J. J.-C., E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng, “Mining user similarity from semantic trajectories,” in: Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN’10, p. 19–26, 2010.

Yuan H., Y. Qian, R. Yang, and M. Ren, “Human mobility discovering and movement intention detection with gps trajectories,” *Decision Support Systems*, 63: 39–51, 2014.

Yuan Y., and M. Raubal, “Measuring similarity of mobile phone user trajectories a Spatio-temporal edit distance method,” *International Journal of Geographical Information Science*, 28 (3):496–520, 2014.

Zheng Y., “Trajectory data mining: An overview,” *ACM Trans. Intell. Syst. Technol*., 6 (3):29:1–29:41, 2015.

Zheng Y., X. Z. (Eds.), Computing with Spatial Trajectories, Springer, Berlin, 2011.

Zheng Y., and X. Xie, “Learning location correlation from gps trajectories,” in: Proceedings of the 2010 Eleventh International Conference on Mobile Data Management, MDM’10, p. 27–32, 2010.

Zheng Y., L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, “Recommending friends and locations based on individual location history,” *ACM Trans. Web*,5(1) 5:1–5:44, 2011.