# Finding Similar Users Using Category-Based Location History

Xiangye Xiao[1,2], Yu Zheng[1], Qiong Luo[2], Xing Xie[1]

1Microsoft Research Asia, 4F Sigma Building, No.49 Zhichun Road, Haidian District, Beijing 100190, China
2Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

xiangye.xiao@gmail.com, {yuzheng, xingx}@microsoft.com, luo@cse.ust.hk

## ABSTRACT

In this paper, we aim to estimate the similarity between users according to their GPS trajectories. Our approach first models a user's GPS trajectories with a semantic location history (SLH), e.g., *shopping malls* → *restaurants* → *cinemas*. Then, we measure the similarity between different users' SLHs by using our maximal travel match (MTM) algorithm. The advantage of our approach lies in two aspects. First, SLH carries more semantic meanings of a user's interests beyond low-level geographic positions. Second, our approach can estimate the similarity between two users without overlaps in the geographic spaces, e.g., people living in different cities. We evaluate our method based on a real-world GPS dataset collected by 109 users in a period of 1 year. As a result, SLH-MTM outperforms the related works [4].

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - *data mining, Spatial databases and GIS.*

## General Terms

Algorithms, Measurement, Experimentation.

## Keywords

User similarity, location history, GPS trajectory, MTM.

## 1. INTRODUCTION

In recently years, an increasing number of people start using GPS-enabled devices to log their outdoor movements with GPS trajectories [4][8][10][12]. These trajectories do not only record users' location histories in the physical world but also imply their personal interests and preferences [2~13]. Figure 1 demonstrates the mobility of four individuals (*A*, *B*, *C*, and *D*) who respectively recorded a one-day trip with a GPS trajectory. According to the outdoor movement, we can observer the following three insights:

1) *Geographic overlaps*: People having similar outdoor location histories in the geographic spaces could share some similar life interests. For instance, the users *A* and *B* might share some similar interests as both of them have visited the same cinema, museum, coffee shop, and shopping mall.
2) *Semantic overlaps*: People could share some similar interests if they have similar mobility patterns in the space of semantic locations. For example, though the user *C* does not access the same locations with *B*, the semantic meanings (categories) of the locations (museum, cinema and coffee) are the same with that of *B*. That is, they could still share similar interests.
3) *Location sequence*: Although the user *D* also visited a museum, a coffee shop, a shopping mall, and a cinema, the sequence between these locations is different from that of the users *B* and *C*. Thus, the similarity between *D* and *B* might not be as significant as that between *C* and *B*.

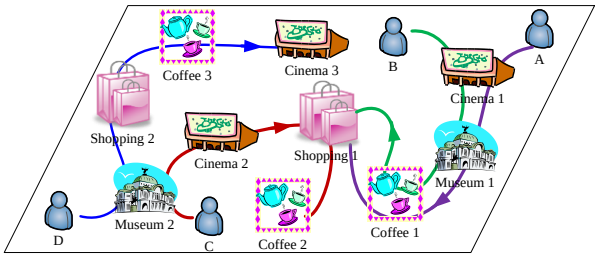



**Figure 1 GPS trajectories and user interests**

In this paper, we aim to estimate the similarity between users according to the semantic location histories (SLH) inferred from their GPS trajectories. This similarity can enable friend and location recommendation, and bridge the gap between the physical world with online social networks. For instance, as shown in Figure 1, if knowing the users *B* and *C* are similar according to their location histories, we can recommend the museum 1 and cinema 1 to the user *C*, and provide the user *B* with the museum 2 and cinema 2 as a recommendation.

The two essential steps of finding similar users are 1) modeling users' interests from their historical GPS trajectories and 2) measuring the similarity between them based on their location histories. To address these problems, we first construct each user a SLH based on their historical GPS trajectories. Then, we compute the similarity between different users in terms of their SLHs, considering the sequence, granularity and popularity features mentioned above. The contributions of our work include:

1) The SLH models a user's interests and the uncertainty of the semantic meanings of a place where a user stayed.
2) The MTM algorithm finds out the maximal subsequence matches (between two sequences) by considering both the visiting order and travel time between two locations.
3) We evaluated our approach using a real-world GPS data collected by 109 users over a year. The dataset is released to the public [1][5][10].


GIS '10, 03-NOV-2010, San Jose CA, USA

## 2. PRELIMINARY

**Definition 1** (*GPS Trajectory*) A GPS trajectory *Tra* is a sequence of time-stamped points, $Tra = p_0 \rightarrow p_1 \rightarrow, \ldots, \rightarrow p_k$, where $p_i = (x, y, t)$ $(i = 0,1, \ldots, k)$; $(x, y)$ are latitude and longitude respectively, and $t$ is a timestamp. $\forall 0 \leq i \leq k,\ p_{i+1}.t > p_i.t$.

**Definition 2 (***Stay Point*) A stay point $s$ is a geographical region where a user stayed over a time threshold $\theta_t$ within a distance threshold $\theta_d$. In a trajectory, $s$ is characterized by a set of consecutive points $P = \langle p_m, p_{m+1}, \ldots, p_n \rangle$, where $\forall m < i \leq n$, $Dist(p_m, p_i) \leq \theta_d$, $Dist(p_m, p_{n+1}) > \theta_d$ and $Int(p_m, p_n) > \theta_t$. Therefore, $s = (x, y, t_a, t_l)$, where

$$s.x = \sum_{i=m}^{n} p_i.x/|P|, \quad (1)$$

$$s.y = \sum_{i=m}^{n} p_i.y/|P|, \quad (2)$$

respectively stands for the average $x$ and $y$ coordinates of the collection $P$; $s.t_a = p_m.t$ is the user's arriving time on $s$ and $s.t_l = p_n.t$ represents the user's leaving time.

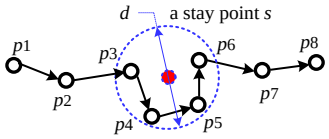



**Figure 2 A GPS trajectory and a stay point**

As depicted in Figure 2, $\{p_1, p_2, \ldots, p_8\}$ formulate a GPS trajectory, and a stay point would be detected from $\{p_3, p_4, p_5, p_6\}$ if $d \leq \theta_d$ and $Int(p_3, p_6) \geq \theta_t$. In contrast to a raw point $p_i$, a stay point carries a particular semantic meaning, such as a shopping mall or a restaurant a user accessed.

Figure 3 presents the architecture of our work. Given 1) GPS trajectories of multiple users and 2) a POI database, our objective is to infer the similarity score between each pair of users. Later, this similarity can be used by some existing clustering algorithms, like K-means, as a distance function to cluster users into groups.

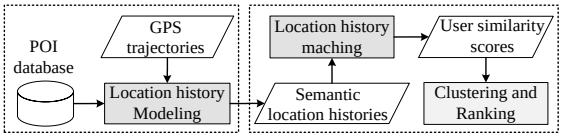



**Figure 3 The architecture of similar user discovery**

In order to make different users' location histories comparable, we first put all users' GPS trajectories together and create a shared framework of location history. Here, a POI database is employed to transfer a user's location history from geographic spaces into the semantic spaces. The POI database contains a corpus of POI entities, each of which includes the properties of category, latitude and longitude, etc. Then, based on the framework we can respectively build a location history for each user. Later, for each pair of users, we explore their similarity by matching their location histories.

## 3. LOCATION HISTORY MODELING

Figure 4 shows the process of modeling location history for each user, and Figure 6 gives a demonstration. This step is comprised of three components denoted as grey boxes in Figure 4.

**Stay Points Representation:** In this component, we first extract stay points from each user's GPS trajectories by using a stay point detection method proposed in paper [4][11]. However, it is almost impossible to identify the exact POI a user visited according to a stay point, given the GPS positioning error and crowded distribution of POIs in a city. In practice, a GPS reading usually have a 10-meter or more error to the real position. Sometimes, there could be multiple POIs pertaining to different categories exist in such a distance range, while the nearest POI to the stay point may not be the real place that a user visited. Therefore, in this work, we represent a stay pint as a $[s.x - \gamma, s.x + \gamma] \times [s.y - \gamma, s.y + \gamma]$ region, where $\gamma$ is a parameter related to the GPS positioning error.

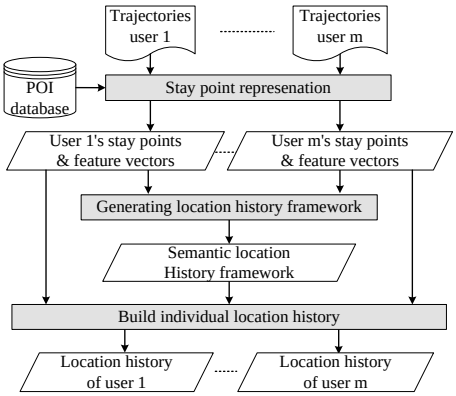



**Figure 4 The procedure of modeling user location history**

We construct a feature vector for each stay region according to the POIs fallen in the region. Here, we employ the ideal of TF-IDF (term frequency-inverse document frequency), and design the feature vector for a stay region as follows.

**Definition 3** (*Feature Vector*) The feature of a stay region $r$ in a collection of regions $R$ is $f_r =< w_1, w_2 \ldots w_F >$, where $w_i$ is the weight of POI category $i$ in region $r$. $F$ is the number of unique POI categories in a POI database.

$$w_i = \frac{n_i}{N} \times log \frac{|R|}{|\{Regions\ containing\ i\}|}, \quad (1)$$

Where $n_i$ is the number of POIs of category $i$ located in region $r$, $N$ stands for the total number of POIs in region $r$, and $|R|$ is the number of regions in the collection.

According to Equation 1, we can represent a stay region with a feature vector (refer to the top part of Figure 6). The feature vector reflects on the uncertainty of accessed categories while bypasses the difficulties in identifying the exact POI visited.

**Generating Location History Framework:** In this component, we cluster the stay regions into some groups according to their feature vectors. The stay regions in the same cluster can be regarded as locations having similar semantic meanings. Intrinsically, we are more capable of discriminating similar users given categories with a finer granularity. For example, "restaurant" identifies users dining outside while "Japanese restaurant" differentiate people interested in different types of foods. So, as shown in the middle part of Figure 6, we hierarchically cluster the feature vectors in a divisive manner and build a tree-structured sematic location hierarchy, where clusters at the same layer share the same granularity and a lower layer denotes a finer granularity.

**Definition 4** (*Semantic Location*) A semantic location $c$ is a feature vector cluster and represents a set of stay regions sharing similar semantic meanings of a certain granularity.

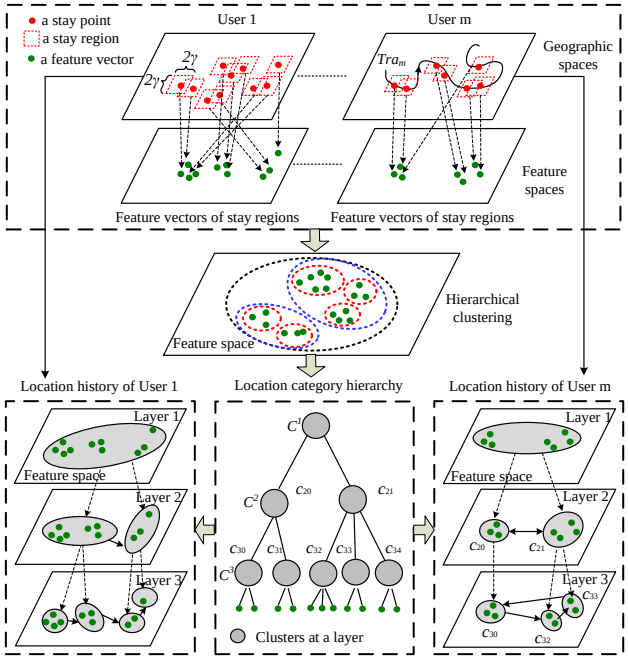



**Figure 6 The demonstration of location history modeling**

**Definition 5** (*Semantic Location Hierarchy*) A semantic location hierarch $\mathcal{F}$ is a tree-structured framework in the feature vector space, $\mathcal{F} = \bigcup_{l=1}^{L}\{C^l\}$, where $L$ is the total number of layers; $C^l = \{c_{l1}, c_{l2}, \dots, c_{lk}\}$ is the set of semantic locations at layer $l$, and $c_{lk}$ denotes the *k*-th semantic location on the *l*-th layer.

**Building Individual Location History:** In this component, we construct a location history for each user based on the semantic location hierarchy $\mathcal{F}$ and the user's stay points. Originally, a user's location history in the geographic spaces is represented by a sequence of stay points with traveling time between each two consecutive stay points. Then, on each layer of the semantic location hierarchy $\mathcal{F}$, we respectively substitute a stay point with the semantic location that the stay point's feature vector pertains to. Now, different users' location histories become comparable.

**Definition 6** (*Semantic Location History*) A user's semantic location history is a sequence of semantic locations on each layer of $\mathcal{F}$, $H = \bigcup_{l=1}^{L}\{S^l\}$, where $S^l = (c_{l0} \xrightarrow{\Delta t_1} c_{l1} \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_k} c_{lk})$ is the sequence on the *l*-th layer of $\mathcal{F}$. Suppose having two consecutive stay points $s_{k-1}$ and $s_k$, $s_{k-1} \in c_{l,k-1}$ and $s_k \in c_{lk}$, then $\Delta t_k = s_k.t_a - s_{k-1}.t_l$ is the traveling time from $c_{l,k-1}$ to $c_{lk}$.

As demonstrated in the up-right part of Figure 6, according to trajectory $Tra_m$ user $m$'s location history can be represented by

$$H = (s_1 \xrightarrow{\Delta t_1} s_2 \xrightarrow{\Delta t_2}, \dots, \xrightarrow{\Delta t_7} s_7),$$

Later, by replacing a stay point with the cluster ID (semantic location) the point's feature vector pertaining to, we can obtain two sequences, $S^2$ and $S^3$, on the second and third layer of $\mathcal{F}$ separately.

$$S^2 = (c_{20} \xrightarrow{\Delta t_1} c_{20} \xrightarrow{\Delta t_2} c_{21} \xrightarrow{\Delta t_3} c_{21} \xrightarrow{\Delta t_4} c_{21} \xrightarrow{\Delta t_5} c_{21} \xrightarrow{\Delta t_6} c_{20}),$$

$$S^3 = (c_{30} \xrightarrow{\Delta t_1} c_{30} \xrightarrow{\Delta t_2} c_{32} \xrightarrow{\Delta t_3} c_{32} \xrightarrow{\Delta t_4} c_{33} \xrightarrow{\Delta t_5} c_{33} \xrightarrow{\Delta t_6} c_{30}),$$

So, user $m$'s location history can be represented as $H = \{S^2, S^3\}$.

# 4. LOCATION HISTORY MATCHING

## 4.1 Finding Maximal Travel Matches

**Definition 7** (*Sub-sequence*) Given a sequence $c_1 \to c_2 \to \dots \to c_m$), we denote the *i*-th item of $S$ as $S[i]$(e.g., $S[1] = c_1$) and represent its subsequence as $S[a_1, a_2, \dots a_k]$ where $1 \le a_1 < a_2 < \dots \le m$.

For instance, $S[1,3,6,7] = c_1 \to c_3 \to c_6 \to c_7$ in the above definition. Note that, we allow *holes* in a sub-sequence, i.e., discontinuous, for a better sequence match.

**Definition 8** (*Travel Match*) Given a temporal constraint factor $\rho \in [0, 1]$ and sub-sequences $S_1[a_1, a_2, \dots a_k]$ and $S_2[b_1, b_2, \dots b_k]$ from two sequences $S_1$ and $S_2$ respectively, these two sub-sequences formulate a *k*-length travel match if they hold the following two conditions.

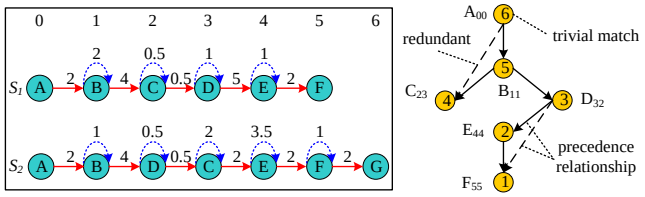
1. $\forall\, i \in [1,k],\ a_i = b_i$, and
2. $\forall\, i \in [2,k],\ \frac{|\alpha_i - \alpha_{i'}|}{Max(\alpha_i, \alpha_{i'})} \le \rho$, where $\alpha_i = S_1[a_i].t_a - S_1[a_{i-1}].t_l$ and $\alpha_i{}' = S_2[b_i].t_a - S_2[b_{i-1}].t_l$, i.e., the travel time between two locations.

In the latter of this paper, we represent the travel match as $(a_1, b_1) \to (a_2, b_2) \to \dots \to (a_k, b_k)$.

**Definition 9** (*Maximal Travel Match*) A travel match $(a_1, b_1) \to (a_2, b_2) \dots \to (a_k, b_k)$ between two sequences $S_1$ and $S_2$ is a maximal travel match if,

1. No left increment: $\nexists a_0 < a_1, b_0 < b_1$, s.t., $(a_0, b_0) \to (a_1, b_1) \to (a_2, b_2) \dots \to (a_k, b_k)$;
2. No right increment: $\nexists a_{k+1} > a_1, b_{k+1} < b_k$, s.t., $(a_1, b_1) \to (a_2, b_2) \dots \to (a_k, b_k) \to (a_{k+1}, b_{k+1})$, and
3. No internal increment: $\forall i \in [1,k]$, $\nexists a_i < a_{i'} < a_{i+1}$ and $b_i < b_{i'} < b_{i+1}$, s.t., $(a_1, b_1) \to (a_2, b_2) \dots \to (a_i, b_i) \to (a_{i'}, b_{i'}) \to (a_{i+1}, b_{i+1}) \to \dots \to (a_k, b_k)$.

Figure 7 (a) demonstrates an example of the maximal travel match between two sequences $S_1$ and $S_2$. Here, a node stands for a semantic location and the letter in a node represents the ID of the location. The numbers on the top of the box denotes the index of a node in a sequence. The number appearing on a solid edge means the travel time between two consecutive nodes, and the number shown on a dashed edge denotes the stay time in a location.




(a) Two stay category sequences (b) Precedence graph

**Figure 7 An example of the maximal travel match**

Let $\rho$ = 0.2 in this example. First, $A \to B$ is a travel match, because the travel times $(A \to B)$ in $S_1$ and $S_2$ are identical, |2-2|/2=0. Then, we find that $B \to C$, also satisfies the conditions defined in Definition 8. Though $B$ and $C$ is not directly connected in $S_2$, the travel time between these two locations is 4+0.5+0.5=5, which is very similar to that of $S_1$. In short, |5-4|/5=0.2. However, both $A \to B$ and $B \to C$ are not the maximal travel match in this example as they are contained in $A \to B \to C$. Later, $C \to E$ and

$C \rightarrow F$ cannot formulate travel matches due to the difference between corresponding travel times. Using the same approach, we find $A \rightarrow B \rightarrow D \rightarrow E \rightarrow F$ is another maximal travel match. Overall, we detect two maximal travel matches, $A \rightarrow B \rightarrow C$ and $A \rightarrow B \rightarrow D \rightarrow E \rightarrow F$ from $S_1$ and $S_2$.

We prove that the maximal matches shown in Figure 7 (a) are equivalent to the maximal length paths in the graph $G$ shown in Figure 7 (b). Here, a path $P$ is a maximal path in $G$ if the first node of $P$ has zero in-degree and the last node has zero out-degree, e.g., $A_{11} \rightarrow B_{22} \rightarrow C_{34}$. Figure 8 presents the algorithm for building the graph. In line 9 of this algorithm, $v_l$ is a precedence of $v_t$ if the user reached $v_l$ before $v_t$ and $v_l \rightarrow v_t$ holds condition 2 of Definition 8. Later, we can find out the maximal travel matches by searching $G$ for the maximal length paths.

**Algorithm 3** BuildGraph ( $S_1$, $S_2$)

**Input**: Two semantic location sequences $S_1$ and $S_2$
**Output**: A directed acyclic graph $G'$.

1: **For** $\forall i \in [1, |S_1|], \forall j \in [1, |S_2|]$
2: **If** $S_1[i].c = S_2[j].c$
3: Add the node $(i, j)$ into a list $\Psi$;
4: $\Psi \leftarrow$ Sort $(\Psi)$; //sort in a decreasing lexicographical order. //Suppose $\Psi = (v_1 = (i_1, j_1), \dots, v_k = (i_k, j_k))$.
5: **For** $l$ from 2 to $k$
6: For $t$ from $l-1$ down to 1
7: **if** $v_l$ is white
8: **if** $v_l$ is a precedence of $v_t$
9: Build an edge $v_l \rightarrow v_t$ in $G'$.
10: Mark all nodes reachable from $v_t$ black
11: **Return** $G'$;

**Figure 8** Building refined graph directly based on two sequences

## 4.2 Calculating Similarity

Given two users' location histories $H_1$ and $H_2$, we compute the similarity between them by summarizing the weighted similarity of semantic location sequences detected at each layer of the hierarchy $\mathcal{F}$, according to Equation 2, 3, 4.

$$SimUser(H_1, H_2) = \sum_{l=1}^{L} f_w(l) \times SimSq(S_1^l, S_2^l); \quad (2)$$

$$SimSq(S_1, S_2) = \frac{\sum_{j=1}^{m} sg(t_j)}{|S_1| \times |S_2|}, \quad (3)$$

$$sg(s) = g_w(k) \times \sum_{i=1}^{k} iuf(c_i); \quad (4)$$

We use a function $f_w(l)$ to assign a higher weight to the similarity of sequences occurring at a lower layer, e.g., $f_w(l) = 2^{l-1}$. Then, the similarity between two semantic location sequences $S_1$ and $S_2$ at a layer, $SimSq(S_1, S_2)$, is represented by the sum of the similarity score, $sg(t_j)$, of each maximal match between $S_1$ and $S_2$. Here, $m$ is the total number of maximal matches. Meanwhile, $SimSq(S_1, S_2)$ is normalized by the production of the lengths of the two sequences, since a longer sequence have a high probability to have long matches. Later, we calculate the similarity of a maximal travel match $t$, $sg(t)$, by summing up the $iuf$ of each semantic location $c$ in $t$ and weighting $sg(t)$ in terms of the length $k$ of $t$, e.g., $g_w(k) = 2^{k-1}$.

# 5. EVALUATION

Using a real-world GPS data collected by 109 users over a year, we evaluate the following 3 aspects of our method. 1) Compare the semantic location history with a physical one, i.e., SLH vs. HGSM. 2) The hierarchy of our method. 3) SLH-MTM vs. other similarity functions, such the Cosine similarity and Pearson correlation 4) The weights used in the similarity function. The results are shown in Figure 9, 10, 11, 12 respectively. Here, we set $\theta_d$ =200m, $\theta_t$ =30, $\gamma$ =200m, $\rho$=0.2, and use *nDCG* as a metric.

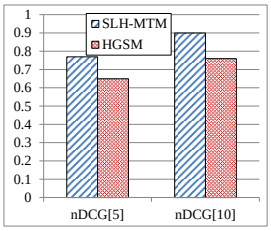



**Figure 9. SLH vs. HGSM**

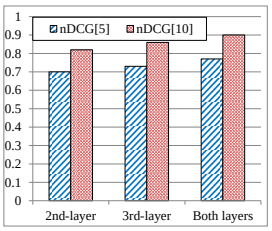



**Figure 10. Study the hierarchy**

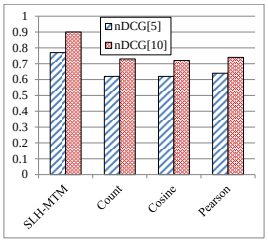



**Figure 11. SLH-MTM vs. other similarity functions**

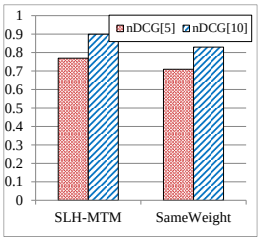



**Figure 12. Study the weights used in similarity score**

# 6. CONCLUSION

In this paper, we estimate the similarity between users in terms of the semantic location history learned from their historical GPS trajectories. The experimental results show that users sharing 1) a finer semantic location, 2) a longer sequence of locations and 3) less popular semantic locations would be more similar to each other. Future work can be personalized location recommenders and community discovery based on this user similarity.

# 7. REFERENCES


[1] GeoLife GPS trajectories: http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx

[2] Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. PUC, 10(4): 255–268, 2006.

[3] Hung, C., Chang, C., Peng, W.: Mining trajectory profiles for discovering user communities: In ACM GIS workshop on location-based social networks, 2009

[4] Li, Q., Zheng, Y. Xie, X. Chen, Y., Ma, W.: Mining user similarity based on location history. In: Proceedings of ACM GIS (2008).

[5] Zheng, Y., Chen, Y., Xie, X., Ma, W.: GeoLife2.0: A Location-Based Social Networking Service. In: Proceedings of MDM (2009).

[6] Zheng, V., W., Cao, B., Zheng, Y., Xie, X., Yang, Q.: Collaborative Filtering Meets Mobile Recommendation: A User-centered Approach. in AAAI (2010).

[7] Zheng, Y., Liu, L., Xie, X.: Learning transportation mode from raw GPS data for geographic applications on the web. In WWW (2008).

[8] Zheng, Y., Xie, X.: Learning Location Correlation from GPS trajectories. In MDM 2010.

[9] Zheng, Y., Xie, X.: Learning travel recommendations from user-generated GPS traces, ACM Trans. on Intelligent Systems and Technologies. 1,1, 2010

[10] Zheng, Y., Xie, X., Ma, W.: GeoLife: A collaborative social networking service among user, location and trajectory. IEEE Data Engineering Bulletin. 33, 2, 2010, pp. 32-40.

[11] Zheng, Y., Zhang, L., Xie, X.: Recommending friends and locations based on individual location history. ACM Trans. on the Web, 2011.

[12] Zheng, Y., Zhang, L., Xie, X., Ma, W.: Mining interesting locations and travel sequences from GPS trajectories. In Proc. of WWW 2009.

[13] Zheng, V. W., Zheng, Y. Xie, X., Yang, Q.: Collaborative location and activity recommendations with GPS history data. In Proceedings of WWW 2010.