GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory

Yu Zheng, Xing Xie and Wei-Ying Ma
Microsoft Research Asia, 4F Sigma Building, NO. 49 Zhichun Road, Beijing 100190, China
{yuzheng, xingx, wyma}@microsoft.com

Abstract

People travel in the real world and leave their location history in a form of trajectories. These trajectories do not only connect locations in the physical world but also bridge the gap between people and locations. This paper introduces a social networking service, called GeoLife, which aims to understand trajectories, locations and users, and mine the correlation between users and locations in terms of user-generated GPS trajectories. GeoLife offers three key applications scenarios: 1) sharing life experiences based on GPS trajectories; 2) generic travel recommendations, e.g., the top interesting locations, travel sequences among locations and travel experts in a given region; and 3) personalized friend and location recommendation.

1 INTRODUCTION

The advance of location-acquisition technologies like GPS and Wi-Fi has enabled people to record their location history with a sequence of time-stamped locations, called trajectories. These trajectories imply to some extent users' life interests and preferences, and have facilitated people to do many things, such as online life experience sharing [12, 13], sports activity analysis [4] and geo-tagging multimedia content [3]. Using these trajectories, we cannot only connect locations in the physical world but also bridge the gap between users and locations and further deduce the connections among users. That is, we are able to understand both users and locations based on the trajectories.

In this paper, we introduce our project GeoLife [1, 2, 5 14], which is a social networking service incorporating users, locations and user-generated GPS trajectories. As demonstrated in the left part of Figure 1, people access a sequence of locations in the real world and generate many trajectories in a form of GPS logs. Based on these GPS trajectories, we can build three graphs: a location-location graph, a user-location graph, and a user-user graph. In the location-location graph, a node is a location and a directed edge between two locations stands for that a least some users have consecutively traversed these two locations in a trip. In the user-location graph, there are two types of nodes, users and locations. An edge starting from a user and ending at a location indicates that the user has visited this location for some times, and the weight of the edge is the visiting times of the user. Further, we can infer the user-user graph where a node is a user and an edge between two nodes represents that the two users have visited the same location in the real world for some times (the edge weight is the times sharing the same locations).

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

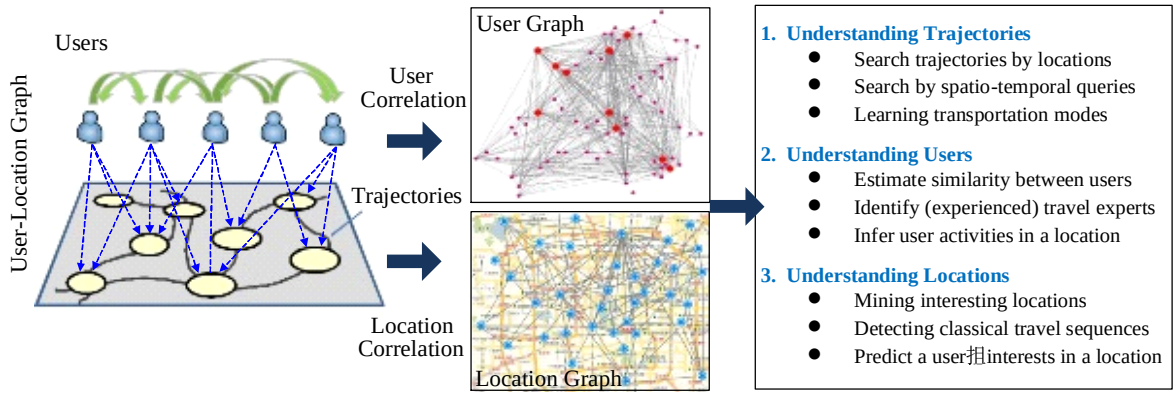

Figure 1: The philosophy and research points of GeoLife

Using the graphs mentioned above, we conduct three aspects of research listed in the right part of Figure 1. 1) *Understanding trajectories*: For example, we provide a method to search for the trajectories by giving a set of locations of interests [16], or by issuing a spatial query range combined with a temporal interval [1] as a query. Meanwhile, we infer the transportation modes that a user took when generating a GPS trajectory [8, 11, 15]. 2) *Understand users*: First, we estimate the similarity between each pair of users in terms of their location histories represented by GPS trajectories [2]. Second, using an iterative inference model, we compute the travel experience of each user based on their location history, and then find out travel experts in a geo-region by ranking users according to the experiences [9]. Third, we can infer the activities that a user can perform in a location based on multiple users' location histories and the associated comments [5][6]. 3) *Understanding locations*: On example is that we mine the top interesting locations and travel sequences in a given region from a large number of users' GPS trajectories. Another example is we predict a user's interests in an unvisited location by involving the GPS trajectories of the user and that of others.

In the rest of this paper, we present three application scenarios, each of which includes several research points mentioned above. These scenarios are 1) sharing life experiences based on GPS trajectories, 2) generic travel recommendation, and 3) personalized friend and travel recommendation.

2 SHARE LIFE EXPERIENCES BASED ON GPS TRAJECTORIES

Recently, a branch of GPS-trajectory-sharing applications [3, 4, 12, 13] appeared on the Internet. In these applications, people can record their travel routes using a GPS-enabled device and then share travel experiences among each other by publishing these GPS trajectories in a Web community. Photos, comments and tips can also be associated with the related locations in a trajectory. GPS-trajectories-sharing offers a more fancy and interactive approach than text-based articles to better express people's travel experiences, which provide users with valuable references when planning a travel itinerary. However, as a large-scale GPS tracks have been accumulated, how to manage these GPS data is an important issue in these applications. Obviously, users need an efficient approach to retrieve the specific GPS trajectories they are interested in, since nobody has time and patience to browse trajectories one-by-one. Moreover, people intend to learn information about user behaviors as well as user intentions behind the raw data. Thus, this section first presents two approaches that can facilitate users to search for trajectories efficiently. Then, we introduce a machine-learning algorithm inferring the transportation modes of a GPS trajectory. This technique can provide us with deep understanding of a trajectory and achieve the automatic categorization of GPS trajectories by transportation modes, and further enable trajectory filtering by transportation modes.

2.1 Search Trajectories by Location

When traveling to an unfamiliar city, people usually have a set of places of interests in their mind while they do not know how to travel among these places. For example, most of people visiting Beijing know or hear of Tiananmen Square, the Summer Palace and the Bird's-Nest in the Olympic Park. However, they have no idea how to travel in each place and what a proper visiting sequence among these places is. At this moment, a user can give the names of these three locations as a query, or directly point out the locations of these places on a map (if they know where these places are). Then, our method searches the existing database for some trajectories that were generated by other users and traverse these places. In short, this query can benefit travelers planning a trip to multiple places of interest in an unfamiliar city by providing similar routes traveled by other people.

In this work, we propose a new type of trajectory query called the k Best-Connected Trajectory (k-BCT) query for searching trajectories by multiple geographical locations. Generally, the k-BCT query is a set of locations indicated by coordinates like (latitude, longitude), which could be a famous attraction, a nameless beach, or any arbitrary place approximated by the center location. The user may also specify a preferred visiting order for the intended places, in which case the order of a trajectory needs to be taken into account as well. To answer the k-BCT query, we first define a similarity function, which should consider the distance from the trajectory to each query location, to measure how well a trajectory connects the query locations. Second, we propose an Incremental k-NN based Algorithm (IKNN), which retrieves the nearest trajectory points with regard to each query location incrementally and examines the k-BCT from the trajectory points discovered so far. In this algorithm, the pruning and refinement of the search are conducted by using the lower bound and upper bound of trajectory similarity that are derived from the found trajectory points. The retrieval of nearest trajectory points is based on the traditional best-first and depth-first k-NN algorithms over an R-tree index. Refer [16] for details.

2.2 Search Trajectories by Spatio-Temporal Queries

Sometimes, people are interested in some GPS trajectories showing the travel experiences within a particular geo-region and in a specific time interval. For instance, what is fun in the Olympic Park of Beijing in the weekend? Where would be interesting in the downtown Beijing during Christmas? These questions can be represented by a spatio-temporal query, with which we can retrieve some existing trajectories providing reference travel experiences. In these online GPS-trajectory-sharing applications, we observe that users tend to upload GPS trajectories of the near past more frequently than the trajectories of the distant past. For instance, users are more likely to upload GPS trajectories of today or yesterday than those of months ago. Thus, traditional spatio-temporal indexing schemes, like R tree or its variants, are not optimal to handle the skewed nature of accumulative GPS trajectories.

We propose Compressed Start-End Tree (CSE-tree) [1] for the GPS data sharing applications based on users' uploading behavior. In this scheme, we first partition the space into disjoint cells that cover the whole spatial region, and then maintain a flexible temporal index for each spatial cell. To insert a new GPS track, we divide the track into segments according to spatial partition. Then each segment is inserted into the temporal index of corresponding spatial grid. For all segments in a temporal index, they are divided into several groups according to end time of the segment. We observed that for different groups, the frequency of new updates is different. CSE-tree uses B+ tree index for frequently updated groups and sorted dynamic array for rarely updated ones. The update frequency of a group may change as time goes by, so we transform a B+ tree into a sorted dynamic array if update frequency of a group drops below a threshold. Our contribution lies in the following two aspects.

- A stochastic process model is proposed to simulate user behavior of uploading GPS tracks to online sharing applications. This model can also be applied to other data sharing applications on the Web.
- A novel indexing scheme is optimized to the user behavior of uploading GPS tracks. Our scheme requires less index space and less update cost while keeping satisfactory retrieval performance.

2.3 Learning transportation modes based on GPS Data

As a kind of human behavior, people's transportation modes, such as walking and driving, can provide pervasive computing systems with more contextual information and enrich a user's mobility with informative knowledge. With the transportation modes of a GPS track, people can obtain more reference knowledge from others' trajectories. Users can know not only where other people have been but also how these people reach each location. Meanwhile, such knowledge enables the classification of GPS trajectories by transportation modes. So, we can perform smart route recommendations/designs for a person based on the person's needs. For instance, a system should return a bus line rather than a driving route to an individual intending to move to somewhere by a bus.

We designed an approach based on supervised learning to automatically infer users' transportation modes, including driving, walking, taking a bus and riding a bike, from raw GPS logs. Our approach consists of three parts: a change point-based segmentation method, an inference model and a graph-based post-processing algorithm. First, we propose a change point-based segmentation method to partition each GPS trajectory into separate segments of different transportation modes. The key insight of this step is that people need to walk for a while before transferring to another transportation mode. Second, from each segment, we identify a set of sophisticated features, which are not affected by differing traffic conditions (e.g., a person's direction when in a car is constrained more by the road than any change in traffic conditions). Later, these features are fed to a generative inference model to classify the segments of different modes. Third, we conduct graph-based post-processing to further improve the inference performance. This post-processing algorithm considers both the commonsense constraints of the real world and typical user behaviors based on locations in a probabilistic manner. Refer to papers [8, 11, 15] for details.

The advantages of our method over the related works include three aspects. 1) Our approach can effectively segment trajectories containing multiple transportation modes. 2) Our work mined the location constraints from user-generated GPS logs, while being independent of additional sensor data and map information like road networks and bus stops. 3) The model learned from the dataset of some users can be applied to infer GPS data from others. Using the GPS logs collected by 65 people over a period of 10 months, we evaluated our approach via a set of experiments. As a result, based on the change-point-based segmentation method and Decision Tree-based inference model, we achieved prediction accuracy greater than 71 percent. Further, using the graph-based post-processing algorithm, the performance attained a 4-percent enhancement.

3 GENERIC TRAVEL RECOMMENDATION

3.1 Mining Travel Experts, Interesting Locations and Travel Sequences

This recommender provides a user with the top n experienced users (experts), interesting locations and the classical travel sequences among these locations, in a given geospatial region. To define interesting location, we mean the culturally important places, such as Tiananmen Square in Beijing and the Statue of Liberty in New York (i.e. popular tourist destinations), and commonly frequented public areas, such as shopping malls/streets, restaurants, cinemas and bars. With the information mentioned above, an individual can understand an unfamiliar city in a very short period and plan their journeys with minimal effort.

However, we will meet two challenges when conducting this recommendation. First, the interest level of a location does not only depend on the number of users visiting this location but also lie in these users' travel experiences. Intrinsically, different people have different degrees of knowledge about a geospatial region. For example, the local people of Beijing are more capable than overseas tourists of finding out high quality restaurants and famous shopping malls in Beijing. Second, an individual's travel experience and interest level of a location are relative values (i.e., it is not reasonable to judge whether or not a location is interesting), and are region-related (i.e., conditioned by the given geospatial region). A user, who has visited many places in a city like New York, might have no idea about another city, such as Beijing.

To achieve this generic recommendation, we first propose a tree-based hierarchical graph (TBHG) that models multiple users' travel sequences on a variety of geospatial scales. As demonstrated in Figure 2, three steps need to be performed when building a TBHG. 1) Detect stay points: We detect from each GPS trajectory some stay points [2] where a user has stayed in a certain distance threshold over a time period. 2) Formulate a tree-based Hierarchy H: We put together the stay points detected from users' GPS logs into a dataset. Using a density-based clustering algorithm, we hierarchically cluster this dataset into some geospatial regions (a set of clusters C) in a divisive manner. Thus, the similar stay points from various users would be assigned to the same clusters on different levels. 3) Build graphs on each level: Based on the tree-based hierarchy H and users' location histories, we can connect the clusters of the same level with directed edges. If consecutive stay points from one trip are individually contained in two clusters, a link would be generated between the two clusters in a chronological direction according to the time serial of the two stay points.

Based on the TBHG, we propose a HITS-based model to infer users' travel experiences and interest of a location within a region. This model leverages the main strength of HITS to rank locations and users with the context of a geospatial region, while calculating hub and authority scores offline. Therefore, we can ensure the efficiency of our system while supporting users to specify any geo-regions as queries. Using the third level of the TBHG shown in Figure 2 as a case, Figure 3 illustrates the main idea of our HITS-based inference model. Here, a location is a cluster of stay points, like c_{31} and c_{32}. We regard an individual's visit to a location as an implicitly directed link from the individual to that location. For instance, cluster c_{31} contains two stay points respectively detected from u_1 and u_2's GPS traces, i.e., both u_1 and u_2 have visited this location. Thus, two directed links are generated respectively to point to c_{31} from u_1 and u_2. Similar to HITS, in our model, a hub is a user who has accessed many places, and an authority is a location which has been visited by many users. Therefore, users' travel experiences (hub scores) and the interests of locations (authority scores) have a mutual reinforcement relation. More specifically, a user's travel experience can be represented by the interest levels of the visited locations. In turn, the interest level of a location can be calculated by the experiences of the users who have accessed this location. Using a power iteration method, we can generate the final score for each user and location, and find out the top n interesting locations and experience users in a given region.

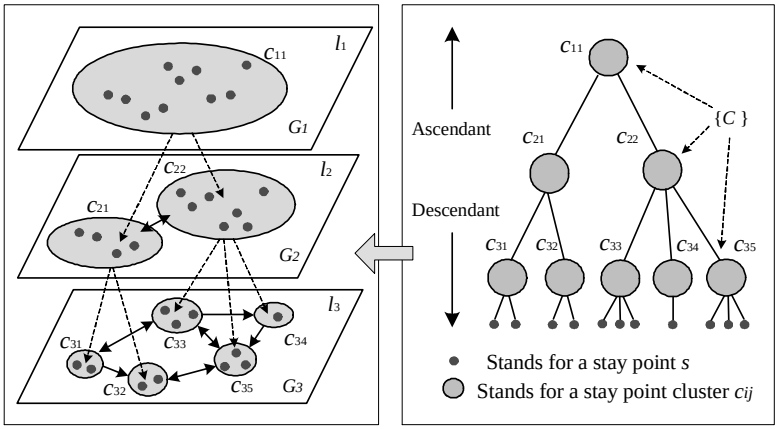

Figure 2: Building a tree-based hierarchical graph

With users' travel experiences and the interests of locations, we calculate a classical score for each location sequence within the given geospatial region. The classical score of a sequence is the integration of the following three aspects. 1) The sum of hub scores of the users who have taken this sequence. 2) The authority scores of the locations contained in this sequence. 3) These authority scores are weighted based on the probability that people would take a specific sequence. Refer to papers [9, 14] for details.

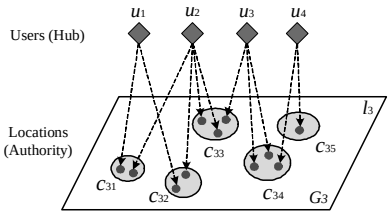

Figure 3: Our HITS-based inference model

3.2 Collaborative Location and Activity Recommendation

Typically, people would have the following two types of questions in their mind when traveling. One is that, if we want to do something like sightseeing or food-hunting in a large city,

where should we go? The other is, if we have already visited some places, such as the Olympic park of Beijing, what else can we do there? In general, the first question corresponds to location recommendation given some activity query (where "activity" can refer to various human behaviors such as food-hunting, shopping, watching movies/shows, enjoying sports/exercises, tourism, etc.), and the second question corresponds to activity recommendation given some location query.

In this work, for the first question, we can provide a user with a list of interesting locations, such as Tiananmen Square and Bird's Nest. For the second question, if a user visits Bird's Nest, we can recommend her to not only go sightseeing but also experience some outdoor exercise facilities or try some nice food nearby. We put both location recommendation and activity recommendation together in the knowledge mining, since locations and activities are closely related in nature. As we mentioned, to better share life experience, a user can add some comments or tips to a point location in a trajectory. Such comments and tips imply a user's behavior and intentions when visiting a location. With the available user comments, we can get the statistics about what kinds of activities the users performed on a location, and how often they performed these activities. As shown in the middle part of Figure 4, by organizing this statistics' data in a matrix form, we can have a location-activity matrix, with rows as locations and columns as activities. An entry in the matrix denotes the frequency for the users to perform some activity on some location. However, the location-activity matrix is incomplete and very sparse. Therefore, we leverage the information from another two matrices, the location-feature and activity-activity matrices, respectively shown in the left and right part of Figure 4.

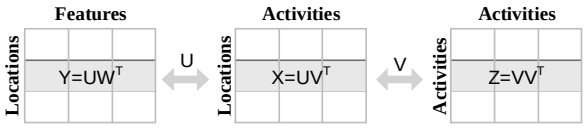

Figure 4: Collaborative location-activity learning model

Location-feature matrix: We exploit the location features with the help of POI category database. The database is based on the city yellow pages, and it can provide us the knowledge that what kinds of POIs we have in an area. This helps us to get some sense of this location's functionalities, so that we can use them as features for better recommendations. Similarly, by organizing the data in a matrix form, we can have a location-feature matrix, where each entry of the matrix denotes some feature value on that location.

Activity-activity matrix: We exploit the World Wide Web to get the knowledge about the activity correlations. With this knowledge, we may better infer that if a user performs some activity on a location, then how likely she would perform another activity. By organizing the data in a matrix form, we have an activity-activity matrix, where each entry of the matrix denotes the correlation between a pair of activities.

After the data modeling, we have the location-activity, location-feature and activity-activity matrices. Our objective is to fill those missing entries in the location-activity matrix with the information learned from the other two matrices, so as to get a full matrix for both location and activity recommendations. Here, we provided a collaborative filtering (CF) approach based on collective matrix factorization to take these information sources as inputs and train a location and activity recommender. By defining an objective function and using the gradient descent to iteratively minimize the objective function, we can infer the value of each missed entry. Based on the filled location-activity matrix, we can rank and retrieve the top k locations/activities as recommendations to the users. Refer to papers [5, 6] for details.

4 PERSONALIZED FRIEND & LOCATION RECOMMENDATION

Besides the generic recommendation, an individual also wants to visit some locations matching her travel preferences (personalized). Actually, people's outdoor movements in the real world would imply rich information about their life interests and preferences. For example, if a person usually goes to stadiums and gyms, it denotes that the person might like sports. According to the first law of geography, everything is related to everything else, but near things are more related than distant things, people who have similar location histories might share similar interests and preferences. The more location histories they share, the more correlated these two users

would be. Therefore, based on users' GPS trajectories we conduct a personalized friend & location recommender, which provides an individual with some similar users in terms of location histories and recommends some places that could interest the individual while having not been found by the individual.

4.1 Friend Recommendation

To achieve the friend recommendation, we first build a shared framework shown in Figure 2 according to the first two steps mentioned in Section 3.1. By feeding each user's GPS trajectories into this framework, we can build a personal hierarchical graph for each user. Based on this graph, we propose a similarity measurement, referred to as a hierarchical-graph-based similarity measurement (HGSM), which takes into account the following three factors to estimate the similarity between users:
1) *The sequence property of people's outdoor movements*: The longer similar sequences matched between two users' location histories, the more related these two users might be.
2) *The hierarchical property of geographic spaces*: Users who share similar location histories on geographical spaces of finer granularities might be more correlated.
3) *The visited popularity of a location*: Similar to inverse document frequency, two users who accessed a location visited by a few people might be more correlated than others sharing a location history accessed by many people.

Given two users' hierarchical graphs, we can calculate a similarity score for them by using the HGSM. Later, a group of people, called potential friends, with relatively high scores will be retrieved for a particular individual. Refer to paper [2] for details.

4.2 Personalized Location Recommendation

This recommender uses a particular individual's visits on a geospatial location as their implicit ratings on the location, and predicts a particular user's interest in an unvisited location in terms of their location history and those of other users. To achieve this recommender, we first formulate a matrix between users and locations, where rows stand for users and columns represent users' ratings on locations. We incorporated a content-based method into a user-based collaborative filtering algorithm, which uses HGSM as the user similarity measure, to estimate the rating of a user on an item. Later, some unvisited locations that might match their tastes can be recommended to the individual. Refer to paper [10] for details. Though the CF model using HGSM as a similarity measurement is more effective than those using the Pearson correlation and the Cosine similarity, it consumes a lot of computation since the HGSM-based method needs to calculate the similarity between each pair of users while the number of users could keep on increasing as a system becomes popular. To address this problem, we propose an item-based CF model regarding locations as items. In this work, we first mine the correlation among locations from multiple users' GPS traces in terms of 1) the sequences that the locations have been visited and 2) the travel experiences of the users creating these sequences. Then, the location correlation is incorporated into a CF-based model that infers a user's interests in an unvisited location based on her locations histories and that of others. The item-based CF model using location correlation is slightly less effective than the HGSM-based one while is much more efficient than the latter. Papers [7, 14] offer details.

5 CONCLUSION

User-generated GPS trajectories do not only connect locations in the physical world but also bridge the gap between people and locations. In GeoLife, we aim to understand trajectories, users, and locations in a collaborative manner, and perform three key application scenarios. The first scenario, sharing life experience based on GPS trajectory, focuses on understanding GPS trajectories. In the second scenario, generic travel recommendation, we infer the travel experience of a user (understand users) and the interest level of a location (understand locations) in an iterative manner. Later, the classical travel sequences among locations (location correlation) are

detected based on the inferred user experience and location interest. Meanwhile, we learn user activities in a location (user-location correlation) and enable a location-activity recommender. In the third scenario, personalized friend and location recommendation, we estimate the similarity between users (user correlation) based on the similarity between their location histories (location correlation). Later, a personalized recommender, which predicts a user's interests in an unvisited location (user-location correlation) by integrating this user similarity into a CF model, is conducted. Overall, user, location and trajectory have a collaborative and mutual reinforcement relationship among each other.

References

[1] Longhao Wang, Yu Zheng, Xing Xie, Wei-Ying Ma. A Flexible Spatio-Temporal Indexing Scheme for Large-Scale GPS Track Retrieval, In MDM 2008.

[2] Quannan Li, Yu Zheng, Yukun Chen, Xing Xie. Mining user similarity based on location history. In ACM SIGSPATIAL GIS 2008.

[3] Scott Counts, M. Smith. Where were we: Communities for sharing space-time trails. In ACM GIS 2007

[4] SPORTSDO. 2007. http://sportsdo.net/Activity/ActivityBlog.aspx.

[5] Wenchen Zheng, Bin Cao, Yu Zheng, Xing Xie, Qiang Yang. Collaborative Filtering Meets Mobile Recommendation: A User-centered Approach. In AAAI 2010.

[6] Wencheng Zheng, Yu Zheng, Xing Xie, Qiang Yang. Collaborative Location and Activity Recommendations With GPS History Data. In WWW 2010.

[7] Yu Zheng, Xing Xie. Learning Location Correlation from User-Generated GPS trajectories. In MDM 2010.

[8] Yu Zheng, Like Liu, Longhao Wang, Xing Xie. Learning Transportation Modes from Raw GPS Data for Geographic Application on the Web, In WWW 2008.

[9] Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma. Mining interesting locations and travel sequences from GPS trajectories. In WWW 2009

[10] Yu Zheng, Lizhu Zhang, Xing Xie. Recommending friends and locations based on individual location history. To appear in ACM Transaction on the Web, 2010.

[11] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie. Understanding Mobility Based on GPS Data. In UbiComp 2008.

[12] Yu Zheng, Yukun Chen, Xing Xie, Wei-Ying Ma. GeoLife2.0: A Location-Based Social Networking Service. In MDM 2009.

[13] Yu Zheng, Longhao Wang, Xing Xie, Wei-Ying Ma. GeoLife-Managing and understanding your past life over maps, In MDM 2008.

[14] Yu Zheng, Xing Xie. Learning travel recommendation from user-generated GPS trajectories. To appear in ACM Transaction on Intelligent Systems and Technologies (ACM TIST).

[15] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, Wei-Ying Ma. Understanding transportation modes based on GPS data for Web applications. ACM Transaction on the Web. 4(1), January, 2010. 1-36.

[16] Zaiben Chen, Heng Tao Shen, Xiaofang Zhou, Yu Zheng, Xing Xie. Searching Trajectories by Locations - An Efficiency Study, In ACM SIGMOD 2010