Mining Interesting Locations and Travel Sequences from GPS Trajectories

Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma
Microsoft Research Asia, 4F, Sigma building, No.49 Zhichun road, Haidian District, Beijing 100190, China
{yuzheng, v-lizzha, xingx, wyma}@microsoft.com

ABSTRACT

The increasing availability of GPS-enabled devices is changing the way people interact with the Web, and brings us a large amount of GPS trajectories representing people's location histories. In this paper, based on multiple users' GPS trajectories, we aim to mine interesting locations and classical travel sequences in a given geospatial region. Here, interesting locations mean the culturally important places, such as Tiananmen Square in Beijing, and frequented public areas, like shopping malls and restaurants, etc. Such information can help users understand surrounding locations, and would enable travel recommendation. In this work, we first model multiple individuals' location histories with a tree-based hierarchical graph (*TBHG*). Second, based on the *TBHG*, we propose a HITS (Hypertext Induced Topic Search)-based inference model, which regards an individual's access on a location as a directed link from the user to that location. This model infers the interest of a location by taking into account the following three factors. 1) The interest of a location depends on not only the number of users visiting this location but also these users' travel experiences. 2) Users' travel experiences and location interests have a mutual reinforcement relationship. 3) The interest of a location and the travel experience of a user are relative values and are region-related. Third, we mine the classical travel sequences among locations considering the interests of these locations and users' travel experiences. We evaluated our system using a large GPS dataset collected by 107 users over a period of one year in the real world. As a result, our HITS-based inference model outperformed baseline approaches like *rank-by-count* and *rank-by-frequency*. Meanwhile, when considering the users' travel experiences and location interests, we achieved a better performance beyond baselines, such as *rank-by-count* and *rank-by-interest*, etc.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - *data mining*. H.5.2 [Information Interface and Presentation]: User Interface. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, retrieval model*.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Spatial data mining, GPS trajectories, Location recommendation.

1. INTRODUCTION

GPS-enabled devices, like GPS-phones, are changing the way people interact with the Web by using locations as contexts. With such a device, a user is able to acquire present locations, search the information around them and design driving routes to a destination. In recent years, many users start recording their outdoor movements with GPS trajectories for many reasons, such as travel experience sharing, life logging, sports activity analysis and multimedia content management, etc. Meanwhile, a branch of Websites or forums [1][2][3], which enable people to establish some geo-related Web communities, have appeared on the Internet. By uploading GPS logs to these communities, individuals are able to visualize and manage their GPS trajectories on a Web map. Further, they can obtain reference knowledge from others' life experiences by sharing these GPS logs among each other. For instance, a person is able to find some places that attract them from other people' travel routes, hence, plan an interesting and efficient journey based on multiple users' experiences.

With the pervasiveness of the GPS-enabled devices, a huge amount of GPS trajectories have been accumulating unobtrusively and continuously in these Web communities. However, almost all of these applications still directly use raw GPS data, like coordinates and time stamps, without much understanding. Hence, so far, these communities cannot offer much support in giving people interesting information about geospatial locations. What's more, facing such a large dataset in a community, it is impossible for a user to browse each GPS trajectory one by one.

Typically, people would desire to know which locations are the most interesting places in a 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, bars etc. Further, given these interesting locations in a geospatial region like a city, users might also wonder what the most classical travel sequences are among them. For example, an individual would be more likely to go to a bar *after* visiting a cultural landmark than they would *before*, making landmark-to-bar a classical travel sequence.

With the information mentioned above, an individual can understand an unfamiliar city in a very short period and plan their journeys with minimal effort. Meanwhile, such information would enable mobile guides [6][14]; given the recommendation of the interesting places and travel sequences around them, mobile users are more likely to enjoy a high quality travel experience while saving lots of time for location finding and trip planning.

However, it is not easy to infer the interest of a location because of the following two reasons. 1) The interest of a location does not only depend on the number of users visiting this location but also lie in these users' travel experiences. Intrinsically, various people have different degrees of knowledge about a geospatial region. In a journey, the users, with more travel experiences about a region, would be more likely to visit some interesting locations in that region. For instance, the local people of Beijing are more capable than overseas tourists of finding out high quality restaurants and famous shopping malls in Beijing. 2) An individual's travel

WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

experience and interest 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. Likewise, the most interesting restaurant in a district of a city might not be the most interesting one of the whole city (as other restaurants from the remaining districts might outperform it).

In this paper, based on multiple users' GPS trajectories, we aim to mine the top *n* interesting locations and the top *m* classical travel sequences in a given geospatial region, by taking into account users' different travel experiences as well as the correlation between locations. At the same time, we are able to infer the most *k* experienced users in a geo-related community. Here, we regard a user's visit to a location as an implicitly directed link from the user to that location, i.e., a user would point to many locations and a location would be pointed to by many users. Further, these links are weighted based on different individuals' travel experiences in this region. Therefore, we are able to involve the key idea of the HITS model to infer users' travel experiences and the relative interest of a location.

In this HITS-based model, a geospatial region corresponds to a topic; an individual's hub score stands for their travel experiences, and the authority score of a location represents the interest of the location. Users' travel experiences and the interest of a place have a mutual reinforcement relationship. Intuitively, the user with rich travel experiences in a region might visit many interesting places in that region, and a very interesting place in that region might be accessed by many users with rich travel experiences. For simplicity's sake, in the remainder of this paper, we call the user with rich travel experiences (i.e., relatively high hub score) in a region, an experienced user of that region, and a location that attracts people's profound interests (relatively high authority score) is denoted as an interesting location. Further, considering a user's experience of travel and the interest of a location, we mine the classical travel sequences from people's GPS logs.

The work reported in this paper is a step towards enhancing mobile Web by involving the knowledge mined from multiple users' location histories. Also, this is an approach to improve the location-based services by integrating social networking into mobile Web. The contributions of this paper lie in four aspects:

- We propose a tree-based hierarchical graph (*TBHG*), which can model multiple users' travel sequences on a variety of geospatial scales based on GPS trajectories.
- 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 allowing users specify any regions on a map.
- Considering individuals' travel experiences and location interests as well as people's transition probability between locations, we mine the classical travel sequences from multiple users' location histories.
- We evaluated our methodology using a large GPS dataset, which was collected by 107 users over a period of one year in the real world. The number of GPS points exceeded 5 million and its total distance was over 160,000 kilometers.

The remainder of this paper is organized as follows. Section 2 gives an overview of our system. Section 3 presents the algorithms regarding location history modeling. Section 4 details the processes of location interest inference and classical travel sequence mining. In Section 4, we report on major experimental results and offer some discussions. Finally, in Section 5, we draw our conclusions and present the future work.

2. OVERVIEW OF OUR SYSTEM

In this section, we first clarify some terms used in this paper. Then, the architecture of our system is briefly introduced. Finally, we demonstrate the application scenarios of our system on desktops and GPS-phones using some snapshots of its user interfaces.

2.1 Preliminary

In this subsection, we will clarify some terms; including GPS log (*P*), GPS trajectory (*Traj*), stay point (*s*), location history (*LocH*), and tree-based hierarchical graph (*TBHG*).

Definition 1. GPS log: Basically, as depicted in the left part of Figure 1, a GPS log is a collection of GPS points $P=\{p_1, p_2, \ldots, p_n\}$. Each GPS point $p_i \in P$ contains latitude ($p_i.Lat$), longitude ($p_i.Lngt$) and timestamp ($p_i.T$).

Definition 2. GPS trajectory: As shown in the right part of Figure 1, on a two dimensional plane, we can sequentially connect these GPS points into a curve based on their time serials, and split this curve into GPS trajectories (*Traj*) if the time interval between consecutive GPS points exceeds a certain threshold ΔT. Thus, $Traj = p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$, where $p_i \in P$, $p_{i+1}.T > p_i.T$ and $p_{i+1}.T - p_i.T < \Delta T$ $(1 \leq i < n)$.

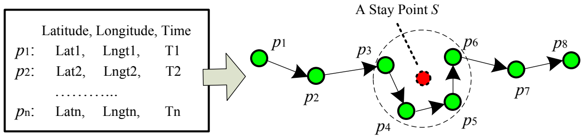


Figure 1. a GPS log, a GPS trajectory and a stay point

Definition 3. Stay point: A stay point *s* stands for a geographic region where a user stayed over a certain time interval. The extraction of a stay point depends on two scale parameters, a time threshold (T_{threh}) and a distance threshold (D_{threh}). Thus, like the points $\{p_3, p_4, p_5, p_6\}$ demonstrated in Figure 1, a single stay point *s* can be regarded as a virtual location characterized by a group of consecutive GPS points $P=\{p_m, p_{m+1}, \ldots, p_n\}$, where $\forall m < i \leq n$, $Distance(p_m, p_i) \leq D_{threh}$ and $|p_n.T - p_m.T| \geq T_{threh}$. Formally, conditioned by *P*, D_{threh} and T_{threh}, a stay point $s=(Lat, Lngt, arvT, levT)$, where

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

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

respectively stand for the average latitude and longitude of the collection *P*, and $s.arvT = p_m.T$ and $s.levT = p_n.T$ represent a user's arrival and leaving times on *s*.

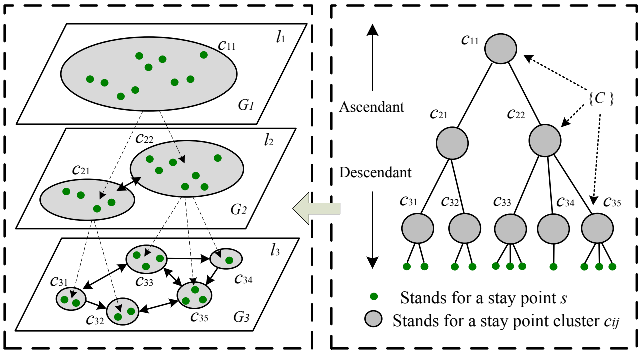
Typically, these stay points occur in the following two situations. One is that an individual remains stationary exceeding a time threshold. In most cases, this status happens when people enter a building and lose satellite signal over a time interval until coming back outdoors. The other situation is when a user wanders around within a certain geospatial range for a period. In most cases, this situation occurs when people travel outdoors and are attracted by the surrounding environment. As compared to a raw GPS point, each stay point carries a particular semantic meaning, such as the shopping malls we accessed and the restaurants we visited, etc.

Definition 4. Location history: Generally, a location history is a record of locations that an entity visited in geographical spaces over a period of time. In this paper, an individual's location history (*LocH*) is represented as a sequence of stay points (s) they visited with corresponding arrival and leaving times.

$$LocH = (s_1 \xrightarrow{\Delta t_1} s_2 \xrightarrow{\Delta t_2}, \dots, \xrightarrow{\Delta t_{n-1}} s_n);\ \Delta t_i = s_{i+1}.arvT - s_i.levT.$$

However, the location histories of various people are inconsistent and incomparable as the stay points pertaining to different individuals are not identical. To address this issue, we propose a structure, called tree-based hierarchical graph (*TBHG*), to model multiple users' location histories. Generally speaking, a *TBHG* is the integration of two structures, a tree-based hierarchy H and a graph G on each level of this tree. The tree expresses the parent-children (or ascendant-descendant) relationship of the nodes pertaining to different levels, and the graphs specify the peer relationships among the nodes on the same level.

As demonstrated in Figure 2, in our system two steps need to be performed when building a *TBHG*. 1) *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 (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. 2) *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 on one journey 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.



A Tree-Based Hierarchical Graph A Tree-Based Hierarchy

Figure 2. Building a tree-based hierarchical graph

Definition 5. Tree-Based Hierarchy H: H is a collection of stay point-based clusters C with a hierarchy structure L. $H = (C, L)$, $L = \{l_1, l_2, \dots, l_n\}$ denotes the collection of levels of the hierarchy and $C = \{c_{ij} | 1 \le i \le |L|, 1 \le j \le |C_i|\}$ means the collection of clusters on different levels. Here, c_{ij} represents the jth cluster on level $l_i \in L$, and C_i is the collection of clusters on level l_i.

Definition 6. Tree-Based Hierarchical Graph (***TBHG***): Formally, a *TBHG* is the integration of H and G, *TBHG*=(H, G). H is defined in *Definition 5*, and G={ $g_i = (C_i, E_i), 1 < i \le |L|$}. On each layer $l_i \in L$, $g_i \in G$ includes a set of vertexes C_i and the edges E_i connecting $c_{ij} \in C_i$.

Notations: In the rest of this paper, we use the following notations to simplify the descriptions. $U = \{u_1, u_2, \dots, u_n\}$ represents the collection of users in a community, $u_k \in U, 1 \le k \le |U|$denotes the kth user, and $P^k, Traj^k, S^k$and $LocH^k$ respectively stand for the u_k's GPS logs, GPS trajectory, stay points and location history.

2.2 Architecture

Figure 3 shows the architecture of our system, which is comprised of the following three parts; location history modeling, location interest and sequence mining, and recommendation. The first two operations can be performed off-line, while the last process should be conducted on-line based on the region specified by a user.

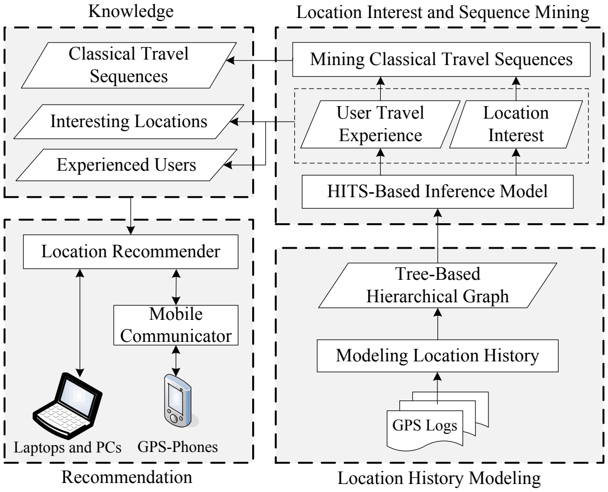

Knowledge
Location Interest and Sequence Mining
Classical Travel Sequences
Mining Classical Travel Sequences
Interesting Locations
User Travel Experience
Location Interest
Experienced Users
HITS-Based Inference Model
Location Recommender
Tree-Based Hierarchical Graph
Mobile Communicator
Modeling Location History
GPS Logs
Laptops and PCs
GPS-Phones
Recommendation
Location History Modeling

Figure 3. Architecture of our system

Location history modeling: Given multiple users' GPS logs, we build a *TBHG* off-line. In this structure, a graph node stands for a cluster of stay points, and a graph edge represents a directed transition between two locations (clusters). In contrast to raw GPS points, these clusters denote the locations visited by multiple users, hence would carry more semantic meanings, such as culturally important places and commonly frequented public areas. In addition, the hierarchy of the *TBHG* denotes different geospatial scales (alternatively, the zoom level of a Web map), like a city, a district and a community. In short, the tree-based hierarchical graph can effectively model multiple users' travel sequences on a variety of geospatial scales.

HITS-based inference model: With the *TBHG*, we propose a HITS-based inference model to estimate users' travel experiences and location interests in a given region. In this model, an individual's visit to a location (cluster) is regarded as a directed link from the individual to that location. Thus, a user is a hub if they have visited many locations, and a location is an authority if it has been accessed by many users. Further, a user's travel experience (hub score) and interest of a location (authority score) have a mutual reinforcement relationship. Using a power iteration method, we can generate the final scores for each user and location, and find out the top n interesting locations and the top k experience users in a given region. (See Section 4.2 for details)

As a user's travel experience is region-related, we need to specify a geospatial region as the context for the inference model. Actually, each cluster of the *TBHG* specifies an implied region for its descendant clusters (locations). Therefore, we are able to mine in advance each individual's travel experience and interests of locations conditioned by the regions of clusters on different levels. In other words, a user would have multiple hub scores based on different regions, and a location would have multiple authority scores specified by their ascendant clusters on different levels. This strategy takes the advantage of a HITS model in ranking locations and users based on a region context (query topic), while making the calculations of authority and hub scores offline.

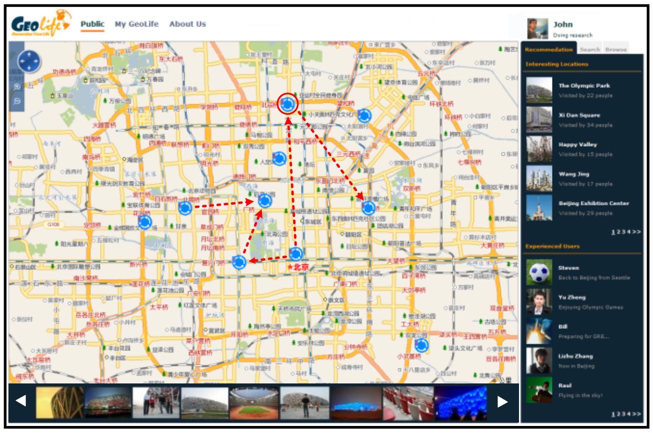
Mining classical travel sequence: We calculate a classical score for each location sequence within a given region considering two factors; the travel experiences of the users taking this sequence and the interests of the locations contained in the sequence. Since there would be multiple paths starting from a location, the interest of this location should be shared among all the paths, with which it points to other locations. The interest of a location to different paths is based on the probability of users' taking these paths. Later, the sequences with relatively high classical score will be retrieved as classical travel sequences. As people would not travel to too many places in a journey, classical sequences containing two or three locations would be more useful than longer ones.

Recommendation: By changing the zoom level of a Web map and moving the map, an individual can specify any geospatial regions with the present view of the map. This region can cover a whole country or a part of a city. With the received zoom level, our recommender can find out the corresponding level of hierarchy in the *TBHG*, and then collect the locations (clusters) fall in the given region on this level. The hub and authority scores conditioned by the first shared ascendant cluster by these locations will be used to rank locations and users (refer to Figure 9). Finally, the most *k* experienced users, top *n* interesting locations and top *m* classical travel sequences within the specified region can be returned to the users with desktop PCs and mobile devices like GPS-phones.

2.3 Application Scenarios

The work reported in this paper is an important component of our project GeoLife [12][19][20][21], whose prototype has been internally accessible within Microsoft since Oct. 2007. So far, we have had 107 individuals using this system.

Figure 4 shows its user interface for desktop computers. In the right part of this figure, we can view the top five interesting locations and the most five experienced users in the region specified by the present view of the Web map shown in the left part. Meanwhile, the top five classical travel sequences within this region are displayed on the Web map. By changing the zoom level and/or moving this Web map, an individual can retrieve such results within any regions. In addition, the photos taken at an interesting location will be presented on the bottom of the window after a user clicks the icon representing the location on the map. Also, the sub-locations contained in this region will be presented.



John
Doing research

Interesting Locations

The Olympic Park
Visited by 22 people

Xi Dan Square
Visited by 34 people

Happy Valley
Visited by 15 people

Wang Jing
Visited by 17 people

Beijing Exhibition Center
Visited by 29 people

1 2 3 4 >>

Experienced Users

Steven
Back to Beijing from Seattle

Yu Zheng
Enjoying Olympic Games

Bill
Preparing for GRE...

Lizhu Zhang
Now in Beijing

Raul
Flying in the sky!

1 2 3 4 >>

Figure 4. The user interface regarding location recommendation

As shown in Figure 5, a user with a GPS-phone can find out the top five interesting locations as well as the most five classical sequences nearby their present geographic position (the red star). In addition, when the user reaches a location, our system would provide them with a further suggestion by presenting the top three classical sequences start from this location.

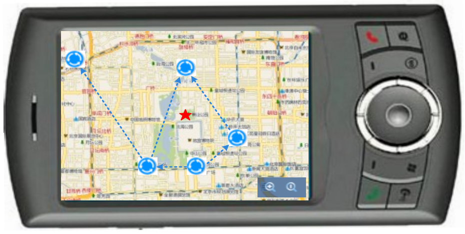


Figure 5. Location recommendations on a GPS-phone

3. Modeling Location History

In this Section, we detail the process of modeling multiple users' location histories with a tree-based hierarchical graph. Figure 6 gives a formal description of this operation.

```
Algorithm LocHisModeling(φ, $D_{thresh}$, $T_{thresh}$)
Input: The collection of users' GPS logs: $\varphi = \{P^k, 1 \le k \le |U|\}$.
Output: a tree-based hierarchical graph: TBHG
Foreach $u_k \in U$ do
    $Traj^k$= LogParsing($P^k$);
    $S^k$= StayPointDetection( $Traj^k$, $D_{thresh}$, $T_{thresh}$);
    $LocH^k$=PersonalLocHis($S^k$);  // individual location history
    SP.Add( $S^k$);                   //the collection of stay points
H= HierarchicalClustering (SP);
Foreach $l_i \in H.L$ do              // build a graph on each level
    $g_i.C_i = H.C_i$;
    Foreach $u_k \in U$ do
        $g_i$ = GraphBuilding( $g_i$, $LocH^k$ );
    G.Add($g_i$);
TBHG=(H, G);
Return TBHG;
```

Figure 6. The procedure of modeling users' location histories

First, for each user $u_k \in U$, we parse their GPS logs (P^k) into GPS trajectories ($Traj^k$), and extract stay points (S^k) from each trajectory by seeking the spatial regions where u_k spent a period exceeding a certain threshold (refer to [12] for details). Then, u_k can formulate a location history ($LocH^k$) with these stay points.

Second, we put these stay points together into a dataset $SP = \{S^k, 1 \le k \le |U|\}$. Using a density-based clustering algorithm, this dataset SP will be hierarchically clustered into several geospatial regions C in a divisive manner. Thus, the similar stay points from various users will be assigned to the same clusters on different levels of the hierarchy. In addition, we would filter away the clusters, which might represent users' homes. If an individual has visited to a cluster with a frequency exceeding a threshold, we believe this cluster may be the individual's home or working place.

Third, with the tree-based hierarchy H and each user's location history $LocH^k$, we build connections among the clusters on the same level. A directed link would be generated for two clusters if they contain consecutive stay points pertaining to an individual's location history.

4. LOCATION INTEREST INFERENCE

In this Section, we first give a brief introduction on the key idea of HITS. Second, we describe *our* HITS-based inference model. Third, by involving such inference results, we mine the classical travel sequences from each graph of the *TBHG*.

4.1 Basic Concepts of HITS

HITS stands for hypertext induced topic search, which is a search-query-dependent ranking algorithm for Web information retrieval.

When the user enters a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings for the expanded set of pages, authority ranking and hub ranking. For every page in the expanded set, HITS assigns them an authority score and a hub score. As shown in Figure 7, an authority is a Web page with many in-links, and a hub is a page with many out-links. The key idea of HITS is that a good hub points to many good authorities, and a good authority is pointed to by many good hubs. Thus, authorities and hubs have a mutual reinforcement relationship. More specifically, a page's authority score is the sum of the hub scores of the pages it points to, and its hub score is the integration of authority scores of the pages pointed to by it. Using a power iteration method, the authority and hub scores of each page can be calculated. The main strength of HITS is ranking pages according to the query topic, which may provide more relevant authority and hub pages. However, HITS needs some time consuming operations, such as on-line expanding page sets and calculating the hub and authority scores.

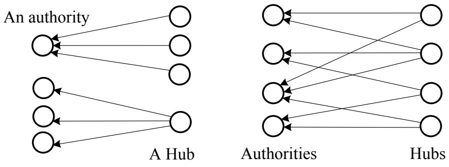


Figure 7. The basic concept of HITS model

4.2 Our HITS-Based Inference Model

4.2.1 Model Description

Using the third level of the *TBHG* shown in Figure 2 as a case, Figure 8 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 trajectories, 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 relationship.

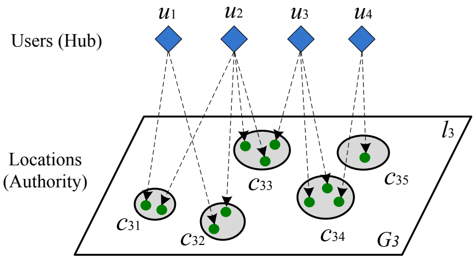


Figure 8. Our HITS-based inference model

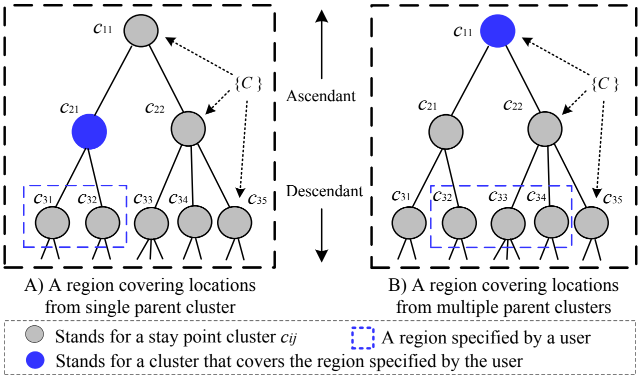
4.2.2 Strategy for Data Selection

Intrinsically, a user's travel experience is region-related, i.e., a user who has much travel knowledge in a city might have no idea about another city. Also, an individual, who has visited many places in a part of a city, might know little about another part of the city (if the city is very large, like New York). This feature is aligned with the query-dependent property of HITS. Thus, before conducting the HITS-based inference, we need to specify a geospatial region (a topic query) for the inference model and formulate a dataset that contains the locations falling in this region. However, using an online data selection strategy, (i.e., specify a region based on an individual's input), we need to perform lots of time consuming operations, which may reduce the feasibility of our system. Actually, on a level of the *TBHG*, the shape of a graph node (cluster of stay points) provides an implicit region for its descendent nodes. These regions covered by clusters on different level of the hierarchy might stand for various semantic meanings, such as a city, a district and a community. Therefore, we are able to calculate in advance the interest of every location using the regions specified by their ascendant clusters. In other words, a location might have multiple authority scores based on the different region scales it falls in. Also, a user might have multiple hub scores conditioned by the regions of different clusters.

Definition 7. Location Interest: In our system, the interest of a location (c_{ij}) is represented by a collection of authority scores $a_{ij} = \{a_{ij}^1, a_{ij}^2, \dots, a_{ij}^l\}$. Here, a_{ij}^l denotes the authority score of cluster c_{ij} based on the region specified by its ascendant nodes on level l, where $1 \le l \le i-1$.

Definition 8. User Travel Experience: In our system, a user's (e.g., u_k) travel experience is represented by a set of hub scores $h^k = \{h_{ij}^k | 1 \le i < |L|, 1 \le j \le |C_i|\}$ (refer to definition 5), where h_{ij}^k denotes u_k's hub score conditioned by the region of c_{ij}.

Figure 9 gives a demonstration of these definitions. In the region specified by cluster c_{11}, we can respectively calculate an authority score (a_{21}^1 and a_{22}^1) for cluster c_{21} and c_{22}. Meanwhile, within this region, we are able to infer authority scores (a_{31}^1, a_{32}^1, a_{33}^1, a_{34}^1 and a_{35}^1) for cluster $c_{31}, c_{32}, c_{33}, c_{34}$ and c_{35}. Further, using the region specified by cluster c_{21}, we can also calculate another authority score (a_{31}^2 and a_{32}^2) for c_{31} and c_{32}. Likewise, the authority scores (a_{33}^2, a_{34}^2 and a_{35}^2) of c_{33}, c_{34} and c_{35} can be re-inferred with the region of c_{22}. Therefore, each cluster on the third level has two authority scores, which would be used in various occasions based on users' inputs. For instance, as depicted in the Figure 9 A), when a user selects a region only covering location $c_{31}$and c_{32}, the authority score a_{31}^2 and a_{32}^2 can be used to rank these two locations. However, as illustrated in Figure 9 B), the region selected by a user covers the locations from two different parent clusters (c_{21} and c_{22}). At this moment, the authority value a_{32}^1, a_{33}^1 and a_{34}^1 should be used to rank these locations.


c_{11}
$\{C\}$
c_{21}
c_{22}
c_{31}
c_{32}
c_{33}
c_{34}
c_{35}
Ascendant
Descendant
● Stands for a stay point cluster c_{ij}
▭ A region specified by a user
● Stands for a cluster that covers the region specified by the user

A) A region covering locations from single parent cluster

B) A region covering locations from multiple parent clusters

Figure 9. Some cases demonstrating the data selection strategy

The strategy, which sets multiple hub scores for a user and multiple authority scores for a location, has the following two advantages. First, we are able to leverage the main strength of HITS to rank locations and users with the contexts of geospatial region (query topic). Second, these hub and authority scores can be calculated offline. Therefore, we can ensure the efficiency of our system while allowing users specify any regions on a map.

4.2.3 *Inference*

Given the set of locations pertaining to the same ascendant cluster, we are able to build an adjacent matrix *M* between users and locations based on the users' accesses on these locations. In this matrix, an item v_{ij}^k stands for the times that u_k (a user) has visited to cluster c_{ij} (the *j*th cluster on the *i*th level). Such matrixes can be built offline for each non-leaf node. For example, the matrix *M* formulated for the case shown in Figure 8 can be represented as follows, where all the five clusters pertain to c_{11}.

$$M = \begin{array}{c c} & \begin{array}{ccccc} c_{31} & c_{32} & c_{33} & c_{34} & c_{35} \end{array} \\ \begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \qquad (3)$$

Then, the mutual reinforcement relationship of user travel experience h_{ij}^k and location interest a_{ij}^l is represented as follows:

$$a_{ij}^l = \sum_{u_k \in U} v_{ji}^k \times h_{lq}^k; \qquad (4)$$

$$h_{lq}^k = \sum_{c_{ij} \in c_{lq}} v_{ij}^k \times a_{ij}^l; \qquad (5)$$

Where c_{lq} is c_{ij}'s ascendant node on the *l*th level, $1 \leq l < i$. For instance, as shown in Figure 9, c_{31}'s ascendant node on the first level of the hierarchy is c_{11}, and its ascendant node on the second level is c_{21}. Thus, if $l = 2$, c_{lq} stands for c_{21} and $(c_{31}, c_{32}) \in c_{21}$. Also, if $l = 1$, c_{lq} denotes c_{11}, $(c_{31}, c_{32}, \ldots, c_{35}) \in c_{11}$.

Writing them in the matrix form, we use ***a*** to denote the column vector with all the authority scores, and use ***h*** to denote the column vector with all the hub scores. Conditioned by the region of cluster c_{11}, ***a***=$(a_{31}^1, a_{32}^1, \ldots, a_{35}^1)$, and ***h***=$(h_{11}^1, h_{11}^2, \ldots, h_{11}^4)$.

$$\boldsymbol{a} = \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{h} \qquad (6)$$

$$\boldsymbol{h} = \boldsymbol{M} \cdot \boldsymbol{a} \qquad (7)$$

If we use $\boldsymbol{a}_n$ and $\boldsymbol{h}_n$ to denote authority and hub scores at the *n*th iteration, the iterative processes for generating the final results are

$$\boldsymbol{a}_n = \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{M} \cdot \boldsymbol{a}_{n-1} \qquad (8)$$

$$\boldsymbol{h}_n = \boldsymbol{M} \cdot \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{h}_{n-1} \qquad (9)$$

Starting with $\boldsymbol{a}_0 = \boldsymbol{h}_0 = (1,1,\ldots,1)$, we are able to calculate the authority and hub scores using the power iteration method.

```
Algorithm LocationInterestInference ($TBHG$, $LocH$)
Input:   A tree-based hierarchy graph $TBHG$=($H$, $G$), and
         collection of users' location histories $LocH$
Output: the collection of users' hub scores, $\boldsymbol{h}$, and the collection of
         locations' authority scores, $\boldsymbol{a}$.
$\boldsymbol{h}$=$\boldsymbol{a} = \emptyset$;
For $i = 1; i < |L|; i++$                //for each level
    For $j = 1; j \leq |C_i|; j++$       // for each cluster on this level
        For $x = i + 1; x \leq |L|; x++$ //search the descendant levels
            $C_x'$=LocationCollecting ($x$, $c_{ij}$, $H$);
            $M$=MatrixBuilding($C_x'$, $LocH$);
            ($\{h_{ij}^k\}$,$\{a_x^i\}$)=HITS-Inference($M$);
            $\boldsymbol{a} = \boldsymbol{a} \cup \{a_x^i\}$;
            $\boldsymbol{h} = \boldsymbol{h} \cup \{h_{ij}^k\}$;
Return ($\boldsymbol{h}$, $\boldsymbol{a}$);
```

Figure 10. The algorithm for inferring the authority and hub scores

Figure 10 depicts an off-line algorithm for inferring each user's hub scores and the authority scores of each location conditioned by the different regions. Here C_x is the collection of clusters on *x*th level. $C_x' \subset C_x$ denotes the collection of c_{ij}'s descendant clusters on the *x*th level. For instance, the C_2' of c_{11} is $\{c_{21}, c_{22}\}$, and C_3' of c_{11} is $\{c_{31}, c_{32}, \ldots, c_{35}\}$. $\{a_x^i\}$ represents the collection of authority scores of the locations contained in C_x conditioned by their ascendant node on the *i*th level.

4.3 Mining Classical Travel Sequences

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.

Using a graph of *TBHG*, Figure 11 demonstrates the calculation of the classical score for a 2-length sequence, A→C. In this figure, the graph nodes (A, B, C, D and E) stand for locations, and the graph edges denote people's transition sequences among them. The number shown on each edge represents the times users have taken the sequence.

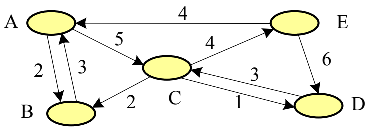


Figure 11. Demonstrating classical sequence mining with a graph

Equation (10) presents the classical score of sequence A→C, which includes the following three parts. 1) The authority score of location A (a_A) weighted by the probability of people's moving out by this sequence (Out_{AC}). Clearly, there are seven (5+2) links point out to other nodes from node A, and five out of seven of these links direct to node C. So, $Out_{AC} = \frac{5}{7}$, i.e., only five sevenths of location A's authority (a_A) should be offered to sequence A→C, and the rest of a_A should be provided to A→B. 2) The authority score of location C (a_C) weighted by the probability of people's moving in by this sequence (In_{AC}). 3) The hub scores of the users (U_{AC}) who have taken this sequence.

$$\begin{aligned} S_{AC} &= \sum_{u_k \in U_{AC}} (a_A \cdot Out_{AC} + a_C \cdot In_{AC} + h^k) \\ &= |U_{AC}| \cdot (a_A \cdot Out_{AC} + a_C \cdot In_{AC}) + \sum_{u_k \in U_{AC}} h^k \\ &= 5 \times \left(\frac{5}{7} \times a_A + \frac{5}{8} a_C\right) + \sum_{u_k \in U_{AC}} h^k. \qquad (10) \end{aligned}$$

Following this method, we are able to calculate the classical score of sequence C→D, $S_{CD} = 1 \times \left(\frac{1}{7} \times a_C + \frac{1}{7} a_D\right) + \sum_{u_k \in U_{CD}} h^k$. Thus, the classical score of sequence A→C→D equals to:

$$S_{ACD} = S_{AC} + S_{CD}. \qquad (11)$$

Using this paradigm we are able to calculate the classical score of any *n*-length sequences. Later, the top *m n*-length sequences with relatively high scores can be retrieved as *n*-length classical travel sequences. However, it is not necessary to find out the classical sequences with long length, as people would not visit many places in a trip. Moreover, the process of searching for *n*-length classical sequences is time consuming, although this operation can be performed offline. Thus, in this paper, we start with mining 2-length classical sequences, and then try to find out some 3-lenth classical sequences by extending these 2-length sequences.

5. EXPERIMENTS

In this Section, we first present the experimental settings. Second, we introduce the evaluation approaches. Third, some major results are reported followed by some discussions.

5.1 Settings

5.1.1 GPS Devices and Data Collectors

Figure 12 shows the GPS devices we chose to collect data. They are comprised of stand-alone GPS receivers (Magellan Explorist 210/300, G-Rays 2 and QSTARZ BTQ-1000P) and GPS phones. Except for the Magellan 210/300, these devices are set to receive GPS coordinates every two seconds. Carrying these GPS-enabled devices, 107 users (49 females and 58 males) recorded their outdoor movements with GPS logs from May 2007 to Oct. 2008. Figure 13 shows these users' demographic statistics.

Figure 12. GPS devices used in our experiment

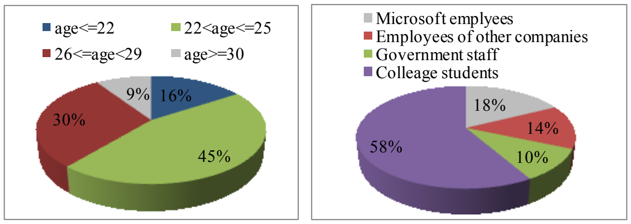
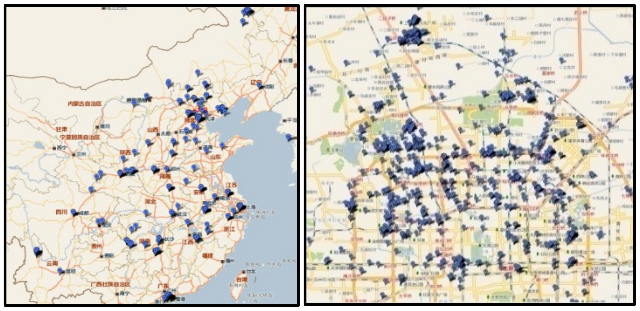


Figure 13. Demographic statistics of our experiment

5.1.2 GPS Data

Figure 14 depicts the distributions of the GPS data we used in the experiments. Most parts of this dataset were created in Beijing, China, and other parts covered 36 cities in China as well as a few cities in the USA, South Korea, and Japan. The volunteers were motivated to log their outdoor movements as much as possible by the payments based on the distance of GPS trajectories collected by them; the more data collected by them, the more money they obtained. As a result, the total distance of the GPS logs exceeded 166,372 kilometers, and the total number of GPS points reached 5,081,369. Considering the privacy issues, we use these datasets anonymously.



A) Data distribution in China B) Data distribution in Beijing

Figure. 14 Distribution of the GPS dataset we used in this experiment

5.1.3 Parameter Selection

Stay point detection: In this experiment, we set T_{threh} to 20 minutes and D_{threh} to 200 meters for stay point detection. In other words, if an individual stays over 20 minutes within a distance of 200 meters, a stay point is detected. These two parameters enable us to find out some significant places, such as restaurants and shopping malls, etc., while ignoring the geo-regions without semantic meaning, like the places where people wait for traffic lights or meet congestion. As a result, we extracted 10,354 stay points from the dataset.

Clustering: We use a density-based clustering algorithm, OPTICS (Ordering Points To Identify the Clustering Structure), to hierarchically cluster stay-points into geospatial regions in a divisive manner. As compared to an agglomerative method like K-Means, the density-based approach is capable of detecting clusters with irregular structures, which may stand for a set of nearby restaurants or shopping streets. In addition, this approach would filter out a few sparsely distributed stay points, and ensure each cluster has been accessed by some users. As a result, a four-level *TBHG* is built based on our dataset (see Table I for details).

Table I. Information of the *TBHG* used in the experiment

	Num. of Clusters	**Ave. size of clusters KM**	**Ave. num of user/cluster**	**Ave. num stay points/cluster**
Level 1	1	11,450.7	107	10,354
Level 2	32	14.5	6.7	267.5
Level 3	70	2.1	8	112.7
Level 4	159	0.26	6.5	46.2

5.2 Evaluation Approaches

5.2.1 Framework of the Evaluation

Figure 15 illustrates the framework of the evaluation, in which we respectively explore the effectiveness of location & travel sequence recommendation by performing a user study. In this study, 29 subjects (14 females and 15 males), who have been in Beijing for more than 6 years, were invited to answer the evaluation questions. Using the region specified by the fourth ring road of Beijing, we retrieved the top 10 interesting locations and the top 5 classical travel sequences based on a variety of approaches, including our methods and some baselines. As the subjects are familiar with this region, we are more likely to find out common ground truths shared by them.

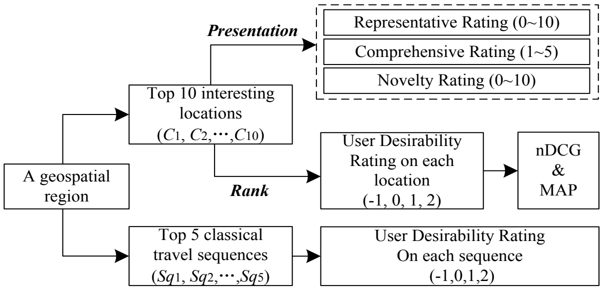

Presentation
Representative Rating (0~10)
Comprehensive Rating (1~5)
Novelty Rating (0~10)
Top 10 interesting locations (C_1, C_2,⋯,C_{10})
A geospatial region
Rank
User Desirability Rating on each location (-1, 0, 1, 2)
nDCG & MAP
Top 5 classical travel sequences (Sq_1, Sq_2,⋯,Sq_5)
User Desirability Rating On each sequence (-1,0,1,2)

Figure 15. Framework of the evaluations

Regarding the interesting locations, we conduct the following two aspects of evaluations. One is the ***Presentation***, which stands for the ability of the retrieved interesting locations in presenting a given region. The other is the ***Rank***, which represents the ranking performance of the retrieved locations based on relative interests.

1) Presentation. In this aspect, each subject had to answer the following evaluation questions:

- *Representative*: How many locations in this retrieved set are representative of the given region (0-10)?
- *Comprehensive*: Do these locations offer a comprehensive view of the given region (1-5)?
- *Novelty:* How many locations in this retrieved set have interested you even though they only appeared recently (0-10)? In the study, the subjects were able to view the points of interests (POIs) falling in each location as well as the photos taken there.

2) Rank. Each subject had to individually rate the interest of each retrieved location with a value (-1~2) shown in table II. Then, we aggregated these subjects' ratings for each location, and select the mode of the ratings for the location. If the mode of two rating levels is identical, we prefer the lower ratings. For example, if fifteen subjects rated a location with 2 and ten rated it with 1, we regard the subjects' rating on this location as 2 (since 2 is the most frequently occurring rating in the dataset). However, if the number of subjects rating a location with 2 is the same with those rating it with 1, the final aggregated rating should be 1.

Table II. Users' interests in a location

Ratings	Explanations
2	I'd like to plan a trip to that location.
1	I'd like to visit that location if passing by.
0	I have no feeling about this location, but don't oppose others to visit it.
-1	This location does not deserve to visit.

With regard to evaluating the retrieved classical travel sequences, we required the subjects to rate each sequence in the set with the scores shown in table III. Also, we aggregated these subjects' rating for each sequence like the method mentioned above.

Table III. Users' interests in a travel sequence

Ratings	Explanations
2	I'd like to plan a trip with this travel sequence.
1	I'd like to take that sequence if visiting the region.
0	I have no feeling about this sequence, but don't oppose others to choose it.
-1	It is not a good choice to select this sequence.

5.2.2 Measurements

Measurements for presentation: We compare our method with the baselines using the mean score of the ratings offered by the subjects. In addition, we perform a T-test for each comparison to justify the significant advantages of our method.

Measurements for ranking: We employ two criteria, *nDCG* (normalized discounted cumulative gain) and *MAP* (Mean Average Precision), to measure the ranking performance of the retrieved interesting locations. *MAP* is the most frequently used summary measure of a ranked retrieval run. In our experiment, it stands for the mean of the precision score after each interesting location is retrieved. Here, a location is deemed as an interesting location if its interest level equals to 2. For instance, the *MAP* of an interest rating vector, $G =< 2, 0, 2,0,1, 0, 0, 2, 0, -1 >$, for the top 10 location, is computed as follows.

$$MAP = \frac{1 + 2/3 + 3/8}{3} = 0.681$$

nDCG is used to compute the relative-to-the-ideal performance of information retrieval techniques. The discounted cumulative gain of *G* is computed as follows: (In our experiments, b = 3.)

$$CG[i] = \begin{cases} G[1], & if\ i = 1 \\ DCG[i-1] + G[i], & if\ i < b \\ DCG[i-1] + \dfrac{G[i]}{log_b i}, & if\ i \geq b \end{cases}$$

Given the ideal discounted cumulative gain *DCG'*, then *nDCG* at *i*-th position can be computed as $NDCG[i] = DCG[i]/DCG'[i]$.

Measurement for classical sequence: We used the mean score of these subjects' ratings, along with a T-test for each comparison, to distinguish our method from baselines. At the same time, we investigated the *classical rate*, which represents the ratio of sequences with a score of 2 in the set, of different methods.

5.2.3 Baselines

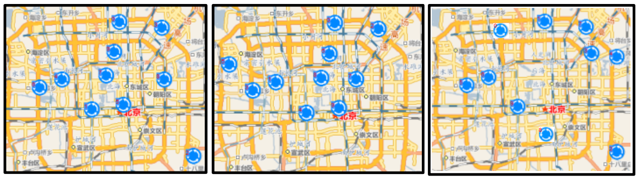
Baselines for mining interesting locations: Here, we explore the effectiveness of two baseline methods, *rank-by-count* and *rank-by-frequency*. Regarding the former one, the more users visiting a location the more interesting this location might be. In the latter, the more frequent people accessed a location the more interesting this location might be. The visited frequency of a location is the ratio between the number of the users visiting this location and the time span, from the first day one user accessed this location to the last day at least one individual visited it.

Baselines for mining classical travel sequences: We compare our method with three baselines; *rank-by-count*, *rank-by-interests* and *rank-by-experience*. With regard to the first baseline, we rank a sequence based on the number of the users who have taken this sequence. Regarding the second one, we only take into account the interests of the locations contained in a sequence to rank the travel sequences. In the third baseline method, we only consider the experiences of the users who have taken this sequence.

5.3 Results

5.3.1 Results Related to Interesting Locations

Presentation ability: Figure 16 illustrates the top 10 interesting locations, which were respectively inferred out by our method and two baselines using the region within the fourth ring road of Beijing (the zoom level corresponds to the 3rd level of the *TBHG*).



A) Our method B) *Rank-by-count* C) *Rank-by-frequency*

Figure 16. Top 10 interesting locations of different approaches

Based on these results, 29 subjects individually answered the evaluation questions with the ratings mentioned in Table II. As shown in Table IV, our method is more capable than the baselines of finding out representative locations in the give region (T-test result: p_1<0.01, p_2<0.01). Meanwhile, the top 10 locations retrieved by our method presented a more comprehensive view of this region over the baselines (p_1<<0.01, p_2<<0.01). In addition, using our method, more novel locations that interest the subjects have been retrieved (p_1<0.01, p_2<0.01). These regions represent the development of new Beijing, while having not been noticed by many people. Regarding the baselines, *Rank-by-count* outperformed *rank-by-frequency* in finding out the representative locations (p<0.01) and presenting a comprehensive view of the region (p<0.01). However, the former method does not show a clear advantage beyond the latter in detecting the novel interesting locations (p>0.2).

Table IV. Comparison on the presentation ability of different methods

	Ours	*Rank-by-count*	*Rank-by-frequency*
Representative	**5.4**	4.5	3.1
Comprehensive	**4**	3.4	2.3
Novelty	**3.4**	2.4	2.2

Ranking ability: Table V depicts the ranking ability of different methods using *nDCG*@5, *nDCG*@10 and *MAP* as measurements. Although the set of interesting locations retrieved by our method and *rank-by-count* had a 60 percents overlap, our method showed clear advantages beyond baseline methods in effectively ranking this location set.

Table V. Ranking ability of different methods

	Ours	***Rank-by-count***	***Rank-by-frequency***
nDCG@5	**0.823**	0.714	0.598
nDCG@10	**0.943**	0.848	0.859
MAP	**0.759**	0.532	0.365

5.3.2 Results Related to Classical Sequences

Using two measurements (mean score and classical rate), Table VI differentiates the performance of our method from the baselines in finding out the classical sequences in the given region. Clearly, our method considering both users' travel experiences and location interests outperformed *rank-by-count* (p<<0.01), *rank-by-interest* (p<0.01) and *rank-by-experience* (p<0.01). Meanwhile, when respectively taking into account users' travel experiences (p<0.01) or location interests (p<0.01), the performance of *rank-by-counts* had been significantly improved. These results proved that user travel experience and location interests respectively play an important role in retrieving the classical travel sequences and offered a greater contribution when being used together. (See 5.2.2 for the meaning of *classical rate*)

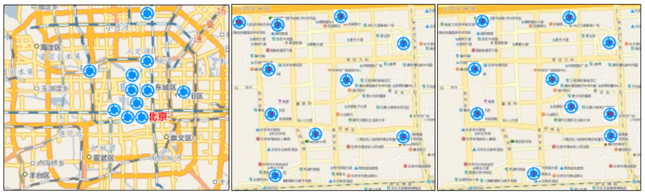
Table VI. Performance of different methods in finding classical sequences

	Ours (Interest + Experience)	***Rank-by-counts***	***Rank-by-interest***	***Rank-by-experience***
Mean score	**1.6**	1.2	1.4	1.5
Classical Rate	**0.6**	0.3	0.4	0.4

5.3.3 Investigations into Our Method

The advantages of the hierarchy of the *TBHG* lie in two aspects. 1) It would offer a comprehensive view of a large region (a city) and help users understand the region step-by-step (level-by-level). 2) The hierarchy can be used to specify users' travel experiences in different regions. Hence, we are more likely to effectively retrieve interesting locations in a region.

First, as shown in Figure 17 A), 8 out of the top 10 interesting locations (within the fourth ring road of Beijing) have fallen into the town center if these locations are ranked according to their authority scores inferred out on the bottom level of the *TBHG* (i.e., no hierarchy). In contrast to Figure 16 A), Figure 17 rushes users into low level details of Beijing, while ignoring a more comprehensive view of this city. Failing in understanding the city step-by-step, people would not gain some high-level conceptual information related to Beijing at the very beginning.



A) Inferring the top 10 interesting locations without using hierarchy

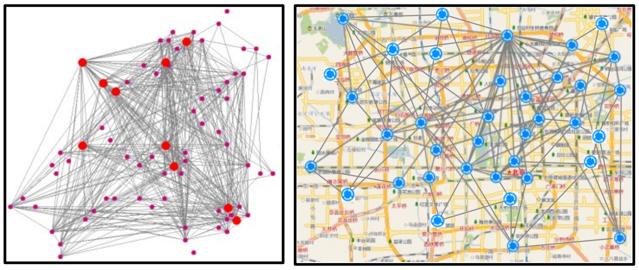
B) Ranking the locations using the authority scores of the region

C) Ranking locations using their authority scores of the whole Beijing

Figure 17. Investigations into our method

Second, Figure 17 B) presents the top 10 interesting locations ranked by their authority scores conditioned by the region shown in this figure (these locations pertain to the same ascendant cluster on the 3rd level). As compared to Figure 17 C) where the locations are ranked in terms of the authority scores conditioned by the whole city (i.e., without hierarchy), Figure 17 B) provided a more *comprehensive* view of this region (4.1>3.1, p<0.01) and retrieved more *representative* locations (6.8>4.7, p<0.01) in the region. In addition, the results shown in Figure 17 B) outperformed that of Figure 17 C) in effectively ranking the retrieved locations (*nDCG@5* 0.86>0.67). Intrinsically, users' travel knowledge is region-related, e.g., some individuals, who are familiar with most parts of Beijing, might know little about the region shown in Figure 17 B). Thus, we cannot directly use these individuals' hub scores to infer the authority scores of the locations in this region.

Figure 18 A) demonstrates the correlations between users in the GeoLife community. A link was generated to connect two users (a node) if they had visited 5+ of the same locations. The relatively big nodes denote the top ten experienced users in Beijing. Figure 18 B) shows the correlation among locations using the sequences generated by multiple users in Beijing. Such information would further enable us to explore the social networking in geo-related community and understand locations based on users' GPS logs.



A) Relations between users B) Correlations between locations

Figure 18. Correlation between locations and users

5.4 Discussions

5.4.1 Discussion on Interesting Locations

With data shown in Table IV, we observe that users' travel experiences are useful in not only retrieving representative locations in a region but also finding out more novel and interesting locations beyond baseline methods. Intuitively, some interesting places, which contain high-quality restaurants or nice shopping malls developed recently, would not be visited by many people. However, a location covering some landmarks, which is not that interesting but with a relatively long history, might be accessed by more people. Hence, the *rank-by-count* cannot handle this kind of problem well. Meanwhile, a user would frequently access the restaurant nearby their working place for convenience rather than food quality or having fun. Therefore, a location frequently visited by people might not be interesting.

5.4.2 Discussion on Classical Sequences

The results shown in Table V justify the contributions of users' travel experiences and location interests in mining classical travel sequences. First, intuitively, without considering such information, the sequence from a railway station to a nearby hotel might be detected as a classical travel sequence because some tourists would live nearby the station. Obviously, this is not a good recommendation for users. Second, if only using individuals' travel experiences, we would mine out some life routine of an experienced user rather than classical sequences. For instance, sometimes, an experienced user would have dinner at a restaurant nearby their home and then go to a supermarket not far away from this restaurant. Since the user has a relatively high hub score, their life routine, like from the restaurant to the supermarket, might be detected as a classical travel sequence. Third, if only considering location interest, some ineffective sequences would be found out. For example, the Summer Palace and the Forbidden City are two very interesting locations in Beijing. An experienced user would not visit them in a sequence as they are far from each other and everyone deserves a one day tour. However, a few tourists without much travel knowledge about Beijing might carelessly visit these two places in a sequence, hence make this sequence classical.

6. RELATED WORK

6.1 Mining Location History

***Mining individual location history*:** During the past years, a branch of research [5][9][11][13] has been performed based on individual location history represented by GPS trajectories. These works include detecting significant locations of a user [5][9], predicting the user's movement among these locations [5][13], and recognizing user-specific activities at each location [15]. As opposed to these works, we aim to model multiple users' location histories and learn patterns from numerous individuals' behaviors.

***Mining multiple users' location histories*:** Gonotti et al. [8] mined similar sequences from users' moving trajectories, and Mamoulis et al. [16] proposed a framework for retrieving maximum periodic patterns in spatio-temporal data. MSMLS [11] used a history of a driver's destinations, along with data about driving behavior extracted from multiple users' GPS trajectories, to predict where a driver may be going as a trip progresses. Eagle et al [7] aimed to recognize the social pattern in daily user activity from the dataset collected by 100 users with a Bluetooth-enabled mobile phone. In contrast to these techniques, we extend the paradigm of mining multiple users' location histories from exploring users' behaviors to understanding locations as well as modeling the relationship between users and locations.

6.2 Location Recommenders

***Recommenders based on real-time location*:** Mobile tourist guide systems [4][6][14][17] typically recommend locations and sometimes provide navigation information based on a user's real-time location. Previously, such kinds of systems were somehow naïve as they always returned the information close to an individual without understanding the individual and the nearby locations. Recently, some researchers aim to filter away from the returned results the invisible entities occluded by the nearby building [6][17]. Meanwhile, another branch of work [4][14] started involving a user's location history in these systems to provide the user with a more personalized recommendation. In contrast to these techniques, we aim to integrate social networking into the mobile tourist guide systems, by helping each individual deeply understand the locations around them with the knowledge mined from multiple users' location histories.

***Recommenders based on location history*:** Using multiple users' real-world location histories, some recommender systems, such as *Geowhiz* [10] and *CityVoyager* [18], etc, have been designed to recommend geographic locations like shops or restaurants to users. Horozov et al. [10] proposed an enhanced collaborative filtering solution to generate the recommendation of a restaurant. Takeuchi et al. [18] attempted to recommend shops to users based on their individual preferences estimated by analyzing their past location histories. The major difference between these works and ours lies in two aspects. First, we differentiate the travel experiences of various users. Second, we consider the relationship between locations and users' travel experiences, e.g., the mutual reinforcement relationship and the region-related constraints.

7. CONCLUSION

In this paper, using the GPS trajectories generated by multiple users, we mined interesting locations and classical travel sequences within a given geospatial region. Such information can help us understand the correlation between users and locations, and enable travel recommendation as well as mobile tourist guidance. In this work, we regard an individual's visit to a location as a link from the individual to the location, and weight these links in terms of users' travel experiences in various regions. A HITS-based model is proposed to infer a user's travel experience and the interest of a location considering the following two aspects. One is the mutual reinforcement relationship between location interest and user travel experience. The other is that user travel experience as well as location interest are region-related. Later, we detected the classical travel sequences in a specified region using location interests and users' travel experiences. We evaluated our method with a real-world GPS dataset created by 107 users over a period of 1 year. As a result, our method showed clear advantages beyond *rank-by-count* and *rank-by-frequency* by providing a better presentation ability and ranking performance. Meanwhile, when employing both users' travel experiences and location interests, we achieved the best performance.

In the future, we would like to improve the efficiency of sequence mining. Also, grouping users based on their location histories or clustering locations in terms of people's visits are potential works.

8. REFERENCES

[1] Bikely: http://www.bikely.com/

[2] GPS Track route exchange forum: http://www.gpsxchange.com/

[3] GPS sharing: http://gpssharing.com/.

[4] Abowd, G. D. Cyberguide: a mobile context-aware tour guide, wireless network, 3(5), 421-433.

[5] Ashbrook, D., and Starner, T. Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7(5), 275-286.

[6] Beeharee, A. et al. Exploiting real world knowledge in ubiquitous applications. Personal and Ubiquitous Computing 11(6), 429-437.

[7] Eagle, N. et al. Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255-268.

[8] Gonotti, F., et al. Trajectory pattern mining. In Proceedings of KDD'07 (San Jose USA, Aug. 2007), ACM Press, 330-339

[9] Hariharan, R. et al. Project Lachesis: Parsing and Modeling Location Histories, In Proceedings of GIScience, (Park Utah, October 2004), ACM Press: 106-124.

[10] Horozov, T., et al. Using Location for Personalized POI Recommendations in Mobile Environments. In Proceedings of SAINT, (Phoenix, USA, Jan. 2006), IEEF Press: 124-129.

[11] Krumm, J.et al. Predestination: Inferring Destinations from Partial Trajectories. In Proceedings of the Ubicomp'03, (Orange County USA, September 2003). Springer Press: 243-260.

[12] Li, Q. and Zheng, Y. et al. Mining user similarity based on location history. In Proc. of GIS'08 (Santa Ana, CA, Nov. 2008). ACM Press: 298-307

[13] Liao, L., et al. Building Personal Maps from GPS Data. In proceedings of IJCAI MOO05, Springer Press(2005): 249-265

[14] Park, M., H. Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices. In Proc. UIC'07 (Hong Kong, China, July 2007). Springer Press:1130-1139

[15] Patterson, D., J. et al. Inferring High-Level Behavior from Low-Level Sensors. In Proc. of Ubicomp'03, Springer Press (2003), 73-89

[16] Mamoulis, N. et al. Mining, Indexing and Querying Historical Spatiotemporal Data. In Proceedings of KDD'04 (Seattle USA, August 2004), ACM Press: 236-245.

[17] Simon R., et al. A Mobile Application Framework for the Geospatial Web. In Proceedings of WWW '07 (Banff Canada, May 2007). ACM Press: 381-390.

[18] Takeuchi, Y. et al. CityVoyager: An Outdoor Recommendation System Based on User Location History. In Proceedings of UIC'2006, (Berlin, 2006), Springer Press: 625-636.

[19] Zheng, Y, et al. Learning transportation modes from raw GPS data for geographic applications on the Web. In Proceedings of WWW 2008, (Beijing China, April 2008), ACM Press: 247-256.

[20] Zheng, Y., et al. GeoLife: Managing and understanding your past life over maps. In Proceedings of MDM'09, (Beijing China, April 2008), IEEE Press: 211-212

[21] Zheng, Y. et al. Understanding mobility based on GPS data. In Proc. Ubicomp'08, (Seoul Korea, Sept. 2008), ACM Press: 312-321