Exploring Application Usage Patterns of Smart Phones for Discovering Personal Semantic Regions

Student： Chun-Ta Lu Advisors：Dr. Wen-Chih Peng

Institute of Computer Science
National Chiao Tung University

ABSTRACT

Discovering personal semantic regions that referring user preferences on locations is a crucial prerequisite for personal recommendation, location-based service, or data cleaning. A considerable amount of research effort has been elaborated on mining semantic meanings from user trajectories. However, previous studies discover semantics of regions only from geographic information provided. Given a region, since each user may have different kinds of activities within the region, the personal semantic of this region should reflect individual user's intension. Due to rapidly increasing amount of mobile applications on smart phones, in which various applications could be executed assist user's activities in different situations, making it possible to automatically identify individual user's preferences from application usage patterns. Thus, in this paper, we propose a personal semantic region discovery framework based on users trajectories and application usage logs from smart phones. Explicitly, we explore density concepts in spatial and temporal domains to discover regions of interest from user trajectories. Then, the semantics of the discovered regions are identified by both geographic information and application usage patterns. The application usage patterns discovered from application usage logs summarize the representative applications, which reveal the user's behavior preferences on the regions. In this study, we conduct a comprehensive experimental study on real datasets collected in smart phones and trajectory community. The results demonstrate the effectiveness of our approach to discover personal semantic regions.

Contents

1 Introduction **1**

2 Related Work **7**

2.1 Mining Semantic Meaning of User Locations 7

2.2 Algorithms for Determining Regions of Interest 8

2.3 Mining User Behavior from Mobile Phone Usage 10

3 A Framework for Discovering Personal Semantic Regions **12**

4 Regions of Interest Extraction **15**

4.1 Spatial-Temporal Density Measurement 15

4.2 Trajectory Partition 17

4.3 Sequential Density Clustering Algorithm 18

4.4 Regions of Interest Discovery 20

5 Mining Personal Semantics of ROIs **22**

5.1 Geographic Semantic Enhancement 22

5.2 Behavior Semantic Extraction 23

5.2.1 Location Dependency of Applications 24

5.2.2 Mining Application Usage Pattern 25

6 Experiments **29**

6.1 Experimental Setup 29

6.2 Case Study: Personal Semantic Regions 33

6.3 Effectiveness of App Usage Pattern 34

6.4 Effectiveness of Extracting ROI 37
6.5 Effect of Region Discovery Parameter 38
6.6 Accuracy of ROI Discovery 41

7 Conclusion **43**

List of Figures

1.1 An example of personal semantic regions of different users 2
1.2 An example of app usage spatial distribution 4

2.1 An example of discovering regions from a trajectory 9
2.2 An example of discovering regions by the stay point approach 9

3.1 Overview of discovering personal semantic regions 13

4.1 Overlap partition 17
4.2 Sequential density clustering over trajectory density distribution 18

6.1 Usage behavior statistic of user user 30
6.2 Stay duration distribution 31
6.3 Relative App Entropy of Behaviors 31
6.4 Relative App Entropy of Different Categories 32
6.5 Case study of user U_3 and user U_7 33
6.6 Precision for determining representative apps. 36
6.7 Comparison of different approaches. 36
6.8 Region coverage 38
6.9 Comparison under different folds 39
6.10 Comparison under different thresholds 40

Chapter 1

Introduction

Different from mobile phones in past time, smart phones are not limited to their built-in functions such as making a phone call but allow the users run myriad mobile applications, abbreviated as Apps. With the rapid technologies adoption and enormous amount of Apps for smart phones, people can easily find appropriate Apps to fit their requirement in their life. In particular, smart phones can serve as ideal sensor devices to observe user behavior in different locations because people habitually carry their smart phones with them and use various Apps as a medium to assist their activities. As the availability of location and App usage records from smart phone prospers, knowledge discovery on human behavior becomes a promising research area. In the past decade, several works have been elaborated on trajectory data, which is a sequence of a user's daily locations as time changed, for capturing relationship between people movement behavior and geographic locations, such as modeling user movement[25, 5], learning human mobility patterns [11, 17, 13]. Furthermore, location based activity [19, 18, 26, 32] or semantic meaning [2, 4] can be annotated from geographic information on a discovered region where a user often stays. However, the same region always infers the same semantic meanings to different users if only considering the relation between location and geographic information. Fortunately, by analyzing an individual user's App usage logs and exploring App usage patterns on the region, the user's behavior can be inferred and the personal semantic meaning can be annotated accordingly. Thus, in this paper, we focus on discovering personal semantic regions by exploring movement behavior and App usage patterns from users' smart

phone.

More specially, our target is to discover regions that are semantically meaningful to each individual user and then annotate the personal semantics to each region for a user by learning user behavior preferences. Automatically learning user preferences is important and critical to understand user habits on different locations. For example, it can be applied to personalized recommendation services such as searching and advertisement. The personal adjusted searching results could be better fitting for user requirement according to the user's preferences. The advertisement could find potential customers and avoid distributing inappropriate information to the wrong persons.

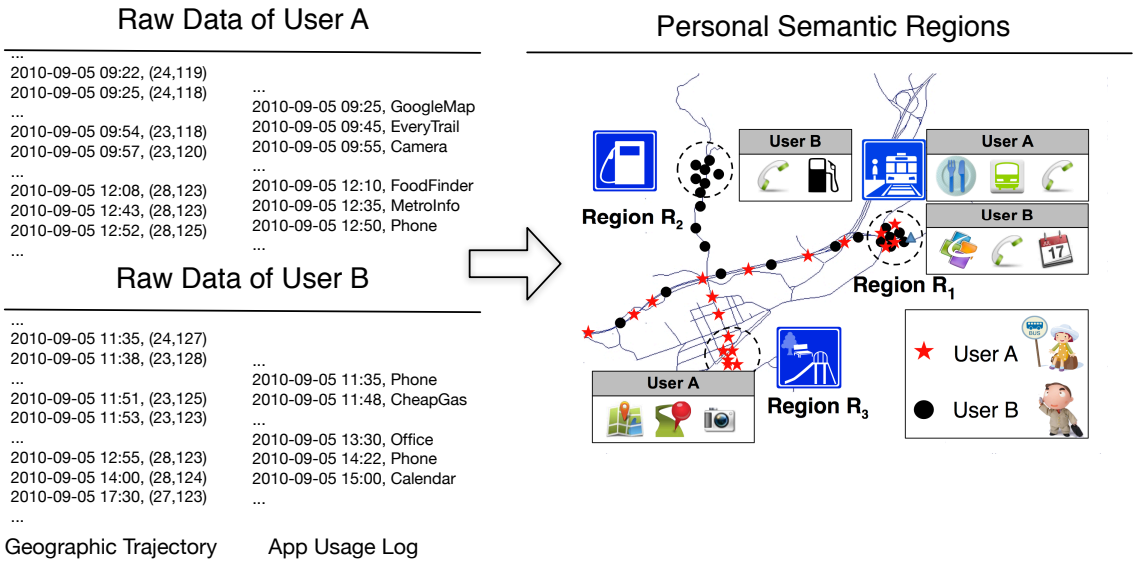


Figure 1.1: An example of personal semantic regions of different users

In general, from existing works [26, 32, 18, 2, 4], there are two tasks for semantic region discovery: identify regions of interest (abbreviated as ROIs) and annotate semantic meanings to the regions discovered. Such a region summarizes a set of location points from historical trajectory data that are close enough in the spatial space. The associated semantics of the discovered regions which are determined by manually labeling [18, 32] or by referring to points of interest, abbreviated as POIs (e.g., restaurants, shopping centers, landmarks or sightseeings), within these regions [2, 26, 32, 4]. In particular, the authors in [4] propose a framework for mining semantic locations from GPS data. Explicitly, by exploring the concept of stay points in [31], the authors first extract and cluster stay points as the regions of interest.

Then, they apply a reverse geocoder to obtain street addresses and then annotate the semantics of the ROIs by using a yellow pages directory. Note that POIs within ROIs are referred to the semantics of ROIs. Consider an example in Figure 1.1, where three regions (R_1,R_2, and R_3) can be extracted as ROIs. The semantic of each region are determined by the POIs (a train station, a gas station, and a national park, respectively) within the region.

However, it is not feasible to infer the personal semantic meaning of regions to a user only from his/her movement trajectory. Additional data of personal behavior observation will be needed to accomplish the personal semantics of regions. In previous works, the semantic of a region extracted from a user's manual comments or extracted from geographic information system such as POI categories or yellow pages directory is applied to all users, i.e., the semantic of a discovered region expresses the same meaning to different users. In fact, even the same region may bear different semantic meanings to different users, depending on what kind of behavior they did on the region. The semantics annotating problem could be better understood by Figure 1.1. In Figure 1.1, R_1 is a train station where both user A and B visited. User A traveled along region R_3 and R_1 and used some applications related to local tour service, while user B drove pass region R_2 to refuel and checked calendars or processed office for work on region R_1. For user A, R_1 is a tourist spot, whereas R_1 is one working place for user B. Therefore, we suppose that the same discovered region may have different semantic meanings to different users. If only the geographic information is taken into consideration, it will be the same semantic meaning to both users. Ask each user to manually annotate the semantics for each personal region can be helpful to solve this problem, but it is not practical because users may only offer a small amount of time and effort to label the regions or leave the comments denoting about the regions. Automatically recognizing the individual's behavior preference will be necessary for identifying personal semantic on a region.

As mentioned above, due to the increasing amount of Apps, nowadays, people download and execute Apps to meet their requirements and assist their activities. Thus, we claim that by exploring App usage patterns, the semantic meanings could be further annotated. For instance, user A in Figure 1.1 referred GoogleMap to find the direction to famous spots, recorded trips by EveryTrail, and took pictures on the spots. When she went to the train station, she used FoodFinder to search nearby restaurant for lunch and used MetroInfo to check the train schedule. The usage of these Apps is highly related to the user's behavior.

As a result, analyzing mobile application usage logs along users' trajectories can offer an opportunity for understanding personal behavior on different locations.

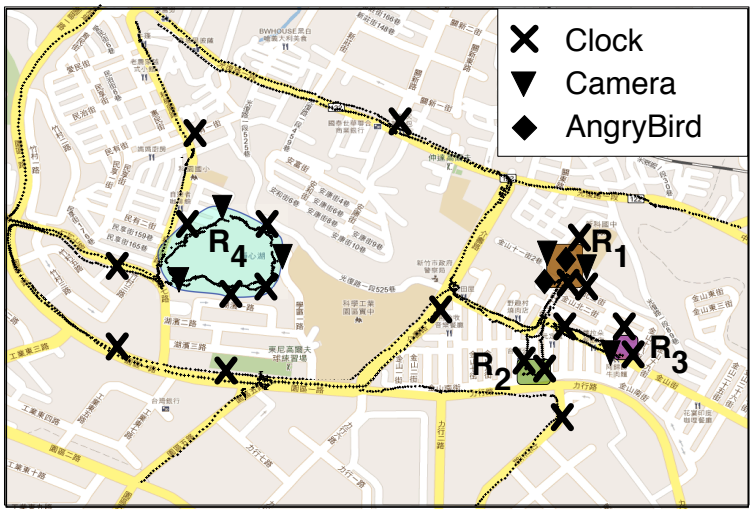


Figure 1.2: An example of app usage spatial distribution

To illustrate the research tasks and issues faced to achieve our goal, we consider Figure 1.2, where App usage records attached along an individual user's trajectories are given. For inferring user behavior preference on each region, one would like to identify the App usage pattern by mining the representative Apps of the region. Intuitively, the more frequently used Apps can be viewed as more representative to the user behavior on the region. Actually, some Apps are used anywhere but cannot represent for a specific location.

	R_1	R_2	R_3	R_4
Clock	3	2	2	4
Camera	2	0	1	3
Angry Bird	2	0	0	0

Table I: An example of application-region distribution

For instance, Table I summarizes the number of App usage on each region in Figure 1.2. In Table I, Clock was the most frequently used App on all regions, but Clock cannot show the user's behavior on any region. As a result, when measuring the representativeness of Apps to regions, not only the usage frequency but also the relation between Apps and regions should be considered. Expressly, when the frequency of an App to be executed on a region is much higher than that on other regions, the more possible the App could represent the user's behavior

preference of the region. Consider an example of the representativeness comparison on R_1, where Angry Bird would be more representative than Camera because Angry Bird was only executed on region R_1 while Camera was executed on three regions. From above observation, we found that representativeness of an application on a region is similar to the importance of a term in a document. The term importance increases proportionally to the number of the term appears in the document, but is offset by the frequency of the term in the corpus. Similarly, the representativeness of an App increases proportionally to the usage frequency in the region, but decreases when the number of locations in which the App to be recorded increases. As a result, to extract the behavior semantic, a representativeness measure could be designed based on an information retrieval measurement TF-IDF (term frequency-inverse document frequency), first introduced in [15], to evaluate the relevance between applications and regions.

In order to discover the individual semantics of ROIs for each user, we propose a framework to discover personal semantic regions. Our proposed personal semantic of a region contains both geographic semantic and behavior semantic, where a geographic semantic contains the ROI or POIs' information, and a behavior semantic consists of a set of Apps that could infer users' behaviors from the categories of Apps. The framework first discovers the regions of interest from trajectory data and then annotates the personal semantic meaning of each region discovered.

Different from previous works, the semantically meaningful regions we want to find are where a user may frequently stay for some behaviors not where a user frequently passes by. By matching the timestamps the Apps were recorded at to the stay durations a user stays in a region, we could decide whether the Apps was executed in the region and learn the user behaviors from the Apps. To discover regions attracting a user to stay, there is a need to sequentially follow the user's movement trajectory to observe whether the user stays in the regions or not. Therefore, we propose a sequential density clustering approach to extract regions from user trajectories, where the spatial-temporal density is designed to reflect the local configuration of trajectory points in the spatial-temporal data space. We argue that when a user stays or lingers in a region, the spatial-temporal density of the cluster of the sequential trajectory points is likely to be much higher than that when the user enters or leaves the region. Thus, our clustering method is proposed to extract the dense regions from

the cluster of sequential trajectory points. Then we further cluster the regions where the user frequently stays as ROIs from multiple trajectories.

To annotate the personal semantic meaning, in light of these discovered ROIs, we characterize not only the geographic semantics by nearby POIs but the behavior semantics by user's App usage patterns on the regions. To identify the App usage pattern, we propose a ranking model to rank the representative Apps. In this model, we first filter out non-location dependent Apps, which are executed not depending on any location and cannot represent for any region, by the location entropy of Apps. Then, several ranking functions based on TF-IDF are proposed to rank the representativeness of Apps to the region. The app usage pattern is represented by a ranking list of top-K representative Apps of the region and the behavior semantic is implied from the pattern. Finally, the personal semantics of a region discovered can be precisely represented by geographical and behavior semantics.

In this study, we prove our proposed framework by real datasets collected in smart phones and trajectory community. The experiments demonstrate the effectiveness of discovering personal semantic regions. The main contributions of this paper are summarized below.

- We propose a personal semantic region discovery framework based on users' trajectories and App usage logs from smart phones.
- We propose a sequential density clustering to discover ROIs by considering both spatial and temporal domain.
- We define a concept of annotating semantic meanings from both geographic information and App usage patterns.
- We present comprehensive experimental results over various real datasets. The results demonstrate that our techniques are more accurate than existing schemes.

The rest of papers is organized as follows: In Section 2, we reviews the related literatures. Section 3 gives an overview of the proposed framework. In Section 4, we briefly introduce how to derive regions of interest. The semantic extraction model is shown in Section 5. Section 6 discusses the experimental evaluations by real datasets and Section 7 concludes. For reference convenience, the symbols used in the paper are listed in Table I.

Chapter 2

Related Work

In this section, we first survey the research works for mining semantic meaning of user locations and then review some existing algorithms for determining regions of interests. Finally, we discuss related works on mining user's behavior from mobile phone usage.

2.1 Mining Semantic Meaning of User Locations

Since the locations attracting users to pass by or stay could carry particular semantic meanings, a variety of research works for mining semantic meaning of user location have been studied. Some research works address on the user locations shared in location-based social network services, e.g. Foursquare, Whrrl, Google Latitude. For example, Mao *et al.* in [27] developed a semantic annotation technique to automatically annotate all checked-in places with category tags. Based on the check-in behavior of users, the authors extracted features of places from explicit patterns of individual places and implicit relatedness among similar places. Upon these features, the semantic of each places could be classified by the probability of category tags to be labelled to the place. In [28], Ying *et al.* provided a friend recommendation service by user similarity on semantic trajectories, where a semantic trajectory consists of a sequence of locations with semantic tags to capture the landmarks passed by. The authors proposed a similarity measurement taking not only the geographical closeness but the semantics of trajectories into account.

Some researchers claim that the human activities are highly related to locations, such as eating in a restaurant or watching movies in a cinema. From user trajectories, Liao *et al.* in [19] utilized a conditional random field model to recognize user activities. Also, in [26], Xie *et al.* argued that activities can only happen when trajectory is geographically near for a proper period of time for these activities. Based on the idea, the authors proposed a spatio-temporal join approach that considers both geographical and temporal influence to identify user activities. Moreover, since user locations may not be repeated exactly in similar trajectories, a number of research works attempt to discover regions of interest (ROIs) from trajectories for replacing the exact locations where users often pass by or stay.

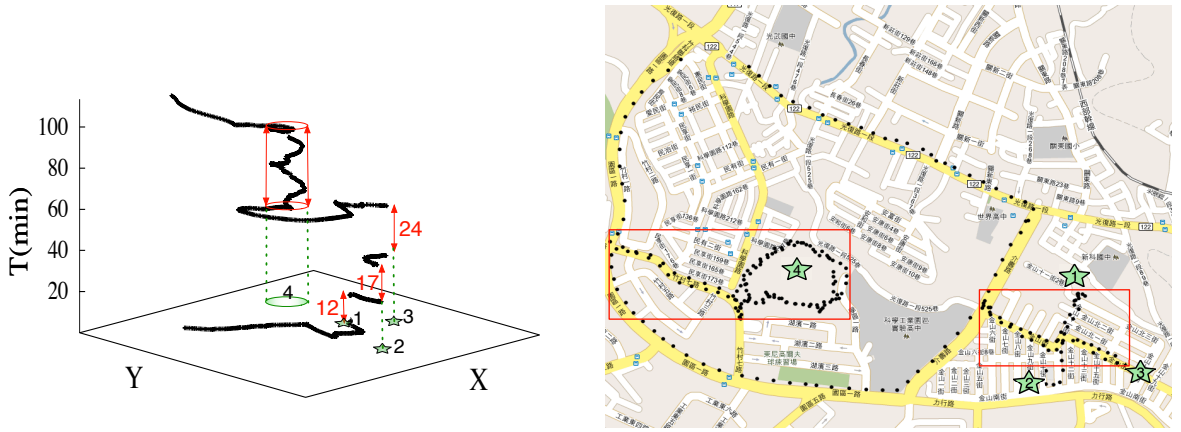
However, the semantics of regions are extracted only by mapping spatial locations on background geographic information in previous works. That is to say, each discovered region is assigned the same semantic meaning to different users. None of these works consider the difference of personality on the same location. In contrast, our approach identifies semantics not only by referring the geographic information but by mining user behavior from personal mobile application usage. Furthermore, to the best of our knowledge, this is the first work for mining user behavior from App usage logs on discovered unknown regions.

2.2 Algorithms for Determining Regions of Interest

Region discovery has been widely discussed in the field of trajectory data analysis [10, 14, 31, 30, 33, 2, 23, 22, 26]. Most of proposed methods employ density based clustering techniques to group a set of trajectory points into a cluster as a region, such as DBSCAN [9] and OPTICS [3]. Giannotti *et al.* [10] adopted grid density region to discover popular regions as ROIs where dense cells in space are detected and merged if they are neighbors. The authors in [14] extracted frequent regions by applying the clustering method DBSCAN. In spite of advantage of DBSCAN that clusters in arbitrary shape can be detected, they have to decompose a cluster when it is too large to describe correlations between frequent regions.

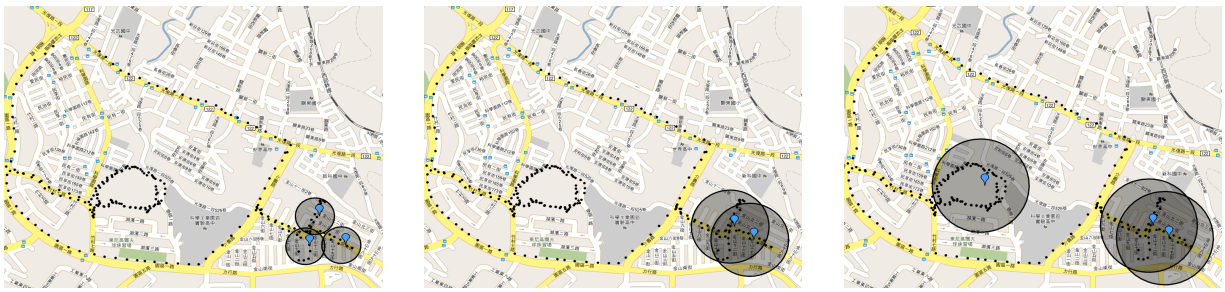
Recently, some works [16, 31, 32] aimed to applying sequential constraint to a single trajectory for extracting locations where a user may stay. The authors in [16] proposed an idea of stay point and claimed that can stand for a geographic region and carry a particular semantic

meaning. Unlike density based clustering, a stay point is detected when the consecutive points of a examined point do not exceed the predefined distance threshold during the specified period of time threshold. In [29], Zhheng *et al*. proposed a framework to discover interesting locations and possible activities that can be performed for tour recommendations by using the location data based on GPS and users' comments on various locations. After that, they parsed the POI category database and the users' comments from WWW, which in general are texts, to extract the possible activities in these detected regions. Cao *et al*. in [4] detected locations from the stay points by adopting the existing clustering algorithms OPTICS and K-means. Then, they revised the geocode to obtain street addresses from geographic information system, and obtained region semantics by using a yellow pages directory.



(a) The original trajectory (b) Regions discovered by density-based approach

Figure 2.1: An example of discovering regions from a trajectory



(a) 100 Meters (b) 200 Meters (c) 250 Meters

Figure 2.2: An example of discovering regions by the stay point approach

However, all of the above techniques have common deficiencies for semantic region discovery from trajectory data. By applying an universal spatial distance limit for cluster discovery may either miss regions with different density or merge non-related regions. Because a user

could perform different kinds of activities in a region and each activity could have different distance density and speed distribution, the size of a region varies with the user's activity in the region. Figure 2.1 and Figure 2.2 show an example to illustrate the deficiency of using universal distance limit to identify ROIs. Consider a trajectory from Figure 1.2, as shown in Figure 2.1(a), there are three indoor regions (region 1, 2 and 3) labeled with the temporal duration gaps as user stay durations on these regions, and one region (region 4) where the user was walking around. The temporal transition gap between sampled location points is generated due to the loss of satellite signal when GPS-embedded location recorder is inside a building. If clustering by a spatial density based algorithm DBSCAN (Minpts = 7, Eps = 50 meters), as shown in Figure 2.1(b), only two regions where represented by minimum bounding rectangles (MBRs) are discovered. While the user stays in the region 2 for 17 minutes, DBSCAN cannot discover region 2. This is because that region 2 does not have a sufficient amount of points to form a cluster. Besides, road-sections and intersections, where often passed but carries non-semantic meaning to the user, are included in both discovered regions.

Figure 2.2 shows regions discovered by the stay points approach, where the time threshold is fixed to 10 minutes and the stay point extraction results of three different distance thresholds are compared. When distance threshold is set to 100 meters, in Figure 2.2(a), there are three stay points mapping to three regions, but the region 4, with much larger area of activity, cannot be detected. The regions cannot be detected until the distance threshold is not less than 250 meters, but the other three regions have been mixed up and their coverage have been overlapped (shown in Figure 2.2(b) and Figure 2.2(c)) before then. Thus, to detect regions with a variety of activities, the methods that using universal distance limit cannot discover all kind of regions. An algorithm for extracting various regions without a distance threshold is needed.

2.3 Mining User Behavior from Mobile Phone Usage

Since mobile phones can serve as ideal devices to observe the users' behavior, a significant amount of research works has been conducted on the usage of mobile phones. Most of these works focus on mining user behavior from user's locations such as user movement modeling

[25, 5], user mobility patterns discovering [11, 17, 13], and so on. In [25], a large scale study attempted to explore the limits of predictability in human mobility. By measuring the entropy of each individual's trajectory, the authors found 93% potential predictability in user mobility from cell based location log data across 50,000 mobile phone users. Also, an information-theoretic approach was proposed in [5] to quantify the difficulty for predicting a mobile phone user's location when the user's location at selected time points are known. In [11], Gonzalez *et al*. studied the trajectory of 100,000 mobile phone users for a six-month period and found that human trajectories show a high degree of temporal and spatial regularity and a significant probability to return to a few highly frequented locations. After correcting for differences in travel distance and the inherent anisotropy of each trajectory, they also found that the individual travel patterns collapse into a single spatial probability distribution, indicating that, despite the diversity of their travel history, human follow simple reproducible patterns.

Moreover, various software usage behaviors from smart phones has been another avenue of research. From spatial-temporal data collected in Reality Mining project, Eagle *et al*. in [7] identified structure inherent in daily behavior by the principal components of the complete behavioral dataset, a set of characteristic vectors termed eigenbehaviors. In their model, an individuals behavior over a specific day can be approximated by a weighted sum of his or her primary eigenbehaviors. In [6], Eagle *et al*. analyzed the mobile application usage statistic on different locations. However, the number of studied applications is less than 10, major on built-in functions, and the locations involved are few and coarse (only home, working place.) Instead, in this paper, we analyze various Apps (over 500 Apps and 20 kinds of App categories) usage statistics on each discovered unknown regions. Similar to [11], few frequently stayed locations (ROIs) can be discovered from users' trajectories. In order to realize the semantic of ROIs, we further propose a ranking model for mining representative Apps of each discovered regions. Compared with existing literatures above, our approach is the first work to explore the semantic of ROIs from the combination of user locations and App usage logs.

Chapter 3

A Framework for Discovering Personal Semantic Regions

From a user's smart phone usage logs (user's geographic trajectories and App usage logs), we aim at discovering personal semantic regions by extracting regions of interest and annotating personal semantic of each region. A user's trajectory is represented as a sequence of points $Traj = \{p_1, p_2, ..., p_n\}$, and the set of user's trajectories is denoted as S_{Traj}. Each point $p_i (1 \leq i \leq n)$ contains user's position (x_i, y_i) and timestamp (t_i). A user's App usage logs is represented as a set of usage logs for each application $U = \langle U_{App_j} | \forall App_j \in App \rangle$, where U_{App_j} is recorded as a sequence of timestamps of the application App_j to be executed at $U_{App_j} = \{t_{j1}, t_{j2}, ..., t_{jm}\}$.

As shown in Figure 3.1, our framework consists of three parts, including regions of interest extraction, geographic semantic enhancement, and behavior semantic extraction. The first part is designed to extract ROIs, and the second and third parts are designed to annotate the geographic semantic and behavior semantic, respectively.

Prior to annotating the personal semantics, the first step is to discover the regions of interest (ROIs) where a user may frequently stay for some behaviors. As discussed in Section 2.2, a region extraction algorithm using an universal distance limit cannot extract various regions. Thus, without setting any distance threshold, we propose a spatial-temporal sequential density clustering to extract the dense region as ROI candidates and then cluster similar

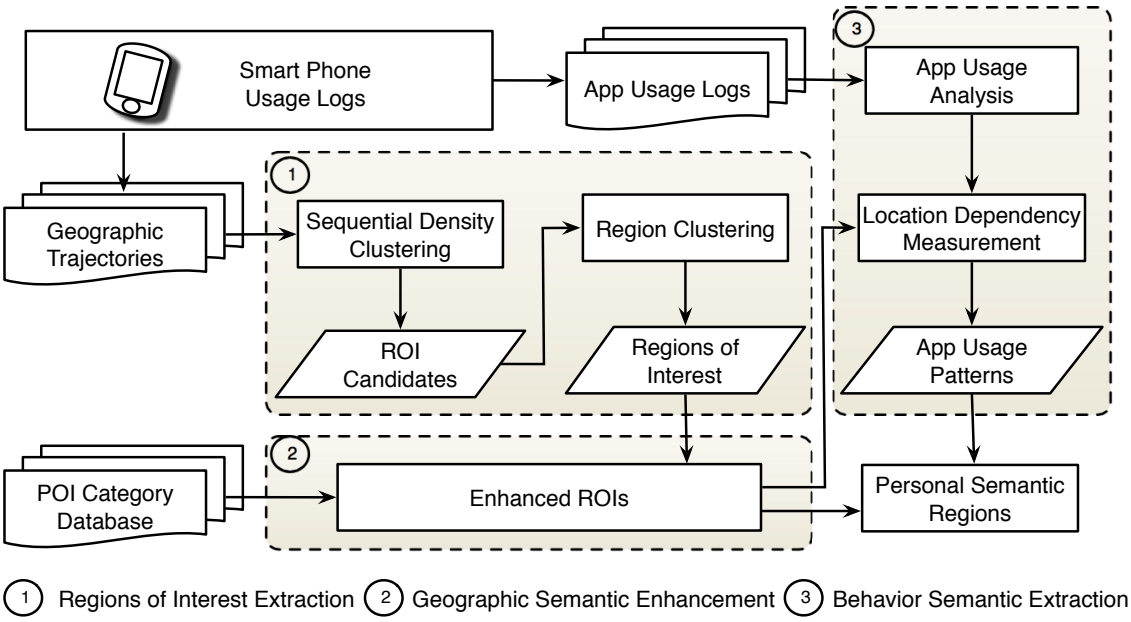


Figure 3.1: Overview of discovering personal semantic regions

candidates as a ROI R and denote the set of ROIs as S_R. Based on the discovered regions S_R, we further annotate the personal semantics of ROIs, including geographic semantics and behavior semantics.

Second, geographic semantic enhancement is the process to annotate the geographic semantics of each discovered region by extracting the semantics from points of interesting (POIs) in geographic information database. Ideally, each region R in S_R matches to a unique geographic semantic Sem_l (decided by nearby POIs in POI category database) and corresponds to a real world location, such as a train transition or a park. By adopting the SEM-CLS method, proposed in [4], we enhance each region R in S_R with geographic semantic as a enhanced ROI ℓ and denote the set of enhanced ROIs as S_ℓ.

Third, entering the behavior semantics extraction, we investigate the App usage patterns on ROIs for extracting behavior semantics of the regions. The usage of Apps (such as Local Directory, EveryTrail, or Camera) can imply the user's behavior ("TRAVEL") in the region. Thus, we propose a ranking model to identify the App usage patterns, which is a ranking list of representative Apps of the regions. The representative App is an App executed much frequently on the location than other locations. It means that the App is location dependent and can represent the user's App usage preference on the location. As a result, in our proposed

model, we first measure the location dependency of applications and filter out non-location dependent applications. Next, we propose several measures to evaluate and rank the representativeness of Apps for mining the App usage pattern of each enhanced ROI. From a App usage pattern, the user's behavior semantic Sem_b can be driven from the Apps in the pattern. Consider user A in R_1, the Sem_b of the App usage pattern (Local Directory, EveryTrail, Camera) is "TRAVEL".

Finally, from the discovered ROIs and the extracted semantics, the personal semantic regions is defined as follows:

Definition 1 ***Personal Semantic Region:*** *A personal semantic region is represented by a three-tuple* $\langle MBR, Sem_l, Sem_b \rangle$*, where* MBR *is the minimum bounding rectangle boundary of the region,* Sem_l *is the geographic semantic and* Sem_b *is the behavior semantic.*

For example, a personal semantic region ⟨ {(22,186),(24,120)},"Taiwan National Park", "TRAVEL" ⟩ has boundary {(22,186),(24,120)}, "Taiwan National Park" as the geographic semantic, and "TRAVEL" as the behavior semantic.

Symbols	Descriptions
R	region of interest (ROI)
ℓ	geographic semantic enhanced ROI
Sem_l	geographic semantic
Sem_b	behavior semantic
T	stay duration threshold
$fold$	trajectory partition smoothing parameter
ξ	spatial-temporal density difference threshold
$MinSR$	minimum support for ROI
d_h	maximum search range for ROI
Γ_ℓ	App usage pattern of ℓ
K	maximum size of Γ

Table I: Table of Notations

Chapter 4

Regions of Interest Extraction

In this section, we propose a sequential density clustering to discover regions of interest from users' trajectories. We first introduce the spatial-temporal density measurement to evaluate the closeness of trajectory data. For each trajectory, we will partition the trajectory into a set of trajectory segments and sequentially measure the density of each segment to observe its density distribution, where the distribution is a sequence of segment densities in order. Then, the ROI candidates can be extracted from the clusters of sequential dense segments. Finally, we will discover ROIs from the user's trajectories by clustering ROI candidates.

4.1 Spatial-Temporal Density Measurement

The spatial-temporal density is defined to reflect the local configuration of the points in the spatial-temporal data space and a cost function is used as a density measurement. Generally, the cost function is designed to represent the penalty of dissimilarity of the points within a segment. Prior work [20] defined the cost of a segment as the sum of squared Euclidean distance between points and its spatial centroid, where the cost is also called the variance of the segment. Without loss of generality, the squared Euclidean distance function is adopted as given below to measure the dissimilarity between two points.

$$D_{E^2}(p_i, p_j) = (x_i - x_j)^2 + (y_i - y_j)^2. \tag{1}$$

However, it only counts the spatial dissimilarity without considering the temporal feature such as the duration of user staying in a location. When devising a cost function, we should take both spatial and temporal features together. Because a trajectory does not involve only one activity in real world, the distance between points could vary with different activities in spatial domain and the temporal interval from a point p_i to its succeeding point p_{i+1} could vary from seconds to hours. If we directly measure the spatial density of the segment around this area, we cannot detect its significance. Thus, we take temporal feature as a weight compounded with spatial relation to measure the dissimilar cost of a segment, i.e, the spatial-temporal density of a segment.

Given a segment $s_l = p_i, p_{i+1}, ...p_j$, the definition of weighted cost function is stated as follows.

$$Cost(s_l) = \frac{\sum_{k=i}^{j} w_k * D_{E^2}(p_k, c)}{\sum_{k=i}^{j} w_k}, \tag{2}$$

$$c = (\frac{\sum_{k=i}^{j} w_k * x_k}{\sum_{k=i}^{j} w_k}, \frac{\sum_{k=i}^{j} w_k * y_k}{\sum_{k=i}^{j} w_k}), \tag{3}$$

$$w_k = \frac{(t_k - t_{k-1}) + (t_{k+1} - t_k)}{2} \tag{4}$$

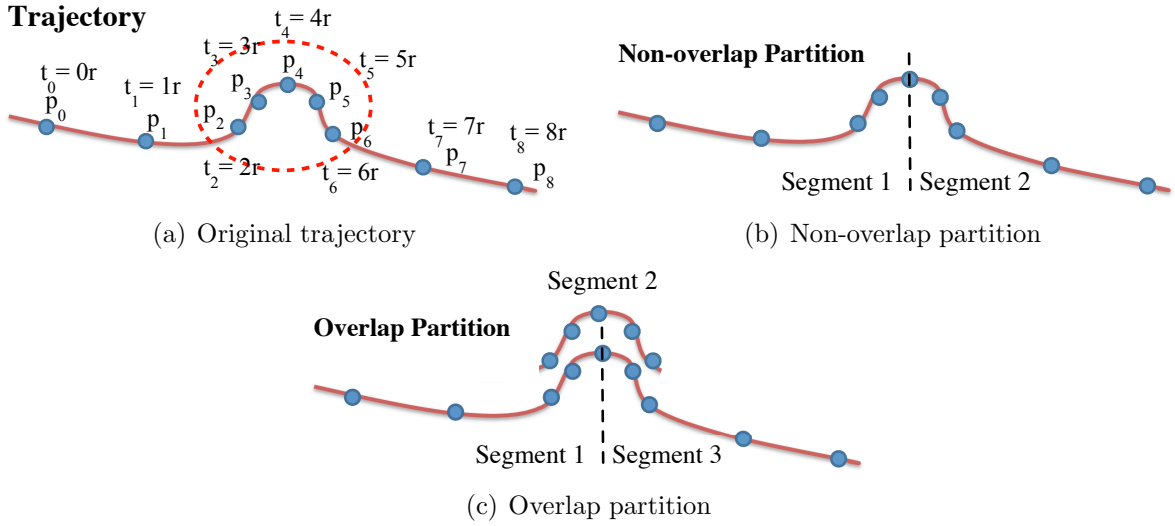
where w_k is the weight of point p_k, c is the weighted centroid of segment s_l, respectively. Furthermore, because there are different activities processing in a trajectory, the $Cost(s_l)$ can vary in a wide range. To normalize the density of clusters with different activities, the density function is measured as the logarithm of one over the cost. The definition of spatial-temporal density function is stated as follows.

$$Density(s_l) = \log_e(1 + \frac{1}{Cost(s_l) + \gamma}), \tag{5}$$

where the $Density(s_l)$ is in the boundary of $[0, log_e(1 + \frac{1}{\gamma})]$ and γ is a constant (given as $\frac{1}{e-1}$ in this paper to keep the maximum density equals to 1)

4.2 Trajectory Partition

Given an user's trajectory $Traj$, we aim to analyze its spatial-temporal density distribution to extract the region where the trajectory movement is more dense than the neighboring regions, i.e. the density in this region is a local maximum in the trajectory density distribution. Unlike the problem in [20], we are not pursuing to partition a trajectory such that the total cost of partitioned segments is minimized. Instead, we partition the trajectory in order to compare the density variance between sequent segments in an efficient way. To simplify the description of the spatial-temporal density distribution of a trajectory, each trajectory is periodically partitioned into $\lfloor \frac{p_n.t - p_1.t}{T} \rfloor$ trajectory segments, and the density of each sequent segment is computed. We assume such a sequential density set can be used to describe the density distribution of the trajectory.



(a) Original trajectory

(b) Non-overlap partition

(c) Overlap partition

Figure 4.1: Overlap partition

However, while the density distribution of a trajectory can be represented by the density distribution of a sequence of segments, it may occurs a partition loss that a dense region of a trajectory is lost because of partition. The reason is a dense region may be separated into several segments by partitioning. Under this condition, the density of each segment is smaller than the density of the dense region. As shown in Figure 4.1(a), the dense region of a trajectory is in the center, marked within a circle. Given the time interval of each point to its neighbor point is r and the period T is set as $4r$, the partitioned segments are S_1 and S_2 shown in Figure 4.1(b). As a result, the dense region in the center of this trajectory is split

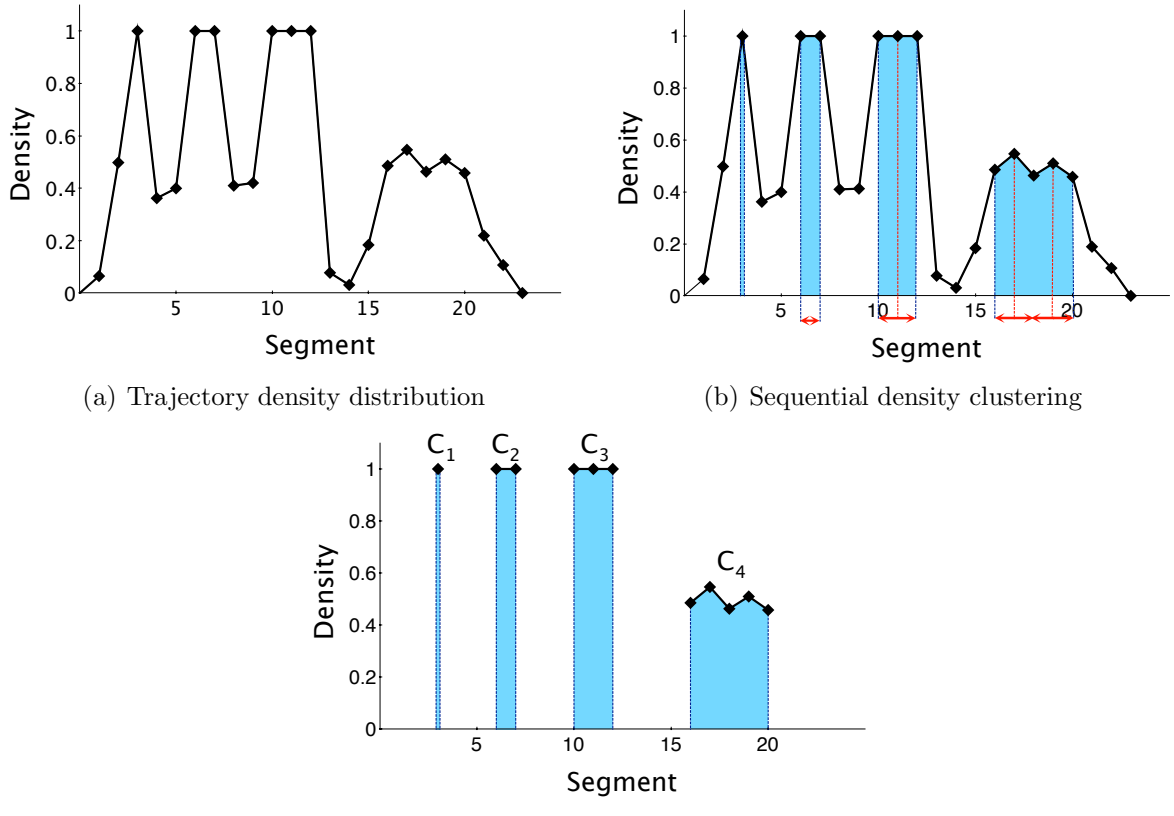
into two segments and the dense region cannot be detected.

To solve this problem, overlapping partition is implemented to smooth the region-split property when partitioning the trajectory. The time interval of the trajectory in Figure 4.1(c) is set as $[0, 4r), [2r, 6r), [4r, 8r)$ corresponding to segment S_1, S_2, S_3, respectively. The time interval function of overlapping partition is given as follows.

$$[t_{start_k}, t_{end_k}) = [\frac{(k-1) * T}{fold}, \frac{(k-1) * T}{fold} + T),$$

where $fold$ is a parameter to smooth the partition. In this paper, $fold$ can be set as a fixed integer and our experiment shows the result change slightly when $fold \geqslant 3$.

4.3 Sequential Density Clustering Algorithm



(a) Trajectory density distribution

(b) Sequential density clustering

(c) Clusters

Figure 4.2: Sequential density clustering over trajectory density distribution

We now present our sequential density clustering algorithm for ROI candidates extraction. Given a set of sequential trajectory segments S, our algorithm generates a set of sequent dense segments clusters as ROI candidates. It requires a parameter ξ, density threshold for similarity measurement. Before clustering, each density D_k of partitioned segment S_k is calculated by spatial-temporal cost. In a trajectory density distribution, a segment with a local maximum can correspond to a dense region of a trajectory. The segments with similar density are grouped into a cluster if they are adjacent to each other. Finally, the boundary of a region, i.e. a cluster, is extracted at where the density dramatically change. Thus, the ROI candidates extraction involves grouping the segments (if they belong to the same dense region) and setting boundary of dense region.

Algorithm 1: Sequential Density Clustering

Input: A set of trajectory segments S, a density threshold ξ
Output: A set of sequent dense segments clusters SDC

```
1  Compute D = Density(S) for each segment in S
2  Sequentially record the local max and local min from D to an array E
3  foreach local max E_i in E do
4      Take nearby two local mins and E_i as a group G
5      and take the last group in GroupSet as G'
6      if the density difference between G and G' is smaller ξ then
7          Merge G with G' and put into GroupSet
8      else
9          Add G into GroupSet
10 foreach group G in GroupSet do
11     foreach D_j from local max to local mins in G do
12         if |D_j − D_{j−1}| ⩽ ξ and (|D_j − D_{j+1}| ⩽ ξ) then
13             Add S_j into of a sequent dense segment cluster C
14     if boundary(C) ≠ boundary(G) then
15         Add cluster C into SDC
16 return SDC
```

For instance, we let $T = 10$, $fold = 2$ to partition the trajectory in Figure 2.1(a) into 23 segments and compute the density for each segment. The density distribution D of the periodically two-fold-partitioned trajectory is shown in Figure 4.2(a). Such a sequence of trajectory segments is the input of the algorithm. Algorithm 1 shows the sequential density clustering to detect ROI candidates from the density distribution. Initially, the local extremes

(maxima and minima) are identified and recorded as a set of group G. Each G is a group of local maxima E_i and its nearby local minima. The algorithm consists two steps. In the first step (Line 3-9), the algorithm computes the density similarity between two adjacent groups. If density difference between two adjacent groups is equal or smaller than density threshold ξ, these groups are sequentially similar. The algorithm performs the clustering to merge them into a new group. For example, there are two connected groups $G_1 : \{D_{14}, D_{17}, D_{18}\}$ and $G_2 : \{D_{18}, D_{19}, D_{23}\}$ in Figure 4.2(b). Given $\xi = 0.1$, G_1 and G_2 are similar ($|D_{17} - D_{18}| \leqslant \xi$ and $|D_{19} - D_{18}| \leqslant \xi$) and can be merged into a new group $G' : \{D_{14}, D_{17}, D_{23}\}$. The clustering results are added to *GroupSet* as a sequence of groups. In the second step (Line 10-15), the boundary of a cluster is extracted from each group G. The precise boundary of a cluster C is extended from the local maximum in G to its nearby local minima until the density difference between two continuous segment is more than ξ. The cluster $C_4 : \{S_{16}, S_{17}, S_{18}, S_{19}, S_{20}\}$ is extracted from group $\{D_{14}, D_{17}, D_{23}\}$ as shown in Figure 4.2(c). Only clusters with significant activity change are taken as ROI candidates. If there are no continuous density changes more than ξ inside a group, this implies the region enclosed in the group can be viewed as a non-semantic area.

4.4 Regions of Interest Discovery

While ROI candidates represent the location where an individual proceeds with some kind of high dense activities in a trajectory, it does not imply that those locations "frequently" appeared. Thus, given a set of trajectory data, we want to find out the region where an object frequently stays or lingers around for a certain activity, i.e, a region of interest. A region of interest is a summary of a set of similar ROI candidates from different trajectories. To define the similarity between candidates and to discover ROI, we adopt the definition of shared nearest neighbor (SNN) [12] and SNN density-based clustering [8]. That is, the similarity between a pair of points is measured by the number of their shared nearest neighbors. In graph terms, a link is created between a pair of nodes if both have each other in their K nearest neighbor (KNN) lists and an SNN similarity graph is created. Clusters are simply the connected components of the SNN graph. The discovery of ROI is similar to find clusters.

For each ROI candidate, it can be viewed as a node in SNN graph. However, if nodes are not close enough, they do not stay in the same region apparently.

Algorithm 2: Regions of Interest Discovery Algorithm

Input: A set of nodes, distance threshold d_h, minimum support $MinSR$
Output: a set of clusters

1 Find the $MinSR$-nearest neighbors in d_h of all nodes.
2 Construct the shared nearest neighbor similarity graph.
3 For every node in the graph, calculate the number of links.
4 Identify core nodes which has more or equal to $MinSR$ links.
5 Identify noise nodes which is neither a core node nor linked to a core node and remove them.
6 Take connected components of nodes to form clusters.
7 **return** the union of all clusters

When applying SNN density based clustering to discover regions of interest, we constrain the searching range of nearest neighbors is a radius d_h around the examined node. The setting of d_h is determined by demeaned range size of regions. We define a region is a ROI if each ROI candidates inside which contains at least $MinSR$ number of neighbors in the distance radius d_h. The nodes without $MinSR$ nearest neighbors are viewed as non-frequent regions and discarded. All the connected components in the resulting graph are clusters finally. These clusters can be considered as regions of interest where an object often visits for some behaviors.

We develop a regions of interest discovery algorithm (Algorithm 2) based on the property described in new SNN clustering algorithm [8]. The nodes that have at least $MinSR$ connectivity in the SNN graph are candidates for core nodes since they tend to be located well inside the natural cluster, and the nodes with connectivity lower than $MinSR$ and not connected to any core node are identified as noise nodes. As a result, a cluster is detected if there exists a connected component in SNN graph. The cluster which has at least $MinSR$ similar ROI candidates is regarded as a region of interest. Notice that the number of clusters is not considered as a parameter. Depending on the nature of the data, the algorithm finds the nature number of clusters for given set of parameters, $MinSR$ and d_h.

Chapter 5

Mining Personal Semantics of ROIs

In this section, we extract the personal semantics of ROIs in the combination of geographic semantic and behavior semantic. We first enhance ROIs with geographic semantics fetched from POIs close to the ROIs. Then, for extracting the behavior semantics of enhanced ROIs, we introduce the concept of location dependency of applications and mine the App usage pattern for each region.

5.1 Geographic Semantic Enhancement

Given regions of interest discovered from the user's trajectories, we aim at extracting the geographic semantics in the regions from their spatial relation with points of interest of geographical information database. One idea is to apply a reverse geocoder and fetches the POIs for each ROI candidates by external APIs (Google Place). The result returns for a candidate includes the street address and the corresponding POIs of the address. When the ROI candidates in a region of interest are mapped to the same address, the geographic semantic of the region can be extract from the POI.

However, a region may contain different semantic locations since the ROI candidates in the region cloud have different semantics. To enhance the semantic purity of ROI, we adopt the SEM-CLS (semantics-enhanced clustering) method in [4]. The method involves two steps: split and merge. In the split step, we call external APIs to obtain the street address and the

semantic of ROI candidates. If the ROI candidates in a ROI have different semantics, we split the region since it may contain multiple semantic locations. In the merge step, we merge the ROIs that may represent the same semantic locations. In order to determine when to merge two ROIs,each ROI is represented by a three-tuple $(t_{ae}, t_{as}, \overrightarrow{Sem_l})$ of features: the average stay entry time t_{ae}, which is the average starting time of all the ROI candidates; the average stay duration t_{as}, which is the average duration of all the ROI candidates; and the semantic list vector $\overrightarrow{Sem_l}$, which contains all the semantics that may apply to the ROI candidates. If two close ROIs exhibit the similar patterns, the two ROIs are possibly the same. We compute the similarity of two ROIs based on the features summary of the ROIs. The vector space model is used to compute the similarity between two ROIs R_1 and R_2 as follows:

$$Sim(R_1, R_2) = \frac{\overrightarrow{Sem_{l1}} \cdot \overrightarrow{Sem_{l2}}}{\| \overrightarrow{Sem_{l1}} \| \| \overrightarrow{Sem_{l2}} \|} + \frac{min(t_{ae1}, t_{ae2})}{max(t_{ae1}, t_{ae2})} + \frac{min(t_{as1}, t_{as2})}{max(t_{as1}, t_{as2})} \quad (1)$$

The similarity of the average stay duration is computed as the smaller one divided by the larger one, and the same is done for the average entry time.

5.2 Behavior Semantic Extraction

Given App usage logs U and enhanced ROIs S_ℓ, we propose a ranking model to find the App usage pattern for extracting behavior semantics of the regions. For each enhanced ROI ℓ, the App usage logs U are matched to the stay duration of the ROI candidates associated to ℓ. When an application record (App_i, t_j) is logged during one of the ROI candidates, App_i is executed inside the ROI. When the usage frequency of an App on ℓ is much higher than that on other regions, the more possible the App could represent the user's behavior preference of ℓ. However, if the Apps are executed independent to location, it cannot represent for any specific region. Thus, in this model, we measure the location dependency of applications by the location entropy of application usage and filter out the non-location dependent applications. After that, several measures are proposed based on TF-IDF to evaluate the representativeness of Apps to the region ℓ, and the App usage pattern is discovered as the ranking list of top-K representative Apps. Finally, the behavior semantic can be extracted from the App usage

pattern.

5.2.1 Location Dependency of Applications

The location dependency of an application is measured by the Shannon's conditional entropy [24].

$$H(\ell|App_i) = -\sum_{\ell_j \in \ell} P_{\ell|App_i}(\ell_j|App_i) log_2 P_{\ell|App_i}(\ell_j|App_i), \quad (2)$$

where $P_{\ell|App_i}(\ell_j|App_i)$ is the probability of App_i to be executed on ℓ_j. Intuitively, $H(\ell|App_i)$ measures uncertainty expressed as the average number of bits that are needed to determine the location for a given application App_i. Since we want to find the applications which are highly related to the location, $H(\ell|App_i)$ should be lower than $H(\ell)$, where $H(\ell)$ is the location entropy of the user to execute any application. The value of $H(\ell)$ for a user is determined using the following method. Let U' be a randomly selected sequence from application usage logs U. Since U' should not exhibit relation to any application, even the minimum entropy of $H(\ell|U')$ should be higher than $H(\ell)$. Therefore, we record its minimum entropy as $\widehat{H}(\ell)$, and only the application with entropy smaller than $\widehat{H}(\ell)$ may carry the behavior semantic of locations. To provide a 99% confidence level on which application is location dependent, we repeat the above random selection experiment 100 times and record the minimum entropy. The 99-th smallest value of these 100 experiments will serve as a good estimator for location entropy.

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
App_1	10	10	10	10	10
App_2	1	5	1	0	0
App_3	7	7	7	0	0

Table I: An example of application-location distribution

Example 1 *(Location dependency measurement) Table I shows an example of application-location distribution, where each entry denotes the usage frequency of an application on a given location. The location dependency of each Apps are calculated as below:*

$$H(\ell|App_1) = -5 * \frac{10}{50} * log_2(\frac{10}{50}) = 2.32$$
$$H(\ell|App_2) = -\frac{5}{7} * log_2(\frac{5}{7}) - 2 * \frac{1}{7} * log_2(\frac{1}{7}) = 1.15$$
$$H(\ell|App_3) = -3 * \frac{7}{21} * log_2(\frac{7}{21}) = 1.58$$

Assume the location entropy $H(\ell)$ of the user is 2.0, the App_1 will filtered out as a non-location dependent App.

5.2.2 Mining Application Usage Pattern

For fully extracting the behavior semantic of a region, the App usage pattern is mined as the top-K ranked representative Apps. By adjusting the value of K, the App usage pattern can be described more precisely to fit the user behavior. Basically, when ranking representative Apps, we can simply measure the representativeness of applications by the frequency of occurrences on a enhanced ROI. However, such a solution could have bias in evaluating the relation between applications and locations. The more frequently used Apps on a location will be viewed as more important to the user at the location. In the real world, some Apps are only used at some specific locations but, conversely, some Apps can be used anywhere. For instance, in Figure 1.1, user B used a communication App to talk with friends on every locations. By contrast, he utilized the office App only on the train station (region R_1). Even though the usage frequency of office App is lower than that of communication App in R_1, the office App can reflect the activity of user B in R_1 and has stronger location dependence than communication App has. To user B, the semantic meaning of R_1 is a working place which could be inferred from his log on R_1. Thus, we should consider not only the usage frequency in a specific location but also the relation between applications and locations.

From above observation, we found that representativeness of an application on a location is similar to the importance of a term in a document. Specifically, each location with a set of applications executed on it can be treated as a document with a set of terms in the document.

The representativeness of an App increases proportionally to the usage frequency in the region, but decreases when the number of locations in which the App to be recorded increases. As a result, we proposed AF-ILF (application frequency-inverse location frequency) based on TF-IDF to evaluate the relevance between applications and locations. The AF-ILF function can show the relevance between application App_i and location ℓ_j:

$$AF\text{-}ILF_{i,j} = AF_{i,j} * ILF_{i,j} \tag{3}$$

The application frequency (AF) is a normalized count of an application App_i on the given location ℓ_j. We define AF as follows:

$$AF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

where $n_{i,j}$ is the number of occurrences of the considered application App_i in location ℓ_j, and the denominator is the sum of number of occurrences of all applications in location ℓ_j. If the user executed App_i more frequently on location ℓ_j, the importance of App_i to location ℓ_j is higher.

The inverse location frequency (ILF) is a measure of the relation of App_i to ℓ_j. It is inversely proportional to the number of locations, obtained by dividing the total number of locations ($|\ell|$) where the user had ever executed any application. The formal definition of ILF is given as:

$$ILF_{i,j} = \log_2(\frac{|\ell|}{|\{l : App_i \in \ell\}|}) \tag{5}$$

where $|\{l : App_i \in \ell\}|$ is the number of locations where the application App_i appears (i.e. $n_{i,j} \neq 0$). If App_i is executed on many locations, the relation of App_i to location ℓ_j is smaller.

However, AF-ILF will underestimate some Apps when the Apps are occasionally executed once on a location. Due to the small number of personal locations ($|\ell|$), the AF-ILF value of App_i will decreases quickly along to the increase of the number of locations App_i appears ($|\{l : App_i \in \ell\}|$). To solve it, we modify the ILF to the relative location frequency (RLF),

which is the overall percentage of the App_i recorded in ℓ_j.

$$RLF_{i,j} = \frac{n_{i,j}}{\sum_q n_{i,q}} \tag{6}$$

Because $RLF_{i,j}$ can measure the dependency of the App_i to the location ℓ_j, the value will decrease when the recorded locations of App_i are sparse. Similar to AF-ILF, we propose another measure AF-RLF as the product of AF and RLF.

$$AF\text{-}RLF_{i,j} = AF_{i,j} * RLF_{i,j} \tag{7}$$

The idea is that a location ℓ_j has high representative score for the App_i executed more often by the user (high $AF_{i,j}$), discounting for those Apps that are executed not depended on the location (low $RLF_{i,j}$).

Definition 2 ***App Usage Pattern:*** *An App usage pattern Γ_ℓ is a ranking list of top K representative Apps on enhanced ROI ℓ, where K is the maximum size of list Γ_ℓ.*

Example 2 *(App usage pattern) Assume that an App usage pattern is the top-1 representative App, we take the pattern of ℓ_2 in Table I to illustrate the result of each measures. If ranking by the usage frequency, App_1 will represent for ℓ_2. If ranking by AF-ILF, the score of App_1, App_2 and App_3 are $\frac{10}{10+5+7} * log_2(\frac{5}{5})$, $\frac{5}{10+5+7} * log_2(\frac{5}{3})$, $\frac{7}{10+5+7} * log_2(\frac{5}{3})$, respectively. The final scores of App_1 to App_3 are 0, 0.167 and 0.234, and the App usage pattern ranking by AF-ILF is App_3. If ranking by AF-RLF, the scores of App_1 to App_3 are $\frac{10}{22} * \frac{10}{50}$, $\frac{5}{22} * \frac{5}{7}$ and $\frac{7}{22} * \frac{7}{21}$, respectively. The final scores of App_1 to App_3 are 0.09, 0.162 and 0.106, and the App usage pattern ranking by AF-RLF is App_2.*

For each discovered location, we can measure the AF-ILF$_{i,j}$ and AF-RLF$_{i,j}$ score for each application App_i recorded on the location ℓ_j and form the App usage pattern by a ranking

list ranked by scores. The behavior semantic Sem_b is represented by the categories of the top ranked applications.

Chapter 6

Experiments

The experiments in this study are designed for two objectives. First, we evaluate the effectiveness of discovering personal semantic regions. We first introduce the dataset collected from real traces of smart phone users in Section 6.1. In Section 6.2, we perform a case study to illustrate the personal semantic regions. Section 6.3 evaluates the effectiveness of mining behavior semantics from App usage logs. Second, we verify the coverage ratio of ROI candidates to the stay locations where some activity happened. Section 6.4 describes the precision comparison between Sequential Density Clustering (SDC) and Stay Point (SP). Section 6.5 investigates the effect of region discovery parameters. We conduct experiments on our prototype which was implemented in the python and PHP on CarWeb [21], a traffic data collection platform on Ubuntu 9.10 operating system.

6.1 Experimental Setup

Behavior Semantic	App Category
COMMUNICATION	Communication
OFFICE	Finance, Business, Productivity, Tools, Personalization
LEISURE	Entertainment, Games, Photograph, Media & Video Music & Audio, Books & Reference, News & Magazine
OUTDOOR	Social, Transportation, Travel & Local, Weather, Health & Fitness, Shopping, Sports

Table I: List of behavior semantic

We track 14 users' application usage logs over 10 months on smart phones running android 2.2 platform. To collect their usage logs, a background monitor records for each application when it starts to be executed and shown on screen, and the monitor also records the user's locations via GPS, Wi-Fi and GSM. We use the subset of 10 users with sufficient log data during 2010-07 to 2010-12 for evaluating the proposed methods. To extract the behavior semantics from Apps, we summarize the category of Apps in Table I. From the frequency of each App to be recorded, we analyze the general smart phone usage behavior statistic of each user, in Figure 6.1. Not surprising, COMMUNICATION comprises the most percentage of applications usage for almost all the users in this study, since it serves as the most basic and still the most important function of the smart phones. The others usage behaviors hold different importance depended on individuals.

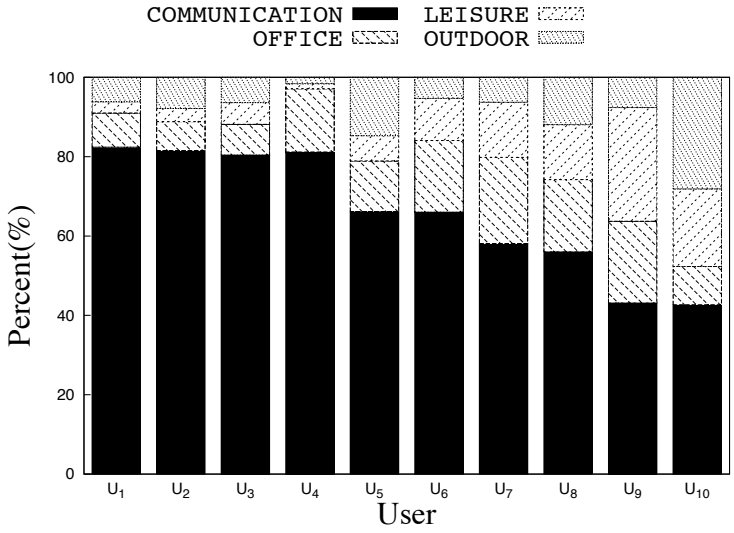


Figure 6.1: Usage behavior statistic of user user

Next, we apply the ROI discover methods to find personal ROIs for each user. Following the approach of ROI candidate discovery with stay period set as 15 minutes, we obtain 3,446 candidates from 0.3 million location points. From ROI candidates, we discover total 92 enhanced ROIs , where a user stays at least three times, in the dataset. The stay duration distribution is shown in Figure 6.2. Obviously, the most of stay occur between half hour to one hour and the number of stay decreases when the stay duration is longer, because users usually not stop in a place for a long time. It also imply that approximately 80% of the period for a user to stay in a location does not exceed over 6 hours.

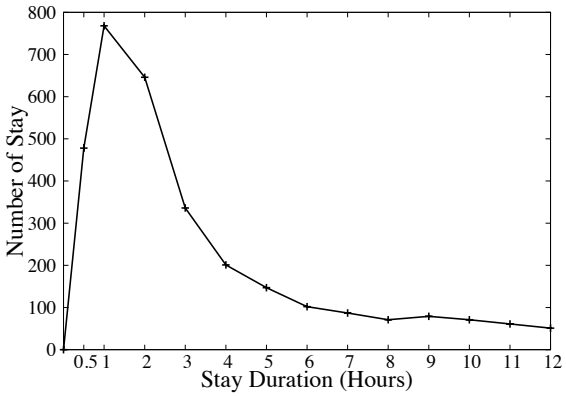


Figure 6.2: Stay duration distribution

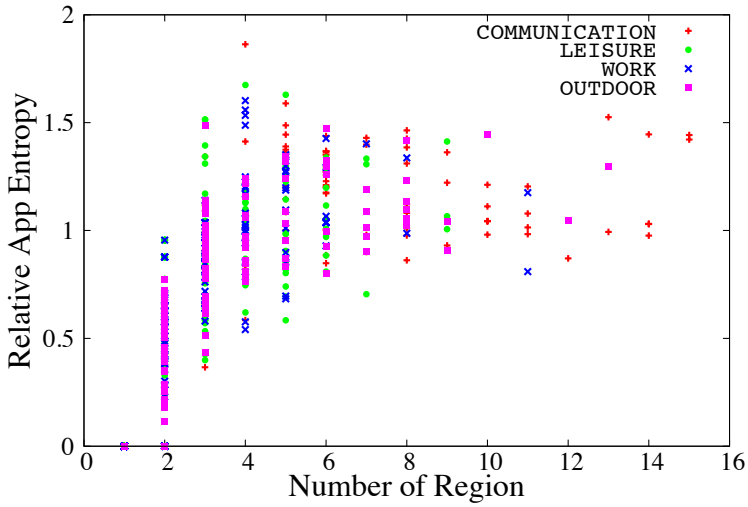
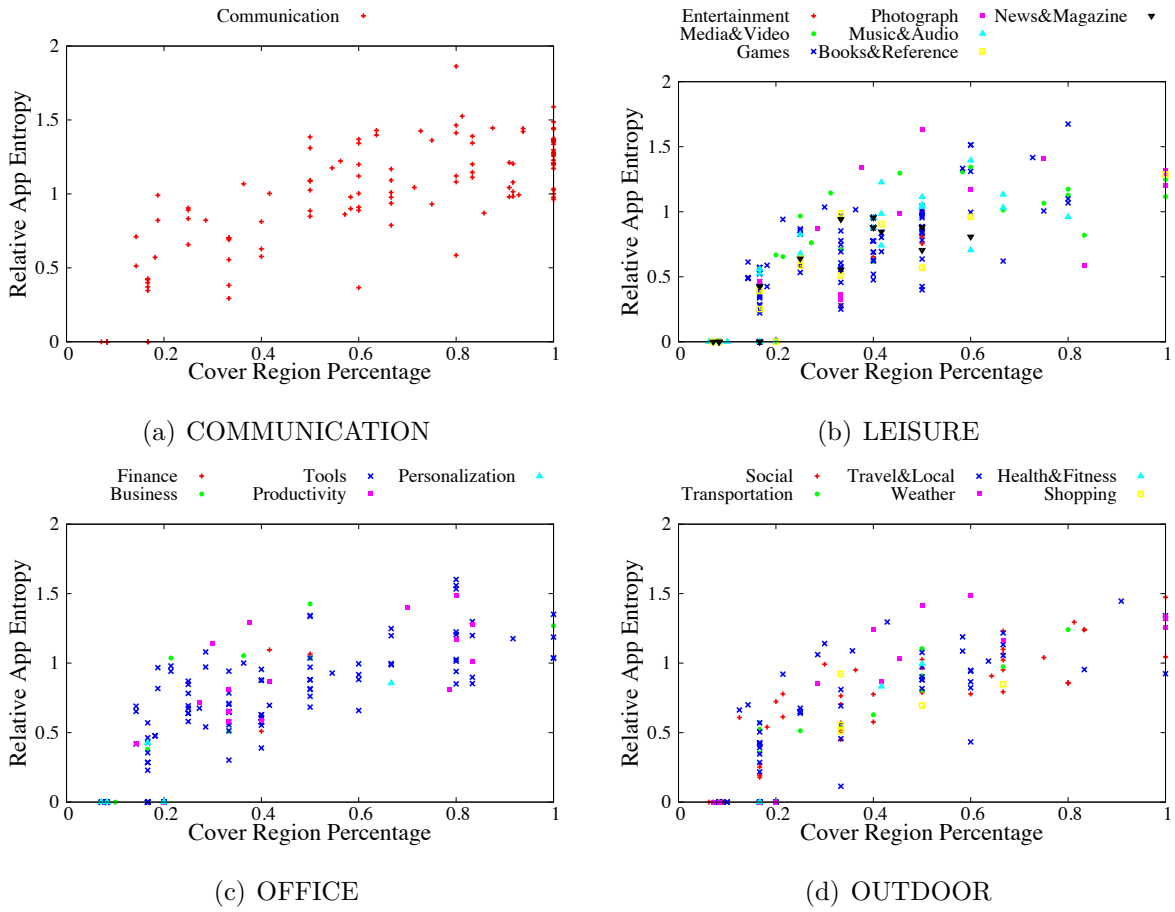


Figure 6.3: Relative App Entropy of Behaviors

Based on discovered ROIs, we also analyze the property of each behaviors by plotting the relative Apps entropy of each App App_i belong to the behavior, where the relative App entropy is the value of the App entropy divided by the location entropy (e.g. $\frac{H(\ell|App_i)}{H(\ell)}$). The App_i will be a non-location dependent application when the relative App entropy of App_i is larger than 1. Figure 6.3 illustrates that the relative App entropy increases proportionally to the number of the regions the App appears (e.g. $|\{l : App_i \in \ell\}|$). It fits our observation that when an App is executed on more regions, the less possible for the App to be utilized depending on location and to represent the user's behavior preference on it.

In Figure 6.4, we further investigate the relation between the relative App entropy and the cover region percentage of Apps under different behaviors, where the cover region percentage



(a) COMMUNICATION

(b) LEISURE

(c) OFFICE

(d) OUTDOOR

Figure 6.4: Relative App Entropy of Different Categories

is the number of the regions the App appears over the number of ROIs (e.g. $\frac{|\{l:App_i \in \ell\}|}{|\ell|}$). In Figure 6.4(a), the most of COMMUNICATION Apps locate at top-right area (relative App entropy > 1 and cover more than half ROIs), which means COMMUNICATION usually happens in most regions the user stays and not focus on any particular region. While most of LEISURE Apps in Figure 6.4(b) locate at bottom-left area, which means our users are usually LEISURE at some specific regions. However, few LEISURE such as listening music, watching videos, playing games could also happen in more than half regions. The OFFICE behavior in Figure 6.4(c) is similar to LEISURE behavior, but some Tools (like Clock) can be used on many regions. In Figure 6.4(d), the relative App entropy of OUTDOOR is proportional to the cover region percentage, which means our users do not execute OUTDOOR Apps on particular regions. Because when a user is familiar to a region, he/she has no need to utilize OUTDOOR Apps to help explore the region.

6.2 Case Study: Personal Semantic Regions

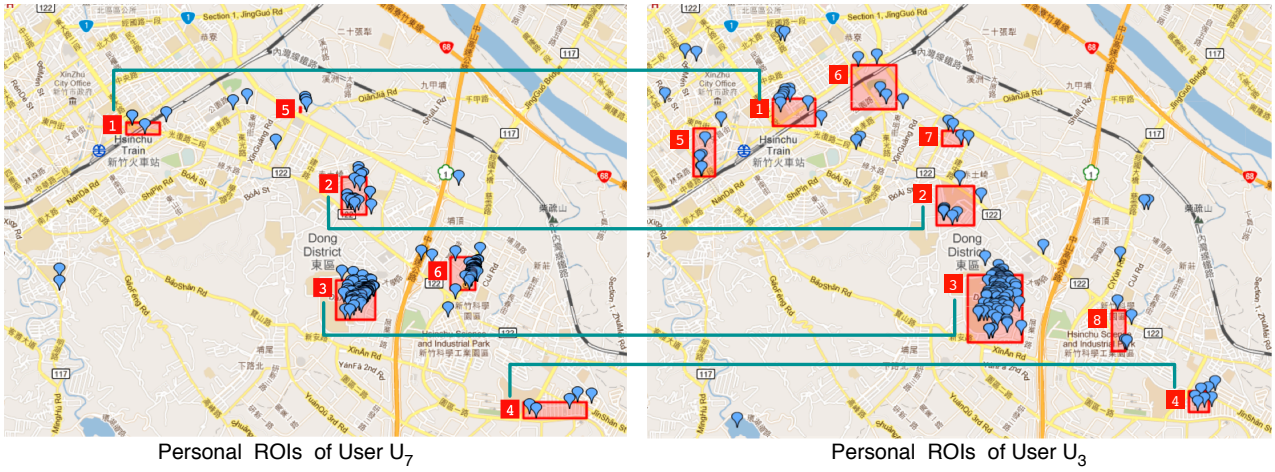


Figure 6.5: Case study of user U_3 and user U_7

Region	User U_3		User U_7	
	App Usage Pattern	AF-RLF Score	App Usage Pattern	AF-RLF Score
ℓ_1	PATS (Travel & Local) NubiNews (News & Magazine) FriendStream (Social)	0.023 0.017 0.004	GoogleMap (Travel & Local) Fxcamera (Photograph) AngelPiano (Games)	0.052 0.049 0.044
ℓ_2	Album (Media & Video) Skype (Communication)	0.013 0.01	Facebook (Social) Calendar (Productivity) Skype (Communication)	0.035 0.023 0.017
ℓ_3	Clock (Tools) NubiNews (News & Magazine) Facebook (Social)	0.036 0.018 0.017	MMS (Communication) Calendar (Productivity) Bump (Social)	0.027 0.025 0.023
ℓ_4	NubiNews (News & Magazine) FriendStream (Social)	0.045 0.014	SkyMap (Books & Reference) Bubbles (Games) Naruto (Games)	0.015 0.009 0.008

Table II: App usage patterns on common regions

In this subsection, we perform a case study by using App usage pattern to compare the personal semantic of different smart phone users on the same locations. Figure 6.5 shows parts of personal regions of user U_3 and user U_7 in common place. The blue marks show the locations where the user stays over 15 minutes, and the red rectangles show the personal ROI where the user stays at least three times. Both users share four common regions ($\ell_1 - \ell_4$) which are linked in two maps. Because the different users could be attracted by different parts of a region, the common region shared by both users will not cover exactly the same. However,

by referring the geographic semantic of the region, the common region can be identified. The App usage patterns ($K = 3$) on the four common regions are presented in Table II. Each App in a App usage pattern can vote its score to the behavior semantic corresponding to the category of the App. After voting, the behavior semantic can be determined by comparing the summation score. Take the behavior semantic of user U_3 on ℓ_1 as an example, PATS is belong to Travel&Local, so it votes its score 0.023 to OUTDOOR, and NubiNews votes 0.017 to LEISURE, and FriendStream votes 0.004 to OUTDOOR, respectively. As a result, the behavior semantic in this example is OUTDOOR.

Region	Geographic Semantic	Behavior Semantic	
		User U_3	User U_7
ℓ_1	Hsinchu Train Station	OUTDOOR	LEISURE
ℓ_2	NTHU night market	LEISURE	OUTDOOR
ℓ_3	NCTU Campus	WORK	COMMUNICATION
ℓ_4	JinShan st.	LEISURE	LEISURE

Table III: Personal Semantics on common regions

The personal semantics of user U_3 and user U_7 on four common regions are presented in Table III. Obviously, in a common region (like ℓ_1,ℓ_2 or ℓ_3) with the same geographic semantic, different users could have different behaviors even there are only four types of user behaviors. Moreover, if we investigate the App usage pattern on the region (ℓ_4) where both users have the same behavior type, the Apps they used are much different. In ℓ_4, user U_3 prefer to read news, while user U_7 prefers to play games. Thus, to extract the personal semantic, not only the geographic information but the user behavior should be considered. Experimental evaluation of the effectiveness of extracting personal semantic is discussed in the following.

6.3 Effectiveness of App Usage Pattern

On each personal enhanced ROI, our system will find the App usage patterns on the ROI. To evaluate the effectiveness of our methods, users fill some questionnaires honestly to reflect how they used these Apps and answer whether an App is a representative App. Thus, we are fully aware of the ground truth about users usage behaviors from their feedback. Since each

representative App record could reflect the user's behavior, we randomly select representative App records on each region and testify the effectiveness of App usage pattern provided by each method. If the representative App is in the App usage pattern, the method can reflect the behavior semantic for the region. We take *precision* and *nDCG* as our performance metrics. Here, *precision* is used to evaluated how many Apps we predicted are exact the representative Apps. It is formulated as the number of correct predictions divided by the number of representative Apps. Moreover, to measure the effectiveness of a representative App, the validation is tested by the Normalized Discounted Cumulative Gain (nDCG), which is a well-known measure of effectiveness of a Web search engine algorithm. The premise of *nDCG* is that representative Apps appearing lower in the ranking list should be penalized. Its value is reduced logarithmically proportional to the position of the list. For a transaction that the representative *App* is at position p, the value of *nDCG* is computed as:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(1+i)}, nDCG_p = \frac{DCG_p}{IDCG_p} \tag{1}$$

, where *IDCG* is an ideal *DCG* with the value 1.

Intuitively, the more frequent an App occurs on a given region, the more important the App could be and also the more possible the App to be the representative App. We compare our two approaches AF-ILF and AF-RLF with a frequency-based approach which provides a list ranking by the App usage frequency on the region. Figure 6.6 shows the comparison results of these approaches in terms of *precision*. Each approach will provide top 6 Apps as its ranking list. In general, AF-RLF dominates other two approaches for almost all the users. Notably, application frequency approach find much less representative App and the precision of the application frequency approach is much lower than other approaches, because most of the top frequently used Apps are usually executed on many locations and not able to represent for a particular place.

In Figure 6.7, we also compare the result with the value of K varied. In general, when the size of list K increases, precision values of every methods linearly increase in Figure 6.7(a).

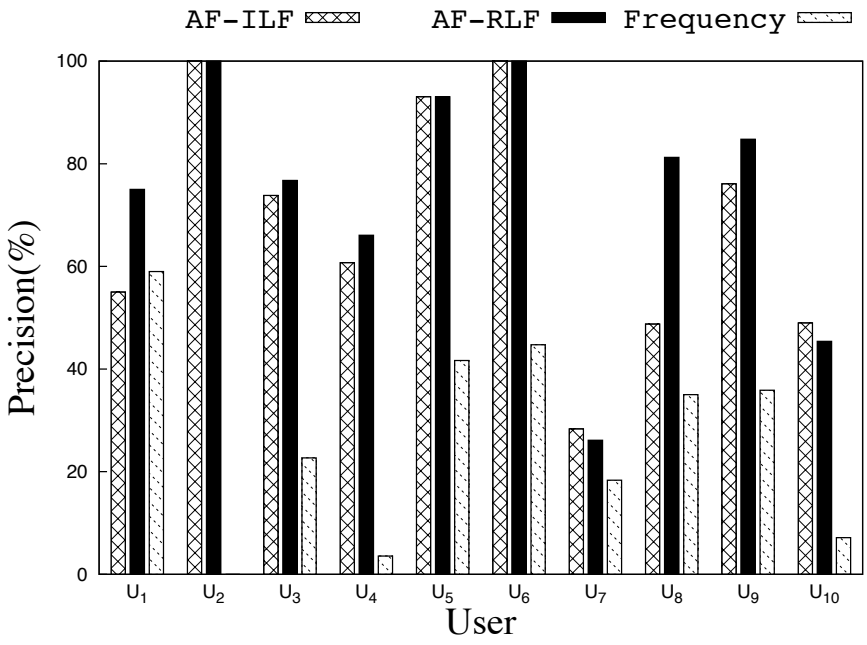
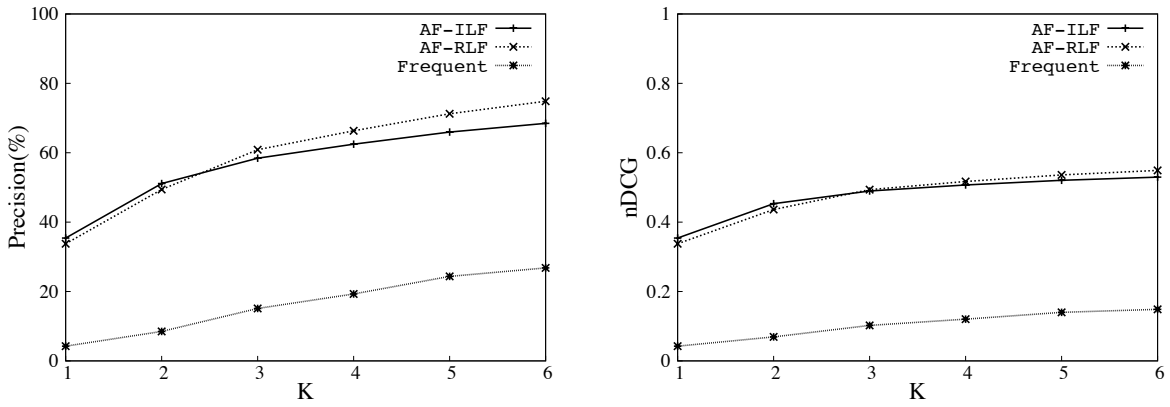


Figure 6.6: Precision for determining representative apps.



(a) Precision (b) nDCG

Figure 6.7: Comparison of different approaches.

Moreover, AF-RLF outperforms others approaches when $K \geqslant 3$. The reason for AF-RLF can achieve better result than AF-ILF is that some representative Apps, which only occur once in a region but still has lower score in AF-ILF function, should be ranked higher in the ranking list. Figure 6.7(b) shows the nDCG results of different approaches. Similarly, our approaches still work well when the position of the representative App in the ranking list is considered. However, AF-ILF and AF-RLF increase little when $K \geqslant 3$, mainly because they have already retrieved the most representative Apps in the beginning.

6.4 Effectiveness of Extracting ROI

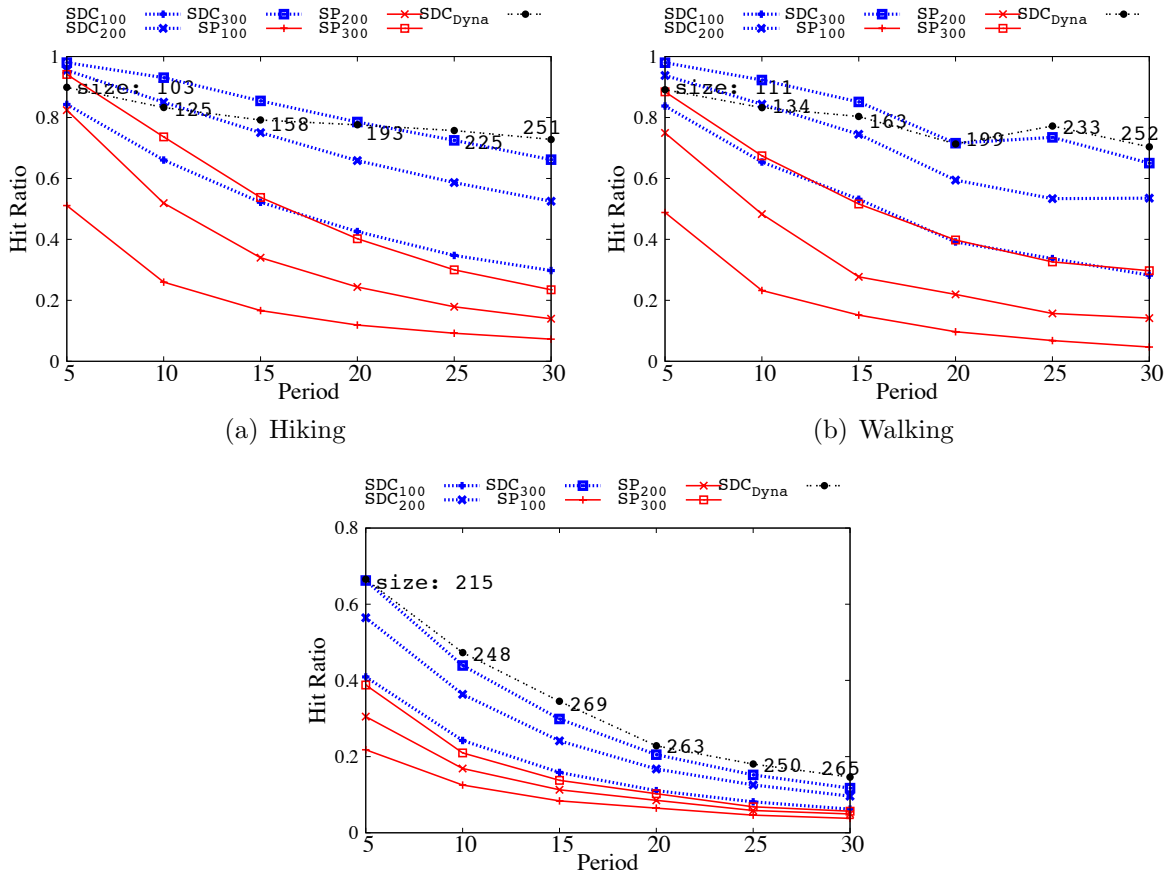
We evaluate the effectiveness of extracting ROIs with another real dataset from EveryTrail [1] in California. Each data includes an labeled activity trail (a trajectory) and a set of photos with geographic information where are taken by user. We assume the ground truth that location with photo is where the activity happen at. Each photo represents a interesting of the user (photo taker) and each region containing the photos can be considered as a interesting region. Three kind of activity (Hiking, Road Biking, Walking) in California are selected. The major difference between each activity is the average speed (Road Biking > Walking > Hiking). Table IV shows the total number of trajectories and photos for each activity.

Activity	# Trajectory	# Photo
Hiking	3839	33065
Road Biking	5032	11968
Walking	955	4685

Table IV: Dataset of each activity in California

In order to evaluate the effectiveness of ROI candidates discovery, we compare the region coverage of SDC with that of SP under varying conditions. A region coverage is measured as the hit ratio of the photos enclosed by discovered region to total photos for each activity. We set SDC parameters as follows: the partition smoothing parameter $fold = 3$, the density threshold $\xi = 0.02$ for all datasets. There are two parameters setting for SP: distance and time thresholds. For comparison with SP fairly, the dynamic size of a region is constrained as a fixed size of SP. Thus, We set various distance thresholds (100, 200, 300 meters) of stay point as the radius of the region around stay point and also as the radius around mean point discovered via SDC. In additional, we compare above regions of fixed size with the regions of dynamic size discovered via SDC. The time threshold of SP is set as the period of activity for SDC and varied from 5 minutes to 30 minutes.

For each activity, in Figure 6.8, the hit ratio of our method is much higher than that of SP. As expected, SDC shows the coverage of discovered region with dynamic size is better than that with fixed size while the average size (the size number marked with SDC curve in Figure 6.8) is smaller than the fixed size, especially in datasets of slow-speed activity (Hiking and Walking). The reason is our method can vary the covering shape and size of discovered region according to the physically passed region. Besides, the hit ratio is much lower in high-speed



(a) Hiking

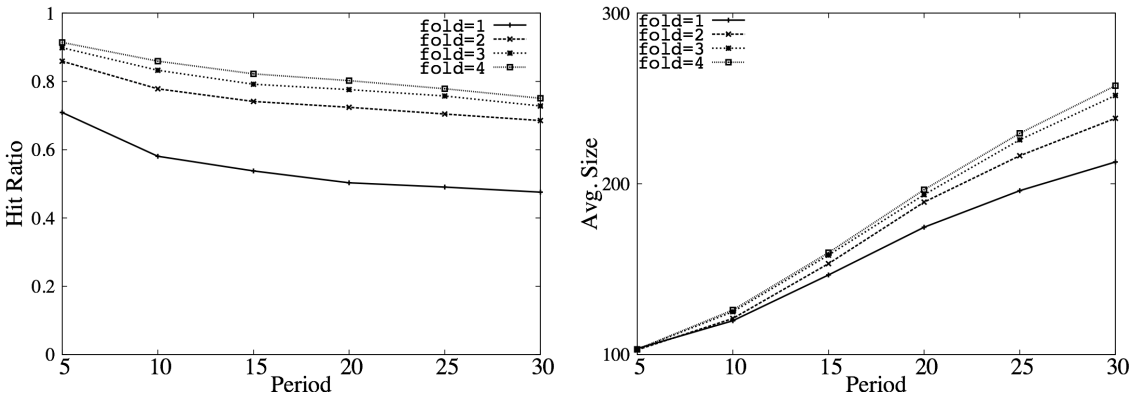
(b) Walking

(c) Biking

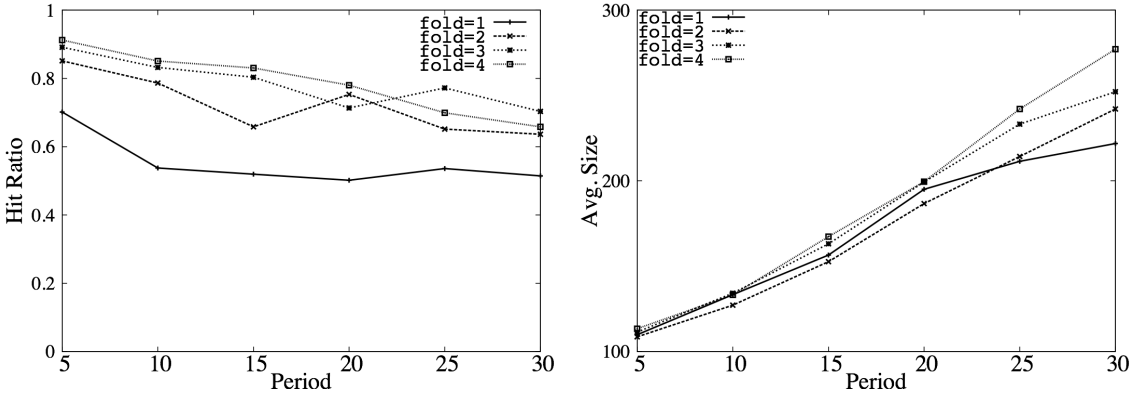
Figure 6.8: Region coverage

activity than in low-speed activity, since the staying region is much harder to be obtained when the activity has higher average speed and has many sudden changes of direction or speed. These results prove that using SDC for ROI candidates discovery is obviously more precise than using SP under different average speed. Another observation demonstrates that hit ratio decreases when the period of activity (time threshold) increases. Because the period of activity is a user-defined parameter which indicates the minimum duration of an activity, less region will be discovered when the activity we want to found in the region is expected to keep running longer.

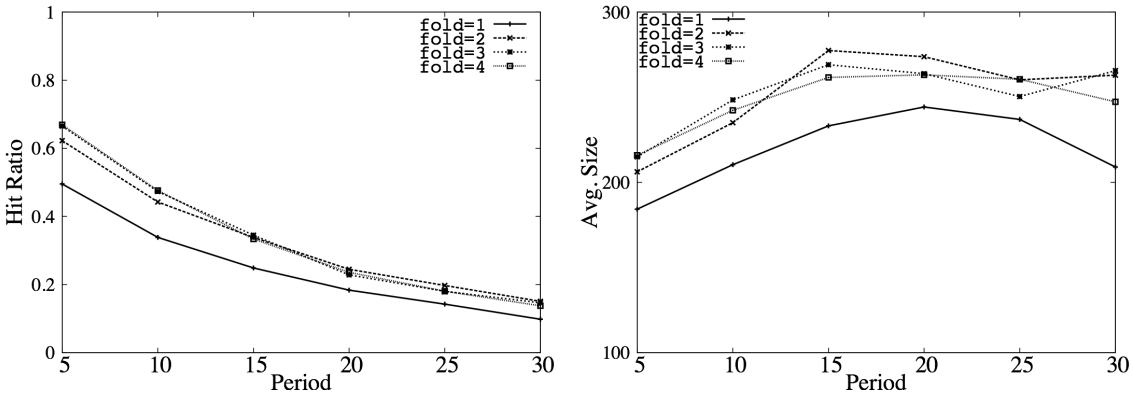
6.5 Effect of Region Discovery Parameter



(a) Hiking



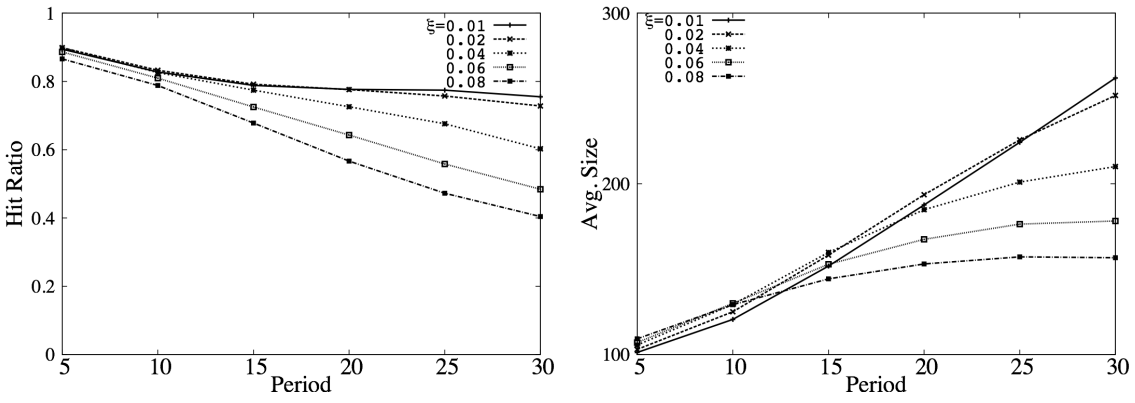
(b) Walking



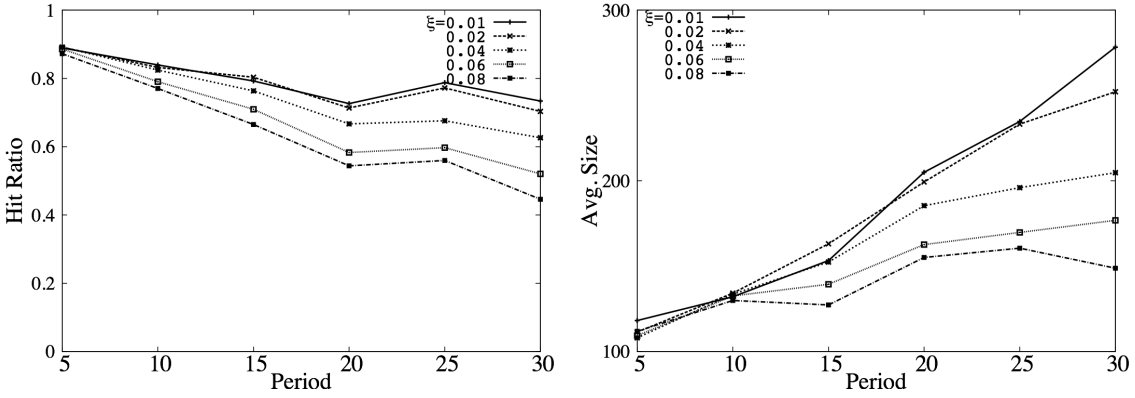
(c) Biking

Figure 6.9: Comparison under different folds

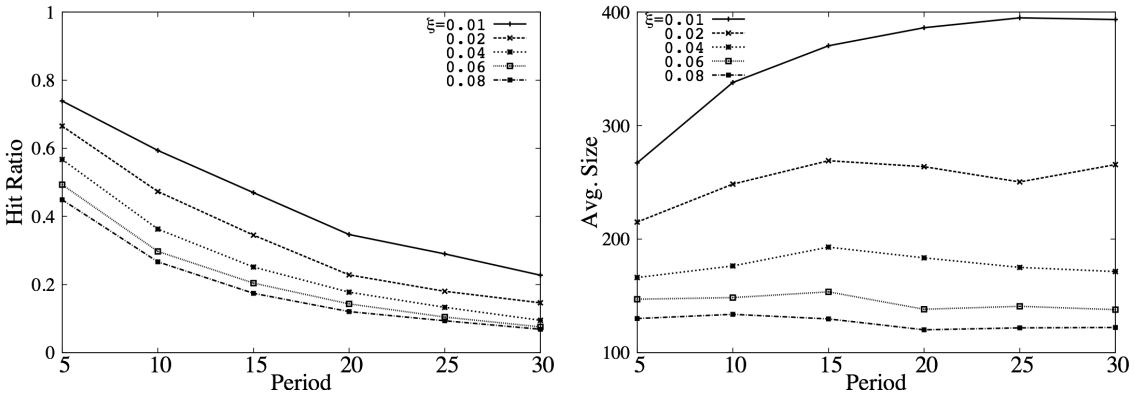
We next study the effect of partition smoothing parameter $fold$ for region discovery. Figure 6.9 shows the hit ratio with smoothing increases at least 25% compares to that with no smoothing ($fold = 1$). The result indicates the partition smoothing can reduce the trajectory partition loss and achieve high region discovery precision. Besides, it changes slightly when $fold \geq 3$ because there is no more dense region split into different regions.



(a) Hiking



(b) Walking



(c) Biking

Figure 6.10: Comparison under different thresholds

The effect of density difference threshold ξ is studied too, where the ξ is closely related to the average size of discovered region. When ξ decreases, the boundary of the discovered region is more difficult to be distinguished. Because it may keep the margin of discovered region, which will be removed at expending step in Algorithm 1. Although it can catch more activity and achieve better hit ratio, in Figure 6.10, the average size will increase dramatically.

Especially for the high-speed activity (Biking), the average size dramatically grows when $\xi \geq 0.02$. On the other hand, with appropriate density threshold, we implement $\xi = 0.02$ for comparison, it can maintain the small region and not loss much hit ratio.

6.6 Accuracy of ROI Discovery

To show the accuracy of regions of interests discovery, we obtained a user's trajectory over one week and labeled the top five frequent regions. We then generated 1000 different trajectory dataset each have 100 similar trajectories to the original trajectory. For each trajectory, we set the period of activity $T = 15$ minutes to discover ROI candidates. We take the ROI candidates as nodes in a $5 * 5$ map. Frequency and radius of each frequent semantic regions are stated in Table V.

# Region	Frequency	Radius
2	50%	0.5
2	80%	0.1
1	30%	0.8

Table V: Dataset of regions of interests

$MinSR$	Precision	Recall	F-measure
10%	0.856 ± 0.068	0.995 ± 0.013	0.919 ± 0.04
20%	0.903 ± 0.065	0.884 ± 0.005	0.894 ± 0.036
30%	0.916 ± 0.066	0.441 ± 0.074	0.592 ± 0.068

Table VI: Impact of minimum support

We take F-measure to analyze the accuracy of discovered ROIs. Precision is defined as the overlapped area discovered in labeled regions divided by the total discovered area, and recall is defined as the overlapped area discovered in labeled regions divided by the total area of existing labeled regions. The definition of F-measure is the harmonic mean of precision and recall:

$$F = 2 * \frac{precision * recall}{precision + recall}$$

A higher precision score means the higher representative of discovered regions, while a higher recall score means the higher coverage of labeled regions. Although a larger region can cover

more labeled regions and obtain high recall, it is hard to distinguish these labeled regions and has low precision.

In Table VI, we fix the radius d_h as 0.5 and report the performances of our model under different minimum support ($MinSR$) requirement for a ROI. The entry value in Table VI denotes the mean and standard deviation of precision, recall and F-measure. As shown in the table, our method can achieve high precision under different $MinSR$. However, when the requirement of $MinSR$ increase, it is much harder to find regions of low frequency in a large radius.

Chapter 7

Conclusion

In this paper, we investigate the personal semantic region discovery problem, which aims to automatically discover regions of interests and annotate personal semantics to regions discovered along with user's trajectories. Such personal semantic regions are a crucial prerequisite for personal recommendation, location-based service, or data cleaning. In our framework, we first propose a sequential density clustering approach to extract ROI candidates based on both the spatial and the temporal domains of trajectories. Then we further adopt a shared nearest neighbor (SNN) clustering to explore ROIs from ROI candidates extracted from multiple trajectories. In light of these discover ROIs, we characterize not only the geographic semantics by nearby POIs but the behavior semantics by App usage patterns, where the patterns is summarized as a ranking list of representative applications which reveal the user's behavior semantics. To identify App usage patterns for extracting behavior semantics, non-location dependent applications are filter out by location entropy of applications, and several measures based on TF-IDF are proposed to evaluate the representativeness of applications to the regions. Finally, we present comprehensive experimental results over various real datasets collected in smart phones and trajectory community. The results demonstrate the effectiveness of our approach to discover regions from user's trajectories and infer personal semantics from application usage patterns.

Bibliography

[1] Everytrail – gps travel community. `http://www.everytrail.com`, 2010.

[2] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macêdo, B. Moelans, and A. A. Vaisman. A Model for Enriching Trajectories with Semantic Geographical Information. In *Proc. of GIS*, page 22, 2007.

[3] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proc. of SIGMOD*, pages 49–60, 1999.

[4] X. Cao, G. Cong, and C. S. Jensen. Mining Significant Semantic Locations From GPS Data. In *Proc. of VLDB*, volume 3, pages 1009–1020, 2010.

[5] D. Choujaa and N. Dulay. Predicting human behaviour from selected mobile phone data points. In *Proc. of UbiComp*, pages 105–108, 2010.

[6] N. Eagle and A. Pentland. Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[7] N. Eagle and A. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, May 2009.

[8] L. Ertoz, M. Steinbach, and V. Kumar. A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In *2nd SIAM International Conference on Data Mining*, 2002.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of KDD*, pages 226–231, 1996.

[10] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory Pattern Mining. In *Proc. of KDD*, pages 330–339, 2007.

[11] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[12] R. A. Jarvis and E. A. Patrick. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. on Computers*, 22(11):1025–1034, 1973.

[13] H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A Hybrid Prediction Model for Moving Objects. In *Proc. of ICDE*, pages 70–79, 2008.

[14] H. Jeung, H. T. Shen, and X. Zhou. Mining Trajectory Patterns Using Hidden Markov Models. In *Proc. of DaWaK*, pages 470–480, 2007.

[15] K. S. Jones. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28:11–21, 1972.

[16] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining User Similarity Based on Location History. In *Proc. of GIS*, page 34, 2008.

[17] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining Periodic Behaviors for Moving Objects. In *Proc. of KDD*, pages 1099–1108, 2010.

[18] L. Liao, D. Fox, and H. A. Kautz. Location-Based Activity Recognition. In *Proc. of NIPS*, 2005.

[19] L. Liao, D. Fox, and H. A. Kautz. Location-Based Activity Recognition using Relational Markov Networks. In *Proc. of IJCAI*, pages 773–778, 2005.

[20] C.-R. Lin and M.-S. Chen. On the Optimal Clustering of Sequential Data. In *Proc. of SDM*, 2002.

[21] C.-H. Lo, W.-C. Peng, C.-W. Chen, T.-Y. Lin, and C.-S. Lin. CarWeb: A Traffic Data Collection Platform. In *Proc. of MDM*, pages 221–222, 2008.

[22] E. H.-C. Lu and V. S. Tseng. Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In *Proc. of MDM*, pages 273–278, 2009.

[23] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A Clustering-Based Approach for Discovering Interesting Places in Trajectories. In *Proc. of SAC*, pages 863–868, 2008.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27:379–423, July 1948.

[25] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010.

[26] K. Xie, K. Deng, and X. Zhou. From Trajectories to Activities: A Spatio-temporal Join Approach. In *Proc. of GIS-LBSN*, pages 25–32, 2009.

[27] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the Semantic Annotation of Places in Location-Based Social Networks. In *Proc. of KDD*, 2011.

[28] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng. Mining User Similarity from Semantic Trajectories. In *Proc. of GIS-LBSN*, pages 19–26, 2010.

[29] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative Location and Activity Recommendations with GPS History Data. In *Proc. of WWW*, pages 1029–1038, 2010.

[30] Y. Zheng and X. Xie. Learning Travel Recommendations from User-generated GPS Traces. *ACM TIST*, 2(1):2, 2011.

[31] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining Interesting Locations and Travel Sequences From GPS Trajectories. In *Proc. of WWW*, pages 791–800, 2009.

[32] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Collaborative Location and Activity Recommendations with GPS History Data. In *Proc. of WWW*, pages 26–30, 2010.

[33] C. Zhou, N. Bhatnagar, S. Shekhar, and L. G. Terveen. Mining Personally Important Places from GPS Tracks. In *Proc. of ICDE Workshops*, pages 517–526, 2007.