A Framework of Mining Semantic Regions from Trajectories

Chun-Ta Lu[1], Po-Ruey Lei[2], Wen-Chih Peng[1], and Ing-Jiunn Su[2]

[1] National Chiao Tung University, Hsinchu, Taiwan, ROC
{lucangel, wcpeng}@gmail.com
[2] Chung Cheng Institute of Technology, National Defense University, Taoyuan, Taiwan, ROC
{kdboy1225, suhanson}@gmail.com

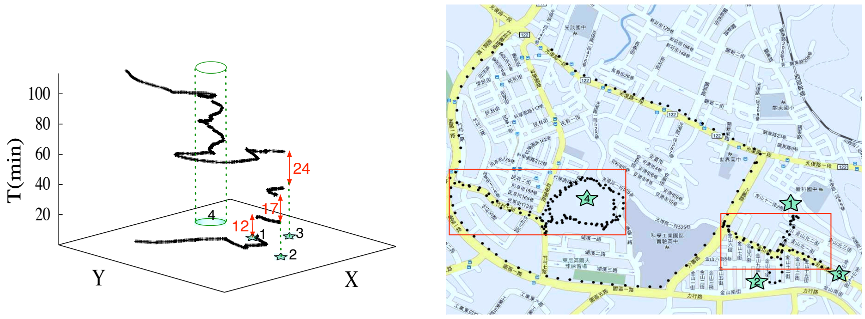
Abstract. With the pervasive use of mobile devices with location sensing and positioning functions, such as Wi-Fi and GPS, people now are able to acquire present locations and collect their movement. As the availability of trajectory data prospers, mining activities hidden in raw trajectories becomes a hot research problem. Given a set of trajectories, prior works either explore density-based approaches to extract regions with high density of GPS data points or utilize time thresholds to identify users' stay points. However, users may have different activities along with trajectories. Prior works only can extract one kind of activity by specifying thresholds, such as spatial density or temporal time threshold. In this paper, we explore both spatial and temporal relationships among data points of trajectories to extract semantic regions that refer to regions in where users are likely to have some kinds of activities. In order to extract semantic regions, we propose a sequential clustering approach to discover clusters as the semantic regions from individual trajectory according to the spatial-temporal density. Based on semantic region discovery, we develop a shared nearest neighbor (SNN) based clustering algorithm to discover the frequent semantic region where the moving object often stay, which consists of a group of similar semantic regions from multiple trajectories. Experimental results demonstrate that our techniques are more accurate than existing clustering schemes.

Keywords: Trajectory pattern mining, sequential clustering and spatial-temporal mining

1 Introduction

Knowledge discovery from spatial-temporal data has risen as an active research because of the large amount of trajectory data produced by mobile devices. A trajectory is a sequence of spatial-temporal points which records the movement of a moving object. Each point specifies a moving location in space at a certain instant of time. The semantic knowledge may contain in some re-appear trajectories and can be applied in many applications, such as trajectory pattern mining for movement behaviors [6, 19, 8], predicting user location [10, 18], and location-based activity discovery [13, 14, 7]. Unfortunately, locations may not be repeated exactly in similar trajectories. The common preceding task for the above works is to discover the regions for replacing the exact locations where moving objects often pass by or stay. Such a region summarizes a set of location points from different trajectories that are close enough in the spatial space. Then, the relation between regions can be extracted for knowledge analysis. Intuitively, the quality of regions directly affects the analysis result of trajectory data. Thus, in this paper, we focus on effectively and precisely discovering regions from trajectory data where can imply the potential of users are likely to have some kinds of activities, called semantic regions.

Traditionally, regions are extracted from trajectory points by density-based clustering methods (e.g., DBSCAN [4]). Given the definition of distance (i.e., measure of dissimilarity) between any two points, regions with higher density are extracted in terms of clustering similar data points in the spatial domain.



(a) The original trajectory

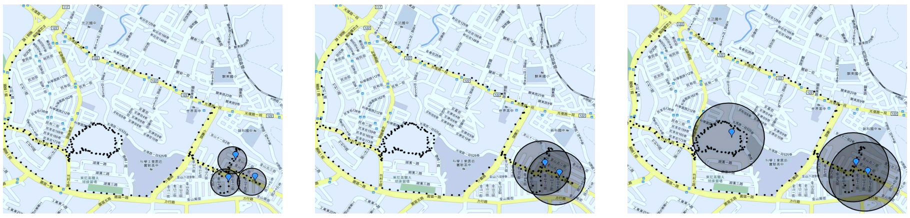
(b) Semantic region discovered by density-based approach

Fig. 1. An example of discovering semantic regions from a trajectory

However, such regions, extracted by clustering spatial points without considering sequential relation, only focus on the geometrical properties of trajectories. Consider an example in Figure 1, where there are four activities involved in this trajectory. Each region associated with one activity is marked with a star in Figure 1(b). We observed that, from the original trajectory based on spatial-temporal domain, there should be three indoor regions (region 1, 2 and 3) because of the appearance of temporal transition gaps which are labeled with stay durations in Figure 1(a). 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 (e.g., restaurant, home or office). In additional, in Figure 1(b), there is an outdoor activity (i,e., in region 4) where the user is walking around a lake. Two regions where represented by minimum bounding rectangles (MBRs) are discovered by a spatial density based clustering algorithm, DBSCAN ($Minpts = 7$, $Eps = 50$ meters) . There are three problems in this example. First, some semantic regions are missing. By verifying with the ground truth (i.e., four regions with stars in Figure 1(b)), only two regions are detected by DBSCAN and region 2 is missing. As shown in Figure 1(b), while this 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 GPS data points to form a cluster. Second, granularity problem causes the indistinguishability between region 1 and 3. Third, road-sections and intersections, where an object often passes but carries non-semantic meaning to the user, are included in both discovered regions. The above example indicates that only exploring density-based approaches in the spatial domain of data points in trajectories cannot discover semantic regions.

Recently, the authors in [21] proposed the concept of stay point detection to discover the stay regions. Unlike density based clustering, 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. The authors claimed that a stay point can stand for a geographic region and carry a particular semantic meaning. However, a trajectory usually contains more than one activity, such as driving, walking, sightseeing, staying and so on. Each activity has different distance density and speed. In other words, the density of trajectory points vary from different activities. Thus, the traditional density clustering approach or stay point detection, which using universal parameter to detect the clusters only for a certain density, cannot discover all semantic regions. Figure 2 shows regions discovered by the stay points approach, where the time threshold is fixed to 10 minutes and three distance thresholds are set to 100 meters, 200 meters and 250 meters. When distance threshold is set to 100 meters, in Figure 2(a), there are three stay points mapping to three semantic regions, but the semantic region 4 (lake), with much larger area of activity, cannot be detected. The regions are not detected completely until the distance threshold is larger than 250 meters. On the other hand, the other three regions have been mixed and their coverage have been overlapped shown in Figure 2(b) and Figure 2(c). As such, the stay point approach considers both the temporal and spatial thresholds for detecting regions. However, the stay point approach is highly dependent to thresholds. Consequently, to detect regions with a variety of activities, the stay point approach may need to have different settings of thresholds.

Consequently, in this paper, we first propose a sequential density clustering approach to extract candidate semantic regions based on both the spatial and the temporal domains for GPS data points in trajectories. The density is measured by cost function to analyze the density distribution of a trajectory. The cost function reflects the local configuration of the trajectory points in spatial-temporal data space. In light of candidate semantic regions, we further propose shared nearest neighbor (SNN) clustering to extract frequent semantic regions from a set of candidate semantic regions. Our approach is nonexclusive to be applied in many different activity scenarios, not being to one single application.



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

Fig. 2. An example of regions discovered by the stay point approach

Our experiments demonstrate that semantic regions can be extracted precisely as well as efficiently. The main contributions of this paper are summarized below.

- The scheme of region extraction is proposed for effectively and precisely semantic region discovery.
- We propose a sequential density based clustering method to discover semantic regions from a trajectory. The clustering method takes both spatial and temporal domain into account.
- We define the similarity between semantic regions and develop a shared nearest neighbor based clustering algorithm to discover frequent semantic regions from trajectory dataset.
- We present comprehensive experimental results over various real datasets. The results demonstrate that our techniques are more accurate than existing clustering schemes.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Our framework of mining semantic regions is proposed in section 3. In section 4, we evaluate our framework by real trajectory datasets. Finally, section 5 concludes this paper.

2 Related Work

Hot region detection has been widely used in the field of trajectory data analysis such as trajectory pattern mining [17, 2, 6, 11, 8], moving objects' location prediction [10, 18], location-based activity discovery [14, 13, 21] and so on. 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 [4] and OPTICS [1]. In density based clustering, clusters are regions of high density separated by regions of low density. Based on density based clustering algorithms [6, 11, 10, 18], the regions are extracted only according to the density in spatial domain without considering the density in temporal domain. Giannotti et. al. [6] adopted grid density region to discover popular regions as ROIs where dense cells in space are detected and merged if they are neighbors. It implies that popular regions can be extremely large. Thus, they have to give additional constrains to select significant and limited regions to represent ROIs. The authors in [11] 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. The hybrid location prediction model proposed by [10] that divides a trajectory into several periodic sub-trajectories. Then, frequent regions of the same time offset are extracted by using DBSCAN to cluster locations from sub-trajectories.

For the purpose of knowledge discovery of the ROIs which contains activity-related meaning to users, few existing works [12, 20, 21] aimed to applying sequential constraint to a single sequence. The authors in [12, 20, 21] proposed a stay point and claimed that can stand for a geographic region and carry a particular semantic meaning. A stay point is the mean point of a sub-sequence where the consecutive points of a examined point do not exceed the distance threshold during the period of time threshold. Each stay point contains information about mean coordinates, arrival time and leaving time. In addition, the authors in [21] proposed stay regions extracted from stay points via grid based clustering algorithm.

All of the above techniques have some deficiencies for discovery ROIs from trajectory data. First, traditional clustering approach only considers similarity in one domain, i.e.,spatial domain only. They have focused on geometric properties of trajectories, without considering the temporal information or sequential relation. The region extraction for semantic analysis has to consider both spatial and temporal domains. Second, applying a universal density threshold for cluster discovery may either miss regions with different density or merge non-related regions. In this paper, the challenge is that trajectories may consist of different activities and each activity has different distance density and speed distribution. We want to extract significant and precise regions with semantic meaning from trajectories and these regions can imply certain activity of moving objects by spatial-temporal clustering approach.

3 A Framework of Mining Semantic Regions

3.1 Overview

We propose an effective and precise algorithm to discover semantic regions from trajectory data based on spatial-temporal density model and sequential density clustering, and we develop a shared nearest neighbor based clustering method to discover frequent semantic regions from multiple trajectories. Figure 3 outlines the framework for semantic region discovery. On the process of semantic region discovery, each trajectory is first partitioned into a set of trajectory segments. The spatial-temporal density of each segment is computed by cost function. Then, the sequential density clustering is applied to sequentially group the segments with similar density. The region where users may have some kinds of activities locates in the cluster with local maxima density. Finally, while each trajectory is transformed into a sequence of semantic regions, a set of similar

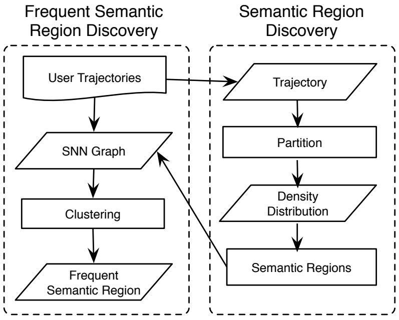


Fig. 3. Overview of extracting semantic regions

semantic regions is clustered to indicate the major frequent semantic regions from multiple trajectories.

3.2 Problem Formulation

Given a trajectory dataset of a moving object, our algorithm generates a set of clusters as semantic regions of each trajectory and a set of frequent semantic regions from the trajectory dataset. An object's trajectory is represented as a sequence of points $\{p_1, p_2, ..., p_i, ..., p_n\}$. Each point $p_i (1 \leq i \leq n)$ contains location (x_i, y_i) and timestamp (t_i). A trajectory can be partitioned into continuous segments $\{s_1, s_2, ..., s_l, ..., s_m, ...\}$ according to user-defined parameter T. Let T be an integer called period of activity that is the minimum duration of activity proceeding time we are interested in. For example, T can be set to 30 minutes for sightseeing at an interesting spot or 2 hours for exercising at the gym.

A semantic region is a spatial-temporal based cluster and is denoted as SR. The cluster C_k is a set of trajectory segments $\{s_l, s_{l+1}, s_{l+2}, ..., s_m\}$, where $m \geq l$. The cluster C_k is a semantic region if (1) the stay duration of each segment in C_k is not less than T (i.e. $|t_j - t_i| \geq T$) (2) and spatial-temporal density of C_k is higher than that of its adjacent clusters (C_{k-1} and C_{k+1}) by a predefined threshold ξ. A frequent semantic region is a representative region which indicates that this region appears in a sufficient number of trajectories. Such a sufficient number is defined as MinSR.

The spatial-temporal density of a segment mentioned above 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. Previous work [15] defines 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 a moving object staying in a location or lingering around some places. Our main idea of this research is to extract the region with semantic information where involves some activities of user. Because a trajectory does not involve only one activity in real world, the distance between location points can vary with different activities in spatial domain and the temporal interval from a point p_i to its succeeding point p_{i+1} can vary from seconds to hours. Furthermore, most location-acquisition technologies cannot localize and record current location under some condition. For example, when a GPS-embedded object enters a building or a cave, the GPS tracking device will lose satellite for a time interval until coming back outside and few points are recorded on such place. If we directly measure the spatial dissimilarity of the segment around this area, we cannot detect its significance. It implies that both spatial and temporal feature can affect the result of semantic region discovery. 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. 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 density function is stated as follows.

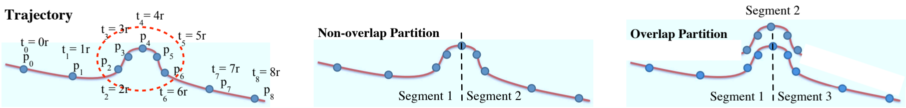
$$Density(s_l) = \log_e(1 + \frac{1}{Cost(s_l) + \gamma}), \tag{5}$$

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)

3.3 Discovering Semantic Regions

3.3.1 Trajectory Partition

Given an object's trajectory $\{p_1, p_2, ..., p_i, ..., p_n\}$, 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 [15], 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, where T is a period of activity, i.e., a minimum duration of activity we are interested in and the density of each sequential 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

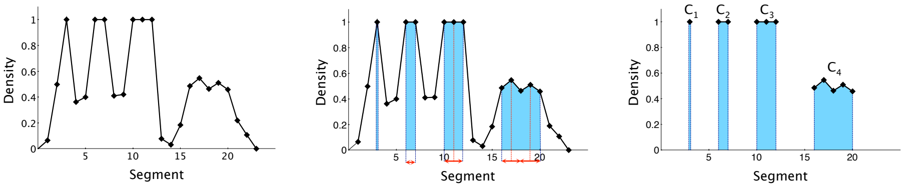
Fig. 4. Overlap partition

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(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(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(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 $fold \geqslant 3$.

3.3.2 Sequential Density Clustering Algorithm

We now present our sequential density clustering algorithm for semantic region discovery. Given a set of sequential trajectory segments S, our algorithm generates a set of clusters as semantic regions. We define a cluster as a sequential density-connected set. It requires a parameter ξ, density threshold for similarity measurement. Before clustering, each density D_k of partitioned segment S_k is



(a) Trajectory density distribution (b) Sequential density clustering (c) Clusters

Fig. 5. Sequential density clustering over trajectory density distribution

```
Algorithm 1: Sequential Density Clustering
  Input: A set of trajectory segments $S$, a density threshold $\xi$
  Output: A set of sequential density cluster $SDC$

Compute $D = Density(S)$ for each segment in $S$
Sequentially record the local max and local min from $D$ to an array $E$
foreach local max $E_i$ in $E$ do
    Take nearby local min and $E_i$ as a group $G : \{E_{i-1}, E_i, E_{i+1}\}$ and take the
    local max of last group in GroupSet as $E_{last}$
    if $|E_i - E_{i-1} \leqslant \xi|$ and $|E_{last} - E_{i-1} \leqslant \xi|$ then
        Merge $G$ with last group in GroupSet
    end
    else
        Add $G$ into GroupSet
    end
end
foreach group $G$ in $GroupSet$ do
    foreach $D_j$ from local max to local min in $G$ do
        if $|D_j - D_{j-1}| \leqslant \xi$ and $(|D_j - D_{j+1}| \leqslant \xi)$ then
            Add $S_j$ into of a density cluster $C$
        end
    end
    if $boundary(C) \neq boundary(G)$ then
        Add cluster $C$ into $SDC$
    end
end
return $SDC$
```

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 semantic region, i.e. a cluster, is extracted at where the density dramatically change. Thus, the semantic region discovery involves grouping the segments (if they belong to the same dense region) and setting boundary of dense region.

For instance, we let $T = 10$, $fold = 2$ to partition the trajectory in Figure 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 5(a). Such a sequence of density D is the input of the algorithm. Algorithm 1 shows the sequential density clustering to extract semantic regions 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 E_{i-1} and E_{i+1}, i.e., $G : \{E_{i-1}, E_i, E_{i+1}\}$. The algorithm consists two steps. In the first step (Line 3-11), 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 5(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 12-21), 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 5(c). Only regions with significant activity change are taken as semantic regions. 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 an non-semantic area.

3.4 Mining Frequent Semantic Regions

While semantic regions represent the location where a moving object proceeds with some kind of high dense activities in duration of time from a trajectory, it does not imply that those semantic regions are an object's "frequently" appearing at. 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 frequent semantic region. A frequent semantic region is a summary of a set of similar semantic regions from different trajectories. To define the similarity between semantic regions and discover the frequent semantic regions, we adopt the definition of shared nearest neighbor (SNN) [9] and SNN density-based clustering [3]. 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 frequent semantic regions is similar to find clusters. For each semantic region, 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. When applying SNN density based clustering to discover frequent semantic regions, we constrain the searching range of nearest neighbors is a radius D_h around the examined node. We define a semantic region is a frequent semantic region if each semantic region of 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 frequent semantic regions where an object often visits for certain activities.

```
Algorithm 2: Frequent Semantic Region Discovery Algorithm
  Input: A set of nodes, distance threshold $D_h$, minimum support $MinSR$
  Output: a set of clusters
1 Find the $MinSR$-nearst 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
```

We develop a frequent semantic region discovery algorithm (Algorithm 2) based on the property described in new SNN clustering algorithm [3]. 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 is regarded as a frequent semantic region. For each semantic region which has at least $MinSR$ similar semantic regions, it will be included in a frequent semantic region. 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.

4 Experiments

The experiments in this study are designed for two objectives. First, we compare the semantic region coverage of our method, Sequential Density Clustering (SDC), with Stay Point (SP) that is the method considering the sequential constraint in literature. Second, we verify the accuracy of frequent semantic region discovery. We conducted experiments on our prototype which was implemented in the python language on CarWeb [16], a traffic data collection platform on Ubuntu 9.10 operating system.

Table 1. Dataset of each activity in California

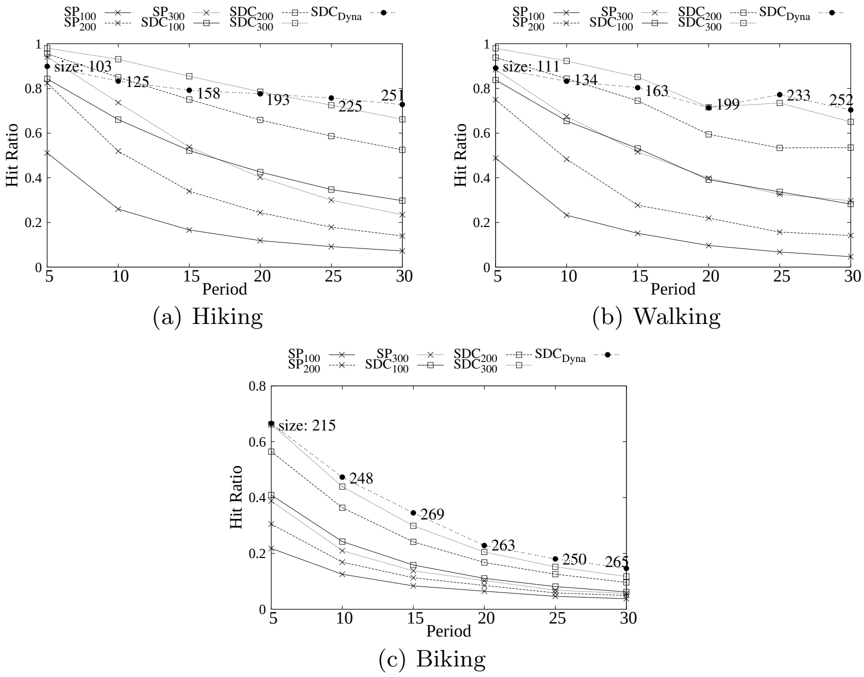
Activity	# Trajectory	# Photo
Hiking	3839	33065
Road Biking	5032	11968
Walking	955	4685

We evaluate the experiment with real dataset from EveryTrail [5] in California. Each data includes an labelled 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 (semantic) 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 1 shows the total number of trajectories and photos for each activity.

4.1 Evaluation of Semantic Regions

In order to evaluate the effectiveness of semantic region discovery, we compare the semantic region coverage of SDC with that of SP under varying conditions. A semantic 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 sematic 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, 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) 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 activity than in low-speed activity, since the semantic 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 semantic region discovery is obviously more precise than using SP under different average speed. Another observation demonstrates that hit ratio decreases when



(a) Hiking

(b) Walking

(c) Biking

Fig. 6. Semantic region coverage

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 semantic region will be discovered when the activity we want to found in the region is expected to keep running longer.

4.2 Accuracy of Frequent Semantic Regions

To show the accuracy of frequent semantic region discovery, we obtained a user's trajectory over one week and labeled the top five frequent semantic 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 $T = 10$ minutes to discover semantic regions. We take the semantic regions as nodes in a $5 * 5$ map. Frequency and radius of each frequent semantic regions are stated in Table 2.

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

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

Table 2. Dataset of frequent semantic regions

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

Table 3. Impact of minimum support

$MinSR$	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

A higher precision score means the higher representative of discovered regions, while a higher recall score means the higher coverage of labelled regions. Although a larger region can cover more labelled regions and obtain high recall, it is hard to distinguish these labelled regions and has low precision. In Table 3, we fix the radius D_h as 0.5 and report the performances of our model under different minimum support ($MinSR$) requirement for a frequent semantic region. The entry value in Table 3 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.

5 Conclusion

In this paper, we propose the concept of semantic region that indicates regions along with trajectories where users may proceed with some activities. First, spatial-temporal cost is introduced to model the density distribution of a trajectory. Then, we adopt a sequential density clustering algorithm to extract the semantic regions. Based on semantic region discovery, we define the similarity between semantic regions and devise a SNN based clustering algorithm to discover frequent semantic regions from multiple trajectories. Finally, to show the preciseness and effectiveness of our framework, we present comprehensive experimental results over various real datasets. The results demonstrate that our framework is able to accurately extract semantic regions.

In the future, we intend to investigate the user activities on the semantic regions and mine the relations between them. Moreover, we would like to build the common user behaviour on frequent semantic regions. Developing novel applications, such as personalized recommendation based on user behaviour, is also a task we aim to accomplish. We consider these as promising future works.

Acknowledgments. Wen-Chih Peng was supported in part by ITRI-NCTU JRC program, Project No. 99EC17A05010626.

References

1. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. In: SIGMOD, pp. 49–60 (1999)
2. Cao, H., Mamoulis, N., Cheung, D.W.: Mining Frequent Spatio-Temporal Sequential Patterns. In: ICDM, pp. 82–89 (2005)
3. Ertoz, L., Steinbach, M., Kumar, V.: A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In: 2nd SIAM International Conference on Data Mining (2002)
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: KDD, pp. 226–231 (1996)
5. Everytrail – gps travel community, `http://www.everytrail.com`
6. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory Pattern Mining. In: KDD, pp. 330–339 (2007)
7. Hung, C.C., Chang, C.W., Peng, W.C.: Mining Trajectory Profiles for Discovering User Communities. In: GIS-LBSN, pp. 1–8 (2009)
8. Hung C.C., Peng, W.C.: Clustering Object Moving Patterns for Prediction-Based Object Tracking Sensor Networks. In: CIKM, pp. 1633–1636 (2009)
9. Jarvis, R.A., Patrick, E.A.: Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Trans. Comput., 22(11), 1025–1034 (1973)
10. Jeung, H., Liu, Q., Shen, H.T., Zhou, X.: A Hybrid Prediction Model for Moving Objects In: ICDE, pp. 70–79 (2008)
11. Jeung, H., Shen, H.T., Zhou, X.: Mining Trajectory Patterns Using Hidden Markov Models. In: DaWaK, pp. 470–480 (2007)
12. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining User Similarity Based on Location History. In: GIS (2008)
13. Liao, L., Fox, D., Kautz, H.A.: Location-Based Activity Recognition. In NIPS (2005)
14. Liao, L., Fox, D., Kautz, H.A.: Location-Based Activity Recognition using Relational Markov Networks. IJCAI, pp. 773–778 (2005)
15. Lin, C.R., Chen, M.S.: On the Optimal Clustering of Sequential Data. In: SDM (2002)
16. Lo, C.H., Peng, W.C., Chen, C.W., Lin, T.Y., Lin, C.S.: CarWeb: A Traffic Data Collection Platform. In: MDM, pp. 221–222 (2008)
17. Mamoulis, N., Cao, H., Kollios, G., Hadjieleftheriou, M., Tao, Y., Cheung, D.W.: Mining, Indexing, and Querying Historical Spatiotemporal Data. In: KDD, pp. 236–245 (2004)
18. Monreale, A., Pinelli, F., Trasarti, R., Giannotti. F.: WhereNext: a Location Predictor on Trajectory Pattern Mining. In: KDD, pp. 637–646 (2009)
19. Yang, J., Hu, M.: TrajPattern: Mining Sequential Patterns from Imprecise Trajectories of Mobile Objects. In: EDBT, pp. 664–681 (2006)
20. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining Interesting Locations and Travel Sequences From GPS Trajectories. In: WWW, pp. 791–800 (2009)
21. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Collaborative Location and Activity Recommendations with GPS History Data. In: WWW, pp. 26–30 (2010)