A Method for Detecting Significant Places from GPS Trajectory Data

Chuyen Luong, Son Do, Thang Hoang, and Deokjai
Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea
Email: luongchuyen0789@gmail.com, dchoi@jnu.ac.kr

***Abstract*—Detecting significant places are necessary for learning patterns of human behavior. Moreover, the Global Positioning System is the high accurate estimation of positioning method for mobile tracking. In this paper, we propose a method for detecting significant places based on GPS data. It is difficult to determine significant places relying on the clusters from using distance and time thresholds because of the difference of noises. Therefore, we introduce a method based on EIDBSCAN algorithm to detect arbitrary shape clusters with different densities, namely MKEIDBSCAN. We also propose the way to estimate input parameters of density-based clustering algorithms including the radius and the minimum number of points. As a result of using our proposal, significant places are detected more accurately and the running time is reduced.**

***Index Terms*—significant places, density-based clustering algorithm, GPS-collected points**

I. Introduction

Nowadays, many of popular mobile applications and services are using the exact location or location information such as ‘Glympse’ application for sharing your location, ‘Locale’ application for your changed behavior in the different locations, ‘Where’ for location recommendation, ‘NeverLate’ for changing time to go to the other location and so on. Furthermore, smartphones that embed many sensors allow identifying the user’s coordinates based on GPS, Wi-Fi with high accurate estimation of positioning method.

Conceptually, a significant place is a region where moving objects pause or wait or slow in order to complete important activities such as home, places of work, restaurants, shops, etc. GPS enabled device can record user’s location as time stamped sequences of latitude and longitude coordinates. Therefore, the significant places are formed from GPS points.

There are many methods to detect the significant places based on stay points or density of GPS points. Stay points are detected in [1]-[5] by extracting consecutive GPS points from trajectories with a satisfaction of stay time and distance thresholds. These stay points are then grouped into Region of Interest (RoI) using density-based clustering. The focus of these methods is to detect places where the user stays during a certain time period. Normally, human has a tendency to stay at significant places. Therefore, the density of GPS points in these regions is denser than other regions. A method based on density is revealed in [6]. Data space is divided into grid and RoIs are groups of nearby dense regions without considering stay time.

Our contributions: In this paper, we propose the method of detecting the significant places using density-based clustering algorithm, namely MKEIDBSCAN algorithm. First of all, we present an effective way to estimate two input parameters in EIDBSCAN including the fixed radius value ε and the varied number of point MinPts in a circle with that radius value, using the proposed MinPts-value method. Moreover, we use K-medoid algorithm to specify elite points. These elite points are also using as initial points for formed new clusters in MKEIDBSCAN. MKEIDBSCAN algorithm is based on EIDBSCAN with multi density to discover the clusters. This algorithm uses seeds which are the closest points to eight Marked Boundary Objects like IDBSCAN algorithm. One circle is divided into 8 regions. Seeds in the three consecutive regions are considered for keeping or deleting. Thus, the processing time will be reduced. To determine the significant places, we extract cluster information. If this clusters’ information satisfies to comparison thresholds, they are considered as significant places.

The rest of this paper is organized as follows. We start with a discussion of related work in Section II. Some basic concepts are defined in section III. Details of proposed algorithm for estimating the input parameters and detecting the significant places are presented in section IV. The experimental results and comparisons are described in section V. Finally, section VI presents a conclusion.

II. Related Work

A method for detecting the significant place is shown in papers [1]-[4]. They use Geolife dataset. Time and distance thresholds are considered as two scale parameters. If a consecutive GPS point sequence satisfies two thresholds, a stay point is formed. There are two situations for bad detections. They are the loosed satellite signal case and user’s moving out of certain geospatial range for a period case. And then a density-based clustering algorithm, OPTICS, applied to compute RoIs from stay points. The RoI construction algorithm is

Manuscript received December 11, 2014; revised April 14, 2015.

doi: 10.12720/jait.6.1.44-48

applied by considering the density value which is the number of GPS trajectories into a grid in the paper [6]. RoI is a combination of nearby dense cells on condition that the average density of the region is higher than a threshold. Paper [5] shows two drawbacks of this algorithm. There are RoI determinations without considering the stay time and the density estimation including moving points in the cell. After detecting stay points, the authors of paper [5] use Local Outlier Factor to extend them into RoI. They also remove a certain percentage of stay points with the largest LOF values to get regions clearly.

In general, in a significant place, people either tend to move slowly or don't move. Thus, the density distribution in these places is much denser than others. Otherwise, these locations' sharp is different. Those are reasons why multi density clustering algorithms are considered in the papers [7]-[9]. There are two problems including the estimation of two input parameters and the processing time for clustering based on multi density method. For estimating two input parameters, papers [8], [9] use k-dist method. Firstly, they calculate the average distance from a point to K-closest neighbors. Next, the average distances are sorted in the ascending order. Then they plot them and choose changed sharps. The average distance values at changed points are considered input radius matrices. Finally, they will calculate the average number of points in a circle with any point center and a radius in radius matrices respectively. For reducing the running time, IDBSCAN [10], KIDBSCAN [11] and EIDBSCAN [12] are revealed. Most of them consider a seed which is the closest distance to eight MBOs to expand the cluster. KIDBSCAN use K-mean algorithm to find elite points for initializing new clusters. EIDBSCAN divide a circle into 8 regions and they consider seeds in the three consecutive regions. If these considered seeds exist, the seeds in the second region will be deleted.

III. Basic Concepts

Basic concepts are defined as follows:

- The neighborhood within a radius of a circle ε of a given object is considered as the ε-neighborhood of the object.
- The object is a core point, if the ε-neighborhood of an object contains at least the minimum number of point within the circle ε-radius, MinPts object.
- The points in the ε-neighborhood of a core point are considered as border points.
- A noise point is any point without a core point or a border point.
- An object p is directly density-reachable from object q if p is within the ε-neighborhood of q and q is a core object.
- An object p is density-reachable from object q, if there is a chain of object where and such that is directly density reachable from.
- An object p is density-connected to q with respect to ε and MinPts if there is an object o such that both p and q are density reachable from o.
- A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. This is illustrated in Fig. 1.
- MBOs: the eight distinct points are considered as marked boundary objects. Assuming that the core point is P (0, 0), the eight marked objects may be defined as in Fig. 2.
- The closest points from MBOs in the ε-radius circle are considered as seeds.

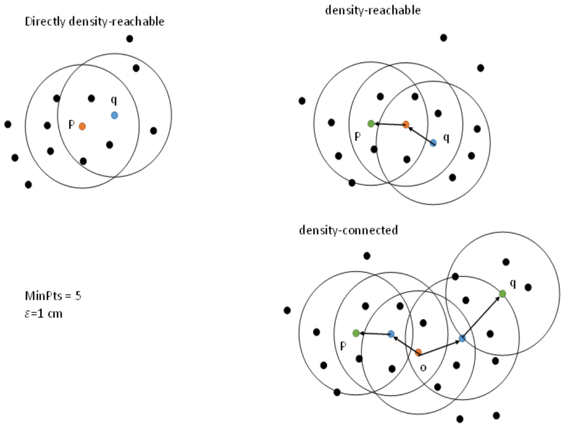


Figure 1. Basic concepts in density-based clustering algorithm

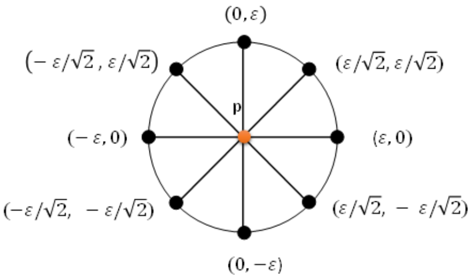

$(0, \varepsilon)$
$(-\varepsilon/\sqrt{2}, \varepsilon/\sqrt{2})$
$(\varepsilon/\sqrt{2}, \varepsilon/\sqrt{2})$
p
$(-\varepsilon, 0)$
$(\varepsilon, 0)$
$(-\varepsilon/\sqrt{2}, -\varepsilon/\sqrt{2})$
$(\varepsilon/\sqrt{2}, -\varepsilon/\sqrt{2})$
$(0, -\varepsilon)$

Figure 2. The eight marked boundary objects

- The closest points from MBOs in the ε-radius circle are considered as seeds.

IV. Methodology

A. Parameter Estimation

The ε-radius and MinPts values are estimated.

1) Estimate the ε-radius value

The density is the number of points within a region. Some specific cases have the same density with different couple of the radius value and MinPts. To consider the different densities, we fix the radius value and change the MinPts value. We consider the bandwidth in the density estimation method as the diameter of circle. In the paper [13], [14] show that the density of human movement in the urban is Gaussian distribution. Thus, the optimal bandwidth value in [15] is assigned to the ε value and calculated by equation:

$$\varepsilon = \left(\frac{2\left(\sigma_x + \sigma_y\right)^{\left(\frac{5}{2}\right)}}{3n} \right)^{\frac{1}{5}} \quad (1)$$

where n is the number of GPS records of a user and σ_x, σ_y are the standard deviations of the whole GPS sequence in two dimensions, respectively.

2) Estimate MinPts

To estimate MinPts, we calculate the number of points within a ε-radius and a center point of the elite points. Then, the changes of densities are detected. The method is illustrated by following the below steps:

Step 1: Using K-medoid to find the coordinate of K elite points (K-center points) in input data.

Step 2: Calculate the number of points (Minpts-value) in a circle with fixed radius and K elite points.

Step 3: plot sorted Minpt-values in ascending order.

Step 4: Finding sharp change corresponding with suitable value of Minpts for each density level.

For example, Fig. 3 shows the ascending order number of points MinPts-value in a circle with ε radius and elite point coordinate. The point A scans all MinPts- value points to find shape changes considered as MinPts. There are three relatively smooth lines which describe three density levels. Line f shows the densest density, line b shows the sparsest density and line a shows the MinPts-value of outlines. Take line a and b as a sub-MinPts-value plot to select MinPts1, line c and d as a sub-MinPts-value plot to select MinPts2, line e and f as a sub-MinPts-value plot to select MinPts3.

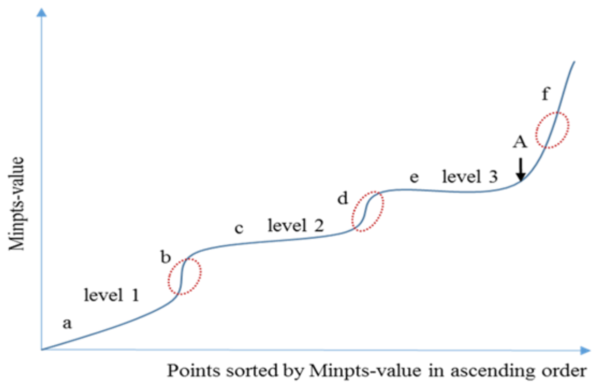


Figure 3. Minpts-value plot of sample dataset

B. MKEIDBSCAN Algorithm

The clustering method helps for detecting the significant places. In the reality, the significant places have arbitrary sharps. If we consider the GPS point densities within a region, they are different densities depending on the occurrence frequency and time stay. For this reason, GPS point cluster method with multi densities is brought forward. Moreover, because the number GPS points are large, the processing time should be necessarily considered. MKEIDBSCAN algorithm is proposed to solve these problems. Firstly, this algorithm detects initial points via using elite points which are center points in K-medoid algorithm. Moreover, these elites are used to estimate input parameters illustrating in section IV. Secondly, seeds also identify to expand the regions with core points of elite points. It is similar with EIDBSCAN algorithm for reducing the number of expansion seeds if three consecutive regions are existed. Therefore, the running time is reduced significantly.

MKEIDBSCAN is implemented by steps as follows:

Step 1: Discovery K elite points based on apply K-mediod algorithm. K value is set by a simple rule of thumb.

Step 2: Determine the input parameter relied on equation 1 and above MinPts-value method.

Step 3: Sort MinPts in descending order and implement cluster with each MinPts and radius ε in succession.

Step 4: Cluster GPS point by applying EIDBSCAN, but the initial points for forming new cluster is the K elite point.

C. Significant Place Decision

This is final step to determine clusters into significant places. This is illustrated in the Fig. 4. All of points in a trajectory are scanned. A cluster is a significant place when three factors including average velocity stay time and cluster radius are satisfied simultaneously. We extract the movement information of the user in each cluster. If an average velocity of the consecutive points in a cluster in a trajectory is less than the velocity threshold, a stay time is longer than the stay time threshold and a cluster radius smaller than the radius threshold, the cluster is considered as the significant place.

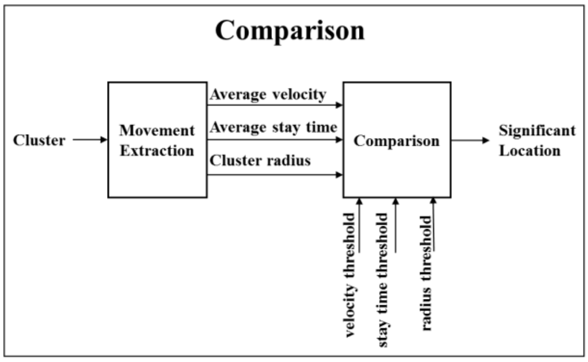


Figure 4. A comparison to determine the significant place

V. EXPERIMENT

A. Geolife Data Descriptions

Geolife dataset [16] collected by 182 users from April 2007 to August 2012. Dataset is collected in over 30 cities of China, some cites of USA, and Europem. But, the majority of dataset was created in Beijing, China. Different GPS collector information is contained in many trajectories. GPS information is record at every 1~5 seconds or every 5~10 meters. Moreover, the user transportation modes were collected such walk, bike, bus and car, etc.

Every user's GPS log file stores many trajectories which named by their starting time. A trajectory is formatted as follows:

Line 1…6 are useless in the dataset, and can be ignored. Points are recorded in following line.

Field 1: Latitude in decimal degrees

Field 2: Longitude in decimal degrees.

Field 3: All set to 0 for this dataset

Field 4: Altitude in feet (-777 if not valid)

Field 5: Date-number of days (with fractional part) that have passed since 12/30/1989

Field 6: Date as a string.

Field 7: Time as a string.

Example:

39.906631, 116.385564, 0, 492, 40097.5864583333, 2009-10-11, 14:04:30

B. *Experiment Results*

For achieving the clustering results, we combine the input parameter estimation and the reducing running time method. We also apply k-dist value and Minpts-value to estimate the input parameters of KIDBSCAN, EIDBSCAN and MKEIDBSCAN algorithms. Table I shows the clustering results using different algorithms. If the input parameters are different, the number of classes is also different. The priority of classes depends on their different density. Therefore, the number of classes of MKIDBSCAN is more than in other algorithms. Because of using k elite points to make new clusters and seeds to expand the cluster, the running time of MKIDBSCAN is reduced significantly.

TABLE I. THE CLUSTERING RESULTS USING DIFFERENT COMBINED ALGORITHM

		Average class number	Running time
k-dist value	KIDBSCAN	18	266.4318
	EIDBSCAN	18.9268	52.1157
	MKEIDBSCAN	19.0732	46.5827
Minpts-value (My proposal)	KIDBSCAN	27.4146	183.0485
	EIDBSCAN	29.0976	36.7736
	MKEIDBSCAN	30.0732	34.34

After clustering, we extract average velocity, stay time and radius of cluster. Specially, stay time values are time intervals of consecutive time sequences. Then, they are compared to thresholds to determine the significant locations. In my paper, I setup these thresholds in Table II.

TABLE II. THE THRESHOLD VALUES FOR COMPARING TO CLUSTER INFORMATION

Threshold name	Threshold value	Note
Velocity	1.5 m/s	Below the average walker's speed
Stay time	20 minutes	The minimal time to stay in a place
Radius	200 m	The radius of a building

In Fig. 5, the intersections, places in which the GPS signal is lost or the user wanders, are detected as stay points and then they are clustered into interesting locations. Using the MKEIDBSCAN algorithm, these incorrect interesting locations will be removed. This is illustrated in Fig. 6.

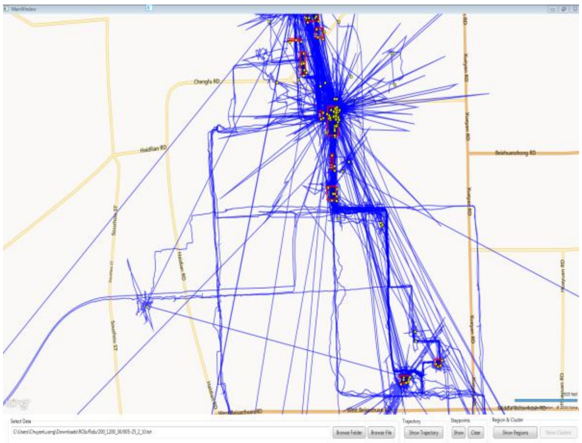


Figure 5. Significant locations of user 005 in the paper [5]

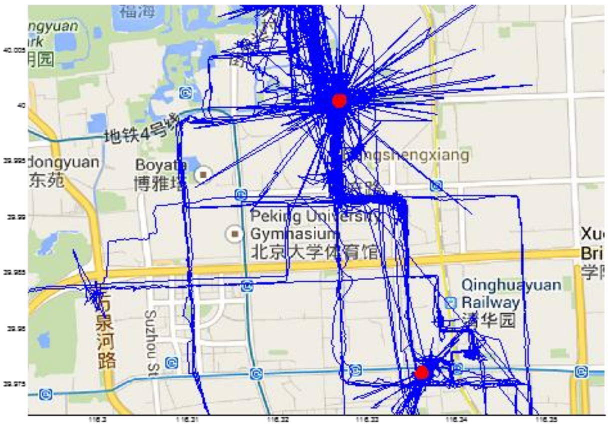


Figure 6. The significant locations of user 005 using KMEIDBSCAN algorithm

VI. CONCLUSIONS

In this paper, we introduce an algorithm to detect the significant places using GPS traces with Geolife dataset. GPS points are grouped into clusters and noise. The input parameters are estimated based on suitable calculated methods. Then, some specific clusters are considered as significant places if they satisfy threshold conditions simultaneously. These conditions include the velocity threshold of 1.5 m/s, stay time threshold of 20 minutes and cluster radius threshold of 200 meters. We found that MKEIDBSCAN is an effective algorithm compared to other algorithms to detect the significant places.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2012R1A1A2007014).

REFERENCES

[1] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Y. Ma, "Mining user similarity based on location history," in *Proc. 16th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, 2008, pp. 34.

[2] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. the 18th International Conference on World Wide Web*, ACM, 2009, pp. 791--800.

[3] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *Proc. of WWW*, 2010, pp. 1029-1038.

[4] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Y. Ma, "Recommending friends and locations based on individual location history," in *ACM TWEB*, vol. 5, no. 1, 2011.

[5] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles for location-based services," in *Proc. the 28th Annual ACM Symposium on Applied Computing*, March 18-22, Coimbra, Portugal, 2013.

[6] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. SIGKDD*, ACM Press, 2007, pp. 330–339.

[7] 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. 2nd int. Conf. on Knowledge Discovery and Data Mining*, Portland, Oregon, AAAI Press, 1996.

[8] M. T. H. Elbatta and W. M. Ashour, "A dynamic method for discovering density varied clusters," in *Proc. Int. J. of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 1, February, 2013.

[9] M. N. Gaonkar and K. Sawant, "AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset," in *Proc. Int. J. on Advanced Theory and Engineering*, vol. 2, no. 2, 2013.

[10] B. Borah and D. K. Bhattacharyya, "An Improved sampling-based DBSCAN for large spatial databases," in *Proc. International Conference on Intelligent Sensing and Information*, 2004, pp. 92–96.

[11] C. F. Tsai and C. W. Liu, "KIDBSCAN: A new efficient data clustering algorithm," *Lecture Notes in Artificial Intelligence*, vol. 4029, pp. 702-711, 2006.

[12] C. F. Tsai and C. W. Liu, "EIDBSCAN: An extended improving DBSCAN algorithm with sampling techniques," in *Proc. Int. J. of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 94–111, 2010.

[13] X. H. Chen and J. Pang, "Protecting query privacy in location-based services," *Geoinformatica*, vol. 18, no. 1, pp. 95-133, January 2014.

[14] A. Noulas, S. Scellato, R, Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: Universal patterns in human urban mobility," in *PloS one, Public Library of Science*, vol. 7, 2012.

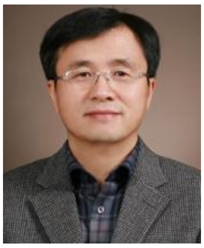
[15] D. Guo, H. Zhou, Y. Zou, W. Yin, H. Yu, Y. Si, J. Li, Y. Zhou, X. Zhou, and R. J. S. Magalhaes, "Geographical analysis of the distribution and spread of human rabies in china from 2005 to 2011," *PLoS ONE 2013*, vol. 8, no. 8.

[16] GeoLife GPS Trajectories. [Online]. Available: http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

Chuyen Luong, she received Engineering degree in School of Electronics and Telecommunications from Hanoi University of Sciences and Technology, Vietnam in 2012. She got MS Degree in School of Electronics and Computer Engineering, Chonnam National University, South Korea in 2014. Her research interests are mainly in the field of context awareness, pattern recognition.

Son Do, he received BS degree in Department of Math and Computer Science, University of Science, VNU-HCMC in 2011. He got MS Degree in School of Electronics and Computer Engineering, Chonnam National University, South Korea in 2014. His research interests are context awareness, and pattern recognition.

Hyukro Park, he received BS degree in Department of Computer Science, University of Science, VNU-HCMC in 2010. He got MS Degree in School of Electronics and Computer Engineering, Chonnam National University, South Korea in 2014. His research interests are context awareness, and pattern recognition.



Deokjai Choi, he is full professor of Computer Engineering Department at Chonnam National University, South of Korea. He received BS degree in Department of Computer Science, Seoul National University, in 1982. He got MS degree in Department of Computer Science, KAIST, South Korea in 1984. He got PhD degree in Department of Computer Science and Telecommunications, University of Missouri-Kansas City, USA in 1995. His interest on research spans from context awareness, pervasive computing, sensor network, future Internet and IPv6.