A Spatio-Temporal Distance Based Clustering Approach for Discovering Significant Places from Trajectory Data

Sungjun Lee, Yerim Choi, Seongmin Lim, and Jonghun Park

***Abstract*—**With the growing number of trajectory data produced by mobile devices, discovering the frequently visited places of a user, *significant places*, is crucial for supporting location-based services and applications. Most existing works have focused on the spatial distance between trajectory points. However, temporal aspects of trajectories such as moving speed and visit periodicity are important information as well. In this paper, we propose a spatio-temporal distance based clustering approach for discovering significant place from trajectory data. Our approach clusters multiple points that indicate the same place into a significant place while being robust to GPS sensing errors. We evaluated the proposed approach with real-world data, and experimental results show that it outperforms baselines.

***Keywords*—**Trajectory data, significant place discovery, spatio-temporal distance based clustering

I. Introduction

In the last few years, large amount of trajectory data is produced with the rapid proliferation of mobile devices. There has also been an explosion of location-based services and applications based on the information of the places which users frequently visit and stay, *significant places*. The trajectory data produced by mobile devices is in the form of trajectory which is composed of points, one of which represents a position in space in a certain instant of time. Therefore, to close the gap between trajectory data and significant places, discovering the significant places from the trajectory data is important.

Discovering significant places from trajectory data such as GPS is a well-explored area in the literature [1], [4], [7]. For instance, [8] extended the density-based clustering algorithm, DBSCAN [2], to DJ-Cluster and used the extended algorithm to discover significant places from GPS points. A grid-based clustering algorithm for detecting places-of-interests from GPS while dealing with missing trajectory data was proposed in [3]. [6] Suggested a clustering-based approach for discovering interesting places by separating trajectories as a set of stops and moves. Unfortunately, there have been little consideration to temporal dimension of trajectory points.

Sungjun Lee, Yerim Choi, Seongmin Lim, and Jonghun Park are with the Department of Industrial Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, Korea (corresponding author: Seongmin Lim; phone: +82-10-3174-3579; e-mail: hovern@snu.ac.kr).

To this end, in this paper we propose a spatio-temporal distance based clustering approach for discovering significant places from GPS trajectory data. Our approach has three stages as follows. In the first stage, we extract *stay* where users stayed certain amount of time from GPS trajectory data by a clustering algorithm which is an extension DBSCAN. Not only the distance but also the moving speed between two points are considered in this algorithm. In the second stage, the stays are clustered into *elementary places* using Chebyshev′s inequality-based method. In the final stage, we calculate the spatio-temporal distance between two elementary places to discover significant places. If the distance between two elementary places is shorter than a predefined value, these places are merged into a single place, and the clustered places are discovered as significant places after the three.

The rest of the paper is organized as follows. In the next section we present an algorithm for extracting stays from GPS trajectory data. In Sections Ⅲ and Ⅳ, methods for clustering stays into elementary places and the spatio-temporal distance measure for merging elementary places into significant places are proposed. In Section Ⅴ, we show the experimental results on real-world data. Finally, we conclude this paper in Section Ⅵ.

II. Extracting Set of Stays

The input to our problem is a sequence of N trajectory point consists of GPS, longitude and latitude, along with a timestamp. The input trajectory is in the form of $P=\{p_1, p_2, ..., p_N\}$, where each point p_i is denoted by $p_i=(x_i, y_i, t_i)$, composed of GPS point (x_i, y_i) and corresponding timestamps $t_i, i=1,2,3,...,N$. Among these trajectory points, there are some points generated when a user is moving as well as some points when the user is staying. Therefore, in this section, we propose an algorithm to detect a set of stays, $S=\{s_1, s_2, ..., s_M\}$, where M is the number of stay, $s_i \subseteq P$ and $s_1 \cap s_2 \cap ... \cap s_M = \phi$.

The intuition of our algorithm is that if there are some points in a trajectory, where speed of users moving between each other

points is slow, the points are likely to be parts of a significant place. The speed of users might be slower while visiting a significant place than while moving from a significant place to another. From this consideration, we present a moving speed based clustering algorithm.

[6] Proposed an algorithm considering spatial distance between two points, called CB-SMoT. To consider moving speed between points, we have extended CB-SMoT by revising Eps-linear-neighborhood to include moving speed parameter. We define the revised Eps-linear-neighborhood (RELN) of a point p_k as the set of points before and after p_k in the trajectory whose spatio-temporal distance from p_k is less or equal to Eps. Eps is a positive number that represents the maximum spatio-temporal distance between a point p_i and its RELN in the trajectory. Following this, the RELN of a point p_k is the maximal set of points p_i such that:

$$\sum_{i=m}^{k-1}\left(\rho_{spd}(p_i,p_{i+1})\cdot dist(p_i,p_{i+1})\right)\le Eps$$
$$\cup\ \sum_{i=k+1}^{n}\left(\rho_{spd}(p_{i-1},p_i)\cdot dist(p_{i-1},p_i)\right)\le Eps \qquad (1)$$
$$\text{where } t_1 \le t_m < t_k < t_n \le t_N$$

where $dist(p_i,p_{i+1})$ is the Euclidean distance between p_i and p_{i+1}. In the RELN, speed parameter ρ_{spd} is additionally considered, which is calculated as follows.

$$\rho_{spd}(p_i,p_{i+1}) = \frac{dist(p_i,p_{i+1})}{t_{i+1}-t_i} \qquad (2)$$

Using RELN concept, we propose the DBSCAN based stay detection procedure, which is shown in Algorithm 1. In the line 1 to 9, neighbors of a point from trajectories are grouped when the two constraints in the line 5 are satisfied. To consider temporal distance between the point and the neighbors, we added the *minTime* constraint. The *SeqTime* of the neighbors is defined as follows.

$$SeqTime(NB) = \sum_{\forall p_i \in NB}\left(t_j - t_i\right) \qquad (3)$$
where p_j is the immediate succesor of p_i

In the line 10 to 24, same procedure of above is applied to the neighbors of each neighbor point. As a result, the set of stays consists of the neighbor point clusters who satisfy the two conditions is extracted from the trajectories.

Algorithm 1 Stay detection algorithm

Input: *P*, *eps*, *minTime*, *minPts*
Output: *S*

1: $S \leftarrow \phi$
2: **for each** unvisited location point p in trajectory P **do**
3: mark p as VISITED
4: $NB \leftarrow$ getRELN(p, *Eps*)
5: **if** $|NB| <$ *minPts* **and** seqTime(NB) < *minTime* **then**
6: mark p as NOISE
7: **else**
8: $C \leftarrow$ a new cluster; $C.id \leftarrow$ a new id
9: add point p to cluster C
10: **for each** neighbor $p' \in NB$ **do**
11: **if** p' is not visited **then**
12: mark p' as VISITED
13: $NB' \leftarrow$ getRELN(p, *Eps*)
14: **if** $|NB'| >=$ *minPts* **and** seqTime(NB') >= *minTime*
15: **then**
16: $NB \leftarrow NB \cup NB'$
17: **end if**
18: **end if**
19: **if** p' is not yet member of any cluster **then**
20: add p' to cluster C
21: **else**
22: $C.id$ = the id of cluster holding p'
23: **end if**
24: **end for**
25: $S[C.id] \leftarrow C$
26: **end if**
27: **end for**

III. Detecting Elementary Places

In the previous section, we extract a set of stays from trajectory points. Due to the GPS sensing error, there is a possibility that two different stays are actually parts of the same significant place of a user. Fig. 1 shows an example of these stays not clustered into a single stay as expected. p_8 is generated because of GPS sensing error, and as the spatio-temporal distances between p_8 and other points are longer than the given Eps, s_1 and s_2 are not clustered into a single stay.

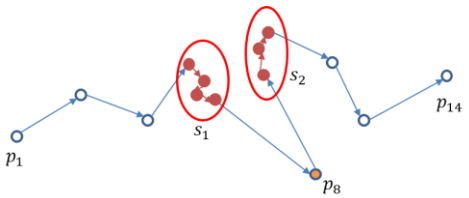

Fig. 1 An example of incorrectly extracted stays due to a noise point, p_8, caused by GPS error

To solve this problem, we present a method that merges stays into a set of elementary places $E = \{e_1, e_2, ..., e_K\}$, where K is the number of elementary place, $e_i \subseteq S$ and $e_1 \cap e_2 \cap ... \cap e_K = \phi$. The basic idea of the method is as follows. If there exists a point p_i in s_a or s_b which is possibly a part of both stays, s_a and s_b can be merged into a single place. Based on Chebyshev's inequality, we propose a merging constraint of stays which is defined as follows.

$$(p_i - \mu_a)^T V_a^{-1} (p_i - \mu_a) < k^2 \cap (p_i - \mu_b)^T V_b^{-1} (p_i - \mu_b) < k^2 \quad (4)$$

where μ_i is the mean of trajectory points vector in s_i, V is the covariance matrix. Following the proposed method, if there exists a point p_i in s_a or s_b that satisfies (4), s_a and s_b are merged to compose a single elementary place. We proposed this method to deal with the incorrect trajectory points generated when GPS sensing error occurs.

IV. Discovering Significant Place

From trajectory data, we extracted a set of stays and merged the stays into elementary places to resolve GPS sensing error. At last, multiple elementary places are to be grouped to discover a significant place with a single semantic. For example, two different gates of a big building could be extracted as two different elementary places, as GPS sensor doesn't work inside buildings. In addition, if a user visit a place periodically but not sequentially, this place could be discovered as two different elementary places due to the temporal distance between each visit.

To this end, in this section, we propose a method for discovering a set of significant places, $L = \{l_1, l_2, ..., l_H\}$, where H is the number of significant place, $l_i \subseteq E$ and $l_1 \cap l_2 \cap ... \cap l_H = \phi$, with considering the spatio-temporal distance which is not handled in the previous stages. In the first step, we define the spatial distance $d_\sigma(e_a, e_b)$ between two elementary places e_a and e_b as follows, based on the Haversine formula, which is an equation for calculating the shortest distance between the two points on a sphere from their longitudes and latitudes.

$$d_\sigma(e_a, e_b) = \Sigma_{p_a \in e_a, p_b \in e_b} 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_{p_b} - \phi_{p_a}}{2}\right) + \cos(\phi_{p_a}) \cdot \cos(\phi_{p_b}) \sqrt{\sin^2\left(\frac{\lambda_{p_b} - \lambda_{p_a}}{2}\right)}}\right) \quad (5)$$

where p_a is the points in e_a, p_b is the points in e_b, ϕ_{p_i} is a latitude of point p_i, λ_{p_i} is a longitude of point p_i, r is the radius of the sphere, earth, .

In the second step, we define the temporal distance $d_\tau(e_a, e_b)$ between two elementary places e_a and e_b as follows.

$$d_\tau(e_a, e_b) = \Sigma_{p_a \in e_a, p_b \in e_b} \left(w_{p_a p_b} \cdot \theta(p_a, p_b)\right) / \left(\Sigma_{p_a \in e_a, p_b \in e_b} w_{p_a p_b}\right) \quad (6)$$

where $\theta(p_a, p_b)$ is the time difference between points p_a and p_b. To consider the periodical visit of the place, we define $w_{p_a p_b} = e^{-r|day(p_b) - day(p_a)|}$, where $day(p_i)$ takes the value from 0 to 6 consecutively, when the day that point p_i was collected is from Monday to Sunday. Taking both distance together, we define the spatio-temporal distance as follows.

$$d(e_a, e_b) = d_\sigma(e_a, e_b) + \beta \cdot d_\tau(e_a, e_b) \quad (7)$$

where β is weighting parameter. If there exists two elementary places e_a and e_b where $d(e_a, e_b)$ is lesser than a predefined value, then e_a and e_b are considered as a single significant place.

Fig. 2 and Fig. 3 illustrate the concept of a significant place discovery. In this example, elementary places, C8 and C9 in Fig. 2, have become a single significant place, C48 in Fig. 3, after the merging procedure is performed. From the example, it can be shown that the proposed method successfully discovered the significant places, since there are two different gates of the building.

.

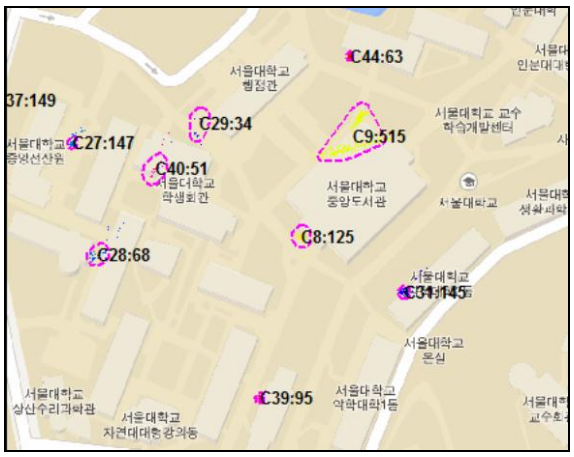

Fig. 2 Elementary places before merging

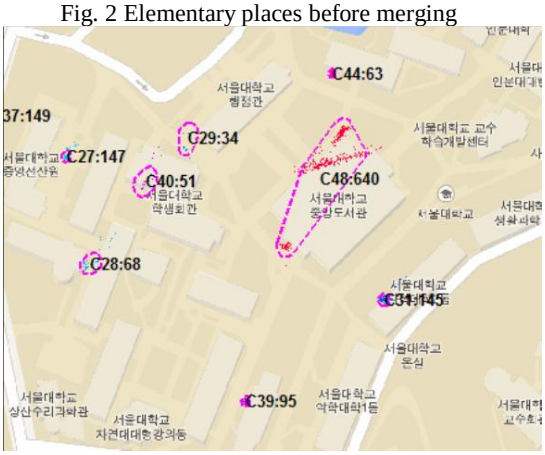

Fig. 3 Significant places after merging

V. Experiments and Evaluation

We have implemented an application for data collection, and using the application, collected trajectory data of 46 college volunteers. At the sampling rate of two minutes, a total of 996,467 trajectory points were acquired during four consecutive weeks. Because of the weak network connection or discharged state of mobile devices, the collected data is not totally sequential.

To evaluate the effectiveness of the proposed framework, we manually labeled significant places based on Google Maps since collecting ground truth from users is almost impossible task. For the trajectory points generated within campus, we labeled significant places in a detailed fashion, e.g. "Library" or "Cafeteria". For the points outside campus, we labeled the significant places in a relatively abstract manner, e.g. "Nokdu Street" or "Ginko Avenue".

To evaluate the proposed significant place discovery method, we used an evaluation system partially based on the one presented in [3]. Significant places discovered by using the proposed approach and by manual labeling process are called *Found and Labeled, respectively*. Places found and labeled are called *Correct*. Places found but not labeled are called *Discovered*, while places labeled but not found are called *Missed*.

We compare the proposed approach with methods proposed in [5] and [6], which are ones for discovering significant places from trajectory data. To evaluate performances, a set of measures, P, R, and F, where P, R and F denote Precision, Recall, and F-measure, respectively, have been defined as follows (# stands for 'the number of').

$$P = \frac{\#Correct}{\#Discovered}$$
$$R = \frac{\#Correct}{\#Labeled} \quad (8)$$
$$F = 2 \cdot \frac{P \cdot R}{(P + R)}$$

With the above definitions, Table I shows the performance comparison results obtained from the three approaches. The proposed approach showed better results in terms of both Precision and Recall. Moreover, the proposed approach finds more discovered places than the others do, meaning that if there exists some places that have not been labeled but are significant to users, the proposed approach has more potential to find these places than the others methods. From this table, we can see that the proposed approach outperforms [5] and [6] with respect to the effectiveness of discovering significant places.

TABLE I
Results Obtained for all Users and all Days

	Fou.	*Lab.*	*Cor.*	*Disc.*	*Mis.*	*P*	*R*	*F*
[5]	119	107	64	27	32	0.538	0.598	0.566
[6]	160	107	70	29	15	0.438	0.654	0.524
Proposed	156	107	92	47	15	0.590	0.860	0.700

VI. Conclusion

In this paper, we proposed a spatio-temporal distance based clustering approach for discovering significant places from trajectory data. First, we extracted *stay* where user used to stay and cluster them from GPS trajectories by considering moving speed and spatial distance between points. Then, to deal with the GPS sensing error, we mined elementary places from the set of stays based on the probability concept. Finally, we discovered the significant places from the elementary places by leveraging the spatio-temporal distance between elementary places. We have evaluated the proposed approach with real-world data, and performance comparison results showed that the proposed approach outperforms two baselines.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (No. 2013R1A2A2A03013947)..

REFERENCES

[1] D. Ashbrook and T. Starner, "Learning Significant Locations and Predicting User Movement with GPS", In Proc. of the 6th ISWC, 2002.

[2] M. Ester, H. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise", In Proc. of the SIGKDD, pp. 226-231, 1996.

[3] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin, "Discovering semantically meaningful places from pervasive rf-beacons", In Proc. of the Ubicomp, 2009.

[4] L. Liao, D. Fox, and H. Kautz, "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields", IJRR, pp. 119–134, 2007.

[5] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data", In Proc. of the 3rd ICMU, 2010.

[6] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories", In Proc. of the SAC'08, pp. 863-868, 2008.

[7] G. Yang, "Discovering Significant Places from Mobile Phones - A Mass Market Solution", In Proc. of the MELT, pp. 34–49, 2009.

[8] Z. Zhou, D. Frankowsk, P. Ludford, S. Shekhar, and L. Terveen, "Discovering Personal Gazetters: An Interactive Clustering Approach", In Proc. of the 12th ACMGIS, pp. 266-273, 2004.