D-010

User's Next Place Prediction based on User's Attributes and Significant Points Extracted from Trajectories

Xiasi Liu† Mutsumi Suganuma‡ Wataru Kameyama‡

1. Introduction

Location based services and many other proactive services that take advantage of mobile devices are highly demanded nowadays.

The discovery of significant places for users is one of the important issues in this field. [1] proposes a method based on spatial density of all GPS records in trajectories, and [2] applies spatial density based method by taking into account the temporal information of all records. [3] proposes to preprocess the data to find candidates for significant places, as prior methods often result in pointing out meaningless cluster of location. Moreover, another important issue is to predict where people go next with contextual data, which may also make a large difference on prediction result but are seldom considered in related researches.

To enhance the performance of next location prediction against the issues, we propose a new framework to find significant places from users' trajectories and a new model of next place prediction. In finding significant places, we take three-stage procedure, where at first segmenting the trajectory, then detecting stay segments by time and distance thresholds, and finally clustering the stay segments while considering the density and reoccurrence of users. For next place prediction, we take into account the context information such as weather and users' moving speeds, in addition to the trajectory information.

2. Experiments

2.1 Dataset

Our experiments have been performed on the Microsoft GeoLife GPS Trajectories Dataset[4]. It contains GPS trajectory data of 182 users' outdoor movements recorded from April 2007 to August 2012 mostly in Beijing. Every record includes the information of latitude, longitude and time stamp.

2.2 Significant Places

Significant places are defined as places where a user stays for a period of time and revisits with a high probability. And our proposed algorithm predicts one of the significant places as the next place to go.

2.2.1 Stay Segment

To find out these places, we segment the trajectories using modified Spatio-Temporal Kernel Window (STKW) statistic[5]. Given that the GeoLife data have a variety of sampling rates, we modify the way to calculate the STKW values by assessing how much time the user spends within a threshold of 25 meters in both directions of every point rather than counting the number of points in this range.

Suppose a segment has n points and their geometry center is GC, point P_i and the duration satisfy the following formulas.

$$Distance(P_i, GC)_{max} < 200\ [m] \quad (1)$$

$$Duration = t_n - t_1 > 5\ [min] \quad (2)$$

Where t_n is the N-th timestamp in the segment starting from 1.

Then, we consider this segment as a stay segment. The threshold is assumed to be enough to distinguish stay segments where people spend their time from moving segments considering people prefer to walk at 1.4 [m/s].

2.2.2 Clustering for Significant Places

Stay segments and the last points of every trajectory are seen as the candidates of significant points. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[6] is applied to group the significant points and detect the outliers. Fig. 1 shows the extracted significant places of user No. 167 in northern Beijing on map as an example with the parameters of min_cluster_size=2 and min_samples=5. In this figure, every colored square represents a cluster. Due to different scale of places, it's difficult to distinguish them by one threshold. However, HDBSCAN performs better by applying hierarchy to find reasonable clusters. Then each cluster is assigned a unique label as the predicting target.

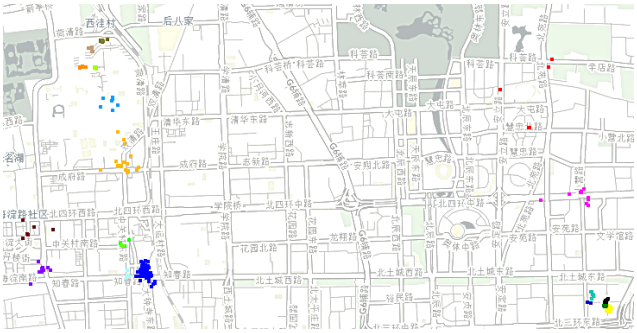

Fig. 1 Clusters of Significant Places of User No. 167 in Northern Beijing

2.3 Features

All features we exploit are shown in Table 1. Five successive points simulate user's current spatial movement in a short period of time. Weather information is retrieved from the online database of National Centers for Environmental Information (NCEI) of United States[7], which is one of the context that we consider might influence on people's activity. It's also assumed that the average speed from the first to the other four points can infer users' transportation mode and direction change.

2.4 Results

2.4.1 Classification

† Graduate School of Fundamental Science & Engineering, Waseda University
‡ Faculty of Science & Engineering, Waseda University

Some places like home, work office, school are frequently visited by users, and this makes the data extremely unbalanced. Therefore, we choose an ensemble method of classification for this problem, which combine multiple weak models to yield a better one with less bias. We pick up 2 algorithms: Gradient Boosted Regression Trees (GBRT) and Random Forest to do the prediction. Fig. 2 and Fig. 3 show the overview of the accuracy of all users by applying GBRT and Random Forest Classifier.

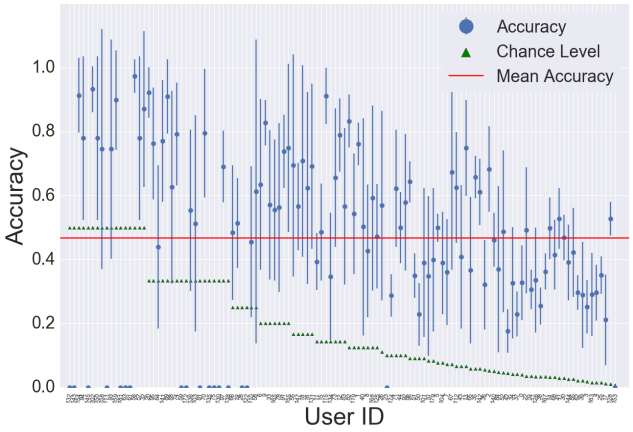

Fig. 2 Accuracy Distribution of Users by GBRT

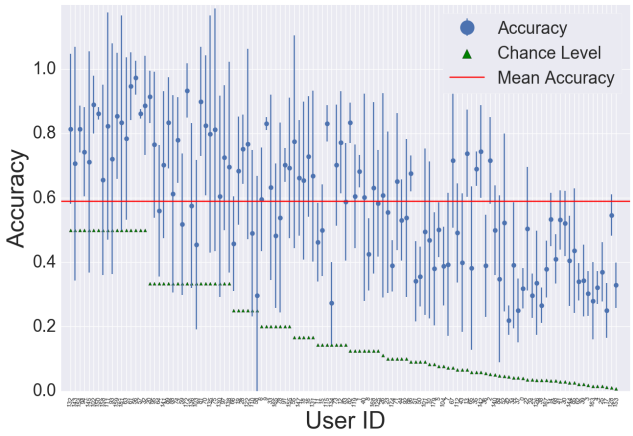

Fig. 3 Accuracy Distribution of Users by Random Forest Classifier

Users are sorted by ascending order of number of significant places. For users with no detected significant places or only have one, the prediction cannot be performed in the proposed algorithm. Therefore, such users are omitted from those figures. Chance level is defined as $1/n$ where n is the number of significant places. The circles and the lines show the mean and the standard deviation of accuracy calculated by conducting 6-fold cross validation on time series, which first split the data into 6 blocks, then use the first block of data as training data and test on the second block, after that use the first and second blocks of data as training data and test on the third block, and so on. The mean accuracy for all users by using GBRT is about 46%, while Random Forest delivers about 60%. However, in both methods, there is a huge variance between users. By looking into the size of data and users' neighborhood, we cannot find strong relations between these factors and the unpredictability.

Table 1 Feature's Description and Frequency in Contribution Ranks

Features	Description	Number of times the feature			
		1st	2nd	3rd	4th
lat1	Latitude of the first point	3	10	11	21
lon1	Longitude of the first point	7	9	56	26
lat2	Latitude of the second point	5	4	13	50
lon2	Longitude of the second point	11	14	17	23
lat3	Latitude of the third point	5	12	9	20
lon3	Longitude of the third point	2	11	24	40
lat4	Latitude of the fourth point	7	14	23	23
lon4	Longitude of the fourth point	8	19	34	37
lat5	Latitude of the fifth point	14	49	28	42
lon5	Longitude of the fifth point	12	25	28	38
windspeed		16	28	37	76
visibility		19	29	42	68
temperature		91	183	85	46
v1	Speed from first to second point	0	0	0	0
v2	Speed from first to third point	0	0	2	0
v3	Speed from first to fourth point	0	0	0	1
v4	Speed from first to fifth point	0	0	0	0
day of week		48	80	137	67
time of day		342	103	44	12

2.4.2 Feature Importance

To evaluate our model, we check the feature importance of Random Forest. Columns at the right in Table 1 show the frequency of each feature whose score of importance ranks as first, second, third and fourth, respectively. It shows that periodicity information embodying day of week and time of day, and weather make the most contribution to prediction in this case.

3. Conclusion and Future Work

Our method of finding significant places not only considers spatial and temporal information of trajectory, but also reoccurrence of users. Our model of predicting user's next place achieves 60% accuracy and reveals that to an extent people's daily routine and weather information plays an important role in this model.

The performance of prediction varies highly between users which is a problem we need to continue our research on to make the model more robust. We hypothesize that better performance can be achieved by grouping users according to certain criterion if there are larger user data.

References

[1] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: An interactive clustering approach", in Proc. ACMGIS, pp. 266–273, 2004
[2] M. Umair, W. S. Kim, B. C. Choi and S. Y. Jung, "Discovering personal places from location traces", 16th ICACT, Pyeongchang, 2014, pp. 709-713
[3] Daniel Ashbrook and Thad Starner, "Using GPS to learn significant locations and predict movement across multiple users", Personal Ubiquitous Comput., 7(5):275–286, 2003
[4] https://www.microsoft.com/en-us/download/details.aspx?id=52367 (last visited on Jun. 27, 2017)
[5] K. Siła-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham, "Analysis of human mobility patterns from GPS trajectories and contextual information", IJGIS, 30:5, pp.881-906 (2016)
[6] L. McInnes, J. Healy, S. Astels, "hdbscan, Hierarchical density based clustering", JOSS, The Open Journal, Vol. 2, No. 11. 2017
[7] NNDC Climate Data Online, NCEI, United States, DS3505, https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd (last visited on May. 7, 2017)
[8] OpenStreetMap contributors, https://www.openstreetmap.org (last visited on Jun. 27, 2017)