Clustering of location sequences

Yujin YAN[1,2], Arnaud KNIPPEL[1], Alexandre PAUCHET[2]

[1] Normandie University, INSA Rouen Normandie, LMI.

[2] Normandie University, INSA Rouen Normandie, LITIS.

{yujin.yan, arnaud.knippel, alexandre.pauchet}@insa-rouen.fr

Keywords : *mobile data, location sequence, routine pattern, clustering.*

1 Introduction

This work analyzes the daily behaviour of users by using their mobile data. With the popular use of mobile phones, an increasing number of large-scale mobile datasets [1, 2] have emerged. Mobile phones are one of the few devices that people carry around in today society.

Several studies [3, 4] have demonstrated that it is possible to mine users' routines from mobile data. Although routines do not mean that humans do exactly the same things as robots every day, daily life can reflect the state and interests of users. There are various ways [3, 4] to define a routine. In this paper, we define the routine to be the pattern of users appearing at certain places at certain times of day. The aim of this article is to extract routine patterns by analyzing the location sequences obtained from users' mobile data. Since the users' location sequence is sparse and we want to obtain the users' activity pattern during a single day, we analyze the users' location sequence for each day of the week. First, the mobile data is pre-processed to obtain the average one-day location sequence of each user. Then, a model is built to obtain the dissimilarity between the one-day location sequences of users. Finally, all one-day location sequences are clustered based on the obtained dissimilarity.

2 Problem Definition

Suppose the dataset contains $N+1$ different location labels. The set of these location labels is $L \cup \{U\}$, $L = \{l_1, l_2, ..., l_N\}$, where $l_i \in L$ denotes a location containing corresponding semantic information and U denotes a location whose semantic information is unknown (it could be any location in the set L or location containing other semantic information). By counting all the location sequence data of the users, we are able to obtain the average one-day location sequence of each user. The one-day location sequence consists of elements from the set $L \cup \{U\}$. The sequence can be modified by adding, deleting or replacing elements.

Funded on the characteristics of the different location labels in the dataset [2], we define the dissimilar properties between the location labels and the dissimilarity properties between two sequences. According to the dissimilarity properties we defined, we choose the Levenshtein distance to calculate the dissimilarity value between two sequences. The score of substituting elements is from zero to q, $C_{sub} \in [0, q]$, because adding or removing the same element after an element of the location sequence indicates that the user has spent more or less time in that point, compared to adding or removing a different element which produces less dissimilar score. The score of adding or deleting an different element after an element in the location sequence is $C_{add} = C_{del} = q$, and the score of adding or deleting an identical element after an element is $C_{add} = C_{del} = q/2$. Then the dissimilarity between two sequences is defined as the minimum score needed to transform one sequence into another by these operations.

Based on the above definition of adding, deleting and substituting operations, we obtain the calculation of the dissimilarity score between a pair of location sequences seq_A and seq_B (with i and j elements, respectively):

$$
S_{A,B}(i,j) = \begin{cases} q * max(i,j) & if \quad min(i,j) = 0 \\ min \begin{cases} S_{A,B}(i-1,j) + C_{add}(A_{i-1}, B_j) \\ S_{A,B}(i,j-1) + C_{del}(A_i, B_{j-1}) \\ S_{A,B}(i-1,j-1) + C_{sub}(A_{i-1}, B_{j-1}) \end{cases} & otherwise \end{cases} \tag{1}
$$

$$
C_{del}(x,y) = C_{add}(x,y) = \begin{cases} q & x \neq y \\ q/2 & x = y \end{cases} \tag{2}
$$

With these choices, the dissimilarity values can be computed by dynamic programming in $O(m,n)$, where m and n are the lengths of the two sequences.

3 Clustering

Based on the dissimilarity values between location sequences, these sequences can be clustered to analyze the characteristics of different classes of sequences. We are interested with two clustering methods: spectral clustering [5] and an exact clustering method using integer linear programming [6]. In the spectral clustering approach, a weighted undirected graph is partitioned according to eigenvectors of the Laplacian matrix of the graph. In the second approach, we solve an integer linear program using the software VIESA [7] and CPLEX.

Acknowledge

The research is supported by China Scholarship Council (CSC).

References

[1] A Pentland, N Eagle, and D Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, 2009.

[2] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. Technical report, 2012.

[3] Tian Qin, Wufan Shangguan, Guojie Song, and Jie Tang. Spatio-temporal routine mining on mobile phone data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(5):1–24, 2018.

[4] Mingqi Lv, Ling Chen, and Gencai Chen. Mining user similarity based on routine activities. *Information Sciences*, 236:17–32, 2013.

[5] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in analysis*, pages 195–200. Princeton University Press, 2015.

[6] Zacharie Ales, Arnaud Knippel, and Alexandre Pauchet. Polyhedral combinatorics of the k-partitioning problem with representative variables. *Discrete Applied Mathematics*, 211:1–14, 2016.

[7] Zacharie Ales, Arnaud Knippel, and Alexandre Pauchet. Viesa project. `https://github.com/ZacharieALES/viesa`.