Transportation Research Part C 96 (2018) 122–143

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Similarity analysis of frequent sequential activity pattern mining

Zhenyu Shou[a], Xuan Di[a,b,*]

[a] Department of Civil Engineering and Engineering Mechanics, Columbia University, United States
[b] Center for Smart Cities, Data Science Institute, Columbia University, United States

ARTICLE INFO

Keywords:
Similarity
Frequent pattern mining
Clustering
Connected vehicles

ABSTRACT

Activity pattern classification is extensively studied using multi-person single-day mobile traces. However, human mobility exhibits intra-personal variability and thus single-day activity may not fully capture one's activity patterns. This paper creates a methodological framework to analyze similarity of activity patterns using frequent sequential pattern mining when multi-person multi-day data is available. Frequent sequential pattern mining discovers the frequently occurring ordered subsequences and is a natural approach of analyzing multi-day travel patterns. PrefixSpan algorithm is implemented to extract frequent patterns for each individual. Then similarity measures are defined to describe the extent to which two travelers' activity patterns are alike and the regularity of how one repeats her activity patterns from day to day. Based upon the pairwise similarity values between two individuals, hierarchical clustering is conducted to divide travelers into communities. To illustrate these methodologies, 349 travelers' 19,130 travel activity sequences are extracted from the world's first connected vehicle testbed in Ann Arbor, Michigan. Three major clusters are identified. Coupled with demographics, these clusters are characterized as "seniors", "working class", and "parents", respectively. Multinomial logistic regression is employed to model to what extent the similarity of socio-demographics can explain that of travel patterns. This work can be extended to either infer an unknown user's demographics (or customer profiling) based on her activity patterns, or to reconstruct an unknown user's frequent activity patterns based on her demographics and other similar travelers' patterns.

1. Introduction

Human mobility is fundamentally stochastic (Castellano et al., 2009) because it exhibits both *inter-personal variability* and *intra-personal variability* (Pas and Koppelman, 1986). Inter-personal variability refers to high heterogeneity across the population, which are primarily dictated by socia-demographics (including gender, age, income, household size), preferences and tastes, spatio-temporal constraints, and social relationships (Cho et al., 2011; Hasan et al., 2016). Intra-personal variability (also called "day-to-day variability") refers to variations in one's activity patterns over time, which is highly influenced by a person's daily life needs, moods, weather, traffic conditions, and other external factors (Pas and Koppelman, 1986). Mobility's uncertainty and unpredictability makes its modeling and prediction challenging. Built on our understanding of intra- and inter-personal variability, one's mobility pattern is usually reconstructed using either her own historical patterns or homogeneous counterparts' patterns.

Human mobility highly depends on historical behaviors (Lu et al., 2013). Regularity or predictability depicts the extent to which one repeats her mobility patterns over time (Goulet-Langlois et al., 2018). To quantify the interplay between regular and random components in intra-personal variability, Song et al. (2010) investigated the fundamental limits that characterize the predictability of

* Corresponding author at: Department of Civil Engineering and Engineering Mechanics, Columbia University, United States.
E-mail address: sharon.di@columbia.edu (X. Di).

https://doi.org/10.1016/j.trc.2018.09.018
Received 4 August 2017; Received in revised form 13 September 2018; Accepted 18 September 2018

human mobility using three main measures related to entropy: (i) random entropy, the simplest entropy concept assuming every event in the time series is performed with equal probability; (ii) temporal-uncorrelated entropy (or entropy in Goulet-Langlois et al. (2018)), which is commonly used to measure the occurrence frequency of isolated events in a time series; and (iii) actual entropy (or entropy rate in Goulet-Langlois et al. (2018)) computes the occurrence frequency of time-ordered subsequence, which captures the sequential information in a sequence. Lu et al. (2013) found that the maximum predictability is not only a fundamental theoretical limit for potential predictive power, but also an approachable target for actual prediction accuracy. Markov chain models were implemented to predict one's next visited locations and produced a prediction accuracy of at least 87%.

One's next visited location can also be reconstructed by her social network's location histories thanks to the coupling of inter-personal social relationship (Toole et al., 2015b). To capture inter-personal variability, the travel activity pattern classification clusters human movements into a set of relatively homogeneous groups within which people resemble in travel patterns. Within each cluster, regression techniques including relational Markov network (Widhalm et al., 2015), conditional random field (Allahviranloo and Recker, 2015), or hidden Markov model (Allahviranloo and Recker, 2013; Liu et al., 2015) is employed to relate individual's socio-demographics to travel patterns and urban infrastructure space (Recker et al., 1985; Paul et al., 2014).

Travel activity pattern classification has been extensively studied for its critical role in numerous applications: (1) to characterize the inter-personal variability in travel activity patterns (Joh et al., 2001, 2002); (2) to measure the deviation of the predicted travel activity patterns generated by some activity-based simulation from the observed ones (Sammour et al., 2012); (3) to relate classification to socio-demographics (Recker et al., 1985; Jiang et al., 2013); and (4) to infer and predict activities of unknown persons within one group, which can be used to generate travel activities and travel demand of one specific group in the activity-based simulation (Allahviranloo et al., 2014; Toole et al., 2015b; Li and Lee, 2017).

The rationale underlying travel activity pattern classification is to compute a distance between every two travelers' patterns using a predefined similarity measure and apply some classification method to separate these patterns based on distances. To ensure the efficiency of similarity computation, dimension reduction techniques are sometimes used to project the original travel activity space to a transformed space before any similarity computation can be conducted. In summary, travel activity pattern classification usually involves three steps: (i) dimension reduction: principal component analysis or eigen decomposition (Jiang et al., 2013), feature extraction (Recker et al., 1985); (ii) similarity measure specification: cosine (Toole et al., 2015b), Euclidean distance (Joh et al., 2001; Jiang et al., 2013), (multi-dimensional) sequence alignment, agenda dissimilarity (Allahviranloo et al., 2014); and (iii) classification: K-means (Jiang et al., 2013; Toole et al., 2015b).

Among these steps, defining an appropriate similarity measure is the key. Sequence alignment method is one of the most widely used similarity measure in transportation literature (Joh et al., 2001, 2002; Sammour et al., 2012; Allahviranloo et al., 2014). The idea is borrowed from biology when two DNA sequences need to be compared. The distance between two strings is characterized by the minimum operational distance to equate these two strings, where each operation (i.e., deletion, insertion, substitution, identity) is assigned a cost. Table 1 summarizes the existing similarity measures over different dimensions of travel activity patterns. These dimensions include: (i) activity types (i.e., distribution of frequencies of activity types accomplished), (ii) order or sequential information (i.e., which activity follows which), (iii) spatio-temporal characteristics (i.e., where and when an activity is performed), (iv) attribute interdependency, and (v) pattern frequency (i.e., how often a subsequence of one activity pattern is conducted).

The existing similarity measures have several issues: (1) Intra-personal variability is not usually taken into account in activity clustering due to lack of multi-day high-resolution travel data. Accordingly, frequency information is not captured using the existing similarity measures; (2) When similarity is calculated, one's single-day activity sequence is used in most literature. These activity sequences are usually formatted as a vector of labels by dividing a 24-h day into fixed time intervals with most elements "home" due to the longer stay duration of staying at home activity, which may lead to high similarity value across the population and fail to capture differences among other activities; (3) The existing similarity measures are generally not normalized and may not be easy to compare across a population;

Table 1
Similarity measure comparison.

Similarity measure	Pattern dimensions					Range	Data format	Reference
	Activity type	Order	Spatio-temporal	Attribute interdependency	Frequency			
Euclidean distance	✓					[0, ∞)	Single-day equal-spaced equal-length	Recker et al. (1985), Joh et al. (2001), Toole et al. (2015b)
Cosine	✓					[0, 1]	Single-day equal-spaced equal-length	Toole et al. (2015b), Joh et al. (2001)
Sequence alignment	✓	✓		✓		[0, ∞)	Single-day equal-spaced equal-length	Joh et al. (2001, 2002)
Agenda dissimilarity			✓			[0, ∞)	Single-day equal-spaced equal-length	Allahviranloo et al. (2014)
Frequency similarity	✓	✓			✓	[0, 1]	Multi-day unequal length combining repetitive labels	This paper