Transportation Research Part B 91 (2016) 511–524

Contents lists available at ScienceDirect

Transportation Research Part B

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

Understanding urban mobility patterns with a probabilistic tensor factorization framework

Lijun Sun[a,b,*], Kay W. Axhausen[b,c]

[a] The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[b] Future Cities Laboratory, Singapore-ETH Centre, Singapore 138602, Singapore
[c] Institute for Transport Planning and Systems (IVT), ETH Zürich, Zürich CH-8093, Switzerland

ARTICLE INFO

Article history:
Available online 29 June 2016

Keywords:
Human mobility
Urban computing
Smart card data
Tensor decomposition
Data-driven

ABSTRACT

The rapid developments of ubiquitous mobile computing provide planners and researchers with new opportunities to understand and build smart cities by mining the massive spatial-temporal mobility data. However, given the increasing complexity and volume of the emerging mobility datasets, it also becomes challenging to build novel analytical framework that is capable of understanding the structural properties and critical features. In this paper, we introduce an analytical framework to deal with high-dimensional human mobility data. To this end, we formulate mobility data in a probabilistic setting and consider each record a multivariate observation sampled from an underlying distribution. In order to characterize this distribution, we use a multi-way probabilistic factorization model based on the concept of tensor decomposition and probabilistic latent semantic analysis (PLSA). The model provides us with a flexible approach to understand multi-way mobility involving higher-order interactions—which are difficult to characterize with conventional approaches—using simple latent structures. The model can be efficiently estimated using the expectation maximization (EM) algorithm. As a numerical example, this model is applied on a four-way dataset recording 14 million public transport journeys extracted from smart card transactions in Singapore. This framework can shed light on the modeling of urban structure by understanding mobility flows in both spatial and temporal dimensions.

1. Introduction

Urban transportation systems are the backbone of cities, allowing people to have diverse types of interactions through social activities such as work, school, shopping and leisure. To provide people-centric transportation and to enhance the efficiency of spatial interaction become the primary goals of transportation planning. As the first step towards an efficient transportation system, understanding and modeling urban transportation demand is crucial to fleet management, infrastructure design, epidemic control, urban planning and policy making. In the last decades, the developments of transportation demand modeling have created a variety of landmark works, including the classic four-step model and the recent activity-based model (Axhausen and Gärling, 1992; Bhat and Koppelman, 2003). These sophisticated approaches normally involve the estimation of a parametric spatial interaction model using available data and then apply the estimated model to predict

* Corresponding author.
E-mail addresses: sunlijun@mit.edu (L. Sun), axhausen@ivt.baug.ethz.ch (K.W. Axhausen).

http://dx.doi.org/10.1016/j.trb.2016.06.011
0191-2615/

demand at a population scale. To bring these models into practice, collecting credible data recording individual's transportation activities from travel diaries and household surveys becomes the first critical step. However, the developments of novel transportation demand models still suffer from the small sampling shares, high cost, infrequent periodicity and limited accuracy of the survey data.

With the fast development of information and communication technologies (ICT) and ubiquitous mobile computing, large quantities of digital traces that register individual human activities at both spatial and temporal scales have become available. These datasets not only help planners and researchers understand cities as complex systems, but also allow practitioners to understand cities through data-centric technologies (Batty et al., 2012). In terms of urban transportation, such data set allows us to further reveal the spatial-temporal structure of cities inherited from human mobility (González et al., 2008; Roth et al., 2011; Sun et al., 2015).

The emergence of such mobility data indeed brings us new opportunities to integrate more information into decision making. However, the complexity of the data also increases with the dimensions of its contents, which imply complex dependence and higher-order interactions among space, time and individual social-demographic attributes. Given the increasing volumes and complexity of datasets, retrieving important information and critical features from them becomes challenging. It is crucial to developing advanced data-driven approaches and models to enrich our understanding about urban transportation systems and human mobility patterns from the large amount of data.

This paper is dedicated to providing a data-driven approach to characterize the collective mobility patterns from high-dimensional structured datasets. Considering the multivariate nature of urban mobility data, we focus on depicting the complex dependence and interactions with a probabilistic framework, which provides us with a natural way to represent the structural properties and uncertainty in the data. To this end, we concentrate on a multi-way categorical setting, which allows us to summarize the high-dimensional data into contingency tables and tensors (Agresti, 2002; Kolda and Bader, 2009). In a probabilistic setting, collective mobility data can be modeled in the sense that each mobility record is a sample generated from a universal multivariate distribution. The main contribution of this paper is to introduce an analytical framework to deal with multi-dimensional transportation/mobility data. In so doing, we adopt statistical techniques to better interpret human behavior and urban dynamics. To better understand the structure of urban mobility, we apply a factorization model to decompose high-dimensional mobility data into important patterns, from which we can extract key information by reasoning about the semantics of regions and activities. Based on this information, we can further reveal the dynamics of cities. Using a real large-scale dataset from public transport systems in Singapore, we find both travel behavior and spatial configuration of the city are well structured.

The remainder of this paper is organized as follows. In Section 2, we review relevant papers on revealing urban structure using various spatial-temporal datasets and in particular the developments of applying tensor decomposition on high-dimensional urban data analytics. Section 3 introduces the framework of probabilistic factorization on multivariate transportation data based on the concept of Tucker decomposition and multi-way probabilistic latent semantic analysis (PLSA). The expectation-maximization (EM) algorithm is applied to efficiently infer the model. In Section 4, we present a case study of applying the model on a four-way dataset (time, passenger type, origin zone and destination zone) extracted from public transport smart card transactions in Singapore. We also discuss the implications and insights of the experiment. Finally, Section 5 summarizes our key findings and discusses the potential application of such a data-driven approach in helping us better understand urban dynamics.

2. Literature review

With the recent advances in ICT, large quantities of individualized data—such as call detail record (CDR), geo-referenced social media data, Global Positioning System (GPS) trajectories and transit smart card transactions—are generated through various social activities. Thanks to the rich spatial-temporal information in such datasets, analyzing and modeling the spatial-temporal structure of cities from a variety of proxies on human activities have become an emerging topic in geography, urban and transportation research (e.g., Calabrese et al., 2011; Jiang et al., 2012; Sun et al., 2012; Yuan et al., 2012; Coffey and Pozdnoukhov, 2013; Allahviranloo and Recker, 2013; Wang et al., 2014; Sun et al., 2015; Han and Sohn, 2016). In the meanwhile, the dimensions of information also become richer and richer. For example, smart card transactions contain not only spatial-temporal information of trips, but also some basic social-demographic profiles of cardholders. In most cases, such data is well-structured with pre-defined fields and each record can be considered a draw of multivariate variables.

A tensor representation allows us to summarize multivariate categorical data into a multi-dimensional array (Kolda and Bader, 2009). The goal of tensor decomposition is to efficiently reproduce the complex dependence and higher-order interactions between different modes in multivariate data by using simple structures with relatively few parameters. As a natural choice in dealing with multi-dimensional data, tensor decomposition becomes increasingly important in interpreting the underlying structure of complex datasets. It has been successfully applied in a variety of fields, such as signal processing, computer vision, online recommendation, web data mining, psychometrics and survey analysis (Kolda and Bader, 2009; Sun et al., 2006).

Given its strength in retrieving information from large datasets, tensor decomposition also attracts more and more attentions in the field of transportation data analysis. For example, Tan et al. (2013) integrated Tucker decomposition and EM algorithm to efficiently impute the missing values in a four-way tensor of traffic data (by link, day, hour and a five-minute domain). The numerical experiments suggest that the tensor-based method shows superior performance compared with