Transportation Research Part C 89 (2018) 19–34

Contents lists available at ScienceDirect

Transportation Research Part C

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

Individual mobility prediction using transit smart card data

Zhan Zhao[a], Haris N. Koutsopoulos[b], Jinhua Zhao[c,*]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, United States
[c] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Keywords:
Individual mobility
Next trip prediction
Smart card data
Bayesian *n*-gram model

ABSTRACT

For intelligent urban transportation systems, the ability to predict individual mobility is crucial for personalized traveler information, targeted demand management, and dynamic system operations. Whereas existing methods focus on predicting the next location of users, little is known regarding the prediction of the next trip. The paper develops a methodology for predicting daily individual mobility represented as a chain of trips (including the null set, no travel), each defined as a combination of the trip start time t, origin o, and destination d. To predict individual mobility, we first predict whether the user will travel (*trip making prediction*), and then, if so, predict the attributes of the next trip (t,o,d) (*trip attribute prediction*). Each of the two problems can be further decomposed into two subproblems based on the triggering event. For trip attribute prediction, we propose a new model, based on the Bayesian *n*-gram model used in language modeling, to estimate the probability distribution of the next trip conditional on the previous one. The proposed methodology is tested using the pseudonymized transit smart card records from more than 10,000 users in London, U.K. over two years. Based on regularized logistic regression, our trip making prediction models achieve median accuracy levels of over 80%. The prediction accuracy for trip attributes varies by the attribute considered—around 40% for t, 70–80% for o and 60–70% for d. Relatively, the first trip of the day is more difficult to predict. Significant variations are found across individuals in terms of the model performance, implying diverse travel behavior patterns.

1. Introduction

Individual mobility prediction is a critical enabler for various applications that support intelligent urban transportation systems, such as personalized traveler information, targeted demand management, and dynamic system operations. The success of these applications ultimately leads to enhanced customer experience and improved system performance. The prevalence of personal mobile devices (e.g., mobile phones, smart cards) makes it possible to trace individual digital footprints and potentially discover diverse and complex mobility patterns. However, the problem of predicting individual mobility remains challenging, because travel behavior concerns multiple dimensions (most notably the temporal and spatial dimensions), exhibits longitudinal variability for an individual, and varies across individuals.

Recent years have seen a considerable amount of work dedicated to human mobility modeling based on individual digital traces. The most commonly used data in these studies is mobile phone network data (Pappalardo et al., 2015; Schneider et al., 2013; Song et al., 2010b; Eagle and Pentland, 2009; González et al., 2008). Other data sources, such as GPS data (Zhao et al., 2015), Wi-Fi data

* Corresponding author.
E-mail address: jinhua@mit.edu (J. Zhao).

https://doi.org/10.1016/j.trc.2018.01.022
Received 13 August 2017; Received in revised form 18 December 2017; Accepted 20 January 2018
0968-090X/

(Sapiezynski et al., 2015) and social media check-in data (Colombo et al., 2012; Hasan et al., 2013), have also been adopted for mobility studies. This type of data is sensitive and needs to be handled in accordance with privacy legislation. One common trait of these data sources is that they are generated by non-transportation activities, and cannot be interpreted as travel behavior directly. For example, mobile phone network data are generated from cellular network activities including voice calls, text messages, cellular data activities and cell tower handovers. Thus, there is a critical mapping required to bridge the discrepancy between telecommunication behavior and travel behavior. This is not straightforward because the two aspects of behavior are driven by correlated but distinct personal preferences that vary across individuals (Zhao et al., 2016). In this paper, we refer to this type of data as *extrinsic mobility data*.

Extrinsic mobility data capture individual mobility by sampling the user's positions over time. The specific sampling process of a given data source is determined by its data-generating events. Under this data collection mechanism, individual mobility is represented as a series of time-stamped locations, and the prediction problem is framed as that of predicting an individual's next location. Next location prediction is a well studied problem, as will be discussed in Section 2, but it has some limitations in transportation applications. People generally spend most of their time staying at a location, rather than traveling between locations. Thus, over the short term, a user's location will, more often than not, stay the same. Based on the analysis of mobile phone network data, Song et al. (2010b) adopted a representation of individual mobility as a sequence of locations at hourly intervals, and reported a 93% potential predictability of the next location. However, this does not reflect the true predictability of travel behavior. In transportation, the few hours when people travel to a different location are much more important than the majority of hours when they do not. Prediction accuracy of individual locations on an hourly basis tends to be higher when people travel less frequently. For example, Hawelka et al. (2017) found that human mobility is most predictable between 02:00 and 08:00, when most people are asleep. For these reasons, we argue the next location prediction problem is not particularly helpful for transportation applications.

On the other hand, *intrinsic mobility data* are directly collected from urban transportation systems, such as transit smart card data (Goulet-Langlois et al., 2017; Zhong et al., 2015; Hasan et al., 2012) and bike sharing data (Purnama et al., 2015). Unlike extrinsic mobility data, intrinsic mobility data are generated by travel events with each record typically indicating the start or end of a trip. Therefore, intrinsic mobility data provide direct information about individual mobility as a series of trips. The individual mobility prediction problem in this case can be framed as that of predicting an individual's next trip. Unlike time-stamped locations, trips reflect critical travel decision moments, and thus match the actual behavior process of individual mobility. Although some knowledge of people's locations is useful for various location-based services (e.g., recommending a nearby restaurant), trips are more relevant to transportation applications. Intrinsic mobility data are usually mode-specific, and thus the trips captured by such data can be directly mapped to a certain transportation system.

Although a number of algorithms have been proposed for next location prediction, there is no existing model for next trip prediction. The objective of this paper is to define and formulate the specific problems that constitute individual mobility prediction based on intrinsic mobility data, propose a suitable methodology to solve these problems, and examine the predictability of individual mobility across different behavior dimensions using the proposed methodology. The method requires the historic sequence of individual trip records, and is agnostic to any particular mode. In this paper, the rail-based public transportation system in London, U.K. is chosen as a case study and the transit smart card data is the primary data source.

The remainder of the paper is organized as follows. A literature review of the related work on individual mobility prediction is presented in Section 2. Section 3 proposes a new methodological framework, based on Bayesian *n*-gram models, for individual mobility prediction using intrinsic mobility data. Section 4 demonstrates the application of the proposed methodology using pseudonymized transit smart card records of more than 10,000 users over two years. Section 5 concludes the paper with a summary of the main findings, and a discussion of future research directions and potential implications.

2. Literature review

A plethora of methods have been proposed to solve the problem of next location prediction. This problem may be formulated in two ways—predicting the location that the user will visit next (Gambs et al., 2012; Mathew et al., 2012; Noulas et al., 2012), or predicting the location of the user in the next time interval (often set as an hour) (Hawelka et al., 2017; Alhasoun et al., 2017; Lu et al., 2013; Calabrese et al., 2010). The main difference is that the former treats individual mobility as a sequence of locations, whereas the latter treats it as a sequence of time intervals with associated locations. As time is a critical dimension of mobility, the latter problem formulation is more valuable for practical applications. However, it requires data with a high data sampling frequency (e.g., the frequency in which new data arrive should be no less than the frequency of the prediction).

The majority of existing methods are based on modeling sequential patterns of individual location histories. One commonly used approach is to model the location sequence of a given user as a Markov Chain (MC). Simple MC models have been shown to be able to achieve high prediction performance (Lu et al., 2013). Gambs et al. (2012) experimented with a previously proposed model called Mobility Markov Chain (MMC) with k previous visited locations. They found that the next location could be predicted with accuracy of 70–95% when $k = 2$, although this did not improve much when $k > 2$. Note that when k increases, the number of transition probabilities grows exponentially, making model estimation increasingly difficult, especially when individual-level data is sparse. As a result, individually fitted MC models are prone to overfitting, and unable to predict locations that users have never visited before. One way to address these issues is to utilize the location histories of other users, in addition to that of the user in focus, for model estimation. The Mixed Markov chain Model (MMM), proposed by Asahara et al. (2011), is an intermediate approach between individual and universal models. MMM predicts the next location by identifying the group to which a particular individual belongs and applying the MC model for that group. Similarly, Mathew et al. (2012) presented a hybrid method of clustering location histories