Transportation Research Part C 44 (2014) 363–381

Contents lists available at ScienceDirect

Transportation Research Part C

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

# Urban activity pattern classification using topic models from online geo-location data

Samiul Hasan, Satish V. Ukkusuri *

*Purdue University, West Lafayette, IN 47907, USA*

ARTICLE INFO

*Article history:*
Received 8 January 2014
Received in revised form 8 April 2014
Accepted 9 April 2014

*Keywords:*
Activity pattern classification
Activity-based modeling
Social computing
Location-based data
Big data
Social media
Topic modeling
Machine learning

ABSTRACT

Location-based check-in services in various social media applications have enabled individuals to share their activity-related choices providing a new source of human activity data. Although geo-location data has the potential to infer multi-day patterns of individual activities, appropriate methodological approaches are needed. This paper presents a technique to analyze large-scale geo-location data from social media to infer individual activity patterns. A data-driven modeling approach, based on topic modeling, is proposed to classify patterns in individual activity choices. The model provides an activity generation mechanism which when combined with the data from traditional surveys is potentially a useful component of an activity-travel simulator. Using the model, aggregate patterns of users' weekly activities are extracted from the data. The model is extended to also find user-specific activity patterns. We extend the model to account for missing activities (a major limitation of social media data) and demonstrate how information from activity-based diaries can be complemented with longitudinal geo-location information. This work provides foundational tools that can be used when geo-location data is available to predict disaggregate activity patterns.

© 2014 Elsevier Ltd. All rights reserved.

## 1. Introduction

The introduction of location-based services in smartphone applications has enabled people to share their activity related choices (typically via "check-in") in their online social networks (e.g. Facebook, Foursquare, Twitter, etc.). Market analysis predicts 1.75 billion smartphones in the world by 2014 (eMarketer, 2014). These smartphones are typically equipped with ubiquitous location-based technologies suitable for check-in services offered by most of the social media applications. A survey on smartphone usage in the USA found that nearly one in five smartphone users are tapping into check-in services like Facebook Places, Foursquare (Comscore, 2011). The research community is realizing the potential to harness the rich information provided by these ubiquitous devices. The availability of this data has the potential to impact many areas including mobility and activity behavior analysis (Cheng et al., 2011; Noulas et al., 2011), marketing (Gao et al., 2012; Collins et al., 2013), social network analysis (Cho et al., 2011), urban planning (Cranshaw et al., 2012), and health monitoring (De Choudhury et al., 2013). As such, there are opportunities to develop fundamental tools to analyze this data at various levels of spatial and temporal resolution. Transportation researchers, also, have realized the potential of this data for travel demand modeling and analysis (Collins et al., 2013; Hasan, 2013; Cebelak, 2013; Hasan et al., 2013; Jin et al., 2014; Yang

* Corresponding author. Tel.: +1 7654942296.
*E-mail addresses:* samiul.hasan@gmail.com (S. Hasan), sukkusur@ecn.purdue.edu (S.V. Ukkusuri).

http://dx.doi.org/10.1016/j.trc.2014.04.003
0968-090X/© 2014 Elsevier Ltd. All rights reserved.

et al., 2014; Ni et al., 2014; Alesiani et al., 2014). A key challenge however is the lack of appropriate methodologies to handle such large data including the ability to address the limitations of this data.

The focus of this work is to *develop methodologies to understand individual activity patterns using large-scale location based data* obtained from social media check-in services. Through these services, individuals share their activities with the specific information on geo-location and timing of where and when they participate in those activities. Also, geo-location data can be collected in large-scale by recording the GPS coordinates from smartphones. Activity types and time of participation over multiple days can be observed from geo-location information either shared in online social media or collected by smartphones. This large-scale geo-location data can be useful to understand human activity behavior due to the extensive coverage that was unimaginable before. With the availability of new data sources like social media check-in services and smartphone GPS devices, there is a profound interest in understanding and modeling individual activity behavior. As such, geo-location data from these emerging sources has created opportunities to develop complex probabilistic models inferring the patterns of activity behavior. This paper investigates the idea of using large-scale geo-location data to infer individual activity patterns. To the best of our knowledge, this is one of the first methodological works for classifying user activity patterns using large-scale social media data.

Geo-location data from online social media and GPS devices has the potential to understand activity behavior in urban areas. Until now, many of the studies in activity-travel analysis have relied primarily on small-scale but detailed records of individual activity participation and have correlated socio-demographic information with activity participation behavior. Such modeling approaches have enormous importance in long-term policy-level analysis and planning applications. However, with the availability of big data such as the data from social media and smartphones we can observe activity participation of a large number of people over many days. We can analyze these observations to infer activity patterns up to the level of an individual without drawing any correlation with the socio-demographic attributes. Our analysis is motivated by three major shifts in our thinking about traditional approaches of activity analysis. *First*, geo-location data from a large number of individuals over many days provides the ability to analyze vast amounts of data instead of settling for small-scale data sets – accepting "the unreasonable effectiveness of data" (Halevy et al., 2009). *Second*, this approach embraces the real-world messiness in the data (e.g., missing observations) rather than depending on comprehensive data and develops methods that can account for the noisiness in these data sets. *Third*, it focuses on predicting behavior rather than explaining behavior or drawing correlations between individual activity participation and socio-demographic attributes.

Such an approach is very useful, particularly for geo-location data, to provide collective and individual patterns. This analysis may help to measure the travel demand of a region on a short-term basis; for instance, using these patterns we can compute the real-time origin–destination matrices for transportation operation models. In the context of travel demand analysis, large-scale geo-location data can provide valuable information in addition to the data from traditional surveys, and, as such, it does not replace the existing data sources but strongly complements them. Using correct analytics, such large-scale geo-location data can be used to gather complementary insights and will lead to a transformative understanding of urban travel behavior.

Geo-location data comes with larger sample size for longer period (e.g., for a year) without any significant costs and provides the location and timing of individual activity participation. However, three key limitations limit the use of traditional econometric tools for these data sets. These limitations include: (i) it does not have the detail descriptions of the activities as the start times and the end times of the activities are not reported; thus most of the current methodological approaches for modeling individual time-use behavior are not appropriate for this data; (ii) individuals are recognized by only the identification numbers without any detailed information on individual socio-economic characteristics (e.g., income, age, race, etc.); (iii) the data has missing activities, since we observe only the activities that an individual shares in social media.

However, recent advancements in machine learning techniques have made it possible to analyze large-scale geo-location data to find the spatio-temporal patterns. Specifically, hierarchical mixture modeling (popularly known as topic modeling) has emerged as a powerful methodological approach to find patterns and structure in large collections of data. These models find the latent patterns from a collection of data points where a pattern means a probability distribution over a set of predefined items. In the beginning, these topic models were used to find the underlying patterns or topics of words from a large-collection of documents (Blei et al., 2003). These patterns can be used for clustering, searching, summarizing and predicting a large corpus of documents. Later topic models were used to find patterns in images, audio and speech, genetic data, computer code and mobile phone location-sequence data. In this paper, we present an activity pattern recognition model based on a topic modeling approach.

Specifically, we find the embedded patterns found in a collection of individual activities. Such an approach can be used to classify urban activity patterns- a topic of interest in activity-travel analysis for a long time (Pas et al., 1982; Recker et al., 1985; Joh et al., 2001, 2002; Wilson, 2008; Allahviranloo and Recker, 2013). Although predicting activity travel behavior is one of the major objectives of activity-based modeling, activity pattern recognition or classification provides the basis for more theoretical or empirical analysis (Joh et al., 2001). Most of the previous approaches (Pas et al., 1982, 1983; Koppelman and Pas, 1983; Recker et al., 1985; Joh et al., 2002), used activity patterns as predefined and mainly focused on similarity measures between patterns for two reasons. *First*, typically in activity-travel behavior analysis individual activities are correlated with the socio-demographic and/or spatial contexts (Hanson and Hanson, 1981; Pas, 1984; Allahviranloo and Recker, 2013). Therefore similarities among the observed activity patterns are used to classify individuals so that representative activity patterns can be correlated with individual characteristics. *Second*, similarity values can be used as goodness-of-fit statistics to measure how well an activity-based model can predict the observed activity patterns

Download English Version:

https://daneshyari.com/en/article/524805

Download Persian Version:

https://daneshyari.com/article/524805

Daneshyari.com