Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data

Md Mehedi Hasnat, Samiul Hasan*

Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 12800 Pegasus Drive, Orlando, FL 32816, United States

A B S T R A C T

Reliable travel behavior data is a prerequisite for transportation planning process. In large tourism dependent cities, tourists are the most dynamic population group whose size and travel choices remain unknown to planners. Traditional travel surveys generally observe resident travel behavior and rarely target tourists. Ubiquitous uses of social media platforms in smartphones have created a tremendous opportunity to gather digital traces of tourists at a large scale. In this paper, we present a framework on how to use location-based data from social media to gather and analyze travel behavior of tourists. We have collected data of about 67,000 users from Twitter using its search interface for Florida. We first propose several filtering steps to create a reliable sample from the collected Twitter data. An ensemble classification technique is proposed to classify tourists and residents from user coordinates. The accuracy of the proposed classifier has been compared against the state-of-the-art classification methods. Finally, different clustering methods have been used to find the spatial patterns of destination choices of tourists. Promising results have been found from the output clusters as they reveal most popular tourist spots as well as some of the emerging tourist attractions in Florida. Performance of the proposed clustering techniques has been assessed using internal clustering validation indices. We have analyzed temporal patterns of tourist and resident activities to validate the classification of the users in two separate groups of tourists and residents. Proposed filtering, identification, and clustering techniques will be significantly useful for building individual-level tourist travel demand models from social media data.

1. Introduction

Collecting high-resolution data to observe individual travel behavior of a large population is an essential but challenging task for transportation planners. Conducting travel surveys is an expensive step in collecting travel behavior data. However, traditional travel surveys are limited in terms of sample size, area of coverage and updating frequencies. Particularly in large cities with a significant number of tourists, conducting travel surveys is difficult as tourists are the most dynamic population group whose size and travel choices change rapidly compared to the residents. Although census statistics reveal total inflow and outflow of tourists, it only presents as a macro level data considering over large regions. For major tourism dependent cities, it is essential to understand tourist travel behavior since tourism related traffic cause huge pressure on their transportation systems (Cho et al., 2011; Gursoy et al., 2002).

With millions of active users, social media platforms such as Facebook, Twitter, Instagram, Flickr etc. have become potential big data sources of individual behavior. Nearly 80 percent of Americans use social media while two third of the global internet population visits social networks (Perrin, 2005). Thus, ubiquitous uses of social media platforms have created a tremendous opportunity to gather digital traces. Analyzing millions of user footprints, it is possible to extract travel behavior at a scale unimaginable before (Hendrik and Perdana, 2014). However, not all social network data are available and have rich information. Twitter is a potential data source as it is available through simple web scraping and has a wide range of information within each post (tweet). Twitter has become a popular communication platform with 317 million monthly active users (67 million users from the USA) sending 500 million tweets per day ("Twitter Facts," 2017). Despite being unstructured, tweets provide important clues about latent user attributes and activities- absent in GPS logs and mobile phone records (Cao et al., 2015). From Twitter, we can extract spatial (geo-tagged) and temporal (time-stamped) information for a longer period and large number of users without invading user privacy (Frias-Martinez et al., 2012; Hasan and Ukkusuri, 2015).

Florida has a number of famous tourist spots attracting millions of tourists from home and abroad every year. In 2016, Florida hosted more than 113 million visitors from around the world, which supported 1.4 million jobs and making a spending of 109 billion USD (FDOT, 2017). Central Florida region had 68 million visitors in 2016 with Orlando being one of the top destinations among the global tourist cities and second in annual tourist spending among the domestic cities (FDOT, 2017). A study conducted by Florida Department of Transportation (FDOT) also found that in 2010, nearly 8% of Florida's vehicle miles travelled were comprised of tourism related travel (Florida Transportation Trends and Condition 2012, 2012). Individual movement, route choice, origin and destination of this large number of seasonal population have a significant impact on transportation infrastructure. Such information can provide vital insights for transportation and city planners.

In this paper, we present a data mining framework for understanding out of state tourists' travel behavior from social media data. We have gathered data using Twitter's search interface and followed several filtering steps to create a reliable sample. We propose a classification method to identify the users who are non-native to a particular area. After identification and validation of the tourist accounts, we find the spatial patterns of tourist destination choices to discover the most visited locations in the study area.

Thus, this paper has several contributions. First, it shows the potential of social media data for understanding tourist travel behavior. Second, it presents several filtering steps to create a reliable sample from noisy social media data. Third, it presents a simple but effective method to classify resident users and visitors within a study area. Finally, it presents and compares available spatial clustering methods to determine destination choice patterns of tourists. The proposed identification and clustering techniques will be significantly useful for building individual level travel behavior models based on social media data.

2. Literature review

Location-based social media data have been used to better understand different aspects of mobility and activity patterns. The easy availability and wide range of applications have made the data valuable for researchers in multiple fields including social science, marketing, public health, computer science, and transportation science. Social media data have been used in activity recognition (Lian and Xie, 2011), finding mobility and activity choices (Chen et al., 2017; Hasan and Ukkusuri, 2014; Hu and Jin, 2017), classification of activity choice patterns (Cheng et al., 2011), role of friendship on mobility (Hasan et al., 2016; Sadri et al., 2018, 2017), modeling activity sequence (Hasan and Ukkusuri, 2017) and travel behavior analysis from longitudinal travel data (Zhang et al., 2017a). Several studies have discussed the potential and shortcomings of using Twitter data for longitudinal travel behavior analysis at length (Kuflik et al., 2017; Zhang et al., 2017b). In transportation planning, researchers have used this data to estimate urban travel demand (Lee et al., 2017; Liu et al., 2014; Vij and Shankari, 2015), destination choice (Gkiotsalitis and Stathopoulos, 2015; Huang et al., 2017), traffic flow (Liu et al., 2014; Vlahogianni et al., 2015; Wu et al., 2014), event related transit passenger flow (Cottrill et al., 2017; Ni et al., 2016) and traffic flows (Ni et al., 2014) etc. Most recently location based social network data are being used to detect real time traffic incidents such as traffic accidents (Gu et al., 2016; Zhang et al., 2018), non-recurrent traffic surge (Zhang et al., 2016), traffic disruptions and anomalies due to special events (D'Andrea et al., 2015; Pan et al., 2013; Steiger et al., 2016) etc. Thus, social media data has a significant potential for travel demand models, traffic operations and management and long term transportation planning purposes (Rashidi et al., 2017).

Analyzing tourist travel behavior is a difficult task compared to normal residents due to the non-repetitive movements made by the tourists. Several studies have analyzed geo-tagged photos to discover patterns of tourist mobility around the world (Chareyron et al., 2013). Girardin et al. (2007) used geo-tagged photos to reveal activity and flows of tourists in space and time. Similarly, Majid et al. (2013) and Sun et al. (2013) used geo-tagged photos from Flickr to find tourists' preferable places. Despite its wide adoption, few studies have investigated tourist behavior using Twitter data. Abbasi et al. (2015) considered tourists who are traveling both into and outside of Sydney within four weeks. Analyzing geo-tagged tweets, they could identify the most visited places by local residents and tourists. Using geo-tagged tweets, Lee et al. (2016a, 2016b) demonstrated the growth of activity space of 116 Twitter users over 17 weeks and determined their major activity locations. Various studies have used individual GPS data to understand activity types and activity locations. Scherrer et al. (2018) used mobile app data to cluster the users of two Australian cities (Sydney and Melbourne) in separate groups based on 32 different features and tagged the most prominent two clusters as locals and visitors. Ye et al. (2009) explored the potential of using GPS data to define individual life patterns based on their most frequently visited locations. Similar works have also been found in Liao (2006), Zheng et al. (2009), where GPS data have been used to analyze daily activities and interesting locations. However, collecting GPS data for large geographical area is time consuming and costly. Therefore, being a source of location-based data, Twitter has a tremendous potential to be used for tourist travel behavior analysis.