Transportation Research Part C 77 (2017) 245–256

Contents lists available at ScienceDirect

Transportation Research Part C

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

Travel analytics: Understanding how destination choice and business clusters are connected based on social media data ☆

Arthur Huang [a,*], Luciano Gallegos [b], Kristina Lerman [b]

[a] *Department of Engineering and Computer Science, Tarleton State University, Stephenville, TX 76401, United States*
[b] *Information Science Institute, University of South California, Marina del Rey, CA 90292, United States*

ARTICLE INFO

Article history:
Received 18 October 2015
Received in revised form 7 October 2016
Accepted 30 December 2016

Keywords:
Social media
Check-in data
Destination choice
Land use

ABSTRACT

Understanding how destination choice and business clusters are connected is of great importance for designing sustainable cities, fostering flourishing business clusters, and building livable communities. As sharing locations and activities on social media platforms becomes increasingly popular, such data can reveal destination choice and activity space which can shed light on human-environment relationships. To this end, this research models the relationship between characteristics of business clusters and check-in activities from Los Angeles County, California. Business clusters are analyzed via two lenses: the supply side (employment data by industry) and the demand side (on-line check-in data). Spatial and statistical analyses are performed to understand how land use and transportation network features affect the popularity of the identified clusters and their relationships. Our results suggest that a cluster with more employment opportunities and more types of employment is associated with more check-ins. A business cluster that has access to parks or recreational services is also more popular. A business cluster with a longer road network and better connectivity of roads is associated with more check-ins. The visualization of the common visitors between clusters reveals that there are a few clusters with outstanding strong ties, while most have modest ties with each other. Our findings have implications on the influence of urban design on the popularity of business clusters.

1. Introduction

Understanding the relationship between business clusters and individuals' travel behavior has been an important and challenging endeavor. Such understanding can not only help unravel travel decisions and human nature, but also shed light on how to foster flourishing business clusters and design livable cities. Obtaining such understanding is a challenge in that the forces that shape travel and the built environment are multifaceted and complex. To answer complex behavioral questions requires more microscopic travel data.

With the advancement of computing technologies, the method of collecting travel data has evolved from traditional diary household surveys (Stopher and Greaves, 2007), GPS studies (Wolf et al., 2003; Bohte and Maat, 2009; Gong et al., 2012; Huang and Levinson, 2015), smartphone-based surveys (Cottrill et al., 2013; Nitsche et al., 2014), urban sensing data (Calabrese et al., 2011; Lu and Liu, 2012; Calabrese et al., 2013; Louail et al., 2014), to social media data (Wu et al., 2014;

☆ This article belongs to the Virtual Special Issue on "Smart City".
* Corresponding author.
E-mail addresses: ahuang@tarleton.edu (A. Huang), luciano.gallegos@gmail.com (L. Gallegos), lerman@isi.edu (K. Lerman).

http://dx.doi.org/10.1016/j.trc.2016.12.019

Jurdak et al., 2015; Cao et al., 2015). The increasing popularity of social media especially provides new opportunities for collecting behavioral data. Research has shown that 62% of all American adults (ages 18+ years) use Facebook, 22% use linkedin, 21% use Instagram, and 20% use Twitter (Duggan, 2015). Individuals can broadcast where they are, what they do, and how they feel on various social media platforms. Certain social media sites such as Facebook, Foursquare, and Gowalla grant users to announce their current locations by "checking-in" to services. Such announcements can be shared on various social media sites with linked accounts (Gallegos et al., 2016). The greatest advantage of such social media data is that they are free and comprise large volumes of almost real-time data with geo-locations across the world (Hawelka et al., 2014). Such spatio-temporally explicit data have advantages in capturing revealed real-time human behavior and human-environment relationships at a large scale with low cost. Social media data have been used to investigate park usage (Torres and Costa, 2014), understand emotions (Gallegos et al., 2016), predict the wellbeing of large populations (Schwartz et al., 2013), investigate network disruptions (Chan and Schofer, 2014; Pender et al., 2014) and major disasters (Ukkusuri et al., 2014), classify urban activity patterns (Hasan and Ukkusuri, 2014), and detect influenza epidemics (Aramaki et al., 2011). Despite the unequal geographic distribution and socio-demographic biases (Hawelka et al., 2014), geo-tagged social media data may complement regular household travel survey data to validate findings on mobility and urban structure and to reveal new insights on human nature and behavior.

This research aims to uncover the relationship between destination choice and the built environment based on online check-in data and land use data through spatial and statistical analyses. Our approach is data-driven. We are interested in detecting business clusters from both the demand side (check-ins) and the supply side (jobs by industry). Statistical models are applied to modeling how various urban design factors influence the popularity of business clusters. One key contribution of this research is to investigate the methods for analyzing and visualizing geo-tagged social media data to understand destination choice and business clusters. We further discuss how social media data can be used to supplement travel survey data to explore human-environment relationships.

The rest of this paper is organized as follows: Section 2 reviews related literature. Section 3 introduces the data sets used in this research. Section 4 defines business clusters based on the check-in points. Section 5 models the relationship between clusters' land use characteristics and their popularity. Section 6 visualizes the common visitors between the identified business clusters. Section 7 discusses the implications of our findings and concludes this paper.

2. Literature review

There has been a growing interest in studying location-based datasets to understand human nature. The emergent research topics range from cluster analysis, social networks, mobility, emotions, depression analysis, network disruptions, to the distribution of certain types of food. For instance, Järv et al. (2014) studied monthly variability in human activity spaces based on cell phone data. They identified modest monthly variation in the number of activity locations but great variations in the size of activity space. Widener and Li (2014) combined Census tracts data and geo-tagged tweets to explore the prevalence of healthy and unhealthy food across the United States. Their findings uncovered that lower income Census tracts are associated with a lower proportion of tweets about healthy foods and a higher proportion of tweets about unhealthy foods. Yang and Mu (2015) proposed a procedure to identify depressed users in Twitter and visualized the distribution of depression. Goncalves and Sanchez (2014) used the *k*-means clustering algorithm with geo-tagged Spanish tweets to map the spatial distribution of Spanish-speaking population around the world. Noulas et al. (2011) analyzed Foursquare check-ins with 700,000 users over 100 days to predict the amount of travel. They found that most users moved between 1 and 10 km and spent between 100 and 2000 min traveling. Lenormand et al. (2014) analyzed the tweets posted from roads and rails in Europe and found a positive correlation between the number of tweets on the road and the AADT on highways in the U.K. and France. Mitchell et al. (2013) and Gallegos et al. (2016) assessed the spatial distribution of happiness by performing sentiment analysis on the tweets. Himelboim et al. (2013) developed a Selective Exposure Cluster (SEC) method to identify clusters of interconnected users through the topic of the U.S. President's State of the Union speech in 2012 by analyzing each cluster's hub users, hash tags, hyperlinks, and top-mentioned usernames. Chan and Schofer (2014) examined the use of social media by transit agencies in New York City during Hurricane Sandy in 2012 to reveal the disruption and restoration of the transit network. Yin et al. (2012) extracted situation awareness information from Twitter messages during disasters and crises and pinpointed the correlation of Twitter traffic and the happening of disasters and aftershocks. Cho et al. (2011) investigated the relationship between friendship and human mobility based on social network check-in data and cellphone data. It was found that short-range activity space is more influenced by the social network structure while long-term travel is more influenced by the social network ties.

In transportation research, geo-tagged location data have been used to shed light on land use patterns, human mobility, and the spatial structure of a city. For example, Frias-Martinez and Frias-Martinez (2014) identified the relationship between where people tweet and land use patterns. This research found a dramatic increase of the number of tweets when people arrived at work, went for lunch, and were to leave for the day, which can be used to pinpoint where popular businesses are located. Hasan and Ukkusuri (2014) used geo-tagged check-in data to capture user-specific patterns and identify the top users who contribute to specific patterns. Frias-Martinez and Frias-Martinez (2015) investigated geo-tagged twitter activities to characterize land use patterns in London (U.K.) and Madrid (Spain). Cranshaw et al. (2012) used approximately 18 million geo-tagged check-ins to study the structure and composition of a city. The resulting communities are named as