Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Identifying spatiotemporal urban activities through linguistic signatures

Cheng Fu[a,*], Grant McKenzie[a,b], Vanessa Frias-Martinez[c], Kathleen Stewart[a]

[a] LeFrak Hall, Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA
[b] Burnside Hall Building, Department of Geography, McGill University, Montreal, Quebec H3A 0B9, Canada
[c] Hornbake Building, College of Information Studies, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Twitter
Natural language processing
Big data
Human activity modeling
Urban dynamics

ABSTRACT

Identifying the activities that individuals conduct in a city is key to understanding urban dynamics. It is difficult, however, to identify different human activities on a large scale without incurring significant costs. This study focuses on modeling the spatiotemporal patterns of different activity types within cities by employing user-contributed, geosocial content as a proxy for human activities. In this work, we use linguistic topic modeling to analyze georeferenced twitter data in order to differentiate different activity types. We then examine the spatial and temporal patterns of the derived activity types in three U.S. cities: Baltimore, MD., Washington, D.C., and New York City, NY. The linguistic patterns reflect the spatiotemporal context of the places where the social media content is posted. We further construct a method to link what people post online to the activities conducted within a city. We then use these derived activities to profile the characteristics of neighborhoods in the three cities, and apply the activity signatures to discover similar neighborhoods both within and between the cities. This approach represents a novel activity-based method for assessing similarity between neighborhoods.

1. Introduction

Urban life involves a variety of activity types that are an intrinsic part of urban dynamics, including commuting, shopping, dining out, etc. Exploration and analysis of these different types of activities leads to a better understanding of the pulse of the urban landscape, e.g., transportation, economic, and social behaviors. People's activities in the street comprises Jane Jacobs' "sidewalk ballet" (Jacobs, 1961). Activities also help to delimitate *places*. From *structuration* theory, places are established only if they are locations of constant and reiterative activity (Cresswell, 2014). Poststructuralist *assemblage* theory that refers to the emerge of new unique wholes from the interactions between parts also highlights that the dynamics in a city contribute to an emerging sense of place (Dovey, 2012). Therefore, understanding differences in activity types, and the magnitude of these activities at different locations in a city provides information on the intrinsic nature of different places. Sensed activities can be utilized for decision-making in urban planning or for improving services.

One conventional method for characterizing parts of a city, i.e., neighborhoods, is to use demographic data. For example, the ESRI Tapestry[1] project categorizes residential neighborhoods in the United States into 67 types by employing Census data. Census data, however, does not reflect how people actually interact with urban spaces, and does not cover the socio-economic aspects of the neighborhoods that incorporate, for example, commercial areas, since a census only surveys residents. Using a derived activity distribution among the neighborhoods, we can categorize neighborhoods from an activity-based perspective, and compare the similarity of neighborhoods based on this new perspective.

Sensing human activities in a city can be financially expensive and time consuming. Given the complexity of modern survey techniques, researchers in different fields often survey only a sampled group of individuals with some denoted types of activity that are closely related to their study theme. For example, studies in transportation mainly utilize transportation activity surveys such as the U.S. National Household Transportation Survey (NHTS, Cervero & Kockelman, 1997; Chalasani, Denstadli, Axhausen, & Engebretsen, 2005) or equip a limited number of enrolled vehicles with GPS loggers to track vehicle movements (Wolf, Guensler, & Bachman, 2001). Studies on public health also utilize travel surveys, for example, to link eating activities with a geographical context (Kestens, Lebel, Daniel, Thériault, & Pampalon, 2010; Widener, Farber, Neutens, & Horner, 2015).

Recently, socially-sensed geospatial data sets (*social sensing*, Liu et al., 2015) have been used as proxies of human activities. Socially sensed geodata includes geographic information that are voluntarily contributed by individuals (volunteered geographic information, VGI,

* Corresponding author.
E-mail address: cfu@terpmail.umd.edu (C. Fu).
[1] http://www.esri.com/landing-pages/tapestry

https://doi.org/10.1016/j.compenvurbsys.2018.07.003
Received 13 July 2017; Received in revised form 13 March 2018; Accepted 13 July 2018

Goodchild, 2007), such as the geospatial data of OpenStreeetMap (OSM), georeferenced accident reports on Waze, and geospatial data that is collected but not purposely contributed by the individuals who generate the data (McKenzie & Janowicz, 2014), such as georeferenced taxi trajectories, call detailed records (CDRs), check-in (Cranshaw, Schwartz, Hong, & Sadeh, 2012), and georeferenced microblog posts from Twitter, a social network service (SNS). A georeferenced Tweet is a short message (typically text-based) limited to 140[2] characters from a Twitter user that includes metadata such as a location and a timestamp. In this work, we show how these tweets can be used to represent activities that are being undertaken by individuals in multiple cities.

Previous studies that utilized Tweets as proxies for human activities typically only model *posting a Tweet (tweeting),* as an identical activity, and used the variation of tweet volume only to characterize the social function of a region without fully utilizing the text in tweets that may provide further detailed activity type information. Projects, such as UrbanTick[3] by Neuhaus, relied on a change in the volume of tweets (spatially and temporally) to characterize the activity rhythm, or "the pulse of the city" (Batty, 2010). Such variations in tweet volumes are also used to characterize regions' social functions in a city by combining machine learning approaches (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012; Lee, Wakamiya, & Sumiya, 2012; Wakamiya, Lee, & Sumiya, 2011).

The textual content of a tweet contains useful, descriptive information that is often overshadowed by the spatiotemporal meta data. Within the content of a tweet, people often explicitly or implicitly express their thoughts and feelings related to activities they are conducting when they are tweeting. Text analytics can thus extract place references and meaningful information from georeferenced tweets and construct place characterizations (MacEachren, 2017). One approach that has been taken previously is to filter related tweets by keywords, for example, Tsou et al. (2013)'s analysis on candidate names in the 2012 U.S. Presidential Election and Yang et al. (2016)'s system for exploring human dynamics based on people's interests.

Keyword analysis, however, may only expose specific events that involve a limited set of keywords closely related to the event. There may be new terms created to refer to a new event or a new type of activity that cannot be identified by a predefined set of keywords. Alternatively, an approach such as topic modeling that derives latent topics in text by a word-based statistical modeling approach can be used for knowledge discovery without predetermined keywords (Hofmann, 1999)

One of the most prevalent topic-modeling approaches is latent Dirichlet allocation (LDA Blei, Ng, & Jordan, 2003). LDA assumes that each document in a corpus is associated with numerous latent topics that can be characterized by a unique word probability distribution. LDA and its variants on classification (Blei & McAuliffe, 2008; Ramage, Hall, Nallapati, & Manning, 2009) have been used extensively in previous spatial and place-based research (Adams, McKenzie, & Gahegan, 2015; Chae et al., 2012; Hu & Ester, 2013) but the standard LDA approach is arguably not a good model for tweets given the limited text length in a typical tweet. One solution is to aggregate tweets as one long document based on locations or time intervals to fit into the standard LDA model (Eisenstein & O'Connor, 2010; Jenkins, Croitoru, Crooks, & Stefanidis, 2016; McKenzie, Janowicz, Gao, & Gong, 2015; Mehrotra, Sanner, Buntine, & Xie, 2013; Puniyani, Eisenstein, Cohen, & Xing, 2010). As alternatives Twitter-LDA (Zhao et al., 2011) and Single Topic LDA (ST-LDA Hong, Yang, Resnik, & Frias-Martinez, 2016) assume that: 1) only one topic is involved in each tweet post due to Twitter's length limitation; and 2) multiple authors are involved in writing a collected tweet dataset. Such assumptions are similarly reasonable for this study and for this reason ST-LDA is used as the primary means for topic modeling as it has also been applied to analyze resident-government communication pattern in disaster (Hong, Torrens, Fu, & Frias-Martinez, 2017). Besides LDA models Deep Learning frameworks on topic modeling have also been applied to the same task (Wang et al., 2016)

This research uses the volume profile of different activities as a quantitative means to retrieve knowledge about and the sense of places. This research explores the value of using a large user-contributed georeferenced dataset as a proxy for activities within and between cities on the east coast of the United States, and identifies and compares regions with respect to their activity profiles over several months. Using ST-LDA to build the model that links tweets to activities allows us to explore how activities are distributed both in time and space. This distribution can help us in two ways: First, the temporal and spatial patterns are used to validate the accuracy of the topic model in representing meaningful activities. Second, the overall distribution of the topics is employed to characterize places, such as different neighborhoods. The new computational model also provides feasibility to analyze the activity patterns with finer granularity in time and space as there is no pre-processing geographical or temporal units for aggregating the tweets to form a long text for fitting into a standard LDA model.

In this study, two major research objectives are addressed:

RO1. . An natural language processing (NLP) workflow is applied to derive meaningful activity types from a large number of Twitter posts, and the resulting activity types are evaluated based on their spatial and temporal distributions. We specify a null hypothesize (H1) that the topics derived from georeferenced tweet are identically distributed in space and time. In this work we will demonstrate how this null hypothesis is falsified.

RO2. . The derived activities are used to profile the activity signatures of neighborhoods in three U.S. cities as a novel approach to characterizing the neighborhoods. The activity signatures are further employed to find similar neighborhoods both within and between cities. We specify a null hypothesis (H2) that aggregated topics, as proxies for activities, offer identical signatures that cannot differentiate one neighborhood from another. In this work we will nullify this hypothesis by showing that there are statistically significant differences in the topic signatures.

The remainder of this paper is organized as follows: Section 2 introduces the Twitter dataset collected from three cities in the U.S. for an empirical study. Section 3 discusses the approach used to extract activities from text in Tweets, and validates the set of derived topics via their spatio-temporal distributions. Section 4 shows how the neighborhoods are characterized by the derived activities and how the similar neighborhoods are found. Section 5 takes a neighborhood in Washington D.C. as a case study to show the effectiveness of the model presented in Section 4. The conclusions are presented in Section 6 of this paper, along with a discussion addressing potential limitations, and suggestions for future work.

2. Data

Twitter allows users to register anonymously and to post messages, labeled *tweets*, with rich metadata, including a unique ID for the tweet, a user ID identifying which user posted a message, a time stamp indicating the time when the message was posted, to name a few. Within the content of a tweet, a user can use a hashtag (#) as the prefix to highlight a keyword to summarize the theme of the message or to draw others' attention. If users post tweets from a location-embedded mobile device, Twitter also allows users to include the device's coordinates as part of the tweet's metadata. Twitter provides a set of freely-accessible

[2] The character limit changed to 280 after September 2017

[3] http://urbantick.blogspot.com/2010/01/new-city-landscapes-interactive.html