Computers, Environment and Urban Systems 54 (2015) 255–265

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data

Enrico Steiger *, René Westerholt, Bernd Resch, Alexander Zipf

GIScience Research Group, Institute of Geography, Heidelberg University, Germany

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 8 September 2015
Accepted 17 September 2015
Available online xxxx

Keywords:
Crowdsourcing of human activities
LBSN
Twitter
Spatial autocorrelation
Semantic topic modeling

ABSTRACT

Detailed knowledge regarding the whereabouts of people and their social activities in urban areas with high spatial and temporal resolution is still widely unexplored. Thus, the spatiotemporal analysis of Location Based Social Networks (LBSN) has great potential regarding the ability to sense spatial processes and to gain knowledge about urban dynamics, especially with respect to collective human mobility behavior. The objective of this paper is to explore the semantic association between georeferenced tweets and their respective spatiotemporal whereabouts. We apply a semantic topic model classification and spatial autocorrelation analysis to detect tweets indicating specific human social activities. We correlated observed tweet patterns with official census data for the case study of London in order to underline the significance and reliability of Twitter data. Our empirical results of semantic and spatiotemporal clustered tweets show an overall strong positive correlation in comparison with workplace population census data, being a good indicator and representative proxy for analyzing workplace-based activities.

1. Introduction

Cities are multifunctional complex systems serving as major hubs for a number of human social activities. With more than half of the world's population living in urban areas and a continuing urban growth (United Nations Population Fund, 2008), the capability to provide viable service infrastructure (roads, public transport, energy supplies, etc.) for the majority of people is a rising challenge. The characterization of urban structures can facilitate urban and transportation planning processes providing valuable information, which helps to predict the increased pressure on existing urban infrastructures. Regular commuting from workplaces to places of residence, and activities originating from these areas, are major examples of daily routines within urban areas, influencing human mobility and affecting transportation planning. In the UK in 2013, a person on average made 145 trips with 19% of all trip purposes related to business and commuting activities (Department for Transport, 2014).

Determining the frequency and spatial distribution of travel origins and destinations for every trip purpose is a principal quantitative study area currently carried out by mobility surveys (Morris, Humphrey, & Tipping, 2014). However, they are expensive in terms of the required labor input and usually lead to limited sample sizes. Thus, the investigation of typically larger spatiotemporal human activity clusters obtained from crowdsourced information may help to understand commuting patterns and reveal specific urban structures such as workplace concentrations.

In this context, emerging, inexpensive and widespread sensor technologies have created new possibilities to infer mobility data for exploring urban structures and dynamics. This growing availability of mobile devices equipped with GPS sensors having broadband internet access, allows users to actively participate and create content through mobile applications and location-based services (ITU, 2014).

Particularly georeferenced Twitter data is a promising opportunity to understand geographic processes inside online social networks. The enormous potential of interactive social media platforms like Twitter has been increasingly recognized by numerous research domains over the last years. Although there is a growing research body using Twitter data to analyze urban processes, empirical research towards the validation of human social activities revealing urban structures and human mobility patterns using crowdsourced information is still widely unexplored (Resch, Beinat, Zipf, & Boher, 2012).

In a previous study we introduced a semantic and spatial analysis method, through which we were able to extract human mobility flows from uncertain Twitter data (Steiger, Ellersiek, & Zipf, 2014). However, it remains to be investigated whether we can find similar semantic layers that represent collective human behavior in co-occurrence with underlying social activity.

Therefore, research question (RQ1) investigates the possibility of exploring urban structures through characterizing spatiotemporal and semantic patterns of human social activities. Hence, we extract topics covering work-related and home-related activities that reflect typical collective human behavior (e.g., city-scale human mobility). Thus, the

* Corresponding author at: Institute of Geography, Heidelberg University, Berliner Straße 48, D-69120 Heidelberg, Germany.
E-mail address: enrico.steiger@geog.uni-heidelberg.de (E. Steiger).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.007
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

first research question aims to find evidence for the reflection of collective behavior in tweets. In a further step, the second research question (RQ2) seeks to validate these findings against reliable census data. In particular, we examine associations and correlations between tweets as a proxy indicator of human social activities and available census populations. Summarizing, the main goal of this paper is to validate the detected human social activity clusters from RQ1 with official UK census data.

We have chosen London to be a reasonable study site, given the vast number of Twitter users in this city, providing us with a large enough data sample for our research. This second research question is particularly important against the background of a broad range of uncertainties that arise with Twitter data analysis (s. sub-section 2.1). To the best of our knowledge, no available study has conducted this kind of validation between semantic information extracted from Twitter data and official census information. We aim to provide a first empirical ground truth on how representative and trustworthy tweets for the inference of social activities indicating human mobility are. We propose a suitable methodological approach for answering these questions.

2. Background

The dataset used in this analysis is collected from Twitter. Within online social networks like Twitter, individuals can create an online profile and communicate with other users by sharing common ideas, activities, events or interests (Boyd & Ellison, 2007). Twitter further enhances existing social networks by adding a spatial dimension becoming a LBSN and allows users to exchange details of their personal location as a key point of interaction (Zheng, 2011). Users can post short status messages, namely tweets with up to 140 characters. With the permission of the user, each tweet contains a corresponding geolocation acquired from the GPS sensor within the mobile device. Therefore, user posts in Twitter represent a spatiotemporal digital footprint (geolocation and timestamp of tweet) with a semantic information layer (content of tweet message).

Georeferenced tweets correspond to particular locations and are influenced by each user's individual perception of urban space (Fig. 1). Thus, Twitter data and specific contextual information might serve as an indicator on how strongly the virtual and physical worlds are connected with each other. However, unlike with Foursquare where users can “check in” at predefined venues (restaurants, hotels, etc.), we do not have any a priori knowledge regarding underlying human social activities in Twitter. One interesting question is therefore concerned with investigating whether single tweets denoting a specific semantic incident tend to co-occur with similar other tweets being close in geographic space and time. Such clustering behavior might provide converging evidence about underlying social activity. Our given example tweet “I'm at work” (see Fig. 1), for instance, indicates a particular human social activity which may characterize an underlying urban structure. In this case a possible indication of a workplace.

2.1. Potential limitations of Twitter data analysis

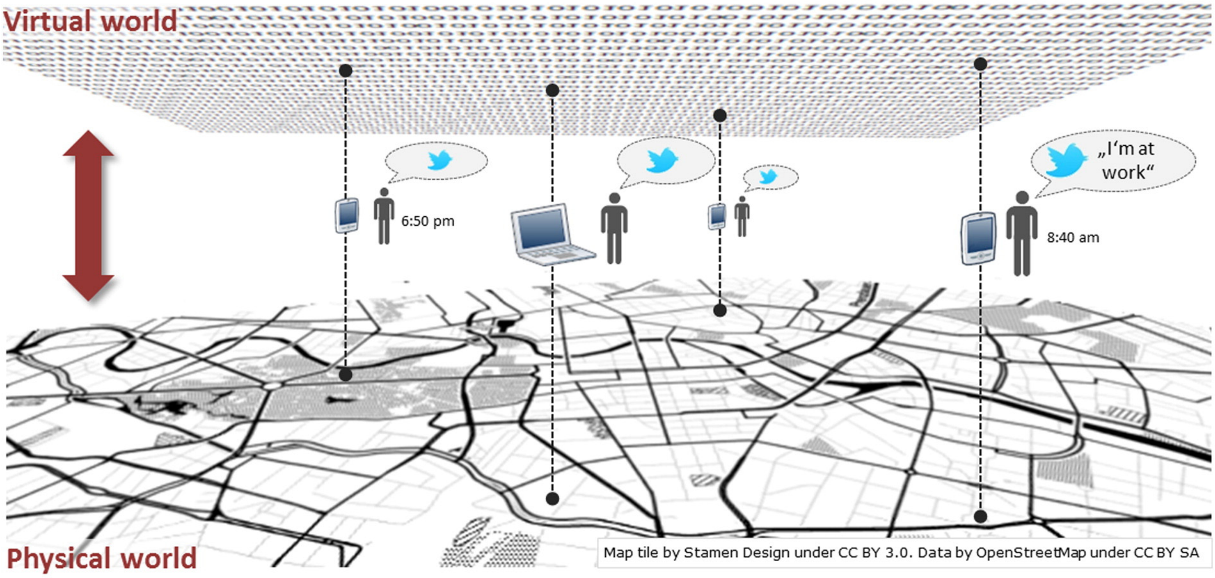
When analyzing spatiotemporal and semantic information from Twitter, we face several data-specific uncertainties, including a number of components such as the location information, the extracted knowledge and the applied methodology.

2.1.1. Location uncertainty

The location information retrieved by built-in GPS receivers might be inaccurate due to different effects. These include intrinsic effects like adverse mobile device characteristics, but also extrinsic factors such as the built environment or the GPS dilution of precision (Zandbergen & Barbeau, 2011). Furthermore, users can individually choose to add their precise location to a tweet or just a general attached location information (such as a city or neighborhood). This might result in imprecise and coarse location information of geotagged tweets.

2.1.2. Sampling Biases

The spatial distribution of tweets in location-based online social networks is also spatiotemporally heterogeneous as users do not contribute records equally across space and time. Particularly the spatial distribution of tweets strongly varies on different real-world scale levels (country, city, borough, etc.) and might be too sparse in some geographical areas (Sengstock & Gertz, 2012). Moreover, when focusing on the ratio between the number of active Twitter users and the overall population, there is also a mismatch between the population and the sampling frame. This effect might lead to exclusion or under/over-representation of certain population groups (Heckman, 1979). In consequence, unrepresentative subsets and different sample sizes from the whole amount of tweets might be generated depending on the Twitter information and analysis approaches (e.g., only georeferenced tweets).



Map tile by Stamen Design under CC BY 3.0. Data by OpenStreetMap under CC BY SA

Fig. 1. Information layers according to Resch et al. (2012).