Computers, Environment and Urban Systems 61 (2017) 108–118

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Deriving retail centre locations and catchments from geo-tagged Twitter data

Alyson Lloyd, James Cheshire *

Department of Geography, UCL, London, UK

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 20 August 2016
Accepted 28 September 2016
Available online 20 October 2016

Keywords:
Social media
Retail
Twitter
Consumer data
Human mobility

ABSTRACT

This investigation offers an initial foray into the application of geo-tagged Twitter data for generating insights within two areas of retail geography: establishing retail centre locations and defining catchment areas. Retail related Tweets were identified and their spatial attributes examined with an adaptive kernel density estimation, revealing that retail related Twitter content can successfully locate areas of elevated retail activity, however, these are constrained by biases within the data. Methods must also account for the underlying geographic distribution of Tweets to detect these fluctuations. Additionally, geo-tagged Twitter data can be utilised to examine human mobility patterns in a retail centre context. The catchments constructed from the data highlight the importance of accessibility on flows between locations, which have implications for the likely commuting choices that may be involved in retail centre journey decision-making. These approaches demonstrate the potential applications for less conventional datasets, such as those derived from social media data, to previously under-researched areas.

1. Introduction

In the UK online retailers have achieved a market share of 16.8% (Retail Research, 2016). This growing percentage represents one of a number of challenges facing traditional British high-streets and town centres, with consumers increasingly substituting certain types of town-centre retailers for these online alternatives (Weltevreden, 2007). Those most affected have been music, video and book retailers, travel agents and a wide range of other retailers that have been susceptible to the effects of online options that are increasingly reliable and easy to use (Wrigley et al., 2015). In order to remain successful, high-streets and shopping centres – or the stores within them – need a greater understanding of the digital footprints of their customers in order to better engage with them.

In light of this, many major retailers are seeking to harness digital platforms in order to attract customers to physical stores and develop closer relationships with them. However, more robust evidence is needed to understand the effectiveness of adopting on-the-go technologies as a means to boost town centre vitality (Wrigley et al., 2015). Here we present two examples where social media data offer a proxy for the digital engagement of consumers in the UK. First we seek to identify areas of high retail activity based on the content and location of consumers' Tweets. Second we combine Twitter data with an established map of retail centres in the UK to discern their digital footprints. It is hoped that these analyses will be of interest not just to the retailers themselves, but also to local authorities and policy makers seeking to reinvigorate many high streets to better meet the challenge of online retail.

The use of geo-tagged social media data as indicators of human activity has received considerable attention in recent years as researchers and businesses seek out alternative data from which to derive insights into population dynamics. Twitter has been most widely utilised since it benefits from an extensive online community of around 15 million active users within the UK (estimated by Twitter in 2015) and it offers a public application programming interface (API) that enables anyone to request a sample of Tweets according to a particular search criteria. Crucially in the context of geographical analysis, this API provides the location of where the Tweet was sent if a user has consented to revealing such information.

Within the retail sector, the value of social media data are well recognised (McKinsey & Company, 2011), with an estimated 62% of Twitter users following their favourite brands, and major retailers receiving an average of 821 direct mentions and 114 replies per day (Brandwatch, 2015). Retailers frequently use such data to improve brand awareness, listen to customer sentiments and improve customer services (Brennan & Schafer, 2010). However, the geographical components of these data have received less attention. Yet, it is estimated that 80% of Twitter users access the platform via a smartphone (Twitter, 2015) and that 25% use the service whilst shopping (Nielsen Media Research, 2014). Furthermore, Cheng, Caverlee, Lee, and Sui (2011)

* Corresponding author.
E-mail addresses: alyson.lloyd.14@ucl.ac.uk (A. Lloyd), james.cheshire@ucl.ac.uk (J. Cheshire).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.09.006
0198-9715/

found shops and restaurants amongst the top 5 places that people are likely to 'check in' on Twitter. It's clear there is the potential to provide insight into the activities and mobility patterns of consumers within the Twitter population.

We argue that the use of the geographical component of the Tweets will serve to inform knowledge of the digital footprints of online customers and engagement with online platforms. In addition, existing retail centre location data are sparsely available and have been primarily constructed using centroid locations of formerly derived retail cores. For example, commonly used centre and boundary data developed by the Department of Community and Local Government (DCLG) State of the Cities Report were primarily defined using the underlying economic activity and locations of anchor stores in 2004. Therefore, contemporary geographical definitions of centre locations require exploration. Furthermore, the data could provide insights in the study of retail centre catchment analysis, which refers to the areal extent from which the main patrons of a store or retail centre are typically found (Birkin, Clarke, & Clarke, 2010). Notably, there is already a large body of literature exploring retail catchments, which have applied methods such as drive-time (where patrons are expected to go to the closest or most logistically convenient location), or incorporated measures of centre attractiveness into more complex models (i.e. Dolega, Pavlis, & Singleton, 2016). However, there has been little exploration of data-driven methods into flows between retail centre locations.

Despite the range of positive applications, Twitter data have a number of disadvantages that make them inherently hard to interpret and analyse - particularly in relation to their representativeness of the broader population. For instance, it is estimated that 23% of the UK population use the service (Pew Research, 2015) and only an estimated of these 1% opt to share their locations. Still, research has demonstrated that when obtaining these data using Twitter's public API, it is possible to extract over 90% of all geo-tagged posts (Morstatter, Pfeffer, Liu, & Carley, 2013) due to the rate of geo-tagged Tweets roughly corresponding to the limitation rate of the stream. The data are also susceptible to demographic biases such as an over-representation of younger cohorts between 15 and 30 (Longley & Adnan, 2016) and suffer from contribution bias, meaning that a small proportion of users generate a large percentage of the Tweets (Nielsen, 2006). Nevertheless, the data do offer a number of advantages, such as having a high temporal granularity that is international in scale and a selective but numerically large representation (Adnan, Lansley, & Longley, 2013). In addition, data-driven methods can provide benefits in comparison to traditional approaches (such as surveys), as they are able to provide unique information about the social dynamics of places that are not easily and inexpensively obtainable on such a large scale (Li, Goodchild, & Xu, 2013).

In this paper we explore two hypotheses. Firstly, that the content of Tweets would have an identifiable correspondence to locations of retailing activities, and secondly, that the data have the potential to inform the creation of retail catchment areas by evaluating the mobility patterns of Twitter users across different locations. Our motivations were to understand the potential applications and limitations of these data within a retail centre context and place them within the broader framework of promoting town centre resilience in this digital era of online uncertainty.

2. Data treatment

2.1. Twitter data

Tweets were obtained through Twitter's filtered streaming API between December 2012 and January 2014. The Tweet locations were predominantly recorded using the integrated Global Positioning System (GPS) on users' smartphones and typically accurate to within several metres (Li et al., 2013). Whilst the API is assumed to collect these Tweets at random, the methods that Twitter employ to sample these data are currently unknown. In total 99,139,622 Tweets sent by 1,777,873 users were collected. As is common to almost all social media datasets the frequency of Tweets per user was positively skewed, with the most active user sending 68,389 Tweets, yet a median of 7 Tweets per user.

Twitter data can also be susceptible to "bots" that characteristically send multiple spam messages (Hawelka et al., 2014). Therefore, measures were taken to clean the database. For example, one account returned 47,132 Tweets such as *"Entrepreneurs: 5 Reasons Why Your Products are Not Selling - http://t.co/BWovUzLL"*. Other users considered unrepresentative of normal Tweeting behaviour were those who had sent repeated posts to gain followers or attention (i.e. from celebrity accounts), for example, 15,081 Tweets from a single user such as *"@Real_Liam_Payne LIAM, please follow me and we love you so much x1"*. In order to filter such cases, the following procedures were applied:

1. A threshold of 3000 Tweets per user over the duration of the dataset, to avoid the large amount of contribution bias dominating the analysis and to remove prolific spam accounts (Lansley & Longley, 2016).
2. Users who had posted identical messages more than three times, as these were likely to be fake accounts (Wang, 2010).
3. Messages containing 'spam' trigger phrases.

The threshold removed 5,206,922 Tweets from 1032 users (5.25% of the full sample, but only 0.05% of users). Messages with high counts eliminated 236,208 Tweets from 173 users. The 'spam trigger phrases' then aimed to further identify the repeated messages that had been modified to avoid detection. Phrases were obtained from Mequoda (2015) and were edited so that they were relevant for the Twitter data. For example, terms such as 'credit' returned many non-spam messages within the database. However, terms such as 'no obligation', and 'ecommerce' were useful to identify spam content. There were 11 spam terms in total (see Appendix A). A total of 9467 Tweets from 33 users were removed at this stage. This left a final sample of 93,687,025 Tweets from 1,776,635 users for analysis.

2.2. Identifying retail tweets

Interactions with major retailers were identified from the cleaned data. An initial list of 366, major UK retailers were obtained from the IRUK Retailing Top 500 Annual Report (IRUK, 2016), across 11 categories (see Table 1). Only primarily high-street retailers were selected, based on the assumption that these interactions may be spatially representative of retail centre activities. This was extended manually to include as many as possible within the UK, including leisure categories (i.e. food and drink).

Due to the informal nature of Twitter, abbreviated or incorrect spelling variations needed to be accounted for. Common variations could be identified by manually observing the sample and the live Twitter feed. For example, "Marks and Spencer's" had 11 variations that produced relevant Tweets (see Appendix B). However, some major retailers could not be utilised in a general query, such as 'Next' and 'Boots', as there was no way of differentiating between relevant Tweets and general usage. For these only mentions of their official Twitter handle were included (i.e. '@NextOfficial' and '@BootsUK'). The final retailer mentions subset comprised 621,946 Tweets from 277,177 users. Therefore, of the cleaned sample, 15.61% of users had interacted with a retailer, but only 0.66% of Tweets were considered retail related.

2.3. Extracting retail centre tweets

In order to create the Twitter catchments, retail centre location and boundary data were obtained from the Local Data Company (LDC), a commercial research consultancy specialising in retail locations. The data consisted of 1287 location centroids and boundaries that defined retail centre spatial extents (see Fig. 1). These were derived from the underlying economic activity defined by the Department for Communities