Developing a streaming data processing workflow for querying space–time activities from geotagged tweets

Monica Wachowicz a,⁎, M. Dolores Arteaga a, Sangwhan Cha a, Yves Bourgeois b

a People in Motion Lab, University of New Brunswick, Canada
b Urban and Community Studies Institute, University of New Brunswick, Canada

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form 24 November 2015
Accepted 3 December 2015
Available online 30 December 2015

Keywords:
Space–time activity
Twitter
Streaming data processing
Cloud computing

ABSTRACT

The critical dimensions in describing space–time activities are “what”, “where”, “when”, and “who”, which are frequently applied to collect data about basic functions people perform in space in the course of a day. Collecting data about these dimensions using activity-based surveys has presented researchers with a number of technical and social limitations, ranging from the restricted period of time participants have to record their activities to the level of accuracy with which participants complete a survey. This paper proposes a new streaming data processing workflow for querying space–time activities (STA) as a by-product of microblogging communication. It allows exploring a large volume of geotagged tweets to discover STA patterns of daily life in a systematic manner. A sequence of tasks have been implemented using different cloud-based computing resources for handling over one million of daily geotagged tweets from Canada for a period of six months. The STA patterns have revealed activity choices that might be attributable to personal motivations for communicating an activity in social networks.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Space–time activities are a sequence of context dependent functions describing when people allocate their time for doing things, where they spend it, and with whom. Previous efforts in developing techniques for capturing space–time activity (STA) data have focussed on developing activity-based surveys to obtain accurate and comprehensive information that is relevant to transportation, socio-economic, and urban studies (Miller, 1991; Kwan, 1998; Jonsson, Karlström, Oshyani, & Olsson, 2014). In spite of several decades of research the capture of STA data continues to be problematic. The main challenge is to develop an approach for retrieving what any person is doing anywhere at any time for a long period of time (Crosbie, 2006; Harvey & Spinney, 2011).

Our underlying research premise is that microblogging communication can be used as a new source of STA data because a large population can post any message about their allocation of time. In particular, the Twitter social network has become a way of massive communication with its large and international user base of approximately 284 million active users sending 100,000 messages every minute (Twitter Statistics, 2014). Twitter offers functionalities to send original 140-character messages, and also reply to and retweet (share) any message posting. It also allows a user to follow other target Twitter users and send out messages automatically to them. Approximately 77% of accounts are outside the United States (Twitter Statistics, 2014), and they have produced flows of information that have been proven to be valuable to researchers as they have been the data source of scholarly research including discovering geographic topics (Hong et al., 2012), detecting epicenters of earthquakes (Earle, Bowden, & Guy, 2012), tracking human mobility (Azmadian, Singh, Gelsey, Chang, & Maheswaran, 2013), community happiness (Quercia, Ellis, Capra, & Crowcroft, 2012) and characterizing population-scale measures such as obesity rates (Mitchell, Frank, Harris, Dodds, & Danforth, 2013).

Geotagged tweets offer the opportunity to collect a larger volume of STA data than previous existing activity-based surveys. However, they represent a challenging source of STA data because of the connection between daily activities and a spatial structure. They are unquestionable as a data source about the type of activity (“what”) and the individual person involved in this activity (“who”). However, there is a level of uncertainty related to “when” and “where” the activity is carried out in view of the fact that Twitter users may post the message at the exact place and time where an activity is taking place. Therefore, the main contribution of this paper is two-fold: first, to develop an approach for streaming a large volume of geotagged tweets and allowing an automated collection of STA information; and second, to investigate the uncertainty of the “when” and “where” dimensions of STA patterns.

Current GIS is not set up for streaming, storing and querying unstructured text of millions of geotagged tweets sent every day (Croitoru, Crooks, Radzikowski, & Stefanidis, 2013, Steiger, Albuquerque, & Zipf, in press). Therefore, we have designed a cloud-
based platform based on the Hadoop ecosystem, which handles not only the large volume of geotagged tweets but also the STA query using the Stanford natural language processing library. The new streaming data processing workflow consists of three tasks that have been developed for data ingestion, data management, and data querying.

The remainder of this article is structured as follows. Section 2 describes previous research work on activity-based surveys and cloud computing. A detailed description of our streaming data processing workflow is provided in Section 3. Section 4 explains the cloud computing implementation of such a workflow and the STA patterns found in Canada for a period of six months. The article concludes with assessing the implications of “what”, “where”, “who” and “when” dimensions in finding STA patterns in microblogging communications.

2. Related work

Space–time activity dimensions are typically described as “what”, “where”, “when”, and “who”. Miller (2005) distinguishes four types of activity-based surveys developed for collecting data about these dimensions. Recall methods require participants to remember the activities that happened in the course of their day and report them during a specified time period. Stylized recall methods require participants to select the activities that happened during some typical time period. Diary methods require participants to record activities in a diary, either in a free-format manner or at pre-determined time periods. Prospective methods are typically game-based and are employed in conjunction with other methods to investigate the effect of potential changes on the activity environment. Amongst these methods, the self-administered activity diary has been the main method used for capturing STA data, often focusing on using a specific coding scheme of activities relevant to the research study at hand, such as household activity patterns (Gan & Recker, 2013) or personal travel diaries (Lyons & Kenyon, 2003).

The Multinational Time Use Project has been proven to be the most functional and extensive collection of STA data from the general population at an international level (Szalai, 1966). An activity diary was designed for participants to reply only two questions: “What did you do?” (Primary Activity) and “And what else?” (Secondary Activity). One of these project outcomes was the design of a coding scheme for harmonizing diary data from twenty-four countries over several decades. However, a very limited pre-coded activity list of 40 categories was developed due to the challenges of achieving a taxonomic harmonization. Second, the disparities in the production of diaries across the different days of the week require several pre-processing steps in order to manage the activity data sets. And finally, the accessibility of these data sets varies widely in terms of countries. Nevertheless, two recommendations described in this project report are very relevant in handing STA data using current technologies such as microblogging communications. They are: (a) activities should be reported by participants in their own words and free form for later coding, and (b) activities should be collected at a national level. Our research work has followed these recommendations for developing a query to retrieve space–time activities from Twitter posts.

Activity-based surveys have also been made available in personal digital devices coupled with GPS for improving the STA data recording in terms of reducing and under-reporting the activities and avoiding possible positioning errors (Doherty, Papinski, & Lee-Gosselin, 2006; Bohle & Maat, 2009). However, a vast majority of these surveys have also used a pre-coded list of activities. One example includes a three-tier ontology-based approach for designing a spatio-temporal scheduler of daily people’s activities as a sequence of events in combination with their expected locations (Stewart, Fan, & White, 2013). The coding scheme contains only the activities that are scheduled and associated with a university campus, for example, making deliveries to campus buildings or managing meetings. The approach is promising to improve accessibility to buildings by maximizing connectivity between scheduled activities.

Another example is the study conducted by McKenzie, Deutsch, and Raubal (2013) that combines an activity diary and a microblogging network to explore how space–time activities in the physical world are reported on social media. The study lasted for three months, and twenty-eight participants have recorded the type, location and time frames of their pre-coded daily activities using their Facebook’s accounts. The Chapin coding scheme (Chapin, 1974) was used to categorize the activities from the self-administered activity diaries as well as the activities posted in Facebook. Although GPS was not used for locating the activities, the results suggest that space–time activities in the real-world contribute to on-line reported activities, in particular, when they occur with less frequency.

Despite these efforts, the main limitations still continue to be the restricted number of participants, the use of pre-coded list of activities, and the short period of time the surveys are conducted, usually keeping the length of a survey to a minimum in order to increase the response rate. The key contribution of this paper is to tackle these limitations by proposing a streaming data processing workflow to query a large volume of geotagged tweets in a systematic manner at a national level and for a long period of time, ranging from hours and days to months and years. Twitter as a new source of STA data will provide a free form of reporting the “what” and “who” dimensions in a meaningful way.

Several efforts can be found in the literature to find the “what” dimension in Tweet posts for applications such as user profiling, content recommendation and topic tracking. Most of the proposed approaches combine topic modeling with geographical diversity. The models are based on text processing libraries for classification, tokenization, stemming, tagging, and parsing (Steiger et al., in press), Wang, Wang, Xie, and Ma (2007) proposed a Latent Dirichlet Allocation-style probabilistic graphical model to explicitly represent the relationships between location labeling and words found in a tweet. Moreover, Hong et al. (2012) address the problem of modeling geographical topical patterns on Twitter by introducing a Sparse Additive Generative Model with zero-mean Laplace distributions. These models usually have complex inference mechanisms that lead to the problem of over-parameterization in order to accurately perform the inference. We propose a space–time query that is both flexible enough to use text processing libraries for tagging geotagged tweets and search for the “gerund verb” to discover space–time activities. By avoiding the complexity of previous models, our approach scales up to millions of geotagged tweets.

The representativeness of the “who” dimension in Twitter has been discussed by Hale, Gaffney, and Graham (2012) who argue that the division of geotagged tweets and non-geotagged tweets is almost certainly biased by factors such as socio-economic status, location and education, and as a result, geotagged tweets are unlikely to form a representative sample of the broader universe of the population. Currently very few research efforts can be found in the literature in dealing with this issue. Some preliminary results show that there are times when the Twitter data is representative and times when it is biased, especially in sparse data situations and for applications that require to collect the same data but from different regions (Morstatter, Pfeffer, & Liu, 2014). Gupta et al. et al. (2014) propose the use of machine learning to address fake messages by evaluating the credibility of Twitter messages. More approaches are expected to be developed in the near future.

The “where” and “when” dimensions of space–time activities cannot be extracted directly from geotagged tweets. Only a small portion of tweets have latitude and longitude coordinates of either the exact location of the device or an approximate location specified by a set of coordinates of a Minimum Boundary Rectangle (MBR). The coordinates and a timestamp are usually obtained from the GPS receiver of a user’s device itself or the user’s Internet Protocol (IP) address. They represent the user’s true physical location at the time a geotagged tweet has been sent. Therefore, they are the proxy location of a space–time activity. Previous research has already shown the benefit of using proxy data
دانلود مقاله

http://daneshyari.com/article/506269

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات