Computers, Environment and Urban Systems 51 (2015) 70–82

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

A scalable framework for spatiotemporal analysis of location-based social media data

Guofeng Cao [a,*], Shaowen Wang [b,c,*], Myunghwa Hwang [b], Anand Padmanabhan [b,c], Zhenhua Zhang [b], Kiumars Soltani [b]

[a] Department of Geosciences, Texas Tech University, Lubbock 79409, TX, USA
[b] Cyberinfrastructure and Geospatial Information Laboratory, Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana 61801, IL, USA
[c] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana 61801, IL, USA

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 8 January 2015
Accepted 8 January 2015

Keywords:
Big data
CyberGIS
Data cube
OLAP
Social media

ABSTRACT

In the past several years, social media (e.g., *Twitter* and *Facebook*) has experienced a spectacular rise in popularity and has become a ubiquitous location for discourse, content sharing and social networking. With the widespread adoption of mobile devices and location-based services, social media typically allows users to share the whereabouts of daily activities (e.g., check-ins and taking photos), thus strengthening the role of social media as a proxy for understanding human behaviors and complex social dynamics in geographic spaces. Unlike conventional spatiotemporal data, this new modality of data is dynamic, massive, and typically represented in a stream of unstructured media (e.g., texts and photos), which pose fundamental representation, modeling and computational challenges to conventional spatiotemporal analysis and geographic information science. In this paper, we describe a scalable computational framework to harness massive location-based social media data for efficient and systematic spatiotemporal data analysis. Within this framework, the concept of space–time trajectories (or paths) is applied to represent activity profiles of social media users. A hierarchical spatiotemporal data model, namely a spatiotemporal data cube model, is developed based on collections of space–time trajectories to represent the collective dynamics of social media users across aggregation boundaries at multiple spatiotemporal scales. The framework is implemented based upon a public data stream of *Twitter* feeds posted on the continent of North America. To demonstrate the advantages and performance of this framework, an interactive flow mapping interface (including both single-source and multiple-source flow mapping) is developed to allow real-time and interactive visual exploration of movement dynamics in massive location-based social media data at multiple scales.

1. Introduction

Social media represents "a group of Internet-based applications that are built on the ideological and technological foundations of web 2.0, and that allow the creation and exchange of user generated content" (Kaplan & Haenlein, 2010). Typical examples include Twitter, Facebook, Foursquare, and Flickr. In recent years, these online applications have attracted hundreds of millions of users for everyday social networking and content sharing while also collecting a huge amount of user-generated social media data (e.g., text messages, photos, videos, and structures of social relationships). Twitter, for example, has grown at an exponential rate since its founding. As of December of 2013, the monthly number of active Twitter users has surpassed 3.9% of the global population and 17.9% of the United States, and they have sent more than 300 billion so-called *tweets* (individual user posts). On another front, with the widespread usage of smart mobile devices and location-based services, location-aware mobile devices have become prevalent access points to social media services. From the perspective of geographic information science (GIScience), hundreds of millions of smartphone users could be viewed as ubiquitous "citizen sensors" that move in geographic spaces, sensing and sharing the surrounding environment using various social media features. The inclusion of location or spatial dimensions blurs the interface between the cyberspace of social media and the geographic space of the real world (Tsou & Leitner, 2013), and together with the

* Corresponding authors at: Department of Geosciences, Texas Tech University, Lubbock 79409, TX, USA (G. Cao), Cyberinfrastructure and Geospatial Information Laboratory, Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana 61801, IL, USA (S. Wang).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.01.002
0198-9715/

temporal dimension, makes social media a promising proxy for understanding the social dynamics in geographic spaces.

With accesses to fine-grained social media footprints at the individual level, location-based social media data provide a new set of lenses and tremendous opportunities to examine complex social dynamics. In behavioral sciences, for example, massive individual geo-tagged social media records can be used to study human activity (e.g., mobility) patterns (e.g., Backstrom, Sun, & Marlow, 2010; Sadilek & Krumm, 2012), and the effects on human life (Frank, Mitchell, Dodds, & Danforth, 2013). At an aggregate level, a careful aggregation of social media footprints for a subpopulation (e.g., a geographic region) could lead to a better understanding of this subpopulation (Cranshaw, Schwartz, Hong, & Sadeh, 2012; Li & Goodchild, 2012) and their connections with others (e.g., Wu, Zhi, Sui, & Liu, 2014). In public health surveillance, studies have shown that, for certain diseases (e.g., influenza), a careful analysis of geo-located Twitter messages could provide surveillance capabilities comparable to the reports of official surveillance agencies but in a much more timely manner (e.g., Nagel et al., 2013; Signorini, Segre, & Polgreen, 2011).

Initial successes in exploiting location-based social media data demonstrate the great potential to gain new scientific insights. Distinct characteristics of location-based social media data, however, pose fundamental representation, modeling and computational challenges to GIScience, spatiotemporal databases and spatiotemporal analysis. As described in Wang, Cao, Zhang, and Zhao (2013), location-based social media data generated by a massive number of social media users are often big and produced continuously at an ever faster rate. Evidently, location-based social media data and other user-generated geospatial contents are becoming an important contributing source of big data (Manyika et al., 2011). While GIScience is shifting rapidly to embrace the data-intensive and computation-intensive paradigm (Wang, 2010; Wright & Wang, 2011), the big data nature of location-based social media is well beyond the capability of mainstream geographic information systems (GIS). Furthermore, the dynamic and real-time characteristics of social media data hinder direct applications of conventional GIS, which tends to represent the real world as static forms instead of dynamic processes (Goodchild, 2004). In addition, social media contents are usually produced as unstructured forms of media (e.g., texts, photos and videos) in contrast to the well-structured, ready-to-use geospatial data sources. Extra efforts, such as data retrieval and data mining processes, are often necessary to make the data meaningful and practical.

To address these challenges, this paper presents a scalable computational framework to harness massive location-based social media data to support systematic and efficient analysis of spatiotemporal dynamics. In the presented framework, location-based social media data are firstly regularized in terms of *space–time trajectories or paths* to represent the activity profile of each social medial individual. To exploit the unstructured contents of social media, specific data mining methods can be plugged into the described framework to gain valuable information of specific interests. As a particular example, this paper examines the chance of influenza like illness (ILI) infection by monitoring the text messages of Twitter posts. Within the context of data warehousing and online analytical processing (OLAP) (Inmon, 2005), a data cube model for space–time trajectories is designed, constructed and regularly maintained to support systematic and efficient spatiotemporal analysis of massive location-based social media data. Specifically, this data cube frames the spatiotemporal dynamics of location-based social media in a multidimensional space (or a cube) of location, time and social media users, and decomposes this multidimensional space (cube) into a multi-scale, hierarchical structure of *cuboids*. A set of measures that characterize the spatiotemporal dynamics of location-based social media is specifically defined for each cuboid (e.g., number of social media users and activities) and each pair of cuboids (e.g., number of travels from one cuboid to another) of the data cube. The cuboids and associated measures can be flexibly merged or split according to the dimensional intervals of interest (e.g., administrative boundaries). With the data cube model decomposed into arrays of cuboids, one can exploit the collective spatiotemporal dynamics in particular regions of interest at multiple levels of spatiotemporal scales (scale effects) and different aggregation boundaries (zoning effects) in a very efficient manner. The presented framework thus transforms the massive, dynamic and unstructured location-based social media data into flexible geospatial datasets that could be easily compatible with the high performance analytical environment of cyberGIS (Wang, 2010) and the typical work-flows of conventional GIS analysis. Implementation details of the framework are described based on open access to a Twitter post stream. An online visual analytical interface, including single-source and multiple-source flow mapping, is developed to allow near real-time, interactive visual exploration of multiple scales of distribution and movement dynamics in massive location-based social media data.

The remainder of this paper is laid out as follows. Key concepts of data representation, particularly space–time trajectories, are first introduced in Section 2. Section 3 introduces the spatiotemporal data cube model for efficient analysis of location-based social media data. Based on a public data stream of *Twitter* feeds posted on the continent of North America, the implementation details of the presented framework are discussed in Section 4. In Section 5, the online flow mapping interface is introduced and demonstrated to showcase the advantages and effectiveness of the proposed framework. Section 6 summarizes the paper and discusses future work.

2. Space–time trajectories

Consider a set of N individuals frequently sharing their activities (e.g., message posts and check-ins) through a location-based social media platform that exhaustively collects activities of users. To ease the privacy and security concerns of individual users, we suppose that the location-based social media platform is designed to collect these data anonymously, that is, the social media platform is unaware of the identities of individual users, and no names or other personal identifiers are shared. Each individual is assumed to move continuously in geographic spaces, either freely in a Euclidean space or restrictedly in a regularized network space, e.g., roads, railways, or airways, and frequently share messages via social media channels.

The concept of *space–time paths or trajectories* has long been used as a simple and effective means for representing and characterizing human mobility pattern (Hägerstraand, 1970) and spatial trajectory analysis (Zheng & Zhou, 2011). In this paper, we assume that each user $u_{id}(id \in [1, N])$ corresponds to a continuously moving, lifetime-long space–time trajectory T_{id} in a geographic space. This "true" trajectory T_{id} is measured and approximated by TS_{id}, a series of footprint tuples of location (s_{id}), timestamp (t_{id}) and message content (m_{id}) in social media, i.e.: $TS_{id} = \{(s^0_{id}, t^0_{id}, m^0_{id}), (s^1_{id}, t^1_{id}, m^1_{id}), \ldots, (s^i_{id}, t^i_{id}, m^i_{id}), \ldots\}$, where $t^0_{id} \leqslant t^1_{id} \leqslant \ldots t^i_{id} \leqslant \ldots$. Different from conventional trajectories of moving objects (Zheng & Zhou, 2011) where measurements are often abundant and sampled at regular time intervals, measurements for trajectories of location-based social media TS_{id} (i.e., user activities) are often very temporally sparse and irregular (Gao & Liu, 2013). Inactive social media users could have long sedentary periods before their next social media activities, and yet due to privacy concerns, users can choose to disable location options when posting activities. Consequently, the intermediate positions between measurements on TS_{id} cannot be reliably reconstructed by commonly used methods