Computers, Environment and Urban Systems 66 (2017) 23–37

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data

Qunying Huang [a,*], Guido Cervone [b], Guiming Zhang [b]

[a] Department of Geography, University of Wisconsin-Madison, Madison, WI, 53706, United States
[b] Department of Geography and Institute for CyberScience, Pennsylvania State University, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 3 September 2015
Received in revised form 22 February 2017
Accepted 22 June 2017
Available online xxxx

Keywords:
Disaster coordination and relief
Disaster management

ABSTRACT

Social media streams and remote sensing data have emerged as new sources for tracking disaster events, and assessing their damages. Previous studies focus on a case-by-case approach, where a specific event was first chosen and filtering criteria (e.g., keywords, spatiotemporal information) are manually designed and used to retrieve relevant data for disaster analysis. This paper presents a framework that synthesizes multi-sourced data (e.g., social media, remote sensing, Wikipedia, and Web), spatial data mining and text mining technologies to build an architecturally resilient and elastic solution to support disaster analysis of historical and future events. Within the proposed framework, Wikipedia is used as a primary source of different historical disaster events, which are extracted to build an event database. Such a database characterizes the salient spatiotemporal patterns and characteristics of each type of disaster. Additionally, it can provide basic semantics, such as event name (e.g., Hurricane Sandy) and type (e.g., flooding) and spatiotemporal scopes, which are then tuned by the proposed procedures to extract additional information (e.g., hashtags for searching tweets), to query and retrieve relevant social media and remote sensing data for a specific disaster. Besides historical event analysis and pattern mining, the cloud-based framework can also support real-time event tracking and monitoring by providing on-demand and elastic computing power and storage capabilities. A prototype is implemented and tested with data relative to the 2011 Hurricane Sandy and the 2013 Colorado flooding.

1. Introduction

Every year extreme weather and climate events, such as cyclones, floods, tornados and geological events such as volcanic eruptions, earthquakes or landslides, claim thousands of lives and cause billions of dollars of damage to property and severely impact the environment (Velev & Zlateva, 2012). Disasters and their effects have been increasing both in frequency and severity in the 21st century because of climate change, increasing population and their reliance on aging infrastructure. In fact, the first decade of the 21st century witnessed 3496 natural disasters including floods, storms, droughts and heat waves, nearly five times as many disasters as the 743 catastrophes reported during the 1970s.[1] Therefore, an urgent need exists to understand spatiotemporal patterns and the general dynamics that contribute to the occurrences of disasters. These combined studies are necessary to develop effective strategies to mitigate their destructive effects, and to respond and coordinate efficiently to protect people, properties and the environment.

Social media have been primarily used as an intelligent "geo-sensor" network to detect extreme events and disasters such as hurricanes and earthquakes, and to gain situational awareness for emergency responders and relief coordinators during crises by monitoring and tracking citizens feedbacks (Sutton, Palen, & Shklovski, 2008). Additionally, they are widely used by scientists to study public risk perception, and people's reactions during disasters (Mandel et al., 2012). On the other hand, remote sensing data are paramount during disasters and have become the de-facto standard for providing high resolution imagery for damage assessment and the coordination of disaster relief operations (Cervone et al., 2016; Cutter, 2003; Joyce, Belliss, Samsonov, McNeill, & Glassey, 2009). Using high resolution imagery from commercial and research air- and space-borne instruments, it is possible to obtain data within hours of major events, frequently including 'before' and 'after' scenes of the affected areas (Cervone & Manca, 2011). These 'before' and 'after' images are quickly disseminated through scientific portal and news channels to assess damage and inform the public. In addition, first responders rely heavily on remotely sensed imagery for coordination of relief and response efforts as well as the prioritizing of resource allocation.

* Corresponding author.
E-mail addresses: qhuang46@wisc.edu (Q. Huang), cervone@psu.edu (G. Cervone), gzhang45@wisc.ed (G. Zhang).

[1] http://www.theguardian.com/environment/blog/2014/jul/14/8-charts-climate-change-world-more-dangerous

http://dx.doi.org/10.1016/j.compenvurbsys.2017.06.004
0198-9715/

Despite the wide availability of large remote sensing datasets from numerous sensors, specific data might not be collected in the time and space most urgently required. Geo-temporal gaps result due to satellite revisit time limitations, atmospheric opacity, or other obstructions. Recently, stream data from social media and remote sensing are being fused for disaster analysis and assessment. Specifically, social media are used to fill in the gaps when remote sensing data are lacking or incomplete (Schnebele & Cervone, 2013; Schnebele, Cervone, Kumar, & Waters, 2014; Schnebele, Oxendine, Cervone, Ferreira, & Waters, 2015).

However, current studies on using social media and remote sensing data for disaster analysis are performed on a case-by-case basis. The approaches typically start with identifying a specific disaster event, and then filters (e.g., keywords, spatiotemporal information) are designed to select and retrieve relevant stream data. These efforts are time-consuming. For example, identifying the tweet hashtags associated to a specific event, may take from hours to days for manual examination of hundreds of tweets to include relevant hashtags so we can use them to filter out non-relevant tweets during a disaster. Furthermore, these efforts need to be duplicated when analyzing a different event. As stated earlier, it is necessary to complete a comprehensive database that can display the historical events with relevant metadata (e.g., event type, severe category, damages, locations, and temporal spans) to allocate resources for analysis. Additionally, from the basic metadata, it is also needed to automatically derive relevant information (e.g., hashtags), which can then be used to retrieve relevant messages from long-term accumulated social media.

With multi-sourced data streams from a multitude of channels, identifying authoritative sources and extracting critical, validated messages information can be quite challenging, especially during a crisis. The volume, velocity, and variety of accumulated stream data produce the most compelling demands for computing technologies from big data management to technology infrastructure (Huang & Xu, 2014). To address these big data challenges, various types of computational infrastructures are designed, from the traditional cluster and grid computing to the recent development of cloud computing and CPU/GPU heterogeneous computing (Schadt, Linderman, Sorenson, Lee, & Nolan, 2010). Specifically, cloud computing has been increasingly viewed as a viable solution to utilize multiple low-profile computing resources to parallelize the analysis of massive data into smaller processes (Huang & Cervone, 2016).

This paper addresses these problems by proposing a novel system to support both historic disaster event analysis and upcoming event monitoring. Wikipedia is exploited as a source to build a disaster event database, which is then applied to retrieve relevant information for a specific disaster from massive social media data accumulated daily. Cloud computing is proposed to serve as the underlying infrastructure that offers the capability of providing on-demand and flexible computing resources to meet the dynamic computing requirements of real-time disaster analysis. The following contributions are made in this research:

1. An integrated system framework is proposed for historical disaster analysis based on multi-sourced data with limit, if any human interaction. To analyze and understand the public behaviors or reactions captured by social media data, our system does not rely on human identification of filtering criteria to retrieve relevant messages. An automatic system based on text mining, and geocoding technologies are developed to derive these information.
2. An event database is built based on Wikipedia. Such a database is useful for scientists easily selecting a relevant event for analysis or selecting disasters of a specific type to identify their patterns, and linking it to other GIS data (e.g., socioeconomic data), climate data, and environment data to understand the driving factors that contribute to the occurrences of these disaster events.
3. Within the proposed system, cloud computing is used as the underlying infrastructure to provide flexible computing power to address the computing challenges posed by the massive data processing and a real-time operational system for emergencies response and disaster coordination. Such a system is suitable for online services and systems where a number of texts, and remote sensing images are dynamically streaming.
4. A prototype is implemented, and recent flooding events are used as a case study to demonstrate the feasibility of the proposed system.
5. This paper provides a general methodology that it is not event specific, and can be used both for retrospective analysis and for real time monitoring and decision making. The proposed framework sheds light on integrating various emerging data sources to support scientific applications of significant interests that go beyond disaster management.

2. Related work

2.1. Social media for disaster management

As social media applications are widely deployed in various platforms from personal computers to mobile devices, they are becoming a natural extension to human sensory system. The synthesis of social media with human intelligence has the potential to be the intelligent sensor network that can be used to detect, monitor and gain situational awareness during a hazard with unprecedented scale and capacity. By monitoring tweets, for example, an earthquake can be detected by developing a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location (Sakaki, Okazaki, & Matsuo, 2010).

By mining social media data, it is possible to establish situation awareness for disaster response and relief (Ashktorab, Brown, Nandi, & Culotta, 2014; Gao, Barbier, & Goolsby, 2011; Huang & Xiao, 2015; Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013; Kumar, Barbier, Abbasi, & Liu, 2011). Using Hurricane Sandy as an example, Huang and Xiao (2015) coded social media messages into different themes within different disaster phases during a time-critical crisis, and a classifier based on logistic regression is trained and used for classifying the social media messages into various topic categories during various disaster phases. Imran et al. (2013) extracted information from disaster-related messages posted on Twitter into several categories including warnings, casualties and damage, donations, and information sources. The coded information can be further analyzed over space and time to inform the situational awareness of the incidences as they unfold. The Australian Government developed an Automated Web Text Mining (ESA-AWTM) system that analyzes Twitter messages to provide incidence identification, near real-time notification, and monitoring (Cameron, Power, Robinson, & Yin, 2012). A web application, "TweetTracker", was developed by Kumar et al. (2011) to track, analyze, and monitor tweets for disaster relief. It can report separately geo-referenced and non-geo-referenced tweets, support keyword search, and generate and display trends of keywords specified by the user.

However, all published methods rely on a case-by-case analysis of historical events, or support simple real-time data searching and analysis functions (Kumar et al., 2011). There is no systematic approach proposed to support both historical event and real-time event tracking capabilities.

2.2. Synthesizing multi-sourced data for disaster management

In a time of disaster, multi-sourced data can be integrated to assess the situation. Such integration results in new approaches to support disaster management in a new way that cannot be done previously and significantly improve the analysis and capability of a single data source. For example, while social media data have been successfully used to detect and track locations of disaster events (e.g., earthquake, tornado, and wildfire; Sakaki et al., 2010; De Longueville, Smith, & Luraschi, 2009; Jain, 2015), disaster detection is not always possible using single-sourced data (e.g., tweets) alone and there is a need to integrate