

DisasterMapper: A CyberGIS framework for disaster management using social media data

Qunying Huang¹, Guido Cervone², Duangyang Jing¹, Chaoyi Chang¹

¹Department of Geography
University of Wisconsin-Madison, Madison, WI 53706
{[ghuang46](mailto:ghuang46@wisc.edu), [djing3](mailto:djing3@wisc.edu), [cchang45](mailto:cchang45@wisc.edu)}@wisc.edu

²Geoinformatics and Earth Observation Laboratory
Department of Geography and Institute for CyberScience
The Pennsylvania State University, University Park, PA 16802
cervone@psu.edu

ABSTRACT

Traditional GIS tools and systems are powerful for analyzing geographic information for various applications but they are not designed for processing dynamic streams of data. This paper presents a CyberGIS framework that can automatically synthesize multi-sourced data, such as social media and socioeconomic data, to track disaster events, to produce maps, and to perform statistical analysis for disaster management. Within our framework, Apache Hive, Hadoop, and Mahout are used as scalable distributed storage, computing environment and machine learning library to store, process and mine massive social media data. The proposed framework is capable of supporting big data analytics of multiple sources. A prototype is implemented and tested using the 2011 Hurricane Sandy as a case study.

Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioral Sciences

General Terms

Human Factors, Experimentation

Keywords

Disaster management, social media, social networks, Big Data, GIS

1. INTRODUCTION

GIS plays an important role in disaster management by producing various hazard maps to support different activities, operations and decision making processes. For example, first responders and emergency staff on the ground rely on street maps for search and rescue operations. During emergencies resulting from disasters (e.g., by floods or earthquakes), decision makers need to use “situational awareness” maps for understanding areas that are

either damaged or likely to become so, mapping road closures and access restrictions, identifying locations of shelters, kitchens, and medicine distribution points, and the locations of environmentally hazardous sites. Similarly, disaster managers can use hazard maps for risk identification and assessment.

Traditionally, the maps used to deal with emergencies and support decision making were primarily created by using static GIS data, such as topography, land use, demographic and socioeconomic conditions (e.g., poverty and education), statistical information on damage and loss caused by historical disasters, and information about existing infrastructure (e.g., housing conditions, public facilities) [1]. Depending on the specific disaster type, other relevant data may be incorporated. For example, for weather-induced disaster, hydrological information and meteoroidal data are also used [2, 3]. However, the information of real-time damage and resource demand (e.g., water, food, and medicine), which can augment our understanding of the overall disaster situation, and facilitate the decision-making towards a better response strategy, cannot be effectively integrated since such information is difficult to obtain due to limitations in data acquisition and techniques in processing the data efficiently in near real time.

With the massive popularity of social networks and their real time production of data, social media streams have emerged as a new source for disaster management. For instance, social media networks have even become widely used as an intelligent “geo-sensor” network to detect and monitor extreme events or disasters such as earthquakes [4, 5]. Such intelligent sensor webs are useful for operations such as event detection, where timely information is most valuable and hard to obtain [6, 7]. Additionally, it has been widely acknowledged that Humanitarian Assistance and Disaster Relief (HA/DR) responders can gain valuable insights and situational awareness by monitoring social media-based feeds from which tactical, actionable data can be mined from text [8-12].

While these novel streams offer new opportunities during emergencies, using traditional GIS techniques it is impractical, if not impossible, to handle such massive heterogeneous, real-time datasets. This limitation is mainly due to traditional GIS tools and systems being designed and implemented to work on standalone computers to process small-sized and static data. There are a number of limitations associated with this approach: the data storage is normally constrained; the data model and management are formatted yet incapable of managing unstructured data and inflexible for expansion; the data processing is pronounced yet its

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data 2015, Seattle, WA, USA

Copyright 2015 ACM 978-1-4503-3974-2...\$15.00

capabilities are restricted to spatial components (locations) and associated attributes of geographic information, but limited to handling text messages. Specially, social media data pose several grand challenges when integrated with a traditional GIS system:

- **Volume:** Social media are streams of data often massive in volume. The volume and data rate leave a conundrum for traditional GIS on standalone machines. For example, just ingesting global daily tweets for a single year generates such a magnitude of data that is well beyond the capability of any mainstream GIS on a commercial standalone computer [13]. Analyzing this volume of data is not feasible for traditional GIS because of the computational requirements involved that go beyond what a typical workstation can support.
- **Velocity:** Social media data are generated dynamically and continuously. Users of social media services at different locations frequently update their status and post messages and pictures online, some of which can be shared (e.g., retweeted on the Twitter) exponentially. These real-time data, complemented with official and authoritative data sources, are valuable and require an adaptable system to dynamically adjust to the different data rates, including peaks during which the volume of the data can greatly exceed the normal traffic [14]. A traditional GIS, however, are not capable to adjust to such dynamic processes. It simply lacks the scalable computing infrastructures to handle streaming social media data. As a consequence, the GIS framework must be extended to cope with this dynamic information flow.
- **Variety:** Compared to well-structured data formats in traditional GIS, such as vector data and raster data, social media data are often produced in unstructured forms. Some posts can be extremely short in content and small in size while others may contain images, videos, and external links that are rich in content and relatively large in size [15]. Given this asymmetry of data flow, conventional GIS data structures and storage systems can be inefficient in storing the data, causing a waste of resources and potentially slowing down data retrieval. In addition, to study trends and processes hidden in these social media streams, additional data mining processes are often necessary because the text include lexicon and linguistic patterns that are fragmented and unstructured. Traditional GIS is unable to provide the analytical capability to query and process spatiotemporal events and understand their patterns. Furthermore, unlike the conventional GIS data that are well organized and presented at certain government websites (e.g., Census¹), social media data often require advanced processing for data collection, such as designated accesses at multiple time points and regions via provided application programming interfaces (APIs) [16]. Issues of uncertainty and noise in accessing the data further hinder the collection process.

In the geospatial fields, progress has been made to archive, manage, index, process and analyze the unstructured social media data, such as messages from Twitter and Google+, check-in's from Foursquare, videos from YouTube, and photos from Flickr, in a data driven framework for different geospatial applications [17, 18]. In these studies, each social media data entry is treated as a spatiotemporal (ST) point, and traditional spatial data analysis algorithms (e.g., spatial clustering), statistics, and geovisualization

methods are applied to understand the patterns of these massive ST points. Simple text processing analysis, such as using text match based approach and sentiment analysis, may be applied on the text content to filter out non-relevant messages and to understand public behavior and perceptions towards a specific event [17]. However, most of the time, text components of social media are mostly ignored or under studied. In fact, much useful information in social networks such as Twitter is textual as people post text messages related to damage, offering help, requesting services, fundraising or donation. From the text, "actionable data" can be harvested and extracted to establish situational awareness or produce a crisis map [7, 12]. As a result, many computer and information scientists are attempting to extract information from disaster-related messages for emergency coordination and disaster relief using machine learning techniques [8-12]. However, in these studies, locations of social media data are not well considered to establish geographic situational awareness (GSA).

According to Huang and Xiao (2005), GSA is defined as knowing what is happening in the ambient space during an event. Users with location services enabled on smart mobile devices can post content (e.g., text messages or photos) with geo-tags (locations), where an event occurred. The locations along with the place names mentioned in the content text are then used to identify the areas of damaged infrastructure, affected people, evacuation zones, and the communities who are in need of aid resources. Therefore, in order to establish GSA, both content and locations of social media data must be considered.

Faced with this challenge, this paper aims to present a CyberGIS based framework that can synthesize multi-sources data, spatial data mining, text mining, geovisualization, big data management, and distributed computing technologies in an integrated environment to support disaster management and analysis. This proposed framework uses Apache Hive as a scalable distributed storage to ingest and archive massive amounts of social media data. Due to its key characteristics of reliability, flexibility, cost, and scalability, Hadoop is employed to process and analyze archived social media data to make it usable. Apache Mahout, an open-source library that implements scalable machine learning algorithms, is used as the underlying library to support big data analytics. The proposed solution is very fast because of its seamless integration with other popular open-source Apache libraries, such as Hadoop and Lucene (a high-performance, full-featured text search engine library). Mahout has been widely used to perform various text mining tasks, such as grouping together similar documents by using various clustering algorithms [12]. Various geovisual tools are also developed and presented in a web interface that allows users to customize analysis and view multi-sourced data in different types of maps and plots for disaster management.

2. A CYBERGIS FRAMEWORK

Figure 1 shows the developed framework to integrate multi-sourced data for disaster management. Overall, there are three steps or modules in our framework, including data collection, data process, and data geovisual analysis. Below we elaborate each module in detail.

¹ <http://www.census.gov/>

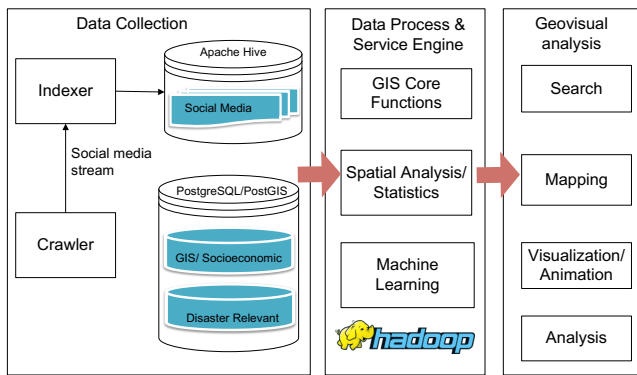


Figure 1. Design of the CyberGIS framework for disaster management using multi-sourced data

2.1 Data collection

Twitter publishes real-time tweet stream through open APIs. By registering an account and applying for access keys, third parties can receive tweets in real-time. In general, Twitter allows to collect about 1% of the daily available tweets. As Twitter only allows users limited access to historical data, tweets have to be archived in a local database. Therefore, the proposed approach starts with archiving data to be processed for future analysis. For the data collection, a crawler based on Twitter4j² is implemented to collect geo-tagged tweets.

Within this framework, an unstructured database, Apache Hive³ is used to store the datasets. MapReduced based systems have emerged as a new computing paradigm for massively parallel data process [19]. Hadoop, the most widely used implementation of MapReduce, has been successfully applied in large-scale Internet services to support big data analytics. Hive, a data warehouse package built on top of Hadoop, facilitates querying and managing large datasets residing in distributed storage [20]. Hive supports queries expressed in a SQL-like declarative language, called HiveQL, which are implemented as map-reduce jobs that are executed using Hadoop. For many years, traditional SQL databases (such as Oracle, MySQL, PostgreSQL, SQLite, and MS SQL Server) have been used for storing different types of data. However, Hive is currently being widely adopted as a scalable data warehousing solution by many enterprises, including Facebook.

By using Hive as the underlying database system for social media storage and management, the framework can leverage scalable and distributed file systems, and Hadoop parallel computing paradigm. Within Hive, all tables are stored as Hadoop distributed file system (HDFS) files in different formats. Using such databases, data are archived and duplicated in across multiple servers with each server containing a subset of the accumulated data. As a result, parallel computing can be applied to query and process data from each server independently. Text files, for example, are stored in the TextInputFormat and binary files can be stored as SequenceFileInputFormat. However, Hive does not impose any restrictions on the type of file input format for the data are stored in [20].

In the proposed framework, RCFileInputFormat, designed for clusters with MapReduce and a step up over standard text files, is used as the storage format that can be defined while creating the

table for storing social media data. RCFileInputFormat stores the data in a column oriented manner. Such an organization can greatly speed up queries that do not access all the columns of the table [20]. For each tweet entry harvested from Twitter, all metadata about the tweet message are stored, such as the user name, time stamp and location when the tweet was created, source generating the tweet, text content, hashtags, etc. However, only one or several fields are queried and retrieved for a specific application. Therefore, RCFileInputFormat is a good storage option for this particular application. Because searching billions of social media records is time consuming, indices are created for several commonly queried fields, such as text content, hashtags, and time information.

While traditional SQL databases cannot efficiently store and manage massive social media datasets, they provide robust spatial query and operation support (e.g., retrieving data within a specific boundary). Therefore, other types of data, such as socioeconomic data downloaded from Census, have relatively structured information, and are organized in a PostgreSQL/PostGIS database, an open-source spatial database solution allowing storage and query of geographic objects.

2.2 Data process and service

Data process and service component retrieves data from the databases, and performs necessary data process, analytical, or mining functions to generate response results requested from the web interface. This module provides basic GIS data processing and analytical functions, such as geospatial data reprojection, data format conversion, and spatial clustering.

Additionally, this module also provides different spatial data and text mining capabilities. To achieve high performance, Hadoop platform, a widely used scalable distributed computing environment, is used in this framework to process social media data. Specially, Mahout, an open source machine-learning package for Hadoop, is used as underlying library for text mining. Many classic algorithms for data mining, such as naïve Bayes, Latent Dirichlet Allocation (LDA) [21], and logistic regression are implemented as MapReduce jobs. LDA is one of most crucial algorithm used in this framework to discover the emerging “hot topics” that are discussed over the social media.

First introduced by Blei et al. (2003), LDA is an example of a topic model for analyzing a large number of unlabeled data. LDA can be used to cluster words into “topics” and documents into mixtures of “topics” by uncovering the hidden thematic structure (or “topics”) in a large collection of documents. In LDA, each document is represented as a probability distribution of various topics, which are in turn distributions over words. Each word could belong to one or more topics.

As each tweet is limited to 140 characters and it is highly unstructured, including a large number of abbreviations and hashtags, unspaced phrases prefixed with the sign “#”. Any user who wants to create a concept category and to discuss and share specific information about a subject can create a hashtag. A Hashtag is an identifier unique to Twitter and is often used to search for tweets that have a common topic. Therefore, in our study, each tweet is a document and only hashtags are extracted as words to represent the document while modeling the tweeting topics with LDA.

An advantage using Mahout is that the LDA algorithm is implemented as a MapReduce job, which can be run in a large Hadoop clusters. Therefore, we can leverage Hadoop clusters

² <http://twitter4j.org/en/index.html>

³ <https://hive.apache.org/>

(Figure 1) to speed up the process of topic detection by adding more nodes into the computation. After running LDA model, an output of the computed topics with each topic being represented as a set of words (hashtags) with certain probability is produced.

2.3 Data geovisual analysis

The Web interface provides an online graphic user interface with geovisual analytical tools to customize analysis and view data in various maps and plots. Users from different communities can request the data search, analysis, visualization, or animation services through the web interface. Geovisual analytical tools normally contain multiple interactive tools, dynamic graphs and live-linked views of data representation [22, 23]. All the functions on the web interface are implemented using dojo framework⁴, HTML5 and JavaScript.

Using the interface, public users can search against the resource catalog to explore and manipulate multi-sourced data for disaster management and analysis. Within this framework, two key capabilities are included through the data geovisualization and analysis component, including 1) tracking real-time events, and 2) analyzing historic events, which are discussed in details and demonstrated in Section 3.

3. DEMONSTRATION

A system prototype is implemented in JAVA based on the JAVA Server Page (JSP) and R to automatically harvest and analyze various types of data. Several open-sources are used for the prototype development. Due to the advanced capabilities in dynamically visualizing multidimensional geographical data online, Google Map is selected for displaying maps of multi-sourced data at the client side.

3.1 Case study and data

Hurricane Sandy, which made landfall in the North-Eastern US on 29 October 2012, is selected as a case study. Sandy began as a tropical wave in the Caribbean on October 19th. It quickly developed, becoming a tropical depression and then a tropical storm. On October 28th, President Obama signed emergency declarations for several states expected to be impacted by Sandy, allowing them to request federal aid and make additional preparations in advance of the storm. On October 29th, Sandy made landfall in the United States (U.S.), striking near Atlantic City, New Jersey, with winds of 80 mph. It affected 24 states in U.S., including the entire eastern seaboard from Florida to Maine and west across the Appalachian Mountains to Michigan and Wisconsin, with particularly severe damage in New Jersey and New York.

For section 3.2, geo-tagged messages posted on Twitter during Oct 10th, 2012 and Nov 27th, 2012, from Gnip (<http://gnip.com/>) are used, and the downtown New York area is chosen as the study region. A total of 1,763,141 geo-tagged tweets were collected by sending a geographic query with the boundary of the selected study area. After filtering out non-relevant tweets using a set of predefined hashtags and keywords, 38,224 tweets are included for the next step analysis [12]. The tweets crawled using Twitter stream API between Oct 27th, 2012 to Nov 07th, 2012 are used for the demonstration of Section 3.3.

3.2 Historical event analysis

The developed web portal has several functions to support the analysis of the historical events temporally and spatially by integrating social media with other socioeconomic and disaster relevant data (Figure 2). The left panel allows users to analyze a specific event by selecting the available events stored in the database, and configuring spatial and temporal scopes. The middle panel shows different types of data. Instead of using one map as traditional desktop and online analytical tools, maps can be attached to multiple windows. Users can compare different variables extracted from different data sources and types (e.g., socioeconomic, historic damage records, climate change, social media, or remote sensing) in parallel to acquire knowledge from the spatial patterns. The map views presented are always focused on the same area. Comparing them, users can learn, for example, underlying factors contribute to the generation of tweets by linking social media data with socioeconomic data in different map windows.

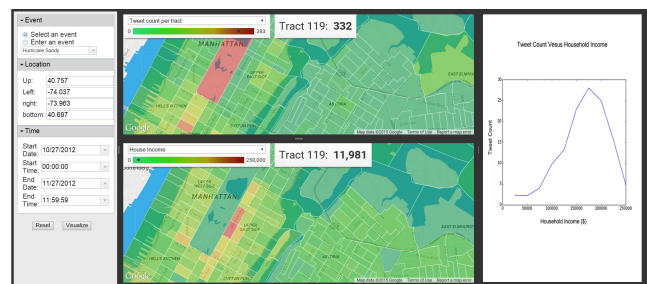


Figure 2. Tweet density versus household income

As an example, Figure 2 middle panel shows the spatial distribution of tweet frequency (upper window) and estimated median house value in each NYC Census tracts (lower window). The median household income in the study area has a unimodal distribution, showing a great diversity ranging from less than \$10,000 to \$250,000 (Figure 3 left). The average median household income is about \$56,000. Most tracts have a household income between \$30,000 and \$110,000, very few tracts have household income more than \$150,000. These wealthy areas are mainly in Manhattan along the Hudson River (Figure 2 middle panel lower window).

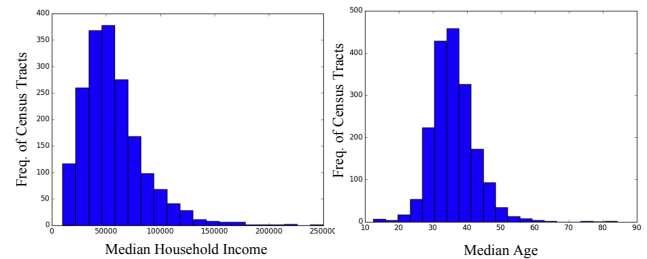


Figure 3. Histograms showing the unimodal distribution of the frequency of NYC census tracts for median house value (left) and median age (right).

Besides displaying maps of various variables, the system is able to return the statistical results as statistical plots to the client. By interactively manipulating the data, dynamic statistical plots with maps are directly linked to the knowledge discovery process. Figure 2 (right panel), for example, shows how people's contribution of relevant tweets differs as their wealth changes. The result is also a unimodal distribution. When household

⁴ <https://dojotoolkit.org/>

income is lower than around \$175,000, people’s contribution to hurricane-related tweets increase significantly as income goes up. After that peak point, the tweet frequency drops dramatically. This pattern indicates that wealthy people tend to more actively participate the online discussion of the disaster and share useful information about the disaster situation in the online community. But for the top very wealthy people (income above \$175,000), this pattern is disrupted, and there seem to be a decreased interested in this process. The analysis result confirms with the findings by Xiao et al. (2015) using statistics models to examine the spatial heterogeneity in the generation of tweets after a major disaster [24].

In addition to comparing tweet density with median household income, we can also understand how age may impact the generation of useful tweets that may contribute to situational awareness during a disaster by analyzing the correlation between age and tweet frequency in each Census tract (Figure 4).

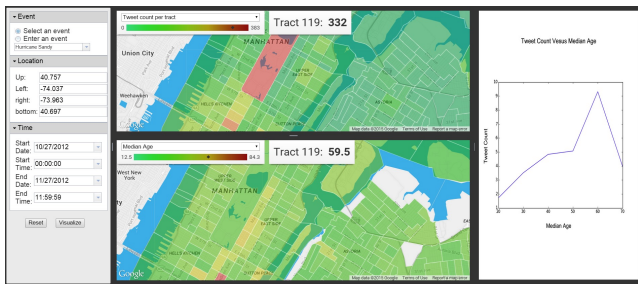


Figure 4. Tweet density versus media age

3.3 Real-time event tracking

The system supports automatic detection of the potential events by integrating topic modeling and spatial clustering algorithms. Figure 5 (left panel) shows that disaster managers or public users can set up relevant parameters for detecting potential hashtags that are associated with a specific event. One of the critical parameters is “track interval”, which indicates the time span for which tweets are monitored and used for detection. For example, if the interval is set to one hour, then tweets posted in the past one hour will be used as the input for the LDA modeling (Section 3.1).

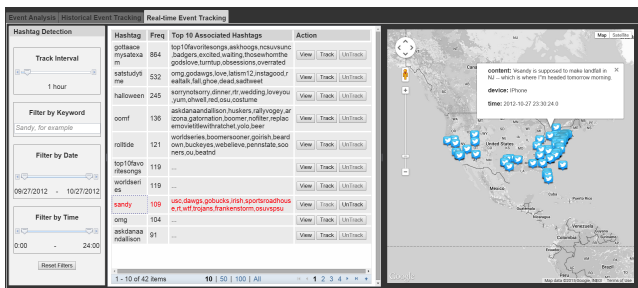


Figure 5. Hashtag detection and real-time event Tracking (For better display purpose, the hashtag sign “#” is removed)

The system consistently calculates the frequency of each hashtag with our developed program, and run the LDA algorithm as a MapReduce job, which can discover topics discussed over the Twitter at the predefined tracking interval. For each hashtag (e.g., Sandy), the system will find the topic that it belongs, and associated words (e.g., HurricaneSandy) in the topic, which are hashags that are posted together with the hashtag, can be detected. This information is automatically stored in the database and can

be accessed and displayed through the web interface (Figure 5 middle panel). For this demonstration purpose, the data between “2012-10-27 22:33:27”, when the first geo-tagged tweet including the hashtag “#sandy” in the message was posted on Twitter, and “2012-10-27 23:59:59” were used. Within this period, it was detected that there are nine hashtags that have the frequency more than 100. It can be observed that several topics, such as SAT test and studies (using the hashtags #gottaacemysatexam, and #satstudytime), Halloween parties (using the hashtag #halloween) also emerged in the Twitter.

For each topic, the top 10 associated hashtags produced by LDA modeling are displayed in the third column of the table (Figure 5 middle panel). The model output is intuitive, and it can be seen that SAT test hashtag (#gottaacemysatexam) is mostly associated with relevant hashags, such as #excited (about the results), or #waiting (for the results) etc. Results also show that tweets related to Hurricane Sandy and associated with the hashtag “#Sandy” have also emerged. After discovering these “hot topics” through the LDA model, the words (hashtags) in each topic are used to match a predefined list of keywords related to natural hazards, such as “hurricane”, “storm”, “quake”.. If a match is detected, an alert is posted by the system to provide early warnings about a potential disaster. Since the both hashtags “#Sandy” and “#frankensstorm” are associated with the word “storm”, which itself is one of the predefined keywords related to natural hazards, the system will post a warning alert and will start monitoring and track all relevant tweets.

Through the user interface, users can also view the spatial distribution of the tweets (Figure 5, right panel) with a specific hashtag included (Figure 5 middle panel). If the hashtag is relevant to a disaster, then users can click “Track” button, the system will then continually monitor the streaming tweets. If it turns out to be a false alarm after monitoring for a while, users can click “UnTrack” button, and a revoke process will trigger the system to remove the monitoring task.

4. CONCLUSION AND FUTURE WORK

In this study, a CyberGIS framework is proposed to synthesize multi-sourced data, including stream data from social media for disaster management. To manage massive social media data, Hadoop Hive is used as the scalable storage solution. Meanwhile, to address big data challenges, Hadoop platform is used as a scalable distributed computing environment to process social media data. Mahout is leveraged to support big data analytics. Using such interactive and high performance framework that can automatically synthesize multi-sourced data to produce different types of maps and plots, effective strategies can be developed to mitigate the potential effects, respond and coordinate efficiently and effectively during disasters.

With the rise of social media data in various volume and types, it is always a challenge to seamlessly integrate the data and analyze them to produce meaningful results that can be used in real time for situation awareness. The presented methodology attempts to pave the way for a systematic integration of new data streams, along with the new computing paradigms, e.g., GPU computing [25], and cloud computing [26-28] to support effective big data management. This paper shows that a CyberGIS framework is effective for disaster mapping and real-time event tracking. The proposed framework has great potential to handle multi-sourced data streams for a variety of application that go beyond disaster coordination and relief.

5. REFERENCES

1. Tran, P., et al., *GIS and local knowledge in disaster management: a case study of flood risk mapping in Viet Nam*. Disasters, 2009. 33(1): p. 152-169.
2. Schnebele, E., et al., *Real time estimation of the Calgary floods using limited remote sensing data*. Water, 2014. 6(2): p. 381-398.
3. Schnebele, E., G. Cervone, and N. Waters, *Road assessment after flood events using non-authoritative data*. Natural Hazards and Earth System Science, 2014. 14(4): p. 1007-1015.
4. Sakaki, T., M. Okazaki, and Y. Matsuo. *Earthquake shakes Twitter users: real-time event detection by social sensors*. in *Proceedings of the 19th international conference on World wide web*. 2010. ACM.
5. Sutton, J., L. Palen, and I. Shklovski. *Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires*. in *Proceedings of the 5th International ISCRAM Conference*. 2008. Washington, DC.
6. Verma, S., et al. *Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency*. in *ICWSM*. 2011. Citeseer.
7. Vieweg, S., et al. *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*. in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010. ACM.
8. Ashktorab, Z., et al., *Tweedr: Mining twitter to inform disaster response*. Proc. of ISCRAM, 2014.
9. Gao, H., G. Barbier, and R. Goolsby, *Harnessing the crowdsourcing power of social media for disaster relief*. IEEE Intelligent Systems, 2011(3): p. 10-14.
10. Imran, M., et al. *Practical extraction of disaster-relevant information from social media*. in *Proceedings of the 22nd international conference on World Wide Web companion*. 2013. International World Wide Web Conferences Steering Committee.
11. Kumar, S., et al. *TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief*. in *ICWSM*. 2011.
12. Huang, Q. and Y. Xiao, *Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery*. International Journal of Geo-Information, 2015. 4(3): p. 19.
13. Wang, S., et al., *CyberGIS software: a synthetic review and integration roadmap*. International Journal of Geographical Information Science, 2013. 27(11): p. 2122-2145.
14. Wang, S., *A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis*. Annals of the Association of American Geographers, 2010. 100(3): p. 535-557.
15. Xu, C., D.W. Wong, and C. Yang, *Evaluating the "geographical awareness" of individuals: An exploratory analysis of Twitter data*. Cartography and Geographic Information Science, 2013. 40(2): p. 103-115.
16. Riteau, P., et al. *A cloud computing approach to on-demand and scalable cyberGIS analytics*. in *Proceedings of the 5th ACM workshop on Scientific cloud computing*. 2014. ACM.
17. Huang, Q. and C. Xu, *A data-driven framework for archiving and exploring social media data*. Annals of GIS, 2014. 20(4): p. 265-277.
18. Padmanabhan, A., et al. *FluMapper: an interactive CyberGIS environment for massive location-based social media data analysis*. in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*. 2013. ACM.
19. Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters*. Communications of the ACM, 2008. 51(1): p. 107-113.
20. Thusoo, A., et al. *Hive-a petabyte scale data warehouse using hadoop*. in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. 2010. IEEE.
21. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation*. the Journal of machine Learning research, 2003. 3: p. 993-1022.
22. MacEachren, A.M. and I. Brewer, *Developing a conceptual framework for visually-enabled geocollaboration*. International Journal of Geographical Information Science, 2004. 18(1): p. 1-34.
23. Roth, R.E., *Interactive maps: What we know and what we need to know*. Journal of Spatial Information Science, 2015(6): p. 59-115.
24. Xiao, Y., Q. Huang, and K. Wu, *Understanding social media data for disaster management*. Natural Hazards, 2015.
25. Li, J., et al., *Visualizing 3D/4D environmental data using many-core graphics processing units (GPUs) and multi-core central processing units (CPUs)*. Computers & Geosciences, 2013. 59: p. 78-89.
26. Huang, Q., et al., *Utilize cloud computing to support dust storm forecasting*. International Journal of Digital Earth, 2013. 6(4): p. 338-355.
27. Huang, Q., et al., *Evaluating open-source cloud computing solutions for geosciences*. Computers & Geosciences, 2013. 59: p. 41-52.
28. Yang, C., et al., *Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?* International Journal of Digital Earth, 2011. 4(4): p. 305-329.