Big Data Driven Crisis Response System Leveraging Social Media

R. Ieshwarya[1], Dr. Kalaimani shanmugam[2], S. Sridharani[3], M.sureka[4]

[1,3,4]*Student Members, Computer Science and Engineering Department,* [2]*Proffessor/Head of Computer Science and Engineering Department, Arasu Engineering College, Kumbakonam, Tamil Nadu, India.*

Abstract **-** Emergent uses of social media enable qualitative situational analysis before, during and after disasters. In times of disaster, the online users generate a tremendous amount of data, some of which are exceedingly valuable for relief efforts. Data analytics are opening avenues to powerful new disaster-forecasting tools, giving scientists and first responders more advance notice. In this paper, a novel Intelligent GeoCyber model is proposed that can automatically synthesize multi- sourced data, such as social media and socioeconomic data in the emergent situation to perform statistical analysis for disaster management. Big Data analytics a new technological paradigm is employed with machine learning library to store, process and mine massive social media data. The main purpose of the proposed approach is to gain valuable insights and situational awareness by monitoring social media-based feeds from which tactical, actionable data related with hydrological information and flood records that can be mined efficiently. We argue that these emergent uses of social media are pre- cursors of border future changes to the institutional and organizational arrangements of disaster response.

Index Terms **-** Big Data analytics, Disaster Management, Geographic Information System (GIS), Social media.

I. INTRODUCTION

Social media, such as social network (e.g., *Facebook*), microblogs (e.g. *Twitter*) have experienced a spectacular rise in popularity, and attracting hundreds of millions of users generating unprecedented amount of information. *Twitter*, for example, has rapidly gained approximately 500 million registered users as of 2012, generating 340 million tweets daily. Although each tweet is limited to only 140 characters, the aggregate of millions of tweets may provide a realistic representation of landscapes for a certain topic of interest. Furthermore, with widespread use of location aware mobile devices, users are sharing their whereabouts through social media services.

Natural disasters have their greatest impact at local level, especially on the lives of ordinary people. Current disasters are becoming more complex and climate change poses a greater potential for adverse impacts (Aalst and Burton, 2002). The damage caused by natural disasters at the community level in Chennai has increased exponentially in the past 12 years, despite the great efforts that the National Climate Centre and India Meteorological department, Interdisciplinary Centre for Water Research and Indian Institute of Science, of NASA's Global Precipitation Measurement (GPM), local communities have put into many disaster prevention programmes.

Government and other organisations have insufficient human and financial resources to implement comprehensive disaster prevention programmes at the family level in disaster prone areas. Even if they did, mobilising local capacities and partnership with communities should be considered an essential component of any disaster management plan (Norton and Chantry, 2002). Communities have shown themselves to be a source of strength, contributing innovative ideas and local knowledge which, when mobilised and used appropriately, can lead to solutions that can make a fundamental contribution to mitigating the negative impacts of natural disasters.

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. This proposed GeoCyber model 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 model 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. 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. The purpose was to prepare detailed flood hazard maps for commune planners, villagers and other stakeholders, to identify the magnitude and extent of past flood disasters, and to make recommendations based on local knowledge and needs to local authorities and decision makers regarding flood risk reduction activities.

II. MOTIVATION AND OBJECTIVES

The proposed approach aims to present a GeoCyber model 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. The proposed model encompasses Apache Hive, Hadoop, and Mahout as scalable distributed storage, computing environment and machine learning library to store, process and mine massive social media data and it is capable of supporting big data analytics of multiple sources.
The main objectives of the proposed system:

1. To extract, transform and integrate social media data with Apache Hive as a scalable distributed storage.
2. To develop a Hadoop environment to process and analyze archived social media data to make it usable.
3. To propose machine learning algorithms with Apache Mahout that serves as an open-source library that supports map reduce tasks with the dataset.
4. To develop and present various geovisual tools as a web interface that allows users to perform customized analysis and view multisource data in different types of maps and plots for disaster management.
5. To evaluate and deploy the proposed system in real-time environment.

III. CHARACTERISTICS OF THE STUDY AREA AND ITS PROBLEMS

Chennai is a coastal city, where the two rivers, namely Adyar and Cooum, flow through. The Chennai Metropolitan Area (CMA) can broadly be divided into three parts. The northern part lies above the Cooum River, middle region is between Cooum and Adyar Rivers and the southern portion is below Adayar River. These two river discharges are added to the urban floods (flood due to high intensity of rainfall within the city) and inundated the city.

Figure 1: Adyar and Cooum Rivers in the Chennai City

Chennai has experienced major floods during the last three decades in the years 1976, 1985, 1996, 1998 and 2005 and then the devastating floods that hit Chennai city and other parts of Tamil Nadu during November - December 2015 have claimed more than 400 lives, 18 lakh (1.8 million) people were displaced and caused enormous economic damages. With estimates of damages and losses ranging from nearly 200 billion to over 1 trillion the floods were the costliest to have occurred in 2015, and were among the costliest natural disasters of the year. Adyar River (748 sq.km approx.) and Cooum River (1266 sq.km. approx.) that flows through the Chennai city caused floods due to high intensity of rainfall.

Thus the extreme high intensity rainfall event that occurred over Chennai was an outcome of a depression generated over a warm Bay of Bengal (BoB) which brought huge moisture from BoB and resulted in heavy precipitation over the South-East coast of India. The spatial distribution of Mean Sea Level Pressure (MSLP) till November, 27, 2015 shows a wide spread low pressure over the South of BoB, which became concentrated over Sri Lanka and brought huge moisture over Chennai region on Dec.1.

The drainage system has been found inadequate because of several reasons. These include:

(i) reduction in the vent way caused by the construction of bridges,
(ii) sand bar formation at the mouths of rivers,
(iii) clogging of the drains due to indiscriminate dumping of solid waste and construction debris,
(iv) inadequate design capacity,

(v) lack of connectivity of storm sewers with macro drainage,
(vi) Encroachments.

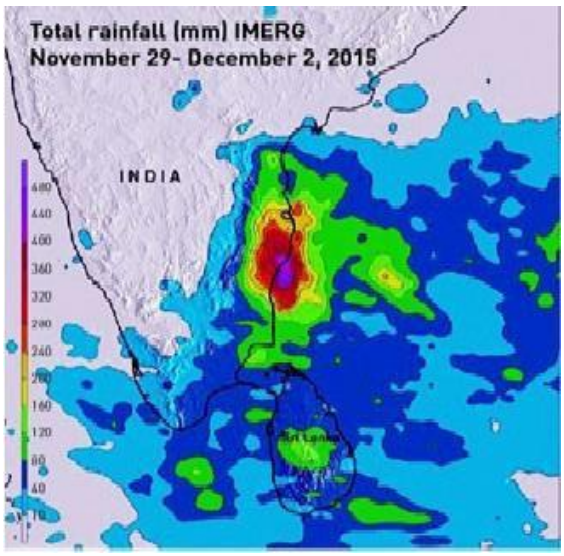

Figure 2: Accumulated rainfall between November 29 and December 2 over Chennai and neighbourhood measured by NASA's GPM satellites (Frontline, 2015b)

The long, medium and short term issues affecting urban floods, discussed in this study with specific reference to the Chennai floods of 2015 are schematically shown in Figure 3. Recommendations arising out of this rapid assessment are summarised in this study to help policy makers, researchers and other stakeholders.

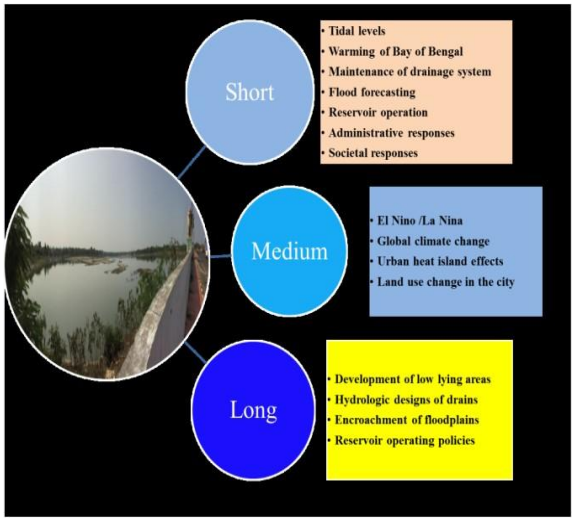

Figure 3: Spatial and Temporal Range of Issues Addressed in the study

While successive governments have focused on dredging of rivers and desilting of major drains, maintenance of minor drains is neglected due to scarcity of funds as well as public apathy. In this context it is important to bring out the effect of bad solid waste management on the condition of drainage channels, both major and minor.

IV. METHODOLOGY

The developed model is integrating multisourced data for disaster management. Overall, there are three steps or modules in our GeoCyber model, including data gathering, data process, and data geovisual analysis.

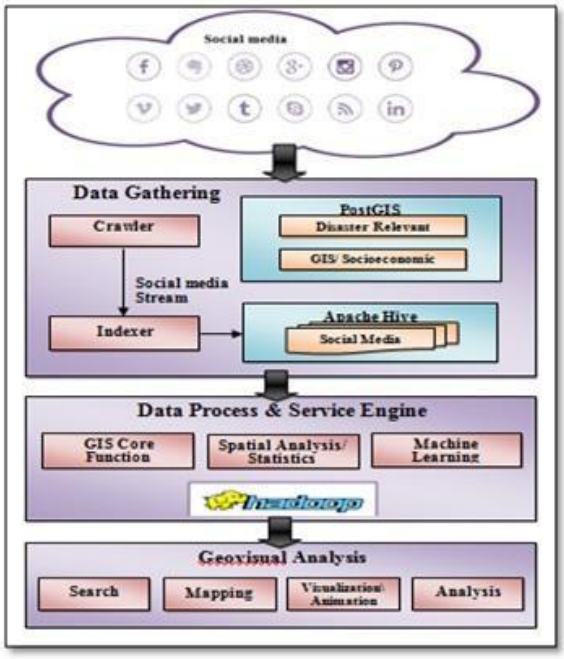

Figure 4: Design of the GeoCyber model for Disaster Management using multi-sourced data

A. Data Gathering

Data are gathered from social media (e.g. *Twitter).* 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 collecting 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.

V. Focussed Crawler:

A Web crawler is an Internet bot (i.e., software application that runs automated tasks/scripts over the Internet) which systematically browses the World Wide Web, typically for the purpose of web indexing. Web search engines and some other sites use web crawling software to update their web content or indices of others sites' web content. Web crawlers can copy all the pages they visit for

later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently. Crawlers consume resources on the systems they visit and often visit sites without tacit approval.

Therefore, the proposed system starts with archiving data to be processed for future analysis. For the data gathering, a crawler based on Twitter is implemented to collect geo-tagged tweets.

VI. Apache Hive:

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce. It stores schema in a database and processed data into HDFS. Hive is designed for OLAP and it is familiar, fast, scalable, and extensible. It provides SQL type language for querying called HiveQL or HQL which are implemented as map-reduce jobs that are executed using Hadoop.

In this proposed model, an unstructured tweet dataset, Apache Hive is used to process the generated dataset. For many years, traditional SQL databases (such as Oracle, MySQL, PostgreSQL, SQLite, and MSSQL 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 system can control scalable and distributed file systems, and Hadoop parallel computing paradigm. MapReduce based systems have emerged as a new computing paradigm for massively parallel data process in the Hadoop environment.

Within Hive, all tables are stored as Hadoop distributed file system (HDFS) files in different formats. Hadoop Distributed File System, the most widely used data store that has been successfully applied in large-scale Internet services to support big data analytics. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

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 SquenceFileInputFormat.

In the proposed system, 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. 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.

VII. PostGIS/PostgreSQL:

PostGIS is an open source, freely available, and fairly OGC compliant spatial database extender for the PostgreSQL Database Management System. PostGIS is very similar in functionality to SQL Server Spatial support, ESRI ArcSDE, Oracle Spatial, and DB2 spatial extender except it has more functionality and generally better performance than all of those. 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, are spatial database solution allowing storage and query of geographic objects.

B. Data Process and Service Engine

Normally, Data processing is the process of collecting and manipulating the data to produce meaningful information.

VIII. GIS Core Function:

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 model also provides different spatial data and text mining capabilities.

IX. Hadoop:

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop File System was developed using distributed file system design. HDFS is highly fault tolerant and designed using low-cost hardware.

To achieve high performance, Hadoop platform, a widely used in this model to process social media data. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality – nodes manipulating the data they have access to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

X. Apache Mahout:

Apache Mahout is an open source project that is primarily used in producing scalable machine learning algorithms. Mahout is such a data mining framework that normally runs coupled with the Hadoop infrastructure at its background to manage huge volumes of data. Many classic algorithms for data mining, such as Naive Bayes, Latent Dirichlet Allocation (LDA) k-means, fuzzy k-means, Canopy, Mean-Shift, and logistic regression are implemented as MapReduce jobs. The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment. Therefore, we can leverage Hadoop clusters to speed up the process of topic detection by adding more nodes into the computation. Mahout uses the Apache Hadoop library to scale effectively in the cloud.

Latent Dirichlet Allocation (LDA)

LDA is one of most crucial algorithm used in this model to discover the emerging "hot topics" that are discussed over the social media. In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. After running LDA model, an output of the computed topics with each topic being represented as a set of words (hash tags) with certain probability is produced.

C. 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.

Using the interface, public users can search against the resource catalogue to explore and manipulate multi-sourced data for disaster management and analysis. Within this model, two key capabilities are included through the data geovisualization and analysis component, including 1)
tracking real-time events, and 2) analyzing historic events.

XI. CONCLUSION AND FUTURE WORK

This paper is to study the significance of Big Data analytics to the flood occurrences in the TamilNadu state. Inferences were made on the relationship of the rainfall data with the water level of a river. By combining those information related to flooding from Twitter and satellite observations for built a real-time map of location, timing, and impact of floods. Social media enables qualitative situational analysis for the sessions - before, during, and after the disaster. Floodtags (a social media analytics platform) was employed to extract information from Twitter, enabling the filtering, visualization, and mapping of social media content based on location and keywords.

Satellite data came from the Global Flood Detection System (GFDS), which provides a service for rapid identification of inundated areas through daily passive microwave satellite observations. Hence the proposed work is expecting an early warning alert that will be produced from the result.

In future, we demonstrate Hashtag detection and real time event tracking by Java implementation and proposed to use the state of the art tool R for analyzing various types of data. Further, the proposed approach is not only confined to flood information, it may also extend to infer information pertaining to other natural disaster data such as hurricane and earthquake.

REFERENCE

[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]. ISRO/nrsc, Hydrological Simulation Study of Flood Disaster in Adyar and Cooum Rivers, TamilNadu. 2015 http://www.tutorialspoint.com/mahout/"Mahout"and http://www.tutorialspoint.com/hadoop "Hadoop information".

[3]. "ANNUAL CLIMATE SUMMARY – 2015" Issued by Government of India, Ministry of Earth Sciences, Earth System Science Organization, India Meteorological Department, Pune.

[4]. David M. Blei, Andrew Y. Ng, Michael I. *University of California Berkeley, CA 94720, US.*