Ain Shams Engineering Journal (2016) xxx, xxx–xxx

Ain Shams University

Ain Shams Engineering Journal

www.elsevier.com/locate/asej
www.sciencedirect.com

A review on application of data mining techniques to combat natural disasters

Saptarsi Goswami [a], **Sanjay Chakraborty** [a,*], **Sanhita Ghosh** [a], **Amlan Chakrabarti** [b], **Basabi Chakraborty** [c]

[a] *Institute of Engineering and Management, Kolkata, India*
[b] *A.K.Choudhury School of Information Technology, Kolkata, India*
[c] *Faculty of Software and Information Science, Iwate Prefectural University, Japan*

Received 17 August 2015; revised 26 November 2015; accepted 16 January 2016

KEYWORDS

Natural disaster;
Data mining;
Twitter;
India;
Big Data

Abstract Thousands of human lives are lost every year around the globe, apart from significant damage on property, animal life, etc., due to natural disasters (e.g., earthquake, flood, tsunami, hurricane and other storms, landslides, cloudburst, heat wave, forest fire). In this paper, we focus on reviewing the application of data mining and analytical techniques designed so far for (i) prediction, (ii) detection, and (iii) development of appropriate disaster management strategy based on the collected data from disasters. A detailed description of availability of data from geological observatories (seismological, hydrological), satellites, remote sensing and newer sources like social networking sites as twitter is presented. An extensive and in-depth literature study on current techniques for disaster prediction, detection and management has been done and the results are summarized according to various types of disasters. Finally a framework for building a disaster management database for India hosted on open source Big Data platform like Hadoop in a phased manner has been proposed. The study has special focus on India which ranks among top five counties in terms of absolute number of the loss of human life.

1. Introduction

Natural disasters affect human and animal lives and properties all around the globe. In many cases the reasons are not in our control. As noted in [1], for the three decades namely 1970–80 (rank 2nd), 1980–90 (rank 4th), 1990–00 (rank 2nd), India ranks in first 5 countries in terms of absolute number of the loss of human life. It is not only the immediate effect as observed in [2], and exposure to a natural disaster in the past months increases the likelihood of acute illnesses such as diarrhea, fever, and acute respiratory illness in children under

* Corresponding author. Mobile: +91 9038205310.
E-mail addresses: saptarsi.goswami@iemcal.com (S. Goswami), sanjay.chakraborty@iemcal.com (S. Chakraborty), sanhita.ghosh@iemcal.com (S. Ghosh), amlanc@ieee.org (A. Chakrabarti), basabi@iwate-pu.ac.jp (B. Chakraborty).
Peer review under responsibility of Ain Shams University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.asej.2016.01.012
2090-4479 © 2016 Faculty of Engineering, Ain Shams University. Production and hosting by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article in press as: Goswami S et al., A review on application of data mining techniques to combat natural disasters, Ain Shams Eng J (2016), http://dx.doi.org/10.1016/j.asej.2016.01.012

5 year by 9–18%. The socioeconomic status of the households has a direct bearing on the magnitude and nature of these effects. The disasters have pronounced effects on business houses as well. As stated in [3] 40% of the companies, which were closed for consecutive 3 days, failed or closed down within a period of 36 months. The disasters are not infrequent as well. Only for earthquake [4], there are as many as 20 earthquakes every year which has a Richter scale reading greater than 7.0. The effects of the disasters are much more pronounced in developing countries like India.

Meteorologist, Geologists, Environmental Scientists, Computer Scientists and Scientists from various other disciplines have put a lot of concerted efforts to predict the time, place and severity of the disasters. Apart from advanced weather forecasting models, data mining models also have been used for the same purpose. Another line of research, has concentrated on disaster management, appropriate flow of information, channelizing the relief work and analysis of needs or concerns of the victims. The sources of the underlying data for such tasks have often been social media and other internet media. Diverse data are also collected on regular basis by satellites, wireless and remote sensors, national meteorological and geological departments, NGOs, various other international, government and private bodies, before, during and after the disaster. The data thus collected qualify to be called 'Big Data' because of the volume, variety and the velocity in which the data are generated.

A brief technical description of some of the major natural disasters is as follows:

- *Earthquake:* A sudden movement of the earth's crust, causing destruction due to violent activities caused due to volcanic action underneath the surface of the earth. 55% of India's landmass are in seismic zone III–V.
- *Landslide:* A sudden collapse of the earth or mass of rock from mountains or cliff due to vibration on the earth's surface. In India the northern sub-Himalayan region and Western Ghats are prone to landslides.
- *Cloudburst:* It is an extreme form of unpredicted rainfall in the form of thunder storm, hail storm and heavy precipitation which is short lived. Unseasonal heavy rainfalls are common in India. A devastating effect of it was the flash flood in North India in 2013 that killed thousands of pilgrims and animals.
- *Storm:* A bad weather in the form of rain or snow caused by strong winds or air currents formed due to unexpected changes in air pressure on the earth's surface. Cyclones are common in various parts of India, especially the coastal regions that leave long lasting and expensive damages to human lives and properties.
- *Flood:* An overflow of huge water masses beyond normal limits over dry land. Every year, millions of human lives, cattle and agricultural crops are destroyed in India due to lack of planning and improper weather forecasting.
- *Tsunami:* High sea waves that are large volumes of displaced water, caused due to an earthquake, volcanic eruption or any other underwater explosions. The 2004 Tsunami that hit parts of the southeastern coast of India had devastating effects on the mainland and Andaman and Nicobar Islands.
- *Volcanic eruption:* It is a sudden, violent discharge of steam, gases, ashes, molten rocks or lava from the surface of the earth that are ejected to heights and spread for several miles. Underwater volcanoes on the islands surrounding the landmass of India are common. However, they have not imposed significant damages to the mainland till now.

The unique contributions of the paper are as follows:

- A comprehensive summary of different data mining techniques applied to various tasks pertaining to the natural disasters.
- A detailed account of various types and sources of data for each category of task and disaster.
- A brief account of disaster management 'status-quo' from Indian context.
- A brief review of suitability of 'Twitter' as a data source.
- A presentation of proposed architecture to streamline disaster management.

The organization of the paper is as follows: In Section 2, natural disasters have been discussed with focus on India; a brief description of the existing disaster management structure is also outlined. In Section 3, the broad categorizations of the tasks that can be achieved with respect to natural disaster are presented in detail. In Section 4, granular levels of tasks are enlisted with respect to the major type of tasks discussed in Section 3. Details of the tasks, data used in the task, data mining methods used, country or region have also been discussed. In Section 5, a structured view of different types of required data and their corresponding sources has been discussed. A short review of twitter and other Internet resources as data source has been discussed, along with their application for natural disaster in Section 6. In Section 7, a process flow and architecture of a disaster management system have been proposed. Section 8 contains, conclusion with the direction of future work.

2. Natural disaster from an Indian context

India is vulnerable to various natural disasters due to its unique geo-climatic condition as a result of its geographical location. This subcontinent is surrounded by water bodies on three sides and the Himalayas on the North. The country has been hit approximately by 8 natural calamities per year and there has been about 5 times increase in frequency of natural disasters in the past three decades. The calamities that affect the country can be categorized as follows: 57% landmass is prone to earthquakes, 12% floods (about 40 million hectares of land is vulnerable to floods) and 8% are prone to cyclones. Table 1 records such disasters for last 15 years.

Asia, tops in terms of number of disaster events among the continents. Close to 60% of the disasters in Asia are originated in South Asia and 40% are originated in India. In the below figure (Fig. 1), the above statistics are displayed.

India is a victim of natural disasters every year and the loss of lives and properties adds up to millions of rupees which this country cannot afford to lose. There are certain reasons for such poor disaster management procedures followed in this country.

Please cite this article in press as: Goswami S et al., A review on application of data mining techniques to combat natural disasters, Ain Shams Eng J (2016), http://dx.doi.org/10.1016/j.asej.2016.01.012