ISSN No. 0976-5697

Volume 8, No. 2, March 2017 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Using Big Data For Computational Epidemiology In India

Shweta Chaudhary
Food Safety and Standards Authority of India
e-mail:shweta.fssai@gmail.com

Abstract: The myriad epidemic data and factors contributing to epidemic outbreaks in India are scattered so unmanageably that it hardly leads to any conclusions or analysis for future predictions. In a developing country like India, the dual burden of communicable and chronic diseases hits a large population every year without sparing enough time to Government Bodies to issue any health advisories in time and develop adequate vaccinations/medications to control the epidemic.

Big Data poses as a promising technology to help combat epidemics by making combined use of enormous sources of data like physical health data and trends on the open web.

Computational epidemiology relies on the access and analysis of massive health data. When analyzed, these provide for possibilities of constructing flexible and dynamic systems with attractive real-time properties; such uses include early warnings, halting or mitigation of disease spread, simulations and scenario-based reasoning relating to health policies, and real-time decision support to first responders.

Keywords: Big Data, Analytics, Public Data, Computational Epidemiology, Health advisory

LITERATURE REVIEW

'Epidemiology' is a study of distribution and determinants of health related states or events and application of the study to control of diseases and other health problems. [1]

In contrast with traditional epidemiology, **'Computational Epidemiology'** also looks for patterns in unstructured sources of data, such as social media. It can be thought of as the hypothesis-generating antecedent to hypothesis-testing methods such as national surveys and randomized controlled trials. [2]

However, the use of Big Data for epidemic control started sometime back with the Ebola virus outbreak in West Africa offering strong evidence of the important role big data can play in epidemic tracking. When news of the epidemic began breaking in March of 2014, most of the official forecasts on the spread of Ebola came from the US Centers for Disease Control (CDC) and the World Health Organization (WHO). While initially, both the CDC and WHO relied primarily on conventional epidemiological approaches and measures to arrive at their estimates of how far and how quickly the disease was likely to spread, CDC recognized that these traditional tools were not adequate and leveraged big data to hone their insights.

They had been piloting BioMosaic, a tool that merges health, population and movement data to predict the spread of disease, for several years, and realized that Ebola fit remarkably well with the tool. BioMosaic provided CDC with near real-time availability of the global air transportation network, and enabled them to identify the at-risk populations and create a mosaic map of the diaspora population both on the move from affected areas as well as statically in terms of the US resident population. As the outbreak threatened to become a global pandemic, other organizations and sources also turned to real-time data elements to see if they could detect patterns that would help better predict how, and where, the disease might be spreading next.

Thereafter and in parallel were being developed and studied other similar epidemic control/advisory systems. HealthMap, a disease-monitoring website maintained by a team of researchers and epidemiologists at Boston Children's Hospital, is one such source of big data analytics. The site provides early detection and real-time surveillance on emerging health threats by aggregating and analyzing data from multiple sources, including social media streams, online news stories, official reports, travel sites, and official sources. CDC unveiled a new software tool, the Epi Info viral hemorrhagic fever (VHF) application, to help identify people exposed to the virus faster than traditional reporting methods allowed.

The flu is another illness that Big Data is helping monitor and abate. When a recent flu epidemic in Boston and New York had infected hundreds and killed 18, app developers and health officials turned to Big Data for help. [3]

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu. By today, similar projects such as the flu-prediction project by the institute of Cognitive Science Osnabrück carry the basic idea forward, by combining social media data e.g. Twitter with CDC data, and structural models that infer the spatial and temporal spreading of the disease.

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offered for declared research purposes. Google Flu Trends was not really a success because they relied on data from only one source– their search engine queries. [4]

In both the Ebola and the flu epidemics, social media played a large part in supplying massive data sets that helped identify outbreaks, forecast where the diseases might spread next, and gave clues as to where new outbreaks may be developing. Patients sharing symptoms, “checking in” at medical clinics on apps and social media sites, and wearable’s add up as other health markers.

The influenza H1N1 epidemic initiated at Mexico in March 2009, and spread globally. The influenza surveillance program in India monitored for patients with influenza symptoms. Daphne Lopez, M. Gunasekaran, and B. Senthil Murugan’ researched and presented an ecological niche model based on geographically weighted regression to predict the incidence and prevalence of H1N1 influenza in different regions of Vellore, India, thereby assisting in prioritizing high risk areas for implementation of optimal prevention interventions. [5]

Many of the research projects attempting to solve the Big Data problem for epidemiologists are already underway.

PROPOSED MODEL/SOLUTION

The proposed model/solution aims to conclude that despite multiple epidemic data/factors and lack of health data accuracy maintained across India, these data sources when combined together and analyzed using Big Data Analytics are bound to lead to a better health advisory system and epidemic control in the country.

Already, new data streams—structured and unstructured—are cascading into the healthcare realm from fitness devices, genetics and genomics, social media research and other sources. But very small percent of this data can presently be captured, stored and organized so that it can be manipulated by computers and analyzed for useful information.

Fortunately, advances in data management, particularly virtualization and cloud computing, are facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data.

The analytics associated with big data is described by three primary characteristics: volume, velocity and variety.

Integrating Big Data with Cloud Computing

Using cloud computing, multiple users can access a single server to retrieve and process their data without having to buy licenses for different applications. Also, the user need not worry about technical issues as these are handled by the third party and hence, can focus on other useful tasks. There are three types of clouds: private, public and hybrid. After taking into account various factors such as cost, security, workload, etc, an organization can decide on the type of infrastructure to be deployed and accordingly, avail the necessary services. It is also possible to implement hybrid models which combine certain selective features of private clouds, such as privacy with those of public clouds, such as scalability or interoperability.

Big data, due to its vast size, variability and high velocity requires the use of multiple servers that work in a parallel manner. Since cloud environments already make use of multiple servers and allocate resources on demand, it would be highly beneficial for organizations to make use of cloud services for big data analysis. The facility of parallel computing offered by cloud services can augment the efficiency with which big data is processed. It also has the potential to replace most batch processing systems with real time processing systems. The intersection between cloud and big data is still relatively untapped. Yet, utilizing a cloud system to store big data has long term benefits to both, the insights yielded, as well as, the performance of the IT sector. Data without analysis is worthless. Big data requires advanced analytic techniques to deal with the extensive amounts of data. Cloud systems are typically based on remote servers, which are able to handle extensive amounts of data with rapid response time for real time processes.

Utilizing IaaS, cloud systems further reduces data center rental costs. Thus, Cloud computing infrastructure enables more efficient use of hardware and software investments. Pooling these resources forces costs down and improves utilization. Problems of analysis and storage can be solved through a hosted cloud system which provides secure, scalable solutions for managing data. Elasticity, pay-per-use, low upfront investment, and low security risks are some of the major features that make cloud computing an ideal platform for big data analysis, which would not have been economically viable on traditional infrastructure. [6]

Over time, health-related data will be created and accumulated continuously, resulting in an incredible *volume* of data. Data is accumulated in real-time and at a rapid pace, or *velocity*. The ability to perform real-time analytics against such high-volume data in motion and across all specialties would revolutionize healthcare. Therein lies *variety*.

Research in epidemiology aims to identify the distribution, incidence, and etiology of human diseases to improve the understanding of the causes of diseases and to prevent their spread. Traditionally, epidemiology has been based on data collected by public health agencies through health personnel in hospitals, doctors' offices, and out in the field. In recent years, however, newer data sources have emerged where data are frequently collected directly from individuals through the digital traces they leave as a consequence of modern communication and an increased use of electronic devices.

This research will be based upon some public sources of health data combined with data on social media platforms. Some common metrics to measure epidemiology can be considered as Mortality, Morbidity, Risk factors, etc as listed below.

TABLE 1: PUBLIC SOURCES OF DATA

Source	Metrics
Integrated Disease Surveillance Project (IDSP)	Mortality & causes of death
	Morbidity & health status
National Vector Borne Diseases Control Programme	Mortality & causes of death
	Morbidity & health status
National Family Health Surveys	Mortality
	Morbidity & health status
	Risk factors
	Service provision

World Health Survey	Mortality
	Morbidity & health status
	Risk factors
	Service provision
UNICEF Multiple Indicator Survey	Morbidity & health status
	Risk factors
	Service provision
Ministry of Health and Family Welfare "Annual Report"	Health infrastructure
Global Burden of Disease and Risk Factors Study	Mortality
	Morbidity & health status

For instance, Integrated Disease Surveillance Programme (IDSP), operated through India's National Center for Disease Control (NCDC), is today active in all Indian states. Weekly surveillance data on 18 epidemic-prone diseases, including viral hepatitis, are collected through this program. All 28,850 government-run primary health care centers and hospitals and 2,923 designated private facilities serve as reporting units, which collect and report data on hepatitis cases (any acute onset of jaundice) and outbreaks, and report them to district surveillance units each week. These reports are submitted as aggregate data to IDSP through a web portal (http://www.idsp.nic.in). However, no demographic information, risk factors, or other data are collected or reported [7].

IDSP has acknowledged helping this research by letting make use of their data sets for cloud computation.

The social media has also been a great resource for big data in which innovations and research can be done to deduce conclusive reports on infectious and chronic disease trends in populations. Given the vast amount of people who are connected online, at one given point of time, someone is publically sharing volumes of personal and community health information. Social media allows for reciprocity; the more one person shares, the more others share in return (Kass-Hout & Alhinnawi, 2013). An example is a study by Salathe and Kandelwal (cited in Kass- Hout & Alhinnawi, 2013) where the researchers assessed vaccination sentiments on the social media during the H1N1 pandemic. Sentiment trends could be seen in some parts of the network and was figured that negative sentiments spread more effectively than positive sentiments [8].

At the time of writing this paper, the research is being carried forward in the direction of unstructured data like open web trends and social media. Guidance on SAS® Social Media Analytics has been sought from NEGD, Ministry of Electronics and Information Technology, Government of India, such that social media factors can effectively be included to generate dynamic epidemic reports.

RESULTS AND DISCUSSIONS

This paper reviews the dimensions of potentials that big data can contribute to the public health domain. Potential include real-time data analysis, a more rigorous research and development arena as well as valuable information from open web. However, there are areas of concern and challenges in applying big data to public health. These challenges include the current lack of universal standardization and classification that may render big data to be of poor use. There are also privacy and security concerns where individuals will lose the right to their private information, leading to a future with no secrets.

Another challenge is the need for platforms and powerful tool to analyze the large and rapidly growing amount of data. However, these challenges can be overcome with good leadership, training, specialization, advocacy and contemporary policies to support the development of public health informatics. The use of big data in health care differs from its use in other industries such as marketing or product development. These differences can be seen in the need for regulations, ethical standards, privacy boundaries and some form of standardization in the diversity of data sources and in their differing goals. There is a need for more skilled health care informatics professionals and leaders to deal with big data confidently and to address the challenges that arise from the use of big data. As in any science, data is evidence and knowledge. The ability to harmonize and analyze data enables the sharing of knowledge across disciplines and to gain insight into the complex and challenging field of public health. Big data, used wisely, is a Pandora's Box for health care.

CONCLUSION

We live in the era of cloud computing and big data. Discussions and posts on various social media platforms on topics relating epidemiology act as valuable sources in addition to traditional sources of health data collected by public departments for developing an effectively updated atlas of infectious diseases.

REFERENCES

[1] http://www.who.int/topics/epidemiology/en/.

[2] https://en.wikipedia.org/wiki/Computational_epidemiology.

[3] http://govdatadownload.com/2015/01/27/big-data-enables-epidemic-tracking/.

[4] https://en.wikipedia.org/wiki/Google_Flu_Trends.

[5] Spatial Big Data Analytics of Influenza Epidemic in Vellore, India, Daphne Lopez, M. Gunasekaran, and B. Senthil Murugan.

[6] Integration of Big Data and Cloud Computing, 2014, Castelino, Gandhi, Narula, Chokshi

[7] The potentials and challenges of big data in publicHealth, Rena N. Vithiatharan, 2014.

[8] A monthly surveillance report from Integrated Disease Surveillance Programme National Health Mission, April 2016.