Journal of Glycomics & Lipidomics

ISSN: 2153-0637

Extended Abstract

Big data for research and development

Alex V Vasenkov

Abstract

This talk will focus on Big Data for Research and Development (R&D). There are several definitions of Big Data which create confusion about this subject. There is even more confusion about synthetic Big Data that can be defined as a collection of research articles, PhD theses, patents, test reports and product description reports. Such data have emerging attributes like high volume, high velocity, high variety, and veracity that make an analysis of synthetic data difficult. There is an emergent need for a framework that can synergistically integrate search or information retrieval (IR) with information extraction (IE). Traditional IR-based text searching can be used for a quick exploration of large collections of synthetic data. However, this approach is incapable of finding specific R&D concepts in such collections and establishing connections between these concepts. Also, the IR models lack an ability to learn concepts and relationships between the concepts. In contrast, the IE models are too specific and typically require customization for a domain of interest. A novel framework will be presented and its feasibility to mine synthetic data will be shown. It was found possible to partially or fully automate analysis of synthetic data to find labeled information and connecting concepts. The present framework can help individuals to identify non-obvious solutions to R&D problems, to serve as an input for innovation, or to categorize prior art relevant to a technological concept or a patent application in question.

With the explosion of social media sites and proliferation of digital computing devices and Internet access, massive amounts of public data is being generated on a daily basis. Efficient techniques/algorithms to analyze this massive amount of data can provide near real-time information about emerging trends and provide early warning in case of an imminent emergency (such as the outbreak of a viral disease). In addition, careful mining of these data can reveal many useful indicators of socioeconomic and political events, which can help in establishing effective public policies. The focus of this study is to review the application of big data analytics for the purpose of human development. The emerging ability to use big data techniques for development (BD4D) promises to revolutionalize healthcare, education, and agriculture; facilitate the alleviation of poverty; and help to deal with humanitarian crises and violent conflicts. Besides all the benefits, the large-scale deployment of BD4D is beset with several challenges due to the massive size, fast-changing and diverse nature of big data. The most pressing concerns relate to efficient data acquisition and sharing, establishing of context (e.g., geolocation and time) and veracity of a dataset, and ensuring appropriate privacy. In this study, we provide a review of existing BD4D work to study the impact of big data on the development of society. In addition to reviewing the important works, we also highlight important challenges and open issues.

In the modern world we are inundated with data, with companies such as Google and Facebook dealing with petabytes of data. Google processes more than 24 petabytes of data per day, while Facebook, a company founded a decade ago, gets more than 10 million photos per hour. The glut of data, buoyed by fast advancing technology, is increasing exponentially due to increased digitization of all aspects of modern life (using technologies such as the Internet of Things (IoT) –which uses sensors, for example in the shape of wearable devices, to provide data related to human activities and different behavioral patterns). It is estimated that we are generating 2.5 quintillion bytes per day (we note here that a quintillion bytes, or an exabyte, is equal to 1018 bytes.
The presence of "big data", or this massive amount of increasing data, offers both an opportunity as well as a challenge to researchers. A lot of progress has been made in developing the capability to process, store, and analyze big data: In addition to the big data computing capability (in terms of processing and storing big data in a distributed fashion on a cluster of computers the rapid advances in using intelligent data analytics techniques—drawn from the emerging areas of artificial intelligence (AI) and machine learning (ML)—provide the ability to process massive amounts of diverse unstructured data that is now being generated daily to extract valuable actionable knowledge. This provides a great opportunity to researchers to use this data for developing useful knowledge and insights

From the perspective of big data for development (BD4D), an important quandary is gaining access to important people-related data, which is often in the exclusive access of the government in the form of paper documents. Fortunately the emerging trend known as "open data", which promotes open public sharing of data from various public and private sector entities in searchable and machine-readable formats is a boon for BD4D research. Governments worldwide are increasingly adopting open data projects to fuel innovation and transparency. In addition, open source platforms have been developed that facilitate the creation and gathering of digital data from mobile platforms. While open data can be rightly regarded as a subset of all the available big data: the nuance is in the liquidity of big data. Open data also promotes a culture of creativity and public wellbeing as is evident by different hackathons that are being organized to tap the potential of open data in terms of useful mobile applications (e.g., the local government of Rio de Janeiro has created the Rio Operation Center. aimed at harnessesing the power of technology and big data to run the city effectively in terms of transport management, natural disaster relief, mass movement and management of slum areas). In a recent report from McKinsey Global Institute. the net worth of open data was estimated to be $3 trillion. In this report, the importance of open data is highlighted for seven particular sectors: education, health, transportation, consumer products, electricity, oil and gas, and consumer finance. In 2009, the Secretary-General of the United Nations (UN), Ban Ki-moon started the UN Global Pulse (UNGP) initiative, with the explicit goal of harnessing big data technology for human development.

The Global Pulse program is aimed at forming a network of innovation centers, called the Pulse Labs, all over the world. Ideally, these Pulse labs will bring together people from different fields of life together to make use of the free and open source computing methods/ software toolkits to analyze data to help the development and humanitarian operations especially in the developing countries. In [14], Kirkpatrick, the director of the UN Global Pulse innovation initiative, presents the case for deploying big data techniques and analytics in the field of human development. It is highlighted that data—especially from mobile phone and social media—can be utilized in fighting hunger, disaster and poverty. This report talks about "data philanthropy" where the companies, whose businesses revolve around data, can collaborate with the UN in predicting imminent humanitarian crises and help take possible steps to avoid situations that can lead to disasters. The report also discusses the issues and challenges faced by the UN in terms of data access, user privacy and the integration of big data techniques into the various UN humanitarian systems.

This work is partly presented at 2nd International Conference on Big Data Analysis and Data Mining 30-December 01, 2015 San Antonio, USA

Alex V Vasenkov,
Multi Scale Solutions, USA, E-mail: avv@multiscalesolutions.com