Chapter 17
Big Data Analytics in Bioinformatics

Dharmendra Trikamlal Patel
Charotar University of Science and Technology, India

ABSTRACT

Voluminous data are being generated by various means. The Internet of Things (IoT) has emerged recently to group all manmade artificial things around us. Due to intelligent devices, the annual growth of data generation has increased rapidly, and it is expected that by 2020, it will reach more than 40 trillion GB. Data generated through devices are in unstructured form. Traditional techniques of descriptive and predictive analysis are not enough for that. Big Data Analytics have emerged to perform descriptive and predictive analysis on such voluminous data. This chapter first deals with the introduction to Big Data Analytics. Big Data Analytics is very essential in Bioinformatics field as the size of human genome sometimes reaches 200 GB. The chapter next deals with different types of big data in Bioinformatics. The chapter describes several problems and challenges based on big data in Bioinformatics. Finally, the chapter deals with techniques of Big Data Analytics in the Bioinformatics field.

1.0 INTRODUCTION TO BIG DATA ANALYTICS

In last several years, the volume, variety and velocity of data generating in any organization whether big or small reaches to unprecedented level. Such kind of voluminous data is termed as Big Data. Few major trends that is responsible and management of this situation are:

- Internet that has made the information available to everyone. Through Internet people send e-mails, likes face book comments, send tweets, upload photos in face book etc. The data transit via Internet is in voluminous amount and has in variety of types. Internet has given the birth of Internet of Things (IoT) that is one kind of network of manmade things embedded with electronics products, software and sensors which enable things to collect and exchange data. According to Gartner, information technology research and advisory company, revenue generated for IoT things would exceed $300 billion by 2020. IoT and big data are basically considered as two sides

of the same coin. Extraction and management of IoT data is the biggest challenge for most of all organization in the world. To manage and extract IoT data properly, organization should set up a proper analytics platform that give performance and can be expanded in future.

- Ubiquitous devices have helped people to connect with everyone and for personal computing activities such as social networking, photo sharing, micro blogging etc. The rise of big data increasingly demanded that data analytics should be supported through ubiquitous devices to enable decision making from anywhere without barrier of specific workplace.
- Cloud Computing has provided infinite computing power with cheapest rate. The cloud is an ideal place for big data because of its scalable storage space, infinite computing power and elastic resources.

Big Data (Bryant et al.,2008) is broad and very complex term for data sets(J. Becla et al.,2009) so traditional way of analytics is inadequate for that. Big Data Analytics is a new way of analytics that is used to make analysis of Big Data. Following is a complete definition of Big Data Analytics.

Big Data Analytics is the proactive process of examining quantitative and qualitative big data to identify useful, hidden, valid, previously unknown patterns and correlations that can be used to make better decisions for any organization.

Basically, analytics have four main approaches and they falls either in reactive or proactive category.

1. Business Intelligence is a reactive approach to generate standard and ad hoc business reports based on certain analytics when they are demanded. This kind of analytics generates reports based on static past.
2. Big Data Business Intelligence (McAfee et al.,2012) is a similar approach as Business Intelligence but applicable on huge data sets.
3. Big Analytics is a proactive approach of analytics (Sandryhaila et al.,2014) that does optimization, text mining, forecasting, statistical analysis, predictive analysis etc. However, it is not applicable on Big Data as traditional storage environment (Balazinska et al.,2009).
4. Big Data Analytics is a proactive approach that is applicable on Big Data. It is helpful for determining hidden, unknown and useful information from terabytes, petabytes and exabytes.

Big Data Analytics deals with Big Data so understanding of Big Data terminology is vital for analysis and interpretation. Next section will deal with the Big Data thoroughly.

1.1 Big Data Overview

There is no specific rule about exactly what size of dataset requires to be considered as "Big". However certain things require defining Big Data such as (a) the requirement for new techniques (Chen et al.,2014) and tools (Shvachko et al.,2010) in order to be able to process it (b) the requirement of software programs that span multiple physical or virtual machines in order to process data efficiently and timely.

Big Data is characterized by four important vectors – volume, variety, velocity and veracity.

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/big-data-analytics-in-bioinformatics/197708

Related Content

Answering Top-k Keyword Queries on Relational Databases

Myint Myint Theinand Mie Mie Su Thwin (2012). *International Journal of Information Retrieval Research (pp. 36-57).*

www.irma-international.org/article/answering-top-keyword-queries-relational/78313

Text Clustering using Distances Combination by Social Bees: Towards 3D Visualisation Aspect

Hadj Ahmed Bouarara, Reda Mohamed Hamouand Abdelmalek Amine (2014). *International Journal of Information Retrieval Research (pp. 34-53).*

www.irma-international.org/article/text-clustering-using-distances-combination-by-social-bees/127000

Optimal Query Generation for Hidden Web Extraction through Response Analysis

Sonali Guptaand Komal Kumar Bhatia (2014). *International Journal of Information Retrieval Research (pp. 1-18).*

www.irma-international.org/article/optimal-query-generation-for-hidden-web-extraction-through-response-analysis/126326

SPAM: An Effective and Efficient Spatial Algorithm for Mining Grid Data

Ritu Chauhanand Harleen Kaur (2018). *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications (pp. 1323-1337).*

www.irma-international.org/chapter/spam/198601

Interactive IR in Online Database Environments

Iris Xie (2008). *Interactive Information Retrieval in Digital Environments (pp. 53-82).*

www.irma-international.org/chapter/interactive-online-database-environments/24524