Chapter 8
Mining Big Data and Streams

Hoda Ahmed Abdelhafez
Suez Canal University, Egypt

ABSTRACT

Mining big data is getting a lot of attention currently because businesses need more complex information in order to increase their revenue and gain competitive advantage. Therefore, mining the huge amount of data as well as mining real-time data needs to be done by new data mining techniques/approaches. This chapter will discuss big data volume, variety, and velocity, data mining techniques, and open source tools for handling very large datasets. Moreover, the chapter will focus on two industrial areas telecommunications and healthcare and lessons learned from them.

INTRODUCTION

Mining big data is getting lot of attention currently because the businesses need more complex information in order to increase their revenue and gain competitive advantage. the growing of the telecommunication data traffic according to Cisco annual forecasting will reach 8.6 zettabytes by the end of 2018 up from 3.1 zettabytes per year in 2013 (Cisco analysis, 2014). Therefore, mining the huge amount of data as well as mining real-time data needs to be done by new data mining techniques/approaches. Big Data is a new term used to identify the datasets that are of large size and have grater complexity (Bifet, 2013). Data mining (DM) is the process of searching large volumes of data automatically for patterns such as association rules (Gupta et al., 2014). Big data mining is defined as the capability of extracting valuable information from large datasets or streams of data that due to its characteristics it is not possible before to do it (Fan & Bifet, 2013). This chapter will discuss the challenges of big data, new data mining techniques compared with traditional techniques and the main DM tools for handling very large datasets. Moreover, the chapter will focus on two industrial areas telecommunications and healthcare and lessons learned from them.

DOI: 10.4018/978-1-5225-7598-6.ch008

BACKGROUND

The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. We need new algorithms, and new tools to deal with all of these data. Therefore, the use of big data is becoming a crucial way for leading companies. For example, in healthcare, data pioneers are analyzing the health outcomes of pharmaceuticals when they were widely prescribed, and discovering benefits and risks that were not evident during necessarily more limited clinical trials (McGuire, 2012). We selected some significant articles that discussed challenges, techniques and tools for mining big data. Yadav et al. (2013) presented a review of several algorithms from 1994-2013 necessary for handling big data set. It gives an overview of architecture and algorithms used in large data sets, various tools that were developed for analyzing them as well as various security issues and trends. Bifet (2013) discussed data stream mining and how it offers many challenges and also many opportunities. Che et al. (2013) presented an overview of mining big data and its challenges include heterogeneity, scalability, speed, accuracy, trust, provenance and privacy. This paper also provides an overview of the platforms for processing and managing big data as well as platforms and libraries for mining big data. Jovic et al. (2014) discussed several data mining tools including RapidMiner, R, Weka, KNIME, Orange, and scikit-learn. Fan and Bifet (2013) presented big data challenges, applications of mining big data, Apache Hadoop and other open sources for big data mining and big graphic mining. Singh (2014) discussed machine learning techniques to capturing the value hidden in big data. He presented supervised learning using neural networks, Support Vector Machines (SVMs) and Naive Bayes classifiers, and also unsupervised learning using k-Means, hierarchical clustering and self-organizing maps.

CHALLENGES OF BIG DATA SYSTEMS

Big data has five key elements: Volume, Velocity, Variety, Veracity and value. These 5 V's are considered challenges of Big Data systems (Yin & Kaynak, 2015; Ishwarappa & Anuradha, 2015; Marr, 2015).

Volume refers to the huge amount of data. Many companies have large archived data in the form of logs but do not have the capacity to manipulate and analyze that data using traditional database technology. Now big data technology can help store and use these datasets in order to gain benefits from them.

Velocity represents the speed at which data generated and the speed at which data moves around. The speed at which credit card transactions is checked for fraudulent activities and the social media messages going to viral in seconds. Thus, big data technology can be used to analyze the data while it is being generated without putting it into databases.

Variety means different data types or format. Traditional database can store and process structured data that fit into tables such as financial data. Now 90% of data generated is in unstructured form and it cannot easily be put into relational databases such as photos, video sequences or social media updates. Big data technology can now harness various types of data like messages, photos, sensor data, and social media conversations and bring them together with more structured data.

Veracity refers to the trustworthiness of the data. The quality and accuracy of big data are less controllable because there will be dirty data. For instance, twitter posts with hash-tags, abbreviations, typos and colloquial speech. Big data analytics now allows us to work with these types of data. The volume, variety and velocity of data often make up for the lack of quality or accuracy.