Chapter 14
Collaborative and Clustering Based Strategy in Big Data

Arushi Jain
Ambedkar Institute of Advanced Communication Technologies and Research, India

Vishal Bhatnagar
Ambedkar Institute of Advanced Communication Technologies and Research, India

Pulkit Sharma
Ambedkar Institute of Advanced Communication Technologies and Research, India

ABSTRACT

There is a proliferation in the amount of data generated and its volume, which is going to persevere for many coming years. Big data clustering is the exercise of taking a set of objects and dividing them into groups in such a way that the objects in the same groups are more similar to each other according to a certain set of parameters than to those in other groups. These groups are known as clusters. Cluster analysis is one of the main tasks in the field of data mining and is a commonly used technique for statistical analysis of data. While big data collaborative filtering defined as a technique that filters the information sought by the user and patterns by collaborating multiple data sets such as viewpoints, multiple agents and pre-existing data about the users' behavior stored in matrices. Collaborative filtering is especially required when a huge data set is present.

INTRODUCTION

A huge surge in the amount of data being generated that needs to be stored and analyzed quickly has been witnessed in the recent years. Walmart handles millions of transactions per hour while Facebook handles 40 billion photos uploaded by its users each day. Big data has become important part of data analytics market. Big Data can be defined using five v's. These are:

1. **Volume**: This refers to the amount of data. While volume is indicatory of more data, it is the particulate nature of the data that is exclusive. For example data logs from twitter, click streams

DOI: 10.4018/978-1-5225-7501-6.ch014

of web pages and mobile apps, sensor-enabled equipment capturing data, etc. It is the task of big data for converting data into useful information so that valuable action could be taken.

2. **Velocity**: This refers to the rate at which data is generated, captured and received. For example, to make lucrative offers ecommerce applications combines mobile location and personal choices of the buyer.
3. **Variety**: This refers to various types of structured, unstructured and semi- structured data types. Unstructured data consist of files such as audio and video. Unstructured data has many of the requirements similar to that of structured data, such as summarization, audit ability, and privacy. This data is generated from varied sources such as satellites, sensors, social networks, etc.
4. **Value**: This refers to the intrinsic value that the data may possess, and must be discovered. There is wide variety of techniques to derive value from data. The advancement in the recent years have led to exponential decrease in the cost of storage and processing of data, thus providing statistical analysis on the entire data possible, unlike the past where random samples were analyzed to draw inferences.
5. **Veracity**: This refers to the abnormality in data. Veracity in data analysis is one of the biggest challenges. This is dealt with by properly defining the problem statement before analysis, finding relevant data and using proven techniques for analysis so that the result is trustworthy and useful. There are various tools and techniques in the market for big data analytics.

Some of the challenges of big data are:

1. The biggest challenge in big data is to aggregate data from heterogeneous sources and analyzes it to get useful information out of it to improve various aspects of functioning and business process of organizations. The data may come from various social networks, with each having a different format.
2. One of the main characteristics of big data is Autonomous where data source works independently without being dependent on centralized control. For example World Wide Web generates function correctly without involving other servers.
3. Another challenge is complexity. The complexity of Big Data is due to multiple data; the data is collected in very different contexts (multi-source, multi-view, multi-tables, sequential, etc.).
4. Big data is always evolving, thus evolution of complex data which poses a big challenge. The typical example is when a customer posts a review on a page of social networking, it has to be extracted over specific periods of time so that the algorithm can operate and provide relevant information to the users.

LITERATURE SURVEY

To manage the growing demands, there is a need to increase the capacity and performance of tools and methods employed for analysis of data. Chen et al. (2014), in their work "Big data: A survey" focused on big data and reviewed related technologies and examined the application of big data in various fields. Al-Jarrah et al. (2015), in their work "Efficient Machine Learning for Big Data: A Review" reviewed the data modeling in large scale data intensive field relating to model efficiency and new algorithm approaches. Hoffmann and Birnbrich (2012) to protect their customer from third party fraud proposed