Chapter 5
Big Data Collection, Filtering, and Extraction of Features

Ganesh B. Regulwar
https://orcid.org/0000-0001-9320-7040
Vardhaman College of Engineering, India

Ashish Mahalle
G.H. Raisoni College of Engineering, Nagpur, India

Raju Pawar
G.H. Raisoni College of Engineering, Nagpur, India

Swati K. Shamkuwar
G.H. Raisoni College of Engineering, Nagpur, India

Priti Roshan Kakde
https://orcid.org/0000-0002-5951-8723
G.H. Raisoni College of Engineering, Nagpur, India

Swati Tiwari
https://orcid.org/0009-0006-1063-5538
G.H. Raisoni College of Engineering, Nagpur, India

ABSTRACT

Big data is a term used to describe data sets that are too large or intricate for traditional data processing systems to handle. Big data collection, filtering, and feature extraction are significant procedures in data science that enable organizations to scrutinize vast amounts of data to obtain insights and make well-informed decisions. Following filtration, feature extraction is executed to identify vital patterns and relationships in the data using techniques such as clustering, principal component analysis, and association rule mining. The primary objective of big data collection, filtering, and feature extraction is to identify valuable information that can aid in decision-making, enhance operations, and develop new products and services. These processes are essential for organizations that aspire to remain competitive and at the forefront of the constantly changing data landscape.

DOI: 10.4018/979-8-3693-0413-6.ch005

1. INTRODUCTION

In today's digital era, we find ourselves immersed in an unprecedented flood of data. This overwhelming abundance of information is commonly referred to as "big data." To tap into the immense potential of this data ocean, it's crucial to understand how to effectively collect, filter, and extract valuable features from it.

his chapter delves into the essential aspects of handling big data, focusing on three vital stages: data collection, data filtering, and feature extraction. We will explore the methodologies and techniques used to gather data from diverse sources, refine it by eliminating noise and irrelevant details, and uncover valuable insights by extracting meaningful features (Ray, 2019).

Whether you are an enthusiastic data explorer, a researcher, or a professional seeking to harness big data's power for your organization, this chapter will provide you with a solid groundwork to navigate through the realm of abundant data effectively. Let's embark on this enlightening journey to unveil the hidden treasures concealed within big data!

1.1 Overview of Big Data

The concept of big data in the rapidly evolving digital world. Big data refers to massive volumes of both structured and unstructured data, produced at a high speed and encompassing various data types. It is characterized by three fundamental attributes known as the three Vs: Volume, Velocity, and Variety. The exponential growth of big data is driven by factors such as the proliferation of connected devices, social media platforms, and the Internet of Things (IoT).

Organizations across industries recognize the potential value hidden within big data. By analyzing and extracting insights from these vast datasets, businesses can make data-driven decisions and gain a competitive edge. Governments and research institutions also utilize big data for critical applications, ranging from healthcare advancements to urban planning (Perer, 2019).

However, managing big data comes with challenges as traditional data processing tools struggle to cope with the scale and complexity of these datasets. Scaling storage, processing capabilities, and ensuring data privacy and security become vital concerns (Desai, 2020).

The chapter aims to delve into the methodologies and techniques employed for big data collection, filtering, and feature extraction, empowering researchers and organizations to unlock the potential of big data and drive innovations in today's data-centric world.

1.2. Importance of Data Collection, Filtering, and Feature Extraction

Data collection, filtering, and feature extraction hold immense significance in the domain of big data analytics. Each stage plays a vital role in transforming raw data into valuable insights and actionable knowledge. Here are the key reasons that underscore the importance of these processes:

Data Collection: High-quality and comprehensive data collection is crucial for accurate and reliable analysis. Incomplete or biased data can result in erroneous conclusions and flawed decision-making. Data collection from diverse sources enables the identification of hidden patterns, trends, and correlations, leading to a deeper understanding of complex phenomena. Well-executed data collection empowers organizations to make informed decisions based on evidence rather than intuition or guesswork.

Data Filtering: Filtering out irrelevant data and noise improves data quality and enhances the precision and dependability of subsequent analyses. Reducing data volume and eliminating unnecessary