Data Mining for Visualizing Polluted Gases

Yas A. Alsultanny
https://orcid.org/0000-0002-6211-7074
Uruk University, Iraq

INTRODUCTION

Big Data Mining (BDM) and Data Visualization (DV) are very important topics in the field of knowledge extraction. Big data required considerable data processing and storage capacity. The big data can be visualized and analyzed to extract knowledge. Big data can be used as a useful tool to enhance decision making (Shumway, 2014). The visual analytical tools have steadily improved during the last years to work with big data. The data age, where data grows exponentially, is a significant struggle to extract knowledge (Zhwan & Zeebaree, 2021). Visual analytics enables the exploration of air quality influence among various traffic scenarios by proper visual means (Bachechi, Po, & Rollo, 2022).

Big data is a term used to describe some of current directions in information technology, as a concept that take into consideration data analysis. The amount of data in the world is huge, in 2020, every person generated 1.7 megabytes per second (Petrov, 2021). It is important to note that most of the big data is unstructured data, where it is not organized and does not fit the usual databases (Smallcombe, 2020).

Data Mining is the technique to get useful knowledge out of databases; data mining requires preprocessing and analytic approach for finding the value. Data mining requires many operations such as data integration, data selection, and so on (Han, Kamber, & Jian, 2012). Selecting a suitable method of data mining is best method for knowledge extraction and forecasting the future (Alsultanny, 2011).

Visual analytic first defined by Thomas and Cook in 2005 as, the science of analytical reasoning facility by interactive visual interface. Murray in 2013 described Data Visualization as; "fortunately, we humans are intensely visual creatures. Few of us can detect patterns among rows of numbers, but even young children can interpret bar charts, extracting meaning from those numbers' visual representations. Visualizing data is the fastest way to communicate it to others." Data Visualization are valuable for the introduction of data in graphical form (Thanuj, Vinitha, & Sumathi, 2021).

Air pollution levels raised risk for diseases such as heart disease, stroke, chronic obstructive pulmonary disease, cancer, and pneumonia, the death every year is 4.2 million due to exposure to ambient (outdoor) air pollution (World Health Organization, 2021a). Air pollution is important in our live; most of the pollutants in the air are a result of emissions from cars, trucks, buses, factories, refineries, and other resources.

The objective of this chapter is to highlight the aspects of Big Data miming to visualize air pollution concentrations and it is relative to meteorological parameters. The data for this chapter collected from stations for monitoring pollution gases. These stations usually have an hourly reading to measure concentrations of gases such as ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), carbon dioxide (CO_2), particulate matter (PM_{10} and $PM_{2.5}$), moreover these stations have an hourly reading for meteorological parameters such as Temperature (Temp), Humidity (Hu), Wind Speed (WS), Wind Direction (WD), and Air Pressure (AP). RapidMiner was used in this chapter to show visually the pollution gases distribution.

DOI: 10.4018/978-1-7998-9220-5.ch077

Background

“Big data” was first known in the 1990s (Hiter, 2021). Big data rises with the huge growth of data. It refers to the storing, processing, and analyzing the vast amounts of data, the speed of information data growth is faster and faster (Guo, Tang, Liu, & Gu, 2021). Big data brings new challenges to visualization because of the speed, size, and diversity of data. One of the most common definitions of big data is data that have volume, variety, and velocity (Dion, AbdelMalik, & Mawudeku, 2015; De Mauro, Greco, & Grimaldi, 2016). The Big data generated may be structured data, Semi Structured, data or unstructured data (Fan & Bifet, 2014).

The term “Big Data” is surrounded by a lot of advertising, where many software vendors claim to have the ability to handle big data with their products (Oracle, 2021). Innovations in hardware technology such as those in network bandwidth, memory, and storage technology have assisted the technology of Big Data. The new innovations coupled with the latent need to analyze the massive unstructured data that stimulated their development (Bhagattjee, 2014). Big data analytics was adopted by many organizations for constructing valuable information from big data to improve operational efficiency (Sivarajah, Kamal, Irani, & Weerakkody, 2017). Large scale data visualization is the best method of utilizing traffic and environmental big data, to improve the efficiency of data analysis, and human-computer interaction (Cao, Wang, & Liu, 2020).

Big data have characteristics 5 Vs (volume, variety, velocity, veracity, and value) (Marr, 2015). These characteristics extended to 10 Vs by adding more 5 Vs (variability, validity, vulnerability, volatility, and visualization) (Firican, 2017). Ability to utilize big data to visualize, analyze, and predict is changing the humanitarian operations and management dramatically (Akter & Wamba, 2019). Big data can be easily, efficiently processed by Map Reduce (Thanekar, Subrahmanyam, & Bagwan, 2016). The global big data market is supported by the growth of big data, which attained USD 208 billion in 2020. The market is expected to grow to attain USD 450 billion by 2026 (Expert Market Research, 2021).

Data analytics helps all types of public and private sector organizations to make better, quicker, and more efficient decisions (Barbero et al., 2016). Data Mining is the field of discovering novel and potentially useful information from large amounts of data (Cheng, Liu, Shi, Jin, & Li, 2016). Data mining defined as the use of analytical tools to discover knowledge in a database. The analytical tools may include machine learning, statistics, artificial intelligence, and information visualization (Redpath, 2000). Data mining categorized into seven categories as Fayyad, Piatetsky-Shapiro, & Smyth, stated in 1996. The important categories of data mining are regression, clustering, summarization, dependency modeling, link analysis, and sequence analysis. Knowledge Discovery in Databases (KDD) is the processing steps used to extract useful information from large collections of data (Frawley, Piatetsky-Shapiro, & Matheus, 1991).

Data mining mainly have two methods: classification is assigns items in a collection to target categories or classes, and clustering is a form of unstructured learning method. Decision trees are types of classifications such as: Reduced Error Pruning (REP) tree, K Nearest Neighbors (KNN), the J48 based on C4.5 algorithm, and M5P algorithm is an improvement of the Quinlan’s M5 algorithm (Tan, Steinbach, & Kumar, 2006; Neeb & Kurrus, 2009; Kantardzic, 2011; Witten, Frank, Hall, & Pal, 2017).

Visualization has two meanings, “To form a mental image of something” refers to a cognitive, internal aspect whereas “to make something visible to the eye” refers to an external, perceptual role (Oxford English Dictionary, 2009). Visualization is any kind of technique to present information (Chen, Hardle, & Unwin, 2008; Keim et al., 2008). Data visualization refers to any graphic representation that can examine or communicate the data in any discipline (Few, 2009). Data visualization is the presentation of