Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 7770-7771

Research Article

# A Review Paper on Big Data and Data Mining Concepts and Techniques

**Varsha Singh**

**Abstract**

In the digital era like today the growth of data in the database is very rapid, all things related to technology have a large contribution to data growth as well as social media, financial technology and scientific data. Therefore, topics such as big data and data mining are topics that are often discussed. Data mining is a method of extracting information through from big data to produce an information pattern or data anomaly.

**Keywords**-component; data, big data, data mining.

## INTRODUCTION

Since the digital era began and the internet began to be used extensively in the early 1990s, it has produced tremendous amounts of data transactions. Even long before the internet era of things a few decades earlier several studies had discussed the data of the washouse, big data and data mining. Analogously, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.[1] Data mining is used to analyze and explore large amounts of data to find a valuable result from the extraction. there is a lot of information that can be extracted from a collection of databases that can be used for several needs, for example a company engaged in the commercial sector can utilize the transaction data to find optimal sales patterns. Thus, the income from a company can be increased through the use of data mining.

## CONCEPTS

A Big Data Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three Vs.

• Volume The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a webpage or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

• Velocity Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

• Variety Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video require additional preprocessing to derive meaning and support metadata. B. Data Mining Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.

## ALGORITHM

A. Decision Tree algorithm Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data). The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

B. Genetic Algorithms (Complex Adaptive Systems) A genetic or evolutionary algorithm applies the principles of evolution found in nature to the problem of finding an optimal solution to a Solver problem. In a "genetic

algorithm," the problem is encoded in a series of bit strings that are manipulated by the algorithm; in an "evolutionary algorithm," the decision variables and problem functions are used directly. Most commercial Solver products are based on evolutionary algorithms. An evolutionary algorithm for optimization is different from "classical" optimization methods in several ways

C. Neural Networks An artificial neural network is made up of a series of nodes. Nodes are connected in many ways like the neurons and axons in the human brain. These nodes are primed in a number of different ways. Some are limited to certain algorithms and tasks which they perform exclusively. In most cases, however, nodes are able to process a variety of algorithms. Nodes are able to absorb input and produce output. They are also connected to an artificial learning program. The program can change inputs as well as the weights for different nodes.

## CHALLENGES AND ISSUES

A. Challenges

Efficient and effective data mining in large databases poses numerous requirements and great challenges to researchers and developers. The issues involved include data mining methodology, user interaction, performance and scalability, and the processing of a large variety of data types. Other issues include the exploration of data mining applications and their social impacts.

B. Issues

• Information poorness

▪ The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation

▪ Data collected in large data repositories become “data tombs”

▪ data archives that are seldom visited

• Decision Making

• Important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker’s intuition

• The decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data

## CONCLUTION

With very rapid data growth, research in the field of big data and data mining is still growing rapidly too, those who struggling in big data will always find an increasement of complexity from big data. Therefore we conclude that the approach of data mining algorithms can still be improved.. With the many problems that still exist now and issues that occur the possibility is still widely open.

## REFERENCES

1. Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining : Concepts and Techniques”, ISBN 978-0-12-381479-1, 225 Wyman Street, Waltham, MA 02451, US, 2012
2. Ian H. Witten, Eibe Frank, “Data Mining Practical Machine Learning Tools and Techniques”, 500 Sansome Street, Suite 400, San Francisco, CA 9411, 2015
3. Bart van der Sloot, Dennis Broeders, Erik Schrijvers, ”Exploring the Boundaries of Big Data”, Amsterdam, 2016
4. Nikita Jain, Vishal Srivastava, “Data Mining Techniques: A Survey Paper”, eISSN: 2319-1163 | pISSN: 2321-7308, Rajasthan, India, 2013
5. Nirmal Kaur, Gurpinder Singh,” A Review Paper On Data Mining And Big Data”, ISSN No. 0976-5697, Jalandhar, Punjab, India, 2017