Journal of
Computer Science and Software Development

JScholar
Fostering Scholarly Communication

Research Article Open Access

Big Data Mining and Its Challenges: A Review

Amjad A Alaskar*, Hailah A Almesned, Nouf N Almuqati and Dr. Mohammad M. Hassan

King Saud University, Riyadh, Saudi Arabia

***Corresponding author:** Amjad A Alaskar, King Saud University, Riyadh, Saudi Arabia, Tel: +966114695202, E-mail: 439203736@student.ksu.edu.sa

Received Date: September 19, 2022 **Accepted Date:** October 19, 2022 **Published Date:** October 22, 2022

Citation: Amjad A Alaskar, Hailah A Almesned, Nouf N Almuqati, Dr. Mohammad M. Hassan (2022) Big Data Mining and Its Challenges: A Review. J Comput Sci Software Dev 2: 1-8.

Abstract

In recent years, a massive amount of data being generated, these data are too big, move in fast and have various types and cannot be processed using traditional database systems or data mining algorithms, and need special approaches to handle it, this data known as big data. Data mining aims to discover, explore, analyze relevant data from a huge data source by use different techniques and algorithms in order to obtain extracted patterns and useful information. Mining big data refers to the process of extracting knowledge and patterns from massive datasets. Due to the characteristics of big data new requirements and techniques need to manage and mining big data and thus makes new research challenges. This paper gives overview of techniques used to mine big data and some challenges that face the process.

Keywords: Big Data; Data Mining; Big Data Mining; Hadoop; HDFS; MapReduce

Introduction

The term big data arise due to the huge amount of data which became difficult to analyze, store and process in traditional ways, because of its complexity. The increasing in the amount of data is correlated with the increasing of different data generating sources, such as internet, mobile devices social media and sensors. The volume, variety and velocity are the main characteristics of big data due to these characteristics the traditional techniques cannot handle the management and mining processes of the data. New techniques are needed to manage, process and mine the big data. Big data mining is the ability to discover and extract valuable and useful information from very large datasets. Big data mining technologies and algorithms opened door for new research areas.

This paper presents the concepts of big data, data mining and its algorithms. Also discuss techniques used to big data mining and the challenges that face it.

The rest of this paper is structured as follows: Section 2, presents big data and its characteristics. The data mining concept reviewed in section 3. Section 4, describe the data mining algorithms. An overview of mining big data presented in section 5. In section 6, we discuss the challenges in mining big data. Section 7, summarizes the techniques for big data mining. Section 8, presents the evolution to big data analytics techniques. Finally, we conclude this paper in section 9.

Big Data and Its Characteristics

Big Data is huge amount and variety of data needs to be stored, processed, distributed and analyzed to enhance decision making and processed optimization. Big Data is generated and collected from myriad data sources like: Search Engine Data, Social Media Data and Stock Exchange Data.

The concept of Big Data is based on 3V which they: volume means huge amount of data, velocity means high speed of data in and out, and variety means different data types and sources [1]. Also, other papers add, veracity indicates to the uncertainty of the data, variability is not similar the same as variety. if the meaning is constantly changing it can have a huge impact on data homogenization. and value, it is the most important thing, big data should contain valuable data to support businesses [2].

The big data can be divided into three parts: Unstructured Data type (like Word, Text, PDF and video, audio), Structured Data type (like relational data and Databases) and Semi-Structured Data type (like XML file).

Data Mining

Data mining (DM) can be defined as the process of exploring valuable and hidden knowledge and laws from unknown data; this is from a technical view point. From business perspective, data mining is the extraction, processing and analysis of large amounts of data, producing some values access to critical information and knowledge that can support making business decisions [18]. Data mining is also known as Knowledge Discovery in Data (KDD). Data mining can be defined as the process of discovering patterns from huge data and making predictions to obtain new data [13].

The KDD process has five steps as shown in Figure 1

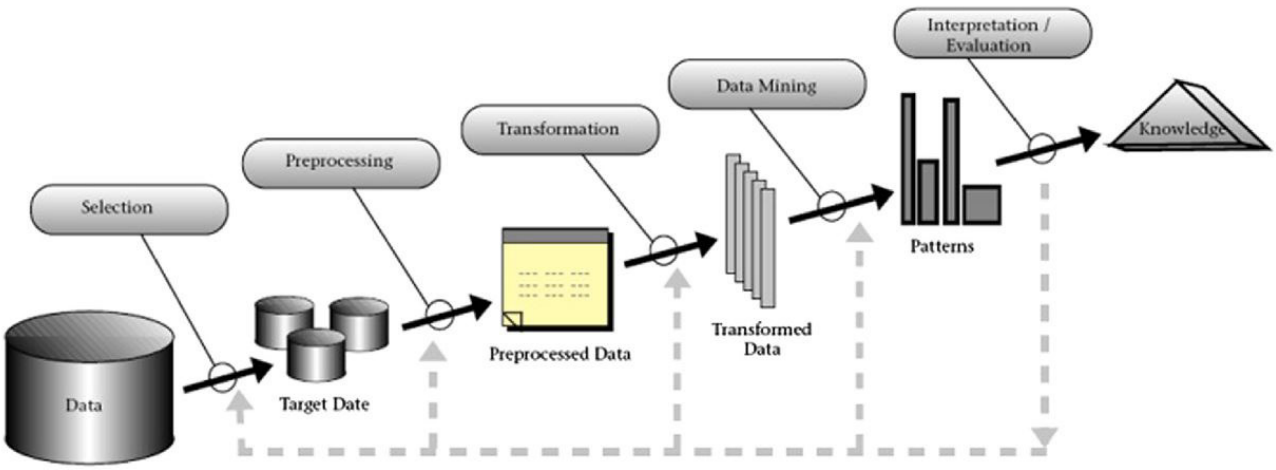

Figure 1: The KDD process [19].

1. Data Selection: selecting proper data and relevant variables, on which discovery has to be performed.

2. Data Processing: this step aims to make the data clean by replacing missing values, removing noise and outliers.

3. Data Transformation: reducing and projecting the data in order to obtain a suitable form that data mining algorithms can be implement.

4. Data Mining: choosing a proper data mining method (classification, clustering or regression), suitable algorithm to perform the task, and extracting the patterns.

5. Evaluation and Interpretation: this is the last step, the patterns extracted and now the user interprets and extracts the knowledge from the patterns. This step includes visualization of extracted patterns and models, or visualization of data using the extracted models [19,20].

Data Mining Algorithms

In present's world of big data, a large database is becoming a necessity. Just imagine there present a database with many terabytes. As Facebook alone handles 600 terabytes of new data every single day. Also, the primary challenge of big data is how to make sense of it. Moreover, the big volume is not the only problem. Also, big data need to diverse, unstructure and fast changing. Consider audio and video data, social media posts, 3D data or geospatial data. This kind of data is not easily categorized or organized. additional, to meet this challenge, a many of algorithms for extracting information or data mining. In this section, we discuss a variety of learning algorithms including k-means, decision trees, classification algorithms, neural network, Naive Bayes, K Nearest Neighbors Algorithm, association, regression, and ID3 algorithm. And here, We'll talk about the details of the most commonly used algorithms:

Classification

Classification is a more complex data mining algorithm that forces you to collect various attributes together into discernible categories, which you can then use to draw further conclusions, or serve some function. For example, if you are evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as low, medium, or high credit risks. You could then use these classifications to learn even more about those customers

Decision Trees

A graphical representation of a collection of classification rules. Given a data record, the tree directs the record from the root to a leaf. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

K Nearest Neighbors Algorithm KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN is a non-parametric, simple learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

Naive Bayes

The Naive Bayes algorithm is based on the Bayesian theorem. It is particularly used when the dimensionality of the inputs is high. The Bayesian algorithm is capable of calculating the possible output. That is based on the input. Naive Bayes can often outperform more sophisticated classification methods.

ID3 Algorithm

This Data Mining Algorithm starts with the main set as the root hub. On every cycle, it emphasizes through every unused attribute of the set and figures. That the entropy of attribute. At that point chooses the attribute. That has the smallest entropy value. The set is S then split by the selectedattribute to produce subsets of the information. This Data Mining algorithms proceed to recurse on each item in a subset. Also, considering only items never selected before.

Mining Big Data

In recent years, massive amounts of data are generated every moment in different fields such as Internet, bank, health care, social media and physical systems, this is known as big data. Valuable information can be extracted from this big data by using data mining. Traditional data mining techniques can find out potential useful information, valuable relationships and patterns in the data. The extractedinformation supports decision

making process and makes some predictions [16,17]. Multiple applications gain benefits from data mining such as education, science, health and smart cities [13]. The traditional data mining techniquesunsuitable to deal with big data due to limitations of these techniques in dealings with characteristics of big data [16,17]. Big data mining is the ability to extracting valuable and beneficial information from huge datasets that is due to its heterogeneity, volume and velocity, it was not possible todo it [13]. New requirements and techniques need to manage and mining big data, in order to fulfil theserequirements MapReduce and Hadoop introduced [16].

Challenges in Mining Big Data

The mining of Big Data involves multiple processes that facing a lot of challenges, like [3] [4]:

• Heterogeneity and Incompleteness:

- the data do not have a particular format. it is a mixed data based on different patterns or rules. Data can be both structured and unstructured. 80% of the data generated by organizations are unstructured like images, pdf documents, video, audio etc. and they cannot be stored in row/ column format as structured data. so, it needs sophisticated technology that can deal with heterogeneous data.

- Incomplete data refers to the missing of data field values. While most modern data mining algorithms have inbuilt solutions to handle missing values it seeks to impute missing values in order to produce improved models.

• Scale and complexity:

The traditional technologies are not enough for managing the increasing volumes of data. Big data analysis is considered as a challenge due to scalability and complexity of data that needs to be analyzed.

• Speed/Velocity:

Big data has a speed/velocity in data. and it needs to be processed within a certain period of time, otherwise, the results will become not valuable.

• Privacy and Security:

Security and privacy play a significant role in big data research and technology, especially in social media, bank transactions and health information. Developing algorithms that deal with personal data is a major challenge.

Techniques for Big Data Mining

Big data analysis is the complex process of examining large and varied data sets to get useful information, extract knowledge and hidden patterns, that can help companies or application make informed business decisions and manage their problems. For analysis the huge amount of data requires sophisticated technologies. Emerging technologies such as the Hadoop framework and MapReduce offer new and exciting ways to process and transform complex, unstructured and large amounts of data, into meaningful knowledge [3] [4] [5].

Hadoop Framework

Hadoop is an open source project under Apache Software for developing distributed applications that can process huge of data. It was introduced in 2005. It works in an environment that gives distributed systems and computation across various clusters [7]. Hadoop is developed to scale up from a single server to many servers. Hadoop is designed to process big data efficiently and used to support the processing of big data in a distributed computing Environment.

Hadoop ecosystem consist of multiple components are explained as below [6]:

- Hadoop Distributed File System (HDFS): is responsible for splitting and storing files into computer nodes.
- MapReduce Framework: is responsible for processing each data block in parallel [8].
- HBase: A column oriented distributed NoSQL database for random read/write access.
- Pig: A high level data programming language for analyzing data of Hadoop computation.
- Hive: A data warehousing application that provides a SQL like access and relational model.
- Sqoop: A project for transferring/importing data between relational databases and Hadoop.
- Oozie: An orchestration and workflow management for dependent Hadoop jobs [9].

Mapreduce Framework

A way to speed up the mining of big data is to distribute the training process into several machines in parallel. MapReduce framework is configured as master-slave JobTracker. It is designed for processing extremely big data in parallel mode by splitting the job into various independent tasks. A MapReduce program in general is a combination of two tasks: Map and Reduce. In map phase, the data is filtered and sorted containing a key-value pair.

In reduce phase, they are aggregated for better results. And its advantages as following below [10]:

A. Simplicity: programming jobs to run using MapReduce is simple understanding of system infrastructure is not required.

B. Fault-tolerance: In an environment with thousands of data nodes, defects are expected to occur. MapReduce can deal with this problem, so no loss of results or interruption of work can happen.

C. Flexibility: MapReduce does not require data to be organized in a specific format.

D. Scalability: MapReduce can scale to more of clusters.

With MapReduce parallel programming being applied to many data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model. to mine information from Big data, parallel computing-based algorithms such as MapReduce are used. In such algorithms, large data sets are divided into a number of subsets and then, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to meet the goal of Big Data mining. The data mining algorithms can be converted into big data map reduce algorithm which is based on parallel computing basis [11].

Evolution to Big Data Analytics Techniques

Due to the increment of data volume have made the well-known data mining algorithms unsuitable for such data sizes. Therefore, many studies have currently been directed towards improvements that data mining techniques can handle Big Data. Big data analytic techniques are concerned with several data mining functions, where the most important functions are: association rules mining and classification tree analysis.

In [12] paper, it analyzed the main data mining tasks which can adopt big data analytics techniques and "V" dimensions of big data.

Table 1 represents a summary of the analysis done for the evolution of data mining tasks to big data analytics.

Table 1: Evolution of Data Mining Technique to Big Data Analytics

S. No	Data Mining Task	Technique to be used	Developed to big data analytics	Dimensions covered
1	Classification	K- nearest neibhour	Y	Volume & Varacity
		Decision Tree	Y	Volume, Velocity & Variety
		Support Vector Machine	N	Volume, Velocity & Variety
		Naïve Bayes Classifier	N	Volume, Velocity & Variety
		Ripper	N	Volume, Velocity & Variety
		Neural Network	Y	Volume
2	Association Mining	Apriori	Y	Volume & Velocity
		FP Growth	Y	Velocity
3	Clustering	K-Means Clustering	Y	Volume
		K-Medoids	N	Volume
4	Optimization	Genetic Algorithm	N	-
		Sampling Techniques	N	-
5	Classifiers Ensembles	Bagging	N	-
		Random Forest	N	-
		Rotation Forest	N	-

Conclusion

Now, we are in big data time, and there is a growing demand for tools which can process and analyze it. Big data analytics deals with extracting valuable information from that massive data which can't be handled by traditional data mining tools. In this paper, we discuss big data mining, it characteristics, challenges and algorithms used to deal with big data mining efficiently. Also provide some techniques of big data mining: Hadoop framework and MapReduceframework. Big data mining can be in many different applications in enterprises, social networks and mobile clouds. Finally, we discuss some of big data analytics techniques and its evolution.

Copyright Form

References

1. Kumar Manish, Baluja G, Sahu Dinesh (2017) Conceptualizing Big Data Analytics Through Hadoop," COMPUSOFT an International Journal of Advanced Computer Technology 6: 5.

2. Muttipati Appala, Akkinapalli Koushik, Santhosh Eagala (2017) "Big Data: Challenges and Solutions," International Journal of Computer Science and Engineering 5: 10.

3. Albarznji Kamal, Atanassov Atanas(2016) " A Survey of Big Data Mining: Challenges and Techniques," Proceedings of 24th International Symposium "Control of Energy, Industrial and Ecological Systems".

4. Jaseena KU, Julie M David (2014) "Issues, Challenges, and Solutions: Big Data Mining," Computer Science & Information Technology.

5. Bibhudutta Jena, Mahendra Kumar Gourisaria, Siddharth Swarup Rautaray, Manjusha P (2017) "A Survey Work on Optimization Techniques Utilizing Map Reduce Framework in Hadoop Cluster " I.J. Intelligent Systems and Applications.

6. Dominic Ehiwe, Kayode Akinola, Akpovi Ominike (2016) "Enterprise Big Data: Case Study of Issues and Challenges for Businesses in Finance and Retail Sectors", International Journal of Applied Information Systems 11: 4.

7. Shalika Jaiswal, Amandeep Singh Walia(2017) "Big Data and Hadoop Challenges and Issues", International Journal of Advanced Research in Computer Science 8: 4.

8. Shobha Rani, B Rama (2017) "MapReduce with Hadoop for Simplified Analysis of Big Data " 8: 5.

9. P Nandhini, M Pavithra, R Suganya (2018) " Big Data with Data Mining", IJSRSET 4.

10. Alhaddad Ebraheem, Eassa Fathy (2018)" Performance Improvement Techniques for MapReduce - A Survey," International Journal of Computer Science and Mobile Computing 7: 4.

11. Rohit Pitre, Vijay Kolekar (2014) "A Survey Paper on Data Mining with Big Data", International Journal of Innovative Research in Advanced Engineering 1.

12. Tiju Cherian, Hrushabh Bhadkamkar (2017) "A Study and Survey of Big Data Using Data Mining Techniques", International Journal of Engineering Sciences & Research Technology 3.

13. AS Hashmi, T Ahmed (2016) "Big Data Mining: Tools & Algorithms", International Journal of Recent Contributions from Engineering, Science & IT (iJES).

14. Ashish Bindra, Sreenivasulu Pokuri, Krishna Uppala, Ankur Teredesai (2012) "Distributed Big Advertiser Data Mining", International Conference on Data Mining Workshops.

15. Bina Kotiyal, Ankit Kumar, Bhaskar Pant, RH Goudal (2013) "Big Data: Mining of Log File through Hadoop", International Conference on Human Computer Interactions (ICHCI).

16. Al Aghbari, Zaher (2015) "Mining Big Data: Challenges and Opportunities", International Conference on Enterprise Information Systems, Proceedings.

17. Alotaibi Nojod, Abdullah Manal (2016) "Big Data Mining: A classification perspective", International Conference on Communication, Management and Information Technology ICCMIT'.

18. Jinlong Wang, Jing Liu, Russell Higgs, Li Zhou, Chuanai Zhou (2017) "The Application of Data Mining Technology to Big Data", IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).

19. Petar Ristoski, Heiko Paulheim (2016) "Semantic Web in data mining and knowledge discovery: A comprehensive survey", Journal of Web Semantics 36.

20. Cemil Colak, Esra Karaman, M GokhanTurtay (2015) "Application of knowledge discovery process on the prediction of stroke", Computer Methods and Programs in Biomedicine 119.

21. Francesco Gullo (2015) "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do", Physics Procedia 62.