Processing Big Data using Data Mining technique

L.Mohan
Assistant Professor, Dept. of CSE
Chirstu jyothi Institute of Technology & Science
jangaon, India.

S.Vijaya Laxmi
Assistant Professor, Dept. of CSE
Chirstu jyothi Institute of Technology & Science
jangaon, India.

Abstract-***Big Data relates large-volume, complex, increasing data sets with multiple independent sources. With the rapid evolution of data, data storage and the networking collection capability, Big Data are now speedily expanding in all science and engineering domains. Big Data mining is the ability of extracting constructive information from huge streams of data or datasets, that due to its variability, volume, and velocity. Data mining includes exploring and analyzing big quantity of data to locate different molds for big data. Data mining is extension to data warehouse which derives useful patterns help us to take decision's for business growth .but present volume of data increased comes from different sources along with complex relationship known as big data. This paper presents a work that include the importance, challenges and applications of Big Data in various fields and the different approaches used for Big Data Analysis using Data Mining technique.***

Keywords: *Big data, data mining, processing****.***

I. INTRODUCTION

Data Mining is the method of discovering interesting knowledge, such as pattern, association and important structures, from huge amounts of data stored in database [1], data warehouses, or other information repositories. Due to the large availability of huge amounts of data in electronic forms, and the necessity to turn into useful information and knowledge for broad application including market analysis, big business management, and decision support, data mining has involved a great deal of attention in information industry in recent years . Researchers view data mining as an essential step of knowledge discovery process consists of an iterative sequence of the following steps such as data maintenance, data integration, data choice, data transformation, pattern evaluation.

The origin of the term 'Big Data' [2] is due to the fact that we are creating a huge amount of data every day. The data produced nowadays is estimated in the order of petabytes', and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices and big companies as Google, Apple, Facebook and Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. "Bigdata" is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing.

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data [3]. Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources to help draw a best possible picture to reveal the genuine data. The term Big Data literally concerns about data volumes.

A) There are two types of big data

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically.

B) Big Data Application

Big data can be useful in both commercial & noncommercial purpose. It can be listed as follows:

→ Targeted customer approach

→ Identification & solution of customer issues in real time.

→ Easier information search & knowledge building.

→ Competence building by R&D.

→ Improved analysis capabilities etc.

These applications [6] will help people and organizations to have better services & better customer experience.

II. BIG DATA SCOPE

➢ Big data is a constantly moving target. Traditional data bases systems are not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data [5] continuous to grow.

➢ Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data.

➢ This data driven model involves demand driven aggregation of information sources and mining analysis of the big data.

Big Data is a concept that refers to oceanic volume of data, moving at an unforeseen velocity, with such a high variation in structure and very often veracious in nature that - to be fully exploited, explored and to derive its value.

- ***Volume:*** The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate.
- ***Variety:*** Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more.
- ***Velocity:*** Data in motion. The speed at which data is created, processed and analyzed continues to accelerate.
- **Variability:** There are changes in the structure of the data and how users want to interpret that data.
- **Value:** Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

The purpose of Big data mining [5] is to go beyond the usual request-response processing, market basket analysis or uncovering some hidden relationships and patterns between numerical parameters of data but to design and implement very large scale parallel data mining algorithm. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge in the target domain. However, this brings a series of new challenges to the research community. Overcoming the challenges will reshape the future of the data mining technology, resulting in a spectrum of groundbreaking data and mining techniques and algorithms.

A) The operations related to big data value can be divided in to following 3 processes

- ***Big data collection:*** It focuses on data collection a data control over data access. They contain large datasets, collect and append new data and provide access to those data on request. Companies related to it sell or license data access & data sets. One of the examples is twitter that sales data feeds to Gnip.
- ***Big data Aggregation & Integration:*** The operation focuses on building technical infrastructure for data aggregation, management and backup of Bigdata. They provide software tools to manage and restore useful data from Big data pool and integrate them for decision making. Companies like Oracle sale technical infrastructure service & consulting services. Oracle big data provide appliance as an integrated Bigdata solution.
- ***Big data Analytics:*** The operation focuses on goal oriented analysis of big data sets. The analysis tools extract meaningful dataset from large pool of data and analyze them according to user perspective and for users benefit. Organizations like kaggle, SAP and others sale data analysis and visualization services.

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. As the internet is growing, amount of big data continue to grow. Big data analytics provide new ways for businesses and to analyze unstructured data. Numerous technological innovations are driving the dramatic increase in data and data gathering. The data driven model involves demand driven aggregation of information sources and mining analysis is to be analyzed further to enhance the process of data mining technique in terms of Big Data.

III. OPTIMIZATION DATA MINING

Data mining started to be an interest target for information industry because of the existence of huge data containing large amounts of hidden knowledge. In fact we could say that data mining is a relatively new scientific research area in the field of statistics, machine learning, database management science and visualization to discover and present knowledge in a form which is easily understandable to us. The function of data mining is to discover hidden knowledge from the large volumes of data sets. This extracted knowledge can help and support organizations to make better and more intelligence decisions. Data mining systems have potential for generating millions of patterns and rules. An interesting pattern represents knowledge. Measures of pattern interestingness either objective or subjective can be used to guide the discovery process.

There are many popular models that can be efficiently used in different data mining problems. Optimization is the process of finding the most cost effective or highest achievable performance alternatives under some given constraints by maximizing the desired factors and minimizing the undesired ones. Genetic algorithms are of the most well-known algorithms for optimization and search problems, where a method of “breeding” computer solutions of simulated evolution is used. A population of randomly created individuals initiates the evolution. For every generation, the optimization technique evaluates the fitness of every individual in the population to be selected in the next iteration of the algorithm. The algorithm stops when either a threshold maximum number of generations has been created, or an acceptable fitness level has been achieved for the population [6], [7]. Thus, data mining techniques are used in data preprocessing, where data can be cleaned from outliers by the usage of clustering techniques, and then can be smoothed from noisy values by applying regression techniques. Sampling techniques are one kind of the statistics approaches that are needed in data preprocessing before applying most of the data mining techniques. Sampling is usually used with data mining because processing the entire data set of interest is too expensive and time-consuming [8].

IV. CONCLUSION

Big Data is concerned with the huge amount of data that are continuously growing, besides their unprecedented speed that need to be dealt with in a timely manner. The presence of big data has produced a unique moment in the history of data analysis. With the incremental demand to analyze huge amounts of data, resulting from variant sources and generated at very high rates, researchers at different domains have studied the expansion of the existing data mining techniques to cope with the evolved nature of data and to develop new analytic techniques. In this paper, we provide a detailed comprehensive study of the data mining technique, analyzing the new developments that have been introduced to some of them that have been successfully developed into big data analytic techniques.

REFERENCES

[1] Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013

[2] Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014

[3] Jonathan Stuart Ward and Adam Barker "Undefined By Data: A Survey of Big Data Definitions" arXiv: 1309, 20 Sep 2013.

[4] Laney, Douglas "The importance of „Big Data‟: A definition, Gartner, Retrieved 21 June 2012.

[5] U. Fayyad. Big Data Analytics: Applications and Opportunities, 2012.

[6] Minaei-Bidgoli, Behrouz, and William F. Punch. "Using genetic algorithms for data mining optimization in an educational web-based system." Genetic and Evolutionary Computation—GECCO 2003. Springer Berlin Heidelberg, 2003.

[7] Sastry, Kumara, David Goldberg, and Graham Kendall. "Genetic algorithms." Search methodologies. Springer US, 2005. 97-125.

[8] Hand, David J., "Statistics and data mining: intersecting disciplines." ACM SIGKDD Explorations Newsletter 1.1 (1999): 16-19.