A Review Paper On Data Mining with Big data analysis algorithms, Tools, Applications and Challenges

[1]Sarangam Kodati, [2]Nara Sreekanth, [3]Rajaram Jatothu

[1] Assistant Professor, Department of Computer Science and Engineering, Brilliant Institute of Engineering & Technology - [BRIL], Hyderabad, Telangana ,India.

[2.] Assistant Professor, Department of Computer Science and Engineering, BVRIT Hyderabad College of Engineering for Women, Hyderabad,Telangana, India.

[3] Assistant Professor, Department of Computer Science and Engineering, Brilliant Group of Institutions (Integrated Campus)-[BRIG], Hyderabad, Telangana, India.

ABSTRACT : Huge amounts of data are nowadays collected and stored by organizations with the hope of them being useful in the future. Big data is not only used by organizations to seek better insights for improving the quality of their service and profit, but it can be also used to achieve a variety of targets where success is dependent on the smart analysis. The Big Data introduce unique computational or statistical challenges, including scalability or storage bottleneck, noise accumulation, spurious correlation and measure errors. These challenges are special and require recent computational and statistical paradigm. This paper presents the literature criticism about the Big data Mining and the problems and challenges including emphasis on the distinguished features of Big Data. It also discusses some techniques after deal together with big data. This paper offers an overview regarding big data along with its type, source or characteristics. A review concerning a number of big data mining platforms, algorithms or challenges is additionally discussed in this paper. On the other hand, Big Data additionally arises including many challenges, such as much difficulties among data capture, data storage, data analysis and data visualization. This paper is aimed according to demonstrate a close-up view about Big Data, such as Big Data applications, Big Data opportunities and challenges, as well as like the state-of-the-art methods and technologies we currently adopt to deal with the Big Data problems.

Keywords *— Big data, Bata mining, Big data mining algorithms, big- data challenges, Big-data Tools, Big data applications.*

1. INTRODUCTION

The era of petabyte has come and gone, leaving us to angrily face/stand up to the exabytes time in history now. Technology revolution has been helping millions of people by creating huge/extreme data via ever-increased use of digital devices and especially remote sensors that create continuous streams of digital data, resulting in what is known as "big data". It has been a confirmed important thing/big event that huge amounts of data have been being constantly created at never-before-seen and ever increasing scales. For example according to a survey, Google receives over 2 million queries, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content etc. The main challenge before us is collecting useful information from this huge amount of data. Various technologies are coming up to cope up with these requirements like Cloud computing, Google's model i.e. MapReduce etc. From data mining point of view mining of data from Big Data is a major challenge before us. The extracted information can be useful for making various business decisions and for predicting the future trends. Organizations can make knowledge driven decisions. Various data mining techniques are available for discovery of knowledge from databases and these techniques are often applied with parallel processing architectures and distributed storage systems to improve the performance.

2. BIG DATA

Data is the gathering of values and factors that are related in some sense and contrasting in some other sense. However, lately, the sizes of databases have expanded quickly[7]. Big data is defined by[8] as a data that is mind-boggling as far as volume, variety, velocity as well as its connection to other data, which makes it difficult to deal with utilizing the traditional database management or instruments. A dataset can be called huge information in the event that it is considered to perform securing, curation, analysis, and perception using current innovations[9]. The progression of innovation turns out with the new framework and systems in the organization's information management. The effectiveness of decision-making will progressively be driven by analytic-generated awareness. In this way, the more precise and convenient these are, the better possibility a (human or computerized) chief needs to predict what changes should be done that link with the end goal of the organization[10]. According to[11], "most importantly, big data is a multidisciplinary and developmental combination of new innovations in a mix with new measurements in data stockpiling and handling (volume and speed), another period of a data source variety and the challenge of managing information quality sufficiently (veracity)." However,[12] has defined the characteristics of big data based on 5 Vs, which is summarized as below:

There are many properties associated with bigdata. The prominent aspects are Volume, Velocity Variety, Vercacity and Value.

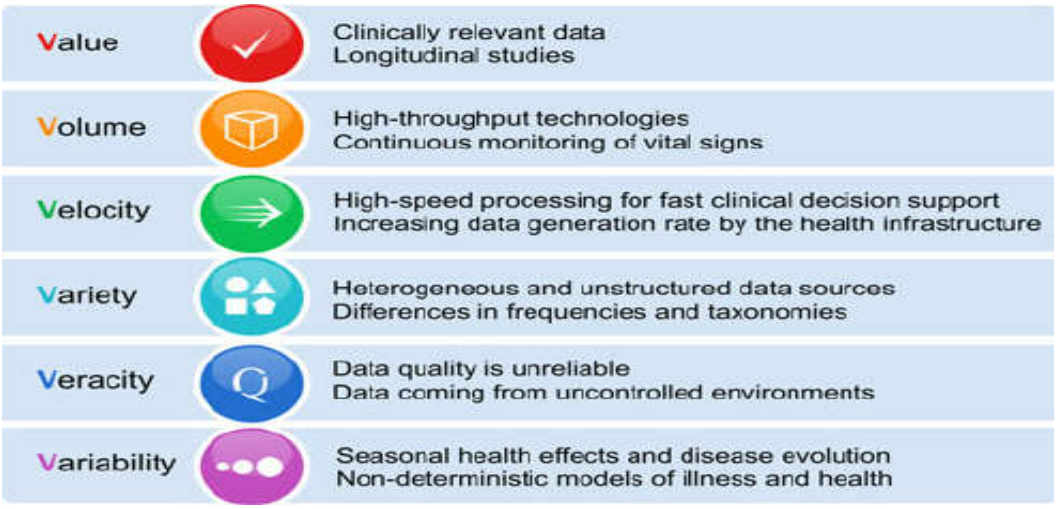

Fig1: Six V's of big data

Volume: The volume of big data is exploding exponentially day to day. The data accumulated through social websites and sensor networks going to cross from petabytes to Zetabytes.

Velocity: This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly.

Variety: Data produced are from different categories, consists of unstructured, standard, semi structured and raw data which are very difficult to be handled by traditional systems.

Vercacity: It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

Value: All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly

they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

3. BIG DATA ISSUES AND CHALLENGES

Big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis regarding Big Data involves more than one distinct phases which include data acquisition or recording, data extraction or cleaning, data integration, quantity and representation, query processing, data modeling or analysis and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

4. BIG DATA AND DATA MINING

Data stored at the server of Facebook according to that amount is used by people into daily existence the place we upload a number concerning kinds on data as pictures, videos and entire about it data stored regarding the warehouse regarding data at the Facebook servers, we called such big- data due to its complexity[4].Big-data is nothing but a data available at autonomous and heterogeneous sources of extreme huge amount which gets up to date inside a fraction over second. Another example concerning big- data we can take like analyzing performed from an electronics microscope of the universe. Now the term Data mining be able stay defined namely extraction of useful data from the collected and gathered data or we execute speech extraction of knowledge from database[5]. So big-data mining is a close on view that contains a bunch on useful detailed facts on big-data.

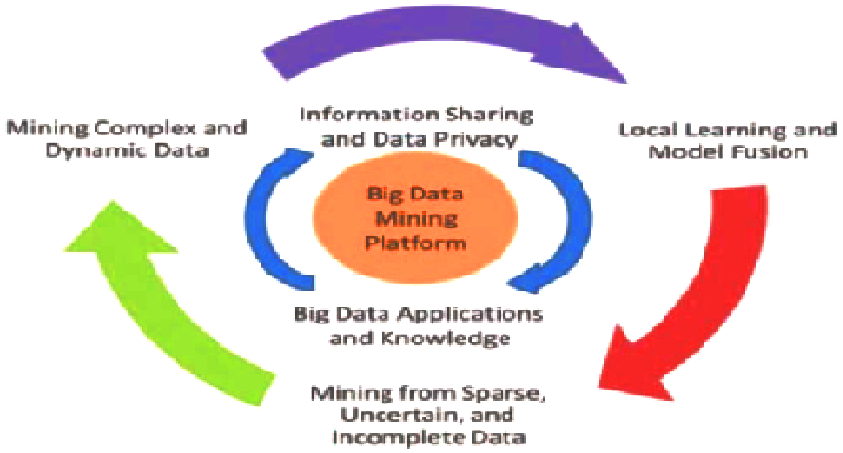

Fig 2: Cycle of big-data mining platform

ADVANTAGE OF BIG-DATA IN VARIOUS APPLICATIONS

- Big-data included: Enterprise data
- Transaction data
- Social media
- Public data
- Sensor data

5.BIG DATA TOOLS.

Large numbers concerning tools are available after technique big data. In this section, we discuss some current strategies for analyzing big data with emphasis of some important emerging tools namely Cassandra.

- Hadoop.
- Plotly.
- Bokeh.
- Neo4j.
- Cloudera.
- OpenRefine.
- Storm.

5.1DATA MINING TASK:

- Classification (Predictive)
- Clustering (Descriptive)
- Association Rule Discovery(Descriptive)
- Sequential Pattern Discovery (Descriptive).
- Regression (Predictive).
- Deviation Detection (Predictive).
- Collaborative Filter (Predictive).

6.DATA MINING TOOLS

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. Furthermore, we propose criteria for the tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. Every tool has its own advantages and disadvantages. [1] Within data mining, there is a group of tools that have been developed by a research community and data analysis. They are offered free of charge using one of the existing open source licenses. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. There are number of open
source tools available for data mining

- Rapid Miner. Availability: Open source.
- Orange. Availability: Open source.
- Weka. Availability: Free software.
- KNIME. Availability: Open Source.
- Sisense. Availability: Licensed.
- Apache Mahout.
- Oracle Data Mining.
- DataMelt.

ADVANTAGE OF DATA MINING IN VARIOUS APPLICATIONS:-

- Banking
- Marketing
- Health Care

- Manufacturing and Production
- Insurance
- Law
- Government and Defense
- Computer hardware and software
- Airlines
- Brokerage and Securities trading.

CHALLENGES FACED BY DATA MINING:-

- Data quality
- Privacy preservation
- Network Setting
- Data Ownership and distribution
- Complex and Heterogeneous data
- Scalability
- Streaming Data
- Dimensionality.

7.BIG DATA ANALYSIS ALGORITHMS MINING ALGORITHMS FOR SPECIFIC PROBLEM

Since the enormous information issues have showed up for almost ten years, in [13], Fan and Bifet brought up that the expressions "big data" [14] and "big data mining" [15] were first exhibited in 1998. The big data and big data mining nearly showing up in the meantime clarified that discovering something from enormous information will be one of the real errands in this domain. Data mining algorithms for data analytics likewise assume the key part in the big data analysis, as far as the calculation cost, memory prerequisite, and precision of the final products. In this area, we will give a short discussion from the viewpoint of examination and pursuit calculations to clarify its significance for big data analytics.

7.1 CLUSTERING ALGORITHMS

In the big data age, conventional bunching calculations will turn out to be considerably more restricted than before in light of the fact that they ordinarily require that every one of the information be in a similar arrangement and be stacked into a similar machine to locate some helpful things from the entire information. Despite the fact that the issue [2] of breaking down large-scale and high-dimensional dataset has pulled in numerous analysts from different traits in the most recent century, and a few arrangements have been exhibited as of late, the attributes of big data still raised a few new difficulties for data clustering issues.

7.2 CLASSIFICATION ALGORITHMS

Like the clustering algorithm for big data mining, a few investigations likewise endeavored to alter the conventional classification algorithms to influence them to take a shot at a parallel figuring condition or to improve new classification algorithms which work normally on a parallel computing environment. In [3], the outline of classification algorithm considered as the information that are assembled by distributed data sources and they will be handled by a heterogeneous set of learners.

7.3 FREQUENT PATTERN MINING ALGORITHMS

The greater part of the researchers on frequency pattern mining (i.e., association rules and sequential pattern mining) were centered around taking care of large-scale dataset at the earliest reference point

since some early methodologies of them were endeavored to analyze the information from the transaction data of big shopping mall. Since the quantity of transactions are more than "tens of thousands", the issues about how to deal with the large scale data were examined for quite a long while, for example, FP-tree [1] utilizing the tree structure to incorporate the frequency pattern to additionally diminish the calculation time of association rule mining.

7.4 COMMUNITY DETECTION ALGORITHMS

Researches on community detection were focused on handling small group dataset at the very beginning because some early approaches of them were attempted to analyse the data. The combination of multiple algorithms depends on top down or bottom up approach many researchers has proved the efficient time complexity in detecting similar communities in wide range of data.

8.CONCLUSION AND FUTURE WORK

Big data is going to continue growing during the next years and each data scientist will have to manage much more amount of data every year. The data is going to be larger, diverse and faster. Many technical challenges like implementations and visualizations are to be taken into consideration in future. This is just the survey paper which shows the demand of big data and how big companies are taking interest in it. We are at the beginning of a new era where big data mining will help us to discover knowledge that no one has discovered before. To manage and analyze edge data explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge system and the potential use for data. This paper reviews about the various big data mining platforms and algorithms. To support big data mining, high-performance computing platforms are required. It is understood that interestingness of discovered patterns, developing a global unifying theory, building efficient mining platforms or algorithms, privacy preserving, security, trust and data integrity are the major challenging issues in the current big data mining scenario. Big Data is becoming the new area for scientific data research and for business applications. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. It can be concluded from the findings to that amount Enterprise are still looking for the right infrastructure tools so much will enable to them after effectively deal with theirs big-data, among row including theirs business needs. Most companies are already the use of dedicated big-data tools however all still see gaps within capabilities or hold concern related to the suit between these tools or their current and expected needs.

REFERENCES

[1] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000. pp. 1–12.

[2] Chiang M-C, Tsai C-W, Yang C-S. A time-efficient pattern reduction algorithm for k-means clustering. Inform Sci. 2011;181(4):356 31. Russom P. Big data analytics. TDWI: Tech. Rep ; 2011.

[3] Tekin C, van der Schaar M. Distributed online big data classification using context information. In: Proceedings of the Allerton Conference on Communication, Control, and Computing, 2013. pp 14

[4] IDC, Extracting Value from Chaos: http://idcdocserv.com/1142, june 2011

[5] O'Reilly Radar, What is bigdata? http://radar.oreilly.com/2012/01/what-is-big-data.html. January 11,2012.

[6] Peter Buneman, Semistructured Data http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf, 1997.

[7] Jaseena K, David J. Issues, challenges, and solutions: Big data mining. Journal of Computer Science and Information Technology. 2014:131–40.

[8] Mouthaan N. Effects of big data analytics on organizations' value creation; 2012.

[9] Chen C, Zhang C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences. 2014; 275:314–7.

[10] Ericsson White Paper. Big data analytics: actionable insights for the communication service provider; 2015.

[11] Buhl H, Röglinger M, Moser F, Heidemann J. Big data. Wirtschaftsinformatik. 2013; 55(2):63–8.

[12] Lawal Z, Zakari R, Shuaibu M, Bala A. A review: Issues and challenges in big data from analytic and storage perspectives. 2016; 5(3):4–6.

[13]Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD ExplorNewslett. 2013;14(2):1–5.

[14] Diebold FX. On the origin(s) and development of the term "big data", Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, Tech. Rep. 2012. [Online]. Available: http://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf.

[15] Weiss SM, Indurkhya N. Predictive data mining: a practical guide. San Francisco: Morgan Kaufmann Publishers Inc.; 1998.