International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue-01
January 2018

Issues and Challenges in Data Mining

Ankur Gupta
Assistant Professor in Computer Science
RSD College, Ferozepur City

Abstract:- *Data is a crucial a part of each business, organization, economy and individual. The term huge knowledge is employed to outline immense quantity of information. the info is therefore immense that the standard management systems ar unable to store, method and analyze. It introduces several challenges, opportunities and analysis topics that need new tools and techniques for process and analyzing of huge knowledge. This paper presents review on the term huge data processing and presents varied challenges on analyzing huge knowledge.*

Keywords:-Data mining, big data, big data mining, knowledge discovery, Hadoop, MapReduce.

I. INTRODUCTION

This paper presents review on the term massive data processing and presents numerous challenges on analyzing massive Data. The era of computer memory unit has come back and gone, effort North American country to angrily face/stand up to the exabytes time in history currently. Technology revolution has been serving to voluminous folks by making huge/extreme information via ever-increased use of digital devices and particularly remote sensors that make continuous streams of digital information, leading to what's called "big data". It's been a confirmed vital thing/big event that massive amounts of knowledge are being perpetually created at never-before-seen and ever increasing scales. For instance consistent with a survey, Google receives over two million queries, YouTube users transfer seventy two hours of video, Facebook users share over two million items of content etc. the most challenge before North American country is aggregation helpful info from this immense quantity of knowledge. numerous technologies square measure developing to cope up with these necessities like Cloud computing, Google's model i.e. MapReduce etc. From data processing purpose of read mining of knowledge from massive information may be a major challenge before North American country.

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.

II. Data Mining

It is a method of finding hidden information and insights from large quantity of knowledge. It finds varied patterns and relationships hidden during this large quantity of knowledge. varied data processing algorithms and techniques square measure obtainable, however these algorithms aren't scalable and techniques aren't able to match the amount, speed and sort of rising knowledge. but these techniques aren't able to add real time, thus new algorithms and techniques square measure needed to figure with large quantity of knowledge} otherwise price of this information are useless. thus new tools square measure needed that may add parallel, that square measure scalable which will add real time having interaction with users.

III. Big Data Mining

The desires of massive facts mining strategies go beyond fetching the asked information or maybe uncovering some hidden relationships and styles between numeral parameters. Studying fast and big move statistics may additionally result in new valuable insights and theoretical concepts [1]. Evaluating with the results derived from mining the conventional datasets, unveiling the big quantity of interconnected heterogeneous huge data has the capability to maximize our knowledge and insights in the target area. but, this brings a sequence of latest demanding situations to the research community. Overcoming the challenges will reshape the destiny of the statistics mining era, ensuing in a spectrum of groundbreaking information and mining techniques and algorithms. One possible approach is to improve present strategies and algorithms by exploiting vastly

parallel computing architectures (cloud platforms in our thoughts). Big facts mining need to deal with heterogeneity, extreme scale, speed, privacy, accuracy, believe, and interactiveness that current mining strategies and algorithms are incapable of. The need for designing and implementing very-massive-scale parallel machine learning and facts mining algorithms (ML-DM) has remarkably improved, which accompanies the emergence of effective parallel and very-large-scale records processing structures, e.g., Hadoop MapReduce. NIMBLE[2] is a transportable infrastructure that has been mainly designed to permit fast implementation of parallel MLDM algorithms, jogging on top of Hadoop. Apache's Mahout[3] is a library of system studying and information mining implementations. The library is also implemented on top of Hadoop using the MapReduce programming model. some vital components of the library can run stand-alone. the principle drawbacks of Mahout are that its mastering cycle is simply too lengthy and its loss of person-pleasant interaction aid. besides, it does not put in force all of the wanted records mining and system getting to know algorithms. BC-PDM (big Cloud-Parallel data Mining)[4], as a cloud-primarily based statistics mining platform, additionally based on Hadoop, gives access to huge telecom facts and commercial enterprise answers for telecom operators; it supports parallel ETL technique (extract, remodel, and cargo), data mining, social
community analysis, and text mining. BC-PDM attempted to conquer the problem of unmarried feature of different strategies and to be more relevant for enterprise Intelligence. PEGASUS (Peta-scale Graph Mining system)[5] and Giraph[6] each implement graph mining algorithms the use of parallel computing and they each run on pinnacle of Hadoop. GraphLab[7] is a graph-based totally, scalable framework, on which sever all graph-primarily based device learning and records mining algorithms are carried out.

IV. *Applications of Data Mining*

1) *Data Mining Applications in Sales/Marketing*
Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way.

2) *Data Mining Applications in Banking / Finance*
Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.

3) Data Mining in Health Care
It has the potential to solve the problems of health. It uses analytics and different machine learning, multi dimensional databases etc. to find the processes to make sure that the patients receive appropriate care at right time.

4) Data Mining in Market Based Analytics
Data mining helps to find out the things that if a person buys a particular thing, he is most likely to buy other things also.

5) Data Mining in Education
Data Mining can be used by an institution to take accurate decisions, predicting student's future learning behavior etc.

6) Data mining by Crime Agencies
It is being used by Crime Agencies to spot trends across myriads of data etc.

V. Issues and Challenges

- Variety and Heterogeneity:- Variety is that the characteristic of massive knowledge. knowledge is collected from several sources could generate knowledge on its own or may contribute to that. It means that there's selection further as non uniformity within the knowledge. These kinds of knowledge area unit interconnected, interconnected and inconsistent. knowledge could also be structured which can slot in the info, semi-structured which can partly slot into the info or unstructured might not match in the info. therefore mining hidden patterns and information from these heterogeneous knowledge may be a challenge before the information scientists.
- Scalability:- Big data requires high scalability of its data management and mining tools. This data may contain knowledge and information which may not be possible to collect from conventional data.
- Speed/Velocity:- The data mining rule should be able to end the process among a specific time. The info

should be accessed and processed quickly otherwise the results obtained from these can become no-count. The factors that have an effect on the speed. Data mining depends embody data time interval and potency of mining algorithms. correspondence is also enclosed to extend the accessing speed.

- Accuracy and Trust:- With an increase in the amount of data, there are many data sources from where data is collected, which may not be verifiable or trustable. Therefore the accuracy and trust of data source becomes an issue, which may propagate to results as well. Therefore data validation and accuracy becomes an important issue for discovery of useful information.

- Privacy:- It is a vital issue that knowledge should be unbroken personal and invisible to others. Data processing needs personal data so as to supply results. Social media contains all of the data concerning a private and data will be well-mined from that information then the privacy disappears. Therefore this is often the problem that should be thought-about by knowledge scientists and tools should be designed by taking this thought under consideration.
- Interactiveness:- It means feature of the data mining system that allows user interaction by using feedback/guidance. It is an important issue as it allows the users to visualize, evaluate and interpret intermediate and final results.

VI. Conclusion

In this era data is generated at unprecedented speed. This paper presented the limitations in existing data mining techniques used for data mining. More work is required to be done to cope with the challenges related to it. New techniques must be developed and parallelism must be used for improving the speed of analysis and accessing of data. We are in the beginning of era where Big data mining allows us to discover knowledge and use that knowledge for different applications.

References

[1] Berkovich, S., Liao, D.: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York (2012).

[2] Ghoting, A., Kambadur, P., Pednault, E., Kannan, R.: NIMBLE: a Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on MapReduce.In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.334-342, San Diego, California, USA (2011).

[3] Mahout, http://lucene.apache.org/mahout/.

[4] Yu, L., Zheng, J., Shen, W.C., et al: BC-PDM: Data Mining, Social Network Analysis and Text Mining System Based on Cloud Computing. In: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1496-1499 (2012).

[5] Kang, U., Tsourakakis, C.E., Faloutsos, C.: PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations In: 9th IEEE International Conference on Data Mining, pp. 229-238 (2009).

[6] Apache Giraph Project, http://giraph.apache.org/.

[7] Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., Hellerstein, J.M.: Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. VLDB Endowment, vol. 5, no. 8, pp.71-727 (2012).

[8] Madden, S.: From Databases to big data. In: IEEE Internet Computing, vol. 16, no. 3, pp. 4-6. IEEE Computer Society (2012).

[9] B R Prakash, Dr. M. Hanumanthappa. Issues and Challenges in the Era of Big Data Mining, Volume 3, Issue 4,pp 321-325, IJETTCS(2014).

[10] Dunren Che , Mejdl Safran , and Zhiyong Peng. From Big Data to Big Data Mining: Challenges, Issues and Opportunities,pp.1-15.

[11] Jaseena K.U and Julie M. David, ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING, Natarajan Meghanathan et al. (Eds) : NeTCoM, CSIT, GRAPH-HOC, SPTM – 2014 pp. 131–140, 2014.

[12] Fayyad, U.M., Gregory, P.S., Padhraic, S.: From Data Mining to Knowledge Discovery: an Overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1-36. AAAI Press, Menlo Park, CA (1996).