ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

An Analysis and Review on Various Techniques of Mining Big Data

Hitesh Kumar Bhatia
Masters of Computer Applications Sardar Patel Institute of Technology

Manuscript Info

Manuscript History:

Received: 15 May 2015
Final Accepted: 29 June 2015
Published Online: July 2015

Key words:

Big Data; Data Mining; HACE Theorem; Clustering; Parallel computing.

Corresponding Author

Hitesh Kumar Bhatia

Abstract

Data Mining refers to extracting knowledge from the database. In today's world, information stored in database is so large that there is need to develop a technique by which information can be mined efficiently. Data Mining is rapidly growing in all domains. Big Data is collection of large amount of heterogeneous and unstructured data. There are many challenges for big data such as storage, searching, privacy and security of data. In this paper we have compared various mining techniques which can be used for mining of information of Big Data. We have also analysed and described the advantages and disadvantages of various technique used for mining of Big Data which can be used for further developing of various techniques in an efficient manner.

INTRODUCTION

Big Data is described as large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [1]. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.

Data Mining is the technology to extract the knowledge from the pre-existing databases. It is used to explore and analyze the same. The data which is to be mined varies from a small dataset to a large data-set i.e. Big Data. Big data is so large that it does not fit in the main memory of a single machine, and that it need to process big data by efficient algorithms. Modern computing has entered the era of Big Data.

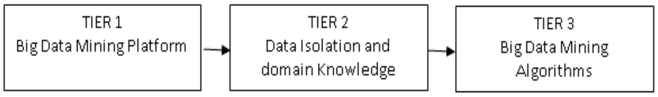
TIER 1
Big Data Mining Platform
TIER 2
Data Isolation and domain Knowledge
TIER 3
Big Data Mining Algorithms

Figure 1: Data Mining Processing (3 TIER Architecture)

In Figure 1, processing of Data Mining is shown in the form of 3 Tier Architecture. Tier 1 deals with parallel computing as single computer cannot handle big data, we require high performance computing. Tier 2 deals with the way data mining algorithm is designed. We need to find feasible business solution for data mining on big data, for this purpose Tier 2 works efficiently. Tier 3 deals with mining partial, complex as well as dynamic data. It also deals with semantic association between the data as well as complex relation in data.

The massive amounts of information available on the Internet enable computer scientists, physicists, economists, mathematicians, political scientists, bioinformaticists, sociologists, and many others to discover interesting properties about people, things, and their interactions. Analyzing information from Twitter, Google, Facebook, Wikipedia, or the

Human Genome Project requires the development of scalable platforms that can quickly process massive-scale data. Such frameworks often utilize large numbers of machines in a cluster or in the cloud to process data in a parallel manner.

Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.
There are two types of big data: **structured and unstructured**.

Structured data are data (may be numbers or words) that can be easily categorized and analysed. These data are generated by various sources such as Smartphone, and global positioning system (GPS) devices, sales figures, account balances, transaction data etc.
Unstructured data have more complex information than structured which includes customer reviews from various websites, photos and other multimedia, and comments from social networking sites. These data are difficult to categories or analyse numerically.
The most important 5 V's of big data are volume, velocity, value, veracity and variety [9].

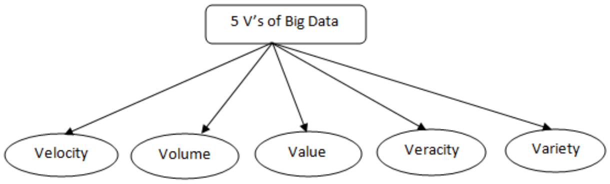

Figure1: 5V's of Big Data

Volume –Volume refers to the vast amounts of data that is generated every second. With large amount of data created all around the world in many years, it's a big challenge to handle large volume of data.

Velocity –Velocity refers to the speed at which new data is generated and the speed at which it moves around. Reacting fast enough and analyzing the streaming data is troubling to businesses, with speeds and peak periods often inconsistent.

Variety – Refers to the different forms of data that we collect and use. Data comes in different formats, such as structured and unstructured. Mostly data over the real world are unstructured and in various formats.

Veracity – Veracity refers to the uncertainty surrounding data, which is due to data inconsistency and incompleteness, which leads to another challenge, keeping Big Data organized.

Value – Through effective data mining and analytics, the massive amount of data that we collect throughout the normal course of doing business can be put to good use and yield value and business opportunities.

By applying various data mining techniques and analytics to expose valuable business information embedded in structured, unstructured, and streaming data and data warehouses, this insight can be used to help revamp supply chains, improve program planning, track sales and marketing activities, measure performance across channels, and transform into an on-demand business.

Literature Review

Big Data is considered as an emerging trend in today's world. To mine big data many techniques have been invented by many authors which gives efficient way to mine the data. Big data is emerging in all fields of science and engineering domains.

Xingdong Wu et al [1], have presented HACE Theorem which describes all the features of how Big Data evolved. Big Data framework is divided into three tier architecture where each tier has its own unique functionality and working. Yanfeng Zhang et al [2], proposed a novel incremental method which is an extension of MapReduce known as I2MapReduce which reduces I/O overhead for accessing preserved fine-grain computation states. It performs key-value pair level incremental processing rather than task level re-computation which makes the novel technique improve performance and give optimized result for mining the data. It merges K-means algorithm with MapReduce for better results.

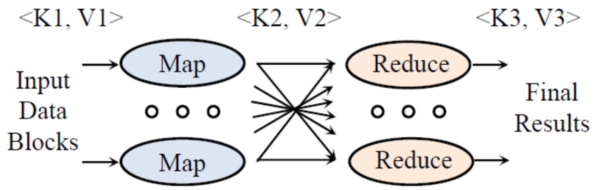

Figure 3 [2]: Map Reduce Computation

Jie Xu et al [5], proposed a technique for mining the information over the web for prediction of user's behaviour. When information is large enough to store, it is difficult to know the user behaviour over the web, however the proposed framework reduces the implementation complexity over the real world data.

Yang Liu et al [8], proposed a Bulk Synchronous Parallel (BSP) model which combines four parallel graph mining algorithm for efficient and better performance than any other mining algorithm. It has the ability to analyze big graph data and achieved a better performance than the Hadoop-based data mining tools BC-PDM and BSP based parallel platform BC-BSP.

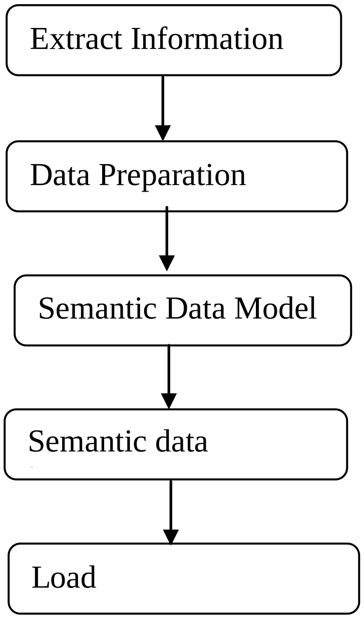

Figure 4: Extraction Transformation and Loading (ETL) Process

Srividya K. Bansal et al [11], uses Extraction, Transformation and Loading (ETL) process to integrate data from heterogeneous sources into meaningful semantic model and then can apply various mining algorithms. The extract transform load (ETL) framework is used for integrating data from multiple sources or applications, possibly even from different domains.

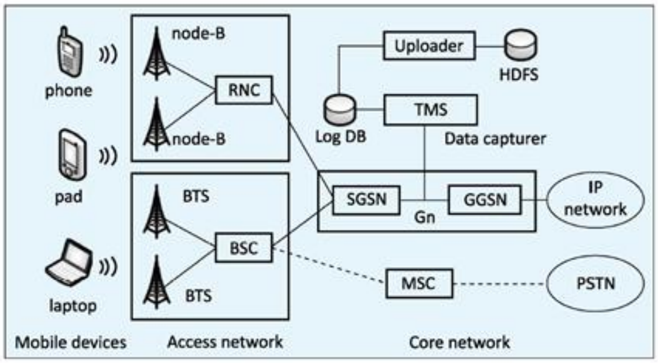

Figure 5 [4]: Cellular network

Liu Jun et al [4], proposed a Zipf-like model which analyses the behaviour of user over the mobile internet era and helps to navigate easily through pages. It uses diurnal pattern clustering algorithm to cluster the data into k-cluster centres. These clustering patterns reveal various usage patterns of the user which has helped to find out hidden patterns and understand user patterns.

Extracting useful knowledge from huge digital datasets requires smart and scalable analytics services, programming tools, and applications[10]. Various data mining techniques and associated tools can help extract information from large, anormous complex datasets that is useful in making informed decisions in many business and applications including tax payment collection, market sales, social studies, biosciences, and high energy physics. Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new insights in a shorter time. Using a synergic approach that integrates the use of clouds and data analysis techniques to investigate key issues will help researchers achieve this goal [10].

Despite clear technological advances, research challenges must be solved to realize a standard large-scale, Quality of Service (QoS) optimized platform for managing streaming big data analytics ecosystem[11]. The data processing technologies are designed to maintain an efficient and fault-tolerant collection of data that is accessed and aggregated only when users issue a query or transaction request by the user.

Comparative study of various Mining algorithm for Big Data Mining

Sr. no	Authors	Technique Used	Advantages	Disadvantages
1	Xindong Wu, Xingquan Zhu,Gong-Qing Wu, and Wei Ding (2014)	HACE Theorem	Works well with huge and diverse data	Complicated to implement.
2	Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu (2014)	I2 Map Reduce	Can significantly reduce the run time for refreshing big data mining results compared to recomputati	Complexity Increases

			onon both plain and iterative MapReduce.	
3	Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, Mihaela van der Schaar (2015)	Context-aware Traffic Prediction	Self-adapting to changing environment, Optimal Prediction	It doesn’t work with distributed environment
4	Yang Liu, Bin Wu, Hongxu Wang, and Pengjiang Ma(2014)	BSP-based Parallel Graph Mining (BPGM)	Better performance over cloud than other hadoop data mining tools.	It limits the scale of data.
5	Srividya K. Bansal, Sebastian Kagemann (2015)	Semantic ETL process	It has significant potential to produce linked data that supports innovative datadrivenapps for smart living.	Sometimes it requires manual data analysis which is time consuming
6	LIU Jun, LI Tingting, CHENG Gang, YU Hua, LEI Zhenming (2013)	Zipf-like model	The clustering method used helps us to better understand user behaviour and pattern.	The patterns generated sometimes are not correct and may lead to incorrect understanding.

Conclusion

Big data deals with large volume of data which is heterogeneous, autonomous and complex. However Mining of big data is a challenging task. We have compared various algorithms and techniques which can be used for mining various kinds of big data. We have analysed and stated advantages and disadvantages of all the techniques described in the paper. The comparison gives a better idea about the techniques and approaches developed so far for mining of big data on various fields and domain which can help researches develop novel techniques in an efficient manner.

References

[1] Xindong Wu, Xingquan Zhu,Gong-Qing Wu, and Wei Ding, “Data Mining with Big Data”, IEEE transactions on knowledge and data engineering, vol. 26, no. 1, January 2014

[2] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu ,”i^2MapReduce: Incremental MapReduce for Mining Evolving Big Data”, IEEE Transactions On Knowledge And Data Engineering, Vol. , No. , January 2014.

[3] Mirco Musolesi, “Big Mobile Data Mining: Good or Evil?”, IEEE Computer Society January/February 2014.

[4] LIU Jun, LI Tingting, CHENG Gang, YU Hua, LEI Zhenming ,“Mining and Modelling the Dynamic Patterns of Service Providers in Cellular Data Network Based on Big Data Analysis”, China Communications , Ict Industry Convergence, December 2013

[5] Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, Mihaela van der Schaar, “Mining the Situation: Spatiotemporal Traffic Prediction with Big Data”, DOI 10.1109/JSTSP.2015.2389196, IEEE Transaction, 2015.

[6] Ming-Syan Chen, Jong Soo Park, Philip S. Yu, “Efficient Data Mining for Path Traversal Patterns” IEEE Transactions On Knowledge And Data Engineering, Vol. 10, No. 2, March/April 1998.

[7] Yang Liu, Bin Wu, Hongxu Wang, and Pengjiang Ma, “BPGM: A Big Graph Mining Tool”, TSINGHUA SCIENCE AND TECHNOLOGY, ISSN l l1007-0214l l04/10l lpp33-38, Volume 19, Number 1, February 2014.

[8] Srividya K. Bansal, Sebastian Kagemann ,“Integrating Big Data: A Semantic Extract- Transform-Load Framework”, Cover Feature Big Data Management, IEEE Computer Society, 2015.

[9] Deepak S. Tamhane, Sultana N. Sayyad, “BIG DATA ANALYSIS USING HACE THEOREM”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 1, January 2015

[10] Domenico Talia ,“Clouds for Scalable Big Data Analytics”

, published by the IEEE Computer Society, 2013.

[11] , Rajiv Ranjan , “Streaming Big Data Processing in Datacenter Clouds”

IEEE Cloud ComputIng published by thE IEEE Computer society