Adaptive Approach of Data Mining Using HACE Algorithm

Miss. Megha A.Chandak
Student, Computer Science and Engineering,
P.R. Patil COET, Amravati,
Maharashtra, INDIA.
megha9890@gmail.com

Prof. Ajay B. Gadicha
Assistant Professor, Computer Science and Engineering,
P.R. Patil COET, Amravati,
Maharashtra, India.
ajjugadicha@gmil.com

Abstract:-Data mining is an interdisciplinary subfield of computer science, is the computational process of discover patterns in large data sets involving methods at the intersection of artificial intelligence , machine learning, statistic, and database systems. Big Data is a new term used to identify the datasets that due to their large size and complication. Data comes from everywhere, sensors used to gather climate information, post to social media sites, digital pictures and videos etc. this data is known as big data. Big Data concern large-volume, difficult, growing data sets with many, independent sources. With the fast development of networking, data storage, and the data group ability, Big Data is now fast expanding in all science and work domains, including physical, biological and bio-medical science. This paper gives brief idea about a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining point of view.

Keywords: *Big data, Volume, Variety, Velocity, Variability, Value, heterogeneous, autonomous, complex, evolving..*

*****

I. INTRODUCTION

Data Mining is the method of discovering interesting knowledge, such as pattern, association, change, anomalie and important structures, from huge amounts of data stored in database, data warehouses, or other information repositories. Due to the large availability of huge amounts of data in electronic forms, and the imminent need for turn-off such data 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.

Data Mining is the extraction, predictive information from huge database is knowledge of discovery using sophisticated blend of technique from a established statistics, artificial intelligence and computer graphics. To best apply the advanced technique, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools at present operate outside the warehouse require extra steps for extract, importing and analyzing the data. Furthermore, when new insight require operational operation, integration with the warehouse simplifies the application of results from data mining. [Sharayu S. Sangekar 2014]

Objective:

1) To provide huge volume of data having heterogeneous and diverse dimensionalities.
2) To provide distributed and decentralized control on data.

Big data:

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges includes capture, screen, storage space, search, sharing, transmit, analysis and visualization. The trend to big data sets is due to the additional information derivable from analysis of a single large set of related data, as compare to separate smaller sets with the same total quantity of data, allow connection to be found to “spot big business trends, determine feature of research, prevent diseases, link , conflict crime, and determine real-time roadway traffic conditions”. [D. S. Tamhane, S. N. Sayyed 2015]

There are two types of big data: structured and unstructured.

Structured data are numbers and words that can be easily categorize and analyze. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also includes thing likes sales information, account balances, and transaction data.

Unstructured data include more difficult information, such as customer reviews from advertisement websites, photos and other multimedia, and comment on social networking sites. These data can not easily be unconnected into categories or analyzed numerically.[B. Thakur,M. Mann 2014]

Fig1. Big Data

II. . LITERATURE REVIEW:

Xindong, et al “Data Mining with Big Data”, proposed HACE theorem that characterize the features of the Big Data revolution, and propose a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information source, mining and analysis, user interest modeling, and security and privacy considerations. We examine the challenging issues in the data-driven model and also in the Big Data revolution [1].

Bharti Thakur, Manish Mann,” Data Mining for Big Data A Review”, proposed data mining is a technique for discover exciting patterns as well as expressive, understandable models from large scale data. Overviewed types of big data and challenge in big data for future [2].

Deepak S. Tamhane, Sultana N. Sayyad,” Big Data Analysis Using Hace Theorem”, proposed HACE theorem that states the independence of the Big Data revolution, and proposes a Big Data processing model from the data mining view. This data-oriented model contains demand-driven aggregation of data sources, mining and study, user knowledge modeling, and safety and privacy issues. [3].

J.Josepha Menandas, J.Jakkulin Joshi,” Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data”, proposed a parallel processing technique (PPT) that characterize the features of big data revolution, reduces complexity, and proposes a large data processing model from the data mining point of view [4].

Vitthal Yenkar, Prof.Mahip Bartere, “Review on Data Mining with Big Data” proposed paper discusses a characterize applications of Big Data processing model and Big Data revolution, from the data mining view. The analysis of big data can be hard because it often involves the collection and storage of mixed data based on different patterns or rules (heterogeneous mixture data). This has made the heterogeneous mixture property of data a very important issue. This paper introduces heterogeneous mixture learning, We study the hard issues in the Big Data revolution and also in the data-driven model [5].

Dinesh D. Jagtap,” Big Data using Hadoop”, “Big Data” is data that becomes big enough that it cannot be processed using conservative method. The term Big Data concerns with the huge volume, complex and rapidly growing data sets with multiple, independent sources .Due to fast development of networking ,data storage and data collection capacity the concept of large data is now rapidly expanding in all science and engineering domains including biological, physical and biomedical sciences[6].

S. S. Sangekar, P. P. Deshmukh,” Data Mining Of Complex Data With Multiple, Autonomous Sources”, The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining soft- ware tools [7].

Smitha T, MCA, V. Suresh Kumar,” Application of Big Data in Data Mining”, The proposed paper mainly focussed different types of big data and its application in information discovery[8].

Sherin A, Dr S Uma, Saranya K, Saranya Vani M(2014)”Survey On Big Data Mining
Platforms, Algorithms And Challenges”, proposed paper overview of big data along with its type, source and characteristic and challenges is also discussed [9].

III. The Five V’s In Big Data :

The term “Big Data” has served as a catch all phrase for the large amount of information available and collected in the digital world. Big Data is being called the increasing power of the 21st century and is helping as much more than a catchword, acting as a high performance Inter-Thread Messaging Library in the technology way. It shows Five V’s in Big Data.

Volume: There is additional data than ever before; its size continues increasing, but not the percent of data that our tools can process.

Variety: There are different types of data, as text, sensor data, audio, video, graph, and more.

Velocity: Data is continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Variability: There are change in the organization of the data and how users want to interpret that data.

Value: Big 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.

[B. Thakur, M. Mann 2014]

IV. Big Data Characterization (HACE theorem):

The characteristics of Big Data are heterogeneous, autonomous sources with disseminated and decentralized control, and seek to explore difficult and evolving relationships among data. These characteristics make it great challenge for discovering useful knowledge from the Big Data.

4.1 Huge data with heterogeneous and diverse dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionality. This is because different information collector prefers their own schemata or protocols for data recording, and the nature of dissimilar applications also results in diverse data representation.

4.2 Autonomous sources with distributed and decentralized control

It is a main characteristic of Big Data application. Being autonomous, each data source is able to generate and collect information without involving any central control. This is parallel to the World Wide Web (WWW) setting where each web server provide a certain quantity of information and each server is able to fully function without of necessity relying on other servers. On the other hand, the huge volumes of the data also make an application weak to attacks or malfunction, if the whole system has to rely on any central control unit. For major Big Data-related application, such as Google, Flicker, Face book, and Wal-Mart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such independent sources are not only the solution of the technological design.

4.3 Complex and evolving relationships

While the volume of the Big Data increases, so do the complexity and the relationship underside the data. In an early stage of data centralized information systems, the focus is on finding greatest feature values to represent each examination. For reducing complexity, most frequently accessed data should be kept in a separate servers, so that to make active completely. For ex, major social network sites, such as Face book or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlation between individuals inherently complicates the whole data representation and any reasoning process on the data. The features used to represent the individuals and the social sites used to represent our connections may also develop with respect to temporal, spatial, and other factor. [J. J. Menandas 2014]

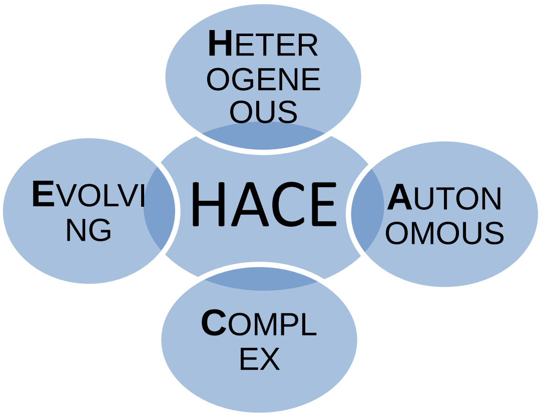

Fig2. HACE diagram

Difference between Data Mining and Big Data

Data mining	Big data
Data mining refers to the activity of going through big data set to look for relevant information	Big data is a term for large data set.
Data mining is the handler which provides beneficial result.	Big data is the asset.
Data mining refers to the operation that involve relatively sophisticated search operation	Big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.

V. CHALLENGES IN HANDLING IN BIG DATA:

It includes in technology needs new architecture, algorithms, techniques for its implementation. It also requires technical skills .So experts are needed for this new technology to deal with big data.

The most important challenges by the enterprises or media when handling Big Data are capturing of big data, its duration, storage, sharing of big data and its analysis, visualization of the massive data etc.

5.1 Application O f Big Data In Data Mining

Data mining a number of different data repository can be concerned. Data mining should be applicable to any kind of data repository as well as to transient data such as data stream. The challenges and techniques of mining may differ for each of the storage systems.

Advanced databases or information repositories require sophisticated services to powerfully store recover and update large amounts of complex data. They also provide fertile grounds to raise many challenging research and implementation issue for data mining.

For data mining in object relational system, techniques need to be developed for handling complex object structures, complex data types, class and sub class hierarchies, property inheritance and method and measures. Data mining technique can be used to find the uniqueness of object evaluation or the trend of changes for objects in the database. Such information can be valuable in decision making and approach planning. For example stock exchange data can be mined to uncover trends that could help to plan investment strategies.

They are useful for vehicle navigation. Spatiotemporal database that change with time is also a big data in which information can be mine. They may be huge infinite volume, dynamically changing in nature.

5.2 Different types of data mining system

There are special types of data mining system which can be used with big data. The main techniques used with data mining are as follows.

5.2.1 Classification

Classification is the process of finding a form or function that describes and distinguishes data classes or concepts, for the function of being able to use the model to predict the class of objects whose class tag is unknown. The derived model is based on the analysis of benefit of preparation data. The model can be representing in various forms such as classification rules, decision tree, mathematical formula or neural networks.

5.2.2. Evolution analysis

Evolution analysis is used with time series data of preceding years. Regularities in such point in time series data is used to predict future trend in stock market prices, causal to decision making with reference to supply investment.

5.2.3. Outlier Analysis

Outlier study may be detected using statistical tests that assume a distribution or probability model for the data or using distance measures where objects that are a substantial distance from any other cluster are considered outliers.

5.2.4. Cluster analysis

In cluster Analysis, there are no class labels in the training data sets. The labels are generating using this technique. The objects in a cluster are grouped based on their correspondence. Then rules are formed from the clusters .The most important cluster methods include portioning methods, hierarchical method, thickness based method, model based method and constraint based clustering method. [Smitha T, V. S. Kumar 2013]

VI. Conclusions

Big data is the term for a group of difficult data sets. It accurately related to data volumes, our heterogeneous, autonomous, Complex and evolving apply the key characteristics of the Big Data are huge with various and diverse data sources, and independent with spread and decentralized control, and hard and developing in data and knowledge associations.

References

[1] Xindong Wu ,Gong-Quing Wu and Wei Ding(2014) “Data Mining with Big data” IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1.

[2] Bharti Thakur, Manish Mann(2014)” Data Mining for Big Data A Review”.International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5.

[3] Deepak S. Tamhane, Sultana N. Sayyad(2015)” Big Data Analysis Using Hace Theorem”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 1.

[4] Josepha Menandas, J.Jakkulin Joshi(2014)” Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data”,Global Journal of advanced research Vol-1,Issue-1PP. 69-80 No 30.

[5] Vitthal Yenkar, Prof.Mahip Bartere(2014) “Review on Data Mining with Big Data”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, pg. 97-102.

[6] Dinesh D. Jagtap(2014)” Big Data using Hadoop”, International Journal of Engineering Research and General Science Volume 2, Issue 6.

[7] S.S.SANGEKAR, P.P.DESHMUKH(2014)” Data Mining Of Complex Data With Multiple, Autonomous Sources”, International Journal Of Pure AndApplied Research In Engineering Andtechnology, Volume 2 (9): 793-799.

[8] SMITHA T, MCA, V. Suresh Kumar(2013)” Application of Big Data in Data Mining” ,International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 7.

[9] Sherin A, Dr S Uma, Saranya K, Saranya Vani M(2014)”Survey On Big Data Mining Platforms, Algorithms And Challenges”. International Journal of Computer Science & Engineering Technology,Vol. 5 No.