Big Data Factors on Data Mining

A.Swarupa Rani
Research Scholar, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, AP, India.
E-mail: swaruparani_kanta@yahoo.com

ABSTRACT--The major aim of this paper is to make a review on the idea Big information and its application in information mining. Big Datais a new term used to identify the datasets thatdue to their large size and complexity, we cannot managethem with our current methodologies or data mining soft-ware tools.Big Data miningis the capability of extractinguseful information from these large datasets or streams ofdata, that due to its volume, variability, and velocity, itwas not possible before to do it. The Big Data challengeis becoming one of the most exciting opportunities for thenext years. We present in this issue, a broad overview ofthe topic, its current status, controversy, and forecast tothe future. The paper principally thinking distinctive sorts of huge information and its application in learning revelation.

Index Terms -- Aggregation, Big data, Data Mining, Data streams, Statistics.

I. INTRODUCTION

Huge information is enormous volume of both organized and unstructured information from different sources, for example, social information, machine produced information, customary venture which is large to the point that it is hard to prepare with conventional database and programming systems. Enormous Data will be information whose scale, differences, and unpredictability require new design, systems, calculations, and investigation to oversee it and concentrate esteem and concealed learning from it. Qualities of Big Data incorporate 4 Vs. They are Volume, Velocity, Variety and Veracity. Huge information is fundamentally used to spot patterns, to decide the nature of research, to anticipate ailment, to interface lawful reference and so on. It is utilized as a part of various applications, for example, Medicine, Physics, Simulation, RFID, Astrometry, Biology and so forth. There are diverse sorts of information, for example, social, basic, literary, semi organized, chart information, spilling information and so on can be incorporated into enormous information. This information can be utilized for Aggregation and Statistics in Data stockroom and OLAP, Indexing, Searching, and Querying, Keyword based looking, Pattern coordinating (XML/RDF), Knowledge revelation in Data Mining and Statistical Modeling.

II. MATERIALS AND METHODS

Enormous information incorporates organized, semi-organized, and unstructured information. This unstructured information contains valuable data which can be mined utilizing appropriate information mining innovation. We can see that the advanced streams that people make are developing quickly. The majority of the general population are utilizing camera all alone portable. Enormous Data are of abnormal state volume, high speed, and high assortment of data that necessities propelled technique to handle the Big Data. Also, the ordinary programming apparatuses are not fit for taking care of such information. Enormous Data requires broad engineering moreover.

Distinctive sorts of information, for example, Social information – Customer input frames for Customer Relationship Management (CRM) in Social media locales, for example, Twitter, Face book, LinkedIn and so on. Machine-created information in sensor readings, satellite correspondence. Conventional undertaking information, for example, and record data. worker data, client data and so on are alluded as large information.

2.1 Characteristics of Big Data

There are mainly four characteristics for big data. They are Volume, Velocity, Variety and Veracity.Volume means vast amount of data generated in every second. It is a scale characteristic. The data is in rest state. Machine generated data are examples for these characteristics. Nowadays data volume is increasing exponentially. The second generated characteristics of big data are velocity or speed. Velocity is the speed at which data generated. The streaming data may not be massive and its state is in motion. It should have high speed data. Example is data created through social media. The data is begin generated fast and need to be processed fast. Online Data Analytics includes these types of big data. E-Promotions and health care monitoring are examples. In e-promotion, based on our current location and our purchase history, what we like will send promotions right now for store next to us. In Healthcare monitoring, sensors monitoring our activities and body. Any abnormal measurements require immediate reaction can be immediately identified through this.

Assortment is another critical normal for enormous information. Different information configurations, sorts, and structures can be alluded here. The kind of information may incorporate distinctive varities, for example, Text, numerical, pictures, sound, video, arrangements, time arrangement, online networking information, multi-diminish clusters, and so on… It likewise incorporates s static information and spilling information. A solitary application can create by gathering many sorts of information. To remove the information every one of these sorts of information should be connected together. Veracity implies information in uncertainty. The vulnerability of information can be found because of the irregularity and deficiency. The chaos of information (Abbreviation, casual discourse and so on) may come about the veracity.

III. CHALLENGES IN HANDLING BIG DATA

The difficulty in taking care of huge information incorporates into innovation. The innovation needs new design, calculations, strategies for its usage. It likewise requires specialized aptitudes .So specialists are required for this new innovation to manage huge information. The adjustment and relationship of information makes more intricacy. The primary difficulties should be confronted by the endeavors or media when taking care of Big Data are catching of huge information, its term, stockpiling, sharing of huge information and its investigation, perception of the monstrous information and so forth. Association and connection of information which depicts more about relationship among the information.

3.1 Application of Big Data in Data Mining

In data mining a number of different data repositories can be involved. Data mining should be applicable to any kind of data repository as well as to transient data such as data streams. The challenges and techniques of mining may differ for each of the repository systems.Advanced databases or information repositories require sophisticated facilities to efficiently storeretrieve 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 methods and procedures. Data mining techniques can be used to find the characteristics of object evaluation or the trend of changes for objects in the database. Such information can be useful in decision making and strategy planning. For example stock exchange data can be mined to uncover trends that could help to plan investment strategies. [13]

Geographic databases have also numerous applications ranging from forestry and ecology planning to providing public service information regarding the location of cables, pipes or sewage system. They are also useful for vehicle navigation. Spatiotemporal database that change with time is also a big data in which information can be mined. [10] Streams of data flow in and out of an observation pattern dynamically. They may be huge infinite volume, dynamically changing in nature.Usually multi level, multidimensional on-line analysis and mining should be performed on stream data. Even if the web pages are fancy and informative to readers, they can be highly unstructured and lack pattern. Data mining can often provide additional help to the web search services which include big data. Data mining are used to specify the kind of patterns to be found in data mining task. The tasks can be classified as predictive and descriptive.

3.2 Different types of data mining system

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

3.2.1 Classification

Order is the way toward finding a model or capacity that depicts and recognizes information classes or ideas, with the end goal of having the capacity to utilize the model to foresee the class of items whose class mark is obscure. The inferred model depends on the examination of benefit of preparing information. The model can be spoken to in different structures, for example, characterization rules, choice tree, scientific formulae or neural systems. Maybe order and expectation ought to be gone before by significance investigation, which endeavors to distinguish traits that don't add to the characterization or forecast handle. These traits can then be barred.

3.2.2. Evolution Analysis

Evolution analysis is used with time series data of previous years. Regularities in such time series data is used to predict future trends in stock market prices, contributing to decision making regarding stock investments. In one activity, entitled “Evolution in Action: Graphing and Statistics,” students are guided through the analysis of this sample of the Grants’ data by constructing and interpreting graphs, and calculating and interpreting descriptive statistics. The second activity. “Evolution in Action: Statistical Analysis,” provides an example of how the data set can be analyzed using statistical tests, in particular the Student’s t-Test for independent samples, to help draw conclusions about the role of natural selection on morphological traits based on measurements.

3.2.3. Outlier Analysis

Outlier analysis 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.

3.2.4. 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 similarity. Then rules are formed from the clusters .The major clustering methods includes portioning methods, hierarchical methods, density based methods, model based methods and constraint based clustering method.If the cluster contains large number of data or big data, then it has to used methods like frequent pattern based clustering or high dimensional data clustering.

IV. TOOLS FOR HANDLING BIG DATA

There are many tools are currently available for handling big data some of them are follows. Map Reduce [2] is a programming model for handling complex combination of several tasks and it was published by Google. It is a batch query processor and can run an ad hoc query for whole dataset and get the results in a sensible manner which has to be transformative. It has two steps. 1. Map: Queries are divided into sub queries and allocated to several nodes in the distributed system and processed in parallel. 2. Reduce: Results are assembled and delivered. Oracle has introduced the total solution for the scope of enterprise which requires Big Data. Oracle Big Data Appliance[3] is a tool to integrate optimized hardware and extensive software into Oracle Database 11 to endure the Big Data challenges. The Real-time application of Big Data can also be in Patient Health Information System on Cloud[4]. Patient Health Record (PHR) is an emerging technique to store the Patient Heath Information Record and exchange the data over the network, which is stored at the cloud for accessing the data log anytime and anywhere. To assure more security individuals are given with their own login and data stored over the cloud would be encrypted. PHR includes variety of data such as structured, unstructured, and semi-structured. In PHR, we propose machine generated data by acquiring the finger print or iris pattern or face of the patient for saving the entire data log of the patient. It uses finger print sensor or Iris scanner or face recognizer for capturing the patient Identification. Finger print or iris pattern or facial expression act as a key for retrieving the data saved in the database.

V. RESULTS

Huge Data are utilized to be incorporated for finding the client conduct, for distinguishing the market patterns, for expanding the advancements, for holding the clients, for playing out the operations proficiently. Surge of information originating from many sources must be taken care of utilizing some non-conventional database devices. It gives more market esteem and orderly for the up and coming era. Huge information has an assortment of utilization and impact in the field of information mining.

VI. CONCLUSION

To implement data mining techniques, we can use big data concept. Big data presents more opportunities for research and reference in the public sector as well in technical progress. The challenges in data analyzing can be overcome by capturing the techniques in big data.

REFERENCES

[1] A. K. Choudhary, J. A. Harding and M. K. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing, Volume 20, Number 5, 501-521, 2008.

[2]ArijayChaudhry and Dr. P.S.Deshpande. Multidimensional Data Analysis and Data Mining, Black Book.

[3]Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. Wadsworth, Belmont. 1984.Classification and Regression Trees.

[4]Charissis, G., V. Moustakis, and G. Potamias. 1993b. Diagnosis of acute abdominal pain children: alication case study. Heraklion, Greece: FORTH.

[5] Clark, P., and R. Boswell. 1991. Rule induction with CN2: Some recent improvements. In Proceedings of the Fifth Working Session on Learning, 151163,

[6] G. SenthilKumar “online message categorization using Apriori algorithm” International Journal of Computer Trends and Technology- May to June Issue 2011.

[7] Green, M. 1980. Pediatric diagnoses. Philadelphia, Pa.: W. B. Saunders.

[8] Hand, D, John Wiley & Sons, Chichester, “Construction and Assessment of Classification Rules.”(1997).

[9]Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000). "Algorithms for association rule mining --- a general survey and comparison". ACM SIGKDD Explorations Newsletter 2: 58.

[10] Hunt, E. B. 1962. Concept learning: An information processing problem. New York: Wiley.

[11]Karmaker et al. “Incorporating an EMApproach for Handling Missing Attribute Values in Decision Tree Induction”.

[12]Kononenko, I. 1991. SemiNaive Bayesian classifier. In Proceedings of EWSL91,206219. Tom M. Mitchell, (1997). Machine Learning, Singapore, McGraw Hill.

[13]Leverington, D.W., 2001. Discriminating Lithology in Arctic Environments from Earth Orbit: An Evaluation of Satellite Imagery and Classification Algorithms,PhDThesis, U.Manitoba,Winnipeg, Manitoba.

[14]MarekKretowski, MarekGizes, Bialystok Technical University, Poland “Classification and Regression Trees”, 1984.

[15]Mucherino A. PetraqpapajorgjiP.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

[16]N.Deepika * et al. ”Association rule for classification of heart attack patients”, (IJAEST) International Journal of Advanced Engineering Sciences And Technologies Vol No. 11, Issue No. 2, 253 – 257

[17]Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

[18] Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442.

[19] Quinlan, J.R. 1986. Induction of Decision trees. Machine Learning

[20] Shelly Gupta et al. “data mining classification techniques applied for breast cancer for diagnosis and prognosis “Indian Journal of Computer Science and Engineering (IJCSE).

[21] UCI Machine Learning Repository http://mlearn.ics.uci.edu/databases. Usama et al. “On the Handling of Continuous Values Attributes in Decision Tree Generation”.University of Michigan, Ann Arbor.

[22] Smitha.T, Dr.V.Sundaram, “Application of Data Mining in Education” in the proceeding of international conference ICNICT2011 on December 2011 at karpagam university, Coimbatore (ISBN NO.978-81-8424-742-8), pp 669-673.

[23] Smitha.T, Dr.V.Sundaram, “Classification Rules by Decision Tree for disease prediction” International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975-8887; pp- 35-37 .

[24] Smitha.T,Dr.V.Sundaram, “Knowledge Discovery from Real Time data base using data mining technique”, International journal of Scientific Research and Publication, (IJSRP) vol 2, issue 4, April 2012, ISSN 2250-3153, pp .74-76.

[25] Smitha.T, Dr.V.Sundaram, “A Case study on High dimensional data analysis using decision tree model”-International journal of computer science Issues, (IJCSI) ISSN 1694-0814 in vol 9, issue 3, May 2012 PP 538-544.

[26] Smitha.T,Dr.V.Sundaram“Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis”- published in International journal of Advances in Engineering & Technology (IJAET) ISSN2231-1963 173 in vol 4, issue2, sept 2012 PP 15-20.

[27] Smitha.T, Dr.V.Sundaram, “Association models for prediction with Apriori Concept” published in International journal of Advances in Engineering & Technology (IJAET) ISSN 2231-1963, in vol 4, issue2, November 2012 PP 354-360.