DATA MINING: APPROCHES, CHALLENGES, ISSUE

Prof.B.B.Vikhe[1]
[1] Computer Engineering, P.R.E.C Loni, Maharashtra , India

ABSTRACT

In the world of internet i.e. Rapid development and popularization of internet large amount of data is generated, collected, stored, transfer and integrated every second. This is an era of big data i.e. data with large size ,heterogeneous, independent, sources in distributed and decentralized network and seek to explore complex and evolving relationship among data. This paper presents various data mining issues with big data and gives data mining techniques for big data that will be useful for big data mining technology future.

Keyword - *Data mining, BIG data mining, Approaches of data mining, Challenges for BIG data mining, BIG data mining techniques, Parallel mining, Map Reduce*

1. INTRODUCTION

Current years had came up with a dramatic increase in our ability to collect data from various sensors, devices, in various formats, from independent or connected applications. The era of Big Data has arrived. Each day, 2.5 Quintillion bytes of data are generated .It is been estimated that 90 percent of the data in the world today were produced within the past two years [26]. Our ability for data generation has never been so powerful and enormous since the invention of the IT in the early 19th century. Furthermore, with mobile phones becoming the sensory gateway to get large data , the large amount of available data that the network can be process to prove our daily life has constantly outpaced our past CDR i,e Call Data Record based on processing for a billing purposes . It can be foreseen that Internet of things i,e IOT applications will led the scale of large dataset to a non-believable level. People and devices are all loosely connected. To help improve quality and efficiency of life in vast amount such type of connected components will generate a huge data ocean, and valuable and relevant information must be invented from the data and make our world a better place.

1.1 Data Mining

Data mining is to extract and analysis of large data sets, for in order to discover meaningful pattern and rules. Data mining is the nothing but component or wider process called knowledge discovery from huge database. It is the process of analysis data from different perspective and summarizing the results as useful information. And information that can be used for increasing revenue, cut costs, or both. Technically, data mining is the method of finding association or patterns among number of fields in amounts of data stored either in databases and data warehouses.

Data mining is an interdisciplinary .i,e integrated database, AI, machine learning, statistics etc. Many areas of technology in current era are databases, artificial intelligence, data mining and statistics are the study of three strong large technology pillars. Data mining is the multi-step process, requires accessing and preparing data for in a mining a data, data mining algorithm, analyzing results and taking appropriate actions. The data which we accessed can be stored in one or more operational databases. In data mining the data can be mined by passing the various processes shown in below fig.1

1.2 Data Mining Process

In a data mining process consisting of an iterative sequence of the following steps:-

1. Data Cleaning

   In data cleaning to remove the noise or irrelevant data.

2. Data Integration

   Whereas the multiple data sources are combined.

3. Data Selection

   Whereas data is selected from large data set.

4. Data Transformation

   In transformation data are transformed or consolidated into forms appropriate to the mining by performing summary and aggregation operations.

5. Data Mining

   In this method data is extracted from large dataset like mining gold from sand and rocks.

6. Pattern Evaluation

   To identify the matching interesting patterns .i,e represents knowledge based on some interesting measures of the database.

7. Knowledge Presentation

   Whereas visualization and knowledge representation approaches are used to represents the mined knowledge to the user from the dataset.

2. Approaches of Data Mining

In data mining the data is mined by using the two learning techniques i.e. supervised learning technique and unsupervised learning technique they are following.

1. Supervised Learning:-

In supervised learning it also called as directed data mining the variables under declaration can be divided into two parts are explanatory variables and one is dependent variables. The goal of the analysis is to specify the relationship in the dependent variable and independent variables it is done in regression survey. To go forward with the directed data mining approach the values of the dependent variables must be known for sufficiently large parts of the data set.

2. Unsupervised Learning:-

In unsupervised learning, all the variables are treated as same way, there is no difference between dependent and explanatory variables. In contrast to the name undirected data mining approach there is some target to be achieved. This target might be a data reduction as general or more specific i,e clustering. The medium between unsupervised learning and supervised learning is the same that distinguishes discriminating analysis from cluster study. Supervised learning method requires target variables should be well defined and that the sufficient numbers of their values are given. In unsupervised learning typically either the target variable has been recorded for too small a number of parts and the target variable is unknown.
The data mining as a term used for the specific task of six approaches as follows:-
1. Classification
2. Estimation
3. Prediction
4. Association rule
5. Clustering
6. Description

1. Classification:-
Classification is a process of generalizing the data according to the different instances. Several major kinds of classification algorithms in data mining i,e decision tree, k-mean classifier, Naïve- bays, A priori and Ada Boost. Classification consists of examining features of an new presented object and assigning to it a predefined classes. The classification task is characterized by the well-defined classes, and training set consisting of reclassified examples
2. Estimation:-
Estimation deals with continuously valued outcomes, estimation to come up with a value for some unknown

continuous variables such as income, height or credit card balance etc.

3. Prediction:-

It is a statement about the way things will happen in the future, often but not always based on the experience or knowledge. Prediction may be a statement in which some outcomes are expected

4. Association rules:-

Association rule is a rule which implies certain association among a set of objects in the databases

5. Clustering:-

It can be considered the major unsupervised learning problem so, as each other problem of this type it deals with searching and finding a structure in a collection of unlabeled databases should be well defined and that a satisfied number of their values are given. In Unsupervised learning method typically either the target variable has only recorded for the small a number of cases or the target variable is unknown.

3. BIG Data Characteristics

Characteristics of BIG data are as follows

1) Volume:

Volume is matter of size. Volume is related with size of the data sets or size of the databases.

Volume is a most evident characteristic to identify big data. such that, data sets with large volume include medical data sets, educational data-bases and financial data.

2) Velocity:

Velocity is matter of speedy data and streaming data. Velocity of data in terms of how frequently data is stored, generated, how quickly transferred and retrieved and it is associated with the data of flow.

3) Variety:

In Variety is matter of structure of data and its roots. Variety means the data comes from variety of sources and variety of types of data. For example data is gather from social networking web sites appear from variety of varied sources and with number of types of data such as image, texts, audio and video. The data is coming from number of sources and appear in varied formats like multimedia, blogs, text, emails, sensors, etc.

There are other two characteristics introduced by Value and Veracity

4) Value:

Value is the cost of related with data while generating, analyzing, transferring etc. While the data is being produced, gathered, and studied from different quarters, it is eventful to state that today's data has some cost.

5) Veracity:

It is the situation of noise; it is similar with data items in the terms of insignificant data. It is require checking the correctness of the data by reducing the noise through methodologies such as data sanitization and pedigree.

3.1 Challenging Issues With Big Data

There are large challenges in big data. Big data had number of opportunities and it is facing lot of challenges also at the time of handling big data challenges arise in the following areas.

1. Data Capture and Storage.
2. Data Transmission.
3. Data Curation.
4. Data Analysis.
5. Data Visualization.

Challenges of big data mining are divided into three tiers.

The first tier involves: Setup the data mining platforms.

The second tier involves:

1. Information sharing and Data Security.
2. Application and domain Knowledge.

The third tier involves:
1. Local Learning and model fusion for multiple sources.
2. Mining from infrequent, partial data and uncertain data.
3. Mining composite data and active data.

Generally mining of data from different data sources is complex one as the size of data is larger. The big data is stored in distributed manner and gathering those data will be a tedious task and applying basic data mining algorithms will be an obstruction for it. The second case is the privacy of data. Where as in big data platform the data is processed by using parallel algorithms i,e map reduce approach is applied on those data. Later on that the data are combined using summation algorithms.
In these steps the data privacy is very much broken and privacy is a question mark in this case and The third case is mining algorithms.

3.2 Data Mining Techniques For BIG Data

Traditional data mining techniques are not enough for mining BIG data. But if we apply those data mining techniques by using other methods and technology, it may feasible to perform successful mining with big data. Innovative data mining techniques for BIG data mining are introduced such as:

1. Parallel mining for BIG data.
2. Sampling based data mining for BIG data.
3. Machine learning techniques for mining BIG data.

Parallel computing is one of the approach for working with big data. It is use of multiple compute resources at same time to solve a computational problem. Large problem solution are break into multiple smaller solutions and then they are executed in smaller divisions concurrently. Parallel computing techniques are: bit-level, instruction level, and data level and task parallelism.

If data mining techniques are used with parallel computing with Map Reduce framework then data mining algorithms can be useful for mining BIG data efficiently and effectively. In parallel data mining approaches, some algorithms are applied on huge dataset in parallel manner for effective processing. Data mining approaches with parallel technique in Map Reduce area is as follows:

1.Parallel data mining with Map Reduce environment for BIG data:

Large size and dimensionality, heterogeneity of BIG data makes data mining tasks ineffective though many algorithmic improvements are performed. Hence there is a growing need to develop efficient and effective parallel data mining algorithms that can run on a loosely coupled system. Parallel data mining algorithmsfor BIG data are:

i) Parallel pattern mining: It is used for mining patterns that occurs frequently on the data set. There are some effective algorithms in data mining for mining patterns such as Apriori, FPGrowth and Eclat [7]. Mining frequent patterns from BIG data with these algorithms is difficult. So there is need for some innovative algorithms.

Properties of Apriori algorithm are inherited by many parallel data mining algorithms, thus most paralleldata mining algorithms are said to be a variation of Apriori. Main challenges of parallel data mining are minimizing I/O, minimizing synchronization and communication effective load balancing. Parallel Algorithms used are as listed below.

a)Count Distribution – Based on apriori algorithm frequency of patterns is measured and then parallelized.

b) Candidate Distribution – Based on apriori algorithm longer patterns are generated and then parallelized.

c) Hybrid Count and Candidate Distribution – Based on Apriori algorithm, a hybrid algorithm tries to combine the strengths.

d) Sampling with Hybrid Count and Candidate Distribution – Based on Apriori algorithm it only uses sample of the database.

e) MR Éclat [8]: It is based on Map/Reduce framework.

f) Dist-Éclat [9]- It is a Map Reduce implementation where speed is optimized if specific encoding of the data fits into memory.

g) BigFIM [9]- By combining principles of both Apriori and Éclat it truly deals with Big Data optimization by using a hybrid algorithm, also on Map Reduce.

ii) Parallel classification and classification with MapReduce: Classification is supervised learning technique of data mining. It classifies data set into classes for better analysis using some class label. Many classification techniques and algorithms are available for mining data sets, But when these algorithms are applied on BIG data we need some modification else it will give a very poor performance in terms of speed and time as Big data consists of large volume and complex data. So we use some more advanced classification algorithms for mining BIG data. Following algorithms are applied for better efficiency for parallel classification and Map Reduce environment while applying on BIG data.

a) PNB(+) [10]: It is a parallel computing implementation by Naive Bayesian classifier.

b) SVM with Map Reduce [11]: It is a Map Reduce implementation of SVM.

c) C4.5 with Map Reduce[11]: It is a Map Reduce implementation of C4.5.

d) KNN with Map Reduce [11]: It is a Map Reduce implementation by KNN.

e) Naive Bays with Map Reduce [11]: It is a Map Reduce implementation of Naive Bays Classifier.

iii) Parallel clustering: Unsupervised learning technique of data mining is clustering. It makes a group of clustering is based on similar or dissimilar principle. There are many clustering techniques and algorithms for mining data sets, But when these algorithms are applied on BIG data we need some modification else it will give a very poor performance in terms of speed and time as Big data consists of large volume and complex data

Parallel power iteration clustering [13] is a parallel implementation of power iteration clustering algorithm.

2. Sampling based data mining for BIG data:

To deal with the "big data" problem simply sample the "big" data into a smaller one. Then run algorithms on the smaller data set, which will be completed much more quickly.

Sampling is a powerful data reduction technique, applied to a variety of data mining algorithms for reducing computational overhead. Sampling can be used to gather quick preliminary rules in the context of association rules. Sampling speed up the mining process. The validity of the sample is determined by the size of the sample and the quality of the sample.

3. Machine learning for BIG data:

Machine learning, new statistical algorithms analyze big volume of diverse data sources i.e image, sound, video, social network, delocalization etc. in near real time. Computers could learn from data for better future use by using these new types of programs.

Map Reduce is an arrangement of compute tasks enabling relatively easy scaling.

4. CONCLUSIONS

In this paper we represents data mining challenges, issue, characteristics in terms of heterogeneous, independent, and different evolving association in the big data. This paper gives various big data mining techniques which are different from traditional data mining techniques that can be efficiently and effectively mined big data and can increase its quality.

5. REFERENCES

[1]. R. Ahmed and G. Karypis, “Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[2]. M.H. Alam, J.W. Ha, and S.K. Lee, “Novel Approaches to Crawling Important Pages Early,” Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[3]. S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,” Science,vol. 337, pp. 337-341, 2012.

[4]. A. Machanavajjhala and J.P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data,” ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[6] E. Birney, “The Making of ENCODE: Lessons for Big-Data Projects,” Nature, vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011