ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Survey on Classification of Data Mining Using Big Data

Mutcharla Venkata Krishna Subhash
M.Tech Student,
Department of Computer Science and Engineering,
Amalapuram Institute of Management Sciences
and College of Engineering.

Mohammed Alisha
Associate Professor & HOD,
Department of Computer Science and Engineering,
Amalapuram Institute of Management Sciences
and College of Engineering.

ABSTRACT:

Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. Data cube commonly abstracting and summarizing databases. It is way of structuring data in different n dimensions for analysis over some measure of interest. For data processing Big data processing framework relay on cluster computers and parallel execution framework provided by Map-Reduce. Extending cube computation techniques to this paradigm. MR-Cube is framework (based on mapreduce)used for cube materialization and mining over massive datasets using holistic measure. MR-Cube efficiently computes cube with holistic measures over billion-tuple datasets.

Keywords:
Big data, data cube, cube materialization, Map Reduce, Challenging issues, Datasets, Data Mining Algorithms.

I.INTRODUCTION:

Today is the era of Google. The thing which is unknown for us, we Google it. And in fractions of seconds we get the number of links as a result. This would be the better example for the processing of Big Data. This Big Data is not any different thing than out regular term data. Just big is a keyword used with the data to identify the collected datasets due to their large size and complexity? We cannot manage them with our current methodologies or data mining software tools. Another example, the first strike of Anna Hajare triggered number of tweets within 2 hours. Among all these tweets, the special comments that generated the most discussions actually revealed the public interests. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting.In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. upto 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1].for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012[10]. this shows that it is very difficult for big data applications to manage, process and retrieve data from large volume of data using existing software tools. It's become challenge to extract knowledgeable information for future use [15]. There are different challenges of Data mining with Big Data. We overlook it in next section. Currently Big Data processing depends upon parallel programming models like MapReduce, as well as providing computing platform of Big Data services. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter. Due to the large size of data it is becoming expensive to analysis data cube. The Map-Reduce based approach is used for data cube materialization and mining over massive datasets using holistic (non algebraic) measures like TOP-k for the top-k most frequent queries. MRCube approach is used for efficient cube computation. Our paper is organized as follows: first we will see key challenges of Big Data Mining then we overlook some methods like cube materialization, MapReduce and MR-cube approach.

II. BIG DATA MINING:

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress"

[9]. Big Data mining was very relevant from the beginning, as the first book mentioning ’Big Data’ is a data mining book that appeared also in 1998 by Weiss and Indrukya [34] . However, the first academic paper with the words ’Big Data’ in the title appeared a bit later in 2000 in a paper by Diebold [8]. The origin of the term ’Big Data’ is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [11] in his invited talk at the KDD BigMine’12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. Anew large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to find useful patterns to improve user experience. Alex ’Sandy’ Pentland in his ’Human Dynamics Laboratory’ at MIT, is doing research in finding patterns in mobile data about what users do, and not in what people says they do [28]. We need new algorithms, and new tools to deal with all of this data. Doug Laney[19] was the first one in talking about 3 V’s in Big Data management:

- Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process
- Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more
- Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time Nowadays, there are two more V’s:
- Variability: there are changes in the structure of the data and how users want to interpret that data
- Value: 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 Gartner[15] summarizes this in their definition of Big Data in 2012 as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. There are many applications of Big Data, for example the following [17; 2]:
- Business: costumer personalization, churn detection
- Technology: reducing process time from hours to seconds
- Health: mining DNA of each person, to discover, monitor and improve health aspects of every one
- Smart cities: cities focused on sustainable economic development and high quality of life, with wise management of natural resources These applications will allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [17]. So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information. As shown in fig 1 below.

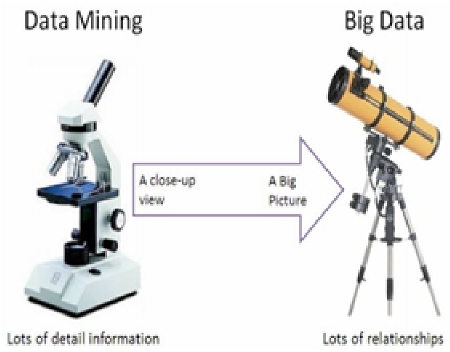

Fig.1 Data Mining with Big Data

III. KEY FEATURES OF BIG DATA:

The features of Big Data are:

- It is huge in size.
- The data keep on changing time time to time.
- Its data sources are from different phases.
- It is free from the influence, guidance, or control of anyone.
- It is too much complex in nature, thus hard to handle.
- It’s huge in nature because, there is the collection of data from various sources together.

If we consider the example of Facebook, lots of numbers of people are uploading their data in various types such as text, images or videos. The people also keep their data changing continuously. This tremendous and instantaneously, time to time changing stock of the data is stored in a warehouse. This large storage of data requires large area for actual implementation. As the size is too large, no one is capable to control it oneself. The Big Data needs to be controlled by dividing it in groups.Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time.

For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

IV. METHOD OVERVIEW

4.1 Data Cube

Data cube provide multi-dimensional views in data warehousing. If n dimensions given in relation then there are 2^n cuboids and this cuboids need to computed in the cube materialization using algorithm[2]which is able to facilitate feature in MapReduce for efficient cube computation. In data cube Dimension and attributes are the set of attributes that user want to analyze. Cube lattice is formed representing all possible groupings of this attributes, based on those attributes. After that by grouping attribute into hierarchies and eliminating invalid cube regions from lattice we get more compact hierarchical cube lattice. Finally cube computation task is to compute given measure for all valid cube groups. There are different techniques of cube computations [3] like multi- dimensional aggregate computation, BUC(Bottom-Up Computation), star cubing for efficient cube computation. There are limitations in these techniques: 1) They are designed for a single node or for a cluster with less nodes [19], so it is difficult to process data with a single or few machines. 2) Many analyses over logs, involve computing holistic measure where as many techniques uses the algebraic measures. 3) Existing techniques failed to detect and avoid data skew. There is need of technique to compute cube in parallel on holistic measure over massive dataset. Hadoop based MapReduce can handle large amount of data in cluster with thousand of machines. So this technique is good option for analysis of data.

4.2 Map Reduce:

MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. The nature of this programming model and how it can be used to write programs which run in the Hadoop environment is explain by this model. Hadoop [11] is an open source implementation for this environment. Map and Reduce are two functions. The main job of these two functions are sorting and filtering input data. During Map phase data is distributed to mapper machines and by parallel processing the subset it produces pairs for each record. Next shuffle phase is used for repartitioning and sorting that pair within each partition. So the value corresponding same key grouped into {v1, v2,….}values. Last during Reduce phase reducer machine process subset pairs parallel in the final result is written to distributed file system.

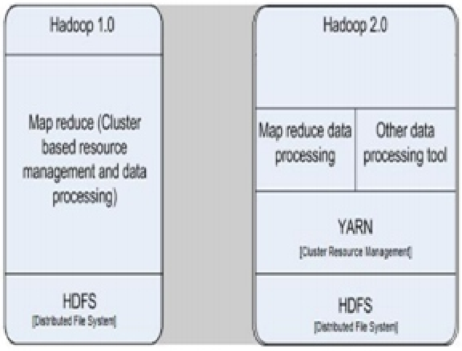
Hadoop 1.0
Map reduce (Cluster based resource management and data processing)
HDFS
(Distributed File System)
Hadoop 2.0
Map reduce data processing
Other data processing tool
YARN
(Cluster Resource Management)
HDFS
(Distributed File System)

Fig.2:- Architecture of Hadoop1.0 and 2.0

MR1 is used in Hadoop1.0 but due to some resource management issues like inflexible slot configuration, scalability. After Hadoop version 0.23, MapReduce changed significantly. Now it known as MapReduce 2.0 or YARN (Yet Another Resource Negotiator). MapReduce 2.0 has two major functionalities of job tracker which are spit into resource management and job scheduling into separate daemons [4]. Fig 1 shows the architecture of both Hadoop versions. In Hadoop1.0 Job Tracker has a responsibility for managing the resources and scheduling jobs across the cluster. But in Hadoop2.0 the architecture of YARN allows the new Resource Manager to manage the usage of resources across all applications. And Application Masters takes the responsibility of managing the job execution. This new approach improves the ability to scale up the Hadoop clusters to a much larger configuration than it was previously possible. In addition to this, YARN permits parallel execution of a range of programming models. This includes graph processing, iterative processing, machine learning, and general cluster computing.

4.3 MR-cube Approach:

MR-Cube MR-Cube is a MapReduce based algorithm introduces for efficient cube computation [5] and for identifying cube sets/groups on holistic measures. MR-Cube algorithm is used for cube materialization and identifying interesting cube groups. Complexity of the cubing task is depending upon two aspects: size of data and size of cube lattice. Size of data impacts size of large group and intermediate size of data, where as the cube lattice size impacts on intermediate data size and it is controlled by the number/depth of dimension. First we identify the subset of holistic measures that can easily compute in parallel than an arbitrary holistic measure. We can call it Partially Algebraic Measures. The technique of partitioning large groups based on algebraic attribute called Value partitioning. Value partitioning is used to effectively distribute the data; we can easily compute it with Naïve algorithm [9]. Value partitioning performs on only on group that are likely reducer friendly and dynamically adjust the partition factor. Partition factor is ratio by which a group is partitioned. There are different approaches for detecting reducer unfriendly groups. One of the approach is sampling approach where we estimate the reducer unfriendliness of cube region based on the number of groups it is estimated and perform partitioning for all small groups within the list of cube region that are estimated to be reducer unfriendly.

4.4 Cube Materialization:

Cube materialization task comes under the MR-Cube approach. Materializing the cube means computing measures for all cube groups satisfying the pruning condition. After materializing cube we can identify the interesting cube groups for cube mining algorithm. The main MR-CUBE-MAP-REDUCE task is perform using annotated lattice. The combine process of identifying and value partitioning unfriendly regions followed by partitioning of regions is referred as annotate. Based on the sampling results cube regions have deemed as reducer unfriendly and require partitioning. Each tuple in dataset the MR-Cube-Map emits key:value pairs for each batch area. In required keys are appended with hash based on value partitioning. The shuffle phase then sorts them by key yielding reducer tasks. The BUC algorithm is then run on each reducer and cube aggregates are generated. The value partitioned group are merged during post processing to produce the final result.

V. CONCLUSION:

In real-world applications managing and mining Big Data is Challenging task, As the data concern large in a volume, distributed and decentralized control and complex. There are several challenges at data, model and system level. We need computing platform to handle this Big Data. The MapReduce framework is one of the most important parts of big data processing, and batch oriented parallel computing model. In earlier versions of MapReduce the components were designed to address basic needs of processing and resource management. Recently, it has evolved into a improved version known as MapReduce 2/ YARN that provides improved features and functionality. With Big Data technologies we able to provide most relevant and accurate social sensing feedback to better understand to society at real-time. MR-Cube efficiently distributes the computation workload across machines and completes the cubing task. Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited to participate in this intrepid journey.

REFERENCES:

[1]. Xindong Wu, Fellow, IEEE, 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]. Zhengkui Wang, Yan Chu, Kian-Lee Tan, Divyakant Agrawal, Amr EI Abbadi, Xiaolong Xu, "Scalable Data Cube Analysis over Big Data" appliarXiv:1311.5663v1 [cs.DB] 22 Nov 2013

[3]. Dhanshri S. Lad #, Rasika P. Saste, "Different Cube Computation Approaches: Survey Paper" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4057-4061

[4]. The Apache Software Foundation"http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html"

[5]. Arnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan, Fellow, IEEE, "Data Cube Materialization and Mining over MapReduce" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 6, NO. 1, JANUARY 2012

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

[7]. D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.

[8]. A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032- 2033,2012.

[9]. A. Nandi, C. Yu, P. Bohannon. And R. Ramakrishnan, "Distributed Cube Materialization on Holistic Measures, " Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.

[10]. F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[11]. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo" Yahoo! Developer Network Blog, 2008.

[12]. "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[13]. A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011.

[14]. K. Yury, "Applying Map-Reduce paradigm for parallel closed cube computation," Proc. First Int'l

[15]. Hadoop. http://hadoop.apache.org/.

[16]. P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquini. Incoop: Mapreduce for incremental computations. In SOCC, 2011.

[17]. Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. Haloop: Efficient iterative data processing on large clusters. PVLDB, 3(1):285–296, 2010.

Author Details:

Mohammed Alisha
is currently working as Associate Professor & Heading the department of Computer Science and Engineering, Amalapuram Institute of Management Sciences &College Of Engineering. He is a postgraduate in Computer Science and Engineering and had 9 years of teaching and research experience. His research interests include Spatial Data Mining, Compuer Networks, Web Mining and Data Warehousing.