www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14160-14163

A Survey on Parallel Method for Rough Set using MapReduce Technique for Data Mining

Ms.Suruchi V.Nandgaonkar, Prof.A.B.Raut

Student ME CSE, HVPM COET
Amravati Maharashtra, India
nsuruchee@gmail.com

Associate Prof. HVPM COET
Amravati, Maharashtra, India
anjali_dhahake@gmail.com

Abstract: ***In this paper present survey on Data mining, Data mining using Rough set theory and Data Mining using parallel method for rough set Approximation with MapReduce Technique. With the development of information technology data growing at an incredible rate, so big data mining and knowledge discovery become a new challenge. Big data is the term for a collection of data sets which are large and complex, it contain structured and unstructured both type of data. Data comes from everywhere, posts to social media sites, digital pictures and videos etc this data is known as big data. Useful data can be extracted from this big data with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data. Rough set theory has been successfully applied in data mining by using MapReduce programming technique. We use the Hadoop MapReduce System as an Implementation platform. The lower and upper approximations are two basic concept of rough set theory. A parallel method is used for the effective computation of approximation and is improving the performance of data mining. With the benefits of MapReduce The MapReduce technique, received more attention from scientific community as well as industry for its applicability in big data analysis it makes our approach more ideal for executing large scale data using parallel method .In this paper we have presented working and execution flow of the MapReduce Programming paradigm with Map and Reduce function. In this work we have also briefly discussed different issues and challenges that are faced by MapReduce while handling the big data. And lastly we have presented some advantages of the Mapreduce Programming model.***

Keywords: Data mining, MapReduce, Rough sets, Approximations, Hadoop,

1. INTRODUCTION

Nowadays, with the degree of knowledge growing at unmanageable rate, massive data processing and data discovery became a new challenge.

Three V's in Big Data

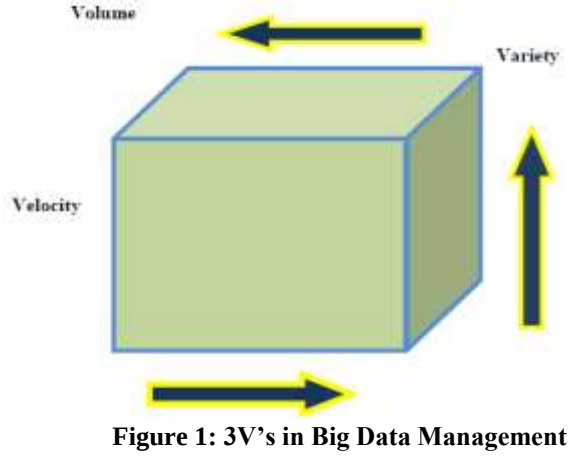

Figure 1: 3V's in Big Data Management

Doug Laney was the first one talking about 3V"s in Big Data Management ***Volume:*** The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate. ***Variety:*** Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more. ***Velocity:*** Data in motion. The speed at which data is created, processed and analyzed continues to accelerate. 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.

Rough set theory for data discovery has been with success applied in data processing. Then MapReduce technique has received a lot of attention from each scientific community and trade for its applicability in massive knowledge analysis. With the event of knowledge technology, great deal of knowledge area unit collected from numerous sensors and devices in numerous formats. Such knowledge processed by freelance or connected applications can typically cross the peta-scale threshold, which might successively raise the process needs. With the quick increase

DOI: 10.18535/ijecs/v4i9.16

and update of huge knowledge in real-life applications, it brings a new challenge to rapidly acquire the helpful information with big data processing techniques. For processing of big data, Google developed a software system framework known as MapReduce to support large distributed data sets on clusters of computers that is effective to analyze large amounts of data. MapReduce has been a well-liked computing model for cloud computing platforms. Followed by Google's work, several implementations of MapReduce emerged and much of ancient strategies combined with MapReduce are given as yet. This paper presents survey on the various fields like data mining data mining with rough set technique and data mining with MapReduce Technique, for rough set approximation calculation of data mining of parallel technique.

In this Paper, section one is totally introduction. Section two describe regarding data processing Technique with the connected work. Section three describe shortly regarding data mining with rough set theory with its connected work. Section four is motivation of this paper, that describe parallel technique for rough set approximation in data mining with its connected work. Section five future works on the parallel technique for rough set approximation.

2. Data Mining Technique

Data mining could be a new developing technology for enterprise information and data integration. It will scale back the operation value, increase profit, and strengthen market competition of the enterprise. Generally, there are two ways to establish data mining application tailor to Associate in enterprise: using business intelligence solutions and product available on the market, or developing data processing algorithms all by oneself. However, each of them are measure impractical in value and time. the previous one prices lots, whereas the latter needs developers to be conversant in each enterprise business and data processing technology Software apply could be a answer to avoid repeated work in the software development. it's thought to be a approach to solve the software crisis and promote potency and quality of software production. As a kennel technique to support software reuse, software component technique gets progressively wide attention. to completely build use of reusable component, and support mass component's production, classification, search, assembly and maintenance, component library is extremely vital. Applying software component technique to data processing, wrapping individual business modules of knowledge mining within the kind of component, and victimization part technique to attain the organization, management and retrieval of the component, might greatly increase the efficacy and quality, and reduce the cost and period of data mining application development.

The demand of variability of knowledge mining tasks may be met still, and therefore the application of knowledge mining technology will be broaden.

The data Mining supported Neural Network and Genetic algorithmic program is researched in detail. the various technology to attain the information mining on Neural Network and Genetic algorithmic program also are surveyed. There are different task in data mining which are 1) Classification 2) Estimation 3) Prediction 4) Grouping or association rule 5) Clustering 6) Description and Visualizations the primary 3 tasks are all example of directed data processing or supervised learning. Consecutive 3 tasks are unit example of undirected data processing. There completely different ways for the classification task in data processing and Rough set theory is one of the classification methodologies. Rough set theory will be used for classification to find structural relationships within vague or noisy data. so main difference between big data mining and data mining is shown in following table.

Table 1

Big data	Data mining
Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information.
Big data is the asset.	Data mining is the handler which beneficial result**.**
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.	Data mining refers to the operations that involve relatively sophisticated search operation.

3. Data Mining Using Rough Set Theory

3.1 Rough set Theory

Rough set theory was developed by Zdzislaw Pawlak in the early 1980's. it's a very powerful mathematical tool for dealing with imprecise information in decision situations. . The main goal of the rough set analysis is induction of approximations of conception additionally it plays a very important role within the fields of machine learning, pattern recognition and data processing. Rough set based mostly knowledge analysis uses data tables known as a decision table, columns of those tables are labeled by attributes, rows – by objects of interest and entries of the table are attribute values. Attributes of decision table are divided into 2 completely different group referred to as condition and decision attributes. Every row of a decision table induces a decision rule that shows decision or outcome. If some conditions are fulfilled, the decision rule is certain.

Once decision rule uniquely identifies decision in terms of conditions. Otherwise the decision rule is unsure. Decision rules are a related to approximations. Lower approximation refers to certain decision rule of decision in terms of conditions, whereas boundary region referred by uncertain decision rule of decisions. A rough set learning algorithmic rule may be obtain to get a group of rules in IF-THEN form, from a decision table. The rough set technique provides a good tool for extracting information from

databases. Here creates a cognitive content, classifying objects and attributes at intervals the created decision tables. Then an information discovery method is initiated to get rid of some undesirable attributes. Finally the information dependency is use out, within the reduced information, to seek out the lowest set of attributes known as reduct.

A classification has been provided based on the various soft computing tools and their hybridizations used, the information mining function enforced. The efficiency of the various soft computing methodologies is highlighted. usually fuzzy sets will use for handling the problems associated with understandability of patterns, incomplete or noisy data, mixed media data and human interaction, and may give calculable solutions quicker. Neural networks are nonparametric, and reveal sensible learning and generalization capabilities in data-rich environments. Rough sets are appropriate for handling differing kinds of uncertainty in data.

Silvia Rissino and Germano Lambert-Torres [4] discussed The rough set approach to processing of incomplete data is based on the lower and the upper approximation, and it is defined as a pair of two crisp sets corresponding to approximations The main advantage of rough set theory in data analysis is that it doesn't would like any preliminary or further data regarding knowledge. Rough set theory has additionally provided the required formalism and concepts for the event of some propositional machine learning systems. Rough set has additionally been used for information illustration; knowledge mining; coping with imperfect data; reducing information representation and for analyzing attribute dependencies. Rough set theory has found several applications like power system security analysis, medical data, finance, voice recognition and image processing; and one of the analysis areas that has with successfully used. Rough Set is the information discovery or data processing in database

4. Data mining using parallel method for rough set approximation with MapReduce Technique

To our data, most of the normal algorithms supported rough sets are the sequential algorithms and existing rough set tools solely run on one pc to take care of massive information sets. To expand the applying of rough sets within the field mining and take care of large data sets, the parallel computation of the rough set approximations is enforced. And this Parallel approximation may be achieved by using MapReduce Technique.

• Map-function takes an input pair and produces a set of key, value pairs. The MapReduce group together all values related to a similar key I and transforms them to the reduction function.

• Reduce-function accepts key I and a set of values for that key. It merges these values along to create a possibly smaller set of values. By doing sorting and shuffling it produces reduced values get from Map function

4.1 MapReduce Technique

Hadoop Distributed file system is that the storage system employed by Hadoop applications. HDFS creates multiple replicas of data blocks and allocates them on data nodes, to enable reliable very fast computations. Hadoop carries with it 2 major elements that are: File storage and Distributed process system. the primary element of file storage is thought as "HDFS (Hadoop distributed file system)". It provides scalable, dependable, relatively low price storage. HDFS stores files across a set of servers in an exceedingly cluster. HDFS ensures knowledge handiness by frequently observance the servers in an exceedingly cluster and therefore the blocks that manage knowledge. The second important elements of Hadoop, is that the parallel processing system referred to as "MapReduce". The MapReduce framework and therefore the Hadoop distributed file system are running on same set of nodes. In MapReduce programming, it permits the execution of java code and conjointly uses software package written in different languages Jeffrey Dean and Sanjay Ghemawat [6] gift the temporary description concerning MapReduce programming model, with completely different programs as example. Also provides execution summary of MapReduce programming model, with Fault Tolerance, Task granularity and locality

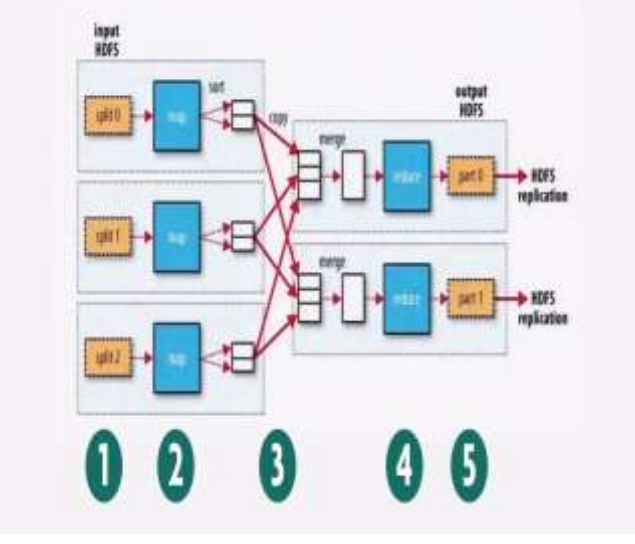

Figure 2 Map Reduce Programming Model

It explains reasons for successful use of MapReduce Programming model at Google. The model hides the small print of parallelization, fault-tolerance, and load balancing

So it is easy to use also a large variety of problem are easily expressible as mapreduce computations. Redundant execution will be wont to cut back the impact of slow machines, and to manage machine failures and data loss Zdzilsaw Pawlak and Andrzej Skowron [7] present basic ideas of rough set theorey,also listed some analysis directions and exemplary applications supported the rough set approach. In this paper it mentioned the methodology supported discernibility and mathematician reasoning for economical computation of various entities together with reducts and decision rules. it's been make a case for that the rough set approach will be used for synthesis and analysis of idea approximations within the distributed surroundings of intelligent systems Zhang, T Li, Loloish pan [8] planned 3 rough set based mostly strategies for data acquisition using MapReduce technique. to evaluate the performances of the

planned parallel strategies used speedup. Comprehensive experimental results on the real and synthetic data sets demonstrated that the proposed methods could effectively process large amount of data sets in data mining. There are 3 algorithms are used for the data acquisition from huge knowledge supported MapRedudce.

Junbo Zhang, Tianrui Li, prosecutor Rusan [9] proposes a parallel methodology for computing rough set approximations. Accordingly algorithms corresponding to the parallel methodology supported the MapReduce technique are put forward to deal with the massive data also used speedup a, scaleup and sizeup to evaluate the performances of the planned parallel algorithms. The experimental results on the real and synthetic data sets showed that the proposed parallel algorithms may effectively wear down large data sets in data processing. . In this paper there are 4 types of algorithms are using which are Equivalence class computing algorithm, Decision class computing algorithm, association algorithm and Indexes of rough set approximation computing algorithm for parallel method.

5. Future Work

To implement the Parallel method for rough set using MapReduce technique in data mining there are four types of algorithms are using, which are 1)Equivalence class computing algorithm, 2) Decision class computing algorithm, 3) association algorithm and 4)Indexes of rough set approximation computing algorithm. Here first two algorithms are implementing parallel and third and fourth in series. This is still a time consuming, so to increase time efficiency instead of using four proposed to use three algorithm by merging last two algorithms.

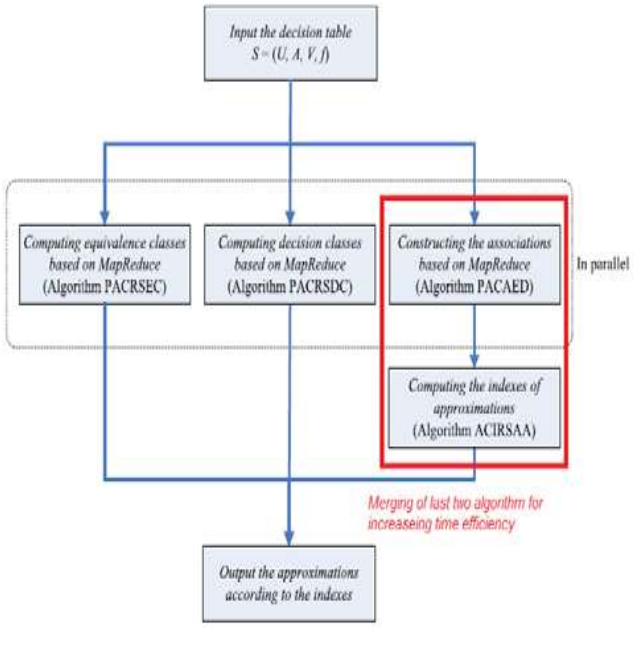

Figure 3:-The totally parallel methods for computing rough set approximations

Also to increase time efficiency and speedup the process proposed to use attribute selection option from the massive data, for more accurate result.

6. Conclusion

In this paper the various fields like data processing, data processing exploitation rough set technique and data processing exploitation MapReduce Technique in short describe, for rough set approximation calculation exploitation parallel technique. Additionally describe the long run work for the present parallel technique for scheming rough set approximation.

ACKNOWLEDGMENT

I am thankful to all the teachers and the principal for their valuable guidance. I would like to sincerely thank Prof. Dr. Anjali Raut whose knowledgeable guidance helped me to make more descriptive. I would also like to thank my parents and friends for their constant support.

References

[1]Sushmita Mitra, Pabitra Mitra, "Data Mining In Soft Computing Framework: A Survey" IEEE Transactions on Neural Networks, Vol. 13, No. 1, January 2002.

[2]Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simple Data Processing on Large Clusters" Proc. To appear in OSDI 2004.

[3]Pavel JURKA, "Using Rough Sets In Data Mining" Proceedings of the 12th Conference and Competition STUDENT EEICT 2006 Volume 4.

[4] Zdzisław Pawlak, Andrzej Skowron "Rudiments of rough sets" Proc. Elsevier accepted in 7 June 2006

[5]Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.

[6]J Zhang, T Li, Yi pan "Parallel Rough Set Based Knowledge Acquisition Using MapReduce from Big Data" Proc. ACM Big Mine '12, August 12, 2012 Beijing, China

[7]NASSCOM Big Data Report 2012.

[8]Neelamadhab Padhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[9]Nikita Jain, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 Nov-2013.

[10]Silvia Rissino and Germano Lambert-Torres, "Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications" Open Access Database www.intechweb.org

[11] Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013.

[12] Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future", Vol 14,Issue 2,2013.

[13] Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013

[14] Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knowledge and Data Engineering Vol 26 No1