International Journal of Engineering Research and Science & Technology

ISSN: 2319-5991
Vol. 1, No. 3
May 2015

International Conference on “Advance Research and Innovation in Engineering, Science, Technology and Management” ICARSM 2015
Organized by
Department of ECE, ARS College of Engg., Sathyamangalam, Chennai, India.

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com
A PARALLEL MATRIX-BASED METHOD FOR COMPUTING APPROXIMATIONS IN INCOMPLETE INFORMATION SYSTEMS

Sathish1*, Muhammed Sadique1, Krishna Raj1 and Sabarinathan1

*Corresponding Author: Sathish

Recent trend extensively shows that usage of large scale data mining increases tremendously, hence knowledge utilizations leads in great challenge, rough set theory is one of the intelligent technique used in the field of pattern recognition and data mining, by using rough set theory can approximate a group of distinct object in lower and upper bound. To establish rough set theory in better way computation should be approximated, which is vital to achieve performance level tuning. Map reduce is the significant and notable technique which gained a focus on large scale data processing. Ancient technique of map reduce is a programmable model which is related for implementing a large set of data with a parallel, distributed technique, but this technique missed a case for finding incomplete information system, to overcome this drawback we go for parallel matrix based method to detect incomplete information system in large scale data base in our proposed method. Comparing our traditional methods our proposed method shows better way of detecting incomplete data than our traditional method of processing data.

INTRODUCTION

With the development of information technology, the volume of data is rapidly increasing at a unique rate. The Map Reduce model supports large distributed data sets on clusters of computers which allows for an efficient analysis of large amounts of data. As one of the most essential and effective cloud computing techniques, Map Reduce has become a important computing model for cloud platforms. Map Reduce framework supports many cloud platforms. For detail, Amazon and Microsoft Azure both support open source Map Reduce Hadoop and Twister runtime systems such as Amazon Elastic Map Reduce and Twister, Azure.

In exceptional velocity information is streaming these days. Enormous information is a well known term which characterizes the development of information which may be organized or unstructured. Enormous information implies vast measure of information which is impossible be transformed in customary way. When we are managing huge datasets, the associations confront a few issues to make, overhaul and deal with the datasets. It has high speed, high volume and verities of data resources which has diverse approach to be prepared.

1 Computer Science and Engineering Dept., SRM University, Chennai, India.

This article can be downloaded from http://www.ijerst.com/International-Conference-on-ICARSM-2015.php#1
Fragmented data implies the dataset which is having obscure information or some information is absent. In this paper the tenets era issue from fragmented data framework we have considered. We took enormous measure of information with deficient data and we managed it. not adapt specifically towards any particular application domain.

Rough set theory can be viewed as new scientific device for flawed information investigation. In choice help area this theory discovered application. Unpleasant set in the sense the information set which is crisped. Each one unpleasant set has limit line cases, which is called lower and upper close estimation. The distinction between the lower and upper estimate constitutes the limit locale of the unpleasant set.

Rough set theory is an influential and well known approach in separating and minimizing standards from choice table. In past paper, unpleasant set technique utilized for little measure of information. At the same time in this paper unpleasant set approach is reached out to expansive databases.

For handling big data we are using hadoop tool. Hadoop can handle different type of data structured or unstructured. Apache hadoop is an open source platform and it has a new way to processing and storing data. Hadoop enables distributed parallel processing of large amount of data and can scale without limit.

**RELATED WORK**

**The Map Reduce Model**

Simplified Data Processing on Large Clusters

Map Reduce is an important model and an combine implementation for preparing and developing large data sets. Users define a map method that processes a key/value pair to process a set of intermediate pairs, and a reduce method that combine all median values associated with the relevant and same intermediate key.

Our development of MapReduce implements on number of clusters and is highly reliable: common MapReduce computation methods many terabytes of data on many numbers of machines. Programmers and the system simple to use: Numbers of MapReduce process has been developed and upwards of one thousand MapReduce jobs are executed on every day Google’s clusters. The run-time system takes care of the specifies of assigning the input data, programming the program’s performance across number of machines, machine failures, and controlling the required inter-machine communication.

**Fast Parallel Clustering Algorithm for Large Spatial Databases**

The clustering algorithm DBSCAN build on clusters based on density notion and is produced to identify clusters of arbitrary shape to distinguish noise. In this paper, propose PDBSCAN, algorithm for a parallel version. We use the ‘shared-nothing’ architecture with number of computers attached through a network. Distributed data structure and fundamental component of a shared nothing system. We introduce the dR*-tree, a distributed spatial index structure in which the data is spread among multiple computers and the indexes of the data are replicated on every computer. We implemented many workstations connected via Ethernet using our method. An experimental result shows that PDBSCAN offers linear speedup and has efficient scale up and size up behaviour.
An efficient accelerator for attribute reduction from incomplete data in rough set framework

Attribute reduction from large-scale incomplete and missing data is a most complex problem in areas such as recognition, machine learning process and data mining. In rough set theory, incomplete data aims for feature selection to retain the power of original features. To solve this issue, many selection algorithms have been proposed and these algorithms are usually computationally time-consuming. To solve this shortcoming, we suggest in this paper positive approximation, we implement rough set theory based on theoretic framework, and can be used to increase a heuristic process for feature selection from incomplete and missing data. A general feature selection algorithm is designed an application, which is proposed accelerator. Rough set theory mainly used for heuristic algorithm, and also obtain several modification by integrating the accelerator. Experiments show original counterparts for modified algorithms. It is important noting that the evaluation of the modified algorithms with larger data sets.

Data Mining with Rough Set Using Map-Reduce

A enormous data mining and knowledge discovery performance a great challenge with the data growing at an effective rate. Retrieve meaningful data using different methods. One of the methods is Rough set and it is based on lower and upper approximation. Existing serial way method calculation for rough set approximation. Anyway we propose a parallel method. Large-scale computation is managed by Map Reduce method. Recently introduced Map-Reduce technique has received much consideration from both scientific community and industry for its applicability in big data analysis. The efficient computation of approximation is essential step in improving the performance of rough set. For mining the vast large data, parallel computing nodes, algorithms and different methods get used in research fields. In this paper, explained a parallel method for computing rough set. We can achieve the same by using Map reduce. We can create rules and abstract attributes of large data because of Map Reduce.

PROPOSED METHODOLOGY

In this world we have vast large number of data and those data are need to analyze and collected daily. Data mining have multiple Structures and it analyze data with structure along with sequences, trees and graphs. Evaluating data with vast amount of attributes with more effective by using our proposed parallel method. Feature selection and rule detection is the key from computing lower and upper approximation when using rough set-based methods by map reduce method.

The Map Reduce model, by Google, is implemented to handle large amount of data sets in an appropriate environment and has become a eminent parallel model in cloud computing. Here processing vast large number of data using map reduce method. Here user send request to admin side for response and data stored in data base. Each request separated and conformed by user and it provides easy way to analyze the data. Main advantages of MapReduce are robustness and feasibility.

Figure 1 shows the MapReduce framework. In this method mainly two phases are used to process in MapReduce model. First phase is Map phase and second is Reduce phase. In each phase uses key and value pairs as input and
output. Users can choose the types of pair attributes based on requirements. Additionally two functions are processed namely Map and Reduce functions performed for Map and Reduce phase. In Map phase input data consider as main node and divided into sub program. A key and value pair(K1,V1) take as input and another pair(k2,V2) as output generated by Map function. The system is collected all intermediate results and produce list of values (K2, a list of (v2)) as an input of reduce function. After first phase function finished its process and output produces as another key and list values pair.

**Rough set theory**

The main concept we are dealing with is rough set analysis in induction of (learning) approximations of concepts. A Rough set constitutes a sound basis for Knowledge Discovery. It provides mathematical tools to discover patterns hidden in data. It can be used for detail selection, extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc. Identifies partial (or) total dependencies in data, eliminates redundant data and gives approach to zero values, missing and dynamic data and others.
Recent extensions of rough set theory: Rough metrology and Ontology-based rough sets have developed new methods for decomposition of large data sets, data mining in distributed, multi-agent systems and granular computing. Rough set theory is a new mathematical approach to imperfect knowledge. The problem of imperfect knowledge has been tackled for a long time by philosophers, logicians and mathematicians. Recently it became also a crucial issue for computer scientists, particularly in the area of AI. There are many approaches to the problem of how to understand and manipulate imperfect knowledge.

Let us describe led the universe and an in-discernibility relation $R \subseteq U \times U$, representing our lack of knowledge about elements of $U$. For the bivalence relation. Let $X$ is a subset of $U$. We want to characterize the set $X$ with respect to $R$. We will need the basic concepts of rough set theory as given below.

The lower approximation of a set $X$ with respect to $R$ is the set of all objects, which can be for certain classified as $X$ with respect to $R$ (are certainly $X$ with respect)

The upper approximation of a set $X$ with respect to $R$ is the set of all objects which can be possibly classified as $X$ with respect to $R$ (are possibly $X$ in view of $R$).

Loading data(Data Loaded Local From HDFS) $>$ normal file system into Hadoop file system using put commands, Hadoop supports processing of many different formats and types of data through Input Format. The Input Format of a Hadoop MapReduce computation generates the key-value pair inputs for the mappers by parsing the input data. The main classes for file manipulation in Hadoop are in the package org.apache.

Hadoop's. In our Hadoop we loaded unstructured network exchange file system. Basically Hadoop file API is generic and can be used for working with filesystems other than HDFS. MapReduce job. Hadoop requires the Mapper and the Reducer to be their own static classes. In PutMerge program, to use the Hadoop file API to both read the local filesystem and write to HDFS normally we load unstructured data to be load in hdfs it will automatically change in structured in hdfs. in our input format for plain text files, which generates a key-value to record for each line of the input text files. Each line of the input data is broken into a key (text) and value (text) pair with replicas. the replicated file to be Applied For the next Phase and Computed values with the logistic regression methods.
In this subsection, we demonstrate a correlation of distinctive parallel methods. We here just analyze the execution time of distinctive parallel techniques as opposed to speed up since Strategies 2 and 3 (S2, in Section 4.2 and S3, in Segment 4.3) utilization an incremental method to quicken the redesign process, whose time unpredictability relies on upon the qualities of the testing information. Figure 6 shows the execution time of Strategies 1, 2 and 3 versus the number of centers. It is clear that S2 and S3 dependably have the preferred execution over S1. That is on account of us . Use the incremental system in S2 and S3 to encourage , The Sub-Merge operation. With the incremental system,

Further procedures of consolidation (overhaul) operation are just executed while the qualities in transitory result framework are 1. Rather, it generally executes $n \times (n$ is the numbers of articles) operations in S1. Along these lines, numerous repetitive methodologies can be skipped in S2 and S3

**CONCLUSION**

In this paper, to process substantial scale fragmented information with harsh set hypothesis, we initially presented the framework representation of lower and upper rough guesses in IIS. As per the qualities of the lattice, we proposed three parallel methods taking into account MapReduce to register close estimations. Every one of them was executed on the MapReduce runtime framework Twister. The aftereffects of testing speedup of S1 demonstrate that the parallel method S1 has a decent execution on transforming huge scale information. In analyzing an incremental system to quicken the handling methodology in S2 and S3, we observed that S2 and S3 dependably have better execution than S1, and, much of the time, S3 beats S2. Yet, in the event that the info information contains tremendous amounts of items also just a couple of traits, S2 would show improvement over S3. Our broad trial assessment shows that our proposed parallel strategies are more productive in breaking down information with substantial measure of characteristics. As we can see, registering lower and upper rough guesses is the way to manage instigation and peculiarity choice when using unpleasant set-based systems. We arrange in our future work to further research the learning found furthermore emphasizes determination from vast scale, inadequate information

**REFERENCES**


