IMPROVING PERFORMANCE OF APRIORI ALGORITHM USING HADOOP

Ravindra Bachate[1], Hyder Ali Hingoliwala[2]

[1]*Department of Computer Engineering, JSPM's JSCOE, Pune, Maharashtra, India*
[2]*Department of Computer Engineering, JSPM's JSCOE, Pune, Maharashtra, India*

Abstract

Spatial data is a data having a geological information. This paper explores the use of Hadoop framework to improve the performance of Apriori algorithm for spatial data mining. FP growth algorithm is better than Apriori but it fails in certain situations. By applying the Apriori algorithm parallely using Hadoop framework to spatial data, we can perform well as compare to FP growth. This paper includes clustering based on geological location, classification based on mineral resource type and spatial coherence between mineral resources. Spatial data mining find out the different association rules by observing the spatial data by using Apriori algorithm. The result of the paper will indicate the accurate prediction of occurrence of commodity with respect to other commodity of mineral resources.

Keywords: *Hadoop, data mining, association rules, clustering, spatial coherence*

----------***----------

1. INTRODUCTION

FP growth algorithm is a very popular algorithm used for the association mining due its performance and less storage requirement. With these pros, it has some limitations also which are considered rarely. This paper focuses on the limitations of FP growth association mining algorithm and trying to suggest improvements in Apriori algorithm which can overcome the limitations of FP Growth association mining algorithm. As we aware about the hardware cost, it's decreasing day by day, so there is no need to concentrate more on storage requirement. FP growth has two limitations – 1. It is difficult to use for interactive mining where the user may change support value as per requirement. 2. It is not suitable where the data has been increasing with time. So to deal with limitations of FP growth, this paper suggest to implement Apriori algorithm using Hadoop for association rule mining.

1.1 Mineral Resources Data System

Mineral Resources Data System is a collection of data describing metallic and nonmetallic mineral resources in the world [7]. It includes resource name, location, commodity, geologic characteristics, resource description, production, reserves, and references. As MRDS contains mineral resources data around the world, it is large and complex. If data size goes beyond the Tera Byte, it is difficult to process and mine using FP growth algorithm. The performance of FP growth algorithm hampers when adding the new records and also by changing the support value. As the mineral resources data set is collected from various regions and people, we need to perform ETL (extract, transform and load) operations on the data for processing and mining.

1.2 Hadoop Map Reduce

To deal with unstructured and big data like mineral resources data system (MRDS), we need a best technology which can cope with it. There are two options available, parallel DBMS and Hadoop Map Reduce technology. Hadoop has another projects also which can be used for the mining purpose like Hive. But internally again all the projects using Hadoop Map Reduce technique [5]. It gives a better data processing performance with minimum cost and time as compare to parallel DBMS because it works with commodity hardware. Hadoop stores data in the form of blocks on Hadoop Distributed File System i.e. HDFS. The Hadoop framework provides a solution for problems of massive data processing; because it runs applications on large cluster built of commodity hardware with failure tolerance [4].Unstructured data can be processed with Hadoop Map Reduce technique which is not possible with RDBMS. Map Reduce provides flexibility and fault tolerance which is not with parallel DBMS. Map Reduce provides automatic parallelization, data partitioning, task scheduling, handling machine failures and manages inter-machine communication. Hadoop is totally transparent from the end user. The rate of growing an unstructured data is much more as compare to the structured data. The unstructured data includes media files, heavy text files, csv files, log files etc.

2. RELATED WORK

Association rule mining algorithm includes to find coverage, support, confidence, lift and interesting [4]. Coverage defines the proportion of case data specified on the Left Haft Hand Side of the rule. Support of an association rule means percentage of task relevant data for which the pattern is true. Confidence gives the trustworthiness associated with the patterns discovered. Lift is a measure of the importance of the association. The term interesting gives the strength of associations between sets of items in the association rule.

Hongyong Yu, Deshuai Wang [1] proposed a system for data processing and mining log data of SaaS cloud using Hadoop. This paper focuses on Hadoop's Map Reduce technique and the algorithm used for data mining by Hongyong Yu, Deshuai Wang. They suggest by applying Apriori algorithm concurrently in the distributed system, performance of Apriori algorithm can be increased in proportional to the number of nodes in the distributed system [1].

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [3].The commonly used software technology cannot cope with massive data and big challenge is to extract an important information from it. Big data has large volume, heterogeneous format and decentralized data control. The example of big data applications are Facebook, Twitter and Google. It is a big challenge to manage and mining a massive data because of its volume, different file formats and growing rate of the data in the world. There are many challenges with big data such as storage, processing, variety and cost.

3. PROPOSED SYSTEM

This paper proposes a system which overcomes the problem faced in FP growth association mining algorithm. To improve the performance of Apriori algorithm, it is implemented parallely using Hadoop map reduce technique. The proposed system has three modules

1. Spatial Clustering
2. Spatial Classification
3. Spatial Coherence

Before implementing these modules, we need to perform Extract, Transfer and Load operations on the raw data. Because the data available may be in the various form and having an unnecessary information into to it. So first we need to extract the required data from raw data and then transfer it in to the csv format. To process this data, it should be loaded on to the HDFS.

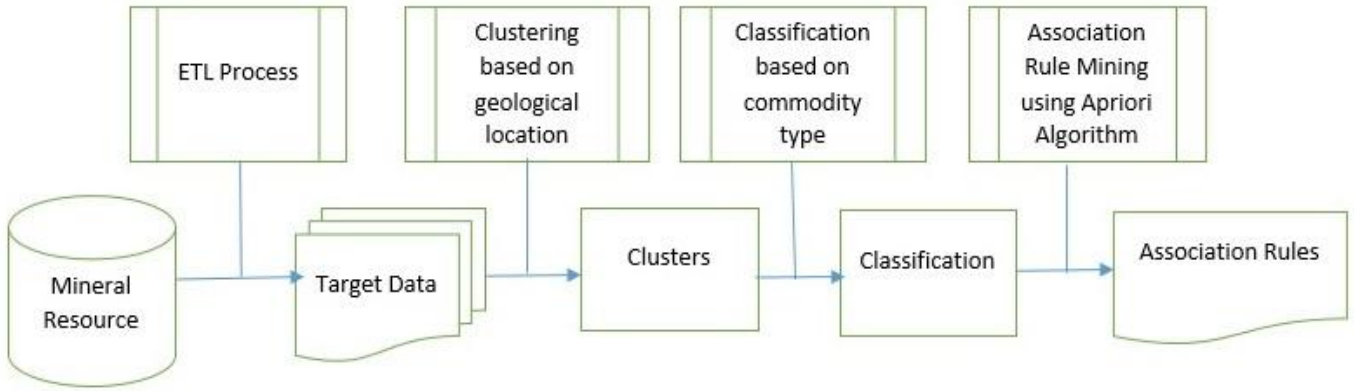

Fig-1 Proposed System

3.1 Spatial Clustering

There are various algorithms available for clustering like k means but here in this paper, Hadoop partition technique is used to perform spatial clustering based on the geological location like Alaska, California etc[2]. It runs parallel on different nodes concurrently. Because of this, time required for spatial clustering is less as compared to k-means clustering algorithm. In this paper, U.S. mineral resources dataset is taken. In effect, there are 50 clusters formed with respect to the states in United State.

These clusters have all the records which belongs to respective state. Intention behind to form the clusters is to make it specialize and reduce the processing time and cost.

3.2 Spatial Classification

Here, the same technique is used for implementing a spatial classification which is used for spatial clustering. In this module, again the data set is clustered according to the type of commodity which is required for the next module. In spatial classification, the records having same commodity put in to the separate cluster to make it more specialize. It means, e.g. Alaska state may have another sub clusters with commodity type like gold, silver, copper etc.

3.3 Spatial Coherence

Objective of this paper is to find out the spatial coherence. Spatial coherence means if we are mining for gold then what are the possibilities of getting other mineral resources at the same location. For this, association mining is required. As we have discussed above, FP Growth algorithms fails in two situations. So to overcome this, this paper introduces parallel Apriori algorithm for spatial association mining. In Apriori algorithm, first of all, we have to find out the candidate keys. After this, we have to perform two task, one is to find out the support values and second is to find out the confidence. Formula's to calculate a support and confidence for gold with copper.

1. Support Formula

$$Support(gold => copper) = \frac{(gold \cup copper)}{N}$$

Where N is all the records in a cluster

2. Confidence Formula

$$Confidence(gold => copper) = \frac{Sup(gold \cup copper)}{Sup(gold)}$$

Here, a formulas for calculating support and confidence of gold is given. Support of (gold => copper) gives the probability of having gold with copper in all the records whereas confidence gives the probability of having of gold with copper with respect to all the gold records in a cluster. The value of support is always less than the confidence.

4. RESULTS

To implement this idea, we have taken a dataset of MRDS which is a spatial data set of mineral resources in U.S.[7]. This sample data has around 3.5 lack records.

Table -1: Support and Confidence

Commodity	Support (%)	Confidence (%)
(Gold -> Copper)	2.89	9.30
(Gold -> Copper, Lead, Silver)	0.72	2.3
(Gold -> Silver)	2.17	6.97
(Gold -> null)	24.63	79.06

Here, the Apriori algorithm is performed on single node and multi node Hadoop to compare the performance. Also we find the association rules for this spatial data for a single cluster.

Table -2: Execution on Single and Multinode System

No. of Node	Execution Time in Sec
1	5.2
2	3.1
3	2.3

The result in Table-1 shows the different association rules with respect to gold.Table-2 shows the performance of Apriori algorithm using single node and multi node system. It is observed that if we implement the Apriori Algorithm parallel on Hadoop, performance is improved.

5. CONCLUSION

FP growth algorithm has two limitations, it cannot be used for dynamic data size and where the support needs to change according to situation. By applying the Apriori Algorithm concurrently using Hadoop, we can overcome the problems faced by FP growth association mining algorithm. Also we can improve the speed of association rule mining for spatial data as compare to the FP growth algorithm as Hadoop is a distributed system.

ACKNOWLEDGEMENTS

I express true sense of gratitude towards my project guide Prof. H.A. Hingoliwala, Associate Professor Computer Department for his invaluable co-operation and guidance that he gave me throughout my project. I specially thank our P.G coordinator Prof. M. D. Ingle for inspiring me and providing me all the lab facilities. I would also like to express my appreciation and thanks to HOD Prof. S.M. Shinde & JSCOE Principal Dr. M.G. Jadhav and all my friends who knowingly or unknowingly have assisted me throughout my hard work

REFERENCES

[1]. Hongyong Yu, Deshuai Wang, “Mass Log Data Processing and Mining Based on Hadoop and Cloud Computing” .The 7th International Conference on Computer Science & Education (ICCSE 2012)July 14-17, 2012. Melbourne, Australia.

[2]. Duck-Ho Bae Coll. of Inf. & Commun., Hanyang Univ., Seoul, South Korea Ji-Haeng Baek ; Hyun-Kyo Oh ; Ju-Won Song ; Sang-Wook Kim, “SD-Miner: A SPATIAL DATA MINING SYSTEM” Network Infrastructure and Digital Content, 2009.

[3]. Xindong Wu,Fellow, IEEE,Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding,Senior Member, IEEE, “Data mining with big data,” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.

[4]. Binbin He, Ying Cui, Jianhua Chen, Pingjing Xie, “A Spatial Data Mining Method for Mineral Resources Potential Assessment,” IEEE 978-1-4244-8351-8/11,2011IEEE.

[5]. Hadoop: The definitive Guide, 3rd ed., O’Reilly, Tom White, 2012

[6]. Hadoop, http://hadoop.apache.org/

[7]. MRDS, http://tin.er.usgs.gov/mrds/