|| Volume 6 || Issue 6 || June 2021 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

AN EFFICIENT FREQUENT PATTERNS MINING BY USING MAP-REDUCE IN HADOOP

HUMA FATIMA SHAIKH MOINUDDIN[1], ASST.PROF. V. KARWANDE[2]

ME Student, Department of Computer Science & Engineering, EESGOI , India[1]
HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI,India. [2]

Abstract:* In many real-life applications, frequent pattern mining is widely employed. Since its debut, many academics have been attracted by the mining of common patterns from exact data. More emphasis has been paid in recent years to the mining of ambiguous data. Items in each transaction of this uncertain data are generally linked to existential probabilities that describe the probability that these articles will be present during the transaction. Compared to the accurate data extraction, the space for the search/solution of the uncertain data is significantly bigger owing to the existential probability. The models provided are based on the widely known Apriori and MapReduce algorithms. The algorithms suggested are split into three primary classes. Two Apriori MapReduce and AprioriMR methods are intended to properly extract patterns in big datasets. These algorithms extract any existing data items irrespective of their frequency. Tape the search space with the antimonotonic characteristic. Two more space trimming methods AprioriMR and top AprioriMR are introduced with the objective of identifying any common data patterns. Maximum common patterns. In addition, we live in the Big Data age. Furthermore, we offer certain improvements to increase its performance further. Experimental findings indicate the efficacy of the MapReduce for Big Data Analytics algorithm and its improvements in mining frequent patterns from unspecified data.

***Keywords**: MapReduce (MR); Hadoop Archives (HAR); Sequential Pattern Mining (SPM); Parallel Frequent Pattern Growth (PFPG).*

I INTRODUCTION

The DATA analysis is more and more interested in various sectors, such as business intelligence, which covers a number of approaches for converting raw data into meaningful and practical information for business analysis. With the rising significance of data in each application, the amount of data to be processed is unmanageable and the performance of these approaches can be reduced. The phrase Big Data is increasingly used to refer to the difficulties and gains resulting from the efficient processing of very large datasets. Model mining is regarded to be an important component of data analysis and data mining. Its objective is to extract subsections, substructures or object sets that reflect any sort of homogeneity and regularity of data, indicating key inherent features. In order to discover frequent groupings of items bought together, this problem was first suggested as part of the market basket analysis. Since its official formulation, a great number of algorithms have been described in the beginning of the 90s. Most of these algorithms are based on Apriori-like approaches, generating a list of things or patterns that consist of any combination of elements. However, as the total quantity of these things rises, the problem of pattern mining becomes a difficult task and more efficient techniques are necessary. Let us imagine a dataset consisting of no single items or singletons to understand the complexity. The number of item-sets that may be created is equal to 2n−1, such that with the growing number of singletons, it gets exceedingly complicated. All this led to a reasonable result that it is not always possible to examine the whole range of options.

However, in many application domains, it is not necessary to create an existing collection of items but just those judged to be of interest, e.g. those covered by a large number of transactions. In order to do this, various approaches were developed, some of which were based on the anti-monotone characteristic as a cutting strategy. It establishes that each sub-pattern in a frequent pattern is likewise common, and any super-pattern in an unusual pattern is never common. This cutting technique reduces the search space as no new pattern must be produced if a pattern is described as unusual. Despite that, large amounts of data have decreased the performance of existing approaches in many domains of application. Traditional pattern mining techniques are not appropriate for genuinely large data and pose 2 major challenges: (1) computer complexity and (2) primary storage needs. In this case, sequential pattern mining algorithms on a single computer cannot manage the entire procedure, and it might be essential to adapt them to new technologies.

MapReduce is a programming approach for big data sets with a parallel algorithm spread across a cluster. Map Reduce the handling of large data when used with HDFS. A key value pair is the basic information unit utilised in MapReduce. All structured and unstructured data types must be converted into this fundamental unit before data is sent to the MapReduce paradigm. MapReduce consists, as the name indicates, of two distinct routines: map function and reduction function. Unlike other data frameworks, MapReduce logic is not limited to simply structured datasets. It is also able to handle unstructured data extensively. The essential phase that makes this feasible is the map stage. Mapper provides an unstructured data structure.

II LITERATURE SURVEY

In this research, Frequent pattern mining in data mining research is an important topic. With the period of big data, the database size rises quickly. How to efficiently measure common patterns from large transaction databases is always a difficulty. A parallel approach to the problem is the mining algorithm. Traditional parallel algorithms have problems, however, with workload balance and failure recovery. A novel parallel method based on MapReduce is therefore offered with three contributions. Firstly, a hybrid mining approach is presented. This automatically changes from broad first to deep first mining and simultaneously carries out broad first mining and depth first mining. Second, in broad-first mining a hybrid vertical mix data format is employed, and a novel approach is suggested to turn a mix set back to a horizontal data display that helps first-depth mining. Thirdly, techniques are given for reducing the number of candidates for the first-largest mining and facilitating the first-largest mining to prevent candidates generating, saving both space and time. The results demonstrate that the suggested method outperforms and is highly scalable, the current MapReduce based techniques [1].

In this research, Pattern mining is a basic data mining approach for finding interesting connections in the data collection. There are numerous different models of mining, for example frequent mining of goods, sequence mining and high utility mining. High utility itemset mining is a new data science activity aimed at extracting information on a domain basis. The usefulness of a pattern indicates that it may be determined on the basis of user priority and domain-specific expertise. The problem of sequential pattern mining (SPM) is extensively explored in many ways. Sequential pattern exploitation lists sequential patterns in the gathering of sequence data. In recent years, researchers have focused increasingly on the frequent pattern mining of ambiguous data for transactions. Mining item sets in big data have gained considerable attention in recent years, based on the architecture of Apache Hadoop and Spark. This study attempts to provide a general overview of the many techniques to pattern mining in the field of big data. We explore initially the topic of pattern mining, and related solutions like Apache Hadoop, Apache Spark, parallel and distributed processing. Then we review key advancements in parallel, distributed and scalable pattern mining, assess them from the standpoint of large data and highlight problems in the design of algorithms. Particularly in the uncertain data, we examine four types of mining of articles: parallel frequent mining of articles, high utility mining, sequential mining patterns and frequent mining of articles. This paper closes with a debate on open concerns and opportunities. It also gives guidance to further improve existing techniques [2].

In this research, Data analysis is a crucial part of the decision-making process. The insights of such pattern analysis offer huge benefits, including more revenue, reduced costs and enhanced competitive advantages. However, the underlying patterns of frequent itemsets are longer to be taken into account when the amount of data rises over time. In addition, the mining of the hidden patterns of the often produced objects requires substantial memory usage owing to intensive algorithm calculations. A powerful algorithm is therefore required to evaluate the hidden patterns of frequent itemsets in a shorter time and with a lower memory use, whereas the data volume rises with time. This study analyses and compares the various FPM methods so that a more efficient FPM algorithm may be devised [3].

In this research, Frequent pattern mining is an effective technique for analysis of mobile trajectory large data in intelligent transport systems with spatio-temporal associations. Although previous parallel methods have been effective in the common pattern of large-scale trajectory data mining, it is a two main problem to deal with Hadoop's intrinsic flaws, including enormous tiny files, and to find out about MapReduce implicitly spatiotemporal patterns. This article offers a MapReduce-based Parallel Frequent Pattern Growth (MR-PFP) technique for analysing taxi space-temporal features utilising large-scale taxi tracks with big small-scale processing strategies on a platform in Hadoop. To address these problems. More precisely, we develop first three ways to overcome Hadoop Archives (HAR), Combine File Input Format (CFIF), and Sequence Files (SF) and then suggest two solutions based on their performance reviews. Next, we include the SF in a Frequent Pattern Growth method and subsequently deploy an optimised MapReduce FP Growth algorithm. Finally, we examine in parallel the features of MR-PFP taxis functioning in spatial and temporal dimensions. The findings show that the MR-PFP is superior in efficiency and scalability than the Parallel FP-growth (PFP) method [4].

In this research, 'Big data' is an emerging subject that has drawn interest in industrial system engineering and cybernetics from many scholars and practitioners. Big data analytics will surely lead many businesses to important insights. Business and risk management activities can benefit, since various data gathering channels are available in the associated industrial systems e.g., wireless sensor networks, Internet-based systems, etc. However, big data research is still in its early stages. Its emphasis is uncertain and related research are not properly combined. This article presents the problems and potential of Big Data Analytics in this specific field of application. Technological progress and advancements for industrial business systems, dependability, security and operational risk management of industrial systems are explored. Important topics will also be explored and exposed for further investigation [5].

In this research, every second huge amounts of data are created with the advent of the Big Data age. In the past, many methods and structures for data processing were suggested to improve the execution of data mining algorithms. One such method extracts patterns from the transactional database most often. Depending on transactions on time and location, the frequent mining work becomes more complicated. The objective of the present study is to discover and extract the frequent patterns from such transactional data. The spatio-temporal dependence of air quality data is utilised mostly to detect contaminants often appearing at several points in Delhi, India's capital. There have been several techniques in the past to effectively extract common patterns, but this study offers a wider strategy that can be used for any numerical spatio-temporal transactional data, including air quality data. In addition, this paper presents a detailed description of the algorithm and a typical example of the air quality data set. The synthetically created data sets, benchmark datasets and actual world datasets are examined in depth. In addition, a comparison is provided with the Spatio-Temporal Apriori and other state-of-the-art non-aprior algorithms. Results indicate that the suggested method exceeded previous approaches in terms of algorithm execution time and memory resources [6].

III. SYSTEMS ARCHITECTURE

The system model architecture, new effective pattern mining algorithms have been presented to work in huge data. All are based on the MapReduce framework and the open-source implementation of Hadoop. Two of these AprioriMR and IAprioriMR algorithms allow the discovery of existing patterns. Two further SPAprioriMR and TopAprioriMR algorithms utilise a cutting technique for common patterns. Finally, a method is also suggested for mining MaxAprioriMR. Chan et al. presented the challenge of high utility pattern mining. However, the definition of high use goods utilised in their study is different from the criteria used in this investigation. The study evaluated utility of different things, but no quantitative values of objects were taken into account in transactions. We have defined the task of high-level mining by considering both quantity and profit.

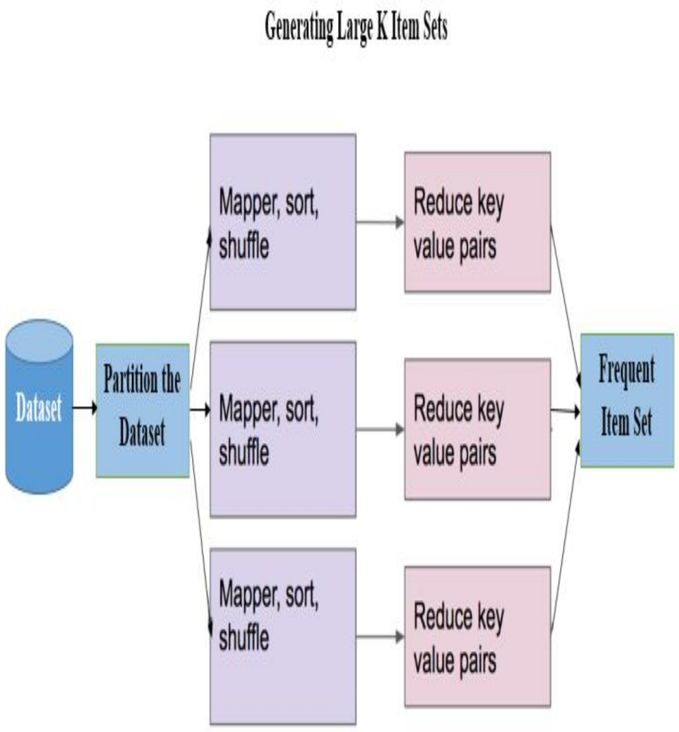

Figure No 3.1: System Architecture

IV EXPERIMENTAL RESULTS

The below are some technologies employed to produce this improvised system operation. The platform to create this project is Java as its indepension platform is Net beans. Net beans are safer, efficient and most efficient.

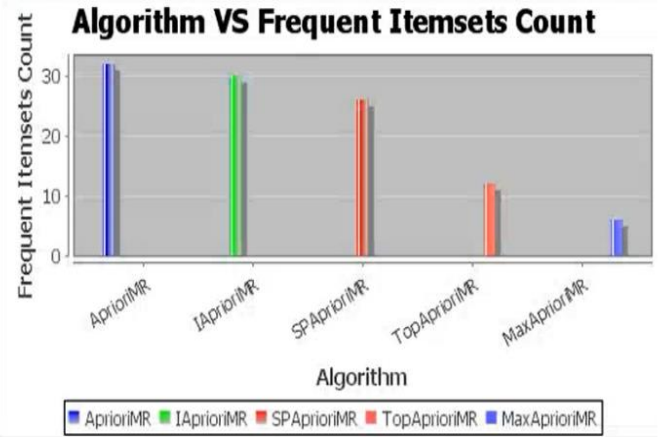

Figure No 4.1: Algorithm vs Frequent Itemsets Count.

V CONCLUSION

Proposed new effective Big Data pattern mining techniques. All models given are based on the well-known Apriori and the MapReduce algorithm. The algorithms suggested are split into three primary classes. The experimental phase include comparisons of both well-known sequential model mining techniques and ideas for MapReduce. The ultimate objective of this study is to provide a foundation for future field research. Results showed the interest of utilising the MapReduce architecture while considering large data. They also have P which indicates that this framework is not appropriate for little data, therefore sequential methods are preferred. No strategy for trimming. Two AprioriMR and IAprioriMR algorithms have been suggested for mining any existing data pattern. Take the search area with anti-monotone properties. Two new SPAprioriMR and TopAprioriMR algorithms have been suggested to detect common data patterns. Maximum common patterns. A final method MaxAprioriMR for mining condensed depictions of common patterns has also been presented.

REFERENCES

1. 1.Junqiang Liu, Xiangcai Yang, Yanjun Hu, Bo Jiang, Yong Zhang and Zhousheng Ye,"Distributed Mining of Frequent Patterns in Big Data by Hybrid Strategies",IEEE International Conference on Data Mining Workshops (ICDMW),2019.

2.Sunil Kumar and Krishna Kumar Mohbey,"A review on big data based parallel and distributed approaches of pattern mining",Journal of King Saud University Computer and Information Sciences,2019.

3.Chin Hoong Chee,Jafreezal Jaafar,Izzatdin Abdul Aziz,Mohd Hilmi Hasan and William Yeoh,"Algorithms for frequent itemset mining: a literature review",Springer,2018.

4.Dawen Xia,Xiaonan Lu,Huaqing Li,Wendong Wang,Yantao Li and Zili Zhang,"A MapReduce-Based Parallel Frequent Pattern Growth Algorithm for Spatiotemporal Association Analysis of Mobile Trajectory Big Data",Hindawi Complexity Volume,2018.

5.Tsan Ming Choi,Hing Kai Chan and Xiaohang Yue,"Recent Development in Big Data Analytics for Business Operations and Risk Management",IEEE Transactions on Cybernetics,2016.

6.Apeksha Aggarwal and Durga Toshniwal,"Frequent Pattern Mining on Time and Location Aware Air Quality Data",IEEE Access,Volume 4,2016.

7.Carson Kai-Sang Leung,Richard Kyle MacKinnon and Fan Jiang,"Finding efficiencies in frequent pattern mining from big uncertain data",Springer,6 September 2016.

8.Chowdhury Farhan Ahmed,Md. Samiullah,Nicolas Lachiche,Meelis Kull and Peter Flach,"Reframing in Frequent Pattern Mining",IEEE 27th International Conference on Tools with Artificial Intelligence,2015.

9.Lan Vu and Gita Alaghband,"Efficient Algorithms for Mining Frequent Patterns from Sparse and Dense Databases",De Gruyter,181-197,September 19, 2014.

10.Sandy Moens, Emin Aksehirli and Bart Goethals,"Frequent Itemset Mining for Big Data",IEEE International Conference on Big Data,2013.

11.Bo Wu, Defu Zhang, Qihua Lan and Jiemin Zheng,"An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure ",IEEE International Conference on Convergence and Hybrid Information Technology,2008.

12.Jiawei Han,Hong Cheng, Dong Xin and Xifeng Yan,"Frequent pattern mining: current status and future directions",Springer Science Business Media,27 January 2007.

13.Gosta Grahne and Jianfei Zhu,"Fast Algorithms for Frequent Itemset Mining Using FP-Trees",IEEE Transactions on Knowledge And Data Engineering, Vol. 17, No. 10, October 2005.