IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A STATISTICAL APPROACH USING AN EFFECTIVE HADOOP FREQUENT PATTERNS MINING

SAYED FARZEEN ISHAKODDIN[1], ASST.PROF. V. S. 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: Frequent pattern mining is widely used in many real-world applications. The mining of common patterns from accurate data has drawn the attention of numerous scholars since its introduction. The mining of unclear data has received more attention in the last few years. Every transaction involving this uncertain data often has items associated with existential probabilities that indicate the likelihood that these articles will be present at the time of the transaction. Because of the existential probability, there is a much larger area available for the search and solution of uncertain data as compared to precise data extraction. The popular MapReduce and Apriori algorithms serve as the foundation for the models that are offered. The proposed algorithms are divided into three main types. The goal of two Apriori MapReduce and AprioriMR techniques is to accurately identify patterns in large datasets. These algorithms retrieve any available data item, regardless of how frequently it occurs. Put tape on the antimonotonic feature of the search space. AprioriMR and top AprioriMR, two more space trimming techniques, are presented with the aim of finding any shared data patterns. maximum number of frequent patterns. We also live in the Big Data era. Moreover, we propose certain enhancements to further boost its efficiency. The MapReduce for Big Data Analytics method is effective at extracting common patterns from undefined data, as demonstrated by experimental results.

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

I INTRODUCTION

The field of business intelligence, which encompasses a variety of methods for transforming unstructured data into useful and applicable information for business analysis, is one that the DATA analysis is becoming increasingly interested in. The volume of data that has to be processed is unmanageable due to the increasing importance of data in every application, and this might lower the effectiveness of these methods. The challenges and benefits that arise from processing extremely large information efficiently are frequently referred to as "big data." Model mining is seen to be a crucial part of both data mining and data analysis. Its goal is to identify important innate properties by extracting subsections, substructures, or object sets that show any kind of homogeneity and regularity of data. This problem was first proposed as a component of the market basket analysis in order to identify recurring groupings of products purchased together. Numerous algorithms have been reported since the formal formulation in the early 1990s. The majority of these algorithms generate a list of items or patterns made up of any combination of components by using techniques similar to those used in Apriori. However, pattern mining becomes a challenging undertaking and more effective strategies are required as the overall amount of these objects increases. To get an idea of the complexity, consider a dataset that has no singletons or single objects in it.It becomes extremely complex as the number of singletons increases, as the number of item-sets that may be generated is equal to $2n-1$. All of this pointed to the logical conclusion that it isn't always feasible to look at every possibility.

But in many application domains, it suffices to construct a collection of objects that are deemed interesting—that is,

those that are covered by a high volume of transactions, for example—rather of creating an entire collection of items. Several methods were devised to do this, some of which relied on the anti-monotone trait as a cutting technique. It proves that every super-pattern in an uncommon pattern is never common, and that any sub-pattern in a frequent pattern is also common.

By describing a pattern as uncommon, this cutting strategy lowers the search space by eliminating the need to create a new pattern. Despite this, the performance of current methods has declined in many application domains because to the abundance of data. There are two main issues with traditional pattern mining approaches: (1) computer complexity and (2) primary storage demands. These techniques are inappropriate for really huge data sets. In this scenario, sequential pattern mining algorithms may need to be modified to accommodate new technologies as they are unable to handle the full process on a single machine.

Programming large data sets using a parallel technique dispersed over a cluster is known as MapReduce. When used with HDFS, Map Reduces the handling of huge data.The fundamental information unit used in MapReduce is a key-value pair. Before data is transmitted to the MapReduce paradigm, it must be transformed into this fundamental unit for both structured and unstructured data types. As the name suggests, MapReduce is made up of two separate routines: the reduction function and the map function. The logic of MapReduce is not restricted to just structured datasets, in contrast to other data frameworks. It can also manage large amounts of unstructured data. The map stage is a crucial step that makes this possible. An unstructured data structure is offered by Mapper.

II LITERATURE SURVEY

In this research, A key area of study in data mining research is frequent pattern mining. The size of the database increases rapidly during the big data era. It's never easy to figure out how to measure frequent patterns from big transaction databases effectively. The mining algorithm is one method of approaching the issue in parallel. However, traditional parallel algorithms suffer from issues related to failure recovery and workload balancing. Thus, a new MapReduce-based parallel approach is presented with three contributions. First, a hybrid mining strategy is explained. This performs both depth first and broad first mining concurrently, switching automatically between broad first and deep first mining. Secondly, in broad-first mining, a unique method is proposed to convert a mix set back to a horizontal data display that facilitates first-depth mining, using a hybrid vertical mix data format. Thirdly, methods are provided for minimizing the quantity of candidates for the first-largest mining and streamlining the mining process to avoid the generation of candidates, therefore conserving time and space. The outcomes show that the recommended method performs better than the existing MapReduce based solutions and is extremely scalable. [1].

In this research, A fundamental data mining technique for identifying intriguing correlations in the data collection is pattern mining. There are many various types of mining models, such as high utility mining, sequence mining, and frequent mining of items. A new field of data science called high utility itemset mining seeks to extract information on a domain-by-domain basis. A pattern's utility suggests that it may be chosen based on user priorities and domain-specific knowledge. Sequential pattern mining (SPM) is a well studied issue in many domains. In the process of acquiring sequence data, sequential patterns are listed using sequential pattern exploitation. Researchers have been concentrating more and more on the regular pattern mining of unclear data for transactions in recent years. This paper aims to give a broad overview of the various approaches to big data pattern mining. First, we look into pattern mining and its associated technologies, such as distributed and parallel processing, Apache Spark, and Hadoop. Next, we examine significant developments in distributed, parallel, and scalable pattern mining, evaluate them in light of massive data, and point out issues with algorithm design. We analyze four forms of article mining: sequential mining patterns, high utility mining, frequent mining, and parallel frequent mining, especially in the uncertain data. This article concludes with a discussion of unresolved issues and prospects. It also provides direction for enhancing current methods even more. [2].

In this research, An essential step in the decision-making process is data analysis. Such pattern analysis's findings may lead to significant advantages such as increased competitive advantages, lower expenses, and more income. When the volume of data increases over time, it becomes necessary to consider the underlying patterns of frequently occurring itemsets. Furthermore, due to intense algorithm calculations, mining the hidden patterns of the often created objects needs significant memory utilization. Thus, when the data amount increases over time, a strong algorithm is needed to assess the hidden patterns of frequently occurring itemsets in less time and with less memory use. In order to create an FPM algorithm that is more effective, this study examines and contrasts the different FPM techniques. [3].

In this research, A useful method for analyzing huge amounts of mobile trajectory data in intelligent transportation systems with spatiotemporal correlations is frequent pattern mining. While prior parallel approaches have demonstrated efficacy in the common pattern of large-scale trajectory data

mining, addressing Hadoop's inherent shortcomings—such as massively small files and uncovering underlying spatiotemporal patterns in MapReduce—remains a dual challenge. This paper presents a MapReduce-based Parallel Frequent Pattern Growth (MR-PFP) method for Hadoop platform analysis of taxi space-temporal characteristics by combining massive small-scale processing techniques with large-scale taxi tracks. to deal with these issues. To be more specific, we create the first three methods of overcoming Sequence Files (SF), Hadoop Archives (HAR), and Combine File Input Format (CFIF) and then provide two solutions depending on their performance reviews.The SF is then included into a Frequent Pattern Growth technique, and an optimized MapReduce FP Growth algorithm is then implemented. In conclusion, we investigate concurrently the characteristics of MR-PFP taxis operating in both temporal and spatial dimensions. The results demonstrate that MR-PFP outperforms Parallel FP-growth (PFP) in terms of efficiency and scalability. [4].

In this research, A growing body of research and practice in industrial system engineering and cybernetics has been captivated by the topic of "big data." Big data analytics will undoubtedly provide crucial insights for a lot of firms. Since the related industrial systems have several data collection channels, such as wireless sensor networks and Internet-based systems, business and risk management operations stand to gain. Big data research is still in its infancy, though. Its focus is ambiguous, and relevant studies are not well integrated. The challenges and possibilities of big data analytics in this particular application area are discussed in this article. Technological developments are examined with respect to industrial business systems, operational risk management, security, and dependability. Crucial subjects will also be investigated and made public for further investigation [5].

In this research, The Big Data era has brought to the creation of enormous volumes of data every second. Several data processing techniques and frameworks have been proposed in the past to enhance the performance of data mining algorithms. The most common pattern extracted from the transactional database is one such way. The frequency and location of transactions determine how complex the mining task gets. Finding and extracting recurrent patterns from such transactional data is the goal of this study. The primary use of the spatiotemporal dependency of air quality data is the detection of pollutants that often emerge at many locations in Delhi, the capital of India. While there have been a number of successful methods for extracting common patterns in the past, this work presents a more comprehensive approach that can be used to any numerical spatio-temporal transactional data, including data on air quality. This paper also includes a representative example of the air quality data set and a full description of the technique. A thorough analysis is conducted on the benchmark datasets, real-world datasets, and artificially generated datasets. Furthermore, a comparative analysis is presented between the Spatio-Temporal Apriori and other cutting-edge non-aprior algorithms. The results show that the recommended strategy outperformed earlier techniques in terms of memory requirements and algorithm execution time. [6].

III. SYSTEMS ARCHITECTURE

The system model architecture, with the system model design, new efficient pattern mining algorithms that can handle large amounts of data have been introduced. They're all built on top of the open-source Hadoop implementation and the Map Reduce framework. It is possible to find preexisting patterns using two of these Apriori MR and I Apriori MR techniques. For frequent patterns, two more SP Apriori MR and Top Apriori MR Algorithms employ a cutting strategy. Lastly, a mining strategy for Max Apriori MR is also proposed. High utility pattern mining was the challenge put up by Chan et al. Nevertheless, the criteria applied in this analysis differ from the concept of high use commodities employed in their study. The study assessed the usefulness of various items; however, no consideration was given to the numerical values of the objects in transactions. The high-level mining task has been defined by taking into both quantity and profit.

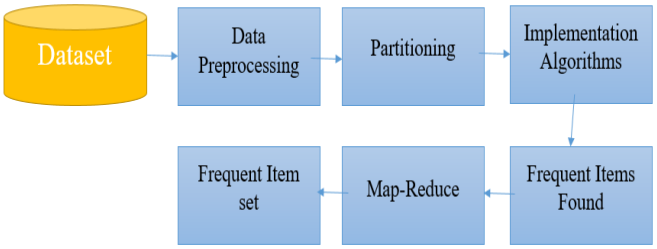

Figure No 3.1: System Architecture

IV EXPERIMENTAL RESULTS

Here are a few of the technologies used to create this makeshift system. This project was created using the Java platform, which is dependent on the Net Beans platform. Net beans are the most effective, safest, and efficient.

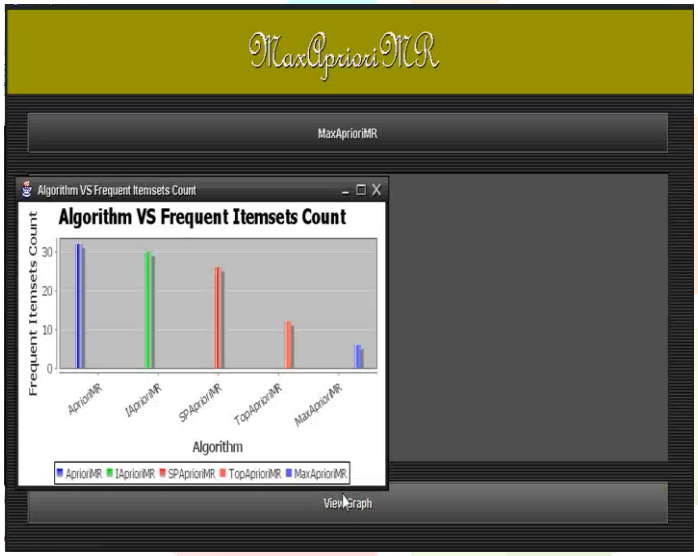

Figure No 4.1: Algorithm vs Frequent Itemsets Count.

V CONCLUSION

Proposed new effective Big Data pattern mining techniques. suggested fresh approaches to pattern mining in big data that work well. Every model provided is based on the popular MapReduce algorithm and Apriori. The proposed algorithms are divided into three main types. During the experimental phase, both popular sequential model mining methods and MapReduce concepts are compared. Ultimately, this study aims to provide the groundwork for subsequent field research. The use of the MapReduce architecture while taking big data into consideration was shown by the results. Additionally, they have P, which suggests that sequential approaches are preferable because this framework is inappropriate for little amounts of data. No plan for cutting. There are two proposed AprioriMR and IAprioriMR algorithms for mining any given preexisting data pattern. Consider the anti-monotone search region. To identify common data patterns, two novel SPAprioriMR and TopAprioriMR algorithms have been proposed. maximum number of frequent patterns. Additionally, a fourth approach, MaxAprioriMR, has been given for mining condensed representations of frequent patterns.

REFERENCES

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

2.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.

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

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

5.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.

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

7.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.

8.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.

9.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.

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

11.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.

12.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.

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