EPGA: Enhanced Pattern Growth Algorithm for Sequential Pattern Mining in Hadoop-Mapreduce Framework for Big Data

Sujit R Wakchaure

Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore(M.P.) 452010

sujitw2777@gmail.com

Dr. Rajeev G Vishwakarma

Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore (M.P.) 452010

rajeev@mail.com

Article Info
Page Number: 12372-12387
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 20 June 2022
Revised: 15 July 2022
Accepted: 10 August 2022

Abstract : Mining frequent itemsets plays a crucial part in mining associations, relations, causality, and other significant data mining activities. This is because Frequent Itemset Mining (FIM) is an essential component of the process of uncovering association rules. Due to factors like high memory consumption, high I/O overhead, and poor processing speed, certain conventional frequent itemsets mining techniques are not able to successfully manage enormous tiny file datasets. Therefore, as the size of the data increases, the single machine FIM method faces the challenges of taking a significant amount of time and using a significant amount of memory. Therefore, a new implementation technique named as the Enhanced Pattern Growth Algorithm (EPGA) is proposed. This technique relies on a MapReduce parallel environment and is designed for mining frequent item sets in order to produce association rules. This method is validated by employing various sizes of real-time big datasets on various nodes in the cluster while simultaneously choosing speedup, reliability as a criterion. The findings indicate that the suggested method is both practicable and reasonable, and that it has the potential to enhance the general performance as well as the effectiveness of the Apriori and FP-Growth algorithms in order to fulfil the requirements of big data association rules mining.

Keywords: Enhanced Pattern Growth algorithm, Apriori, Sequential Pattern mining, Hadoop Map reduce, Big data.

I. Introduction

For more than a decade, one of the primary focuses of study into data mining has indeed been mining for frequent patterns. An abundance of literature has been devoted to this area of investigation, and significant progress has been made in the field. These advancements range from effective and scalable methodologies for FIM in transaction databases to various research frontiers, like correlation mining, sequential pattern mining, organised pattern mining, associative categorization, and frequent pattern-based grouping, as well as the various uses of these techniques [24, 29, 30]. There is a widespread consensus that research on frequent pattern mining has significantly widened the application scope of data evaluation and it will, in the long term, have a significant influence on the methodology and uses of data mining. Nevertheless, until frequent pattern mining can be considered a foundation strategy in data mining methods, there are still several difficult research problems that need to be addressed.

The stage in association rules mining known as "Frequent Itemset Mining" is the one that requires the most effort and time investment. In recent decades, one of the most active research areas within the field of data mining has been focusing on increasing the effectiveness of FIM [19, 20]. The conventional FIM method is a serial process that only requires a single computer in order to extract a small-scale data collection effectively. Furthermore, if the data set's size continues to grow, the serial technique will eventually fail because of a memory bottleneck or because it will be unable to keep up due to the limitations of a single piece of hardware [18]. There are some good big data platforms depending on distributed architecture, like MapReduce and Spark, which present new chances to enhance the efficiency of mining techniques. This is due to the maturation of distributed computing technology [7].

Apriori and Frequent Pattern (FP-growth) methods are the two main categories that can be used to classify the many mining techniques for frequent item sets [16, 17, 26, 27, 28]. The traditional method known as Apriori makes use of a procedure known as generate-and-test, which produces a huge number of candidate item sets [23, 25]. In order to do this, Apriori must continuously scan a complete database. A revolutionary strategy known as FP-growth, which does not generate candidate item sets, was developed in order to cut down on the amount of time needed to scan databases [3]. Because of the increase in the size of the data set, FP-growth and Apriori are unable to complete the computational tasks in a suitable amount of time. Because of the constraints imposed by methods that run on a single machine, a number of academics have developed distributed algorithms that are dependent on the MapReduce framework [21, 22].

In this paper, using the concept of MapReduce programming, the conventional Apriori and FP-Growth algorithms are relocated to the MapReduce environment in to efficiently overcome the existing issues of Apriori and FP-Growth method in the conventional methodologies, and to fulfill the requirements of large-scale data association rules mining.

The remaining parts of this work are structured in the following manner. In the second section, a comprehensive discussion is made on the many frequent itemset mining strategies that have been carried out by researchers. In section 3, research methodology of proposed enhanced pattern growth algorithm is depicted. Then several experimental findings and discusion are presented in Section 4. Lastly, the paper is concluded in Section 5.

II. Literature Survey

An enhanced parallel FP-growth technique that is dependent on Spark was presented by Yuhang Miao et al. [1] in 2019. To begin, matrix technology has made the FP-growth method better. Additionally, the ability to condense data sets into data matrices can lower the amount of memory required. The subsequent step is to synchronise the enhanced FP-growth algorithm using Spark. In conclusion, the proposed method is utilised in thermal power plants in order to improve the effectiveness of steam turbines. The findings demonstrate that the proposed approach is more effective in terms of resource use than the currently used parallel FP-growth technique.

A complete mining approach that is based on the FP-growth algorithm was proposed by Ze-Zhong Wang et al. [2] in the year 2018. The first step was to use the K-means algorithm to cluster and generalise the data on meteorological, solar periods, holidays, and total load collected in the Pudong Region of Shanghai over the course of 546 days. After that, the initial transaction sets were categorised based on the numbers of the clustering result. In addition to this, they were mined using a variety of techniques to an extensive degree. When compared to association rules derived by conventional methods, the comprehensive mining approach has the ability to mine a greater number of association rules while maintaining its correctness and resilience. The exhaustive methodology offers a foundation for load forecasting in addition to distribution network load alerting, both of which are necessary for the operation and administration of smart grids.

By using publically available maintenance data to expose the trends and principles behind the numerous damages would present a workable alternative to the maintenance problems that have been occurring. This would be accomplished in the year 2020 by Lingxizhu et al. [3]. This research obtained container maintenance information and records from China Railway Container Transport Corporation limited for a period of nine years. Creates a maintenance element chain for the maintenance portions of the maintenance process, and transforms it into a sparsely matrix format data set in accordance with the specifications of the methodology. This is done based on the data that were provided. The analysis of the data set is performed with the FP growth technique. The level of the confidence as well as the lift level are adjusted to reflect the current circumstances, and the 178 rules for strong associations are afterwards obtained. According to the findings, the elements of the doors are the areas that are destroyed the most frequently, and there are significant associations between the types of damages that occur in the parts of the back side and the doors.

In the year 2020, Xiaolei Ma et al. [4] employed the FP growth approach to mine frequent itemsets. In order to separate the data and make it easier to interpret, they used event folding window. The findings, which were derived from a correlation analysis of real SCADA data, demonstrate that the technique has a substantial practical technical importance for accelerating fault diagnosis immediately following the occurrence of a problem in the power system.

The Frequent-Pattern Growth approach is used by Xueyi Gao et al. [5] in their analysis of metagenomics data for the purpose of assisted diagnosis of acne illness in 2019. The primary concepts are, first and foremost, the conversion of the data sets into a binary format consisting of either 0 or 1. The actual document for lipids in which the element content is 0 are changed to reflect the value 0, whereas the original data for lipids in which the element content is indeed not 0 are changed to reflect the value 1. After that, the data sets are analysed in order to construct a frequent pattern tree depending on the number of times each element appears and the support it has. The FP-tree is then utilised to complete the process of determining frequent itemsets. The element items that are included in frequent item sets relate to the lipids that have a strong correlation with the data that are being referred to by those element items. The suggested method was tested on a dataset that included normal control, pimples healthy skin and pimple diseased skin. The findings showed that it was able to identify the frequent item sets for each of the three separate sample sets. Lipids that are capable of distinguishing between various states of the skin can also be identified by examining the differences between frequent item sets. This can provide directing assistance for the supplementary evaluation and treatment of skin pimples.

In the year 2020, Yalu Jia et al. [6] propose a technique for the recognition of new vocabulary for microblog short text titled as Frequent Pattern Growth with Part-of-Speech (POS-FS). In the first step of the process, the candidate unidentified words are derived from the combination of the N-grams framework and the frequent item sets. After that, the unidentified word is put through a series of tests to determine whether or not it is accurate, including better mutual information, information entropy, and context dependence. In conclusion, the open verification procedure is applied in order to get the last unknown word. Investigations have shown that the technique enhanced the identification of unknown words in brief texts such as those used in microblogs.

A distributed FP-growth method that is dependent on Spark (DFPS) was proposed by Xiujin Shi et al. [7] in the year 2017. DFPS divides up the computing responsibilities in such a manner that every computing node develops its own conditional FP-tree and separately mines the frequent item sets using a pattern growth approach. During the mining of frequent item sets, DFPS does not need to exchange information among any of the nodes. The performance analysis reveals that the DFPS technique is superior than Yet Another Frequent Set Mining (YAFIM), particularly when the duration of transactions is lengthy, the quantity of items is huge, and the data is huge. This is especially true when the data is enormous. Additionally, DFPS provides tremendous scalability capabilities. The findings of the experiments indicate that DFPS is over ten times quicker than YAFIM when applied to the T10I4D100K dataset and the Pumsb star dataset, respectively.

A strategy for concealing critical sequential patterns was proposed by F. Shahzad et al. [8]. The FP Growth technique serves as the foundation for this strategy. In order to conceal potentially harmful sequential patterns, the FP tree is constrained using anti-monotone and monotone constraints.

Electronic health records (EHRs) can be expressed as series of time-stamped occurrences, as stated by Faezeh Movahedi et al. [9]. Therefore, temporal patterns, like transitions among clinical occurrences throughout time, are capable of being retrieved utilising techniques associated with temporal mining. This offers the advantage of translating enormous temporal data records into a knowledge base that is transparent and straightforward, making it easier to comprehend, and having the potential to assist clinical practise. Furthermore, due to the dynamic, varied, and complicated nature of health data, EHR data presents a number of obstacles that must be overcome. In the context of this project, the development of a structured approach to mine patterns of transitions among adverse events following the implantation of a Left Ventricular Assist Device (LVAD) in individuals suffering from severe heart problems is the primary focus.

According to Di Wang et al. [10] more than 270 kinds of ECG-type data are included in the Chinese Cardiovascular Disease Database which includes 12-Lead electrocardiogram (ECG) data and descriptive features with diagnosis. Every record may correspond to a variety of different diagnosis results. Furthermore, the majority of the approaches that are now in use are designed for single-label data. Only a small number of researchers research multi-label ECG data, particularly the association between ECG type tags. The association rule mining technique is considered to be among the top ten most classic data-mining techniques. This approach may be applied to a vast quantity of data in order to find relevant correlations or relevant information among different item sets. Due to its space-time complexity and Input/Output cost, the association rule mining method is challenging to adapt to jobs involving the processing and evaluation of large amounts of data when it is run on a single system. This article makes extensive use of the robust data processing capacity offered by Spark's cluster in order to investigate the recurrent patterns found in large amounts of ECG data.

In 2018, Pan Zhaopeng et al. [11] generated a new form of FP-tree by utilising the approach of dynamic insert node FP-tree design, in addition to the entire back pointer. The Max-IFP optimum frequent patterns mining methodology is also proposed in this paper. It makes use of the newest generation of FP - tree to expose all of the maximal frequent item sets. The outcomes of the experiments reveal that the improved FP-tree takes up less space, and the technique that is suggested in this study is both shorter and much more efficient than traditional algorithms when it comes to mining the sets of items that occur the most frequently.

An enhanced version of the multi minimum support frequent pattern mining (IMISFP) growth was proposed in 2019 by WU Jia et al. [12]. In the first step of the preprocessing of the items in the transaction database that takes place before the construction of the tree, those items whose support is lower than the least item support are deleted, and multiple support trees are constructed making use of the frequently used items that are still in the database. After that, a fresh approach to building multiple item trees depending on intersection rules is presented as an alternative. This approach does not use a particular standard layout item to produce tree anymore; instead, it builds a tree based on the concept of intersection whenever a novel transaction item set is supplied. In the end, the CFP-growth++ algorithm and the IMISFP growth algorithm are contrasted on 5 distinct databases. According to the findings of the experiments, the revised algorithm outperforms the CFP-growth++ method in terms of the amount of time it takes to execute, the amount of memory it uses, and its ability to scale.

The objective of Norulhidayah Isa et al. [13] was to assist Neddy Enterprise Sdn. Bhd. (NE) in the process of decision-making in the year 2021, particularly with regard to the planning of their supplier purchase. Mining of association rules was done in order to establish a connection between the supplies that were required and the tasks that were completed. In order to accomplish this research, the Cross-Industry Standard Process for Data Mining has been utilised as a project framework. As a consequence of this, a certain category of endeavour requires a plethora of different materials. NE will be able to improve its purchase planning as a result of this research.

Katerina Vrotsou et al. [14] proposed a novel method for interactive visual sequence mining. This method gives the user the ability to direct the implementation of a pattern-growth technique at suitable moments by utilising a potent visual interface. This approach enables for the incremental viewing of patterns that are being mined and introduces the potential of making use of local limitations while the mining process is in progress. It gives the user the ability to direct the algorithm being used for mining in interesting directions. The capability of users to gradually narrow the search area without having to restart computations is greatly improved when local constraints are utilised. This improvement can be seen as a significant positive.

Roshani Patel et al. [15] reviewed a variety of sequential pattern mining techniques, including GSP, FreeSpan, PrefixSpan, and CAI-PrefixSpan, in order to increase the performance of identifying sequential patterns. There have been many other constraints suggested in order to obtain a suitable pattern, such as the spacing between two purchased items, the amount of time that elapses between items, etc. The already-existing CAI-PrefixSpan technique was implemented with a time constraint that gave the time from when the consumer purchased the item for the first time. The comparison study demonstrates the effectiveness of the various algorithms with regard to the parameters.

According to Chunkai Zhang et al. [16] High Utility Sequential Pattern Mining is a relatively new topic of research that belongs to the field of data mining. HUSPM is able to provide comprehension that is more essentially relevant in contrast to the two discussions that came before it. This is because it takes into consideration utility, which suggests the value to the company, as well as sequential, which implies the connection between the various items. Due to the fact that it combines utilitarian and sequence elements, HUSPM is more difficult than the issues that came before it. This is because it presents a greater number of obstacles. These two optimization methods for HUSPM, which are known as HUS-UT and HUS-Par, were developed as a result of this research and are named after their respective initials. It makes use of a new dataset that is known as Utility-Table in order to reduce the utility estimation and make it simpler for the suggested HUS-UT technique to discover the required patterns in a timely manner. This was accomplished by using the Utility-Table. The HUS-Par method is a parallelized variation of the HUSUT technique that is dependent on the thread paradigm. The algorithm makes use of the thread model in addition to utilising two different balance strategies in order to get improved performance. In accordance to this, extensive tests were carried out so that the approaches could be evaluated for their level of effectiveness. The outcomes of the experiments suggest that the procedures are much more effective than the strategies that are now regarded as state-of-the-art approaches.

In 2017, according to Md.Mahamud Hasan et al. [17] The ability to mine common item sets is one that is held by a large number of individuals and is put to use in a wide variety of situations that occur in the real world. In addition to the Frequent Pattern Technique, the Apriori algorithm is used most commonly in the process of extracting frequent item sets. This is because the Apriori algorithm was developed first. On the other hand, in order to obtain frequent item sets through using Apriori and FP Growth technique, there is a challenging issue in determining the minimum support that is essential (the threshold). If the minimum support is lowered, then an excessive number of common item sets will be formed. This could cause the Apriori as well as FP Growth technique to become less effective, or it could even cause memory to be lost. When the least support is exceeded, however, a smaller number of itemsets are found rather than when it is set to a lesser value. This is because the minimum support can only support so many items. A technique that tends to make utilization of the Binomial Distribution (BD) to regard appropriate minimum support in an improved way is presented as a solution to this problem in this body of work. This method is intended to address the issue in question. The process of mining efficient frequent item sets has been simplified, which has led to the proposed approach performing markedly better than the standard that was utilised in the past.

III. Research Methodology

The proposed Enhanced Pattern Growth Algorithm Based on MapReduce is implemented for big healthcare dataset. In the first stage of the process, the transaction database is split into subtransaction databases of similar size before being distributed throughout the many nodes that make up the cluster. In the second stage, every node in the cluster is responsible for independently calculating the total number of supporters for its particular item. These numbers are then summed together on the same node. The findings are presented in a F_list of frequent 1-itemsets which is organised by the support count in descending order. In the third stage, the requisite transaction of creating the frequent item's FP-Tree is dispersed to every frequent item correlating node. This ensures that the global frequent itemsets have been incorporated. Afterwards, EPGA is used to obtain local frequent itemsets that comprise this frequent item. In order to acquire the global frequent itemets, it is necessary to aggregate the local frequent itemsets first.

Implementation Process

The implementation of EPGA based on MapReduce environment comprises of following phases.

Phase1: the TDB is partitioned into STDBs of uniform size and then distributed to a number of different nodes. This partitioning and distribution is an automatic process that is carried out by the Hadoop distributed file system.

Phase 2: F-List for distributed computing.

Phase 3: Parallel computing is performed for the regional frequent itemsets. While the Reduce function produces local frequent itemsets by generating and mining condition sub frequent pattern growth tree, the Map function takes the frequent item matching transaction sets and transfers it to the same node to form list.

Phase 4: Aggregation of Local frequent itemsets is performed from every node and thus, the global frequent itemset is obtained.

IV. Result Analysis and Discussion

The proposed system is implemented in the Hadoop cluster using one master and four slave machines. The big IOT health care dataset is used which is further processed into 64 MB, 128 MB, 256 MB and 512 MB datasets. The overall performance of Apriori and EPGA is evaluated using indicators like speedup, reliability using single and Mapreduce environment. The detail discussion of various experiments is described as follows.

Experiment A:

Comparative analysis of execution time of Apriori and EPGA algorithm in the singlemachine and Mapreduce environment

In this experiment, Apriori and EPGA algorithm is used for association rules mining on the group of 4 datasets (64MB, 128MB, 256MB, 512MB) in the single machine and MapReduce environment respectively.

Consider the following table 1 which depicts the execution time Apriori and EPGA in single machine and Mapreduce environment.

Table 1: The Comparative analysis of execution time of Apriori algorithm and EPGA in the single machine and MapReduce environment

Algorithm		Data Size (MB)	Execution Time (sec)
Apriori	Single	64	710
		128	1715
		256	-
		512	-
	Map Reduce	64	851
		128	1412
		256	2044
		512	2560
EPGA	Single	64	510
		128	1555
		256	-
		512	-
	Map Reduce	64	775
		128	1205
		256	1813
		512	2532

The experimental outcomes are shown in Figure 1.

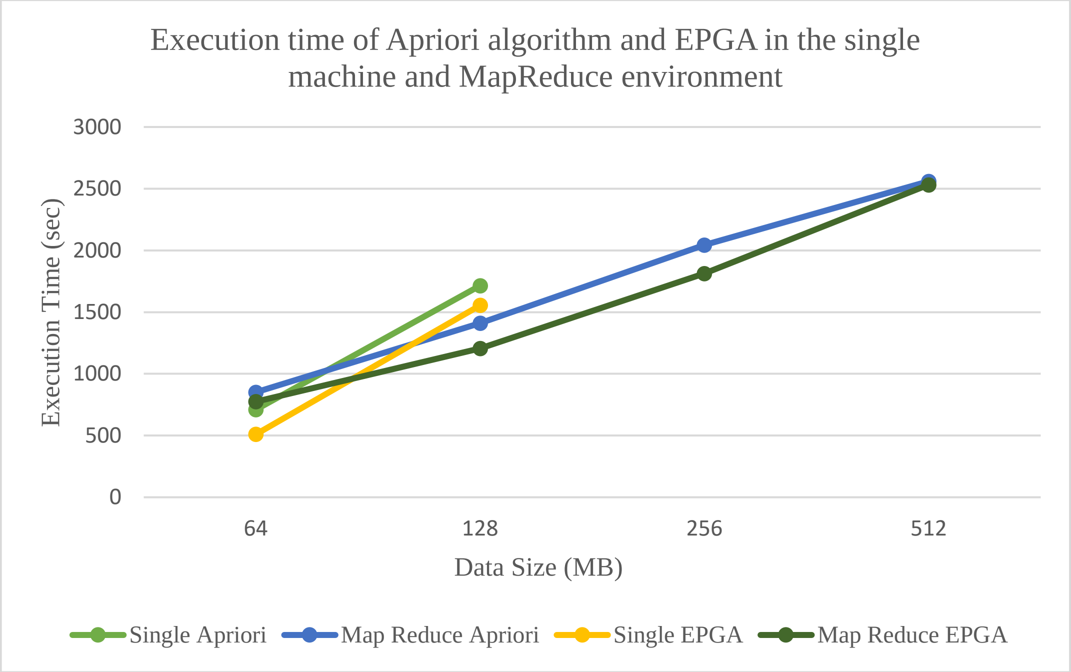

Figure 1. The execution time of Apriori algorithm and EPGA in the single machine and MapReduce environment

The above figure 1 depicts the execution time of Apriori algorithm and EPGA in the single machine environment as well as Mapreduce environment.

Apriori and EPGA, when executed on the same system in the same environment, cause the usage of memory as well as other resources to increase continuously, resulting in a significant decrease in the total performance of the two approaches. This consumption rise is caused by the increasing of the datasets. When working with datasets that are 256 megabytes in size, the "out of memory" notification will arise in the single machine context, and the method will cease operating as a result. On the other hand, Apriori and EPGA in the MapReduce environment are able to successfully execute computing tasks whenever the dataset size is greater than 256 megabytes. The experiment shown above demonstrates that both of these algorithms are workable within the MapReduce environment.

Whenever the datasets are minimal, Apriori and EPGA perform better in an environment with a single machine than they do in a Map Reduce environment. The reason for this is that the response of every node and the communication among nodes produce a certain period of time usage in the cluster environment. Additionally, the time spent on operation and the system's maintenance is considerably longer than the time spent actually during computation tasks.

The efficiency of Apriori and EPGA in a MapReduce context improves progressively as datasets continue to gradually expand, the operating time tends to remain stable, and the advantage of processing efficiency is evident. The investigation presented above demonstrates that both Apriori and EPGA can function correctly within a MapReduce environment.

Experiment B :

Comparative analysis of Speedup of Apriori and EPGA based on Mapreduce environment

In this experiment, testing the Speedup of Apriori and EPGA based on MapReduce environment is performed using group of four datasets (64MB, 128MB, 256MB, 512MB). When the number of Datanodes is steadily raised, the amount of time required to run Apriori and EPGA is evaluated for datasets of varying sizes.

Consider the following table 2 which depicts the speedup of Apriori and EPGA based on MapReduce environment.

Table 2: Comparative analysis of speedup of Apriori algorithm and EPGA based on MapReduce environment

Algorithm	Data Size	Node	Execution Time (sec)
Apriori	64	1	1510
		2	1421
		3	1335
		4	1215
	128	1	1752
		2	1605
		3	1589
		4	1476
	256	1	2809
		2	2445
		3	2260
		4	2191
	512	1	3311
		2	3125
		3	2825
		4	2712
EPGA	64	1	1321
		2	1287
		3	1223
		4	1175
	128	1	1650
		2	1589
		3	1523
		4	1408
	256	1	2741
		2	2587
		3	2278
		4	2189

	512	1	3256
		2	3150
		3	2889
		4	2441

The experimental outcomes are shown in Figure 2.

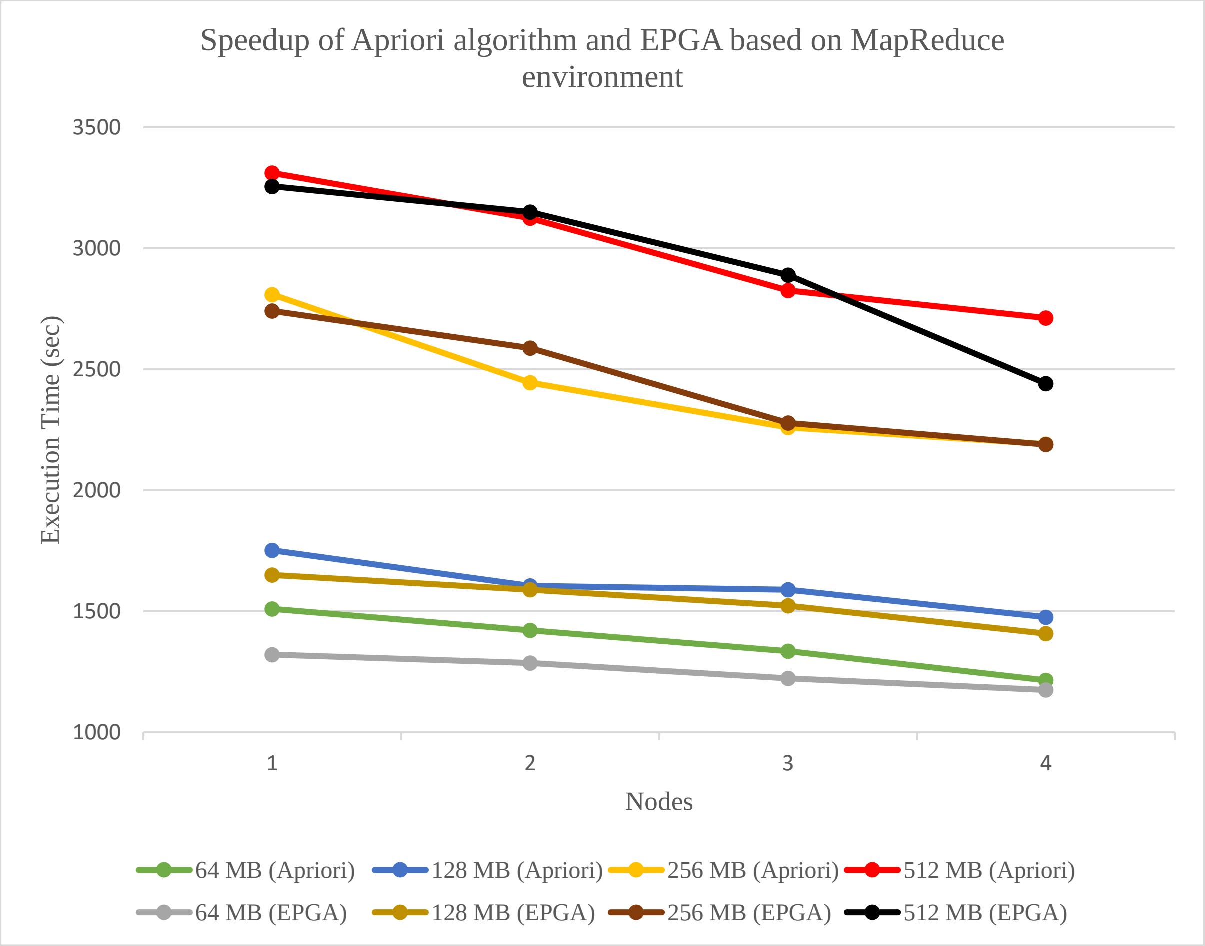

Figure 2. The speedup of Apriori algorithm and EPGA based on MapReduce environment

The above figure 2 depicts the speed up of Apriori algorithm and EPGA in the Mapreduce environment using different node and datasize.

When working in a MapReduce system with datasets of varying sizes, Apriori and EPGA see a reduction in the amount of time it takes to run as the number of Datanodes available inside the cluster increases. The investigation shown above demonstrates that both of the Apriori and EPGA have satisfactory speedup within the MapReduce environment.

Experiment C:

Comparative analysis of Scalability of Apriori and EPGA based on Mapreduce Environment

In this experiment, when 4 data nodes are involved, scalability of Apriori and EPGA is tested using 4 groups of dataset (64 MB, 128MB, 256MB, 512MB) based on Mapreduce environment.

Consider the following table 3 which depicts the scalability of Apriori and EPGA based on Mapreduce environment.

Table 3: Comparative analysis of Apriori and EPGA based on Mapreduce environment

Algorithm	Data Size (MB)	Node	Execution time
Apriori	64	1	1400
	128	2	1812
	256	3	2819
	512	4	2756
EPGA	64	1	1325
	128	2	1658
	256	3	2789
	512	4	2768

The experimental outcomes are shown in Figure 3.

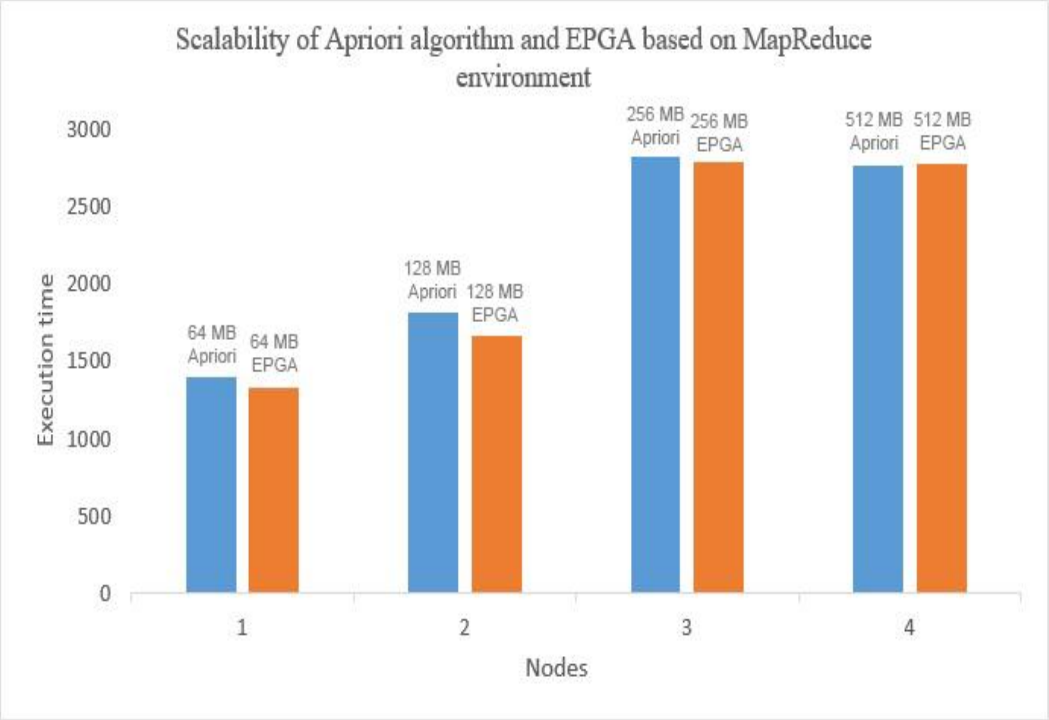

Figure 3. The scalability of Apriori algorithm and EPGA based on MapReduce environment

The above figure 3 depicts the scalabilityof Apriori algorithm and EPGA in the Mapreduce environment using different nodes and data size.

It demonstrates that the Apriori and EPGA have high scalability in the MapReduce environment since the running time of the technique seems to have little change prior to and after the proportionate rise in the number of nodes and the computing datasets.

Experiment D:

Comparative analysis of reliability of Apriori and EPGA Based on MapReduce Environment

The node is shut off by utilising the 256 MB datasets that are included within the cluster of four Datanodes. Both of the MapReduce-based algorithms are tested in an effort to determine whether or not they are capable of regular operation and producing accurate outcomes.

Consider the following figure 4, which depicts the reliability of Apriori and EPGA based on Mapreduce environment.

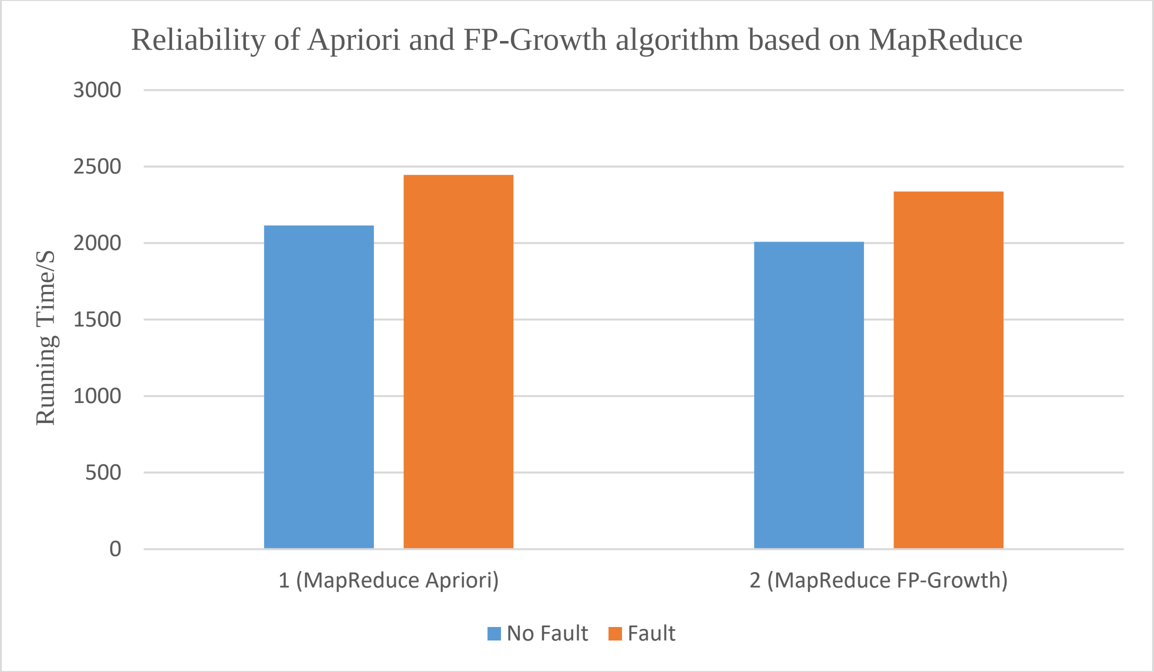

Figure 4: The reliability of Apriori and FP-Growth algorithm based on MapReduce.

The above figure 4 depicts the relability of Apriori algorithm and Enhanced pattern growth algorithm using Mapreduce environment.

In the context of a single machine, if either the software or the hardware of the machine does not work properly, then these two algorithms will not be able to execute normally. This will result in the failure of the entire computing operation, and the dependability will not be able to be guaranteed.

Even though the execution time would increase in the Map Reduce context whenever a Datanode in the cluster is shut down, the Apriori and EPGA can still perform computing operations regularly and produce correct results. The reason for this is that the Hadoop platform has a greater fault tolerance and reliability, and when a failure node happens, the JobTracker will fairly schedule other nodes in the cluster to conduct the Map task or Reduce task of the fault node in the cluster. This is carried out to ensure that computing jobs can be successfully performed. According to the research shown above, these two algorithms offer a high level of dependability when used in a MapReduce setting.

V. Conclusion

The proposed work aims to address the existing problems of the classical algorithm Apriori and FP-Growth when handling large-scale data mining association rules in the single machine environment. An in-depth analysis of the MapReduce environment is presented and Apriori, EPGA is applied to the MapReduce environment to realize association rules mining on large-scale datasets. At the same time, based on MapReduce environment, big healthcare dataset is used with the building of Hadoop platform. The three experimental results were performed using Apriori and EPGA to find scalability, testing performance, speedup and reliability using single and Hadoop environment. It is observed that proposed EPGA improves the overall performance and the efficiency of processing.

References

1. Yuhang Miao, Jinxing Lin and Nuo Xu. “An improved parallel FP-growth algorithm based on Spark and its application”, 2019, Proceedings of the 38th Chinese Control Conference, IEEE.
2. Ze-Zhong Wang and Shuo cao. “A Power Load Association Rules Mining Method Based on Improved FP-Growth Algorithm”, 2018, International Conference on Electricity Distribution, IEEE.
3. Lingxizhu, Yufeiguo and Jingyiwang. “Application of FPGrowth Algorithm of Sequential Pattern Mining on Container Maintenance Components Association”, 2020, 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE.
4. Xiaolei Ma, Yongguang Li, Ran Liu, Yanjun Zhang, Liya Ma and Ziwen Gao. “Frequent Itemsets Mining of SCADA Data Based on FP-Growth Algorithm”, 2020, 4th Conference on Energy Internet and Energy System Integration (EI2), IEEE.
5. Xueyi Gao, Yu Wang, Mengru Sun, Congfen He and Yan Jia. “Assisted analysis of acne metagenomics sequencingdata based on FP-Growth method”, 2019, IEEE.
6. Yalu Jia, Lei Liu and Hao Chen. “An Unknown Words Recognition Method for Microblog Short Text Based on Improved FP-Growth”, 2019, IEEE.
7. Xiujin Shi, Shaozong Chen and Hui Yang. “DFPS: Distributed FP-growth Algorithm Based on Spark”, 2017, IEEE.
8. F. Shahzad and s. Asgha. “Hiding Sequential Patterns Using FP Growth Technique”, 2011, IEEE.
9. Faezeh Movahedi, Yiye Zhang, Rema Padman and James F. Antaki. “Mining temporal patterns from sequential healthcare data”, 2018, International Conference on Healthcare Informatics, IEEE.

10. Di Wang, Jing Ge, Lu Wu and Xiaofeng Song. “Mining Frequent Patterns for ECG Multi-label Data by FP-Growth Algorithm Based on Spark”, 2019, 7th International Conference on Information, Communication and Networks, IEEE.
11. Pan Zhaopeng, Liu Peiyu and Yi Jing. “An Improved FP-tree Algorithm for Mining Maximal Frequent Patterns”, 2018, 10th International Conference on Measuring Technology and Mechatronics Automation, IEEE.
12. Wu Jia, Zhang Lijuan, Cui Wei and Jiang Bohang. “Frequent Pattern Mining Algorithm based on Multi Minimum Support”, 2019, International Conference on Power Data Science (ICPDS), IEEE.
13. Norulhidayah Isa, Siti Khadijah Neddy and Norizan Mohamed. “Association Rule Mining using FP-Growth Algorithm to Prevent Maverick Buying”, 2021, Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE.
14. Katerina Vrotsou and Aida Nordman. “Exploratory Visual Sequence Mining Based on Pattern-Growth”, 2018, IEEE.
15. Roshani Patel andTarunika Chaudhari. “A Review on Sequential Pattern Mining using Pattern Growth Approach”, 2016, IEEE.
16. Chunkai Zhang, Yiwen Zu, Junli Nie and Linzi Du. “Two efficient algorithms for mining high utility sequential patterns”, 2019, IEEE.
17. Md.Mahamud Hasan and Sadia Zaman Mishu. “An Adaptive Method for Mining Frequent Itemsets Based on Apriori And FP Growth Algorithm”, 2017, IEEE.
18. Swati Nagori and Dr. Hemant Kumar Soni. “Issues and Research Challenges in Sequential Pattern Mining”, 2020, International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), IEEE.
19. Bhargav C. Kachhadiya and Prof. Bhavesh Patel. “A Survey on Sequential Pattern Mining Algorithm for Web Log Pattern Data”, 2018, 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE
20. WU Jia, LV Bing and CUI Wei. “An Improved Sequential Pattern mining Algorithm based on Large Dataset”, 2019, International Conference on Power Data Science (ICPDS), IEEE.
21. Talukdar, V., Dhabliya, D., Kumar, B., Talukdar, S. B., Ahamad, S., & Gupta, A. (2022). Suspicious activity detection and classification in IoT environment using machine learning approach. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 531-535. doi:10.1109/PDGC56933.2022.10053312 Retrieved from www.scopus.com
22. Singh, H., Ahamad, S., Naidu, G. T., Arangi, V., Koujalagi, A., & Dhabliya, D. (2022). Application of machine learning in the classification of data over social media platform. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 669-674. doi:10.1109/PDGC56933.2022.10053121 Retrieved from www.scopus.com
23. Sindhwani, N., Anand, R., Vashisth, R., Chauhan, S., Talukdar, V., & Dhabliya, D. (2022). Thingspeak-based environmental monitoring system using IoT. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 675-680. doi:10.1109/PDGC56933.2022.10053167 Retrieved from www.scopus.com

24. Rami Ibrahim and M. Omair Shafiq. “Towards a New Approach to Empower Periodic Pattern Mining for Massive Data using Map-Reduce”, 2018, IEEE.
25. Pallavi V. Nikam and Dr. Deepa S. Deshpande. “New approach in Big Data Mining for frequent itemset using mapreduce in HDFS”, 2018, 3rd International Conference for Convergence in Technology (I2CT), IEEE.
26. Mercy Nyasha Mlambo, Naison Gasela and Michael Bukohwo Esiefarienrhe. “Implementation and Analysis of Enhanced Apriori Using MapReduce”, 2018, IEEE.
27. S.Haseena, S.Manoruthra, P.Hemalatha and V.Akshaya. “Mining Frequent Item sets on Large Scale Temporal Data”, 2018, 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE.
28. Bao Lei. “Apriori-based Spatial Pattern Mining Algorithm for Big Data”, 2020, International Conference on Urban Engineering and Management Science (ICUEMS), IEEE.
29. Muhammad J. Alibasa, Rafael A. Calvo and Kalina Yacef. “Sequential Pattern Mining Suggests Wellbeing Supportive Behaviors”, 2019, IEEE.
30. Ji-Soo Kang, Ji-Won Baek and Kyungyong Chung. “PrefixSpan Based Pattern Mining Using Time Sliding Weight From Streaming Data”, 2020, IEEE.
31. Jerry Chun-Wei Lin, Gautam Srivastava, Yuanfa Li, Tzung-Pei Hong and Shyue-Liang Wang. “Mining High-Utility Sequential Patterns in Uncertain Databases”, 2020, International Conference on Big Data (Big Data), IEEE.
32. Saıd Jabbour, Jerry Lonlac and Lakhdar Saıs. “Mining Gradual Itemsets Using Sequential Pattern Mining”, 2019, IEEE.
33. Chunkai Zhang and Yiwen Zu. “An efficient parallel High Utility Sequential Pattern Mining algorithm”, 2019, 21st International Conference on High Performance Computing and Communications; 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems, IEEE.