Enriching Frequent itemset mining for retail shopping using M-tree

Shubham Devrao Mohurle
Computer Engineering
K J C O E M R, Pune, India
shubhammohurle25@gmail.com

Prof. N. Bogiri
Computer Engineering
K J C O E M R, Pune, India
mail2nagaraju@gmail.com

***Abstract*— In today's business world, be it wholesale, retail or E-commerce earning profit is the most important thing to exist in the trending market. The secret of this lies in always fulfilling the user's requirements based on the fact of the trends or on the past sale data. This is usually done by market analysis and current trends and matching it with the customer requirements. As these trends keep changing due to various reasons such as product launches, seasons etc. and makes it almost impossible to predict the trend. So the next best option remained is to analyze the past sale records and try to estimate the most frequent items sold in the given context of time. By predicting this any business owner will get an idea of the customer's requirement for the past same instance that makes him to grow his business with the predicted frequent items. The most intricate thing lies in this process is to manage the space complexity as most of the traditional frequent itemset algorithms like Apriori and all are suffering from this due to exponential growth in the candidate sets. Many research shows tree based mining of the frequent itemsets is the best choice, based on this thread proposed system uses the M-tree technique to mine the frequent itemsets. The frequent itemsets are analyzed based on the candidate sets formed after analyzing the entropy using the Shannon information gain theory. The conducted experiments reveal that the proposed system yields the good analysis time over some past traditional methods.**

Keywords **-** M-tree, Linear Clusters., Shannon Information gain theory, Entropy analysis, Frequent itemsets. Candidate sets.

I. INTRODUCTION

Frequent Itemset Mining is one of the most important techniques for the characterization of data. This procedure for the characterization of the data called the Frequent Itemset Mining is also misinterpreted as Association Rule Mining, which is a very complex utilization of the Frequent Itemset Mining. It is one of the most important tasks in the Data Mining domain. The main objective of this procedure is the extraction of various interesting patterns from the database.

The main problem that is being solved with this Frequent Itemset Mining technique is the identification and mining of Association Rules. As the Frequent Itemset Mining is one of the techniques that is utilized for fulfilling Data Mining responsibilities, that include identification of characteristics, symptoms, products and items. The need for implementation of such an algorithm and the purpose of Data Mining arose due to the need for analysis of supermarket data.

The supermarket transaction data is a highly useful form of data collection that can ensure various parameters that are responsible for the behavior of the customer. The various purchases made by the customer for different products and their frequency form a pattern that is not visible by the naked human eye but can be analyzed with the help of Frequent Itemset Mining. Therefore, it is imperative to analyze the customer behavior, as it can help provide valuable insight into the inner workings of the store and it can also be used to improve the store considerably.

For the Analysis of the supermarket or any other data mining application, the frequent itemset mining utilizes sequence analysis, which is utilized extensively in various different applications. It can lead to the discovery of various behavioral patterns in an individual that can be attributed to the decision that the person makes. Therefore, It can provide some useful insight into the behaviors of certain people and also the public as a whole.

The Market Basket analysis aims to extract the same result that can be used to categories and bundle various frequently bought together items into a combo package. This is really helpful for the owner as it provides a greater amount of comfort and convenience to the customer who would’ve bought the same items separately, this builds customer rapport and also influences the sales in a positive manner. This technique of market basket analysis is fundamentally different from the techniques utilized by the other algorithms which are quite similar such as the similarity search.

This is a very different branch of Data Mining and this is the reason there are specialized algorithms such as the Apriori algorithm that is utilized for the purpose of Frequent Itemset Mining. The Apriori Algorithm executes by removing the bigger datasets and concentrating on the smaller ones. This is due to the fact that a bigger set would not be frequent unless its smaller sets are frequent. The Apriori algorithm is one of

the most computationally extensive algorithms that require a lot of memory to execute.

Entropy is a very important concept in the field of Physics and it is equally important in the field of computer science. In physics, the Entropy is the measure of disorder or randomness. This is a very complicated section of physics that deals with the randomness of matter. Therefore, if the Disorder is more there is more entropy, if the disorder or randomness is less there is less entropy. Entropy is an essential concept in physics and it used to describe the formation of various celestial objects and it has also been one of the driving forces for the conservation of energy principle.

This differs from the interpretation in the computer science field, it details the purity of a selection of information or data. The purity is based on the representation of a particular attribute in the whole dataset. Therefore, the entropy which is an essential part of information theory is nothing but an indication of the purity of the selection of the data representing a particular attribute.

As expected, there is a need for a larger degree of purity in the data for it to be useful, therefore the Entropy is inversely related to the purity of the data. Hence, the reduction of Entropy would result in the gain in the purity of the data. The process of reducing the Entropy of the data to increase the purity of the selection is called the Information Gain.

Shannon Entropy is basically just the Entropy as discussed before, but the Shannon Entropy is a concept that is quite unique and innovative. The Shannon Entropy utilizes the concept that the Entropy of a dataset would subsequently decrease as it is continually split into smaller datasets, this action would drastically reduce the entropy or the messiness of the data as it is in a much more controllable size. Therefore, this should be the main objective of any classification technique, it should be able to efficiently be able to reduce the entropy of the whole system as a whole.

M tree or Multi way trees are trees that are constructed for the organization of various datasets which are usually large in size. The organization of the large datasets is possible due to a feature in the multi way tree that utilizes a metric space for this purpose. The metric space is defined by a distance function that is utilized to satisfy various properties, such as Triangle inequality postulates, symmetry and positivity at the same time.

The Multi way tree is an example of a balanced tree that can be utilized to index dynamic files. The multi way tree does not need frequent reorganization due to its balanced nature. The M- tree is designed with due influence from the database access methods as well as the properties that define metric trees. Therefore, the M-tree is inherently balanced and can be optimised for performance by consideration of both the Input-Output costs as well as the Distance calculation that is performed by the CPU.

Multi way-Tree is one of the most useful technique for the indexing of multimedia objects. As the M-tree makes the management and searching of the various elements of the tree to be highly accessible and also really useful for reducing the entropy of the dataset. The M-tree is also highly susceptible to overlaps which cannot be avoided. It is one of the most essential concepts in the area of Information Retrieval and has been in extensive use for this purpose for a long time.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

A.Maske [1] states that the Market Basket analysis is an essential concept for the purpose of data mining. This is due to the fact that it can be used for the analysis of the shopping behaviors of the customers. The authors surveyed the procedure and have concluded that the process analyses the data to find a relationship between the different items being placed in various customers “baskets”. This is highly useful to help intelligently organize the stores and increase the efficiency of the business. Usually, an Apriori algorithm is utilized to extract the Frequent Itemset, which is utilized to find the associations. The survey had some limitations as it was only conducted for very few test cases.

J. Zheng explores the field of Frequent Itemset Mining due to the fact it is one of the most important aspects of the Data Mining paradigm. The Frequent itemset Mining is highly useful for a business as it isolates the various relationships between the items in the dataset. The authors have analyzed the Apriori technique for the extraction of Frequent Itemsets [2]. The traditional techniques are highly inefficient as they have a very high requirement of memory and computational power. Therefore, the authors propose an innovative technique for extracting the Frequent Itemsets with the help of DABIE (Distributed Apriori Based on Itemset Encoding). The proposed technique has only been exclusively developed to improve performance for a multi-iterative Frequent Itemset Mining.

X. Han expresses that Frequent Itemset Mining is an important aspect of the data mining field. The authors state that the algorithms that exist now are not capable of handling a lot of data efficiently. The algorithms usually run into problems with insufficient memory and require multiple-pass to extract all the Frequent Itemsets. To ameliorate these effects the authors have proposed a pre computation-based Frequent Itemset Mining algorithm [3]. The presented algorithm is capable of computing massive datasets in a short duration

efficiently. The technique has not been tested extensively to judge the performance benefits accurately.

B. He introduces an innovative concept of the purpose of data mining performed on Big Data. The researchers utilize the MapReduce functions and the FP-tree to improve the efficiency of the mining procedure. The authors propose a technique involving three steps, the first step utilizes MapReduce to distribute the data, the distributed data is utilized by the Frequent Itemset Mining to generate the FP-tree, the results are then combined in the form of a central node. The proposed technique is highly computationally extensive as observed in the paper. [4]

D. Kaur elaborates on the initiatives for the rapid growth in the field of Data Mining as it has gained a lot of traction in the supermarket businesses. The Data Mining technique is very essential for this type of businesses it allows the business to identify the trends and also make effective decisions based on these trends [5]. The Data Mining paradigm allows the owners of the supermarket to gain some valuable insight into the data which can actually help increase the efficiency of the operations. Data Mining can help achieve optimum growth and addition of Machine Learning or Artificial Intelligence will improve the business even further.

S. Solanki [6] states that there has been exponential growth in the Data Mining sector recently, specifically in the area of Frequent Itemset Mining as it is the central component of the Data Mining area. The authors in this paper have extensively elaborated the various benefits and the downsides of performing Frequent Pattern Mining that is being utilized by various organizations currently. The researchers have elaborated on various algorithms that are used for this purpose, such as, Apriori, Eclat and FP Growth, by stating their downsides and their benefits.

S. Jalan expresses that extraction of the Frequent Patterns is one of the most important aspects of the Association Rule Mining. The authors state that there are two algorithms that are highly efficient for the extraction of Frequent Patterns, the first one is the P-tree and the second one is the FP-tree. P-Tree and FP-Tree are algorithms that extract the patterns into trees. The researchers in this paper propose an elegant technique that utilizes both the top-down and bottom-up approaches for generating the FP+ tree [7]. The technique has been tested extensively for performance and efficiency. The non-recursive nature of the algorithm reduces the computational complexity of the proposed system.

S. Tribhuvan introduces the concept of Data Mining and how it is one of the most essential components for owners who want to make practical and efficient decisions for their organizations. As the Data Mining techniques allow for the prediction of various trends and future tendencies, this is highly valuable for any business in achieving very high efficiency. The extraction of Frequent Itemsets is very important for the Data Mining process. In this paper, the authors implement an Improved Apriori Algorithm amalgamated with a MapReduce Framework consecutively to decrease the time taken for the execution [8]. The only drawback this technique faces is the increased memory utilization due to the Apriori algorithm.

M. Hasan explores the concept of Data Mining with respect to the Mining of Frequent Itemsets through the assistance of the Apriori and the FP growth algorithms. The FP-Growth and Apriori are the most frequently used algorithms but they are not without their shortcomings. The traditional algorithms are heavily dependent on the minimum threshold values, therefore, if the values are set to low, it would generate a lot of itemsets and vice-versa. This effect is highly undesirable and needs to be eliminated, hence, the authors present an innovative technique that utilizes Binomial Distribution for the dynamic determination of the minimum threshold. The main drawback for the proposed technique is the lack of support for large databases. [9]

P. Zhaopeng proposes an innovative procedure for the extraction of the Frequent itemsets with the help of an FP Growth algorithm. The conventional FP Growth algorithm that is generally utilized for this purpose generates a lot of items, which eventually slows the system down and drastically reduces the efficiency of the whole system. The authors in this paper present an innovative modification to the FP Growth algorithm that reduces the number of candidate items thereby increasing the efficiency and speed of the system [10]. The proposed Max-IFO Algorithm provides only marginal benefits in terms of performance and accuracy.

A.Gassama states that the importance of Frequent Itemset Mining in the realm of Data Mining cannot be overstated. The Frequent Itemset Mining forms the crux of the Data Mining procedure as it enables an effective insight into the data with the help of various applications, such as, market basket analysis and outlier detection etc. [11] The authors have also noticed that the conventional algorithms suffer when faced with mining a large dataset, therefore, an innovative Scalable, parallel FP-growth algorithm is presented on the Apache Spark Platform. The drawback in this paper is that the proposed FP Growth algorithm still consumes a considerable amount of memory when executed on a large enough database.

L. Juan [12] proposes a novel QFP algorithm for the purpose of performing Association Rule Mining that improves over the commonly used FP-Growth algorithm that is utilized for the Data Mining Tasks. One of the most helpful features of the algorithm is that the QFP algorithm only scans the database once to execute the mining. This reduces the computational complexity drastically. The proposed technique has been extensively tested for performance and efficiency and has produced exceptionally good results.

Y. Yang elaborates on the Data Mining paradigm for the purpose of analyzing customer choices and shopping behavior. Usually, there are a set of items that the majority of the customers tend to buy together almost regularly. This information is highly useful and can be utilized for combining the frequently bought items together to sell as a combo. To achieve this effect, the authors utilized the FP-tree algorithm, which mines the frequently bought items which are then bundled together as a combination to achieve better customer satisfaction. Due to the utilization of the FP-growth, the memory consumption of the system is very high. [13]

H. Wang explains that Data Mining has seen a lot of growth in recent years, as frequent itemset mining is utilized regularly for the process of Association Mining. Most of the times, when there are a large number of candidate items that are mined from the database, it gets increasingly difficult and complex to maintain. Therefore, the researchers have presented an innovative concept of utilizing maximal frequent patterns due to their compressed and smaller size can be easier to understand and manage [14]. The proposed algorithm reduces the computational complexity but takes a longer time to process, which is one of the drawbacks of the technique.

Y. Feng explores a novel concept for the purpose of performing data mining tasks with the help of MapReduce and FP-tree weight computation. Data Mining is one of the most useful concepts that can help optimize a business efficiently and help with the growth of the business too. The authors have noticed that most of the conventional algorithms scan the database multiple times to execute mining, which is highly unnecessary [15]. To stop this unnecessary practice, the authors have presented a technique that utilizes MapReduce running on multiple core clusters in conjunction with an optimized FP-growth algorithm. The presented technique suffers from a high execution time which can be reduced further.

III Proposed Methodology

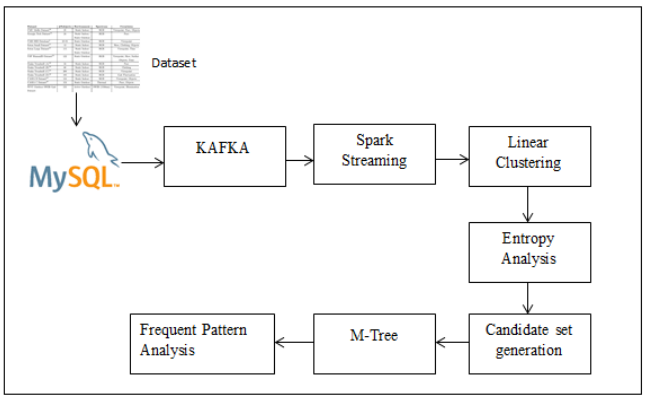

Figure 1: Proposed System Overview

The Proposed model for Frequent pattern analysis, for the retail shopping data is deployed according to the model depicted in figure 1. The steps that are involved in the process of frequent itemset mining are elaborated in the below mentioned steps along with the most important concepts that are used in the deployment of the model.

Key Concepts:

Apache Spark -Apache Spark is a platform for handling Big Data. It is utilized predominantly across most of the real-time applications of Big Data processing. It is composed of a SQL Libraries and data processing engines for performing various tasks such as SQL batch processing and Extract Load Transform activities. Spark is popular because it has inbuilt libraries for most of the operation that can be performed on Big Data.

Apache Kafka -Apache Kafka was originally developed by LinkedIn for their website to make and evaluate connections between their users. It was then donated to where it is now. Kafka is predominantly used for processing stream data. It was designed with the foresight of not being computationally taxing on the computer, therefore it runs on a distributed system. So the framework runs on several machines simultaneously in a distributed architecture.

Mtree- This is a file management utility that is used for maintaining servers of filesystems. It is a very powerful utility that utilizes a config file to check a hierarchical directory for errors or nonconformance to the config file. Mtree can make changes and provide fixes outlined in the config file. It is a lot more convenient for large datasets to correct any discrepancies in the data.

Step 1:Dataset collection - This is the primitive step of the proposed model where a retail dataset is being collected from the publicly available repository kaggle. The URL is https://www.kaggle.com/vijayuv/onlineretail, The Dataset is in the CSV extension which is then stored in the Workbook format. The stored dataset in the workbook format contains some attributes like Invoice no, Stock Code, Description, Quantity, Invoice date, Unit Price, Customer ID ,Country.

Step 2: Streaming - This stored data in workbook is streamed into the Mysql Database. Step 1: This is the initial step of the proposed model where the selected dataset is fed into the system, and thereby it is inserted into the MySQL database. Once the dataset is set into the database, then by using KAFKA the data can inject into the Spak. Kafka is one of the data injection tools. it is used for Stream processing of data.

Step 3: Linear Clustering - The proposed model is developed for a scenario of maintenance of a retail shopping

mall. Where All the above mentioned attributes are needed to perform the transaction. In the proposed model an option is being given to the owner to evaluate the frequent itemset for the stated period of a day or for a month.

As the period is set, then all the sold items for that period is being collected in a list. This list is subjected to hashing process to collect the unique items. Then for each of these unique items respective rows are being collected to create the cluster. This Clustering can be depicted in the following algorithm 1.

Algorithm 1: Linear Clustering

// Input: Unique Itemset U_{ITM}
// Data Set D
// Output: L_C (Linear Clusters)
Function: linearCluster(U_{ITM},D)
Step 0: Start
Step 1: $L_C = Ø$
Step 2: ***for*** i=0 Size of U_{ITM}
Step 3: ITEM=$U_{ITM\,[i]}$
Step 4: $S_C = Ø$ [Single Cluster]
Step 5: ***for*** j=0 Size of D
Step 6: ROW=$D_{[j]}$
Step 7: **IF** ITEM ∈ ROW, **THEN**
Step 8: $S_C = S_C + ROW$
Step 9: **End *for***
Step 10:$L_C = L_C + S_C$
Step 11: **End *for***
Step 12: return L_C
Step 13: Stop

Step 4: Entropy Analysis - Each of the unique item is counted for its presence in the number of the clusters. Then this count is used to estimate the distribution factor of the item using the Shannon information gain theory. This Shannon information gain theory yields a numerical value in between the 0 and 1. The value nearer to 1 indicates the item is more important and nearer to 0 indicates item is having less importance. The estimation of the Shannon information gain is done based on the below mentioned equation 1.

After evaluation of the information gain most important items are being selected based on the highest values of the gain.

IG(E) = - (P / T) log (P / T) - (N / T) log (N / T)
Where
P= Item frequency
N= T-P
T= Number of clusters
IG(E) = Information Gain for the given item

Step 5: Candidate set generation - Here candidate sets of the selected items are being generated based on the power set creation technique, which is depicted in the below algorithm 2.

Algorithm 2: Candidate set Generation

// Input : W_L Item List
// Output : Pattern List P_L
Function : patternList (**W_L**)
Step 0: Start
Step 1: ***for*** i=0 **to** length of W_L-1
Step 2: $W=W_L[i]$
Step 3: ***for*** j=0 **to** length of W_L
Step 4: $W_S=W_L[j]$
Step 5: $C_{SET}=W+W_S$ [Candidate Set List]
Step 6: add C_{SET} to P_L
Step 7: $C_{SET}=\varnothing$
Step 8: repeat (1 TO 5)
Step 9: **End *for***
Step 10: **End *for***
Step 11: return P_L
Step 12: Stop

Step 5: M tree and frequent itemset analysis - As the frequent itemsets are being generated they are subjected to form the M- tree. Here in this process, the initial candidate set is considered as the root frequent itemset. Then the all next candidate sets are assigned their position based on their respective support values.

This process continuously run in recursive manner to generate a well-mannered sorted tree called as M- tree and is depicted in Algorithm 3.

Algorithm 3: M- tree

//input: Candidate Set List C_L
Step 0: Start
Step 1: Create an empty tree as **T**
Step 2: Create the Root Node for first frequent itemset **R_n**
Step 3: FORi=0 to size of **C_L**
Step 4: Compare the distance with the root node **R_n**
Step 5: If (**C_{Li}**support<**R_n**)
Step 6: Add node as left child in **T**
Step 7: Else
Step 8: Add node as Right child in **T**
Step 9: End **FOR**
Step 10: return **T**
Step 11: Stop

This is the last phase of our model where all the candidate sets are traversed in pre Order manner. Where nodes are traversing in ROOT, LEFT CHILD and then Finally RIGHT CHILD manner to collect the similar support candidate sets to form the clusters of the frequent itemsets based on the similar support.

IV RESULT AND DISCUSSIONS

The proposed methodology for the analysis of frequent itemsets has been deployed in a Windows Operating environment running on a machine which is powered by a Core i5 processor assisted by a physical memory of 8 GB used as RAM. The methodology has been coded in the Scala Programming language and utilizes the NetBeans Integrated Development Environment to achieve the deployment efficiently.

The presented technique has been tested for its performance and efficiency on the Online Retail Dataset which has been utilized for performing various experiments depicting the various features of the methodology. The presented technique in this research paper utilizes the M-tree and Shannon Information to achieve frequent itemset mining.

The Shannon information gain is beneficial as it reduces the computational complexity of the process by a large margin when compared with techniques that utilize other algorithms such as Apriori, Eclat, etc. The M-tree technique has been utilized to analyze the frequent itemsets mined through the M-tree technique. As the M-tree traverses in depth it has a greater accuracy and efficiency than that of the conventional algorithm that is in use commonly.

The proposed methodology has been compared to the one presented in [16]. The technique outlined in [16] utilizes an Improved Apriori Algorithm combined with the MapReduce technique to achieve the mining of frequent itemsets. The experiments were performed to measure the time required for the process to mine the frequent itemsets. Both the techniques have been demonstrated on the same dataset and their response time is observed and tabulated in table 1 below.

Size of Dataset (Number of	Process Timing (sec)	
	M-tree Shannon	Improved Apriori
10000	800	1200
20000	950	1600
30000	1000	2100
40000	1350	3000
60000	2800	6500

Table 1: Time Required for Improved Apriori and M-tree Shannon

Therefore, it can be deduced from the table 1 that the methodology proposed in this paper is highly superior as it performs very effectively and records significantly lower time for the generation of the frequent itemsets.

This is due to the fact that the M tree is far better than the Apriori algorithm which works on the horizontal pattern. This is in contrast with the M tree which prefers to traverse in depth. This gain in performance over [16] can be observed in figure 2 below.

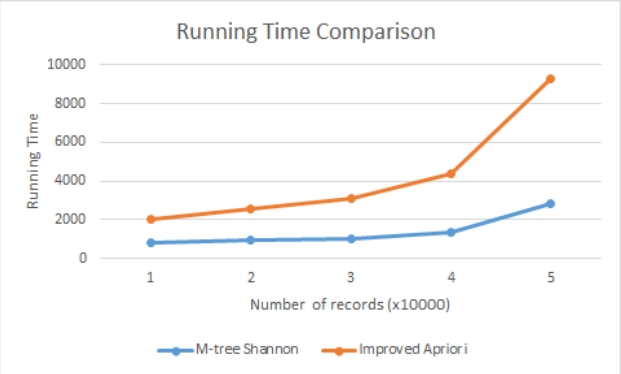

Figure 2: Runtime for different size of records comparison.

The presented technique was further tested and experimented and then compared to another methodology outlined for the extraction of Frequent itemsets given in [17]. The technique elaborated in [17] has utilized the Hadoop Distributed File System (HDFS) for the purpose of mining the frequent itemsets based on the MapReduce model. This technique performs better than the methodology based on Eclat or Apriori Algorithm, due to the resilience of the MapReduce model. This is evident from the experimental comparison of the two techniques given in table 2 below.

Size of Dataset (Number of Records)	Process Timing (sec)	
	M-tree Shannon	Modified Apriori
1000	89	103
2000	98	124
3000	154	176
5000	187	213
10000	247	270

Table 2: Time Required for Modified Apriori (from [2]) and M-tree Shannon

But the technique detailed in [17] cannot outperform the methodology proposed in this paper due to the bottlenecking nature of the HDFS framework where copious amounts of storage are required. This is not the problem with our technique which is based on the M tree which does not encounter memory issues due to very low space complexity of the proposed technique. This is depicted clearly in figure 3 given below which reinforces the superiority of the proposed

M-tree Shannon Information Gain methodology proposed in this paper.

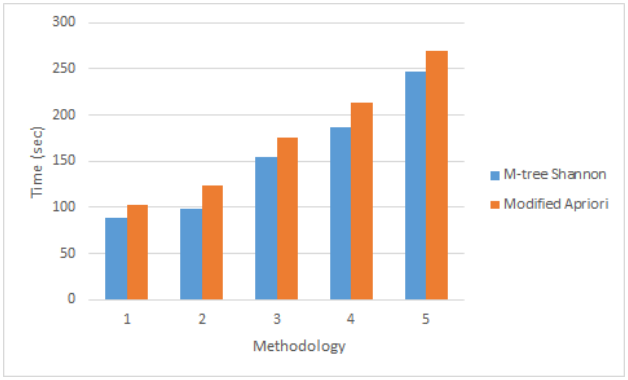

Figure 3: Comparative Graph of the Two Methodologies

V CONCLUSION AND FUTURESCOPE

Frequent Itemset Mining is one of the most essential aspects that can enable greater insight into large amounts of data which plays a vital role in the market analysis and product requirements. For the purpose of extracting the frequent itemsets, the most popular and widely used algorithms such as Eclat, Apriori and FP tree are used. All of these approaches have a lot of shortcomings such as the Eclat algorithm works only on moderate amounts of data in a vertical manner, the Apriori algorithm is very slow and only works in a horizontal pattern and the FP-tree algorithm has a very high space complexity.

Therefore, to overcome these shortcomings and perform efficient extraction of the Frequent itemsets, this research paper utilizes Shannon Information Gain for this purpose. Shannon Information reduces the computational complexity of the process drastically by eliminating the low usage itemsets. This is combined with the M-tree technique which analyses the frequent itemsets extracted by the Shannon process. The addition of M-tree allows the system to traverse depth-wise which increases the accuracy of the system.

The proposed methodology has been tested extensively and compared with prominent researches done in the field of frequent itemset mining. The experimental results asserted the superiority of the proposed technique as it took significantly reduced the time for the extraction of the frequent itemsets. Our technique has been demonstrated to produce frequent itemsets with high accuracy and significantly faster execution times.

For the purpose of future research, this system can be implemented in a cloud environment to reduce the fragmentation and the space requirements. Another direction for future research would include mining of frequent itemsets from a data stream or graph databases which can enable the dissemination of valuable information.

REFERENCES

[1] A. Maske and B. Joglekar, “Survey on Frequent Item-Set Mining Approaches inMarket Basket Analysis”, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

[2] Jingyi Zheng, Xiaoheng Deng andHonggang Zhang, “A Novel Method to Generate Frequent Itemsets inDistributed Environment”, IEEE 37th International Performance Computing and Communications Conference (IPCCC), 2018.

[3] Xixian Han, Xianmin Liu, Jian Chen, Guojun Lai, Hong Gao, and Jianzhong Li, “Efficiently mining frequent itemsets on massive data”, IEEE Access (Volume: 7), 2019.

[4] B. He, H. Zhang and J. Pei, “The Mining Algorithm of Frequent Itemset based on MapReduce and FP-tree”, International Conference on Computer Network, Electronic and Automation, 2017.

[5] D. Kaur and J. Kaur, “Data Mining in Supermarket: A Survey”, International Journal of Computational Intelligence Research, 2017.

[6] S. Solanki and N. Soni, “A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth”, International Journal of Computer Techniques, 2014.

[7] S. Jalan, A. Shrivastava, and G. Sharma, “A non-recursive approach for FP-tree based Frequent Pattern Generation”, Proceedings of Student conference on research and development, 2009.

[8] S. Tribhuvan, N. Gavai and B. Vasgi, “Frequent Itemset Mining using Improved Apriori Algorithm with MapReduce”, International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017.

[9] M. Hasan and S. Mishu, “An Adaptive Method for Mining Frequent Itemsets Based on Apriori And FP Growth Algorithm”, International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.

[10] P. Zhaopeng, L. Peiyu, and Y. Jing, “An Improved FP-tree Algorithm for Mining Maximal Frequent Patterns”, 10th International Conference on Measuring Technology and Mechatronics Automation, 2018.

[11] A. Gassama, F. Camara2, S. Ndiaye, “S-FPG: A Parallel Version of FP-Growth Algorithm under Apache Spark”, 2nd IEEE International Conference on Cloud Computing and Big Data Analysis 2017.

[12] L. Juan and M. De-ting, "Research of An Association Rule Mining Algorithm Based on FP tree", IEEE International Conference on Intelligent Computing and Intelligent Systems, 2010.

[13] Y. Yang and H. Peng, "Modular Order Picking Approach based on FP-Tree Algorithm", Seventh International Symposium on Computational Intelligence and Design, 2014.

[14] H. Wang and C. Hu, "Mining Maximal Patterns Based on Improved FP-tree and Array Technique", Third International Symposium on Intelligent Information Technology and Security Informatics, 2010.

[15] Y. Feng, M. Cho, K. Lu, Z. Ming, H. Zong, W. Cai, and Z. Li, "Optimize the FP-tree based Graph Edge Weight Computation on Multi-core MapReduce Clusters", IEEE 23rd International Conference on Parallel and Distributed Systems, 2017.

[16] S. Tribhuvan, N. Goyal and B. Vasgi, "Frequent Itemset Mining using Improved Apriori Algorithm with MapReduce", International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017.

[17] P. Kulkarni and S. Khonde, "HDFS Framework Efficient Frequent Itemset Mining Using MapReduce", 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017.

*****

Author 1: Mr. Shubham Mohurle has completed hi Bachelor of Engineering degree from K. J. college of Engineering and Management Research, Pune.
Email Id: shubhammohurle25@gmail.com

Author 2: Mr. Nagaraju Bogiri has completed his Masters of Engineering and is working as a Professor at K. J. college of Engineering and Management Research, Pune.
Email Id: mail2nagaraju@gmail.com