Selection of a Promising Frequent Pattern Mining Technique Using Comparative Analysis

Dr. Arpit Solanki[1], Dr. Rajeev G. Vishwakarma[2]

Dr. A.P.J. Abdul Kalam University, Indore
Corresponding Author Email : arpitsolanki@aku.ac.in

Abstract: Frequent pattern mining is a powerful method for identifying useful patterns in databases. Numerous algorithms have been proposed for frequent pattern mining. Because of its many applications in real-world problems, the selection of suitable strategies for mining frequent patterns is therefore a particularly promising and significant research subject. Based on a comparative analysis of findings reported in previous studies, this research identifies a promising frequent pattern mining technique for future improvement. The algorithms Apriori, FP-Tree, and Eclat are taken into consideration for performance study in this paper. The comparative analysis indicates that the Eclat algorithm provides superior execution performance, whereas the Apriori algorithm generates a greater number of frequent rules. Therefore, the selection of an appropriate algorithm depends on the application requirements. Furthermore, the study highlights the potential for improving the Apriori algorithm to enhance its overall efficiency.

Keywords: Frequent pattern mining, Association rule mining, Data mining, Rule mining techniques.

I. INTRODUCTION

In recent years, technology has advanced at a quick pace, and growth in computer technology has altered dramatically. Due to its massive size, data mining is one of the most promising study topics for extracting essential information from large amounts of data. As a result, this critical information is required for many businesses and can aid in decision making.

In data mining, Frequent pattern mining is the process of identifying patterns within a dataset that occur frequently. This is done by analyzing datasets to find items that appear together. This technique is useful for discovering that information which is not visible in data. Additionally, it can find association and correlation among data items or attributes. However, there are several limitations. such as it can generate a large number of patterns with high dimensionality. Additionally, the number of patterns can be very large, making it difficult to interpret the results. Therefore, selection of an appropriate technique for frequent pattern mining is one of the most difficult tasks in data mining.

The fundamentals of frequent pattern mining are covered in this section. The relevant study that was recently included to enhance the frequent pattern analysis method is included in the next part.

II. LITERATURE SURVEY

In this section, the background study related to the frequent pattern mining and association rule mining is discussed.

In order to create a more effective algorithm, C. H. Chee et al. [1] compare several methods for frequent pattern mining. Finding the hidden patterns in the frequently occurring itemsets takes longer when there is more data. Additionally,

RAM is needed since mining the hidden patterns involves a lot of work. As the amount of data increases, a competent computation should be able to mine regular itemsets in a shorter amount of time and with less memory.

F. Min et al. [2] offer a frequent pattern discovery technique for a new sort of pattern by categorising the alphabet as strong, medium, and weak. It's called a tri-pattern. It is more generic and versatile, which makes it more intriguing. Experiments were conducted on data from other disciplines to demonstrate the universality of the pattern. These include protein sequences, time series for petroleum output, and fabricated Chinese literature. The results suggest that tri-patterns are more significant than other sorts of patterns. It enriches the semantics of SPM and its applications.

SPM does not take into account sequential repetition. To tackle this problem, Y developed gap constraint SPM. Wu et al. [3] avoids discovering too many worthless patterns. Non-overlapping SPM indicates that no two occurrences can utilise the same sequence letter in the same location. Non-overlapping SPM can achieve a balance of efficiency and completeness. The present FPM approaches have duplicate patterns. To eliminate them and enhance performance, the author uses closed pattern mining and presents a comprehensive technique called NetNCSP. It has two steps: support calculation and proximity determination. The backtracking approach is used to compute non-overlapping support on the relevant Nettree, hence reducing time complexity. They also suggest three types of pruning: inheritance, prediction, and determination.Prior to the creation of candidates, the pruning techniques can identify duplicate patterns and forecast their frequency and proximity. The results demonstrate that NetNCSP finds closed patterns in addition to being efficient. It mines the SARS-CoV-2 and SARS viruses' closed patterns.

Fractal, a high performance and high productivity system for enabling distributed GPM, is proposed by V. Dias et al. [4]. To adjust workload characteristics, Fractal uses a dynamic load-balancing based on a hierarchical and locality-aware work stealing technique. In order to avoid keeping a lot of intermediate state and increase memory efficiency, it also enumerates sub-graphs by combining a depth-first technique with the scratch processing paradigm. It offers a modular, expressive, and straightforward API for programmer productivity. In many sub-graph querying issues, fractal-based implementations perform better than both current options.

W. Gan et al. [5] expand the occupancy measure to evaluate the usefulness of patterns in order to extract high-quality patterns. They suggest the HUOPM algorithm. It takes into account user preferences such as occupancy, utility, and frequency. The utility occupancy list and FU-table are two data structures that are intended for pruning, together with a FU-tree. This approach does not need candidate generation in order to find the whole collection of high-quality patterns. Numerous datasets have been used for experiments. The patterns are understandable, logical, and acceptable, according to the results, and they perform better in terms of runtime and search space when pruned.

W. Gan, J. C. W. Lin, H. C. Chao, P. Fournier-Viger, V. S. Tseng, and P. S. Yu et al. [6] provided a thorough study of utility-oriented pattern mining (UPM), including the most recent advancements in high-utility pattern mining approaches. The authors divided current systems into four categories: Apriori-based, tree-based,

projection-based, and vertical/horizontal data format-based, and analysed their benefits, drawbacks, and applications. The survey also identified open research challenges, such as improving mining efficiency, reducing computational complexity, and developing scalable algorithms for large and dynamic datasets, all of which provide important directions for future research in frequent and utility-oriented pattern mining.

The goal of C. R. Wijesinghe et al [7] is to uncover frequent workflow patterns in a corpus of Galaxy bioinformatics workloads. FSG algorithm is used to analyse processes. Seventy-one repetitive process patterns were discovered, with at least 5% support. The next idea is to annotate the found frequent patterns and embed them in workflow systems with the goal of increasing usability through a high-level interface for the user. DNA's primary purpose is information storing. The progress of sequencing technology has allowed DNA sequence data to rise at an exponential rate, pushing the subject into big data. Furthermore, Machine Learning (ML) is a technique for analysing vast amounts of data and learning to acquire information.A. Yang et al [8] discuss the development path of sequencing technology, as well as the structure and similarity of DNA sequences. They examine the data mining process, summarise ML methods, and discuss the issues that ML algorithms face while mining biological sequences. Then, consider four ML applications for DNA sequence data: sequence alignment, classification, clustering, and pattern mining. They evaluated the biological application's origins and relevance, as well as its development and issues. Finally, summarise the substance of the review and consider future directions.

F. Wang et al. [9] developed an ARM-based analytic framework to investigate the effects of features on PDR under TOU pricing. First, a customer PDR characterisation model is developed, which uses a difference-in-difference model to quantify the impact and a probability distribution fitting approach to characterise the feature. Then, an ARM analysis utilising the Apriori method is provided to investigate the impacts in four categories: home features, socio-demographic, appliances and heating, and energy attitudes. Finally, results from 2993 records containing smart metering data show that PDR level cannot be determined solely on ownership and consumption. Sociodemographic statistics, Internet access, and home insulation all lead to a rise in PDR levels. The proportion of renewable power also demonstrates a link.The framework enhances the advantages of TOU programs and helps policymakers build effective energy-saving strategies.

Bilharzia, also known as schistosomiasis, is one of the most deadly and deceptive diseases. Predicting and identifying the causes of illness in its early stages may help guide treatment. Data mining tools help medical experts classify disorders. Y. Ali et al [10] explore the recovery and mortality variables that contribute to the schistosomiasis illness dataset, which was gathered in Hubei, China. ARM (Apriori) is a learning approach that is used to identify variables. Various tools were utilised for analysis and model evaluation, with a minimum support and confidence level of greater than 90%, to develop rules. In addition, characteristics suggesting recovery and death of people were found. Strong relationship variables, including as BMI, viability, nutrition, and amount of ascites, were identified and categorised.The results obtained by ARM may be valuable to experts in making accurate treatment decisions.

H. Deng et al. [11] present an approach for reducing the number of database scans. It is based on the downward disclosure attribute, which adds candidate item sets at various points in time during the scan. This dynamic block is produced from the database designated by start points, and unlike earlier Apriori algorithms, the sets of candidates are dynamically changed during the database scan. Unlike Apriori, it does not start the next level scan at the conclusion of the previous level scan; instead, it begins the scan by attaching a starting label to each dynamic partition of candidate sets.This minimises the database search for locating frequent item sets by simply adding a new candidate at any point during the runtime. However, it creates a big number of candidates, and computing their frequencies is the bottleneck of speed, whereas database scans only use a tiny portion of runtime.

According to S. Das et al [12], in 2011, 4,432 pedestrians were killed and 69,000 were wounded in vehicle-pedestrian collisions in Louisiana, United States. Vehicle collisions have become a source of worry due to the high fatality rate. In 2012, pedestrians accounted for 17%. Alcohol was present in roughly 44%. The authors used the 'Apriori' technique to detect patterns. They want to use eight years of data (2004-2011) to uncover trends in vehicle-pedestrian accidents. The findings suggested that nighttime street illumination reduced pedestrian accidents.A few groups of interest were identified: male pedestrians' higher propensity for severe and fatal crashes, younger female drivers (15-24) being more crash-prone, vulnerable impaired pedestrians even on roadways with night lighting, middle-aged male pedestrians (35-54) being more likely to crash, and the dominance of single vehicle crashes. Based on the trends, they offer many methods to improve safety. The data will assist traffic safety specialists in analysing patterns and developing remedies to promote awareness and improve pedestrian safety.

S. Gupta and R. S. Sikarwar et al. [13] provided a thorough examination of association rule mining strategies for efficient frequent pattern identification. The study analyses traditional algorithms like Apriori, FP-Growth, and Eclat in terms of execution time, memory usage, and scalability. The authors addressed the algorithms' merits and weaknesses, emphasising the need for better mining strategies that can successfully handle large-scale datasets while lowering computing complexity.

P. Kallay and T. D. Mihoc et al. [14] performed a comparative analysis of popular pattern mining algorithms, comparing their performance across various datasets and mining settings. The study found that the efficiency of Apriori, FP-Tree, and Eclat depended on the dataset's properties and application needs. The authors found that no single method is superior in all scenarios and recommended more study into hybrid and optimised frequent pattern mining algorithms.

Although numerous studies have compared frequent pattern mining algorithms and proposed improvements in execution time and scalability, limited work has focused on identifying the most suitable algorithm for future enhancement through a consolidated comparative analysis. Therefore, this study analyzes the characteristics of Apriori, FP-Tree, and Eclat based on findings reported in previous studies to identify a promising algorithm for further improvement.

III. FREQUENT PATTERN MINING ALGORITHMS

In this part, the Apriori, FP-tree, and Eclat algorithms are discussed. The quick summary of

the frequent pattern mining algorithm is as follows:

Apriori algorithm: The Apriori algorithm is a basic method for mining frequent item sets in a transactional database. The method examines the database for frequent itemsets, which are then utilised to build k+1- itemsets. Each k-itemset must be more than or equal to the minimal support level to be considered frequent. Otherwise, it is known as candidate itemsets. First, the algorithm scans the database to determine the frequency of 1-itemsets with only one item by counting each item in the database. The frequency of 1-itemsets is used to locate the itemsets in 2-itemsets, which are then used to find 3-itemsets, and so on until no k-itemsets remain. If an itemset is not frequent, then any big subset of it is likewise not frequent; this condition reduces the search space in the database.

The FP-Tree algorithm: Frequent Pattern Tree creates a tree-like structure from the database's original itemsets. The goal of the FP tree is to find the most common pattern. Each node in the FP tree represents an item from the itemset. The root node symbolises null, while the lesser nodes represent the itemsets. The relationship of the nodes with the lower nodes, or the itemsets with the other itemsets, is maintained while the tree is being formed. The frequent pattern growth approach allows us to locate the frequent pattern without creating candidates.

The ECLAT algorithm stands for Equivalence Class Clustering and Bottom-Up Lattice Traversal. It is a common approach to Association Rule mining. It is a more efficient and scalable implementation of the Apriori algorithm. The Apriori algorithm mimics the Breadth-First Search horizontally, whereas the ECLAT method mimics the Depth-First Search vertically. This method to the technique makes it quicker than the Apriori algorithm.The core concept is to utilise Transaction Id intersections to compute a candidate's support while avoiding the production of subsets that are not in the prefix tree. The initial call to the method uses just single items.The function is then executed recursively, with each call verifying and combining each item-tidset pair with the previous one. This procedure is repeated until no candidate item-tidset pairings can be merged.

IV. COMPARATIVE ANALYSIS OF FREQUENT PATTERN MINING ALGORITHMS

This section presents a comparative analysis of the Apriori, FP-Tree, and Eclat algorithms using findings from past research. The comparison is based on execution time, memory utilisation, candidate creation, scalability, and the number of frequent rules created. The goal is to discover the best algorithm for future improvement.

The Apriori algorithm is one of the earliest and most widely used algorithms for frequent pattern mining. It uses a breadth-first search approach to produce candidate item sets repeatedly. Although Apriori is simple to implement and generates a large number of association rules, its repetitive database scans and intensive candidate creation increase execution time and computational complexity, particularly for big datasets. As a result, its performance declines as the number of transactions and items increases. [1],[13].

The FP-Tree technique solves the candidate generation problem by condensing the transactional database into a Frequent Pattern Tree. Compared to Apriori, this technique considerably decreases the number of database scans while improving execution performance. However, generating and maintaining the FP-Tree

may necessitate more memory, especially for big datasets. Overall, FP-Tree strikes a favourable balance between execution time and memory utilisation. [1], [5], [13].

To find common itemsets, the Eclat method uses a vertical database representation using Transaction ID (TID) sets and does a depth-first search. Instead of continually scanning the database, it computes support using TID-set intersections, which reduces execution time and improves scalability. Several studies have found that Eclat outperforms Apriori and FP-Tree on big and sparse transactional datasets due to its efficient search technique and low computing cost. [1], [13], [14].

According to the literature, Eclat generally provides better execution performance, whereas Apriori generates a larger number of association rules.By eliminating candidate creation, FP-Tree provides a good balance between execution speed and memory efficiency. As a result, choosing an effective frequent pattern mining technique is determined by the dataset's properties as well as the individual application needs. Furthermore, the literature suggests that enhancing the Apriori algorithm by lowering candidate creation, minimising execution time, and optimising memory utilisation is an interesting area for future study. [5],[13], [14].

Table I summarises the main features of the Apriori, FP-Tree, and Eclat algorithms. According to prior research, Eclat outperforms other execution systems due to its vertical data format and efficient support computing. FP-Tree simplifies database searches by removing candidate creation, yet Apriori remains a straightforward and understandable method that creates a higher number of association rules. As a result, the algorithm should be chosen based on the dataset's properties and the unique application needs.

Parameter	Apriori	FP-Tree	Eclat
Search Strategy	Breadth-First Search	Pattern Growth	Depth-First Search
Database Format	Horizontal	Horizontal	Vertical
Candidate Generation	Yes	No	No
Database Scans	Multiple	Two	One (initial scan)
Execution Time	High	Moderate	Low
Memory Usage	Moderate	Moderate to High	Low to Moderate
Scalability	Low	Good	Excellent
Rule Generation	High	Moderate	Moderate
Suitable Dataset	Small to Medium	Dense Datasets	Large/Sparse Datasets
Suitable Applications	Market Basket Analysis	Dense Transaction Databases	Large Sparse Databases

Table I. Comparative Analysis of Apriori, FP-Tree, and Eclat Algorithms

V. CONCLUSION

Association rule mining and frequent pattern mining are traditional data mining approaches for identifying significant links between often recurring elements in transactional databases. The fundamental goal of frequent pattern mining is to find all meaningful frequent item sets in a given dataset. Apriori, FP-Tree, and Eclat are the most

popular algorithms for this purpose because to their efficacy and wide range of applications. This study presented a comparative analysis of the performance of these three algorithms in terms of the number of frequently created rules and execution time. The comparative analysis based on previous studies indicates that the Eclat algorithm has the highest execution performance, but the Apriori method creates a greater number of frequent rules. As a result, the right algorithm is determined by the application's unique needs, such as execution speed or rule creation capabilities. Furthermore, despite its increased computational cost, the Apriori method provides several potential for further improvement. Future research may concentrate on improving Apriori by shortening execution time, decreasing information loss, optimising candidate creation, and increasing scalability for large-scale datasets. Such enhancements can improve its efficiency and widen its use in frequent pattern mining and association rule mining jobs.

REFERENCES

[1] C. H. Chee, J. Jaafar, I. A. Aziz, M. H. Hasan, W. Yeoh, “Algorithms for frequent itemset mining: a literature review”,Artif Intell Rev,Vol.52, Page No. 2603–2621, 2019

[2] F. Min, Z. H. Zhang, W. J. Zhai, R. P. Shen, “Frequent pattern discovery with tri-partition alphabets”, Information Sciences 000 (2018) 1–18

[3] Y. Wu, C. Zhu, Y. Li, L. Guo, X. Wu, “NetNCSP: Nonoverlapping closed sequential pattern mining”, Knowledge-Based Systems 196 (2020) 105812

[4] V. Dias, C. H. C. Teixeira, D. Guedes, W. Meira Jr., S. Parthasarathy, “Fractal: A General-Purpose Graph Pattern Mining System”, SIGMOD ’19, June 30–July 5, 2019, Amsterdam, Netherlands, ACM

[5] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, and P. S. Yu, “HUOPM: High Utility Occupancy Pattern Mining,” IEEE Transactions on Cybernetics, vol. 50, no. 3, pp. 1195–1208, 2020.

[6] W. Gan, J. C. W. Lin, H. C. Chao, P. Fournier-Viger, and P. S. Yu, "A Survey of Utility-Oriented Pattern Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1306–1327, Apr. 2021.

[7] C. R. Wijesinghe, A. R. Weerasinghe, “Mining Frequent Patterns in Bioinformatics Workflows”, International Journal of Bioscience, Biochemistry and Bioinformatics, Volume 10, Number 4, October 2020

[8] A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, L. Zhang, “Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA”, Front. Bioeng. Biotechnol. 8:1032, 2020

[9] F. Wang, L. Liu, K. Li, N. Duić, M. S. Khah, J. P. S. Catalão, “Impact Factors Analysis on the Probability Characterized Effects of Time of Use Demand Response Tariffs Using Association Rule Mining Method”, Energy Conversion and Management 197, 2019

[10] Y. Ali, A. Farooq, T. M. Alam, M. S. Farooq, M. J. Awan, T. I. Baig, “Detection of Schistosomiasis Factors Using Association Rule Mining”, IEEE Access VOLUME 7, 2019

[11] Z. H. Deng and S. L. Lv, "PrePost+: an efficient N-Iists-based algorithm for mining frequent item sets via children-parent equivalence pruning”, Expert Systems with Applications, Vol. 42, Issue 13, Page No. 5424-5432, 2015.

[12] S. Das, A. Dutta, R. Avelar, K. Dixon, X. Sun, M. Jalayer, “Supervised association rules

mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures", International Journal of Urban Sciences, 2018

[13] Gupta, S., and Sikarwar, R. S., "A Review of Association Rule Mining Techniques Toward Efficient Frequent Pattern Discovery," Global Journal of Advanced Engineering Technologies and Sciences, vol. 11, no. 1, pp. 1–10, 2024. DOI:10.64149/gjaets.11.1.1-10.

[14] Kallay, P., and Mihoc, T. D., "Comparative Analysis of Frequent Pattern Mining Algorithms," *Acta Universitatis Sapientiae, Informatica*, vol. 17, Article 8, 2025.

[15] "Hemlata Pate and Dhanraj Verma" Data Classification in Web Usage Mining using SVM, SVR and K-NN, Journal

[16] of Innovative Engineering and Research (JIER), Vol.- 4,Issue -2, pp. 18-22, October 2021.

[17] "Kailash Patidar and Dr. Dhanraj Verma" A Survey of Evolutionary Algorithms for Data Mining, Journal of

[18] Innovative Engineering and Research (JIER), Vol.- 4,Issue -1,pp. 1-3, April 2021.

[19] "Sami Abdul Qader Mohammed Al – Ademi and Dr. Sanjay Singh Bhadoriya" Study on Data Mining Tools Open

[20] Source, Journal of Innovative Engineering and Research (JIER), Vol.- 4,Issue -1,pp. 15-18, April 2021.

[21] "Nidhi Nigam Verma and Deepika Pathak" A Brief Study of Various Data Mining Techniques and its Applications

[22] in Internet Banking, Journal of Innovative Engineering and Research (JIER), Vol.- 2,Issue - 1, pp. 1-4, April 2019.

[23] "Deepika Pathak and Dhanraj Verma" Mining Frequent Item Sets over Dynamic Data Streams by Efficiently

[24] Removing Aged Transaction from Current Sliding Window, Journal of Innovative Engineering and Research (JIER),Vol.- 1, Issue - 1, pp. 37-40, April 2018.