A Study on Frequent Pattern Mining and Its Applications

Saravanan.Suba[1], Selvavinayagam.R[2]

[1]*Research scholar,Manonmaniam Sundaranar University,TN,India,*

[2]*Research scholar,Bharathiyar University,TN,India,*

Abstract

Data mining involves finding interesting patterns from huge dataset to maximize the income of the future trade. Association rule mining is the core area in the field of data mining exploration with wide range of applications such as retail industry, Healthcare and Insurance, banking sector, fraud detection and biological data analysis. Determining the frequent patterns in large dataset is the main task of association rule mining and it is frequently used by business decision makers to improve their future business strategy. Various algorithms have been introduced in the literature and tremendous progresses have been done to find frequent patterns. The analysis of literature survey will offer the information about what has been done formerly in the similar area, what is the present tendency and what are the other connected areas. This paper clarifies the concepts of Frequent Pattern Mining and three important approaches that is candidate generation approach, without candidate generation and vertical layout approach. It also explains various frequent pattern mining algorithms and how it can be applied to diverse areas. This paper also surely helps the researches to get clear idea about the application of frequent pattern mining algorithms in various areas.The references given in this paper explain the major theoretical issues and guiding the researcher in interesting research directions that has yet to be explored.

Keywords— *Apriori, Eclat, FIN, FP-Growth ,PrePost*

I.INTRODUCTION

After the invention of internet, there has been an exponential development in the production and management of digital information as various operations of human beings and business are computerized. Most of the organizations have started to understand that the information gathered over years is a significant tactical asset and they also understand that there is potential business information hidden in the huge amount of data. For that, what these organizations want a system that permits them to extract the most appreciated information from stored data. The field of data mining provides such techniques which find relations among the current data and concluding hidden information that would be very useful in future decision making [1].

Data mining is a group of methods for automatic finding of previously unknown, valid, novel, useful and understandable pattern in large dataset [1]. The pattern must be usable so that they can be utilized in the enterprise's future decision making process. It composed of various segments such as Data Cleaning removes noise and further unreliable data which are present in the input dataset, Data Integration methods are engaged to combine data from various sources because the dataset is usually collected from various sources and it is normally termed as Data Warehouse, Data selection stage catches the specific part of whole dataset relevant to mining task in the whole dataset, Data Transformation stage which transfer the selected dataset into format appropriate for data mining and the data mining task applies intelligent methods to mine hidden useful knowledge in given dataset [2].

The Data mining tasks can be commonly categorized as two types such as descriptive and predictive. Descriptive method defines an algorithm in which the vital characteristics of the data in the dataset are represented. The

descriptive procedures comprise tasks like Association, Sequential Mining and Clustering [3]. Predictive jobs are those that execute inference on input dataset to reach at hidden knowledge and produce interesting and beneficial prediction [2]. The predictive techniques include tasks such as Classification, Regression and Deviation [3]. Usually Data mining task is encouraged by decision support problems faced by most business enterprise and is designated as indispensable area of research. Key research disputes or challenges in Data Mining are mining approach, different applications, user interaction, data variety and performance. So the mining methodologies must be competent and scalable fine to the scope of dataset and their run times [2].

The association rule mining is one of the most widespread descriptive Data mining methods [3]. After its introduction, it has attracted great interest among data mining investigator and experts [5]. It finds correlations between variables in huge dataset. For example,

$$\forall p \in persons, buys(p,\text{mobile}) \rightarrow buys(a, headphone) \quad (1)$$

The rule in (1) says that a great ratio of persons who buy mobile will also buy headphone.

The Association Rule Mining problem can be divided into two sub problem such as frequent pattern mining and rule production [2]. There are two statistical measures such as support and confidence that control the process of association rule mining. This can be mathematically given as follows.

Let M = $\{x_1, x_2 \ldots x_n\}$ denotes group of elements. Any subsection of M is called element set. Let DS be a collection of records. Each record X in DS states group of elements such that $X \subseteq M$. Each record has a distinctive identifier (Transaction Number). Let E, F be one or more elements, Association rule can be formed as $E \rightarrow F$, $E \wedge F = \phi$, where E is an antecedent and F is the consequent of the rule. It defines two important statistical methods such as support and confidence that regulate the procedure of association rule mining.

$$Support(E \rightarrow F) = \sum(E \cup F)/N \quad (2)$$

$$Confidence(E \rightarrow F) = \sum(E \cup F)/\sum E \quad (3)$$

The whole process of association mining is organized by user specific parameter such as minimum support and confidence [2]. The first frequent pattern mining task was suggested by Agrawal et al. [5] to determine the exciting links among items in market basket transactions. Various studies have been being conducted after its introduction to address various conceptual, implementation and application topics concerning to this correlation analysis task. The competence of association mining rule can be based on computational method applied to detect the required patterns. The computation model of detecting association rule in dataset can be serial or parallel and on line or batch. This research study mostly focuses on finding frequent patterns based on implementation issues. This paper revises the existing methods for the same and challenges pertaining to this domain are raised.

This paper is organized as follows: Section II presents the literature reviews of frequent pattern mining and compares the various important existing frequent pattern mining algorithms according to its effectiveness and improvements, Section III explains the applications of frequent pattern mining, Section IV discusses the challenging issues in frequent pattern mining and section V describes the conclusion.

II. LITERATURE SURVEY OF EXISTING FREQUENT PATTERN MINING (FPM) METHODS

Data mining methods find interesting hidden knowledge from huge dataset. One of the main data mining methods is Association rule mining because it plays major role in other data mining methods. It normally catches connection among items in huge dataset. It strictly discovers the frequent pattern as its basic task. Various algorithms have been suggested to discover frequent patterns. Typically those procedures can be categorized into

two types: candidate generation method or pattern growth method. They are explained in the following sections. The research activities involves integrating the mining competency into present database technology, Designing competent and scalable algorithms, Handling Domain specific limits and Post processing of mined patterns.

2.1 AIS Algorithm

The very first algorithm to discover frequent pattern was the AIS (Agrawal, Imielinski and Swami) algorithm suggested by Agrawal et al. in 1993 [5]. It mainly motivates to improve the quality of dataset together with necessary functionality to process decision support queries. The main shortcoming of the AIS algorithm is too many candidate itemsets that lastly turned out to be small are made. So it require more space and times and scans.

2.2 Apriori Algorithm

Apriori algorithm was first suggested by Agrawal and Srikant [6]. The AIS is just a direct technique that needs several scans over the dataset and producing lot of candidate itemsets and saving counters of each candidate while most of them will become not frequent. Apriori executes efficiently during the candidate generation process for two reasons, it applies a diverse candidate's generation method and a novel pruning technique.

In Apriori, the candidates are formed by connecting among the frequent itemsets level-wisely and are filtered according to the Apriori property. So it decreases the computation, I/O cost and memory space requirement than AIS. These were implemented using Apriori-gen and GenerateRules functions by Agrawal and Srikant in 1994[6]. Even though Apriori works better than AIS, it still has the drawback of scanning the entire dataset multiple times based on the occurrence of the dataset.

Based on Apriori algorithm, several new have been designed and implemented with some modification and improvements. Usually there were two approaches: one is to lessen the number of scans over the entire dataset another approach uses various filtering techniques to make the number of candidate itemsets much smaller as outcome based on user objectives[7].

2.3 Sampling Approach

Sampling method is one of a main data reduction technique that has been applied to several data mining methods for decreasing the computational overhead. Sampling speeds up the mining process by shrinking the number of transactions in the dataset to reduce I/O costs. Usually good statistical sampling techniques are applied to shrink the dataset based on user objectives [8].

In 1995, the first sampling algorithm to find frequent patterns to generate association rule was introduced by Toivonen [9]. This approach needs two passes to find frequent patterns based on support threshold. The frequent patterns in the sample dataset are obtained during pass1. These patterns obtained from sample dataset are then validated against the entire dataset.

2.4 Partitioning Approach

The partitioning algorithm was suggested by Savasere et al. [10] in 1995 using divide and conquer method to find of frequent patterns based on support threshold. This method divides the input dataset in to several disjoined parts (partitions) and frequent items with in the partitions are produced first. Then those frequent itemsets got from various partitions are joined and find the superset of all frequent itemsets in the complete dataset. The final counts of frequent itemsets are computed during the second read of the dataset. So this method of partitioning may increase the performance of discovering large itemsets in numerous ways. The distributed and parallel algorithms can be implemented with the help of this partitioning method.

2.5 Vertical Layout Approach (Eclat)

Eclat (Equivalence CLAss Transformation) is also uses depth first search to generate all frequent patterns based on user given support threshold and it was suggested by Mohammed J. Zaki in 2000. It uses this approach inside the Apriori method. It presents a structure to represent the dataset in vertical format using Transaction identifier. It determines that all frequent patterns can be counted via simple transaction identifier (TID) list intersections. This algorithm is very useful when the found frequent patterns are big. [11].

2.6 FP Growth Approach

Most of the algorithms debated earlier have basically used a breadth first strategy towards to find the frequent pattern mining. It uses the depth first search to discover the frequent patterns [2]. The Apriori method considerably minimizes the candidate sets size using the Apriori property. Any way it has the following disadvantages such as producing a huge number of candidate sets, continually reading the dataset and testing the candidates by pattern matching to find the targeted frequent patterns.

The FP-Growth algorithm to find all frequent patterns without candidate generation was introduced by Han et al. [12] in 2000 using FP-tree structure. It also uses the divide and conquers principle to find the frequent patterns based on given user support threshold. The construction of FP-tree in the main memory to find frequent pattern is a time consuming process.

2.7 dEclat

This was introduced to minimize the memory requirement of Eclat by Mohammed J. Zaki and Karam Gouda in 2003. It introduced new vertical data representation techniques called *diffset* to find the frequent patterns. It reduces the size of memory needed to store intermediate results. So, it outperforms than Eclat in terms of time and memory required [13].

2.8 PrePost

The PrePost method uses the advantages of FP-Growth and dEclat to improve the performance to produce significant frequent patterns for given support threshold. It was suggested by DENG ZhiHong, WANG ZhongHui & JIANG JiaJian in 2012. It uses one tree like FP-tree in FP-Growth called PPC-tree (Pre order Post order Code). This tree uses vertical data representation called N-list to store the basic information about frequent patterns. The N-list is a shortened data structure since transactions with common prefixes share the same nodes of the PPC-tree. The investigational outcomes of PrePost approve that PrePost is quicker than FP-Growth, Eclat and dEclat [14] to find the frequent patterns based on user specified support threshold.

2.9 FIN (Fast mining frequent itemsets using Nodesets)

In 2014 Zhi-Hong Deng and Sheng-Long Lv have suggested this FIN algorithm to speed up the performance of PrePost using Nodeset. Nodesets uses only the pre-order (or post-order code) of each node to find the required frequent patterns based on given support threshold. So, it saves nearly half of memory space instead of using N-lists. The empirical analysis of FIN proves that it outperforms than PrePost and FP-Growth [15].

2.10 PrePost+

This algorithm was also suggested by Zhi-Hong Deng, Sheng-Long Lv in 2015 to improve the performance of FIN and PrePost by applying a competent pruning strategy named Children–Parent Equivalence pruning to reduce the search space. So It outperforms than PrePost and FIN [16].

2.11 Comparative Analysis of Standard FPM algorithms

The comparison of any existing methods related particular task solving can be suitable to understand the competence and usefulness of any one method with others. It can be helpful to catch and avoid the problem in any of existing method so that to progress the competence and usefulness of that method. The Table I compare the efficiency of the major breakthrough algorithms in frequent pattern mining on the basis of number of dataset scans needed to find frequent patterns, type of search used and type of data structure required.

Table I: Comparisons of important FPM Algorithms

Algorithm	Search Type	No. of Scans	Data Structure
AIS	Breadth First Search	Multiple	List
Apriori	Breadth First Search	Multiple	Hash Table
Sampling	Not specified	2	Not specified
Partitioning	Breath first search	2	Hash table
Elcat	Depth first search	1	List
FP-Growth	Divide and conquer	2	FP-tree
PrePost	Divide and conquer	2	PPC-tree
FIN	Divide and conquer	2	PPC-tree
PrePost+	Divide and conquer	2	PPC-tree

Generally the performance of any algorithm is valued by calculating the mining time needed to run the algorithm. The run time of any mining algorithm is modified based on number of dataset reads required to complete the mining operation, type of search applied , type of data structure applied , number of condition needs to be checked (logic), number of records in the dataset , number of elements in the dataset and actual existence of transactions in the dataset. Based on the above debates, it is concluded that the PrePost+ outperforms than FIN, FIN outperforms than PrePost and PrePost outperforms than FP-Growth and FP-Growth outperforms than Apriori and Apriori outperforms than AIS to generate significant frequent patterns for user specified support threshold.

III. APPLICATION OF FREQUENT PATTERN MINING

Frequent Pattern Mining applications are applied to many different domains such as market basket and risk analysis in commercial situation epidemiology, clinical medicine, fluid dynamics, astrophysics, crime prevention, counter-terrorism, sale campaign analysis, Web log (click stream) analysis, DNA sequence analysis , loss-leader analysis, clustering, Insurance ,disease diagnosis and other areas in which the association between items can offer useful knowledge. The following subsections discuss some of the important applications.

3.1 Crime Detection

P. Dhakshinamoorthy, T.Kalaiselvan has introduced new frequent pattern mining method to investigate the information between the police department and computer science department. It uses frequent pattern detection technique to solve the crimes faster in order to help the policemen [17].

3.2 Network Forensic Investigation

XIUYU ZHONG has suggested network forensic analysis by using Apriori algorithm. To protect the products in the network against intrusion methods, network forensic is required. Lot of data are caught and investigated in network forensics and after catching and filtering network data package, the Apriori algorithm is applied to mine the frequent patterns and association rules according to the evidence relevance and further it lessen the number of matching times significantly and efficiently improve the crime detection. The outcomes display that the application of Apriori algorithm can increase the speed of data analysis for network forensics. This application can aid to resolve the real-time adaptable problems in network forensics [18].

3.3 Network Attacks

S.S.Garasia, D.P.Rana, R.G.Mehta has discussed about Botnet is one of severe threat in cyber-attacks. A botnet is defined as a group of co-operated computer systems which are distantly controlled by hackers to spread various network attacks, such as click fraud, identity theft, and spam and information phishing. Those authors have presented a new method for botnet detection using Timestamp and frequent pattern set produced by the Apriori algorithm. The core benefit of the suggested method is that previous knowledge of Botnets like Botnet signature is not needed to detect the malevolent botnets [19].

3.4 Animal behaviour analysis

Susan P. Imberman, Michael E. Kress and Dan P. McCloskey [20] have introduced the housing environment with animals. They have used housing environment equipped with a system of RFID sensors. RFID transponders have been used to study animal and the naked mole rat. The outcomes have been was analysed using principal component analysis and frequent pattern mining. The results have demonstrated that these methods can classify time periods of high behavioural activity from that of low behavioural activity along with which groups of animals interacted with one another.

3.5 Educational Data

Dr.Vijayalakshmi M N, S.Anupama Kumar and Kavyashree BN [21] have suggested the application of association mining on educational data to understand the knowledge and performance of students. They have applied Apriori algorithm on student log dataset to get the interesting association rules. Those rules can be helpful to evaluate the performance of the students and to predict the quality of education given in the educational institutions. The algorithm produced frequent patterns using support threshold to understand the attention of the students in the course. Interesting rules are produced based on found frequent patterns using confidence threshold in the dataset to predict the future performance of the students.

IV. CHALLENGING ISSUES AND PROPOSED METHOD

This paper has listed recent and important frequent pattern mining algorithms. But there are still numerous critical research problems that need to be solved yet. Most of the methods available in the literature to mine frequent patterns do not offer the flexibility for reusing the computation done during mining process and it still requests research to shrink the size of derived pattern and enhance the quality of found patterns.

Most of the frequent mining algorithms find all frequent patterns based on support threshold but some of the application may need specific patterns than all patterns to analyze the future tendency of the particular applications. It still needs a frequent pattern mining method to find specific frequent patterns directly based on user objective to predict the future action of particular application. So it still needs to do research to find better

frequent pattern mining algorithm to predict the future of any applications which needs frequent pattern mining based on goal of the given application.

The frequent pattern mining algorithms are the base for other datamining techniques such as classification, clustering and sequence discovery. So it still needs research to connect the frequent pattern mining algorithm with other data mining tasks such as classification and clustering to solve the real world efficiently.

It should be proposed a new frequent mining method that solves any one of the above mentioned drawback of existing algorithm with suitable application which changes the society in a better way.

V. CONCLUSION

The core task of association rule mining is frequent pattern mining. It can be very useful in other data mining tasks such as clustering, sequence discovery and classification. The overview of current literature and future research directions of frequent pattern mining have been presented. It has been understood that the frequent pattern mining has proved wonderful progress and ease the solving of many problems in the real world. So it still needs deep research to solve several critical issues mentioned in the challenging issues section.This survey has presented a noteworthy fundamental contributions made within frequent pattern mining research over this time. This can be useful to detect some gap available in the current knowledge for future research in this field. It can be applied proficiently in numerous social and health care related problems to get better life in future.

VI. REFERENCES

[1] N, P, Gopalan & B, Sivaselvan, ”*Data mining Techniques and Trends*”, PHI learning private limited, New Delhi, 2009.

[2] G.K. Gupta, “*Introduction to Data mining with Case Studies*”, PHI learning private limited, New Delhi, 2009.

[3] S.Shankar and T.Purusothaman “Utility Sentient Frequent Item set Mining and Association Rule Mining: A Literature survey and Comparative Study”, International Journal of Soft Computing Applications ISSN: 1453-2277, Issue 4, 2009, pp.81-95.

[4] R.Agrawal, T.Imielinski, and A.Swami, “Mining Association Rules Between Sets Of Items In Large Databases”, In proceedings of the ACM SIGMOD International Conference on Management of data, 1993,pp. 207-216.

[5] M. J. Zaki and C.J. Hsiao “CHARM: An efficient algorithm for closed association rule mining”, Technical Report 99-10, Computer Science Dept., Rensselaer Polytechnic Institute, October 1999.

[6] Rakesh Agrawal and Ramakrishnan Srikant “Fast Algorithms For Mining Association Rules In Large Databases”, In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Sep’1994,pp 487-499.

[7] Anurag Choubey, Ravindra Patel,J. L.Rana “A Survey Of Efficient Algorithms And New Approach For Fast Discovery Of Frequent Item Set For Association Rule Mining”, International Journal of Soft Computing and Engineering, MAY 2011.

[8] V. Umarani et al. ”A Study on Effective Mining of Association Rules from Huge Databases”, International journal of computer science and research, Vol .1, Issue 1,2010.

[9] Toivonen H “Sampling large databases for association rules “, In VLDB Journal, 1996,pp. 134-145.

[10] Savasere A, Omiecinski E, and Navathe S, “An Efficient Algorithm For Mining Association Rules In Large Databases”, In Proceedings of 20th International Conference on VLDB,1995.

[11] Zaki M.J, 'Scalable Algorithms for Association Rule Mining', IEEE Transactions on Knowledge and Data Mining 12(3), May 2000,pp.372-390.

[12] Han,J and Pei.J, " Mining Frequent Patterns By Pattern Growth: Methodology And Implications", SIGKDD Explorations 2,2000, pp.14–20.

[13] Mohammed J. Zaki & Karam Gouda,' Fast Vertical Mining Using Diffsets' *SIGKDD* '03 , Washington, DC, USA, 2003.

[14] DENG, ZhiHong, WANG, ZhongHui & JIANG, JiaJian, 'A new algorithm for fast mining frequent itemsets using N-lists', Science China Press and Springer-Verlag Berlin Heidelberg, Vol. 55 No. 9, 2012,pp. 2008–2030.

[15] Zhi-Hong Deng & Sheng-Long Lv' Fast mining frequent itemsets using Nodeset, Expert Systems with Applications, Elsevier publications, Vol. No. 41, pp. 4505–4512, 2014.

[16] Zhi-Hong Deng & Sheng-Long Lv, 'PrePost+: An efficient N-lists-based algorithm for mining frequent itemsets via Children–Parent Equivalence pruning' Expert Systems with Applications, Elsevier publications, Vol. 42, pp. 5424–5432, 2015.

[17]` M. Dhakshinamoorthy, T.Kalaiselvan, "Crime Pattern Detection Using Data Mining"" , International Journal of Advanced Research in Computer Science and Applications, Vol.1, Issue.1,2013, pp.46-50.

[18] Xiuyu Zhong " The Application of Apriori Algorithm For Network Forensics Analysis", Journal of Theoretical and Applied Information Technology, Vol.50, Issue.2, 2013,pp.430-434.

[19] S.S.Garasia, D.P.Rana, R.G.Mehta "HTTP Botnet Detection Using Frequent Pattern set Mining", International Journal of Engineering Science & Advanced Technology Vol.2, Issue.3,2012, pp.619-624.

[20] Susan P. Imberman, Michael E. Kress, Dan P. McCloskey "Using Frequent Pattern Mining to Identify Behaviours in a Naked Mole Rat Colony", Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, 2012,pp.394-399.

[21] Dr.Vijayalakshmi M N, S.Anupama Kumar, Kavyashree BN. 2014. " Investigating Interesting Rules Using Association Mining for Educational Data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2,2014, pp.268-271.