Comparative Analysis of Single And Multiple Level Association Rule Mining

M.Saranya[1], C.Kavitha[2]
[1]Dept of Computer Science
[2]Assistant Professor, Dept of Computer Science
[1, 2] Sri Kaliswari College(Autonomous) Sivakasi

Abstract- *Association Rule mining is one of the most important fields in data mining and knowledge discovery. This paper proposes an compared in multiple level algorithms. The main aim is to study and analyze the various existing techniques for mining frequent itemsets and evaluate the performance of new techniques and compare with the existing classical Apriori and FP- tree algorithm and Eclat and RFM Analysis.*

I. INTRODUCTION

With the increase in Information Technology, the size of the databases created by the organizations due to the availability of low-cost storage and the evolution in the data capturing technologies is also increasing,. These organization sectors include retail, petroleum, telecommunications, utilities, manufacturing, transportation, credit cards, insurance, banking and many others, extracting the valuable data, it necessary to

This valuable information can help the decision maker to make accurate future decisions. KDD applications deliver measurable benefits, including reduced cost of doing business, enhanced profitability, and improved quality of service. Therefore Knowledge Discovery in Databases has become one of the most active and exciting research areas in the database community.

II. LITERATURE REVIEW

Analyzing all the data that is collected in the data store or warehouse is definitely a necessity for every enterprise because a more proper decision can be made considering all data set[1]s. In order to provide users with information that is more useful for data analysis and decision making, it is important to mine and identify all the significant hidden patterns that exist frequently in a data set. [2]Therefore, this paper analyzes a number of FPM algorithms to provide an overview of the FPM state-of- the-art[3]. The previous works done on FPM algorithms are presented in Sects. 2.1 to 2.10, while Sect. 2.11 presents a table which provides a comparison of the fundamental and significant FPM algorithms that have been proposed by other researchers[5].

This section has been divided into two sections [8–10]. First section single level includes some basic concepts related to data mining, its techniques and specifically association rules mining which are helpful for carrying out present research work. In second section explore the literature related to multiple-level association rules mining algorithms and last section includes miscellaneous research paper which helps the researcher to carrying out the research work directly or indirectly on –design of an improved multiple level association rule algorithm for discovery of frequent patterns.

III. METHODOLOGY

Apriori algorithm

Apriori (Agrawal and Srikant 1994) is an algorithm that mines frequent itemsets for generating Boolean association rules. It uses an iterative level-wise search technique to discover (k + 1)-itemsets from k-itemsets. A sample of transactional data that consists of product items being purchased at different transactions is shown in Table 1. First, the database is scanned to identify all the frequent 1-itemsets by counting each of them and capturing those that satisfy the minimum support threshold.

The identification of each frequent itemset requires of scanning the entire database until no more frequent k-itemsets is possible to be identified. According to Fig. 2, the minimum support threshold used is 2. Therefore, only the records that fulfill a minimum support count of 2 will be included into the next cycle of algorithm processing.

Table 1 Sample of transactional data. Reproduced with permission from (Han et al. 2012)

TID	list of items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I5,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3,I2

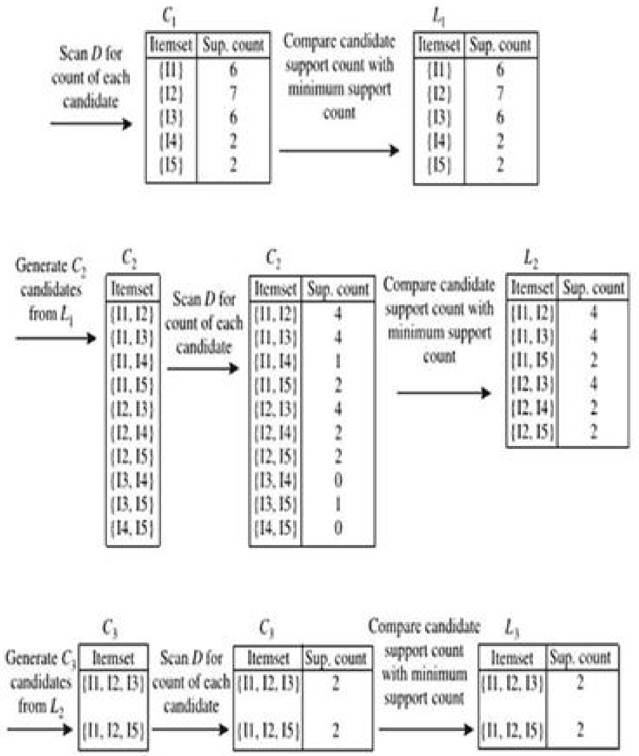


Fig. 2 Generation of candidate itemsets and frequent itemsets. Reproduced with permission from (Han et al. 2012)

In many cases, the Apriori algorithm reduces the size of candidate itemsets significantly and provides a good performance gain. However, it is still suffering from two critical limitations (Han et al. 2012). First, a large number of candidate itemsets may still need to be generated if the total count of a frequent k-itemsets increases. Then, the entire database is required to be scanned repeatedly and a huge set of candidate items are required to be verified using the technique of pattern matching

FP-Growth algorithm :

Frequent Pattern Growth (FP-Growth) (Han et al. 2000) is an algorithm that mines frequent itemsets without a costly candidate generation process. It implements a divide-and-conquer technique to compress the frequent items into a Frequent Pattern Tree (FP-Tree) that retains the association information of the frequent items. The FP- Tree is further divided into a set of Conditional FP-Trees for each frequent item so that they can be mined separately. An example of the FP-Tree that represents the frequent items is shown in Fig. 3. The FP- Growth algorithm solves the problem of identifying long frequent patterns by searching through smaller Conditional FP- Trees repeatedly.

An example of the Conditional FP-Tree associated with node I3 is shown in Fig. 4, and the details of all the Conditional FPTrees found in Fig. 3 are shown in Table2. The Conditional Pattern Base is a “sub- database” which consists of every prefix path in the FP-Tree that co-occurs with every frequent length-1 item. It is used to construct the Conditional FP-Tree and generate all the frequent pattern

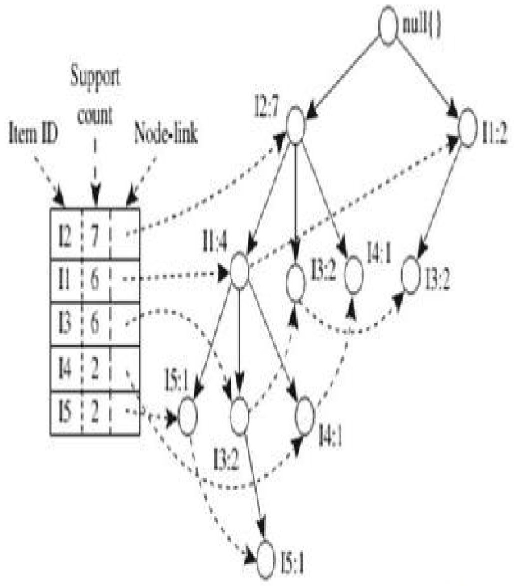


Fig. 3 Frequent pattern tree (FP-Tree). Reproduced with permission from (Han et al. 2012)

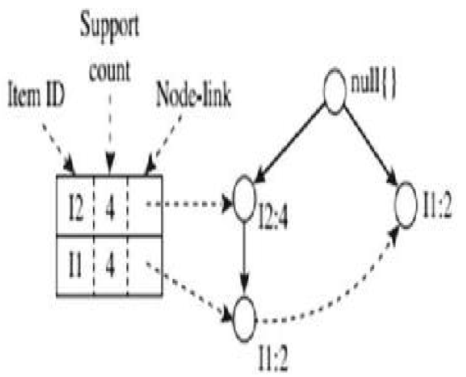


Fig. 4 Conditional FP-Tree associated with Node I3. Reproduced with permission from (Han et al. 2012)

Table 2 Conditional Pattern Base and conditional FP-Tree. Reproduced with permission from (Han et al. 2012)

Item	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	{I2: 2, I1: 2}	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	{I2: 2}	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	{I2: 4, I1: 2}, {I1: 2}	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	{I2: 4}	{I2, I1: 4}

related to every frequent length-1 item. In this way, the cost of searching for the frequent patterns is substantially reduced. However, constructing the FP-Tree is timeconsuming if the data set is very large (Meenakshi 2015)

Eclat algorithm:

Equivalence Class Transformation (EClaT) (Zaki 2000) is an algorithm that mines frequent itemsets efficiently using the vertical data format as shown in Table 3. In this method of data representation, all the transactions that contain a particular itemset are grouped into the same record. First, the EClaT algorithm transforms data from the horizontal format into the

Table 3 Transactional data in vertical data format. Reproduced with permission from (Han et al. 2012)

itemset	TID_set
I1	{T100, T400, T500, T900}
I2	{T100, T200, T300, T800, T900}
I3	{T300, T500, T600, T900}
I4	{T200, T400}
I5	{T100, T800}

Table 4 2-Itemsets in vertical data format. Reproduced with permission from (Han et al. 2012)

itemset	TID_set
{I1, I2}	{T100, T400,
{I1, I3}	{T500, T700,
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600,
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

Table 5 3-itemsets in vertical data format. Reproduced with permission from (Han et al. 2012)

itemset	T
{I1, I2, I3}	{
{I1, I2, I5}	{

vertical format by scanning the database once. The frequent (k + 1)-itemsets are generated by intersecting the transactions of the frequent k-itemsets. This process repeats until all the frequent itemsets are intersected with one another and no frequent itemsets can be found as shown in Tables 4 and 5. For the EClaT algorithm, the database is not required to be scanned multiple times in order to identify the (k + 1)-itemsets.

The database is only scanned once to transform data from the horizontal format into the vertical format. After scanning the database once, the (k + 1)- itemsets are discovered by just intersecting the k-itemsets with one another. Apart from this, the database is also not required to be scanned multiple times in order to identify the support count of every frequent itemset because the support count of every itemset is simply the total count of transactions that contain the particular itemset. However, the transactions involved in an itemset can be quite a lot, making it to take extensive memory space and processing time for intersecting the itemsets.

3.4.RFM Anyalysis:

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer lifetime value, and recency affects retention, a measure of engagement.

IV. DATASET

The data set used for our model is a collection of frequent itemset provide by the organizers of the super market basket.Marketbasketoptimization('../input/market.csv')groceries('../input/groceries- dataset/Groceries_dataset.csv'). The market basket optimization dataset to implement theap,fp-growth and Eclat. Another grocerises dataset to implement the RFM Analysis.

V. RESULT

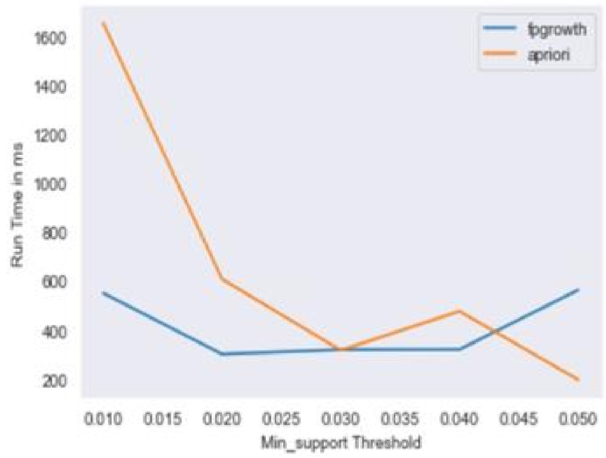
Figure 1 represent the support count of apriori.
Figure 2 represent the support count of fp- growth.
Figure 3 represent the comparision of apriori&fp-growth.
Figure 4 represent the support count of Eclat.
Figure 5 represent the Eclat algorithm.
Figure 6 represent the algorithm of RFM analysis.

	support	itemsets
0	0.087188	(burgers)
1	0.081056	(cake)

Fg:1

	support	itemsets
0	0.238368	(mineral water)
1	0.132116	(green tea)

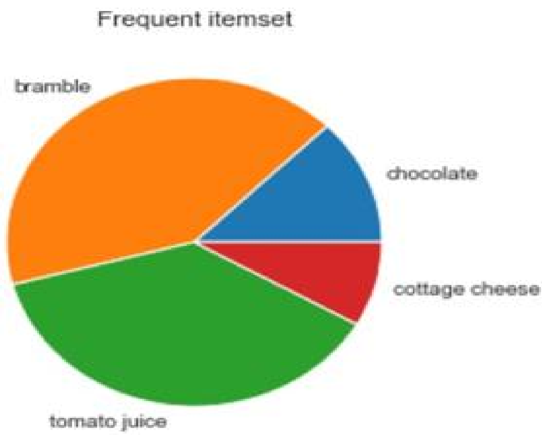
Fg:2



Fg:3

{'chocolate': 0.1638448206905746,
'mineral water': 0.23836821757099053,
'spaghetti': 0.17411011865084655,
'cookies': 0.08038928142914278,}

Fg:4



Fg:5:

Fg:6

VI. CONCLUSION

The knowledge discovery from rapidly growing data can be handled by data mining techniques. There are various data mining techniques like association rules, classification, clustering, correlations, sequential patterns and many more. In this research work the author concentrates on association rules mining algorithm for discovery of frequent patterns. The research work has been implemented data mining using juypter notebook.

In this thesis, we considered the following factors for creating our new scheme, which are the time and the memory consumption, these factors are affected by the approach for finding the frequent itemsets. Work has been done to develop an algorithm which is an improvement over Apriori and FP-tree with using an approach of improved Apriori and FP-Tre algorithm for a transactional database FP Tree algorithm for a transactional database

REFERENCES

[1] Z. Abdulla, T. Herawan, A. Norazia,M.M. Deris, A Scalable Algorithm for, 2014.

[2] Constructing Frequent Pattern Tree', International Journal of Intelligent Information Technologies, 10 (1), pp. 42-56.

[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceeding of the 20th International Conference of Very Large Databases (VLDB), Santigo, Chile, 1994, pp. 487–499.

[4] R.C. Agrawal, C.C. Agrawal, V.V.V. Prasad, Depth first generation of long patterns, in: Proceeding of the 6th International Conference on Knowledge Discovery and Data Mining, 2000, pp. 108– 118.

[5] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), Washington, USA, 1993, pp. 207–216. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1996) 'Fast Discovery of Association Rules', Published in Advances in Knowledge Discovery and Data Mining, AAAI Press, pp. 307–328.

[6] W. AlZoubi, An Improved Graph Based Method for Extracting Association Rules, 2015.

[7] International Journal of Software Engineering & Applications (IJSEA), 6(3), pp. 1–10 E. Ansari, G.H. Dastghaibifard, M. keshatkaran, Distributed trie frequent itemset mining, in: Proceedings of International Multi Conference of Engineers and Computer Scientists, 1, 2008, pp. 978– 988.

[8] A. Appice, M. Berardi, M. Ceci, D. Malerba, Mining and filtering multi-level spatial association rules with ARES, in: Proceedings in 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, 2005, pp. 342–353.

[9] J. Ayres, J. Gehrke, T. Yiu, J. Flannick, Sequential pattern mining using A bitmap representation, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, 2002, pp. 429–435.

[10] N. Balaji Raja, G. Balakrishnan, An improved algorithm of graph and clustering based association rule mining (GCBARM) in discovering of frequent itemsets, J. Manag. Comput. Sci. 6 (2) (2011) 6–10.

[11] D. Barbara, J. Couto, S. Jajodia, L. Popyack, N. Wu, ADAM: detecting intrusions by data mining, in: Proceedings of the IEEE Workshop on Information Assurance and Security Symposium, NDSS'00, 2001, pp. 157–170.