Conceptual Model of Incremental R-Eclat Algorithm for Infrequent Itemset Mining

Mustafa Man[1], Nurul Aqilah Ruslan[2], Julaily Aida Jusoh[3], Wan Aezwani Wan Abu Bakar[4]

[1,2]*Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu, Terengganu, Malaysia*
[3,4]*Faculty of Informatic & Computing, Universiti Sultan Mizan Zainal Abidin, Terengganu, Malaysia*

ABSTRACT

Mining valuable information from database could be very challenging especially for crucial decision-making. This is because mining association rule may require repetitious scanning of large databases that leads to the use of high memory usage and affects the running time. Few methods and algorithms were introduced by researchers to handle the issues in data mining. Rare Equivalence Class Transformation (R-Eclat) algorithm is one of the rule mining techniques using vertical format data repositories for infrequent pattern mining. The main operation in R-Eclat is intersecting tidset. Since the size of tidsets would affect the memory usage and its running time, more memory and time required for a bigger tidsets. Adopting to R-Eclat algorithm, a new incremental approach is introduced called Incremental Rare Equivalence Class Transformation (IR-Eclat). IR-Eclat specifically design for infrequent pattern mining and it is beneficial for dynamic database as the data is increasing in volume from time to time. In conjunction with big data explosion, the end users are at an advantage for the use of this incremental approach.

Keywords : *data mining, infrequent itemset mining, R-Eclat algorithm, incremental R-Eclat.*

I. INTRODUCTION

Data mining is a process of extracting hidden pattern from a large dataset to generate valuable information for decision-making purpose. There are various techniques used in data mining such as artificial intelligence (AI), statistical, and machine learning while the most common techniques that have been use recently including prediction, classification, decision tree, sequential patterns, clustering, and association rule [1]. Data mining is an integrative research field among database system, information system, intelligent system, and industries. Hence, the implementation of data mining would contribute many benefits in various field such as medical, e-commerce, and education.

Association rule is one of the most well-known technique in data mining and it helps in learning behavior, predicting events and making decisions from an abundance of data [2]. The problem of mining association rules was introduce in [3]. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of X ⇒ Y, where X and Y are sets of item. The rule indicates transactions in the database contain items in X that is also contain the items in Y. For example, 95% of customers who buy cereal also buy milk. That 95% is called the confidence of the rule while X ⇒ Y is the percentage of transactions that contain both X and Y is addressed as support. In this case, the challenge is to find all rules that satisfy a user-specified minimum support and minimum confidence [4]. Applications for association rule mining range from decision support to telecommunications alarm diagnosis and prediction [5]. This paper will discuss about the R-Eclat algorithm and introducing a new Incremental R-Eclat algorithm.

II. RELATED WORK

A. Infrequent Itemset Mining

Mining rare pattern from a database is not as popular as frequent ones. However, there are some areas where the rare pattern mining was found to be more important compared to frequent pattern mining [6]. Pattern mining concept was invented by [6] for frequent pattern mining. A set of items is defined as K = { …, }. The itemset K is frequent if and only if its frequency of occurrence in the database is equal to or greater than the user-defined minimum support threshold. In contrast to infrequent pattern mining, the frequency of occurrence in the database must be equal to or less than the user-defined minimum support threshold. Mining rare pattern using traditional frequent pattern mining method is ineffective if the user-defined threshold is set very low [6].

In recent years, there has been an increasing interest for mining infrequent pattern especially in medical and networking field. In network security line, the unfamiliar occurrences might indicate some network failure or security invasion [6]. As for medical field, the use of infrequent pattern mining might contribute in finding the medication of rare disease cases. The aim of infrequent pattern mining is to mine patterns which are not frequent that is the support value less than the threshold. However, the process of extracting rare patterns from the database is quite challenging. Based on recent studies, few methods have been implement for infrequent pattern mining which is Apriori [3], FP-Growth [6] [7], and the latest method is R-Eclat [8]. Each method use different approach and data structures though their resulting sets of rules are all the same. Fig. 1 shows the chronology of Data Mining.

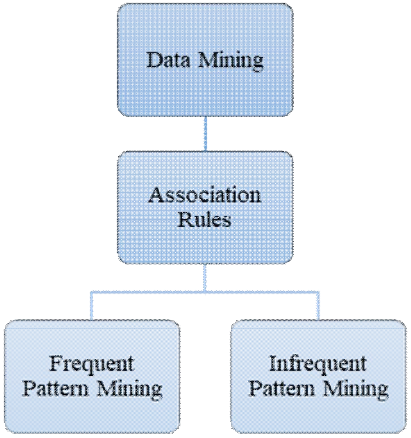


Figure 1. Chronology of Data Mining [8]

B. R-Eclat

Equivalent Class Transformation (Eclat) has been proposed by [9]. Eclat was the first algorithm using a vertical layout of the database combined with a depth-first traversal of the search space that is organized in a prefix-tree [9]. Eclat algorithm consist of four variants; tidset, diffset, sort-diffset, and post-diffset [8]. However, these variants is limited for mining frequent itemsets. By referring to Eclat algorithm, R-Eclat is proposed and the original algorithm has been modified to satisfy the infrequent pattern mining [8]. This algorithm utilizes column-based (vertical) instead of row-based (horizontal) to represent the dataset. It counts the support through determining support of any m-itemsets on the intersecting tidset lists of its m-1 subsets. Transaction id (tids) for each item is generated in the first scanning of database. In the beginning, the items are generated by single item (m=1), it is then increment by 1 generating (m+1)-itemsets. The (m+1)-itemset is generated from the former m-itemset with a depth-first calculation order. This process is done by intersecting tids of the m-itemsets to compute the tidsets of the corresponding (m+1)-itemsets. It is continuously calculated until there is no infrequent pattern appear.

The support for each itemset are count by undergo all transactions in the transaction database while it is inspecting if the transaction contains the related itemsets. R-Eclat is the implementation of intersection tidset, so the bigger the size of tidsets, the higher the time consumed and memory usage of R-Eclat to complete the process. R-Eclat approach consist of two pre-processing stages. At first, all the algorithms in association rule mining related vertical mining is checked to determine the suitable algorithm for itemset mining and the last part of pre-processing, an algorithm modification process has occurred. All four variants of Eclat algorithms are transformed to assure it is compatible for mining infrequent pattern. Fig. 2 shows the conceptual model of R-Eclat.

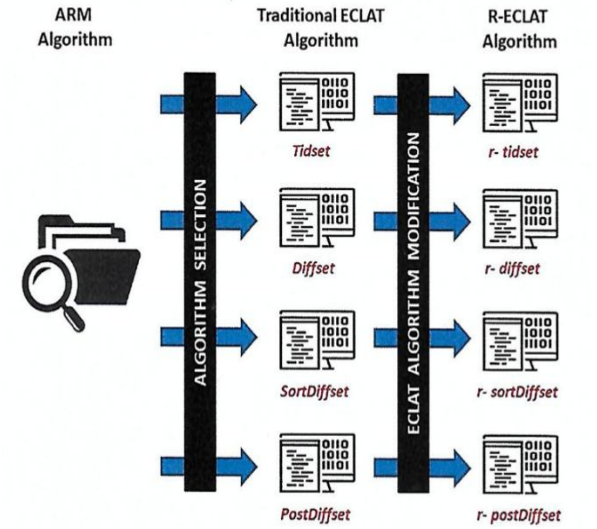


Figure 2. Conceptual Model of R-Eclat [10]

Since R-Eclat is designed based on previous Eclat algorithm that consist of four variants, it is also comes with four varieties; R-tidset, R-diffset, R-sortdiffset, and R-postdiffset. The advantage of R-Tidset is the size of tidsets represents the support and R-Tidset perform vertical intersection of tidlist. In R-Diffset, it only keeps track of differences in tidsets, thus make the intersection faster and less memory usage. In R-Sortdiffset, it is a combination of R-Tidset and R-Diffset, then tidset is sorted in ascending order while diffset is sorted in descending order. There is no need for switching condition, hence reduce its running time and memory usage. Last variant R-Postdiffset is the combination of R-Tidset and R-Diffset, which offer a better performance among the others. In this paper, authors will focus on R-tidset variant only. This would make it easier to understand the basic concept of R-Eclat as the algorithm is more straightforward compared to the other variants. However, the main drawback of R-Tidset is difficult in pruning technique as the longer the tidset, the more time and memory is needed. Fig. 3 illustrated the chronology of infrequent pattern mining.

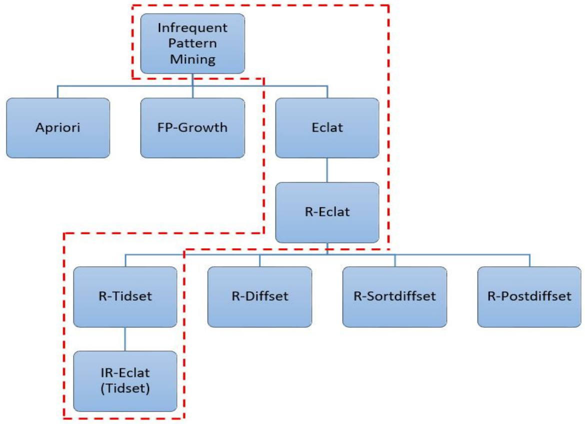
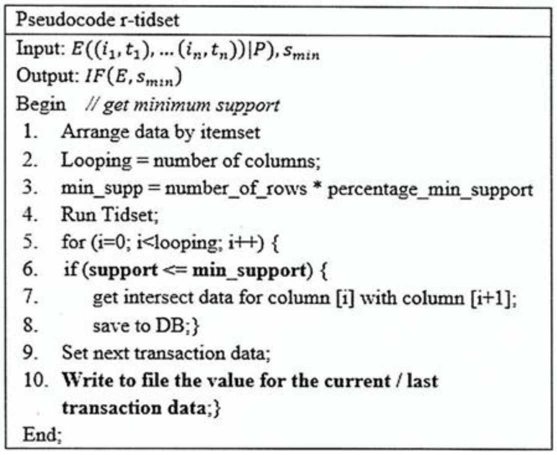


Figure 3. Chronology of Infrequent Pattern Mining

Pseudocode for r-tidset algorithms is presented in Fig. 4. The minimum support threshold value is considered as a benchmark to discover a low occurrence in each dataset. In [11], (1) used to calculate the minimum support, which is determined in terms of percentage.

$$\frac{\delta}{100} * \alpha \qquad (1)$$

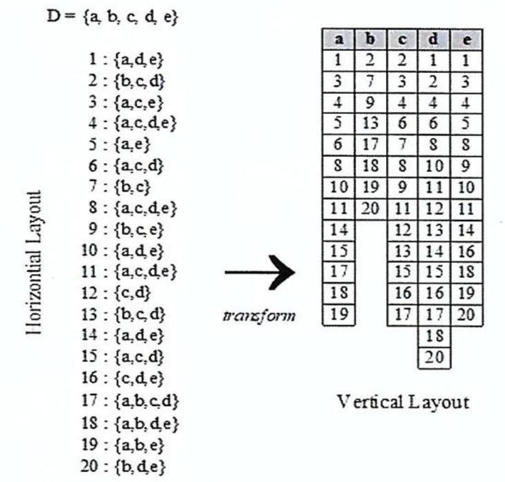
where: δ = User-specified minimum support value
α = Total of records in datasets



```
Pseudocode r-tidset
Input: $E((i_1,t_1),\ldots(i_n,t_n))|P), s_{min}$
Output: $IF(E, s_{min})$
Begin    // get minimum support
  Arrange data by itemset
  Looping = number of columns;
  min_supp = number_of_rows * percentage_min_support
  Run Tidset;
  for (i=0; i<looping; i++) {
   if (support <= min_support) {
      get intersect data for column [i] with column [i+1];
      save to DB;}
  Set next transaction data;
  Write to file the value for the current / last
  transaction data;}
End;
```

Figure 4. Pseudocode for R-Eclat (r-tidset) [8]

The R-Eclat algorithm is slightly different from previous Eclat algorithm. The main contrast between these two algorithms is the support counting that is less than or equal to minimum support threshold value. Another dissimilarity is the results are written in a text file, so that the process is ready for the next row of transaction data. The purpose is to reduce memory usage. Other processes remain unchanged from Eclat algorithm except for two key respects that has been mention earlier. In each loop, begin with the first loop, if the support is less than or equal (<=) to min_supp, then obtain the result of the intersection between i^{th}column and i^{th}+1 column and save the database. Transaction records transformed from horizontal into the vertical layout illustrated in Fig. 5. Based on Figure 5, the construction of association rules in r-tidset is illustrated in Fig. 6. Given minimum support is three (3).



D = {a, b, c, d, e}

Horizontal Layout

1 : {a,d,e}
2 : {b,c,d}
3 : {a,c,e}
4 : {a,c,d,e}
5 : {a,e}
6 : {a,c,d}
7 : {b,c}
8 : {a,c,d,e}
9 : {b,c,e}
10 : {a,d,e}
11 : {a,c,d,e}
12 : {c,d}
13 : {b,c,d}
14 : {a,d,e}
15 : {a,c,d}
16 : {c,d,e}
17 : {a,b,c,d}
18 : {a,b,d,e}
19 : {a,b,e}
20 : {b,d,e}

a	b	c	d	e
1	2	2	1	1
3	7	3	2	3
4	9	4	4	4
5	13	6	6	5
6	17	7	8	8
8	18	8	10	9
10	19	9	11	10
11	20	11	12	11
14		12	13	14
15		13	14	16
17		15	15	18
18		16	16	19
19		17	17	20
			18	
			20	

Vertical Layout

Figure 5. Transformation of the D Database from Horizontal into Vertical [8]

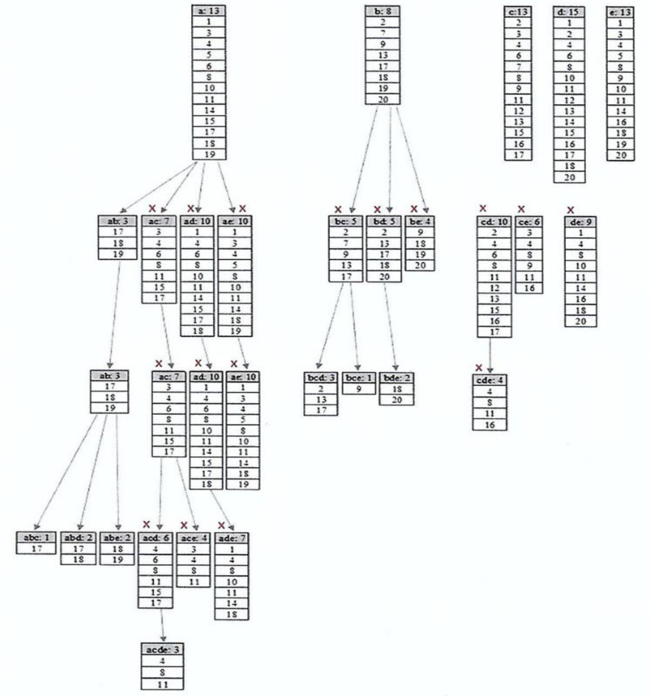


Figure 6. Construction of D Database in R-tidset Variant [8]

III. PROPOSE INCREMENTAL R-ECLAT MODEL

In frequent pattern mining, several algorithms have been implemented specifically for the incremental updating. The former incremental updating strategy was called Fast Update (FUP) [12] and it is intended to deal with new additional transaction data [13]. Another proposals introduced are ZIGZAG [14], FIIU [15], DFIIU [15], CATS Tree [16], CanTree [17], INUP-Tree [18] and the list goes on. However, there are less contribution for infrequent pattern mining on incremental updating.

R-Eclat could be benefit in mining infrequent pattern. This algorithm can be improved by introducing incremental approach that is called Incremental R-Eclat (IR-Eclat). Incremental R-Eclat aimed to minimize the memory and spaces requirement [19,23]. The incremental mining is used in either itemsets or records of transaction. Incremental itemsets means the new items are added into the existing itemsets in database while incremental in record of transaction means new transaction is added into the transaction database. Fig. 7 shows the Conceptual Model of IR-Eclat and pseudocode of the proposed method IR-Eclat is shown in Fig. 8.

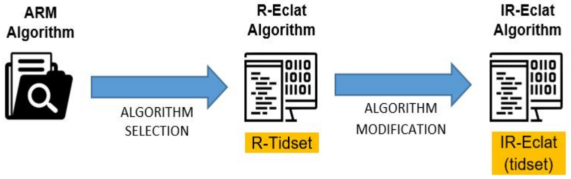


Figure 7. Conceptual Model of IR-Eclat

```
IR-Eclat Pseudocode
Input: $E((i_1, t_1), \dots (i_n, t_n))|P), s_{min}$
Output: $IF(E, s_{min})$

start        //get minimum support
sort data by itemset
looping = num_of_column
r = record_of_transactions
min_supp = num_of_row * percentage_min_supp
run tidset
while r ≠ 0 do        //process itemset by batch
for (i=0; i<looping; i++)
        if (support<=min_support)
                get intersect data for column [i] with column [i+1];
                save to db
        add next transaction data;
        flush value for current/last transaction data;
end
```

Figure 8. Incremental R-Eclat Pseudocode

The beginning steps for IR-Eclat is still same with R-Eclat algorithm. Then, process the dataset by batch. For example, there are 10000 records in a dataset, note that the increment value of records to be processed is 2000 records per transaction. So the first 2000 records is process in first transaction, followed by the next 2000 records until all records in the dataset is processed. For each loop, begin with the first loop, if the support is less than or equal (<=) to min_supp, then get the result of intersection between i^{th} column and i^{th}+1 column and save results to database. Next transaction data will be added and current or last transaction data is flushed before the next transaction data takes place

IV. CONCLUSION

Mining rare pattern remains as one of important parts in data mining. It has received great attention among researchers in developing most efficient method for mining infrequent pattern. In this paper, we have discuss about the R-Eclat and a new Incremental R-Eclat algorithm is proposed to enhance the current R-Eclat algorithm. As mentioned before, the incremental approach is beneficial for dynamic database with subject to new record of transactions or new items being added to the database. It is shown that main operation of eclat is intersection between tidset, so the size of tidset will affect its running time and memory usage. The bigger the tidset, the bigger the memory usage and the longer the execution time. In IR-Eclat, since the dataset is process by batch, it is believe that the execution time will be decrease as memory requirement to process each transaction is more considerable compared to previous r-eclat which records are processed all at one time.

ACKNOWLEDGEMENT

We would like to thank our financial support, Fundamental Research Grant Scheme (FRGS) under Ministry of Education Malaysia for grant with Vot: 59543 A New Parallel R-Eclat Algorithm for Infrequent Pattern Mining. We would also like to show our gratitude to the faculty members and grant collaborators of UMT who reviewed our research and provided meaningful comments and suggestions that greatly assisted the research.

REFERENCES

[1] Data mining, available at http://www.zentut.com/data-mining/data-mining-processes, Retrieved on 6/7/2019.

[2] Adhikary, D., Roy, S. "*Issues in Quantitative Association Rule Mining: A Big Data Perspective*." Book Issues in Quantitative Association Rule Mining: A Big Data Perspective, Springer Singapore. (2016): 377-385.

[3] R. Agrawal, T. Imielinski, A. Swami. "*Mining Association Rules Between Sets of Items in Large Databases*." Proc. of the ACM SIGMOD Conference on Management of Data. Washington, D.C. (1993): 207-216.

[4] Srikant, R., Agraval, R. "*Mining Generalized Association Rules*." Proc. of the 21st VLDB Conference. Zurich, Swizerland. (1995): 407-419.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.). "*Advances in Knowledge Discovery and Data Mining*." AAAI. Menlo Park, California. (1996).

[6] Borah, A., & Nath. "*Rare Pattern Mining: Challenge and Future Perspectives*." Springer International Publishing. (2018).

[7] Han J, Pei J, Yin Y. "*Mining Frequent Patterns Without Candidate Generation*." ACM SIGMOD Record. (2000). 29(2): 1-12.

[8] J. A. Jusoh. "*A New R-Eclat Algorithm for Infrequent Itemset Mining*." PhD Thesis. Malaysia: Universiti Malaysia Terengganu. (2019).

[9] Zaki M, Parthasarathy S, Ogihara M, Li W. "*New Algorithms for Fast Discovery of Association Rules*." Proceedings of the 3rd International Conference on Knowledge Discovery in Databases. (1997). 283-286.

[10] M. Man, J. A. Jusoh, S. I. A. Saany, W. A. W. A. Bakar. "*Analysis Study on R-ECLAT Algorithm in Mining Infrequent Itemsets*." The Conference on Mathematics, Informatics and Statistics (CMIS2018). Kuala Terengganu, Malaysia. (2018).

[11] J. A. Jusoh, M. Man, W. A. W. A. Bakar. "*Performance of R-Eclat Algorithm in Large Dataset*." International Journal of Engineering & Technology. (2018). 7(4.1): 134-137.

[12] Cheung D, Han J, Ng V, Wong C. Y. "*Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique*." Proceeding of the 12th Intl. Conf. on Data Engineering. (1996).

[13] R. Hernandez, J. hernandez, J. A. Carraso-Ochoa, J. Fco. Martinez-Trinidad. "*A Novel Incremental Algorithm for Frequent Itemsets Mining in Dynamic Datasets*". CIARP, Springer Verlag Berlin Heidelberg. (2008): 145-152.

[14] Veloso A, Meira Jr W, de Carvalho M. B, Possas B, Parthasarathy S, Zaki M. "*Mining Frequent Itemsets in Evolving Databases*." Proceedings of the 2nd SIAM Intl. Conf. on Data Mining. Arlington, USA. (2002).

[15] Veloso A, Gusmao W, Meira Jr W, de Carvalho M. B, Parthasarathy S, Zaki M. "*Parallel, Incremental and Interactive Mining for Frequent Itemsets in Evolving Databases*." Intl. Workshop on High Performance Data Mining: Pervasive and Data Stream Mining. (2003).

[16] Cheung W, Zaiane O. R. "*Incremental mining of frequent patterns without candidate generation or support constraint.*" Proceedings of the Seventh IEEE International Database Engineering and Applications Symposium. (2003): 111–116.

[17] Leung C. K, Quamrul I. K, Hoque T. "*CanTree: A Tree Structure for Efficient Incremental Mining of Frequent Patterns*." Proceedings of the Fifth IEEE International Conference on Data Mining. (2005).

[18] Hai T. H, Shi L. Z. "*A New Method for Incremental Updating Frequent Patterns Mining*." Proceedings of the Second International Conference on Innovative Computing, Information and Control. (2007).

[19] W. A. W. A. Bakar, Z. Abdullah, M. Y. M. Saman, M. A. Jalil, M. Man, T. Herawan, A. R. Hamdan. "*Incremental-Eclat Model: An Implementation via Benchmark Case Study*." Springer International Publishing Switzerland. (2016): 35-46.

[20] G. Sophana, V. Joseph. **"***Infrequent Pattern Mining Techniques: A Review*." International Journal of Engineering Research in Computer Science and Engineering. (2018): 5(4).

[21] W. A. W. A. Bakar. "*An Enhanced Eclat Algorithm Based on Incremental Approach for Frequent Itemset Mining*." PhD Thesis. Malaysia: Universiti Malaysia Terengganu. (2015).

[22] J. A. Jusoh, M. Man. "*Modifying iEclat Algorithm for Infrequent Patterns Mining*." Advanced Science Letters. (2018): 24 (3).

[23] Naveed, Q. N., Qureshi, M. R. N., Tairan, N., Mohammad, A., Shaikh, A., Alsayed, A. O., ... & Alotaibi, F. M. (2020). "*Evaluating critical success factors in implementing E-learning system using multi-criteria decision-making***".** Plos one, *15*(5), e0231465.