International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 10 | Oct -2016 www.irjet.net p-ISSN: 2395-0072

Association Rule Mining Algorithms: Survey

Subhash Rohit
Dept. Information Technology
Parul Institute of Engineering & Technology, Limda
Email: subhashrohit10693@gmail.com

Abstract: *In this paper, the detail description of survey focus on various association algorithms. As we know, Data mining is the process of discovering useful, hidden and understandable pattern in terms of information from large amount of data. It has many techniques for discovering the information like Association rule mining, classification, clustering, regression etc. Among them Association rule mining is one of the most important research area in data mining. In past research, many algorithms were developed like Apriori, Fp-Growth, Eclat, Bi-Eclat etc. In this paper we discuss this algorithms in detail.*

Keyword: *Data Mining, Association rule mining, Apriori, Eclat, FP-Growth, Bi-Eclat.*

1. INTRODUCTION

Now a days, industry, commercial product, computerise of many business and government's transactional scientific data has been collected threw large amount of data every day. So the data mining provides the tools to discover knowledge from the data. Data mining is also known as KDD (Knowledge Discovery from Data) [1]. Knowledge Discovery from data means process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases. There are so many other words are used to same article and document like knowledge mining from data base, data archaeology, data analysis, knowledge extraction, data grudging etc. KDD is mainly divided in three processes: Pre-processing, Data mining Process and Post processing [12]. Pre-processing includes data selecting, data cleaning, data integration, and data transformation. Data mining process includes different Algorithms and find hidden knowledge. Post processing, which includes finds the result according to user's requirement and domain knowledge [12]. There are lots of data mining tasks like Association rule mining, regression, clustering, classification etc. [2]. So Among this the existing task of Association rule mining is the most interesting research area in data mining.

2. ASSOCIATION RULE MINING

Association rule mining is generally used to extract the interesting correlation, frequent pattern, association or casual structure among sets of items in database [12] [4]. Association rule mining consist of two main measure: support and confidence [3] [4]. Support can be defined based on minimum value or more than minimum value. Confidence can be defined how many time the number of statement has become true. Association rule mining finds those rules which satisfies the minimum support and confidence. Association rule is an implication form let I= {I1, I2, I3,.In} be a set of item that contain in database and each item have unique transection id T where T ⊂□□. Transection containing set of item like A, B. A→B, where A and B are the antecedent and consequent respectively and X ∩ Y = ϕ.

Association rule mining is a two-step process: First step is to find frequent item that is less than minimum support. Second step is to combine all frequent item [3]. Association rule mining is the technique to knows the human behaviour e.g. National basketball Association is the best example of the association. In the National basketball association (NBA) the team worker capture the motion of the player like up, down, top, left, right etc and store in the database. And that pattern are used to train new player. Now another and most popular example of association rule mining is market basket analysis. Market basket analysis is the process to find the habit of the customer. Market basket analysis is very useful term for improving market strategy. It also defines to store layouts. Like if one item is purchased then it encourage the people to buy other item. For example if any customer by laptop then he will also buy mouse and keyboard at same time. So place laptop near mouse and keyboard, so whenever customer buy laptop then same time can buy mouse and keyboard [5].

3. Apriori Algorithm

Apriori is a one of the famous, most important, and scalable algorithm for mining frequent itemsets and association rule mining. Apriori was introduced by Agrawal and Srikant in 1993 [6] [10]. Apriori algorithm is used to find all frequent itemsets in a given database. Apriori algorithm is to make multiple scan over database. It uses breath first strategy to search over items in the database [10]. Apriori algorithm is being used by so many industry for transactional operation and also it can be used in real time applications like (shopping mall, general store, grocery shop etc.) by collecting the item bought by customer over the time so

frequent item can be generated. Apriori algorithm require two important thing: minimum support and minimum confidence. First we can checks the item whether they are greater than or equal to minimum support and after than we can finds the frequent item set. Second thing is minimum confidence constraint is used to form association rules [7].
Let see the example of Apriori algorithm:

Original Dataset.

T_{ID}	Items
T100	A, C
T200	B, C, A
T300	A, B, C, D
T400	D, C, E
T500	F, A, B

Step1: Count the Number of transaction in which each item occur. Like A Occur in 4 Time

Item	Count
A	4
B	3
C	4
D	2
E	1
F	1

Step2: Remove the item that are not satisfied Minimum Support rule.

Item	Count
A	4
B	3
C	4
D	2

This is the single item that are bought frequently. Now let's say we want to find a pair of item that are bought frequently.

Step3: We start making two pairs from dataset.

Item
A, B
A, C
A, D
B, C
B, D
C, D

Step4: Now Count how many time each pair is bought to gather.

Item	Count
A,B	3
A, C	3
A, D	1
B, C	2
B, D	1
C, D	2

Step5: Remove item pairs with not satisfied the minimum support condition

Item	Count
A,B	3
A, C	3
B, C	2
C, D	2

These are the pairs of items frequently bought together.

Now let's say we want to find a set of three item that are brought together.

Step6: To making the three pair from the dataset.

Item	Count
A,B, C	2
A, B, D	1
A, C, D	1
B, C, D	1

Step7: Remove the item that are not satisfied minimum support.

Item	Count
A,B, C	2

So the final result are A, B, C it means A, B, C are frequently bought together.

Now Generate Strong Association rule from frequent item set which satisfied minimum support.
Strong rule: A→B, C it means when a bought then it will also bought B and C.

Disadvantages of Apriori algorithm are it require multiple database scan and additionally generate many candidate set. It is generally costly to handle huge number of candidate set.

4. FP- Growth Algorithm

FP-Growth algorithm introduced by Hanet in 2000 [8]. It covers the drawbacks of Apriori algorithm. FP-Growth algorithm is one of the fastest algorithm in the association rule mining. It generate frequent item and does not create huge amount of candidate itemsets. FP-Growth algorithm is two-step process: First step is Scanning the database twice. Now in the first step to scan the database and find the support count of each item and infrequent item are deleted from the list and remaining item are store in descending order. In the second step to construct FP tree using frequent tem [8].
FP- Growth algorithm has three important unprejudiced: First the database is scanned only two time and computational cost is decreased dramatically. Second thing is that no candidate itemset are generated. Third and last objective is to use the divide and conquer approach which consequently reduce the search space. FP tree consist of two main property, first is the node link property and second is the prefix path property [8].
Let see the example of FP-Growth algorithm as below:

Original Dataset:

T_{ID}	Items
T100	A, C
T200	B, C, A
T300	A, B, C, D
T400	D, C, E
T500	F, A, B

Step1: Scan the Items and Count their Occurrence.

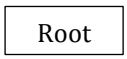
Item	Support
A	4
B	3

C	4
D	2
E	1
F	1

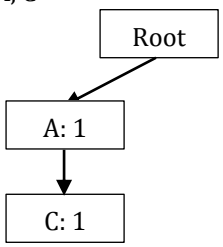
Step2: Set Items in Descending order.

T_{ID}	Item	Order List
T100	A, C	A, C
T200	B, C, A	A, C, B
T300	A, B, C, D	A, C, B, D
T400	D, C, E	C, D
T500	F, A, B	A, B

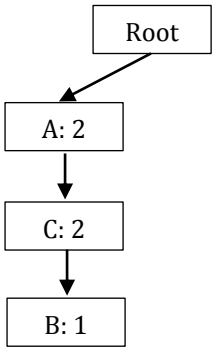
Step3: Construct FP Tree: First Create Root Node:



- A, C→

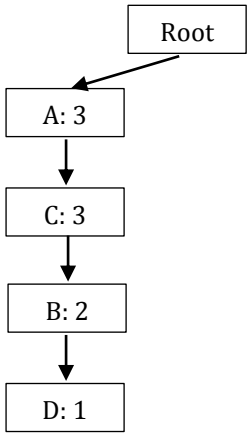


- A, C, B

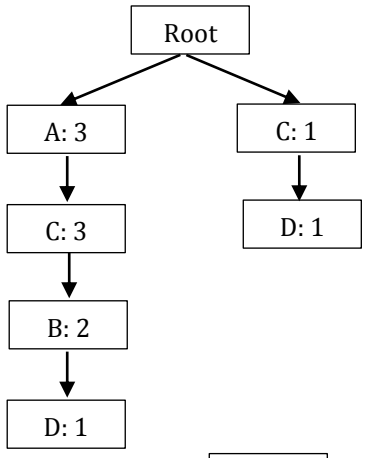


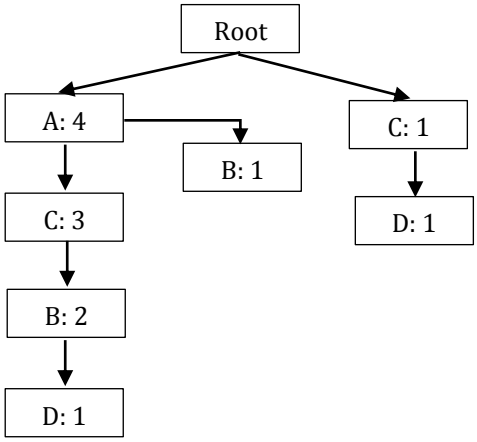
Note: We adds the Items so increment the count of Item.

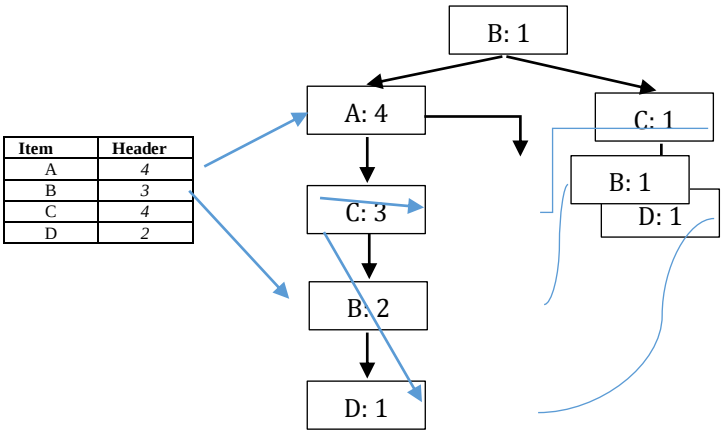
- A, C, B, D



- C, D



- 



Drawback of FP- Growth algorithm: It is difficult to use in incremental mining, as new transections are added to the database, FP tree need to be updated and the whole process need to repeat [8]. Execution time is large due to complex compact data structure.

5. Eclat Algorithm

Eclat algorithm introduced by Mohammed Javeed Zaki. It uses vertical database layout. It cannot uses horizontal dataset. If there is any horizontal dataset, then we need to convert into vertical dataset. There is no need to scan dataset again and again. Eclat algorithm only consider support [5]. The Eclat algorithm has divide in four phase.

The first phase is initialize the phase, second phase is transformation, third phase is asynchronous phase and final phase is reduction phase [11]. Eclat algorithm does not takes the extra computation overhead of building or searching complex data structures, nor does it have to generate all the subsets of each transaction [11]. Eclat algorithm uses depth first search property. All frequent Itemsets can be computed with intersection of Transaction ID list. In the first scan pass of database a Transaction ID list is maintained for each single item. k+1 Itemset can be generated from k Itemset using Apriori property and depth first search computation. (k+1) Itemset is generated by taking intersection of Transaction ID set of frequent k-Itemset. This process continues, until no candidate Itemset can be found [8].It does not scan the original database that is the main advantage of this algorithm. It only counts the current generated database.

Let see the Example of Eclat algorithm:

Original dataset.

T_{ID}	Items
T100	A, C
T200	B, C, A
T300	A, B, C, D
T400	D, C, E
T500	F, A, B

Step1: Transfer Horizontal layout to vertical layout

Item	T_{ID}
A	T100, T200, T300, T500
B	T200, T300, T500
C	T100, T200, T300, T400
D	T300, T400
E	T400
F	T500

Step2: Scan the dataset and remove those items who does not satisfied the minimum Support Condition. For example E and F have only one support count so they are remove from the database.

Item	T_{ID}
A	T100, T200, T300, T500
B	T200, T300, T500
C	T100, T200, T300, T400
D	T300, T400

Step3: Make the two pairs of the itemset.

Item	T_{ID}
A,B	T200, T300, T500
A, C	T100, T200, T300
A, D	T300
B, C	T200, T300
B, D	T300
C, D	T300, T400

Note: Eclat algorithm only focus on the current generated dataset (follow the step 2 dataset). They cannot scan the original dataset (Original dataset.) like Apriori.

Step4: Remove those Items whose occurrence is less than the Minimum Support.

Item	T_{ID}
A,B	T200, T300, T500
A, C	T100, T200, T300
B, C	T200, T300
C, D	T300, T400

Step5: Now makes the three item set pairs.

Item	T_{ID}
A,B, C	T200, T300
A, B, D	T300
A, C, D	T300
B, C, D	T300

Step6: Remove those items whose count is less than the Minimum Support.

Item	T_{ID}
A,B, C	T200, T300

The computation in eclat algorithm is done by intersection of the tid sets of the corresponding k itemsets that is the main drawback of this algorithm. The transection Id sets can be quite large, which takes preferable memory space as well as computation time for intersecting the large sets, thus the total stress available in main memory can increase the length in frequent itemsets [9].

Eclat algorithm does not take full advantage of Apriori property in reducing the number of candidate itemsets explored during frequent itemset generation. Therefore, the number of candidates generated in Eclat is much greater than that in Apriori algorithm. The situation gets worse as the data set consisting many dense and large itemsets [9].

6. Bi- Eclat Algorithm

Bi-Eclat algorithm introduced by Xiamomei at 2014. This algorithm is the up gradation of Eclat algorithm. It follows so many Eclat property like depth first search etc. Bi Eclat algorithm is mainly divided in three phase: First phase is to represent the database in vertical format and sort the transaction id list in descending order of item. Second part is to Prune the item and generate candidate frequent item. Third and final phase is to mine the probabilistic frequent item in candidate database [9]. Bi- Eclat algorithm is first transfer the Horizontal to Vertical layout, scan the database and store all item in descending order. Again it will scan the item bottom to top for removing the item less than minimum support. It will not scan the whole database. That is the big Advantage of Bi- Eclat algorithm and it will reduce the time and scanning of the algorithm.

Let see the example of the Bi-Eclat algorithm:

Original Dataset

T_{ID}	Items
T100	A, C
T200	B, C, A
T300	A, B, C, D
T400	D, C, E
T500	F, A, B

Step1: transfer the Horizontal Layout to Vertical Lay Out and store in Descending order.

Item	T_{ID}
A	T100, T200, T300, T500
C	T100, T200, T300, T400
B	T200, T300, T500
D	T300, T400
E	T400
F	T500

Step2: Scan Bottom to Top and remove the item less than Minimum Support.

Item	T_{ID}
A	T100, T200, T300, T500
C	T100, T200, T300, T400
B	T200, T300, T500
D	T300, T400

Note: If we find the support of item equals to the minimum support than it will not scan above presented item sets. For example if Item D support is equal to the minimum support than it will stop the scanning of the database. So it will reduce the scanning and time.

Step3: Make the two pair of the items and store in descending order.

Item	T_{ID}
A, B	T200, T300. T500
A, C	T100, T200. T300
B, C	T200, T300
C, D	T300, T400
A, D	T300
B, D	T300

Step4: Scan Bottom to Top and remove the item less than Minimum Support.

Item	T_{ID}
A, B	T200, T300. T500
A, C	T100, T200. T300
B, C	T200, T300
C, D	T300, T400

Step5: Makes the three pair of the item and store in descending order

Item	T_{ID}
A, B, C	T300, T200
A, B, D	T300
A, C, D	T300
B, C, D	T300

Step6: Scan Bottom to Top and remove the item less than Minimum Support

Item	T_{ID}
A, B, C	T300, T200

Final Strong rule is A, B, C
This Bi- Eclat algorithm consume more memory and do not handle massive dataset that is the main drawback of this algorithm.

Table 1: Comparative Analysis of Association rule mining algorithm

	Apriori	FP-Growth	Eclat	Bi-Eclat
Technique	Breath First search Technique	Divide and conquer technique	Depth first Search Technique	Depth first Search Technique
Database Scan	Original database scan every time a frequent item set is generated	Database is scan two time only	Database is scan few time.	Database is scan few time and item are store in descending order
Time	Execution time take more.	Execution time take less compare to Apriori.	Execution time take less compare to FP- Growth and Apriori	Execution time take less compare to Eclat
Storage Structure	Array	Tree	Array	Array
Data Format	Horizontal	Horizontal	Vertical	Vertical
Advantage	Apriori suitable for large dataset.	No candidate set are generation and Database scan only two time.	No need to scan original database. Support can be find current generated database	Reduce the scanning of the database.
Disadvantage	Scan original database again and again. Require more memory space	FP tree is expensive to build Consume more memory.	Not suitable for massive dataset. Consume more memory	Not scanning all dataset. Not suitable for large dataset.

7. CONCLUSION

We have comparatively analysed varies association rule mining algorithm like Apriori, FP- Growth, Eclat, Bi- Eclat etc. We also have compared these algorithm using same examples to understand their working. Major issue of all algorithms are execution time and it consumes more memory and we found this Bi-Eclat algorithm in well manner as compared to other algorithm.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jain Pei, "Data Mining Concept and Technique 3rd Edition".

[2] A.H.M. Sajedul Hoque, Sujit Kumar Mondal, Tassnim Manami Zaman, Dr. Paresh Chandra Barman, Dr. Md. AI·Amin Bhuiyan, "Implication of Association Rules Employing FP·Growth Algorithm for Knowledge Discovery", IEEE 14th International Conference on Computer and Information Technology (ICCIT 2011) 22-24 December, 2011, Dhaka, Bangladesh.

[3] Jugendra Dongre, Gend Lal Prajapati, S. V. Tokekar, "The Role of Apriori Algorithm for Finding the Association Rules in Data Mining", IEEE 2014.

[4] Surbhi K. Solanki, Jalpa T. Patel, "A Survey on Association Rule Mining", IEEE Fifth International Conference on Advanced Computing & Communication Technologies 2015.

[5] Manjit kaur, Urvashi Grag, "ECLAT Algorithm for frequent Itemsets Generation", International Journal of Computer Systems Volume 01– Issue 03, December, 2014.

[6] Sallam Osman Fageeri, Rohiza Ahmad, Baharum B. Baharudin, "A Semi-Apriori Algorithm for Discovering the Frequent Itemsets", IEEE 2014.

[7] Akshita Bhandaria, Ashutosh Guptaa, Debasis Dasa, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining",Elsevier, International Conference on Information and Communication Technologies -2014.

[8] Shamila Nasreen, Muhammad Awais Azam, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey", Elsevier, The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks-2014.

[9] Xiaomei Yu, Hong Wang, Xiangwei Zheng, "A Bidirectional Process Algorithm for Mining Probabilistic Frequent Itemsets", IEEE, Ninth International Conference on Broadband and Wireless Computing, Communication and Applications - 2014

[10] Sanjeev Rao, Priyanka Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm", IJCST 2012.

[11] Mohammed Javeed Zaki, Srinivasan Parthasarathy, and Wei Li, "A Localized Algorithm for Parallel Association Mining", ACM 1997.

[12] Qiankun Zhao, Sourav S. Bhowmick, "Association Rule Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, 2003.

BIOGRAPHIES

Mr. Subhash T. Rohit received the Bachelor of Education in Computer Science& Engineering at Gujarat Technological University in 2014 and Master degree on Information Technology at Gujarat Technological University in 2016.