A Comparative Study of Association Mining Algorithms for Market Basket Analysis

Ishwari Joshi[1], Priya Khanna[2] , Minal Sabale[3], Nikita Tathawade[4]
RMD Sinhgad School of Engineering, SPPU
Pune, India
Under Guidance of Prof. Sweta Kale , HOD IT Department ,RMD Sinhgad School of Engineering, SPPU
Pune, India

Abstract

Association Rule Mining (ARM) aims to identify the purchasing patterns of customer. The purpose is to discover the concurrence association among data in large database & to discover interesting association between attributes in databases. The main aspect of ARM is to find frequent item set generation & Association Rule generation. In this paper we concentrate on frequent pattern mining Algorithms. This research paper discusses the comparison between three minig Algorithms i.e. Apriori Algorithm, Eclat Algorithm, and Improved Apriori Algorithm. It also focuses on advantages & disadvantages of these algorithms. The comparison is done w.r.t Market Basket Analysis using Hadoop. Mining of association rules from frequent pattern mining from massive collection of data is of interest for many industries which can provide guidance in decision making process such as cross marketing or arrangement of item in Stores & Supermarkets.

Keywords –*Apriori Algorithm, Eclat Algorithm, Frequent pattern mining, Improved Apriori Algorithm, Market Basket Analysis using Hadoop.*

I. INTRODUCTION

For discovering interesting relation between various variables in huge databases, association rule learning is used. Association rule is an if-then rule which is supported by data. In addition to many other diverse fields, ARM is used widely in marketing & retail communities also. This technique is also referred as Market Basket Analysis. Technically, association rule concept is looking for relationship between each item with another item. For example, a set of items such as Milk & Bread that appear together frequently in a transaction is a frequent item set. Frequent item set is an item set that appears frequently in transaction data. If pattern occurs frequently it is called as frequent pattern. Mining associations is all about finding such frequent pattern thus playing an important role in mining associations, correlation & many other interesting relationships among items of the dataset. This analyzes customer buying habits by finding association rule between different items in customer's "Shopping Basket"[6]. Itemset is a set of items, group of elements that are represented together as a single entity. In frequent pattern mining, to check whether the itemset occurs frequently or not, we can use a parameter called 'support' of an itemset. If an itemsets support count is greater than minimum support count setup initially then that itemset is called as Frequent Itemset.

II. RELATED WORK

Market Basket Analysis is one of the data mining techniques which is used for discovering the system for relation between one item to another item. The positive percentage of combined item use new knowledge i.e. it is very useful for determining decision[2]. The association rule is divided into two steps.

1)To find frequent item set and
2) To association from itemsets.

For import relationship association rule uses the criteria of "Support" and "Confidence".

a) Support :

Support for item set X is number of transactions covered by itemset X. Support is use to find frequent itemset in transactions'. Support is already decided pre-assumed value called as S_0.[3]Item set which has support greater than equal to S0 will be consider as frequent item set.
Formula for Support for one item set:

$$S(A\ ) = \frac{\text{Amount of transaction A}}{\text{Total Transactions}}$$

Formula for support of two item set:

$$S(A\ U\ B) = \frac{\text{Amount of transaction A \& } B}{\text{Total Transaction}}$$

b) Confidence:

The confidence of rule indicates probability of both predecessor and successor appearing in same transaction[3]. Confidence can be expressed in probability notation as follows:
Confidence(A implies B)=Probability(B/A)

$$ie. = \frac{\text{Probability of (A, B)}}{\text{probability of A}}$$

Consider association rule,
If A then B then
Confidence is given by:

$$Confidence = \frac{\text{Support (A U B)}}{\text{Support A}}$$

Where,
Support A= no of transactions covered by A.
Support (AUB)= no of transactions covered by (AUB).

III. ANALYSIS

There are various techniques proposed for generating frequent itemsets so that association rules are mined efficiently. The process of generating frequent itemsets are divided into basic three techniques[2].

1. Apriori Algorithm : Horizontal Layout based
2. Eclat Algorithm : Vertical Layout based
3. Improved Apriori Algorithm : Horizontal layout based

For analysis let us consider a Dataset D.

Transaction ID	List of Items
1	Apples, Milk, Beer
2	Milk, Crisps
3	Milk, Bread
4	Apple, Milk, Crisps
5	Apple, Bread
6	Milk, Bread
7	Apple, Bread
8	Apple, Milk, Bread, Beer
9	Apple, Milk, Bread

Table no. : 1

Association Rule : An association rule is an important implication expression of X→Y, where X & Y are disjoint itemsets ,i.e. X ∩Y=ϕ.The strength of an association rule can be measured in terms of its Support & Confidence.

1. Apriori Algorithm

Apriori Algorithm is best known Algorithm for association rule mining from transactional databases. Apriori Algorithm is popular as well as easy data Mining Algorithm technique. The basic idea of Apriori is to generate candidate itemset of a particular size & then scan the database to count there to see if they are large. Now, Let us assume minimum Support = 2.The Database is scanned i.e. frequency of each item occurring in dataset is counted

C_k = k-itemset candidates
L_k= frequent k itemset, i.e support count >= minimum support count
Therefore, C_1 & L_1 is as follows :

Itemset	Support
{Apple}	6
{Milk}	7
{Bread}	6
{Crisps}	2
{Beer}	2

Table no. : 2

As all the Item set have minimum support >=2, we will combine the itemsets in pairs &scan .
Therefore, for 2-itemset C_2 is as follows :

Itemset	Support
{Apple, Milk}	4
{Apple, Bread}	4
{Apple, Crisps}	1
{Apple, Beer}	2
{Milk, Bread}	5
{Milk, Crisps}	2
{Milk, Beer}	2
{Bread, Crisps}	0
{Bread, Beer}	1
{Crisps, Beer}	0

Table no. : 3

Now, Compare the support value with minimum support value (2)& discard the itemset having minimum support value < 2
Therefore, for 2-itemset $L_2$is as follows :

Itemset	Support
{Apple, Milk}	4
{Apple, Bread}	4
{Apple, Beer}	2
{Milk, Bread}	4
{Milk, Crisps}	2
{Milk, Beer}	2

Table no. : 4

Like 2-Frequent itemset , now we will generate 3-Frequent itemset.
Therefore, for 3-itemset C_3 & L_3 is as follows :

Itemset	Support
{Apple, Milk, Bread}	2
{Apple, Milk, Beer}	2

Table no. : 5

Therefore , from this we obtain the 2 frequently purchased itemsets that are Apple, Milk, Bread & Apple, Milk, Beer

2. Eclat Algorithm

Eclat algorithm is a depth first search based algorithm. It uses a vertical database layout i.e. instead of explicitly listing all transactions; each item is stored together with its cover and uses the intersection based approach to compute the support of an itemset less space than Apriori if itemsets are small in number[2] .It is suitable for small datasets and requires less time for frequent pattern generation than Apriori. Below is the example of Eclat Algorithm for minimum support = 2.

Itemset	Transaction ID set
Apple	{1,4,5,7,8,9}
Milk	{1,2,3,4,6,8,9}
Bread	{3,5,6,7,8,9}
Crisps	{2,4}
Beer	{1,8}

Table no. : 6

2-Frequent itemset
Therefore, C_2& L_2 is as follows :

Itemset	Transaction ID set
{Apple ,Milk}	{1,4,8,9}
{Apple, Bread}	{5,7,8,9}
{Apple, Crisps}	{4}
{Apple, Beer}	{1,8}
{Milk, Bread}	{3,6,8,9}
{Milk, Crisps}	{2,4}
{Milk, Beer}	{1,8}
{Bread, Beer}	{8}

Table no. : 7

Likewise, for 3-itemset C_3 & L_3 is as follows :

Itemset	Transaction ID set
{Apple, Milk, Bread}	{8,9}
{Apple, Milk, Beer}	{1,8}

Table no. : 8

Therefore ,from this we obtain the 2 frequently purchased itemsets that are Apple, Milk, Bread & Apple, Milk, Beer

3. Improved Apriori Algorithm

Let us consider the same database D.

Let us assume minimum support as 2.

For 1-Frequent itemset L_1

Firstly, scan all transactions to get frequent 1-itemset L1 which contains the items and their support count and the transactions ids that contain these items, and then eliminate the candidates that are infrequent or their support are less than the minimum support.

Itemset	Support	Transaction ID set
Apple	6	{1,4,5,7,8,9}
Milk	7	{1,2,3,4,6,8,9}
Bread	6	{3,5,6,7,8,9}
Crisps	2	{2,4}
Beer	2	{1,8}

Table no. : 9

2-Frequent itemset c2

The next step is to generate candidate 2-itemset from L1[4]. To get support count for every itemset, split each itemset in 2-itemset into two elements then use l1 table to determine the transactions where you can find the itemset in, rather than searching for them in all transactions. For example ,let's take the first item (Apple, Milk), in the original Apriori we scan all 9 transactions to find the item (Apple, Milk); but in our proposed improved algorithm we will split the item (Apple, Milk) into Apple and Milk and get the minimum support between them using L1, here Apple has the smallest minimum support .After that we search for itemset (Apple, Milk) only in the transactions 1, 4, 8 and 9.

C_2 is shown as

Itemset	Support	minimum	Transaction ID set
{Apple ,Milk}	4	Apple	{1,4,8,9}
{Apple, Bread}	4	Apple	{5,8,9}
{Apple, Crisps}	1	Crisps	{4}
{Apple, Beer}	2	Apple	{1,8}
{Milk, Bread}	5	Bread	{3,6,8,9}
{Milk, Crisps}	2	Crisps	{2,4}
{Milk, Beer}	2	Beer	{1,8}
{Bread, Beer}	1	Bread	{8}

Table no. : 10

L_2 is shown as

Itemset	Support	minimum	Transaction ID set
{Apple ,Milk}	4	Apple	{1,4,8,9}
{Apple, Bread}	4	Apple	{5,8,9}
{Apple, Beer}	2	Apple	{1,8}
{Milk, Bread}	5	Bread	{3,6,8,9}
{Milk, Crisps}	2	Crisps	{2,4}
{Milk, Beer}	2	Beer	{1,8}

Table no. : 11

3-Frequent Itemset C_3& L_3 are as follows

Itemset	Support	minimum	Transaction ID set
{Apple, Milk, Bread}	2	{Apple ,Milk}	{8,9}
{Apple,Milk, Beer}	2	{Milk, Beer}	{1,8}

Table no. : 12

IV. RESULT

Apriori Algorithm	Eclat Algorithm	Improved Apriori Algorithm
{Apple, Milk, Bread}	{Apple, Milk, Bread}	{Apple, Milk, Bread}

Table no. : 13

V. COMPARISION

Parameters	Apriori Algorithm	Eclat Algorithm	Improved Apriori Algorithm
Concept	Find Support Compare & Prune	No need of Support Determine by Transaction	Find Support & Minimum of the Itemset.
No. of scans	More	More	Less
Compatible Dataset	Large	Small	Large
Memory Utilization	More	Less	Less
Time Consumed	More	More	Less

Table no. : 14

VI. CONCLUSION

In this paper, the comparison is made between three Frequent Pattern generation Algorithms that are Apriori, Eclat & Improved Apriori Algorithms. The comparison tells that the basic concept & the result is same for all the three algorithms but the creation technique is different for all three algorithms. As, this analysis is done with respect to the concept Market Basket Analysis using Hadoop, there is need of using large dataset for BigData. In order to fulfill requirement of Hadoop, Apriori Algorithm & Improved Apriori Algorithms are most feasible.

ACKNOWLEDGEMENTS

We would like to thank all academic staff in our Institute for supporting us in this research Project.

REFERENCES

[1] Kaushal Vyas ,Shilpa Sherasiya. Modified Apriori Algorithm using Hash Based Technique, International Journal of Advance Research and Innovative Ideas in Education ISSN(O)-2395-4396,2016

[2] Siddharajsinh Solanki, NehaSoni. A Survey on Frequent Pattern Mining MethodsApriori,Eclatand FP International Journal of Computer Techniques.
International Journal of Advance Research and Innovative Ideas in Education ISSN (O)-2395-4396, 2016.

[3] Ritu Garg,Preeti Gulia. Comparative Study of Frequent Itemset Mining Algorithms Apriori and FP Growth International Journal of Computer Applications , September 2015.

[4] Minal G Ingle, N. Y Suryavanshi. Association Rule Minig using Improved Apriori Algorithm, International Journal of Computer Application(0975-8887) Volume 112-No 4, February 2015.

[5] Mohammed Al-Maolegi,BassamArkok. An Improved Apriori Algorithm for Association Rules.
International Journal on Natural Language Computing , February 2014.

[6] Wan Faeah Abbas, Nor Diana Ahmad,Nurli Binti Zaini. Discovering Purchasing Pattern of Sport Items Using Market basket Analysis,2013 International Conference on Advance Computer science Applications & Technologies,978-1-4799-2758-6/13,2013 IEEE.

[7] XIE Wen-xiu,Qi Heng-nian,huang Mei-li. Market Basket Analysis based on text segmentation & association & Distributed Computing 978-0-7695-4207-2/10,2010 IEEE.

.