*International Journal of Research in Advent Technology, Vol.2, No.3, March 2014*
*E-ISSN: 2321-9637*

# A Survey on Association Rule Mining for Market Basket Analysis and Apriori Algorithm

Arti Rathod[1]
*Department of Computer Science,*
*Shrinathji Institute of Technical Education,*
*Rajasthan Technical University*
*Nathdwara, India*
artirathod@ymail.com

Mr. Ajaysingh Dhabariya [2]
*Dept. of Computer Science*
*Shrinathji Institute of Technical Education,*
*Rajasthan Technical University*
*Nathdwara, India*

Chintan Thacker[3]
*Dept. of Computer Science*
*HJD Institute of Engineering, kera*
*Gujarat Technological University*
*Bhuj, India*

**Abstract**-Association rule mining discovers the association or relationship between a large set of data items. With huge quantity of data constantly being obtained and stored in databases, several industries are becoming concerned in mining association rules from their databases. The identification of such associations can assist retailers to expand marketing strategies by gaining insight into which items are frequently purchased jointly by customers. It is helpful to examine the customer purchasing behaviour and assists in increasing the sales and conserve inventory by focusing on the point of sale transaction data. Market Basket analysis identifies the buying behaviour of the customer among various items that customer places in their shopping baskets. It is the technique to derive associations between datasets. Apriori algorithm is the classical algorithm for mining association rule.

***Index Terms*** **— Association, Apriori algorithm, Market Basket Analysis, Support, Confidence.**

## 1. INTRODUCTION

The Organizations are increasingly interested in retaining existing customers as well as targeting non customers. Measuring customer satisfaction provides an indication of how successful the organization is at providing products and /or services to the market place. In the present scenario customer satisfaction is the key stone for the success of the every organization .Therefore it is necessary to evaluate the changes happening in the tastes and desires of the customers [1].

Thus Apriori is a classical algorithm for association rule mining which is used to discover knowledge for the purpose of explaining current behaviour, predicting future outcomes and to provide support for Bank's Decision making Processes and also for some other Business Intelligence Purposes [1]. Market Basket Analysis is the best example for the Association rule mining.

## 2. ASSOCIATION RULE MINING

In Data Mining Association rule learning is a method for discovering interesting relations between variables in large database. Association rule discovers the interesting association or correlation among a large set of data items. [2]

Example: The rule {onions, potatoes} => {burger} in the sales of Super market would indicate that if a customer buys onions and potatoes together, he or she is likely to buy burger also. This information will help business to know the behavior of the customers [2]. Goggle auto complete is also the application of Association rule mining where we type a word and it searches frequently associated words that user type after that particular word.

## 3. CONCEPTS OF ASSOCIATION RULE MINING

*Support:* If x and y are two items in database then both comes together. [1]

$$\text{Support}(X, Y) = n(XUY)/N$$

N=Total no .of transactions.

*Confidence:* Ttransactions that contain *X* also contain *Y.[1]*

$$\text{Confidence}(X, Y) = \text{support}(XUY)/\text{support}(X)$$

Table 1 Basic concept of association rules [3]

| Name | Explain | Formula |
|---|---|---|
| Confidence | Probability of set Y appear only if X appear | P(Y\|X) |
| Support | Probability of set X and Y appear simultaneity | $P(Y \cap X)$ |
| Expected | Probability of set Y appear | P(Y) |

| | | |
|---|---|---|
| Confidence | | |
| Lift | Ratio of confidence to expected confidence | P(Y\|X)/P(Y) |

*Item:* It is a field of transactional database.

Consider the following Transactional database Table-I:

TABLE 2
TRANSACTIONAL DATABASE

| Transaction Id | Milk | Bread | Butter |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 |
| 5 | 0 | 1 | 0 |

In Table 2, 1 represent the presence of item and 0 represent the absence of items. Now let's count the support and confidence.

Consider X=milk and Bread, Y = Butter.

Support {milk, Bread}→{ Butter}=Support(X→Y)
=1/5
=0.2(20%)

Confidence {Milk, Bread} → {Butter} =Confidence(X→Y)
=0.2/0.4
=0.5(50%)

Support says that milk butter and bread all purchased together while confidence says that whenever milk and bread purchased there is also possibility of butter.

## 4. MARKET BASKET ANALYSIS

In the market basket analysis we analyse the existing database to identify potentially interesting patterns. The objective is not only to characterize the existing database. What one really wants to do is, first, to make inferences to future likely co-occurrences of items in a basket, and, second and ideally, to make causal statements about the patterns of purchases: if someone can be persuaded to buy item I1 then they are also likely to buy item I2.[4]

Let us call the items currently seen by the customer as X and Item Y is the item associated with the current item(X).If we have two items namely P and Q then the possible association rules are only two [5]: P→Q and Q→P.

If we have three items P, Q and R as follow (Table II).Then we will have 12 possible association rules [5] (Table-III).

TABLE II
TRANSACTIONAL DATABASE

| Transaction Id | P | Q | R |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 |
| 5 | 0 | 1 | 0 |

Based on table-2 we can derive the association rule (Frequent Items) using market basket analysis as follow: 1.Generate all possible association rules 2.Computer the support and confidence of all possible association rules 3.Apply two threshold criteria: Minimum support and Minimum confidence. [5]

TABLE 3
COMBINATION OF PURCHASED ITEMS

| | X | Y | Support | Confidence |
|---|---|---|---|---|
| 1 | P | Q | 0.4(40%) | 0.4(40%) |
| 2 | P | R | 0.2(20%) | 0.5(50%) |
| 3 | P | Q,R | 0.2(20%) | 0.5(50%) |
| 4 | Q | P | 0.4(40%) | 0.66(66%) |
| 5 | Q | R | 0.2(20%) | 0.33(33%) |
| 6 | Q | P,R | 0.2(20%) | 0.33(33%) |
| 7 | R | P | 0.2(20%) | 0.5(50%) |
| 8 | R | Q | 0.2(20%) | 0.5(50%) |
| 9 | R | P,Q | 0.2(20%) | 0.5(50%) |
| 10 | P,Q | R | 0.2(20%) | 0.5(50%) |
| 11 | P,R | Q | 0.2(20%) | 1(100%) |
| 12 | Q,R | P | 0.2(20%) | 1(100%) |

From the table-3 we can give the threshold value to the support and confidence for getting association rule. so let's give the minimum value to support 30% means the frequency of the item X and Y Buy together and minimum value to the confidence 60% means the frequency of the transaction when customer buy item X also buy item Y.

From table 3 the transaction no-4 (Table 4) has the higher then the threshold value of support and confidence. Means the product in the transaction-4 are certainly purchased by the customers.

TABLE 4
OBTAINED TRANSACTIONS AS A ASSOCIATION RULE

| | X | Y | Support | Confidence |
|---|---|---|---|---|
| 1 | Q | P | 0.4(40%) | 0.66(66%) |

X is the Combination of items up to n-1 where n is the number of items. Y is the combination of the set difference between all items and items listed on the Y.

## 5. APRIORI ALGORITHM

Apriori is a algorithm proposed by R.Agrawal and R.Srikant in1994 for mining frequent itemsets for association

rules. The name of the algorithm based on the fact that the algorithm uses prior knowledge of frequent itemsets properties [6].Apriori algorithm employs an iterative approach known as level wise search where k-itemsets are used to explore (k+1) itemsets[7].

Apriori Property: "All nonempty subsets of a frequent itemset must also be frequent" [7][6].

Apriori is one of the association rule mining algorithm which is used to discover all frequent itemsets from transactional database [6]. To understand the Apriori algorithm we need to understand the definition of following terms:

*Itemset*: It's a collection of items in a database[8].

*Transaction*: It's a database entry which contains a collection of items [8].

*Frequent Itemset (Large Itemset ($L_i$)):* The itemsets which satisfies the minimum support criteria are known as frequent itemsets. [8]

*Candidate Itemset ($L_i$):* Items which are only used for the processing. Candidate itemsets are all possible combination of itemsets. [8]

*Minimum Support:* It's a condition which helps to eliminate the non frequent item s from database. [8]

*Support:* Interesting association rule can be measured with the help of support criteria. Support is nothing but how many transactions have such itemsets that match both sides of implications in the association rule.[9]

Support (i) =Count (i)/total transaction

Apriori algorithm works on two concepts: [7]

(i) Joining and
(ii) Pruning

### *5.1 Apriori Algorithm Steps [9]:*

*1)* First, the set of candidate-1itemset is found ($C_1$).
2) Then support is calculated by counting the occurrence of the item in transactional database.
3) After that we will prune the $C_1$ using minimum support Criteria. The item which satisfies the minimum support criteria is taken into consideration for the next process and which is known as $L_1$.
4) Then again candidate set generation is carried out and the 2-itemset which is generated known as $C_2$.
5) Again we will calculate the support of the 2-Itemset ($C_2$). And we will prune $C_2$ using Minimum support and generate L2.
6) This Process Continues till there is no Candidate set and frequent itemsets can be generated.

Let's consider one example to understand the concept of Apriori algorithm: Table 5 shows transactional database having 4 transactions.

TABLE 5

| Transaction | Items |
|---|---|
| 1 | P,R,S |
| 2 | Q,R,T |
| 3 | P,Q,R,T |
| 4 | Q,T |

Performing the first step by scanning the database to identifying the number of occurrence for specific items. After that we will get $C_1$ as shown in Table 6 below:

TABLE 6

| Itemset | Support |
|---|---|
| P | 2 |
| Q | 3 |
| R | 3 |
| S | 1 |
| T | 3 |

The next step is pruning in which we will consider the minimum support criteria=2. The items which does not have minimum support Criteria will be eliminated. And we will get $L_1$.Table 7 shows the pruning step.

TABLE 7

| Itemset | Support |
|---|---|
| P | 2 |
| Q | 3 |
| R | 3 |
| T | 3 |

Now the candidate generation step is carried out and 2-itemset candidates are generated this is denoted as $C_2$. (Table 8)

TABLE 8

| Itemset | Support |
|---|---|
| P,Q | 1 |
| P,R | 2 |
| P,T | 1 |
| Q,R | 2 |
| Q,T | 3 |
| R,T | 2 |

Now pruning has to be done by considering minimum support criteria=2 and then we will get $L_2$. (Table 9)

TABLE 9

| Itemset | Support |
|---|---|
| P,R | 2 |
| Q,R | 2 |
| Q,T | 3 |
| R,T | 2 |

Again we will generate candidate set $C_3$. (Table 10)

TABLE 10

| Itemset | Support |
|---|---|
| P,Q,R | 1 |
| P,Q,T | 1 |
| Q,R,T | 2 |
| P,R,T | 1 |

Now pruning (minimum Support criteria=2) has been done to get $L_3$ As in (Table 11)

TABLE 11

| Itemset | Support |
|---|---|
| Q,R,T | 2 |

As we can see in Table 11 the frequent items are Q, R, and T.

***5.2 Pseudo Code [10]:***

$C_k$ : Candidate itemsets of size k

Lk: frequent itemset of size k

L1= {frequent items};

For (k= 1; Lk! = Lk-1; k++) do begin

Ck+1= candidates generated from Lk;

For each transaction tin database do

Increment the count of all candidates in Ck+1that are contained in t

Lk+1= candidates in Ck+1with min_support

End

***5.3 Drawbacks of Apriori Algorithm [11]:***

1. It takes too much time to scan the database.
2. It generates large number of in-frequent itemsets which increase the space complexity.
3. It has the difficulty to find rarely occurring items
4. Generates large amount of frequent itemsets which Are not efficient.
5. It needs several iterations for mining data.
6. Treats all items in database equally by considering only the presence and absence of an item within the transaction .it does not take into account the significance of item to user or business.

Many researchers has introduced several improved Apriori algorithm to eliminate the limitations of Apriori algorithms like record filter approach [12][15], Intersection approach [13][15], set size frequency approach [14][15], Interest Item approach [15][16] and Utilization of Attributes[15][17].

## 6. CONCLUSION AND FUTURE WORK

Association rules are very efficient in revealing all the interesting relationships in a relatively large database with huge amount of data. Association rule mining is easy to use and implement and can improve the profit of companies. The analysis of Apriori algorithms suggests that the usage of association rule mining for market basket analysis will help in better classification of the huge amount of data. The Apriori algorithm can be modified effectively to reduce the time complexity and enhance the accuracy.

Classical Apriori algorithm for association rule mining has several limitations like scanning time, memory optimization, candidate generation which can be solved by several improved Apriori approaches like record filter approach, intersection approach ,matrix based approach, set size frequency approach, interest item approach .

This classical Apriori treats all the items in database equally by considering only the presence and absence of an item within the transaction. So, Apriori algorithm efficiency can be improved by using quantity, profit attributes and support count which will give the valuable information to customer as well as business.

## 7. REFERENCES

[1]. ”Association Model for market basket analysis, customer behaviour analysis and business intelligent solution embedded with Apriori concept”,J.M Lakshmi ,Mahesh,SCMS school of technology and management Muttom, cochin, Kerala, International journal of research in finance & marketing ,vol 2,Issue 1, January 2012 ISSN:2231-5985.

[2]. "The Survey on Data Mining Algorithm for market basket analysis." , Dr. Dhanabhakyam, Dr.M.Punithavalli, Dr.SNS college of Arts and Science, Global Journal of Computer Science and Technology (IJCSEIT), Vol.11,Issue 11 Version 1.0, July 2012,ISSN:0975-4172.

[3]. "The Research of improved association rules mining apriori algorithm", Huiying Wang , Xiangwai Liu 2011 IEEE Eight International Conference on Fuzzy Systems and Knowledge Discovery(FSKD).

[4]. "Association Rule Mining as a Data Mining Technique",Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1,2008.

[5]. "Association Rule Mining as a Data Mining Technique",Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1,2008.

[6]. "Mining Efficient Association rules through Apriori algorithm using attributes and comparative analysis of various Asssociation rule algorithm", Ms Shweta, Dept. of Computer Science and application, Kurukshetra University, Kurukshetra, India, International Journal of Advance Research in Computer Science and software engineering, Vol-3, Issue-6, june 2013.

[7]. "Ranking and Suggesting Popular Itemsets In Mobile Stores Using Modified Apriori Algorithm", P V Vara Prasad, Sayempu Susmitha,Badduri Divya, Gogineni Riharika, Guntur Vijya Raghu Ram., International Journal of Modern Engineering Research(IJMER),Vol 2,Issue 1,Jan-Feb2012,pp-431-435.

[8]. "Improved Apriori Algorithms-A Survey",Pranay Bhandari, K.Rajeswari,Swati Tonge, MahadevShindalkar, Dept. of Computer Engineering , Pimpri Chinchwad College of Engineering Pune, Maharastra India, International Journal of Advance Computational Engineering and Networking,ISSN:2320-2106,Vol-1,Issue-2-2013.

[9]. "An improved Apriori based algorithm for association rule mining" ,Haun Wu Zhigang Lu,Lin Pan,Rongsheng Xu,Computer center,Institute of High energy physics,chinese academy of sciences,Beijing100049,China, 2009-Sixth international conference on Fuzzy Systems and knowledge Discovery IEEE Exlopre.

[10]."An approach to extract efficient frequent patterns from transactional database" , Mamta Dhanda ,nternational journal of engineering science and technology, vol 3,no.7, july 2011,ISSN:0975-5462.

[11]. *"Drawbacks and solutions of applying association rule mining in learning management system"*, Enrique Garcia, Cristobal Romero, Sebastian Ventura, Toon Calders , Cordoba University, campus Universitario de Rabanales,14071,Cordoba,spain,Eindhoven university of Technology (TU/e),Netherlands, International workshop on Applying datamining in E-learning 2007.

[12]. "*Frequent pattern mining using record filter approach*", D.N Goswami, Anshu Chaturvedi, and C.S Raghuvanshi, International Journal of Computer science issues Vol-7,Issue-4,No 7,July 2010.

[13]. "*An algorithm for frequent pattern mining based on apriori*", D.N Goswami, Anshu Chaturvedi, and C.S Raghuvanshi, International Journal of Computer science and Engineering,Vol-02, No.-4, 2010,942-947, ISSN:0975-3397.

[14]. "*Association rule mining based on apriori algorithm in Minimizing candidate generation*", Sheila A.Abaya International Journal of Scientific & engineering Resesrch, Vo1-3, Issue-7, July 2012, ISSN 2229-5518.

[15]. "*A review Approach on various form of Apriori with association rule mining*", Ms.Pooja Agrawal,Mr.Suresh Kashyap,Mr.Vikas Chandra Pandey,Mr. Suraj Prasad Keshri,International Journal on Recent and Innovation Trenda In Computing and Communication.volume-1,Issue-5,May2013.

[16]. "*Research on improving apriori algorithm Based on interested table*" ,Wu, Kui Gong, Fuliang Guo, Xiaohua Ge, Yilei Shan School of Computer, Wuuhan University, Wuhan, China,2010, IEEE.

[17]. "*Mining efficient association rules through apriori algorithm using attributes*",Mamta Dhanda, Sonali Guglani, Gaurav Gupta, International Journal of Advance Computer Science and Technology,Vol-2,Issue-3,September 2011.