Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.411 – 417

RESEARCH ARTICLE

AN EXHAUSTIVE STUDY ON ASSOCIATION RULE MINING

P.Thangaraju[1], D.Nanthini[2]

[1]Assistant Professor, Department of Computer Application, Bishop Heber College, Tiruchirappalli, TN, India
[2]Research Scholar, Department of Computer Science, Bishop Heber College, Tiruchirappalli, TN, India
[1] thangaraj@bhc.edu.in; [2] nandhu2021@gmail.com

Abstract— Association Rule Mining is a data mining method that enables to find out frequent item set, interesting pattern, interesting correlation among set of items in a transactional database or data repositories. This paper establishes the preparatory of basic concept about Association Rule Mining. Broadly, Association Rule Mining can be classified into Apriori, Partitioning, and Frequent Pattern Tree Algorithms. This study is to show the Benefits and Limitations.

Keywords— Data Mining, Association rule, Apriori, FP Tree, and Partitioning

I. INTRODUCTION

Data mining has recently attracted huge amount of attention in the database research [1].Data mining is defined as the process of non-trivial extraction of previously unknown and potentially useful information from data stored in databases [2]. Data mining is used to find patterns (or itemsets) hidden within data, and associations among the patterns.

Data mining is interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization [3].In contrast to an expert system (which draws inferences from the given data on the basis of a given set of rules) data mining attempts to discover hidden rules underlying the data.Also called data surfing. Data mining requires an algorithm or method to analyze the data of interest. Data may be a sequence data, sequential data, time series, temporal, spatial- temporal, audio signal, video signal to name a few [17].Data mining techniques useful in many fields.

II. LITERATURE REVIEW

A. *Association Rule Mining*

Association Rule Mining [ARM] is one of the most significant and traditional techniques of the Data Mining. Association Rule[AR] are extensive used in different area like market basket analysis, medical application, modern communication network etc.ARM has been attractive topic of research area. So,many researcher focus in the ARM.ARM techniques and algorithm will be scanty introduce.Association Rule that compensate the predefine minimum support and minimum confidence from given database.

Association Rule is involved two types of process: Candidate Large item sets generation process and frequent item sets generation process.Hence,frequent item sets or expected item sets are called candidate item sets. Further, many strategies have been proposed to reduce the number of AR, that only generating for interesting rules and non redundant rule or only satisfy certain other criteria.

B. *Concepts of ARM*

ARM plays dominant role in the field of data mining.ARM is capable method which is used in finding the AR. Let I=I1,I2…..IN be a set of N distinct attributes, T is represents transaction that hold a set of items such that T⊆I,D is represents database with different transaction records Ts. Association Rule is an implication in the form of X→Y, where X,Y⊂I are sets of items called item sets, and X ∩Y=φ.X is a antecedent and Y is a Consequent, the rule means X implies Y [8].AR involves two principals, Support(S) and Confidence(C).

$$\text{Support } X \rightarrow Y = P\ (X \cup Y)$$

The Support(S) as a define rule X→Y is fraction of transaction which contain both X and Y.

$$\text{Confidence } X \rightarrow Y = P\ (Y|X)$$

The Confidence(C) as a define rule X→Y is the fraction of transaction containing X which Also contain Y.

Find all Association Rule X→Y with minimum Support(s) and minimum Confidence(c).

$$\text{Min\_Sup} = \frac{\text{No. of transacting on containing both A \& B}}{\text{Total No. of transaction}}$$

$$\text{Min \_Conf} = \frac{\text{No. of transacting on containing A \& B}}{\text{Transaction containing only A}}$$

Association Rule mining supports many algorithms especially Apriori, Partitioning, FP Tree algorithms processed effectively.

C. *Different ARM Algorithms*

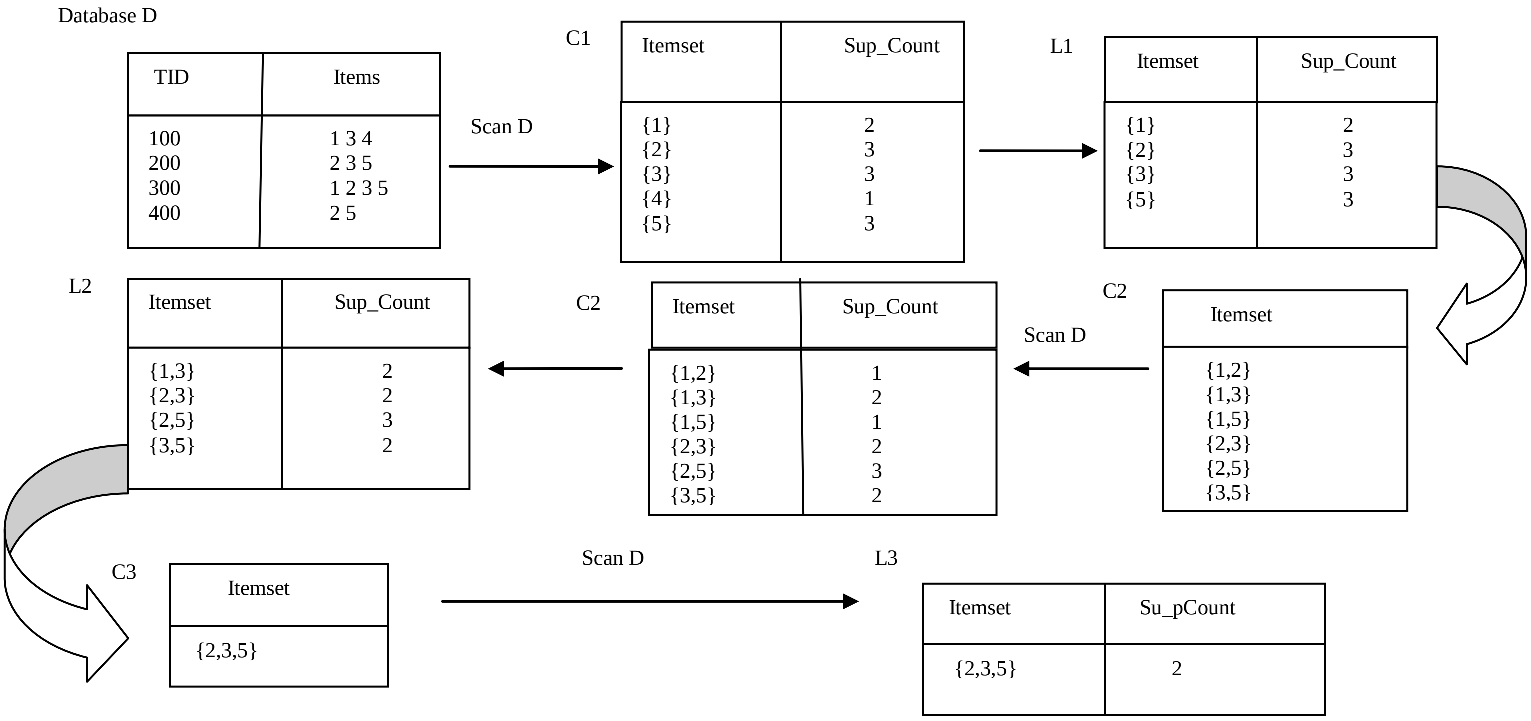
Apriori Algorithm

Apriori algorithm is the most classical association rule mining algorithms and was proposed by Agrawal et al in 1993. This algorithm based on level-wise search techniques and used to reduce the search space. This algorithm divide into two steps: retrieve frequent item sets from candidate frequent item sets, generate rules from frequent item sets [9].It mining frequent item set using pruning technology.

In this algorithm ,if an item set I does not satisfy the minimum support threshold, then I is not frequent, since item set must satisfy both support and confidence threshold value.

TABLE I
APRIORI ALGORITHM EXAMPLE

Minimum support_count=2



Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C1

Itemset	Sup_Count
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L1

Itemset	Sup_Count
{1}	2
{2}	3
{3}	3
{5}	3

C2

Itemset
{1,2}
{1,3}
{1,5}
{2,3}
{2,5}
{3,5}

C2

Itemset	Sup_Count
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

L2

Itemset	Sup_Count
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

C3

Itemset
{2,3,5}

L3

Itemset	Su_pCount
{2,3,5}	2

Apriori Algorithm Involves:

1. Scan full database once for counting candidate item set.
2. Compare candidate support count with minimum support count.
3. Eliminate infrequent items from item set.
4. Generate another candidate item set.
5. Scan the database and count each candidate. This process repeatedly occurs until satisfy the minimum support count threshold [Table1].

Partitioning Algorithm

Partitioning algorithm is basically based on apriori algorithm, but it requires only two complete scans over the database.Figure1 depicts the partitioning approach for frequent item sets mining [4].The partition algorithm is divided into two phases:[18]

Phase1. The database is divided into a number of non overlapping partitions and frequent item sets local to partition are generated for each partition. The database is scanned completely for the first time.

Phase2. Local frequent item sets from each partition are combined to generate global candidate item sets. Then the database is scanned second time to generate global frequent item sets.

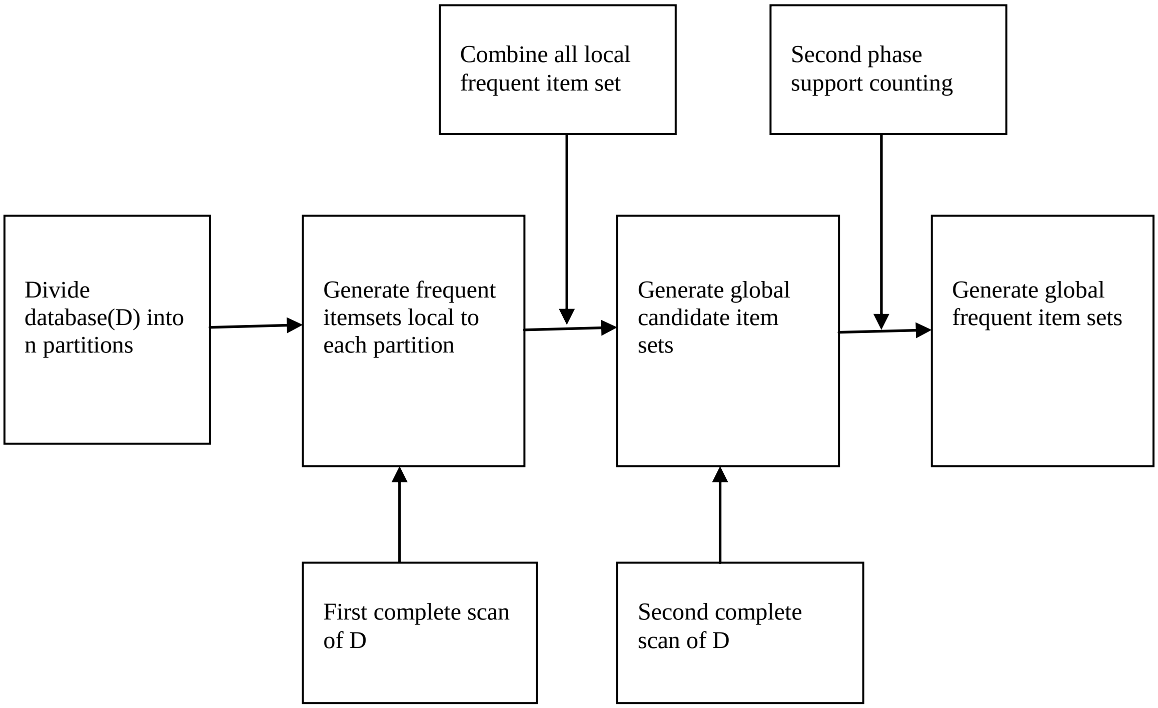

Combine all local frequent item set
Second phase support counting
Divide database(D) into n partitions
Generate frequent itemsets local to each partition
Generate global candidate item sets
Generate global frequent item sets
First complete scan of D
Second complete scan of D

Fig 1. Partitioning Algorithm Example

Partitioning Process Structure:

1. Divide the database into number of partitions.
2. Generate frequent item set local to each partition.
3. Scan whole database.
4. Combine all local frequent item set.
5. Generate global candidate item sets.
6. Scan whole database again.
7. Second phase support counting.
8. Generate global frequent item sets [Fig 1].

Frequent Pattern (FP) Tree

The Frequent pattern tree (*FP-Tree*) is a prefix-tree structure for storing crucial and compressed information about frequent pattern. *FP-Trees* involve the divide and conquer method. The root of the *FP-Tree* is represented as "NULL" value. Childs of the roots are the set of item of data. Conventionally a *FP-Tree* contains three fields- Item name, node link and count.

FP-Tree structure:

1. One root labeled as "null" with a set of item-prefix sub trees as children, and a frequent-item-header table.
2. Each node in the item-prefix sub tree consists of three fields:
   i) Item-name: registers which item is represented by the node.
   ii) Count: the number of transactions represented by the portion of the path reaching the node.
   iii) Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.
3. Each entry in the frequent-item-header table consists of two fields:
   i) Item-name: as the same to the node.
   ii) Head of node-link: a pointer to the first node in the FP-tree carrying the item-name [Table 2] [Fig 2].

TABLE II
EXAMPLE OF TRANSACTION DATABASE IN FP TREE

TID	Items bought	(Ordered) frequent items
101	*a, b, c, d, e, f, g, h*	a, c, b, g, h
102	*b, i, c, a, j, g, k*	a, c, b, i, g
103	*I, a, l, m, k*	a, i
104	i, *c, n, o, h*	c, i, h
105	*b, a, c, p, j ,h, g, q*	a, c, b, g, h

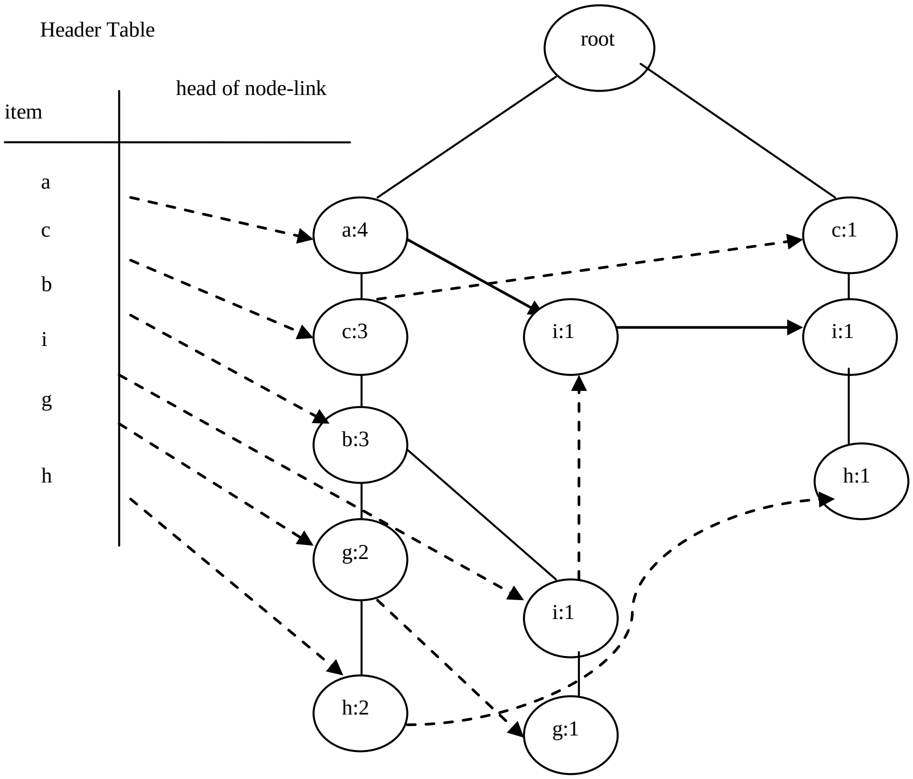


Fig 2. FP Tree Structure

C. Benefits and Limitation

TABLE III
THE COMPARISON OF THREE ALGORITHMS

Algorithms	Benefits	Limitations
Apriori	1.Uses large item set property 2.Easily parallelized 3.Easy to implement	1. Assumes transaction database is memory resident. 2. Requires many database scans. 3. Apriori algorithm can be very slow and the bottleneck is candidate generation.
Partitioning	1. Generates clinically more intuitive models that do not require the user to perform calculations. 2. Allows varying prioritizing of misclassifications in order to create a decision rule that has more sensitivity or specificity. 3. May be more accurate.	1.Does not work well for continuous variables 2. May over fit data.
FP Tree	1. FP-tree is that the algorithm scans the tree only twice. 2. Uses compact data structure. 3. No candidate generation. 4. Much faster than Apriori. 5. Eliminates repeated database scan.	1. FP-Tree is expensive to build. 2. Time is wasted as the only pruning that can be done is on single items. 3. Support can only be calculated once the entire data-set is added to the FP-Tree.

III. CONCLUSION

We have studied three association rule mining algorithm. The algorithms are systemized and their performance is analyzed based on benefits and limitations.The Comparison table shows [Table3] that the FP growth displayed better performance in all the cases.Because it find out frequent itemset without candidate generation and only two scan.

REFERENCES

[1] Shruti Mishraa, Debahuti Mishraa, Sandeep Ku. Satapathya, “Fuzzy Frequent Pattern Mining from Gene Expression Datausing Dynamic Multi-Swarm Particle Swarm Optimizati”, Elsevier, 2011.

[2] Unil Yun a, Hyeonil Shin b, Keun Ho Ryu a, EunChul Yoon c,”An efficient mining algorithm for maximal weighted frequent patterns in transactional databases”, Elsevier, 2012.

[3] T.Karthikeyan, P.Thangaraju,”Analysis of Classification Algorithms Applied to Hepatitis Patients”, International Journal of Computer Applications (0975 – 8887) Volume 62– No.15, January 2013.

[4] J.Han, m.Kamber, ”Data Mining: Concepts and Techniques”,2001: Morgan kaufmann publishers.

[5] Wei, Huang. "Study on a Data Warehouse Mining Oriented Fuzzy Association Rule Mining Algorithm" In Intelligent Systems Design and Engineering Applications (ISDEA),2014 Fifth International Conference on, pp. 935-938. IEEE, 2014.

[6] Deshpande, Deepa S, "A Novel Approach for Association Rule Mining using Pattern Generation" ,(2014).

[7] Priyanka, Er.VinodKumar, Sharma,"AprioriAlgorithm For Minig Frequent Itemsets –A Review", ijcaet, Volume 3-Issue 3, July 2014.

[8] Sotiris Kotsiantis, Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006.

[9] Sonal Vishwakarma, Shrikant lade, "Modern Research Trends in Association Data Mining Techniques"**,** IJCTA, 2013.

[10] A.B.M.Rezbaul Islam, Tae-Sun Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique", IEEE, International Conference on Information Science and Applications (ICISA), 2011.

[11] Gosta Grahne, Member, Jianfei Zhu, Student Member,"Fast Algorithms for Frequent Itemset Mining Using FP-Trees", IEEE, 2005.

[12] Victoria Nebot, Rafael Berlanga," Finding association rules in semantic webdata", Elsevier, 2011.

[13] Bay Vo, Tzung-Pei Hong, Bac Le,"A lattice-based approach for mining most generalization association rules", Elsevier, 2013.

[14] M.Sumani1, T.Anuradha2, K.Gowtham3, A.Ramakrishna4,"A Frequent Pattern Mining Algorithm Based On FP-Tree Structure And algorithm", Research Journal of Computer Systems Engineering-RJCSE, 2011.

[15] Bilal Sowan a, Keshav Dahal a, M.A. Hossain b, Li Zhang b, Linda Spencer b, " Fuzzy association rule mining approaches for enhancing prediction performance", Elsevier,2013.

[16] Rashmi Shikhariya, Prof.Nitin Shukla, "An improved association rule mining tree using positive and negative integration" ,JGRCS,2010.

[17] M.S.B.PhridviRaja, C.V.GuruRaob,"Data mining – past, present and future – a typical survey on data streams", Elsevier, 2013.

[18] A.Sarasere, E.Omiecinsky, S.Navathe,"An Efficient Algorithm for Mining Association Rules in Large Databases", International Conference on VeryLargeDatabases (VLDB), 1995.