Efficient Analysis of Frequent itemset Association Rule Mining Methods

M. Inbavalli

Asst. Professor, MCA

Er. Perumal Manimekalai College of Engineering, Hosur

inbavelu@yahoo.com

Dr. Tholkappia Arasu

Principal, AVS Engineering, College, Salem

[2]tholsg@gmail.com

Abstract

***Data mining* is the analysis of huge amounts of data in order to discover meaningful patterns.The growing demand of finding pattern from huge data makes the association rule mining (ARM) one of the most important data mining tools. It intends to extract interesting frequent patterns, associations, correlations or casual structures among sets of items in the transaction databases or other data repositories.Frequent itemset is groups of items which appear together in a sufficient number of transactions.The Frequent pattern mining helps to discover customer behaviors such that the quality of business decisions can be improved in the tremendous globalization era.Various efficient algorithms support association mining but this paper shows the efficient analysis of only the two very popular approach *Apriori and FP-tree* method.**

Keywords

Data mining ,Association Rule Mining, Frequent Patterns, Apriori , Fp-Tree

1.INTRODUCTION

Decision-making is a vital part of the business world. It is useful for the successful operation of organizational activities. Organisations that perceive themselves as more successful than their peers rate their ability to drive executive decisions with data most highly. This has led to a growing interest in the development of tools

capable in the automatic extraction of knowledge from data for decision making . Data mining (sometimes called data or knowledge discovery) is a tool used the processing of automatic analyzing data from different perspectives and

summarizing it into useful information[1] - information that can be used to increase revenue, and reduce the costs. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

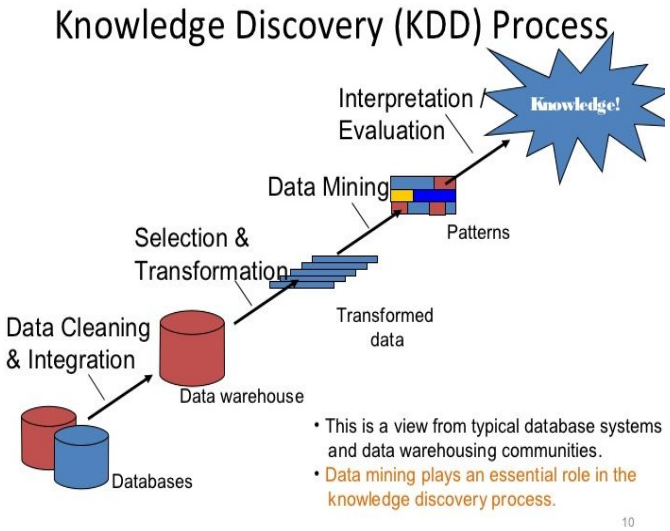


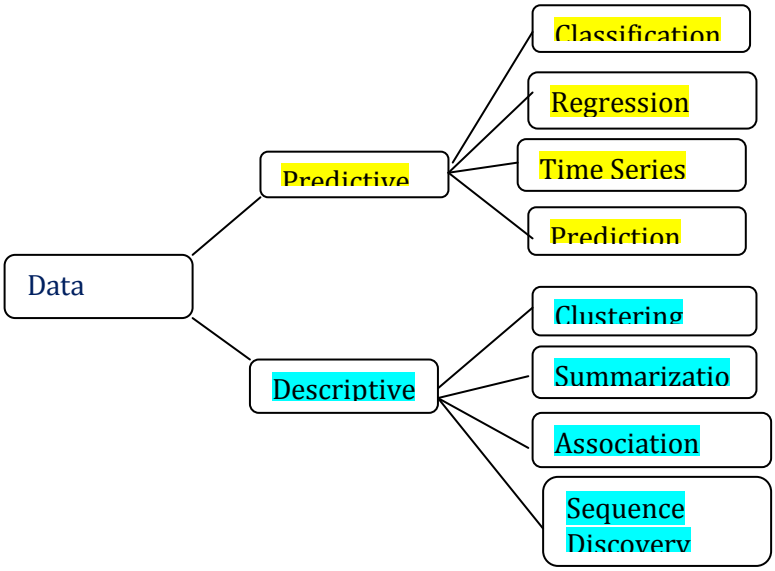
Figure 1. Process of knowledge discovery in databases

2. DATA MINING FUNCTIONALITIES

On the basis of kind of data to be mined there are two kind of functions involved in Data Mining They are Descriptive mining and Predictive mining[2]. Descriptive mining refers to the method to derive patterns (correlation,trends) that summarizes the underlying relationship between data.Frequent pattern mining ,Clustering, Association and Correlations are the main functionalities involved in the descriptive mining techniques tasks. Predictive mining Prediction is the process of predicting some unknown or missing numerical values rather than class labels infer patterns from the data in a similar manner as predictions. Predictive mining refers to predict the value of a specific attribute (target/dependent variable)based on the value of other attributes (explanatory). Techniques include functions like Classification,Decision Tree, Mathematical formulae(Regression and Deviation detection) and neural networks.[3][4]

Figure 2. Descriptive and Predictive Analytics

Figure 3.Predictive and Descriptive Data Mining Algorithms



Association rule mining is an important component of data mining. Association rules are an significant set of methods of finding patterns in data. It is possibly the most important model made-up and broadly studied by databases and data mining community. Association mining has been used in many application domains such as discovering of purchase patterns or association between products, finding patterns in biological databases, extraction of knowledge from software engineering metrics, web personalization, text mining etc. Association rule mining can also play an important role in discovering knowledge from agricultural databases, survey data from agricultural research, data about soil and cultivation, data containing information linking geographical conditions and etc., such knowledge helps to make effective decision making.[21][22].

3. ASSOCIATION RULE MINING

Association Rule Mining is the process of finding interesting correlations, frequent patterns or associations among sets of items in the transaction databases, relational databases or other information repositories[5] .An association rule is an expression in the form of X⇒Y, where X and Y are set of items called itemsets and intersection of X and Y is null [7]. The portion of the rule to the left of the implication (⇒) is known as the antecedent (X), whereas the

right side of the implication is known as the consequent (Y). A rule may contain more than one item in antecedent and consequent part. Association rule mining tends to produce a large number of rules. The goal is to find the rules that are useful to users[5][23]. There are four important performance measures for association rules: Support , Confidence, Minimum support threshold and Minimum confidence threshold.

Support is the percent of the transactions that contain X U Y (i.e. both X and Y) to the total number of transactions in database. Suppose the support of an item is 0.2%, it means only 0.2 percent of the transaction have that item [6].

Confidence is the percent of the transactions that contain X U Y to the total number of transactions that contain X. Suppose the confidence of the association rule X⇒Y is 70%, it means that 70% of the transactions that contain X also contain Y [6].

$$support(A \Rightarrow B) = P(A \cup B)$$
$$confidence(A \Rightarrow B) = P(B|A).$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

On discovering interesting association rule, the support and the confidence of the rule should satisfy a user-specified support threshold called minsup and a confidence threshold called minconf, respectively.[8]

For example, consider the database shown in the following table over the set of items

I = {milk, bread, chips, cocacola}:

Item Set	Tid	Support	Confidence (Frequency)
{}	{100, 200, 300, 400}	4	100%
{milk}	{100, 200}	2	50%
{bread}	{100,200,300}	3	75%
{chips}	{300,400}	2	50%
{cocacola}	{100,300}	2	50%
{milk, bread}	{400}	1	25%
{milk, cocacola}	{100}	1	25%
{bread,chips}	{400}	1	25%
{bread, cocacola}	{100}	1	25%
{chips, cocacola}	{300}	1	25%
{milk,bread,cocacola}	{100}	1	25%

Table 1. Set of items

The following table shows all frequent sets in Tid with respect to a minimal support threshold equal to 1, their cover inTid, plus their support and frequency:

Table 2 .Frequent Sets based on minimal support threshold

If we are given the min_sup,(minimum support threshold) then every frequent set X also characterize the trivial rule X=> {} which holds with 100% confidence. [8]

The task of discovering all frequent sets is quite challenging. The search space is exponential in the number of items occurring in the database and the targeted databases tend to be massive, containing millions of transactions. Both these characteristics make it a worthwhile effort to seek the most efficient techniques to solve this task.[5][8].

3.1 .ASSOCIATION RULE PROBLEM

Given a set of items represents $I=\{I_1,I_2,\ldots,I_m\}$ and a database of transactions database $D=\{t_1,t_2, \ldots, t_n\}$ where $t_i=\{I_{i1},I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the Association Rule Problem is to identify all association rules X ➔ Y with a ***minimum support and confidence***.

Association rule mining techniques : [9]

1) Finding all the frequent itemsets that satisfies support thresholds.

2) Generating interesting association rules from these frequent itemsets.

4. FREQUENT ITEMSET MINING METHODS

The first algorithm for mining all frequent itemsets and association rules was the AIS algorithm. Shortly after that the algorithm was improved and renamed Apriori. Apriori is an algorithm for frequent item set mining and association rule learning from the transactional databases. It progress by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules.[10]

5. APRIORI ALGORITHM

Apriori is designed to operate on databases containing transactions. Each transaction contains set of items called itemset. It is a decisive algorithm, which uses an repetitive approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. This set contains items that satisfy minimum support and is denoted by L1. This set is used for generating new itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually frequent and they are used in the next pass. Therefore, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. An important property called Apriori property is used to reduce the search space which is described as: "All nonempty subsets of a frequent itemset must also be frequent" . How Lk-1 is used to find Lk is consisting of two steps, join and prune actions as followed: [5][8][24][25].

1. **Join Step**: Join Lk-1 with itself to obtain the candidate itemset Ck.
2. **Prune Step**: Scan the database to determine the count of each candidate in Ck. When the count is less than the minimum support count, it should be removed from the candidate itemsets. Meanwhile, if any (k-1) subset of candidate k-itemset is not in Lk-1 then the candidate cannot be frequent either and so can be removed. After this, we get k-itemset which is denoted by Lk. [8]

Algorithm Apriori(*T*)[8]

$C_1 \leftarrow$ init-pass(T);

$F_1 \leftarrow \{f \mid f \in C_1, f.\text{count}/n \geq minsup\}$; // n: no. of transactions in T

for ($k = 2$; $F_{k-1} \neq \emptyset$; k++) **do**

$C_k \leftarrow$ candidate-gen(F_{k-1});

for each transaction $t \in T$ **do**

for each candidate $c \in C_k$ **do**

if c is contained in t **then**

$c.count$++;

end

end

$F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$

end

return $F \leftarrow \bigcup_k F_k$;

Function candidate-gen(F_{k-1})

$C_k \leftarrow \emptyset$;

forall $f_1, f_2 \in F_{k-1}$

with $f_1 = \{i_1, \ldots, i_{k-2}, i_{k-1}\}$

and $f_2 = \{i_1, \ldots, i_{k-2}, i'_{k-1}\}$

and $i_{k-1} < i'_{k-1}$ **do**

$c \leftarrow \{i_1, \ldots, i_{k-1}, i'_{k-1}\}$; // join f_1 and f_2

$C_k \leftarrow C_k \cup \{c\}$;

for each (k-1)-subset s of c **do**

if ($s \notin F_{k-1}$) **then**

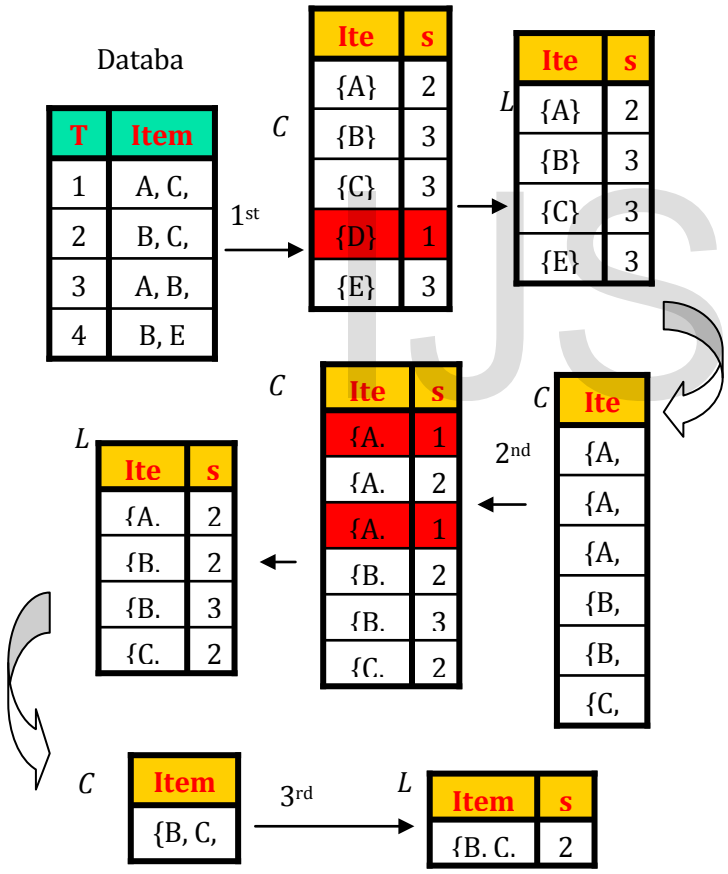
delete c from C_k; // prune

end

end

return C_k;

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



Tid	Set of items
100	{milk, bread, cocacola}
200	{milk,bread}
300	{chips,cocacola}
400	{bread, chips}

Figure 4. Example for Apriori Algorithm

5.1 DRAWBACKS

- It takes more time, space and memory for candidate generation process.
- To generate the candidate set, it requires multiple scans of transaction database.
- Generating huge number of candidates need to generate more association rules.
- Tedious workload for support counting for candidates.

In order to overcome the drawbacks of Apriori Algorithm, FP-Growth algorithm has been developed. The main difference between the two approaches is that the Apriori algorithm generates the candidate itemsets but FP-Growth does not generate the candidate itemset.

6. FP-TREE ALGORITHM

FP-Growth algorithm is an efficient algorithm for producing the frequent itemsets without generation of candidate itemsets.It Compress a large database into a compact, Frequent-Pattern tree (FP-tree) structure.It is Highly strong, but complete for frequent pattern mining and it avoids costly database scans.[8]

This method adopts a divide-and-conquer strategy as follows: first it compresses the database representing frequent items into frequent-pattern tree, or FP-tree, which maintain the itemset association information. It then segregates the compressed database into a set of conditional database, each associated with one frequent item or pattern fragment, and mines each such database separately. [8][26].

Let us create the FP-tree for the following:

Support	Confidence	Speed	
		Apriori	FP-Tree
0.1	0.2	44s	10s
0.15	0.2	19s	10s
0.25	0.2	10s	8s
0.5	0.2	8s	6s
1.0	0.2	8s	6s
1.5	0.2	6s	5s

Table 3 : Transaction Database

- First we scan the database and determine the set of frequent items (1-itemsets) and their support counts(frequencies):
  L={{I2:7},{I1:6},{I3:6},{I4:2},{I5:2}}
- Then we create the root of the FP-tree and label it with "null"
- We take each transaction, sort the items according to descending support count, and create a branch for it. For example the scan of the first transaction "T100:I1, I2, I5", which contain tree items: I2, I1 and I5 in sorted descending, leads to the construction of the first branch of the tree: (I2:1), (I1:1), (I5:1).
- The second transaction T200 contains the items I2 and I4. This would result a branch where I2 is linked to the root and I4 is linked to I2. However this branch would share a common prefix, i2, with the existing path for T100. Therefore we instead increment the count of the 12 node by 1 and create a new node (I4:1),which is linked as a child of (I2:2).

In general when considering the branch to be added for a transaction, the count of each node along a common prefix is incremented by 1and nodes for the items following the prefix are created and linked accordingly.

To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. In this way the problem of mining frequent pattern in database is transformed to that of mining the FP-tree.

The FP-tree is mined as follows:

Start from each frequent length-1 pattern, as an initial suffix pattern, construct its conditional pattern base, a sub-database, which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern, then construct its conditional FP-tree and perform mining recursively on such a tree.

The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

Algorithm:

1.Scan DB once, find frequent 1-itemset (single item pattern)

2.Order frequent items in frequency descending order, called ***L order***: (in the example below: F(4), c(4), a(3), etc.)

3.Scan DB again and construct FP-tree

a. Create the root of the tree and label it null or {}
b. The items in each transaction are processed in the L order (sorted according to descending support count).
c. Create a branch for each transaction
d. Branches share common prefixes

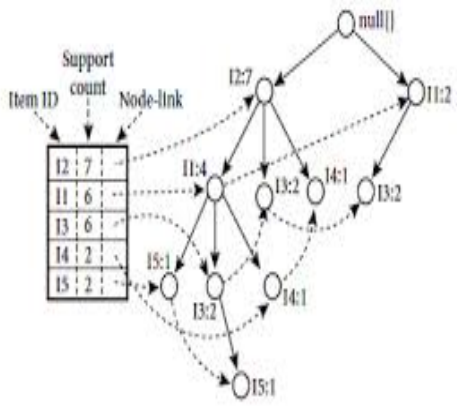


Figure 5. FP-Tree

6.1 ADVANTAGES

The FP-Tree method transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.

6.2 LIMITATIONS

It is difficult to be used in an interactive mining system. During the interactive mining process, users may alter the threshold of support according to the rules. However for FP-Tree the changing of support may lead to repetition of the whole mining process.Also FP-Tree is not suitable for incremental mining because with changing time new datasets may be inserted into the database which may lead to a repetition of the whole process using this algorithm.[v4]

7. COMPARISON OF APRIORI AND FP-TREE ALGORITHMS

Characteristics	**Apriori**	**Fp-Growth**
Data Support	Limited	Very Large
Execution Time	Slow	Fast
Accuracy	Less	More Accurate
Memory consumption	High	Less
Data Structure and mining methods	Easy to use	More complicated
Number of Scans	Increase based on dimensions of the candidate item sets	Atmost two scans
Effective	For small databases with large support factor	For large databases
Generating Frequent itemsets	Fast	Comparatively Slow

Table 4 . Comparison of Apriori and Fp-Tree Miniing Algorithms

8. EXPERIMENTATION RESULT

A syntactic data set has been used with 500 items for analysis.a set of Association rules were generated using Apriori and Fp-Tree algorithm.By analysing the data with different support and confidence values, we obtain different rules.During analysis we found that Fp-Growth is faster than Apriori algorithm for large number of transactions. But it takes less time to generate frequent itemsets.The transaction database are tested on java and the platform used is Pentium dual core processor with 2 GHz speed and I GB-RAM,Windows 2000.Table 3 shows that the execution time is based on the support factor .The superlative performance is obtained by FP-Tree algorithm.

Table 5. Experimentation Results of Apriori and Fp-Tree based on support and confidence

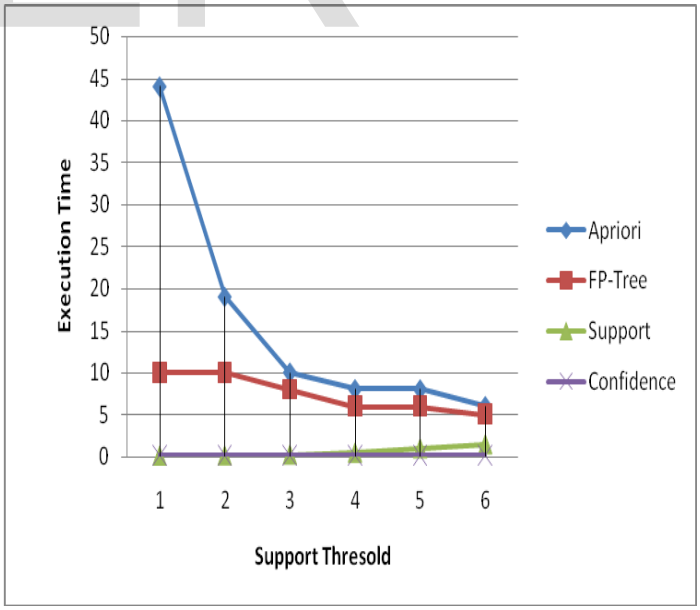


Figure 6. Comparison chart performance of Apriori and Fp-Tree based on support and confidence

CONCLUSION

The most important tasks of frequent pattern mining approaches are : itemset mining and association rule mining. An efficient data mining algorithms exist in the literature for mining frequent patterns. we have performed a efficient study of Apriori and Fp-tree that exists for the mining of frequent patterns.With the experimental a nalysis we found that Apriri algorithm takes more time to compute association rules with the same number of transactions.Fp-tree is muct faster than Apriori because there is no candidate generation and also it uses compact data structure, it eliminates repeated scans.

REFERENCES

[1] U.M. Fayyad, et al.: "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*:1-34, AAAI Press/ MIT Press, 1996, ISBN 0-262-56097-6.

[2] *Data mining: A knowledge discovery approach*. New York,NY: Springer, 2012.

[3] Thabet Slimani, Amor Lazzez ," *Efficient Analysis of Pattern and Association Rule Mining Approaches*", International Journal of Information Technology and Computer Science(IJITCS) Vol. 6, No. 3, February 2014 ISSN: 2074-9007

[4] Marek Wo, Krzysztof Ga, Krzysztof Ga. "*Concurrent Processing of Frequent Itemset Queries Using FP-Growth Algorithm*", Proc. of the 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'05),2005,Tallinn, Estonia.

[5] Gagandeep Kaur* Shruti Aggarwal , "*Performance Analysis of Association Rule Mining Algorithms* ",International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[6] R.Divya and S.Vinod kumar, "*Survey on AIS, Apriori And FP-Tree Algorithms*", International Journal of Computer Science and Management Research, Volume 1, Issue 2, pp. 194- 200, September 2012.

[7] Sotiris Kotsiantis and Dimitris Kanellopoulos, "*Association Rules Mining: A Recent Overview*", International Transactions on Computer Science and Engineering, Volume 32 (1), pp. 71-82, 2006.

[8] J. Han, M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[9]R. Agrawal, T. Imielinski, and A. Swami, "*Mining association rules between sets of items in large databases*",Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data (1993) 207-216

[10]Thabet Slimani, Amor Lazzez "*Efficient Analysis of Pattern and Association Rule Mining Approaches*" *College of Computer Science and Information Technology , Taif University , KSA*

[11]. C.S.Kanimozhi Selvi, and A.Tamilarasi, "*An Automated Association Rule Mining Technique With Cumulative Support Thresholds* ",Int. J. Open Problems in Compt. Math, Vol. 2, No. 3, September 2009 ISSN 1998-6262.

[12]. Archita Bonde, Deipali V. Gore," Comparative Study of Association Rule Mining Algorithms with Web Logs", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6153-6157 ISSN:0975-9646.

[13] Alva Erwin, Raj P.Gopalan, N.R.Achuthan, "*A Bottom – Up Projection Based Algorithm for Mining High Utility Itemsets*", Proceedings of the 2nd International workshop on Integrating Artificial Intelligence and Data Mining – Volume 84.

[14]. Jyotsana Dixit, Abha Choubey," *A Survey of Various Association Rule Mining Approaches*", International Journal of Advanced Research in

Computer Science and Software Engineering, Volume 4(3), March 2014 ISSN: 2277 128X

[15]. Rachna Somkunwar," *A study on Various Data Mining Approaches of Association Rules*", International Journal of Advanced Research in Computer Science and Software Engineering 2 (9), September- 2012, pp. 141-144

[16] Charanjeet Kaur*," Association Rule Mining using Apriori Algorithm: A Survey*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013 ISSN: 2278 – 1323

[17] Parita Parikh Dinesh Waghela," *Comparative Study of Association Rule Mining Algorithms*", UNIASCIT, Vol 2 (1), 2012, 170-172 ISSN 2250-0987.

[18]. Vimal Ghorecha," *Comparative Evaluation of Association Rule Mining Algorithms with Frequent Item Sets*", *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 5 (Mar. - Apr. 2013), PP 08-14*

[19]. Trupti A. Kumbhare, Prof. Santosh V. Chobe," An *Overview of Association Rule Mining Algorithms*", International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 927-930 ISSN 0975-9646.

[20]. Shikha Maheshwari,Pooja Jain, "*The Research on Top Down Apriori Algorithm using Association Rule*", International Journal of Advanced Research in

Computer Science and Software Engineering, Page 839 Volume 4 (4),April 2014 ISSN: 2277 128X.

[21] Web: http://www.iasri.res.in/ebook/win_school_aa/notes/association_rule_mining.pdf

[22]. Web :http://shodhganga.inflibnet.ac.in/bitstream/10603/36008/1/chapter1.pdf

[23]. Web: http://www.academia.edu/648890/Support_vs_Confidence_in_Association_Rule_Algorithms.

[24].Web:http://www.academia.edu/7930414/Implementation_of_Association_Rule_Mining_using_Reverse_Apriori_Algorithmic_Approach

[25]. Web: Hongyan Geng, Xiaoqi L, "Design and Implementation of Association Rules Mining based on Apriori Arithmetic".

[26]. J. Bhatia,, Anu Gupta," Mining of Quantitative Association Rules in Agricultural Data Warehouse: A Road Map ",International Journal of Information Science and Intelligent System,3(1): 187-198, 2014.

Frequent pattern mining has been a focused topic in data mining research with a good number of references in literature

COMPARISON OF APRIORI AND FP-GROWTH ALGORITHMS ON DETERMINATION OF ASSOCIATION RULES IN AUTHORIZED AUTOMOBILE SERVICE CENTRES

The paper first presents the basic concept of association rule mining, then discuss a few different types of association rules mining including multi-level association rules, multidimensional association rules, weighted association rules, multi-relational association rules, fuzzy association rules.

Mining association rules is an important task for knowledge discovery. We can analyze past transaction data to discover customer behaviors such that the quality of business decisions can be improved. Various types of association rules may exist in a large database of customer transactions. The strategy of mining association rules focuses on discovering large item sets, which are groups of items which appear together in a sufficient number of transactions. We propose a graph-based approach to generate various types of association rules from a large database of customer transactions. This approach scans the database once to construct an association graph and then traverses the graph to generate all large item sets. Empirical evaluations show that our algorithms outperform other algorithms which need to make multiple passes over the database

It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories

The process of data mining produces various patterns from a given data source. The most recognized data mining tasks are the process of discovering frequent itemsets, frequent sequential patterns, frequent sequential rules and frequent association rules. Numerous efficient algorithms have been proposed to do the above processes. Frequent pattern mining has been a focused topic in data mining research with a good number of references in literature and for that reason an important progress has been made, varying from performant algorithms for frequent itemset mining in transaction databases to complex algorithms, such as sequential pattern mining, structured pattern mining, correlation mining. Association Rule mining (ARM) is one of the utmost current data mining techniques designed to group objects together from large databases aiming to extract the interesting correlation and relation among huge amount of data. In this article, we provide a brief review and analysis of the current status of frequent pattern mining and discuss some promising research directions. Additionally, this paper includes a comparative study

between the performance of the described approaches.

Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. Innovation of association rules among the huge number of item sets is observed as a significant feature of data mining. The always growing demand of finding pattern from huge data improves the association rule mining. The main purpose of data mining provides superior result for using knowledge base system. Researchers presented a lot of approaches and algorithms for determining association rules. This paper discusses few approaches for mining association rules. Association rule mining approach is the most efficient data mining method to find out hidden or required pattern among the large volume of data. It is responsible to find correlation relationships among various data attributes in a huge set of items in a database. Studying Apriori algorithm, it is a illustration of an enhanced association rule mining algorithm, which supports to avoid the replication of same items. This paper discusses an enhanced version of Apriori algorithm that is concentrated on four characteristics namely, First data preparation and chooses the desired data, second produce itemsets that decides the rule constraints for knowledge, third mine k-frequent itemsets using the new database and fourth produce the association rule that sets up the knowledge base and offer better results. Another approach discussed in this paper are the HASH MAPPING TABLE and HASH_TREE tactics used to optimize space complexity and time complexity

Let us consider I5, which is the last item in L. I5 occurs in two branches of the

FP-tree:

(I2, I1, I5:1)

(I2, I1, I3, I5:1)

1 I5 is a suffix, so its corresponding two prefix paths are

¡ (I2, I1:1)

¡ (I2, I1, I3:1)

1 Its conditional FP-tree contains only a single path : (I2:2, I1:2); I3 is removed because its support count of 1 is less than the minimum support count

l The single path generates all the combinations of f

requent patterns:

¡ {I2,I5:2}

¡ {I1,I5:2}

¡ {I2, I1, I5:2}

l

1 For I4 exist 2 prefix path, which form the conditional pattern base:

¡ {{I2, I1:1},{I2:1}}

1 This generates a single-node conditional FP-tree:

¡ (I2:2)

L The frequent pattern: {I2, I1:2}

Association rule mining, one of the most important techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. This paper represents comparative study of association rule mining algorithm.

Keywords–Association rule mining, Apriori, AprioriTid, AprioriHybrid

Data mining is the process of extracting knowledge from large amount of databases whether database is relational, temporal, spatial, multimedia etc. Data mining is very popular technology some companies like IBM, ORACLE, and TCS are working on data mining. Data mining engine having different kinds of

approach like classification, clustering, association**,** outliers analysis**.** Each approach has a different technology like classification based on decision tree induction whether association based on **interesting patterngeneration**. Some people are saying data mining is the part of knowledge discovery in databases (KDD). We can say Data mining tools perform data analysis and may uncover important data patterns. This paper has a full attention on association rule. Association rule is very important tool for mining process it has two special chartersticis first one is **support** and another is **confidence.** Support gives total number of transaction of any particular item are occurring in datasets while confidence gives strength of a data in a dataset, we can say support is probability of A and B while confidence is conditional probability. Association rule based on these two characteristics. Different algorithms support association rule but this paper show only two very popular approach *Apriori and FP-tree* method. Both approach gives frequent patterns, and candidate generation; frequent pattern means those item sets that satisfy minimum support value. We can say in one-word association rule gives an "interesting pattern"; this paper give and tries how association rules generate interesting patterns form a huge amount of databases.

Info

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships*. Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

Mining association rules is an important task for knowledge discovery. We can analyze past transaction data to discover customer behaviors such that the quality of business decisions can be improved. Various types of association rules may exist in a large database of customer transactions. The strategy of mining association rules focuses on discovering large item sets, which are groups of items which appear together in a sufficient number of transactions. We propose a graph-based approach to generate various types of association rules from a large database of customer transactions. This approach scans the database once to construct an association graph and then traverses the graph to generate all large item sets. Empirical evaluations show that our algorithms outperform other algorithms which need to make multiple passes over the database

The process of data mining produces various patterns from a given data source. The most recognized data mining tasks are the process of discovering frequent itemsets, frequent sequential patterns, frequent sequential rules and frequent association rules. Numerous efficient algorithms have been proposed to do the above processes. Frequent pattern mining has been a focused topic in data mining research with a good number of references in literature and for that reason an important progress has been made, varying from performant algorithms for frequent itemset mining in transaction databases to complex algorithms, such as sequential pattern mining, structured pattern mining, correlation mining. Association Rule mining (ARM) is one of the utmost current data mining techniques designed to group

objects together from large databases aiming to extract the interesting correlation and relation among huge amount of data. In this article, we provide a brief review and analysis of the current status of frequent pattern mining and discuss some promising research directions. Additionally, this paper includes a comparative study between the performance of the described approaches.

Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. Innovation of association rules among the huge number of item sets is observed as a significant feature of data mining. The always growing demand of finding pattern from huge data improves the association rule mining. The main purpose of data mining provides superior result for using knowledge base system. Researchers presented a lot of approaches and algorithms for determining association rules. This paper discusses few approaches for mining association rules. Association rule mining approach is the most efficient data mining method to find out hidden or required pattern among the large volume of data. It is responsible to find correlation relationships among various data attributes in a huge set of items in a database. Studying Apriori algorithm, it is a illustration of an enhanced association rule mining algorithm, which supports to avoid the replication of same items. This paper discusses an enhanced version of Apriori algorithm that is concentrated on four characteristics namely, First data preparation and chooses the desired data, second produce itemsets that decides the rule constraints for knowledge, third mine k-frequent itemsets using the new database and fourth produce the association rule that sets up the knowledge base and offer better

results. Another approach discussed in this paper are the HASH MAPPING TABLE and HASH_TREE tactics used to optimize space complexity and time complexity

Association rule mining, one of the most important techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. This paper represents comparative study of association rule mining algorithm.

Keywords–Association rule mining, Apriori, AprioriTid, AprioriHybrid

ABSTRACT:

Data mining is the process of extracting knowledge from large amount of databases whether database is relational, temporal, spatial, multimedia etc. Data mining is very popular technology some companies like IBM, ORACLE, and TCS are working on data mining. Data mining engine having different kinds of approach like classification, clustering, association**,** outliers analysis**.** Each approach has a different technology like classification based on decision tree induction whether association based on **interesting patterngeneration**. Some people are saying data mining is the part of knowledge discovery in databases (KDD). We can say Data mining tools perform data analysis and may uncover important data patterns. This paper has a full attention on association rule. Association rule is very important tool for mining process it has two special chartersticis first one is **support** and another is **confidence.** Support gives total number of transaction of any particular item are occurring in datasets while confidence gives strength of a data in a dataset, we can say support is probability of A and B while confidence is conditional probability. Association rule based on these two characteristics. Different algorithms support association rule but this paper show only two very popular approach *Apriori and FP-tree* method. Both approach gives frequent patterns, and candidate generation; frequent pattern means those item sets that satisfy minimum support value. We can say in one-word association rule gives an "interesting pattern"; this paper give and tries how association rules generate interesting patterns form a huge amount of databases.

INTRODUCTION:

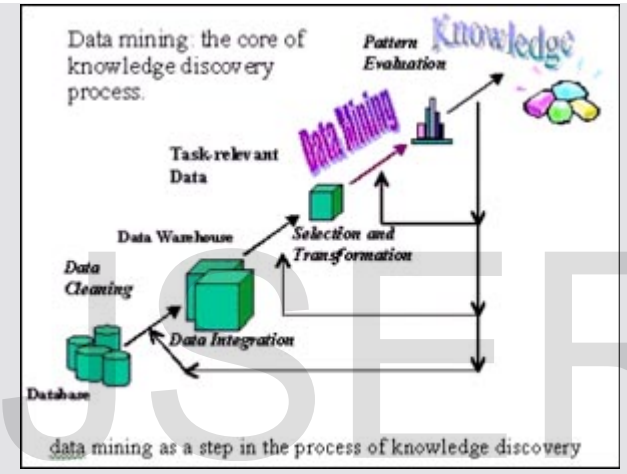
The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from business management, production control, and market analysis, to engineering design and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive

information that experts may miss because it lies outside their expectations. Data mining used different types of engine such as **classification, clustering, association, and outliers** .In classification class label is known whether in clustering class label is unknown, association gives to interesting pattern of the data while outliers give fraud detections.

What is Data Mining?

Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledgebase's, and scientific and medical research. The widening gap between data and information calls for a systematic development of *data mining tools.* So simply you can say**data mining** refers to *extracting* or *"mining"* knowledge from large amount of data.

KDD (Knowledge Discovery in Databases): Many people treat data mining as a synonym for another popularly used term, i.e. KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. KDD process is depicted in figure.



1. **Data Cleaning:** To remove noise and inconsistent data.

2. **Data integration:** Where multiple data source may be combined.

3. **Data Selection:** Where data relevant to the analysis task are retrieved from the database.

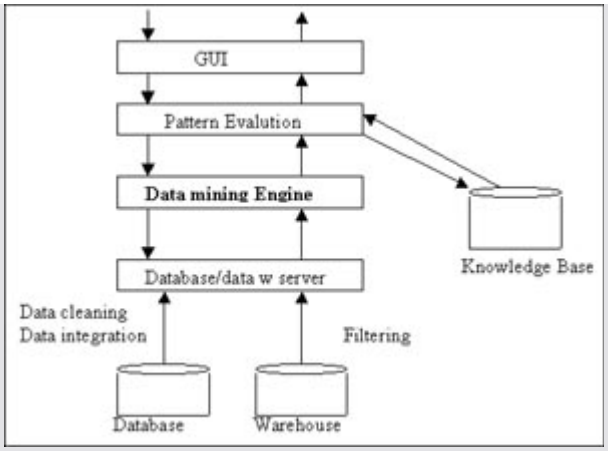
4. **Data Transformation:** Where data are transformed

5. **Data mining:** An essential process where intelligent methods are applied in order to extract data pattern.

6. **Pattern Evaluation:** To identify the interesting pattern.

7. **Knowledge presentation:** Where visualization and knowledge representation techniques are used.

Architecture of a data mining system: The Typical data mining architecture may have following major component.



The data mining components are

- **Database, data warehouse, or other information repository**
- **Database, data warehouse server**
- **Data mining engine**
- **Patterns evaluation model**
- **Graphical user interface**

What is association analysis? Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transition data analysis.

Association rule having two main important properties.

- **Support**
- **Confidence**

The definition of the **support and confidence** is

$$Support\ (AB) = P\ (AUB)$$

$$Confidence\ (AB) = P\ (B|A).$$

If we correlate support and confidence then

$$Confidence\ (AB) = P\ (B|A) = Support\_count\ (AUB)/Support\_count\ (A)$$

Where Support_ count (AUB) is the number of transaction containing the item sets AU B, and Support_

count (A) is the number of transactions containing the item set A.

"How are association rules mined from large databases?" **Association rule** mining is a two –step process:

1. **Find all frequent item sets:** By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count.

2. **Generate strong association rules from the frequent item sets:** By definition, these rules must satisfy minimum support and minimum support and minimum confidence.

APRIORI ALGORITHM: (Finding Frequent Item sets Using Candidate Generation)

Apriori is an influential algorithm for mining frequent item sets. The name of the algorithms is based on the fact that the algorithm uses *prior knowledge* of frequent item sets properties. Apriori employs an iterative approach known as a *level-wise* search.

To improve the efficiency of the level-wise generation of frequent item sets, an important property called the **apriori property, i.e.**" *all nonempty subsets of a frequent item sets must also be frequent."*

Apriori algorithms having a two-step process.

- **The join step:** To find L_k , a set of candidate k item sets is generated by joining L_{k-1} with itself. This set of candidate is denoted C_k.
- **The prune step:** C_k is the superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the databases to determine the count of each candidate in C_k would result in the determination of L_k. (i.e. all candidates having a count no less than the minimum support count are frequent by definition, and therefore belongs to L_k)


```
procedure AprioriAlg()
begin

    L1 := {frequent 1-itemsets};
    for ( k := 2; Lk-1 0; k++ ) do {
       Ck= apriori-gen(Lk-1) ; // new candidates
    for all transactions t in the dataset do {
      for all candidates c Ck contained in t do
            c:count++
      }
      Lk = { c Ck | c:count >= min-support}
    }
    Answer := k Lk

end
```

It makes multiple passes over the database. In the first pass, the algorithm simply counts item occurrences to determine the frequent 1-itemsets (itemsets with 1 item). A subsequent pass, say pass

k, consists of two phases. First, the frequent itemsets L_{k-1} (the set of all frequent (k-1)-itemsets) found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the apriori-gen() function. This function first joins L_{k-1} with L_{k-1}, the joining condition being that the lexicographically ordered first k-2 items are the same. Next, it deletes all those itemsets from the join result that have some (k-1)-subset that is not in L_{k-1} yielding C_k.

The algorithm now scans the database. For each transaction, it determines which of the candidates in C_k are contained in the transaction using a hash-tree data structure and increments the count of those candidates. At the end of the pass, C_k is examined to determine which of the candidates are frequent, yielding L_k . The algorithm terminates when L_k becomes empty.

FP-TREE GROWTH ALGORITHM:

Apriori algorithms suffer from the following two shortcomings:

1. It is costly to handle large numbers of candidate sets. For instance, 10^4 frequent 1-itemsets, then approximately, 10^7 candidate 2-itemsets are generated.

2. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching.

Keeping this in mind, a new class of algorithms has recently been proposed which avoids the generation of large numbers of candidate sets. We describe one such method, called the FP-tree growth algorithm. It is proposed by Han et al. The main idea of the algorithm is to maintain a frequent pattern tree of the databases.

A frequent pattern tree (or FP-tree) is a tree structure consisting of an item-prefix-tree and a frequent item-header table.

- Item- prefix-tree:
    - o It consists of a root node labeled null
    - o Each non-root node consists of three fields:
        - * Item name
        - * Support count
        - * Node link.

- Frequent-item –header-table: It consists of two fields;
    - * Item name

* Head of node link which points to the first node in the FP-tree

* Carrying the item name.

Association rules should not be used directly for prediction without further analysis or domain knowledge. They do not necessarily indicate causation. They are however a helpful starting point for further exploration, making them a popular tool for understanding data.

CONCLUSION:

The discovery of association relationship among huge amount of data is useful in selective marketing, decision analysis, and business management. A popular area of application is **market basket analysis**, which studies the buying habits of customers by searching for sets of item that are frequently purchased together. **Association rule mining** consists of first finding **frequent** item sets, from which **strong**association rules in the form of A B are generated. Association rule is the important tool for data mining engine. It is very popular technology now days. This paper have used only two approaches for association rule for mining the process named apriori and fp-tree. These two algorithms are very popular for mining rule. Last but not least we can say association rule gives interesting pattern to our customers.

References:

* http://www.kdnuggets.com/

* http://www.dmg.org/

* http://www.almaden.ibm.com/software/quest/

* http://www.data-mine.com/bin/site/templates/splash.asp

* http://www.thearling.com/

* www.indianmba.com

* **Data mining a book by A.K.PUJARI (University of Hyderabad)**

* **Data mining book by Dunham**

* **Research papers by Rakesh Aggrawal and S Srikanth**

Efficient Analysis of Pattern and Association Rule Mining Approaches

Association Rule Mining using Apriori Algorithm: A Survey

A study on Various Data Mining Approaches of Association Rules

An Overview of Association Rule Mining Algorithms

Comparative Evaluation of Association Rule Mining Algorithms with Frequent Item Sets

Performance Analysis of Association Rule Mining Algorithms

Comparative Study of Association Rule Mining Algorithms

Identifying Best Association Rules and Their Optimization Using Genetic Algorithm