RESEARCH ARTICLE OPEN ACCESS

Comparative Study of Apriori Algorithm Performance on Different Datasets

Prof. Prashasti Kanikar[1], Twinkle Puri[2], Binita Shah[3], Ishaan Bajaj[4], Binita Parekh[5]

[1](Assistant Professor, Department of Computer Engineering, MPSTME, MUMBAI)
[2](IV B.Tech, Department of Computer Engineering, MPSTME, MUMBAI)
[3](IV B.Tech, Department of Computer Engineering, MPSTME, MUMBAI)
[4](IV B.Tech, Department of Computer Engineering, MPSTME, MUMBAI)
[5](IV B.Tech, Department of Computer Engineering, MPSTME, MUMBAI)

ABSTRACT
Data Mining is known as a rich tool for gathering information and apriori algorithm is most widely used approach for association rule mining. To harness this power of mining, the study of performance of apriori algorithm on various data sets has been performed. Using Java as platform implementation of Apriori Algorithm has been done and analysis is done based on some of the factors like relationship between number of iterations and number of instances between different kinds of data sets. Conclusion is supported with graphs at the end of the paper.

***Keywords* -** Association Rule Mining, Confidence, Data Mining, Data Warehousing, Knowledge Discovery Process

I. INTRODUCTION

This review paper aims at studying data mining its association rules (Apriori), at depth to identify the relationship between several factors. The efficiency of an algorithm for data mining shall be measured by reducing the number of iterations and set generation for determining the strong association rules. These refined algorithms will be then applied on random sample data sets and results shall be compared to determine consistency and long term performance.

II. WHAT IS DATA WAREHOUSE

Warehousing – A Large collection of operational data-marts which consolidates data from several data sources which also includes some flat files, Redo & Archive Logs

Characteristics of a Data Warehouse –

- Subject Oriented
- Non Volatile
- Time Variant
- Integrated

Data Warehousing works hand in hand with the Data Mining Tool, it is essential to implement an efficient data mining system for any Data Warehouse to bring out the optimum results required from that chunk of Data.

III. WHAT IS DATA MINING

Mining – Strategic Information derived from random Data has always been of prime importance for all Decision Support Systems and Top Level Management of Organizations. Mining is that fundamental component acting upon a Data Warehouse that helps a lay user extract vital information from the huge chunk of data which holds relevance to the kind of information one is looking out for. Mining helps the user narrow down to that specific item in the Data Set that is of consideration for the particular scenario. It is extremely important to implement an efficient mining utility. In simple terms Mining is similar to query firing on a normal database table, But the difference lies in the size of the Data, for a regular database table there might be a few hundred entries relevant to a particular topic and sizing to around a few thousand Kilobytes, whereas considering a Data warehouse which is a primary archival source of data the size of the Data is in a few thousand Megabytes, spreading equally far breadth and depth wise. Looking up one line of data in this type of a store is equivalent to finding a needle in the hay bundles, nearly impossible. Well in the case of digital query firing, retrieving the data would not be so difficult but it would take immensely long amount of time to complete the query which is not feasible in this day and age of technology as the data that is being fetched has to be fresh and up to date.

Data Mining comes into play, similar to its younger counterpart, mining is like query firing but embedded with intelligence of its own to take better decisions thus saving time and iterations in looking up the Data.

Based on various algorithmic techniques several types of data mining algorithms are implemented and studied constantly to serve the purpose and simultaneously improvise their performance in the near future.

Various Data Mining Algorithms –

1. Association Rule Mining
2. Clustering
3. Anomaly Detection
4. Classification
5. Regression
6. Summarization

This research review paper shall concentrate on the technicalities and performance measure of the Apriori Algorithm that is designed under the Association Rule Mining Category.

A. Goals of Data Mining

1. To extract information from a Data Set.
2. Transform it to an understandable structure for better use.

B. Raw Analysis involved in Data Mining

1. Extract Data from all Valid Sources.
2. Cleaning and Formatting Data at the Staging area to make it compatible with each other and ready to be loaded into the Data Warehouse.
3. Transformation – Data is transformed into Tabular Form.
4. Now Patterns are to be identified from the Data using various mining algorithms like FP-Tree, Clustering etc.
5. Data Visualization and Interpretation of outcome takes place.

This process of Raw Analysis is the first step at making the Raw Data in the Data Warehouse useful with the means of Data Mining Algorithms, these help a common man easily interpret and utilize the given information.

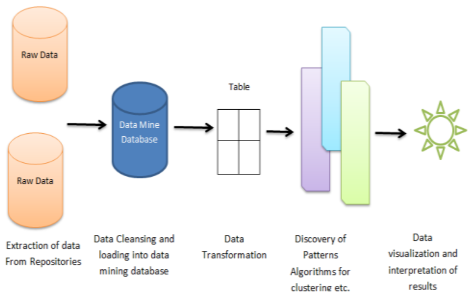

Raw Data
Raw Data
Data Mine Database
Table
Extraction of data From Repositories
Data Cleansing and loading into data mining database
Data Transformation
Discovery of Patterns Algorithms for clustering etc.
Data visualization and interpretation of results

Fig 1.1 – Steps in Data Visualization & Analysis

IV. Association Rule Mining (Apriori Algorithm)

As the name itself speaks, Association rule mining is based on the relational or more precisely conditional aspect of patterns occurring in the environment or data set.

For understanding purpose with respect to this algorithm a very commonly observed day to day list of activities –

1. Say a person goes to a Mobile store and Purchases a New Mobile, so what are the chances of that person for choosing the next available accessories for that mobile? This choice is not a matter of sheer marketing of the salesperson but it will also be based on the utility, need and budget of the purchaser, let us see for ourselves .
   a) Mobile Cover-INR 400-For a INR 20,000Mobile handset it is evident that a person will purchase a Cover/Case to protect his valuable phone from the wrath of the physical environment. (STRONG RULE).
   b) Premium Stereo Headset – INR 1200 – If the Person is not so much into music and is satisfied with the OEM headset packed along with the phone, there is no point he would dive in for one of these (WEAK RULE)
   c) Power Bank – INR 2000 – Now this is an accessory where Utility v/s Value for Money comes in, if a person doesn't have heavy mobile usage the utility for him is negligible, on the other hand for a person with more usage it is a matter of calculating whether keeping a power bank or a spare battery is more effective and based on that the decision might be related (Adequate confidence support required).

Thus from the above example it can be deduce the functioning of the Association Rule Mining, and for a matter of fact, this rule varies with diversity in population, region, individual likes and dislikes and several other parameters, therefore in spite of being a simple algorithm to implement the outcomes are always varied based on the Data Set onto which the algorithm is applied taking into account environmental factors.

A. Association Rule Mining – The Algorithm

The major applications of Association Rule Mining are :

a) Basket Data Analysis
b) Cross Marketing

Let us consider a given set of items for understanding the Association Algorithm,

TID	Items
1	**Bread, Butter, Milk, Jam**
2	**Bread, Jam, Milk**
3	**Bread, Butter**
4	**Bread, Milk**
5	**Bread**

Table 1 – Association Example Data Set

With these given transactions we observe that the most commonly purchased transaction from the store is BREAD, thus our data mining approach shall be oriented around the fact that what all shall go along with bread and what item has the maximum demand.

The Association Rules for the following shall be thus made as follows –
{Bread} □ {Butter}
{Bread} □ {Milk}
{Bread} □ {Jam}

And based on their repeated occurrence in the transactions, it can deduce strong and weak rules for the given Item Set.

B. APRIORI ALGORITHM

Apriori Algorithm is used for large transactional databases and it is very influential algorithm over other algorithms as other algorithms are derived from this algorithm.

Over the years several improvisation techniques have been applied on Apriori and a resultant of which is the newer algorithm known as FP-Tree

Apriori is a Bottom-up generation of Frequent item set combinations whereas on the other hand FP-Tree Generation of Item Set is based on either Divide and Conquer or Partitioning Rule.

After discovering the disadvantages of FP-Tree Algorithm, "Győrödi" introduced an improvised version of the same known as DYN FP-Tree.

APRI	FP
Principle : Uses generate and test approach *If an item set is frequent ,then all its subset must be frequent	**Principle :**Allows frequent set generation without candidate set generation
Advantages : For large Transactional Databases	**Advantages :** Compact Data structure is created
Disadvantages : *Generation of Candidate sets is Expensive *Support Counting	**Disadvantages:** Resulting FP-Tree is not unique for same logical

Table 2 – Comparison between Apriori & FP-Tree

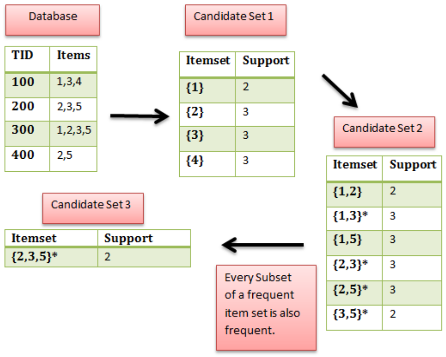
V. IMPLEMENTING THE ALGORITHM

After consideration of all aspects of Association Rule Mining and our algorithm under consideration, Apriori, now let us proceed to test the algorithm in a real time analysis on a sample database containing random data sets.

The algorithm for the following program implemented in APRIORI is as follows –

a) Data Accumulation – The initial stage for running any mining algorithm is data accumulation, which is done in the following manner for Association Rule Mining :

a. Enter Number of Transactions
b. Enter Items per transactions
c. Enter Minimum Support Count
d. Enter Threshold Confidence Level



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

Itemset	Support
{1}	2
{2}	3
{3}	3
{4}	3

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

Itemset	Support
{2,3,5}*	2

Fig 1.2 – Example for Understanding APRIORI Algorithm

b) Data Mining – The major part about the programs role to process the stored data lies in this section. The working of APRIORI algorithm is based on the rule of sufficient repeat occurrence of a particular item in a transactional data set, enough to maintain the minimum support count requirement. The fundamental pattern of APRIORI Working is described ahead –

Step 1 – For the given transactional dataset the algorithm checks for the number of items in singularity and the frequency of occurrence for the same.

Step 2 – As seen in the image, the ITEMS {1, 2, 3, 5} are identified and the first candidate set is made that lists down its frequency of occurrence.

Step 3 – This candidate set is then checked with the minimum support count given to identify the ITEMS that are above the support count.

Step 4 – After eliminating the ITEMS that fallout from the requirement the remaining ITEMS are then clubbed to create another candidate set that will check their occurrence in a group of TWO, THREE, and FOUR and so on as far as possible.

Step 5 – Each time a candidate set is compared to the support count, it is checked that there are ITEMS or Group of ITEMS that surpass the minimum threshold value in order to make the new candidate set, once this cannot be achieved the generation of candidate steps is terminated and the Association rules are derived from the last generated candidate set and the first ITEM set

Step 6 – Identification of STRONG and WEAK association rules, an association rule that is generated from the ITEM Sets derived from the above process is then compared to the threshold value of the particular transactional DATASET. The best understanding of the Threshold value can be explained as the minimum tolerance level of an association not taking place below the acceptable amount. For example, if there is a Rule that says that 8 of 10 people who bought Milk also purchased Bread states that if Milk □ Bread rule has to be made and since Bread occurred in 8 Transactions of Milk it is 80% Confidence and given threshold is 70% thus, it is affirm that this rule is a STRONG rule. On the other hand if the number of people purchasing Bread would have been 6 the confidence level of the transaction would have been 60% which is below the given threshold level and thus the rule stands to be a WEAK Rule.

The APRIORI Algorithm – Program Flow can also be described with the help of a logical flowchart as defined below-

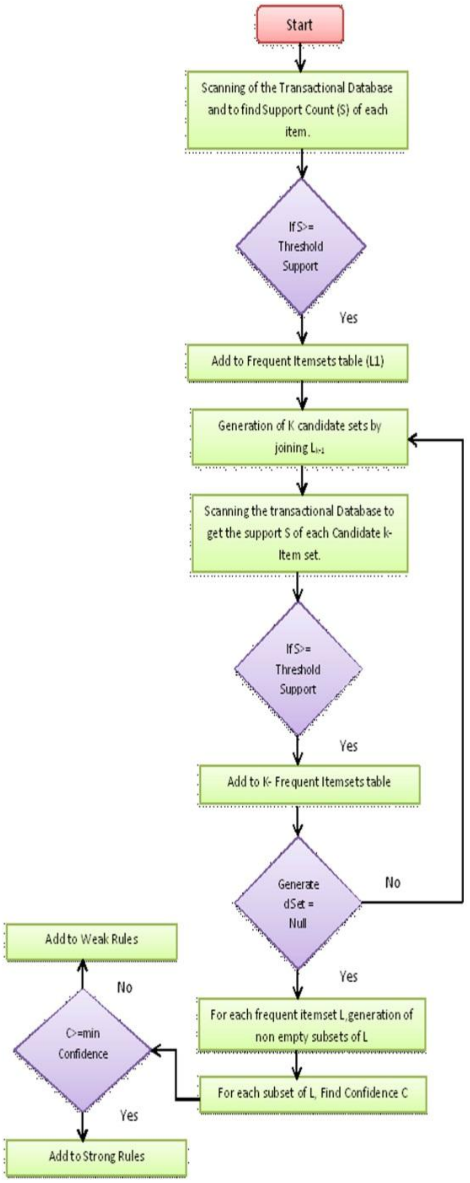

Start
Scanning of the Transactional Database and to find Support Count (S) of each item.
If S>= Threshold Support
Yes
Add to Frequent Itemsets table (L1)
Generation of K candidate sets by joining L_{k-1}
Scanning the transactional Database to get the support S of each Candidate k-Item set.
If S>= Threshold Support
Yes
Add to K- Frequent Itemsets table
Generate dSet = Null
No
Yes
For each frequent itemset L,generation of non empty subsets of L
For each subset of L, Find Confidence C
C>=min Confidence
No
Add to Weak Rules
Yes
Add to Strong Rules

Fig 1.3 – Flowchart Decision Making (APRIORI)

VI. Analysis on Random Data Sets

Large Dataset

It consists of the detailed products which are bought in a supermarket. The example taken is a large dataset of the same which consists of 4627 instances.

Some of the Attributes include:

a) bread and cake
b) biscuits
c) tea
d) canned vegetables etc.

Using these attributes and various combinations of the food items purchased by the customer so as to analyze the buying patterns and plan the marketing strategies accordingly.

The purpose for using these datasets is to understand the relationship between the no of instances in a dataset and the no of iterations or cycles.

Medium dataset

It consists of the detailed products which are bought in soya bean database. The example taken is a medium dataset of the same which consists of 683 instances.

Some of the Attributes include:

a) date
b) plant-stand
c) precip
d) temp
e) area-damaged
f) severity etc.

Using these attributes and various combinations of them this has been analyzed the characteristic patterns of germination and planting of soya bean seeds.

The purpose that these datasets have been used is to understand the relationship between the no of instances in a dataset and the no of iterations or cycles.

Small dataset

It consists of the detailed products which are bought in contact-lens database. The example taken is a medium dataset of the same which consists of 24 instances.

Some of the Attributes include:

a) spectacle-prescrip
b) astigmatism
c) tear-prod-ratetables etc

Using these attributes and various combinations of them the characteristic patterns of contact lenses have been analyzed. The purpose that these datasets have been used is to introduce the relationship between the no. of instances in dataset and no.of iteractions or cycles.

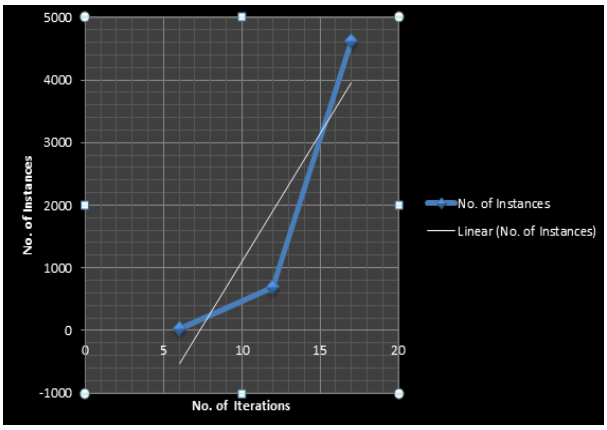


Fig 1.4 Graphical Representation of Relationship between no. of Instances and no. of Iterations

Name of Dataset	Size of dataset	Number of instances	Number of cycles(iterations)
Contact-lens	Small	24	6
Soya bean	Medium	683	12
Super-Market	Large	4627	17

Table 3-Descriptive Analysis of Datasets Used

C. OUTCOME

The study has been made and the following results were obtained in regards to relationship between Space occupied and the number of candidate sets generated.

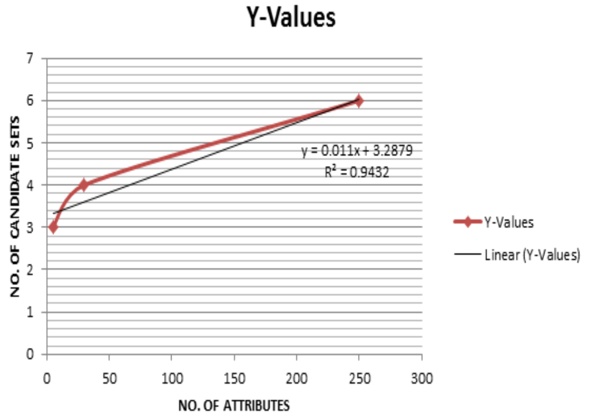


Fig 1.5 – Graphical Representation of relationship between no. of Attributes and no. of candidate sets

Conclusion from the above graph is the directly proportional relationship between the two factors and thus bolsters the conclusion that the larger the attributes the greater are the no of candidate sets generated.

Example: Larger number of attribute then Larger will be candidate sets

VII. CONCLUSION

After studying the algorithm on Association Rule Mining using the APRIORI technique, it has been analyzed that this technique is useful for only small datasets ,on larger datasets it takes lots of space as well as lots of time in generation of candidate sets and finding frequently occurring items from huge datasets is a big problem as seen in the above graphs, as the size of Data increases the algorithm takes more number of Data Scans and eventually greater number of iterations for deriving the strong association rules. Also it was noticed more the no. of instances more will be the no of iterations performed. This is not feasible for further development of the algorithm to be used to mine larger volumes and large file size worth of data in several formats, as the load on the system and algorithm shall increase tremendously resulting in frequent crashes and erroneous results.

Based on our comparative study of Data Mining algorithms a consideration for improvising the APRIORI algorithm is the Divide and Conquer Depth wise method used in FP-TREE Algorithm, the same dataset shall be passed through the FP-Tree and Dynamic FP-Tree algorithmic programs to compare and improvise the search outcome and hence develop a better application for larger volume based quicker data mining.

VIII. ACKNOWLEDGEMENTS

This experiment was successfully accomplished under the sole guidance of our esteemed mentor Prof. Prashasti Kanikar, Our Sincere Thanks and deep gratitude to her for her excellent mentorship, motivating support and thorough aid in understanding the subject and its complexities.

REFERENCES

[1] A Comparative study of Association rule Mining Algorithm – Review Cornelia Győrödi*, Robert Győrödi*, prof. dr. ing. Stefan Holban

[2] Image Processing and Image Mining using Decision Trees - KUN-CHE LU AND DON-LIN YANG

[3] IMAGE MINING TECHNIQUES: A LITERATURE SURVEY - Mahendra S.Makesar

[4] An Experiential Survey on Image Mining Tools, Techniques and Applications – C. Laxmi Devasena

[5] http://ijana.in/papers/6.11.pdf

[6] http://www2.cs.uregina.ca/~dbd/cs831/notes/itemsets/itemset_apriori.html

[7] http://staffwww.itn.liu.se/~aidvi/cou rses/06/dm/lectures/lec7.pdf

[8] http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=image+mining+algorithms&x=0&y=0

[9] http://www2.ims.nus.edu.sg/preprints/2005-29.pdf4

[10] http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec7.pdf

[11] http://archive.ics.uci.edu/ml/

[12] http://www.kdnuggets.com/datasets/