Comparative Exploration on Association Rule Mining Algorithms

Neha Walia[1], Arvind Kalia[2]
[1]*Research Scholar, Department of Computer Science, Himachal Pradesh University, Shimla, Himachal Pradesh*
[2]*Professor, Department of Computer Science, Himachal Pradesh University, Shimla, Himachal Pradesh*

Abstract: Rapid technological enhancements resulted in existence of oversized raw data in this digital age. Extraction of patterns/associations with existing simple models is a cumbersome task. One of the technique/algorithms to find correlation among the databases is achieved through association rule mining.The core emphasis of this paper is to analyse and compare the association rule mining algorithms namely Apriori and FP-Growth. The parameters considered are efficiency, techniques used, memory usage and so on.The study reveals that FP growth algorithm is more efficient, less time and space consuming as compared to Apriori algorithm.

Keywords: Data Mining, Association Rule Mining, Apriori Algorithm, FP-Growth algorithm, Frequent Itemsets, FP-Tree.

I. INTRODUCTION

Due to the rapid enhancements in both hardware and software in this era, enormous data in digital format is produced in every next second andavailable on our fingertips. Transactional databases, relational databases, multimedia databases, temporal databases, spatial databases, www to name the few are the main sources of this gigantic data. One of the major challenges of this age is to search trends, patterns and relationships on this data through simple models, making crispand useful information and further knowledge. Since the valuable and interesting knowledge are not mined for decision making from this unprocessed data with the prevailing traditional data analysis tools and methods, so there is need to develop a technique through which analysis of such raw data is possible. Data mining is the computing method of analysing data to determine unknown and unidentified patterns and predict valuable trends. Data mining includes various steps which comprises of collection, access and warehousing of data and data mining as a final point. Data mining is also defined as sequence of action for the exploration of data in innumerable means to discern unknown associations which are unfamiliar and possibly valuable information for decision making. This is vital in direction to envisage yet to come trends and behaviours and to make proactive decisions [10]. Different data mining techniques have been studied to process and analyse several types of data patterns, where the most popular data mining tasks are Classification, Summarization, Association Rule Mining and Clustering [6]. One of the significant techniques of Data mining is Association Rule Mining [16] .Extractions of interesting associations and patterns among interrelated itemsets from gigantic databases are accomplished by this approach. Association rules are extensively used in innumerable areas such as market basket analysis[12], medical diagnosis[3][5], census data[2][4] , protein sequences [1][13], Customer Relationship Management[7][12][17] to list a few. Generation of the frequent itemsets from the dataset is the prerequisite condition for Association Rule Mining. Association Rule Mining involves frequent itemset generation followed by discovering the rules. Those itemsets in the dataset which exceeds the minimum support are considered as frequent itemset and rules are identified from these frequent itemsets having confidence greater than minimum support [14][18].Association rules are if/then statements consisting of an antecedent (if) and a consequent (then) part [8].

II. ALGORITHMS OF ASSOCIATION RULE MINING

Numbers of algorithms for Association Rule Mining areexisting. To name a few AIS [6], SETM [11], AprioriTID [15], AprioriHybrid[15]. In this paper, only two very significantalgorithms are analysed for the generation of best rules from the gigantic databases.

2.1 Apriori Algorithm

Apriori is a conventional algorithm, anticipated by Rakesh Agrawal and RamakrishnanSrikant in the year 1994 for finding frequent itemsets in a dataset for boolean association rule[15]. It was named as Apriori as the algorithm uses previous knowledge of properties of frequent itemset. Apriori algorithm works on anti-monotonicity property of the support measure. It adopts that all subsets of a frequent itemset must be frequent. It uses bottom up approach. Algorithm followed by Agrawal and Srikant are illustrated in figure1.

L1 ={large 1-itemsets};
for (k=2; Lk-1≠ϕ ;k++)

```
   do begin
Ck=                                                    Apriori-gen9(Lk-1);
forall transactions tϵ  D
        do begin
          Ct =subset(Ck ,t);  Candidates contained in t
forall candidates c ϵ  Ct
            do
c.count++;
       end
       Lk = {c ϵ  Ck | c.count>= minsup}
end
Answer=U k Lk ;
```

Figure 1: Algorithm for Apriori[15]

Features

- Works on generate and test approach.
- Uses bottom up approach.
- Downward closure property [15].
- Generation of candidate itemset.
- Apriori property.
- Use of union and pruning technique.
- Involves two step approachi.e., frequent itemset generation and rule generation.
- Database is read for eachiteration.

Advantages

- Simple and easy to implement.
- Generate rules on gigantic databases.
- Supported by varied tools.

Disadvantages

- Not an efficient algorithm.
- Generation of candidate itemset is expensive.
- Support calculation is costly.
- Execution time is high.
- Requires more storage space.

2.2 Fp-Growth Algorithm

As there were number of restrictions in Apriori algorithm, in the year 2000 Jiawei Han, Jian Pei and Yiwen Yin proposed Frequent Pattern Growth algorithm. Algorithm uses the concept of tree generation for frequent itemsets insteadof candidate frequent itemset generation. It is based on three main conventions i.e. construction of FP tree, FP tree based pattern fragment growth mining method and partitioned based divide and conquer technique.Efficacy of this algorithm is highly increased because it uses FP direct pattern generation from single tree and last frequent events as suffix [9]. Figure 2 and 3depicts the algorithm proposed by Han et al.

Algorithm 1 (FP-tree construction)
Input: A transaction database DB and a minimum support threshold.
Output: Its frequent pattern tree, FP-tree
Method: The FP-tree is constructed in the following steps:
1. Scan the transaction database DB once. Collect the set of frequent items F and their supports. Sort F in support descending order as L, the list of frequent items.
2. Create the root of an FP-tree, T, and label it as “null". For each transaction Trans in DB do the following.

Select and sort the frequent items in Trans
According to the order of L. Let the sorted
Frequent item list in Trans be [p|P], where p is the first element and P is the remaining list. Call insert tree([p|P], T)

Figure 2: Algorithm for FP –Tree generation [9]

Algorithm 2 (FP-growth: Mining frequent patterns with FP-tree by pattern fragment growth)

Input: FP-tree constructed based on Algorithm 1, using DB and a minimum support threshold .
Output: The complete set of frequent patterns.
Method: Call FP-growth (FP-tree , null).

```
Procedure FP-growth (Tree, α)
{
if Tree contains a single path P
then for each combination (denoted as β)
of the nodes in the path P do
   generate pattern βUα with support =
   minimum support of nodes in β ;
else for each ai in the header of Tree
 do
 {
   generate pattern β= ai Uα  with
   support = ai.support;
   construct β 's conditional pattern base
   and then β 's conditional FP-tree Treeβ;
   if Treeβ ≠ϕ  ;
   then call FP-growth (Treeβ, β)
  }
```

Figure 3: FP-Growth Algorithm[9]

Features

- * Works on the concept of depth-first approach. .
- Uses frequent pattern tree for frequent itemset generation and generate rules directly from this tree.
- Apply divide and conquer strategy.
- Reads the file only twice.

Advantages

- Fast and efficient algorithm.
- No candidate generation.
- Execution time is small.
- Requires less storage space.

Disadvantages

- Not easy and simple algorithm.
- For large databases, FP tree does not fit in memory.

Generation of the tree is costly.

III. COMPARISON OF ASSOCIATION RULE MINING ALGORITHMS

Two widespread algorithms for association data mining namely Apriori algorithm and FP growth algorithm along with their important features, advantages and disadvantages are considered for this study. Comparisons of these algorithms are organizedin tabularizedlayout.

Table 1: Apriori v/s FP- Growth Algorithm

Parameters/Factors	Apriori	FP-Growth
Processing speed	Slow	Fast
Internal storage structure	Array	Tree
Searching Technique	Breadth first	Depth first
Database scan	Multiple	Only two
Candidate generation	Yes	No
Memory	Large	Small
Technique	Apriori property	Construction of pattern tree
Runtime complexity(with increase in dataset size)	Exponential	Linear
Efficiency	Low	High
Approach	Generate and test	Divide and conquer
Operations	Count accumulation and prefix count adjustment	Candidate set generation and pattern matching

Comparison of Apriori algorithm and FP-Growth algorithm are depicted in Table1. Comparison is done on the basis of various parameters/factors such as speed, searching technique, candidate generationand efficiency and so on.

IV. CONCLUSION AND FUTURE SCOPE

The paper contributes a crispexplanation to the data mining and association rule mining. Two very important algorithms: Apriori and FP Growth were considered in the study. Important characteristics along with their pros and cons of these algorithms are discussed followed by comparative analysis. The comparison reveals that FP growth algorithm is more efficient, less time and space consuming as compared to Apriori algorithm.In future, these algorithms will be compared with more existing algorithms.

V. REFERENCES

[1] C. Branden and J. Tooze, “Introduction to Protein Structure”, Garland Publishing inc, New York and London, 1991.

[2] D. Malerba, F. Esposito and F.A. Lisi, “Mining Spatial Association Rules Iin Census Data”, In Proceedings of Joint Conference on New Techniques and Technologies for Statistics and Exchange of Technology and Know-how, 2001.

[3] D. Gamberger, N. Lavrac and V. Jovanoski, “High Confidence Association Rules for Medical Diagnosis”, In Proceedings of IDAMAP99, pp. 42-51.

[4] G. Saporta, “Data Mining and Official Statistics”, In Proceedings of Quinta ConferenzaNationale di Statistica, ISTAT, Roma, 15 November, 2000.

[5] G. Serban, I. G. Czibulaand A. Campan, "A Programming Interface for Medical Diagnosis Prediction", StudiaUniversitatis, "Babes-Bolyai", Informatica,LI(1), pp. 21- 30, 2006.

[6] Han Jiawei, Micheline Kamber and Jian Pei, “Data Mining: Concepts and Techniques” Third Edition, ISBN 978-0-12-381479-1, Morgan Kaufmann Publishers, 2011.

[7] H. S. Song, J. K. Kim and S. H. Kim, “Mining the Change of Customer Behavior in an Internet Shopping Mall”, Expert Systems with Applications, 2001.

[8] Https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining Retrieved on December 20, 2019.

[9] J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”,ACM SIGMOD Record, Vol. 29, No.2, pp. 1–12,2000.

[10] M.Dunham, “Data Mining Introductory and Advanced Topics”. ISBN 978-0-13-088892-1, Prentice Hall, 2003.

[11] M. Houtma and A. Swami, “Set-Oriented Mining of Association Rules”. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.

[12] M. J. Shaw, C. Subramaniam , G. W. Tan and M. E. Welge, “Knowledge Management and Data Mining for Marketing”, Decision Support Systems, Vol.31, No.1, pp. 127-137, 2001.

[13] N. Gupta, N. Mangal, K. Tiwari and P. Mitra, “Mining Quantitative Association Rules in Protein Sequences”, In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining – AUSDM, 2006.

[14] P. Tan, V. Kumar and J Srivastava (2004). Selecting the Right Interestingness Measure for Association Patterns”,In Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , pp. 32-41.

[15] R. Agrawal and R. Srikant,“Algorithms for Mining Association Rules”. In Proceedings of the 20th International Conference.on Very Large Databases (VLDB '94), Santiago, Chile, June 1994.

[16] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”.In Proceedings of the ACM SIGMOD International Conference on Management of Data - SIGMOD '93, pp. 207 Washington, D.C., May 1993.

[17] R. S. Chen, R. C. Wu and J. Y. Chen, “Data Mining Application in Customer Relationship Management of Credit Card Business”, In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05), Vol. 2, pp. 39-40.

[18] S.J. Lee and S. Keng,“A Review of Data Mining Techniques”, Industrial Management & Data Systems, Vol. 101 No. 01, pp. 41-46, 2001.