www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 10 Oct 2015, Page No. 14649-14652

Quantitative Analysis of Apriori and Eclat Algorithm for Association Rule Mining

Tanu Jain* Dr. A.K Dua Varun Sharma
Amity University Jaipur
Hod (Cse), Amity University Jaipur
Amity University Jaipur
Tanujain.Engg@Yahoo.Com
Akdua@Jpr.Amity.Edu
Vsharma@Jpr.Amity.Edu

Abstract—***Apriori and Eclat algorithms are the mostly used algorithms in the area of association rules mining. They are generally used for mining of frequent item sets and to discover associations between these frequent item sets. R is a domain specific language for data analysis and analytics. It is already being used across different disciplines from Computer Science to Social Sciences. In this research a qualitative and quantitative analysis of Apriori and Eclat algorithms is done using R Environment. Different R-Packages and libraries are used for the access of different datasets and their connectivity with R. In this research, both algorithms have been implemented using different data sets and are further analysed on the basis of their performance. The performance analysis is based on total execution time taken by these algorithms in order to identify their quantitative performance and speedup with different volume of datasets.***

Keywords— Apriori, Eclat, Association Rule Mining, R language, R Environment

INTRODUCTION

Association rule learning is a standout amongst the most prevalent and generally utilized process as a part of request to locate the intriguing relations among variables inside of substantial databases and datasets. We generally need to discover solid standards to be found in expansive databases utilizing distinctive measures of performance matrices. Quantitative Analysis just is the study of an occasion, basically a money related business, by method for complex scientific and factual demonstrating according to the standard definition.

.

A. *Apriori Algorithm*

This algorithm has been often utilized for mining of frequent item sets and to find associations. The real distinction in Apriori is the less hopeful item sets it produces for testing in every database pass. The quest for association guidelines is guided by two parameters: support and confidence.
Apriori gives back an association guideline on the off chance that its support and confidence qualities are above client characterized limit values. It is a breath first search algorithm. The yield is requested by confidence. On the off chance that few principles have the same certainty then they are requested by support. In this manner Apriori supports more certain tenets and portrays these guidelines as additionally intriguing.

B. *Eclat Algorithm*

Eclat creates less number of succession tables which sets aside less time for frequent accessed patterns to examples when contrasted with Apriori. In apriori if huge data is their then it takes colossal time to create the successive frequent accessed patterns.
Eclat execution speaks to the arrangement of exchanges as a bit network and meets columns give the backing of thing sets. It takes after a profundity first traversal of a prefix tree.

C. *R-Programming*

R is a space specific dialect for information investigation, which is, as indicated by a few measures, the most prevalent such stage. The dialect grammar has establishes in the 1970's at Bell Labs, and was produced specifically as a dialect for factual information investigate.
R is a free, open-source dialect and vitally, is now utilized generally crosswise over orders (from Computer Science to Social Sciences -, for example, Political Science). All things considered, giving uninhibitedly available programming to utilize established calculation frameworks as a component of a R bundle will permit specialists and analysts from an extensive variety of regions to make utilization of the examination grew by the apriori and éclat research group who thus can consolidate the conceivably endless measure of criticism and true sending data into their exploration.

D. *Problem Statement*

Apriori and Eclat both are the well-known and highly used algorithms. We need to find that which one is better in which scenario? When to apply Apriori algorithm and when to apply

Eclat algorithm while keeping their performance at their best. Which algorithm should be apply to achieve the various requirements of association rules mining that includes confidence, support, lift and other performance matrices.

Literature Survey

The procedure of finding the incessant thing sets inside of an arrangement of exchanges is a surely understood technique for an issue essentially known as business sector wicker bin investigation, this is utilized as a part of request to discover regularities or regular examples from the shopping propensities for buyers of shopping entries, grocery stores, organizations that backings mail-request, on-line shopping locales and so forth. Specifically, it is expected to recognize the arrangements of items that are purchased together all the time.

The working of Apriori calculation is decently relies on the Apriori property which expresses that" All nonempty subsets of a regular item sets must be frequent".

The real issue of distinguishing the regular item sets, or the item sets that exists in a client determined transactions, is that there are different conceivable sets, which renders naıve approaches incapable as a result of their costly execution time. Among all these most prominent and modern systems two algorithms known under the names of Apriori and Eclat are generally utilized. Both algorithms works on a top down looking approach in the subset grid of items.

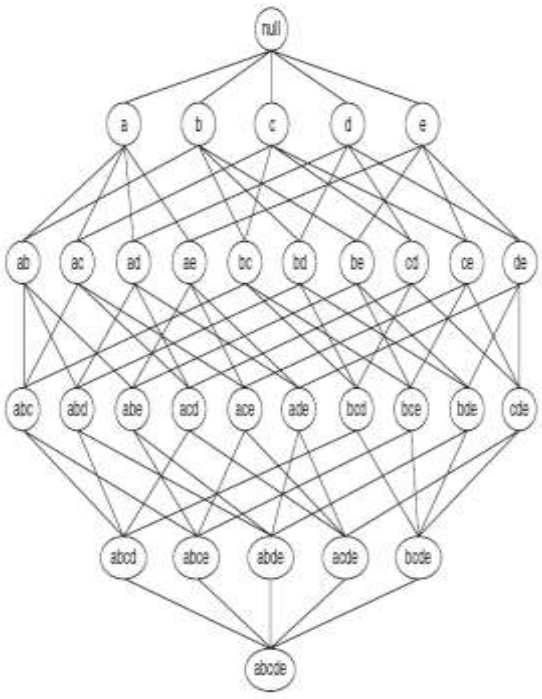

Fig. 1: A subset lattice for five item sets

The edges in above figure show subset relations among different item sets. The **significant contrasts in the middle of Apriori and Eclat** is the methodology that how they navigate the prefix tree and how they compute the support for an item set, that is the aggregate number of exchanges the item set contained inside.

Apriori essentially traverses the tree in BFS (breadth first search), it implies it first checks for the item set of size 1 and after that further for the item arrangement of size 2 and so on. Apriori surveys the support of item sets may be by checking each of item set which the exchanges contains inside, or by navigating for an exchange each subset of the most as of late handled size and expanding the related item set counters.

Eclat, explores the prefix tree top to bottom first demand, being backwards to apriori. It just implies that, it broadens a thing set prefix until and unless it reaches to the limit in the middle of the rare and continuous item sets and afterward further backtracks to process the nearing prefix. Eclat figures the support of each and every item set by making the rundown of the considerable number of identifiers of exchanges that contain the item sets. It utilizes the methodology of crossing two arrangements of exchange identifiers for two distinctive item sets just by a solitary item or together frame the thing set as of late prepared.

E. *The R Programming*

R is popular among organizations as a statistical and programming language. It is a software environment used in graphics and data analysis. The R language is mostly used among the statisticians and data scientist for the development of statistical applications and analysis of data. [7] R is a free and open source software environment used for statistical computing. It compiles and executes on a wide range of UNIX based operating systems, Windows based and Mac based operating systems. [8]

In the past decade, the momentum coming from academia as well as from industry has settled the R programming language as one of the most important tool for data analysis, computational statistics, and data visualization and for data science.

From all over the world, millions of data scientists and statisticians uses R to solve their most challenging problems related to data in fields ranging from computational biology, data science to quantitative marketing analysis.

R is one of the most popular language for data science as well as an essential tool for Finance analysis and analytics-driven companies such as Google, Facebook, and LinkedIn.

F. *Why R?*

1. R is free programming. R is an official GNU extend and circulated under the Free Software Foundation General Public Permit (GPL).
2. R is an effective information investigation bundle with numerous standard and forefront measurable capacities. See the Far reaching R Archive Network (CRAN's) Task Views to get a thought of what you can do with R [8].
3. R is a programming dialect, so its capacities can without much of a stretch be reached out through the utilization of client characterized capacities. A huge gathering of client contributed capacities and bundles can be found in CRAN's Contributed Packages.
4. R is generally utilized as a part of political science, measurements, econometrics, actuarial sciences, humanism, fund, and so on.
5. R is accessible for all real working frameworks (Windows, Mac OS, GNU-Linux).

6. R is item situated. For all intents and purposes anything (e.g., complex information structures) can be put away as an R object.
7. R is a framework dialect.
8. R language structure is considerably more deliberate than Stata or SAS punctuation.
9. R can be introduced on your USB stick

III. Experimental Setup & Approach

For the implementation of Eclat & Apriori algorithms we've used the R Environment with various user defined and system defined libraries.

1. Setting up R-Environment:

We need to install required packages in R environment: **User Libraries:**

- arules
- arulesViz

System Libraries:

- matrix
- grid

2. Finding Association Rules:

1. We'll use Apriori() and Eclat() functions from arules package to mine the association rules from the datasets, While setting the Parameter specification and algorithmic control.

2. We'll inspect the rule and select from them as per the required parameter specifications we want.

3. Data Sets:

We've used four datasets consisting of 5000, 10000, 20000 and 40000 records respectively and have performed Apriori and Eclat algorithms with no parameter and parameterized mode. To analyze the performance of these algorithms based of CPU utilization or the time consumed by CPU to execute these algorithms with different datasets and different modes.

IV. Results and Analysis

1. Experiment-1 (Simple Algorithm):

Time taken by CPU to execute the algorithms in Seconds. Algorithms performed with no parameters.

Steps to be repeat for each dataset:

A. Load Data:

```
results<-
read.csv("C:/Users/Tanu/Desktop/R/DataSet40000.csv")
```

B. Applying the Apriori algorithm:

```
System.time(rules<- apriori (results))
```

C. Applying the Eclat algorithm:

```
System.time(rules<- eclat (results))
```

Data Sets (Records)	Apriori	Eclat
5,000	0.11	0.06
10,000	0.13	0.12
20,000	0.25	0.24
40,000	0.56	0.47

Table 1: Execution time of Apriori and Eclat algorithms for different Datasets

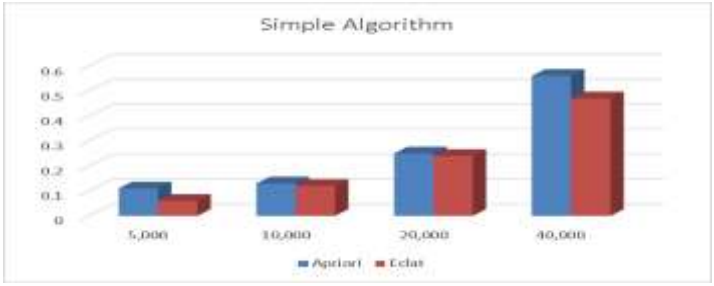

Figure 2: Comparison of Execution time of Apriori and Eclat algorithms

2. Experiment-2 (Algorithms used with Parameters):

Time taken by CPU to execute the algorithms with two parameters in Seconds. Algorithms performed with two parameters minlen and support.

Steps to be repeat for each dataset:

A. Load Data:

```
results<read.csv("C:/Users/Tanu/Desktop/R/DataSet40000
.csv")
```

B. Applying the Apriori algorithm:

```
system.time(rules <- apriori(results, parameter =
list(minlen=2, supp=0.05)))
```

C. Applying the Eclat algorithm:

```
system.time (rules <- eclat(results, parameter =
list(minlen=2, supp=0.05)))
```

Data Sets (Records)	Apriori	Eclat
5,000	0.07	0.05
10,000	0.13	0.11
20,000	0.25	0.23
40,000	0.51	0.49

Table 2: Execution time of Apriori and Eclat algorithms for different Datasets

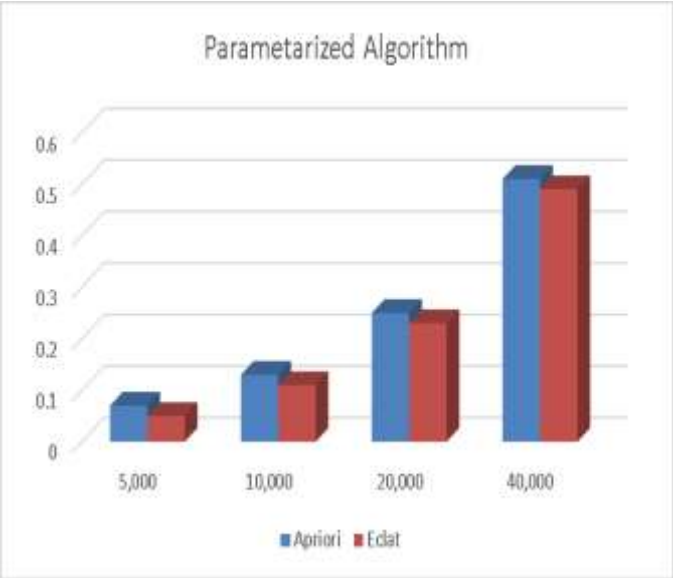

Figure 3: Comparison of Execution time of Apriori and Eclat algorithms

V. Conclusion

In this paper, we have done the quantitative analysis of association rules mining algorithms and discusses some problems of generating frequent item sets from the algorithm. This paper will adopt efficient sequential pattern mining techniques using the Apriori and Eclat algorithm for the filtered data set. Both the algorithms helps to find out the navigation behaviour of the user based on the previous visits and also shows the comparison of the two techniques adopted for predicting user access behaviour using R-programming.

The Performance of Apriori algorithm improves as the number of records in datasets increases but in all the scenarios performance for Eclat algorithm is better for association rules mining in terms of execution time.

Discovering the frequent item sets and association rules from the two algorithms, our experiments shows that Eclat algorithm serves better for the large datasets as well as for the comparatively small datasets despite Apriori algorithm as it generates less tables and therefore less time it takes to perform the analysis.

References

[1] O. R. Zaiane, J. Han, and H. Zhu. Mining Recurrent Items in Multimedia with Progressive Resolution
Refinement. ICDE'00, 461-470, San Diego, CA, Feb. 2000

[2] Rahul Mishra et. al. "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data." (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4), 2012, Pp. 4662 – 4665.

[3] Sachin Sharma, Vidushi Singhal and Seema Sharma, "A SYSTEMATIC APPROACH AND ALGORITHM FOR FREQUENT DATA ITEMSETS", Journal of Global research in computer science, Volume 3, No. 11, November 2012.

[4] Christian Borgelt,"Efficient Implementations of Apriori and Eclat", Department of Knowledge Processing and Language Engineering School of Computer Science, Otto-von-Guericke-University of Magdeburg Universitatsplatz 2, 39106 Magdeburg, Germany

[5] Sathish Kumar et al. "Efficient Tree Based Distributed Data Mining Algorithms for mining Frequent Patterns" International Journal of Computer Applications (0975 – 8887) Volume 10– No.1, November 2010.

[6]The R Project for Statistical Computing, http://www.r-project.org/

[7] Haoyuan Li,Yi Wang,Dong Zhang, Ming Zhang,Edward Chang 2008."Pfp: parallel fp-growth for query recommendation Proceedings of the 2008 ACM conference on Recommender systems Pp. 107-114.

[9] Fox, John and Andersen, Robert (January 2005). "Using the R Statistical Computing Environment to Teach Social Statistics Courses", Department of Sociology, McMaster University. Retrieved 2006-08-03.,http://en.wikipedia.org/wiki/R_%28programming_language%29