International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 124 - 126

Frequent Item Set Mining In Data Mining: A survey

Ishita Rana
Department of Computer Science& Engineering
Parul Institute of Engineering & Technology , Limda
Gujarat,India
ishita.rana92@gmail.com

Amit Rathod
Department of Information Technology
Parul Institute of Engineering & Technology, Limda
Gujarat
rathod.amit.h@gmail.com

***Abstract*—** Data mining is process of extracting useful information from different perspectives. Frequent Item set mining is widely used in financial, retail and telecommunication industry. The major concern of these industries is faster processing of a very large amount of data. Frequent item sets are those items which are frequently occurred. So we can use different types of algorithms for this purpose. Frequent Itemset mining can be performed Apriori, FP-tree, Eclat, and RARM algorithms. For the work in this paper, we have analyzed widely used algorithms for finding frequent patterns with the purpose of discovering how these algorithms can be used to obtain frequent patterns over large transactional databases. This has been presented in the form of a comparative study of the following algorithms: Apriori, Frequent Pattern (FP) Growth, Rapid Association Rule Mining (RARM) and ECLAT algorithm frequent pattern mining algorithms. This study also focuses on each of the algorithm's advantages, disadvantages and limitations for finding patterns among large item sets in database systems.

Keywords*-Association rule, Data Mining, Frequent item*

*****

I. INTRODUCTION

Data mining is a Process of analyzing data from different perspectives and summarizing it into useful information. Data mining has evolved in to an important area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real world databases. There are many technique of data mining such as Classification, Clustering, Association rule mining, Regression etc.

Data Mining Task [13]

Data Mining is the semi-automatic discovery of patterns, associations, changes, anomalies and statistically significant structures and events in data. Data Mining as a term used for the specific set of tasks or activities as follow:

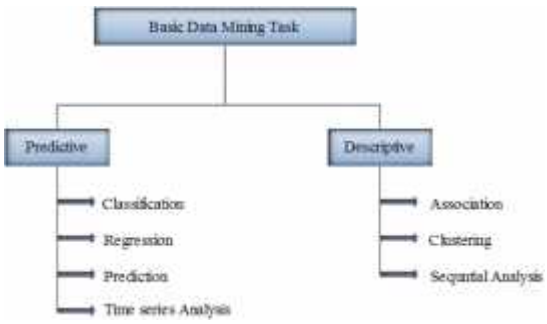

Figure 1: Data Mining Task [13]

Association Rule Mining

In Data mining, the definition of association rule mining finds interesting association or correlation relationships among a large set of data items. Association rule mining finding frequent pattern, correlations among the items or object in transactional database, or relational database. Association rule can create analyzing data for frequent pattern using the criteria Support & Confidence to identify the relationship. Support is indicating of how frequently the item appears in the database. Confidence indicates the number of time has been found. There are many algorithms used in association rule mining.

The main goal of association rule mining is:
Frequent item set generation
Rule generation (Find large item set)

Frequent Itemset Mining
Frequent pattern mining has been an important subject matter in data mining from many years. A remarkable progress in this field has been made and lots of efficient algorithms have been designed to search frequent patterns in a transactional database. Frequent pattern mining can be used in a variety of real world applications. It can be used in super markets for selling, product placement on shelves, for promotion rules and in text searching.

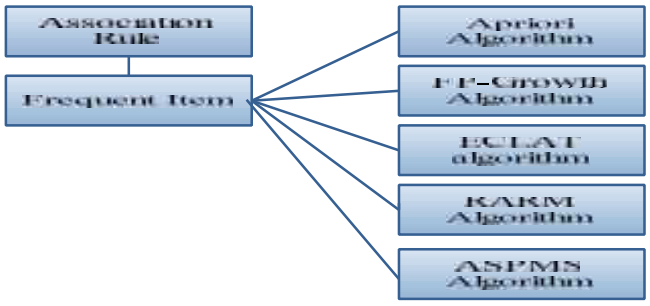

Figure 2: various types of Algorithm used in association rule [1]

II. REALATED WORK

Since its introduction by Agrawal et al [1], it has received a great deal of attention and various efficient and sophisticated algorithms have been proposed to do frequent itemset mining. Among the best-known algorithms are Apriori, Eclat and FP-Growth.

The Apriori algorithm [2] uses a breadth-first search and the downward closure property, in which any superset of an infrequent itemset is infrequent, to prune the search tree. Apriori usually adopts a horizontal layout to represent the transaction database and the frequency of an itemset is computed by counting its occurrence in each transaction.

FP-Growth [3] employs a divide-and-conquer strategy and a FP-tree data structure to achieve a condensed representation of the transaction database. It is currently one of the fastest algorithms for frequent pattern mining.

Eclat[4] takes a depth-first search and adopts a vertical layout to represent databases, in which each item is represented by a set of transaction IDs (called a tidset) whose transactions contain the item. Tidset of an itemset is generated by intersecting tidsets of its items. Because of the depth-first search, it is difficult to utilize the downward closure property like in Apriori. However, using tidsets has an advantage that there is no need for counting support, the support of an itemset is the size of the tidset representing it. The main operation of Eclat is intersecting tidsets, thus the size of tidsets is one of main factors affecting the running time and memory usage of Eclat. The bigger tidsets are, the more time and memory are needed.

Zaki and Gouda[5] proposed a new vertical data representation, called Diffset, and introduced dEclat, an Eclat-based algorithm using diffset. Instead of using tidsets, they use the difference of tidsets (called diffsets). Using diffsets has reduced drastically the set size representing item sets and thus operations on sets are much faster. Eclat had been shown to achieve significant improvements in performance as well as memory usage over Eclat, especially on dense databases[5]. However, when the dataset is sparse, diffset loses its advantage over tidset. Therefore, Zaki and Gouda suggested using tidset format at the start for sparse databases and then switching to different set format later when a switching condition is met.

TABLE 1. Analysis of various Algorithm used in Frequent item

	Apriori	**RARM**	**ECLAT**	**FP-Growth**	**ASPMS**
Techniques	Breadth first search and Apriori property(for pruning)	Depth first search	Depth first Search & intersection of t-id	Divide and Conquer	BSM(Branch Sort Method) using merge method
Database Scan	Database is scanned for each time a candidate item set is generated	Database is scanned few times to construct.	Database is scanned few times	Database id scanned two times only	Database is scanned only one time
Drawback	-requires large memory space. -Too many candidate item set.	-Difficult to use in interactive system mining	It requires the virtual memory to perform the transaction.	FP-Tree is expensive to build consumes more memory.	----
Advantages	-Easy to implement. -Use large item set property.	-No candidate generation.	-No need to scan database each time	-Database is scanned only two times.	-It requires less memory. -Highly suitable for interactive mining
Data format	Horizontal	Horizontal	Vertical	Horizontal	Horizontal
Storage structure	Array	Tree	Array	Tree(FP Tree)	Tree(Asp tree)
Time	More execution time	Less execution time as compared to Apriori & FP Growth	Execution time is less than Apriori algorithm	Less time as compared to Apriori algorithm	Less execution time as compared to FP growth algorithm

III. CONCLUSION

We have analyzed the comparative study of various frequent patterns mining algorithms. A comparison framework has developed to allow the flexible comparison of Apriori, Eclat and FP-growth algorithms. Using this framework this paper presented the comparative performance study of these algorithms such as, Apriori, Eclat and FP-growth. This study also focuses on each of the algorithm's advantages, disadvantages and limitations for finding patterns among large item sets in database systems.

REFERENCES

[1] Shamila Nasreen, Muhammad Awais Azamb, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar "Frequent Pattern mining algorithm finding associated frequent patterns for Data Streams: A Survey" 2014, Science Direct

[2] Xiaofeng Zheng [a], Shu Wang [a*] "Study on the Method of Road Transport Management Information Data mining Based on Pruning Eclat Algorithm and Map Reduce "2014, Science Direct

[3] Zhigang Zhang, Genlin Ji[*], Mengmeng Tang "MREclat: an Algorithm for Parallel Mining Frequent Item sets" 2013, IEEE

[4] Marghny H. Mohamed • Mohammed M. Darwieesh "Efficient mining frequent item sets algorithm" 2013, Springer

[5] Dr. S.Vijayarani, Ms. P. Sathya "An Efficient Algorithm for Mining Frequent Item Sets in Data Streams" 2013, International Journal of Innovative Research in Computer and Communication Engineering

[6] Kan Jin "A new Algorithm for Discovering Association Rules" 2010, IEEE

[7] Mingjun Song, and Sanguthevar Rajasekaran "A Transaction Mapping Algorithm for Frequent Item Sets Mining" Member, IEEE

[8] R. Agrawal, T. Imielinski, and A.N. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD International Conference on Management of Data, Washington, 1993.

[9] R. Agrawal, and R. Srikant,"Fast algorithms for mining association rules," in 20th International Conference on Very Large Data Bases, Washington, 1994.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD International Conference on Management of Data, Texas, 2000.

[11] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in Third International Conference on Knowledge Discovery and Data Mining, 1997.

[12] Avani M. Sakhapara, Bharathi H. N. "Comparative Study of Apriori Algorithms for Parallel Mining of Frequent Itemsets," International Journal of Computer Applications, 2014.

[13] Tipawan Silwattananusarn1 and Assoc.Prof. Dr. KulthidaTuamsuk2 "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.

[14] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.