Comparative Study of Revised FP – Growth, Weighted Apriori and Fuzzy Apriori Algorithm

A.Arthi Priadharsni [1], Dr. E. Ramaraj[2]

Research Shcoloar [1], Professor [2]

Department of Computer Science and Engineering,

Alagappa university, Karaikudi

Abstract -**Association Rule Mining is considered as one of the crucial step in finding the frequent Itemsets, for the purpose of extracting association rules from high voluminous relational databases. Many algorithms were developed to find the frequently occurred Itemsets. The association rules were considered as better, since they are useful at the level of decision making. . The main benefit of the Apriori- algorithm is that it doesn't need to generate conditional patterns Iteratively and adds the pruning step to eliminate the irrelevant data. FP-tree is highly a compact representation of all relevant frequency information in the data set. This paper presents a summarization and a comparative study of the available algorithms namely the Weighted Apriori, Revised FP-growth (Frequent Pattern), Frequent pattern growth and the fuzzy Apriori algorithm with their variations in mining the association rules to get frequent Itemsets.**

***Keywords*- Association Rule Discovery, Frequent Pattern Mining, FP-Growth algorithm, Weighted Apriori, Revised Fp-Growth.**

I. INTRODUCTION

Data Mining is the process of extracting potentially useful, but unknown information from high voluminous datasets. Mining of data includes formative process such as the data preprocessing (i.e) the removal of noise and irrelevant data, removal of duplicate and redundant data etc. Then the data is converted to the required format it should be able to access. This transformation is done with the help of various methods and tools. The Data Mining step includes the processing of gaining exclusively unknown information, which is valid from the given relational database is to be extracted.

There exists much number of methods and techniques for data mining such as the classification, clustering, feature selection, Association between the data etc., The Association Rule discovery and generation is most important part of every data mining task. The Association represents the type of relationship between the antecedent and the consequent. Data mining, one of the steps in process of knowledge discovery, "consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns over the data". Data mining is typically a bottom-up knowledge engineering strategy ,Knowledge discovery involves the additional steps of target data set selection, data pre processing, and data reduction (reducing the number of variables), which occur prior to data mining. Association rules are used to identify relationships among a set of items in databases. These relationships are not based on inherent properties of the data themselves, but rather based on co-occurrence of the data items [1] .

Two important parameters were considered in the association rule mining, support and confidence. An association rule is of a relation A=>B, where support is the percent of transaction that contains A and B, whereas confidence is the percent of transaction contains both A and B values. Associate rules are called strong if they satisfy both the minimum support threshold and the minimum confidence threshold.

Apriori is the Association rule mining algorithm that is most widely considered. This is also undergoes several process and many new algorithm based on Apriori such as Improved Apriori, Custom built Apriori, Weighted Apriori Etc where in existence. The Apriori property shows that every subsets of the frequent itemset must also be a frequent item set.

Next comes the FP – Growth (Frequent Pattern) growth algorithm, which produces considerably more number of frequent Itemsets, require iterative scanning of the database. At the first scanning step, 1- itemset is generated, followed by removal of infrequent itemset at the second scan. The revised

FP- growth algorithm includes a minor change that the use of entropy heuristic and new batch search mechanism is introduced.

The revised FP-growth algorithms is that the revised one uses the entropy heuristic during the generation of FP-tree and replaces the one-by-one scanning search mechanism with the batch search mechanism[7].

The rest of the paper is presented as follows. Section 2 consists of various related works carried out in the same area. Section 3 comprises of the description of Weighted Apriori and Revised Fp-Growth, FP- Growth and Fuzzy Apriori algorithms. Section 4 contains the comparative analysis results obtained. Finally, the conclusion is drawn at section5.

II. Related Work

In the work, extends by Ankit.R et.al., the process of Apriori and its function is discussed. The work, shows that apriori is the influential algorithm to find out the frequent item sets from high volume of data. Once the frequent item sets are generated then, straightforwardly the association rules are incorporated to show its support and confident between one another. All frequent item sets are found out at first step. The frequent item set is the item set that is included in at least minimum support transactions.

Lei Wang et. Al [7] extends their discussion on the mining of association rule based on weighted Apriori and the Revised FP- growth algorithm. The shows the new weighted based fp- tree generation on the huge volume dataset and generate rules.

In [8], data mining techniques have been used for text analysis by extracting co occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed.

Jia wei Han et al. [6]in their proposed work, introduces a novel frequent pattern tree structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth [4].

Rakesh Agrawal and Ramakrishan Srikant[2] consider the problem of discovering association rules between items in a large database of sales transactions. Ning Zhong et.al,[8] proposed a new method for effective pattern discovery to text mining. They presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

R.Sridevi and E. Ramaraj,[10] in their proposed work paper introduces a method to handle the categorical attributes and the numerical attributes in an efficient way. The conversion of data into quantitative method is taken place. Using the binary patterns frequent patterns are identified using FP growth. The conversion reveals all the frequent patterns from the database.

M.Suman et.al [9] extends their views on A Frequent Pattern Mining Algorithm Based on Fp-Tree Structure and Apriori Algorithm, in which on Apriori algorithm and the FP-tree structure is presented to mine frequent patterns. The advantage of the Apriori-Growth algorithm is that it doesn't need to generate conditional pattern bases and sub- conditional pattern tree recursively.

III. Weighted Apriori , FP- Growth , Fuzzy Apriori and Revised FP – Growth Algorithm

The following section briefly explains the basic algorithms the Apriori, weighted Apriori, FP-Growth, the Revised Frequent Pattern Growth algorithm and fuzzy Apriori with their working process and the description.

3.1 FP- Growth Algorithm

Han, Pei et al [5] proposed a data structure called FPtree (frequent pattern tree). FP-tree is a highly compact representation of all relevant frequent information in the given data set . Every path of FP -tree represents a frequent item set and the nodes in the path are stored in decreasing order of the frequency of the corresponding items.

The main advantage of the fp- growth algorithm is that overlapping Itemsets share same path in the node creation. Hence, the compression is done easier while using the algorithm for the dataset transaction. It scans the data set twice and candidate

itemset is no longer needed. An FP-tree has a header table. The nodes in the header table link to the same nodes in its FP-tree. Single items and their counts are stored in the header table by decreasing order of their counts.

3.2 Weighted Apriori Algorithm

While using the Apriori algorithm each transaction in the database is considered as a record and each dataset record is the item. The Apriori is the classic algorithm that is most widely used for the discovery of association rules. The association rules mining are divided into two steps, first to find all the frequently occurred sets in the database, by using the frequent itemset generate strong association rules that satisfy the minimum support and confidence value.

The Apriori is based on the record sets that are frequently occurred in the dataset. Apriori algorithm is actually a layer-by-layer iterative searching algorithm, where k-itemset is used to explore the (k+ 1)-itemset [11]. It first scans the database to find number of occurrences of each item on the given database. The Itemsets which satisfy the minimum support and the confidence value is set as the frequent 1 – itemset.

Now, the frequent 1 –Itemsets were joined together to form the k – itemset. The combination of the generated candidate set, the pruning step is taken place, which will remove the irrelevant items, that does not satisfy the minimum support and the confidence value. The weighted Apriori as such follows the classic rules, except provide the weights to each data item.

The weights were assigned in such a way that, after the candidate item sets generation, the separation of the attribute taken place. Each itemset is provided with weights, that is minimum, average and maximum value for the itemset for easier division of database. Based on the weights threshold the dataset item is pruned, and their association rules were generated. Based on the generated association rules, frequently occurred Itemsets were easily mined.

3.3 Revised FP – Growth Algorithm

The revised frequent pattern growth algorithm overcomes the drawbacks in the existing Apriori algorithm. For the Fp –growth algorithm, the fp – tree with rules are generated, based on that the frequent item sets were mined.

In the revised FP –Growth algorithm, uses the entropy heuristic during the generation of FP-tree and replaces the one-by-one scanning search mechanism with the batch search mechanism.[Xing].

The procedure for the revised FP –Growth algorithm is as follows: The transaction database is set with minimum support and confidence threshold. Then, the frequent Itemsets were scanned and sorted in their descending order with minimum weights. Now, the Fp- tree list is constructed.

It is then followed by recursion. If the tree is constructed as the list, each list is constructed with at least two nodes, in which each data item sets is compared with entropy value and the search mechanism is applied hierarchical order. The optimization of the frequently occurred pattern is found from the revised frequent pattern growth algorithm, since it holds only the Itemsets that are higher in order, with entropy heuristic.

3.4 Fuzzy Apriori Algorithm

Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains. A mining approach with the Apriori algorithm associated with the weighting of dataset is considered as fuzzy Apriori Algorithm. Basically, it consists of two divisions, the input data file and the fuzzy logic values for each dataset.

There are two forms of input data: (i) binary valued data for Weighted ARM only and (ii) fuzzy valued data for Fuzzy ARM. The first part consists of binary valued (0, 1) based dataset. In fuzzy method, Weightings are loaded in a separate text file. The assumption is that the weighting for each attribute is constant. Weightings are expressed in the form of real numbers between 1.0 and 0.0. The

format is one weighting per line. Thus there should be as many lines in the weighting file as attributes.

There are three different mechanisms for calculating support according to the nature of the algorithm chosen namely, Weighted Support, Fuzzy Support and Weighted Fuzzy Support. In the third division, for single items sets the support is the sum of the product calculation for each weighting/fuzzy member ship pair (w*f). For 2-itemsets and larger the support is the sum of the products of all the weightings and fuzzy membership calculations.

IV. COMPARATIVE ANALYSIS OF THE ALGORITHMS

This section consists of the comparative analysis of the provided algorithms based on their time limit and the total number of association rules generated for each of the algorithm. The dataset accounted here is the student database with 7 attributes and large number of records. The following table shows the results obtained from the 4 algorithms.

The table 1 consists of the time variation taken place with the change in the number of records. The fuzzy Apriori shows better results when number of records becomes huge. The time value is calculated in milliseconds (ms). The table 2 shows the total number of association rule mining being evolved from the algorithms.

TABLE 1. COMPARATIVE ANALYSIS OF ALGORITHMS BASED ON TIME LIMIT WITH RECORD VARIATION

Number of Records	Revised FP – Growth (ms)	Weighted Apriori (ms)	Fuzzy Apriori (ms)
500	7.0	5.0	4.0
1000	7.0	6.0	7.0
2000	11.0	10.0	9.0
3000	27.0	24.0	14.0

TABLE 2. COMPARATIVE ANALYSIS OF ALGORITHMS FOR NUMBER OF RULES GENERATED

Association Rule Mining	Revised FP –Growth	Weighted Apriori	Fuzzy Apriori
Number of Rules Generated	148	72	72

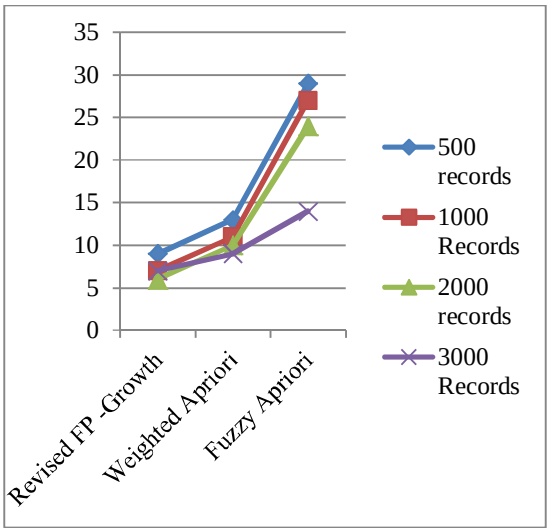

Fig 4.1 Graph shows the time variation for algorithms

V. CONCLUSION

The proposed comparative study, shows with the given student database, after the preprocessing the dataset, transactions were being implemented with both the algorithms stated above. Based on the results obtained, for the weighted Apriori algorithm the association rules found was 72 and for the revised FP- Growth it seems to be 148 rules. Hence, the weighted Apriori yields better results, compared to other. The weights indicated by the algorithm shows the complete variation between the rules. For the time limit analysis apart from the other algorithms, the fuzzy Apriori is more efficient when the number of records becomes high. Within limited time period it yields better results

REFERENCES

[1] *Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, ISBN 1-55860-153-8.*

[2] *R. Agrawal, Tomasz Imielinski, Arun Swami, "Mining association rules between sets of items in lager databases", in Proceeding of ACM SIGMOD international conference of management of data, Washington, DC, May 1993, 207-216.*

[3] *H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo,"Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries(ADL '98), pp. 2-11, 1998.*

[4] *Ankit R Kharwar,Viral Kapadia,Nilesh Prajapati,Premal Patel,"Implementing APRIORI Algorithm on Web serve log",National Conference on Recent Trends in Engineering & Technology,B.V.M. Engineering College, V.V.Nagar,Gujarat,India,13-14 May 2011.*

[5] *Han J., Pei J., Yin Y. and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach" Data Mining and Knowledge Discovery, 2004.*

[6] Jiawei H, Micheline K, "Data Mining:Concepts and Techniques" Morgan Kaufinann, 1 st edition 2000.

[7] Lei Wang, Xing-Juan Fan, Xing-Long Lw, Huan Zha0, "Mining Data Association Based On A Revised Fp-Growth Algorithm", in Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012,pp:91-95.

[8] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Efftective Pattern Discovery for Text Mining", in IEEE transaction on Knowledge and Data Engineering, Vol 24, No 1, Jan 2012, pp: 30 -43.

[9] M Suman ,T Anuradha ,K Gowtham, A Ramakrishna, "A Frequent Pattern Mining Algorithm Based On Fp-Tree Structure Andapriori Algorithm", in International Journal of Engineering Research and Applications, Vol. 2, Issue 1, Jan-Feb 2012, pp.114-116.

[10] R.Sridevi and Dr.E.Ramaraj, "Finding Frequent Patterns Based On Quantitative Binary Attributes Using FP-Growth Algorithm", in Int. Journal of Engineering Research and Applications, Vol. 3, Issue 6, Nov-Dec 2013, pp.829-834.

[11] Xing-Juan Fan, Xing-Long Lw, Huan Zha, "Mining Data Association Based On A Revised Fp-Growth Algorithm", in Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012 , pp:91- 95.