Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 9-12

An Enhanced Approach to Mine Maximal Frequent Itemset using Maximal Frequent Itemset Prima Algorithm (MFIPA)

R. Smeeta Mary[1] and K. Perumal[2]
[1]Assistant Professor, Department of Computer Applications, Fatima College, Madurai, Tamil Nadu, India
[2]Professor, Department of Computer Applications, School of Information Technology,
Madurai Kamaraj University, Madurai, Tamil Nadu, India
E-Mail: smeetamaryr@gmail.com, perumalmala@gmail.com

***Abstract* - In data mining finding out the frequent itemsets is one of the very essential topics. Data mining helps in identifying the best knowledge for different decision makers. Frequent itemset generation is the precondition and most time-consuming method for association rule mining. In this paper we suggest a new algorithm for frequent itemset detection that works with datasets in distributed manner. The proposed algorithm brings in a new method to find frequent itemset not including the necessitate to create candidate itemsets. The proposed approach could be implemented using horizontal representation for transaction datasets and allocating prime value. It explores all the frequent itemset that is present in the input and according to the support the maximum frequent itemset is identified. It was applied on different transactions database and compared with well-known algorithms: FP-Growth and Parallel Apriori with different support levels. The try out showed that the proposed algorithm attain major time improvement over both algorithms.**
Keywords: **Data Mining, Itemset, Prima Algorithm**

I. INTRODUCTION

Data mining is a technique that takes input as data and knowledge as outputs. One of the main and most important definitions of the data mining, which highlights the distinctive characteristics, is provided by Fayyad, Piatetsky-Shapiro and Smyth (1996), who define it as "the nontrivial process of identifying potentially useful valid, novel and eventually understandable patterns in data."

Frequent itemsets discovery's that it is one of the most important techniques in data mining. It is used to find the association relationships among data objects or events that are concealed in the data, although the associated objects or events are not related. By understanding the relationships which helps to analyze events and may disclose useful patterns for marketing policies ,decision support, even medical diagnosis and many other applications[2]. There are various frequent itemset mining techniques for finding frequent pattern in large datasets. Literature includes many approaches that solve the FIM problem like Apriori, FP-Growth and Pobpa.

Frequent itemsets plays an essential role in many data mining tasks which finds the interesting patterns from databases, such as association rules, correlations, episodes, classifiers, sequences,clusters and many more of this mining of association rules is the trendiest problems. Mining frequent itemset is considered as the main problem in association mining [3].

Frequent Item Mining explains how the items are obtained together. For example, a supermarket contains 200,000 customer transactions. The purchases of diapers were 4,000 transactions and purchases of beer were 5,500 transactions. By including both the purchase of diapers and beer was 1.75%. The diaper purchases 87.5% includes beer also which indicates an association between diapers and beer.

This paper is structured as follows. Section 2 defines the problem statement. Section 3 briefly explains a literature overview. Section 4 our proposed approach. Section 5 describes our experimental results. Finally, section 6 summarizes the conclusions and future work.

II. PROBLEM STATEMENT

Data mining is a process which is used to extract data from a larger set of any raw data. Using one or more softwares it analyses data patterns in large volume of data. Data mining has many applications in science and research. As an example of data mining, business learn more about their customers and develop effective strategies which are related to business functions. Frequent Itemset Mining is an important subject matter in data mining from many years. An outstanding growth in its field was made and lots of algorithms were designed to search maximal frequent items in a database. In real world applications various frequent pattern mining can be used. It can be used in product placement on shelves, super markets for selling and for promotion rules and in text searching. As a result of this the need for new frequent itemset mining algorithms can be designed to tackle the new trends that are emerged. In this paper we propose a new algorithm for FIM that could deal with data sets exploiting the multicore features.

III. LITERATURE REVIEW

In this section an overview of well-known FIM algorithms is presented. It includes: Apriori, the AIS algorithm and FP-Growth that depends on representation of prime numbers.

"The AIS algorithm" was the first algorithm projected by Agrawal, Imielinski, and Swami for mining association rule"[5]. AIS algorithm scans the databases many times to get the frequent itemsets. "The support count of each individual item was gathered during the first pass over the database. Depending on the threshold of support count the items which have the count less than its minimum value are removed from the list of items. Next is the Candidate 2-itemsets which are generated by extending frequent 1-itemsets with others transactions. In the second pass, the support count of those candidate 2-itemsets are obtained and checked against the support threshold. Similarly this process is repeated for the entire item in the same transaction. Until any one of them becomes empty the generation of candidate itemsets and frequent itemsets are generated. Since all the candidate itemsets and frequent itemsets are assumed to be stored in the main memory, memory management is not enough.

Apriori Algorithm is one of the famous algorithms for finding out the mining frequent itemsets. The algorithm uses the prior knowledge of frequent itemsets. Apriori scans the transaction data base to get the support of S then compare S with min_sup, and get a support of 1-itemsets, L1.The algorithm then uses Lk−1 join Lk−1 to generate a set of candidate k-itemsets and unfrequented itemsets are pruned from the set. Scan the database to get the support S and compare it with the minimum support to get the frequent item set. And set the value of candidate key as NULL. Generate the nonempty set and check the particular rule and find the confidence level.

"FP-Growth Algorithm is the next level of frequent itemset. The main of this algorithm was to remove the difficulties of the Apriori algorithm in generating and testing candidate sets and thus it got the name as frequent pattern tree or FP-tree. FP-Growth uses both the vertical and horizontal database to store the values of the database in main memory. Rather than storing ID for every transaction in the database, it actually stores the database in a tree structure and linked list connects all the items that have its own item[6].

IV. THE PROPOSED ALGORITHM (MFIPA)

In this paper, a new approach Maximal Frequent Itemset Prima algorithm is introduced and works more efficiently with the transactional database. The new algorithm, calledMFIPA for short "Maximal Frequent Itemset Prima algorithm". It depends on greatest common divisor calculation between different transactions in the transaction database. Generally speaking, MFIPA consists of two main steps:

1. Data Preparation

MFIPA explore the transaction dataset and finds the frequency of all the transactional items. Also it finds a list that how many times the items are repeated. The obtained result is arranged in descending order with the transaction ID and its count.

2. Frequent Itemsets Deduction (Greatest Common Divisor calculation (GCD)

According to the count we assign the prime value for the entire transactional ID. These values are substituted in the transaction items and the total is found out. Then by using repeated subtraction from the total we get the frequent itemsets in the transaction dataset.

A. Data Preparation

MFIPA uses horizontal representation for transaction datasets where all the transactions are represented as a row in the database which contains transaction identifier (TID) as shown in Table I.

TABLE I REPRESENTATION OF THE DATABASE

Tid	Titems
T1	A,C,T,W
T2	C,D,W
T3	A,C,T,W
T4	A,C,D,W
T5	A,C,D,T,W
T6	C,D,T

The first and foremost step in the data preparation phase is finding out the frequency of all individual items in the dataset. Once the frequencies of the items are determined then the algorithm will prepare to work Table II.

TABLE II FREQUENCY OF THE ITEMS

Titems	Count
A	4
C	6
D	4
T	4
W	5

According to the frequency the items are sorted in the descending order and the highest frequency is assigned the least prime value Table III.

TABLE III ASSIGNING PRIME VALUES

Titems	Prime values
C	2
W	3
A	5
D	7
T	11

The prime values that are assigned to the transaction items are substituted to the input and the total of the transaction of the items are being found out. As in Table III the transaction C has the highest frequency and the prime value is 2 starts subtracting the value from the total value. Check is there any exit criteria if not so take the next prime value and subtract it from the present total value.

B. Frequent Itemsets Deduction Using MFIPA

The proposed algorithm depends on calculating the total by assigning the prime values to the input items. The main advantage of this method is that repeated subtractions are performed until exit criteria value is reached. For example the frequency count of the transaction items C, W is 5 and 6. So the prime values 2 and 3 are assigned to it. If the prime count are substituted to the input then the count will be 21 and 12 are first two transaction. Now the transaction item C having the prime value 2 is subtracted from the total so the values are changed as 19 and 10. Since the exit criteria are not reached the same procedure is followed by taking next prime value and continuous subtraction until exit criteria is reached.But one important note is that if suppose the frequency values are same for three items means then only the previous total must be taken and subtracted from all three. If suppose the exit criteria is reached then the linked list is followed and the frequent item set that are present are listed. These frequent items are checked with the input to know how many times they are repeated. According to the support the frequent item is calculated and displayed Table IV.

TABLE IV FREQUENCY ITEM SET DEDUCTION USING MFIPA

Trans	Total	C	CW	CWA	CWAD	CWAT
T1	21	19	16	11	4	0
T2	12	10	7	-	-	-
T3	21	19	16	11	4	0
T4	17	15	12	7	0	-
T5	28	26	23	18	11	7
T6	20	18	15	10	3	-

From Table IV it is understood that CWAT and CWAD are frequent item set and it is checked with the input thus CWAD is repeated 2 times and CWAT is repeated 3 times. If suppose, the mini support is specified as 3 then CWAT is considered as the maximal frequent item.

V. EXPERIMENTAL RESULTS

By experimenting withvarious datasets and different applications the results are given. These datasets were acquired from the UCI repository of machine learning databases [7]. Our algorithm has theabilitytowork large number of database. We used Retail dataset in R programming language to show thepotential of the proposed algorithm MFIPA.

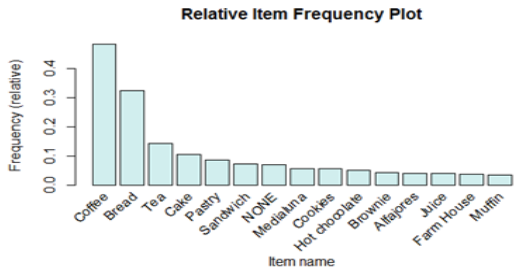

Fig. 1 Relative Item Frequency plot

The relative frequency of the items are displayed in Fig.1

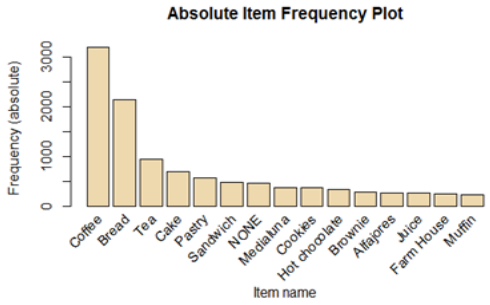

Fig. 2 Absolute Item Frequency plot

The absolute frequency of the items are displayed in Fig.2

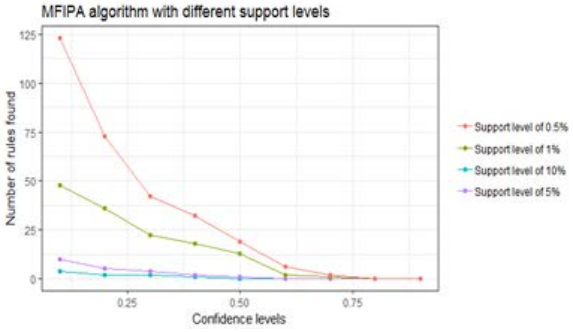

Fig. 3 MFIPA algorithm with different support levels

The different support levels of the items are displayed in Fig.3

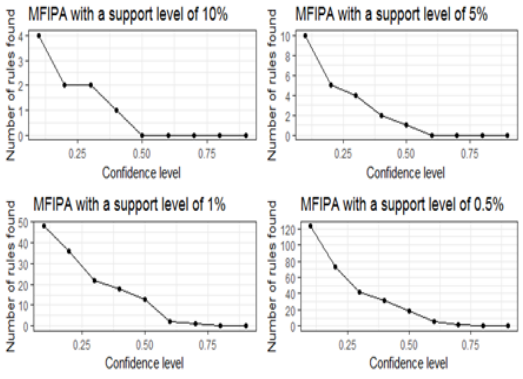

Fig. 4MFIPA with a support level

Compared the various support levels of the proposed algorithm fig. 4.Thus it was shown that MFIPA made a tremendous development especially with large volume of the datasets items. We applied our experiments on the datasets using 10% to 0.5% minimum supportthreshold.

VI. CONCLUSION

In this paper a new technique for Frequent Item Mining is introduced. Our approach, MFIPA, proved to do extremely well than other approaches in the literature. The algorithm can be applied to different areas. It aims to work effectively and optimize more work. It depends on repeated subtraction. It assign the prime number to the frequency item which guarantees less storage and computational. The processing speed is more than FP-Growth and Parallel Apriori.

REFERENCES

[1] U.Fayyad, G. Piatetsky-Shapiro and P. Smyth, "*From data mining to knowledge discovery in databases. AI Magazine*", Vol. 17, No. 3, pp. 37-54, 1999.

[2] S. Pramod, and O. P.Vyas, "Survey on Frequent Item set Mining Algorithms Survey on Frequent Item set Mining Algorithms Survey on Frequent Item set Mining Algorithms", *International Journal of Computer Applications*, 2015.

[3] JayantKayastha, and N. R. Wankhade "A Survey Paper on Frequent Itemset Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering Research*, Vol. 6, No. 2277 128X, 2016.

[4] RanaIshita, and Amitrathod, "Frequent Itemset Mining in Data Mining: A Survey", *International Journal of Computer Applications*, Vol. 139, No. 9, pp. 0975 – 8887, 2016.

[5] R. Agrawal, TImielinksi and A.Swami "Mining association rules between sets of items in large database", *The ACM SIGMOD Conference*, 1993.

[6] J. Han, HPei and YYin, "Mining Frequent Patterns without Candidate Generation", *Conf. on the Management of Data (SIGMOD'00, Dallas, TX)*, 2000.

[7] C. L. Blake, C. J.Merz, "UCI Repository of Machine Learning Databases", *In: CA, USA: Dept. of Information and Computer Science 1998.*