An Efficient Mining of Infrequent Weighted Itemset and Optimization using Transaction Mapping Technique

B. Indumathi and Dr.R. Jeevarathinam

Abstract--- Mining infrequent Itemset is fundamental method for mining association rules as well as for many other frequent Itemset mining tasks. In existing methods for Mining frequent and infrequent has been implemented using FP growth or Apriori algorithm. Also numerous experimental results have demonstrated that these techniques are scalable to the mining process. In this paper, a novel Transaction Mapping Technique for filtering the association rules for infrequent Itemset in the transaction dataset is proposed. This Proposed technique produces improved performance for sparse data items. Furthermore, an open and closed Itemset extraction of rules using optimization techniques is presented. The experimental results prove that this proposed technique is highly scalable and consume less memory compared to the state of art techniques.

Index Terms--- Association Rules, Frequent Itemset, Data Mining

I. INTRODUCTION

Mining Infrequent Item set is an exploratory data mining technique widely used for discovering valuable correlations among data which are weighted high in frequent and infrequent items in the transaction as infrequent Itemset utilized in low numbers. In order to identify infrequent Itemset in the different transaction Itemset, Itemset mining was focused on discovering frequent item sets with valuable Correlations using FP tree [1] i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold [2]. However, many traditional approaches ignore the influence/interest of each item/transaction within the analyzed data. To allow treating items/transactions differently based on their relevance in the frequent itemset mining process[3], the notion of weighted itemset has also been introduced with optimization criteria's[4][5]. A weight is associated with each data item and characterizes its local significance within each transaction through FP growth algorithm [6][10]. Reporting an infrequent Itemset which has an infrequent proper subset is redundant, since the former can be deduced from the latter. Hence, it is essential to report only the minimally infrequent itemsets. For instance, in [7], [8] algorithms for discovering minimal infrequent itemsets, i.e., infrequent itemsets that do not contain any infrequent subset, have been proposed. Infrequent itemset discovery is applicable to data coming from different real-life application contexts such as (i) statistical disclosure risk assessment from census data and (ii) fraud detection [9]. However, traditional infrequent item set mining algorithms still suffer from their inability to take local item interestingness into account during the mining phase. FP growth algorithm [10] is used to mine the frequent item . If the support threshold is too high, then less number of frequent item sets will be generated resulting in loss of valuable association rules. On the other hand, when the support threshold is too low, a large number of frequent item sets and consequently large number of association rules are generated, thereby making it difficult for the user to choose the important ones. As Part of this problem,

B. Indumathi, Research Scholar, Department of Computer Science, SNR Sons College.E-mail:b.indumathimca@gmail.com

Dr.R. Jeevarathinam, MCA., MPhil., PhD., Associate Professor, Department of Computer Science, SNR Sons College. E-mail:mani-jeeva-2000@yahoo.com

mining the items with different characteristic through weighted aggregation function is possible transaction mapping algorithm [11]. Transaction Mapping is algorithm proposed in this work to generate infrequent item sets irrespective of the length of the itemset. Weighted item support and confidence quality indexes are defined accordingly and used for driving the item set and rule mining phases as it is optimized aggregate function. This paper differs from the above-mentioned approaches because it focuses on mining infrequent itemsets from weighted data instead of frequent ones counting of itemsets is performed by intersecting these interval lists in a depth-first order along the lexicographic tree. When the compression coefficient becomes smaller than the average number of comparisons for intervals intersection at a certain level, the algorithm switches to transaction id intersection. Hence, different pruning techniques are exploited. A related research issue is probabilistic frequent itemset mining [12], [13]. It entails mining frequent itemsets from uncertain data, in which item occurrences in each transaction are uncertain. The objective of discovering item sets whose frequency of occurrence in the analyzed data is less than or equal to a maximum threshold is considered as frequent Itemset. The Infrequent Itemset-support measure is defined as a weighted frequency of occurrence of an Itemset in the analyzed data. Occurrence weights are derived from the weights associated with items in each transaction by applying a given cost function. The Infrequent Itemset -support- min measure, which relies on a minimum cost function, i.e., the occurrence of an Itemset in a given transaction is weighted by the weight of its least interesting item. The Infrequent Itemset measure, which relies on a maximum cost function, i.e., the occurrence of an itemset in a given transaction is weighted by the weight of the most interesting item. Note that, when dealing with optimization problems, minimum and maximum are the most commonly used cost functions. Hence, they are deemed suitable for driving the selection of a worthwhile subset of infrequent weighted data correlations. The remainder of the paper is structured as follows. In section 2, related work of the frequent item set algorithms and technique are discussed. Section 3 describes the proposed based on infrequent weighted item set mining technique .Experimental evaluation of the proposed system is presented in the Section 4. Finally, section 5 concludes the work and points out future research directions.

II. RELATED WORK

Frequent Item Set Mining

Frequent itemset mining is a widely used data mining technique for obtaining the association between items in the transaction dataset. Weighted association rules (WAR), which include weights denoting item significance. However, weights are introduced only during the rule generation step after performing the traditional frequent item set mining process to drive the association of items greater or lesser than threshold values .

Infrequent Item Set Mining

Weighted item support and confidence quality indexes are defined accordingly and used for driving the Itemset and rule mining phases. Probability of occurrence of an item within a transaction may be totally uncorrelated with its relative importance. A different item pruning strategy tailored to the traditional FP-tree structure to perform Infrequent Itemset mining efficiently. An attempt to exploit infrequent item sets in mining positive and negative association rules has been calculated through support count.

III. INFREQUENT ITEMSET MINING TECHNIQUE

Pre-processing of Transaction Dataset

A data set is a collection of transaction data records, usually presented in tabular form. Each column represents particular variables such as item name and a data set has several characteristics which define its structure and properties of the item. In these modules, it is essential to understand pre-process the data is represented in DB to extract the unique records from collection of different group of multidimensional data items to ease the discovering of the frequent Itemset in the infrequent list. Weighted of the associated data is calculated such as items of from thousands records of transaction for its significance While this is a a challenging task, to obtain the solution, dimensionality reduction offers a middle ground, which generally results in faster computational time, while yielding reasonable accuracy through the following Frequent pattern growth algorithm and transaction mapping algorithm for mapping the transaction records of the different tables .

Discovering Frequent Weighted Item Set Using FP Tree Algorithm

Frequent Weighted items are Mined based on the Threshold conditions and correlation conditions for mining the frequency of the items. The rule generation function based on Support- Maximum is used calculate the frequent items in the transaction table. It is designed to identify frequency of occurrence in the analyzed data is less than or equal to a maximum threshold. One of the pruning measures used in FP model is *interest* or influence measure to determine the similarity or Coorelation strength of the data. Correlation measure that is given as follows

The occurrence of itemset A is independent of the occurrence of itemset B if

$$P(AUB) = P(A)P(B)$$

Otherwise, itemset A and B are dependent and correlated as events.

The cluster of the frequent correlated data of the frequent and infrequent items of the single dataset is calculated

Estimating the Infrequent Data Correlations

Discovering item sets whose frequency of occurrence in the analyzed data correlated based on the different criteria as follows

- Data items is occurred in specified intervals
- Data items occurred less than or equal to a maximum threshold is correlated into Itemset clusters,
- Similar Transaction items is grouped based on the Transaction ids
- Sorting is carried on the each criteria to calculate the weight of the data items

Resultant cluster will yield as infrequent weighted items for single cluster

Infrequent Weighted Item Set Using Transaction Mapping

Transaction mapping driven by the maximum -support constraint, i.e., items belonging to different clusters based on different rules is grouped based on the aggregation function which yields the average resultant weight for items in the clusters.

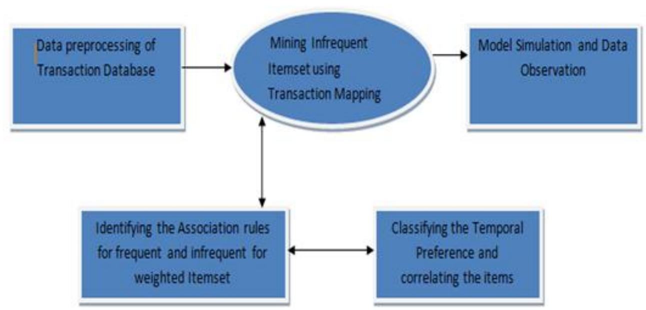

Figure 1: Architecture Diagram of the Infrequent Item Set Mining Using Transaction Mapping

Figure 1 explains the architecture of the infrequent item set mining process against the transaction dataset. It is an algorithm for infrequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

The transaction tree is similar to FP-tree but there is no header table or node link. The transaction tree has compact representation of all the transactions in the database. Each node in tree has an id corresponding to an item and a counter that keeps the number of transactions that contain this item in this path. Here transaction for each Itemset to continuous intervals by mapping transaction ids into a different space to a transaction tree is grouped sparsely.

Algorithm: Infrequent Weighted Itemset Mining Using Transaction Mapping

Input: -Database DB

Output: -all infrequent item sets

1. Scan the database and identify the infrequent item sets combinations using infrequent item mining.
2. Construct the transaction tree with the count for each node.
3. Construct the transaction interval lists for transactions, create an FP tree for each transaction.
3. Construct the lexicographic tree in a depth first order keeping only the minimum amount of information necessary to complete the search.
4. Iterative the transaction until all transaction are traced
5. Create Conditional pattern base calculate value
6. Obtain infrequent Itemset

IV. EXPERIMENTAL RESULTS

In this section, infrequent weighted Itemset mining and Transaction Mapping is evaluated for Mining the Coorelation of the different parameters of the items to yield the weight of aggregation function. The performance of the mining results is carried by means of the following metrics,

- Support threshold
- Confidence

Support Threshold

The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactions contain $X \cup Y$.

$$Sup = Pr(X \cup Y)$$

Confidence

The rule holds in *T* with confidence *conf* if *conf*% of transactions that contain *X* also contain *Y*.

$$Conf = \Pr(Y \mid X)$$

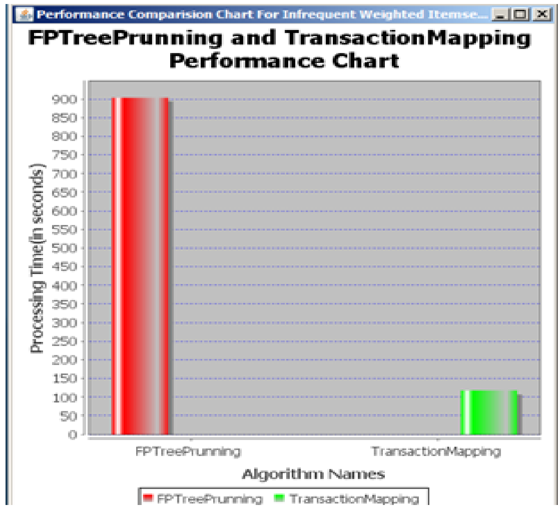

Figure 2: Explains the Performance of the System against the Processing TIME

Infrequent weighted mining using the FP tree algorithm and optimized technique using transaction mapping is explained in chart to evaluate the results. The optimized technique yields better results in terms of processing time as it is observed in the figure 2

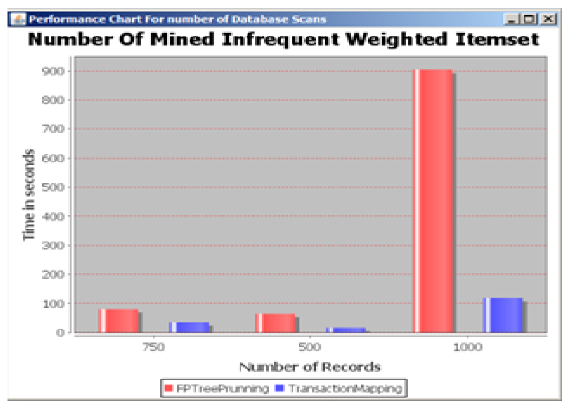

Figure 3: Evaluation of the Mining against Different No of Records

In figure 3. Infrequent weighted item set in transaction items with different record set using optimized technique is evaluated against the FP tree association rule when *X* occurs, *Y* occurs with certain probability.

Where support and Confidence of the optimized technique is given by

Support count: The support count of an Itemset *X*, denoted by *X.count*, in a data set T is the number of transactions in T that contain X. Assume T has n transactions.

$$confidence = \frac{(X \cup Y).count}{X.count}$$

$$support = \frac{(X \cup Y).count}{n}$$

As the support thresholds are reduced, the pruning condition becomes activated and leads to reduction in search space. Above the neutral point, the pruning condition is not effective.

Table 1: FP Tree Pruning

TransId	CPUUsageA	CPUUsageB	CPUUsageC	CPUUsageD	ThresHoldValue
309	26	26	31	73	25
633	26	36	45	45	25
248	26	41	41	44	25
753	26	51	82	82	25
202	26	56	42	88	25
807	26	57	57	30	25
295	27	27	27	27	25
751	27	27	30	37	25
884	27	27	59	34	25
593	27	33	33	36	25
36	27	35	39	39	25
822	28	26	59	59	25

Table 1 represents the FP tree pruning of the infrequent Itemset in the transaction database of the CPU utilizations.

FP-tree node pruning driven by the maximum IWI-support constraint, i.e., early discarding of part of the search space thanks to a novel item pruning strategy.

Table 2: Transaction Mapping Algorithm

NodeId	Count	Item	StartId	EndId
RA	10	CPUUsageA	1	10
RB	6	CPUUsageB	11	16
RC	12	CPUUsageC	2	4
RD	4	CPUUsageD	2	4
RAB	11	CPUUsageB	3	8
RAC	8	CPUUsageA	9	16
RAD	7	CPUUsageD	17	23
RBD	1	CPUUsageC	24	24
RCD	1	CPUUsageB	25	25
RABC	2	CPUUsageC	4	5
RABD	5	CPUUsageC	2	6
RACD	1	CPUUsageA	9	9
RBCD	2	CPUUsageB	5	6
RABCD	3	CPUUsageD	11	13
R	1	Root	1	1

Table 2 represents the Transaction Mapping of the infrequent Itemset in the transaction database of the CPU utilization

The algorithm is implemented by creating a tree structure where the infrequent item sets are mined using aggregation function and compared with the threshold value to yield the infrequent item records with different characteristics .

V. CONCLUSION & FUTURE WORK

Design and implementation infrequent mining method which has carried out using Infrequent weighted Itemset mining and transaction Mapping algorithm by discovering infrequent item sets in the different categories of the group is aggregated by using weights for differentiating between relevant items and not within each infrequent weighted Itemset transaction list. Existing algorithms evaluated on dense of the sparse dataset. The usefulness of the discovered patterns has been validated on data with multi level minimum support along the domain expert. As a future work, integrating the decision making algorithm for generation of Association rules to mine the infrequent weighted dataset that supports domain expert's targeted actions based on the characteristics of the discovered Infrequent Itemset set with Missing values. Furthermore, the different aggregation functions on the Itemset mining will be analysed and implemented based different dynamic strategies for high utility Itemset mining.

REFERENCE

[1] RLuca Cagliero and Paolo Garza "Infrequent Weighted Itemset Mining using Frequent Pattern Growth", IEEE Transactions on Knowledge and Data Engineering, Pp. 1- 14, 2013.

[2] M.L. Antonie, O.R. Zaiane, and A. Coman, "A pplication of Data Mining Techniques for Medical Image Classification," Proc. Second Intl. Workshop Multimedia Data Mining in Conjunction with seventh ACM SIGKDD (MDM/KDD '01), 2001.

[3] G. Cong, A.K.H. Tung, X. Xu, F. Pan, and J. Yang, "Farmer: Finding Interesting Rule Groups in Microarray Datasets," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '04), 2004.

[4] W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and data Mining (KDD '00), Pp. 270-274, 2000.

[5] F. Tao, F. Murtagh, and M. Farid, “Weighted Association Rule Mining Using Weighted Support and Significance Framework,” Proc. nineth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’03), Pp. 661-666, 2003.

[6] R.Jeevarathinam and Antony Selvadoss Thanamani, “An Implementation of FP-growth algorithm for Software Specification Mining”, CIIT Int’l Journal of Data mining and Knowledge Engineering, Vol-1, Pp.18-23, April 2009.

[7] A. Manning and D. Haglin, “A New Algorithm for Finding Minimal Sample Uniques for Use in Statistical Disclosure Assessment,” Proc. IEEE Fifth Int’l Conf. Data Mining (ICDM ’05), Pp. 290-297, 2005.

[8] Diti Gupta and Abhishek Singh Chauhan “Mining Association Rules from Infrequent Itemsets: A Survey”, International Journal of Innovative Research in Science, Engineering and Technology, ISSN: 2319-8753, Vol. 2, Issue 10, Pp. 5801 – 5808, 2013

[9] D.J. Haglin and A.M. Manning, “On Minimal Infrequent Itemset Mining,” Proc. Int’l Conf. Data Mining (DMIN ’07), Pp. 141-147, 2007.

[10] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data, Pp. 1-12, 2000.

[11] M.Song and S. Rajesekaran, “ A Transaction Mapping Algorithm for Frequent Item sets Mining”, IEEE Trans. On Knowledge and Data Engineering, Vol. 18, No. 4, pp. 472- 481. April 2006.

[12] C.-K. Chui, B. Kao, and E. Hung, “Mining Frequent Itemsets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD ’07), Pp. 47-58, 2007.

[13] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, “Probabilistic Frequent Itemset Mining in Uncertain Databases,” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’09), Pp. 119-128, 2009.

[14] C.K.-S. Leung, C.L. Carmichael, and B. Hao, “Efficient Mining of Frequent Patterns from Uncertain Data,” Proc. Seventh IEEE Int’l Conf. Data Mining Workshops (ICDMW ’07), pp. 489-494, 2007.

[15] Asha Rajkumar and G.Sophia Reena “Frequent Item set Mining Using Global Profit Weight Algorithm”, International Journal on Computer Science and Engineering, Vol. 02, No. 08, Pp. 2519-2525, 2010.