An Enhanced Method to Mine Rare Item Sets using Multiple Item Sets Support based on CP-Tree

Acharya Isha Umeshbhai
Parul Institute of Technology
P.O: Limda, Ta: Waghodia
Dist.: Vadodara

Ankur N. Shah
Parul Institute of Technology
P.O: Limda, Ta: Waghodia
Dist.: Vadodara

ABSTRACT
Rare Association rule is an association rule consisting of rare items.

Frequent Pattern (FP)-growth is an approach for utilizes the preceding knowledge providing by the user at the time of input and discovers frequent patterns with a two scan on the transactional dataset. We are presented a CP-tree (Compact-pattern tree), that capture database information with one scan (Insertion phase) and provided the same mining performance as the FP-growth method (Restructuring phase) by dynamic tree restructuring process. CP-tree can give functionalities for interactive and incremental mining with single database scan with our CP-tree outperforms in denominate of both execution time and memory requirements.

Hence, we are going to present a generated MIS-tree based on CP-tree.

General Terms
MCCFP Growth Algorithm and MIS CP-Tree Algorithm.

Keywords
Data mining, association rule mining, rare item sets, frequent pattern, MCCFP-growth, MIS CP-tree.

1. INTRODUCTION
Data Mining is the extraction of hidden, predictive information patterns from large databases [1].

Data Mining is especially useful present days when there is massive amount of data and identifying the useful section of it can be a tedious job in itself. Data mining allows us to be proactive about situations rather than backdated this means that we can now try and envision the future trends rather than identifying them after they have already taken place.

Rare Association Rule Mining (RARM) refers to an association rule mining forming between frequent rare items or among rare items and later are not found in algorithm as they suffer with "rare item problem".[7] That is, at high minimum support (min_sup) value, frequent pattern involving rare items could not be expected as rare items fail to satisfy to minimum support value. Therefore in order to search the rare items the minimum support value would be set to small which also results in producing many frequent items.

Applications of Rare association rule mining

- Gold Mining,
- Social networking services,
- Coal Mining-Iron.

Fig 1: Rare Association Rule Mining (RARM)

2. RELATED WORK

MCCFP-Growth approaches:

1. MCCFP-growth use multiple min_sup for tracking down complete set of frequent pattern.

2. MCCFP-Growth required only one scans on the dataset by MIS-value.

3. In MCCFP-growth, tree is fabricated with both frequent items as well as rare items in a transaction dataset.

MIS Tree Approach:

In "Maximum Constraint Model", each item is specified with a support constraint, called **Minimum Item Support (MIS)**. MSapriori algorithm can search rare item rules without producing a big no of meaningless rule minimum item support changed into it. Some subsidiary of frequent item set may not be frequent form rules so their support will be missing. Finally, we can say that MIS-Tree is avoids rescanning & search different frequent item sets.

CP-Tree Approach:

It's functionality for interactive and incremental mining with single database scans tree builds a compact prefix-tree with one scan & performance as Frequent Pattern (FP)-growth technique by efficient tree restructuring process's Tree increase outperforms on execution time and memory requirement.

Steps:
1] Insertion phase: Staring this phases into it inserts transection into CP-Tree according to sort order of I-list and update Freq_count (Frequency Count) of respectively items in I-list. Process into that item having higher count values are rearranging at upper most section of the tree.

2] Restructuring phase: Finishing with this phase into that rearranges the I-list before frequency descending order of items and restructure the tree node in to new I-list & the end of database.

3. MINIMUM ITEM SUPPORT TREE

3.1 Compact-MIS Algorithm

1: Let tree is the prefix-tree is MIS-Tree.

2: for each item i_j in MIS-list of MIS-Tree do

3: if f (i_j) < MIS(i_j) then

4: POP item from the MIS-list with item = i_j.

5: /* Tree Pruning Operation */

6: for each node in the node-link of i_j in Tree do

7: if node is a leaf node then

8: POP node directly;

9: else

10: POP node and then its parent node is linked to its child node(s).

11: end if condition

12: end for loop

13: end if condition

14: /* Tree Merging Operation */

15: for each item i_j in the compact MIS-list do

16: if there are child nodes with the identical item name then

17: merge these nodes and set the count as the addition of these nodes' counts;

18: end if condition

19: end for loop

20: end for loop.

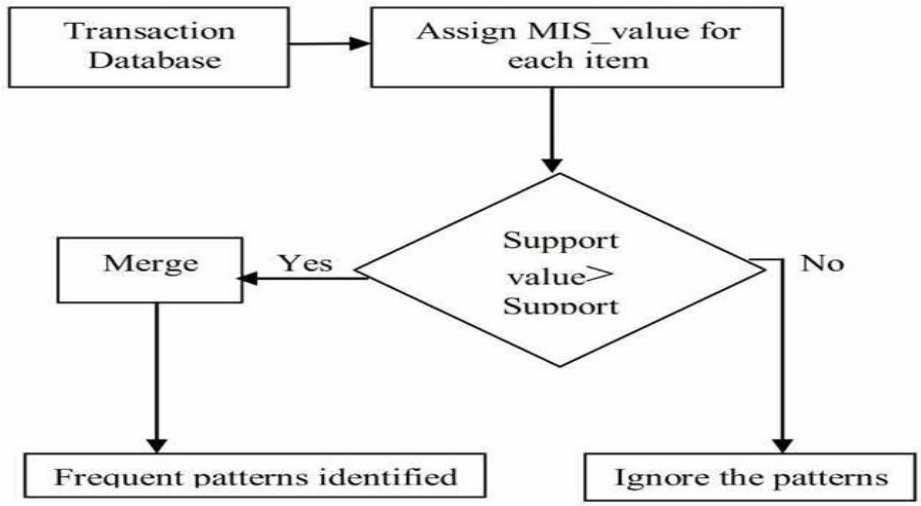

Fig 2: Flow Diagram for MIS-tree [8]

3.2 CP-Tree Algorithm

Let take MIS-List as list for storing details of each items MIS-values.

Let I-list maintain frequent of items while scanning the Data base.Phase-1. Insertion phase

Step-1. For each item of transection form database D.
- Insert item into MIS-tree according to current order of I-list.
Step-2. Update frequency count of respective items in I-list.
End for.

Phase-2. Restructure phase

Step-1. Rearrange the I-list according to frequent Descending to frequent descending order of Items.
Step-2. Restructures the tree nodes according to new I-list.

4. IMPLEMENTATION METHOD

4.1 TOOLS AND TECHNOLOGY

In Computer Programming, Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-ins system for customizing the environment. It written mostly in Java language, Eclipse can be used to develop applications. By using various plug-ins, Eclipse may also be used to develop applications in other programming languages like: C, C++, JavaScript, Perl, PHP, Python, and Ruby. It can also be used to develop packages for the Mathematical Software.

Development of environments include the Eclipse like Java development tools (JDT) for Java, Eclipse CDT for C/C++ and Eclipse PDT for PHP, etc.

4.2 DATASET DESCRIPTION

T10I4D100K is a larger sparse dataset with 100000 transaction and 870 distinct items, Pumsb_star is a larger sparse dataset with 49046 transaction and 2088 distinct items and kosarak has larger sparse dataset with 990002 transaction and 41270 easily distinguishable items and chess is a dense dataset with 3196 transaction and 75 distinct items and mushroom is a smaller dense dataset with 8124 transaction and 119 distinct items.

Table: 1 Dataset characteristics. The term "max," "avg," and "trans" respectively denote maximum, average, transactions.

Dataset	Transactions	Distinct Items	Max. Trans. Size	Avg. Trans. Size
Mushroom	8124	119	23	23
Chess	3196	75	28	20
Kosarak	990002	41270	2498	8.1
T10I4D100K	100000	870	29	10.102
Pumsb_star	49046	2088	988	50

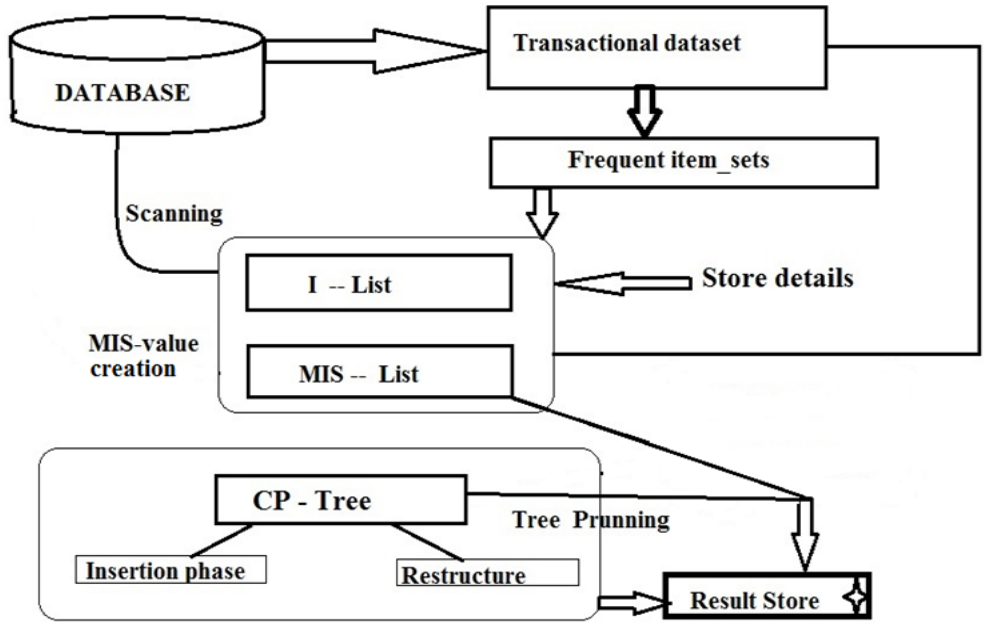

Fig 3: Work Flow of Proposed Approach

5. EXPERIMENTAL RESULT

5.1 IMPLEMENTATION METHOD

Implementation is the phase of the thesis when the theoretical design is turned out into the working system. Thus, it can be considered to be the most critical phase in achieving the successful new system and in giving to user, confidence that generated new system will work and be effective. The implementation phase involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

5.2 RESULT ANALYSIS

Result analysis shows the comparison between MCCFP Growth algorithm and MIS-CP Tree algorithm. Below figure shows the comparison chart between MCCFP Growth and MIS-CP Tree for β value {0.5, 5}.The chart shows how both algorithm consume memory to process for same β values. Here, β values are same for both algorithm and we get memory usage for given β value. The chart demonstrates that for different β values MIS-CP Tree consumes less memory instead of MCCFP Growth.

6. CONCLUSION AND FUTURE WORK

FP-Tree has some limitation like, it required two scan on the dataset, and Frequent Pattern (FP)-Tree is established with

only frequent items in a transaction dataset, with support descending order of items & use only single min_sup. Rare items based on their item support process to reduce both rule missing & rule explosion problem. The benefit of MIS-tree is it avoids rescanning & search the different item sets. CP-tree support interactive mining & incremental mining, it can extract Frequent Pattern without need to rescan database. CP-tree techniques gives advantage of saving Memory & Time. So, we are going to generate MIS-tree based on Compact Pattern-tree (CP-Tree). Future Work: We are going to investigate how popular Non-Iterative approaches and iterative approaches can be extended to assign min_sup and max_sup to each item to extract frequent item sets involving and also we analyze the behavior of various interesting measurement like accuracy, efficiency, performance, scalability on mining rare association rule..

7. REFERENCES

[1] Neelamadhab Padhy, Dr.Pragnyaben Mishra, "The Survey of Data Mining Applications and Feature Scope", DOI: 10.5121, IJCSEIT-2012. Vol.2, No.3.

[2] R.Uday Kiran, P. Krishna Reddy, IN: DASFAA-2010,"Mining Rare Association Rules in the Datasets with Widely Varying Items' Frequencies".

[3] Weimin Ouyang, Qinhua Huang, Mining Direct and Indirect Association Patterns with Multiple Minimum Supports", H: IEEE-2010.

[4] Azadeh Soltani and M. R. Akbarzadeh, "Confabulation-Inspired Association Rule Mining for Rare and Frequent Item sets", IN: IEEE Transaction on Networks and Learning System-2014.

[5] Sandeep Singh Rawat, Lakshmi Rajamani,"Probability Apriori based Approach to Mine Rare Association Rule. "IN: IEEE-2011, 3RD Conference on Data Mining and Organization (DMO).

[6] N. Hoque, B. Nath, D. K Bhattacharyya, "A New Approach on Rare Association Rule Mining" International Journal of Computer Application (0975-8887) vol-53, no.-3. September-2010.

[7] R. Uday Kiran, P. Krishna Reddy, IN: IEEE-CIDM-2009,"An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules".

[8] M. Sinthuja, S. Sheeba Rachel and G. Janani, "MIS -Tree Algorithm for Mining Association Rules with Multiple Minimum Supports"-Binfring International Journal of Data Mining, Vol -1, December -2011

[9] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Young-Koo Lee," CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining" Springer-2008.

[10] R.Uday Kiran and Polepalli Krishna Reddy, "An Efficient Approach to Mine Rare Association Rules Using Maximum Items' Support Constraints" Springer-Verlag Berlin Heidelberg-2012.

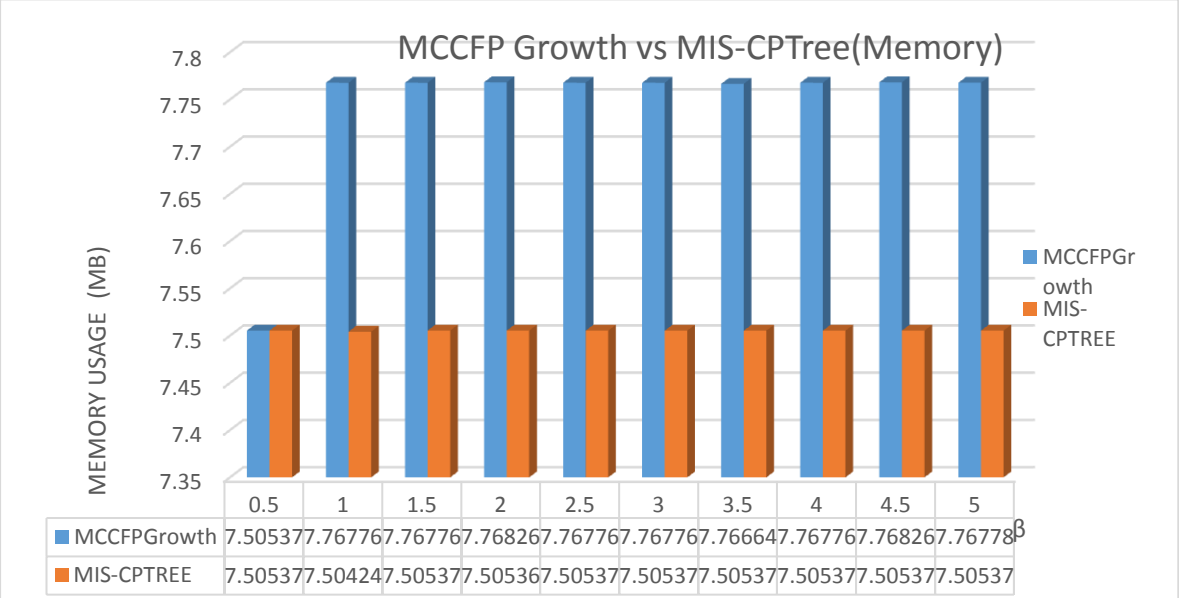

Fig 4. MCCFP GROWTH VS MIS CP TREE