A Novel preprocessing Algorithm for Frequent Pattern Mining in Multidatasets

Dr.K.Duraiswamy *kduraiswamy@yahoo.co.in*
K.S.Rangasamy College of Terchnology,
Tiruchengode -637 209, Tamilnadu, India

B.Jayanthi (Corresponding Author) *sjaihere@gmail.com*
P.G.Department of Computer Science,
Kongu Arts and Science College,
Erode – 638 107, Tamilnadu, India

Abstract

In many database applications, information stored in a database has a built-in hierarchy consisting of multiple levels of concepts. In such a database users may want to find out association rules among items only at the same levels. This task is called multiple-level association rule mining. However, mining frequent patterns at multiple levels may lead to the discovery of more specific and concrete knowledge from data. Initial step to find frequent pattern is to preprocess the multidataset to find the large 1 frequent pattern for all levels. In this research paper, we introduce a new algorithm, called CCB-tree i.e., Category-Content-Brand tree is developed to mine Large 1 Frequent pattern for all levels of abstraction. The proposed algorithm is a tree based structure and it first constructs the tree in CCB order for entire database and second, it searches for frequent pattern in CCB order. This method is using concept of reduced support and it reduces the time complexity.

Keywords: Frequent Patterns, Multiple-level, Association Rule, CCB-tree, Minimum Support.

1. INTRODUCTION

Association rule mining is an important research subject put forward by Agrawal in reference [1]. Association Rule mining techniques can be used to discover unknown or hidden correlation between items found in the database of transactions. The problem of mining association rule could be decomposed into two sub problems, the mining of frequent itemsets/Patterns and the generation of association rules. [1][3].Finding frequent itemsets becomes the main work of mining association rules [2] many applications at mining associations require that mining be performed at multiple levels of abstraction [6].For example; a transaction in the database consists of a set of items. An example of such an association rule might be “80% of customers who buy itemset X also buy itemset Y”. The support count of an itemset is the number of transactions containing an itemset and support of an itemset is the fraction of those transactions besides, finding 80 percent of customers that purchase milk may also buy purchase bread, it is interesting to allow users to drill-down and show that 75 percent of people buy wheat bread if they buy 2 percent milk [10]. The association relationship in the latter statement is expressed at a lower level of abstraction but carries more specific and concrete information than in the former. Therefore a data mining should provide efficient methods for mining multiple-level association rules. To explore multiple-level association rule mining, one needs to provide: 1) data at multiple levels of abstraction, and 2) efficient methods for multiple-level rule mining. In many applications, taxonomy information is either stored implicitly in the database. Therefore, in this study, we generate category-content-brand tree i.e., CCB-tree to find frequent pattern at all levels of abstraction. The proposed algorithm has the following advantages. 1) It generates a frequent pattern at all levels. 2) If follows Top-down deepening Search method. So that searching time is reduced for lower level tree if ancestors are not at minimum support count. It also reduces the execution time.

The rest of the paper is organized as follows. Section gives the basic concept related to multiple level association rules. Section 3 gives the view of the related works. Section4 gives the statement of problem. Section presents the Apriori Algorithm Section6 presents the frequent pattern generation algorithm. Section7 gives the example of the proposed algorithm. Section8 shows the experimental results of the performance of the algorithm. Section9 Concluding remarks of the proposed research work.

2. MULTIPLE-LEVEL ASSOCIATION RULES

We assume that the database contain 1) an item dataset which contain the description of each item in **I** in the form of (A_i, description), where A_i € **I** and 2) a transaction dataset, T, which consist of a set of transaction (T_i { A_p,…. A_q,}), where T_i is a transaction identifier and A_i € **I** for (for **I** = p….q).

To find relatively frequent occurring patterns and reasonably strong rule implications, a user or an expert may specify two thresholds: minimum support, σ' and minimum confidence, φ. For finding multiple-level association rule, different minimum support and/or minimum confidence can be specified at different levels.

Definition 1: The support of an item A in a set S, $\sigma(A/S)$, is the number of transactions(in S) which contain A versus the total number of Transactions in S.

Definition 2: The confidence of $A \rightarrow B$ in S, $\varphi(A \rightarrow B/S)$, is the ratio of $\sigma(AUB/S)$ versus $\sigma(A/S)$, i.e., the probability that item B occurs in S when item A occurs in S.

The definition implies a filtering process which confines the pattern to be examined at lower level to be only those with large support at their corresponding high level. Based on this definition, the idea of mining multiple- level association rules is illustrated below.

TABLE1: A sales transaction table

transaction_id	Bar_code_set
351428	{17325, 92108, 55349…}
982510	{92458, 77451, 60395…}
----	----

Example 1: Let the query to be to find multiple-level association rule in the database in Table 1 for the purchase patterns related to Category, Content and Brand of the food which can only be stored for less than three weeks.

TABLE 2: A sales_item (description) relation

Bar_code	Category	Brand	Content	Size	Storage_pd	price
17325	Milk	Foremost	2%	1(ga)	14(days)	$3.89
----	----	----	---	----	----	----

TABLE 3 : A generalized sales_item description table

GID	Bar_Code_Set	Category	Content	Brand
112	{17325, 31414, 91265}	Milk	2%	Foremost
----	----	----	---	----

The relevant part of the sales item description relation in Table 2 is fetched and generalized into a generalized Sales_item description table, as shown in Table 3, in which is tuple represent a generalized item which is the merge of a group of a tuples which share the same values in the interested attributes. For example, the tuple with the same category, content and brand in Table 2 are merged into one, with their bar codes replace by a bar-code set. Each group is then treated as an atomic item in the generation of lowest level association rules. For example, the association rule generated regarding to milk will be only in relevance to (at the low concept levels) brand (such as Dairyland) and Content (such as 2%) but not to size, producer, etc.

The taxonomy information is provided in table 3. Let Category (such as “milk”) represent the first-level concept, content (such as “2%”) for the second level one and brand (such as “Foremost”) for the third level one. The table implies a concept tree like Fig.1.

The process of mining Multiple-level association rules is actually will be starting from top-most concept level. Let the minimum support at this level be 5% and the minimum confidence is 50%. One may fine the Large 1-itemset: “bread (25%), meat (10%), and milk (20%), Vegetable (30%).

At the second level, only the transactions which contain the large items at the first level are examined. Let the minimum support at this level be 2% and the minimum confidence is 40%. One may find frequent 1-itemsets: “lettuce (10%), Wheat bread (15%), white bread (10%, 2% milk (10%)...”The process repeats at even lower concept level until no large patterns can be found.

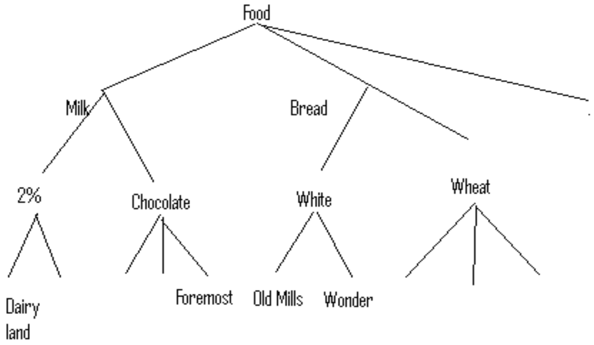

FIGURE 1: taxonomy for the relevant data items.

2. RELATED WORK

Since it was introduced in [1](R.Agrawal,T.Imielinski and A.N.Swami,1993). The problem of frequent itemset mining has been studied extensively by many researchers. As a result, a large number of algorithms have been developed in order to efficiently solve the problem [2][3](R.Agrawal, R.Srikant, 1994, J.Han, J.Pel, Y.Yin, 2000).In practice; the number of works has been focused on mining association rules at single concept level. Thus there has been recent interest in discovering Multiple Level Association rule. A new approach to Find Frequent pattern for multi-level datasets has to be considered. Work has been done in adopting approaches originally made for single level datasets into techniques usable on multi-level datasets. The paper in [4] Han & Fu (1995) shows one of the earliest approaches proposed to find frequent itemsets in multi-level datasets and later revisited in [5] Han & Fu (1999). This work primarily focused on finding frequent itemsets at each level in the dataset. The paper in [11] (Thakur, Jain & Pardasani 2006) proposed to find cross-level frequent itemsets. The paper in (8) (Pratima Gautham & K.R. Pardasani 2010) proposed efficient version of Apriori approach to find large 1 frequent pattern. The paper in [9] (Popescu, Daniela.E, Mirela Pater 2008) proposed AFOPT algorithm. The paper in [12] (Yinbo Wan, Yong Liang, Liya Ding 2009) proposed a novel method to extract multilevel rules based on different hierarchical levels by organizing and extracting frequent itemsets mined from primitive data items. The paper in [7](Mohamed Salah Gouider, Amine Farhat 2010) proposed a technique for modeling and interpretation of constraints in a context of use of concept hierarchies. However, even with all this work the focus has been on finding the large 1 frequent pattern using Apriori algorithm method. This work attempts to find the Large 1 frequent pattern for all levels with new approach i.e., CCB-tree using reduced support.

3. PROBLEM STATEMENT

The problem of mining multiple-level association rules was introduced in [4](Han & Fu (1995)), [5]Han & Fu(1999), [11](Thakur, Jain & Pardasani 2006), [8](Pratima Gautham & K.R. Pardasani 2010), [9] (Popescu, Daniela.E, Mirela Pater 2008), [12] (Yinbo Wan, Yong Liang, Liya Ding 2009), [7](Mohamed Salah Gouider, Amine Farhat 2010). There are two steps in association rule mining. First step is to find Large 1 frequent patterns for all level and then Large2...LargeK frequent pattern and Second step is to generate Association rules. We focus on first step i.e., finding large 1 Frequent Patterns at all levels. The objective of this work is to construct category-content-Brand tree (CCB-tree) in depth first order and it search for the large 1 frequent pattern in the same order so that it reduces the searching time. In this work, an algorithm CCB-tree is proposed, to find the frequent patterns for different levels. More specifically, given a transaction database TD, a different minimum Support for each level.

4. PROPOSED ALGORITHM

```
Algorithm CCB-tree construction and mining:
Input:
1. Transaction Database TD, minimum support (min_sup) for all levels
Output:
  Large 1 Frequent pattern for all levels.
Steps:
1. Create the root of the CCB-tree T with label “Null”
2. For each transaction Trans in TD do the following
3. Select items in Trans
4. Let item list in Trans be [p/P], where p is the first element and each element has a
   dimension d and P is the remaining list
5. Call Insertion ([p/P], T)
6. Call mining(T)
7. End for
8. Function Insertion ([p/P],T)
9. //Search a tree T for Key Value $P^1$,.. $P^d$. It is assumed that branching is determined by
    the dimension d of the key value//
10. For i = 1 to d by 1 do
11. If T has a child $N^i$ such that $N^i$.itemName = $p^i$.itemName
12. Then $N^i$.Count = $N^i$.Count + 1 and Trans_id = TID
13. Else
14. If i <d Create a new node with 3 fields i.e., item.name, Count, Trans_id
15. Then $N^i$.itemName = $p^i$.itemName , $N^i$.Count = $N^i$.Count + 1 and Trans_id =
    TID
16. Else Create a new node with 2 fields i.e., item.name, Count
17. Then $N^i$.itemName = $p^i$.itemName , $N^i$.Count = $N^i$.Count + 1
18. End If
19. Increment i and perform steps from 9 to 16.
20. End For.
21. Function mining (T)
22. Put the initial node in T on a list search
23. If initial node. count>=min_sup print its item.name, count and
24. Move towards its descendents i.e., next level by level of the same parent and
25. Print its item.name, count
26. Else move to the successors of initial node
27. End If
28. End For
```

5. EXAMPLE

This Section shows the example to demonstrate the proposed algorithm to mine Large 1 frequent pattern in multidatasets, which uses a hierarchy information encoded transaction table [5]. This based on the following consideration, first a data mining is usually in relevance to only a portion of the transaction database, such as food instead of all the items. It is beneficial to collect the

relevant set of data and then work repeatedly on the task-relevant set. Second, encoding can be performed during the collection of task-relevant data and thus there is no extra “encoding pass” required. Third, an encoding string, which represents a position in a hierarchy, required fewer bits than the corresponding object identifier or bar-code.
An abstract example, which simulates the real life example of Example 1, is analyzed as follows:

Example 2: The taxonomy information for each (grouped) item in Example 1 is encoded as a sequence of digits in the transaction table4. For example, the item ‘2% Foremost milk’ is encoded as ‘112’ in which digit, ‘1’ represents ‘milk’ at level-1, the second, ‘1’, for ‘2%(milk)’ at level-2 and the third,’2’, for the brand ‘Foremost’ at level-3. Similar to Agrawal and Srikant [2], repeated items at any level will be treated as one item in one transaction.The derivation of large 1 itemsets at all levels proceed as follows.

TABLE4: Sample Data

TID	Items
T1	{111, 121, 211, 211}
T2	{111, 211, 222, 323}
T3	{112, 122, 221, 411}
T4	{111, 121}
T5	{111, 122, 211, 221, 413}
T6	{113, 323, 524}
T7	{131, 231}
T8	{323, 411, 524, 713}

CCB-Tree Construction:
Let T1 = {111, 121, 211, 211} and p be a data with 3 dimensions, i.e., 1-category, 2-content and 3-Brand.Consider level 1(dimension 1 of first item) search a tree for key value. It is assured that level is determined by the dimensions d of p. If key values are not in tree, create a node with item.name, count and transaction id.

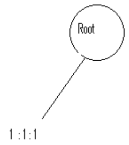

FIGURE 1: First level 1: item.name 1 : count and 1: trans_id

Consider level 2 (dimension 2 of first item) searches a tree for key value. If key values are not in tree, create a node with item.name, count and transaction id.

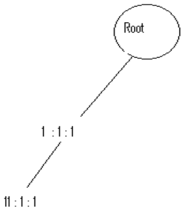

FIGURE 2: Second level

Consider level 3 (dimension 3 of first item) searches a tree for key value. If key values are not in tree, create a node with item.name, count.

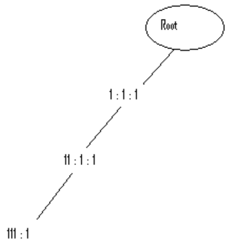

FIGURE 3: Third level

After T1 is over the appearance of CCB- Tree:

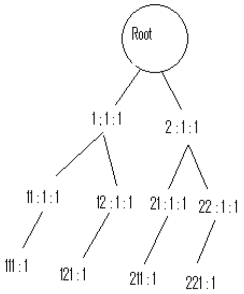

FIGURE 4: CCB-tree for T1

After the complete construction of CCB-Tree for the Table4:

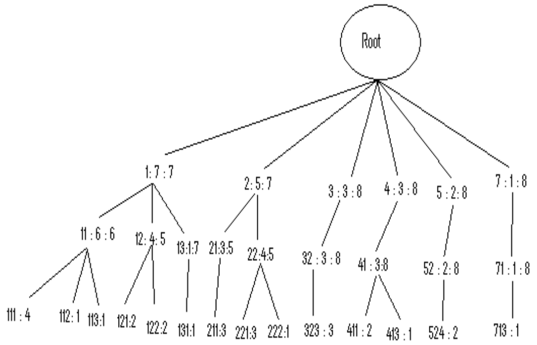

FIGURE 5: CCB-tree for Table4

CCB-Tree Mining Process:

Minimum support for all levels is 4, 3, and 3:

Mining starts from the left most initial node i.e., from 1**: 7 > min_sup and its descendents 11*:6>3 and 111>3. But 112,113<3 so it's considered to be a large 1 frequent pattern.

Finally frequent pattern for level 1: 1**, 2** Level 2: 11*, 12*, 21*, 22* Level 3:111,211,221.

6. EXPERIMENTAL ANALYSIS

Here, we study the experimental analysis of CCB-tree algorithm to mine large-1 frequent pattern.

As far as we know, the Apriori algorithm [1 – 5, 11,14] is the only other algorithm that has been designed to mine large-1 frequent pattern. So the first set of experiments we conduct is to compare our algorithm CCB-tree with Apriori.

We also provide the following results for CCB-tree with different choices of the Threshold for different levels; the performance as database size scales.

Finally, we examine the performance of CCB-tree with respect to a synthetic transactional database generated by IBM Quest Market-Basket Synthetic data generator [13]. We used 5000 datasets with three levels; top level of tree has 10 items.

The algorithms were implemented in C language and executed on a Windows machine with Intel CPU.

Threshold	Minimum support thresholds
1	[50, 40, 30]
2	[40, 30, 30]
3	[30, 20, 20]

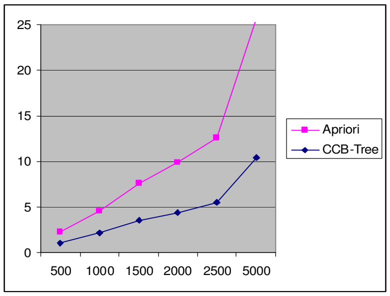

FIGURE 6: Threshold 1

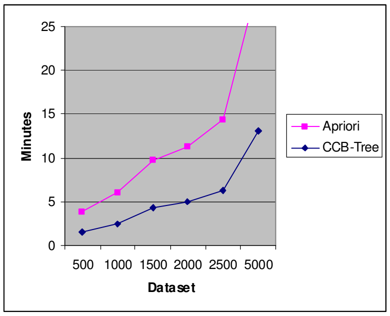

FIGURE 7: Threshold 2

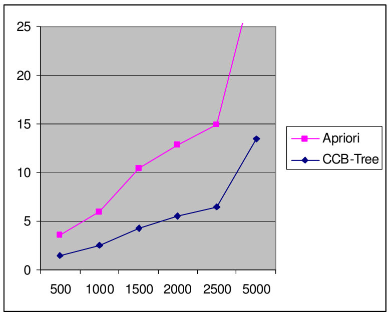

FIGURE 8: Threshold 3

Fig 6 - 8 shows performance measurements for mining large-1 frequent pattern using CCB-tree and Apriori algorithm. The running time and the number of transactions are shown to different minimum support thresholds for different levels ranging from 50 to 20.The above three figures shows two interesting features. First, the relative performance of the two algorithms under any setting is relatively independent of the number of transactions used in the testing, which indicates that the performance is highly relevant to threshold setting. Second, the CCB-tree algorithm have relatively good 'scale-up' behavior since the increase of the number of the transactions in the database will lead to approximately the linear growth of the processing of large transaction databases.

7. CONCLUSION AND FUTURE WORK

Transaction databases in many applications contain data that has built-in hierarchy information. In such databases, uses may be interested in finding association rules among items only at the same level or association rules that span over multiple levels in the hierarchy. In this paper, we presented an efficient preprocessing algorithm for Frequent Pattern Mining in Multidatasets. This algorithm can be used as initial processing step to find frequent pattern generation. As a result, its execution time is much smaller than that of Apriori-based algorithm so that overall time complexity for frequent pattern generation can be reduced.. We conducted extensive experiments and the results confirmed our analysis. In future an efficient algorithm can be generated for frequent pattern mining in multidatasets based on transaction reduction concept.

REFERENCES

[1] Agrawal R,Imienlinski T,Swami A,(1993).Mining association rules between sets of items in large databases. In Proc. Of the ACM SIGMOD Int. Conf. on Management of Data, Pages 207-216.

[2] Agrawal R, and Srikant R, (1994). Fast algorithms for mining association rules. In Proc. Of the 20th Int. Conf. on very Large Databases. Pages 487-499.

[3] Han .J ,Pei .J, and Yin .Y,(2000) Mining Frequent patterns without candidate generation. In Proc. Of ACM-SIGMOD Int. Conf. on Management of Data, pages 1-12.

[4] Han, J., Fu, Y., Discovery of Multiple-Level Association Rules from Large Databases, in Proceedings of the 21st Very Large Data Bases Conference, Morgan Kaufmann, P. 420-431, 1995.

[5] Han, J., Fu, Y., Mining Multiple-Level Association Rules in Large Databases, in IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 5, September/October 1999.

[6] Mehmet Kaya, Reda Alhajj, " Mining Multi-Cross-Level Fuzzy Weighted Association rules", Second IEEE International Conference on Intelligent Systems.Vol.1,pp.225-230, 2004

[7] Mohamed Salah Gouider, Amine Farhat, "Mining Multi-level Frequent Itemsets under Constraints", International Journal of Database Theory and Application Vol. 3, No. 4, December, 2010

[8] Pratima Gautham, Pardasani, K. R., "Algorithm for Efficient Multilevel Association Rule Mining", International Journal of Computer Science and Engineering, Vol.2 pp. 1700-1704, 2010.

[9] Popescu, Daniela.E, Mirela Pater, "Multi-Level Database using AFOPT Data Structure and Adaptive Support Constraints", Int. J. of Computers, Comm. & Control, Vol.3,2008.

[10] Rajkumar.N, Karthik.M.R, Sivanada.S.N, "Fast Algorithm for mining multilevel Association Rules,"IEEE Trans. Knowledge and Data Engg., Vol.2 pp. 688-692, 2003.

[11] Thakur, R. S., Jain, R. C., Pardasani, K. R., Mining Level-Crossing Association Rules from Large Databases, in the Journal of Computer Science 2(1), P. 76-81, 2006.

[12] Yinbo WAN, Yong LIANG, Liya DING, "Mining Multilevel Association Rules from Primitive Frequent Itemsets", Journal of Macau University of Science and Technology, Vol.3 No.1, 2009

[13] Synthetic Data generation Code for Associations and Sequential Patterns (IBM Almaden Research center). http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html.

[14] Gavin Shaw, 'Discovery & Effective use of Quality Association Rules in Multi-Level Datasets ", Ph.D-Thesis, Queensland University of Technology, Brisbane, Australia,2010.