Different Strategies for Mining Erasable Itemsets

Suchi shah[1] Jayna B Shah[2]
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]SVIT, Vasad

Abstract— Mining Erasable itemsets allows the manager to consider their production plan carefully and also ensure the stability of the factory. There are many strategies for finding the erasable itemsets based on Apriori, VME , Mining based on NC Sets ,Difference of NC Sets, Mining Top Rank –K erasable itemsets, MEI and current issues are studied.

Key words: Data mining, Erasable itemsets mining

I. INTRODUCTION

Data mining is the boon of IT industry. Upto now we are using the frequent pattern mining to find that which items are frequent. For e.g. in retail store we see the buying habits and see that bread and butter are purchased frequently. But, now when we see a production plan than a problem of mining erasable itemsets originates. In a manufacturing factory, the products are produced on a very large scale. Also, each product is formed with a few components or materials. The factory has to spend the large amount to purchase and also store this component to manufacture their products. But, the factory cannot purchase all the needed components when there is a financial crisis. So, now a big question for the managers is that how to decide the production plan due to the limited money. As they cannot purchase all the components due to the financial crisis, hence one needs to stop the manufacturing of some products.

So, managers can create the new production plan by finding which itemsets can be erased i.e. eliminated. This is known as erasable itemsets. But, by stopping the manufacturing of some products the loss of factory's profit should be controllable. The organization of the paper is as section 2 gives the erasable itemset mining problem. Section 3 briefly explains the different erasable itemset mining algorithms. .Section 4 gives the conclusion.

II. ERASABLE ITEMSETS MINING PROBLEM

Here for finding erasable itemsets the problem is decomposed into two sub-problems:

(1) The first sub-problem is again divided into two. The first is to compute the sum of all products and second is to find the gain of itemsets.[7]
(2) 2. Second sub-problem is to find the erasable itemsets by comparing the gain of itemsets and the threshold.[7]

Product	Items	Val(million $)
P1	a,b,c	2100
P2	a,b	1000
P3	a,c	1000
P4	b,c,e	150
P5	b,e	50
P6	c,e	100
P7	c,d,e,f,g	200
P8	d,e,f,h	100
P9	d,f	50
P10	b,f,h	150
P11	c,f	100

Table 1: Consider the Product database

Let I= {i1,i2…, im} be a set of all items, which shows the components of products. Here DB={P1,P2,….Pn) represents the product database, where Pi($1 \le i \le n$) is a product presented in the form of <Items, Val>,where items are the components and the Val is the profit that the factory obtains by selling the product Pi. In this database, {a, b, c, d, e, f, g, h} is the set of items (components) used to create all products {P1, P2, …, P11}. Product P2 is made from two components {ab}. The factory earns 1000 million dollars by selling product P2. [6]

A. *The Gain Of Itemset:*

The gain of itemset X can be calculated by formula as

$$g(X)= \sum_{\{Pk / X \cap Pk.Items \neq \phi\}} Pk.Val$$

The gain of itemset X is the sum of profits of the products which include at least one item in itemset X. For instance, let X ={ac} be an itemset. From the example database in Table 1, { P1, P2, P3, P4, P6, P7, P11} are the products which include {a}, {c}, or {ac} as components. Therefore, g(X) = P1.Val + P2.Val + P3.Val + P4.Val + P6.Val + P7.Val + P11.Val =4650 million dollars.[6]

B. *When Items Are Erasable:*

Given a threshold ξ and a product database DB, let T be the total profit of the factory. An itemset X is erasable if:

$g(X) \le T * \xi$

Where T is computed as follows

$$T= \sum_{Pk \in DB} Pk.Val$$

The total profit of the factory is the sum of profits of all products. From the example database in Table 1, the total profit of the factory T is 5000 million dollars. An itemset X is called an EI if $g(X) \le T*\xi$. For example, let the threshold be 16% (ξ =16%). According to Definition 1, g ({e}) = 600 million dollars. Item e is called an EI with ξ = 16% because g ({e}) =600≤5000*16%=800. This means that the factory does not need to buy and store item e. In that case, the factory will not manufacture the products P4, P5, P6, P7, and P8.[6]

III. LITERATURE SURVEY

There are various algorithms for mining erasable itemsets.

A. *META:*

META Algorithm is used for Mining Erasable iTemsets with the Anti-monotone property. It adopts the Apriori

All rights reserved by www.ijsrd.com

algorithm and uses iterative approach with level wise search. In this level wise approach, first step is to find the set of erasable 1-itemsets.This set is denoted by E1.Next step is to find the set of erasable 2-itemsets from the erasable 1-itemsets.This is denoted by E2. Similarly, we need to find k- erasable itemset upto no more sets are found. Also, it uses the horizontal data format for finding the gain .But, when we find k-erasable itemsets it scans the database k+1 times. And only the items with same prefix are combined. So, the itemsets may be redundant.[1]

B. VME:

VME Algorithm uses the same approach as META, but the data structure used is vertical for mining erasable itemsets. It uses the PID_list for storing the identification number of products, which stores in the format <Item, gain>. It can remove the irrevalant data easily. In level wise approach, first step is to find the set of erasable 1-itemsets.This set is denoted by E1.Next step is to find the set of erasable 2-itemsets from the erasable 1-itemsets.This is denoted by E2. Similarly, we need to find k- erasable itemset upto no more sets are found.

Item	Inverted List
a	<1,2100>,<2,1000>,<3,1000>
b	<1,2100>,<2,1000>,<4,150>,<5,50>,<10,150>
c	<1,2100>,<3,1000>,<4,150>,<<6,100>,<7,200>, <11,100>
d	<7,200>,<8,100>,<9,50>
e	<4,150>,<5,50>,<6,100>,<7,200>,<8,100>
f	<7,200>,<8,100>,<9,50>,<10,150>,<11,100>
g	<7,200>
h	<8,100>,<10,150>

But, scanning the database takes a lot of time and memory. VME stores each value which leads to data duplication.[2]

C. MERIT:

MERIT algorithm is used for mining fast erasable itemsets using NC-Sets. To keep the track of complete information the new structure is used called as NC-Sets (node code sets) as it is a compact structure. Also, it can efficiently remove the irrelevant data. Using NC-Sets, the gain of itemsets can be computed in a linear time complexity. Without generating candidate itemsets it can find erasable itemsets sometimes. The NC-Sets builds the WPPC Tree and WPP-Code with <(pre-order, post-order):weight>.But, MERIT uses the union strategy and hence requires the large memory. It scans the database many times and so takes lot of time. Also it stores the value which leads to the data duplication.[4]

D. MIKE:

MIKE algorithm is used for mining top rank k erasable itemsets, where k is the maximum value to be mined. It is the task of finding the erasable itemsets whose ranks is no greater than k. Here the unwanted results is removed and only the required results are generated. So, the search space is also reduced to a larger extent. But, MIKE can only find the top rank k erasable itemset. So ,deciding the value of k is very difficult.[3]

E. MERIT+:

MERIT+ is an algorithm which is used for mining erasable itemsets using difference of NC-Sets. It is similar to MERIT and which is then established for the foundation of dMERIT+.To improve the mining time the weight index, a hash table and the difference of Node Code Sets (dNC-Sets) are used. Also, the memory usage is reduced to a large extent. But, not capable to say whether it is erasable or not by checking all its subsets.[5]

F. MEI:

MEI algorithm uses the divide and conquer strategy and dpidset i.e. difference of two pidsets concept for mining erasable itemsets. As only the difference is stored so the memory usage and time are also reduced. So from erasable 1 –itemset we can find erasable 2-itemsets and so on upto k-erasable itemsets. Hence, we can efficiently find the erasable itemsets. But, the problem is that it cannot be used for incremental approach.[6]

IV. CONCLUSION

Here the erasable mining problem is divided into two sub-problems by calculating the gain of itemsets and by finding the erasable itemsets. So, vertical data format is appropriate than horizontal data format. And also top rank k approach is more efficient then the threshold approach. In future work , mining erasable itemsets from huge database. Some issues related to mining closed/maximal erasable itemsets, mining erasable itemsets from incremental databases, mining of rules from erasable itemsets , also how to use the erasable itemsets in recommendation systems can be studied.

REFERENCES

[1] Zhi-Hong Deng,Guo-Dong Fang,Zhong- Hui Wang,Xiao-Ran Xu, 2009, Mining erasable itemsets.
[2] Zhihong Deng,Xiaoran Xu,2010, An efficient algorithm for mining erasable itemsets.
[3] Zhihong Deng,Xiaoran Xu,2011,Mining top-rank-k erasable itemsets.
[4] Zhi_Hong Deng,Xiao_Ran Xu,2012, Fast mining erasable itemsets using NC_sets.
[5] Tuong Le,Bay Vo,Frans Coenen,2013,An efficient algorithm for mining erasable itemsets using the differnence of NC-Sets.
[6] Tuong Le,Bay Vo,2013,MEI:An efficient algorithm for mining erasable itemsets.
[7] Ms Shewta,Dr.Kanwal Garg,2013,Searching best strategies of mining erasable itemsets.
[8] Thein-Phuong Le, Tzung-Pei Hong, Bay Vo, Bac Le,2012,An efficient incremental mining approach based on IT-tree.
[9] Jain Pei, Jiawei, Runying Mao, Closet:An efficient algorithm for mining frequent closed itemsets.
[10] Doug Burdick, Manuel Calimlim, Jason Flannick, Johannes Gehrke,2005,Mafia: A maximal frequent itemset algorithm.

All rights reserved by *www.ijsrd.com*