ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 6, June 2015

Review Paper: Generation of Erasable Itemset by Economical Method

Anuradha Panjeta[1], Chhavi Miglani[2]

CSE & Kurukshetra University[1,2]

Abstract: This paper presents the work carried out is an MVME which is an improved version of VME(Vertical format based).MVME is an algorithm which is used to find erasable itemset. To find erasable itemset we need to do mining of data. Need for mining erasable itemset originate from production planning problem. Data mining refers to extracting or mining knowledge from large amounts of data. VME algorithm is doing the process in same manner as APRIORI algorithm do, but there is difference that APRIORI is used to find frequent itemset and it shows result on the basis of minimum support value where as VME is used to find erasable itemset and it shows result on the basis of maximum threshold value. MVME which is an improved version of VME also works on the approach of APRIORI , there are two major difference between MVME and VME. In MVME after generating candidate we directly calculate the gain of itemset where as in VME , it checks the subsets first and if subsets found then it will calculate the gain of itemset .Second difference is that, in MVME we compare the gain with threshold value at the time of candidate generation where as in VME it store all the candidate of a level and return this candidate set to calling procedure , In calling procedure candidate compare with threshold value and gets the final result. So conclusion is that MVME produces the same results but in lesser time.

Keywords: Data Mining, MVME, VME,APRIORI.

I. INTRODUCTION

DATA MINING

Data Mining is core part of Knowledge Discovery process (KDD). The KDD process consist of data selection, data cleaning, data transformation, pattern searching (data mining) and finding pattern evaluation. Focusing specially, on the definition of data mining, it has been described as " the task of discovering interesting patterns from large amount of data where the data can be stored in databases, data warehouses or other information repositories" [1].

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining.

Thus, data mining should have been more appropriately named "knowledge mining from data," which is sometime known as knowledge mining. Many other term carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and knowledge discovery from data or KDD.

But data mining is the central part of the KDD process. Knowledge discovery as a process consists of sequence of following steps as shown in Fig 1.1:

1. Data integration (where multiple data sources may be combined).

2. Data selection (where data relevant to the analysis task are retrieved from the Database).

3. Data preprocessing (to remove noise and inconsistent data).

4. Data transformation (where data are transformed into forms appropriate for mining by performing summary or aggregation operation).

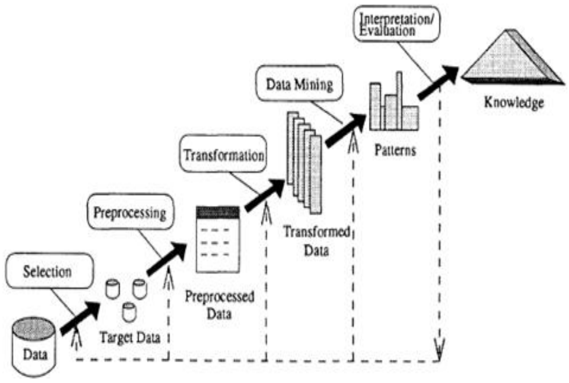


Fig 1.1 Knowledge Discovery in Database (KDD) Process [25]

5. Data mining (an essential process where intelligent methods are applied in order to extract data pattern).

6. Pattern evaluation (to identify the truly interesting pattern representing knowledge based on some measure).

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

Data mining is the process of discovering interesting knowledge from large amounts of data stored in the database, data warehouse, or other information repositories.

Data mining is only one step of the process, involving the application of discovery tools to find interesting patterns from targeted data. Flow of the data mining process can be shown by Fig 1.2

DOI 10.17148/IARJSET.2015.2625

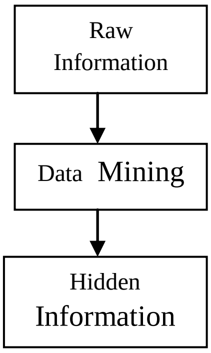


Fig 1.2 Flow of the data mining process

A data mining session is usually an interactive process of data mining query submission, task analysis, and data collection from the database, interesting pattern search, and findings presentation [1].

II. PROCESS FOR MINING THE DATA

An important concept is that building a mining model is part of a larger process that includes everything from defining the basic problem that the model will solve, to deploying the model into a working environment. This process can be defined by using the following six basic steps.

- Defining the problem
- Preparing data
- Exploring data
- Building models
- Validating models
- Updating models

Need of Data Mining

You will need to automate:

- The right offer
- To the right person
- At the right time
- Through the right channel

The right offer means managing multiple interactions with your customers, prioritizing what the offer will be while making sure that irrelevant offers are minimized. The right person means that not all customers are cut from same cloth. Your interactions with them need to move towards highly segmented marketing campaigns that target individual wants and needs. The right time is result of the fact that interactions with customers now happen on continues basis. This is significantly different from the past, when quarterly mailings were cutting edge marketing. Finally, the right channel means that you can interact with your customers in variety of ways (direct mail, email, telemarketing, etc).

III. WHAT CAN DATA MINING DO FOR US

- **Identify our best prospect:** - By concentrating marketing efforts only on the best prospects Data Mining can save time and money, thus increasing effectiveness of the marketing operation.
- **Predict cross sell opportunities and make recommendations:** - Whether there is a traditional or web based operation, information can be provided quickly to the customers, which helps in increasing the value of each communication with the customers.
- **Learn parameters influencing trends in sales and margins:** - One may think this can be done with OLAP (Online Analytical Processing) tools.
- **Segment markets and personalize communications:** - There might be distinct groups of customers, patients, or natural phenomena that require different approaches in their handling.

IV. LITERATURE REVIEW

In this chapter different types of algorithms based on horizontal format Vertical are discussed briefly.

David W. Cheung et al., (1996) presents an efficient algorithm named MLUp (multiple level association rules update) that updating of discovered multi-level association rules. This algorithm is applicable only to a database which allows frequent or occasional updates restricted to insertions of new transactions [2].

Kishore B. Kumar et al.,(2006) proposes a new algorithm, efficient hierarchical online rule mining, which optimizes the time requirements of the earlier reported algorithm HORM. The proposed new algorithm incorporates two specific enhancements: hierarchy-aware counting and transaction reduction in the Phase I sub-problem. For Phase II, we have modified in a natural way the algorithm of to model the generation of hierarchical association rules. HORM algorithm takes as input the transaction database *D*, and a set of classes or sub-classes of interest, denoted *SIC*. After one scan over the database, the count-arrays of all classes or sub-classes of interest contain the number of occurrences, i.e. support values ,in the transaction database of all the subsets of the classes or sub-classes of interest. The time complexity of this algorithm is $O(jDjK2M)$ The memory requirement of HORM is K2M, since each element of SIC requires an array of size 2*M*. [3].

MVME & VME ALGORITHMS

VME is an algorithm for mining erasable itemset

In MVME first we scan the database : By this process we find the "Sum-value" i.e. total value by selling all the products, as we do in VME.

RELATIVE MVME & VME ALGORITHMS

These components are also called as erasable itemsets. There is an existing algorithm for mining erasable itemsets called Vertical-format-based algorithm for Mining Erasable itemsets[1].VME algorithm is inverse of Apriori algorithm. In Apriori the database is represented in horizontal format,where as in VME the database is in Vertical format. So we don't need to check the subsets.VME is an algorithm for mining erasable itemset. Mining erasable itemset means to find out items from an itemset which can be erased. In VME, to erase an itemset, we check the gain of an itemset and if it is less than the threshold value then we will add that item into list of erasable itemsets. In this problem we have products and each product consist of some items. Each product is represented in this form <PID,ITEMS,VALUE>. PID is product identifier, ITEMS are all those items which make

DOI 10.17148/IARJSET.2015.2625

product, VALUE is the profit that a organization obtain by selling product.

PRODUCTS	**PID**	**ITEMS**	**VALUE**
P1	1	i1,i3,i5	1200
P2	2	i4,i5,i6	2200
P3	3	i2,i3,i9	1100
P4	4	i2,i7	700
P5	5	i1,i8,i10	2500
P6	6	i7,i8,i9,i10	2600
P7	7	i2,i4,i6	2200
P8	8	i6,i8	2500

Table 1: (Database in horizontal format)

Items	**INVERTED LIST**
i1	<1,1200>,<5,2500>
i2	<3,1100>,<4,700>,<7,2200>
i3	<1,1200>,<3,1100>
i4	<2,2200>,<7,2200>
i5	<1,1200>,<2,2200>
i6	<2,2200>,<7,2200>,<8,2500>
i7	<4,700>,<6,2600>
i8	<5,2500>,<6,2600>,<8,2500>
i9	<3,1100>,<6,2600>
i10	<5,2500>,<6,2600>

Table 2: (Vertical format of Table 1)

Let's take an example of 'P5' product <5,{i1,i8,i10},2500>. Here '5' is identifier of product and items are { i1,i8,i10} and have a profit value 2500. But we have to erase items, so we need to calculate the gain of items.

TO calculate the gain of itemset, we need to sum the profit of all the products that have at least one item as their component. After calculating gain of an itemset we compare that gain with Threshold value, if gain is less than Threshold value then we add that itemset in the list of erasable itemset.

Calculating gain of an itemset , Let {i1} be an itemset and we want to calculate the gain of itemset{i1}, we need to sum all the products value that have {i1} itemset as their component, so P1 and P5 are products which have {i1} as their component . So gain of itemset{i1} is 3700(1200+2500) .

In VME instead of using horizontal data format it uses a vertical data representation. Vertical data representation is nothing but just the inverse of horizontal data representation. The working of VME is explained as below:-

Let TABLE 2 as Database and a threshold value of **30 %**

First we scan the database to find the set of erasable 1-itemset and for that we need to check the gain of each item and compare it with threshold value and if it less than threshold value than that itemset will be add in table 3 .

i1	<1,1200>,<5,2500>
i2	<3,1100>,<4,700>,<7,2200>
i3	<1,1200>,<3,1100>
i4	<2,2200>,<7,2200>
i5	<1,1200>,<2,2200>
i7	<4,700>,<6,2600>
i9	<3,1100>,<6,2600>

Table 3: (List of Erasable 1-itemset)

Product	**PID**	**Items**	**VAL(Thousand $)**
P1	1	{*i*2, *i*3, *i*4, *i*6}	50
P2	2	{*i*2, *i*5, *i*7}	20
P3	3	{*i*1, *i*2, *i*3, *i*5}	50
P4	4	{*i*1, *i*2, *i*4}	800
P5	5	{*i*6, *i*7}	30
P6	6	{*i*3, *i*4}	50

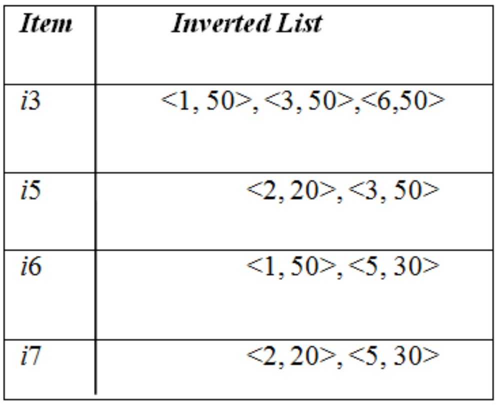
Table 4: (Product Database for MVME)

As we already know that we use vertical data representation, Vertical representation is just inverse of above table. Here we show a table 5 in which data is represent in vertical format. First scan the database and find the sum value which is 1000 in that case and we take 18% as threshold value which is 180.So any itemset whose gain is less than180,will come in the list of erasable itemset. So erasable 1-itemsets are table 6.

Item	***Inverted List***
*i*1	<3, 50>, <4, 800>
*i*2	<1, 50>, <2, 20>, <3, 50>,<4,800>
*i*3	<1, 50>, <3, 50>, <6, 50>
*I*4	<1, 50>, <4, 800>, <6, 50>
*I*5	<2, 20>, <3, 50>
*I*6	<1, 50>, <5, 30>
*I*7	<2, 20>, <5, 30>

Table 5: (Vertical Format of Table 4)

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 6, June 2015



Item	Inverted List
i3	<1, 50>,<3, 50>,<6,50>
i5	<2, 20>, <3, 50>
i6	<1, 50>, <5, 30>
i7	<2, 20>, <5, 30>

Table 6: (List of erasable-1 itemset in MVME)

It is clearly shown from the above results that for varies threshold values and number of transaction the execution time to generate erasable itemsets of MVME are less than VME. It is concluded from the result table that MVME algorithm is 5% to 10% faster than VME algorithm. The results are same at each level for both algorithm, but execution time is less in MVME.

Jian Pei et al., (2001) Methods for efficient mining of frequent patterns have been studied extensively by many researchers. However, the previously proposed methods still encounter some performance bottlenecks when mining databases with different data characteristics, such as dense vs. sparse, long vs. short patterns, memory-based vs. disk-based, etc. In this study, we propose a simple and novel hyperlinked data structure, H- struct, and a new mining algorithm, H-mine, which takes advantage of this data structure and dynamically adjusts links in the mining process.

A distinct feature of this method is that it has very limited and precisely predictable space overhead and runs really fast in memory-based setting. Moreover, it can be scaled up to very large databases by database partitioning, and when the data set becomes dense, (conditional) FP-trees can be constructed dynamically as part of the mining process. Our study shows that H-mine has high performance in various kinds of data, outperforms the previously developed algorithms in different settings, and is highly scalable in mining large databases.

This study also proposes a new data mining methodology, space-preserving mining, which may have strong impact in the future development of efficient and scalable data mining methods [4].

Yves Bastide et al., (2001) finds that data mining has been well studied for several decades. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data and the science of extracting useful information from large data sets or databases.

" In general, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. And technically, it is the process of finding correlations or patterns among dozens of fields in large relational databases [5].

N. Rajkumar et al., (2003) purposes two algorithms AprioriNewMulti and AprioriNewSingle, for Data Mining multilevel and single level association rules in large database respectively. The algorithms introduce a new concept called multi minimum support i.e. minimum support will vary for different length of the itemset. Unlike other algorithms AprioriNewMulti does not depend on number of levels in concept hierarchy i.e. it doesn't scan the database for each level of abstraction for finding association rules. This paper extends the scope of the study of mining from single level to multiple level association rules from a large transaction databases. Mining multiple-level association rules may lead to progressive mining of refined knowledge from the data and have interesting applications for knowledge discovery in transaction-based, as well as other business or engineering databases. With the help of multi minimum support concept, finding interesting frequent itemset in higher length of itemsets made easy.

Jiawei Han et al.,(2004) Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an *Apriori*-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist a large number of patterns and/or long patterns. In this study, we propose a novel frequent-pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, *FP-growth*, for mining *the complete set of frequent patterns* by pattern fragment growth. Efficiency of mining is achieved with three techniques: (1) a large database is compressed into a condensed, smaller data structure, FP-tree which avoids costly, repeated database scans, (2) our FP-tree-based mining adopts a pattern-fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space. Our performance study shows that the *FP-growth* method is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the *Apriori* algorithm and also faster than some recently reported new frequent-pattern mining methods [6].

Christian Borgelt (2005) Recursive elimination is an algorithm for finding frequent item sets, which is strongly inspired by the FP-growth algorithm and very similar to the H-mine algorithm. It does its work without prefix trees or any other complicated data structures, processing the transactions directly. Its main strength is not its speed (although it is not slow, even outperforms Apriori and Eclat on some data sets), but the simplicity of its structure. Basically all the work is done in one simple recursive function, which can be written with relatively few lines of code [7].

DOI 10.17148/IARJSET.2015.2625

V. CONCLUSION

Data mining has attracted a great deal of attention in the information and knowledge gained can be used for application ranging from business management, production control, marketing analysis to, engineering design and science exploration Association rule mining is the discovery of correlation among objects. Erasable itemset is an itemset that meets the maximum threshold requirement, it means that itemset value is less than threshold value.

This study demonstrates that mining knowledge is both practical and desirable .In the field of erasable itemset mining , most of the proposed method for generating candidate set adopt approaches similar to Apriori algorithm. VME an existing algorithm in mining erasable itemset uses apriori approach for generating itemset. The scope of the study of mining erasable itemset developed new method (MVME) that generate erasable itemset without checking of their subsets , It is an improved version of VME. This algorithm generate erasable itmeset as we do in VME, so an input file the results are same for both VME and MVME, but MVME take less time in executing the program. It takes input, generate candidate and calculate their gain and compare it with max threshold value and add into the list of erasable itemset. It works level by level. i.e for erasable 1-itemset then erasable 2-itemset then erasable 3-itemset and level goes on until there is no more candidate.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, *"Mining association rules between sets of items in large databases"*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC,1993

[2] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme and Lotfi Lakhal, *"Mining Frequent Patterns with Counting Inference"*. In proceeding of ACM SIGKDD, pp. 68-75, December 2000.

[3] G. Cormode, *"Fundamentals of Analyzing and Mining Data Streams"*, WORKSHOP ON DATA STREAM ANALYSIS, San Leucio, Italy, 2007.

[4] M.Kantardzic, "*Data Mining: Concepts, Models, Methods, and Algorithms"*, Wiley-Interscience, Hoboken, NJ, 2003.

[5] N.Rajkumar, M.R.Karthik, and S.N.Sivanandam, *"Fast Algorithm for Mining Multilevel Association Rules"*,IEEE, 2003.

[6] M.H.Margahny and A.A.Mitwaly, *"Fast Algorithm for Mining Association Rules",* Proceedings of AIML 05 Conference, CICC, Cairo, Egypt, 2005.

[7] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. *"Mining Data Streams: A Review"*, ACM SIGMOD Record Vol. 34, No. 2, Australia, 2005.

DOI 10.17148/IARJSET.2015.2625