Searching for the Best Strategies of Mining Erasable Itemsets

Ms Shweta, Dr. Kanwal Garg

Abstract: This paper discusses few approaches for mining erasable itemsets. In this paper, author decomposes the original problem into two smaller sub problems: First, Computing the gain of itemset and second is, Searching for erasable itemsets. The existing solutions based on horizontal data layout to this problem make repeated scans of database. Extensive studies proposed different strategies for efficient erasable itemset mining based on vertical format, horizontal format, threshold, top rank-k etc. It is the appropriate time to ask "what are the advantages and limitations of strategies (vertical layout vs. horizontal layout, threshold vs. top rank-k based)? And what and how can one pick and integrate the best strategies to get better performance in general cases". Author clear the above doubts by a systematic study of the erasable itemsets mining strategies. Finally, the paper concludes with the idea that how the integration of both vertical format based and top rank-k based approach can give better results.

Index Terms*:* Data mining, Frequent Pattern Mining (FPM), Database, Erasable Itemsets, META, VME.

1. INTRODUCTION

Data mining is the key process of "Knowledge Discovery Process". It is the process of searching the useful patterns from the large database [9].

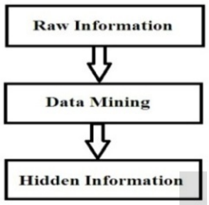

Fig. 1 Data Mining

Frequent pattern mining (FPM) is the important data mining concept which discovers frequent itemset in database. Frequent patterns are the patterns that appear in database frequently. Frequent pattern mining concept used in many fields such as association rule mining, clustering, pattern based classification, finding correlated item, erasable itemset mining etc [3]. In this paper author mainly focus on erasable itemset mining. Erasable itemset mining is helpful in production planning. Problem of Erasable itemset mining is the new data mining task first introduced by G.D. Fang and Z. Deng in [2] in 2009. In any manufacturing industry which manufacture some products. These products constitute of some components known as items. Sometimes a manufacturing factory due to financial crises may not be able to purchase all the components. Then the components that can be erased are known as erasable items. The original motivation for finding erasable itemset has been raised from the need to control the loss in profit due to absence of some component used for manufacturing products.

The organization of the paper is as follow: Section 2 gives literature review. Section 3 defines the erasable itemset mining problem. Section 4 and 5 will define sub problem 1(computation of gain) and 2(discovering erasable itemsets). Section 6 briefly describes the different erasable itemset mining algorithms. And finally author concludes in section 7 with a discussion of future work.

2. REVIEW OF LITERATURE:

In this section author has discussed some research papers which had been previously undertaken in the field of association rule mining, frequent pattern mining and erasable itemset mining.
Goswami D.N. and Chaturvedi Anshu have discussed the three different FPM approach i.e. record filter approach, intersection approach and combination of these two. Based on the concept of Apriori algorithm Zhi H. Deng, Guo D. Fang proposed a new algorithm META for mining erasable itemsets. Sanjeev rao and Priyanka gupta had discussed the disadvantages of Apriori algorithm. Jiawei Han, J. Wang and Y. Lu had elaborated the new mining task top-k frequent patterns and conclude that mining top-k frequent closed pattern without min-support is more preferable than the min-support based mining. Z. Deng and X. Xu had presented VME for erasable itemset mining. Z. Deng and X. Xu gave a new mining approach MIKE i.e. mining top rank-k erasable itemsets. In 2012 Miss Poushali Sarkar has introduced a mining task called Taking out High Valued/Utility erasable itemsets.

3. ERASABLE ITEMSET MINING PROBLEM:

Erasable itemset mining problem can be decomposed into two sub-problems:
1. First sub problem is further composed of two sub problems-one is to find the sum of value of all products, second is, to finding the gain of itemsets.
2. Second sub problem is to generate erasable itemset by comparing the gain of itemset and threshold. Itemsets whose gain is less than the user defined threshold are known as erasable itemsets.

- *Ms Shweta currently pursuing masters degree program in computer science and engineering from Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India, PH-09728387500. E-mail: shwetabidhan@gmail.com*
- *Dr. Kanwal Garg currently working as Assistant Professor in Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India. His area of expertise is Data Bases, Data Mining and Data Warehousing. E-mail: gargkanval@gmail.com*

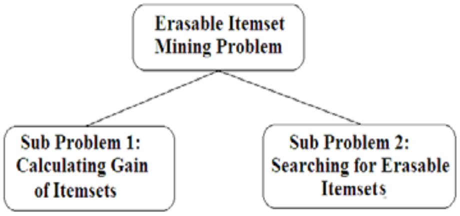
Erasable Itemset Mining Problem
Sub Problem 1: Calculating Gain of Itemsets
Sub Problem 2: Searching for Erasable Itemsets

Fig. 2 Erasable Itemset Mining Sub problems

Let I = {i_1, i_2, i_3, i_4…….i_m} is the set of m different literals known as items also called universal itemset. PDB = {P_1, P_2, P_3, P_4 …....Pn} is the product database over I. Means each product contains a set of items {i_1, i_2, i_3, i_4...i_j} ⊆ I. and have identifier PID. As shown in example P_4 constitute of i_1, i_2, i_4 and have a product identifier PID i.e. 4.

Table 1 Product Database

Product	PID	Items	Value
P_1	1	{i_2, i_3, i_4, i_6}	90
P_2	2	{i_2, i_5, i_7}	60
P_3	3	{i_1, i_2, i_3, i_5}	90
P_4	4	{i_1, i_2, i_4}	1200
P_5	5	{i_6, i_7}	70
P_6	6	{i_3, i_4}	90

3.1 How the Erasable Itemset Mining Different From Frequent Pattern Mining: First, Backgrounds of FPM and erasable itemset mining are different. Erasable itemsets mining play important role in planning the production of products in the manufacturing industry. However, FPM focuses on finding the collection of items in the retail trade. Second, Basic computation unit of erasable itemset is the values of itemsets. Whereas, FPM have counts. Third, constraints for mining of Erasable itemsets mining and FPM are different. In erasable itemsets mining, the constraint is that the value of an erasable itemset must be ≤ threshold. Whereas, the constraint in FPM is that the count of a frequent pattern must be ≥ threshold [4]. Thus erasable itemsets mining and FPM are different mining problems, but many concept used for FPM can applied to erasable itemset mining as explained in further sections.

3.2 Gain and Threshold: P_i (i ∈ [1…n]) is a type of product and is represented as: < PID, Items, Val >, *PID* is the unique identifier of P_i. *Items* are items or components that are used to form P_i. *Val* is gain or profit that a manufacturing industry gets by selling all P_i-type products [4]. Formal definition of Gain of an itemset is:

Definition 1: Let *S* (⊆ *I*) is an itemset (i.e. set of items), the gain of *S* is defined as:

$$\text{Gain (S)} = \sum_{\{P_k \mid S \cap P_k.\text{Items} \neq \emptyset\}} P_k.\text{Val} \qquad (1)$$

It means that the gain of itemset S is the sum of profits/value of all products that are made up components that constitute at least one item present in itemset S as their components [4]. For example, S (= {i_5, i_7}) be an itemset. Therefore according to the above equation (1) the gain of S is the sum of P_2.Val, P_3.Val and P_5.Val. Thus gain of S is 220(P_2.Val (60) + P_3.Val (90) + P_5.Val (70) =220).

3.3 Erasable Itemsets:

Definition 2: Given a predefined/ user defined threshold ξ and a product database *DB*, an itemset S is erasable if [7]

$$\text{Gain (S)} \le \sum_{\{P_k \in DB\}} (P_k.\text{Val}) \times \xi \qquad (2)$$

Thus, the problem of mining erasable itemsets is the problem of searching for the itemset whose gain is less then threshold %.

For example, suppose % threshold is ξ be 15%. Then threshold in terms of gain is:

$$\text{Net gain} = \sum_{\substack{1 \le k \le 7 \\ \{P_k \in DB\}}} (P_k.\text{Val})$$

$$= P_1.\text{Val} + P_2.\text{Val} + P_3.\text{Val} + P_4.\text{Val} + P_5.\text{Val} + P_6.\text{Val} (1000)$$

Now according to equation (2): Threshold = 1000 X 15%=150; Now find itemsets whose gain is less than or equal to 150. It can be said that itemset {i_6} is an erasable, as its net gain is 80(P_1.Val (50) +P_5.Val (30) = 80), which is less then threshold that is 150. Similarly you can obtain erasable 2-itemset and so on; by taking the union of gain of these 1-itemset. Erasable itemsets is helpful for manufacturer to decide how one can purchase raw material or help to select which components can be rejected used for manufacturing products in case of some economic problem.

4. SUB PROBLEM 1 - COMPUTATION OF GAIN OF ITEMSET:

The key point of erasable itemset mining problem is how one can compute the gain of an itemset. Computation of gain depends on the following two data formats:

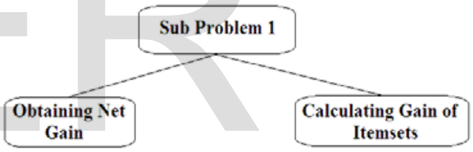

Fig.3 Sub Problem 1- Calculating Gain

4.1 Obtaining Gain Using Horizontal Data Format: For example, table 1 represents the data in horizontal format, for finding the gain of an itemset S{i_1}, repeatedly scan the product database check if the product includes that item as its component. Check P_1, it does not constitute i_1 as component then check P_2 and so on. Gain(S) = P_3.Value+ P_4.Value (90+1200 = 1290).

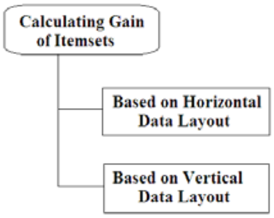
Calculating Gain of Itemsets
Based on Horizontal Data Layout
Based on Vertical Data Layout

Fig.4 Calculating Gain of Itemset

4.2 Obtaining Gain Using Vertical Data Format: MAFIA, CHARM, algorithm for frequent pattern mining use vertical representation of dataset and shows better performance over A-Close, Pascal algorithm that used horizontal data layout [8].

Table 2 Database (Vertical Data Format)

Items	Inverted List
i_1	<3, 90>, <4, 1200>
i_2	<1,90>,<2, 60>,<3, 90>,<4, 1200>
i_3	<1, 90>, <3, 90>, <6, 90>
i_4	<1, 90>, <4, 1200>, <6, 90>
i_5	<2, 60>, <3, 90>
i_6	<1, 90>, <5, 70>
i_7	<2, 60>, <5, 70>

In vertical data format gain of itemset is obtained by adding the second part of the "inverted list"(as shown in Table 2). For example gain of i_6 is: Here 1 and 5 is the product id. Gain (i_6) = sum of second part of (<1, 90>, <5, 70>) (90+70 = 160).

4.3 Vertical Data Format vs. Horizontal Data Format: (i) Gain of itemsets can be obtained very easily. For example gain of i_6 is calculated by summing the second part of <1, 90> and <5, 70> i.e. 160 (90+70). (ii) Vertical data format automatically trim irrelevant data [7]. For example $\{i_1\}$, if the gain is obtained by database scanning, it is compared with each product of database in Table 2. Comparing $\{i_1\}$ with P_1, P_2, P_5, P_6 and P_7 is nonsense because they do not take i_1 as their component. But by using inverted list of $\{i_1\}$, P_1, P_2, P_5, P_6 and P_7 are filtered automatically.

5. SUB PROBLEM 2- DISCOVERING ERASABLE ITEMSETS:

there are two different techniques for searching erasable itemsets.

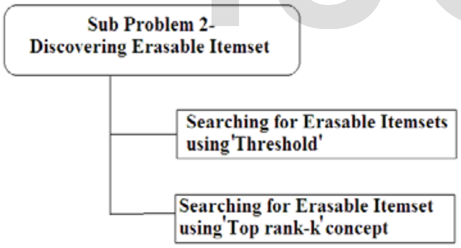
Sub Problem 2-
Discovering Erasable Itemset
Searching for Erasable Itemsets using'Threshold'
Searching for Erasable Itemset using'Top rank-k'concept

Fig.5 Sub Problem 2- Searching for Erasable Itemsets

6. ERASABLE ITEMSETS MINING ALGORITHMS:

In this section different algorithm related to frequent pattern mining and erasable itemset mining are briefly described:

6.1 Apriori Algorithm for Frequent Pattern Mining: This algorithm is proposed by R.Agrawal and R. Srikant [1] for frequent pattern mining. The key concept of the algorithm is to use the prior knowledge i.e. the knowledge it obtained by previous iterations [10]. This algorithm is suffered from some weaknesses: first, generation of large number of candidate itemsets. Second, number of database passes is equal to the maximum length of frequent itemset [11].

6.2 META: META is the abbreviation for Mining Erasable iTemsets using Anti-monotone property. It is most classical and important algorithm for erasable itemset mining. The idea of the META algorithm is taken from Apriori algorithm [Agrawal and Srikant, 1994]. It uses an iterative approach known as breadth first search (level-wise appoarch) through search space (Where k-itemsets are used to explore (k+1) itemsets). It uses Horizontal data format for finding gains [4].

6.3 VME: VME is the abbreviation for Vertical-format-based algorithm for Mining Erasable itemsets [7]. As in META, VME also uses, Anti-monotone property based on downward closure property states that "all non empty subsets of a erasable itemset must also be erasable and vice-versa [[21]]". This algorithm uses vertical data format.

Table 3 Database in horizontal format

PRODUCTS	PID	ITEMS	VALUE
P_1	1	i_1,i_3,i_5	1200
P_2	2	i_4,i_5,i_6	2200
P_3	3	i_2,i_3,i_9	1100
P_4	4	i_2,i_7	700
P_5	5	i_1,i_8,i_{10}	2500
P_6	6	i_7,i_8,i_9,i_{10}	2600
P_7	7	i_2,i_4,i_6	2200
P_8	8	i_6,i_8	2500

Vertical data representation is nothing but just the inverse of horizontal data representation.

Table 4 Vertical format of Table3

Items	INVERTED LIST
i_1	<1,1200>,<5,2500>
i_2	<3,1100>,<4,700>,<7,2200>
i_3	<1,1200>,<3,1100>
i_4	<2,2200>,<7,2200>
i_5	<1,1200>,<2,2200>
i_6	<2,2200>,<7,2200>,<8,2500>
i_7	<4,700>,<6,2600>
i_8	<5,2500>,<6,2600>,<8,2500>
i_9	<3,1100>,<6,2600>
i_{10}	<5,2500>,<6,2600>

Now, the gain of {*i*5} is the sum of the second part <1, 1200>, <2, 2200>, (1200+2200=3400).

The detailed idea of the *level-wise*-based approach is described as: First, the set of erasable 1-itemsets is found, denoted as *E*1. *E*1 is used to find *E*2, which is the set of erasable 2-itemsets, and then *E*2 is used to find *E*3, and so on, until no more erasable *k*-itemsets can be found.

For example Table4 and threshold value of 30 %. Scan the database to find the set of erasable 1-itemset and for that check the gain of each item and compare it with threshold value and if it less than threshold value than that itemset added in table 5.

Table 5 List of Erasable 1-itemset

i_1	<1,1200>,<5,2500>
i_2	<3,1100>,<4,700>,<7,2200>
i_3	<1,1200>,<3,1100>
i_4	<2,2200>,<7,2200>
i_5	<1,1200>,<2,2200>
i_7	<4,700>,<6,2600>
i_9	<3,1100>,<6,2600>

Table5, have all those items whose threshold value is less than 4500 because profit value is 15000 and threshold value is 30%. Now create candidates for level-2 and so on.

6.5 Mining Top-Rank-K Erasable Itemsets: Algorithms explained till now are threshold based to obtain erasable itemsets. In this section a new mining task called mining top-rank-k erasable itemsets is explained, where k is the biggest rank value of all erasable itemsets to be mined.

Definition 3: (The Rank of an itemset) [8]: Given product database DB and a pattern S, R_S, the rank of S, is defined by

$$\text{Rank } (S) = | \{\text{Gain}(X) \mid X \subseteq I \text{ and } \text{Gain}(X) \leq \text{Gain}(S)\}| \quad (3)$$

|Y| -is the numbers of elements in Y. Thus, given a transaction database DB and a threshold k, an itemset S is called to be a top-rank-k erasable itemset if and only if R_S is no longer that k. that is, $R_S \leq k$ [8]. Based on these definitions, mining top-rank-k erasable itemsets can be defined as, given a transaction database DB and a threshold k, the top-k erasable itemsets mining is to find the complete set of erasable itemsets whose ranks are less than k.

Suppose threshold k is 5, Table 6 shows all top-rank-5 erasable itemsets [8].

Table 6 All top-rank-5 erasable itemsets

PID	Items	Value
1	{ i_3}	2300
2	{ i_7}	3300
3	{ i_5}	3400
4	{ i_1}, { i_9}	3700
5	{ i_2}	4000

6.6 MIKE (An Efficient Algorithm for Mining Top-Rank-k Erasable Itemsets):

This algorithm is based on mining high-value k or top rank-k erasable itemsets [14]. It does not require setting of threshold. In this during the mining process undesired itemsets are trimmed and erasable itemsets are selected to generate other longer hopeful itemsets.this approach reduces the search space [8].

Threshold vs. Top rank-k based: There are problems in setting a percent threshold: setting threshold is quite difficult, if threshold is too small, then this d may lead to generation of thousands of itemsets. Whereas if threshold is too big, then it may leads to generate no answer i.e. no erasable itemsets. Another advantage of top rank-k based approach over threshold based approach is that, in top rank-k based has small search space then threshold based. Search space can be reduced by using top rank-k based approach [12].

7. CNCLUSION:

In this paper, author gave formal definition erasable itemsets mining. Data mining is called Erasable Itemsets Mining, when applied to the finding the erasable itemsets in production planning of any manufacturing industry. Author divides the erasable itemset mining problem into two sub problems. Calculating gain of itemset (horizontal vs. vertical layout) and discovering of erasable itemsets (Threshold vs. top rank-k). Thus the aim of the work is to find the techniques which search erasable itemsets in reasonable time. Author concludes that algorithm using vertical layout are better than based on horizontal data layout. And also finds that top rank- k based approach more efficient than threshold based approach. This work will have tremendous significance for the application of Erasable Itemsets mining techniques in production planning. **Future Work:** vertical data layout approach is better that horizontal layout approach in terms of time efficiency. And the concept based on taking out high utility erasable itemsets is performed over threshold based erasable itemset mining. For future work, adopting the idea of integration of these approaches can give more efficient concept for erasable itemsets mining. Researcher can also extend these algorithms to deal with the case of mining erasable itemsets from distributed databases. Erasable itemsets mining can be extended by using the ideas from studies for mining maximal frequent, closed frequent patterns and top-k frequent patterns in many recent research papers.

REFERENCES:

[1] Han, J., Pei, and Yin: “Mining frequent patterns without candidate generation”. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 1–12. ACM Press, New York (2000).

[2] en.wikipedia.org/wiki/association_rule_lear
ning.

[3] Zhi-Hong Deng, Xiao-ran Xu, “Fast
Mining erasable itemsets using NC_sets”,key lab of Machin perception, SEECS, Peking University, Beijing, china.

[4] Deng, Z., Fang, G., Wang, Z., Xu, X.: “Mining Erasable Itemsets”. In: 8th IEEE International Conference on Machine Learning and Cybernetics, pp. 67–73. IEEE Press, New York (2009).

[5] Bernecker, T., Kriegel, H., Renz et.al. “Probabilistic Frequent Itemset Mining in Uncertain Databases”. In: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 119–127. ACM Press, New York (2009).

[6] www2.cs.uregina.ca/~dbd/cs831/notes/
kdd/1_kdd.html

[7] Zhihong Deng and Xiaoran Xu, “An efficient Approach for Mining Erasable Itemsets”, in ADMA 2010, Part I, LNCS 6440, pp. 214–225, 2010.

[8] Zhihong Deng and Xiaoran Xu, China, “Mining Top-rank-K Erasable Itemsets”, in ICIC Express Letters, ICIC International @ 2011 ISSN 1881-803X, Volume 5, Number 1, January 2011.

[9] Rakesh Agrawal, T. Imieliński, A. Swami, "Mining association rules between sets of items in large databases". In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. pp. 207-216.

[10] Sanjeev Rao, Prianka Gupta, “Implimenting Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm”, In: preceending of IJCST, VOL.3, Issue 1, Jan-March 2012.

[11] Goswami D.N., Chaturvedi Anshu, and Raghuvanshi, “An Algorithm for Frequent Pattern Mining Based On Apriori”, In: (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947.

[12] Jiawei Han, Jianyong Wang, Ying Lu, Petre Tzvetkov, “Mining Top-K Frequent Closed Patterns without Minimum Support”, In: Proceedings of the 2002 IEEE International Conference on Data Mining**.**

[13] A.M.J. Md. Zubair Rahman and P. Balasubramanie, “An Efficient Algorithm for Mining Maximal Frequent ItemSets”, In: proceeding of Journal of Computer Science: 638-645, 2008.

[14] Miss Poushali Sarkar, “Taking out High Utility Erasable Itemsets”, In International Journal of Computer science an Application, Issue 2, Volume 3(June 2012).