The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Reduction of Erasable Itemset Mining to Frequent Itemset Mining

Tzung-Pei Hong[*1,2], Chun-Ho Wang[*2], Chia-Che Li[*2], Wen-Yang Lin[*1]

[*1] National University of Kaohsiung, Kaohsiung, Taiwan

[*2] National Sun Yat-sen University, Kaohsiung, Taiwan

Frequent-itemset mining and erasable-itemset mining are two commonly seen and useful techniques in data mining. Although the two mining problems look contrary, they are actually close to each other. In this paper, we will show the erasable itemset mining problem can be reduced into the frequent-itemset mining problem and can be solved by the existing algorithms of finding frequent itemsets. By this way, the variants of erasable-itemset mining can be easily designed out based on the frequent itemset mining.

1. Introduction

The frequent-itemset mining is a very important step in the search for association rules [1][2][10]. It is mainly determined based on the occurrence frequency of the itemsets in the transactions. The erasable itemset mining problem, proposed by Deng et al in 2009 [5], aimed for the purpose of analyzing the factory production plan [5]. The problem assumes that a factory has some products to produce, and each product requires some kinds of raw materials and can gains some profits. The erasable itemset mining can be used to find out which combination of raw materials will not reduce the profit of the factory too much if they are removed. Although the frequent itemset mining and the erasable itemset mining seem to be relatively different, we will show in this paper that both of them are similar in nature and thus we can reduce the erasable itemset mining problem to the frequent itemset mining problem. The solutions for the frequent itemset mining problem can then be transformed back to form the results for the erasable itemset mining problem. We hope that based on this reduction, the transformation process can provide a novel and interesting design strategy for data analysis methods. The rest of this paper is organized as follows. Some related work is briefly reviewed in Section 2. The reduction process is proposed in Section 3. The correctness of the reduction is shown in Section 4. Conclusion is given in Section 5.

2. Related Work

2.1 Frequent Itemset Mining

Finding frequent itemsets and association rules from a transaction database is a very important research topic in data exploration nowadays. Among the existing algorithms, most of which were based on the Apriori algorithm [1][2] and the FP-tree algorithm [8]. The former approach generates and tests candidate itemsets level-by–level. It is composed of two parts. The first part is to find out the itemsets with large counts, called frequent itemsets or large itemsets; the second part utilizes the conditional probability to find the association rules from the large itemset. If the conditional probability of a possible rule is greater than or equal to the given minimum confidence, it is desired. An incremental mining approach is also proposed [3]

On the other hand, the FP-tree approach uses a tree structure called the frequent-pattern-tree (FP-tree) for efficiently mining association rules without generation of candidate itemsets [8]. The FP-tree can be thought of a compression structure for the database and stores only large items. It is also composed of two parts. The first part is to construct the FP-tree from the database by processing the transactions one by one; the second part utilizes a recursive mining procedure called FP-Growth to derive frequent patterns from the FP-tree.

2.2 Erasable Itemset Mining

Erasable itemset mining was proposed by Deng et al. in 2009 for analyzing production planning [5]. There are several algorithms for erasable itemset mining, such as META[12], MERIT [6], MERIT+ [4], MEI [11], and VME [7]. Some variants are also proposed [9].

Formally, let I is a set of m items. A product dataset, DB, contains a set of n manufactured products. Each product is represented by a subset of I and its profit. The gain of an itemset $X \subseteq I$ is the sum of the profits of the products which include at least one item in X. A set X is called an erasable itemset if its gain is equal to or less than a given threshold of the total profits of all products.

3. Reducing Erasable Itemset Mining to Frequent Itemset Mining

In this section, we reduce erasable itemset mining to frequent itemset mining. The transformation procedure we proposed is described with some examples interweaved in order to express the concept more easily.

3.1 The transformation procedure

In this subsection, we describe the transformation procedure for reducing the erasable itemset mining to frequent itemset mining. The procedure is described as follows.

STEP 1: Transform the product database into a corresponding transaction database.

STEP 2: Transform the threshold for erasable itemset mining into the one for frequent itemset mining.

STEP 3: Find the solutions of frequent itemsets.

Contact: Tzung-Pei Hong, National University of Kaohsiung, Taiwan, tphong@nuk.edu.tw

STEP 4: Transform the frequent itemsets into the erasable itemsets.

3.2 An example

Assume Table 1 is a product database with three fields.

Table 1: An example of a product database *P*

P		
PID	Items	Value
P_1	*CE*	150
P_2	*BCD*	150
P_3	*CD*	50
P_4	*ADE*	150

The items in the transaction database are just the complement of the items contained in the product. Taking P_1 as an example, since P_1.*Value* is 150, there will be 150 identical transactions, T_1 to T_{150}, generated and put into the transaction database *T*. Since the items in P_1 are *C* and *E*, and the set of all the items in the product database are {*A*, *B*, *C*, *D*, *E*}, each derived transaction thus includes *A*, *B* and *D*. The final derived transaction database is shown in Table 2.

Table 2: The derived transaction database *T*

T							
TID	Items	TID	Items	TID	Items	TID	Items
T_1	*ABD*	T_{151}	*AE*	T_{301}	*ABE*	T_{351}	*BC*
T_2	*ABD*	T_{152}	*AE*	T_{302}	*ABE*	T_{352}	*BC*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
T_{150}	*ABD*	T_{300}	*AE*	T_{350}	*ABE*	T_{500}	*BC*

Next in Step 2, the minimum threshold ratio r^T for frequent itemset mining is set. Assume in this example, the maximum threshold ratio, r^P, in erasable itemset mining is set at 0.6, then the minimum threshold ratio r^T is set as $1 - r^P$ (= 0.4).

In Step 3, any frequent itemset mining algorithm such as Apriori [2] or FP-Tree [8] can be used to mine all frequent itemsets from the derived transaction database. The mining results from the transaction database in Table 2 are shown in Table 3.

Table 3: Erasable itemsets mined from Table 2

Erasable itemsets	
Itemsets	gain
A	150
B	150
E	300
AB	300
AE	300

Finally in Step 4, the frequent itemsets are thought of as the erasable itemsets for the original product database.

4. The Correctness of the Reduction

In the section, we give a theorem to prove the correctness of the reduction process. That is, we will show the erasable itemsets from both the original erasable mining and from transformed frequent itemset mining are the same.

Theorem: The erasable itemsets from the original erasable mining are the same as the frequent itemsets obtained by the frequent itemset mining after the transformation procedure.

For proving the itemsets obtained by the two mining strategies (directly erasable itemset mining and transformed frequent itemset mining) are the same, the following two cases need to be considered.

(1) If an itemset is erasable, then it is a frequent itemset by the transformed frequent itemset mining.

(2) If an itemset is not erasable, then it is not a frequent itemset by the transformed frequent itemset mining.

Both of them can be proven from the definitions of the two mining problems based on the transformation process. The formal proof is skipped here. Below, we show the results of the product database in Table 1 directly mined by the META algorithm [5], which is commonly used in the erasable itemset mining. The results are shown in Table 4.

Table 4: The erasable itemsets directed mined by META

Erasable itemsets
A
B
E
AB
AE

Comparing the above results in Tables 3 and 4, we may find that the two results are identical, which indirectly show the correctness of the proposed procedure.

5. Conclusion

The erasable itemset mining looks for itemsets whose gain values are below a user-specified threshold, so that we can sacrifice the items in the erasable itemsets to plan the manufacture of products with limited money. While the frequent itemset mining is used to mine itemsets whose frequency is above a user-specified threshold, so that we can see that these items appear frequently. Although the frequent itemset mining and the erasable itemset mining seem to be relatively different, they have been shown that the two problems are similar in nature and can be designed in an interchangeable way. Note that we do not mean to actually solve the erasable itemset mining problem by the frequent itemset mining algorithms in real applications, but provide an effective design way to find out good algorithms to solve erasable itemset mining and its variants.

References

[1] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases”, The ACM SIGMOD International Conference on Management of Data, pp. 207–216, 1993.

[2] R. Agrawal and R. Srikant, “Fast algorithm for mining association rules”, The 20th International Conference on Very Large Data Bases, pp. 487–499, 1994.

[3] D. W. Cheung, J. Han, V. T. Ng and C. Y. Wong, “Maintenance of discovered association rules in large databases: Agrawal approach”, The 12th IEEE International Conference on Data Engineering, pp. 106–114, 1996.

[4] T. Le, F. Coenen and B. Vo, “An efficient algorithm for mining erasable itemsets using the difference of NC-Sets”, The IEEE International Conference on Systems, Man, and Cybernetics Manchester, pp. 2270–2274, 2013.

[5] Z. H. Deng, G. D. Fang, Z. H. Wang and X. R. Xu, “Mining erasable itemsets”, The 8th International Conference on Machine Learning and Cybernetics, pp. 12–15, 2009.

[6] Z. H. Deng and X. R. Xu, “Fast mining erasable itemsets using NC_sets,” Expert Systems with Applications, Vol. 39, pp. 4453–4463, 2012.

[7] Z. H. Deng, “Mining top-rank-k erasable itemsets by PID_lists”, International Journal of Intelligent Systems, Vol. 28, Issue 4, pp. 366–379, 2013.

[8] J. Han, R. Mao, J. Pei and Y. Yin, “Mining frequent patterns without candidate generation: a frequent-pattern tree approach”, Data Mining and Knowledge Discovery, Vol. 8, Issue 1, pp. 53–87, 2014.

[9] B. Vo, T. P. Hong and B. Le, “A dynamic bit-vector approach for fast mining frequent closed itemsets,” Expert Systems with Applications, Vol. 39, Issue 8, pp. 7196–7206, 2012.

[10] W. Li, M. Ogihara, S. Parthasarathy and J. Zaki, “New algorithms for fast discovery of association rules”, The 3th International Conference on Knowledge Discovery and Data Mining, 1997.

[11] T. Le and B. Vo, “An efficient algorithm for mining erasable itemsets”, Engineering Applications of Artificial Intelligence, Vol. 27, pp. 155–166, 2014.

[12] T. Le, G. Nguyen and B. Vo, “A survey of erasable itemset mining algorithms”, Wires Data Mining Knowledge Discovery, pp. 356–379, 2014.