DMA:
Matrix Based Dynamic Itemset Mining Algorithm

Damla Oguz, Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey & Department of Computer Engineering, Ege University, Izmir, Turkey

Baris Yildiz, Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey & Department of Computer Engineering, Dokuz Eylul University, Izmir, Turkey

Belgin Ergenc, Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey

ABSTRACT

Updates on an operational database bring forth the challenge of keeping the frequent itemsets up-to-date without re-running the itemset mining algorithms. Studies on dynamic itemset mining, which is the solution to such an update problem, have to address some challenges as handling i) updates without re-running the base algorithm, ii) changes in the support threshold, iii) new items and iv) additions/deletions in updates. The study in this paper is the extension of the Incremental Matrix Apriori Algorithm which proposes solutions to the first three challenges besides inheriting the advantages of the base algorithm which works without candidate generation. In the authors' current work, the authors have improved a former algorithm as to handle updates that are composed of additions and deletions. The authors have also carried out a detailed performance evaluation study on a real and two benchmark datasets.

Keywords: Algorithms, Dynamic Itemset Mining, Itemset Mining, Matrix Apriori, Operational Database

1. INTRODUCTION

Association rule mining discovers interesting relations among sets of items in databases. It is composed of two steps: finding all frequent itemsets and generating association rules from the itemsets discovered. The number of occurrences of an itemset is called its support count. An itemset becomes frequent when its support count exceeds a predefined threshold. Finding frequent itemsets in a given dataset is non-trivial because datasets can be very large and may contain many items. On the other hand, the second step of the association rule mining is straightforward. Therefore, the general performance of any association rule mining algorithm is determined by the first step (Han & Kamber, 2005).

Apriori and FP-Growth are known to be the two important association rule mining algorithms each having a different approach to find frequent itemsets (Agrawal & Srikant, 1994; Han, Pei, & Yin, 2000). The Apriori Algorithm uses Apriori Property in order to improve the

DOI: 10.4018/ijdwm.2013100104

efficiency of the level-wise generation of frequent itemsets. On the other hand, candidate itemsets generation and multiple database scans are the drawbacks of the algorithm. FP-Growth creates signatures of transactions on a tree structure to eliminate the database scans and outperforms Apriori (Han et al., 2000). A recent algorithm called Matrix Apriori, which combines the advantages of Apriori and FP-Growth, was proposed by Pavón, Paulo and Viana (2006). The algorithm eliminates multiple database scans by creating signatures of itemsets in the form of a matrix. Yildiz and Ergenc (2010) showed that Matrix Apriori provides a better overall performance than FP-Growth for the specified datasets and decreasing minimum support values.

Although all these algorithms handle the problem of association rule mining, they ignore the dynamic nature of the databases. When new transactions arrive, the entire process needs to be done from the beginning. The solution to this problem is dynamic itemset mining which proposes the idea of keeping frequent itemsets up-to-date when the database is updated. Dynamic itemset mining has four challenges: i) handling database updates without re-running the frequent itemset mining algorithms, ii) allowing new item appearances in updates, iii) being flexible to support changes during entire process and iv) handling deletions as well as additions in updates.

Dynamic itemset mining algorithms can be categorized in four groups. The first group is Apriori based algorithms (Cheung, Han, Vincent, & Wong, 1996; Cheung, Lee, & Kao, 1997). The main goal of these algorithms is to reduce the number of candidate sets and the need of scanning the original database when new transactions arrive. The second group can be dedicated to FP-Growth based algorithms (Cheung & Zaïane, 2003; Hong, Lin, & Wu, 2008; Adnan, Alhajj, & Barker, 2008; Li & Li, 2010; Pradeepini & Jyothi, 2010). These algorithms try to keep every itemset of the original database in a tree and modify the tree with each update. The third group is Border based algorithms (Aumann, Feldman, Lipshtat, & Manilla, 1999; Taha, Gharib, & Nassar, 2011) where the idea is to keep track of potential itemsets which can be frequent at anytime. The last group of the dynamic itemset mining algorithms use different data structures as tries (Woon, Ng, & Das, 2001) or matrices (Oguz & Ergenc, 2012) to keep the signatures of the transactions in the original database and modify them when the new updates arrive.

In this paper, we focus on improving our previous work (Oguz & Ergenc, 2012) in which the Incremental Matrix Apriori (IMA) Algorithm was presented. IMA is capable of providing solutions to the first three challenges of dynamic itemset mining algorithms; keeping frequent itemsets up-to-date without scanning the original dataset, being flexible to support changes and appearance of new items in the updates. However, it does not cover the last challenge which is handling deletions. Therefore, the enhancement of this paper can be summarized as;

1. IMA is modified as to handle deletions in the updates. This new version is named as Dynamic Matrix Apriori (DMA).
2. A detailed performance study on a real retail and the two benchmark datasets is presented with different update scenarios that include addition and deletion of transactions in order to find out the intervals of update sizes where DMA is advantageous.

This paper is organized as follows; Section 2 reviews dynamic itemset mining algorithms. Section 3 presents the proposed algorithm DMA for dynamic itemset mining. In Section 4, test results and performance evaluations are discussed. Finally, Section 5 covers our final remarks related with this study and the direction for future work.

2. RELATED WORK

The analysis of transactional databases is taken into consideration with the growth in the amount of data. Consequently, a new data mining technique "association rule mining" was proposed by Agrawal, Imielinski and Swami (1993).