Abstract

Association rule mining is one of the most widely used data mining techniques. To achieve a better performance, many efficient algorithms have been proposed. Despite these efforts, many of these algorithms require a large amount of main memory to enumerate all frequent itemsets, especially when the dataset is large or the user-specified support is low. Thus, it becomes apparent that we need to have an efficient main memory handling technique, which allows association rule mining algorithms to handle larger datasets in the main memory. To achieve this goal, in this chapter we propose an algorithm for vertical association rule mining that compresses a vertical dataset in an efficient manner, using bit vectors. Our performance evaluations show that the compression ratio attained by our proposed technique is better than those of the other well-known techniques.
Introduction

One of the most widely used data mining techniques is association rule mining. Association rule mining algorithms iterate a dataset many times to enumerate frequent itemsets that exist in the transactions of a given dataset. However, a dataset scan is considered as an I/O exhaustive process (Zaki, 1999). A single scan can take a significant amount of time when it is large. Therefore, the performance degrades if the mining algorithm requires multiple dataset scans.

Many algorithms have been proposed to reduce the cost of dataset scan (Zaki, 1999; Zaki, 2003; El-Hajj & Zaiane, 2003; Han, Pei, & Yin, 2000; Doug, Manuel, & Johannes, 2001; Shenoy et al., 2000). These algorithms use various techniques, such as compression (Han et al., 2000), intersection (Zaki, 1999, 2003; Doug et al., 2000; Shenoy et al., 2000), and indexing (El-Hajj & Zaiane, 2003), to name a few. In spite of these efficient techniques, they still incur a common problem: all of these algorithms are inherently dependent on the amount of main memory (Geothals, 2003). Since the size of the main memory is limited, when the dataset is large and/or the support is low, the main memory size required by these algorithms never seems to be enough. Consequently, it is not surprising that these algorithms are unable to finish the mining task, or have to downgrade the performance significantly (Geothals, 2003).

Since main memory plays a significant role in association rule mining performance, in recent years several novel techniques have been proposed (Zaki, 2003; El-Hajj & Zaiane, 2003; Han et al., 2000; Doug et al., 2000; Shenoy et al., 2000) in order to efficiently use main memory. These techniques generally cut down the dataset size, so that the mining algorithms will be able to finish the mining task on bigger datasets, or with a low support.

We are motivated by the abovementioned fact that main memory is an important resource and, to improve performance, we need to use it in an efficient way without exceeding its capacity. To enhance main memory capacity, in this chapter we propose an algorithm that uses a bit-oriented approach to compress vertical tid dataset. The proposed technique keeps track of the differences between two tids and converts the differences into a bit format; finally, it stores these bits into a bit vector in an efficient way, so the resultant bit vector has only a few unused bits.

The important outcome of this method is that the proposed technique is not biased, which means that it does not depend on a particular dataset characteristic (i.e., dense or sparse) or the user-specified support. Rather, it has the ability to compress the original dataset regardless of dataset size, type, or user-specified support. Our performance evaluation also shows that it achieves a good compression ratio in all scenarios. Therefore, it is able to keep large datasets and allows the mining algorithms to perform mining tasks on such datasets.
A Mathematical Database to Process Time Series
Cyrille Ponchateau, Ladjel Bellatreche, Carlos Ordonez and Mickael Baron (2018).
*International Journal of Data Warehousing and Mining* (pp. 1-21).
www.igi-global.com/article/a-mathematical-database-to-process-time-series/208690?camid=4v1a

Clinical Data Mining in the Age of Evidence-Based Practice: Recent Exemplars and Future Challenges
www.igi-global.com/chapter/clinical-data-mining-age-evidence/44295?camid=4v1a