MISQ Archivist

Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing

Syam Menon and Sumit Sarkar

Abstract

Scalability and privacy form two critical dimensions that will eventually determine the extent of the success of big data analytics. We present scalable approaches to address privacy concerns when sharing transactional databases. Although the benefits of sharing are well documented and the number of firms sharing transactional data has increased over the years, the rate at which this number has grown is not quite what it could have been; concerns about revealing proprietary information have prevented some retailers from sharing, despite the obvious advantages in an increasingly networked economy. In the context of sharing transactional data, sensitive information is typically based on relationships derived from frequently occurring itemsets—a result of surprisingly successful promotions by the retailer, or unexpected relationships identified by the retailer while mining the data. Prior work in this area includes optimal approaches based on integer programming to maximize the accuracy of shared databases, while hiding all sensitive itemsets. While these approaches were shown to solve problems involving up to 10 million transactions, many transactional databases in the big data context are considerably larger and the existing integer programming-based procedures do not scale well enough to solve these larger problems. Consequently, there is no effective solution procedure for such databases in extant literature.

In this paper, we first present an optimal procedure leveraging intuition from linear programming based *column generation*. Next, we identify a common structure that exists in these problems, and show how it can be taken advantage of through an approach based on sorting and column generation to make the process more efficient. We then illustrate how this structure can be incorporated into the column generation based procedure to develop an effective, scalable heuristic. Computational experiments are conducted on databases with 50 and 100 million transactions, involving problems that could not be solved using existing optimal procedures. These experiments show that the optimal column generation based procedure can solve problem instances significantly larger than those tackled previously, and that the scalable heuristic identifies near-optimal solutions quickly in all instances where the optimal solution is known. We investigate the impact of hiding sensitive itemsets on the quality of a rule-based recommender system derived from the shared data. As expected, recommendation quality decreases as the number of sensitive itemsets increases; however, recommendation accuracy stays above 80% of the original rate when using the unmodified data even when there are 1,000 sensitive itemsets to hide. The effect on recommendation accuracy from using the heuristic relative to the optimal approach was very small: the accuracies with the heuristic were over 97% of the corresponding accuracies with the optimal approach in every experiment, and over 99% in the vast majority.

Keywords: Data quality, optimization, column generation, itemset hiding, recommendation quality