Chapter 3.12
Rapid Privacy Preserving Algorithm for Large Databases

K. Anbumani
Karunya Institute of Technology and Sciences (Deemed University), India

R. Nedunchezhian
Sri Ramakrishna Engineering College, India

ABSTRACT

Data mining techniques have been widely used for extracting non-trivial information from massive amounts of data. They help in strategic decision-making as well as many more applications. However, data mining also has a few demerits apart from its usefulness. Sensitive information contained in the database may be brought out by the data mining tools. Different approaches are being utilized to hide the sensitive information. The proposed work in this article applies a novel method to access the generating transactions with minimum effort from the transactional database. It helps in reducing the time complexity of any hiding algorithm. The theoretical and empirical analysis of the algorithm shows that hiding of data using this proposed work performs association rule hiding quicker than other algorithms.

INTRODUCTION

Data Mining (DM) techniques have been widely used in many areas especially for strategic decision-making (Agarwal, Imielinski, & Swami, 1993; Agarwal and Srikant, 1994; Fayyad, 1996; Han & Kamber, 2001; Han, Pei, & Yin, 2000; Hidber, 1999; Lin & Kedam, 2002; Webb, 2000). Apart from its usual benefits, it also has a few disadvantages associated with it. Experts say that data mining in the wrong hands will end up in destruction. The main threat of data mining is to privacy and security of data residing in large data stores (Agrawal & Srikant, 2000; Agrawal & Aggarwal, 2001; Ashrafi, Taniar, & Smith, 2005; Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios, 1999; Clifton, 1999; Lee, Chang, & Chen, 2004; O'Leary, 1991). Some of the information considered as private and secret can be brought out with advanced data mining tools. It is a real concern of people working in the field of database technology. Different research efforts are under way to address this problem of preserving security and privacy.

Sensitive information contained in a database can be extracted with the help of non-sensitive information. This is called the inference problem (Clifton & Marks, 1996; Marks, 1996; Verykios, Elmagarmid, Bertino, Saygin, & Dasseni, 2003). Different concepts have been proposed to handle the inference problem. The process of modifying the transactional database to hide some sensitive information is called sanitization. By sanitizing the original transactional database, the sensitive information can be hidden. In the sanitization process, selective transactions are retrieved and modified before handing over the database to a third party.

Modification of transaction involves removing an item from a transaction or adding an element to the transaction. In some cases, transactions will be either added to or removed from the database as suggested in (Clifton & Marks, 1996). The modified database is called sanitized database or released database.

A number of approaches have been proposed to hide sensitive data with good accuracy. The efficiency of a privacy-preserving algorithm is measured based on: (1) the time taken to hide the data, (2) the number of new rules introduced because of the hiding process, and (3) the number of legitimate rules lost or unable to be extracted from the released database.

The task of locating a transaction for sanitization from a massive amount of data is not a trivial process and it will certainly be a time consuming one. In many research efforts, the highly time consuming process of retrieving the transactional database is not taken into account efficiently. This article proposes a method to hide sensitive association rule in a fast manner. It takes the advantages of Frequent Pattern Growth Tree to identify and retrieve the generating transactions directly from the transactional database without exhaustive search. An array is used to keep track of the identifiers of the required transactions for sanitization. The article is organized as follows: first we discuss the existing related works. The proposed approach is discussed in the next section. The performance of the algorithms is discussed in subsequent section and the final section concludes the article.

RELATED WORK

Typically, the hiding process involves two steps: (a) generation of association rules and (b) hiding of association rules. Association rule mining is one of the functionalities of data mining. The process of producing association rules consists of: (1) the frequent itemsets generation and (2) rule generation. Frequent itemsets generation is a tedious process because it performs the time consuming task of the generation of candidates and pruning of unnecessary itemsets.

A number of *Apriori* derivative algorithms (Agarwal et al., 1993; Han & Kamber, 2001; Hidber, 1999; Lin & Kedam, 2002; Nedunchezhian & Anbumani, 2004; Webb, 2000) are available to improve efficiency of association rule mining. The algorithms for mining association rule differ in the approaches they use. Bottom-up (Agarwal & Srikant 1994; Agarwal et al., 1993), top-down, or a combination of both approaches (Lin & Kedam, 2002) is used for generating frequent itemsets. The approaches differ in terms of how the transactions of database are scanned. Few algorithms generate frequent itemsets without the costly candidate generation. The frequent pattern growth tree (FPT) (Han et al., 2000, 2001) generates the frequent itemsets without candidate generation. The tree is constructed based on the occurrence of the frequent items. Each transaction in the database is reordered before adding the items of the transaction to the FPT. It proposes a novel frequent pattern structure. The