Chapter 6

Secure Two-Party Association Rule Mining Based on One-Pass FP-Tree

Golam Kaosar
Victoria University, Australia

Xun Yi
Victoria University, Australia

ABSTRACT

Frequent Path tree (FP-tree) is a popular method to compute association rules and is faster than Apriori-based solutions in some cases. Association rule mining using FP-tree method cannot ensure entire privacy since frequency of the itemsets are required to share among participants at the first stage. Moreover, FP-tree method requires two scans of database transactions which may not be the best solution if the database is very large or the database server does not allow multiple scans. In addition, one-pass FP-tree can accommodate continuous or periodically changing databases without restarting the process as opposed to a regular FP-tree based solution. In this paper, the authors propose a one-pass FP-tree method to perform association rule mining without compromising any data privacy among two parties. A fully homomorphic encryption system over integer numbers is applied to ensure secure computation among two data sites without disclosing any number belongs to themselves.

1 INTRODUCTION

Data mining, often referred as the major part in knowledge discovery in database (KDD) is the process of discovering knowledge for decision making in business by utilizing patterns or models existed in data. Data mining has been one of the hot research areas for at least last two decades. Data mining process is a challenging work because of many reasons, such as various kind of knowledge is required for different databases and administrations, in many cases data sources are distributed, format of the data is diversified (e.g., text, audio, video, image, etc.), privacy preservation of the data, efficiency of the mining process, presentation of the outcome, etc. Some
of the applications of data mining includes but not limited to: advertising, bioinformatics, customer relationship management (CRM), database marketing, fraud detection, e-commerce, health care services, manufacturing, process control, sports and entertainment, telecommunications, web applications, etc.

Association rule mining (ARM) first introduced in 1966 (Hajek, Havel, & Chytil, 1966) is one of the straightforward data mining process which can be further computed in various methods such as Apriori (Agrawal & Srikant, 1994), FP-tree (Han, Pei, & Yin, 2000), Eclat (Zaki, 2000), One-attribute-rule, etc. FP-tree is a method to compute association rule mining (ARM) which was first proposed by Han et al. (2000) and Han, Pei, Yin, and Mao (2004). FP-tree based solution is advantageous than Apriori based solution in some cases due to following reasons:

- Apriori technique requires multiple scanning of database in each iteration. FP-tree requires two scans only. One scan generates frequent items and second scan generates a FP-tree from which frequent itemsets can be generated.
- Apriori technique generates enormous number of candidate itemsets in each iteration which require lot of computation. FP-tree does not create such candidate itemsets.

Ensuring data privacy is a compulsory requirement before integrating data to be mined. Definition of privacy itself varies from context to context. In the age of information technology the definition of privacy may be referred as the right of concealing one’s information and have control over the information in some extent (Ackerman, Cranor, & Reagle, 1999; Cockcroft & Clutterbuck, 2001). Another definition for information privacy (Victorian Government, Department of Human Service, 1999) is “Information privacy refers to a group of related rights regarding an individual’s control over the collection, use, release, and disposal of their personal information. Personal information is information which allows an individual to be identified, and it can appear in any form (for example, sound, image, text, biological-based) and be recorded in any medium (for example, print or electronic)”. Most countries have established privacy acts to preserve data privacy. This privacy can be achieved in basic two ways - statistical method and cryptographic method. Cryptographic method is capable to maximize both privacy and accuracy by the cost of high mathematical computation.

Homomorphic encryption system is most successfully used in preserving privacy in ARM algorithm. Original homomorphic cryptosystems, such as RSA (Rivest, Shamir, & Adleman, 1978), El Gamal (1985), Benaloh (Clarkson, 1994), and Paillier (1999), fully homomorphic encryption for integers (Dijk, Gentry, Halevi, & Vaikuntanathan, 2010), etc., are used in various privacy preserving ARM algorithms. A homomorphic encryption based secure two-party computation technique is used in preserving privacy in ARM in Ouyang and Huang (2006). Homomorphic encryption based association rule mining algorithm is also proposed in Kantarcioglu and Clifton (2004) which minimizes the information sharing with minimum mining overhead. Paillier cryptosystem based homomorphic encryption system and semi-trusted mixer based privacy preserving distributed association rule mining algorithm is proposed in Yi and Zhang (2007).

There are various other techniques in preserving privacy in ARM algorithms too. Luo, Zhao, and Le (2009) in survey on privacy preserving ARM algorithm has divided privacy preserving techniques into three classes - heuristic-based, reconstruction-based, and cryptography-based. Data perturbation techniques (Evmtievski, Gehrke, & Srikant, 2003; Rizvi & Haritsam, 2002) which, basically propose a technique to estimate associa-