Expert Systems with Applications 42 (2015) 1149–1165

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A fast perturbation algorithm using tree structure for privacy preserving utility mining

Unil Yun [*], Jiwon Kim

Department of Computer Engineering, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Available online 9 September 2014

Keywords:
Privacy preserving
Utility pattern mining
Perturbation
Frequent pattern mining
Data mining

ABSTRACT

As one of the important approaches in privacy preserving data mining, privacy preserving utility mining has been studied to find more meaningful results while database privacy is ensured and to improve algorithm efficiency by integrating fundamental utility pattern mining and privacy preserving data mining methods. However, its previous approaches require a significant amount of time to protect the privacy of data holders because they conduct database scanning operations excessively many times until all important information is hidden. Moreover, as the size of a given database becomes larger and a user-specified minimum utility threshold becomes lower, their performance degradation may be so uncontrollable that they cannot operate normally. To solve this problem, we propose a fast perturbation algorithm based on a tree structure which more quickly performs database perturbation processes for preventing sensitive information from being exposed. We also present extensive experimental results between our proposed method and state-of-the-art algorithms using both real and synthetic datasets. They show the proposed method has not only outstanding privacy preservation performance that is comparable to the previous ones but also 5–10 times faster runtime than that of the existing approaches on average. In addition, the proposed algorithm guarantees better scalability than that of the latest ones with respect to databases with the characteristics of gradually increasing attributes and transactions.

1. Introduction

Data mining is an effective technique that extracts significant, available information within massive, complicated data. Frequent pattern mining (Gonzaalez, Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2013; Pyun, Yun, & Ryu, 2014) discovers recurring relationships in large repositories of data, called patterns or itemsets. Apriori (Agrawal & Srikant, 1994) is a representative method of the frequent pattern mining, which uses the generate-and-test approach to discover useful information. In addition, there are many variations of the frequent pattern mining such as incremental pattern mining considering dynamic databases (Ahmed, Tanbeer, Jeong, Lee, & Choi, 2012), closed or maximal pattern mining for extracting representative patterns (Yun, Lee, & Ryu, 2014; Yun, Pyun, & Yoon, in press; Yun & Ryu, 2013; Zaki & Hsiao, 2002), sliding window-based pattern mining considering data streams and time flows (Lee, Yun, & Ryu, 2014), erasable pattern mining for solving financial crises in industrial environments (Deng, 2013; Le & Vo, 2014), itemset mining based on lattice structures (Hashem et al., 2014; La, Le, & Vo, 2014), and top-k pattern mining for finding interesting top-k patterns without any predefined threshold (Pyun & Yun, 2014).

Utility pattern mining (Feng, Wang, & Jin, 2013; Ryang, Yun, & Ryu, 2014; Yun, Ryang, & Ryu, 2014), one of the interesting pattern mining techniques, can analyze business relationships in market data including utility issues since such utility pattern mining approaches can deal with non-binary data such as quantities of products and consider relative importance of items such as their profits. However, these methods may also cause privacy concerns by violating the privacy of data holders because they can mine sensitive information of data holders (Kantarcioglu, Jin, & Clifton, 2004). In business environments, privacy breaches (or risks) signify that specific information data holders want to hide is unexpectedly or illegally discovered by data analyzing methods. Especially, utility mining is more likely to cause privacy breaches in the process of business data analysis since it can find meaningful information including sensitive data hidden from large-scale business data and companies may malevolently utilize the information to enhance their own profits. For example, companies are usually reluctant to expose their own important, sensitive information to their partners even though they cooperate with each other for better profits because data holders may be damaged when someone who has a malevolent purpose uses the information. Sensitive

[*] Corresponding author.
E-mail addresses: yunei@sejong.ac.kr (U. Yun), jiwonkim@sju.ac.kr (J. Kim).

http://dx.doi.org/10.1016/j.eswa.2014.08.037
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

information is a series of significant data that can lead to privacy breaches of data holders when the information is disclosed by malevolent people (Yeh & Hsu, 2010); for this reason, the concept of privacy preserving utility mining (PPUM) (Yeh & Hsu, 2010), which is an integration of utility pattern mining (Ryang & Yun, 2015; Tseng, Wu, Shie, & Yu, 2010) and privacy preserving data mining (PPDM) (Agrawal & Srikant, 2000; Verykios et al., 2004), has been proposed, and various PPUM approaches (Yeh & Hsu, 2010) have been devised. Compared to PPDM, PPUM can deal with more important, sensitive information because of the characteristics of utility pattern mining. A majority of the efficient PPUM methods have applied database perturbation (Amiri, 2007), one of the PPDM techniques, for effectively hiding sensitive utility items or patterns by removing certain items from original databases.

Previous state-of-the-art PPUM algorithms such as HHUIF and MSICF (Yeh & Hsu, 2010) iterate database scanning operations until their own perturbation processes are completely finished. Therefore, they require an enormous number of database accesses in general; for this reason, inefficient cases can frequently be caused especially when a given utility database has a large-scale volume. In the example of retail applications, massive data are accumulated by numerous customers. Moreover, such data also continue to become larger because of the retail market's characteristics such as increasing transactions, customers, products, etc. Let us consider performing the database perturbation process with the previous PPUM approaches from such large-scale database. Then, they have to scan the database many times – even in the worst case, hundreds or thousands of times. Furthermore, at a lower minimum utility support threshold, sensitive items or itemsets are more frequently contained in the database; hence, the number of required scanning times becomes much larger. Motivated by the fatal problem, we propose a fast perturbation algorithm Using a Tree structure and Tables (called FPUTT), which can more quickly conduct the perturbation process with only three database scans (two times for constructing the tree and one time for updating the modified transactions with sensitive itemsets to the original database) by using a newly proposed tree structure and tables. The main contributions of this paper are summarized as follows.

1. Devising a new tree structure, called FPUTT-tree, which has the following novelty and characteristics: (1) a different form from the trees used in the previous utility pattern mining algorithms (Feng et al., 2013; Tseng, Shie, Wu, & Yu, 2013), (2) a novel approach for PPUM that has never been used in this area, (3) it is possible to construct FPUTT-tree with only two database scans, and (4) once the tree construction is finished, there is no longer need to perform further database scans for the perturbation process – except for one database scan for updating modified transaction information in the last phase of the process.
2. Proposing a novel perturbation algorithm guaranteeing much faster execution time and reliable PPUM performance, which is a result considering not only the developed tree structure and its techniques but also the following two table structures suggested in this paper: a Sensitive Itemset table (*SI-table*) utilized for reducing useless tree traversal operations by allowing the algorithm to search for FPUTT-tree more efficiently, and an Insensitive Item table (*II-table*) used for preventing the integrity of the original database from being violated, i.e., for protecting any loss of original database information.
3. Providing extensive, comprehensive experimental results and analysis between the proposed algorithm and the state-of-the-art PPUM methods, HHUIF and MSICF, which include theoretical comparison in terms of time complexity and empirical comparison with regard to execution time, privacy preservation performance, and scalability based on real and synthetic datasets.

The remaining parts of this paper are organized as follows. Section 2 provides information on the previous works of PPDM and utility pattern mining, and Section 3 provides preliminaries for utility pattern mining. Next, Section 4 describes the details of FPUTT that include a model, data structures, techniques, and theoretical analysis for the proposed algorithm. Section 5 shows the effectiveness of our method through experimental results on real and synthetic datasets. Finally, Section 6 concludes this paper and discusses our future research directions.

2. Related work

2.1. Utility pattern mining

As a fundamental descriptive mining method of the PPUM research, utility pattern mining (Feng et al., 2013) extracts useful knowledge from databases with utility issues. The concept of utility is to consider the quantity and profit of each item; thus, this approach can solve the problem that does not deal with market data in general frequent pattern mining. That is, frequent pattern mining simply analyzes the ratio of each itemset to the whole database, i.e., support, whereas utility pattern mining considers not only the frequency or support for each itemset but also their different importance. Therefore, utility pattern mining can conduct analysis of non-binary data composed of transactions reflecting characteristics of customers. Therefore, results of utility pattern mining can include more important but private, sensitive information; for this reason, it is more essential to preserve such information from malicious users by applying PPUM approaches that specialize in the utility pattern mining rather than traditional pattern mining.

In addition, various studies have been conducted to improve the performance of utility pattern mining by decreasing the number of database scans and necessary memory and runtime resources. The representative high performance algorithms based on tree structures are UP-Growth (Tseng et al., 2010) and Up-Growth+ (Tseng et al., 2013). The UP-Growth approach elicits mining results by using its own tree construction strategies: *DLU*, *DLN*, *DNU*, and *DNN*. UP-Growth+ conducts utility pattern mining through a tree structure utilizing an additional element and effective methods for pruning candidate itemsets. Generally, UP-Growth+ is regarded as the fastest algorithm among approaches based on tree structures. As another utility pattern mining method, HUM-UT (Tseng et al., 2010) is a one-scan algorithm and can extract useful information from transactional data streams. The HUM-UT approach presents a new tree structure, *UT-tree*, for an effective tree reconstruction method. CHUD (Wu, Fournier-Viger, Yu, & Tseng, 2011) was introduced for mining closed high utility itemsets. It uses the following three effective techniques: *REG*, *RML*, and *DCM*. HURM (Lee, Park, & Moon, 2013) is an association rule mining approach based on utility conditions. This method works for a cross-selling environment as a marketing solution and measures specific business profits of firms.

Note that the types of utility pattern mining algorithms do not affect our final result. The reason why any utility pattern mining algorithm can be used is as follows. Each utility pattern mining method has different efficiency in terms of time, space, the number of candidates, and so on, but its utility itemset results are equal to one another. Moreover, the utility itemset results are only used to progress PPUM. For this reason, we use the state-of-the-art fundamental utility algorithm, UP-Growth+, as a utility pattern mining method. The UP-Growth+ approach shows the fastest performance among utility pattern mining algorithms based on tree structures.