Engineering Applications of Artificial Intelligence 55 (2016) 269–284

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining

Jerry Chun-Wei Lin [a,*], Tsu-Yang Wu [a], Philippe Fournier-Viger [b], Guo Lin [a], Justin Zhan [c], Miroslav Voznak [d]

[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[b] School of Natural Sciences and Humanities, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[c] Department of Computer Science, University of Nevada, Las Vegas, USA
[d] Department of Telecommunications, VSB-Technical University of Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 23 August 2015
Received in revised form
17 May 2016
Accepted 14 July 2016

Keywords:
Privacy preserving data mining
Utility mining
Minimum side effects
Maximum sensitive utility
PPUM

ABSTRACT

High-Utility Itemset Mining (HUIM) is an extension of frequent itemset mining, which discovers itemsets yielding a high profit in transaction databases (HUIs). In recent years, a major issue that has arisen is that data publicly published or shared by organizations may lead to privacy threats since sensitive or confidential information may be uncovered by data mining techniques. To address this issue, techniques for privacy-preserving data mining (PPDM) have been proposed. Recently, privacy-preserving utility mining (PPUM) has become an important topic in PPDM. PPUM is the process of hiding sensitive HUIs (SHUIs) appearing in a database, such that the resulting sanitized database will not reveal these itemsets. In the past, the HHUIF and MSICF algorithms were proposed to hide SHUIs, and are the state-of-the-art approaches for PPUM. In this paper, two novel algorithms, namely Maximum Sensitive Utility-MAximum item Utility (MSU-MAU) and Maximum Sensitive Utility-MInimum item Utility (MSU-MIU), are respectively proposed to minimize the side effects of the sanitization process for hiding SHUIs. The proposed algorithms are designed to efficiently delete SHUIs or decrease their utilities using the concepts of maximum and minimum utility. A projection mechanism is also adopted in the two designed algorithms to speed up the sanitization process. Besides, since the evaluation criteria proposed for PPDM are insufficient and inappropriate for evaluating the sanitization performed by PPUM algorithms, this paper introduces three similarity measures to respectively assess the database structure, database utility and item utility of a sanitized database. These criteria are proposed as a new evaluation standard for PPUM.

1. Introduction

In recent years, the amount of data generated and transferred between companies, governments and other organizations has massively increased. Furthermore, with the rapid growth of data mining technologies (Han et al., 2004; Chen et al., 1996; Agrawal and Srikant, 1994a; Lin et al., 2015; Duong et al., 2014), hidden relationships between items in databases can now be uncovered with ease, for the purpose of decision making or to discover users' preferences. As a result, a major issue that has arisen is that knowledge found by data mining techniques may implicitly reveal confidential, private or sensitive information (e.g., personal identification numbers, home addresses, social security numbers, and credit card information).

This issue is especially concerning when organizations publish information publicly. In this case, the use of data mining technologies may lead to privacy threats against individuals or the misuse of their data (Atallah et al., 1999; Agrawal and Srikant, 2000; Verykios et al., 2004). A similar problem can occur when data is shared between organizations that are collaborating with the goal of increasing their benefits or achieving higher profits. Data may be analyzed by collaborators or competitors to discover sensitive or strategic knowledge in the data, which may decrease a company's benefits or result in security threats.

To address these problems, techniques for Privacy-Preserving Data Mining (PPDM) have been proposed. They consist of perturbing a database to sanitize it (Amiri, 2007). Numerous algorithms have been proposed to hide confidential patterns that appear in binary databases such as association rules and frequent itemsets (Verykios et al., 2004; Dasseni et al., 2001; Giannotti et al., 2012). Most of these algorithms delete transactions or itemsets from an original transactional database to respectively reduce the support or confidence of the sensitive patterns during the sanitization process.

* Corresponding author.
E-mail addresses: jerrylin@ieee.org (J.-W. Lin), wutsuyang@gmail.com (T.-Y. Wu), philfv@hitsz.edu.cn (P. Fournier-Viger), linguo.hit@gmail.com (G. Lin), justin.zhan@unlv.edu (J. Zhan), miroslav.voznak@vsb.cz (M. Voznak).

http://dx.doi.org/10.1016/j.engappai.2016.07.003
0952-1976/

Bertino et al. first proposed a framework to evaluate the performance of sanitizing algorithms in PPDM (Bertino et al., 2005). Fayyad et al. published a systematic introduction to data mining, while drawing attention to the problem of privacy (Fayyad et al., 1996). Lindell and Pinkas (2000) designed a protocol focusing on the decision tree learning with the popular ID3 algorithm to address the issue of PPDM. Verykios et al. presented three strategies and five algorithms to hide sets of association rules in transaction databases (Verykios et al., 2004). Lin et al. presented the PSO-based algorithm to sanitize the database in PPDM (Lin et al., 2016).

Recently, High-Utility Itemset Mining (HUIM) has emerged as a critical data mining task. It allows to reveal itemsets yielding a high profit in transactions databases. For this reason, HUIM is more practical than frequent itemset mining for retailers in real-life situations (Chan et al., 2003; Yao and Hamilton, 2006). HUIM considers both the unit profits of items and their quantities in transactions.

Given the aforementioned considerations of PPDM, Privacy-Preserving Utility Mining (PPUM) has become an important topic in recent years. But few studies have addressed the issue of PPUM. Moreover, most of them hide sensitive high-utility itemsets (SHUIs) by reducing the quality of the database or by deleting transactions. A SHUI is a high-utility itemset that is viewed as confidential or sensitive and needs to be hidden before a database is published or shared. Yeh and Hsu (2010) first proposed the Hiding High Utility Itemset First (HHUIF) algorithm and the Maximum Sensitive Itemsets Conflict First (MSICF) algorithms to hide SHUIs. The HHUIF algorithm sanitizes a database with respect to a sensitive itemset by identifying its item having the maximal utility, and deleting that item or decreasing its quantity in transactions. The MSICF algorithm adopts a similar approach but considers the item with the maximal occurrence frequency for deletion. Lin et al. presented a genetic algorithm based approach for hiding SHUIs through transaction insertion (Lin et al., 2014b). Yun et al. proposed the Fast Perturbation algorithm Using a Tree structure and Tables (FPUTT) algorithm (Yun and Kim, 2015) to speed up the sanitization process using a tree structure and its associated index table. The above studies have utilized criteria from PPDM to assess the performance of the developed sanitization algorithms. As it will be explained, these criteria are insufficient for evaluating the performance for PPUM.

In this paper, two algorithms, namely Maximum Sensitive Utility-MAximum item Utility (MSU-MAU) and Maximum Sensitive Utility-MInimum item Utility (MSU-MIU) are proposed to efficiently hide SHUIs. Since the criteria used in PPDM (Bertino et al., 2005) to evaluate the performance of sanitization algorithms are unsuitable for PPUM, three novel similarity measures are also proposed in this paper to assess the effectiveness and efficiency of the proposed algorithms. The major contributions of this paper are threefold:

1. Two algorithms named MSU-MAU and MSU-MIU are developed to efficiently hide sensitive high-utility itemsets (SHUIs) by considering the utility measure in PPUM, and ensure a lower missing cost and higher database integrity.
2. The designed algorithms only perform decrease or delete operations to modify quantities of items in a database. Thus, the algorithms do not introduce an artificial cost.
3. In addition to the traditional evaluation criterion of PPDM, this paper introduces three similarity measures named Database-Structure Similarity (DSS), Database-Utility Similarity (DUS), and Itemsets-Utility Similarity (IUS) to better evaluate the effectiveness and efficiency of the sanitization algorithms for PPUM.

The rest of this paper is organized as follows. Related work is described in Section 2. Preliminaries and problem statement are stated in Section 3. The developed sanitization algorithms are presented in Section 4. Results from an extensive experimental evaluation are reported in Section 5. Finally, a conclusion is drawn and future work is discussed in Section 6.

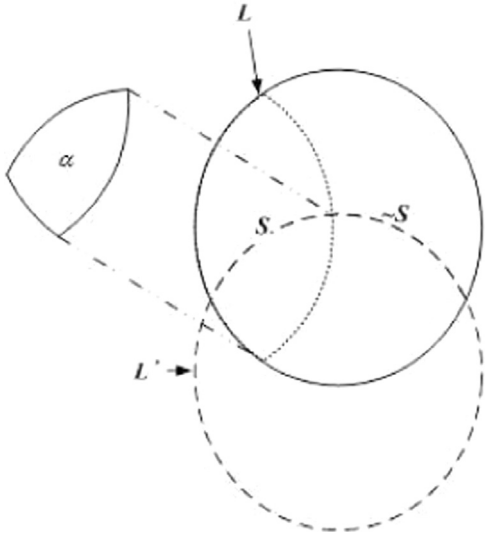

Fig. 2. The set of sensitive itemsets that the PPDM process failed to hide.

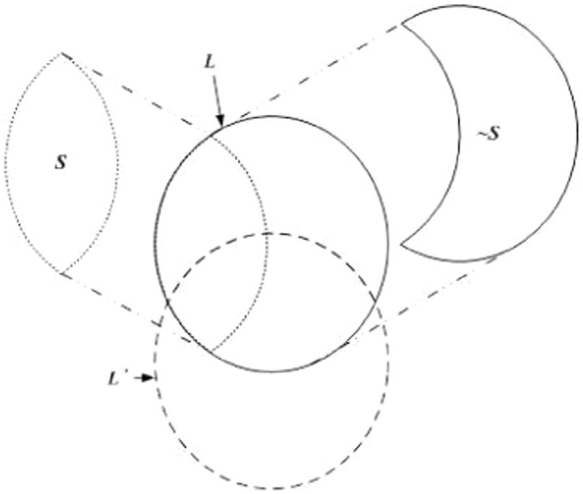

Fig. 1. Relationship between itemsets before and after the PPDM process.

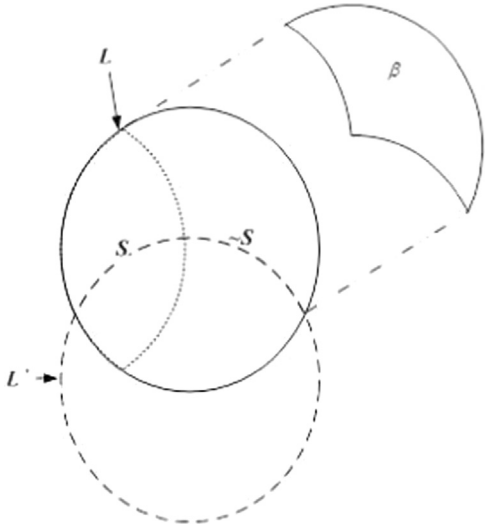

Fig. 3. The missing cost resulting from the sanitization process.