Perspectives in Science (2016) 7, 74—80

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc

High utility-itemset mining and privacy-preserving utility mining☆

Jerry Chun-Wei Lin[a,*], Wensheng Gan[a], Philippe Fournier-Viger[b], Lu Yang[a], Qiankun Liu[a], Jaroslav Frnda[c], Lukas Sevcik[c], Miroslav Voznak[c]

[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 Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 17. listopadu 15, 708 00 Ostrava-Poruba, Czech Republic*

Received 27 October 2015; received in revised form 27 October 2015; accepted 11 November 2015
Available online 10 December 2015

KEYWORDS
Data mining;
High-utility itemset;
Privacy preserving;
PSO algorithm;
GA algorithm;
Evolutionary algorithms

Summary In recent decades, high-utility itemset mining (HUIM) has emerging a critical research topic since the quantity and profit factors are both concerned to mine the high-utility itemsets (HUIs). Generally, data mining is commonly used to discover interesting and useful knowledge from massive data. It may, however, lead to privacy threats if private or secure information (e.g., HUIs) are published in the public place or misused. In this paper, we focus on the issues of HUIM and privacy-preserving utility mining (PPUM), and present two evolutionary algorithms to respectively mine HUIs and hide the sensitive high-utility itemsets in PPUM. Extensive experiments showed that the two proposed models for the applications of HUIM and PPUM can not only generate the high quality profitable itemsets according to the user-specified minimum utility threshold, but also enable the capability of privacy preserving for private or secure information (e.g., HUIs) in real-word applications.

Introduction

With the rapid growth of information techniques and various applications, the Knowledge Discovery in Database (KDD) which is also called data mining, has become a powerful technique and commonly be used to discover interesting and useful knowledge from massive data. The discovered

☆ This article is part of a special issue entitled ''Proceedings of the 1st Czech-China Scientific Conference 2015''.

* Corresponding author.
E-mail address: jerrylin@ieee.org (J.C.-W. Lin).

http://dx.doi.org/10.1016/j.pisc.2015.11.013

knowledge from data mining can be generally classified as frequent patterns (FIs) or association rules (ARs) (Agrawal et al., 1993a,b; Han et al., 2004), high-utility patterns (HUPs) (Ahmed et al., 2009; Chan et al., 2003; Liu et al., 2005; Yao et al., 2004), sequential patterns (SPs) (Agrawal and Srikant, 1995; Pei et al., 2004), clustering (Berkhin, 2006) and classification (Kotsiantis, 2007), among others. Among them, association-rule mining (ARM) (Agrawal et al., 1993a,b; Han et al., 2004) is the fundamental knowledge which can be commonly used to analyze the purchase data of customers. However, ARM only considers whether or not the item or itemset is present in a transaction. The other factors in real-life applications, such as weight, profit, quantity, risk or other measures, are not considered in ARM. Thus, high-utility itemset mining (HUIM) (Ahmed et al., 2009; Chan et al., 2003; Liu et al., 2005; Yao et al., 2004) has emerging a critical issue in recent years since it can be used to reveal profitable itemsets (high-utility itemsets) by considering both purchase quantity and distinct profit factors.

Due to quick proliferation of massive data from government, corporations and organizations, the discovered knowledge may, however, implicitly contain confidential, private or secure information (i.e., personal identification numbers, address information, social security numbers, or credit card numbers), which leads to privacy threats if they are misused (Agrawal and Srikant, 2000; Verykios et al., 2004). Generally, collaboration among industries can work together to share information for achieving higher benefits and profits in business. However, the shared information can be extracted and analyzed by the other collaborators or competitors, thus decreasing its own benefits and causing the security threats. Privacy-preserving data mining (PPDM) was thus proposed to address the above limitations by perturbing the original database and producing a sanitized one (Amiri, 2007). Many algorithms have been extensively proposed to hide the private or secure information (i.e., FIs or ARs) from the different type databases (Dunning and Kresman, 2013; Han and Ng, 2007). As the similar consideration of PPDM, privacy preserving for high-utility itemset mining (PPUM) has also become an important topic in recent years. Fewer studies have addressed the issue of PPUM and most of them are processed to reduce the quality or delete transactions for hiding sensitive high-utility itemsets (SHUIs). Several related algorithms for PPUM have been extensively studied, such as the Hiding High-Utility Itemset First (HHUIF) algorithm and Maximum Sensitive Itemsets Conflict First (MSICF) algorithm to hide the SHUIs (Yeh and Hsu, 2010), the GA-based algorithm for hiding SHUIs through transaction insertion (Lin et al., 2014a), and the Fast Perturbation algorithm Using a Tree structure and Tables (FPUTT) algorithm (Kannimuthu and Premalatha, 2015) to speed the sanitization process with an aided tree structure and the associated index table. The above approaches, however, are insufficient since whether PPDM or PPUM belongs to the NP-hard problem. It is necessary to hide the sensitive information but discover the required information in decision making.

In this paper, we address the research issues by proposing efficient algorithms applied in HUIM and PPUM. We firstly propose a PSO-based algorithm for HUIM, and secondly propose a GA-based approach to evaluate the effectiveness and efficiency of PPUM. Key contributions of this paper are present below. (1) We present two evolutionary algorithms, the PSO-based algorithm for efficiently mining HUIs and the GA-based privacy preserving algorithm in PPUM. (2) Not only the mining performance for HUIM, but also a trade-off between mining performance and its privacy preserving for the SHUIs can be ensured. (3) Extensive experiments conducted on several real-life and synthetic datasets showed that the two proposed models for HUIM and the applications of PPUM have better results than the previous works whether in HUIM or PPUM.

Background

High-utility itemset mining

The concept of high-utility itemset mining was first proposed by Chan et al. (2003), and the mathematical mode was formed by Yao et al. (2004). The problem of high-utility itemset mining (HUIM) was defined to find the rare frequencies itemsets but with high profits (Yao et al., 2004). Since the downward closure property is no longer kept for HUIM, a Two-Phase model (Liu et al., 2005) was presented to keep the transaction-weighted utilization downward closure (TWDC) property for discovering HUIs. Several tree-based algorithms were also proposed, such as the HUP-tree-based IHUP algorithm for mining HUIs in incremental databases (Ahmed et al., 2009), the HUP-growth algorithm (Lin et al., 2011) to find HUIs without candidate generation, and the efficient UP-tree-based two mining algorithms, UP-growth (Tseng et al., 2010) and UP-growth+ (Tseng et al., 2013). Liu et al. then proposed the HUI-Miner algorithm (Liu and Qu, 2012) to build utility-list structures and to develop a set-enumeration tree to directly extract HUIs without either candidate generation or an additional database rescan. The improved FHM algorithm was further designed by enhancing the HUI-Miner for analyzing the co-occurrences among 2-itemsets (Fournier-Viger et al., 2014).

Besides, many interesting other issues on HUIM have also been extensively studied, such as up-to-date HUIs mining which aims at discovering recent HUIs which may be more useful and interesting (Lin et al., 2015d); HUIs mining in dynamic environment (e.g., data insertion (Lin et al., 2014b), data deletion (Lin et al., 2015a), and data modification (Lin et al., 2015c)), HUIs mining in stream data environment (Li et al., 2008); on-shelf HUIs mining (Lan et al., 2011); top-*k* HUIs mining (Wu et al., 2012); HUIs mining with multiple minimum utility thresholds (Lin et al., 2015b); HUIs mining with or without negative unit profit value (Fournier-Viger, 2014).

Privacy preserving for high-utility itemsets mining

Although data mining various techniques can be used to find the implicit information, the confidential or secure information is, however, required to be hidden before it is published in the public place or shared among the collaborators. Privacy-preserving data mining (PPDM) has thus arisen as an important topic in recent years (Agrawal and Srikant, 2000; Han and Ng, 2007; Verykios et al., 2004). A novel reconstruction procedure was presented to accurately estimate the distribution of original data values, thus building the