Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 59 (2006) 378–396
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Association rules mining in vertically partitioned databases ☆

Boris Rozenberg, Ehud Gudes *

Department of Computer Science, Ben-Gurion University, 84105 Beer-Sheva, Israel

Received 21 June 2005; received in revised form 13 September 2005; accepted 13 September 2005
Available online 10 October 2005

Abstract

Privacy concerns have become an important issue in Data Mining. This paper deals with the problem of association rule mining from distributed vertically partitioned data with the goal of preserving the confidentiality of each database. Each site holds some attributes of each transaction, and the sites wish to work together to find globally valid association rules without revealing individual transaction data. This problem occurs, for example, when the same users access several electronic shops purchasing different items in each. We present two algorithms for discovering frequent itemsets and for calculating the confidence of the rules. We then analyze the algorithms privacy properties, and compare them to other published algorithms.

Keywords: Data mining; Privacy; Association rules; Distributed databases

1. Introduction

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, and association rule mining. Traditionally, all these algorithms have been developed within a centralized model, with all data being gathered into a central site, and algorithms being run against that data. Privacy concerns can prevent this approach. There may not be a central site with authority to see all the data, or if the data is partitioned among several sites, sites may not want to disclose to each other their individual database, even if they are willing to disclose some information in order to enable the extraction of globally valid knowledge. A typical example using the traditional "market basket" example, may involve multiple internet shops which serve the same community of users, one site may contain grocery purchases, while another has clothing purchases. Using a unique key such as customer identification, one may like to mine rules concerning the relationships between the purchasing of the various items. A similar example may involve two different hospitals serving the same community. Clearly if all participating sites are willing to share their data, or trust a third party to do the mining, the

☆ Partially supported by the Lynn and Frankel Center for Computer Science and by the Paul Ivanir Robotics Center.
* Corresponding author.
E-mail addresses: rozenbu@cs.bgu.ac.il (B. Rozenberg), ehud@cs.bgu.ac.il (E. Gudes).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2005.09.001

privacy problem is solved. We deal here with the problem that each site tries to protect the confidentiality of its database, without compromising most of the knowledge discovery process.

Informally, the problem is to mine association rules across several databases, where the columns in the table are at different sites. In [3] Vaidya and Clifton presented some successful solutions for this problem (for two sites—called here VDC) and in [18] (for $n > 2$ sites—called here VDCN). Their algorithms require the intensive use of secure computation in order to preserve privacy. Secure computation [5] allows computing functions whose input is provided by the participants to the computaiont without revealing their private input (see Section 2.2). However, as will be shown later in this paper, VDC/N algorithms have the potential for inferring private information based on the results in certain cases. This was one of the main motivations for our algorithms.

Generally, the mining of association rules is divided into two phases: finding frequent itemsets with support above a threshold, and finding rules with confidence above a threshold. In [16] we presented two algorithms for solving the first problem. In [21] we presented an updated version of our Two-party algorithm from [16], and a new version of the *N*-parties algorithm, and briefly described the confidence computing algorithms. In this paper we present the above algorithms in detail, give more precise analysis of our algorithms, and present a comprehensive comparison of them to the VDC/N algorithms.

The rest of the paper is structured as follows. Section 2 presents the background and related work. Section 2.1 reviews the Apriori algorithm for discovering association rules in non-partitioned databases [1]. Section 2.2 discusses some works addressing Secure Multiparty Computation. Sections 2.3 and 2.4 review the VDC and VDC/N algorithms. In Section 2.5 we analyze the VDC/N algorithms and show example cases where they can disclose private information. Section 3 discusses our two two-party algorithm and *N*-parties algorithm for calculating frequent itemsets and the corresponding algorithms for calculating the confidence of the rules. Section 4 analyzes the privacy, security and complexity of the algorithms and compares them to the VDC/N algorithms. Section 5 is the conclusions and future work section. Note that the analysis in Section 4 is a concise version of the full analysis, which is given in [17].

2. Background

Agrawal and Srikant [1] have proposed the apriori algorithm for discovering all significant association rules between items in a large (not distributed) database of transactions. However, this work does not address privacy concerns. Later in [11], the authors propose a procedure in which some or all the numerical attributes are perturbed by a randomized value distortion so that both the original values and their distributions are changed. The proposed procedure then performs a reconstruction of the original distribution. This reconstruction does not reveal the original values of the data, and yet allows the learning of decision trees. Another paper [15] shows a reconstruction method, which does not entail information loss with respect to the original distribution.

Other randomization techniques were proposed in order to provide association rules mining without revealing sensitive information about individuals [19,20]. These techniques are based on probabilistic distortion of user data in the way that can provide a high degree of privacy and retain a high level of accuracy of the result. For example, in [19], the value of the attribute is retained with probability p and flipped with probability $1 - p$. The presented experimental results showed that distortion probability of $p = 0.1$ is ideally suited to provide both privacy and good mining results. We'll come back to this technique in Section 4.3.1.

When the data is partitioned among several sites, one may use the existing data mining algorithms for a centralized database at each site independently and combines the results [2], but this will often fail to achieve a globally valid result. As pointed out in [3], the issues that can cause a disparity between local and global results include:

- Values for a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations.
- The same item may be duplicated at different sites, and will be over-weighted in the results.
- Data at a single site is likely to be from a homogeneous population, hiding geographic or demographic distinctions between that population and others.