第 35 卷 第 7 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.7 Computer Engineering April 2009
·安全技术· 文章编号：1000—3428(2009)07—0153—03 文献标识码：A 中图分类号：TP309

基于可逆方阵的隐私保护关联规则挖掘

田 宏[1,2]，王亚伟[2]，王秀坤[1]
(1. 大连理工大学电子与信息工程学院，大连 116024；2. 大连交通大学软件学院，大连 116028)

摘 要：数据隐私问题引起人们的广泛关注，如何在分布式数据库的环境下挖掘关联规则成为研究的热点。该文探讨在垂直划分数据库中，如何在保护各方隐私数据的前提下挖掘全局频繁项集。各分布式数据库包含全局数据库的一部分属性，共同参与全局挖掘，同时各方不向外泄漏隐私数据。在商品服务器模型的研究基础上，提出一种基于可逆方阵的加密协议，对于垂直划分的分布式数据库，该协议具有较好的隐蔽性、高效性和准确性。

关键词：关联规则；分布式数据库；隐私；可逆方阵

Invertible Matrix-based Privacy-preserving Association Rules Mining

TIAN Hong[1,2], WANG Ya-wei[2], WANG Xiu-kun[1]
(1. School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024;
2. Software Institute, Dalian Jiaotong University, Dalian 116028)

【Abstract】With the growing concern over data privacy-preserving problem, how to discover association rules from distributed databases becomes one of the hot topics of this field. This paper is devoted to study the problem of discovering global frequent itemsets from distributed vertically partitioned databases with the goal of preserving the confidentiality of each database. All sites are worked together to find global frequent itemsets without revealing private data, each one holds some attributes of global database. The paper presents an invertible matrix-based encryption protocol based on the research of commodity server model, which protocol is of great confidentiality, effectiveness and correctness for distributed vertically partitioned databases.

【Key words】association rules; distributed database; privacy; invertible matrix

1 概述

数据挖掘是指从大量的数据中发现潜在的、新颖的、有价值的、可用的、能被用户理解的模式和信息的过程。目前数据挖掘已经在金融业、零售业、医学分析、生产控制、工程设计等领域得到了广泛的应用。随着数据的大量积累，这些蕴含着知识的数据通常分布在不同的站点，传统的数据挖掘方法只能对集中式数据库进行挖掘，所有的数据都必须集中存放在一个站点。由于数据隐私保护问题受到越来越多的重视，因此这种集中式的挖掘方法已经不再有效，目前有大量的事例说明对隐私保护数据挖掘的研究十分紧迫。

文献[1]提出一个有关汽车安全的事例：一款福特汽车和某一型号的凡士通轮胎在特定情况下会产生轮胎面出现裂口的问题。如果及早发现问题，可以使至少 800 人幸免于难。因为这种轮胎装配其他型号的汽车没有问题，这款汽车和其他厂商的轮胎组装也没有问题，两方都觉得没有责任。他们都有自己的数据库，各自的数据库都是需要隐私保护的，如果联合进行数据挖掘，就可以找到问题的原因。

数据挖掘需要准确的原始数据，而数据的隐私性又要求对原始数据进行保密，因此，两者存在矛盾。在这种情况下，基于隐私保护的数据挖掘技术成为未来研究的重点。

2 相关工作

依据对原始数据处理方式的不同，基于隐私保护的数据挖掘可分为 2 种类型：(1)输出隐私保护，即通过最小程度的改变原始数据，使得挖掘的结果能够很好地保护隐私信息。对于这种类型的隐私保护已经形成多种技术，例如干扰、分块、聚集、交换以及采样等。(2)输入隐私保护，通过对原始数据的转换，使得挖掘结果不受影响或影响很小。例如基于加密技术的安全多方计算只允许访问原始数据的一个子集，而使全局挖掘结果不变。

2.1 CSM模型

CSM(Commodity Server Model)模型[2-3]中引入第三方的商品服务器，唯一的前提是第三方不与任何一方相互串通。这种模型有如下优点：(1)第三方不参与计算过程，它只为各方提供数据用于隐藏原始数据。(2)第三方提供的数据不依赖于参与者的隐私信息，因此，第三方不需要知道隐私数据。基于此模型的两方和多方协议比其他两方和多方协议的效率高，但是如果第三方提供的随机数据超出了隐私数据的值域，有可能导致隐私数据的泄漏。

2.2 安全多方计算

安全多方计算 SMC(Secure Multi-party Computation)是指在一个互不信任的多用户网络中，2 个或多个用户能够在不泄漏各自私有输入信息时联合执行某项计算任务。通俗地说，安全多方计算是指在一个分布式网络中，多个用户各自持有一个秘密输入，他们希望共同完成对某个函数的计算，而要求每个用户除计算结果外均不能够得到其他用户的任何输入信息，安全多方计算不信赖第三方。文献[4]首先提出了 SMC

基金项目：辽宁省自然科学基金资助项目(20062114)
作者简介：田 宏(1968－)，女，副教授、博士研究生，主研方向：数据挖掘；王亚伟，硕士研究生；王秀坤，教授、博士生导师
收稿日期：2008-09-13 **E-mail：** th@djtu.edu.cn

问题，后来被文献[5]证明对于任何一个多项式函数都有 SMC 解决方法。他们都采用一个类似的方法：每一个函数 F 用一个布尔电路表示，各方在每一个门电路运行协议。这种方法简单通用，生成的协议依赖于电路的规模，电路的规模依赖于输入和函数 F 的复杂程度。如果函数 F 过于复杂，这种方法对于数据量大的输入不再适用。

文献[6]对一些特殊问题采用了简单处理，其中之一就是向量点积协议。这种协议采用一种 1-out-of N OTP 协议处理两方的安全计算，在其中一方进行挖掘的时候，利用生成的随机向量和随机数对原始数据进行加密，另一方负责接收这些加密后的数据并对数据计算，将计算结果返回。这种方法简单实用，保密性很好，不足在于在计算过程中需要生成大量的随机向量和随机数，而且通信过程中也要传递很多对计算结果无任何影响的随机向量，加大了系统开销。

综上所述，在分布式数据挖掘中，设计一个优良的保护隐私的挖掘算法需要考虑以下几个方面：正确的挖掘结果，计算开销，通信代价和安全强度。

3 基于可逆方阵的频繁项集挖掘

关联规则挖掘的目的是寻找在大量的数据项中隐藏着的有趣的关联，即数据库中的知识模式。关联规则是由 R.Agrawal 等人首先提出的，关联规则挖掘就是在事务数据库中找到满足用户给定的最小支持度和最小置信度的强关联规则，其中计算项集的支持度是关键。

3.1 问题的定义

在垂直划分的分布式数据库环境下，假设关联规则挖掘是一个多方参与的过程，各方的私有数据集 $D_1, D_2, \cdots, D_n$，定义全局数据集 $D=(D_1 \cup D_2 \cup \cdots \cup D_n)$。各方数据集 D_i 包括不同的属性集(项集)，这些属性的并集构成全局数据库 D 属性集，即 $Arr(D_i) \cap Arr(D_j)=\varnothing,\ (i \neq j)$，并且 $Arr(D_1) \cup Arr(D_2) \cup \cdots \cup Arr(D_n)=Arr(D)$，其中，$Arr(D_i)$表示数据集 D_i 包含的属性集。

基于隐私保护的联合关联规则挖掘问题，就是在全局数据集 D 上进行关联规则挖掘，同时保证不能泄漏各方的隐私信息。考虑到问题的复杂性，本文做如下假设：

(1)第三方不能与任何一方相互串通交换不必要信息。

(2)各参与方必须诚实履行既定协议。

(3)网络通信安全畅通。

(4)为了降低计算开销，各方可以保存临时数据。

3.2 一种简单的两方协议

在安全性要求不是很高的应用场合，提出一种简单的两方协议，基于这种协议的挖掘算法计算结果准确，计算开销小而且通信代价低。假设有 2 个站点(Alice,Bob)，对应数据集 D_1 和 D_2，它们包含的事务列表 TIDs 完全相同。如何计算 $c.count$(c 表示候选频繁项集，$c.count$ 表示项集 c 的支持度)是求频繁项集的关键。若候选频繁项集所有属性属于同一方，计算完全可以在这一方内进行，不需要担心隐私保护的问题。若候选频繁项集属性属于两方，这就需要两方分别构造相应的布尔向量 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 用来计算 $c.count$。

首先合作的两方就各自向量 $\boldsymbol{C}_1$, $\boldsymbol{C}_2$ 进行加密处理，例如位移运算、奇偶位互换等，然后将加密结果 $e(\boldsymbol{C}_1)$和 $e(\boldsymbol{C}_2)$发送到第三方(只信赖第三方进行计算和协调各方通信，没有权限访问各方隐私数据)进行支持度计算，最后将计算结果返回两方。例如采用循环位移运算，左移 n 位，然后计算支持度。类似的简单加密方法有很多，计算开销和通信开销可以忽略不计。总结出这类加密方法满足的通式：

$$c.count = e(\boldsymbol{C}_1) \cdot e(\boldsymbol{C}_2)$$

经过适当的改进，这种两方协议可以很容易得到扩展，进而处理多方隐私保护问题。不足之处在于，在这种简单加密的情况下，第三方有可能推断出隐私数据，因此在安全性要求较高的背景下，这种简单的两方协议并不能满足要求。

3.3 一种安全高效的两方协议

为了改进简单加密协议安全性不足的问题，提出采用可逆方阵对原始数据进行加密，用矩阵的乘积运算代替向量点积运算，计算出候选频繁项集的支持度，由于在运算过程中不需要生成大量的随机向量和随机数，而且通信过程中只传递加密后的向量信息，因此相对于 SMC 来讲，计算开销和通信开销要小很多。

协议两方：Alice 和 Bob, $\boldsymbol{C}_1 \in$ Alice, $\boldsymbol{C}_2 \in$ Bob

第三方：Nike

可逆矩阵：$\boldsymbol{Q}$ (两方认可的公用密匙)

候选频繁项集：c

协议 1 (两方协议)

Alice 处理步骤：

(1) Alice 对向量 $\boldsymbol{C}_1$ 加密：$\boldsymbol{C}_1\boldsymbol{Q}$；

(2)将计算结果发送到 Nike。

Bob 处理步骤：

(1) Bob 对向量 $\boldsymbol{C}_2$ 加密：$\boldsymbol{Q}^{-1}\boldsymbol{C}_2^{\mathrm{T}}$；

(2)将计算结果发送到 Nike。

第三方计算矩阵的乘积 $(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_2^{\mathrm{T}})$，计算结果是一阶方阵，也就是 $c.count$，并将其返回 Alice 和 Bob。

3.3.1 正确性分析

第三方收到来自于 Alice 的加密行向量 $\boldsymbol{C}_1\boldsymbol{Q}$ 和来自于 Bob 的加密列向量 $\boldsymbol{Q}^{-1}\boldsymbol{C}_2^{\mathrm{T}}$，然后计算两者乘积 $(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_2^{\mathrm{T}})$，由于

$$(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_2^{\mathrm{T}}) = \boldsymbol{C}_1\boldsymbol{C}_2^{\mathrm{T}}$$

向量 $\boldsymbol{C}_1, \boldsymbol{C}_2$ 是 2 个 $1 \times n$ 矩阵，$\boldsymbol{C}_1\boldsymbol{C}_2^{\mathrm{T}}$是一个一阶方阵，即是：$\sum_{i=1}^{n} \boldsymbol{C}_1[i] \times \boldsymbol{C}_2[i]$，因此，$c.count$=$\boldsymbol{C}_1\boldsymbol{C}_2^{\mathrm{T}}$。

3.3.2 系统开销分析

在分布式数据库环境下进行关联规则挖掘的最优情况是不考虑隐私问题，两方直接用原始数据交替进行计算，省去加密和解密过程，这种情况下系统开销是最小的。本文提出的基于可逆矩阵的协议的系统开销包括：(1)各方对原始数据加密运算 $O(n^2)$；(2)有限的通信开销 $O(n)$；(3)第三方支持度计算的开销 $O(n)$。整个协议不产生随机向量，不传递冗余信息，各方可以保存临时数据，减少了重复计算，所以，系统开销较少。

3.3.3 安全性分析

每一次候选频繁项集支持度的计算都是在可以信赖安全计算的第三方进行，协议两方并没有相互交换加密的隐私信息，因此，两方都不能获得对方的隐私数据。第三方负责对收到的数据进行矩阵乘积运算，这些数据都是两方用矩阵加密后的数据，由于第三方不能得知加密矩阵 $\boldsymbol{Q}$，因此不能推断出两方的隐私信息。

3.4 一种安全高效的多方协议

上述讨论了两方协议，本节将把两方协议进一步推广到多方共同参与的多方协议。假设第一方对应私有向量 $\boldsymbol{C}_1$，另一方对应私有向量 $\boldsymbol{C}_2$，…，第 m 方对应私有向量 $\boldsymbol{C}_m$。为简

化，用 P_i 表示第 i 方。

在多方协议中，各方使用相同的加密方法，即用同一可逆方阵 $\boldsymbol{Q}$ 对原始数据加密。除了对 P_1 和 P_m 进行特别处理以外，其他各方采用相同的处理步骤。

协议方：$P_1, P_2, \cdots, P_m$ ($\boldsymbol{C}_1 \in P_1, \boldsymbol{C}_2 \in P_2, \cdots, \boldsymbol{C}_m \in P_m$)

第三方：Nike

可逆矩阵：$\boldsymbol{Q}$ (各方认可的公用密匙)

候选频繁项集：c

协议 2 (多方协议)

P_1 处理步骤：

(1)P_1 对向量 $\boldsymbol{C}_1$ 加密：$\boldsymbol{C}_1\boldsymbol{Q}$；

(2)将计算结果发送到 Nike。

$P_2, P_3, \cdots, P_{m-1}$ 处理步骤：

(1)各方对私有向量进行转化，将 $1\times n$ 行向量 $\boldsymbol{C}_i$ 转化为 $n\times n$ 对角矩阵 $\boldsymbol{M}_i, i=(2,3,\cdots,m-1)$，对角线上的元素与向量元素一一对应；

(2)计算矩阵乘积 $\boldsymbol{Q}^{-1}\boldsymbol{M}_i\boldsymbol{Q}$；

(3)将计算结果发送到 Nike。

P_m 处理步骤：

(1)P_m 对向量 $\boldsymbol{C}_m$ 加密：$\boldsymbol{Q}^{-1}\boldsymbol{C}_m^{\mathrm{T}}$；

(2)将计算结果发送到 Nike。

第三方 Nike 收到来自于各方的加密数据后，计算矩阵的乘积：

$$(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_2\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_3\boldsymbol{Q})(\cdots)(\boldsymbol{Q}^{-1}M_{m-1}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_m^{\mathrm{T}})$$

其结果是一阶方阵，也就是 $c.count$，并将其返回各方。

3.4.1 正确性分析

第三方负责将来自各方的矩阵作乘积运算，除了 P_1 和 P_m 的加密数据在计算中的位置确定以外，其他各方的加密数据可以相互交换位置，不影响计算结果。$\boldsymbol{C}_i \Rightarrow \boldsymbol{M}_i$ $i=(2,3,\cdots,m-1)$，转换过程如下：

$$(\boldsymbol{C}_i[1]\ \boldsymbol{C}_i[2]\ \cdots\ \boldsymbol{C}_i[n]) \Rightarrow \begin{pmatrix} \boldsymbol{C}_i[1] & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{C}_i[n] \end{pmatrix} \quad (\text{记：} diag(\boldsymbol{C}_i[1]\ \boldsymbol{C}_i[2]\ \cdots\ \boldsymbol{C}_i[n]))$$

定义 1 设 2 个 n 维布尔向量 $\boldsymbol{U}, \boldsymbol{V}$，$\boldsymbol{U}=(u_1\ u_2\ \cdots\ u_n)$，$\boldsymbol{V}=(v_1\ v_2\ \cdots\ v_n)$，则有 $\boldsymbol{U}\cap\boldsymbol{V}=(u_1\times v_1\ u_2\times v_2\ \cdots\ u_n\times v_n)$。

性质 1 设 2 个 n 维布尔向量 $\boldsymbol{U}=(u_1\ u_2\ \cdots\ u_n)$，$\boldsymbol{V}=(v_1\ v_2\ \cdots\ v_n)$，由向量 $\boldsymbol{V}$ 转换的对角矩阵 $\boldsymbol{M}=diag(v_1\ v_2\ \cdots\ v_n)$，易证：$\boldsymbol{U}\boldsymbol{M}=\boldsymbol{U}\cap\boldsymbol{V}$。

性质 2 由性质 1 可知：

$$(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_2\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_3\boldsymbol{Q})(\cdots)(\boldsymbol{Q}^{-1}M_{m-1}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_m^{\mathrm{T}}) = c.count$$

推导过程如下：

$$(\boldsymbol{C}_1\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_2\boldsymbol{Q})(\boldsymbol{Q}^{-1}M_3\boldsymbol{Q})\cdots(\boldsymbol{Q}^{-1}M_{m-1}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\boldsymbol{C}_m^{\mathrm{T}}) =$$

$$\boldsymbol{C}_1M_2M_3\cdots M_{m-1}\boldsymbol{C}_m^{\mathrm{T}} =$$

$$(\boldsymbol{C}_1\cap\boldsymbol{C}_2\cap\cdots\cap\boldsymbol{C}_{m-1})\boldsymbol{C}_m^{\mathrm{T}} =$$

$$\sum_{i=1}^{n}\prod_{j=1}^{m}\boldsymbol{C}_j[i]$$

又由于 $c.count=\sum_{i=1}^{n}\prod_{j=1}^{m}\boldsymbol{C}_j[i]$，因此性质 2 成立。

3.4.2 系统开销分析

多方协议的系统开销要比两方协议大，主要包括通信开销和计算开销。通信开销包括：(1)多方同步控制信息，协调各方进行同步挖掘，可忽略不计。(2)各方向第三方发送的加密信息，这部分通信开销依据候选项集涉及的参与方个数不同而不同，最坏情况完成一次计算需要通信开销接近 $O(mn^2)$ 位。计算开销包括：(1)各方加密计算所用开销，由于各站点可以并行计算且在联合挖掘之前挖掘本方频繁项集，并对频繁项集加密保存，因此这部分开销可以忽略不计。(2)第三方作矩阵乘积运算，计算量最大的情况是：一个候选频繁项集包含所有参与方的私有属性，这种情况下第三方最多需要做 m 个 $n\times n$ 矩阵的乘积来计算项集支持度，复杂度 $O(mn^3)$，这部分是系统主要开销。为降低计算复杂度，可以将 n 维向量分成[n/k]个子向量，根据以上协议，只需用 $k\times k$ 可逆方阵，计算复杂度降为 $O(mnk^2)$，通信开销降为 $O(mnk)$位，但是完成一次计算的通信轮数由原来的 1 增加到[n/k]，增加了这部分的通信开销。通过调整 k 值，使总的系统开销达到最小。

3.4.3 安全性分析

第三方可以从所有站点获得加密后的数据，由于它不知道加密方阵 $\boldsymbol{Q}$，也不能从已知向量推断出对向量进行加密的方阵 $\boldsymbol{Q}$，因此不能得知各站点的隐私数据。而其他各方互相不交换加密数据，因此也不能获得其他站点的隐私数据。由于第三方只做计算，不考虑隐私安全问题，它将向各方返回候选频繁集的支持度，如果某一方或几方拥有不愿泄漏的项集信息，它要在挖掘之前，降低项集支持度，使之小于最小支持度要求，这样在联合数据挖掘过程中就不会被挖掘出来，这种情况的研究，本文将不做深入讨论。

4 结束语

本文讨论在分布式环境下基于隐私保护考虑的关联规则挖掘问题，提出一种简单高效的两方协议，但从隐私安全角度来看，这种协议不能应用到安全要求较高的背景中。随后提出一种基于可逆方阵的两方和多方安全协议，基于这种协议可以设计高效的挖掘算法，在保护隐私数据的前提下，尽可能提高挖掘的效率，这将是下一步工作的主要任务。

参考文献

[1] Vaidya J, Clifton C W. Privacy Preserving Association Rule Mining in Vertically Partitioned Data[C]//Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining. Alberta, Canada: [s. n.], 2002: 639-644.

[2] Beaver D. Commodity-based Cryptography(Extended Abstract)[C]// Proceedings of the 29th Annual ACM Symposium on Theory of Computing. TX, USA: [s. n.], 1997: 217-221.

[3] Beaver D. Server-assisted Cryptography[C]//Proceedings of the 1998 New Security Paradigms Workshop. Charlottesville, VA, USA: [s. n.], 1998: 92-106.

[4] Yao A C. How to Generate and Exchange Secrets[C]//Proceedings of the 27th IEEE Symposium on Foundations of Computer Science. Los Alamitos, USA: [s. n.], 1986: 162-167.

[5] Goldreich O, Micali S. How to Play any Mental Game——A Completeness Theorem for Protocols with Honest Majority[C]// Proceedings of the 19th ACM Symposium on the Theory of Computing. New York, USA: [s. n.], 1987: 218-229.

[6] Atallah M J. Secure Multi-party Computational Geometry[C]// Proceedings of the 7th International Workshop on Algorithms and Data Structures. Providence, Rhode Island, USA: [s. n.], 2001: 165-179.

编辑　索书志