第32卷 第13期 计 算 机 工 程 2006年7月
Vol.32 № 13 Computer Engineering July 2006
·博士论文· 文章编号：1000—3428(2006)13—0012—03 文献标识码：A 中图分类号：TP391

分布式环境下保持隐私的关联规则挖掘算法

黄毅群，卢正鼎，胡和平，李瑞轩
（华中科技大学计算机学院，武汉 430074）

摘 要：保持隐私是未来数据挖掘领域的焦点问题之一，如何在不共享精确数据的条件下，获取准确的数据关系是保持隐私的数据挖掘的首要任务。该文介绍了分布式环境下保持隐私的数据挖掘的基本问题和措施，研究了一种基于向量点积的关联规则挖掘算法，给出了一种安全的向量点积协议。对于垂直划分的分布式数据库，该协议既可用于搜索频繁项集，又能保持各方数据的隐私。

关键词：保持隐私；分布式数据挖掘；关联规则；频繁项集；点积

Privacy Preserving Distributed Data Mining Association Rules of Frequent Itemsets

HUANG Yiqun, LU Zhengding, HU Heping, LI Ruixuan
(School of Computer Science & Technology, Huazhong University and Technology, Wuhan 430074)

【Abstract】 There has been growing interests in private concerns for future data mining research. Privacy preserving data mining concentrates on developing accurate models without sharing precise individual data records. This paper addresses basic ideas and solutions for secure data mining over distributed data. An algorithm based on dot product for distributed mining association rules is presented. It also gives a protocol of secure dot product computation which is effective to discover frequent itemsets on vertically partitioned data. It can provide good data privacy.

【Key words】 Privacy preserving; Distributed data mining; Association rules; Frequent itemsets; Dot product

数据挖掘的任务是从大量数据中提取隐含的、未知的、潜在有用的知识。然而，这些数据通常分布存储在多个不同的站点中，传统的方法需要将所有的数据存放在一个集中的数据仓库或集市中，因此不能保障数据的私有性和保密性。

随着人们对隐私的广泛重视，数据的集中处理备受争议。有调查显示[1]，82%的受访网络用户非常重视隐私保护措施；仅有14%的人认为只要有收益，隐私保持无关紧要。在日本，公众反对将过去分散在各辖区内的户籍信息进行全国性汇总。在美国，为了防止有关机密信息的泄露，国会引入了“2003年度暂停数据挖掘法案”，该法案将会禁止美国国防部对数据挖掘进行研究和开发；除此之外，所有美国的政府机构在开展数据挖掘项目之前都必须向国会通报其保护隐私的措施。

然而，大部分的私人信息仍然存放在各种组织和政府机构之中，对数据挖掘的限制仅仅阻止了数据的进一步集中，却妨碍了用于正当目的的数据挖掘任务，如研究流行性疾病的模式、进行多国合作等。因此，未来数据挖掘的焦点之一将是考虑隐私保护问题[2]。

由于数据挖掘算法需要精确的数据，而数据的隐私性却要求对确切的数据进行保密，因此隐私与数据挖掘就成为一对矛盾。分布式环境下如何能够既在不共享精确数据，又获取准确的数据关系就成为保持隐私的数据挖掘的首要任务。

1 保持隐私的数据挖掘

1.1 隐私的定义

隐私，是指“一种免受非法侵入的权利”。通常人们担心私有信息被滥用。对数据挖掘而言，隐私又分为以下两种[3]：

(1)“单个隐私”，即可追溯到个人的数据的私有性。一旦数据挖掘所使用的数据是个人可识别的，就会引起争议。

(2)“联合隐私”，即数据集合所揭示的信息的私有性。它通常指数据挖掘结果本身是保密的。

因此，“单个隐私”针对的是指数据挖掘的各方输入数据，“联合隐私”讨论的是数据挖掘的输出模型。只要改变数据的收集方式，屏蔽敏感的模型结果就可以实现保持隐私的数据挖掘。

1.2 保持隐私的数据挖掘

保持隐私的数据挖掘需要结合以下几方面来讨论：

(1)数据挖掘结果，即开采模型或模式，如分类树、关联规则、聚类等。

(2)分布式。对于不同的数据划分形式，分为水平和垂直两种。

(3)隐私限制，即不同的隐私类型。

对于每一种挖掘任务、不同的数据分布方式和隐私限制，都需要不同的解决方案。目前，在分布式环境下，保持隐私的数据挖掘有两种方式：基于数据干扰技术的集中式数据挖掘和基于多方安全计算的分布式数据挖掘。

数据干扰技术的主要思想是：首先加密各站点的数据，使其不可辨认；然后将修改后的数据存放到数据仓库中进行统一开采。方法有：数据交换方法和数据随机化等。前者通过交换不同记录之间的数值来隐藏记录所属对象与数值间的对应关系；后者在保持数据的原始分布不变的条件下给原记

基金项目：国家自然科学基金资助项目（60403027）
作者简介：黄毅群(1970—)，女，讲师、博士生，主研方向：数据挖掘，信息安全及隐私，应用密码学和SMC技术；卢正鼎，教授、博导；胡和平，教授；李瑞轩，副教授
收稿日期：2005-08-17 **E-mail：**yqhuang2005@163.com

录加入适量的随机噪声。然而，这种技术的主要问题仍然是从变化了的数据中能否得出正确的模型。

多方安全计算是基于如下设想：进行协同计算的各方即不信任其它任何一方，也不信任彼此之间的通信渠道；但是，它们都希望在保持隐私的条件下，通过传递必要的公用信息，获取计算结果。最初，YAO 于 1986 年提出了两方安全计算；随后 Goldreich 将其推广为对于任何函数都成立的多方安全计算方法，他指出安全地计算即隐私地计算[4]。

在分布式数据挖掘中，采用多方安全计算来保持隐私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。

2 基于向量点积的分布式关联规则挖掘

2.1 基本概念

定义 1 关联规则可描述如下：设 I={i1, i2,…, im}是项的集合；DB 是一个事务数据库，其中每个事务 T 是项的集合，存在 T⊆I，每个事务有一个标识符 TID。一个项目的集合称为项集，在一个项集中项目的数量称为项集的长度，一个长度为 k 的项集称为 k-项集。设 X 为一个项集。如果项集 X⊆T 且 X⊆I，则称事务 T 包含 X。关联规则是形如 $X \Rightarrow Y$ 的蕴涵式，其中 X⊆T，Y⊆T，且 $X \cap Y = \phi$。如果 D 中包含 X 的 c%的事务同时也包含 Y，那么规则 $X \Rightarrow Y$ 在事务集 DB 中有置信度 c（confidence）；如果 DB 中 s%的事务包含 $X \cup Y$，那么规则 $X \Rightarrow Y$ 在事务集 D 中有支持度 s（support）。

关联规则的挖掘就是要从数据库 DB 中找到具有用户给定的最小支持度（min_sup）和最小置信度（min_conf）的强关联规则，其中的关键又是计算项集的支持度。

定义 2 针对数据的不同划分方式，分布式数据库可分为垂直划分和水平划分两种。设有两个垂直划分的分布式事务数据库 A 和 B，每个交易被分割在两个站点中，其中 A 有 n 条事务和 m 个属性，B 有 n 条事务和 l 个属性。

对于每个事务 T，若用 1 代表属性的出现，0 代表不出现，那么数据库 A 和 B 可分别表示为一个 $n \times m$ 和 $n \times l$ 的布尔型矩阵 $A_{n\times m}$、$B_{n\times l}$。计算各项集的支持度 s 就转化为统计该项集中所有属性（项）为 1 的个数。

定义 3 设向量 $\vec{X}$ 和 $\vec{Y}$，$\vec{X}=(x_1,\ldots,x_n)$，$\vec{Y}=(y_1,\ldots,y_n)$，向量 $\vec{X}$ 和 $\vec{Y}$ 的点积定义为：$\vec{X}\cdot\vec{Y}=\sum_{i=1}^{n} x_i * y_i$。

性质 1 若用向量 $\vec{X}=(x_1,\ldots,x_n)$和 $\vec{Y}=(y_1,\ldots,y_n)$分别代表 $A_{n\times m}$、$B_{n\times l}$ 中的任一属性列，其中，$x_i=1$ 或 $y_i=1$，表明数据库 A 或 B 中的第 i 行的该属性值为 1；否则为 0；则点积 $\vec{X}\cdot\vec{Y}\geq$min_sup可判断 2-项集<x, y>是否频繁。

性质 2 若用向量 $\vec{X}=(x_1,\ldots,x_n)$和 $\vec{Y}=(y_1,\ldots,y_n)$分别构造 $A_{n\times m}$、$B_{n\times l}$ 中的内部项集，其中，$x_i=\prod_{j=1}^{s} a_{ij}$，$s\in[1,m]$，$y_i=\prod_{j=1}^{t} b_{ij}$，$t\in[1,l]$，$a_{ij}$、$b_{ij}$ 代表 $A_{n\times m}$、$B_{n\times l}$ 中的任一元素；则判断 $\sum_{i=1}^{n} x_i \geq$min_sup和 $\sum_{i=1}^{n} y_i \geq$min_sup可分别搜寻出 A、B 内部的所有频繁项集；判断 $\vec{X}\cdot\vec{Y}\geq$min_sup可搜寻出 A、B 间的所有频繁项集，<x, y>为(s+t)-项集。

2.2 基于安全点积的频繁项集挖掘算法

对于垂直分割的分布式数据库 A 和 B，基于性质 2 来搜索频繁项集，算法主要分为以下几步：

(1)由 $x_p=\sum_{j=1}^{n} a_j \geq$ min_ sup 和 $y_q=\sum_{j=1}^{n} b_j \geq$ min_ sup，$p\in[1,m]$，$q\in[1,\ell]$生成 A、B 中的频繁 1-项集 L_{1_A}、L_{1_B}；

(2)分别用向量 $\vec{X}$ 和 $\vec{Y}$ 构造 A 和 B 内部的所有项集，其中 $x_i=\prod_{j=1}^{s} a_{ij}$，$s\in[1,m]$，$i\in[1,m]$；$y_i=\prod_{j=1}^{t} b_{ij}$，$t\in[1,l]$，$i\in[1,\ell]$；并且 $a_{ij}\in L_{(k-1)A}$ 和 $b_{ij}\in L_{(k-1)B}$。根据 $\sum_{i=1}^{n} x_i \geq$min_sup和 $\sum_{i=1}^{n} y_i \geq$min_sup生成 A、B 内部的频繁项集 L_{k_A}和 L_{k_B}。

(3)由安全点积协议判断 $\vec{X}\cdot\vec{Y}\geq$ min_sup 生成 A、B 间的频繁项集 L_{k_AB}。

2.3 安全的两方点积协议及其分析

由于 A、B 双方各自拥有保密向量 $\vec{X}$ 和 $\vec{Y}$，为了保持它们的隐私，基于多方安全计算的思想来计算点积 $\vec{X}\cdot\vec{Y}$，协议如下式：

$$\begin{aligned}\vec{X}\cdot\vec{Y}=\sum_{i=1}^{n} x_i * y_i &= \left[\sum_{i=1}^{n}\left(x_i+\sum_{j=1}^{n} c_{i,j} * R_j\right)_A * y_i\right]_B \\ &- \left[\sum_{j=1}^{n} R_j * \left(\sum_{i=1}^{n} c_{i,j} * y_i + R'_{1+\mathrm{int}\left[\frac{(j-1)}{n/r}\right]}\right)_B\right]_A \\ &+ \left[\sum_{i=1}^{r}\left(\sum_{j=1}^{\frac{n}{r}} R_{\left[(i-1)*\frac{n}{r}+j\right]}\right)_A * R_i'\right]_B \end{aligned} \quad (1)$$

式(1)中，A 方生成随机保密向量 $\vec{R}=(R_1,\ldots,R_n)$ 以隐藏向量 $\vec{X}$；B 方生成随机保密向量 $\vec{R}'=(R_1',\ldots,R_r')$ 来隐藏向量 $\vec{Y}$。A 和 B 共享系数矩阵 C（即 $c_{i,j}$，$i,j\in[1,n]$）。角标 A 和 B 代表相关部分由 A 方和 B 方分别计算。算法分为以下几步：

(1) A 产生随机向量 $\vec{R}=(R_1,\ldots,R_n)$；B 产生随机向量 $\vec{R}'=(R_1',\ldots,R_r')$。

(2) B 将向量 $\vec{R}'$从 r 个分量扩充为 n 个分量，即向量 $\vec{R}'_{r\to n}$，其中 $(R'_{r\to n})_i = R'_{1+\mathrm{int}\left[\frac{(i-1)}{n/r}\right]}$；

另外，B 计算向量 $\vec{Y}'=(y_1',\ldots,y_n')$，其中 $Y_j'=\sum_{i=1}^{n} c_{i,j} * y_i + (R'_{r\to n})_i$。将向量$\vec{Y}'$和 r 传给 A。

(3) A 计算向量 $\vec{X}'=(x_1',\ldots,x_n')$，其中 $X_i' = x_i + \sum_{j=1}^{n} c_{i,j} * R_j$。

另外，A 将向量 $\vec{R}$ 分成 r 组，求各组的和以形成新的向量 $\vec{R}_{n\to r}$，其中 $(R_{n\to r})_i = \sum_{j=1}^{\frac{n}{r}} R_{\left[(i-1)*\frac{n}{r}+j\right]}$；

计算点积 $\vec{R}\cdot\vec{Y}'$，并将向量 $\vec{X}'$、$\vec{R}_{n\to r}$ 和 $\vec{R}\cdot\vec{Y}'$结果传给 B。

（4）B 分别计算点积 $\vec{R}'\cdot\vec{R}_{n\to r}$、$\vec{X}'\cdot\vec{Y}$ 和最终结果 $\vec{X}\cdot\vec{Y}=\vec{X}'\cdot\vec{Y}-\vec{R}\cdot\vec{Y}'+\vec{R}_{n\to r}\cdot\vec{R}'$，将结果传给 A。

上述协议的安全性是基于如下原则：n 个未知数无法从 n 个以下的方程中求得。

式(1)中，A 向 B 公开的方程包括：

$$\begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_{n-1}' \\ x_n' \end{pmatrix}_{n\times 1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}_{n\times 1} + \begin{pmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,n} \end{pmatrix}_{n\times n} \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_{n-1} \\ R_n \end{pmatrix}_{n\times 1} \quad (2)$$

$$\begin{pmatrix}(R_{n\to r})_1\\(R_{n\to r})_2\\\vdots\\\vdots\\(R_{n\to r})_r\end{pmatrix}_{r\times 1}=\begin{pmatrix}\overbrace{1\ \cdots\ 10\ \cdots\ 0\,0\ \cdots\ 00\ \cdots\ 0}^{\frac{n}{r}}\ \overbrace{\cdots}^{\frac{n}{r}}\ \overbrace{\cdots}^{\frac{n}{r}}\\0\ \cdots\ 01\ \cdots\ 1\,0\ \cdots\ 00\ \cdots\ 0\\\vdots\ \ddots\ \vdots 0\ \ddots\ \ddots\ \ddots\ \vdots\vdots\ \ddots\ \vdots\\0\ \cdots\ 00\ \cdots\ 0\,0\ \cdots\ 01\ \cdots\ 1\end{pmatrix}_{r\times n}\begin{pmatrix}R_1\\\vdots\\R_{\frac{n}{r}}\\R_{\frac{n}{r}+1}\\\vdots\\R_{\frac{2n}{r}}\\\vdots\\R_n\end{pmatrix}_{n\times 1}\tag{3}$$

B 向 A 公开的方程包括：

$$\begin{pmatrix}y_1'\\y_2'\\\vdots\\y_{n-1}'\\y_n'\end{pmatrix}_{n\times 1}=\begin{pmatrix}c_{1,1}&\cdots&c_{n,1}\\\vdots&\ddots&\vdots\\c_{1,n}&\cdots&c_{n,n}\end{pmatrix}_{n\times n}\begin{pmatrix}y_1\\y_2\\\vdots\\y_{n-1}\\y_n\end{pmatrix}_{n\times 1}+\begin{pmatrix}R_1'\\\vdots\\R_1'\\R_2'\\\vdots\\R_2'\\\vdots\\R_r'\end{pmatrix}_{n\times 1}\tag{4}$$

因此，为了保持 A、B 双方的数据隐私，必须满足以下关系：(1)为了使各方程成为线性独立的方程，对于公共的系数矩阵 C，其秩为 n。(2) n<n+r<2n。当 r=n/2 时，A、B 具有相同的安全性和隐私性。

另外，A、B 间共需传递 2n+r+2 个数值，完成 $2n^2+2n+r$ 次加法和乘法，通信代价和计算开销分别为 $o(n)$ 和 $o(n^2)$。由于是简单的算术运算，因此该算法具有合理的通信代价和计算开销。如果要求 A、B 具有相同的隐私性，只有当 1/2 的数据被破解时，所有数据才会被泄露。

可将式(1)简化为

$$\begin{aligned}&\vec{X}\cdot\vec{Y}=\sum_{i=1}^{n}x_i*y_i\\&=\left[\sum_{i=1}^{n}\left(x_i+\sum_{j=1}^{r}c_{i,j}*R_j\right)_A*y_i\right]_B\\&-\left[\sum_{j=1}^{r}R_j*\left(\sum_{i=1}^{n}c_{i,j}*y_i\right)_B\right]_A\end{aligned}\tag{5}$$

其中，A 向 B 公开的方程包括：

$$\begin{pmatrix}x_1'\\x_2'\\\vdots\\x_{n-1}'\\x_n'\end{pmatrix}_{n\times 1}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_{n-1}\\x_n\end{pmatrix}_{n\times 1}+\begin{pmatrix}c_{1,1}&\cdots&c_{1,r}\\\vdots&\ddots&\vdots\\c_{n,1}&\cdots&c_{n,r}\end{pmatrix}_{n\times r}\begin{pmatrix}R_1\\R_2\\\vdots\\R_{r-1}\\R_r\end{pmatrix}_{r\times 1}\tag{6}$$

B 向 A 公开的方程包括：

$$\begin{pmatrix}y'_1\\y'_2\\\vdots\\y'_{r-1}\\y'_r\end{pmatrix}_{r\times 1}=\begin{pmatrix}c_{1,1}&\cdots&c_{n,1}\\\vdots&\ddots&\vdots\\c_{1,r}&\cdots&c_{n,r}\end{pmatrix}_{r\times n}\begin{pmatrix}y_1\\y_2\\\vdots\\y_{n-1}\\y_n\end{pmatrix}_{n\times 1}\tag{7}$$

同样，为了保持 A、B 双方的数据隐私，必须满足以下关系：

(1) 对于公共的系数矩阵 C，其秩为 r。

(2)r <n <n+r。当 r=n/2 时，A、B 具有相同的安全性和隐私性。

由于式(5)只需要产生一个随机向量 $\vec{R}$，并且不需要将 $\vec{R}$ 的每 n/ r 个分量合为一组，因此降低了通信代价和计算开销。

3 结论

随着网络、存储和处理器等技术的飞速发展，数据库的存放趋于分布式。为了适应分布式数据挖掘对数据的安全性和隐私性的要求，基于隐私保持的分布式数据挖掘正成为新的研究方向。本文总结了该领域的有关概念及相关技术，讨论了一种基于向量点积的分布式关联规则挖掘算法，给出了一种安全的向量点积协议。对垂直划分的分布式数据库，该协议既可用于搜索频繁项集，又能保持各方数据的隐私。

参考文献

1 Crannor L F, Reagle J, Ackerman M S. Beyond Concern: Understanding Net Users' Attitudes About Online Privacy[R]. Technical Report TR 99.4.3, AT&T Labs-Research, 1999-04.

2 Agrawal R, Srikant R. Privacy-preserving Data Mining[C]. Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, 2000: 439-450.

3 Clifton C, Kantarcioglu M, Vaidya J. Defining Privacy for Data Mining[C]. National Science Foundation Workshop on Next Generation Data Mining, 2002: 126-133.

4 Goldreich O. The Foundations of Cryptography[M]. General Cryptographic Protocols, Cambridge University Press, 2004-05.

5 Vaidya J, Clifton C. Privacy Preserving Association Rule Mining in Vertically Partitioned Data[C]. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002: 639-644.

（上接第 11 页）

参考文献

1 Bribiesca E. A New Chain Code[J]. Pattern Recognition, 1999, 32(2): 235-251.

2 Bribiesca E. A Chain Code for Representing 3D Curves[J]. Pattern Recognition, 1999, 33(5): 765-765.

3 李星原, 高　文. 一种鲁棒性的结构未知表格分析方法[J]. 软件学报, 1999, 10(11): 1216-1224.

4 IllingworthJ, Kittler J. A Survey of the Hough Transform[J]. Computer Vision, Graphics, and Image Processing, 1998, 44(1): 87.

5 Liu W Y. From Raster to Vectors: Extracting Visual Information from Line Drawing[J]. Pattern Analysis and Application, 1999, 2(1): 11-21.

6 郑冶枫, 刘长松, 丁晓青等. 基于有向单连通链的表格框线检测算法[J]. 软件学报, 2002, 13(4): 790-796.

7 顾国庆, 许彦冰. 数字图像区域标定的方法[J]. 上海理工大学学报, 2001, 23(4): 295-299.

8 张圣希, 张　薇, 李国强等. 利用顶点链编码探测表格的斜率[J]. 华东师范大学学报, 2004, 3(2): 54-58.