Purpose 融合:基于风险 purpose 的隐私查询访问控制

刘逸敏[1),2)] 周浩峰[1)] 王智慧[1)] 汪 卫[1)]

[1)](复旦大学计算机科学技术学院 上海 200433)

[2)](第二军医大学第三附属医院信息科 上海 200438)

摘 要 基于 purpose 的查询技术是关系数据库对隐私数据访问控制的基础,目前大多数研究都仅仅关注在独立隐私保护数据库环境下,如何建立有效的基于 purpose 的隐私数据访问控制模型.但随着分布应用整合和数据共享需求的日益增长,如何合并独立应用下基于 purpose 访问控制模型的问题就应运而生.为解决这一问题,文章提出了多应用系统集成环境下基于融合 purpose 的隐私数据访问控制实现机制.文章首先分析了由于合并独立 purpose 模型而引起的潜在隐私数据查询泄漏,提出了合并后的 purpose 树是一棵隐私泄漏风险树,并给出了树结点隐私风险度的计算模型.其次,将隐私泄漏风险树分解成一棵所有结点风险度为 0 的风险平衡树以及一组由风险度不为 0 的结点组成的风险路径.这样,一个查询可被改写为先对风险平衡 purpose 树的查询,再对风险路径查询,以此达到隐私泄漏风险最小的安全查询结果.文章给出了 3 组实验的结果:(1)对于同一用户同一查询,在不同 purpose 模型下的查询时间比较.文章提出的模型并没有在查询时间上带来更大的开销.(2)对 RPPAAC 模型防止隐私数据泄漏的有效性验证.文中的模型可降低由应用整合引起访问控制机制不平衡所带来的隐私数据泄漏风险.(3)不同情况下 purpose 融合的执行时间比较.文章与相关研究的不同之处是将 purpose 作为隐私数据的载体、purpose 树的路径作为隐私数据的传递通道,引入了显性隐私度和隐性隐私度的计算模型,评估基于一个 purpose 查询可能带来的隐私泄漏风险,进而提出了多应用系统集成环境下基于融合 purpose 的隐私数据访问控制实现机制.

关键词 隐私;隐私保护数据库;purpose;隐私度;访问控制

中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2010.01339

Purpose Fusion: The Risk Purpose Based Privacy-Aware Data Access Control

LIU Yi-Min[1),2)] ZHOU Hao-Feng[1)] WANG Zhi-Hui[1)] WANG Wei[1)]

[1)](*School of Computer Science, Fudan University, Shanghai* 200433)

[2)](*Information Department of the Third Affiliated Hospital, The Second Military Medical University, Shanghai* 200438)

Abstract The purpose based query technology is the basis of the privacy-aware data access control in relational databases. Most researches focus on how to effectively build a purpose based privacy-aware data access control mechanism for an independent privacy-preserving database system. However, with the popularity of application integration and data sharing, how to merge the purpose based access control mechanisms in different applications and systems becomes a key issue. To address the problem, this paper presents the purpose fusion based privacy-aware data access control mechanism for the integration of multiple applications and systems. It analyzes the potential leakage risks of privacy-aware data due to the fusion of multiple purposes, and evaluates the leakage risks of nodes by considering a merged purpose tree as the risk purpose tree. Then, it split the risk purpose tree into a risk balanced purpose tree with the privacy degree of 0 for all nodes, and a set of risking paths containing the nodes with non-zero privacy degrees. Therefore, a query can be answered by checking the risk balanced purpose tree and then the risking paths,

收稿日期:2010-06-11.本课题得到高等学校博士学科点专项科研基金(200802461146)、国家科技重大专项(2008ZX10002-018)资助.**刘逸敏**,女,1963 年生,博士研究生,高级工程师,主要研究方向为数据库、隐私保护和数据挖掘. E-mail: liuyiminzsh@yahoo.com.cn. **周皓峰**,男,1975 年生,博士,副研究员,主要研究方向为隐私保护、数据挖掘、数据库、软件质量和测试等. **王智慧**,男,1975 年生,博士,讲师,主要研究方向为数据安全与隐私保护、数据挖掘和数据库系统等. **汪 卫**(通信作者),男,1970 年生,博士,教授,主要研究领域为安全数据库、数据挖掘. E-mail: weiwang1@fudan.edu.cn.

thus safe query results can be obtained with minimized privacy leakage risks. Three sets of experimental results have been given in this paper: (1) the query time comparison between different purpose based models for a same user and query; The RPPAAC model presented in this paper does not lead to a larger time overhead; (2) validity checking for the RPPAAC model in terms of the disclosure of private data; (3) the comparison of execution time for purpose fusion in different instances. Different from related works, this paper considers purposes as the carrier of privacy-aware data, the paths of purpose tree as transmission channels of privacy-aware data, and by introducing the public risk and the hidden risk, it evaluates the potential leakage risks of privacy-ware data during query answering, and presents the purpose fusion based privacy-aware data access control mechanism for integrating multiple applications and systems.

Keywords privacy; privacy-preserving database; purpose; privacy score; access control

1 引 言

信息网络技术的快速发展提供了许多基于网络平台的增值服务,如医疗机构间电子健康档案的共享和网络在线预定服务等,然而,用户在享受便捷的同时,个人隐私数据泄漏的风险也随之不断增加,因此隐私数据访问控制技术的研究与应用迫在眉睫.

自 P3P(The W3C's Platform for Privacy Preference)[1]和 OECD[2]对隐私数据访问规范发布以来,出现了许多对基于 purpose 的隐私数据访问控制模型的研究,并在不同领域中得到应用,如在关系数据库中以树结构描述 data purpose 的 generalization/specialization 特性,访问控制模型以查询用户的 access purpose 是否与隐私数据的 data purpose 匹配机制去实现[3-9];还有些控制模型借鉴了安全数据库用户访问等级的概念,将 purpose 作为数据访问的安全等级,分别对应不同的泛化的隐私数据,该控制机制是同一用户在不同的 purpose 等级下,查询不同的隐私数据[6,10-11];第三类模型是一项业务由可选的多组 purpose 组成,那么完成一项业务的最小隐私数据泄漏是找出一组 purpose,使基于该组 purpose 的可访问隐私数据最小[12-13];还有些模型拓展了传统的基于角色的访问控制机制,对角色的授权从原先的对数据库对象(关系)的访问扩展到基于不同 purpose 下的对不同隐私数据(关系中的属性)的访问[14-16].但随着分布应用的整合和数据共享的需求的增加,如何把原先不同应用已建立的基于 purpose 访问控制模型进行合并的问题就应运而生,并逐渐成为一个亟待解决的问题,而已有的模型大都没涉及如何解决该类问题.

本文基于对隐私数据的最大查询结果和最小查询隐私泄漏需求[9,17-19],研究了分布应用被整合环境下由于合并独立 purpose 模型而引起的潜在隐私数据查询泄漏问题,提出了合并后的 purpose 树是一棵隐私泄漏风险树,并给出了结点隐私风险度的计算模型.基于风险 purpose 的隐私查询访问控制模型的基本思想是将隐私泄漏风险树分解成一棵平衡树和一组风险路径,一个查询被改写为先对平衡 purpose 树的查询,若结点隐私度不为 0,再对风险路径查询,以此达到隐私泄漏风险最小的安全查询结果.

本文的主要贡献是:(1)提出了计算由于相同 purpose 结点合并而可能引起的该结点访问隐私属性增加的隐私度以及计算由于合并后存在路径传递而引起的隐私属性增加的隐私度的方法.(2)提出分裂风险结点、建立一棵隐私度为 0 的平衡 purpose 树以及存储风险路径算法,使控制模型的查询能返回隐私泄漏风险最小的安全查询结果.(3)实验表明,本文提出的隐私数据库访问控制模型能有效减少由应用整合引起访问控制机制不平衡所带来的隐私数据泄漏风险.

本文第 2 节是相关工作;第 3 节详细介绍融合 purpose 树结点隐私度的计算模型和基于风险 purpose 查询的隐私数据库访问控制模型;第 4 节是模型实验和讨论;最后第 5 节总结本文工作并给出展望.

2 相关工作

与本文研究内容相关的领域有隐私保护数据库、访问控制技术和风险评估.本文对隐私数据与隐私策略的定义遵循了 P3P[1]的工业标准,将 purpose 作为隐私数据访问与管理的元素,并且 purpose 具

有层次特性. 本文涉及的 purpose 树融合，是以 HDB[4] 提出的 purpose 作为隐私数据标签、以 purpose 树定义隐私策略[5] 为基础，假设在隐私数据库中已存在了描述隐私访问规则的 purpose 树.

在对 purpose 层次定义方面本文与原先工作相同的是 purpose 树是一棵 Gengealization/Specilization 树[3,5-6,10-11]，不同的是本文进一步定义了 purpose 树结点之间隐私数据访问的传递关系. 在现有工作中将基于 purpose 的访问归结为对元数据表的链接操作，但元数据表属性的设置没有反映 purpose 的层次结构，会引起元数据的冗余. 另一方面现有元数据模式的设计没有考虑不同应用之间元数据的整合，这样就不能解决基于多 purpose 融合查询的访问控制模型合并问题. 有些工作[12-13,15] 与本文定义的 purpose 层次概念不同，它们把 purpose 树看成是一项任务的分解，路径构成了完成任务的每个节点，隐私控制就是找出一条路径，使路径上可访问的隐私数据权重值最小.

在访问控制策略研究方面，现有的工作[5,14,16] 大都是在单应用环境下基于 purpose 树的查询控制，但一旦在多应用环境下 purpose 树结构变化，现有的控制机制就有隐私数据查询泄漏的风险. 本文在给出 purpose 隐私度的计算模型方面，基于了社会网络评估用户风险的概念[20-22]，在现有的工作中都将用户的配置信息作为用户的隐私数据，通过评估配置信息的隐私程度来给出用户在社会网络下的风险度. 这些工作与本文计算隐私度的目的不同，前者是评估风险，而本文是通过隐私度计算去实现数据库的访问控制机制.

3　基于风险 purpose 的隐私查询访问控制

当二棵独立 purpose 树合并时，由于相同结点的合并导致 purpose 下可访问隐私属性的增加，打破了原先独立 purpose 树的风险平衡，合并后的 purpose 树(Combined Purpose Tree，CPT)是一棵有隐私数据泄漏风险的 purpose 树，所以需要重建基于 CPT 查询的访问控制模型. 本节先以实例说明基于 CPT 查询的隐私泄漏过程，提出风险 purpose 树结点隐私度的计算模型；其次定义融合 purpose 树(Joint Purpose Tree，JPT)为一棵结点带有隐私度的 CPT，并给出算法分裂隐私度不为 0 的结点，将 JPT 模型分解成一棵平衡树和一组风险路径；最后基于 JPT 的查询改变为先对平衡 purpose 树的查询，若结点隐私度不为 0，再对风险路径查询，以此来实现基于风险 purpose 的隐私查询访问控制(Risk Purpose Based Privacy-Aware Access Control，RPPAAC).

3.1　问题描述

设已知应用 $s1$ 和 $s2$ 的隐私数据库 purpose 模型 $pt1$、$pt2$ 和 purpose 的元数据如图 1 和图 2 所示，表示 $user1$ 和 $user2$ 分别基于 $pt1$ 和 $pt2$ 的 purpose 查询允许访问的隐私数据. Purpose 树中有部分结点相同，但相同 purpose 结点下可访问的隐私数据不完全相等，根据 purpose 模型的 generalization/specilization 的特性[3-5,13,19]，使 $pt1$ 和 $pt2$ 基于 $purpose=p_3$ 的访问有如下性质('∷'表示等同于)：

$$pt1(purpose=p_3)\text{∷}pt1(purpose=(p_3,p_4,p_5))$$

$$pt2(purpose=p_3)\text{∷}pt2(purpose=(p_3,p_{4'}))$$

Tname	*attr*	*purpose*
*T*1	*p_id*	p_0
	orders	p_1
	drug	p_2
	result	p_3
	operation	p_4
	disease	p_5

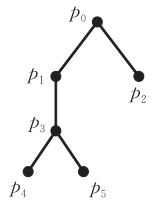

图 1　$pt1$ 的元数据和 purpose 树

Tname	*attr*	*purpose*
*T*2	*p_id*	p_0
	Ins_type	$p_{1'}$
	mobile	p_3
*T*1	*result*	p_3
*T*2	*msn*	$p_{4'}$
	mailbox	p_5
*T*1	*history*	p_5

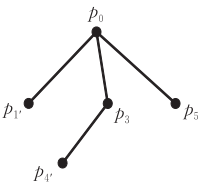

图 2　$pt2$ 的元数据和 purpose 树

如其中基于 $pt2$ 下 $purpose=p_3$ 的查询能访问隐私数据{*mobile*，*result*}，同样基于 $pt2$ 的 $purpose=p_{4'}$ 的查询除了可访问数据 msn 外，也能访问{*mobile*，*result*}.

当分布应用 $s1$ 和 $s2$ 被整合时，就需要合并两棵满足不同应用需求的 purpose 树，并且要求基于融合后 purpose 树的查询要控制隐私数据泄漏的风险. Purpose 合并有两种方式，一是通过根结点合并将 $pt2$ 作为 $pt1$ 的子树，这种方式由于直接合并导致存在重复的 purpose 结点，以 purpose 为访问目的的查询会引起歧义. 另一种方式是先将相同 purpose 结点合并生成 $pt3$，当 $pt2$ 有剩余结点时，在 $pt3$ 中找到剩余结点的父结点，并将其作为该父结点的子节点插入 $pt3$，已知图 3 是 $pt1$、$pt2$ 合并后 purpose 模型 $pt3$ 的树结构(p_0(p_1(p_3($p_{4'}$，p_4，

p_5)),p_2,$p_{1'}$)和合并后 $purpose=\{p_0,p_1,p_3,p_5\}$的元数据,虚线圈表示合并过程,从元数据表可发现由于树结构的变化,*user*1 基于合并 *pt*3 的 $purpose=p_3$ 的查询结果是{*result*,*mobile*},而合并前基于 *pt*1 的 $purpose=p_3$ 的查询结果是 *result*,由此会引起隐私数据的查询泄漏.同样 *user*2 基于 *pt*3 的 $purpose=p_5$ 的查询结果{*mailbox*,*history*,*disease*}多于基于 *pt*2 的 $purpose=p_5$ 的查询结果,当考虑了 *purpose* 的 *generalization*/*specilization* 的特性后,基于 *pt*3 的 $purpose=p_5$ 还可访问数据{result,mobile},即存在 $p_3\geq p_5$ 的路径,而在 *pt*2 中不存在这样的访问路径.所以合并后 purpose 树 *pt*3 是一个带有隐私泄漏风险的访问控制模型.

Tname	*purpose*
p_id	p_0
*T*1.*orders*	p_1
…	…
*T*1.*result*,*T*2.*mobile*	p_3
*T*2.*mailbox*,*T*1.*history*,*T*1.*disease*	p_5

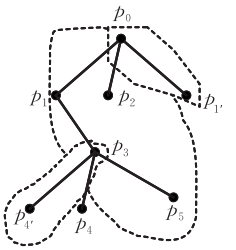

图 3 *pt*3 元数据($purpose=p_0,p_1,p_3,p_5$)和 purpose 树 *pt*3

3.2 purpose 融合

Purpose 融合涉及 purpose 树的合并以及合并后 purpose 结点隐私度的计算.隐私度计算的基本思想是计算由于相同 purpose 结点合并而可能引起该结点访问隐私属性增加的风险度以及计算由于合并后存在路径传递而引起的隐私属性增加的风险度.

3.2.1 隐私度计算模型

定义 1(purpose 树 PT,Purpose Tree). 一个系统的隐私数据访问目的(purpose)用一棵层次 purpose 树 PT 表示,*pi* 和 *pj* 是 PT 的结点,每个结点对应了其可读取的隐私属性集.若 p_i 是 p_j 的父结点,表示 p_i 和 p_j 是泛化和特例(generalization/specilization)的关系,其路径表示基于 p_i 的访问等同于基于$\{p_i,p_j\}$的访问.设$\{ui\}$和$\{vj\}$分别是基于 p_i 和 p_j 可查询的隐私属性集,那么基于 p_j 的可访问数据集是$\{ui\}\cup\{vj\}$,如图 1 应用 *s*1 的 purpose 树是 *pt*1,基于 p_0 和 p_1 可访问的隐私属性分别是{*p*-*id*},{*p*-*id*,*orders*}.

定义 2(合并 purpose 树 CPT,Combined Purpose Tree). 设 *pti*、*ptj*$\in PT$,*pti*[*m*]和 *ptj*[*n*]分别是 *pti* 和 *ptj* 的一个结点,那么 $CPT=pti\cup ptj$.设 $CPT[r]$ 是 CPT 的一个结点,当 $pti[m]=ptj[n]$,那么 $CPT[r]=pti[m]$;基于 $purpose=CPT[r]$ 的查询可访问数据是$\{ui\}\cup\{vj\}$;否则 $CPT[r]=PARENT(pti[m])$.

图 4 是 *pt*2 被并入 *pt*1 的分解图.

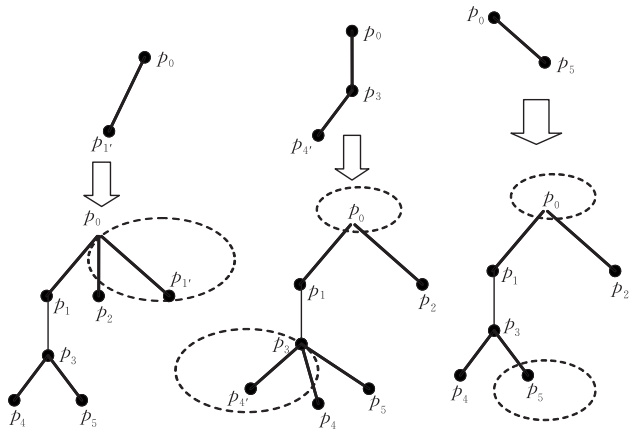

图 4 *pt*1 和 *pt*2 合并过程

定义 3(隐私度,Privacy Risk). Purpose 树结点的隐私度是指基于该结点的查询存在隐私数据泄漏风险的程度.由于 purpose 树结点间有 generalization/specialization 的关系,结点的隐私度不仅要考虑基于结点本身的查询存在隐私泄漏的风险,还要考虑结点所有查询路径上隐私数据泄漏的风险,因此结点隐私度的度量是结点显性隐私度和结点隐性隐私度的函数.隐私度和隐私度计算有如下性质.

性质 1(显性隐私度 *pr*,public risk). 结点的显性隐私度是由于相同 purpose 结点合并而可能引起该结点下可访问隐私属性增加的风险程度,设 *pt*1、*pt*2$\in PT$,p_i 和 p_j 分别是 *pt*1 和 *pt*2 的结点,$\{ui\}$和$\{vj\}$分别是基于 p_i 和 p_j 可查询的隐私属性集,$n=(count|\{ui\}\cup\{vj\}|-count|\{ui\}\cap\{vj\}|)$,表示属性集中不相同隐私属性的数量.当 $pi=pj$ 且二结点合并后的结点是 cpt_l,此时基于 $purpose=cpt_l$ 的查询可访问数据集$\{ui\}\cup\{vj\}$,那么 cpt_l 的显性隐私度 $pr(cpt_l)=n$,若$\{ui\}=\{vj\}$,则 $n=0$.

p_i 的显性隐私度 $pr(p_i)$ 与 p_i 所能访问隐私属性的数量是否增加有关,若增加数量越多,说明隐私数据库用户 *u* 基于 p_i 的查询所能访问的隐私属性就越多,而这些隐私属性中的一部分信息可能已超出了合并前访问控制模型返回的内容范围.由图 3 所示 p_3 的显性隐私度 $pr(p_3)=1$,$pr(p_5)=3$;显然没有合并的结点的显性隐私度 $pr(p_4)=pr(p_{4'})=pr(p_0)=0$.

性质 2(隐性隐私度 *hr*,hidden risk). 结点的隐性隐私度是由于合并后存在路径传递而引起的访问隐私属性增加的风险程度,设 p_i 是 *pt*1 的任意结点,*k* 是结点 p_i 的度,那么 $hr(p_i)=k$.

p_i 隐性隐私度的计算与 p_i 下的子结点数量有关，表示一旦 p_i 可访问隐私数据增加，那么 p_i 所有路径上的子结点，会因层次 purpose 树的继承关系也额外增加了本不可访问的隐私数据，所以 p_i 的隐性隐私度以树结点 p_i 的度来衡量. 图 3 中 $hr(p_3)=3$，说明若 p_3 可读隐私属性增加，那么其 3 个子结点的可读隐私属性也随之增加，加大了数据泄漏的风险，而 $hr(p_5)=0$.

性质 3（隐私度计算 r，Privacy Risk Computation）. purpose 树结点的隐私度是关于 pr 和 hr 的函数 $f(pr,hr)$，设 $p_i\in pt1$ 那么 p_i 的隐私度计算有如下模型，

$$r(p_i)=pr(p_i)\oplus hr(p_i)$$

$$\begin{cases}0, & \text{如果 } pr(p_i)=0\\ pr(p_i)+hr(p_i), & \text{否则}\end{cases}$$

模型表示当且仅当 $pr(p_i)\neq 0$ 时，才存在结点隐私度，因为 $pr(p_i)=0$ 时，表示基于 p_i 可访问隐私数据没有增加，由此也不会影响其子结点可读隐私属性的变化. $r(p_i)$ 越大，表明用户基于 p_i 或 p_i 子结点的查询，其隐私数据泄漏的风险就越高. 图 3 所示 $pt3$ 树中 $r(p_3)=4$，$r(p_5)=3$，而根据定义，模型其余结点的隐私度为 0.

定义 4（风险树 RPT、平衡树 BPT 和融合树 JPT，Risk Purpose Tree、Balance Purpose Tree and Joint Purpose Tree）. 设 $pt1$、$pt2\in PT$，p_i 是 PT 的任意一个结点，$1\leq i\leq n$，n 是 PT 的结点数

RPT 是一棵风险 purpose 树，当且仅当 $\exists p_i$，有 $r(p_i)\neq 0$；

BPT 是一棵平衡 purpose 树，当且仅当 $\forall p_i$，有 $r(p_i)=0$，表明任何用户基于 BPT 树中 purpose 的查询，其返回结果不会造成隐私数据的泄漏；

JPT 是一棵融合树，$Jpt=pt1\cup pt2$，表示融合树 JPT 是一棵已计算结点隐私度的 CPT，且也是一棵风险树.

图 3 所示 $pt3$ 是一棵带有隐私数据泄漏风险的融合 purpose 树.

3.2.2　融合树生成算法

算法 1 为融合 purpose 树生成算法，首先 Combined _purpose()是 purpose 树合并算法，其次 Joint_purpose()是计算合并树的每个结点隐私度，并生成融合 purpose 树.

算法 1.　融合 purpose 树生成算法.

```
Purpose 树合并算法 Combined _purpose()
输入：pt1，pt2 //pt1，pt2 是分布应用 s1 和 s2 的 purpose 树
输出：pt3 //pt3 是一棵合并树，由 pt2 并入 pt1
      //n，m 是 pt1 和 pt2 的结点数，p1[i]和 p2[j]分别是 pt1 和 pt2 的一个结点，该结点下的隐私数据集分别是 D1[i]和 D2[j]，pt3[i]是合并树 pt3 的一个结点，该结点下的隐私数据集是 A[i]
BEGIN
1.  DO WHILE 1≤i≤n
2.    j=1
3.    DO WHILE 1≤j≤m
4.      IF p1[i]=p2[j] THEN //比较 pt1，pt2 结点，若相等则合并隐私数据集
5.        A[i]=D1[i]∪D 2[j]，pt3[i]=pt1[i]
6.        break
      //否则比较 pt2 父结点 parent[p2[j]]，若匹配，那么 p2[j]作为 p3[i]的子结点
7.      ELSEIF p1[i]=Parent[p2[j]] THEN
8.        Child[p3[i]]=p2[j]
9.        break
10.     ENDIF
11.     j=j+1
12.   ENDDO
13.   i=i+1
14. ENDDO
END
```

```
Purpose 树融合算法 Joint_purpose()
输入：A[pt1]，A[pt2]//pt1，pt2 树模型的隐私数据表
输出：JPT(pt3) //结点带有隐私度的融合树 pt3，
//显性隐私度、隐性隐私度和隐私度计算公式分别是 pr、hr 和 r，p1[i]、p2[j]和 pt3[k]分别是 pt1、pt2 和 pt3 的一个结点，l 是 pt3 的结点数
BEGIN
1.  Combined _purpose()
      //返回合并树 pt3 及其 pt3 隐私数据集 A[pt3]
2.  DO WHILE 1≤k≤l
3.    DO WHILE 1≤i≤n OR 1≤ j≤m
4.      IF pt3[k]=p1[i] OR pt3[k]=p2[j] THEN
            //在 pt1 或 pt2 中存在匹配 pt3 树的结点
          IF A[pt3[k]]≠A[pt1[i]] OR A[pt3[k]]≠A[pt2[j]] THEN //若结点合并前后的隐私属性集有变化，则计算 pr，hr，r
5.          计算pr(pt3[k])，hr(pt3[k]，r(pt3[k])=pr(pt3[k])+hr(pt3[k])
6.        ELSE pr(pt3[k])=0，r(pt3[k])=0//若结点隐私属性集无变化，那么该结点的隐私度为 0
7.        ENDIF
8.        Break //继续匹配 pt3 的下一个结点
9.      ELSE i=i+1 OR j=j+1
10.     ENDIF
11.   ENDDO
12.   l=l+1
```

```
13.  ENDDO
END
```

3.2.3 应用实例

已知一临床应用 $s1$,用户 *doctor* 可查询隐私关系模式 $T1(p\_id\ \sharp, result, drug, disease, history, orders, operation)$,其访问控制模型如图 1,描述了基于 purpose 的隐私数据访问控制规则,设 $p_3=report$,说明以查阅报告($purpose=report$)为目的的查询,可读取患者的隐私属性{$p\text{-}id, orders, result$}.已知应用 $s2$,用户 manager 可查询隐私关系模式 $T2(p_id\ \sharp, ins_type, m\text{-}phone, msn, mailbox, charges)$为患者提供服务,其访问控制模型如图 2,如新增报告推送功能,同样以 $purpose=report$ 为访问目的,可查询数据{$mobile, result$}.设 *doctor* 和 *manager* 分别发出查询 Q_1 和 Q_2,Q_1:Select * From $T1$ for p_3; $Q2$:Select * From $T1$ for p_5;基于原访问模型规则分别返回结果 $r1=\{p\text{-}id, orders, result\}$和 $r2=\{p\text{-}id, history\}$.但当 $s1$ 和 $s2$ 合并后,$pt1$、$pt2$ 的合成 purpose 模型 $pt3$ 如图 3 所示,设 doctor 发出基于 $pt3$ 的试探查询 $Q1'$:Select * From $T2$ for p_3, 返回 $r1'=\{p\text{-}id, mobile\}$.而{$mobile$}是只有 manager 才可读取的隐私数据.同样 manager 再次发出基于 $pt3$ 的相同查询 $Q2'$: Select * From $T1$ for p_5, 返回 $r2'=\{p\text{-}id, orders, result, disease, history\}$,而其中{$orders, result, disease$}不应是 manager 在 $purpose=p_5$ 下可查询到的隐私数据.可见融合后的 $pt3$ 是带有隐私泄漏风险的访问控制模型,由于 $r(p_3)=4$,$r(p_5)=3$,其余结点为 0,所以需分裂结点 p_3、p_5,需要建立如图 5 所示的一棵风险平衡树 $pt4$ 和一组隐私路径.

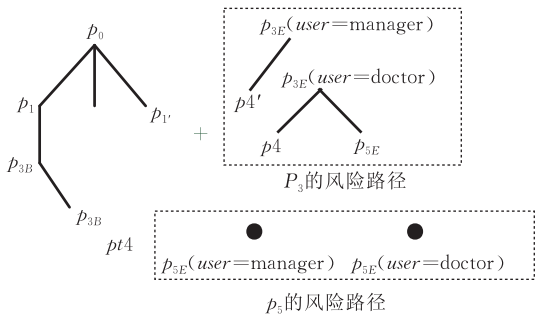

图 5 风险平衡树 $pt4$ 和风险路径图

3.3 基于风险 purpose 的隐私查询访问控制模型(RPPAAC)

模型控制方法的主要思想是分裂每个 $r(p_i)\neq 0$ 的结点,将风险融合树分解为一棵每个节点 $r(p_i)=0$ 的平衡树,任何用户基于平衡树 purpose 的查询没有隐私数据泄漏问题.其次是存储风险路径的元数据,即风险路径 purpose 下可读取的隐私属性.当有一个带有 purpose 的查询请求时[23],先执行基于平衡树 purpose 的查询,得到结果 $r1$,再以用户(或角色)为参数执行基于风险路径 purpose 的查询,得到结果 $r2$,那么最终的查询结果 $r=r1+r2$.

3.3.1 模型定义

RPPAAC 模型满足返回隐私泄漏风险最小的安全查询结果.

定义 5(安全查询结果,Safe Result,SR). 已知 $pt1$、$pt2\in PT$,分别是应用 $s1$ 和 $s2$ 的 purpose 树,pu 是 $pt1$ 和 $pt2$ 的 purpose 集合,已知 Jpt 是融合树,$Jpt=pt1\cup pt2$,$R(D_{Jpt})$是基于 Jpt 树的查询结果,用户 u 发出基于 purpose pu 的对数据库 D 的查询是 Q,那么用户 u 基于融合树 Jpt 的 purpose 的安全查询结果满足 $SR\in R(D_{Jpt})\mid R(D_{Jpt})=R(D_{p_1})\cup R(D_{p_2})$,表示用户基于融合 purpose 树的查询结果应与基于合并前 purpose 模型的查询结果一致.

定义 6(风险路径,Privacy Path). 已知 $path$ 是 Jpt 的一条路径,且 $path\in pt1$ 或 $path\in pt2$,$R(D_{path})$是基于 $path$ 的查询结果,风险路径是指所有用户 u 基于 $path$ 查询的结果 PR 不是安全查询结果,即有 $PR\neq SR$.已知路径 $Path$、路径下可访问的隐私属性集 $Data$ 以及查询用户 $User$,那么风险路径是一个三元组 $G(Path, Data, User)$.

图 4 中 G2 所示的是 $pt2$ 的一条风险路径 $p_0\geq p_3\geq p_{4'}$被合并到 $pt1$ 的过程,已知 $pt2$ purpose 的元数据如图 2 所示,基于 $pt2$ 的查询用户是 manager,那么此风险路径模型为 $G_{G2}[(p_0(p_3(p_{4'}))),(pid, mobile, result, msn), manager]$.

定义 7(基于风险路径 purpose 的隐私属性计算). 已知 R_{Jpt}、R_{Bpt}是融合风险模型 JPT 和平衡模型 BPT 的元数据关系模式,$pt1$、$pt2\in PT$, 式(1)是风险路径上基于 $purpose=p$ 可查询隐私属性的计算定义

$$\pi_{attr}(\sigma_{purpose=p}(R_{Jpt}))-\pi_{attr}(\sigma_{purpose=p}(R_{Bpt})) \quad (1)$$

平衡模型 BPT 的隐私属性可通过式(2)获得:

$$\pi_{attr}(\sigma_{purpose=p}(R_{pt1}))\cap\pi_{attr}(\sigma_{purpose=p}(R_{pt2})) \quad (2)$$

式(2)是找出控制模型 $pt1$ 和 $pt2$ 重合的那些隐私属性,这些属性与查询用户无关.

按照上述定义,基于 purpose pu 的对数据库 D 在融合树访问控制模型下的查询是 Q,那么其隐私泄漏风险最小的查询结果是在未融合 pt 树环境下基于 purpose pu 对数据库 D 查询结果的一个子集,有 $R(D_{BPT})+R(D_{Path})\subseteq R(D_{pt})$.算法 2 和算法 3

分别是基于风险 purpose 树查询算法和 purpose 融合访问控制算法.

3.3.2　算法描述

算法 2.　基于风险 purpose 树查询算法.

融合 Purpose 分解算法 decomposed_purpose()

//目的将风险树分解为平衡树和风险路径

输入：$JPT(pt3)$，u //具有隐私度的 $pt3$，原 $pt1$、$pt2$ 的访问用户 u 算法

输出：$Tpt4$，$Tpath$ //$pt3$ 的风险平衡关系 $Tpt4$，风险路径关系 $Tpath$

//设 $pt3[i]$是 $pt3$ 的一个结点，l 是 $pt3$ 的结点数

BEGIN

1. DO WHILE $1 \leq i \leq l$
2. IF $r(pt3[i])=0$，THEN
3. $pt3[i]$，$A[i]$存入 $Tpt4[i]$ //结点是非隐私结点
4. ENDIF
5. IF $r(pt3[i])\neq 0$，THEN //分裂结点 $pt3[i]$
6. 计算“式(1)”，得到 $A_E[i]$，$pt3[i]$，$PARENT[pt3[i]]$，$u[i]$ 存入 $Tpath[i]$
7. 计算“式(2)”，得到 $A_B[i]$，$pt3[i]$，$u[i]$存入 $Tpt4[i]$
8. ENDIF
9. $i=i+1$
10. ENDDO

END

融合 Purpose 分解查询控制算法 query _purpose()

输入：Q，U，$Tpt4$，$Tpath$，$JPT(pt3)$//Q 是用户 U 的带有 purpose 的查询

输出：R //基于融合 purpose 访问控制的查询结果

BEGIN

1. 已知 Q 的查询 $purpose=p$，$A[p]$是 $purpose=p$ 下的隐私属性集
2. IF $r(pt3[p])=0$ THEN //查询是基于平衡树的查询
3. IF $Tpt4[p]=pt3[p]$ THEN
4. $R=A[p]$
5. ELSE
6. $R=$null
7. ENDIF
8. ENDIF
9. IF $r(pt3[j])\neq 0$ THEN //查询是基于平衡树和风险路径的查询
10. IF $Tpt4[p]=pt3[p]$ THEN $r1=A_B[i]$ ELSE $r1=$null
11. IF $Tpath[p]=pt3[p]$ THEN $r2=A_E[i]$ ELSE $r2=$null
12. $R=r1+r2$

13. ENDIF

END

算法 3.　purpose 融合隐私查询访问控制算法.

风险 purpose 访问控制 purpose_control(input：Q，u，$pt1$，$pt2$，$A[pt1]$，$A[pt2]$；output：R)

BEGIN

//用户 u 的查询为 Q，$pt1$，$pt2$ 是独立的 purpose 树，$A[pt1]$，$A[pt2]$是 $pt1$ 和 $pt2$ 的隐私数据集，R 是查询返回的安全结果

1. Combined _purpose($pt1$，$pt2$)
2. Joint_purpose($A[pt1]$，$A[pt2]$)
3. decomposed _purpose($JPT(pt3)$，u)
4. query _purpose(Q，$Tpt4$，$Tpath$)

END

3.3.3　应用实例

图 5 是应用实例的融合 purpose 模型被分解成平衡树和隐私路径，若基于 $purpose=p_3$ 的查询，p_3 结点被分裂成 p_{3B} 和 p_{3E}，其中 $r(p_{3B})=0$，基于式(2)，p_{3B}的隐私属性是$\{result\} \cap \{result, mobile\}=\{result\}$，而基于式(1)，$p_{3E}$的隐私属性是$(\{p\text{-}id, orders, result, mobile\}-\{result\})=\{p\text{-}id, orders, mobile\}$，$\{p\text{-}id, orders\}$是 p_3 由于层次关系继承了 p_1、p_0 的隐私属性. 同样 p_5 结点被分裂成 p_{5B} 和 p_{5E}，其中 p_{5B}是$\varnothing$，若不考虑结点访问的继承性质，p_{5E}的隐私属性是$\{mailbox, history, disease\}$. 按风险路径定义(定义 6)，风险路径的隐私属性如图 6 所示. 再分析 doctor 的查询 $Q1$ 和 $Q1'$，基于 $pt3$ 的查询已知 $r(p_3)\neq 0$，所以先进行基于 $pt4$ 中 $purpose=p_{3B}$的查询，返回 $r1=(result)$；再基于风险路径 $purpose=p_{3E}$ 的查询，返回包含继承的隐私属性 $r2=(p\text{-}id, orders)$，那么结果 $r=r1+r2=(p\text{-}id, result, orders)$，而 $Q1'$无查询结果返回，查询路径如图 6 实线圈，所以模型控制了由合并 purpose 带来的隐私数据泄漏风险. 同样分析 manager 的查询 $Q2$ ($Q2=Q2'$)，由于 $purpose=p_{5B}$为$\varnothing$，查询基于风险路径 $purpose=p_{5E}$，查询的结果 $r=(p\text{-}id, history)$，查询路径如图 6 虚线圈. 从实例可见相同查询下基于新访问控制模型的查询结果等于分别基于未融合前 $pt1$ 和 $pt2$ 的查询结果，满足最大返回结果和最小隐私数据泄漏. 若基于 $purpose=p_1$ 的查询，

attr	*purpose*	*parent*	*user*
	p_{3E}	p_1	doctor
T2.mobile	p_{3E}	p_0	manager
T2.mailbox, *T1.history*	p_{5E}	p_0	manager
T1.disease	p_{5E}	p_{3E}	doctor
operation	p_4	p_{3E}	doctor
msn	$p_{4'}$	p_{3E}	manager

图 6　风险路径访问的隐私属性

因为 $r(p_1)=0$，直接在平衡树 $pt4$ 查询；若基于 $purpose=p_4$ 查询，因为 $r(p_4)=0$ 且 $p_4 \notin pt4$，直接在风险路径中查询.

4 实验与讨论

4.1 实验设置

实验中涉及的 purpose 树、元数据关系和用户的查询都基于了文章中应用实例的应用场景，实验环境是 Windows XP，AMD ATHLON 1640B 2.71GHz，1GB 内存，Sqlsever 2000，T1 和 T2 的数据记录数 $n=8400$. 本文共设置了 3 组实验：

(1) 基于不同 purpose 模型下相同查询所需时间的比较. 本组实验比较了相同用户在基于风险 purpose 和基于有访问控制 purpose 模型环境下，运行相同查询所需要的时间，并且从返回结果来观察 purpose 融合访问控制机制的执行效果.

(2) RPPAAC 模型防止隐私数据泄漏的有效性验证. 本组实验主要检验基于融合 purpose 访问控制机制的查询是否能防止隐私数据的泄漏，实验中的返回结果说明来自第 3 节的应用实例.

(3) purpose 融合的效率验证. 本组实验的目的是为了了解 purpose 融合所需的时间.

4.2 基于不同 purpose 模型下相同查询所需时间的比较

本组实验基于如下的应用场景：

(1) 应用 $s1$ 具有 purpose 树 $pt1$，其用户是 doctor，发出查询 $Q1$ 和 $Q1'$：

$Q1$：Select * From $T1$ for p_3；$Q1'$：Select * From $T2$ for p_3

(2) 应用 $s2$ 具有 purpose 树 $pt2$，其用户是 manager，发出查询 $Q2$：

$Q2$：Select * From $T1$ for p_5

本组实验主要比较了如下两种 purpose 模型下的查询执行时间：① purpose 模型是按照定义 1 将两棵独立 purpose 树合并而成，如 $pt1$、$pt2$ 合并成 $pt3$. ② 融合 purpose 访问控制机制的 purpose 模型由一棵平衡 purpose 树 $pt4$ 和一组隐私泄漏风险路径组成(图 5). 本组实验中 Doctor 和 manager 的查询都是基于 $pt3$(图 3)访问规则的查询. 图 7 是 doctor 和 manager 基于上述两种 purpose 下的查询时间比较，为清晰比较以 $Q+$、$Q1'+$ 和 $Q2+$ 代表在情况②的查询，其查询时间以阴影柱状图表示.

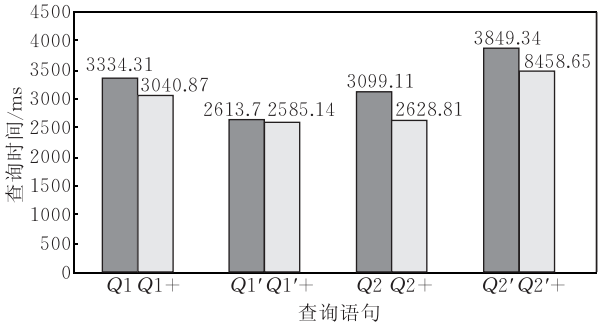

图 7 doctor 和 manager 基于不同 purpose 下的查询时间

从图 7 可以看出，在相同查询条件下基于融合 purpose 控制机制的查询($Q+$，$Q1'+$ 和 $Q2+$)时间性能开要稍优于基于一般的合并 purpose 模型的查询($Q1$、$Q1'$ 和 $Q2$)，查询时间分别从 3334.31ms、2613.7ms 和 3849.34ms 减少到了 3040.8ms、2585.14ms 和 3458.65ms. 究其原因在于查询时间取决于用户查询 purpose 与 purpose 树结点匹配的时间，这与 purpose 树的深度有关. 在情况①下 purpose 树深度 $h=4$，结点数 $n=8$；而情况②尽管分解后的平衡树深度不变，但需匹配的结点数减少到 $n=5$. 由于风险路径表示为一个关系模式，其元组数是路径上的 purpose 个数，数量较少，因此基于风险路径的查询时间可以忽略. 实验结果符合树的查找时间是 $\Theta \lg(n)$ 的特性，所以本文提出的融合 purpose 访问控制机制并没有在查询时间上给应用带来更大的开销.

4.3 RPPAAC 模型防止隐私数据泄漏的有效性验证

本组实验主要检验基于融合 purpose 访问控制机制的查询是否能防止隐私数据的泄漏，实验中的返回结果说明来自第 3 节的应用实例. 图 8 是用户发出同一查询在 3 种不同 purpose 模型下返回的查询结果.

Q1′查询结果(*user*=doctor)		
	p_id	*mobile*
情况1	√	√
情况2	√	
基于*pt1*	√	

Q2查询结果(*user*=manager)					
	p_id	*result*	*orders*	*disease*	*history*
情况1	√	√	√	√	√
情况2	√				√
基于*pt2*	√				√

图 8 相同查询下用户在不同 purpose 模型下返回的查询结果

doctor 基于 $pt1$ 的查询是指在独立应用 $s1$ 中的查询，已知隐私数据 *mobile* 只有用户 manager 才可读取，但图 8 所示的返回结果显示基于情况①模型下 doctor 的查询有隐私泄漏风险，而基于情况②模型下 doctor 的查询结果等于原先独立应用下的查询结果，这说明了泄漏风险原因由不同用户的相同 purpose 的合并引起. 同样 manager 在基于情况②模型下的查询结果等于原先独立应用 $s2$ 下的查询结果，而情况①多返回了结果，结果 $\{result, orders, disease\}$ 只有 doctor 在 $purpose=\{p_3, p_5\}$ 下才可读取，引起泄漏的原因在于 purpose 树相同结点合并后用户继承了其它用户父节点可访问的隐私数据. 由此从返回查询结果看，融合 purpose 的访问控制机制能控制由 Generalization/Specilization 特性和相同 purpose 合并带来的潜在隐私泄漏风险.

4.4　purpose 融合的效率验证

本组实验的目的是为了了解决定 purpose 融合所需时间的关键因素. 图 9 是基于图 4 中 G1、G2 和 G3 3 种合并模式下的融合时间，G1 所需的时间最少，因为是路径的直接合并. G3 的时间稍多于 G1，尽管 G1 和 G3 合并路径上的结点数一样($n=2$)，但 G3 要对 $purpose=p_5$ 做合并，需要额外的匹配时间. G3 模式开销的时间最长，因为被合并的路径最长. 可见融合时间取决于树的深度 h 和被合并路径的结点数 n，时间开销是 $O(nh)$.

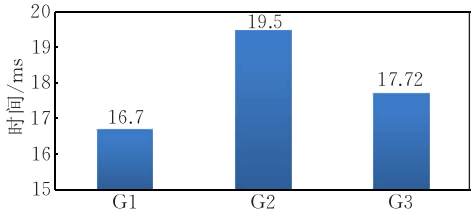

图 9　不同模式下的合并时间比较

4.5　本文与相关工作的不同点

本文将 purpose 作为隐私数据的载体、将 purpose 树的路径作为隐私数据的传递通道，引入了显性隐私度和隐性隐私度的计算模型，进而评估基于一个 purpose 查询可能带来的隐私泄漏风险.

本文讨论的是多应用环境下基于融合 purpose 的访问控制，尽管都是基于对 purpose 的匹配，但本文给出了带有隐私泄漏风险的 purpose 树模型和基于带有 purpose 隐私度查询的访问控制实现机制.

5　总结与展望

系统内隐私数据访问规则的变化、系统间功能的整合以及分布式应用的需求等应用场景都迫切需要隐私保护数据库多 purpose 融合，本文率先提出了融合 purpose 树结点的隐私度计算模型，从而评估每个 purpose 结点由于融合后 Generalization/Specilization 关系变化所引起可访问隐私数据泄漏的风险程度. 进而基于 RPPAAC 模型提出了风险 purpose 隐私查询的访问控制模型，从而返回隐私泄漏风险最小的安全查询结果. 在下一步的工作中，将隐私数据属性 retention、external-recipients 引入隐私度计算模型，研究其对隐私数据泄漏的影响.

参　考　文　献

[1] Lorrie F C, Lawrence L. Web Privacy with P3P. O'Reilly Media, Incorporated, 2002

[2] OECD. Report on the cross-border enforcement of privacy laws. Oecd/Ocde 2006, 2006

[3] Kabir M E, Wang H. Conditional purpose based access control model for privacy protection//Proceedings of the 20th Australasian Database Conference (ADC2009). Wellington, New Zealand, 2009: 137-144

[4] Agrawal R, Kiernan J, Srikant R, Xu Y. Hippocratic databases// Proceedings of the 28th International Conference on Very Large Data Bases (VLDB'02). Hong Kong, China, 2002: 143-154

[5] Byun J W, Bertino E, Lui N. Purpose-based access control for privacy protection in relational database systems. Purdue University, CERIAS Technical Report 2004-52, 2004

[6] Byun J-W et al. Purpose based access control of complex data for privacy protection//Proceedings of the 10th ACM Symposium on Access Control Models and Technologies (SACMAT'05). Stockholm, Sweden, 2005, 102-110

[7] LeFevre K, Agrawal R, Ercegovac V. Limiting disclosure in hippocratic databases//Proceedings of the 30th International Conference on Very Large Data Bases(VLDB'04). Toronto, Canada, 2004, 30: 108-119

[8] Editorial. Some issues in privacy data management. Data & Knowledge Engineering, 2007, 63(3): 591-596

[9] Agrawal R, Kini A. Managing healthcare data hippocratically//Proceedings of the 2004 ACM SIGMOD international Conference on Management of Data. Paris, France, 2004: 947-948

[10] Byun J-W, Bertino E. Micro-views, or on how to protect privacy while enhancing data usability — Concepts and challenges. ACM SIGMOD Record, 2006, 35(1): 9-13

[11] Li M, Wang H, Plank A. Privacy-aware access control with generalization boundaries//Proceedings of the 32nd Australasian Computer Science Conference (ACSC 2009). Wellington, New Zealand, 2009: 93-100

[12] Li M, Sun X, Wang H, Zhang Y. Optimal privacy-aware path in hippocratic databases. Database Systems for Advanced Applications, 2009, 5463: 441-455

[13] Massacci F et al. Hierarchical hippocratic databases with minimal disclosure for virtual organizations. The International Journal on Very Large Data Bases, 2006, 15(4): 370-387

[14] Ni Qun, Alberto Trombetta et al. Privacy-aware role based access control//Proceedings of the 12th ACM Symposium on Access Control Models and Technologies (SACMAT'07). Sophia Antipolis, France, 2007: 41-50

[15] Yasuda M et al. A purpose-oriented access control model for information flow management//Proceedings of the International Conference on Information Security (SEC'98). Budapest, Hongrie, 1998: 230-239

[16] Ni Q, Lin D, Bertino E, Lobo J. Conditional privacy-aware role based access control//Proceedings of the 12th European Symposium on Research In Computer Security. Dresden, Germany, 2007: 72-89

[17] Wang Q et al. On the correctness criteria of fine-grained access control in relational databases//Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07). Vienna, Austria. 2007: 555-566

[18] Grandison T, Ganta S R, Braun U, Kaufman J. Protecting privacy while sharing medical data between regional healthcare entities//Proceedings of the 12th World Congress on Medical Informatics (Medinfo'07). Brisbane, Australia, 2007: 484-488

[19] Samarati P, Sweeney L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Computer Science Laboratory, SRI International: Technical Report SRI-CSL-98-04, 1998

[20] Liu K et al. A framework for computing the privacy scores of users in online social networks//Proceedings of the IEEE International Conference on Data Mining (ICDM 2009). Miami, Florida, USA, 2009

[21] Gross R et al. Information revelation and privacy in online social networks//Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society (WPES'05). Alexandria, VA, USA, 2005: 71-80

[22] Shehab M et al. Beyond user-to-user access control for online social Networks//Proceedings of the 10th International Conference on Information and Communications Security (ICICS'08). Birmingham, UK, 2008: 174-189

[23] Rizvi S, Mendelzon A, Sudarshan S, Roy P. Extending query rewriting techniques for fine-grained access control//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. Paris, France, 2004: 551-562

LIU Yi-Min, born in 1963, Ph. D. candidate, senior engineer. Her research interests include privacy preservation, data mining and database, etc.

ZHOU Hao-Feng, born in 1975, Ph. D., associate researcher. His research interests include privacy preservation, data mining, database, software quality and testing, etc.

WANG Zhi-Hui, born in 1975, Ph. D., lecturer. His research interests include data security and privacy, data mining, and database systems, etc.

WANG Wei, born in 1970, Ph. D., professor. His research interests include security database, data mining, database, etc.

Background

This work was supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (grant No. 200802461146), and the Major Projects for the National Science and Technology Foundation of China (grant No. 2008ZX10002-018). The projects are involved in the key technologies of privacy preservation in data management and publishing. To meet the business requirements, the data, which are maintained by multiple organizations, and distributed in multiple systems, have to be integrated together. However, there are often some data involved with personal privacy, especially in the business of medical care, finance, and social insurance, etc. Sharing data without control will lead to privacy disclosure. Therefore, the owners may not want to share their data. To preserve privacy in data sharing, the research group has been working on many aspects in the area of data privacy and security, and published many papers. The results of these projects will benefit to the information security and the utility of data in information sharing, and facilitate the data exchange and business cooperation.

Privacy-aware data access control is one of the key technologies for privacy-preserving database systems with broad applications in medical care, finance, and social insurance, etc. Traditional works are mainly concerned with privacy-aware data access control in an independent environment. However, the fusion of purposes from multiple sources is still a problem to be addressed, especially in the scenario of system integration. In this paper, the authors analyze the potential leakage risks of privacy-aware data due to the fusion of multiple purposes, and present a novel access control mechanism, called Risk Purpose Based Privacy-Aware Access Control (RPPAAC). Using node split, RPPAAC builds a risk balanced purpose tree with the privacy degree of 0, and the mechanism for storing the risking paths. Experimental results show that RPPACC can reduce effectively the leakage risks of privacy-aware data during system integration.