面向 SaaS 应用的数据组合隐私保护机制研究

张 坤 李庆忠 史玉良
(山东大学计算机科学与技术学院 济南 250101)

摘 要 软件即服务(SaaS)模式下,业务应用和数据库部署在非完全可信的服务运营商的平台上,租户数据的隐私保护成为 SaaS 模式应用和推广中一个极大的问题和挑战.基于明文状态下不同 SaaS 数据属性组合泄露隐私程度的不同,提出一种面向 SaaS 应用的数据组合隐私保护机制.该隐私保护机制支持租户自定制隐私约束,用来描述数据组合隐私保护需求,将 SaaS 数据属性切分到不同的数据分块中,利用可信第三方实现数据切片间关联关系的混淆和重构,并基于伪造数据,确保同一数据分块内部数据切片分布的均衡化,实现 SaaS 数据组合隐私保护和实用性的有效结合.通过分析,证明了隐私保护机制的合理性,并通过实验验证了该隐私保护机制的实用性.

关键词 软件即服务;隐私保护;数据组合隐私;数据分块
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2010.02044

Research on Data Combination Privacy Preservation Mechanism for SaaS

ZHANG Kun LI Qing-Zhong SHI Yu-Liang
(*School of Computer Science and Technology, Shandong University, Jinan* 250101)

Abstract In Software-as-a-Service (SaaS) model, business applications and databases are both deployed at the platform of untrustworthy service providers. Data privacy leakage has become the biggest problem and challenge hindering application and adoption of SaaS model. Based on the privacy leakage degree of different plainx data combination in SaaS model, this paper proposes a data combination privacy preservation mechanism for SaaS. This mechanism supports customizing privacy constraint, which is used to describe the requirements of data combination privacy, and fragments the SaaS data attribute into the different data chunks. Based on the trusted third party, association between data shares from different data chunks could be hidden and reconstructed, fake data are also used to assure the balancing of the data shares in data chunks, which combines data privacy preservation and data usability. It is proven that the privacy preservation mechanism is effective and feasible through analysis and experiments.

Keywords software as a service; privacy preservation; data combination privacy; data chunk

1 引 言

随着网络技术的发展和应用软件的成熟,软件即服务(Software-as-a-Service,SaaS)作为一种新型软件服务形式逐渐兴起.SaaS 模式下,业务应用和数据库都部署在非完全可信的服务运营商的平台上,数据处理和存储在租户非完全可控的环境中进行,服务运营商可以通过直接查看数据库导致租户数据隐私泄露.SaaS 数据隐私保护成为一个比较关

收稿日期:2010-06-08;最终修改稿收到日期:2010-08-31.本课题得到国家自然科学基金(90818001)、国家科技支撑计划(2009BAH44B02)、山东省自然科学基金(ZR2010FQ026,2009ZRB019YT,Y2007G38)及山东省科技公共项目(2010GGX10105)资助.**张 坤**,男,1983 年生,博士研究生,主要研究方向为服务计算、可信计算.**李庆忠**(通信作者),男,1965 年生,博士,教授,主要研究领域为数据库、可信计算、面向语义的应用集成和信息集成.E-mail: lqz@sdu.edu.cn.**史玉良**,男,1978 年生,博士,讲师,主要研究方向为服务计算、可信计算、数据库.

注的问题，是 SaaS 模式面临的一个极大的挑战，影响着 SaaS 模式的进一步应用和推广.

例如，采用 SaaS 模式的客户关系管理系统中，公司的客户信息保存在服务运营商端的数据库中. 服务运营商不一定完全可信，有可能将公司的客户信息泄露给其它竞争公司或广告公司以获取暴利，这些泄露的信息包括客户年龄、性别、地址、邮政编码等.

目前已有的方法主要是通过对数据值的保护实现隐私保护，包括数据值加密和数据值混淆等. 数据值加密可有效防止数据隐私泄漏，但密文数据处理效率相对较低；数据值加密也可通过硬件实现，基于密码协同处理器的抗攻击特征，使用部署在非可信端的密码协同处理器作为可信处理器，在其内部实现对密文数据的加解密和处理；数据值混淆不同于传统加密，它保留了混淆数据的部分特征，可以在混淆数据上进行某些计算，但数据处理效率也有待提高. SaaS 模式下，使用传统的数据加密和数据混淆保护方法使得数据处理效率相对较低；使用密码协同处理器等硬件设备，增加了 SaaS 系统设计的复杂度.

文献[1]首次提出 Database-as-a-Service，并使用额外的索引信息检索加密的数据，但是密文处理效率相对较低，极大阻碍了其在 SaaS 模式中的大规模应用. 文献[2]给出了面向数据库应用的隐私保护技术综述，主要介绍了数据挖掘和数据发布中的隐私保护技术，包括匿名、泛化、扰动等，确保在保护隐私的同时不影响数据应用. 但是文献[2]提到的隐私保护技术主要针对分析型数据，不适于事务型数据处理 SaaS 应用. 文献[3]从数据的机密性、数据的完整性、数据的完备性、查询隐私保护以及访问控制策略等方面给出了数据库服务(Database-as-a-Service)在安全与隐私保护方面的研究进展，其中数据机密性主要从数据加密和基于数据分布两个方面展开分析. 但是 SaaS 模式下，使用加密方式和基于数据分布方式使得数据处理效率相对较低，违背了 SaaS 的初衷.

针对加密隐私保护技术的不足，研究者提出通过保护数据间关系而不是数据值的方式防止隐私泄漏. 文献[4]提出了盲目监视者机制，在同一个服务器中使用分块机制，将数据属性分在不同的数据分块中，通过隐藏数据之间的关系来实现隐私保护. 文献[5]提出了一种基于有损分解的隐私保护方法，但其面向的是分析型数据处理. 文献[6]使用信息分解与合成保护数据隐私，将数据分解后形成的数据分块重新组合，进一步保护租户数据的隐私，其中各个数据分块之间的关联关系保存在客户端. 但是，由于数据分块之间的关联关系保存在客户端，数据处理时服务器需要与客户端协作处理，通信开销相对较大，不适合 SaaS 模式. 文献[7]处理关系隐藏问题，通过独立伪装技术完全隐藏两个数据子集之间的关系，实现数据发布中的隐私保护，无法直接应用在事务型数据处理 SaaS 应用中.

文献[8-9]有效结合了关联关系分离和数据机密两种保护方法，提出使用隐私约束的概念来实现信息分解，提出隐私约束的概念，用来描述需要经过加密保护的数据属性和同时出现回泄露隐私的数据属性组合，根据这些隐私约束，经过信息分解，得到满足要求的分块模式，其中各个数据分块之间的关联关系保存在客户端. 但是文献[8-9]中隐私约束定义不全面，没有表示某些数据属性同时出现不泄露隐私的隐私约束.

针对云隐私，文献[10-11]给出了一种基于客户端的隐私管理器，通过混淆实现数据的隐私保护，并支持混淆数据的同态处理，但数据混淆处理效率有待提高. 文献[12]提出了基于理想格的全同态加密技术，支持在密文数据上的同态处理，但效率较低. 文献[13]提出了 Privacy as a Service 概念，基于密码协同处理器的抗攻击特征，使用部署在云中的密码协同处理器作为可信处理器，在密码协同处理器内部用来处理租户的敏感数据和受保护程序，防止租户敏感数据和受保护程序的泄露. 但是，通过混淆或是全同态加密，在保护数据隐私的同时，降低了数据处理的效率，增大了数据处理的开销，违背了 SaaS 的初衷. 文献[14]针对云计算环境下的隐私保护需求，提出数据隐私和操作隐私的概念，并提出了对应的解决方案. 在云计算环境中，文献[14]假定应用是可信的，并针对数据隐私提出了隐私保护框架，通过对数据的混淆保护隐私，并基于令牌通过业务应用实现混淆数据的反混淆，实现数据隐私保护和实用性的有效结合. 但文献[14]只是提出了一个框架，没有具体的解决方案.

针对以上问题，为实现 SaaS 模式下数据处理效率和隐私保护的有效结合，需要研究明文状态下的数据隐私保护机制，通过保护数据间的关联关系而不是数据值从而防止隐私泄露. 例如在采用 SaaS 模式的客户关系管理系统中通过隐藏客户的年龄、性别、地址、邮政编码等数据属性之间的关联关系，防止泄露客户的组合信息，实现对客户信息的隐私保护. 本文将能够确定一个特定客户但客户不希望泄漏的组合信息定义为数据组合隐私，提出一种面向 SaaS 应用的数据组合隐私保护机制，将组合隐私属

性切分到不同的数据分块中，混淆不同数据分块中数据切片间的关联关系，保护租户数据的组合隐私；针对数据分块中数据切片分布导致的隐私泄露问题，提出基于伪造数据的保护方法；同时，构建混淆数据的重构机制，实现 SaaS 数据隐私保护和实用性的有效结合.

文献[15]为本文之前关于 SaaS 数据隐私保护的研究成果，使用信息分解，将数据分解到不同的数据分块中，隐藏数据之间的关系，从而保护 SaaS 数据隐私. 文献[15]扩充了隐私约束的概念，重新定义了隐私约束的概念，加入了相容隐私约束，用来表示同时出现的不泄露隐私的数据组合，并以相容隐私约束为基础，基于贪婪算法实现快速信息分解，同时文献[15]提出基于元数据驱动的多租户数据共享存储模式，通过租户定制将数据分块之间的关联关系保存在服务运营商端，只有租户才能够重构数据. 本文基于文献[15]的研究基础，提出了(k,α,β,γ)隐私保护机制，在使用分块保护 SaaS 数据隐私的同时，进一步规范化了数据分解过程，并针对数据对象物理存储中数据切片分布导致的隐私泄露问题，提出了均衡化的概念，使用插入伪造数据的方式防止隐私泄露.

本文第 2 节给出面向 SaaS 应用的数据组合隐私保护模型；第 3 节给出基于分块的数据关系混淆方法；第 4 节针对数据分块中数据切片分布导致的隐私泄露问题，给出基于伪造数据的保护方法；第 5 节给出混淆数据的重构机制；第 6 节给出实验及结果；第 7 节给出相关工作；第 8 节进行总结.

2 SaaS 数据组合隐私保护模型

本节给出数据组合隐私保护的相关概念，提出面向 SaaS 应用的数据组合隐私保护模型.

2.1 数据组合隐私

数据隐私保护的一种方法是隐藏数据值，主要包括数据加密和混淆处理，防止数据值的泄露；另一种方法是隐藏数据之间的关联关系，防止数据间关联关系的泄露，即防止某些数据组合带来的隐私泄露. 首先，给出数据组合隐私相关的概念.

定义 1. 数据组合隐私(Data Combination Privacy，DCP). 数据组合隐私为个体不希望暴露的一系列数据属性的组合，其对应数据值可以确定特定个体. 具体的，数据组合隐私为 $DCP\{A_1,A_2,\cdots,A_m\}$，其中 $A_i(1\leq i\leq m)$ 为构成数据组合隐私的数据属性.

如在图 1 所示的例子中，$DCP\{Age, Sex, Zipcode\}$ 是一个数据组合隐私，因为这 3 个属性可

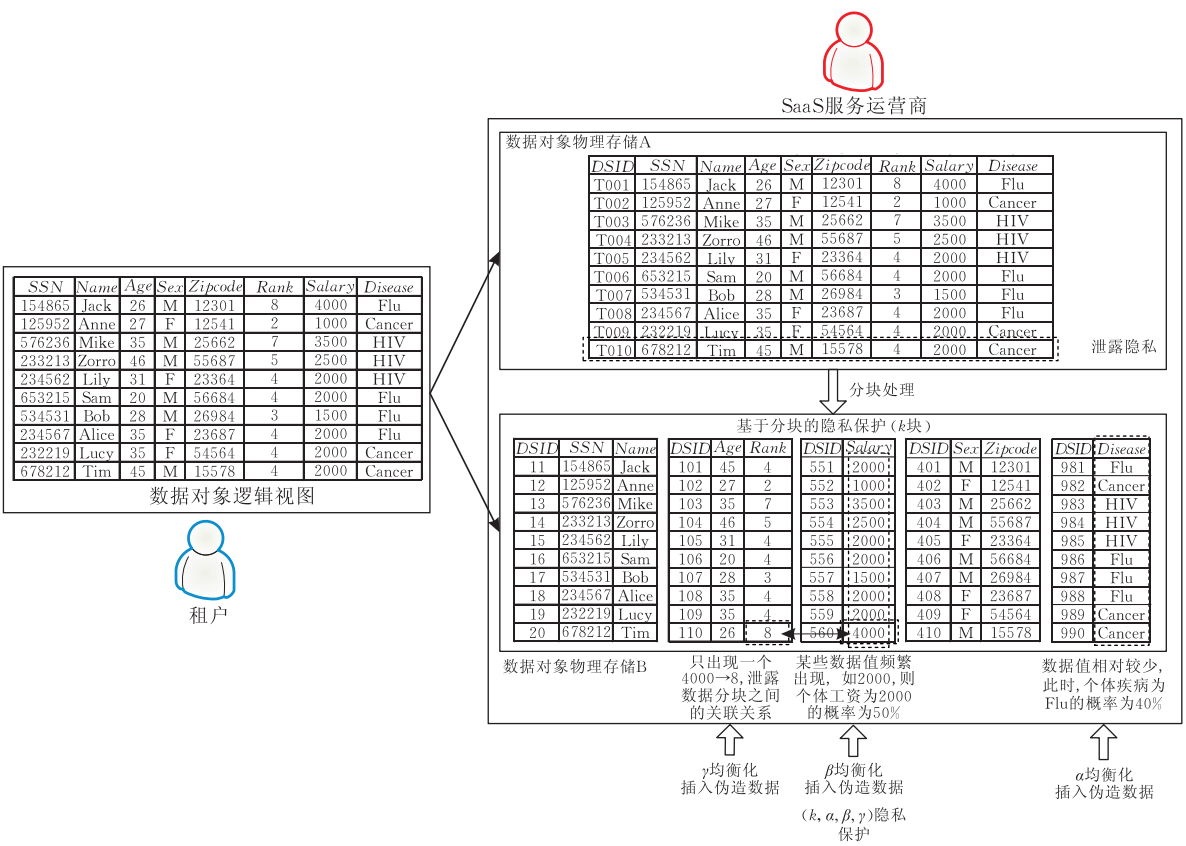


图 1 SaaS 数据组合隐私泄露及保护示意图

以作为一个准标识符，从而可以确定个体.

SaaS 模式下，给定具体数据存储 D，一个具体数据组合隐私 DCP 的泄露概率为 $P_{DCP}=P(DCP|D)$，该泄露概率为一个条件概率，其中，D 为已知的 SaaS 数据物理存储，DCP 为泄露的数据组合隐私. SaaS 模式下，租户数据采用明文存储，按照图 1 中数据对象物理存储 A 所示，则 $P_{DCP\{Age,Sex,Zipcode\}}=1$.

针对数据组合隐私，规定数据组合隐私保护阈值 $T_{\text{threshold}}(0\leq T_{\text{threshold}}\leq 1)$. 对于任何一个数据组合隐私 DCP_i，若 $P_{DCP_i}\leq T_{\text{threshold}}$，即给定 SaaS 数据具体物理存储 D，非完全可信的服务运营商发现任何一个具体数据组合隐私的概率均不大于给定的阈值时，则认为该 SaaS 数据具体物理存储 D 保护了租户的数据组合隐私.

为描述租户的数据组合隐私保护需求，给出隐私约束的概念.

定义 2.　隐私约束(Privacy Constraint，PC). A 为租户数据对象的属性集合，$A=\{A_1,A_2,\cdots,A_n\}$，隐私约束为 $PC\{AS(\text{Attribute Set}),PP(\text{Privacy Policy})\}$，其中 AS 为隐私约束涉及的属性集合，为 A 的子集；PP 为对应的隐私策略，包括 Non-Compatible 和 Compatible 两种，其中 Non-Compatible 表示 AS 中的属性若同时出现会导致数据组合隐私泄露，Compatible 表示 AS 中的属性同时出现不会导致数据组合隐私泄漏. 隐私约束中隐私策略必须选择一种.

其中，不相容隐私约束(non compatible Privacy Constraint)为 $ncPC\{AS,\text{Non-Compatible}\}$、相容隐私约束(compatible Privacy Constraint)为 $cPC\{AS,\text{Compatible}\}$.

2.2　基于数据分块的物理存储

为实现其数据组合隐私保护，基于隐私约束，将 SaaS 数据属性切分到不同的数据分块中，利用可信第三方实现数据切片间关联关系的混淆. 为此，首先给出基于数据分块的物理存储等概念，作为面向 SaaS 应用的数据隐私保护的基础.

定义 3.　数据对象逻辑视图(Data Object Logical View，DOLV). SaaS 模式下，租户数据对象逻辑视图为 $DOLV\{A_1,A_2,\cdots,A_n\}$，其中 $A_i(1\leq i\leq n)$ 为租户数据对象的数据属性.

定义 4.　数据分块物理存储(Data Chunk Physical Storage，DCPS). SaaS 模式下，数据分块在服务运营商平台上数据库服务器中的实际物理存储为 $DCPS\{DSID,A_1,A_2,\cdots,A_x\}$，其中，$DSID$ 为数据切片标记，用来标识该数据分块中的数据切片，$A_i(1\leq i\leq x)$ 为数据属性，该数据分块物理存储中对应的属性的集合为数据对象逻辑视图中属性集合的子集.

基于数据分块物理存储的概念，给出数据对象物理存储的概念.

定义 5.　数据对象物理存储(Data Object Physical Storage，DOPS). SaaS 模式下，租户数据对应的物理存储为 $DOPS\{(DCPS_1,DCID),(DCPS_2,DCID),\cdots,(DCPS_k,DCID)\}$，其中各个数据分块物理存储中的数据属性不存在交集，即 $DCPS_i\cap DCPS_j=\varnothing(i\neq j)$，而且数据分块物理存储中属性集合的并集为对应数据对象逻辑视图的属性集合，即 $\sum DCPS_i=DOLV$. 在数据对象物理存储中的数据分块物理存储是有顺序的，使用 $DCID$ (Data Chunk ID)标记数据分块物理存储的次序，其中 $1\leq DCID\leq k$.

对于租户的数据对象逻辑视图 $DOLV$，对应的数据对象物理存储为 $DOPS$，当 $DOPS$ 中包含的数据分块物理存储 $DCPS$ 的数目为 k 时，称该数据对象物理存储 $DOPS$ 为数据对象逻辑视图的 k 分块化. 此时，数据对象逻辑视图 $DOLV$ 中的一条数据记录被切分为 k 个不同的数据切片，加上对应的数据切片标记 $DSID$，分到 k 个对应的数据分块物理存储 $DCPS$ 中.

数据切片标记 $DSID$ 在数据组合隐私保护机制中的作用十分重要，一方面，属于同一数据记录的不同数据切片的 $DSID$ 不同，隐藏了数据切片之间的关联关系，保护了租户的数据组合隐私；另一方面，利用可信第三方，通过数据切片的 $DSID$，进行数据记录的重构，即数据对象逻辑视图的重构，实现数据隐私保护与实用性的有效结合.

2.3　均衡化

针对某一具体的使用数据分块的数据对象物理存储，非完全可信的服务运营商可根据其数据切片的分布情况泄漏租户的数据组合隐私.

(1) 当数据分块物理存储中某个属性的取值个数较少时，如图 1 中数据对象物理存储 B 中，$Disease$ 属性对应的数据值相对较少，此时，个体疾病为 Flu 的概率为 40%.

(2) 当数据分块物理存储中某个属性中某个取值所占的比例较大时，如图 1 中数据对象物理存储 B 中，$Salary$ 属性中某些数据值频繁出现，如 2000，则个体工资为 2000 的概率为 50%.

(3) 租户数据对象物理存储中数据属性之间存在完全函数依赖时，如针对完全函数依赖 $Salary\rightarrow$

$Rank$,租户数据对象物理存储 B 中对应的两个数据分块物理存储{$DSID$, Age, $Rank$}和{$DSID$, $Salary$}只出现一个 4000→8,泄露了对应数据分块之间的关联关系.

因此,针对数据对象物理存储中数据切片分布导致的隐私泄漏问题,本文提出均衡化的概念,确保每个数据分块物理存储中各种数据切片出现的概率尽可能地平均,防止非完全可信的服务运营商以较大的概率发现数据组合隐私.

定义 6. 数据分块物理存储的 α 均衡化. 数据分块满足 α 均衡化,满足以下条件:数据分块中至少包含 α 个不同的数据切片组合.

定义 7. 数据分块物理存储的 β 均衡化. 数据分块满足 β 均衡化,满足以下条件:数据分块中每种数据切片组合出现的最大频率不超过 $\beta(0<\beta<1)$.

通过 β 均衡化,对应数据分块物理存储组成的具体组合隐私泄露的概率 P_{DCP} 最多为 β,即 $P_{DCP}\leq\beta$.

对于数据属性之间存在的完全函数依赖引起的隐私泄露问题,进一步提出 γ 均衡化的概念.

定义 8. 对于完全函数依赖 F 的 γ 均衡化. 对于完全函数依赖 F:$A\rightarrow B$,数据属性 A 和 B 存储在不同的数据分块物理存储 $DCPS_A${$DSID$, $AttributeSet$}和 $DCPS_B${$DSID$, $AttributeSet$}中,数据分块 $DCPS_B$ 除了包含数据属性 B 外,还包括其它数据属性. 给定数据库实例,对于任何数据记录 t, $t.b=F(t.a)$,若 $t.a$ 在数据分块物理存储 $DCPS_A$ 中出现,则在数据分块 $DCPS_B$ 中,对应有 n_b 个数据切片选择,在这 n_b 个不同的数据切片中,每种数据切片出现的最大频率不超过 $\gamma(0<\gamma<1)$.

通过 γ 均衡化,针对完全函数依赖 F, F 中数据属性涉及的数据分块物理存储组成的具体组合隐私泄露的概率 P_{DCP} 最多为 γ,即 $P_{DCP}\leq\gamma$.

基于伪造数据可有效地实现均衡化,具体细节见第 4 节. 为避免伪造数据对 SaaS 业务应用的影响,需要实现伪造数据的标识和识别,这是通过数据切片标识 $DSID$ 实现的.

2.4 数据对象逻辑视图重构

对于租户数据的实用性,要求能够从数据对象物理存储经过某种映射得到数据对象逻辑视图,即 SaaS 数据的实用性要求能够实现数据间关联关系混淆的逆操作. 首先,给出数据对象逻辑视图重构的概念.

定义 9. 数据对象逻辑视图的重构(Data Object Logical View Reconstruction, DOLVR). 给定数据对象物理存储 $DOPS$,对于数据对象物理存储 $DOPS$ 中的 k 个数据分块物理存储 $DCPS$,能够从 $DCPS_i(1\leq i\leq k)$ 中选择合适的数据切片,得到对应的租户数据对象逻辑视图 $DOLV$.

值得注意的是,非完全可信的服务运营商无法快速、准确地完成数据对象逻辑视图的重构. 在基于分块的数据对象物理存储中,非完全可信的服务运营商可以通过暴力攻击穷尽所有的可能,但是仍无法确定哪些数据分块组合是真实的,哪些数据分块组合是不存在的,从而有效防止了数据组合隐私泄露.

2.5 数据组合隐私保护模型

为方便讨论 SaaS 数据组合隐私保护,假定服务运营商无法窥探内存、SaaS 应用是可信的.

首先,给出数据对象物理存储的(k,α,β,γ)隐私保护定义.

定义 10. 数据对象物理存储的(k,α,β,γ)隐私保护. 数据对象物理存储若满足以下要求则称其满足(k,α,β,γ)隐私保护:(1) 数据对象逻辑视图 $DOLV$ 经过 k 分块化后得到数据对象物理存储 $DOPS$;(2) 每个数据分块物理存储都满足 α 均衡化;(3) 每个数据分块物理存储都满足 β 均衡化;(4) 对于任何完全函数依赖 F,数据对象物理存储 $DOPS$ 均满足 γ 均衡化.

面向 SaaS 应用的数据组合隐私保护模型涉及租户、可信第三方和部署在非完全可信的服务运营商端的 SaaS 数据隐私保护模块. 在租户应用过程中,SaaS 数据隐私保护模块利用可信第三方提供的隐私保护支持功能实现服务运营商端的租户数据组合隐私的保护,这里认为该 SaaS 数据隐私保护模块是可信的.

可信第三方实现认证管理、隐私保护策略管理等功能. 其中,认证管理用来防止服务运营商冒充合法租户身份进行混淆数据的重构;隐私保护策略管理模块维护租户的隐私策略,包括数据切片标识 $DSID$ 的关联策略、伪造数据构建策略、伪造数据切片 $DSID$ 的生成和识别策略等.

SaaS 数据隐私保护模块实现数据组合隐私保护定制、数据关系混淆、均衡化管理、混淆数据重构等功能. 其中数据组合隐私保护定制支持租户按照需求定制隐私约束和数据组合隐私保护阈值等;数据关系混淆模块根据租户定制的隐私约束进行分块处理,并与可信第三方的隐私保护策略管理模块协作,根据数据切片标识 $DSID$ 的关联策略为各个数据切片生成对应的 $DSID$,实现数据切片之间的关联关系的隐藏,并将对应的数据切片保存到数据对象物理存储中;均衡化管理监视 SaaS 数据存储是否

满足租户的均衡化需求，若不满足，则与可信第三方协作的隐私保护策略管理模块协作，根据伪造数据构建策略和伪造数据切片 *DSID* 的生成和识别策略，构建并插入一定量的伪造数据，非完全可信的服务运营商无法确认该类数据的来源，即无法确认插入的数据是真实的还是伪造的；混淆数据重构模块实现数据对象逻辑视图的重构，该模块需要与可信第三方的隐私保护策略模块协作，根据对应策略，去除伪造数据切片，并实现数据对象逻辑视图的重构.

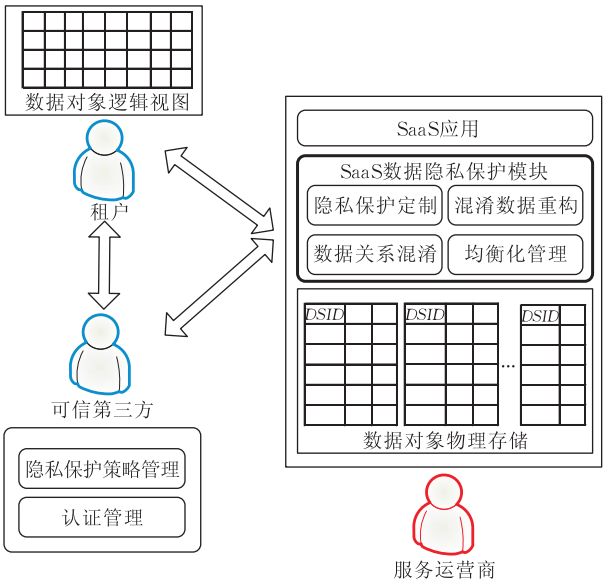


图 2　SaaS 数据组合隐私保护模型示意图

SaaS 数据隐私保护模型的整个工作流程如下所示：首先，租户将一些关于隐私保护的策略信息发送给可信第三方，授权可信第三方进行一定的隐私保护支持操作；然后租户进行隐私保护定制，包括隐私约束定制、隐私保护阈值等，涉及到 k、α、β、γ 等参数；基于租户的隐私保护定制，利用可信第三方的隐私保护策略，SaaS 数据隐私保护模块根据租户定制的隐私约束进行分块处理，得到满足租户隐私需求的 k 个数据分块，并根据租户定制的 α、β、γ 等参数进行各种均衡化处理，SaaS 数据隐私保护模块基于租户数据的分布，与可信第三方协作，产生合适的伪造数据，实现数据对象物理存储的均衡化. SaaS 数据隐私保护模块同时可利用可信第三方的隐私保护策略进行伪造数据的识别和混淆数据的重构，从而实现 SaaS 数据组合隐私保护和实用性的有效结合.

3　数据关系混淆

为保护 SaaS 数据组合隐私，基于隐私约束实现分块处理，将泄漏租户隐私的组合隐私属性切分到不同的数据分块中，并通过 SaaS 数据隐私保护模块与可信第三方的协作，实现数据切片之间的混淆，防止数据组合隐私的泄漏.

3.1　隐私约束检查

SaaS 应用支持租户个性化定制，由于租户水平的良莠不齐，定制的隐私约束可能存在着冗余及冲突，需要隐私保护定制模块进行必要的检查.

隐私约束冗余，即针对相同类型的隐私约束 PC_i、PC_j，对应的属性集合 AS_i、AS_j 为包含关系，则称这两个隐私约束 PC_i、PC_j 存在隐私约束冗余. 为减少隐私约束冗余，对于不相容隐私约束，保留属性集合较小的不相容隐私约束；对于相容隐私约束，保留属性集合较大的相容隐私约束.

隐私约束冲突，即不相容隐私约束 $ncPC_i$ 和相容隐私约束 cPC_j 中，对应的属性集合 AS_i 为 AS_j 的子集，即租户定制的不能同时出现的属性集合出现在相容隐私约束中，则认为隐私约束 $ncPC_i$ 和 cPC_j 产生冲突. 为保护租户数据隐私，以不相容隐私约束为准，去除冲突的相容隐私约束.

3.2　数据分块

基于隐私约束，对数据对象逻辑视图进行分块处理，实现满足要求的数据对象物理存储. 该算法的主要思想为：经过冗余和冲突检查，基于租户定制的隐私约束，进行分块处理. 以相容隐私约束为基础，在不违背不相容隐私约束的前提下，不断加入新的数据属性，否则创建新的数据分块，直到所有的数据属性都分到合适的数据分块中，最终得到满足隐私约束的数据对象物理存储. 基于隐私约束的数据分块算法如算法 1 所示.

算法 1.　基于隐私约束的数据分块算法.

输入：数据对象逻辑视图 $DOLV\{a_1,a_2,\cdots,a_n\}$；
　　不相容隐私约束集合 $NCPC\{ncPC_1,ncPC_2,\cdots,cPC_x\}$；
　　相容隐私约束集合 $CPC\{cPC_1,cPC_2,\cdots,cPC_y\}$
输出：满足隐私约束的数据对象物理存储 *DOPS*
步骤：
不相容隐私约束 *NCPC* 冗余检查及处理；
相容隐私约束 *CPC* 冗余检查及处理；
隐私约束冲突检查及处理；
集合 *AvailableCompatibleFragement*＝{*cPC*｜经过优化后的相容隐私约束集合(任何两个有交集的相容隐私约束，保留数据属性多的，否则随机选取一个)}；
集合 *ToBeSolved*＝{*ai*｜*ai* 为数据对象逻辑视图中的数据属性且没有被分块处理}；
如果 *ToBeSolved* 不为空
　从 *ToBeSolved* 中随机选出一个数据属性 *a*；
　从集合 *AvailableCompatibelFragment* 中选取可用

的最大的数据分块；

若将 a 加入不违背不相容隐私约束，则加入；

否则，选取次大的加入；

如果没有合适的，则创建新的数据分块$\{a\}$，并将其加入到集合 $AvailableCompatibleFragment$ 中；

对于 $AvailableCompatibleFragment$ 中的数据分块，检查是否能够合并；

如果能，则进行合并，直到合并会破坏隐私约束为止；

将 $AvailableCompatibleFragment$ 转换为数据对象物理存储 $DOPS$ 返回.

3.3 数据切片间关系混淆

为实现 SaaS 数据组合隐私保护，需要混淆不同数据分块中各个数据切片间的关联关系，即属于同一数据记录的各个数据切片的关联关系，这主要是通过数据切片标记 $DSID$ 与可信第三方实现的. 下面给出数据切片标记的具体定义.

定义 11. 数据切片标记(Data Share ID，DSID). 数据切片标记用来标记数据分块物理存储中的各个数据切片，其具体定义如下所示：

$$DSID=E_{key}(ID_{\mathrm{TTP}}\oplus a^{DCID}),$$

其中 E 为租户隐私保护策略中选择的乘法同态加密函数；key 为函数 F 对应的私钥；ID_{TTP} 为按照隐私保护策略为数据对象逻辑视图中每个数据记录设置的一个唯一标记；$DCID$ 为数据切片所属数据分块物理存储的次序；a 为隐私保护策略中指定的某个循环群的生成元.

部署在非完全可信的服务运营商端的 SaaS 数据隐私保护模块与可信第三方协作，根据租户的隐私保护策略，为同一数据记录的不同数据切片生成对应的 $DSID$.

整个数据对象逻辑视图 $DOLV$ 的混淆方法如算法 2 所示.

算法 2. 数据对象逻辑视图 $DOLV$ 的混淆方法.

输入：数据对象逻辑视图 $DOLV$；

SaaS 原始数据集 D；

隐私约束 PC

输出：混淆后的数据对象物理存储 $DOPS$

步骤：

利用 $DOLV$ 和 PC，调用算法 1，得到满足隐私约束的数据分块；

数据关系混淆模块与可信第三方协作，得到租户对应的隐私保护策略 P；

对于 D 中的任何一条数据记录

根据 $DOPS$ 进行数据切分，得到 k 个数据分块的切片

根据隐私保护策略 P 和定义 11，为每个数据切片构建 $DSID$；

为每个数据切片标注 $DSID$，并将其加入到 $DOPS$ 中

返回 $DOPS$.

4 基于伪造数据的均衡化调整

基于伪造数据的方法确保数据对象物理存储满足各种均衡化，保护 SaaS 数据组合隐私. 该均衡化工作检查和伪造数据生成都是由 SaaS 数据隐私保护模块完成的. 其主要过程为 SaaS 数据隐私保护模块与可信第三方协作，根据租户的隐私保护策略生成伪造数据和对应的 $DSID$，并一起存储到数据对象物理存储中. 当需要进行数据真伪判别时，SaaS 数据隐私保护模块根据隐私保护策略进行验证.

4.1 伪造数据生成

为满足各种均衡化，利用最大熵，使用插入伪造数据的方法，使得均衡化后的数据切片分布尽可能地平均. 简单起见，假设数据对象各个属性空间是有限离散的. 基于数据对象物理存储，可以得到不同数据属性组合所有取值的概率分布，方便起见，使用出现次数代替概率分布，其中，某些数据组合的出现次数可能为 0. 生成均衡化需要的伪造数据时，每次选择出现次数最少的数据组合作为伪造数据，直到产生满足均衡化的数目为止. 其具体的伪造数据生成算法如算法 3 所示.

算法 3. 均衡化伪造数据生成.

输入：伪造数据个数 m；

待均衡化的数据组合属性(Data Combination Attributes) DCA；

DCA 中所有可能数据值的出现次数 C_{DCA}

输出：待插入的伪造数据集合 T

步骤：

T 置为空；

For (int $i=0$; $i<m$; $i++$)

{

从 C_{DCA} 中选择出现次数最少对应的数据组合 d_i；

将数据组合 d_i 插入到 T 中；

根据插入的数据组合 d_i 更新 C_{DCA}；

}

返回 T.

针对数据分块物理存储 $DCPS$ 的 α 均衡化，检查 $DCPS$ 中不同数据切片的个数 $|DCPS|$，若 $\alpha>|DCPS|$，则调用算法 3 ($\alpha-|DCPS|$, $DCPS$, C_{DCPS}).

针对包含 n' 个数据切片的数据分块物理存储 $DCPS$ 的 β 均衡化，检查 $DCPS$ 中各个数据切片 ds_i 的出现次数 c_{ds_i}，若 $c_{ds_i}/n'>\beta$，则调用算法 3 ($(c_{ds_i}/\beta)-n'$, $DCPS$, C_{DCPS}).

针对数据分块物理存储 $DCPS$ 的 γ 均衡化，假定函数依赖 $F:A\rightarrow B$ 对于数据分块物理存储 $DCPS_B$ 不满足 γ 均衡化，其中数据属性 B 处于数据分块物理存储 $DCPS_B$ 中，A 处于数据分块物理存储 $DCPS_A$ 中，假定该数据分块中除去数据属性 B 之外的其它属性集合为 B'. 则为使得数据分块 $DCPS_B$ 满足 γ 均衡化，需要检查函数依赖 $F:A\rightarrow B$ 中所有的情况. 当某个 $a_i\rightarrow b_j$ 不满足 γ 均衡化后，获得数据分块物理存储 $DCPS_B$ 中，数据属性 B 为 b_i 的数据切片个数 n_b，在这 n_b 个数据切片中出现频率超过 γ 的数据切片标记为 ds，其数目标记为 $|ds|$，调用算法 3 ($|ds|/\gamma-n_b,B',C_{B'}$).

4.2　分　析

在该隐私保护机制下，分析租户数据组合隐私的泄露概率.

SaaS 模式下，租户数据对象物理存储 $DOPS$ 满足 (k,α,β,γ) 均衡化，其中包括各个数据分块物理存储的切片个数为 $n_i(1\leq i\leq k)$，其针对组合隐私为 $DOPS$ 中的属性集合的隐私泄露概率 $P_{DCP(DOPS)}$ 的取值范围为

$$\left[\min\left(\prod(1/n_i),\beta^k,\gamma^k\right),\max\left(\prod(1/n_i),\beta^k,\gamma^k\right)\right].$$

分析：若所有的数据分块都满足 α 均衡化和 β 均衡化，则每个数据分块中重复数据切片所占的最大比例为 β，对应于 k 个数据分块，不考虑函数依赖的情况下，则数据对象物理存储数据组合隐私的泄露概率最大为 β^k.

仅考虑函数依赖，在极端情况下，任何两个数据分块之间都存储函数依赖，且都满足 γ 均衡化，则通过函数依赖，则数据对象物理存储数据组合隐私的泄露概率最大为 γ^k.

若对应的数据分块中，对应的 n_i 个数据切片都是唯一的，则数据对象物理存储数据组合隐私的泄露概率为 $\prod(1/n_i)$.

综上，若数据对象物理存储满足 (k,α,β,γ) 均衡化，则 $P_{DCP(DOPS)}$ 的取值范围为

$$\left[\min\left(\prod(1/n_i),\beta^k,\gamma^k\right),\max\left(\prod(1/n_i),\beta^k,\gamma^k\right)\right].$$

由此可见，在不存在重复数据切片的情况下，隐私保护与数据分块个数和数据切片个数有很大的关系，分块越多、数据切片越多，隐私保护效果越好. 若存在重复数据切片，则 β,γ 值越小，k 值越大，隐私保护效果越好.

5　混淆数据重构

SaaS 模式下，通过 (k,α,β,γ) 隐私保护能有效防止租户数据组合隐私的泄露. 为确保 SaaS 数据的实用性，提出利用可信第三方的混淆数据重构机制.

5.1　重　构

由第 2 节数据分块物理存储和数据对象物理存储的定义可知，数据切片标记可以用来进行对应数据记录的重构，即数据对象逻辑视图的重构，同时，还能用来进行数据切片真伪的判别，从而支持数据对象逻辑视图的重构.

当 SaaS 应用需要重构租户数据对象逻辑视图时，SaaS 数据隐私保护模块从可信第三方得到隐私保护策略，从而进行计算，得到满足要求的租户对象逻辑视图. 其对应的算法如算法 4 所示.

算法 4.　数据对象逻辑视图重构.

输入：数据对象逻辑视图个体 t 的全局标识 ID_{TTP}；
　　数据对象物理视图 $DOPV$；
输出：个体 t 的对应的各个数据分块物理视图的数据切片标记
步骤：
混淆数据重构模块与可信第三方协作，得到隐私保护策略 P；
得到 k=数据对象物理视图的数据分块物理视图个数；
数据切片标记数组 $DSID[k]$；
For(int $i=1$; $i<=k$; $i++$)
{
　根据定义 11 进行计算，得到对应各个数据切片的 $DSID$；
}
Return $DSID[]$.

基于数据切片标记 $DSID$，可以实现数据对象逻辑视图的重构. 算法基于可信第三方的隐私保护策略，可以进行对应的重构. 除了根据数据记录的全局标识进行数据对象逻辑视图的重构外，还可以从任意一个数据切片开始进行对应数据对象逻辑视图的重构. 此时，SaaS 数据隐私保护模块根据数据隐私保护策略，对数据切片标记 $DSID$ 进行解密等操作后，得到该数据切片所属数据记录的全局数据标记 ID_{TTP}，然后再调用算法 4 进行该数据切片对应的数据记录的数据对象逻辑视图的重构.

5.2　分　析

假定数据混淆存储是可信的，能够返回所有满足要求的数据，服务运营商不会删除、篡改、伪造数据混淆存储中的数据. 通过数据对象逻辑视图重构，

利用可信第三方,验证数据的真伪,并得到不同数据切片之间的关联关系,从而实现反混淆,最终能够得到满足要求的无损的数据对象逻辑视图.

6 实 验

6.1 实验环境

针对 SaaS 数据组合隐私保护,通过仿真模拟验证所提隐私保护模型的实用性.

数据库采用 MySQL 5.1.22,编程环境为 Eclipse-SDK-3.5.2-win32,编程语言采用 Java 5,操作系统为Windows XP Professional Service Pack 2,CPU 为 Inter Core2 Duo,主频 2.33GHz,内存 2GB.

6.2 数据分块开销实验

为有效验证数据分块算法针对各种不同隐私约束的有效性,设计了 3 种不同类型的租户隐私保护需求,如表 1 所示,不同类型中相容隐私约束和不相容隐私约束所占的比重不同.假定租户定制的隐私约束总数与数据对象逻辑视图中的数据属性个数相等.

表 1 租户不同类型的隐私需求

	相容隐私约束/%	不相容隐私约束/%
类型 A	90	10
类型 B	10	90
类型 C	50	50

对于不同大小的数据对象逻辑视图,随机生成 3 种不同的隐私约束类型,并利用其进行分块处理,对应的时间开销如图 3 所示.

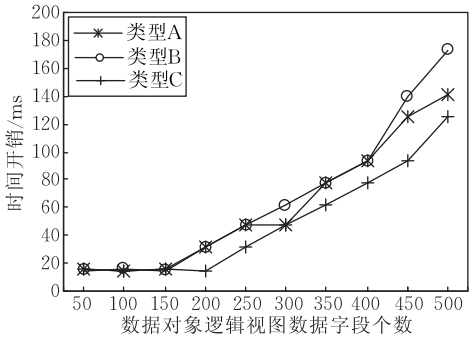


图 3 数据分块处理开销

由实验结果表明:当数据对象逻辑视图中数据属性个数较大时,针对 3 种不同的隐私定制需求,时间开销较小.

6.3 均衡化开销实验

当数据分块个数不同时,各种均衡化的检查时间开销如图 4、图 5 所示.

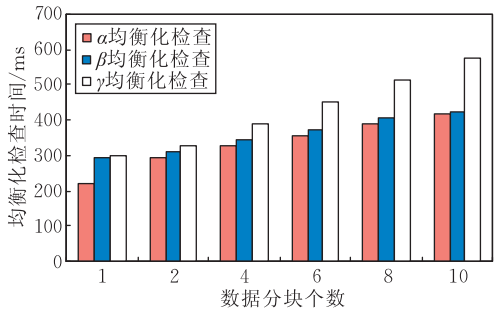


图 4 固定数据切片数目均衡化检查时间开销

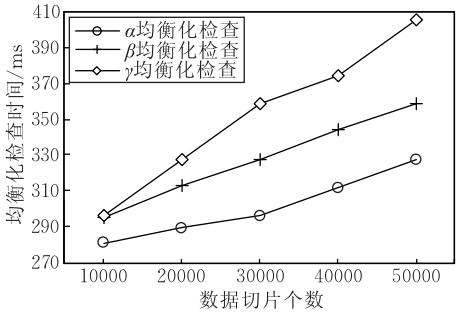


图 5 单个数据分块均衡化检查时间开销

图 4 给出了为当租户数据对象逻辑视图包含固定数目的数据记录(10000 条数据记录)时,在各个不同的分块条件下,各种均衡化的时间开销.图 5 是针对单个数据分块,当租户数据对象逻辑视图数据记录个数不同,即对应数据分块物理存储中数据切片数目不同时的均衡化时间开销.由图 4 和图 5 可知,3 种均衡化检查的时间开销相对较小,实用性较强.同时,在同样的条件下,3 种均衡化检查的时间开销为 γ 均衡化$>\beta$ 均衡化$>\alpha$ 均衡化.

当均衡化检查完毕,对于不满足要求的数据分块,本文提出的隐私保护机制需要根据数据分块中数据切片的分布情况插入适当数量的伪造数据,图 6 中的实验展示了伪造数据的生成开销和插入开销,实验表明伪造数据的生成开销和插入开销相对较小.

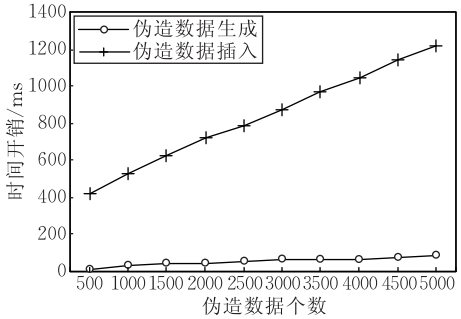


图 6 伪造数据生成及插入时间开销

6.4　数据处理开销实验

本测试暂时只考虑单租户逻辑视图的情况，使用 TPC-E 中的 CUSTOMER 表模式，它包含了经济公司的客户信息，共包括 24 个数据属性. 为验证采用分块带来的开销影响（暂时不考虑伪造数据），并与数据加密方式进行比较，构建对应的测试，其中每个测试数据集中对应的数据对象逻辑视图都包含 10000 条数据记录. 对于 $k>1$ 的情况，数据库会为每个数据分块添加两列，分别为 *GUID* 和 *DSID*，其中，*GUID* 用来在数据库中唯一判定一个数据分块；*DSID* 用来标识数据切片，属于同一数据记录的不同数据切片的 *DSID* 不同，但数据隐私保护模块可通过与可信第三方协作，获得租户的隐私保护策略，从而实现租户数据对象逻辑视图的重构. 对于数据加密方式，假定加密方式辅助索引桶的大小为 10，则 10000 条加密数据记录一共 1000 个桶.

测试如表 2 所示，其中，I 为租户数据对象逻辑视图与数据对象物理存储相同的情况，即不做任何隐私保护措施，直接按照逻辑视图模式存储租户数据. 本文将该测试 I 作为测试基准，用来比较本文提出的隐私保护机制和密文保护机制在实用性方面的性能开销. 简便起见，仅考虑简单的数据操作，包括插入、删除、更新、查询 4 种.

表 2　针对数据操作的测试

测试	隐私保护机制	分块个数	具体分块
I	未保护	1	{24}
II	数据分块	2	{12,12}
III	数据分块	4	{6,6,6,6}
IV	数据分块	6	{4,4,4,4,4,4}
V	数据分块	8	{3,3,3,3,3,3,3,3}
VI	数据分块	10	{3,3,3,3,2,2,2,2,2,2}
VII	数据加密	1	{24}

由图 7 可知，使用基于分块的隐私保护机制，数据处理开销相对加大，主要是由于采用分块机制后，不仅需要处理多个数据分块，同时还需要处理数据切片的关联重构. 分块数 k 越多，对应的开销也越大. 同时，针对数据加密保护方式，本文提出的隐私保护机制的开销相对较低.

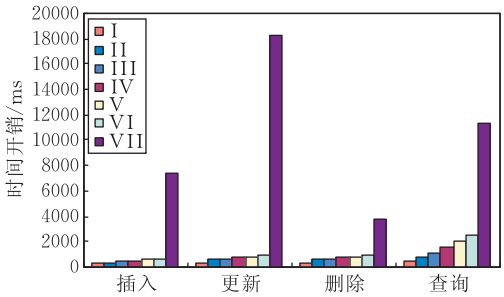


图 7　数据操作时间开销

7　总　结

本文针对 SaaS 模式下业务应用和数据库都部署在非完全可信的服务运营商端的情况，研究明文状态下的数据组合隐私保护机制，提出一种基于分块的 (k,α,β,γ) 隐私保护机制，根据租户定制的隐私约束将租户数据中的组合隐私分解到不同的数据分块中，隐藏数据之间的关联关系，实现了明文状态下的 SaaS 数据隐私保护；通过插入伪造数据的方式，确保 k 个不同数据分块满足均衡化，防止服务运营商泄漏租户的数据组合隐私；构建了数据对象逻辑视图的重构机制，在保证数据组合隐私的前提下确保 SaaS 数据的高实用性. 该隐私保护机制能够保护明文状态下的数据组合隐私，同时确保了数据处理的高效性和实用性，实现了数据隐私保护与实用性的有效结合.

未来工作主要包括：基于数据组合隐私保护的解决，研究单隐私属性的保护方法；研究面向 SaaS 应用的数据组合隐私保护机制中三方协作的安全机制，防止非完全可信的服务运营商冒充合法用户重构数据泄漏隐私；研究面向 SaaS 应用的隐私感知的数据优化处理技术，在保护数据隐私的同时提高数据处理效率，减少隐私保护机制给 SaaS 应用带来的数据处理开销，提高用户体验；根据 SaaS 模式的多租户特征，如何利用其他租户的隐私定制结果进行协同过滤推荐，辅助租户进行个性化隐私定制，降低隐私保护的复杂度和难度.

参　考　文　献

[1] Hacigümüs H, Mehrotra S, Iyer B. Providing database as a service//Proceedings of the 18th International Conference on Data Engineering (ICDE). San Jose, California, USA, 2002: 29-38

[2] Zhou Shui-Geng, Li Feng, Tao Yu-Fe, Xiao Xiao-Kui. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5): 847-861(in Chinese)
(周水庚，李丰，陶宇飞，肖小奎. 面向数据库应用的隐私保护研究综述. 计算机学报，2009，32(5)：847-861)

[3] Tian Xiu-Xia, Wang Xiao-Ling, Gao Ming, Zhou Ao-Ying. Database as a service — Security and privacy preserving. Journal of Software, 2010, 21(5): 991-1006(in Chinese)
(田秀霞，王晓玲，高明，周傲英. 数据库服务——安全与隐私保护. 软件学报，2010，21(5)：991-1006)

[4] Motro A, Parisi-Presicce F. Blind custodians: A database service architecture that supports privacy without encryption//Sushil J, Duminda W eds. Proceedings of the 19th An-

nual IFIP WG 11.3 Working Conference on Data and Applications Security. LNCS 3654. Berlin: Springer, 2005: 338-352

[5] Liu Yu-Bao, Huang Zhi-Lan, Ada Wai-Chee-Fu, Yin Jian. A data privacy preservation method based on lossy decomposition. Journal of Computer Research and Development, 2009, 46(7): 1217-1225(in Chinese)
(刘玉葆, 黄志兰, 傅慰慈, 印鉴. 基于有损分解的数据隐私保护方法. 计算机研究与发展, 2009, 46(7): 1217-1225)

[6] Li Xian-Wei, Liu Guo-Hua, Yuan Ying, Ma Hui-Dong. A database encryption method based on information dissociation and association. Computer Engineering & Science, 2007, 29(10): 54-56, 60(in Chinese)
(李现伟, 刘国华, 苑迎, 麻会东. 一种基于信息分解与合成的数据库加密方法. 计算机工程与科学, 2007, 29(10): 54-56, 60)

[7] Tao Yufei, Pei Jian, Li Jiexing, Xiao Xiaokui, Yi Ke, Xing Zhengzheng. Correlation hiding by independence masking//Proceedings of the 26th International Conference on Data Engineering (ICDE). Long Beach, California, USA, 2010: 964-967

[8] Ciriani V, Vimercati S, Foresti S, et al. Fragmentation and encryption to enforce privacy in data storage//Joachim B, Javier L eds. Proceedings of the 12th European Symposium on Research In Computer Security. LNCS 4734. Berlin: Springer, 2007: 171-186

[9] Ciriani V, Vimercati S, Foresti S, et al. Enforcing confidentiality constraints on sensitive databases with lightweight trusted clients//Ehud G, Jaideep V eds. Proceedings of the 23rd Annual IFIP WG 11.3 Working Conference on Data and Applications Security. LNCS 5645. Berlin: Springer, 2009: 225-239

[10] Mowbray M, Pearson S. A client-based privacy manager for cloud computing//Proceedings of the 4th International ICST Conference on Communication System Software and Middleware (COMSWARE). Dublin, Ireland, 2009

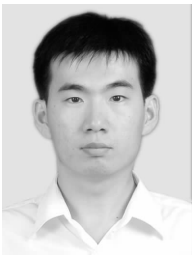
[11] Pearson S, Shen Y, Mowbray M. A privacy manager for cloud computing//Martin G J, Gansen Z, Chunming R eds. Proceedings of the 1st International Conference on Cloud Computing. LNCS 5931. Berlin: Springer, 2009: 90-106

[12] Gentry Craig. Fully homomorphic encryption using ideal lattices//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC). Bethesda, MD, USA, 2009: 169-178

[13] Itani W, Kayssi A, Chehab A. Privacy as a service: Privacy-aware data storage and processing in cloud computing architectures//Proceedings of the 8th IEEE International Symposium on Dependable, Autonomic and SecureComputing (DASC). Chengdu, China, 2009: 711-716

[14] Gu Lin, Cheung Shing-Chi. Constructing and testing privacy-aware services in cloud computing environment: Challenges and opportunities//Proceedings of the 1st Asia-Pacific Symposium on Internetware. Beijing, China, 2009: 1-10

[15] Zhang Kun, Shi Yuliang, Li Qingzhong, Bian Ji. Data privacy preserving mechanism based on tenant customization for SaaS//Proceedings of the International Conference on Multimedia Information Networking and Security (MINES). Wuhan, China, 2009: 599-603



ZHANG Kun, born in 1983, Ph. D. candidate. His research interests include service computing, trusted computing.

LI Qing-Zhong, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include database, trusted computing, semantic oriented application integration and information integration.

SHI Yu-Liang, born in 1978, Ph. D., lecturer. His research interests include service computing, trusted computing and database.

Background

Software as a Service, i. e. SaaS, came a new software delivery model with the development of network and maturity of application software. In SaaS model, business applications and databases are both deployed at the platform of untrustworthy service providers. Data privacy becomes the biggest challenges from adoption of SaaS model.

However, the traditional data privacy preserving, such as data encryption and data obfuscation are not applicable for SaaS model, because of the inefficiency of data processing. The technology and approaches proposed in privacy-preserving data mining and privacy-preserving data publishing are used for data analysis and not appropriate for transactional SaaS model.

To combine the data privacy preservation and data usability in SaaS model, the authors propose a data combination privacy preservation mechanism for SaaS based on the privacy leakage degree of different plain data combination in SaaS model. This mechanism supports customizing privacy constraint, which is used to describe the requirements of data combination privacy, and then fragments the SaaS data attribute into the different data chunks. Based on the trusted third party, association between data shares from different data chunks could be hidden and reconstructed, fake data are also used to assure the balancing of the data shares in data chunks.

The research is supported by the National Natural Science Foundation of China under Grant No. 90818001, the National Key Technologies R&D Program under Grant No. 2009BAH44B02, the Natural Science Foundation of Shandong Province of China under Grant Nos. ZR2010FQ026, 2009ZRB019YT and Y2007G38, the Key Technology R & D Program of Shandong Province under Grant No. 2010GGX10105.