关于外包数据库完整性验证的研究

刘 媛[1],涂晓东[1],张 兵[2]
(1.电子科技大学 通信与信息工程学院,四川 成都 611731;
2.新疆时代石油工程有限公司,新疆 克拉玛依 834000)

摘 要:数据库外包是一种重要的新兴的趋势,它让数据所有者把他们的数据管理工作委托给一个外部服务商。服务商管理客户的数据库,为客户提供安全可靠的机制来创建、存储、更新和访问他们自己的数据库。这种模式引出了数据安全的研究议题,文中通过对比几种不同的签名方式,提出了有效的数据完整性机制模型和安全有效的压缩RSA方案来确保数据的完整性和真实性。它在单一客户端和多查询者模型中运行良好,同时保证了计算量和带宽损耗在最低范围内。

关键词:外包数据库(ODB);完整性;压缩RSA
中图分类号:TP309.2 **文献标识码**:A **文章编号**:1673-629X(2010)05-0150-04

Research on Integrity Verification of Outsourcing Database

LIU Yuan[1], TU Xiao-dong[1], ZHANG Bing[2]
(1. School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. Xinjiang Times Petroleum Engineering Corporation, Karamay 834000, China)

Abstract: Database outsourcing is an important emerging trend which involves data owners delegating their data management needs to an external service provider. The service provider hosts clients' databases and offers seamless mechanisms to create, store, update and access their databases. This model introduces several research issues related to data security. By comparing the signatures of several different ways, conclude with effective mechanisms for data integrity model and a safe and efficient condensed-RSA program to ensure data integrity and authenticity. It works well in a unified client and muti-querier model, while ensuring that the computational and bandwidth overhead at the lowest range.

Key words: outsourced database; integrity; condensed-RSA

0 引 言

随着互联网的持续增长和网络技术的进步,数据库外包[1]成为一种新兴的趋势。如果让外部服务商来管理外包数据,企业就可以把精力放在更重要的工作上,只需通过互联网进行业务往来操作,而不用耗费大量的硬件、软件和人工成本来维护企业数据库。数据库外包就是这一趋势的重要表现。在外包数据库(ODB)模式中,服务商必须具备足够的软硬件和网络资源来容纳客户的数据库,而客户也能高效地创建、更新和访问外包数据。可是这种模式也带来很多问题,比如:整体性能,可用性和可扩展性等。目前最重要的问题就是怎样保证存储数据的安全性,服务商必须提供完备的安全措施来保护数据不受外部和服务商本身的恶意攻击。安全就意味着数据的完整性和私密性,尽管有关数据私密性的一些研究工作已经进行,但是如何有效保证数据的完整性还没有得到更多关注。这篇文章主要研究用最少的计算量和带宽损耗来保证数据可靠性和完整性的安全有效的方法。

1 系统模型

ODB模式是一个客户端-服务器模型。现实中,客户端可能是计算能力薄弱,存储空间有限的设备,如手机或无线掌上电脑。我们将分析以下三种ODB模式:最基本的就是ODB由一个单一的客户端实体来创建、操作和查询,把它称作单一客户端模式(见图1)。还有比较先进的多查询者模式(见图2),它包含两种类型的客户端:数据所有者和查询者,前者是真实数据的拥有者,负责新增、删除和更新数据库记录,而查询

收稿日期:2009-09-11;修回日期:2009-12-26
作者简介:刘 媛(1980-),女,硕士研究生,研究方向为存储网络技术、数据通信与计算机网络;涂晓东,副教授,博士,研究方向为通信网与宽带通信技术、存储网络技术、数据通信与计算机网络、嵌入式系统设计。

者只能对数据库或其中一部分进行读访问。第三种是最普遍的 ODB 模式,一个数据库可以有多个所有者,称为多所有者模式(见图 3)。多查询者和多所有者模式之间的差别很细微,前者是一个单独的安全主体创建和操纵数据库记录,后者是不同的安全主体创建不同的记录,而这两种模型都可以有多个查询者。

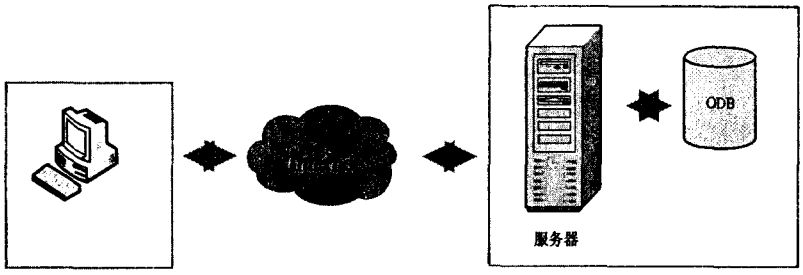

图 1　ODB 单一客户端模式

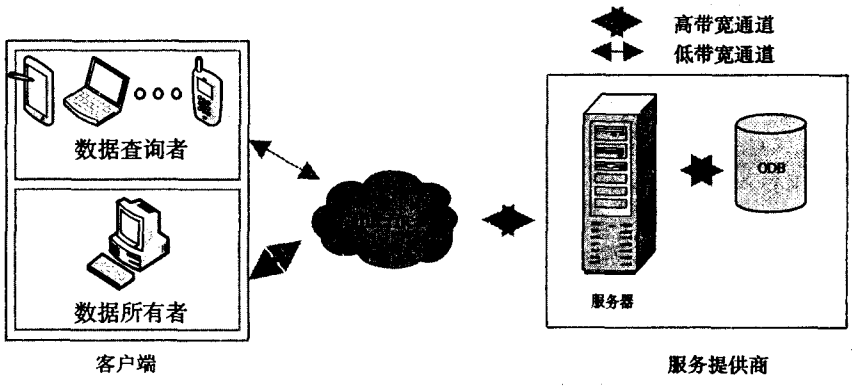

图 2　ODB 多查询者模式

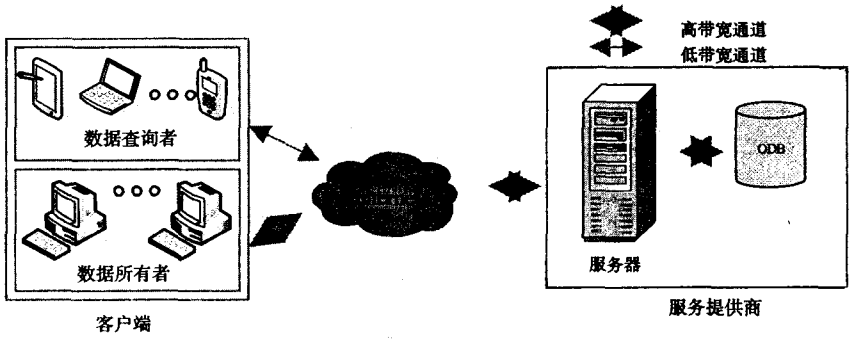

图 3　ODB 多所有者模式

2　假设条件和方案选择

为保证讨论的一般性,不区分存储数据被完全加密、部分加密或完全不加密的情况[2],只是简单假设数据所有者根据数据的性质以适当的形式存储数据,并制定可以远程执行的查询机制。使用数字签名[3,4]还是消息认证码(MAC－s)? 对记录水平的完整性来说,最自然直观的方案就是 MAC－s,因为它能高效计算和验证而且带宽损耗很小,但它只适用客户端和查询者是相同实体的单一客户端模式。在这儿要讨论的是更普遍的多查询者－多所有者模型,假设每个客户端有很多潜在查询者,在这种环境中 MAC－s 无法发挥作用,因为它们需要 MAC 密钥在所有的数据所有者和合法查询者之间共享,很显然这对查询者来说不可能实现。所以唯一的选择是使用公钥数字签名,然而数字签名的存储、带宽占用和计算需要很大的开销。

3　成本因素

具体目标是尽量减少 5 种成本因素(按重要性排列):

(1) 查询者的计算量:核实查询回复中的一组记录的完整性/真实性;

(2) 查询者消耗带宽:发送/接收完整数据(除了发送/接收回复包含的实际记录带来的开销);

(3) 服务器的计算量:查询回复中服务器端对信息完整性的操作;

(4) 数据所有者计算量:计算外包数据库中资料的完整性;

(5) 服务器存储能力:一个外包数据库需要存储完整资料的空间。

考虑上述因素,用数字签名机制来构建理想的或接近理想的解决方案,它允许多个单独的签名聚合(或合并)到一个统一的签名中,这样验证统一签名就相当于验证了每个单独签名。

4　适用的签名方案

4.1　压缩的 RSA

压缩 RSA 方法是标准 RSA 方法[3]的简单扩展。RSA 的著名特性之一是它的乘法同态性质,这种特性适合将单一签名者生成的签名聚合成一个"压缩"签名。在成功验证压缩签名后,验证者可以确保压缩签名涵盖了每个个体信息,通过拥有个人信息签名的任何一方"逐步"压缩 RSA 数字签名。

标准 RSA:一个实体拥有公钥 pk = (n,e) 和私钥 sk = (d),这里 n 是两个 $k/2$－bit 的随机大素数 p 与 q 的乘积,公钥和私钥的指数 $e,d \in Z_n^*$ 满足 $ed \equiv 1 \bmod \phi(n)$,这里 $\phi(n) = (p-1)(q-1)$,在现今的加密理论中 k 至少是 1024bits,RSA 加密系统的安全性还是依赖于大整数素因子分解。事实上,一个 RSA 签名是通过输入信息的哈希值计算出来的,让 $h()$ 指定一个合适的加密哈希函数(比如 MD5 或 SHA－1),它们

用一个变长的输入 m 产生一个定长输出 $h(m)$。信息 m 上的标准 RSA 签名计算如下：$\sigma = h(m)^d (\mathrm{mod}n)$。RSA 签名确认包括检查 $\sigma^e \equiv h(m) \mathrm{mod}n$，签名生成和确认都要计算指数。

压缩 RSA：给定 t 个输入消息 $\{m_1, \cdots, m_t\}$ 和它们的相应签名 $\{\sigma_1, \cdots, \sigma_t\}$（由同一个签名者生成），一个压缩 RSA 签名由单个签名的乘积得出：

$$\sigma_{1,t} = \prod_{i=1}^{t} \sigma_i (\mathrm{mod}n) \tag{1}$$

得出的签名 $\sigma_{1,t}$ 与标准 RSA 签名的大小相同。当验证一个压缩签名时，验证方需要把所有输入信息的哈希值相乘后再进行检查：

$$(\sigma_{1,t})^e \equiv \prod_{i=1}^{t} h(m_i) (\mathrm{mod}n) \tag{2}$$

RSA 签名的批量验证[5]：压缩的 RSA 验证与 RSA 的批量验证相似，批量帮助我们降低了配置中的计算复杂度，这样就要许多签名验证（或其它计算密集型工作）同时完成。RSA 的批量验证旨在通过减小指数的数值加快验证进程。给出一批签名集合 $\{\sigma_1, \cdots, \sigma_t\}$ 和不同的消息集合 $\{m_1, \cdots, m_t\}$，一个 RSA 批量验证（也叫快速筛选[7]）需要检查以下等式：

$$\left(\prod_{i=1}^{t} \sigma_i\right)^e \equiv \prod_{i=1}^{t} h(m_i) (\mathrm{mod}n) \tag{3}$$

压缩的 RSA 和 RSA 的批量验证之间的主要区别在于，后者是由有权访问个人签名的验证方计算出个人签名的乘积，如果批量验证失败还能进行验证后的安全审计，也就是它能筛选个人签名确定故障所在。相比之下，压缩的 RSA 只有一个单一的聚合签名，这使后续审计工作不能进行，因此在核查发生故障时，必须采取进一步措施（可能涉及脱机机制）。

压缩 RSA 的安全性：我们认为对自适应性选择消息攻击来说压缩 RSA 无法被伪造，因为如果一个对手 A 破坏了压缩 RSA，那么利用这个对手可以构造一个伪造者 B 成功创建批量签名集合通过批量验证测试，而这些都不需要在每个输入信息中拥有有效的个人签名。从定义一个对手破坏压缩 RSA 开始，如果针对满足等式(2)而对 t 个输入信息没有有效个人签名的消息集合 $\{m_1, \cdots, m_t\}$ 产生一个有效聚合签名，A 就能成功破坏压缩 RSA，通过论证说明压缩 RSA 的安全性至少像 RSA 批量验证一样安全，相反，在假设 RSA 是单向函数集合的情况下压缩 RSA 是安全的。注意：我们假设用全域哈希函数（FDH）[6]计算，FDH 是一个哈希函数 $H_{\mathrm{FDH}}: \{0,1\}^* \rightarrow Z_n^*$。

要点：A 默认输入 $(m_1, \cdots, m_t)$ 和 $\theta = (\sigma_1, \cdots, \sigma_\omega)$，这里 $\sigma_i = H_{\mathrm{FDH}}(m_i)^d \mathrm{mod}n$，并且 $\omega < t$，A 通过输出一个有效的压缩 RSA 签名 $\sigma_{1,t}$ 破坏压缩 RSA，现在构造一个伪造者 B 破坏 RSA 的批量验证。

细节：伪造者 B，在输入 $(m_1, \cdots, m_t)$ 和 θ 时输出 $(s_1, \cdots, s_t)$，这里 $(\prod_{i=1}^{t} s_i)^e \equiv \prod_{i=1}^{t} H_{\mathrm{FDH}}(m_i) (\mathrm{mod}n)$，B 由以下方式生成：

1. 产生随机数 $s_i \in_R Z_n^*$，在这里 $1 \leqslant i \leqslant (t-1)$。
2. 消息 $(m_1, \cdots, m_t)$ 和 θ 传到 A。
3. 让伪造的压缩 RSA 签名通过 A 返回记为 X。
4. 计算 $s_t = (\prod_{i=1}^{t-1} s_i)^{-1} * \mathrm{X} (\mathrm{mod}n)$。
5. 对应信息集 $(m_1, \cdots, m_t)$ 输出一组批量签名 $(s_1, \cdots, s_t)$。

论断 1：伪造者 B 产生一个签名集合满足批量验证测试。注意到 $\prod_{i=1}^{t} s_i \equiv \prod_{i=1}^{t} \sigma_i (\mathrm{mod}n)$，所以等式(3)满足，批量验证测试成功。

论断 2：研究表明在假设 RSA 是单向的情况下 RSA 的批量验证是安全的，那么压缩 RSA 至少能像批量验证 RSA 一样安全。这里得出：在 RSA 是单向函数集合的条件下，压缩 RSA 是安全的。

成本因素：比较压缩 RSA 和 RSA 批量验证的成本，在这里计算哈希值的成本可以忽略不计，因为它相对于整个模块的计算是微不足道的。在标准 RSA 签名中查询者需要接收处理 t 个签名，每个签名对应每条查询回复记录，为了验证这些签名，必须执行 t 次 RSA 验证，带宽消耗 $t * |n|$ bits。RSA 批量验证过程包括计算所有消息哈希值乘积和消息签名的乘积，做 $2(t-1)$ 次 RSA 验证，带宽消耗与标准 RSA 相同。比较而言，压缩 RSA 的带宽消耗只是一个单一的签名消耗（$|n|$ bits），计算所有消息哈希值的乘积只做 $(t-1)$ 次 RSA 验证，压缩 RSA 节省了 $(t-1) * |n|$ bits 带宽。

ODB 环境中的压缩 RSA：显然，压缩 RSA 适用于单一客户端模式和多查询者模式。服务器执行一个客户端查询需要执行以下操作：选择符合查询条件的记录；取得这些记录对应的签名；聚合这些签名（如上文所述将它们相乘），并和查询结果的记录一起以聚合签名形式发回。在多所有者模式中，服务器可以聚合每个签名者的签名并分别发送，这使查询者的带宽损耗与签名人数成线性关系，客户端可以通过验证每个签名者的一个签名来验证聚合签名。

4.2 BGLS

Boneh 等人提出了一个聚合签名方案，称作 BGLS[7]，它把不同消息上的不同签名者生成的签名聚合到一个基于椭圆曲线和双线性映射的简短签名中，这个方案在 Gap Diffie－Hellman(GDH)群里操作，下面简单介绍相关参数。

1. G_1 和 G_2 是两个 p 阶素数循环群;

2. g_1 是 G_1 的生成元,g_2 是 G_2 的生成元;

3. Ψ 是从 G_2 到 G_1 的可计算的同构体,$\Psi(g_2)=g_1$;

4. e 是可计算的双线性映射:$G_1\times G_2\rightarrow G_T$,这里 $|G_1|=|G_2|=|G_T|$,满足以下特性:

(1) 双线性:$\forall p_1,p_2\in G_1,q\in G_2,e(p_1p_2,q)=e(p_1,q)\cdot e(p_2,q)$及 $a,b\in \mathbf{Z}$,则 $e(p^a,q^b)=e(p,q)^{ab}$;

(2) 非退化性:$e(g_1,g_2)\neq 1$。

这两个特性意味着:$\forall p_1,p_2\in G_1,q\in G_2,e(p_1p_2,q)=e(p_1,q)\cdot e(p_2,q)$和 $\forall p,q\in G_2,e(\Psi(p),q)=e(\Psi(q),p)$。

BGLS方案:BGLS 利用全域哈希函数 $h():\{0,1\}^*\rightarrow G_1$,密钥生成包括挑选一个随机数 $x\in \mathbf{Z}_p$,计算 $v=g_2^x$。公钥中 $v\in G_2$,私钥 $x\in \mathbf{Z}_p$ 中。一个消息 m 的签名包括计算 $h=h(m)$,这里 $h\in G_1,\sigma=h^x$(实际签名是 σ)。为验证一个签名要计算 $h=h(m)$并检查 $e(\sigma,g_2)=e(h,v)$。

BGLS聚合:为聚合 t 个 BGLS 签名,第一要计算单个签名的乘积,如下:$\sigma_{1,t}=\prod_{i=1}^{t}\sigma_i$,这里 σ_i 对应消息 m_i 的签名,聚合签名 $\sigma_{1,t}$ 与一个单独 BGLS 签名大小相同,都是 $|p|$ bits。与压缩 RSA 类似,聚合可以被任何人执行很多次。

一个聚合 BGLS 签名 $\sigma_{1,t}$ 的验证包括计算所有消息哈希值的乘积,还要验证以下等式:$e(\sigma_{1,t},g_2)=\prod_{i=1}^{t}e(h_i,v_i)$。由双线性映射的特性,可以扩展等式左边如下:$e(\sigma_{1,t},g_2)=e(\prod_{i=1}^{t}h_i^{x_i},g_2)=\prod_{i=1}^{t}e(h_i,g_2)^{x_i}=\prod_{i=1}^{t}e(h_i,g_2^{x_i})=\prod_{i=1}^{t}e(h_i,v_i)$。

BGLS性能:分析 BGLS 签名验证的成本时,要区别两个操作:乘法和双线性映射计算。对于单一签名者的 BGLS 签名和 t 个输入信息,要计算双线性映射之后的消息哈希值的乘积($t-1$个数的乘法),验证成本会增加。对多个签名者的 BGLS 签名(有 k 个签名者和 t 个签名),计算($k*t-1$)个数的乘法就像 $k+1$ 个双线性映射一样验证成本也会增加。

ODB模式中的 BGLS:BGLS 对以上三种模式都是适用的。服务器通过选择符合查询条件的记录(包含签名)执行一个客户端查询操作,发回一个包含所有查询结果的聚合签名。BGLS 在单一客户端和多查询者模式下都能减少查询者计算量和带宽,在多所有者模式中,BGLS 能有效使用带宽,但它不能压缩太多查询者计算量。像上面提到的,BGLS 签名的验证包括 k 个不同签名者的信息,涉及计算 $k+1$ 次双线性映射,这是相当昂贵的。

5　讨　论

在这部分将比较以上两种签名方案的成本因素。用基本的加密操作(如:乘法,求倒,求幂)估计成本,通过对每个操作插入时间机制来显示方案产生的真实损耗。使用的测试平台是一个带开放 SSL 库[8]的 P3-977Mhz Linux 机器,在 RSA 中用 1024-bit 标准的 n,对 BGLS 用一个域 F_p,这里 $|p|=512$。用表 1 中的标记法记各成本项。假设查询回复结果包含 $k*t$ 条记录,这里 k 表示签名者数量(数据所有者),t 表示每个签名者生成的签名数(数据记录)。

表 1　标记法

QC	查询者计算量
QB	查询者消耗带宽
SC	服务器计算量
OC	数据所有者计算量
SS	服务器存储能力
$\text{Mult}^t(k)$	$\|k\|$ 的 t 次模乘
$\text{Exp}^t_l(k)$	$\|k\|$ 的 $\|l\|$ 次幂的 t 次模乘
$\text{Inv}^t(k)$	$\|k\|$ 的 t 次模倒数
$\text{BM}(t)$	t 次双线性映射

表 2 描述了每个方案执行加密操作产生的成本(计算量、存储空间和带宽),我们的主要目标就是使 QC(查询者计算)和 QB(查询者带宽)最小化。这个表提供了涵盖第 4 部分提到的成本因素的全部成本统计分析,按重要性递减的顺序排列。

表 2　ODB 模式中成本比较

比较项	压缩 RSA	BGLS
QC	$\text{Mult}^{k*(t-1)}(n)+\text{Exp}^k_e(n)$	$\text{Mult}^{k*t-1}(p)+\text{BM}(k+1)$
QB	$k*n$	p
SC	$\text{Mult}^{k*(t-1)}(n)$	$\text{Mult}^{k*t-1}(p)$
OC	$\text{Exp}^1_d(n)$	$\text{Exp}^1_0(p)$
SS	$k*t*n$	$k*t*p$

单一客户端和多查询者模式的签名者人数 $k=1$,这两种模式的压缩 RSA 和 BGLS 都有恒定的带宽需求,与签名数量无关,由于这两种模式都是含有 t 的乘法运算,QC 的开销与签名数量线性相关,而且验证一个聚合签名涉及到压缩 RSA 中的单一幂指数运算和 BGLS 中的双线性映射运算,这样压缩 RSA 更有效。

由于 QC 的开销与签名数量线性相关,在多所有者模式中,BGLS 产生恒定的 QB 开销,而在压缩 RSA 中 QB 开销只与签名者数量(k)线性相关。因为模乘运算的效率比双线性映射高得多,在 Z_n^*(这里 $|n|=1024$)中模乘运算的成本也低得多,因此压缩 RSA 比 BGLS 好些。

(下转第 157 页)

小为 19187 个字节,嵌入秘密文本信息后的载密文档字符数仍为 3806,大小仍为 19187 个字节。图 3 和图 4 分别为未嵌入秘密信息的载体文档和嵌入秘密信息的载密文档,图 5 为从载密文档中提取的秘密信息。

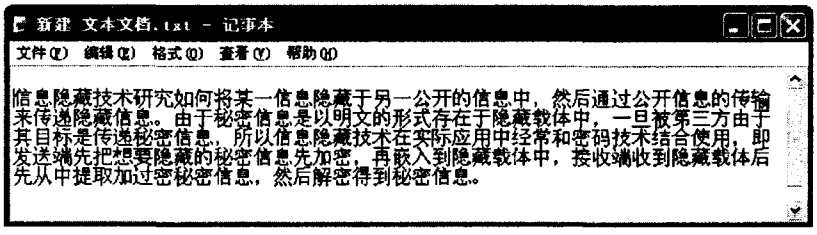

图 5　提取的秘密信息

从实验结果看,嵌入秘密信息后的载密文档与原始的载体文档视觉效果完全一样,感觉不到隐藏秘密信息的存在。

4　结束语

与其他文本文档隐藏方法相比,该隐藏算法具有实现简单、透明性好的优点,缺点是鲁棒性比较脆弱。虽然算法本身比较简单,但通过该算法可以看到,基于 Word 2007 文档的信息隐藏技术和基于 xml 文档的信息隐藏技术是有交叉的,后者的隐藏方法很可能在前者同样适用,这给以后研究基于 Word 2007 文档的信息隐藏提供了一个新思路。

参考文献:

[1] 王炳锡,陈　琦,邓峰森.数字水印技术[M].西安:西安电子科技大学出版社,2003.

[2] Walkthrough: Word 2007 XML Format [EB/OL]. 2008 - 08 - 25[2008 - 10 - 25]. http://msdn.microsoft.com/en-us/library/bb266220.aspx.

[3] Khodami A A, Yaghmaie K. Persian Text Watermarking [C]//PCM 2006. Berlin, Heidelberg: Springer - Verlag, 2006:927 - 934.

[4] Brassil J T, Low S, Maxemchuk N F. Copyright Protection for the Electronic Distribution of Text Documents[J]. Proceedings of IEEE, 1999, 87(7):1181 - 1196.

[5] 李向辉,钟　诚.提高 Word 文本文档信息隐藏容量的方法研究[J].计算机技术与发展,2006,16(9):97 - 99.

[6] 陈　萍,郭水旺,陈华丽.基于字体颜色的文本信息隐藏算法[J].科学技术与工程,2007,7(14):2544 - 2546.

[7] 耿红琴.基于 word 文本文档的信息隐藏技术研究[J].科学技术与工程,2007,7(11):2686 - 2688.

[8] 陈　芳,王　冰.基于文本字体的信息隐藏算法[J].计算机技术与发展,2006,16(1):20 - 22.

(上接第 153 页)

表 3 显示了在这两个方案中生成和验证一个单独签名,一个签名者的多个签名以及多个签名者的多个签名所需的真实时间(t 是签名数,k 是签名人数)。设 *RSA* 的公用指数 e = 3,用中国剩余定理加速签名,但是在任何验证操作中还没有最优化的技术。

表 3　成本比较:验证和签名

		压缩 RSA	BGLS
Sign	1 signature	6.82	3.54
Verify	1 signature	0.16	62
	t = 1000, k = 1	44.12	184.88
	t = 100, k = 10	45.16	463.88
	t = 1000, k = 10	441.1	1570.8

6　结束语

ODB 模式的数据库安全是一个较新的研究课题,笔者在此文中研究提供了有效的 ODB 数据完整性机制模型和安全有效的压缩 RSA 方案,它在单一客户端和多查询者模型中运行良好,只是还不能聚合不同签名者的签名,因此不适用多所有者模式。另一方面,虽然 BGLS 签名方案能把不同用户的签名聚合成一个短签名,但计算复杂度相当高。因此,今后工作的重点就是研究适合多所有者模式的有效实用的签名方案。

参考文献:

[1] Hacigümüs H, Iyer B, Mehrotra S. Providing Database as a Service[C]//In International Conference on Data Engineering. Washington: IEEE Computer Society, 2002:29 - 40.

[2] 赵晓峰,叶　震.几种数据库加密方法的研究与比较[J].计算机技术与发展,2007,17(2):219 - 222.

[3] 徐茂智,游　林.信息安全与密码学[M].北京:清华大学出版社,2007.

[4] 王平水,赵俊杰.多用户环境中签名方案的安全性研究[J].计算机技术与发展,2009,19(1):157 - 160.

[5] Bellare M, Garay J, Rabin T. Fast batch verification for modular exponentiation and digital signatures[C]//In Advances in Cryptology - EUROCRYPT '98, LNCS1403. Berlin: Springer - Verlag, 1998: 191 - 204.

[6] Bellare M, Rogaway P. Random oracles are practical: a paradigm for designing efficient protocols[M]. New York: ACM Press, 1993: 62 - 73.

[7] Boneh D, Gentry C, Lynn B, et al. Aggregate and Verifiably Encrypted Signatures from Bilinear Maps[C]//In Advances in Cryptology - EUROCRYPT '2003, LNCS2656. Berlin: Springer - Verlag, 2003:416 - 432.

[8] OpenSSL Project[EB/OL]. 2009 - 04 - 21. http://www.openssl.org.