计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):257−260 [doi: 10.15888/j.cnki.csa.005881]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

外包数据隐私保护环境中数据完整性检测协议①

廖　勇, 方　睿, 方　欣, 周　敏, 吴　强

(成都信息工程大学 信息安全工程学院, 成都 610225)

摘　要: 随着大规模云计算数据中心服务器在全球范围内的广泛部署, 其低投入, 可扩展性强等特点, 为这些拥有大数据量的公司或事业单位存储数据提供了便利, 并节约了构建 IT 环境的资金成本, 但是在这种数据外包环境下, 这必将涉及到信息安全与隐私保护问题. 在保证安全性和准确性的前提下, 本文提出了基于保护隐私的数据外包存储模型. 与数据隐私保护挖掘算法相结合, 提出了基于数据通信的数据完整性检测协议, 该协议使用数据安全技术, 从安全的多角度出发, 包括数据传输安全, 数据挖掘安全, 完整性安全等, 使得第三方服务器变得可信.

关键词: 外包数据; 安全; 完整性协议; 数据挖掘; 隐私保护

引用格式: 廖勇,方睿,方欣,周敏,吴强.外包数据隐私保护环境中数据完整性检测协议.计算机系统应用,2017,26(8):257–260. http://www.c-s-a.org.cn/1003-3254/5881.html

Data Integrity Checking Protocol in the Privacy Protection Environment of Outsourcing Data

LIAO Yong, FANG Rui, FANG Xin, ZHOU Min, WU Qiang

(Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: With the large-scale cloud computing data center server globally deployed, its advantages such as low cost ,strong expansibility etc. provide convenience for those companies or business units with large amount of data ,and save the cost of capital on the construction of the IT environment. But in this data outsourcing environment, this will involve information security and privacy protection. On the premise of guaranteeing the safety and accuracy, this paper presents an outsourcing data storage model based on the protection of privacy .Combined with data privacy protection of mining algorithm, it puts forward the data integrity detection protocol based on data communication. The protocol uses data security technology, starting from the angles of safety, including the security of data transmission, data mining safety, integrity, security etc, making the third party server trustful.

Key words: outsourcing data; security; integrity protocol; data mining; privacy protection

随着云计算技术的发展以及数据中心服务器的部署, 对于大数据量的企业或事业单位提供一种其低投入, 高安全, 可扩展性强的数据存储方法, 为这些单位存储数据提供了便利, 同时在 IT 资源方面节约了成本. 而解决用户存储和计算任务抽象外包给第三方不可信服务器上时出现的隐私安全问题, 是这些单位愿意把数据存储到第三方服务器上的前提. 对于大数据量的外包数据, 可使用数据挖掘技术来获取有益于单位的知识, 但同时对个人隐私以及团体隐私也造成了很大的危害, 因此隐私与知识形成了对立的矛盾, 所以要保护数据隐私前提下进行有效的数据挖掘, 即带有隐私保护的数据挖掘方法(privacy-preserving data mining, PPDM)已成为研究热点[1], 而保证数据完整性是数据外包的前提.

1　基于隐私保护的外包数据存储模型

根据外包数据挖掘项目的隐私保护需求、涉及的参与方、项目流程和所需要实现的数据隐私保护要求,

① 收稿时间: 2016-12-01; 采用时间: 2016-12-19

本文提出了基于隐私保护的外包动态数据存储模型, 如图 1 所示.

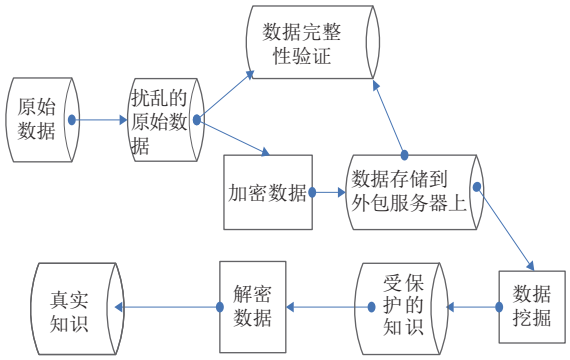

图 1 隐私保护的外包数据存储模型

数据存储在第三方服务器上, 存在隐私安全包括原始数据安全, 数据传输安全, 数据挖掘安全, 数据完整性安全等[1-3].

该项目流程共有两个参与方, 分别为拥有原始数据记录集的所有方和被委托进行数据存储的第三方. 流程主要分为安全传输和数据安全存储以及数据处理等功能, 整个数据安全存储第三方处理模块分为四个模块, 保证数据存储在第三方是安全可靠的. 第一个模块是实施在所有方对原始数据进行数据挖掘隐私保护算法处理, 以保证原始数据的隐私性, 同时又需要保证变换后的数据集合依然可以进行数据挖掘分析; 第二个模块是对处理后的原始数据进行随机扰动和反扰动的加密处理, 以及数字签名, 来保证数据安全传输以及验证数据获取方是所有方存储数据的第三方; 第三个模块是对存储在第三方的数据进行完整性检测, 保证数据存储在第三方服务器上没有被更改; 第四个模块是第三方将保护的挖掘结果回传到所有方后, 由其进行解密交换, 从而得到真实的挖掘结果.

2 研究现状

目前已经提出很多隐私保护数据挖掘算法, 而这些算法基本上是基于: (1) 处理原始数据中的敏感信息; (2) 排除可以通过数据挖掘得到的敏感知识. 从而达到数据经过数据挖掘之后敏感数据和知识不会被泄露, 但又能完成正确的数据挖掘. 例如面向数据外包中的隐私保护算法的研究中, 文献[1]提出的外包聚类挖掘隐私保护技术的的设计, 文献[2]提出了基于布隆过滤器的外包关联规则挖掘算法的保护技术改进等算法. 将数据存储在不信任的第三方服务器上, 对原始数据进行隐私保护算法处理后不破坏数据规则效能, 然后外包公司的相关人才进行数据挖掘后获取到未知的、潜在的、有价值的规则, 但这些规则对于外包数据第三方服务器的公司是保密的.

数据完整性是保证外包服务商诚实地按照合约提供存储服务, 确保存储在不可信第三方服务器上的用户数据没有被恶意地伪造、篡改或删除. 目前研究成果主要集中在: 数据持有性证明技术(PDP)和数据恢复性技术(PDR)两个方面. 前者是一种用于远程完整性验证的方案, 可以检测到外包数据中大于某个比例的数据损坏. 文献[4]基于前人研究的基础上提出了 B+树和 SBT 两种不同数据完整性验证机制的设计. B+数对于动态数据集可以达到更高效更新时间, 可以将一个结点中的所有元素与硬盘中的同一个物理分区相对应, 有利于 I/O 操作更加高效, 但 B+树的结果难于编码, 将数据存储在叶子结点之中, 其他上层结点并不存储实际数据, 只进行检测数据完整性而无法进行纠错处理; 而 SBT 打破了将数据存储于叶子结点的结构局限, 减少了存储空间, 哈希函数简单高效, 但也无法弥补数据无法纠错的缺陷, 同时由于哈希函数固定易进行重放攻击[5]. 文献[4]提出的 POR 方案是典型的基于哨兵技术的方案, 在原始数据中插入不可区分的哨兵, 通过检测哨兵的有效性和正确性, 确信服务器存储的是原始数据, 并在一定范围内进行纠错, 由于是在原始数据中插入哨兵, 破坏了原始数据的结构性, 不能通过数据挖掘获取到有用知识.

从以上研究现状可知, 对于外包数据安全性而言, 目前研究现状主要对数据隐私挖掘保护算法与数据完整性检测算法分开研究, 而没有把这两者相结合, 因此都存在数据安全性问题. 本文主要研究是数据隐私挖掘算法对原始数据做出处理后, 将数据传输到第三方远程服务器上, 对扰乱后的数据进行数据完整性检测的完整性检测协议.

3 协议介绍

该协议是根据数据挖掘隐私保护算法处理后的数据具有安全性特点, 提出本地数据通过数据挖掘隐私保护算法处理后将数据安全可靠的传输到第三方服务器上, 并能对存储在第三方服务器上的数据进行完整

性检测, 验证数据的正确性, 防止拜占庭错误的出现. 能安全的将数据存储到第三方服务器上, 很好的预防网络窃听. 协议流程如图 2 所示.

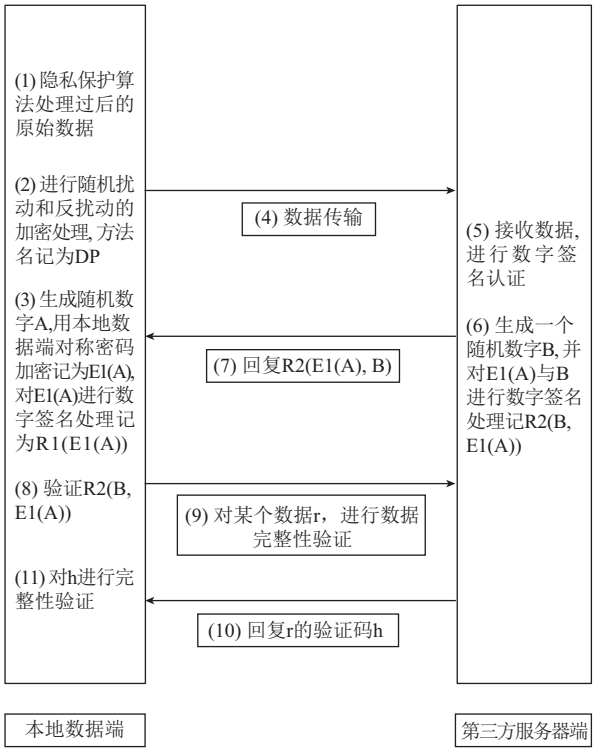

图 2 协议流程

(1) 本地数据通过隐私保护算法处理后, 将数据随机进栈 a, 然后每次出栈三个数看成三维坐标系中不同面的三个坐标, 思想源于 ART 算法[6].

(2) 然后将每个三个数据看成(a1, a2, a3)利用随机扰动和反扰动加密技术进行数据转换.

(3) A 用本地公钥进行加密, 进行数字签名, 本地所有的数据生成可进行完整性检测的平衡二叉树, 以便以后进行数据完整性检测.

(4) 开始传输数据, 数据包括反扰动方法、完整性检测约定参数、R1(E1(A)), 以及外包存储的数据, 并生成一个时间 t(该时间为两方协商后确定的时间), 防止数据丢失, 若时间为 0 时候, 说明数据丢失, 本地数据端继续发生该数据.

(5) 第三方获取到数据后, 进行数字签名认证, 证明发送数据方为协商的本地数据端, 获取到反扰动方法参数, 调用反扰动方法获取数据隐私挖掘算法处理后的原始数据, 存储到第三方服务器上, 并按双方完整性检测方案, 利用获取到的约定参数, 生成可完整性检测的平衡二叉树.

(6) 将 E1(A)与 B 用第三方服务器端的公钥加密, 记为 R2(E1(A), B).

(7) 将 R2(E1(A), B)传输到本地数据端.

(8) 本地数据端用第三方服务器的私钥解密, 获取 B 和 E1(A), 用本地端对称密钥对 E1(A)进行解密获取为 A, 并与 A 进行比较, 相同说明数据完好无损的存储在第三方服务器, 证明第三方服务器是本地数据端信任的, 同时取消时间 t. 若不同, 追究其第三方服务器端责任, 并重新发生该数据.

(9) 每隔一段时间, 本地端请求对存储在第三方服务器上的某个数据 r 进行完整性检测.

(10) 第三方服务器回复该数据的验证码 h.

(11) 对 h 与本地存储的该数据的验证码进行对比, 相同则说明存储在第三方服务器上的数据未被篡改或丢失. 否则, 追究第三方服务端责任.

4 协议分析

数据挖掘隐私保护算法平衡隐私保护和知识获取这对矛盾的可调节机制, 同时消除先验知识对隐私的威胁. 而数据完整性是存储在数据库中的数据值均正确, 要求数据在授权的情况下不被修改或丢弃. 该协议将两者结合起来, 既能达数据隐私的保护, 本地端又能获取到有用的知识, 同时能验证数据存储在第三方服务器上数据是完整的, 提高了效率.

4.1 安全分析

现有的数据挖掘隐私保护算法越来越成熟, 各种研究证明了这些算法对于数据隐私保护有良好的准确性和安全性[1,2], 存储在第三方服务器上数据是隐秘的. 通过该协议来检验通过数据挖掘隐私保护算法处理后的数据的完整性. 通过随机扰动和反扰动加密技术, 有效解决了网络窃听问题[6], 加入数字签名能相互认证是信任, 同时加入时间 t, 能有效避免数据丢失引起的数据完整性错误. 通过随机生成的 A 与 B, 对 A 进行对称加密生成 E1(A), 再对 E1(A)用本地端的公钥进行加密, 第三方服务器用本地端私钥进行解密, 证明数据传输方是约定的本地端, 进行了数字签名. 然而在整个传输过程中 A 都是加密的, 并且密钥只有本地端知道, 解密后与本地端存储的 A 进行比较, 从而说明数据在存储过程中未被更改, 存储方为约定的第三方服务器. 认证信息需要双方来认证, 服务器不会成为黑客攻击对

象. 即使服务器被攻破, 相对而言也是安全的.

平衡二叉树完整性检测方案从图 2 可知, 完整性检测安全性主要基于 hash 的安全性, 基于前人对 hash 函数的安全性研究, 文献[7]通过假设证明, 可知攻击者在不知 hash 函数的前提下, 篡改元组成功通过验证的概率可忽略.

该协议结合加密算法, 数字签名技术, 完整性检测方案, 隐私保护算法, 数据从本地端到存储到第三方服务器上, 依次对数据进行安全性保护, 保证数据存储在第三方服务器上是保密的, 可信的, 完整的.

4.2 效率分析

首先, 该协议不需要对所有数据进行加密处理, 仅对反扰动方法和数字签名认证的随机数字 A 进行加密处理, 明显能提高加密时间效率. 其实, 利用平衡二叉树进行完整性检测, 每个结点存储原始数据, 减少了服务器上的空间开销, 而且可以降低树的高度, 从而降低数据插入删除等基本操作的时间复杂度[3].

5 小结

通过对现有的外包数据隐私保护数据挖掘算法和数据完整性检测方法的研究与分析, 提出了基于保护隐私的数据外包存储模型, 该模型反映出数据存储在第三方服务器上面临的威胁包括有传输威胁、数据挖掘威胁、数据篡改或丢失威胁, 数据窃听威胁等. 只有解决了这些数据安全威胁, 本地端才可放心的存储在第三方服务器上. 因此, 基于这些安全威胁, 本文提出了外包数据隐私保护算法环境中基于数据通信的数据完整性检测协议, 对存储在第三方服务器上的数据在不丢失不篡改的情况下, 保证数据挖掘获取知识的正确性. 结合隐私保护算法的特点, 提出了该协议的数据流程, 从多角度出发, 保证数据流程是安全的, 可靠的. 通过协议分析, 可知该协议在整个数据处理过程中, 数据是安全的, 效率是高效的.

参考文献

1 任静涵. 外包数据挖掘隐私保护算法研究和改进[硕士学位论文]. 上海: 上海交通大学, 2009.
2 刘泓. 面向外包服务的关联规则挖掘隐私保护方法研究[硕士学位论文]. 重庆: 重庆大学, 2013.
3 耿纪昭. 云存储中数据完整性验证机制的研究与实现[硕士学位论文]. 成都: 电子科技大学, 2013.
4 张亮. 云存储数据完整性检测技术研究[硕士学位论文]. 大连: 大连理工大学, 2014.
5 孙岚, 吴英杰, 罗钊, 等. 路网环境下防止重放攻击的位置隐私保护算法. 华中科技大学学报(自然科学版), 2013, 41(S2): 285–290.
6 刘英超. 面向分布式的数据挖掘隐私保护方法研究[硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2013.
7 咸鹤群, 冯登国. 外包数据库模型中的完整性检测方案. 计算机研究与发展, 2010, 47(6): 1107–1115.
8 张二勇, 李云峰, 王玮. Surfer 软件绘图接口的开发及应用. 地下水, 2005, 27(3): 212–214.