

大数据安全保护技术综述

魏凯敏¹, 翁健¹, 任奎²

(1. 暨南大学网络空间安全学院, 广东 广州 510632;
2. 纽约州立大学布法罗分校计算机科学与工程学院, 纽约 布法罗 14260)

摘要: 目前, 大数据受到社会各界的广泛关注。受数据体量大、结构多样化、处理迅速快等因素影响, 大多数传统的数据安全保护技术不再适用于大数据环境, 着使得大数据安全问题日益严重。为此, 近些年提出了大量的大数据安全保护技术。从加密算法、完整性校验、访问控制技术、密文数据去重和可信删除、密文搜索等视角, 对当前大数据安全保护关键技术的研究现状进行分类阐述, 分析其优缺点, 并探讨它们未来发展趋势。

关键词: 大数据; 安全保护技术; 算法; 搜索

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00046

Data security and protection techniques in big data: a survey

WEI Kai-min¹, WENG Jian¹, REN Kui²

(1. School of Information Science and Technology, Jinan University, Guangzhou 510632, China;
2. Computer Science and Engineering, State University of New York at Buffalo, Buffalo 14260, United States)

Abstract: Big Data has attracted tremendous attention from all over the world nowadays. Its sheer volume, complex structure and realtime processing requirements often obsolete traditional technologies when coming to provide sufficient security protection. To address this challenge, significant research efforts have been carried out by the research community since recent years. Different technical aspects of big data security, including encryption algorithms, data integrity auditing, access control, secure data duplication, assured deletion, and secure search were surveyed, and in-depth discussions on their pros and cons were provided. Various future research directions were also discussed.

Key words: big data, data security and protection, algorithm, search

1 引言

随着社会信息化和网络化的快速发展, 物联网、互联网等信息技术在人类政治、经济、文化、生活等不同领域不断渗入和交叉融合, 所产生的数据超越已往任何年代所产生的数据总和, 且数据类型和相互关系复杂化。据 IBM 统计^[1], 目前, 世界上每天产生大约 250 亿字节的数据。国内互联网

公司腾讯, 其经压缩处理后的数据量达到 100 PB 左右, 并且这一数据还在以日增 200~300 TB, 月增 10% 的速率不断增长。

目前, 大数据受到政府机关、工业界和学术界的广泛关注和高度重视。美国政府将大数据比喻为“未来的新石油”, 一个国家拥有的数据规模和运用数据的能力将成为一个国家综合实力的象征, 对数据的占有和控制将深刻

收稿日期: 2015-03-09; 修回日期: 2015-04-02。通信作者: 翁健, cryptjweng@gmail.com

基金项目: 国家自然科学基金资助项目 (No.61272413, No.61472165, No.61502202); 教育部高等学校博士学科点专项基金资助项目 (No.20134401110011)

Foundation Items: The National Natural Science Foundation of China (No.61272413, No.61472165, No.61502202), The Research Fund for the Doctoral Program of Higher Education of China (No.20134401110011)

影响时代的发展。大数据具有着巨大的商业价值潜力,其表现出的数据整合与控制力量远远超过以往任何时代。比如在2009年,Google公司在大数据方面的业务,对美国经济的贡献就达540亿美元^[2]。此外,《Nature》、《Science》等学术期刊出版了大数据对经济、环境、互联网等人类社会各方面的影响和挑战的研究成果。

相对于传统的数据,大数据有5V特征^[3],即体量大(volume)、速度快(velocity)、多样化(variety)、难辨识(veracity)和价值密度低(value)。大数据规模不断增长,已经达到PB级别,未来将达到EB、ZB,这要求对大数据处理的效率必须高。大数据类型多样化,包含不同类型的数据,有结构化的、半结构化的和非结构化的,且数据真伪难以识别。这些特征使传统的数据安全保护方法无法直接应用在大数据环境中,给大数据研究和发展带来极大的挑战。

2 大数据安全保护的概念和安全需求

2.1 大数据概念和特征

什么是大数据,迄今并没有公认的定义。维基百科给出的大数据定义:大数据^[4]又称为海量数据,是指所涉及的数据量规模巨大到无法通过人工或者计算机,在合理时间内达到截取、管理、处理、并整理成为人类所能解读的形式信息。与传统数据相比,大数据具有如下几个特征^[3]。

1) 体量大。体量大是大数据的基本特征。目前,大数据规模从传统大型数据集的TB级增长到至少PB级。

2) 速度快。速度快是大数据区别于传统数据挖掘的显著特征。根据IDC预测,到2020年全球数据使用量将达到35.2ZB(1ZB=210EB)。在规模如此庞大的大数据中,数据处理效率特别关键。

3) 多样化。多样化是大数据的内在特性。大数据的数据类型多样化,有结构化的、非结构化数据的和半结构化的。结构化数据是指可以用二维表结构来逻辑表达实现的数据;非结构化数据

是指不能用数据库二维逻辑表来表现的数据,包括图片、音频、视频等。

4) 难辨识。数据真伪难辨是大数据应用的最大挑战。

5) 价值密度小。大数据由大量碎片化信息数据组合而成,而通过整合这些碎片化的数据能够挖掘到更多有价值的信息。大数据蕴含的巨大价值受到产业界、学术界和政府部门的高度关注与重视。

根据数据来源不同,大数据可以分为3类^[5]:

1) 人类活动,人在使用互联网(包括移动互联网)过程中所产生的各类数据;2) 计算机,各种计算机信息系统产生的数据,多以文件、数据库、多媒体等形式存在;3) 物理世界,各类数字设备所采集的数据,比如气象系统采集设备所收集的海量气象数据、视频监控系统产生的海量视频数据等。

2.2 大数据安全需求

目前,大数据受到严重的安全威胁,其安全需求主要体现如下几个方面。

1) 机密性

数据机密性是指数据不被非授权者、实体或进程利用或泄露的特性。为了保障大数据安全,数据常常被加密。常见的数据加密方法有公钥加密、私钥加密、代理重加密、广播加密、属性加密、同态加密等。然而,数据加密和解密会带来额外的计算开销。因此,理想的方式是使用尽可能小的计算开销带来可靠的数据机密性^[6]。

在大数据中,数据搜索是一个常用的操作,支持关键词搜索是大数据数据安全保护的一个重要方面。已有的支持搜索的加密只支持单关键字搜索,并且不支持搜索结果排序和模糊搜索。目前,这方面的研究集中在明文中的模糊搜索、支持排序的搜索和多关键字搜索等操作。如果是加密数据,用户需要把涉及到的数据密文发送回用户方解密之后再行,这将严重降低效率。

2) 完整性

数据完整性是指数据没有遭受以非授权方式的篡改或使用,以保证接收者收到的数据与

发送者发送的数据完全一致,确保数据的真实性。在大数据存储中,云是不可信的。因此,用户需要对其数据的完整性进行验证。远程数据完整性验证^[7]是解决云中数据完整性检验的方法,其能够在不下载用户数据的情况下,仅仅根据数据标识和服务器对于挑战码的响应对数据的完整性进行验证。此外,在数据流处理^[8]中,完整性验证主要来源于用户对云服务提供商的不信任性。在这种情况下,确保数据处理结果的完整性也是至关重要的。

3) 访问控制

在保障大数据安全时,必须防止非法用户对非授权的资源和数据等的访问、使用、修改和删除等各种操作,以及细粒度的控制合法用户的访问权限。因此,对用户的访问行为进行有效地验证是大数据安全保护的一个重要方面。

3 大数据安全保护技术研究进展和未来趋势

3.1 加密算法

为了保障大数据的机密性,使用加密算法对数据加密。传统的 DES、AES 等对称加密手段,虽能保证对存储数据的加解密速度,但其密钥管理较为复杂,不适合有大量用户的大数据环境中。而传统的 RSA 等非对称加密手段,虽然其密钥易于管理,但算法计算量太大,不适用于对不断增长的大数据进行加解密。数据加密增加了计算开销,且限制了数据的使用和共享,造成了高价值数据的浪费。因此,开发快速加解密技术成为当前大数据安全保护技术的一个重要研究方向。

3.1.1 属性加密

Shai 等^[9]提出了第 1 个属性加密方案,公钥、私钥都和属性相关联。当用户私钥具备解密数据的基本属性时,用户才能够解密出数据明文。例如:用户 1 的私钥有 a、b 2 个属性,用户 2 的私钥有 a、c 2 个属性,若有一份密文解密的基本属性要求为 a 或 b,则用户 1 和用户 2 都可以解密出明文;同样,若密文解密的基本属性要求为 a 和 b,则用户 1 可以解密出明文,而用户 2 无法解密此密文。Goyal 等^[10]提出了一个细粒度访问

控制的属性加密方案,在私钥当中嵌入接入策略。只有当用户属性满足接入策略时,密钥才可以恢复,从而解密消息。该方案可以支持任意单调的包含与/或限门的接入公式。Yu 等^[11]则提出了一个安全、可扩展的细粒度访问控制的属性加密方案。Li 等^[12]提出一个支持用户审计的细粒度访问控制的属性加密方案,用来防止大数据中非法的密钥共享。

Bethencourt 等^[13]提出了密文策略的属性加密方法,将接入策略嵌入在密文当中,而解密私钥只与属性集合相关。当密文的接入策略发生改变时,密文重新加密,且无需重新分配属性对应的解密私钥。该方法需要一个属性授权中心,对属性以及属性对应的解密私钥进行管理。Chase^[14]提出了一个多授权中心属性加密系统,每个授权中心管理不同的属性域。当属性撤销或者属性中一个用户撤销时,密钥更新就产生问题。Hur 等^[15]讨论了这一问题并分析了现有撤销方案的不足,其提出的方案引入了对称密钥的群组密钥管理方案。

基于属性加密和解密的计算比较复杂,通常涉及多个双线性对运算,特别是接入策略的表述比较复杂的情况,将部分计算外包到云服务提供商有助于减少数据使用者终端的开销。Green 等^[16]给出了一个属性加密的外包解密方案,将复杂的解密操作由云服务提供商转化为一个普通的 ElGamal 解密问题,终端只需要一次模指数运算,可有效降低终端的解密计算量。

大数据中属性加密的研究重点,不是加密技术本身,而是如何提高服务效率与质量。目前,基于属性的加密算法的时间复杂度很高,使这种加密方式在大数据中的应用并不广泛。随着大数据研究的进一步深入,加密算法的时间复杂度进一步降低,属性加密将应用在未来的大数据中。

3.1.2 代理重加密

代理重加密算法(PRE, proxy re-encryption)指允许第三方改变数据发送方加密后的密文,使数据接收方可以解密,而第三方并不知道原文。代理重加密的一般化流程如图 1 所示。

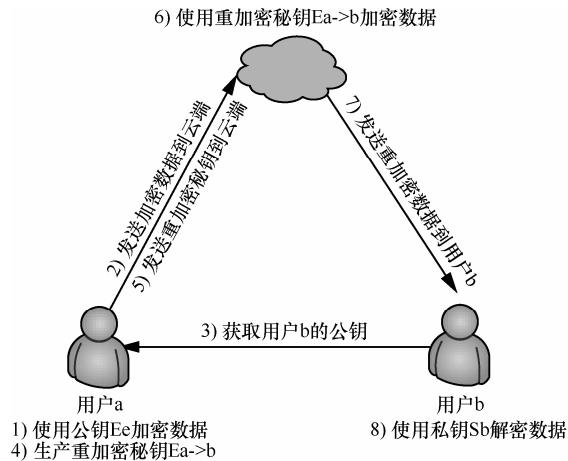


图 1 代理重加密的一般化流程

1988 年, Blaze 等^[17]提出了第一个代理重加密方案, 该方案允许一个半可信的代理者将数据发送者的密文转换为数据接受者可以解密的密文, 同时不泄漏数据发送者的明文消息。Green 等^[18]提出了一种基于身份的代理重加密方案 (IBPRE, identity-based proxy re-encryption), 该方案以用户的唯一身份信息作为公钥参与重加密, 具有单向性、非传递性、非交互性等特点。Mizuno 等^[19]对 IBPRE 方案进行改进, 优化了重加密密文空间的大小并隐藏了代理的身份, 同时证明其具有抵抗选择明文攻击 (CPA, chosen-plaintext attack) 的能力。Zhao 等^[20]提出了分类代理重加密技术, 该技术使数据分发者能够对密文委托权实施细粒度的分类控制。Wu 等^[21]则给出了无证书的代理重加密算法以及基于身份的密钥托管协议。Liang 等^[22]研究了基于身份的可撤销代理重加密机制。

Weng 等^[23]提出的条件代理重加密方案 (CPRE, conditional proxy re-encryption) 引入了访问控制机制, 只有当重加密密钥和指定密文条件同时满足时, 解密操作才被允许。在此基础上, Fang 等提出了支持关键词检索的匿名条件代理重加密方案^[24]和模糊条件代理重加密方案^[25], 提高了算法的执行效率。Chow 等^[26]提出随机预言机模型下的单向代理重加密方案。Lan 等^[27]利用秘密共享机制和双线性对原理构造出多条件代理重加密方案。目前, 大部分现有方案大多仅限于关键字条件, 如何构造支持布尔条件的条件代理重加密算法仍需进一步研究。

3.1.3 同态加密

同态加密 (homomorphic encryption)^[28]是一种加密形式, 对密文做特定的代数运算后得到加密的结果, 与对明文同样的运算结果一样。同态加密可以无需对数据解密而进行各种操作, 比如对加密数据的检索、比较等, 从根本上解决将数据及其操作委托给第三方时的保密问题。因此, 同态加密可以运用在各种云计算中。

2009 年, IBM 研究人员 Gentry^[29]使用“理想格 (ideal lattice)”构建了全同态数据加密方案。该方案包括 3 个关键步骤: 1) 构建一个受限同态加密算法, 该算法支持密文的低阶多项式运算; 2) 将解密操作“打散”成低阶多项式运算; 3) 将受限同态加密算法转变成全同态加密算法。该方案密文处理效率很低, 离实际应用还有较长时间。在此基础上, Smart 等^[30]利用中国剩余定理设计了一个完全同态加密方案。在该方案中, 密钥和消息长度都较小。采用明文打包的方法, Gentry 等^[31]将 R-LWE 问题设计为一个时间开销仅为多项式对数 (polylog) 的完全同态加密方案。之后, Gentry 等^[32]提出了一个完全同态加密算法。该算法将模设置为 2 的幂的近似值, 以提高效率。这些完全同态加密方案使用的明文打包技术仅限于 R-LWE 问题, Brakerski 等^[33]提出了一种基于标准 LWE 问题的完全同态加密方案。该方案简单且更安全。

上述完全同态加密方案都是基于理想格中的一种判定问题和稀疏子集求和问题, 而这 2 个问题只能规约到平均情况下的困难问题。为了提高完全同态加密方案的安全性, Gentry^[34]设计了密钥生成算法, 该算法将完全同态加密方案的安全性建立在稀疏子集求和问题和理想格中一种最坏情况下的困难问题上。Brakerski 等^[35]基于标准 LWE 问题设计了新的完全同态加密方案, 该方案有更高的安全性, 因为 LWE 问题都可规约到最坏困难问题。Lopez 等^[36]提出了一种允许多个密钥参与的完全同态加密方案。该方案基于理想格, 比传统的完全同态加密方案更加灵活、实用。

与基于格问题的全同态加密不同, Van 等^[37]在整数环上设计了一个新的完全同态加密方案, 其安全性依赖于近似最大公约数问题。由于 Van 的

方案中公钥过大, Coron 等^[38]提出了一个改进方案, 该方案具有较短的公钥, 但是安全性基于较强的近似最大公约数假设。为了处理整数环上的完全同态加密, Cheon 等^[39]提出了批处理完全同态加密算法, 该算法的安全性分别依赖于判定近似最大公约数问题和无误近似最大公约数问题。

加密算法是实现大数据安全保护与共享的基础。面对日益增长的大数据, 现有加密算法在加解密效率、密钥管理等方面有着明显的不足。已有研究表明, 完全同态加密算法和基于 LWE 问题的部分同态加密算法能解决大数据安全保护的计算问题, 但是这些算法需要进行大量复杂的指数运算, 大大降低了数据的处理效率。因此, 提高计算效率将是同态加密算法研究的重要方向。

3.2 完整性校验

当大数据存储到云端之后, 用户就失去了对数据的控制权。用户最关心的是, 如果云服务商不可信, 所存储的文件是否被篡改、丢弃等。解决这个问题的最简单方式就是将其全部取回检查, 但该方法不可取, 因为要耗费大量的网络带宽, 特别是当云端数据量非常大时。当前, 对云端大数据完整性进行校验主要依靠第三方来完成。根据是否允许恢复原始数据, 当前的数据完整性校验协议主要可以分为 2 类: 只验证数据完整性的 PDP 协议和允许恢复数据的 POR 协议。数据完整性验证的一般化过程如图 2 所示。

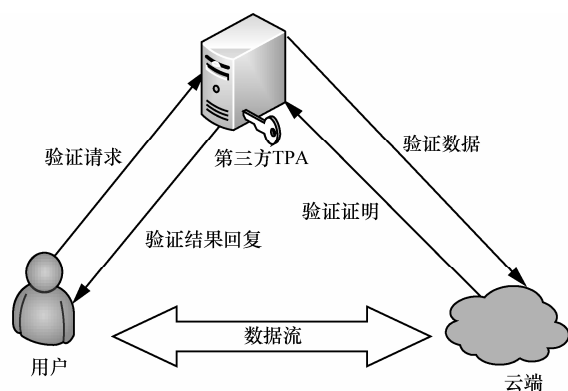


图2 数据完整性验证的一般化过程

Ateniese 等^[40]提出了一种可证明的数据持有 (PDP, provable data possession) 协议。该协议是基于 RSA 问题设计的, 它先从服务器上随机采样相应的数据块, 并生成持有数据的概率证据。

客户端维持着一定数量的元数据, 并利用元数据来对证据进行验证。由于 PDP 协议的通信和存储开销较大, Wang 等^[41]提出了一种支持第三方验证的 PDP 协议, 该协议使用了双线性配对技术基于离散对数问题。Zhu 等^[42]提出了一种支持数据动态变化的第三方验证 PDP 协议。该协议使用双线性配对技术和 Index-hash 表, 允许无限次挑战询问并能够保护数据拥有者的隐私。由于该协议的计算和通信开销较大, Yang 等^[43]利用双线性配对的特点, 设计了一个高效的第三方审计 PDP 协议。该协议能够对用户存在多个云端的数据进行批量校验, 并支持数据的动态操作和审计机构进行无限次挑战询问。

Juels 等^[44]提出可恢复证明 (POR, proof of retrievability) 协议, 该协议使用纠错码技术和消息认证机制来保证远程数据文件的完整性和可恢复性。在该协议中, 原始文件首先被纠错码编码并产生对应标签, 编码后的文件及标签被存储在服务器上。当用户选择服务器上的某个文件块时, 可以采用纠错码解码算法来恢复原始文件。在 POR 基础上, Shacham 等^[45]提出了一个基于双线性对构造的数字签名方案, 支持无限次挑战询问。为了提高效率, Dodis 等^[46]提出了一种允许第三方审计的 POR 协议。该协议使用了困难放大技术, 允许无限次挑战询问。为了支持对数据的动态操作, Wang 等^[47]提出了一个基于短签名和散列的 POR 协议。在此基础上, Wang 等^[48]提出一个允许对多个数据拥有者的数据进行批量校验的 POR 协议, 但这些协议都不能保护数据拥有者的隐私。

目前, 大数据完整性校验算法还不能支持数据动态变化。与 PDP 算法相比, POR 算法具有数据恢复功能和更高的实用性。因此, 研究支持数据动态变化的 POR 算法将是大数据安全保护的研究要点。此外, 数据可能属于不同的所有者且数据规模庞大, 研究支持多主权大数据完整性批量校验也将是未来大数据完整性校验协议的发展趋势。

3.3 访问控制

Sandhur 等^[49]提出了基于角色的访问控制 (role-based access control) 方法, 不同角色赋予不同的访问控制权限。针对云端大数据的时空关联性,

Ray等^[50]提出了LARB(location-aware role-based)访问控制协议,在RBAC(role-based policies access control)的基础上引入了位置信息,通过用户的位置来判断用户是否具有数据访问权限。Zhang等^[51]提出的基于尺度的时空RBAC访问控制模型,使访问控制策略的表达能力得到增强,同时也增强了模型的安全性。

基于属性的访问控制(ABAC, attribute based access control)^[52]是通过综合考虑各类属性,如用户属性、资源属性、环境属性等,来设定用户的访问权限。相对于RBAC以用户为中心,ABAC则是考虑全方位属性,以实现更加细粒度的访问控制。在基于密文属性的加密基础上,Sun等^[53]提出了一种数据安全访问控制方法,该方法将公钥和私钥形式化为读写权限、通过设计密钥来进行访问控制。Ruj等^[54]提出了一种可对云中大数据实现隐私保护和认证的访问控制框架。在这个框架中,在数据存往云端之前对其进行认证,而认证过的用户可以在云端对数据进行ABE,加密存储。

当用户的属性发生变化时,用户就失去了对该加密数据的访问权限。针对这个问题,Liang等^[55]提出基于代理重加密的访问控制方法,该方法通过代理将密文从一种访问结构树加密为另一种访问结构树加密,以达到权限撤销的目的。Hong等^[56]利用CP-ABE算法和公钥密码系统来实现密文访问控制,该方案不足的是,数据拥有者仍然要承受巨大的重加密代价。Yu等^[57]提出了一种基于CP-ABE算法的属性撤销方案,该方案假定云服务商是存在一定可信度的,数据所有者将部分工作交付云服务商执行。在此基础上,Yu等^[58]提出了一种基于KP-ABE技术的细粒度的访问控制,并使用基于重加密技术实现了有效的用户撤销机制。

ABE在授权的时候,用户的每个属性都向授权中心申请签名私钥,这给单个授权中心增加很多工作负担。为此,Chas^[59]首次提出了多授权中心,要求由不同的认证中心来认证每个用户保存属性或访问树,并使用一个可信的中心授权方来管理和约束各个授权中心。在基于多授权中心的属性加密方法中,不同的授权中心管理用户的多

个属性,并对每个属性分配加密密钥。为了防止多个用户勾结非法访问数据,Yang等^[60]提出了一个云环境中多授权中心访问控制模型。在该模型中,可信任的证书签发机构给每个用户分配唯一的用户标识符和唯一的授权标识符。只有通过证书签发机构认证过的用户标识符和密钥一起来使用时才能解密数据,从而保证数据的安全性。在CP-ABE的基础上,Liu等^[61]通过加入属性基签名,提出了一种层次化的属性基访问控制方案。该方案分层处理multi authority,然后每层authority使用不同的CP-ABE算法,以实现粗粒度的访问控制。

大数据在给传统访问控制带来挑战的同时,也带来了机遇。随着大数据的规模不断增长,并在不同领域的应用,将有更多的数据在不同系统中流转,研究可耦合的细粒度访问控制技术迫在眉睫。此外,在大数据中,不同数据的功能和安全需求是不一致的,研究多层次和多级安全的访问控制新技术将是未来大数据访问控制技术的发展方向。

3.4 密文数据去重与可信删除

3.4.1 密文数据去重

存储在云端大数据有很多重复的、冗余的。为了节省存储空间和降低成本,一些重复数据删除(data deduplication)技术^[62]被用来删除在云端的大量重复数据。在云环境中,数据往往是被加密成密文存储,且相同的数据会被加密成不同的密文。因此,很难根据数据内容对重复的安全数据进行删除。密文数据去重技术是近年来数据安全领域中新兴的研究热点,其不仅可以节省存储空间开销,而且可以减少网络中传输的数据量,进而节省网络带宽开销,在大数据时代具有更为广阔的应用价值。各大云存储厂商(如Dropbox、Google Drive等)对去重数据技术都非常重视。

Douceur等^[63]开创性地提出了收敛加密的概念,计算数据的散列值作为密钥,利用传统的对称加密算法对数据加密,这就保证了不同用户共享的相同数据文件必将产生相同的密文。Storer等^[64]提出了一种基于收敛加密方式的重复密文数据删除技术。该技术通过使用相同的加密模式来生产相同的密文数据,以达到使用传统重复数

据删除技术来删除冗余数据的目的。Bellare 等^[65]提出了消息锁定加密 (MLE, message-locked encryption), 其本质是收敛加密的一般化描述, 并给出了 MLE 的严格安全定义和安全模型。然而, 为了解决消息可预测的明文空间易遭受密钥暴力破击攻击的问题, Bellare 等^[66]提出了一个服务器辅助的 MLE 方案 (DupLESS), 通过引入一个可信的密钥服务器, 利用其私钥来帮助用户生成收敛密钥, 从而达到扩大密钥空间的作用。Li 等^[67]提出了支持数据块级去重的密钥外包存储方案 (DeKey), 有效地降低了用户端密钥存储开销。Feng 等^[68]提出了支持多级密钥管理的细粒度安全去重方案 (SecDep), 对文件级和数据块级去重采用不同策略, 有效降低密钥开销带来的效率瓶颈。

密文数据去重是大数据安全保护的重要组成部分。目前, 大数据中密文数据去重研究主要集中在收敛加密方式, 即使用相同的密钥对相同的数据加密产生相同的密文。研究在一般化加密方式中密文数据去重是大数据安全保护的研究重点。

3.4.2 数据可信删除

数据可信删除是近几年大数据安全保护技术的研究热点。大数据存储在云端时, 当用户发出删除指令后, 可能不会被云服务商真正地销毁, 而是被恶意地保留, 从而使其面临被泄露的风险。传统的保护存储在云端数据安全的方法是, 在将数据传输之前进行加密, 则数据可信删除就变成了用户本地密钥安全销毁, 一旦用户安全销毁密钥, 那么存在数据即使被泄露, 被泄露的数据也不能在多项式时间内被解密, 从而保护了数据安全。

很多存储系统采用覆盖的方法删除数据^[69], 而该方法依赖于物理存储介质的性质。一旦数据所有者失去了对数据存储位置的物理控制, 就使基于存储介质的覆盖删除方法不能满足大数据需求。Perlman^[70]首次提出了基于时间的数据可信删除方法, 该方法能实现数据安全删除, 并在规定时间内不被访问。与之类似, Geabasu 等^[71]将数据密钥分成的多个小块存在网络的不同位置, 并在规定时间内删除, 以实现数据可信删除。Peterson 等^[72]提出了在数据块层实现数据的可信删除的方法, 该方法使用全有或全无的转换技术

存储每一个数据块, 然后覆盖其中的一部分, 达到整个数据块不可用的目的。Cachin 等^[73]使用图论、对称加密算法和门限秘密共享技术等相结合来实现基于策略的安全数据可信删除。

由于大数据主要存储在云端, 数据所有者对存在云端的数据失去控制权, 数据可信删除技术在大数据安全保护中是十分关键的。目前, 数据可信删除技术尚在起步阶段, 主要通过第三方来删除密钥来实现。在大数据环境中, 如何实现真正的可信任的数据可信删除是未来大数据安全保护技术的研究要点。

3.5 密文搜索

大数据经常以密文形式存储在云端, 这使数据查询变得困难。此外, 采用一般加密方法加密时, 索引是无法建立的, 从而导致查询效率低。为了保障云端数据的可用性, 可搜索加密技术 (searchable encryption)^[74]被提了出来, 该技术用来实现对密文的有效检索和查询。密文搜索的一般化过程如图 3 所示。数据拥有者将加密后的数据以及对应的可搜索索引上传至云端, 数据使用者随后向云端提出检索请求并发送关键词陷门, 最终由云端安全地返回 (排序后的) 检索结果。该过程需确保云端未窃取到任何与检索操作有关的额外信息。目前, 主要的可搜索加密技术可分为 2 种: 对称可搜索加密技术 (symmetry key cryptography based) 和非对称可搜索加密技术 (public key cryptography based)。

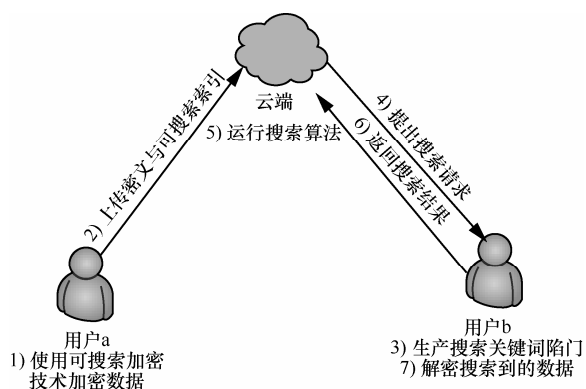


图3 密文搜索的一般化过程

3.5.1 对称可搜索加密

对称可搜索加密技术主要是通过可搜索加密机制建立安全加密索引, 在文件与检索关键词之

间建立检索关联。在密文搜索时,数据拥有者为数据使用者提供陷门,从而完成密文检索。Song等^[74]针对加密数据上的查询问题提出了第一个密文搜索算法。Curtmola等^[75]首次提出了允许多个用户搜索的可搜索对称加密算法,然而该算法在处理动态变化的大数据时效率非常低。在此基础上, Van等^[76]提出了一种可搜索的对称加密算法,该算法搜索效率高,且允许数据系统升级更新。针对云服务商可能主动删除云上的加密数据, Kurosawa等^[77]定义了抗主动攻击的可搜索的对称加密算法,并提出了一个通用可组合安全的高效算法。为了满足多关键字搜索的需求, Chase等^[78]定义了可搜索的结构化对称加密算法和相应模型,并实现了支持多关键字搜索的加密算法。在此基础上, Ballard等^[79]利用秘密分享和双线性对技术提出对其进行改进。

对称可搜索加密算法的检索效率较差,其检索时间与密文数据总长度呈现线性增长关系^[80]。为了提高密文搜索效率, Lu等^[81]提出了一个新的可搜索加密算法,该算法能达到对数时间复杂度。Strizhov等^[82]提出了支持多关键词的相似性搜索的加密算法,其检索时间与文档总数之间存在亚线性关系。上述方案都是针对单一数据源而创建可搜索索引,而在大数据分布在云端的不同存储设备上,由单一数据源所创建的索引必定影响加密数据查询和搜索的效率。针对该问题, Liu等^[83]提出了 MDS-SSE 算法,该算法允许各数据源以分散的方式生成索引。

3.5.2 非对称可搜索加密

非对称可搜索加密技术允许数据发送者以公钥加密数据与关键词,而数据使用者则利用私钥自行生成陷门以完成检索,从而解决服务器不可信与数据来源单一等问题。Boneh等^[84]提出了第1个非对称可搜索加密算法。在该算法中,任何公钥所有者都可以向服务器中写入数据,但只有授权的拥有私钥的用户才能够对密文进行搜索。然而该算法不满足一致性, Abdalla等^[85]针对这个问题重新定义了非对称可搜索加密,并提出了一个新的算法。由于文献[84]中方案需要安全通道, Baek等^[86]提出了一个可搜索的公钥加密算法。该算法使用聚合签名技术,且不需要安全通道,在

随机预言模型下被证明是安全的。为了消除随机预言机, Fang等^[87]提出了首个标准模型下不需要安全通道的可搜索的公钥加密方案。

为了提高非对称可搜索加密技术的效率, Bellare等^[88]提出确定性加密(DE, deterministic encryption)的概念,即对于同一个公钥和明文输出的密文相同。Boldyreva等^[89]提出可证的确定性加密算法,该算法所加密的数据没有额外的限制。由于确定性加密算法能加密较大的数据,而有时候数据间的差异较小。为了提高确定性加密算法的计算效率, Mironov等^[90]首次提出了增量确定性加密,并给出从普通确定性加密算法到增量确定性加密算法的一般过程。

上述非对称可搜索的加密算法往往不支持模糊查询。为此, Boneh等^[91]设计了允许关键词比较、子集查询的非对称可搜索的加密算法。Katz等^[92]提出了基于双线性对技术的公钥加密算法。该算法支持能任意析取连接词、多项式和内积的关键词搜索。

大数据经常以密文形式存储在云端,为了实现这些数据的安全性和可用性,可搜索加密技术研究将集中在支持多样化查询的搜索和相关性排序,以及进一步提升搜索的效率和精度,具体体现在以下3点。

1) 对称可搜索加密技术在大数据环境中,其检索性能显著下降,且可扩展能力差。研究支持多类型的搜索,如短语搜索和邻近搜索等,是未来大数据安全保护技术的发展方向。

2) 当前非对称可搜索加密的查询效率低。研究简单、高效、且安全的非对称可搜索加密算法是未来大数据安全保护技术的研究重点。

3) 目前,可搜索加密算法能实现一般结构数据的动态变化和多关键词的密文搜索。然而,大数据结构十分复杂、类型繁多、搜索需求多样化,研究支持在复杂结构中的多样化查询的加密算法是非常重要的。

4 结束语

继物联网、移动互联网、云计算后,大数据将引起信息产业的又一次颠覆性技术革命。如何保护大数据安全是当前大数据研究的重点和热

点。本文首先概述了大数据的基本概念特征和安全需求。然后从数据使用到销亡的生命周期, 具体包括: 加密、验证、访问、去重与删除、搜索等方面, 分别阐述了当前大数据安全保护技术的研究进展, 分析了它们的优缺点, 并探讨了其未来的发展趋势。

参考文献:

- [1] TAYLOR J. What is big data?[EB/OL]. <http://www-01.ibm.com/software/data/bigdata>.
- [2] AGRAWAL D, BERNSTEIN P. Challenges and opportunities with big data[R]. 2012.
- [3] Big data[EB/OL]. <http://www.villanovau.com/resources/bi/what-is-big-data/>.
- [4] HASHEM I A T, YAQOOB I, ANUAR N B, et al. The rise of “big data” on cloud computing: review and open research issues[J]. *Information Systems*, 2015, 47(47): 98-115.
- [5] LI G J, CHENG X Q. Research status and scientific thinking of big data[J]. *Bulletin of Chinese Academy of Sciences*, 2012, 27(6): 647-657.
- [6] HUDIC A, ISLAM S, KIESEBERG P, et al. Data confidentiality using fragmentation in cloud computing[J]. *International Journal of Pervasive Computing & Communication*, 2013, 9(1): 37-51.
- [7] ATENIESE G, BURNS R, et al. Provable data possession at untrusted stores[C]//The 14th ACM Conference on Computer and Communications Security. c2007:598-609.
- [8] Nirvanix cloud. Why nirvanix[EB/OL]. <http://www.nirvanix.com/company/why-nirvanix.aspx>.
- [9] SAHAI A, WATERS B. Fuzzy Identity-based encryption[R]. *Lecture Notes in Computer Science* 3494, 2004:457-473.
- [10] GOYAI V, PANDEY O, SAHAI A, et al. Attribute-based encryption for fine-grained access control of encrypted data[C]//ACM CSS. c2006: 89-98.
- [11] YU S, WANG C, REN K, et al. Achieving secure, scalable and fine-grained data access control in cloud computing[C]//IEEE Infocom. c2010: 15-19.
- [12] LI J, ZHAO G, CHEN X, et al. Fine-grained data access control systems with user accountability in cloud computing[C]//The 2nd International Conference on Cloud Computing Technology and Science, IEEE Computer Society. c2010:89-96.
- [13] BETHENCOURT J, SAHAI A, WATERS B. Ciphertext-policy Attribute-Based Encryption[C]//IEEE Symposium on Security and Privacy (SP '07), Berkeley. c2007:321-334.
- [14] CHASE M. Multi-authority attribute based encryption[R]. *Lecture Notes in Computer Science* 4392. Berlin: Springer, 2007: 515-534.
- [15] HUR J, NOH D K. Attribute-Based access control with efficient revocation in data outsourcing systems[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2011, 22(7):1214-1221.
- [16] GREEN M, HOHENBERGER S, WATERS B. Outsourcing the decryption of ABE ciphertexts[C]//The 9th USENIX Security Symp. c2011:3-18.
- [17] BLAZE M, BLEUMER G, STRAUSS M. Divertible protocols and atomic proxy cryptography[R]. *Lecture Notes in Computer Science* 1403. Berlin: Springer, 1998:127-144.
- [18] GREEN M, ATENIESE G. Identity-based proxy re-encryption[R]. *Lecture Notes in Computer Science* 4521. Berlin Heidelberg: Springer, 2007: 288-306.
- [19] MIZUNO T, DOI H. Secure and efficient IBE-PKE proxy re-encryption[J]. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, 2011, 94(1):36-44.
- [20] ZHAO J, FENG D G, YANG L, et al. CCA-secure type-based proxy re-encryption without pairings[J]. *Acta Electronica Sinica*, 2011, 39(11):2513-2519.
- [21] WU X X, XU L, ZHANG X W. A certificateless proxy re-encryption scheme for cloud-based data sharing[C]//The 18th ACM Conference on Computer and Communications Security(CCS). c2011:869-871.
- [22] LIANG K T, LIU J K, WONG D S, et al. An efficient cloud-based revocable identity-based proxy re-encryption scheme for public clouds data sharing[C]//The 19th European Symposium on Research in Computer Security (ESORICS). c2014:257-272.
- [23] WENG J, DENG R H, DING X H, et al. Conditional proxy re-encryption secure against chosen-ciphertext attack[C]//The 4th International Symposium on Information, Computer, and Communications Security (ASIACCS). c2009. 322-332.
- [24] FANG L M, SUSILO W, GE C P, et al. Chosen-ciphertext secure anonymous conditional proxy re-encryption with keyword search[J]. *Theoretical Computer Science*, 2012, 462(1):39-58.
- [25] FANG L M, WANG J D, GE C P, et al. Fuzzy conditional proxy re-encryption[J]. *Science China Information Sciences*, 2015, 56(5): 1-13.
- [26] CHOW S, WENG J, YANG Y, et al. Efficient unidirectional proxy re-encryption[R]. *Lecture Notes in Computer Science* 6055. Berlin: Springer, 2010: 316-332.
- [27] LAN C H, WANG C F. A new conditional proxy re-encryption scheme based on secret sharing[J]. *Chinese Journal of Computers*, 2013, 36(4):895-902.
- [28] Homomorphic encryption[EB/OL]. https://en.wikipedia.org/wiki/Homomorphic_encryption.
- [29] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//The 41st Annual ACM Symposium on Theory of Computing (STOC). c2009:169-178.
- [30] SMART N P, VERCAUTEREN F. Fully homomorphic encryption with relatively small key and ciphertext sizes[C]//The Public Key Cryptography (PKC). c2010: 420-443.
- [31] GENTRY C, HALEVI S, SMART N P. Fully homomorphic encryption with polylog overhead[C]//Advances in Cryptology. c2012: 465-482.
- [32] GENTRY C, HALEVI S, SMART N P. Better bootstrapping in fully homomorphic encryption[C]//The Public Key Cryptography (PKC). c2012:1-16.
- [33] BRAKERSKI Z, GENTRY C, HALEVI S. Packed ciphertexts in LWE-based homomorphic encryption[C]//The Public-Key Cryptography (PKC). c2013:1-13.
- [34] GENTRY C. Toward basing fully homomorphic encryption on worst-case hardness[C]//Advances in Cryptology. c2010:116-137.
- [35] BRAKERSKI Z, VAIKUNTANATHAN V. Fully homomorphic

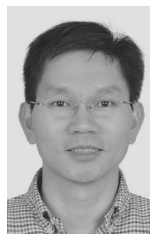
- encryption from ring-LWE and security for key dependent messages[C]//Advances in Cryptology. c2011:505-524.
- [36] LOPEZ A, TROMER E, VAIKUNTANATHAN V. On-the-fly multiparty computation on the cloud via multi-key fully homomorphic encryption[C]//The 44th Annual ACM Symposium on Theory of Computing (STOC). c2012:1219-1234.
- [37] DIJK M V, GENTRY C, HALEVI S, et al. Fully homomorphic encryption over the integers[C]//Advances in Cryptology. c2010: 24-43.
- [38] CORON J, MANDAL A, NACCACHE D, et al. Fully homomorphic encryption over the integers with shorter public key[C]// Advances in Cryptology. c2011: 487-504.
- [39] CHEON J H, CORON J, KIM J, et al. Batch fully homomorphic encryption over the integers[C]//Advances in Cryptology. c2013: 315-335.
- [40] ATENIESE G, BURNS R, CURTMOLA R, et al. Provable data possession at untrusted stores[C]//The 14th ACM Conference on Computer and Communications Security (CCS). c2007: 598-609.
- [41] WANG C, WANG Q, REN K, et al. Privacy-preserving public auditing for data storage security in cloud computing[C]//The 29th IEEE Infocom. c2010: 1-9.
- [42] ZHU Y, WANG H, HU Z, et al. Dynamic audit services for integrity verification of outsourced storages in clouds[C]//The 2011 ACM Symposium on Applied Computing (SAC). c2011: 1550-1557.
- [43] YANG K, JIA X. An efficient and secure dynamic auditing protocol for data storage in cloud computing[J]. IEEE Transactions on Parallel and Distributed Systems, 2013,24(9):1717-1726.
- [44] LI N, LI T, VENKATASUBRAMANIAN S. T-Closeness: privacy beyond k-anonymity and l-diversity[C]//The 23rd IEEE International Conference on Data Engineering (ICDE). c2007: 6-115.
- [45] ZHU Q, ZHAO T, WANG S. Privacy preservation algorithm for service-oriented information search[J]. Chinese Journal of Computers, 2010, 33(8):1315-1323.
- [46] DODIS Y, VADHAN S, WICHES D. Proofs of retrievability via hardness amplification[C]//The 6th Theory of Cryptography Conference (TCC). c2009:109-127.
- [47] WANG Q, WANG C, LI J, et al. Enabling public verifiability and data dynamics for storage security in cloud computing[C]//The 14th European Symposium on Research in Computer Security (ESORICS). c2009:355-370.
- [48] WANG Q, WANG C, REN K, et al. Enabling public verifiability and data dynamics for storage security in cloud computing[J]. IEEE Transactions on Parallel and Distributed Systems, 2011, 22(5): 847-859.
- [49] SANDHU R S, COYNE E J, FEINSTEIN H L, et al. Role-based access control models[J]. Ansi Incits, 2009, 4(3): 554-563.
- [50] RAY I, KUMAR M, YU L. LRBAC: a location-aware role-based access control model[C]//The 2nd International Conference on Information Systems Security. c2006: 147-161.
- [51] ZHANG Y J, FENG D G. A role-based access control model based on space, time and scale[J]. Journal of Computer Research and Development, 2010, 47(7): 1252-1260.
- [52] Attribute-based access control[EB/OL]. https://en.wikipedia.org/wiki/Attribute-based_access_control.
- [53] SUN G Z, DONG Y, LI Y. CP-ABE based data access control for cloud storage[J]. Journal on Communications, 2011,32(7):146-152.
- [54] SUSHMITA R, MILOS S, AMIYA N. Privacy preserving access control with authentication for securing data in clouds[C]//The 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. c2012: 556-563.
- [55] LIANG X H, CAO Z F, LIN H, et al. Attribute based proxy re-encryption with delegating capabilities[C]//The 4th International Symposium on Information, Computer and Communications Security (ASIACCS). c2009:276-286.
- [56] HONG C, ZHANG M, FENG D G. AB-ACCS: a cryptographic access control scheme for cloud storage[J]. Journal of Computer Research and Development, 2010, 47(Z1):259-265.
- [57] YU S C, WANG C, REN K, et al. Attribute based data sharing with attribute revocation[C]//The 5th International Symposium on Information, Computer and Communications Security (ASIACCS). c2010: 261-270.
- [58] YU S C, WANG C, REN K, et al. Achieving secure, scalable, and fine-grained data access control in cloud computing[C]//The Infocom. c2010: 1-9.
- [59] CHASE M. Multi-authority attribute based encryption[C]//The 4th Theory of Cryptography Conference (TCC). c2007: 515-534.
- [60] YANG K, JIA X H. Attribute-based access control for multi-authority systems in cloud storage[C]//The 32nd International Conference on Distributed Computing Systems. c2012: 536-545.
- [61] LIU X J, XIA Y J, JIANG S, et al. Hierarchical attribute-based access control with authentication for outsourced data in cloud computing[C]//The 12th International Conference on Trust, Security and Privacy in Computing and Communications, IEEE. c2013:477-484.
- [62] Data deduplicaiton[EB/OL]. https://en.wikipedia.org/wiki/Data_deduplication.
- [63] DOUCEUR J R, ADYA A, BOLOSKY W J, et al. Reclaiming space from duplicate files in a serverless distributed file system[C]//International Conference on Distributed Computing Systems. c2002:617.
- [64] STORER M W, GREENAN K, LONG D D, et al. Secure data deduplication[C]//The 4th ACM International Workshop on Storage Security and Survivability. c2008: 1-10.
- [65] BELLARE M, KEELVEEDHI S, RISTENPART T. Message-locked encryption and secure deduplication[M]// EUROCRYPT. Berlin Heidelberg: Springer, 2013:296-312.
- [66] BELLARE M, KEELVEEDHI S, RISTENPART T. DupLESS: server-aided encryption for deduplicated storage[C]//The USENIX on Security (SEC), ACM. c2013:179-194.
- [67] LI J, CHEN X, LI M, et al. Secure deduplication with efficient and reliable convergent key management[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 25(6): 1615-1625.
- [68] ZHOU Y, FENG D, XIA W, et al. SecDep: a user-aware efficient fine-grained secure deduplication scheme with multi-level key management[C]// IEEE MASS Storage Systems and Technologies. c2015:1-14.
- [69] JOUKOV N, PAPAXENOPOULOS H, ZADOK E. Secure deletion myths, issues, and solutions[C]//ACM Workshop on Storage Security and Survivability. c2006:61-66.

- [70] CRESECCENZO G D, FERGUSON N, IMAGLIAZZO R, et al. How to forget a secret?[C]//The 16th Symp on Theoretical Aspects of Computer Science (STACS 1999). c1999: 500-509.
- [71] PERLMAN R. File system design with assured delete[C]//IEEE International Security in Storage Workshop. c2005:83-88.
- [72] GEAMBASU R, KOHNO T, LEVY A, et al. Vanish: increasing data privacy with self-destructing data[C]//The 17th USENIX Security Symposium. c2009: 299-316.
- [73] CACHIN C, HARALAMBEV K, HSIAO H C, et al. Policy-based secure deletion[C]//ACM Sigsac Conference on Computer & Communications Security. c2013:259-270.
- [74] SONG D X, WAGNER D, PERRIG A. Practical techniques for searches on encrypted data[C]//The IEEE Symposium on Security and Privacy (S&P). c2000: 44-55.
- [75] CURTMOLA R, GARAY J, KAMARA S, et al. Searchable symmetric encryption: improved definitions and efficient constructions[C]//The 13th ACM Conference on Computer and Communications Security (CCS). c2006: 79-88.
- [76] LIESDONK P V, SEDGHI S, DOUMEN J, et al. Computationally efficient searchable symmetric encryption[C]//The International Workshop on Secure Data Management (SDM). c2010: 87-100.
- [77] KUROSAWA K, OHTAKI Y. UC-secure searchable symmetric encryption[C]//The 16th International Conference on Financial Cryptography and Data Security (FC). c2012:285-298.
- [78] CHASE M, KAMARA S. Structured encryption and controlled disclosure[C]//Advances in Cryptology. c2010: 577-594.
- [79] BALLARD L, KAMARA S, MONROSE F. Achieving efficient conjunctive keyword searches over encrypted data[C]//The 7th International Conference on Information and Communications Security (ICICS). c2005:414-426.
- [80] SONG D X, WAGNER D, PERRIG A. Practical techniques for searches on encrypted data[C]//The IEEE Symposium on Security and Privacy (S&P 2010). c2010:44-55.
- [81] LU Y B. Privacy-preserving logarithmic-time search on encrypted data in cloud[C]//The 19th Network and Distributed System Security Symposium (NDSS 2012). c2012:1-17.
- [82] STRIZHOV M, RAY I. Multi-keyword similarity search over encrypted cloud data[M]//ICT Systems Security and Privacy Protection. Berlin Heidelberg: Springer, 2014:52-65.
- [83] LIU C, ZHU L H, CHEN J J. Efficient searchable symmetric encryption for storing multiple source data on cloud[C]// IEEE Trustcom, Bigdata. c2015:259-261.
- [84] BONEH D, CRESCENZO G D, OSTROVSKY R, et al. Public key encryption with keyword search[C]//Advances in Cryptology. c2004:506-522.
- [85] ABDALLA M, BELLARE M, CATALANO D, et al. Searchable encryption revisited: consistency properties, relation to anonymous IBE, and extensions[C]//Advances in Cryptology. c2005:205-222.
- [86] BAEK J, SAFAVI-NAINI R, SUSILO W. Public key encryption with keyword search revisited[C]//The International Conference on Computational Science and Its Applications (ICCSA). c2008: 1249- 1259.
- [87] FANG L, SUSILO W, GE C, WANG J. A secure channel free public key encryption with keyword search scheme without random oracle[C]//International Conference Cryptology and Network Security(CANS). c2009:248-258.
- [88] BELLARE M, BOLDYREVA A, O'NEILL A. Deterministic and efficiently searchable encryption[C]//International Cryptology Conference on Advances in Cryptology. c2007:535-552.
- [89] BOLDYREVA A, FEHR S, O'NEILL A. On notions of security for deterministic encryption, and efficient constructions without random oracles[C]//International Cryptology Conference on Advances in Cryptology. c2008: 335-359.
- [90] MIRONOV I, PANDEY O, REINGOLD O, et al. Incremental deterministic public-key encryption[M]//Advances in Cryptology. Berlin Heidelberg: Springer, 2012:628-644.
- [91] BONEH D, WATERS B. Conjunctive, subset, and range queries on encrypted data[C]//The 4th Theory of Cryptography Conference (TCC). c2007: 535-554.
- [92] KATZ J, SAHAI A, WATERS B. Predicate encryption supporting disjunctions, polynomial equations, and inner products[C]// Advances in Cryptology. c2008:146-162.

作者简介:



魏凯敏 (1984-), 男, 湖南株洲人, 博士, 暨南大学副研究员, 主要研究方向为移动网络算法分析与设计、移动网络安全。



翁健 (1976-), 男, 广东茂名, 暨南大学教授、博士生导师, 主要研究方向为密码学与信息安全。



任奎 (1978-), 男, 安徽巢湖人, 纽约州立大学布法罗分校系教授, 主要研究方向为云计算中的数据安全、计算服务外包安全、无线系统安全、隐私保护、物联网系统与安全。