软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(11):3563−3575 [doi: 10.13328/j.cnki.jos.006073]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于阈值动态调整的重复数据删除方案*

咸鹤群[1,2,3], 高 原[1], 穆雪莲[1,3], 高文静[1]

[1](青岛大学 计算机科学技术学院,山东 青岛 266071)

[2](信息安全国家重点实验室(中国科学院 信息工程研究所),北京 100093)

[3](广西密码学与信息安全重点实验室(桂林电子科技大学),广西 桂林 541004)

通讯作者: 咸鹤群, E-mail: xianhq@126.com

摘 要: 云存储已经成为一种主流应用模式.随着用户及存储数据量的增加,云存储提供商采用重复数据删除技术来节省存储空间和资源.现有方案普遍采用统一的流行度阈值对所有数据进行删重处理,没有考虑到不同的数据信息具有不同的隐私程度这一实际问题.提出了一种基于阈值动态调整的重复数据删除方案,确保了上传数据及相关操作的安全性.提出了理想阈值的概念,消除了传统方案中为所有数据分配统一阈值所带来的弊端.使用项目反应理论确定不同数据的敏感性及其隐私分数,保证了数据隐私分数的适用性,解决了部分用户忽视隐私的问题.提出了基于数据加密的隐私分数查询反馈机制,在此基础上,设计了流行度阈值随数据上传的动态调整方法.实验数据及对比分析结果表明,基于阈值动态调整的重复数据删除方案具有良好的可扩展性和实用性.

关键词: 重复数据删除;项目反应理论;阈值动态调整;理想阈值

中图法分类号: TP311

中文引用格式: 咸鹤群,高原,穆雪莲,高文静.基于阈值动态调整的重复数据删除方案.软件学报,2021,32(11):3563−3575. http://www.jos.org.cn/1000-9825/6073.htm

英文引用格式: Xian HQ, Gao Y, Mu XL, Gao WJ. Deduplication scheme based on threshold dynamic adjustment. Ruan Jian Xue Bao/Journal of Software, 2021,32(11):3563−3575 (in Chinese). http://www.jos.org.cn/1000-9825/6073.htm

Deduplication Scheme Based on Threshold Dynamic Adjustment

XIAN He-Qun[1,2,3], GAO Yuan[1], MU Xue-Lian[1,3], GAO Wen-Jing[1]

[1](College of Computer Science and Technology, Qingdao University, Qingdao 266071, China)

[2](State Key Laboratory of Information Security (Institute of Information Engineering, Chinese Academy of Sciences), Beijing 100093, China)

[3](Guangxi Key laboratory of Cryptography and Information Security (Guilin University of Electronic Technology), Guilin 541004, China)

Abstract: Cloud storage has become a major application model. As the number of users and data volume increase, cloud storage providers use deduplication technology to reserve storage space and resources. Existing solutions generally use a uniform popularity threshold to process all the data, while the issue is not addressed that different data information should have different privacy levels. A deduplication scheme is proposed based on threshold dynamic adjustment to ensure the security of uploaded data and related operations. The concept of ideal threshold is introduced, which can be used to eliminate the drawbacks of uniform threshold in the traditional schemes. The item response theory is adopted to determine the sensitivity of different data and their privacy scores, which ensures the applicability of data privacy scores, it can solve the problem that some users care little about privacy issues. A privacy score query and response mechanism are proposed based on data encryption. On this basis, the dynamic adjustment method of the popularity threshold is designed

* 基金项目: 国家自然科学基金(61702294); 山东省自然科学基金(ZR2019MF058); 信息安全国家重点实验室开放课题(2020-MS-09)

Foundation item: National Natural Science Foundation of China (61702294); Natural Science Foundation of Shandong Province (ZR2019MF058); Open Project of State Key Laboratory of Information Security (2020-MS-09)

收稿时间: 2018-12-08; 修改时间: 2019-10-08; 采用时间: 2020-04-29

for data uploading. Experiment results and comparative analysis show that the proposed scheme based on threshold dynamic adjustment has sound scalability and solid practicability.

Key words: deduplication; item response theory; threshold dynamic adjustment; ideal threshold

近年来,随着网络技术的快速发展和信息量的逐日增加,用户需要花费大量的资源和时间来存储和管理自己的数据.云存储技术应运而生,并快速发展成为一种主流的存储技术.用户间的数据共享成为一种普遍的应用需求,同时给云存储提供商(cloud storage provider,简称 CSP)带来了新的挑战.随着上传数据量的逐日增加,共享数据所占的比重越来越大,数据冗余程度随之提高.统计数据表明,云存储的数据中有高达 60%的冗余数据.大量的云存储空间和存储资源被冗余数据所占用,这增加了 CSP 对云端数据的存储和维护成本[1].

为了解决上述问题,CSP 普遍采用了重复数据删除技术[2].重复数据删除是基于数据自身的冗余度来检测上传数据流中的相同数据对象,只存储唯一的数据副本,并为其他上传该数据的用户创建数据访问链接.与传统的数据压缩技术相比,重复数据删除技术不仅可以消除文件内部的数据冗余,还能消除共享数据集内文件之间的数据冗余[3−5].然而,部分用户缺乏安全意识,导致大量隐私数据在用户不知情的情况下被共享.近年来,大规模数据泄露事件引发了业界对隐私保护问题的高度关注[6].因此,在提高云端重复数据删除效率的同时,如何更好地保护用户隐私,是一个非常重要的问题[7].Harnik 等人第一次提出了客户端重复数据删除的安全问题[8].文献[9]首次提出了支持上传数据加密的重复数据删除方案——收敛加密,在该方案中,采用数据的散列值作为其加密密钥,从而保证数据和密钥的一一对应,同时对加密密文进行所有权认证.然而,从明文直接获得加密密钥的方式无法达到语义安全要求.文献[10]针对密文重复数据删除问题首次提出了多客户端交叉的重复数据删除方案 Xu-CDE[11],在外部攻击者和诚实且好奇的服务器并存的场景下,保护隐私数据的安全.但在实用性方面,该方案存在加密效率低和认证缺乏实时性的缺点.针对以上缺点,文献[12]提出了 MRN-CDE 方案.该方案通过引入随机数,保证每一次文件所有权认证过程的及时性和有效性,能够避免重放攻击.为了减少加解密过程的运算量和确保数据的安全性,该方案利用 MLE(message locked encryption)方案中的 KP 算法[13]从原始数据中提取密钥,进一步提高了重复数据删除的安全性.此外,一些 CSP 为用户提供客户端加密选项,用户上传数据之前对数据进行加密.这种方法虽然繁琐却可以有效地保护数据隐私.但是,由于客户端加密可能导致相同的明文数据被加密成为不同的密文,为重复数据删除带来了困难.因此,上述方案虽然提高了云存储的安全性,但在存储效率方面仍然有待提高.

针对重复数据删除的效率问题,Stanek 等学者提出了基于流行度划分的方案.该方案根据不同的流行度,采取不同类型的加密方式,可以有效地提高重复数据删除的效率[14].该方案为所有数据分配一个既定的流行度阈值 T,当云端某一数据的副本数量达到 T 时,就认为此数据为流行数据;否则,就将其视为非流行数据.CSP 只对流行数据进行重复数据删除操作,从而在保护用户数据隐私的同时,更好地提高重复数据删除效率.Puzio 等人提出的 PerfectDedup 方案[15]使用 Perfect Hash Function 查询数据信息的流行度,通过可信第三方的协助完成流行数据的重复数据删除操作.但可信第三方的引入增加了 CSP 的通信开销,同时也存在一定的安全隐患.针对以上问题,Liu 等人提出一种不需要第三方服务器的安全重复数据删除方案[16].方案采用口令认证密钥交换(password authenticated key exchange,简称 PAKE)协议,实现了跨用户密钥传递,进而实现跨用户重复数据删除.该方案消除了对第三方服务器的依赖,显著提升了其实用性.但是针对一些流行数据,用户同样需要对其执行对称加密,并且需要与其他用户执行 PAKE 协议,导致了额外的计算开销.从 CSP 的角度考虑,存储空间和存储成本是其最为关注的问题.现有的基于流行度检测的重复数据删除方案,在数据副本总量未达到流行度阈值之前不进行重复数据删除操作,如文献[14,15]中方案.但实际上,用户上传至云端的隐私数据数量庞大,导致非流行数据同样占据云端大量存储空间.为进一步节省云存储空间,针对非流行数据进行重复删除的方案被提出,如文献[17,18]中方案.文献[17]中方案提出了基于椭圆曲线加密数据重复删除方案:该方案采用椭圆曲线加密算法,安全性高、计算量较小;流行数据和非流行数据采用不同的加密方式,对流行数据采用客户端重复数据删除,存储空间和带宽占用都比较少.

了解目前已有的重复数据删除方案后发现,大家均未考虑到一个实际问题——隐私程度不同的数据应该

具有不同的流行度阈值.在目前已有的云存储实际应用中,CSP 给所有上传的数据规定统一的阈值,这种方式会导致许多问题.如果统一的阈值过大,对于本身隐私程度较低的数据而言,在其副本数量未达到阈值之前,需要全部存储在 CSP.大量此类数据的重复存储,会造成较大的存储空间浪费;如果统一的阈值过小,会导致隐私程度较高的数据过早的被执行重复数据删除操作,进而可能增大隐私泄露的安全风险.因此,应该根据数据本身的隐私程度为其设定各不相同的阈值,同时需要考虑用户对其隐私程度的认识.例如,一个常用的软件安装包应具有较小的阈值,使其很快被执行重复数据删除操作,在尽可能减少占用存储空间的同时,也不会对用户隐私造成任何损害;而当某一公司内部机密文件被上传时,根据上传该数据的用户对其隐私程度的认识,CSP 可以为其设定相对较大的阈值,进而有效地避免过早执行重复数据删除,更好地保护用户数据的安全.然而,如何判断每个上传数据的隐私程度,并根据其隐私程度确定一个合理的阈值,仍然是一个复杂而且困难的问题,也是本文重点研究的内容.

本文的主要贡献归纳如下:

(1) 面向云存储场景,提出了一种基于阈值动态调整的重复数据删除方案,确保用户上传的数据及相关操作的安全.

(2) 设计了阈值动态调整机制,运用项目反应理论进行阈值动态调整.结合查询应答机制,根据多数上传用户的反馈为每个上传数据确定一个合理的阈值.

(3) 提出了理想阈值的概念,消除了传统方案中为所有数据分配统一阈值所带来的弊端.

本文第 1 节讨论重复数据删除领域的现状及其发展概况.第 2 节介绍系统模型和设计目标.第 3 节给出方案的预备知识.第 4 节分别从隐私分数查询、数据上传、隐私分数计算及其阈值更新这 3 个部分详细阐述基于阈值动态调整的重复数据删除方案的具体设计.第 5 节给出实验对比及其分析.第 6 节对全文进行总结,并展望未来的研究工作.

1 系统模型和设计目标

1.1 系统模型

本方案的系统模型涉及两类实体,即上传用户和云存储提供商(CSP).在系统建立时,上传用户可以与 CSP 进行数据交互.在交互过程中,上传用户可以扮演两个角色:数据上传者或数据观察者.CSP 为上传用户提供数据存储和数据共享服务,无法获知数据的具体内容.系统模型如图 1 所示.

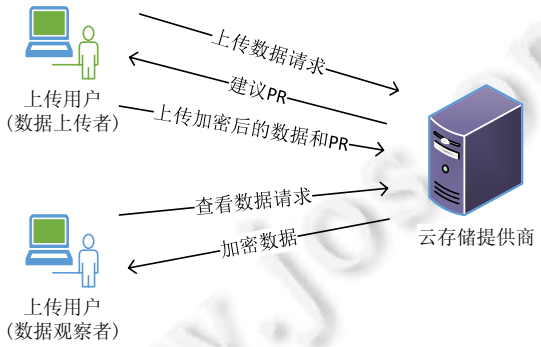

Fig.1 System model

图 1 系统模型

该模型引入了隐私分数(privacy score,简称 PR)[19]的概念.数据 F 的 PR 是隐私风险的指示器,PR 越大,代表该数据的隐私程度越高.在数据上传阶段,首先由上传用户对 CSP 发起上传数据请求,并借助椭圆曲线加密算法计算数据的查询标签.CSP 收到上传请求后,执行数据查询和密文对比等操作,在不泄露数据内容的情况下,检测 F 是否为首次上传.若 CSP 查询到云端已存储该数据,则返回给用户一个建议 PR.用户将加密后的 F 及其 PR 评

分一起上传给CSP.每次数据上传操作完成后,CSP对该数据的*PR*进行调整更新,以便作为后续上传用户的反馈信息.经过不断地上传和调整,每个上传数据*F*均对应一个逐渐趋于稳定且符合多数上传用户要求的PR.CSP根据PR计算*F*的流行度阈值*T*,并根据*T*的实际大小执行重复数据删除操作.这样既降低了存储空间浪费,又避免了隐私数据的泄露.在数据观察阶段,只有上传过该数据的用户才可以对CSP发起查询请求.CSP返回给用户所查询的数据密文.此外,尽管我们假设 CSP 是诚实且好奇的,但它可以进行离线数据分析,以推断额外的信息.因此,从用户隐私保护的角度来看,CSP 是不可信的.

1.2 设计目标

为了更好地保护数据隐私,设计的方案应该具有以下几个特点.

(1) 上传数据的保密性:为保护用户隐私,上传数据需要进行一定的加密操作.

(2) 隐私分数的可查询性:用户上传数据时可在云端查询其合理隐私分数作为参考.

(3) 隐私分数和阈值的可更新性:每个数据的隐私分数和阈值可根据具体上传情况进行动态更新.

2 预备知识

2.1 项目反应理论及其特性

项目反应理论(item response theory,简称 IRT)[20]是一个著名的心理学理论,常被用于问卷调查结果统计和测试数据分析.该理论可以通过度量受测用户的能力和特定测试项目的难度,推断出受测用户正确回答给定问题的概率.文献[21]中已经通过实验证明了,IRT 可以被应用到云场景中.

Rasch模型[22]是最常见的IRT模型之一,它假定正确响应的概率函数仅与θ_i,α_i和β_i有关,问题q_i由一对参数$\xi_i=(\alpha_i,\beta_i)$来表示.其中,θ_i代表受测用户的能力等级,α_i为问题q_i的区分能力,β_i代表测试问题的难度.因其具有计算参数少、构造简单等特性,该模型具有所需样本更小的优势.邀请受测用户j对某一问题q_i进行回答,如果用“正确”或“错误”这种二值标记法来表示问题答案,那么问题q_i被受测用户j正确回答的概率为

$$P_{ij}=\frac{1}{1+\mathrm{e}^{-\alpha_i(\theta_i-\beta_i)}} \tag{1}$$

IRT 具有两个显著的特性.

(1) 群组不变性,即项目的难度是项目自身的性质,与受测用户对该项目的回答无关.或者说单个项目的参数不仅适用于当前受测用户样本,而且对所有类型的受测用户都具有较好的普适性[23].

(2) 受测用户的独立性,即一个受测用户不会影响其他受测用户对某一问题的回答.受测用户对某一问题的回答只取决于其自身.

2.2 通用敏感度计算方法

通常来讲,某一数据的敏感度越高,受测用户就越不想将其公开.如公式(2)所示.

$$\beta_i=\frac{N-|R_i|}{N} \tag{2}$$

用$|R_i|$(即$R(i,j)=1$的个数)来表示愿意公开数据项i的受测用户个数,那么拒绝公开数据项i的受测用户数量与该数据项的敏感度β_i成正比,其中,N为数据项个数;并且数据项i越敏感,β_i的值越高[19].

2.3 通用可见度计算方法

在问题的答案为二值型的情况下,通常采用估计概率 $P_{ij}=prob\{R(i,j)=1\}$来计算数据的可见度.假设测试项目和受测用户之间是相互独立的,即某次测试调查中,受测用户回答每个问题的概率是相同的,我们能够通过将二值矩阵行R_i中1所占的比例和列R^j中1所占的比例相乘积的方式来计算P_{ij}的值.也就是说,如果$|R^j|$表示受测用户j设置的数据项中$R(i,j)=1$的个数,则概率P_{ij}随着信息项敏感度的降低和受测用户共享信息倾向的增加而增加.可见度计算公式如公式(3)所示[24].

$$P_{ij}=\frac{|R_i|}{N}\times\frac{|R^j|}{n} \tag{3}$$

其中,N 为数据项个数,n 为受测用户个数.以上可见度的计算方法是从统计学的角度出发,对所有可能的反应矩阵上根据概率分布进行的采样.实际上,事实可见度是由 $V(i,j)=P_{ij}\times1+(1-P_{ij})\times0=P_{ij}$ 计算得出.

2.4 数据的隐私分数

数据的隐私分数是数据的整体隐私程度的数值化表示.受测用户 j 由数据 i 产生的隐私分数表示为 $PR(i,j)$,其计算公式为

$$PR(i,j)=\beta_i\otimes V(i,j) \tag{4}$$

其中,操作符⊗表示任何关于敏感度和可见度的单调递增的函数组合.

2.5 双线性映射

设$(G_0,+)$,$(G_1,\cdot)$是 p 阶的加法循环群和乘法循循环群,其中,p 是大素数,α是群 G_1 的单位元.Z_p 是模 p 的剩余类整环, Z_p^* 是 Z_p 的可逆元集合,定义双线性映射 $e:(G_0,G_0)\to G_1$,并满足 3 个性质[25].

(1) 双线性.$\forall P,Q\in G_0$ 且 $a,b\in Z_p^*$,都有 $e(P^a,Q^b)=e(P,Q)^{ab}$;

(2) 可计算性.$\forall P,Q\in G_0$,$e(P,Q)$是可计算的;

(3) 非退化性.$\exists P,Q\in G_0$,使得 $e(P,Q)\neq\alpha$.

3 方案设计

3.1 方案概述

当用户向 CSP 上传加密数据时,CSP 可借助椭圆曲线生成数据的查询标签来检查是否已存储该数据,并根据数据库中的已知数据信息反馈给上传用户一个建议 PR.最终,用户将数据及其隐私分数的反馈一起上传给 CSP.CSP 重新聚合隐私分数,并为其更新阈值.根据阈值大小和数据的当前数量,CSP 决定是否对其进行重复数据删除操作.本文假设所有上传用户均为真实可靠的,即不存在恶意用户干扰隐私分数的情况.方案共分为以下 3 个部分:隐私分数查询、文件上传、隐私分数计算和阈值更新,如图 2 所示.

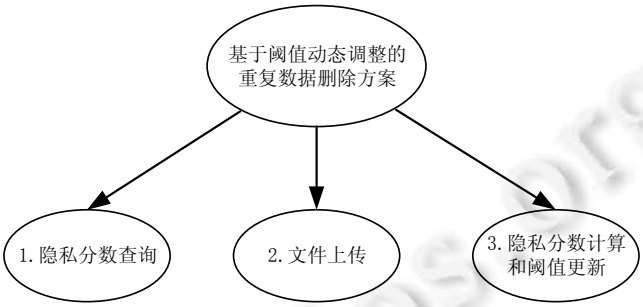

Fig.2 Scheme design

图 2 方案设计

3.2 隐私分数查询

当用户向 CSP 上传数据时,用户可以查询上传数据的当前 PR.基于上下文的隐私分数查询方法是一种可选的解决方案[21].上下文是指从一长串的文字或内容中分析得出的摘要以及大意,甚至可以是由整个上传数据整理出来的具有代表性的关键字组合,其一般形式为($field1=value1,field2=value2,...$).例如,某上传数据包含“Bob 用台式电脑在班级群中共享了本次期末考试成绩”,则归纳的上下文为(共享者=Bob,主题=期末成绩,观察者=同班同学).

文献[21]中详细介绍了一种利用上下文进行数据隐私度查询的方案.用户在查询隐私度之前,将多个虚拟

上下文组成查询集,并发送给 CSP,以隐藏实际请求的上下文.基于上下文的隐私分数查询只需要在上述方案的基础上,将数据隐私度替换成相应的隐私分数即可.当用户借助上下文向 CSP 查询数据的 *PR* 时,用户应该避免直接将上下文发送到 CSP,防止 CSP 将数据共享操作与该上下文相关联,避免隐私数据泄露.

基于上下文的隐私分数查询方式易于实现,但是我们通常假定 CSP 是诚实且好奇的,它可以通过离线分析等操作获得用户上传的具体数据信息.因此,考虑到数据的隐私保护问题,本文采用加密数据的隐私分数查询机制.该机制在本团队已有的研究成果的基础上,利用基于椭圆曲线的文件标签查询方案[17]实现隐私分数查询.文献[17]提出一种无需在线可信第三方的流行度查询协议,通过构造双线性映射查询标签 $s=e(Y,H(F))^{X_1}$ (其中,Y 为加密公钥,X_1,X_2 为辅助密钥,$X_3=X_1+X_2$),在不泄露数据隐私的情况下,使用标签对比的方式快速完成数据的流行度查询.我们将隐私分数与数据标签进行关联,即可使用查询标签对比的方式实现隐私分数的快速查询,避免了采用基于上下文的隐私分数查询方式可能导致的隐私保护方面的问题.

3.3 流行度阈值

为了提高重复数据删除的效率,CSP 为上传数据分配一个流行度阈值.当某一数据 *F* 的上传用户总量大于该阈值时,我们就认为 *F* 为流行数据,则可以采用效率较高的收敛加密,同时对其执行重复数据删除操作;否则,我们认为 *F* 为非流行数据,且具有较高的隐私程度,需要采用语义安全的对称加密对其进行保护.在第 3.5 小节中,我们将给出流行度阈值的动态调整方法.

3.4 数据上传

当用户 *U* 上传数据 *F* 时,首先向 CSP 上传查询标签 *s*,CSP 根据上传数据 *F* 的用户数量 $count(U_F)$与该数据动态阈值 *T* 的大小关系,将数据上传操作分为 3 种情况:上传用户数量小于阈值 $count(U_F)<T$、上传用户数量等于阈值 $count(U_F)=T$ 和上传用户数量大于阈值 $count(U_F)>T$,如图 3 所示.

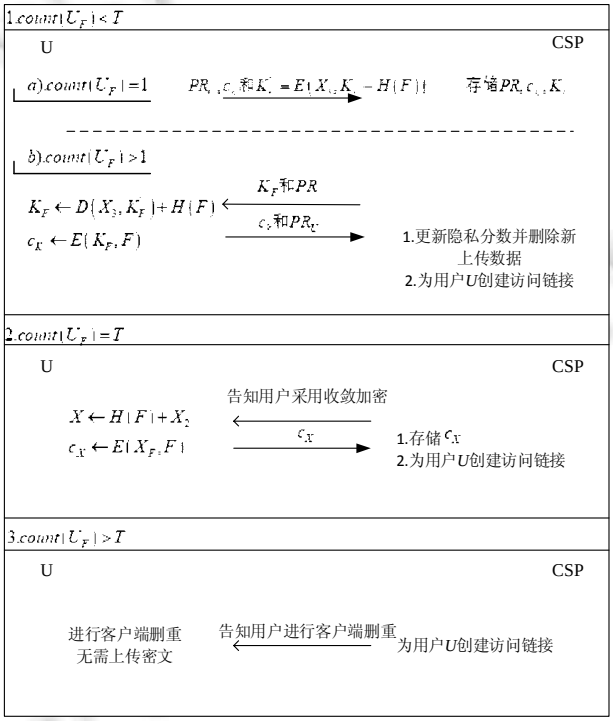

Fig.3 Data upload

图 3 数据上传

- 当 $count(U_F)<T$ 时,CSP 检查是否已经存在相同数据.
  - 若为首次上传,用户 U 加密数据并上传至 CSP,同时给出自己对该数据隐私评分 PR_U.CSP 记录 PR_U,存储加密后的对称加密密钥 $K'_F=E(X_3,K_F-H(F))$ 和密文 $c_K=E(K_F,F)$,如图 3 中 a)部分所示,其中,K_F 是数据加密密钥.
  - 若非首次上传,则 CSP 返回给用户对称加密密钥 K'_F 和建议隐私分数 PR'.U 解密并计算获得 K_F,并用其加密 F,将密文 $c_K=E(K_F,F)$和综合考量的隐私评分 PR_U一同上传至 CSP.CSP 动态更新 PR' 后删除新上传数据信息,并为 U 创建访问链接,如图 3 中 b)部分所示.
- 当 $count(U_F)=T$ 时,CSP 通知用户 U 进行收敛加密.U 上传收敛加密密文 $c_X=E(X,F)$,其中,$X=H(F)+X_2$.CSP 存储该密文.
- 当 $count(U_F)>T$ 时,CSP 告知用户进行客户端删重,并为 U 创建访问链接.

3.5 基于项目反应理论的隐私分数计算和阈值更新

本文采用基于项目反应理论的隐私分数计算方法[26],从而确保数据的隐私分数符合多数上传用户要求.

由于不同的上传数据 F 之间是相互独立的,用于计算其隐私分数 PR 的参数 $\xi=(\alpha,\beta)$可以独立计算,进而使得 PR 的计算能够并发执行.基于项目反应理论(IRT)的 PR 计算仍需要使用公式(3)来估计概率 $P=prob\{R(i,j)=1\}$,其中,$R(i,j)$为二值型矩阵,下文中直接用 R 表示.在云存储环境中,虽然没有受测试者和测试问题,但是有上传用户和上传数据这两类对象.我们将上传用户 U 看作受测试者,上传数据 F 看作信息项.因此,用户的隐私倾向就对应于受测试者的能力:通过隐私倾向参数 θ,量化用户 U 对上传数据 F 隐私的在意程度.θ值越高,表示数据越开放[20].如果我们假设上传数据 F 的隐私问题容易理解且具有完全相同的区分度,那么,问题区分度参数 α就不再是一个需要考虑的变量.因此,可以用常量替代或直接忽略不计.最后,我们用问题的难度参数 β来代表数据信息的敏感度,$\beta\geqslant 0$.

对于每个上传数据的参数 $\xi=(\alpha,\beta)$,可以通过最大似然函数对其进行估计,如公式(5)所示.

$$\prod_{j=1}^{N}P^R(1-P)^{1-R} \tag{5}$$

其中,N 为上传同一数据 F 的用户总数,R 为二值型矩阵.具体的,用户上传数据 F 时会一同上传 PR_U,以代表自己对上传数据隐私程度的评价.本实例在求敏感度 β时,将 PR_U 的值进行简单的推算作为隐私倾向参数 θ,即 $\theta=\frac{PR_U}{100}$.最终,在隐私倾向参数 θ为已知量的基础上求得敏感度 β.

根据上述过程可获知 $\xi=(\alpha,\beta)$的值,确定敏感度 β的值.在此基础上,待上传数据 F 的总体隐私倾向 θ_F 也通过上述似然函数搜索得出.具体的,本实例在敏感度 β为已知量的基础上,选用牛顿-拉夫逊的拓展算法 NR_Attitude_Estimation 算法来搜索似然函数或其对应的 log 似然函数,找到使其最大的隐私倾向参数 θ_F.最终,通过公式(4)对其进行整合,得到建议隐私分数 PR',用户 U 根据建议隐私分数 PR'调整并上传自己的隐私分数评分 PR_U[19].

而对于某一确定数据 F_i 的综合隐私分数表示为 $PR_i=\frac{\sum_{j=1}^{N}PR(i,j)}{N}$,其中,$j$ 为某上传用户的具体表示,$PR_i\in[0,1]$.

最终,CSP 根据每次数据上传结束后的隐私分数变化,结合公式(6)更新数据阈值.

$$T_i=\frac{a}{(1-PR_i)^2}-(a-1) \tag{6}$$

其中,a 为参数,其数值的大小可根据实际情况进行调整.

随着上传用户数量的增加,阈值变化逐渐趋于平稳,即 $\lim_{i\to\infty}|T_{i+1}-T_i|\leqslant\varepsilon$ (其中,ε为可忽略的值).在此,我们提出理想阈值的概念.所谓理想阈值 T^*是指每个文件具有的一个符合用户意愿的阈值.随着用户数量的增加,动态

阈值 T_i 将逐渐趋近于 T^*,即 $\lim_{i\to\infty}|T_i-T^*|\leqslant\varepsilon$.

综上所述,基于 IRT 的隐私分数计算方法具有以下几个优点.

(1) 由于不同上传数据之间相互独立,因此 CSP 可以同时计算多个不同上传数据的 *PR*;

(2) 基于 IRT 的隐私分数计算中所使用的参数是通过似然函数估算得出的,满足群组不变性的性质要求,这使得不同上传数据所对应的 PR 可以直接进行比较.

4 安全性证明与分析

新方案的设计目标是:通过阈值的可调节性,更好地保护隐私数据的安全.攻击者在不掌握原始数据的情况下,仅仅通过数据的查询标签来欺骗 CSP 并获取数据的成功概率是可忽略的.在此,我们主要讨论查询标签的不可伪造性和可区分性,安全性分析如下.

引理 1. 对于安全的散列函数 $H:\{0,1\}^*\to G_1,\forall D_1,D_2\in[0,1]^*,D_1\neq D_2,H(D_1)=H(D_2)$的概率是可忽略的.

即 $P[H(D_1)=H(D_2)|D_1\neq D_2]\leqslant\varepsilon,\varepsilon$为可忽略的值.

定理 1(数据查询标签的不可伪造性). 设初始上传用户 U_0 上传数据 F 的查询标签为 $s_F=e(Y,H(F))^X$.当用户 U_j 上传 F'时,F'的查询标签为 $s_{F'}=e(Y,H(F'))^X$.当且仅当 $s_F=s_{F'}$时 $F=F'$成立,即若 $s_F=s_{F'}$,则 $F\neq F'$的概率是可忽略的.

证明:若 $F\neq F'$,我们从以下两个方面讨论敌手对数据查询标签的攻击.

1. 假设敌手 A 是恶意用户,则 A 能获得参数 X;

2. 假设敌手 A 是 CSP,则对 A 而言,参数 $H(F)$和 X 都是不可预测的.

以上 2 种情况,敌手 A 都无法构造出满足等式 $s_F=s_{F'}$的查询标签.根据引理 1 可以推出:$F\neq F'\Leftrightarrow H(F)\neq H(F')\Leftrightarrow e(Y,H(F))\neq e(Y,H(F'))\Leftrightarrow e(Y,H(F))^X\neq e(Y,H(F'))^X\Leftrightarrow s_F\neq s_{F'}$,即 $P[F\neq F'|e(Y,H(F))^X=e(Y,H(F'))^X]\leqslant\varepsilon$.因此,数据的查询标签具有不可伪造性,根据 $s_F=s_{F'}$即可判断出 $F=F'$.证毕. □

定理 2(查询标签的可区分性). 设数据 F 的初始上传用户 U_0 上传的查询标签为 $s_F=e(Y,H(F))^X$.当用户 U_j 上传 F'时,上传的查询标签为 $s_{F'}=e(Y,H(F'))^X$.若 $F\neq F'$,则 $s_F=s_{F'}$的概率是可忽略的,即 $P[s_F=s_{F'}|F\neq F']\leqslant\varepsilon$.

证明:假设存在 $F\neq F'$使得 $s_F=s_{F'}$.根据双线性映射的性质可得等式:

$$s_F=s_{F'}\Leftrightarrow e(Y,H(F))^X=e(Y,H(F'))^X\Leftrightarrow e(Y^X,H(F))=e(Y^X,H(F'))\Leftrightarrow H(F)=H(F').$$

若使上式成立,则 $H(F)=H(F')$必须成立.

根据引理 1 可得 $F=F'$,与假设矛盾,因此假设不成立,即 $P[s_F=s_{F'}|F\neq F']\leqslant\varepsilon$.

可见,当且仅当 $F=F'$时 $s_F=s_{F'}$成立.即数据的查询标签之间具有可区分性.证毕. □

引理 2(计算 Diffie-Hellman 问题(CDH)). 假设$(G_0,\cdot)$是一个 P 阶的乘法循环群,单位元记为 g.对于给定的 $g,g^a,h\in G_0$,计算 $Q=h^a\in G_0$ 是困难的,其中, $a\in Z_n^*$.

定理 3(数据标签的安全性). CSP 无法对数据的查询标签进行离线穷举攻击,进而获得任何明文信息.

证明:设 CSP 对数据的查询标签 $s_F=e(Y,H(F))^X$ 进行离线穷举攻击.CSP 穷举大量数据$\{F_i\}$,试图找到满足 $s_F=s_{F'}$的数据 F'.CSP 可以计算出 $e(Y,H(F_1))$,但由于不是授权用户,无法从广播中心获得安全参数 X.由引理 2 可知:即使已知 $e(Y,H(F))^X,e(Y,H(F_1))$,计算 $e(Y,H(F_1))^X$ 仍然是困难的.因此,CSP 无法通过穷举攻击的方式从查询标签中获得数据明文信息.证毕. □

5 仿真实验

实验采用 PBC[27]、GMP[28]、PBC_bce[29]和 OPENSSL[30]函数库,使用 C++语言编程实现.选用腾讯云的云存储服务器进行部署,其配置为 4GB 内存,4 核 CPU,1Mb/s 带宽,1T 存储盘.为便于用户理解和操作,本方案在实现隐私分数评分时作出了较为人性化的设计.当上传用户对某一数据进行隐私分数评分时,只需给出一个 1~100 之间的数值作为评分即可.系统自动对其进行换算,并更新该数据隐私分数和阈值.这种符合百分制评分习惯的方式,使用户更直观地理解上传数据的隐私程度,提升了用户体验.考虑到样本选取的难度,本文采用生

成随机数的方式来模拟不同用户对某一上传数据的隐私分数评分.所有实验均重复进行 20 次,取平均值作为最终结果.

5.1 数据集

针对不同数据具有不同理想阈值这一问题,本文对数据的整体隐私分数和阈值变化进行了对比实验.实验分别采用了 3 组不同范围的随机数模拟用户对不同数据的隐私评分.每个数据集由 100 个数据组成:第 1 个数据集由 100 个来自区间[5~15]上的随机数组成,模拟用户对某隐私程度较低数据的隐私评分;第 2 个数据集由 100 个来自区间[90~100]上的随机数组成,模拟用户对某些隐私度极高的数据的隐私评分;第 3 个数据集由来自区间[1~100]的随机数组成,模拟上传用户对某一数据的隐私评分不统一的情况.

在方案性能对比实验中,我们选取 1 000 个大小为 10MB 的文件作为上传数据,其中,隐私程度较低的数据与隐私程度较高的数据所占比约为 3:2.其他对比方案采用统一流行度阈值,并设置为 T=7.

5.2 隐私分数与阈值大小实验分析

首先对上述 3 组实验数据集的数据分别执行数据上传和阈值动态调整的模拟操作,对数据的整体隐私分数和阈值大小的变化进行对比分析.

图 4~图 6 是由区间为[5~15]的数据集得出的,其中,

- 图 4 为整体隐私分数随上传用户数量的变化图,该图中的曲线是由 100 个数据点连接而成,每个点的横坐标为该数据某次上传操作之后,CSP 根据用户的反馈对整体隐私分数的调整结果.其纵坐标为当前上传该数据的用户数量.
- 图 5 为阈值 T 的动态调整值与整体隐私分数 PR 的关系图,该图中的曲线展示了数据的阈值随着整体隐私分数的变化过程,图中每个点的横坐标的含义与图 4 相同,纵坐标为根据整体隐私分数计算得到的该数据的动态阈值,计算公式为公式(6),其中,a 的值可根据需求自行调整.本实验中,取 a=7.
- 图 6 是将图 4 与图 5 放在同一个坐标系下叠加的结果.两条曲线的所有交点的纵坐标最小值处即为实际发生重复数据删除的阈值,大小为 T=9.

同理,图 7~图 9 是由区间为[90~100]的数据集得出的.在该模拟场景中,数据未被执行重复数据删除操作(图 9 的两条曲线没有交点).

图 10~图 12 是由区间为[0~100]的数据集得出的,其实际重复数据删除的阈值大小为 T=37.

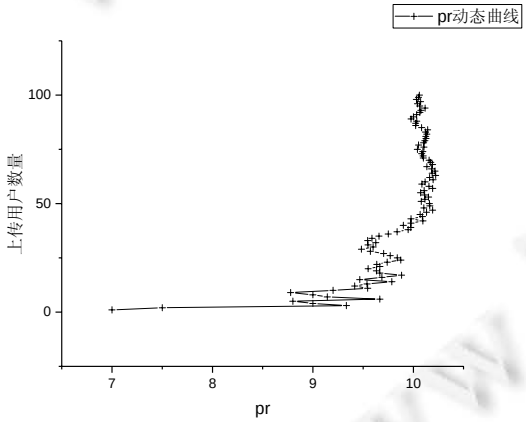

Fig.4 Privacy score with the number of upload users (5~15)

图 4 隐私分数随上传用户数量的变化(5~15)

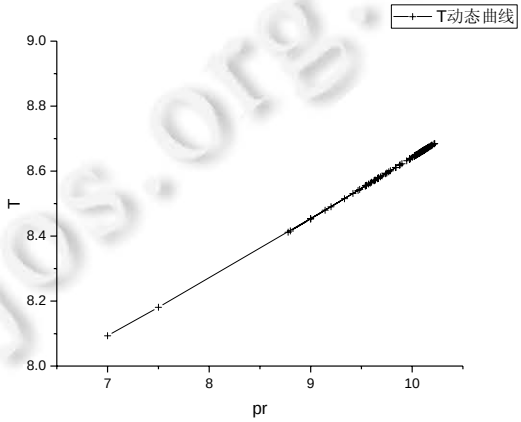

Fig.5 Relationship between threshold and privacy score (5~15)

图 5 阈值与隐私分数的关系(5~15)

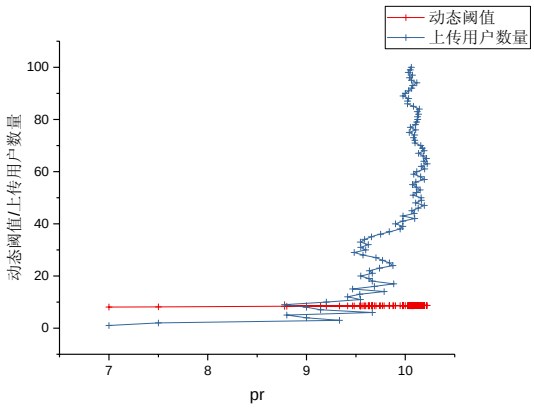

Fig.6 Actual deduplication threshold (5~15)

图 6 实际删重阈值(5~15)

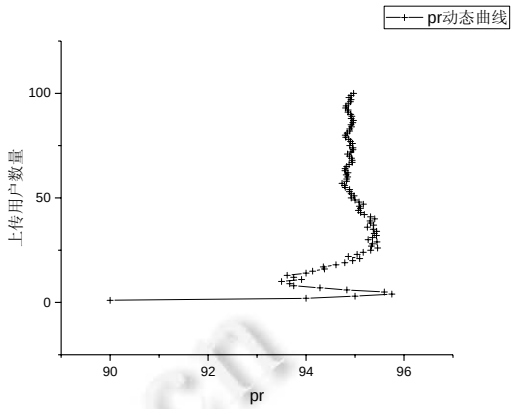

Fig.7 Privacy score with the number of upload users (90~100)

图 7 隐私分数随上传用户数量的变化(90~100)

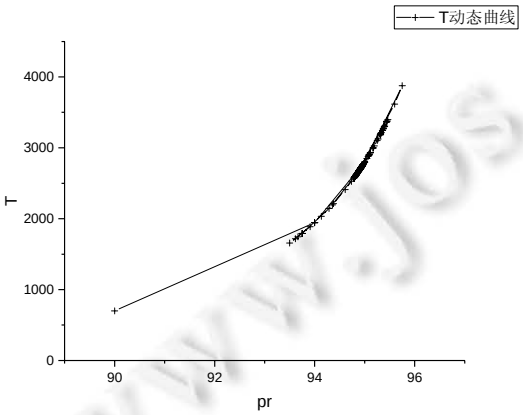

Fig.8 Relationship between threshold and privacy score (90~100)

图 8 阈值与隐私分数的关系(90~100)

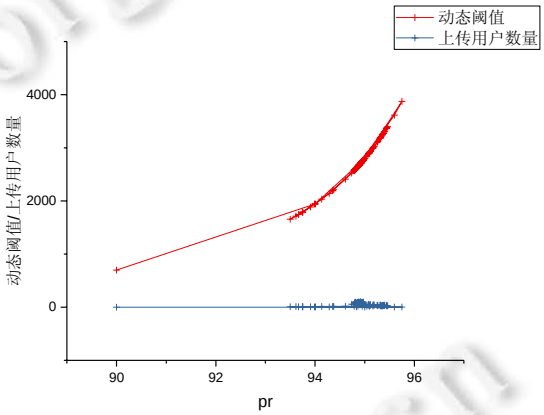

Fig.9 Actual deduplication threshold (90~100)

图 9 实际删重阈值(90~100)

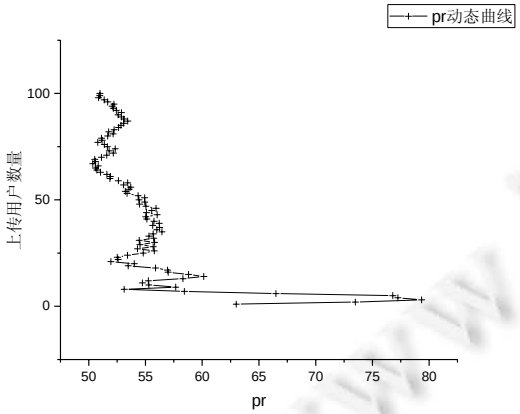

Fig.10 Privacy score with the number of upload users (1~100)

图 10 隐私分数随上传用户数量的变化(1~100)

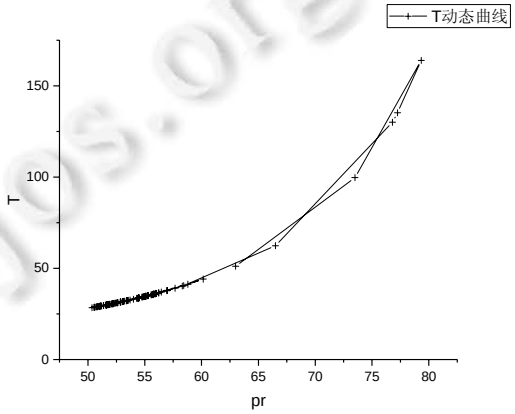

Fig.11 Relationship between threshold and privacy score (1~100)

图 11 阈值与隐私分数的关系(1~100)

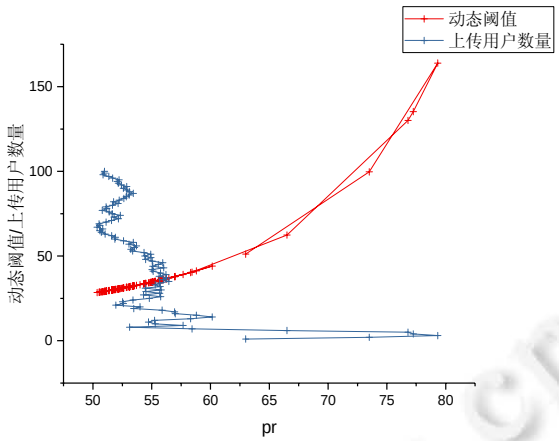

Fig.12 Actual deduplication threshold (1~100)

图 12 实际删重阈值(1~100)

图 4 中曲线表示所有上传用户都认为某上传数据具有较低的隐私程度,但在隐私评分的具体数值大小上仍然存在较小的分歧.我们假设最终每个用户会取来自区间[5~15]的一个数值作为其隐私评分,在上传用户数量较少时,用户上传的隐私评分对整体隐私分数影响较大;随着上传用户数量的增加,单个用户的隐私评分对整体隐私分数影响越来越小;最终,该数据的隐私分数大小稳定在 10 左右.同样地,图 7 代表整体隐私分数评分较高的情况,最终整体隐私分数稳定在 95 左右.图 10 曲线代表的是所有用户对某数据隐私评分不统一时的情况:上传用户数量较少时,隐私分数调整波动较大;随着上传用户数量的不断增加,单个用户的隐私评分对整体隐私分数影响越来越小;数据的整体隐私分数最终会稳定下来.

综合图 4、图 7 和图 10 可以得出,每次上传操作都会对数据的整体隐私分数 PR 带来影响.上传用户数量越少,单个用户上传数据操作对隐私分数的影响越大.当样本数量足够大时,随着用户数量的增加,数据的整体隐私分数会逐渐趋于稳定.图 10 的结果进一步说明,数据最终的隐私分数的大小是由多数用户意愿决定的.该数值只与数据本身的性质和用户对其隐私的在意程度有关,在用户数量较多的情况下,个别用户的态度(体现为评分)对其影响甚微.

由图 5、图 8 和图 11 中曲线可知,数据的阈值与整体隐私分数成正比关系.

由图 6 可知,当数据隐私度很低时,其实际执行重复数据删除的阈值很小.这说明本方案可以有效节约云端存储空间.

由图 9 可知,当数据信息隐私度较高时,两条曲线之间没有交点,即该数据未被执行重复数据删除操作.这说明本方案可以避免统一阈值删重所导致的隐私数据泄露,更有效地保护了隐私数据的安全.

由图 10~图 12 可以看出,当上传用户对某数据的隐私态度存在较大分歧时,系统根据多数人的隐私评分,为数据确定一个合适的阈值.

因此,从总体上看,该方案具有较高的适用性.

5.3 方案性能比较

通过上传 1 000 个大小为 10MB 的数据,计算本方案所需的总时间开销,并与 prefectDedup 方案、普通的基于流行度阈值的重复数据删除方案和 XU-CDE 这 3 种方案进行比较.实验重复进行 10 次,取平均值作为最终结果.实验结果如图 13 所示.在数据加密阶段,4 个方案的时间开销相差不大;在查询阶段,本方案与其他区分流行度的方案相比具有优势.最终,与其他方案相比,本方案在提高了重复数据删除操作安全性的同时,并未产生额外的时间开销.

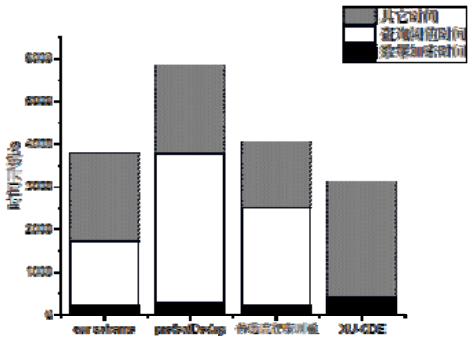

Fig.13 Performance comparison

图 13 性能对比

6 总结与展望

本文解决了云存储环境下对所有数据分配统一阈值进行重复数据删除的问题,提出了一种基于阈值动态调整的重复数据删除方案.提出了理想阈值的概念,并将项目反应理论应用到重复数据删除领域中.通过上传用户对数据隐私程度的反馈,动态调整其隐私分数,由此计算并调整重复数据删除操作的阈值.该方案可以使隐私程度较低的数据更快地达到重复数据删除的条件,而使隐私程度较高的数据得到更好的保护.借助椭圆曲线的重复数据删除方案,对上传数据进行加密处理,保证了数据信息安全.实验结果表明,与其他方案相比,本方案在提升重复数据删除操作安全性的同时,并未造成额外的时间开销,具有较高的实用性.

在确保数据安全性的同时,如何提高重复数据删除效率,是下一步需要研究的问题.

References:

[1] Stanek J, Kencl L. Enhanced secure thresholded data deduplication scheme for cloud storage. IEEE Trans. on Dependable and Secure Computing, 2018,15(4):694−707.

[2] Yuan J, Yu S. Secure and constant cost public cloud storage auditing with deduplication. In: Proc. of the 2013 IEEE Conf. on Communications and Network Security (CNS). IEEE, 2013.

[3] Fu YJ, Xiao N, Liu F. Research and development on key techniques of data deduplication. Journal of Computer Research and Development, 2012,49(1):12−20 (in Chinese with English abstract).

[4] Fan Y, Lin X, Liang W, Tan G, Anda P. A secure privacy preserving deduplication scheme for cloud computing. Future Generation Computer Systems, 2019,101:127−135.

[5] Hou H, Yu J, Hao R. Cloud storage auditing with deduplication supporting different security levels according to data popularity. Journal of Network and Computer Applications, 2019,134:26−39.

[6] Jayapandian N, Rahman AMJMZ. Secure deduplication for cloud storage using interactive message-locked encryption with convergent encryption, to reduce storage space. Brazilian Archives of Biology and Technology, 2018,61:e17160609.

[7] Baracaldo N, Androulaki E, Glider J, Sorniotti A. Reconciling end-to-end confidentiality and data reduction in cloud storage. In: Proc. of the 6th ACM Workshop on Cloud Computing Security (CCSW 2014). ACM, 2014. 21−32.

[8] Cui H, Deng RH, Li Y, Wu G. Attribute-based storage supporting secure deduplication of encrypted data in cloud. IEEE Trans. on Big Data, 2019,5(3):330−342.

[9] Douceur JR, Adya A, Bolosky WJ, *et al*. Reclaiming space from duplicate files in a serverless distributed file system. In: Proc. of the 22nd Int'l Conf. on Distributed Computing Systems. IEEE, 2002. 617−624.

[10] Xu J, Chang EC, Zhou J. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In: Proc. of the 8th ACM SIGSAC Symp. on Information, Computer and Communications Security. ACM, 2013. 195−206.

[11] Tang H, Cui Y, Guan C, Wu J, Wen J, Ren K. Enabling ciphertext deduplication for secure cloud storage and access control. In: Proc. of the 11th ACM on Asia Conf. on Computer and Communications Security. ACM, 2016. 59−70.

[12] Yang C, Ji Q, Xiong SC, Liu MZ, Ma JF, Jiang Q, Bai L. New method for file deduplication in cloud storage. Journal on Communications, 2017,38(3):25−33 (in Chinese with English abstract).

[13] Bellare M, Keelveedhi S. Interactive message-locked encryption and secure deduplication. In: Proc. of the Advances in Cryptology (EUROCRYPT 2013). Berlin Heidelberg: Springer-Verlag, 2013. 374−391.

[14] Stanek J, Sorniotti A, Androulaki E, Kencl L. A secure data deduplication scheme for cloud storage. In: Proc. of the Int'l Conf. on Financial Cryptography and Data Security. Berlin, Heidelberg: Springer-Verlag, 2014. 99−118.

[15] Puzio P, Molva R, Önen M, Sergio L. PerfectDedup: Secure data deduplication. In: Proc. of the Int'l Workshop on Data Privacy Management. Springer Int'l Publishing, 2015. 150−166.

[16] Liu J, Asokan N, Pinkas B. Secure deduplication of encrypted data without additional independent servers. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2015. 874−885.

[17] Zhang SG, Xian HQ, Liu HY, Hou RT. Research on encrypted deduplication method based on offline key transfer in cloud storage environment. Netinfo Security, 2017(7):66−72 (in Chinese with English abstract).

[18] Singh P, Agarwal N, Raman B. Secure data deduplication using secret sharing schemes over cloud. Future Generation Computer Systems, 2018,88:156−167.

[19] Lin XH. Privacy protection in community-based networks [Ph.D. Thesis]. Shanghai: Shanghai Jiaotong University, 2010 (in Chinese with English abstract).

[20] Lord F. A Theory of Test Scores. 1952.

[21] Harkous H, Rahman R, Aberer K. C3p: Context-aware crowdsourced cloud privacy. In: Proc. of the Int'l Symp. on Privacy Enhancing Technologies Symp. Cham: Springer-Verlag, 2014. 102−122.

[22] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: MESA Press, 1993.

[23] Wright BD. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977,14(2):97−116.

[24] Wang DW, Liau CJ, Hsu T. Privacy protection in social network data disclosure based on granular computing. In: Proc. of the 2006 IEEE Int'l Conf. on Fuzzy Systems. IEEE, 2006. 997−1003.

[25] Baker FB, Kim SH. Item Response Theory: Parameter Estimation Techniques. CRC Press, 2004.

[26] Miller VS. The Weil pairing, and its efficient calculation. Journal of Cryptology, 2004,17(4):235−261.

[27] Lynn B. The pairing-based cryptographic library. 2015. http://crypto.Stanford.edu/pbc/

[28] Loukides M, Oram A. Programming with GNU Software. O'Reilly & Associates, 1997,86(3):350−359.

[29] Steiner M. The PBC_bce broadcast encryption library. 2006. https://crypto.stanford.edu/pbc/bce/

[30] Hu XT, Qin ZP, Zhang H, Hao GS. Research and improved implementation of AES algorithm in OpenSSL. Control & Automation, 2009,25(12):83−85.

附中文参考文献:

[3] 付印金,肖侬,刘芳.重复数据删除关键技术研究进展.计算机研究与发展,2012,49(1):12−20.

[12] 杨超,纪倩,熊思纯,刘茂珍,马建峰,姜奇,白琳.新的云存储文件去重复删除方法.通信学报,2017,38(3):25−33.

[17] 张曙光,咸鹤群,刘红燕,侯瑞涛.云存储环境中基于离线密钥传递的加密重复数据删除方法研究.信息网络安全,2017(7):66−72.

[19] 林吓洪.社区化网络中的隐私保护[博士学位论文].上海:上海交通大学,2010.

咸鹤群(1979－),男,博士,副教授,CCF 高级会员,主要研究领域为网络与信息系统安全.

穆雪莲(1995－),女,硕士,主要研究领域为信息安全.

高原(1994－),女,博士生,主要研究领域为信息安全.

高文静(1997－),女,博士生,主要研究领域为信息安全.