DOI:10.20033/j.1003-7241.(2026)08-0151-05

基于区块链的非结构化大数据云存储优化研究

赵维伟[1], 潘建宏[2], 宫 鑫[3], 刘兵兵[3], 张 亮[4]

(1. 福建亿榕信息技术有限公司,福建 福州 350001; 2. 国网吉林省电力有限公司, 长春 130000;
3. 国网山西省电力公司信息通信分公司, 太原 030000; 4. 国网宁夏电力有限公司安监部, 银川 750000)

摘 要:非结构化大数据云存储过程中冗余数据未被有效剔除,导致存储效率低、数据失真率较高。为此,提出一种基于区块链的非结构化大数据云存储优化方法,以提升存储完整性、降低数据冗余与失真率为目标。该方法首先构建基于区块链的冗余数据剔除结构,通过最优分类平面对冗余数据进行分类与剔除;进而引入信息熵评估数据不确定性,结合主成分分析实现数据降维,滤除无效信息;在此基础上,构建区块链存储优化模型,依据行、列读写特性优化数据访问速度与存储完整性;最后设计含第三方仲裁机制的完整性验证方案,用于分析完整性证明篡改与拒绝应答两类存储安全问题。实验结果表明,在完整性证明篡改问题下该方法丢失 22 个标记块,在拒绝应答问题下丢失 30 个标记块,与实际数据块丢失情况完全一致;在 1 MB 至 5 GB 数据量范围内,数据失真率最高仅为 12%,明显低于基于 SDN 技术、粒子群优化算法和局部对称算法的对比方法。所提方法能够有效剔除冗余数据、降低数据维度,提升非结构化大数据云存储的完整性与访问效率,具备良好的实用性与稳定性。

关键词:区块链;非结构化数据;云存储;冗余数据剔除;数据降维;存储优化;完整性验证;数据失真率

中图分类号: TP311.13;TN911.34 **文献标志码**: A **文章编号**: 1003-7241(2026)08-0151-05

Research on optimization of unstructured big data cloud storage based on blockchain

Zhao Weiwei[1], Pan Jianhong[2], Gong Xin[3], Liu Bingbing[3], Zhang Liang[4]

(1. Fujian Yirong Information Technology Co., Ltd., Fuzhou 350001, Fujian, China;
2. State Grid Jilin Electric Power Co., Ltd., Changchun 130000, China;
3. State Grid Shanxi Electric Power Company, Taiyuan 030000, China;
4. Safety Supervision Department, State Grid Ningxia Electric Power Co., Ltd., Yinchuan 750000, China)

Abstract: The ineffective elimination of redundant data in unstructured big data cloud storage leads to low storage efficiency and high data distortion rates. To address this issue, an optimization method for unstructured big data cloud storage based on blockchain is proposed, aiming to improve storage integrity and reduce data redundancy and distortion rates. In this method, a blockchain-based redundant data elimination structure is first constructed to classify and remove redundant data using an optimal separating hyperplane. Information entropy is then introduced to assess data uncertainty, and principal component analysis is employed for dimensionality reduction to filter out invalid information. On this basis, a blockchain storage optimization model is developed, optimizing data access speed and storage integrity based on row and column read-write characteristics. Finally, an integrity verification scheme with a third-party arbitration mechanism is designed to analyze two types of storage security issues: integrity proof tampering and denial of response. Experimental results show that the proposed method loses 22 marker blocks under integrity proof tampering and 30 marker blocks under denial of response, which are completely consistent with actual data loss. The maximum data distortion rate is only 12% across data volumes ranging from 1 MB to 5 GB, significantly lower than the comparison methods based on SDN technology, particle swarm optimization, and local symmetry algorithms. The proposed method effectively eliminates redundant data, reduces data dimensionality, and enhances the integrity and access efficiency of unstructured big data cloud storage, demonstrating strong practicality and stability.

Keywords: blockchain; unstructured data; cloud storage; redundant data elimination; data dimensionality reduction; storage optimization; integrity verification; data distortion rate

目前,互联网上的数据大多具有结构性,但在云平台上,这些非结构性数据无定形,没有固定格式,也没有约定规则,处理难度相对较高。在云计算发展背景下,非结构化数据也出现了迅速增长态势,这让非结构化数据存储负担和代价都有了很大提高,与此同时也对其应用有了更高的要求。

收稿日期:2025-11-22;**录用日期**:2025-12-26

基金项目:国网总部科技项目(SGGSXT00XMXX2100312)

作者简介:赵维伟(1983—),男,助理工程师,研究方向为项目管理及非结构化数据管理。

引用本文:赵维伟,潘建宏,宫鑫,等. 基于区块链的非结构化大数据云存储优化研究[J]. 自动化技术与应用,2026,45(8):151-155. (Zhao Weiwei, Pan Jianhong, Gong Xin, et al. Research onoptimization of unstructured big data cloud storage based on blockchain [J]. Techniques of Automation and Applications, 2026, 45(8): 151-155.)

针对此问题,相关领域专家已经得到了一些较好的研究成果,例如:文献[1]提出了基于 SDN 技术的云存储方法,从软件系统网络结构出发,着重研究软件 Power、数据区域、接口协议等关键技术,并在此基础上搭建软件系统空间数据中心,通过 Hbase 技术扩展软件系统的空间数据存储容量,减少对硬盘的占用。利用四叉树结构组织图像数据,并利用四叉树形结构对单幅图像编码,从而满足多进程空间数据存储要求;文献[2]提出了基于粒子群优化算法,提出了一种新的分布式数据库管理方法,利用基尼系数作为测度变量,构建多任务协作的多目标约束优化模型,并与多任务协作进行对比分析。在此基础上,结合基于粒子群优化多智能体求解多目标约束函数,依次通过节点编码及参数设置、种群初始化、粒子群优化、迭代处理,存储杂乱无章的海量数据;文献[3]提出了局部对称算法,将区块链网络中的节点划分为多个节点子集,每个节点子集生成一个私钥,将该私钥加密到节点子集中,再利用局部对称性等手段对该私钥进行加密,最终将该私钥存储在节点子集中。但是,虽然上述方法有助于减少云平台中的信息传递,但是对信息非结构化特性缺乏针对性。

为此,提出了基于区块链的非结构化大数据云存储优化研究方法。通过最优分类平面求解问题,将冗余数据分类问题转化为最优分类操作方式的问题。这种方法能够针对冗余数据进行高效的分类,并将其剔除,从而减少存储空间的占用。这样可以优化大数据的存储效率和容量。在冗余数据剔除的过程中,充分考虑不同情况下信息源的不确定性,并计算信息熵来评估数据的不确定性。这种方法能够更准确地处理数据,降低信息冗余性,提高存储效率和数据质量。充分考虑到行列读写特点,针对非结构化大数据进行存储优化。这样可以提高非结构化数据在云存储中的速度和完整性,为用户提供更快、更可靠的数据访问服务。

1 基于区块链的冗余数据剔除

不同于传统 Merkle 树的存储数据块结构,在该模式中由于存在着大量的交易数据,所以设计了一个用于非结构化大数据云存储的冗余数据剔除区块链结构,如图 1 所示。

图 1 所示的区块结构与常规结构最大的不同之处是区块体结构和区块头所存储的散列值。数据哈希值用于确保交易数据的完整性,填充字段用于确保各个节点数据的长度相同。因为冗余数据的存在,给数据的正常维护带来了极大的影响,为了保证各个节点的区块头一致,使得各个节点的工作效率不会受到影响,需要对其分类处理[4]。由于采集过程中存在一定网络时滞,所以采用局部特征分析方式通过与相邻地区特性进行对比,以此反映冗余数据特征[5]。

通过最优分类操作方式,将冗余数据分类问题转化为最优分类平面求解问题,函数可表示为

$$f(\varphi)=\sum_{i=1}^{a}\varphi_i\boldsymbol{\omega}_i-\frac{1}{2}\varphi_j\boldsymbol{\omega}_j(\boldsymbol{h}_i\cdot\boldsymbol{h}_j) \tag{1}$$

式中,$f(\varphi)$ 表示分类阈值 φ 的二次判别函数;$\boldsymbol{a}$ 表示最大向量;φ_i、φ_j 分别表示第 i、j 个分类阈值;$\boldsymbol{\omega}_i$、$\boldsymbol{\omega}_j$ 分别表示 $\boldsymbol{h}_i$、$\boldsymbol{h}_j$ 两个向量的权重向量[6-8]。

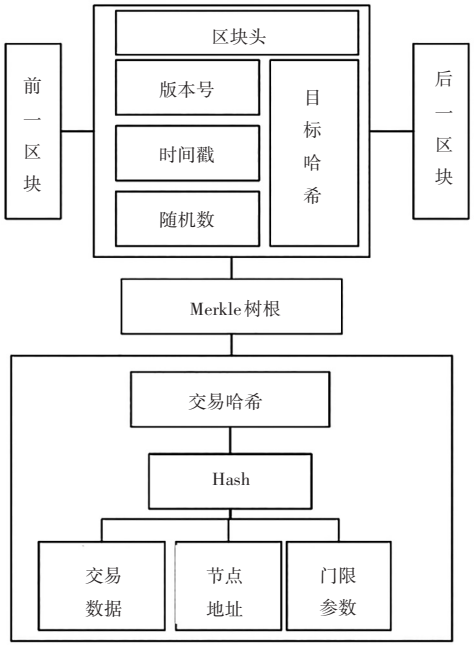


图 1 非结构化大数据云存储冗余数据剔除区块链结构

Fig. 1 The structure of blockchain for eliminating redundant data in unstructured big data cloud storage

假设非结构化大数据云存储中冗余数据特征是非线性转换特征,那么需要通过内积代替最优分类标量,由此构建的平面可表示为

$$g(\boldsymbol{h})=f(\varphi)\operatorname{sgn}\Big(\sum_{i=1}^{a}\varphi_j\boldsymbol{\omega}_j(\boldsymbol{h}_i\cdot\boldsymbol{h}_j)+u\Big) \tag{2}$$

式中,u 表示分类属性[9]。在此基础上,可以获得经过密度进化的非结构化大数据云存储冗余数据,并分类和剔除。

2 区块链大数据云存储优化策略

假定在云环境中面对大量的非结构化数据,总量为 m,持久化集合为 $r=(1,2,\cdots,n)$,集合中 n 为非结构化网络节点数量,即表示存储数据量。基于此,存储数据可表示如下:

$$\boldsymbol{h}_m=g(\boldsymbol{h})\boldsymbol{\alpha}_m[x_1,x_2,\cdots,x_n] \tag{3}$$

式中,$\boldsymbol{\alpha}_m$ 表示域中的行向量;x_n 表示第 n 个节点的数据包[10]。

通过对冗余数据筛选,在主成分分析算法[11]中加入信息熵,实现对网络化大数据的有效降维。在使用主成分分析算法对网络化的稀疏大数据降维前,先利用信息熵滤除其中无效信息,具体步骤如下。

充分考虑不同情况下信息源的不确定性,计算信息熵。

$$\zeta = \boldsymbol{h}_m \sum_{m=1}^{n} p_m \lg p_m \tag{4}$$

式中，p_m 表示存储数据中信号属性对应的概率[12]。设信息熵阈值为 γ，对比分析 ζ 与 γ 之间的关系，如果 $\zeta(\boldsymbol{h}_m) > \gamma$，那么将 $\boldsymbol{h}_m$ 存储到集合中[13]。计算存储在该集合中数据不同维度之间的协方差，得到的协方差矩阵为

$$\boldsymbol{C}_{ov} = \frac{\boldsymbol{h}\boldsymbol{h}^c}{\zeta(\boldsymbol{h},2) - 1} \tag{5}$$

式中，c 表示矩阵阶[14]。将该矩阵正交化分解，使该矩阵中所出现的矢量描述了在本征矢量中的投射长度，该投射长度就是矢量的特征值。在此基础上，分别求出协方差矩阵各部分的本征值，并剔除了其余部分，从而获得了协方差矩阵的特征值。

选取 k 个较大特征值构成特征向量 $\boldsymbol{\vartheta}_{mk}$，由此得到的大数据降维结果为

$$\boldsymbol{h} = \boldsymbol{\vartheta}_{mk}^{c} x \tag{6}$$

根据大数据降维结果，设计如图 2 所示的存储优化模型。

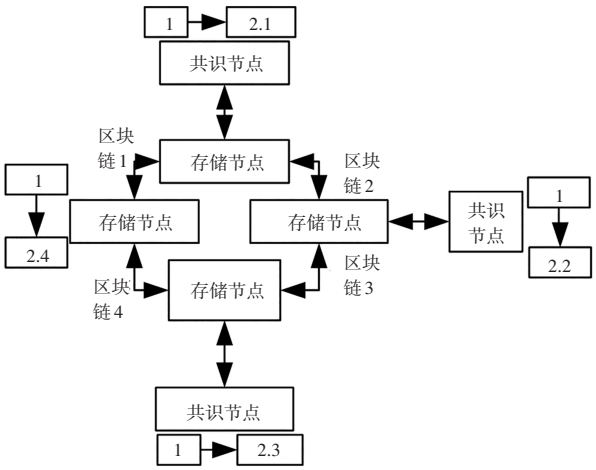


图 2　剔除冗余数据的区块链大数据云存储优化模型

Fig. 2　Optimization model for blockchain big data cloud storage with redundant data removed

根据图 2 可知，构建 n 个数据区块来储存这 n 个小块，然后将 $n-1$ 个区块分配到区块链网络中的 $n-1$ 个存储节点上。储存节点负责数据存储，在该节点接收数据时，必须先验证该节点、再连接[15]。除此 n 个节点外，其他节点均为一致节点，它们需向邻近 n 节点中的某一节点请求数据块，或接收其广播并链接至区块链。

为确保数据的唯一性，行向量 $\boldsymbol{\alpha}_m$ 应该遵循以下分布规律。

$$P_n^{'} = \begin{cases} \boldsymbol{C}_{ov} P^{'}, & d = 1 \\ \boldsymbol{C}_{ov}(1 - P^{'}), & d = 0 \end{cases} \tag{7}$$

式中，$P^{'}$ 表示节点选择概率；d 表示云存储节点间距。

当所有元素均满足存储规则的前提下，根据调度控制域判断数据属性，并实时更新存储规则。

依据该存储规则，需要先确定大数据访问时间复杂程度，在图 2 指定区块链大数据云存储优化模型下，通过如下公式，计算大数据查询所消耗的时间。

$$T = p_a^{''} T_a + p_b^{''} T_b \tag{8}$$

式中，$p_a^{''}$、$p_b^{''}$ 分别表示数据读操作和写操作出现的概率；T_a、T_b 分别表示数据读操作和写操作耗费的时间。

根据行、列读写特点，非结构化大数据云存储优化速度可表示为

$$V = \begin{cases} \dfrac{t}{T} & \text{列存储} \\ 1 & \text{行存储} \end{cases} \tag{9}$$

式中，t 表示查询所消耗时间中需要进行数据排列的时间。通过该计算结果，能够优化非结构化大数据云储存速度。根据区块链结构特点可知，当数据被高速度访问后，在一定时间内数据被访问频率速度也较快，表明该时间内数据能够被完整性存储的概率就高。

3　存储完整性验证

3.1　验证方案

针对非结构化大数据的完整性问题，设计了如图 3 所示的完整性验证过程。

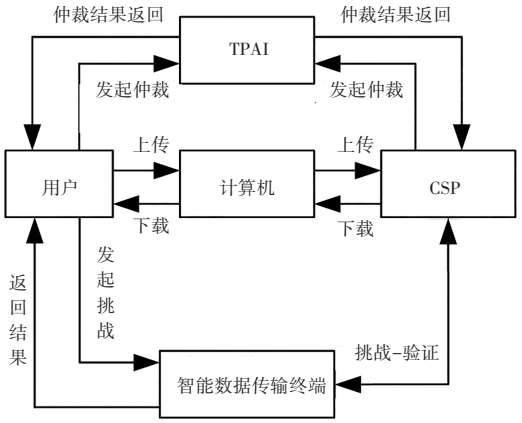


图 3　完整性验证过程

Fig. 3　Integrity verification process

引入第三方仲裁机制，用于解决客户与云计算服务供应商之间的数据完整性问题。该解决方案不仅涉及 TPAI，还涉及用户、CSP 以及区块链等多个方面。所谓的“客户”，可以是个体，也可以是中小企业；CSP 提供了一种高可靠、高安全性、低成本、易于扩充的云端存储服务；BCN 为用户提供了一个存储与验证的函数，将用户的数据完整性与二次验证信息存储在一个块内，并基于该块的抗篡改特性，对数据进行了合法的完整性证明。BCN 中的 Smart 合约将从 CSP 发出一个请求，然后对 CSP 的回复一一确认；TPAI 是一个可信的合法仲裁组织，它能够保证由客户和 CSP 共同提出的仲裁请求信息的完整性。

整个方案可以分为 3 部分。

1）数据上传，用户先对数据进行加密，再进行用户身份验证。然后，利用身份验证的树形结构，将 DIPT、AVIT 等信息传递到区块链上。

2）挑战-应答，在 BCN 中用户发送一个请求，BCN 中

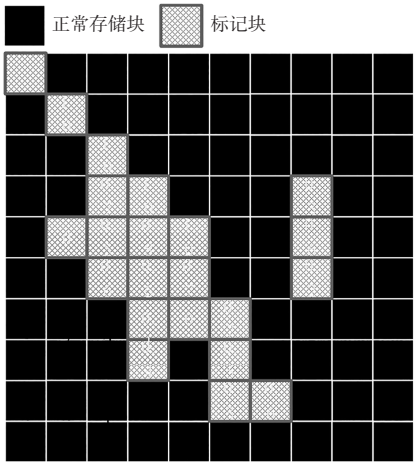
的智能合约将该请求转发给 CSP,CSP 对该请求作出回应,并将该回应信息发送给智能合约。智能合约基于存储在模块中的 AVIT,对应答确认,并将确认的结果传递给用户。

3) 数据完整性仲裁,当用户从智能合约中接收到一项成功的要求后,就会在 TPAI 上执行一项数据完整性的裁决。TPAI 利用 DIPT 在 CSP 服务器端验证数据的正确性,并将决策结果反馈给用户。此外,CSP 也能利用 TPAI 技术对非法用户进行处理。

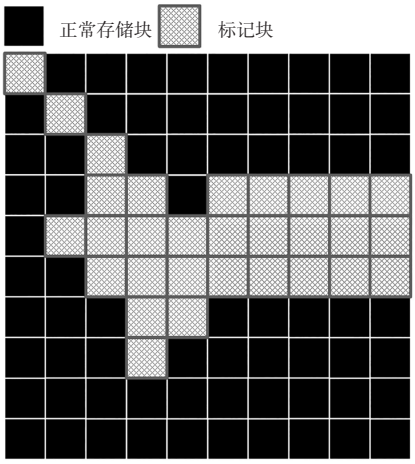
3.2 存储安全问题设置

设置了两类储存安全性问题。一类是可验证的完整性证明篡改问题,另一类是可验证的拒绝应答问题。其中,完整性证明篡改问题是一种被恶意攻击者入侵的内存,从而对内存的完整性验证进行修改的问题。拒绝应答问题是指 CSP 不回应使用者发出的请求,当使用者发出请求后,若未得到回应,则会重新发出该请求;当回复超时时,将回复超时的结果发送给该使用者。

基于上述两种存储安全问题,分析实际存储数据丢失情况,如图 4 所示。



(a) 完整性证明篡改问题



(b) 拒绝应答问题

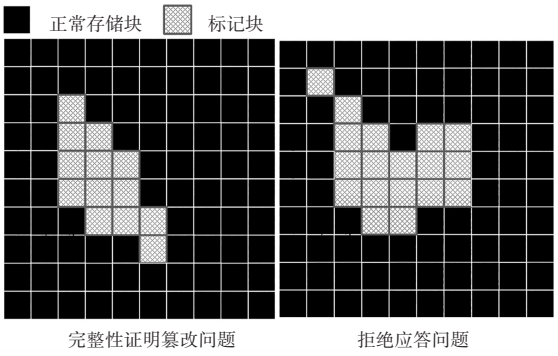
图 4 实际存储数据丢失情况

Fig. 4 Actual loss of stored data

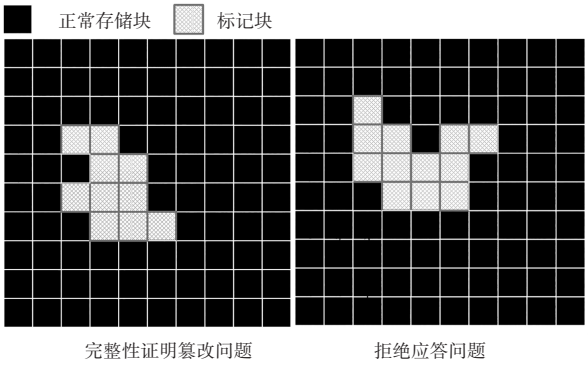
由图 4 可知,标记块越多,说明存储数据丢失越多。在完整性证明篡改问题下,共丢失 22 个标记块;在拒绝应答问题下,共丢失 30 个标记块。

3.3 验证结果分析

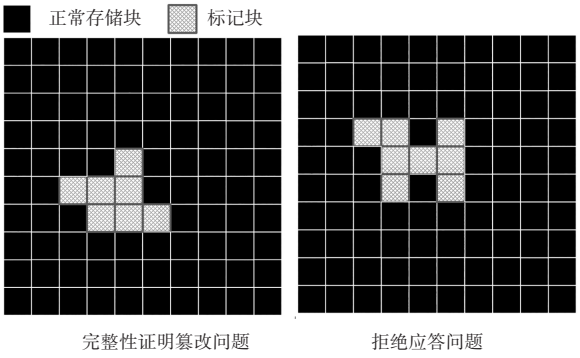
将基于 SDN 技术的云存储方法、文献[2]中的基于粒子群优化算法、文献[3]中的局部对称算法与基于区块链优化方法进行对比分析,获取的存储数据丢失情况如图 5 所示。



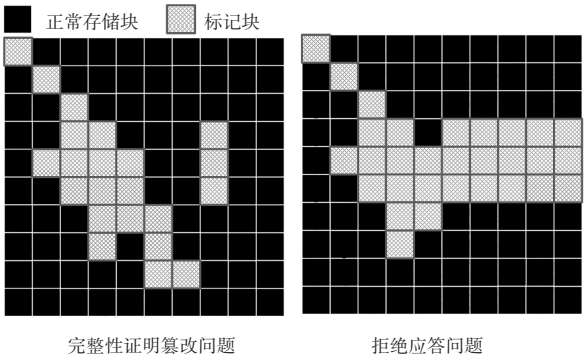
(a) 基于 SDN 技术的云存储方法



(b) 基于粒子群优化算法



(c) 局部对称算法



(d) 基于区块链优化方法

图 5 不同方法标记存储数据丢失情况对比分析

Fig. 5 Comparative analysis of different methods for labeling and storing data loss situations

由图 5 可知,使用基于 SDN 技术的云存储方法在完整性证明篡改问题下,共丢失 13 个标记块;在拒绝应答问题下,共丢失 18 个标记块,与实际数据块丢失存在最大为 12 个的标记误差。

使用基于粒子群优化算法在完整性证明篡改问题下,共丢失 10 个标记块;在拒绝应答问题下。共丢失 12 个标记块,与实际数据块丢失存在最大为 18 个的标记误差。

使用局部对称算法在完整性证明篡改问题下,共丢失 7 个标记块;在拒绝应答问题下,共丢失 8 个标记块,与实际数据块丢失存在最大为 22 个的标记误差。

使用基于区块链优化方法在完整性证明篡改问题下,共丢失 22 个标记块;在拒绝应答问题下,共丢失 30 个标记块,与实际数据块丢失情况一致。

为了进一步分析所研究方法的存储优化效果,再次对比四种方法的数据失真率,对比结果如表 1 所示。

表 1　4 种方法的数据失真率对比分析

Tab. 1　Comparative analysis of data distortion rates for four methods　%

数据量	SDN 技术	数据失真率		
		粒子群	局部对称	区块链优化
1 MB	10	8	9	1
100 MB	15	14	17	3
200 MB	22	18	19	5
500 MB	30	27	26	8
1 GB	45	40	39	10
5 GB	65	58	55	12

由表 1 可知,使用基于 SDN 技术的云存储方法、基于粒子群优化算法、局部对称算法数据失真率随着数据量增加,均较高;使用所研究方法数据失真率随着数据量增加,最高也只为 12%,相比于其他方法数据失真率较低。

4　结论

针对目前云存储中存在的数据一致性不高、数据质量不稳定等问题,研究一种基于区块链的云储存优化方法,以减少数据中的冗余数据,降低数据维度,实现不同数据库间的高效共享。采用了分布式的数据存储原理,将其分为行存储和列存储。通过实验验证结果发现,该方法在数据的存储完整性和数据失真率上都比常规数据处理方式有显著提高,证明了该方法具有一定实用性。

参考文献

[1]潘志安,王金峰,王茂发. 基于 SDN 技术的多进程空间数据云存储方法研究[J]. 计算机仿真,2021,38(5):375-379.
[2]郑力宁,金雪松,云利军. 基于粒子群优化算法的云数据均衡放置策略[J]. 广西师范大学学报(自然科学版),2022,40(3):202-209.
[3]刘云,李公瑾,马儀,等. 局部对称算法在区块链分布式存储中的优化[J]. 兰州大学学报(自然科学版),2023,59(2):190-195.
[4]张星炜,徐鑫乾,王红星,等. 基于数据挖掘的海量激光点云数据快速集成方法[J]. 自动化技术与应用,2024,43(12):93-97.
[5]刘陈伟,孙鉴,雷冰冰,等. 基于改进粒子群算法的云数据中心能耗优化任务调度策略[J]. 计算机科学,2023,50(7):246-253.
[6]李大鹏,李立新,杨清波,等. 云边协同的调控云数据质量优化[J]. 电力系统及其自动化学报,2022,34(3):11-19.
[7]郑力宁,金雪松,云利军. 基于粒子群优化算法的云数据均衡放置策略[J]. 广西师范大学学报(自然科学版),2022,40(3):202-209.
[8]刘云,马儀,田地,等. 可信激励算法对区块链移动节点共识优化研究[J]. 四川大学学报(自然科学版),2022,59(6):68-76.
[9]张秋平,孙胜,刘敏,等. 面向多边缘设备协作的任务卸载和服务缓存在线联合优化机制[J]. 计算机研究与发展,2021,58(6):1318-1339.
[10]文鹏程,沈济南,梁芳,等. 云环境中细粒度的数据安全共享方案[J]. 武汉大学学报(理学版),2022,68(1):93-101.
[11]董俐君,岳恒,李俊峰. 基于主成分聚类分析的电力用户信息自动集成方法[J]. 自动化技术与应用,2025,44(09):124-128.
[12]刘峰,赵俊峰. 基于区块链的云存储数据完整性验证方案[J]. 应用科学学报,2021,39(1):164-173.
[13]曾梦,邹北骥,张文生,等. 多模态医疗数据中海量小文件存储优化方法[J]. 软件学报,2023,34(3):1451-1469.
[14]刘艳,曹晓倩. 海量视频数据分布式存储性能优化方法研究[J]. 计算机应用研究,2021,38(6):1734-1738.
[15]赵越,余红英,王一奇. 一种高速数据存储方法的设计与验证[J]. 数据采集与处理,2021,36(2):384-390.