基于分数阶Fourier变换的云存储系统重复数据删除算法

徐奕奕[1,2] 唐培和[1]
(广西科技大学计算机科学与通信工程学院 柳州 545006)[1]
(武汉理工大学信息工程学院 武汉 430070)[2]

摘 要 云存储系统的重复数据作为大量冗余数据的一种,对其有效及时地删除能保证云存储系统的稳定与运行。由于云存储系统中的干扰数据较多,信噪比较低,传统的重删算法会在分数阶Fourier域出现伪峰峰值,不能有效地对重复数据进行检测滤波和删除处理,因此提出一种改进的基于分数阶Fourier变换累积量检测的云存储系统重复数据删除算法。首先分析云存储系统重复数据删除机制体系架构,定义数据存储点的适应度函数,得到云存储节点的系统子集随机概率分布;采用经验约束函数对存储节点中的校验数据块分存,通过分数阶Fourier变换对云存储系统中的幅度调制分量进行残差信号滤波预处理。采用4阶累积量切片后置算子,把每个文件分为若干个块,针对每个文件块进行重删,进行重复数据检测后置滤波处理,实现存储资源上的重复数据检测及其删除。仿真实验表明,该算法能提高集群云存储系统计算资源的利用率,重复数据准确删除率较高,有效避免了数据信息流的干扰特征造成的误删和漏删,性能优越。

关键词 分数阶Fourier变换,云存储,重复数据
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.038

Duplicate Data Remove Algorithm of Cloud Storage System Based on Fractional Fourier Transform

XU Yi-yi[1,2] TANG Pei-he[1]
(School of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China)[1]
(School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China)[2]

Abstract Duplicate data of cloud storage system is taken as one of a large amount of redundant data, and the effective and timely remove can guarantee the stability and operation of cloud storage system. Because of the interference of data, the SNR is low, the traditional method has false peaks in the fractional Fourier domain, and it cannot effectively detect and remove the duplicate data. An improved duplicate data remove algorithm of cloud storage system was proposed based on fractional Fourier transform cumulant detection. Firstly, the delete system architecture for cloud storage system was taken, the fitness function of data storage point was defined, and system subset random probability distribution function of the cloud storage node was gotten. The constraint function was used for blocking the calibration data of storage nodes, the detection of duplicate data removing processing was taken, and the fractional Fourier transform was used to preprocess the residual signal filtering in cloud storage system. The 4 order cumulanted slice post operator was used to divide each file into blocks. To delete each file block, duplicated data detection post filtering was obtained, and data storage resource detection and deletion were realized. Simulation results show that this algorithm can improve the utilization efficiency of cluster cloud storage system resource, and duplicate data can be accurately removed with higher rate. It can effectively avoid the error removing caused by interference and leakage removing, and it has superior performance.

Keywords Fractional Fourier transform, Cloud storage, Duplicate data

1 引言

随着计算机信息技术的飞跃发展,人类进入了信息化时代,云存储系统作为信息化发展的产物,在物联网、移动互联网和SNS等数据管理和大规模数据集成中扮演着重要角色,成为分布式文件管理系统和各类存储设计集成的重要工具和载体。云系统中的重复数据是各类数据管理成本快速上升过程中留下的冗余数据产物,云存储系统中产生的数据量以几何级数增长,导致网络带宽和存储空间资源的紧缺以及数据管理成本的快速上升,云存储系统中的冗余数据带来了高额

到稿日期:2014-08-29 返修日期:2014-09-28 本文受广西自然科学基金青年基金项目(2013GXNSFBA019268),广西科技大学自然科学基金项目(校科自1261126),广西特色专业建设项目(GXTSZY217),广西教育厅一般项目(YB2014208),广西教育厅立项项目(LX2014182)资助。
徐奕奕(1980—),女,博士,副教授,主要研究方向为网络存储与云存储;唐培和(1964—),男,硕士,教授,主要研究方向为分布式计算与分布式存储。

的硬件和人力成本，降低了系统的性能。为了有效满足爆炸式增长的云存储系统运行数据管理的需求，减轻服务器开销，研究一种有效的云存储系统重复数据删除算法，对消除数据冗余、降低系统能耗和提高存储性能具有重要意义，重复数据删除技术成为云数据存储领域的一个重要的研究热点。

研究表明，在云存储应用系统所保存的数据中，高达60%的数据是冗余的，以重复数据为主导的冗余数据所占比例也将随着时间的推移而上升，重复数据删除作为一项应用于存储系统上的数据管理技术，有必要结合数据特征和存储规模来探讨。结合信号与信息处理的相关理论，采用数据信息检测和数字滤波方法，可实现对重复数据的检测和删除[1]。其中现有的云存储开源重删系统主要有 Lessfs 和 OpenDedup 系统，其建立在应用层之上[2]，通过用户对数据系统状态的识别，避免了用户对底层的管理，但系统没能通过信息处理的方法实现对重复数据的删除，实际效果不好；Wu T Y 等人提出了基于负载均衡的云系统漂移数据删除算法[3]，在概念学习系统的基础上，采用决策树 ID3 算法对重复数据进行剪枝，实现重复数据删除，但该算法计算开销大，应用性不好；文献[4]中蒋海波、王晓京等人提出一种基于水平纠删码的云存储数据布局和重删方法，即采用水平阵列信号处理技术，实现了可容 3 列随机重复数据的纠删码删除算法，实现对云存储数据的优化布局，但该算法需要采用水平纠删码扩展的方法进行线性编译，降低了丢失数据的恢复性能；文献[5]对传统的水平纠删码扩展重复数据删除进行改进设计，对 RS 码进行了改造，在 Galois 域上进行重复查询，采用分数阶 Fourier 变换方法对急剧增加的海量重复数据进行检测滤波处理，然而该算法对高阶累积的重复数据的滤波效果不好，删除性能低；文献[6]采用基于负载均衡的客户/服务器(Client/Server，C/S)双端重复数据删除机制，综合考虑了负载均衡和动态副本管理，该算法对当期重复数据删除效果尚可，但对全局收敛性较低的动态副本数据删除效果较差；文献[7]提出一种基于经验模态分解和粒子滤波的数据检测算法来对云存储系统中的重复数据进行删除，由于没有对累积量特征进行滤波后置检测，导致对云存储系统的重复数据的删除性能不好，且该算法采用单节点架构，扩展性不佳，难以得到大规模应用。

针对上述问题，本文提出一种改进的基于分数阶 Fourier 变换累积量检测的云存储系统重复数据删除算法。首先对云存储系统重复数据删除机制体系架构进行研究，构建云存储系统中重复数据流信息模型与并进行信号特征预处理，采用分数阶 Fourier 变换 4 阶累积量算法实现对重复数据删除算法的改进，仿真实验验证了算法的可行性和优越性。

2 云存储系统重复数据删除机制体系架构与问题描述

2.1 相关预备知识与体系架构

本文设计的云存储系统漂移数据删除体系构架分为 3 个主体：客户端(Client)、控制器(Controller)和服务器(Server)。客户端作为源端设备，是整个系统数据的原始提供者，客户端上传文件到云端；控制器负责管理用户的请求，在元数据块与数据块的修改模式下，对每一个写请求进行基于 Hash 指纹识别的数据冗余判别；服务器便是数据上传的最终存储节点，数据上传至服务器端时，采用延迟删除来提高系统的效率。重复数据删除算法设计中，首先需要进行文件切分，将文件切分为若干数据块(Chunk)，将每个完整的文件当作一个 Chunk 来进行分块，得到的完整数据信息流为：

$$x_{id}(t+1)=wx_{id}(t)+c_1r_1[r_3^{t_0}>T_0\ p_{id}-x_{id}(t)]+c_2r_2[r_4^{t_g}>T_g\ p_{gd}-x_{id}(t)] \tag{1}$$

式中，t_0 和 t_g 分别表示数据块边界偏移的个体极值和全局极值进化停滞步数；T_0 和 T_g 分别表示个体极值和全局极值需要扰动的停滞步数阈值。在云存储系统中，为了适应云存储中的多 QoS 偏好，重新定义数据存储点的适应度函数如下：

$$f_{ij}=w_t\delta_t+w_c\delta_c+w_q\delta_q+w_s\delta_s \tag{2}$$

其中，$w_t+w_c+w_q+w_s=1$，t 代表时间(time)，c 代表代价(cost)，q 代表质量(quality)，s 代表安全(security)，不同应用文件在选择相同的划分策略和指纹提取方法时，各个存储子集间将组合成一个具有层次结构树状图，得到该云存储节点的系统子集随机概率分布函数为：

$$w(e_pk_q)=\alpha\times w(s_pk_q) \tag{3}$$

在云存储系统中，重复数据的双随机概率分布函数的权重主要依据它在所属双随机概率分布函数中出现的概率来计算，而所有双随机概率分布函数都是围绕一个核心概率函数展开的，展开结果为：

$$\tilde{u}_{e|v,k}=\tilde{u}_{e,k}+\tilde{\Sigma}_{ve,k}^{T}\tilde{\Sigma}_{vv,k}^{-1}(v_k-\tilde{u}_{v,k}) \tag{4}$$

如果将 w_k 按照 v_k 和 e_k 的组成原则进行分解，得到重复数据信息流的存储节点权重 $w(e_pk_q)$，以此计算云存储集群系统的校验信息存储子集，计算公式为：

$$\lambda^n(d_{\gamma_0})=\int_{-\infty}^{+\infty}f(t)d_{\gamma_0}^{*}(t)\mathrm{d}t \tag{5}$$

上式表示一个具有 n 个输入控制参量、m 个输出参量分簇云存储器的校验块跨越编码，采用经验约束函数对存储节点中的校验数据块分存，分析云存储系统的校验数据块的重复数据状态向量，由此构建云存储系统重复数据删除机制体系，为进行云存储系统的重复数据删除算法的设计提供模型依据和总体框架。

2.2 云存储系统中重复数据流信息模型与预处理

从上述构建的总体模型分析可见，在云存储系统中，当客户端发出文件存储请求时，文件服务器根据子集校验数据块的任务执行状态和文件读取需求进行数据文件信息分区处理，为提高集群存储系统计算资源的利用率，需要对重复数据进行删除，以提高系统的存储介质性能。构建云存储系统中重复数据流信息模型是进行重删设计的关键[9]，假设云存储系统重复数据流信息的目标端信息分量为：

$$r_1=x(t)-c_1 \tag{6}$$

采用窗函数宽度可变方法对频率分辨力进行调整，重复数据一般都会备份到远端存储节点，可以利用重删系统中文件信息流进行相位信息评估，得到重复数据出现的概率权重为：

$$w_{ij}=\beta\times w(e_p k_q),\beta>1 \tag{7}$$

重复数据信息流通过网络传输到远端存储节点，得到输出向量模型为：

$$x_j'(k)=\frac{1}{1+e^{-u_j'(k)}},j=1,2,3 \tag{8}$$

上式描述了多源进程节点的云存储系统的任务执行模型。在云存储系统中，校验数据块分存到子集云存储系统中，得到重复数据集合为：

$$P=\{p_1,p_2,\cdots p_m\},m\in N \tag{9}$$

对多个任务流中的重复数据进行在线编码调度，云存储系统将启动下一存储子集，客户端向文件服务器请求源数据，得到存储系统生成校验位为：

$$flow_k=\{n_1,n_2,\cdots,n_q\},q\in N \tag{10}$$

式中，q 表示多个云存储节点信息流集合的特征编码位置，n_q 表示数据信息流的数据序列，N 表示信息位总数。通过上述分析，得到了云存储系统的重复数据流信息模型。本文采用信息流检测和滤波处理方法，来实现对重复数据的检测和重删处理，引入分数阶 Fourier 变换信号检测方案，假设云存储系统中重复数据信息流为 $x(t)$，分数阶 Fourier 变换的定义为：

$$X_p(u)=F^p[x(t)]=\int_{-\infty}^{\infty}K_p(t,u)x(t)\mathrm{d}t \tag{11}$$

式中，p 为分数阶 Fourier 域的阶，为实数，旋转角 $\alpha=p\pi/2$。$F^p[\cdot]$ 表示变换算子形式记号，$K_p(t,u)$ 是 FRFT 的变换核。这样，通过对重复数据信息流的 IMF 分量幅度调制的信息，得到信号 $x(t)$ 的第 n 个 IMF 分量，表示为：

$$r_1-c_2=r_2,\cdots,r_{n-1}-c_n=r_n \tag{12}$$

采用分数阶 Fourier 变换对上述云存储系统中的幅度调制分量进行残差信号滤波预处理。通过第一次筛分后去除残差信号，提取出满足固有模态函数的 IMF 分量，得到：

$$x(t)=\sum_{i=1}^{n}c_i+r_n \tag{13}$$

式中，c_i 代表各 IMF 分量，r_n 代表残余均方差的估计值，由此实现了对云存储系统中重复数据信息流的预处理。通过上述分析可见，采用传统的基于分数阶 Fourier 域的残差信号滤波处理还不能有效滤除重复数据，云存储系统中过多的线程将会竞争有限的资源，导致系统性能较低，且根据传统方法解出的信息位不能有效定位重复数据的存储节点和相位信息，需要进行算法改进，以提高对云存储系统的重复数据删除性能。

3 改进的分数阶 Fourier 变换重复数据删除算法的实现

3.1 云存储系统中重复数据信息流检测滤波的改进设计

在上述构架的云存储系统重复数据删除总体模型中，对传统的基于分数阶 Fourier 变换方法的海量重复数据检测滤波和删除算法进行改进，采用分数阶 Fourier 变换高阶累积量算法对重复数据进行检测处理，算法描述如下：

根据式(11)定义的分数阶 Fourier 变换表达式，以及重复数据的丢失信息流特征，进行特征分解，实现对存储系统中的重复数据信息流的分数阶 Fourier 域构造，得到简化后的 Fourier 变换表达式：

$$X=F_a\cdot x \tag{14}$$

式中，

$$X=[X_a(0),X_a(1),\cdots,X_a(N-1)]^{\mathrm{T}} \tag{15}$$

其中，F_a 是 $N\times N$ 维矩阵，存储系统中各个节点需要创建多个线程的信息流特征编码，得到重复数据信息流矩阵各元素为：

$$F_a(m,n)=A_a e^{(j/2)\cot\alpha\cdot m^2\Delta u^2}\cdot e^{(j/2)\cot\alpha\cdot n^2\Delta u^2}\cdot e^{-j\frac{\mathrm{sgn}\sin(\alpha)\cdot 2\pi\cdot m\cdot n}{N}} \tag{16}$$

由此实现对重复数据信息流的检测。为了进一步实现对云存储系统中的重复数据删除性能，本文采用 4 阶累积量后置处理方法，对传统方法进行改进，定义云存储系统中的源端节点存储数据的 4 阶累积量切片表达式为：

$$\hat{c}_{4x}(\tau_1,\tau_2,\tau_3)=\frac{1}{N}\sum_{i=1}^{N}x(i)x(i+\tau_1)x(i+\tau_2)x(i+\tau_3)-\hat{c}_{2x}(\tau_1)\hat{c}_{2x}(\tau_2-\tau_3)-\hat{c}_{2x}(\tau_2)\hat{c}_{2x}(\tau_3-\tau_1)-\hat{c}_{2x}(\tau_3)\hat{c}_{2x}(\tau_1-\tau_2) \tag{17}$$

式中，$\hat{c}_{2x}(\tau)=\frac{1}{N}\sum_{i=1}^{N}x(i)x(i+\tau)$，在目标端重删系统中，假设 $\hat{c}_{4x}(\tau_1,\tau_2,\tau_3)$ 的对角切片表达为 $\hat{c}_{4x}(\tau,\tau,\tau)$，那么多个客户端节点的数据对角切片在分数阶 Fourier 域上的离散时间点 n 处的检测统计量为 $\hat{c}_x^{(N)}(n,\tau)$，则有：

$$\hat{c}_x^{(N)}(n,\tau)=\hat{c}_x^{(N)}(\tau,\tau,\tau)=\langle x(n)x^3(n+\tau)\rangle-3\langle x(n)x(n+\tau)\rangle\langle x^2(n+\tau)\rangle \tag{18}$$

式中，$\langle g(n)\rangle$ 表示均值，即：

$$\langle g(n)\rangle=1/N\sum_{n=1}^{N}g(n) \tag{19}$$

上述改进算法有效利用了 4 阶累积量切片对云存储系统重复数据信息流的能量聚集和噪声抑制的特性，在分数阶 Fourier 变换检测形成后置处理，提高对重复数据的滤波性能。以此为基础，对本文提出的分数级 Fourier 变换的云存储系统重删算法进行改进设计。

3.2 云存储系统的重复数据删除算法的改进实现

根据上述设计的基于 Fourier 变换高阶累积量算法的云存储系统重复数据检测滤波结果，利用 4 阶累积量切片后置聚集处理能力，以滤波处理后的云存储系统节点的重复数据信息流为输入向量，进行重删系统设计和算法实现，重复数据的文件系统层设训练样本集为 $X=[X_1,X_2,\cdots,X_k,\cdots,X_N]^{\mathrm{T}}$，其中任一训练样本为 $X_k=[x_{k1},x_{k2},\cdots,x_{kn},\cdots,x_{kM}]$。在云存储的网络环境中，为了确保数据的可用性与可靠性，数据一般都会备份到远端存储节点，采用上述设计的检测滤波系统，得到云存储系统中的重复数据信息流的离散分数阶 Fourier 逆变换可表示为：

$$x=F_{-a}\cdot X \tag{20}$$

其中，$F_{-a}=F_a^H$。若背景噪声 $\omega(n)$ 具有非高斯性，则其 4 阶混合累积表达为：

$$c_{4\omega}(\tau)=\gamma\sum_{j=0}^{\infty}h(j)h^3(j+\tau) \tag{21}$$

式中，γ 为客户端节点的数据带宽，$h(j)$ 为云存储系统的滤波函数，τ 为重复数据重构时延，采用分数阶 Fourier 变换方法，

结合 4 阶累积量后置处理，得到重复数据信息流的删除后的系统输出为：

$$Y_k=[y_{k1},y_{k2},\cdots,y_{kj},\cdots,y_{kJ}],k=1,2,\cdots,N \tag{22}$$

采用分数阶 Fourier 变换高阶累积特征进行存储资源之间的实际分配关系配比，得到期望输出为：

$$d_k=[d_{k1},d_{k2},\cdots,d_{kj},\cdots,d_{kJ}],k=1,2,\cdots,N \tag{23}$$

综上分析，采用 4 阶累积量切片后置算子，通过 k 次分解后，把每个文件分为若干个块，针对每个文件块进行重删，实现存储资源上的重复数据检测及其删除，提高抗干扰能力，减低误删概率，从而提高了集群云存储系统计算资源的利用率，算法实现的流程框图如图 1 所示。

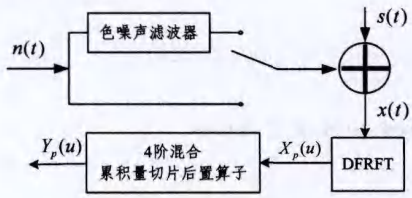


图 1　改进算法设计框图

4　仿真结果及分析

为了测试算法在云存储系统中的重复数据删除性能，进行仿真实验，仿真实验模拟了 1 个客户端和 5 个服务器节点，云存储系统采用 Ubuntu 12.10 进行系统设计，基于 Hadoop 平台进行 C/S 云平台设计。云存储系统中的文件根据内容划分成可变长度的数据块，进行垂直分层，产生 128bit 的信息摘要，通过副本的创建主分量特征建模。仿真实验的硬件环境配置为 CPU：Intel(R) Core(TM) CPU T6600，2.2GHz，双核；内存：2GB，DDR2。云存储系统中包含了大量的云数据，数据格式分别有 DOC、TXT、PPT、VMDK、EXE、PDF 等。在云存储系统中，数据量从 100 MB 到 1 GB，以 100 MB 为单位线性增长。首先采用本文算法进行数据信息流的信号模型构建，得到云存储系统中采集的某一存储数据信息流信号模型如图 2 所示。图 2 中，信息流的采样频率为 500kHz，采样数据为 TXT 数据。

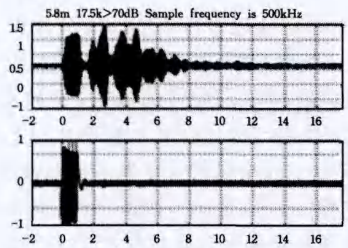


图 2　云存储系统数据信息流信号模型

以上述信息流信号模型为研究对象，作为重删系统的输入参量，进行重复数据删除实验。采用分数阶 Fourier 变换方法，对信息流进行特征检测和数据滤波，以在分数阶 Fourier 域内的归一化投影值为评价指标，得到采用传统的 Fourier 变换得到的重复数据检测结果，如图 3 所示，从图 3 可见，采用传统方法时，由于云存储系统中的干扰数据较多，信噪比较低，色噪声和混响会在分数阶 Fourier 域出现伪峰峰值，因此其不能有效地对重复数据进行检测滤波和删除处理。

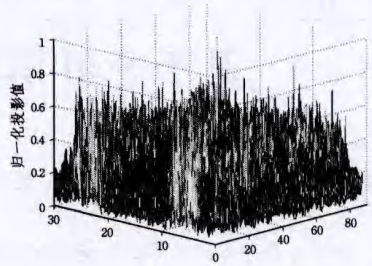


图 3　采用分数阶 Fourier 方法进行云存储系统重复数据检测滤波结果

对此，本文对传统方法进行改进，采用 4 阶累积量切片实现对云存储系统重复数据信息流的能量聚集和噪声抑制，创建多个线程的信息流特征编码，对图 3 所示的输出结果在分数阶 Fourier 域上进行能量聚焦，实现对重复数据的删除，得到的采用本文改进方法处理后的云存储系统重复数据删除的检测输出如图 4 所示。

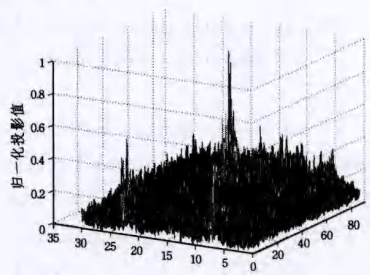
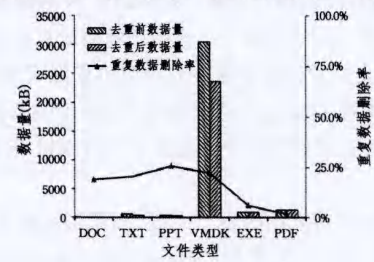
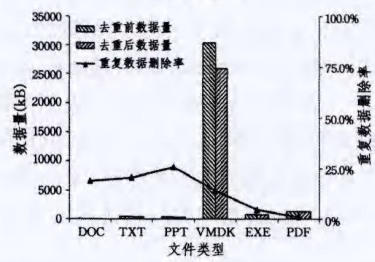


图 4　分数阶 Fourier 累积量后置处理后的重删结果

分析图 4 结果可知，信号在分数阶 Fourier 域可以得到能量聚集而形成冲击函数，通过累积量后置处理，使得伪峰在分数阶 Fourier 域累积，信号幅值大于干扰数据噪声幅值，提高了对重复数据的删除效能。最后，为了定量分析算法性能，进行重复数据删除率对比，得到性能对比结果如图 5 所示。



(a)传统算法



(b)本文算法

图 5　算法应用前后重复数据删除率对比

图 5(a)和图 5(b)中左纵坐标表示不同格式的文件数据

(下转第 209 页)

023115

[12] Sun F Y, Liu S T, Li Z Q, et al. A novel image encryption scheme based on spatial chaos map [J]. Chaos, Solitons & Fractals, 2008, 38(3): 631-640

[13] Rhouma R, Soumaya M, Safya B. OCML-based colour image encryption [J]. Chaos, Solitons & Fractals, 2009, 40(1): 309-318

[14] Guo Q, Liu Z G, Liu S T. Colour image encryption by using Arnold and discrete fractional random transforms in HIS space [J]. Optics and Lasers in Engineering, 2010, 48(12): 1174-1181

[15] Sahar M, Amir M E. Colour image encryption based on coupled nonlinear chaotic map [J]. Chaos, Solitons and Fractals, 2009, 42(3): 1745-1754

[16] Wang X Y, Teng L, Qin X. A novel colour image encryption algorithm based on chaos [J]. Signal Processing, 2012, 92(4): 1101-1108

[17] Zhang W, Wong K W, Yu H, et al. A symmetric color image encryption algorithm using the intrinsic features of bit distributions [J]. Communications in Nonlinear Science and Numerical Simulation, 2013, 18(3): 584-600

[18] Fu C, Lin B, Miao Y, et al. A novel chaos-based bit-level permutation scheme for digital image encryption [J]. Optics Communication, 2011, 284(23): 5415-5423

[19] Liu H J, Wang X Y. Color image encryption using spatial bit-level permutation and high-dimension chaotic system [J]. Optics Communications, 2011, 284(16/17): 3895-3903

[20] Kaneko K. Spatiotemporal intermittency in Coupled Map Lattices[J]. Progress of Theoretical Physics, 1985, 74(5): 1033-1044

[21] Rhouma R, Soumaya M, Safya B. OCML-based colour image encryption [J]. Chaos, Solitons & Fractals, 2009, 40(1): 309-318

[22] Sahar M, Amir M E. Colour image encryption based on coupled nonlinear chaotic map [J]. Chaos, Solitons & Fractals, 2009, 42(3): 1745-1754

[23] Liu H J, Wang X Y. Colour image encryption based on one-time keys and robust chaotic maps [J]. Computers&Mathematics with Applications, 2010, 59(10): 3320-3327

[24] 罗松江，丘水生. 基于时空混沌和S盒的彩色图像加密算法[J]. 电路与系统学报，2010，15(3)：117-122
Luo S J, Qiu S S. Color image encryption algorithm based on spatiotemporal chaos and S-box [J]. Journal of Circuits and Systems, 2010, 15(3): 117-122

[25] He J, Li Z B, Qian H F. Cryptography based on spatiotemporal chaos system and multiple maps [J]. Journal of Software, 2010, 5(4): 421-428

(上接第 177 页)

量大小，横坐标表示6种不同文件类型，右纵坐标表示重复数据删除率的大小。对6种不同格式的文件采用本文算法和传统算法进行云存储系统重复数据删除，结果表明本文算法重复数据准确删除率较高，去重的效果更佳，有效避免了数据信息流的干扰特征造成的误删和漏删，重复数据删除准确性能较好，误删率降低了13.11%，云存储系统的CPU执行时间提高了17.8%，从而展示了算法的优越性能。

结束语 云存储系统中的重复数据是各类数据管理成本快速上升过程中留下的冗余数据产物，云存储系统中产生的数据量以几何级数增长。为了有效面对爆炸式增长的云存储系统运行数据管理的需求，减轻服务器开销，研究了一种有效的云存储系统重复数据删除算法，对消除数据冗余、降低系统能耗和提高存储性能具有重要意义[9]。本文提出一种改进的基于分数阶Fourier变换累积量检测的云存储系统重复数据删除算法，即采用4阶累积量切片实现对云存储系统重复数据信息流的能量聚集和噪声抑制，进行重复数据检测后置滤波处理，创建多个线程的信息流特征编码，实现对重复数据的删除。分析研究和实验结果表明，采用本文算法能有效避免数据信息流的干扰特征造成的误删和漏删，对云存储系统中重复数据的检测性能较好，重复数据删除准确性高，综合性能优于传统算法。

参 考 文 献

[1] 谢平. 存储系统重复数据删除技术研究综述[J]. 计算机科学，2014，41(1)：22-30
Xie Ping. Surey on data deduplication techniques for storage systems[J]. Computer Science, 2014, 41(1): 22-30

[2] Miorandi D, Sicari S, Pellegrini F D, et al. Internet of things: vision, applications and research challenges [J]. Ad Hoc Networks, 2012, 10(7): 1497-1516

[3] Wu T Y, Lee W T, Lin Y S, et al. Dynamic load balancing mechanism based on cloud storage[C]// Computing, Communications and Applications Conference (ComComAp), 2012. IEEE, 2012: 102-106

[4] 蒋海波，王晓京，范明钰，等. 基于水平纠删码的云存储数据布局方法[J]. 四川大学学报(工程科学版)，2013，45(2)：103-109
Jiang Hai-bo, Wang Xiao-jing, Fan Ming-yu. A Data Placement Based on Level Array Codes in Cloud Storage [J]. Journal of Sichuan University (Engineeging Science Edition), 2013, 45(2): 103-109

[5] 敖莉，舒继武，李明强. 重复数据删除技术[J]. 软件学报，2010，21(5)：916-929
Ao Li, Shu Ji-wu, Li Ming-qiang. Data Deduplication Techniques [J]. Journal of Software, 2010, 21(5): 916-929

[6] 付印金，肖侬，刘芳. 重复数据删除关键技术研究进展[J]. 计算机研究与发展，2012，49(1)：12-20
Fu Ying-jin, Xiao Nong, Liu Fang. Research and Development on Key Techniques of Data Deduplicaton[J]. Journal of Computer Research and Development, 2012, 49(1): 12-20

[7] 李渊. 智能PID控制区优化仿真研究[J]. 计算机仿真，2012，29(12)：180-182
Li Yuan. Parameters Optimization of PID Controller[J]. Computer Simulation, 2012, 29(12): 180-182

[8] 谭鹏许，陈越，兰巨龙，等. 用于云存储的安全容错编码[J]. 通信学报，2014，35(3)：109-114
Tan Peng-xu, Chen Yue, Lan Ju-long, et al. Secure fault-tolerant code for cloud storage[J]. Journal on Communications, 2014, 35(3): 109-114

[9] Tang Pei-he, Xu Yi-yi. Resource Scheduling Strategy Based on Credibility in the Enterprise Gloud Strorage[J]. Journal of Convergence Information Technology, 2012, 7(16): 393-400