(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114138552 B
(45)授权公告日 2024.01.12

(21)申请号 202111335541.6

(22)申请日 2021.11.11

(65)同一申请的已公布的文献号
申请公布号 CN 114138552 A

(43)申请公布日 2022.03.04

(73)专利权人 苏州浪潮智能科技有限公司
地址 215100 江苏省苏州市吴中经济开发区郭巷街道官浦路1号9幢

(72)发明人 朱箫鸣 冀国威

(74)专利代理机构 济南舜源专利事务所有限公司 37205
专利代理师 辛向东

(51)Int.Cl.
G06F 11/14(2006.01)

(56)对比文件
CN 109948740 A,2019.06.28
CN 109299727 A,2019.02.01
CN 102323958 A,2012.01.18

审查员 董莉

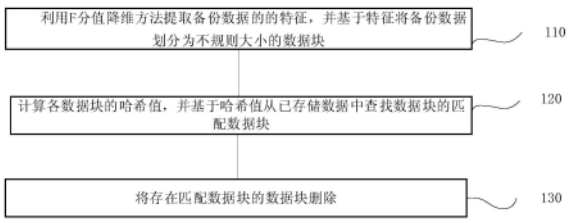
权利要求书2页 说明书8页 附图1页

(54)发明名称

数据动态重删方法、系统、终端及存储介质

(57)摘要

本发明提供一种数据动态重删方法、系统、终端及存储介质，包括：利用F分值降维方法提取备份数据的的特征，并基于特征将备份数据划分为不规则大小的数据块；计算各数据块的哈希值，并基于哈希值从已存储数据中查找数据块的匹配数据块；将存在匹配数据块的数据块删除。本发明解决了固定块分割中存在的问题，大大降低了数据分割处理上对客户端计算资源的占用，能够有效的提升海量小文件备份过程中数据重删率，降低计算资源的占用，同时在数据存储时能够更加节省空间。



1.一种数据动态重删方法,其特征在于,包括:

利用F分值降维方法提取备份数据的特征,并基于特征将备份数据划分为不规则大小的数据块;

计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

将存在匹配数据块的数据块删除;

利用F分值降维方法提取备份数据的特征,并基于特征将备份数据划分为不规则大小的数据块,包括:

利用F分值函数:

$$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

计算备份数据的K个特征的F分值,其中$\bar{x}_i$为第i个特征在整个数据集的平均值,$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值,$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值,$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值,$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值;n_+为正类点的个数,n_-为负类点的个数,n为F分值的个数;

对各特征值F分值按照由大到小的规则进行排序,并根据设定的特征值数量从中选取排名靠前的相应数量的F分值;

将选取的F分值对应的特征值作为备份数据的特征值;

利用F分值降维方法提取备份数据的特征,并基于特征将备份数据划分为不规则大小的数据块,包括:

从备份数据的所有特征值中随机选取一个目标特征值,并将目标特征值作为分割起点;

将所述目标特征值代入分割长度计算函数$S_n=\sum_{i=1}^{n}a_iX_i$,得到分割长度S_n,其中X_i为随机变量序列,a_i是由备份数据的数据结构确定的系数,n为特征值的数量;

切换目标特征值直至遍历所有特征值,得到各特征值对应的分割长度;

根据备份数据的特征值和各特征值对应的分割长度,将备份数据划分为多个数据块。

2.根据权利要求1所述的方法,其特征在于,计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配块,包括:

计算数据块的哈希值,从已存储数据中检索与所述数据块具有相同哈希值的匹配数据块;

获取所述匹配数据块的元数据,将所述元数据作为所述数据块的元数据。

3.一种数据动态重删系统,其特征在于,包括:

数据分割单元,用于利用F分值降维方法提取备份数据的特征,并基于特征将备份数据划分为不规则大小的数据块;

匹配查找单元,用于计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

重复删除单元,用于将存在匹配数据块的数据块删除;

所述数据分割单元用于：

利用F分值函数：

$$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

计算备份数据的K个特征的F分值，其中$\bar{x}_i$为第i个特征在整个数据集的平均值，$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值，$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值，$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值，$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值；n_+为正类点的个数，n_-为负类点的个数，n为F分值的个数；

对各特征值F分值按照由大到小的规则进行排序，并根据设定的特征值数量从中选取排名靠前的相应数量的F分值；

将选取的F分值对应的特征值作为备份数据的特征值；

所述数据分割单元用于：

从备份数据的所有特征值中随机选取一个目标特征值，并将目标特征值作为分割起点；

将所述目标特征值代入分割长度计算函数$S_n=\sum_{i=1}^{n}a_iX_i$，得到分割长度S_n，其中X_i为随机变量序列，a_i是由备份数据的数据结构确定的系数，n为特征值的数量；

切换目标特征值直至遍历所有特征值，得到各特征值对应的分割长度；

根据备份数据的特征值和各特征值对应的分割长度，将备份数据划分为多个数据块。

4.根据权利要求3所述的系统，其特征在于，所述匹配查找单元用于：

计算数据块的哈希值，从已存储数据中检索与所述数据块具有相同哈希值的匹配数据块；

获取所述匹配数据块的元数据，将所述元数据作为所述数据块的元数据。

5.一种终端，其特征在于，包括：

处理器；

用于存储处理器的执行指令的存储器；

其中，所述处理器被配置为执行权利要求1-2任一项所述的方法。

6.一种存储有计算机程序的计算机可读存储介质，其特征在于，该程序被处理器执行时实现如权利要求1-2中任一项所述的方法。

数据动态重删方法、系统、终端及存储介质

技术领域

[0001] 本发明涉及数据存储技术领域,具体涉及一种数据动态重删方法、系统、终端及存储介质。

背景技术

[0002] 随着科学技术的发展,数据开始指数级增长,数据安全也成为政企关注的重点,但在数据的备份保护中,总是充斥着大量冗余数据占用存储空间,为了解决这个问题,人们开始关注“重复数据删除”技术,希望能节约出大量的存储空间。所以,在数据的备份容灾产品中,“重复数据删除”技术也就成了考量产品在技术含量、运行性能、产品质量等方面是否优越的考核指标之一。

[0003] 在数据重删的实现上,厂商一般采取如下方法进行,首先对数据进行分块处理,即把备份数据分割成互不重叠的定长数据块,常用的块大小有4K/8K/16K/32K/128K等,不同的厂商选择的定长数据块大小不一。然后利用哈希算法,为每个数据块建立指纹信息,系统通过计算并检查数据块的“指纹”,判断该数据块是否与已经存在的“元数据”重复:如果重复,则只需保留指向该“元数据”的指针;如果“指纹”显示该数据块是全新的,则保留该数据块,并提取相关信息作为“元数据”保存,以供后续数据检验对比使用。

[0004] 在整个过程中不难发现,切分数据块的大小成为至关重要的问题,数据块的大小将会影响到数据去重处理的运算性能和重删率:数据分块大,数据去重处理运算性能高,但重删率低,精准度下降;数据分块小,数据去重处理运算性能低,但重删率高,精准度提升。同时,对于海量小文件场景,由于数据变化情况复杂,当向源数据对象插入或删除数据时,由于采用了定长的数据块切分,会导致重新进行数据块切分,在增加计算量的同时,会导致重删率更低。

发明内容

[0005] 针对现有技术的上述不足,本发明提供一种数据动态重删方法、系统、终端及存储介质,以解决上述技术问题。

[0006] 第一方面,本发明提供一种数据动态重删方法,包括:

[0007] 利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块;

[0008] 计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

[0009] 将存在匹配数据块的数据块删除。

[0010] 进一步的,利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块,包括:

[0011] 利用F分值函数:

[0012] $$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

[0013] 计算备份数据的K个特征的F分值,其中$\bar{x}_i$为第i个特征在整个数据集的平均值,$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值,$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值,$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值,$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值;

[0014] 对各特征值F分值按照由大到小的规则进行排序,并根据设定的特征值数量从中选取排名靠前的相应数量的F分值;

[0015] 将选取的F分值对应的特征值作为备份数据的特征值。

[0016] 进一步的,利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块,包括:

[0017] 从备份数据的所有特征值中随机选取一个目标特征值,并将目标特征值作为分割起点;

[0018] 将所述目标特征值代入分割长度计算函数$S_n=\sum_{i=1}^{n}a_iX_i$,得到分割长度S_n,其中X_i为随机变量序列,a_i是由备份数据的数据结构确定的系数,n为特征值的数量;

[0019] 切换目标特征值直至遍历所有特征值,得到各特征值对应的分割长度;

[0020] 根据备份数据的特征值和各特征值对应的分割长度,将备份数据划分为多个数据块。

[0021] 进一步的,计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配块,包括:

[0022] 计算数据块的哈希值,从已存储数据中检索与所述数据块具有相同哈希值的匹配数据块;

[0023] 获取所述匹配数据块的元数据,将所述元数据作为所述数据块的元数据。

[0024] 第二方面,本发明提供一种数据动态重删系统,包括:

[0025] 数据分割单元,用于利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块;

[0026] 匹配查找单元,用于计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

[0027] 重复删除单元,用于将存在匹配数据块的数据块删除。

[0028] 进一步的,所述数据分割单元用于:

[0029] 利用F分值函数:

[0030] $$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

[0031] 计算备份数据的K个特征的F分值,其中$\bar{x}_i$为第i个特征在整个数据集的平均值,$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值,$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值,

$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值，$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值；

[0032] 对各特征值F分值按照由大到小的规则进行排序，并根据设定的特征值数量从中选取排名靠前的相应数量的F分值；

[0033] 将选取的F分值对应的特征值作为备份数据的特征值。

[0034] 进一步的，所述数据分割单元用于：

[0035] 从备份数据的所有特征值中随机选取一个目标特征值，并将目标特征值作为分割起点；

[0036] 将所述目标特征值代入分割长度计算函数$S_n = \sum_{i=1}^{n} a_i X_i$，得到分割长度$S_n$，其中$X_i$为随机变量序列，$a_i$是由备份数据的数据结构确定的系数，n为特征值的数量；

[0037] 切换目标特征值直至遍历所有特征值，得到各特征值对应的分割长度；

[0038] 根据备份数据的特征值和各特征值对应的分割长度，将备份数据划分为多个数据块。

[0039] 进一步的，所述匹配查找单元用于：

[0040] 计算数据块的哈希值，从已存储数据中检索与所述数据块具有相同哈希值的匹配数据块；

[0041] 获取所述匹配数据块的元数据，将所述元数据作为所述数据块的元数据。

[0042] 第三方面，提供一种终端，包括：

[0043] 处理器、存储器，其中，

[0044] 该存储器用于存储计算机程序，

[0045] 该处理器用于从存储器中调用并运行该计算机程序，使得终端执行上述的终端的方法。

[0046] 第四方面，提供了一种计算机存储介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行上述各方面所述的方法。

[0047] 本发明的有益效果在于，本发明提供的数据动态重删方法、系统、终端及存储介质，利用F分值降维方法提取备份数据的的特征并基于特征将备份数据划分为不规则大小的数据块，实现对备份数据的动态分割，解决了固定块分割中存在的问题，只针对变化部分的数据做分割处理，这样就大大降低了数据分割处理上对客户端计算资源的占用，能够有效的提升海量小文件备份过程中数据重删率，降低计算资源的占用，同时在数据存储时能够更加节省空间。

[0048] 此外，本发明设计原理可靠，结构简单，具有非常广泛的应用前景。

附图说明

[0049] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0050] 图1是本发明一个实施例的方法的示意性流程图。

[0051] 图2是本发明一个实施例的系统的示意性框图。

[0052] 图3为本发明实施例提供的一种终端的结构示意图。

具体实施方式

[0053] 为了使本技术领域的人员更好地理解本发明中的技术方案,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本发明保护的范围。

[0054] 现有技术也存在滑动窗口分块方案,基于滑动窗口分块方案的首次备份与定长重删的方法一致,它选用固定的长度对整串数据进行分块,并计算各个分块的hash值。选用的这个固定的长度就是窗口的长度,二次备份的时候,利用窗口的滑动来尝试寻找和匹配相同的数据。以数据修改为例,第二个切片发生了数据变化deab-->ddab。首先计算ddab的hash值,该切片的数据是发生了变化的因此无指纹可以匹配。这个时候不着急处理下一个数据切片,而是将窗口向前移动一个单位,继续计算这个窗口下的数据的hash值(f ingerpr int2')并尝试匹配,以此类推,直到找到可以匹配的hash值为止。当某个数据发生覆盖写的时候,其效果与定长重删效果一样,可以获得一样的重删率。这种方法仍需在初始时将数据分为定长的数据块,不适用于小文件的分割;且后续要根据指定的步距滑动窗口,然后进行数据重删,这就需要多次计算数据块的哈希值,计算量多大,导致重删效率低下。

[0055] 在实际业务生产中,由于业务系统的种类不同,数据的大小不一,如结构化数据一般为KB级,非结构化数据为MB级,采用定长数据块分割会存在重删率低的问题,为了解决该问题,本发明公开了一种基于变长数据块切分技术的数据重删方法。该方法主要通过一个不断滑动的窗口来确定数据的分界限,按照其特征函数把备份数据动态分割成不同大小的数据块,从而提升数据重删率。

[0056] 图1是本发明一个实施例的方法的示意性流程图。其中,图1执行主体可以为一种数据动态重删系统。

[0057] 如图1所示,该方法包括:

[0058] 步骤110,利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块;

[0059] 步骤120,计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

[0060] 步骤130,将存在匹配数据块的数据块删除。

[0061] 为了便于对本发明的理解,下面以本发明数据动态重删方法的原理,结合实施例中对数据进行动态重删的过程,对本发明提供的数据动态重删方法做进一步的描述。

[0062] 具体的,所述数据动态重删方法包括:

[0063] S 1、利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块。

[0064] 对于给定的备份数据组成的数据集,假定他的正类点的个数和负类点的个数分别是n+和n_:

[0065] 利用F分值函数:

[0066] $$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

[0067] 计算备份数据的K个特征的F分值,其中$\bar{x}_i$为第i个特征在整个数据集的平均值,$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值,$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值,$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值,$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值。

[0068] 该公式的右侧分子大致反映除了正类点和负类点在第K个特征上差异程度的大小,而分母左面的式子和右面的式子则分别反映出了在第K个特征上的正类点和负类点各自的分散程度,所以,如果F(k)的值越大的话,那么对于第K个特征,就更加能够区分这两类点,这样F-分值这个方法就可以作为选择特征的一个标准。

[0069] 假设预先设定选取的特征数量为d,利用上述公式计算n个F分值,分别为:F(1),F(2),…,F(n);将n个F分值按照由大到小的规则进行排序,并从中选取排名靠前的前d个F分值;取出这些选取出的F分值对应的下标k_i,这些下标对应的特征即为备份数据的特征值。

[0070] 从备份数据的所有特征值中随机选取一个目标特征值,并将目标特征值作为分割起点;

[0071] 将目标特征值代入分割长度计算函数$S_n=\sum_{i=1}^{n}a_iX_i$,得到分割长度S_n,其中X_i为表示未知数的随机变量序列,n为特征值的数量;a_i是由备份数据的数据结构确定的系数,如果备份数据为结构数据则a_i为常数,如果备份数据为非结构数据则a_i为一列系数数组。

[0072] 切换目标特征值直至遍历所有特征值,得到各特征值对应的分割长度;根据备份数据的特征值和各特征值对应的分割长度,将备份数据划分为多个数据块。

[0073] 动态的对数据进行不规则的数据块切分,在海量小文件场景下,数据一般为1~4KB大小不等的文件,基于变长数据块的原理,在备份过程中,依据文件的大小不同划分为不同大小的数据块,比如原文件为1KB,基于变长数据块原理就将1KB作为一个数据块进行哈希算法得出相应的“指纹信息”,如原文件为3KB,则将其划分为两个2KB的数据块,进行哈希算法得出两个相应的“指纹信息”。

[0074] S2、计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块。

[0075] 计算数据块的哈希值,从已存储数据中检索与数据块具有相同哈希值的匹配数据块;获取匹配数据块的元数据,将获取的元数据作为数据块的元数据。

[0076] S3、将存在匹配数据块的数据块删除。

[0077] 步骤S2中已经将匹配数据块的元数据绑定为数据块的元数据,因此将数据块删除已节省存储资源,在查找数据块时根据元数据即可读出与数据块的内容。

[0078] 本实施例有效地解决了固定块分割中存在的问题,当向数据对象中插入数据或者从中删除数据时,如果变化的内容不在数据块的边界内,数据块不发生改变;数据块之间的边界是随机的、动态的,且数据块的部分内容可能是重复的。因此,插入或者删除内容只影响相邻的一个或者两个数据块,其余数据块不会受影响,这样使得数据的去重更为精准。

[0079] 对于非结构化数据,特别是海量小文件的备份,采用了动态的分割技术。因为海量

小文件中,数据变化情况复杂,采用固定块分割,常导致重新对整个备份数据进行重新分块,而采用变长块的分割处理,解决了固定块分割中存在的问题,只针对变化部分的数据做分割处理,这样就大大降低了数据分割处理上对客户端计算资源的占用,并且保障了海量小文件的最佳数据重删效果。

[0080] 如图2所示,该系统200包括:

[0081] 数据分割单元210,用于利用F分值降维方法提取备份数据的的特征,并基于特征将备份数据划分为不规则大小的数据块;

[0082] 匹配查找单元220,用于计算各数据块的哈希值,并基于哈希值从已存储数据中查找数据块的匹配数据块;

[0083] 重复删除单元230,用于将存在匹配数据块的数据块删除。

[0084] 可选地,作为本发明一个实施例,所述数据分割单元用于:

[0085] 利用F分值函数:

[0086] $$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n-1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2}$$

[0087] 计算备份数据的K个特征的F分值,其中$\bar{x}_i$为第i个特征在整个数据集的平均值,$\bar{x}_i^{(+)}$为第i个特征在正类数据集上的平均值,$\bar{x}_i^{(-)}$为第i个特征在负类数据集上的平均值,$x_{k,i}^{(+)}$为第k个正类样本点的第i个特征的特征值,$x_{k,i}^{(-)}$为第k个负类样本点的第i个特征的特征值;

[0088] 对各特征值F分值按照由大到小的规则进行排序,并根据设定的特征值数量从中选取排名靠前的相应数量的F分值;

[0089] 将选取的F分值对应的特征值作为备份数据的特征值。

[0090] 可选地,作为本发明一个实施例,所述数据分割单元用于:

[0091] 从备份数据的所有特征值中随机选取一个目标特征值,并将目标特征值作为分割起点;

[0092] 将所述目标特征值代入分割长度计算函数$S_n=\sum_{i=1}^{n}a_iX_i$,得到分割长度S_n,其中X_i为随机变量序列,a_i是由备份数据的数据结构确定的系数,n为特征值的数量;

[0093] 切换目标特征值直至遍历所有特征值,得到各特征值对应的分割长度;

[0094] 根据备份数据的特征值和各特征值对应的分割长度,将备份数据划分为多个数据块。

[0095] 可选地,作为本发明一个实施例,所述匹配查找单元用于:

[0096] 计算数据块的哈希值,从已存储数据中检索与所述数据块具有相同哈希值的匹配数据块;

[0097] 获取所述匹配数据块的元数据,将所述元数据作为所述数据块的元数据。

[0098] 图3为本发明实施例提供的一种终端300的结构示意图,该终端300可以用于执行本发明实施例提供的数据动态重删方法。

[0099] 其中,该终端300可以包括:处理器310、存储器320及通信单元330。这些组件通过一条或多条总线进行通信,本领域技术人员可以理解,图中示出的服务器的结构并不构成

对本发明的限定,它既可以是总线形结构,也可以是星型结构,还可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0100] 其中,该存储器320可以用于存储处理器310的执行指令,存储器320可以由任何类型的易失性或非易失性存储终端或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。当存储器320中的执行指令由处理器310执行时,使得终端300能够执行以下上述方法实施例中的部分或全部步骤。

[0101] 处理器310为存储终端的控制中心,利用各种接口和线路连接整个电子终端的各个部分,通过运行或执行存储在存储器320内的软件程序和/或模块,以及调用存储在存储器内的数据,以执行电子终端的各种功能和/或处理数据。所述处理器可以由集成电路(Integrated Circuit,简称IC)组成,例如可以由单颗封装的IC所组成,也可以由连接多颗相同功能或不同功能的封装IC而组成。举例来说,处理器310可以仅包括中央处理器(Central Processing Unit,简称CPU)。在本发明实施方式中,CPU可以是单运算核心,也可以包括多运算核心。

[0102] 通信单元330,用于建立通信信道,从而使所述存储终端可以与其它终端进行通信。接收其他终端发送的用户数据或者向其他终端发送用户数据。

[0103] 本发明还提供一种计算机存储介质,其中,该计算机存储介质可存储有程序,该程序执行时可包括本发明提供的各实施例中的部分或全部步骤。所述的存储介质可为磁碟、光盘、只读存储记忆体(英文:read-only memory,简称:ROM)或随机存储记忆体(英文:random access memory,简称:RAM)等。

[0104] 因此,本发明利用F分值降维方法提取备份数据的的特征并基于特征将备份数据划分为不规则大小的数据块,实现对备份数据的动态分割,解决了固定块分割中存在的问题,只针对变化部分的数据做分割处理,这样就大大降低了数据分割处理上对客户端计算资源的占用,能够有效的提升海量小文件备份过程中数据重删率,降低计算资源的占用,同时在数据存储时能够更加节省空间,本实施例所能达到的技术效果可以参见上文中的描述,此处不再赘述。

[0105] 本领域的技术人员可以清楚地了解到本发明实施例中的技术可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解,本发明实施例中的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中如U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质,包括若干指令用以使得一台计算机终端(可以是个人计算机,服务器,或者第二终端、网络终端等)执行本发明各个实施例所述方法的全部或部分步骤。

[0106] 本说明书中各个实施例之间相同相似的部分互相参见即可。尤其,对于终端实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例中的说明即可。

[0107] 在本发明所提供的几个实施例中,应该理解到,所揭露的系统和方法,可以通过其它的方式实现。例如,以上所描述的系统实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结

合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,系统或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0108] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0109] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0110] 尽管通过参考附图并结合优选实施例的方式对本发明进行了详细描述,但本发明并不限于此。在不脱离本发明的精神和实质的前提下,本领域普通技术人员可以对本发明的实施例进行各种等效的修改或替换,而这些修改或替换都应在本发明的涵盖范围内/任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

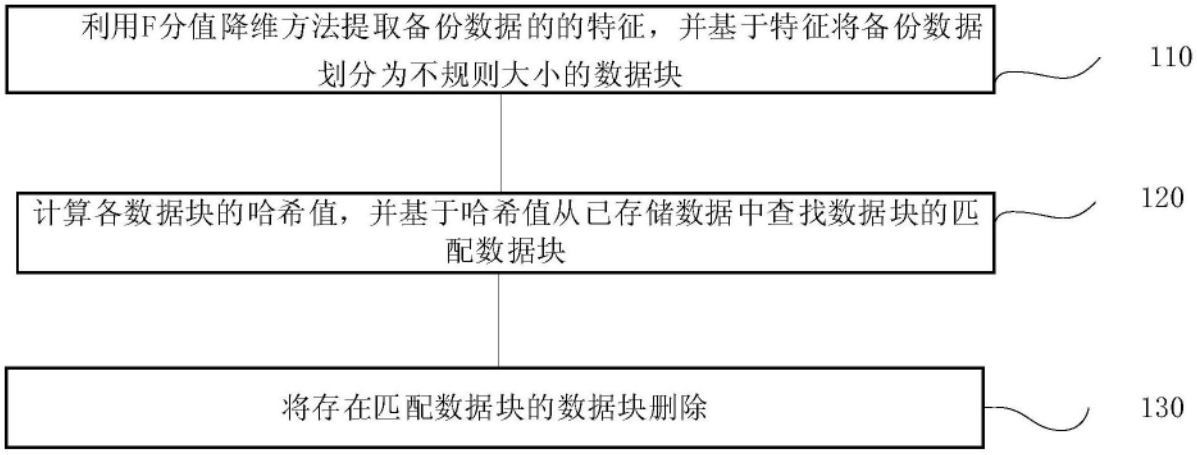

利用F分值降维方法提取备份数据的的特征，并基于特征将备份数据划分为不规则大小的数据块
110
计算各数据块的哈希值，并基于哈希值从已存储数据中查找数据块的匹配数据块
120
将存在匹配数据块的数据块删除
130

图1


系统 200
数据分割单元 210
匹配查找单元 220
重复删除单元 230

图2

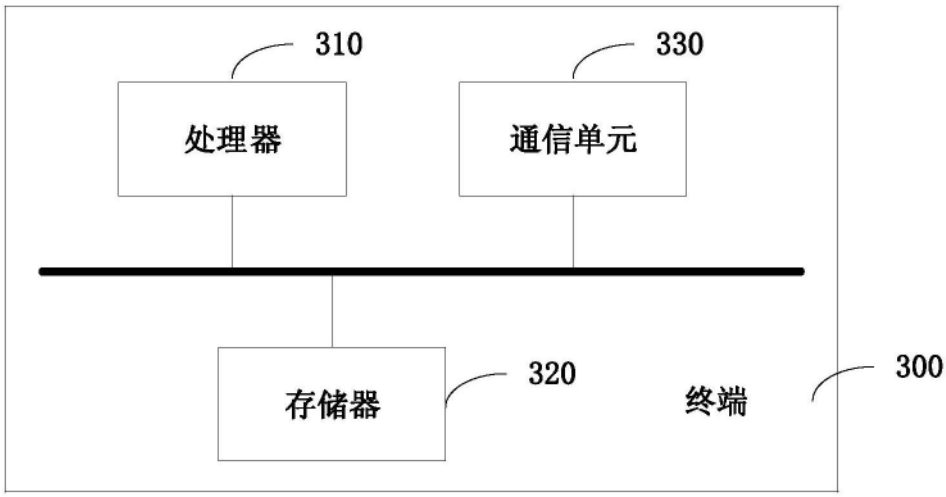

310
330
处理器
通信单元
存储器
320
终端
300

图3