(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111949652 A
(43)申请公布日 2020.11.17

(21)申请号 202010576124.X

(22)申请日 2020.06.22

(71)申请人 联想(北京)有限公司
地址 100085 北京市海淀区上地西路6号2幢2层201-H2-6

(72)发明人 娄婷 段净化

(74)专利代理机构 北京乐知新创知识产权代理事务所(普通合伙) 11734
代理人 周伟

(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/242(2019.01)
G06F 16/2455(2019.01)

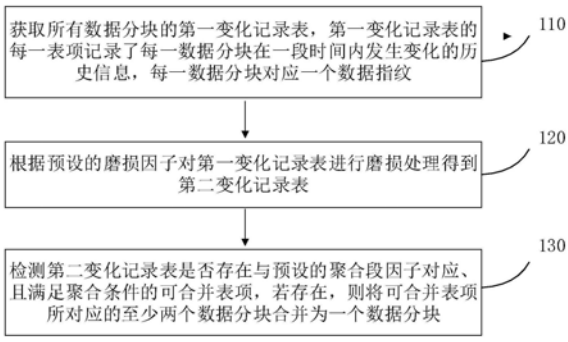
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种数据指纹检测方法、装置及存储介质

(57)摘要

本发明公开了一种数据指纹检测方法、装置及存储介质。该方法采用变化记录表来记录每一数据分块在一段时间内发生变化的历史信息,在符合某一条件时,根据预设的磨损因子对变化记录表进行磨损处理;检测磨损处理后的变化记录表是否存在与预设的聚合段因子对应且满足聚合条件的可合并表项,若存在,则将可合并表项所对应的至少两个数据分块合并为一个数据分块。如此,可以根据每一数据变换的次数或频率等信息,动态的合并不经常变化的数据分块,减少了数据分块的数量,也相应地减少了要处理的数据指纹的数量,从而进一步降低了资源的消耗,并大大提高了系统的处理能力和吞吐量。



CN 111949652 A

1.一种数据指纹检测方法,所述方法包括在符合第一条件时,执行以下操作:

获取所有数据分块的第一变化记录表,所述第一变化记录表的每一表项记录了每一数据分块在一段时间内发生变化的历史信息,所述每一数据分块对应一个数据指纹;

根据预设的磨损因子对所述第一变化记录表进行磨损处理得到第二变化记录表;

检测所述第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项,若存在,则将所述可合并表项所对应的至少两个数据分块合并为一个数据分块。

2.根据权利要求1所述的方法,在所述符合第一条件之前,所述方法还包括:

获取待差分处理的文件;

对所述文件进行分块得到数据分块;

创建所述第一变化记录表,其中所述第一变换记录表的每一表项对应于每一数据分块,用于记录每一数据分块在一段时间内发生变化的历史信息。

3.根据权利要求2所述的方法,在所述创建所述第一变化记录表之后,所述方法还包括在对所述文件进行差分处理时,执行以下操作:

获取所有数据分块的第一数据指纹,所述第一数据指纹是与每一数据分块对应的最新的数据指纹;

获取所有数据分块的第二数据指纹,所述第二数据指纹是与每一数据分块对应的上一次存储的数据指纹;

检测每一数据分块的第一数据指纹和第二数据指纹是否相同,若不同,则更新所述第一变化记录表中相应数据分块所对应的表项。

4.根据权利要求3所述的方法,所述更新所述第一变化记录表中相应数据分块所对应的表项,包括:

获取相应数据分块所对应的所有表项;

依次更新所有表项中的每一表项。

5.根据权利要求1至4任一项所述的方法,所述第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的历史信息,包括:

第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的次数;

相应地,所述更新数据分块所对应的表项,包括:

使所述表项所记录的次数加一。

6.根据权利要求1所述的方法,所述第一条件包括到达预定时间。

7.根据权利要求1所述的方法,所述第一变化记录表包括使用位图bitmap作为存储结构的变化记录表。

8.根据权利要求1所述的方法,在所述根据预设的磨损因子对所述第一变化记录表进行磨损处理得到第二变化记录表之前,所述方法还包括:

确定所述磨损因子和所述聚合段因子。

9.一种数据指纹检测装置,所述装置包括:

第一条件检测模块,用于检测是否符合第一条件;

变化记录表获取模块,用于获取所有数据分块的第一变化记录表;

磨损模块,用于根据预设的磨损因子对所述第一变化记录表进行磨损处理得到第二变化记录表;

聚合段因子检测模块，用于检测所述第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项；

聚合模块，用于将所述可合并表项所对应的至少两个数据分块合并为一个数据分块。

10.一种存储介质，在所述存储介质上存储了程序指令，其中，所述程序指令在运行时用于执行如权利要求1至8任一项所述的数据指纹检测方法。

一种数据指纹检测方法、装置及存储介质

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种数据指纹检测方法、装置及存储介质。

背景技术

[0002] 随着数据处理技术和网络传输能力的不断发展，大文件甚至是超大文件的存储和传输变得越来越普遍。对于大文件特别是超大文件来说，通过比较数据指纹来进行差分备份和传输就显得尤为重要，不仅可以大大减少传输带宽，还可以提高文件存储系统的处理能力

[0003] 在现有方案中，无论文件大小，通常采用固定大小对大文件进行分块，并为每个数据分块建立数据指纹。

[0004] 本发明人发现，对于某些大文件来说，虽然该文件很大，但经常变化的部分却很集中。此时，如果仍采用固定大小对大文件进行分块，则在每次备份或传输时，即使绝大部分的数据分块是不变的，但仍需逐一比较每一数据分块的数据指纹，不仅要花费较长的时间，也会消耗大量的计算机资源。

[0005] 由此可见，如何改进上述大文件的分块策略，提高指纹数据的处理效率是一个尚待解决的技术问题。

发明内容

[0006] 针对以上问题，本发明人创造性地想到：在这种情况下，如果能精确定位经常变化的数据，将不经常变化的数据分块进行合并，就可以大大减少分块的数量，从而缩短数据指纹对比的时间，节约对比数据指纹需要的资源。

[0007] 基于以上发明思路，本发明人提供了一种数据指纹检测方法、装置及存储介质。

[0008] 根据本发明实施例第一方面，一种数据指纹检测方法，该方法包括在符合第一条件时，执行以下操作：获取所有数据分块的第一变化记录表，第一变化记录表的每一表项记录了每一数据分块在一段时间内发生变化的历史信息，每一数据分块对应一个数据指纹；根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表；检测第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项，若存在，则将可合并表项所对应的至少两个数据分块合并为一个数据分块。

[0009] 根据本发明实施例一实施方式，在符合第一条件之前，该方法还包括：获取待差分处理的文件；对文件进行分块得到数据分块；创建第一变化记录表，其中第一变换记录表的每一表项对应于每一数据分块，用于记录每一数据分块在一段时间内发生变化的历史信息。

[0010] 根据本发明实施例一实施方式，在创建第一变化记录表之后，该方法还包括在对文件进行差分处理时，执行以下操作：获取所有数据分块的第一数据指纹，第一数据指纹是与每一数据分块对应的最新的数据指纹；获取所有数据分块的第二数据指纹，第二数据指纹是与每一数据分块对应的上一次存储的数据指纹；检测每一数据分块的第一数据指纹和

第二数据指纹是否相同，若不同，则更新第一变化记录表中相应数据分块所对应的表项。

[0011] 根据本发明实施例一实施方式，更新第一变化记录表中相应数据分块所对应的表项，包括：获取相应数据分块所对应的所有表项；依次更新所有表项中的每一表项。

[0012] 根据本发明实施例一实施方式，第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的历史信息，包括：第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的次数；相应地，更新数据分块所对应的表项，包括：使表项所记录的次数加一。

[0013] 根据本发明实施例一实施方式，第一条件包括到达预定时间。

[0014] 根据本发明实施例一实施方式，第一变化记录表包括使用位图bitmap作为存储结构的变化记录表。

[0015] 根据本发明实施例一实施方式，在根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表之前，该方法还包括：确定磨损因子和聚合段因子。

[0016] 根据本发明实施例第二方面，一种数据指纹检测装置，该装置包括：第一条件检测模块，用于检测是否符合第一条件；变化记录表获取模块，用于获取所有数据分块的第一变化记录表；磨损模块，用于根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表；聚合段因子检测模块，用于检测第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项；聚合模块，用于将可合并表项所对应的至少两个数据分块合并为一个数据分块。

[0017] 根据本发明实施例第三方面，提供一种存储介质，在存储介质上存储了程序指令，其中，程序指令在运行时用于执行上述任一项的数据指纹检测方法。

[0018] 本发明实施例提供一种数据指纹检测方法、装置及存储介质，该方法采用变化记录表来记录每一数据分块在一段时间内发生变化的历史信息，在符合某一条件时，根据预设的磨损因子对变化记录表进行磨损处理；检测磨损处理后的变化记录表是否存在与预设的聚合段因子对应且满足聚合条件的可合并表项，若存在，则将可合并表项所对应的至少两个数据分块合并为一个数据分块。如此，可以根据每一数据变换的次数或频率等信息，动态的合并不经常变化的数据分块，减少了数据分块的数量，也相应地减少了要处理的数据指纹的数量，从而进一步降低了资源的消耗，并大大提高了系统的处理能力和吞吐量。

[0019] 需要理解的是，本发明实施例的教导并不需要实现上面所述的全部有益效果，而是特定的技术方案可以实现特定的技术效果，并且本发明实施例的其他实施方式还能够实现上面未提到的有益效果。

附图说明

[0020] 通过参考附图阅读下文的详细描述，本发明示例性实施方式的上述以及其他目的、特征和优点将变得易于理解。在附图中，以示例性而非限制性的方式示出了本发明的若干实施方式，其中：

[0021] 在附图中，相同或对应的标号表示相同或对应的部分。

[0022] 图1为本发明实施例数据指纹检测方法的实现流程示意图；

[0023] 图2为本发明实施例一应用数据指纹检测方法的具体实现流程示意图；

[0024] 图3为本发明实施例数据指纹检测装置的组成结构示意图。

具体实施方式

[0025] 为使本发明的目的、特征、优点能够更加的明显和易懂，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，而非全部实施例。基于本发明中的实施例，本领域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，都属于本发明保护的范围。

[0026] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0027] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或隐含地包括至少一个该特征。在本发明的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0028] 根据本发明实施例第一方面，一种数据指纹检测方法，如图1所示，该方法包括在符合第一条件时，执行以下操作：操作110，获取所有数据分块的第一变化记录表，第一变化记录表的每一表项记录了每一数据分块在一段时间内发生变化的历史信息，每一数据分块对应一个数据指纹；操作120，根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表；操作130，检测第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项，若存在，则将可合并表项所对应的至少两个数据分块合并为一个数据分块。

[0029] 在操作110中，为了便于对大文件进行差分处理，比如差分备份，通常都会对大文件进行分块以得到多个数据分块，然后为每一个数据分块创建数据指纹。如此，在之后的差分处理中就可以通过对比每个数据分块的数据指纹来获知该数据分块的数据是否发生了变换。

[0030] 为了记录这些数据分块在一段时间内发生变换的历史信息，本发明实施例数据指纹检测方法使用第一变化记录表来记录每一数据分块的变化情况。其中，第一变化记录表的每一表项对应一个数据分块，并记录该数据分块在一段时间内发生变化的历史信息。有了每一数据分块在一段时间内发生变化的历史信息，就可以掌握各个数据发生变化的规律和趋势，并可以动态地将长期不发生变化或变化频率较低的数据分块进行合并，从而实现减少数据分块以及与数据分块相对应的指纹数据数量，进一步提高系统的处理能力和吞吐量。

[0031] 在本实施方式中，并不限定第一变化记录表的具体形式和数据结构，实施者可以根据具体的实施条件选用任何适用的具体形式和数据结构。

[0032] 其中，一段时间是指能够发生两次以上差分处理的一段时间，不然不足以得到与数据变化次数或频率相关的规律，也难以体现本发明实施例数据指纹检测方法的有益效果。从理论上讲，这一段时间累计的数据越多越容易准确定位文件中不经常变化的部分，从而使数据合并的效果越好。但这一段时间也不能太长，因为数据合并是在这段时间之后进

行的，如果这一段时间过长，则会导致不经常变化的数据分块不能及时合并，也不能重复体现本发明实施例数据指纹检测方法的突出效果。因此，这段时间的设定可根据与实际数据较为接近的实验数据进行实验，并根据实验结果进行调整以取得较佳的实际应用效果。

[0033] 其中，发生变化的历史信息，主要是指与每一数据分块在一段时间内数据变化的次数或频率。

[0034] 数据指纹是指可以用来判断数据完整性和同一性的一个序列值，比如，使用某种算法对数据内容进行某种运算得到的、具有标识性的摘要值，其中，常用算法有信息摘要算法(MD5)等。

[0035] 在操作120中，磨损因子指对第一变化记录表中每一表项进行统一磨损处理所依据的一个数值。磨损处理，主要指用每一表项的值减去磨损因子，除以磨损因子，或使用某个以磨损因子为参数的特定函数对每一表项的值进行运算得到一个新的数值。而第二变化记录表就是由这些磨损处理后得到新的数值所组成的表。进行磨损处理，一方面可以突显数据的差异性，另一方面可以更易于找到变化较少的数据分块，减少后续计算的复杂度。

[0036] 在操作130中，聚合段因子主要指进行数据分块合并时一次比较和合并的数据分块的数量。比如，聚合段因子为2的话，就表示每次比较两个数据分块，如果这两个数据分块符合聚合条件时，就将这两个数据分块合并为一个数据分块；聚合段因子为3的话，就表示每次比较三个数据分块，如果这三个数据分块符合聚合条件时，将这三个数据分块合并为一个数据分块；以此类推。此处，聚合段因子要设置为2以上的值，否则没有意义。聚合条件，是指聚合段因子对应的数据分块进行合并要满足的条件，通常这一条件是要求聚合段因子对应的数据分块所记录的变化历史信息相同或相近，且低于某个值，如此就可以找到相邻的、且不经常变化的可合并数据分块。具体实施过程中的聚合条件还可以根据实施条件以及实施者想要达到的目标或效果来自行确定。

[0037] 这一操作是实现数据分块合并的实质性操作，将发生变化的历史信息，特别是一段时间内不经常变化的数据分块进行合并，有利于提高差分处理的效率，减少数据指纹的计算和对比，可大大减少资源消耗，提高系统的吞吐量。

[0038] 根据本发明实施例一实施方式，在符合第一条件之前，该方法还包括：获取待差分处理的文件；对文件进行分块得到数据分块；创建第一变化记录表，其中第一变换记录表的每一表项对应于每一数据分块，用于记录每一数据分块在一段时间内发生变化的历史信息。

[0039] 其中，差分处理指每次仅处理相对于上一次全处理之后新增加的和修改过的数据，例如差分备份等。待差分处理的文件通常都是大文件甚至是超大文件。为了进行差分处理，首先需要对文件进行分块。在没有数据变化的历史信息时，可以先根据一个经验值将文件划分为大小相等的数据分块，并创建一个变化记录表，使用变化记录表中的每一表项来记录每一数据分块在一段时间内的变化情况。这一变化记录表为本发明实施例数据指纹检测方法提供数据分块在一段时间内变化的历史信息，是合并数据分块所依据的数据基础。

[0040] 根据本发明实施例一实施方式，在创建第一变化记录表之后，该方法还包括在对文件进行差分处理时，执行以下操作：获取所有数据分块的第一数据指纹，第一数据指纹是与每一数据分块对应的最新的数据指纹；获取所有数据分块的第二数据指纹，第二数据指纹是与每一数据分块对应的上一次存储的数据指纹；检测每一数据分块的第一数据指纹和

第二数据指纹是否相同，若不同，则更新第一变化记录表中相应数据分块所对应的表项。

[0041] 在这一实施方式中，每次对文件进行差分处理时，会同时记录变化了的数据分块，并更新变化了的数据分块在第一变化记录表中对应的表项。如此，可以记录下每一数据分块在一段时间内，是否发生变化，发生变化的次数或频率等历史信息。

[0042] 根据本发明实施例一实施方式，更新第一变化记录表中相应数据分块所对应的表项，包括：获取相应数据分块所对应的所有表项；依次更新所有表项中的每一表项。

[0043] 当有多个数据分块经过磨损处理后，因符合聚合段因子对应的可合并条件进行合并后，合并后的数据分块就会对应多个表项。此时，需要对该合并后的数据分块所对应的所有表项依次进行更新。

[0044] 根据本发明实施例一实施方式，第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的历史信息，包括：第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的次数；相应地，更新数据分块所对应的表项，包括：使表项所记录的次数加一。

[0045] 在本实施方式中，第一变化记录表的每一表项会记录每一数据分块在一段时间内发生变化的次数。这一记录方式简单易行，且可以很直观地体现每一数据分块在一段时间内的变化次数或频率，以及整个文件不同内容发生变化的不同频率，从而可以动态地、根据数据分块地变化频率合并一些不经常变化的数据分块，从而大大减少数据分块的数量，以及对数据指纹检测的计算和比对等操作。

[0046] 根据本发明实施例一实施方式，第一条件包括到达预定时间。

[0047] 在本实施方式中，进行磨损处理和数据分块的合并是定时进行的。即，每隔一个预设的时长就进行一次磨损处理和数据分块合。这一实施方式，简单易行，且经过实践证明效果较好。如前所述，这一预设的时长是否合理，会影响本发明实施例数据指纹检测方法的实施效果，因此可以通过与实际数据类似的实验数据进行实验，并根据实验结果来确定一个实施效果较好的时长。

[0048] 根据本发明实施例一实施方式，第一变化记录表包括使用位图(bitmap)作为存储结构的变化记录表。

[0049] 在本实施方式中，使用位图作为第一变化记录表的存储结构。所谓的位图就是用一个位(bit)来标记某个元素对应的值(Value)，而键(Key)即是该元素。使用位图还可以把计数排序用的统计数组的每个单位缩小为位级别的布尔数组。由于采用了位为单位来存储数据，可以大大节省存储空间方面，并简化计算的复杂度。

[0050] 根据本发明例一实施方式，在根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表之前，该方法还包括：确定磨损因子和聚合段因子。

[0051] 在本实施方式中，磨损因子和聚合段因子都是预设的值。而将磨损因子和聚合段因子确定为多少合适，可以先根据经验指定几个备选值，然后使用与实际数据类似的实验数据进行实验，并根据实验结果从这几个备选值中选取实施效果较佳的值来设定。

[0052] 下面就结合图2介绍本发明实施例一应用数据指纹检测方法的具体实现流程。在这一应用中，本发明实施例数据指纹检测方法主要应用于大文件的差分备份。为了便于描述过程和体现效果，假设该文件的大小是64M，但实际上该文件的大小可能远远大于64M。首先，假设在实施本发明实施例数据指纹检测方法之前，已经对该文件进行过一次全量备份，

而后续对该文件进行差分备份时，只需传输发生变换的数据分块。其中，进行差分备份并传输差分数据的过程主要包括以下步骤：

[0053] 步骤2010，设置最小数据分块的单位，创建与之对应的变化记录表；

[0054] 假设在执行差分备份的最初，将最小数据分块的单位设置为4M，并对该文件进行等份分块，得到16个数据分块。为该16个数据分块创建一个使用位图作为存储结构的变换记录表，并将每个数据分块对应的位的初始值设为1。此时，该变化记录表的数据状态如表1所示：

[0055]

1(1)	1(2)	1(3)	1(4)
1(5)	1(6)	1(7)	1(8)
1(9)	1(10)	1(11)	1(12)
1(13)	1(14)	1(15)	1(16)

[0056] 表1

[0057] 其中，表1中，每个表格单元中的数值代表每一数据分块对应的表项所存储的值，而数值后括号中的数字则表示数据分块的编号。表1所表示的数据状态就是从第1个数据分块到第16个数据分块中每一数据分块所对应的变化记录表表项的值均为1。

[0058] 步骤2020，按数据分块读取文件；

[0059] 步骤2030，读取数据分块的数据指纹；

[0060] 步骤2040，根据数据指纹的比对结果传输差分数据，并更新变换记录表；

[0061] 假设，根据这一次数据指纹的比对结果，得知第2、6、7、10、13个数据分块发生了变换，则分别对第2、6、7、10、13个数据分块的表项加1，并得到数据状态如表2所示的变化记录表：

[0062]

1(1)	2(2)	1(3)	1(4)
1(5)	2(6)	2(7)	1(8)
1(9)	2(10)	1(11)	1(12)
2(13)	1(14)	1(15)	1(16)

[0063] 表2

[0064] 步骤2050，判断是否到达预设的磨损和合并时间，若是，则继续步骤2060，若否，则回到步骤2020，继续等待下一次的差分处理；

[0065] 在这一次差分备份进行时还未达到预定的磨损和合并时间，于是回到步骤2020，继续等待第二次的差分备份。

[0066] 假设根据步骤2020至步骤2040的指纹检测对比结果，得知在第二次进行差分备份时，第3、7、8、11、14个数据分块发生了变换，则分别对第3、7、8、11、14个数据分块的表项加1，并得到数据状态如表3所示的变化记录表：

[0067]

1(1)	2(2)	2(3)	1(4)
1(5)	2(6)	3(7)	1(8)
1(9)	2(10)	2(11)	1(12)
2(13)	2(14)	1(15)	1(16)

[0068] 表3

[0069] 假设此时已到达预定的磨损和合并时间，继续步骤2060。

[0070] 步骤2060，对变换记录表进行磨损聚合，更新分段信息。之后回到步骤2020，继续等待下一次的差分处理，并下一次差分处理时，使用新的分段信息指定的数据分块来读取文件。

[0071] 假设在本应用中磨损因子为1，先对表3所示的变化记录表进行磨损处理。在该应用中，采用每一表项减去磨损因子的方式来进行磨损处理，得到如表4所示的变化记录表：

[0072]

0(1)	1(2)	1(3)	0(4)
0(5)	1(6)	2(7)	0(8)
0(9)	1(10)	1(11)	0(12)
1(13)	1(14)	0(15)	0(16)

[0073] 表4

[0074] 假设在本应用中，聚合段因子为2。在进行聚合段合并时，首先，读取该聚合段因子对应的数据分块的数值，比如，第1、2数据分块对应的数值(0,1)；第3、4数据分块对应的数值(1,0)；第5、6数据分块对应的数值(0,1)；第7、8数据分块对应的数值(2,0)；第9,10数据分块对应的数值(0,1)；第11、12数据分块对应的数值(1,0)；第13、14数据分块对应的数值(1,1)；第15、16数据分块对应的数值(0,0)。

[0075] 假设在本应用中的聚合条件为聚合段因子对应的聚合段数值都小于等于零。这一聚合条件所代表的实际意义可理解为，在一段时间内，聚合段因子对应的聚合段中的数据分块的变化次数在经过磨损处理后都小小于等于零；换句话说，聚合段因子对应的聚合段中的数据分块的变化次数都小于等于磨损因子所设定的值。根据这一条件，可知在上述的聚合分段中的数据分块中，第15、16数据分块的数值是符合这一聚合条件的，因此第15、16数据分块可以合并为一个数据分块。

[0076] 根据以上的聚合段因子和聚合条件对上述表4进行数据块合并后，得到数据状态如表5所示的变化记录表：

[0077]

0(1)	1(2)	1(3)	0(4)
0(5)	1(6)	2(7)	0(8)
0(9)	1(10)	1(11)	0(12)
1(13)	1(14)	0(15)	0(15)

[0078] 表5

[0079] 其中，原来的第15、16数据分块合并为数据分块15，整个数据分块的数量从16个减少为15个，相应地，在后续的差分备份中，也只需要计算、维护和对比15个数据指纹。

[0080] 之后，就可以回到步骤2020，继续等待下一次的差分备份，并使用新的分段信息指定的数据分块来读取文件，并重复上述过程，持续对变化记录表进行动态地更新、磨损处理和数据分块合并。

[0081] 需要说明的是，在表5所示的变化记录表中，由于第15个数据分块是之前第15、16数据分块合并而成的，所以第15个数据分块会对应两个表项，在根据第15个数据分块的变化更新该变化记录表时，需要获取第15个数据分块所对应的所有表项(即表项15和表项16)，并依次更新第15个数据分块所对应的所有表项。

[0082] 需要说明的是，上述本发明实施例一应用数据指纹检测方法的具体实现流程仅为

示例性说明，实施者可以根据实施条件采取任何适用地实施方法来实现本发明实施例数据指纹检测方法。

[0083] 根据本发明实施例第二方面，一种数据指纹检测装置，如图3所示，该装置30包括：第一条件检测模块301，用于检测是否符合第一条件；变化记录表获取模块302，用于获取所有数据分块的第一变化记录表；磨损模块303，用于根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表；聚合段因子检测模块304，用于检测第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项；聚合模块305，用于将可合并表项所对应的至少两个数据分块合并为一个数据分块。

[0084] 根据本发明实施例一实施方式，该装置30还包括：差分文件获取模块，用于获取待差分处理的文件；分块模块，用于对文件进行分块得到数据分块；变化记录表创建模块，用于创建第一变化记录表，其中第一变换记录表的每一表项对应于每一数据分块，用于记录每一数据分块在一段时间内发生变化的历史信息。

[0085] 根据本发明实施例一实施方式，该装置30还包括：第一数据指纹获取模块，用于获取所有数据分块的第一数据指纹，第一数据指纹是与每一数据分块对应的最新的数据指纹；第二数据指纹获取模块，用于获取所有数据分块的第二数据指纹，第二数据指纹是与每一数据分块对应的上一次存储的数据指纹；数据指纹检测模块，用于检测每一数据分块的第一数据指纹和第二数据指纹是否相同；变化记录表更新模块，用于更新第一变化记录表中相应数据分块所对应的表项。

[0086] 根据本发明实施例一实施方式，变化记录表更新模块包括：表项获取子模块，用于获取相应数据分块所对应的所有表项；表项更新子模块，用于依次更新所有表项中的每一表项。

[0087] 根据本发明实施例一实施方式，第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的历史信息，包括：第一变换记录表的每一表项用于记录每一数据分块在一段时间内发生变化的次数；相应地，变化记录表更新模块具体用于使表项所记录的次数加一。

[0088] 根据本发明实施例一实施方式，第一条件检测模块301具体用于检测是否导电预定时间。

[0089] 根据本发明实施例一实施方式，该装置30还包括磨损因子和聚合段因子确定模块，用于确定磨损因子和聚合段因子。

[0090] 根据本发明实施例第三方面，提供一种存储介质，在存储介质上存储了程序指令，其中，程序指令在运行时用于执行上述任一项的数据指纹检测方法。

[0091] 这里需要指出的是：以上针对数据指纹检测装置实施例的描述和以上针对计算机存储介质实施例的描述，与前述方法实施例的描述是类似的，具有同前述方法实施例相似的有益效果，因此不做赘述。对于本发明对数据指纹检测装置实施例的描述和对计算机存储介质实施例的描述尚未披露的技术细节，请参照本发明前述方法实施例的描述而理解，为节约篇幅，因此不再赘述。

[0092] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有

的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0093] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备和方法，可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的，例如，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，如：多个单元或组件可以结合，或可以集成到另一个装置，或一些特征可以忽略，或不执行。另外，所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口，设备或单元的间接耦合或通信连接，可以是电性的、机械的或其它形式的。

[0094] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的，作为单元显示的部件可以是、或也可以不是物理单元；既可以位于一个地方，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全部单元来实现本实施例方案的目的。

[0095] 另外，在本发明各实施例中的各功能单元可以全部集成在一个处理单元中，也可以是各单元分别单独作为一个单元，也可以两个或两个以上单元集成在一个单元中；上述集成的单元既可以利用硬件的形式实现，也可以利用硬件加软件功能单元的形式实现。

[0096] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储介质、只读存储器(Read Only Memory，ROM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0097] 或者，本发明上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明实施例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，包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本发明各个实施例方法的全部或部分。而前述的存储介质包括：移动存储介质、ROM、磁碟或者光盘等各种可以存储程序代码的介质。

[0098] 以上，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，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以权利要求的保护范围为准。

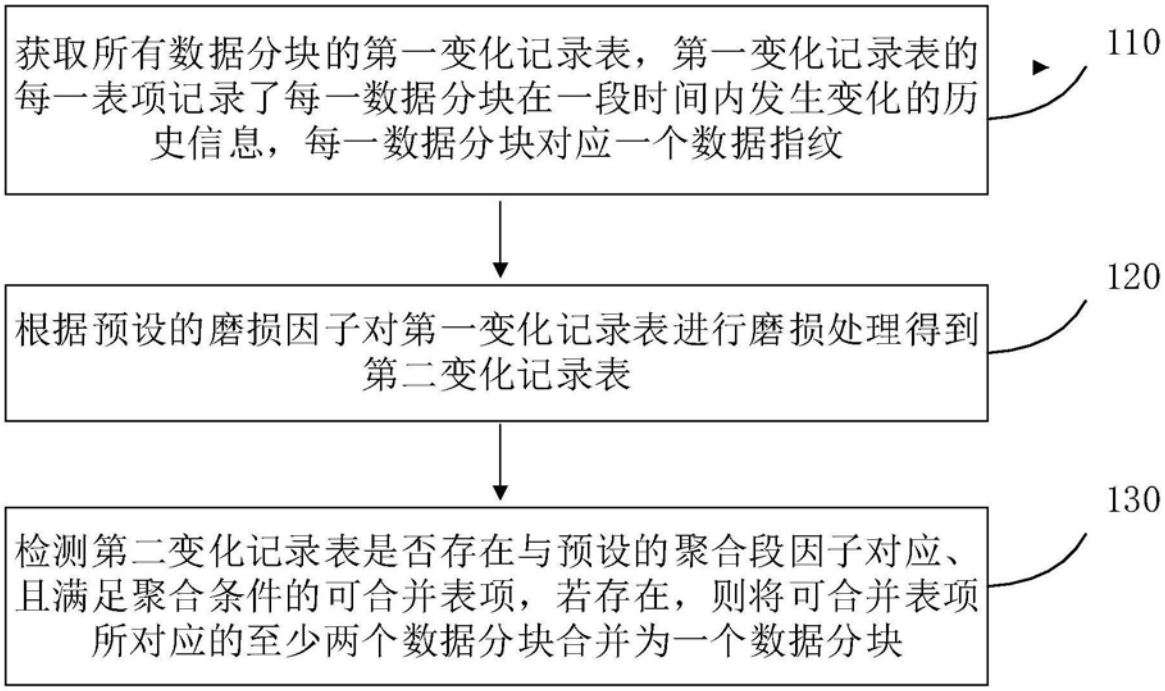

获取所有数据分块的第一变化记录表，第一变化记录表的每一表项记录了每一数据分块在一段时间内发生变化的历史信息，每一数据分块对应一个数据指纹
110
根据预设的磨损因子对第一变化记录表进行磨损处理得到第二变化记录表
120
检测第二变化记录表是否存在与预设的聚合段因子对应、且满足聚合条件的可合并表项，若存在，则将可合并表项所对应的至少两个数据分块合并为一个数据分块
130

图1

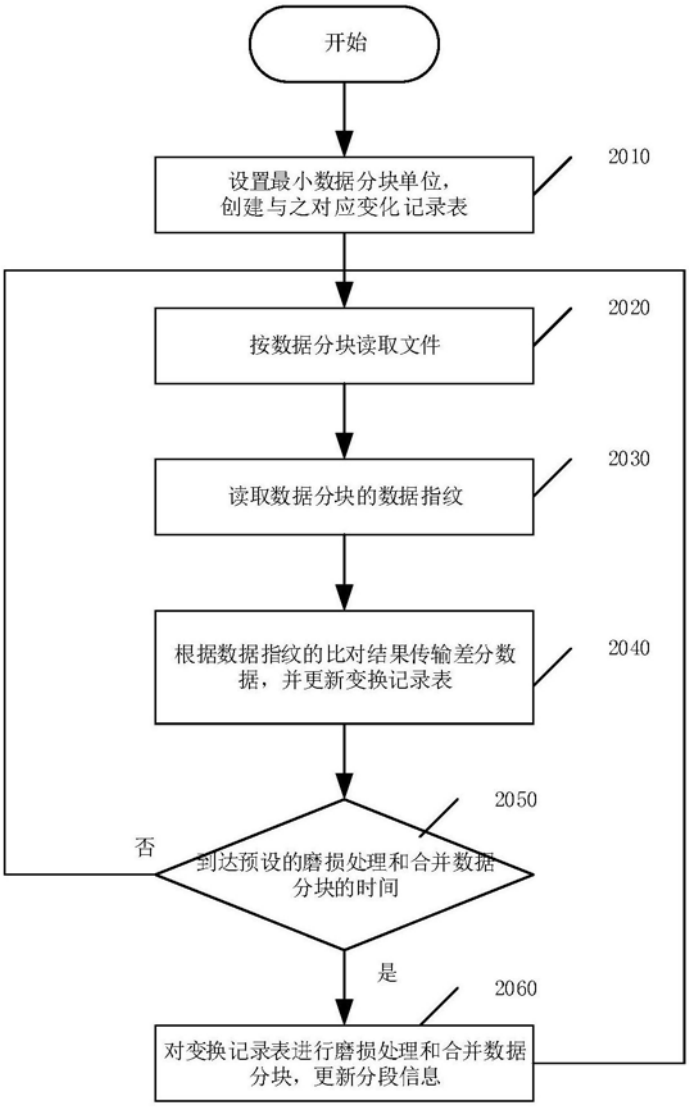

开始
设置最小数据分块单位，
创建与之对应变化记录表
2010
按数据分块读取文件
2020
读取数据分块的数据指纹
2030
根据数据指纹的比对结果传输差分数据，并更新变换记录表
2040
到达预设的磨损处理和合并数据分块的时间
2050
否
是
对变换记录表进行磨损处理和合并数据分块，更新分段信息
2060

图2

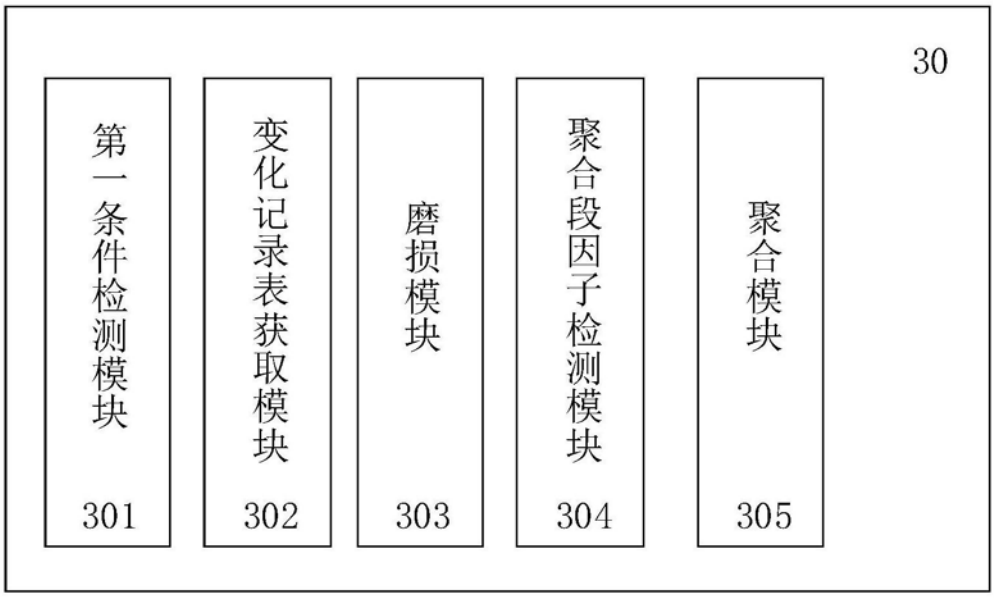

30
第一条件检测模块
301
变化记录表获取模块
302
磨损模块
303
聚合段因子检测模块
304
聚合模块
305

图3