(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103617215 B
(45)授权公告日 2017.02.08

(21)申请号 201310586671.6

(22)申请日 2013.11.20

(65)同一申请的已公布的文献号
申请公布号 CN 103617215 A

(43)申请公布日 2014.03.05

(73)专利权人 上海爱数信息技术股份有限公司
地址 201112 上海市闵行区联航路1188号8幢第2层A-1单元

(72)发明人 叶佑群

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 王民盛 王丽琴

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 103297429 A,2013.09.11,
CN 101546320 A,2009.09.30,
CN 102231727 A,2011.11.02,
CN 103379160 A,2013.10.30,
US 5940507 A,1999.08.17,
US 2008034268 A1,2008.02.07,

审查员 张伯

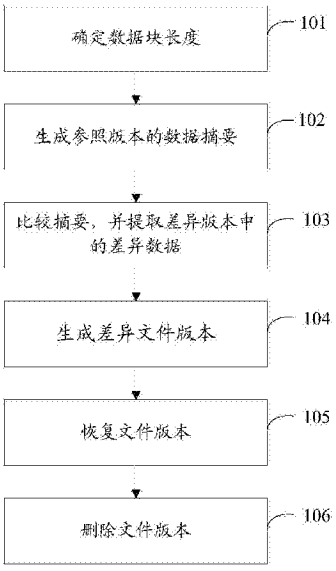
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种利用数据差异算法生成多版本文件的方法

(57)摘要

本申请公开了一种利用数据差异算法生成多版本文件的方法，包括：A、按照预先确定的数据块长度，对参照版本从头至尾依次计算每个数据块的数据摘要值；B、计算在各个偏移量下，与参照版本相同长度的差异版本的数据块的摘要值，将计算得到的差异版本的数据块的摘要值与参照版本的数据块的摘要值进行对比，根据对比结果提取出差异版本中的差异数据，并生成用于存储所述差异数据的差异记录文件。本申请方案可以有效节约多版本文件的存储空间。



CN 103617215 B

1.一种利用数据差异算法生成多版本文件的方法，其特征在于，包括：

A、按照预先确定的数据块长度，对参照版本从头至尾依次计算每个数据块的数据摘要值；包括：对每一个数据块，计算一个长度为N1的快速摘要，以及计算一个长度为N2的慢速摘要，其中N1<N2；快速摘要值相同是数据块相同的必要非充分条件，慢速摘要值相同是数据块相同的充分必要条件；

B、计算在各个偏移量下，与参照版本相同长度的差异版本的数据块的摘要值，将计算得到的差异版本的数据块的摘要值与参照版本的数据块的摘要值进行对比，根据对比结果提取出差异版本中的差异数据，并生成用于存储所述差异数据的差异记录文件，所述参照版本和差异版本作为多版本文件中的不同版本，所述差异记录文件用于作为多版本文件中，从基准版本得到差异版本的依据；

步骤B具体包括：

B1、生成用于存储参照版本的数据块摘要对的哈希数组H，创建一个空白的差异记录文件，然后向这个文件写入16字节的空内容；然后创建一个链表结构K，用于记录差异数据块的比较结果；

B2、将差异版本的文件偏移设置为0；

B3、判断从差异版本当前偏移处开始向后的数据长度是否小于L，若是，执行步骤B4，否则执行步骤B5；

B4、将从差异版本当前偏移处开始向后的数据记录到差异记录文件中，提取差异记录文件当前大小以及链表元素的个数后，将链表K写入到差异记录文件的末尾；然后将文件大小及元素个数写入到差异记录文件中开头的预留字节中，并结束步骤B；

B5、从差异版本当前偏移处开始向后读取长度L的数据块，计算该数据块的快速摘要；

B6、判断是否在哈希数组H中查找与之相同的快速摘要，若找到，执行步骤B8，否则执行步骤B7；

B7、将差异版本的偏移向后移动一个字节，并返回步骤B3；

B8、计算该数据块的慢速摘要，在已经找到快速摘要的哈希数组中的链表中查找是否有相同的慢速摘要，如果找到，执行步骤B9，否则，执行步骤B7；

B9、将该数据块之前已经滑过的数据作为差异数据写入到差异记录文件中，并且生成一个数据块记录项，其类型为“不同”，所述数据块记录项记录这一差异数据在差异记录文件中的偏移，以及数据块的长度，并将记录项插入到链表K的末尾；

B10、将这一个相同数据块的信息，生成一个数据块记录项，其类型为“相同”，所述数据块记录项记录本块数据在参照版本中的偏移以及块长度L，然后将该数据块记录项插入到链表K的末尾；

B11、将差异版本的偏移向后移动L，然后转至步骤B3。

2.根据权利要求1所述的方法，其特征在于，所述数据块长度根据如下公式确定：$L=\left[\left(Log_2 SIZE\right)\left(\sqrt[3]{SIZE}\right)\right]$；其中，中括号表示取整，L表示数据块长度，单位为字节，SIZE表示参照版本文件大小，单位为字节；

若根据公式计算出的L小于200字节，则将数据块长度设置为200字节，若根据公式计算出的L大于512K字节，则将数据块长度设置为512K字节。

3.根据权利要求2所述的方法，其特征在于，所述N1＝4，N2＝16。

4.根据权利要求1所述的方法，其特征在于，步骤B之后进一步包括：

创建一个空的恢复文件；

从差异记录文件中提取数据块记录项；

从数据块记录项中读取数据的偏移以及数据块的长度，然后判断数据块记录项中的数据块类型，若为“不同”，则从差异记录文件中从偏移处读取相应长度字节的块数据，然后写入恢复文件中；若为“相同”，则从参照版本中从偏移处读取相应长度字节的块数据，然后写入恢复文件中。

一种利用数据差异算法生成多版本文件的方法

技术领域

[0001] 本申请涉及计算机数据与存储技术领域，尤其涉及一种利用数据差异算法生成多版本文件的方法。

背景技术

[0002] 随着计算机技术的普及，计算机应用已经渗透到日常生活当中的方方面面。对于文档等各种类型的非结构化数据的处理与存储是我们经常需要面临的问题。特别是当前移动办公方式的兴起，要求在数据传输时具有更小的数量传输量，否则容易导致用户使用成本的提高。但是无论是传统的计算处理与存储还是当前的移动计算处理与存储，都向着集中存储与处理的方向发展。在这种情况下，用户的文档以及其他数据都将只存储于一个集中的数据处理中心，或者存储在用户的本地计算机中。

[0003] 但是无论是集中处理存储与处理还是用户在本地处理，都面临着一个这样的问题，即用户在特定的情况下需要恢复某些文件到某个特定时刻的数据状态。如果用户只在本地存储，那么他将只有一个版本，而在集中处理的方式下，最好也就可能存在两个版本。但是往往这两个版本都不用是用户所需要的数据。针对这种情况，通常的解决方法是在某个条件下会复制一份完全一样的文件数据，并存储于相应的位置。当需要还原这些文件数据时，只要找到相应的某个复本或者与要求的版本最相近的一个复本。

[0004] 显而易见，这种处理方式虽然直接，但是所具有的缺点也是明显的：首先，存储的数据存在大量的冗余。因为文件是完全的复本存储，当需要多少个版本时，其本上就需要多少倍的存储空间。这会造成存储容量需求大增，增加成本支持。为了限制成本的增加，就会导致文件复本存储的数量受到限制，进而影响多复本存储的效率与可用性。其次，当数据的复本是存储于集中数据处理中心时，还会导致在网络上传输的数据量大增，使网络受到较严重的影响。更为严重的是，如果需要处理的文件的尺寸过大时，这两个缺陷所导致的问题会更加明显而难以解决。

发明内容

[0005] 本申请提供了一种利用数据差异算法生成多版本文件的方法，可以有效节约多版本文件的存储空间。

[0006] 本申请实施例提供了一种利用数据差异算法生成多版本文件的方法，包括：

[0007] A、按照预先确定的数据块长度，对参照版本从头至尾依次计算每个数据块的数据摘要值；

[0008] B、计算在各个偏移量下，与参照版本相同长度的差异版本的数据块的摘要值，将计算得到的差异版本的数据块的摘要值与参照版本的数据块的摘要值进行对比，根据对比结果提取出差异版本中的差异数据，并生成用于存储所述差异数据的差异记录文件。

[0009] 较佳地，所述数据块长度根据如下公式确定：$L=\left[\left(Log_2 SIZE\right)\left(\sqrt[3]{SIZE}\right)\right]$；其

中，中括号表示取整，L表示数据块长度，单位为字节，SIZE表示参照版本文件大小，单位为字节；

[0010] 若根据公式计算出的L小于200字节，则将数据块长度设置为200字节，若根据公式计算出的L大于512K字节，则将数据块长度设置为512K字节。

[0011] 较佳地，步骤A所述计算每个数据块的数据摘要值包括：对每一个数据块，计算一个长度为N1的快速摘要，以及计算一个长度为N2的慢速摘要，其中N1<N2；快速摘要值相同是数据块相同的必要非充分条件，慢速摘要值相同是数据块相同的充分必要条件；

[0012] 步骤B包括：

[0013] B1、生成用于存储参照版本的数据块摘要对的哈希数组H，创建一个空白的差异记录文件，然后向这个文件写入16字节的空内容；然后创建一个链表结构K，用于记录差异数据块的比较结果；

[0014] B2、将差异版本的文件偏移设置为0；

[0015] B3、判断从差异版本当前偏移处开始向后的数据长度是否小于L，若是，执行步骤B4，否则执行步骤B5；

[0016] B4、将从差异版本当前偏移处开始向后的数据记录到差异记录文件中，提取差异记录文件当前大小以及链表元素的个数后，将链表K写入到差异记录文件的末尾；然后将文件大小及元素个数写入到差异记录文件中开头的预留字节中，并结束本流程；

[0017] B5、从差异版本当前偏移处开始向后读取长度L的数据块，计算该数据块的快速摘要；

[0018] B6、判断是否在哈希数组H中查找与之相同的快速摘要，若找到，执行步骤B8，否则执行步骤B7；

[0019] B7、将差异版本的偏移向后移动一个字节，并返回步骤B3；

[0020] B8、计算该数据块的慢速摘要，在已经找到快速摘要的哈希数组中的链表中查找是否有相同的慢速摘要。如果找到，执行步骤B9，否则，执行步骤B7；

[0021] B9、将该数据块之前已经滑过的数据作为差异数据写入到差异记录文件中，并且生成一个数据块记录项，其类型为“不同”，所述数据块记录项记录这一差异数据在差异记录文件中的偏移，以及数据块的长度，并将记录项插入到链表K的末尾；

[0022] B10、将这一个相同数据块的信息，生成一个数据块记录项，其类型为“相同”，所述数据块记录项记录本块数据在参照版本中的偏移以及块长度L，然后将该数据块记录项插入到链表K的末尾；

[0023] B11、将差异版本的偏移向后移动L，然后转至步骤B3。

[0024] 较佳地，所述N1=4，N2=16。

[0025] 较佳地，步骤B之后进一步包括：

[0026] 创建一个空的恢复文件；

[0027] 从差异记录文件中提取数据块记录项；

[0028] 从数据块记录项中读取数据的偏移以及数据块的长度，然后判断数据块记录项中的数据块类型，若为“不同”，则从差异记录文件中从偏移处读取相应长度字节的块数据，然后写入恢复文件中；若为“相同”，则从参照版本中从偏移处读取相应长度字节的块数据，然后写入恢复文件中。

[0029] 从以上技术方案可以看出，提取文件版本之间的差异数据，并生成差异记录文件，基于差异记录文件可以从基准文件版本得到差异文件版本。通过这样处理，可以极大地减少需要存储的数据量以及在网络上传输的数据量。有效地利用用户的存储空间以及网络带宽，以及降低用户投资成本。

附图说明

[0030] 图1为本申请实施例提供的利用数据差异算法生成多版本文件的方法流程图；

[0031] 图2为本申请实施例中记录差异文件版本的文件结构示意图；

[0032] 图3为图1所示流程中步骤103的具体实现流程示意图。

具体实施方式

[0033] 本申请方案的基本思想是：通过提取文件版本之间的差异数据，并存储为一个文件的某个版本的复本。通过这样处理，可以极大地减少需要存储的数据量以及在网络上传输的数据量。有效地利用用户的存储空间以及网络带宽，以及降低用户投资成本。

[0034] 为使本申请技术方案的技术原理、特点以及技术效果更加清楚，以下结合具体实施例对本申请技术方案进行详细阐述。

[0035] 本申请实施例提供的利用数据差异算法生成多版本文件的方法，通过比较文件的两个版本之间相同数据块长度的数据块的数据摘要值是否相同，来确定数据块是否相 同，并以此为依据来提取差异数据。具体流程如图1所示，包括如下步骤：

[0036] 步骤101：确定数据块长度。

[0037] 数据块长度是同一个文件的两个不同版本之间进行比较时数据块的划分依据。这两个版本中，其中一个称之为参照版本，即比较时作为基准的文件版本；另一个称之为差异版本，即最终要被用于与参照版本进行比较生成差异数据的文件版本。

[0038] 通常情况下，文件尺寸越大，其被更改的可能性越小，以及其更改的范围越小。反之亦然。为了提高比较的效率，数据块的长度的确定原则是与文件大小相关的：当文件大时，其块长度亦大；反之，当文件越小时，其数据块长度越小。如果定义数据块长度为L，参照版本文件大小为SIZE，两者单位为字节，可以通过公式(1)来确定数据块长度：

[0039] $$L = \left[(Log_2 SIZE) \left(\sqrt[3]{SIZE} \right) \right] \qquad (1)$$

[0040] 其中，中括号表示取整。数据块长度L为参照版本大小取以2为底的对数，并与参照版本文件大小开3次方相乘，并取整。这是一个分析指导下的经验值，本申请实施例中，取数据块长度的下限为200字节，上限为512K字节。若根据公式计算出的L小于200字节，则将数据块长度设置为200字节，若根据公式计算出的L大于512K字节，则将数据块长度设置为512K字节。

[0041] 步骤102：生成参照版本的数据摘要。

[0042] 已经确定好了数据块的长度后，就可以按照这个数据块长度对参照版本从头至尾依次计算每个数据块的数据摘要值。为了在比较时查找的快速，将每个数据块计算两个摘要值，一个是快速摘要，另一个是慢速摘要。快速摘要是一个长度为N1字节的数据块的特征码，用于快速定位有可能相同的数据块。相同的数据块一定有相同的快速摘要，而不相同的

数据块不一定会有不相同的快速摘要值，即快速摘要值相同是数据块相同的必要非充分条件。而慢速摘要是一个长度为N2字节的数据特征码。用于确定所比较的两块数据是否完全相同。相同的数据块，一定有相同的慢速摘要值，而不相同的数据一定有不相同的慢速摘要值，即慢速摘要值相同是数据块相同的充分必要条件。本申请实施例中，N1=4，N2=16。

[0043] 将同一个数据块的快速摘要和慢速摘要称为一个摘要对。较佳地，为了能够快速查找到快速摘要与慢速摘要，可以将摘要对存放于一个哈希数组中，并且以快速摘要的值作为计算数据下标的依据。

[0044] 哈希数组是这样的一种数据结构，存放于某个元素到数组中，是通过某个稳定的方法从元素计算出一个下标值来确定其最终的存储位置。如果在计算出来的位置上已经存在某个元素，则将这个新元素连接到相应位置上最后一个元素的末尾，并且自己成为新的末尾元素。因此具有相同计算出相同数据位置的元素形成一个链表。我们称这个哈希表为H。

[0045] 步骤103：比较摘要，并提取差异版本中的差异数据。

[0046] 为了查找差异数据，必须先查找相同的数据。在步骤102中已经取得了参照版本的所有数据块的快慢摘要值。而在差异版本中，相同的数据块可能出现在文件中的任何一个位置，因此需要计算在各个偏移量下，与参照版本相同长度的差异版本的数据块的摘要值。

[0047] 步骤103的具体实现过程如图3所示，包括如下子步骤：

[0048] 子步骤103-1：生成用于存储参照版本的数据块摘要对的哈希数组H。创建一个空白的差异记录文件，然后向这个文件写入16字节的空内容。这16字节需要用于分析完成后，将分析结果的块记录信息写入到这16字节的空间中。然后创建一个链表结构K，用于记录差异数据块的比较结果。

[0049] 子步骤103-2：将差异版本的文件偏移设置为0。

[0050] 子步骤103-3：判断从差异版本当前偏移处开始向后的数据长度是否小于L，若是，执行子步骤103-4，否则执行子步骤103-5。

[0051] 子步骤103-4：将从差异版本当前偏移处开始向后的数据记录到差异记录文件中，提取差异记录文件当前大小以及链表元素的个数后，将链表K写入到差异记录文件的末尾；然后将文件大小及链表元素个数写入到差异记录文件中开头的预留字节中，并结束本流程。

[0052] 子步骤103-5：从差异版本文件当前偏移处开始向后读取长度L的数据块，计算该数据块的快速摘要。

[0053] 子步骤103-6：判断是否在哈希数组H中查找与之相同的快速摘要，若找到，执行子步骤103-8，否则执行子步骤103-7。

[0054] 子步骤103-7：将差异版本的偏移向后移动一个字节，并返回子步骤103-3。

[0055] 子步骤103-8：计算该数据块的慢速摘要，在已经找到快速摘要的哈希数组中的链表中查找是否有相同的慢速摘要。如果找到，说明找到一块相同数据块，执行子步骤103-9，否则，执行子步骤103-7。

[0056] 子步骤103-9：将该数据之前已经滑过的数据作为差异数据写入到差异记录文件中，并且生成一个数据块记录项，其类型为不同(DIFFER)，该数据块记录项记录这一差异数据在差异记录文件中的偏移，以及数据块的长度，并将记录项插入到链表K的末尾。

[0057] 子步骤103-10:将这一个相同数据块的信息,生成一个数据块记录项,其类型为相同(SAME),该数据块记录项记录本块数据在参照版本中的偏移以及块长度L,然后将该记录项插入到链表K的末尾。

[0058] 子步骤103-11:将差异版本的偏移向后移动L,然后转至子步骤103-3。

[0059] 步骤103完成后即完成了差异数据的提取。差异数据已经写到了差异记录文件中,链表K也记录在差异记录文件中。

[0060] 步骤104:生成差异文件版本。

[0061] 由于在比较差异版本与参照版本时是顺序扫描差异版本的,因此在生成差异数据文件版本时只需要顺序计录差异数据以及相同数据块信息即可。因此很自然的,差异记录文件采用如图2所示的文件结构来记录差异文件版本。其中文件头长度为16字节,记录了差异记录文件当前大小以及链表元素的个数。"差异数据块"记录的是差异版本与参照版本的差异数据。"数据块描述"即链表K中的各个数据块记录项。

[0062] 步骤105:恢复文件版本。

[0063] 当要恢复某些版本的文件时,创建一个空的恢复文件,找到相应版本的差异记录文件,并从中提取数据块记录项,并依次执行如下步骤:

[0064] 从数据块记录项中读取数据的偏移以及数据块的长度,然后判断数据块记录项中的数据块类型。如果数据块类型为差异数据,则从差异记录文件中从偏移处读取相应长度字节的块数据,然后写入恢复文件中。如果数据块类型为相同数据,则从参照版本中从偏移处读取相应长度字节的块数据,然后写入恢复文件中。处理完所有的数据块记录项,即完成了文件版本的恢复。

[0065] 步骤106:删除文件版本。

[0066] 由于每个版本都是与其最初的参照版本进行比较,因此各版本之间不存在关联性,所以直接删除相应的差异记录文件即可。

[0067] 通过本发明可以达到如下效果:

[0068] 1.在进行差异数据比较时,采用的是快速摘要进行初步筛选,慢速摘要最终确认的方法,加快了文件比较的时间,提升了性能。

[0069] 2.在记录差异版本时,只记录差异数据以及相同数据块的块索引信息,因此差异版本的版本记录可以最小化,从而节约了大量的存储空间。

[0070] 3.由于在比较差异数据时,只需要传递相应的参照版本的快慢摘要数据就可以完成,因此如果需要通过网络传递数据时可以大大减少传输的数据量。

[0071] 4.在传输差异版本时,只传输差异数据以及相同数据块的块索引信息,从而节约了大量的网络带宽空间。

[0072] 以上所述仅为本申请的较佳实施例而已,并不用以限制本申请的保护范围,凡在本申请技术方案的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本申请保护的范围之内。

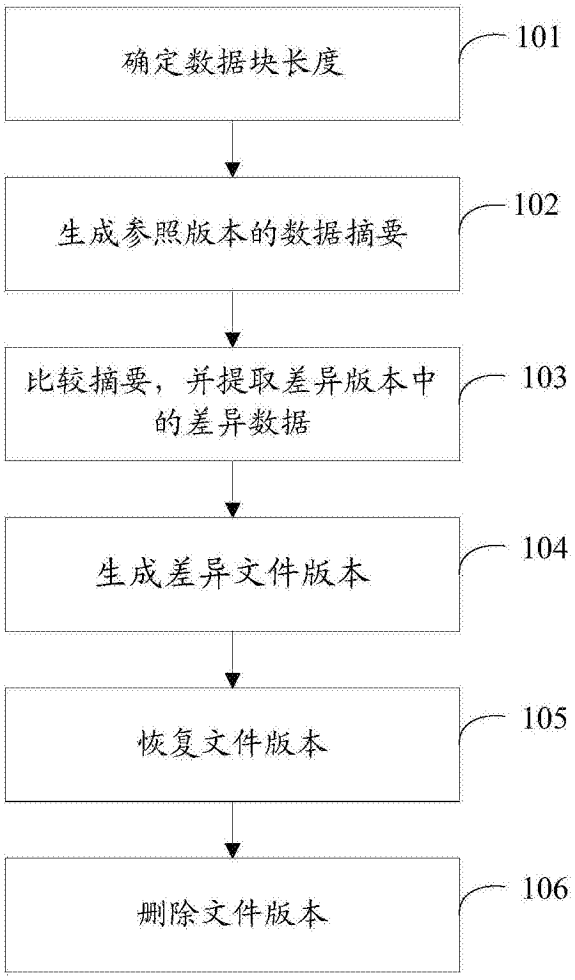

确定数据块长度
101
生成参照版本的数据摘要
102
比较摘要，并提取差异版本中的差异数据
103
生成差异文件版本
104
恢复文件版本
105
删除文件版本
106

图1

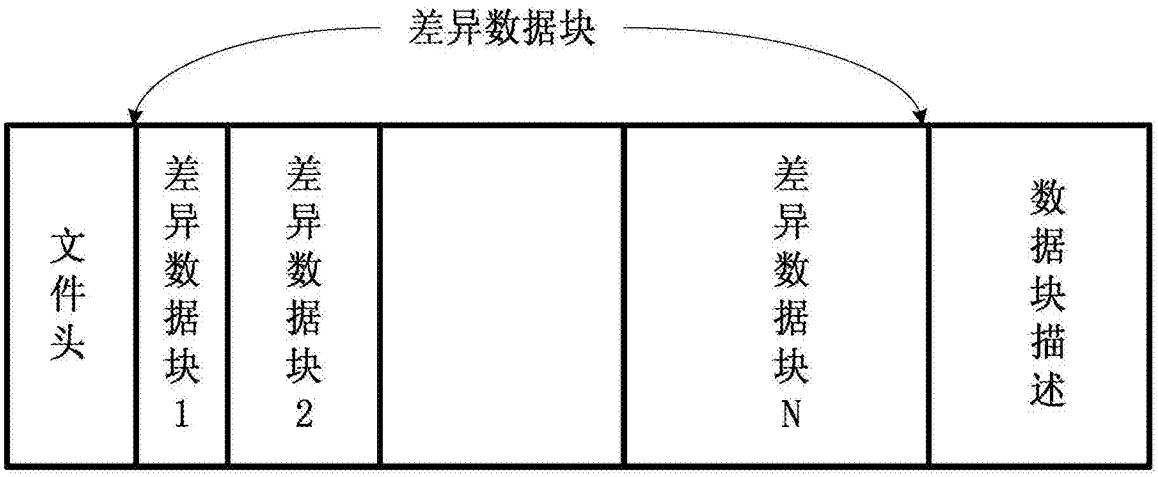

差异数据块
文件头
差异数据块1
差异数据块2
差异数据块N
数据块描述

图2

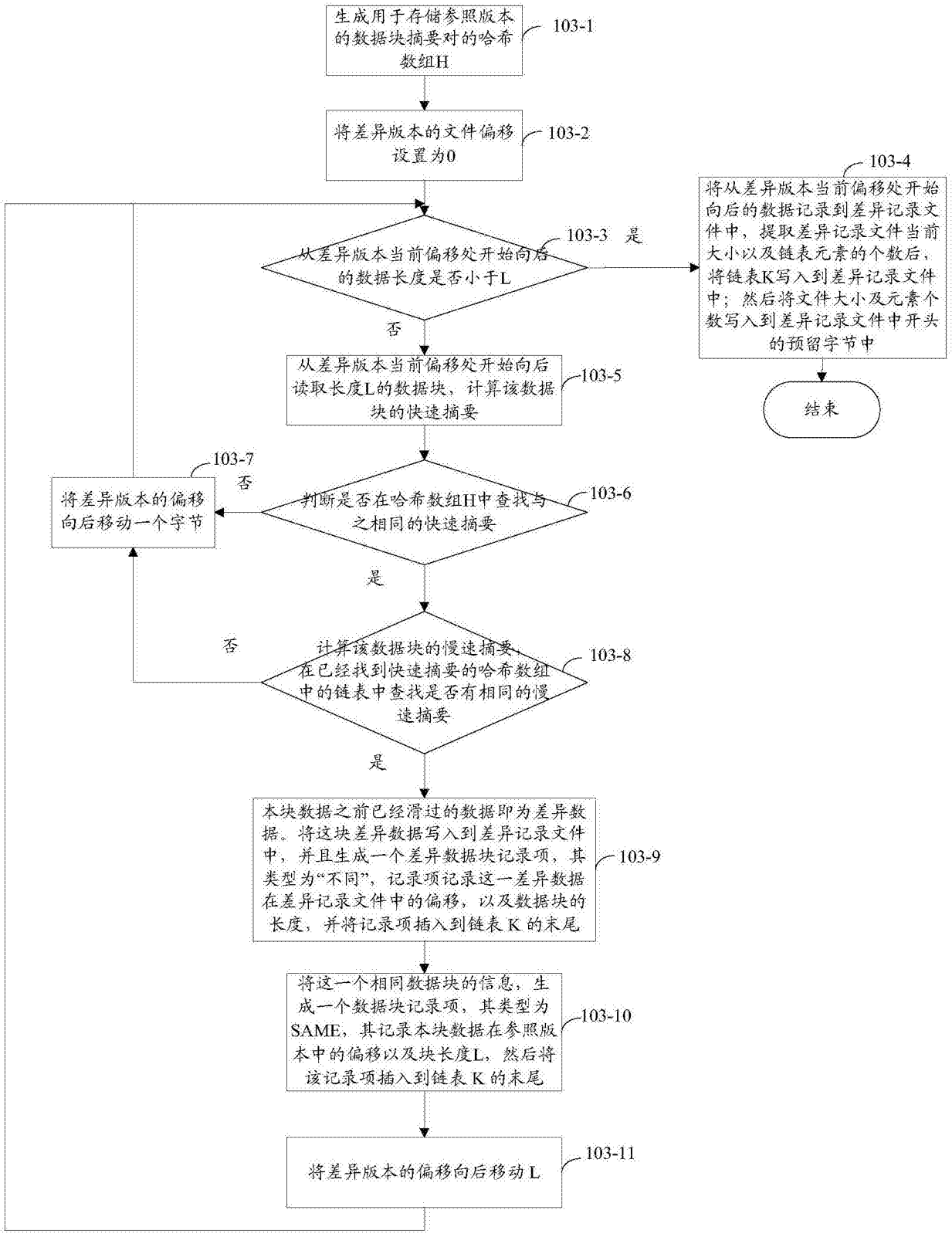

生成用于存储参照版本的数据块摘要对的哈希数组H
103-1
将差异版本的文件偏移设置为0
103-2
从差异版本当前偏移处开始向后的数据长度是否小于L
103-3
是
否
将从差异版本当前偏移处开始向后的数据记录到差异记录文件中，提取差异记录文件当前大小以及链表元素的个数后，将链表K写入到差异记录文件中；然后将文件大小及元素个数写入到差异记录文件中开头的预留字节中
103-4
结束
从差异版本当前偏移处开始向后读取长度L的数据块，计算该数据块的快速摘要
103-5
判断是否在哈希数组H中查找与之相同的快速摘要
103-6
否
将差异版本的偏移向后移动一个字节
103-7
是
计算该数据块的慢速摘要，在已经找到快速摘要的哈希数组中的链表中查找是否有相同的慢速摘要
103-8
否
是
本块数据之前已经滑过的数据即为差异数据。将这块差异数据写入到差异记录文件中，并且生成一个差异数据块记录项，其类型为“不同”，记录项记录这一差异数据在差异记录文件中的偏移，以及数据块的长度，并将记录项插入到链表 K 的末尾
103-9
将这一个相同数据块的信息，生成一个数据块记录项，其类型为SAME，其记录本块数据在参照版本中的偏移以及块长度L，然后将该记录项插入到链表 K 的末尾
103-10
将差异版本的偏移向后移动 L
103-11

图3