(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 101707884 B
(45)授权公告日 2013.11.06

(21)申请号 200880012226.6
(22)申请日 2008.04.09
(30)优先权数据
60/925,624 2007.04.19 US
11/807,204 2007.05.24 US
(85)PCT申请进入国家阶段日
2009.10.16
(86)PCT申请的申请数据
PCT/US2008/004560 2008.04.09
(87)PCT申请的公布数据
WO2008/130504 EN 2008.10.30
(73)专利权人 EMC 公司
地址 美国马萨诸塞州
(72)发明人 R·H·帕特森
(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 张晓冬 李家麟
(51)Int.Cl.
G11C 7/10(2006.01)

(56)对比文件
CN 1302022 A,2001.07.04, 全文.
CN 1512376 A,2004.07.14, 全文.
US 2007/0061266 A1,2007.03.15, 全文.
US 2006/0220790 A1,2006.10.05, 全文.

审查员 王伟

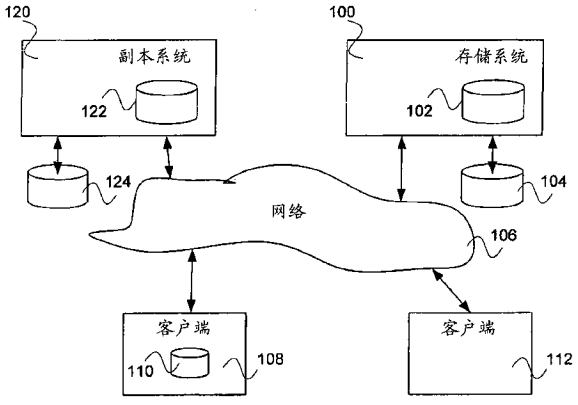
权利要求书2页 说明书6页 附图8页

(54)发明名称

播种复制

(57)摘要

本发明公开了一种播种复制方法。选择存储在重复删除的存储系统上要被复制的一个或多个而非所有文件。确定由所选择的一个或多个而非所有文件所涉及的一个或多个段。创建用于指示至少所述一个或多个段要被复制的数据结构。如果至少部分地基于所述数据结构的指示指示了存储在所述重复删除存储系统上的候选段要被复制,则复制所述候选段。



CN 101707884 B

1. 一种播种复制的方法，包括：

选择存储在重复删除的存储系统上要复制的一个或多个而非所有文件；

确定由所选择的一个或多个而非所有文件所引用的一个或多个段；

创建被用于指示至少所述一个或多个段是要被复制的数据结构；以及

如果所述数据结构指示了存储在所述重复删除存储系统上的所述一个或多个段中的一个是要复制的，则复制所述一个或多个段中的所述一个，

其中，所述数据结构被用于指示是否段被引用以确定该段是否应当被复制。

2. 如权利要求 1 所述的方法，其中，使用基于内容的引用来引用所述一个或多个段。

3. 如权利要求 1 所述的方法，其中，使用非基于内容的引用来引用所述一个或多个段。

4. 如权利要求 1 所述的方法，其中，使用指纹来引用所述一个或多个段。

5. 如权利要求 1 所述的方法，其中，所述数据结构包括段的精确引用的列表。

6. 如权利要求 5 所述的方法，其中，对段的所述精确引用的列表进行重复删除。

7. 如权利要求 5 所述的方法，其中，对段的所述精确引用的列表进行分类。

8. 如权利要求 1 所述的方法，其中，所述数据结构包括概率数据结构。

9. 如权利要求 8 所述的方法，其中，所述概率数据结构包括段的不精确引用的列表。

10. 如权利要求 8 所述的方法，其中，所述概率数据结构包括布隆过滤器。

11. 如权利要求 8 所述的方法，其中，所述概率数据结构指示段可能被所选择的一个或多个而非所有文件引用。

12. 如权利要求 1 所述的方法，其中，创建所述数据结构包括使用哈希函数。

13. 如权利要求 1 所述的方法，其中，所述数据结构响应于播种副本系统的请求而被构建。

14. 如权利要求 1 所述的方法，其中，复制候选段包括将所述段发送给副本系统。

15. 如权利要求 1 所述的方法，其中，复制候选段包括接收所述段被收到或被存储在副本系统上的确认。

16. 如权利要求 1 所述的方法，其中，复制候选段包括向副本系统发送引用以看看所述引用所引用的段是否已存储在所述副本系统上。

17. 如权利要求 1 所述的方法，其中，复制候选段包括在所述段被指示为还未被存储在副本系统上的情况下发送所述段。

18. 如权利要求 1 所述的方法，还包括复制所述一个或多个而非所有文件的一个或多个引用。

19. 如权利要求 18 所述的方法，其中，可以至少部分地基于所述一个或多个引用，重建所述一个或多个而非所有文件。

20. 如权利要求 1 所述的方法，其中，存储在所述重复删除存储系统中的每个段都是用于复制的候选。

21. 如权利要求 1 所述的方法，还包括以次序处理候选段使得所处理的至少两连续段极接近地存储在所述重复删除存储系统中。

22. 如权利要求 1 所述的方法，还包括基本上以候选段在所述重复删除存储系统中的位置的次序处理所述候选段。

23. 一种播种复制的装置，包括：

用于选择存储在重复删除的存储系统上要复制的一个或多个而非所有文件的部件；

用于确定由所选择的一个或多个而非所有文件所引用的一个或多个段的部件；

用于创建被用于指示至少所述一个或多个段是要被复制的数据结构的部件；以及

用于如果所述数据结构指示了存储在所述重复删除存储系统上的所述一个或多个段中的一个是要复制的则复制所述一个或多个段中的所述一个的部件，

其中，所述数据结构被用于指示是否段被引用以确定该段是否应当被复制。

播种复制

[0001] 本申请要求于2007年4月19日提交的、题目为SEEDINGREPLICATION(播种复制)的美国临时专利申请号60/925,624(代理人文案号DDOMP022+)的优先权,为了所有目的在此将其引入以供参考。

背景技术

[0002] 副本系统(replica system)存储与源系统(originating system)相同的数据或相同的数据的一部分。副本系统可以被用于当在源系统中的数据损坏或丢失时恢复数据。为了高效存储,副本系统和源系统都可以是重复删除系统(deduplicating system),在该重复删除系统中,进入数据被分成段,并且如果该段已经存储在该系统中,则存储对该已存储段的引用,而不是再次存储该段。重复删除典型地在该系统存储数据所需要的空间量上导致了相当大的(例如,10×)减少。

[0003] 当首先开始从一个系统到另一个的复制时,如果副本是存储与源系统相同的数据的全部,那么任务是明确的:从头至尾传输所有的数据。由于仅使得文件重建能够实现的引用和被重复删除的段需要被发送,所以对于重复删除系统来说这是高效的。然而,如果副本是只存储起源者系统(originator system)上数据的一部分,那么所存储的段中的哪些需要被发送给副本就不是显而易见的。一种简单的解决方案是匆匆查阅针对要被存储在副本上的那部分数据的段的引用列表,并且询问副本系统涉及的段是否已经被存储。如果该段还未存储到副本系统上,于是该段才会被传送。然而,对于列表中的每个引用而言这需要来回流量(traffic),以及对于段的每个引用而言这需要由副本系统来检查。通过重复删除,可能有比实际数据段多很多倍的这样的引用。能够在不产生流量和检查段的每个引用的情况下,对存储在重复删除的系统中的部分数据进行种子复制会是有益的。

[0004] 当将存储在一个重复删除的系统上的部分数据一次性拷贝到第二重复删除的系统上时存在相似的情形。正被拷贝的那部分数据所引用的所有段都需要被发送到该第二系统。然而,如上所述,检查每个引用以看看相应的段是否要被发送到第二系统,对于需要检查的每个引用引用而言这会在这两个系统之间产生相当大的流量。能够在不产生流量和检查段的每个引用的情况下将存储在重复删除的系统上的部分数据进行拷贝会是有益的。

附图说明

[0005] 在以下的详细说明和附图中公开了本发明的各种实施例。

[0006] 图1是图示用于播种复制的系统的实施例的框图。

[0007] 图2A是图示一部分数据流或数据块的实施例的框图。

[0008] 图2B是图示对一部分数据流或数据块进行分段的实施例的框图。

[0009] 图3是图示文件引用的实施例的框图。

[0010] 图4A是图示数据结构的实施例的框图。

[0011] 图4B是图示数据结构的实施例的框图。

[0012] 图4C是图示数据结构的实施例的框图。

[0013] 图 5 是图示用于播种复制的过程的实施例的流程图。

[0014] 图 6 是图示用于确定引用的段的过程的实施例的流程图。

[0015] 图 7A 是图示用于向数据结构中进行添加的过程的实施例的流程图。

[0016] 图 7B 是图示用于向数据结构中进行添加的过程的实施例的流程图。

[0017] 图 7C 是图示用于向数据结构中进行添加的过程的实施例的流程图。

[0018] 图 8A 是图示用于对段进行复制的过程的实施例的流程图。

[0019] 图 8B 是图示用于对段进行复制的过程的实施例的流程图。

具体实施方式

[0020] 本发明能够以众多种方式予以实现，包括实现为过程、设备、系统、物质的组成、诸如计算机可读存储介质之类的计算机可读介质或者计算机网络，其中程序指令在光或通信链路上被发送。在此说明书中，这些实现方式，或者本发明可以采取的任何其他形式，可以被称为技术。诸如处理器或存储器之类的被描述为配置成执行任务的元件，包括在给定时间临时被配置为执行该任务的通用元件或者被制造成执行该任务的特定元件这两者。总的来说，在本发明的范围之内，可以改变所公开的过程中步骤的次序。

[0021] 下面连同图示本发明原理的附图提供了对本发明的一个或多个实施例的详细描述。本发明是结合这样的实施例予以描述的，但不限于任何实施例。本发明的范围仅由权利要求限定且本发明包括众多可替换例、修改和等价物。为了提供对本发明的透彻理解，在下面的描述中阐述了众多特定细节。提供这些细节是出于举例的目的，并且本发明在没有这些特定细节中的一些或者全部的情况下可以根据权利要求予以实践。为清楚起见，对在本发明有关技术领域中已知的技术资料未进行详细的描述，以便本发明不会被不必要地混淆。

[0022] 公开了播种复制。在重复删除的系统中，针对源系统上共享公共段池的数据的一部分的播种复制，要求确定使得这部分数据的重建得以实现的所引用的段。可以生成包含关于被引用段的信息的列表或数据结构。对于起源者系统上所存储的唯一段而言，可以使用该列表或数据结构对每个进行检查以看看是否该段已经被该部分数据引用。如果该段已经被引用，则将该段发送到副本系统。这允许源系统和副本系统之间来回的通信被减少到段级而不是引用级。再者，就数据结构来说，对存储器的高效使用是可能的，这是因为向量 (vector) 大大小于唯一的被引用段的列表。另外，在未与该副本系统进行任何通信的情况下在起源者系统上生成该列表或数据结构，从而节省了起源者系统和副本系统之间的通信带宽。

[0023] 在一些实施例中，多个文件将涉及相同的存储段，这是因为在这些不同的文件中的段的内容是相同的。这与在相同系统中存储了相同文件的多个版本的情况（例如，当在一个系统上存储了系统的多个快照时）形成对比，其中该相同文件的多个版本中的每一个涉及相同段。文件可以实现文件系统中的文件、逻辑块设备、虚拟磁带盒、数据库表，或者数据存储系统可以存储的任何其它种类的数据对象或这样的数据对象的一部分。

[0024] 图 1 是图示用于播种复制的系统的实施例的框图。在所示的实例中，存储系统 100 为图 1 中由客户端 112 和客户端 108 所表示的客户端存储数据。客户端除本地存储器外还可以具有本地存储设备。例如，客户端 108 具有存储 110，而客户端 112 不具有本地存储设

备。存储系统 100 或者使用内部存储设备 102 或者使用附连的外部存储设备 104 来存储数据。

[0025] 存储系统 100 经由网络 106 与客户端通信。网络 106 包括下述中的一个或多个：局域网、广域网、有线网络、无线网络、因特网、光纤网络，或能够实现通信的任何其他适当网络。存储系统 100 也经由网络 106 与副本系统 120 通信。副本系统 120 包含内部存储设备 122，并连接到外部存储设备 124。

[0026] 副本系统 120 存储在存储系统 100 上所存储的数据的全部或者一部分。最初，存储系统 100 上所存储的数据的全部或一部分通过经由网络 106 发送适当的数据而被复制到副本系统 120 上。在最初的播种之后，副本系统 120 通过从存储系统 100 发送新的数据而被更新。更新可以是持续的，，当数据量超过一定大小时，或在其他适当的时间时，定期一阵 (in bursts) 发送。就数据拷贝来说，最初的播种完成了该任务并且没有后续更新。

[0027] 在各种实施例中，存储设备 102，104，122 和 124 包括单个存储设备、多个存储设备、诸如图书馆系统或者网络附连的存储系统之类的用于存储的系统或者任何其它适当的存储设备或者系统，所述单个存储设备诸如硬盘、磁带机、半导体存储器，所述多个存储设备诸如冗余阵列系统（例如独立磁盘冗余阵列 (redundant array of independent disk，RAID)）。

[0028] 图 2A 是图示数据流或数据块的一部分的实施例的框图。在所示的实例中，示出了数据流或数据块的一部分 200。数据流或数据块的一部分 200 被诸如图 1 的存储系统 100 之类的存储系统接收并存储在其上。数据流或数据块的一部分 200 是从诸如图 1 的客户端 108 或 112 之类的一个或者多个客户端发送的。

[0029] 图 2B 是图示对数据流或数据块的一部分进行分段的实施例的框图。在所示的实例中，数据流或数据块的该部分被分段成多个分段，在图 2B 中由 202，204 和 206 表示。通过从该数据流或数据块创建可以被用于重建该数据流或数据块的多个段而对该数据流或数据块进行分段。当段被用于重建该数据流或数据块时，段可以是重叠的、非重叠的或者是重叠和非重叠的组合。使用文件边界、目录边界、字节计数、基于内容的边界（例如，当窗口中数据的哈希值等于一值时）或确定边界的任何其他适当方法可以确定段边界。数据块、数据流、文件或目录的重建包括使用原始构成该该数据块、数据流、文件或目录的一个或者多个段的引用。诸如 202、204 和 206 之类的段被存储在存储设备 210 中，例如，像分别存储的段 212、214 和 216。

[0030] 图 3 是图示文件引用的实施例的框图。在一些实施例中，文件引用可被用来从存储在诸如存储设备 210 之类的存储设备中的段重建文件。如所示的实例中，文件 300 可以使用段引用 302、304、306 和 308 予以重建。文件 310 可以使用段引用 312、314、316 和 318 予以重建。存储偏移量以使得在存在重叠或者在文件的重建中仅段引用的一部分被使用的情况下使用段引用能够实现文件的重建。对于数据块、数据流、目录或使得使用所存储的段能够实现它们的重建的任何其他数据单元而言，可以使用相似的引用结构。在各种实施例中，段引用包含基于内容的引用、非基于内容的引用、指纹或段的任何其他适当的引用。

[0031] 图 4A 是图示数据结构的实施例的框图。在一些实施例中，图 4 的数据结构被用于指示是否段被引用以确定该段是否应当被复制。在所示的实例中，数据结构 400 包含由 401、402、404、406、408 和 410 所表示的多个比特位。所述比特位被初始化为“0”。

对于被引用的段，计算该段引用的多个哈希－例如，在图 4A 中，该段引用的多个哈希由 Hash(SegRef)412，414，416 和 418 来表示。哈希的值用于指向数据结构 400 中的比特位。所指向的比特位变为“1”。为了检查段是否被指示要复制，计算对该段的引用的相同哈希，并且如果所述哈希所指向的所有比特位都已经被设为“1”，那么假定该段要被引用并且因此应当被复制。在一些情况下，段可能看起来要被引用，这是因为所有比特位都已经被设为“1”，即使对该段的引用未被指示也是如此。当所指示的其他段的组合碰巧将所有位都设为“1”时，这样的冲突就会出现。选择数据结构的长度和哈希函数的数目来平衡数据结构所需的空间、哈希函数所需的计算量和对于段的数据结构中可容忍的冲突数目。想要的是数据结构 400 能够相当精确地指示给定分段是否已经被引用，以使几乎没有不需要复制的段被复制。在各种实施例中，数据结构 400 包括布隆 (Bloom) 过滤器或类似的基于哈希的数据结构。

[0032] 图 4B 是图示数据结构的实施例的框图。在所示的实例中，数据结构 420 可以用于确定对段的不精确引用的列表。多个缩短的引用在图 4B 中由 422 和 424 表示。缩短的引用 422 是从段引用 430 创建的并置于数据结构 420 中。缩短的引用 424 是从段引用 432 创建的并置于数据结构 420 中。例如，缩短的指纹，诸如 m（例如 24）字节段引用中的 n（例如 4）字节，可被用作不精确引用，其中 m 小于 n。具有与该数据结构的任何缩短的指纹的 n 个字节相匹配的具有引用的段被加入到要从起源者系统发送到副本系统的段列表中。n 字节匹配的使用确保了必要的段被复制。可能存在一些附加段也被传输。另一方面，与全部（例如 24 字节）比较相对照，使用缩短的比较（例如 4 字节）减少了存储器空间。

[0033] 图 4C 是图示数据结构的实施例的框图。在所示的实例中，数据结构 440 可以用于确定对段的精确引用的列表。在图 4C 中由 442 和 444 来表示多个引用。引用 442 从段引用 450 创建，并被置于数据结构 440 中。引用 444 从段引用 452 创建并置于数据结构 440 中。精确的段引用的数据结构可被用于在副本中进行播种复制的过程中。

[0034] 图 5 是图示用于播种复制的过程的实施例的流程图。在一些实施例中，图 5 的过程被用于对诸如图 1 的副本系统 120 之类的副本系统进行播种。在一些实施例中，图 5 的过程被用于从一个系统向另一系统拷贝数据集。在所示的示例中，在 500 中，选择要复制的（一个或者多个）文件。在各种实施例中，通过下述来选择（一个或者多个）文件：使用图形用户界面中对一个或多个文件的选择、使用文件列表、通过在窗口中点击、通过选择目录、通过选择多个目录、使用目录列表或者选择文件的任何其他适当方法。在各种实施例中，为了负载平衡、满足服务目标的质量或其他原因可以由系统自动确定（一个或者多个）文件。在各种实施例中，要传输的文件可以实现逻辑块设备、虚拟磁带机或其他类型的数据集或这样的数据集的各部分。在 502 中，对所选文件的文件元数据和段引用进行复制，并且将段引用加入到用于指示至少所确定段要被复制的数据结构中。在一些实施例中，比那些用于（一个或多个）所选定文件的重建所必须的段更多的段被指示为要复制。在 506 中，选择存储在重复删除系统上的第一候选段。在 508 中，确定数据结构是否指示所选候选段要被复制。如果所选候选段要被复制，则在 510 中，复制所选候选段并且控制行进到 512，如果不是，则控制行进到 512。在 512 中，确定是否存在更多的候选段。如果存在，则在 514 中，选择下一个候选段，并且控制行进到 508。在一些实施例中，为了高效访问，当可能时，在 514 中选择的下一个候选段是存储（例如，在重复删除存储系统中）的与先前段极接近

的一个段。在 512 中，如果没有更多的段了，则过程结束。在一些实施例中，存储在重复删除存储系统中的每个段都是用于复制的候选段。在一些实施例中，候选段基本上以他们在重复删除存储系统中所处位置的次序被处理的。这可以通过减少访问段所花费的时间来获得效率（例如，成批地存取段或在下一个段存取之前不要求大量探求 (seek)）。

[0035] 图 6 是图示用于确定被引用段的过程的实施例的流程图。在一些实施例中，图 6 的过程被用于实现图 5 中的 502。在所示的实例中，在 600 中，选择第一文件。在 601 中，复制所选文件的元数据（例如，文件名，文件大小，文件创建日期等）。在 602 中，选择所选文件的第一引用。在 604 中，复制该引用。在 606 中，将段引用添加到用于指示至少确定的段要复制的数据结构中。在 608 中，确定是否存在更多的引用，如果有，则在 610 中，选择下一个引用并将控制行进到 604，如果没有，则在 612 中，确定是否还有更多文件。如果有，则在 614，选择下一个文件并将控制行进到 602。如果没有则过程结束。

[0036] 图 7A 是图示用于向数据结构中进行添加的过程的实施例的流程图。在一些实施例中，图 7A 的过程被用于实现图 6 中的 606。在所示的实例中，在 700 中，确定是否所选引用所涉及的段已经在精确列表中。如果已经存在，则过程结束。如果还没有，则在 702 中，将该段添加到该精确列表中。该精确列表包含对段的明确引用的列表。例如，段指纹或其他类型的引用的列表。在各种实施例中，以下述方式生成该列表：消除或不消除在该列表中出现的重复段，和 / 或在当在起源者存储系统上存储原文件时的时间或以后需要的时间。在一些实施例中，该列表是按段引用来分类的。在一些实施例中，该列表是按存储位置来分类的。

[0037] 图 7B 是图示用于向数据结构进行添加的过程的实施例的流程图。在一些实施例中，图 7B 的过程被用于实现图 6 的 606。在所示的实例中，在 710 中，计算所选引用的多个哈希。在 712 中，如果适当，使用多个哈希中的每一个来改变数据结构。如果该位已经被改变（例如，设为“1”），则不改变该位。在一些实施例中，哈希值被用于确定数据结构中的比特位设为值“1”。数据结构可用于确定段是否被引用。

[0038] 图 7C 是图示用于向数据结构进行添加的过程的实施例的流程图。在一些实施例中，图 7C 的过程被用于实现图 6 的 606。在所示的实例中，在 720 中，创建缩短的引用。在 722 中，将该缩短的引用添加到该数据结构中。在一些实施例中，该数据结构是缩短的引用的长列表。该数据结构可用于确定段是否可能被引用了。

[0039] 在一些实施例中，该数据结构包括如图 6 的过程所生成的要复制的段的列表。

[0040] 图 8A 是图示用于复制段的过程的实施例的流程图。在一些实施例中，图 8A 的过程被用于实现图 5 的 510。在所示的实例中，在 800 中，向副本系统发送段。在这个例子中，操作假设是该副本系统还未具有该段的拷贝，和该段成功到达该副本系统。在一些实施例中，接收到确认，该确认指示已经收到该段和 / 或已经将该段存储在该副本系统上（未示出）。

[0041] 图 8B 是图示用于复制段的过程的实施例的流程图。在一些实施例中，图 8B 的过程被用于实现图 5 的 510。在所示的实例中，在 810 中，向副本系统发送段标识器。在 812 中，接收该段是否被存储在副本系统上的指示。在 814 中，确定是否该指示指示了该段被存储在该副本系统上。如果是，则过程结束。如果不是，则在 816 中，将该段发送给该副本系统。在一些实施例中，接收到确认，其指示该段已经被接收和 / 或已经存储在副本系统上（未示出）。

[0042] 虽然为了清楚理解的目的，前述实施例已经描述了一些细节，但本发明不限于所提供的细节。存在很多实现本发明的可替换方法。所公开的实施例是举例说明性的而非限制性的。

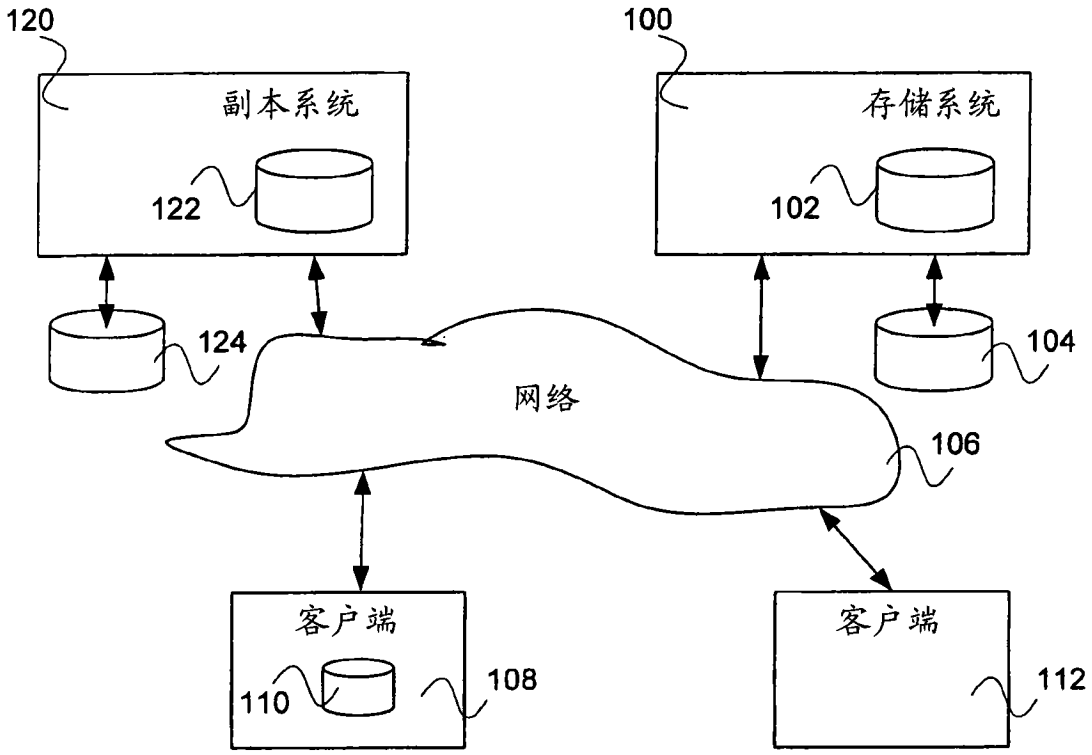


图 1

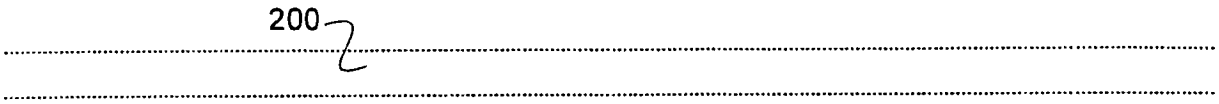


图 2A

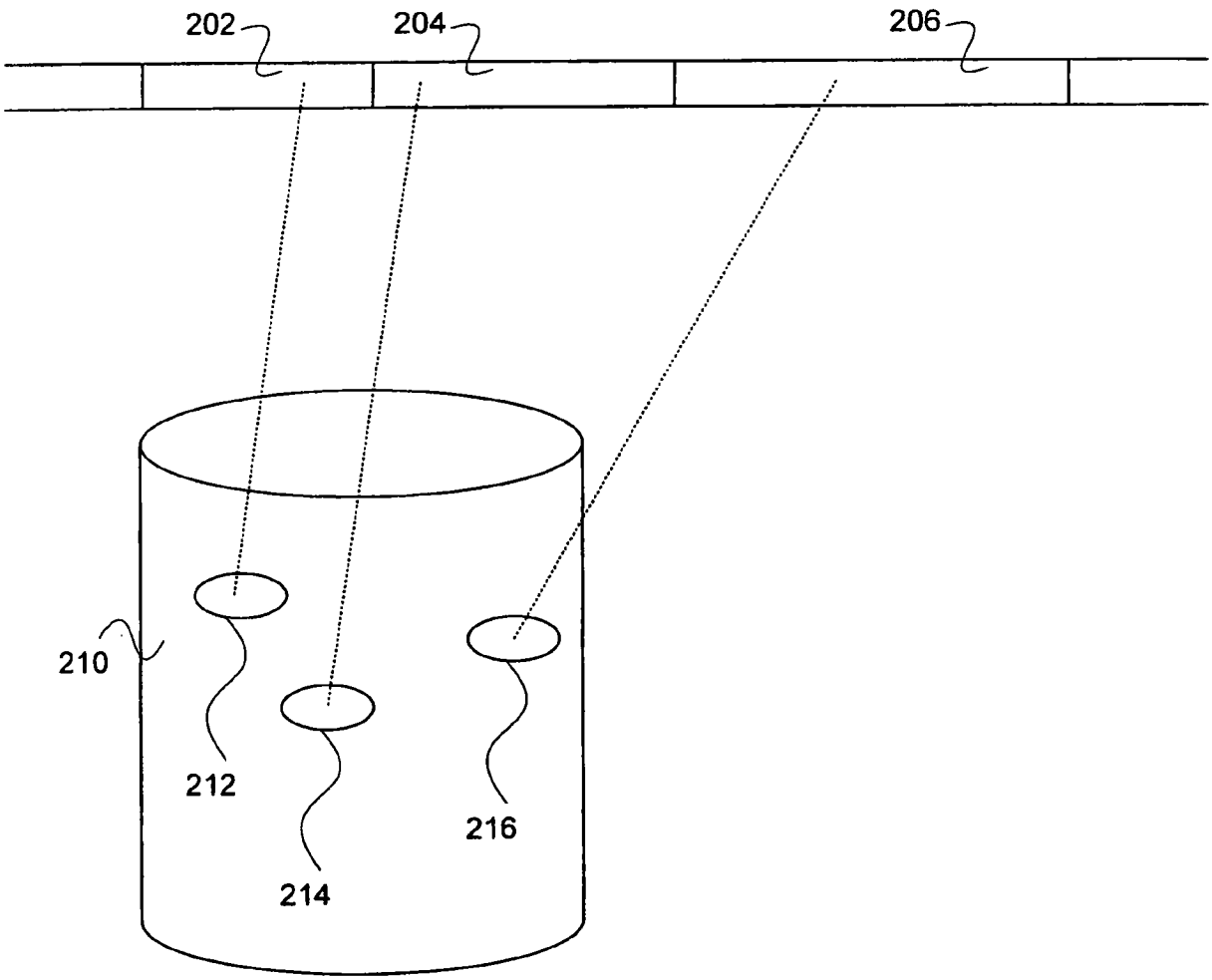

202
204
206
210
212
214
216

图 2B

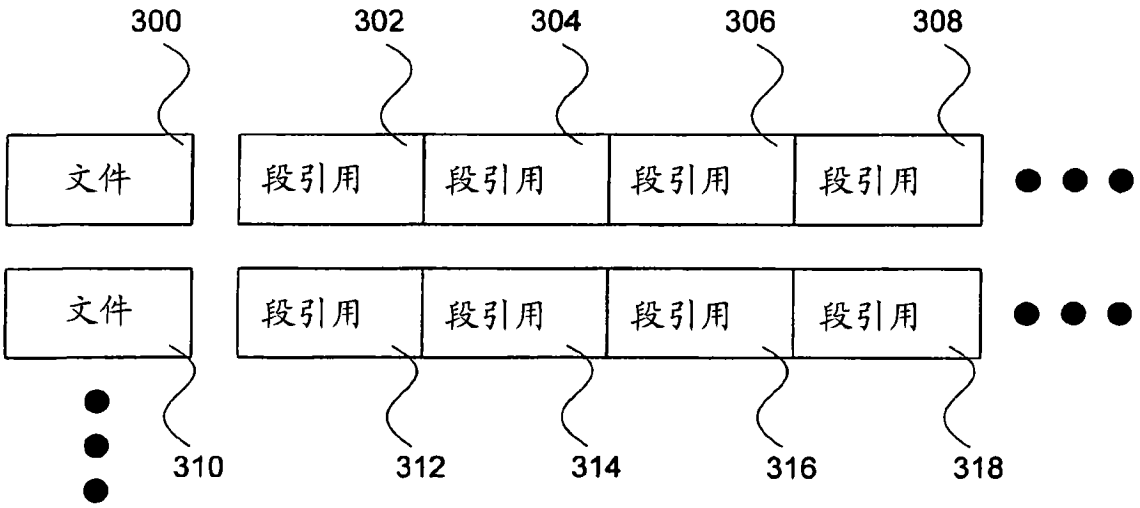

300
302
304
306
308
文件
段引用
段引用
段引用
段引用
文件
段引用
段引用
段引用
段引用
310
312
314
316
318

图 3

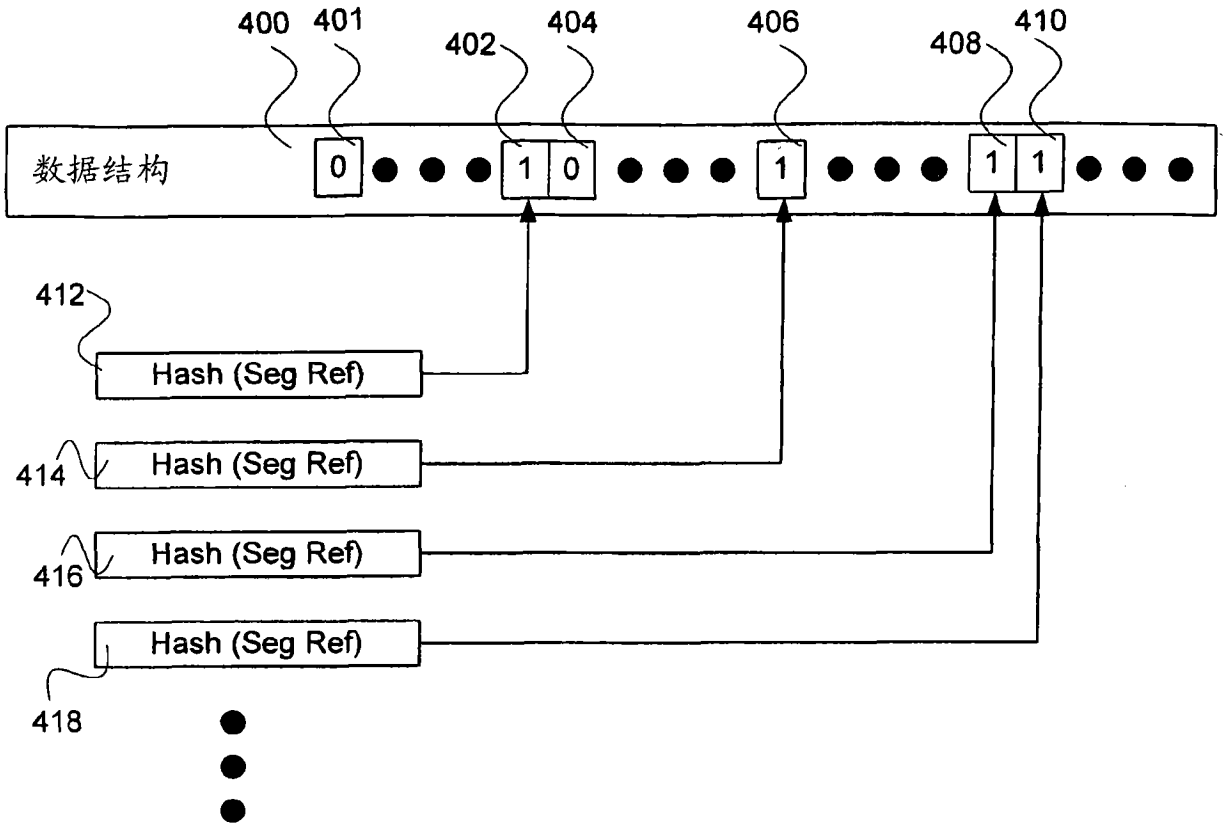

400
401
402
404
406
408
410
数据结构
0
1
0
1
1
1
412
Hash (Seg Ref)
414
Hash (Seg Ref)
416
Hash (Seg Ref)
Hash (Seg Ref)
418

图 4A

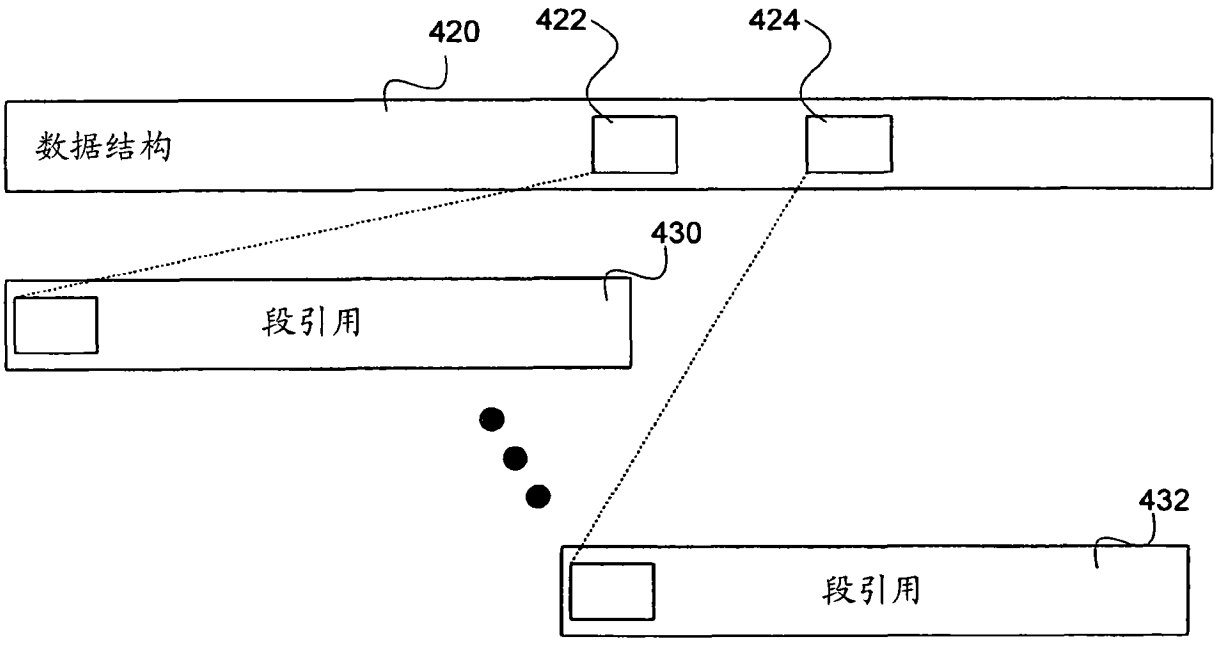

420
422
424
数据结构
430
段引用
432
段引用

图 4B

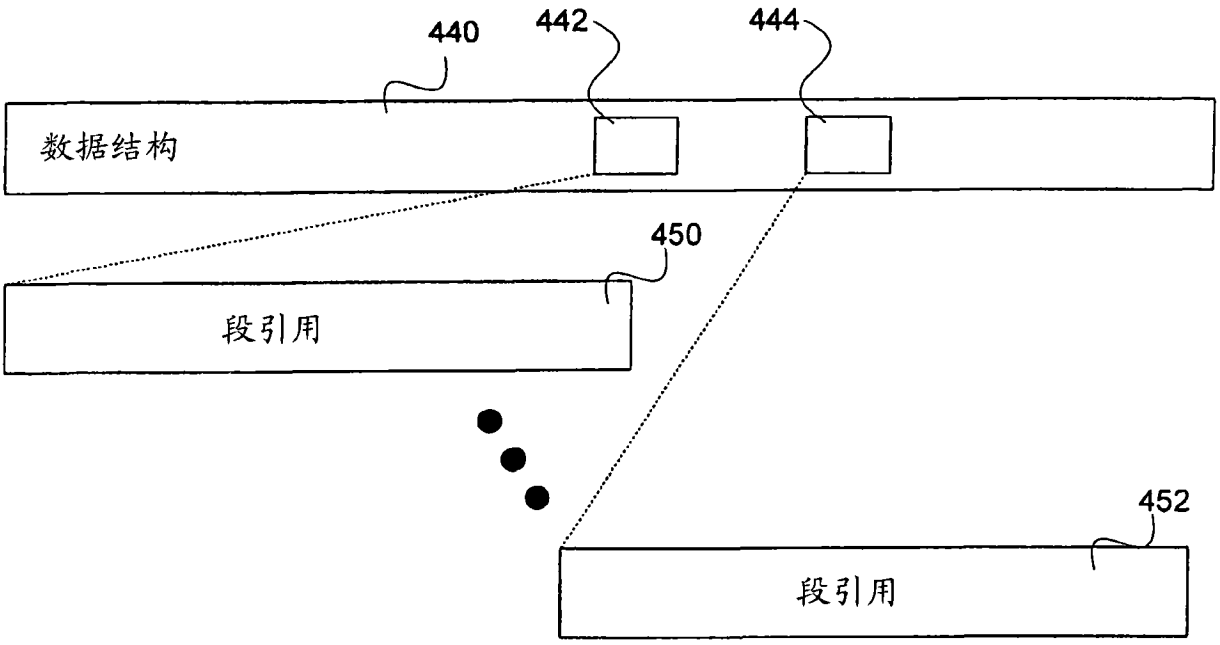


图 4C

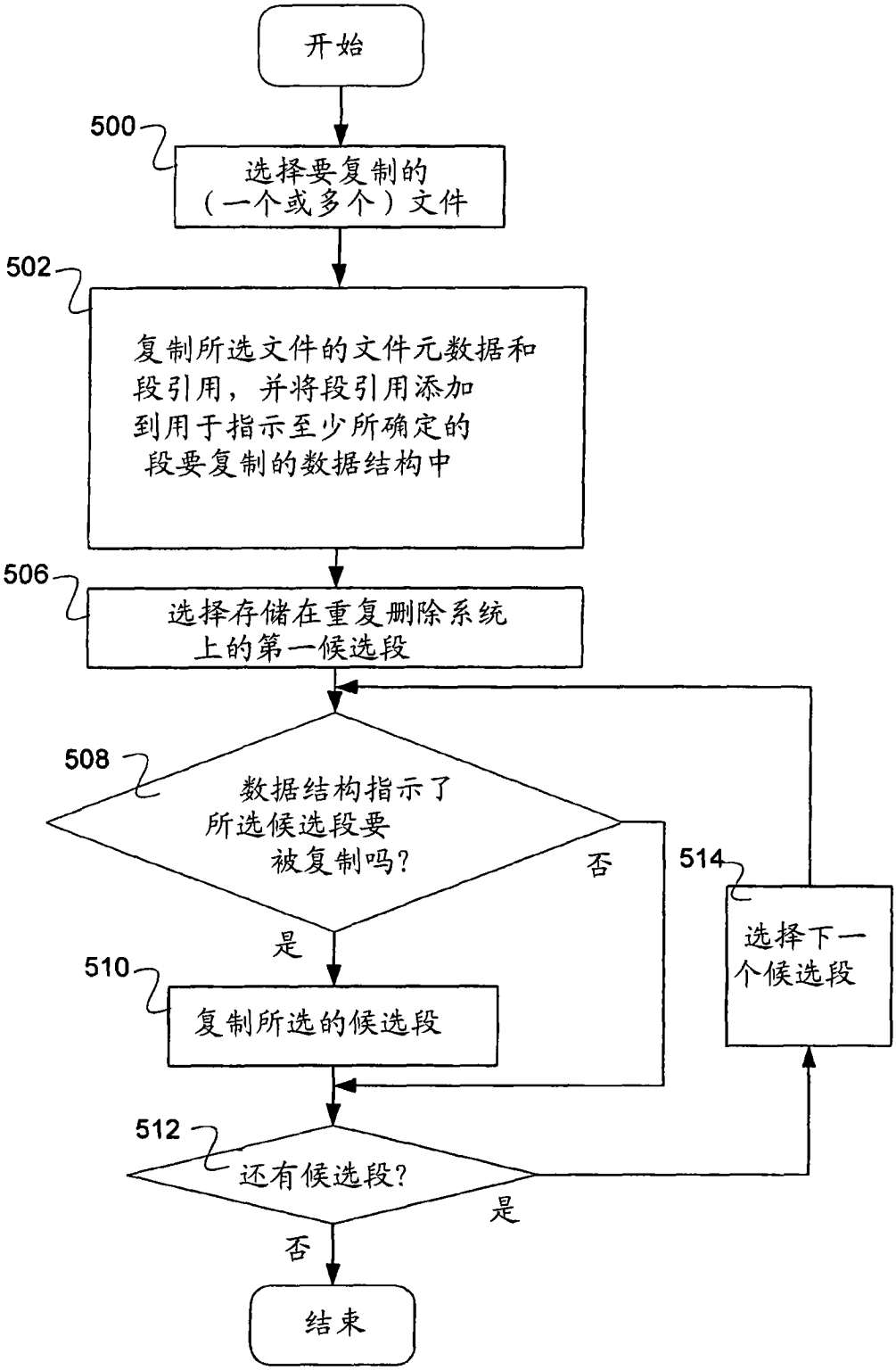

开始
500
选择要复制的（一个或多个）文件
502
复制所选文件的文件元数据和段引用，并将段引用添加到用于指示至少所确定的段要复制的数据结构中
506
选择存储在重复删除系统上的第一候选段
508
数据结构指示了所选候选段要被复制吗？
否
是
514
选择下一个候选段
510
复制所选的候选段
512
还有候选段？
是
否
结束

图 5

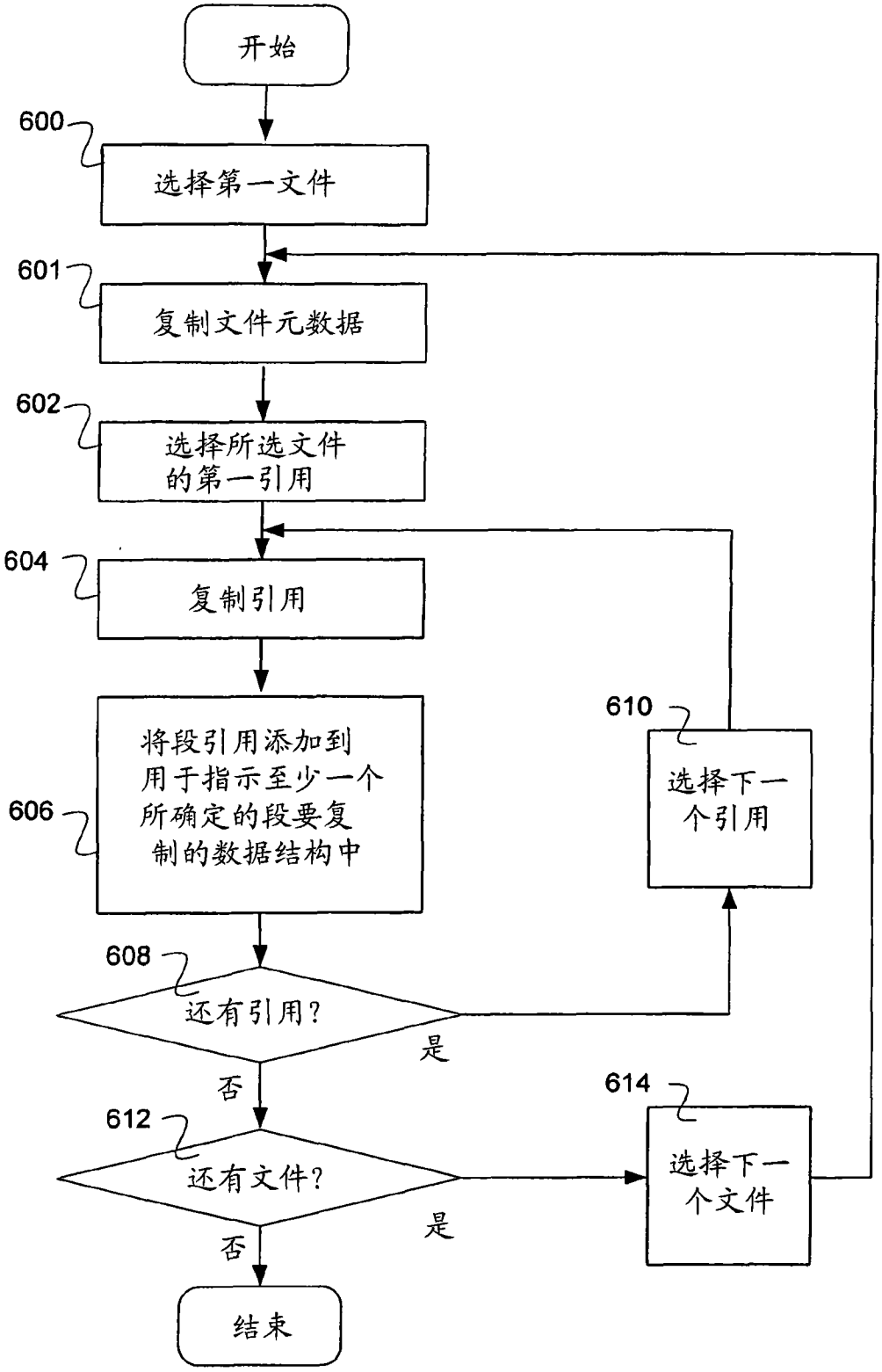

开始
600
选择第一文件
601
复制文件元数据
602
选择所选文件的第一引用
604
复制引用
606
将段引用添加到用于指示至少一个所确定的段要复制的数据结构中
608
还有引用?
是
否
610
选择下一个引用
612
还有文件?
是
否
614
选择下一个文件
结束

图 6

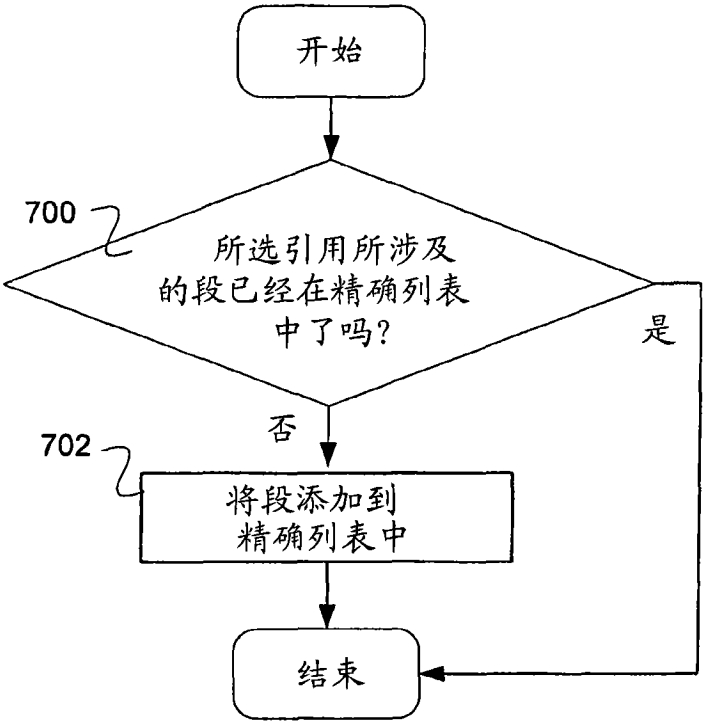


图 7A

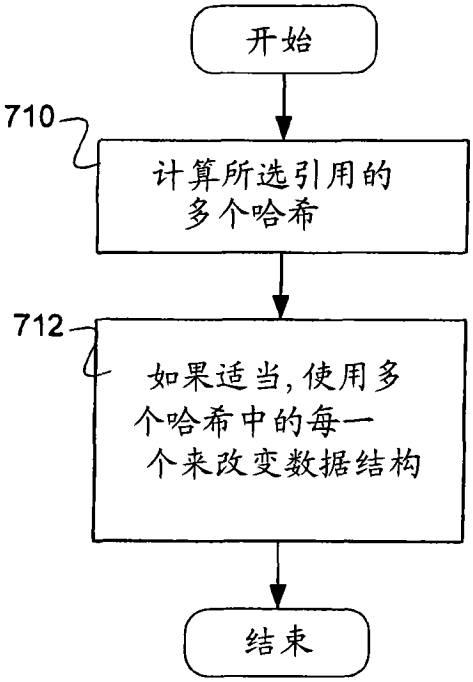


图 7B

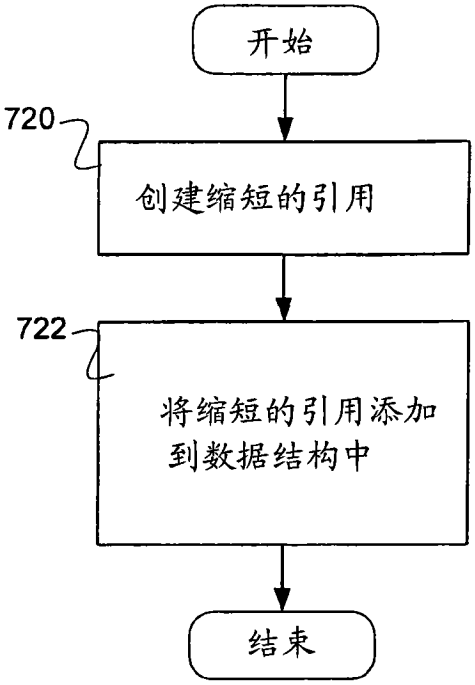


图 7C

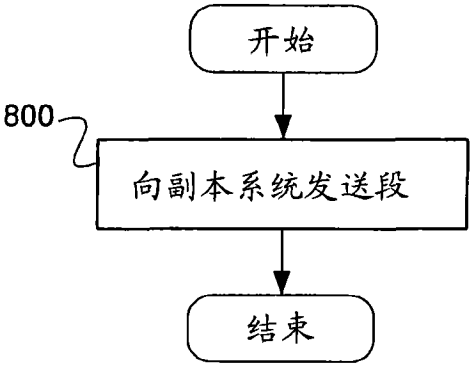


图 8A

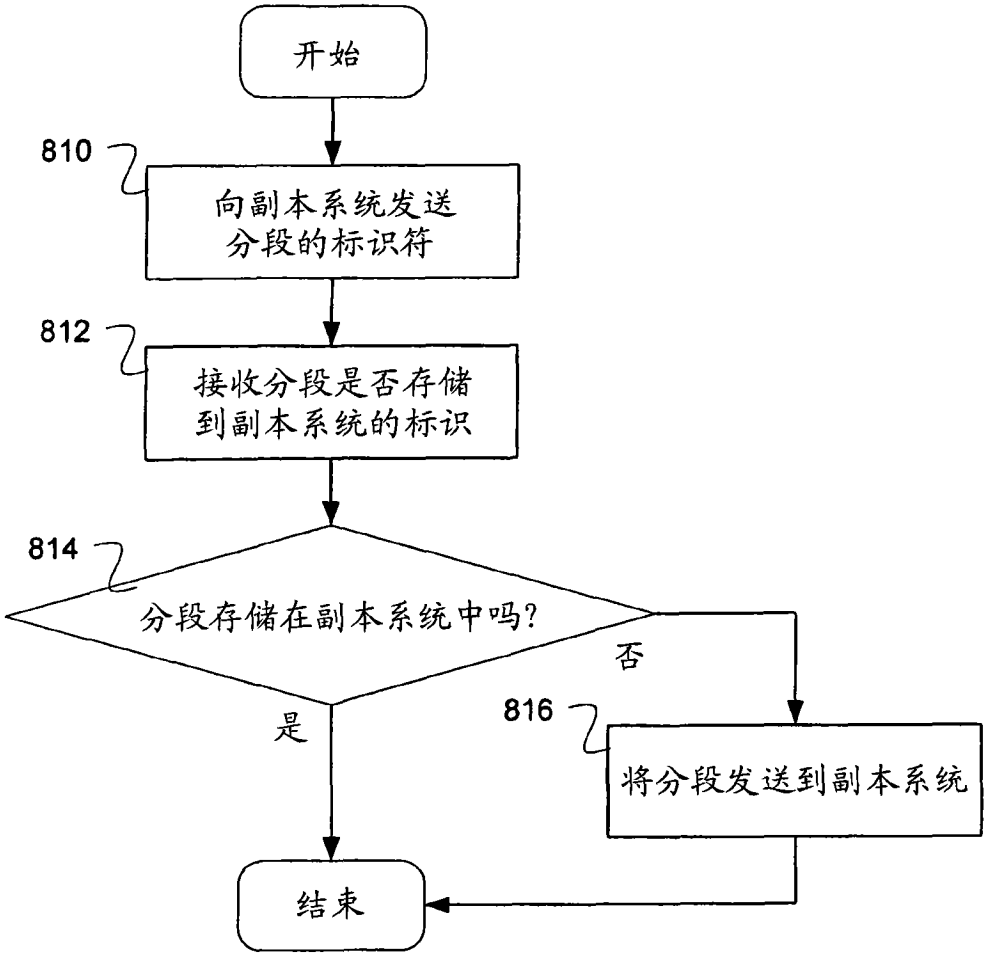

开始
810
向副本系统发送分段的标识符
812
接收分段是否存储到副本系统的标识
814
分段存储在副本系统中吗?
否
是
816
将分段发送到副本系统
结束

图 8B