[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 12/00 (2006.01)

[12] 发明专利说明书

专利号 ZL 03802284.2

[45] 授权公告日 2009 年 12 月 23 日

[11] 授权公告号 CN 100573471C

[22] 申请日 2003.1.15 [21] 申请号 03802284.2

[30] 优先权

[32] 2002. 1.17 [33] US [31] 10/051,999

[86] 国际申请 PCT/US2003/001194 2003.1.15

[87] 国际公布 WO2003/062996 英 2003.7.31

[85] 进入国家阶段日期 2004.7.15

[73] 专利权人 汤姆森特许公司

地址 法国布洛涅

[72] 发明人 马克·A·舒尔茨 林 书

迈克尔·G·凯利

[56] 参考文献

US6292880B1 2001.9.18

WO0161563A1 2001.8.23

WO9632685 1996.10.17

CN1295327A 2001.5.16

CN1308438A 2001.8.15

US20010037323A1 2001.11.1

审查员 刘 琳

[74] 专利代理机构 北京市柳沈律师事务所

代理人 吕晓章 马 莹

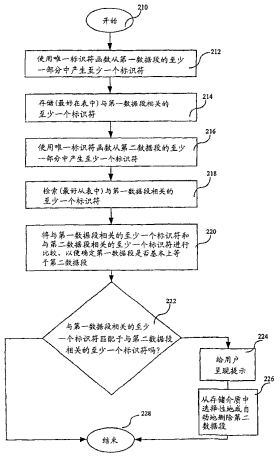
权利要求书 3 页 说明书 6 页 附图 2 页

[54] 发明名称

搜索重复数据的系统和方法

[57] 摘要

本发明涉及一种搜索重复数据的方法(200)和系统(100)。所述方法(200)包括步骤：使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符(212)；使用唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符(216)；以及将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段(220)。



1. 一种搜索重复数据的方法，包括下列步骤：

使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符；

当将第二数据段存储到存储介质中时，使用所述唯一标识符函数从所述第二数据段的至少一部分中产生至少一个标识符；和

将与所述第一数据段相关的至少一个标识符和与所述第二数据段相关的至少一个标识符进行比较，以便确定所述第一数据段是否等于所述第二数据段，

其中所述比较步骤发生在对所述数据段的重复作出任何判定之前。

2. 根据权利要求 1 所述的方法，其中从第一数据段的至少一部分中产生至少一个标识符的步骤包括步骤：当所述第一数据段被记录到存储介质时，使用所述唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符。

3. 根据权利要求 2 所述的方法，其中从所述第二数据段的至少一部分中产生至少一个标识符的步骤包括步骤：当所述第二数据段被记录到不同的存储介质时，使用所述唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符。

4. 根据权利要求 1 所述的方法，其中所述从第一数据段的至少一部分中产生至少一个标识符的步骤发生在所述第一数据段被记录到存储介质之后。

5. 根据权利要求 1 所述的方法，其中所述第一数据段和所述第二数据段是多媒体数据段。

6. 根据权利要求 1 所述的方法，还包括步骤：

在表中存储与所述第一数据段相关的至少一个标识符；和

在所述比较步骤之前，从所述表中检索与所述第一数据段相关的至少一个标识符。

7. 根据权利要求 1 所述的方法，还包括步骤：当与所述第一数据段相关的至少一个标识符匹配于与所述第二数据段相关的至少一个标识符时，给出所述第一数据段基本上等于所述第二数据段的提示。

8. 根据权利要求 1 所述的方法，其中所述第一数据段的至少一部分和所

述第二数据段的至少一部分的大小是基于时间度量的，其中所述第一数据段的至少一部分与所述第二数据段的至少一部分时间对应。

9. 根据权利要求 1 所述的方法，其中所述第一数据段的至少一部分和所述第二数据段的至少一部分的大小是基于位度量的，其中所述第一数据段的至少一部分与所述第二数据段的至少一部分逐位对应。

10. 根据权利要求 1 所述的方法，其中与所述第一数据段相关的至少一个标识符和与所述第二数据段相关的至少一个标识符是散列值，并且所述唯一标识符函数是散列函数，其中当所述第一数据段与所述第二数据段相等时，与所述第一数据段相关的散列值将等于与所述第二数据段相关的散列值。

11. 根据权利要求 1 所述的方法，其中所述比较步骤包括步骤：将与所述第一数据段相关的多个标识符和与所述第二数据段相关的多个标识符进行比较，以便确定所述第一数据段是否基本上等于所述第二数据段。

12. 根据权利要求 1 所述的方法，其中所述比较步骤包括步骤：将与第一组数据段相关的多个标识符和与第二组数据段相关的多个标识符进行比较，以便确定所述第一组数据段是否基本上等于所述第二组数据段。

13. 一种用于搜索重复数据的系统，包括：

控制器，用于从存储介质读取数据，以及将数据写入存储介质；和

连接到所述控制器的处理器，其中所述处理器被编程来：

使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符；

当将第二数据段存储到存储介质中时，使用所述唯一标识符函数从所述第二数据段的至少一部分中产生至少一个标识符；和

将与所述第一数据段相关的至少一个标识符和与所述第二数据段相关的至少一个标识符进行比较，以便确定所述第一数据段是否等于所述第二数据段，

其中所述处理器在对所述数据段的重复作出任何判定之前比较所述标识符。

14. 根据权利要求 13 所述的系统，其中所述处理器还被编程来当所述第一数据段被记录到所述存储介质时，使用所述唯一标识符函数从所述第一数据段的至少一部分中产生至少一个标识符。

15. 根据权利要求 14 所述的系统，其中所述处理器还被编程来当所述第

二数据段被记录到不同的存储介质时，使用所述唯一标识符函数从所述第二数据段的至少一部分中产生至少一个标识符。

16. 根据权利要求 13 所述的系统，其中所述处理器还被编程来在所述第一数据段被记录到所述存储介质后，使用所述唯一标识符函数从所述第一数据段的至少一部分中产生至少一个标识符。

17. 根据权利要求 13 所述的系统，其中所述第一数据段和所述第二数据段是多媒体数据段。

18. 根据权利要求 13 所述的系统，还包括一个表，其中所述处理器还被编程来:

在所述表中存储与所述第一数据段相关的至少一个标识符；和

在所述比较之前，从所述表中检索与所述第一数据段相关的至少一个标识符。

19. 根据权利要求 13 所述的系统，其中所述处理器还被编程来当与所述第一数据段相关的至少一个标识符匹配于与所述第二数据段相关的至少一个标识符时，给出所述第一数据段基本上等于所述第二数据段的指示。

20. 根据权利要求 13 所述的系统，其中与所述第一数据段相关的至少一个标识符和与所述第二数据段相关的至少一个标识符是散列值，并且所述唯一标识符函数是散列函数，其中所述处理器确定与所述第一数据段相关的散列值是否等于与所述第二数据段相关的散列值，表示所述第一数据段与所述第二数据段基本相等。

搜索重复数据的系统和方法

技术领域

本发明的结构通常涉及记录系统，尤其涉及多媒体记录系统，用于将数字编码的信号记录到诸如硬盘驱动器和可记录光盘的盘介质上。

背景技术

目前，可以将多种形式的数据记录到许多不同类型的存储介质。例如，许多消费者将电视节目或音乐记录到光盘介质或硬盘驱动器(HDD)。随着技术的改进，光盘介质和 HDD 的存储容量已经明显增加。事实上，一些 HDD 能够存储多于 50 千兆字节的数据。同样的，消费者可以在这种类型的存储介质上记录大量的节目或歌曲。

当将数据记录到可记录存储介质时，该可记录存储介质装置通常允许用户输入用于识别所记录的内容的标题。当用户希望定位已记录数据的特定块以确定用户先前所记录的那些数据时，这些标题是有用的。然而，值得注意的是，这一搜索处理可能费力、效率较低并且易于出错，因为存储介质可以包含几百甚至上千个标题。如果存储介质是大型 HDD 或者如果某些数据段的标题被给定缺省标题，则这一问题可能尤为尖锐。

甚至认为通过搜索标题能够相对容易地定位存储介质上的数据段，特定标题对于不同的数据段可能是相同的。例如，如果歌曲被记录在存储介质上并根据歌名给定标题，则后来可以记录第二首歌曲，该第二首歌曲具有与第一首歌曲相同的名字。这种冲突是可能发生的，例如，如果两位独立的艺术家记录相同歌曲的不同版本。当记录第二首歌曲时，用户可能检查先前记录的歌曲的标题，并可能错误地认为已经记录了该第二首歌曲。因此，需要一种搜索重复数据的系统和方法，而不增加系统成本或复杂性，并且还减少当搜索和考虑删除重复数据时出错的可能性。

发明内容

本发明涉及一种搜索重复数据的方法。所述方法包括步骤：使用唯一标

识符函数从第一数据段的至少一部分中产生至少一个标识符；使用唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符；和将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段，其中比较步骤发生在对所述数据段的重复作出任何判定之前。

在一种结构中，所述从第一数据段的至少一部分中产生至少一个标识符的步骤包括步骤：当第一数据段被记录到存储介质时或者在第一数据段被记录到存储介质之后，使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符。另外，所述从第二数据段的至少一部分中产生至少一个标识符的步骤包括步骤：当第二数据段被记录到存储介质时，使用唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符。而且，当第二数据段被记录到不同的存储介质时，可能发生从第二数据段的至少一部分中产生至少一个标识符的步骤。

在一方面，第一数据段和第二数据段可以是多媒体数据段。所述方法还可以包括步骤：在表中存储与第一数据段相关的至少一个标识符；和在所述比较步骤之前，从所述表中检索与第一数据段相关的至少一个标识符。另外，所述方法可以包括步骤：当与第一数据段相关的至少一个标识符匹配于与第二数据段相关的至少一个标识符时，给出第一数据段基本上等于第二数据段的提示。

在另一种结构中，第一数据段的至少一部分和第二数据段的至少一部分的大小可以基于时间度量（temporal measurement）或位度量。第一数据段的至少一部分可以与第二数据段的至少一部分时间对应或者逐位对应。在另一方面，与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符可以是散列值，并且唯一标识符函数可以是散列函数，其中当第一数据段与第二数据段相等时，与第一数据段相关的散列值将等于与第二数据段相关的散列值。

而且，所述比较步骤包括步骤：将与第一数据段相关的多个标识符和与第二数据段相关的多个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段。而且，所述比较步骤包括步骤：将与第一组数据段相关的多个标识符和与第二组数据段相关的多个标识符进行比较，以便确定第一组数据段是否基本上等于第二组数据段。

本发明也涉及一种用于搜索重复数据的系统。所述系统包括：控制器，用于从存储介质读取数据，以及将数据写入存储介质；和处理器，其中所述处理器被编程来使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符；使用唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符；和将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段，其中比较步骤发生在对所述数据段的重复作出任何判定之前。所述系统也包括用于实现如上所述方法的合适的软件和电路。

附图说明

图 1 是根据此处本发明结构的能够搜索重复数据的系统的方框图。

图 2 是图解说明根据本发明结构的用于搜索重复数据的操作的流程图。

具体实施方式

在图 1 中以方框图形式示出了根据本发明结构的系统 100，该系统 100 执行各种先进的操作特征。然而，本发明不限于图 1 所图解的特定系统，因为可以用能够接收数字编码信号的任意其他系统来实现本发明。另外，系统 100 不限于从任何特定类型的存储介质读取数据或将数据写入其中，因为能够存储数字编码数据的任意存储介质可以与系统 100 使用。

系统 100 可包括控制器 110，用于从存储介质 112 读取数据和将数据写入其中。控制器也可以从不同的存储介质或存储器 120 读取数据和将数据写入其中。系统 100 也可以包括微处理器 114、表或存储器 116、以及显示器 118。也可以提供控制和数据接口，用以控制控制器 110 和显示器 118 的操作和检索表 116 中存储的信息。在存储器中可以提供合适的软件和固件用以由微处理器 114 执行的常规操作。而且，根据本发明的结构，可以为微处理器 114 提供程序的例行程序。另外，在微处理器 114 中可以使用任何其他合适的软件或电路。

在操作中，控制器 110 可以将第一数据段写入存储介质 112。在一种结构中，当将第一数据段记录到存储介质 112 时，微处理器 114 能够使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符。一旦微处理器 114 从第一数据段的至少一部分中产生至少一个标识符，微处理器 114 能够将

该至少一个标识符发送至表 116。在另一种结构中，在已经将第一数据段记录到存储介质 112 之后，任何时间可以产生与第一数据段相关的至少一个标识符。

微控制器 114 也可以使用唯一标识符函数从第二数据段的至少一个对应部分中产生至少一个标识符。当第二数据段被记录到存储介质 112 时，或者当第二数据段被记录在存储器 120 中时，微处理器 114 可以产生与第二数据段相关的至少一个标识符。应当理解，存储器 120 可以是用于存储数字编码数据的任意适当形式的存储器。

一旦产生，微处理器 114 就能够从表 116 中检索与第一数据段相关的至少一个标识符。然后微处理器 114 可以将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段。如果与第一数据段相关的至少一个标识符匹配于与第二数据段相关的至少一个标识符，则第一数据段基本上相等，即使不完全相等第二数据段。随后微处理器 114 能够通过显示器 118 给用户提示两个数据段相等。下面将更详细地描述本发明的整个操作。

重复数据的搜索

图 2 图解说明了用于论证搜索重复或相同数据的操作的流程图 200。在步骤 210，该处理开始。如在步骤 212 所示，使用唯一标识符函数可以从第一数据段的至少一部分中产生至少一个标识符。当第一数据段被记录到存储介质时，可以产生该标识符。相反，在第一数据段被记录到存储介质之后的任何时刻可以产生标识符。

第一数据段可以是包括基于文本的数据、音频、视频或者它们的任意组合的任何适当类型的数据，或者任何其他适当形式的数据。第一数据段也可以是加密或未加密的数据段。而且，可以从第一数据段的任一部分中产生标识符，该任一部分包括第一数据段的不相邻或不连续部分。而且，从第一数据段中包含的数据的任一部分中可以产生多于一个标识符。从其中产生标识符的数据的部分的大小可以基于时间度量或者位度量。

例如，如果第一数据段是一首歌，可以从这一整首歌中产生标识符，从而至少一部分包括整个第一数据段。再例如，这首歌可以被划分成两个分离的部分：开始部分和结束部分。如果这首歌的这两部分的大小都基于时间度量，则开始部分可以包括这首歌的头 30 秒，结束部分可以包括这首歌的最后

30 秒。根据本发明的结构，可以组合这首歌的这两部分，并且从这种组合中可以产生至少一个标识符。因此，可以使用每个数据段的一个或多个标识符来比较与另一数据段相关的对应数量的标识符。

继续示例，可以从这首歌的两部分中产生标识符，从而从相同的歌曲中产生两个独立的标识符。可选地，可以从开始和结束部分之间的时间度量中产生标识符。而且，如果数据的至少一部分的大小基于位度量，则例如可以从这首歌的第一 1Mb 数据中产生标识符。然而，应当注意，本发明不限于上述示例，因为从任何适当类型数据的第一段(包括当至少一部分包括整个第一数据段)的任意数量的部分中可以产生任意数量的标识符。

返回参考流程图 200，最好在表中存储与第一数据段相关的至少一个标识符，如在步骤 214 所示。在步骤 216，使用唯一标识符函数可以从第二数据段的至少一部分中产生至少一个标识符。与第二数据段相关的至少一个标识符的产生可以根据关于第一数据段讨论的处理(请参阅步骤 212 论述)。然而，为了增加精确度，产生至少一个标识符的第二数据段的一(多个)部分可以对应于第一数据段的至少一部分。这种对应可以是基于时间的或者是基于逐位的。

例如，如果第一数据段是一首歌并且从该整首歌中产生与该第一数据段相关的至少一个标识符(该至少一部分包括整个第一数据段)，随后最大化精确度，从这整首歌中可以产生与第二数据段相关的至少一个标识符(假设第二数据段实际上是一首歌)。再例如，如果第一数据段的至少一部分包括第一 1Mb 数据并且从该部分中产生与第一数据段相关的至少一个标识符，则最好从第二数据段中的第一 1Mb 的数据中产生与第二数据段相关的至少一个标识符。

在一种结构中，当第二数据段被记录到与第一数据段记录到的相同存储介质时，可以产生与第二数据段有关的至少一个标识符。相反，当第二数据段被记录到不同的存储介质时，可以产生与第二数据段相关的至少一个标识符。

在步骤 218，一旦从第二数据段产生一个或多个适当的标识符，就从存储器中，最好是从表中检索与第一数据段相关的至少一个标识符。在步骤 220，可以将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段。如果标识符相同，则第一数据段实际上总是等于第二数据段。在判定块 222，当与

第一数据段相关的至少一个标识符匹配于与第二数据段相关的至少一个标识符时，则可以给用户呈现第一数据段基本上等于第二数据段的提示，如在步骤 224 所示。而且，在步骤 226，用户可以选择地或者自动地删除为了比较步骤而被记录到存储介质的第二数据段的任意部分。该处理在步骤 228 结束。

在一种结构中，至少一个标识符可以是散列值。此外，唯一标识符函数可以是散列函数。当第一数据段和第二数据段相等或基本上相等时，与第一数据段相关的散列值可以等于与第二数据段相关的散列值。可用来实现本发明的几个散列函数的一个示例是异或函数。然而，应当理解，本发明不限于这种特殊散列函数，因为可以使用任何其他合适的散列函数。

虽然结合这里所公开的实施例描述了本发明，但是应当理解，上面的描述旨在举例说明而不用于限定由权利要求所定义的本发明的范围。

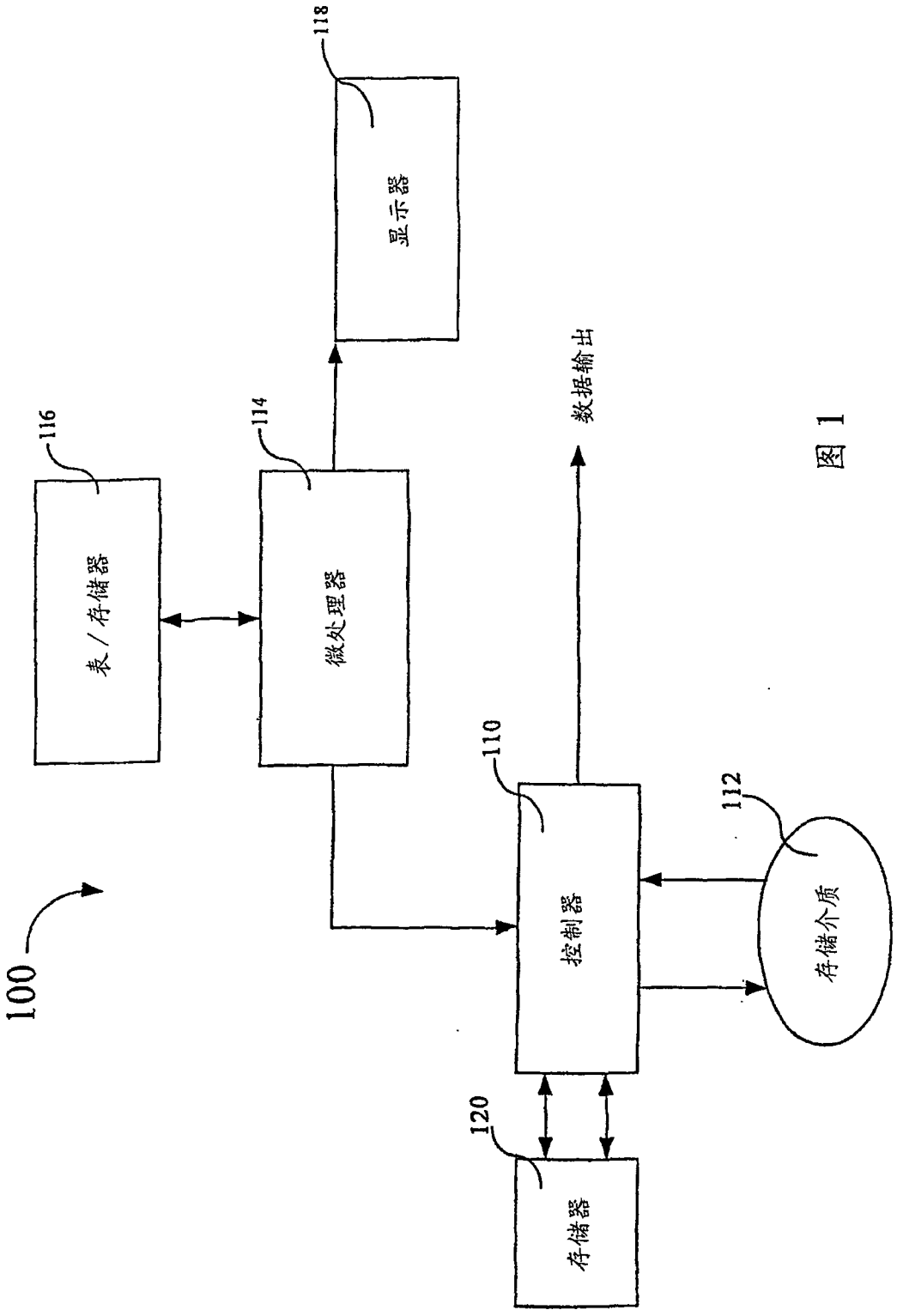


图 1

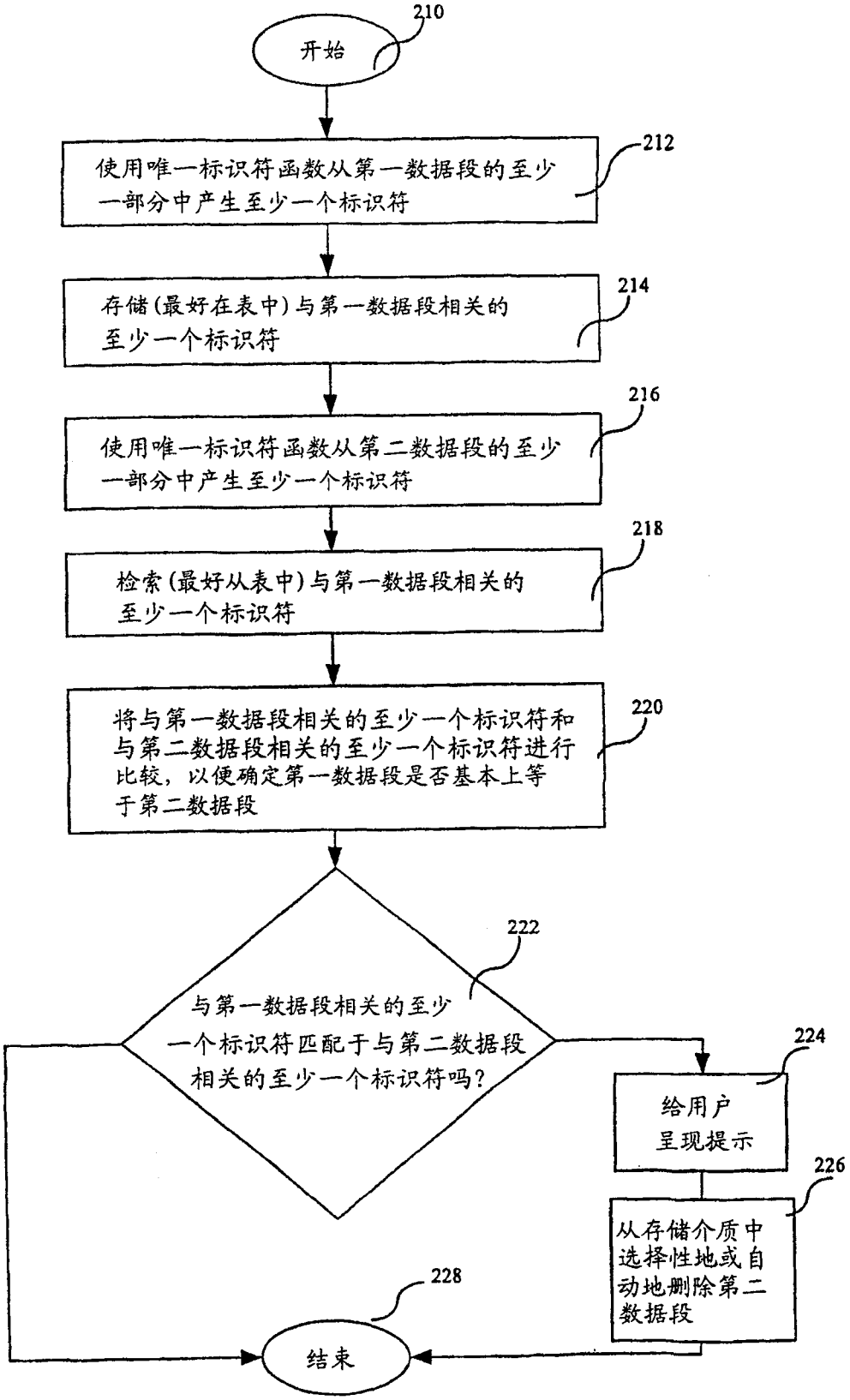

210
开始
212
使用唯一标识符函数从第一数据段的至少一部分中产生至少一个标识符
214
存储(最好在表中)与第一数据段相关的至少一个标识符
216
使用唯一标识符函数从第二数据段的至少一部分中产生至少一个标识符
218
检索(最好从表中)与第一数据段相关的至少一个标识符
220
将与第一数据段相关的至少一个标识符和与第二数据段相关的至少一个标识符进行比较，以便确定第一数据段是否基本上等于第二数据段
222
与第一数据段相关的至少一个标识符匹配于与第二数据段相关的至少一个标识符吗?
224
给用户呈现提示
226
从存储介质中选择性地或自动地删除第二数据段
228
结束

图 2