(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 1868127 B
(45)授权公告日 2011.06.22

(21)申请号 200480030478.3
(22)申请日 2004.10.15
(30)优先权数据
2003905688 2003.10.17 AU
(85)PCT申请进入国家阶段日
2006.04.17
(86)PCT申请的申请数据
PCT/AU2004/001406 2004.10.15
(87)PCT申请的公布数据
WO2005/039057 EN 2005.04.28
(73)专利权人 佩茨拜特软件有限公司
地址 澳大利亚新南威尔士
(72)发明人 布鲁斯·帕克
(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 林锦辉
(51)Int.Cl.
H03M 7/30(2006.01)
(56)对比文件
CN 1257621 A,2000.06.21, 全文.
US 2003179114 A1,2003.09.25, 全文.
JP 2002325252 A,2002.11.08, 全文.
US 6628717 B1,2003.09.30, 全文.
审查员 钟阳雪

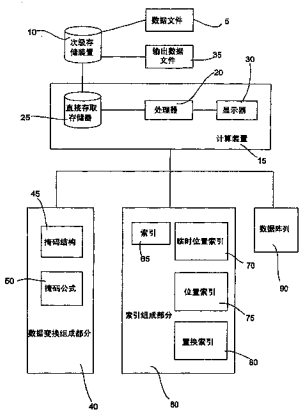
权利要求书 3 页 说明书 10 页 附图 7 页

(54)发明名称

数据压缩系统和方法

(57)摘要

本发明提供一种用于压缩数据文件的方法，所述方法包括步骤：从存储装置检索数据文件；将数据文件存储在直接存取存储器中；计算数据文件的子序列内的惟一字节值的频率，该子序列的长度不超过预定长度；创建该子序列的索引，该索引包括表示计算出的该子序列内的惟一字节值的频率的数据值；以及在该子序列具有低于预定阈值的惟一字节值的频率时，将数据变换应用于该子序列，以增加该子序列中的惟一字节值的频率，并且将表示该数据变换的数据值添加到该索引；在该子序列具有高于预定阈值的惟一字节值的频率时，将表示一个或多个惟一值在所述子序列中的位置的数据值添加到该索引；创建输出数据文件并将该索引添加到该输出数据文件。



CN 1868127 B

1. 一种用于压缩数据文件的方法，所述数据文件包含其长度大于或等于预定长度的字节序列，所述方法包括步骤：

通过一计算装置中包括的处理器从二级存储装置检索所述数据文件；

通过该处理器将所述数据文件存储在该计算装置中包括的直接存取存储器中；

通过该处理器计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；

通过连接至该计算装置的多个索引组成部分创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及

在所述子序列具有低于预定阈值的惟一字节值的频率时，通过连接至该计算装置的多个数据变换组成部分将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；

在所述子序列具有高于预定阈值的惟一字节值的频率时，通过该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；

通过该处理器创建输出数据文件，该数据文件具有文件类型标识符；以及

通过该处理器将所述索引添加到所述输出数据文件。

2. 如权利要求 1 所述的压缩数据文件的方法，其中，所述将数据变换应用于所述子序列的步骤还包括步骤：

在计算机存储器中维持多个变换数据集，所述数据集具有字节值序列并且通过变换数据集标识符而被识别；

从计算机存储器检索一个所述变换数据集，所述检索的变换数据集的长度等于所述数据文件的子序列的长度；并且

根据所述检索的数据集中的相应字节值，将数据变换应用于所述子序列中的各个字节值。

3. 如权利要求 2 所述的压缩数据文件的方法，其中，在基于至少一个所述检索的变换数据集的数据变换之后的所述子序列和所述数据变换之前的所述子序列相同。

4. 如权利要求 2 所述的压缩数据文件的方法，其中，至少一个所述变换数据集包含随机产生的字节速率序列。

5. 如权利要求 2 所述的压缩数据文件的方法，其中，至少一个变换数据集包含预定的字节速率序列。

6. 如权利要求 2 所述的压缩数据文件的方法，其中，至少一个变换数据集包含从除所述数据文件的所述子序列之外所述数据文件的一部分导出的字节值序列。

7. 如权利要求 2 所述的压缩数据文件的方法，还包含步骤：将应用于所述子序列的数据变换数据集的变换数据集标识符添加到所述索引。

8. 一种用于压缩数据文件的方法，所述数据文件包括其长度大于或等于预定长度的字节序列，所述方法包括下述步骤：

通过一计算装置中包括的处理器从二级存储装置检索所述数据文件；

通过该处理器将所述数据文件存储在该计算装置中包括的直接存取存储器中；

通过该处理器计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；

通过该处理器计算所述一个或多个惟一值在所述子序列中的位置；

通过连接至该计算装置的多个索引组成部分创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及

在所述子序列具有低于预定阈值的惟一字节值的频率时，通过连接至该计算装置的多个数据变换组成部分将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；

在所述子序列具有高于预定阈值的惟一字节值的频率时，通过该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；

通过该处理器创建输出数据文件，该数据文件具有文件类型标识符；以及

通过该处理器将所述索引添加到所述输出数据文件。

9. 如权利要求 8 所述的压缩数据文件的方法，其中，所述计算一个或多个惟一值在所述子序列中的位置的步骤还包含步骤：

在计算机存储器中创建临时位置索引；

从所述子序列检索连续字节值；

在检索每个字节值时，确定所述检索的字节值是惟一字节值还是重复值；

在检测到惟一字节值时，将两个位值中的一个添加到所述临时位置索引，或者将所述两个位值中的另一个添加到所述临时位置索引；

根据所述临时位置索引，创建表示所述一个或多个惟一值的位置的位置索引；以及

至少部分地根据所述位置索引，计算表示所述一个或多个惟一值的所述位置的数据值。

10. 如权利要求 9 所述的压缩数据文件的方法，其中，所述子序列中的字节数目等于所述临时位置索引中的位的数目。

11. 如权利要求 9 所述的压缩数据文件的方法，其中，所述位置索引的大小小于所述临时位置索引的大小。

12. 如权利要求 9 所述的压缩数据文件的方法，还包含步骤：

创建表示惟一字节值在所述子序列内的次序的置换索引；以及

根据所述位置索引和所述置换索引，计算表示所述一个或多个惟一值的所述位置的数据值。

13. 如权利要求 12 所述的压缩数据文件的方法，包含步骤：将所述位置索引和所述置换索引并置，以形成表示所述一个或多个惟一值的位置的数据值。

14. 一种用于压缩数据文件的系统，所述数据文件包括其长度大于或等于预定长度的字节序列，其中所述系统包括：

用于存储该数据文件的二级存储装置；

计算装置，连接至该二级存储装置并包括有用于存储该数据文件的直接存取存储器和处理器，该处理器用于从该二级存储装置检索所述数据文件，将所述数据文件存储在该直接存取存储器中，及计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；

多个索引组成部分，连接至该计算装置，用于创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及

多个数据变换组成部分，连接至该计算装置，用于在所述子序列具有低于预定阈值的惟一字节值的频率时，将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；

其中，在所述子序列具有高于预定阈值的惟一字节值的频率时，该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；及

该处理器还用于创建输出数据文件并将所述索引添加到所述输出数据文件，该数据文件具有文件类型标识符。

15. 一种用于压缩数据文件的系统，所述数据文件包括其长度大于或等于预定长度的字节序列，其中所述系统包括：

用于存储该数据文件的二级存储装置；

计算装置，连接至该二级存储装置并包括有用于存储该数据文件的直接存取存储器和处理器，该处理器用于从该二级存储装置检索所述数据文件，将所述数据文件存储在直接存取存储器中，计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度，及计算所述一个或多个惟一值在所述子序列中的位置；

多个索引组成部分，连接至该计算装置，用于创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及

多个数据变换组成部分，连接至该计算装置，用于在所述子序列具有低于预定阈值的惟一字节值的频率时，将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；

其中，在所述子序列具有高于预定阈值的惟一字节值的频率时，该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；

该处理器还用于创建输出数据文件并将所述索引添加到所述输出数据文件，该数据文件具有文件类型标识符。

数据压缩系统和方法

技术领域

[0001] 该发明涉及数据压缩领域，特别是涉及基于因子迭代无损压缩进行数据压缩的系统和方法。

背景技术

[0002] 电二进制文件存在许多不同格式以用于许多不同用途。这些格式包括适用于图像、声音、文本、数据、可执行文件等等的存储的格式。

[0003] 如果未进行加密，那么包含数据的二进制文件，趋向于结构格式。通常存在标题信息、文本、重复 (repetition) 和其它组成部分之间的定位 (positioning)。通常，二进制文件中的开始几个字节包含与该二进制文件兼容的文件类型且因此应用程序的指示符。可执行文件或者用于执行任何类型的功能的文件都具有相当少的结构格式。然而，作为这些文件的结构要素或者必须与操作系统相互作用以执行功能，或者它们是操作系统的一部分。

[0004] 由于通过设计，经过压缩和编码的文件去除文件内的重复值，所以它们具有最小结构。在加密情况下，密钥用来定义替换值。对于压缩而言，“速记 (shorthand)”被用于重复结构。在经过加密或者压缩的文件的情况下，所述文件将不仅内部结构改变，而且特别在压缩情况下，文件的大小也改变。

[0005] 从数学上说，对于大小为 1,048,576 字节 (1Mb) 的二进制文件而言，存在有 $256^{1,048,576}$ 种字节排列的可能结构。而在实际使用中，仅仅使用这个数字的一小部分。根据若干不同文件类型、可执行或者可操作的文件的功能的估计以及可用的压缩和加密例程，实际上使用的数目仅仅是接近于该数目。

[0006] 存在许多现有技术来对数据文件执行数据压缩。一些数据压缩算法是基于索引技术的，并且涉及所述数据文件内的惟一值的计算和索引。在大多数已压缩的数据文件中，在每个 256 字节代码段内存在一些数据值的重复。在平均文件中，每 256 字节代码段仅仅有 160 至 170 个惟一的非重复值。利用这个数目的值，基于阶乘计算的数据压缩技术无法很好地工作。

发明内容

[0007] 在一个方面，本发明提供了一种用于压缩数据文件的方法，该数据文件包含其长度大于或等于预定长度的字节序列，所述方法包括步骤：通过一计算装置中包括的处理器从存储装置检索所述数据文件；通过该处理器将该数据文件存储在该计算装置中包括的直接存取存储器中；通过该处理器计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；通过连接至该计算装置的多个索引组成部分创建所述子序列的索引，该索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及在所述子序列具有低于预定阈值的惟一字节值的频率时，通过连接至该计算装置的多个数据变换组成部分将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；在所述子序列具有高于预定阈值

的惟一字节值的频率时，通过该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；通过该处理器创建输出数据文件，该数据文件具有文件类型标识符；以及通过该处理器将所述索引添加到所述输出数据文件。

[0008] 在本说明书和权利要求中使用的术语“包括(compring)”表示“至少部分地由……构成”。也就是说，当解释本说明书和权利要求中包括该术语的陈述时，每个陈述中由该术语开始的特征所有都需要出现，但是其他特征也可以出现。相关术语（比如“包含”(“comprise”和“comprised”)）也以同样的方式来解释。

[0009] 依照本发明的另一方面，提供了一种用于压缩数据文件的方法，所述数据文件包括其长度大于或等于预定长度的字节序列，所述方法包括下述步骤：通过一计算装置中包括的处理器从存储装置检索所述数据文件；通过该处理器将所述数据文件存储在该计算装置中包括的直接存取存储器中；通过该处理器计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；通过该处理器计算所述一个或多个惟一值在所述子序列中的位置；通过连接至该计算装置的多个索引组成部分创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及在所述子序列具有低于预定阈值的惟一字节值的频率时，通过连接至该计算装置的多个数据变换部分将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；在所述子序列具有高于预定阈值的惟一字节值的频率时，通过该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；通过该处理器创建输出数据文件，该数据文件具有文件类型标识符；以及通过该处理器将所述索引添加到所述输出数据文件。

[0010] 依照本发明的另一方面，提供了一种用于压缩数据文件的系统，所述数据文件包括其长度大于或等于预定长度的字节序列，其中所述系统包括：用于存储该数据文件的二级存储装置；计算装置，连接至该二级存储装置并包括有用于存储该数据文件的直接存取存储器和处理器，该处理器用于从该二级存储装置检索所述数据文件，将所述数据文件存储在直接存取存储器中，及计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度；多个索引组成部分，连接至该计算装置，用于创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及多个数据变换组成部分，连接至该计算装置，用于在所述子序列具有低于预定阈值的惟一字节值的频率时，将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；其中，在所述子序列具有高于预定阈值的惟一字节值的频率时，这些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；该处理器还用于创建输出数据文件并将所述索引添加到所述输出数据文件，该数据文件具有文件类型标识符。

[0011] 依照本发明的另一方面，提供了一种用于压缩数据文件的系统，所述数据文件包括其长度大于或等于预定长度的字节序列，其中所述系统包括：用于存储该数据文件的二级存储装置；计算装置，连接至该二级存储装置并包括有用于存储该数据文件的直接存取存储器和处理器，该处理器用于从该二级存储装置检索所述数据文件，将所述数据文件存储在直接存取存储器中，计算所述数据文件的子序列内的惟一字节值的频率，所述子序列的长度不超过所述预定长度，及计算所述一个或多个惟一值在所述子序列中的位置；多个

索引组成部分，连接至该计算装置，用于创建所述子序列的索引，所述索引包括表示计算出的所述子序列内的惟一字节值的频率的数据值；以及多个数据变换组成部分，连接至该计算装置，用于在所述子序列具有低于预定阈值的惟一字节值的频率时，将数据变换应用于所述子序列，以增加所述子序列中的惟一字节值的频率，并且将表示所述数据变换的数据值添加到所述索引；其中，在所述子序列具有高于预定阈值的惟一字节值的频率时，该些索引组成部分将表示一个或多个惟一值在所述子序列中的位置的数据值添加到所述索引；该处理器还用于创建输出数据文件并将所述索引添加到所述输出数据文件，该数据文件具有文件类型标识符。

附图说明

[0012] 现在将参考附图描述本发明的数据压缩系统和方法的优选形式，其中：

[0013] 图 1 示出本发明的系统的优选形式；

[0014] 图 2,3 和 4 示出本发明的优选形式的压缩过程的流程图；

[0015] 图 5 示出本发明优选实施例的预期压缩结果的表格；

[0016] 图 6 举例说明本发明的另一方面，其涉及多重复字节压缩增强；以及

[0017] 图 7 也举例说明本发明的另一方面，其涉及多重复字节压缩增强。

具体实施方式

[0018] 本发明提供一种数据压缩系统和方法，其旨在应用于数据文件 5。数据文件 5 可以是任何合适的数据格式，其包括 BMP、WAV、DOC、XLS、MDB、ZIP、SIT、ARJ、ZOO、TIF、JPG、GIF、MP3、MP4 等等。数据文件 5 可以存储在形成计算装置 15 的一部分或者至少与其接口的二级存储装置 10 中。计算装置 15 至少包括与直接存取存储器 25 和显示器 30 接口的处理器 20。应当明白的是，计算装置可以包括其它的组成部分或者与其它的组成部分接口，比如数据输入装置（未示出）和输出装置（未示出）。

[0019] 可以预料的是，数据文件 5 包含其长度大于或等于预定长度的字节序列。在本发明的一种优选形式中，该预定长度是 300 字节。

[0020] 在操作中，计算装置 15 的处理器 20 从二级存储装置 10 检索所有或者部分数据文件 5。该检索出的数据文件或者部分被存储在直接存取存储器 25 中。各种操作是针对其中存储的数据文件或者其部分进行的。在直接存取存储器 25 中创建最终得到的输出数据文件 35，并且将其存储在二级存储装置 10 或者其它的二级存储装置中。期望的是，在多数情况下，输出数据文件 35 的大小将小于数据文件 5 的大小。

[0021] 首先检查数据文件 5 的子序列。子序列的长度优选不超过预定长度 300 字节。如果所述识别出的惟一值的数目降到阈值之下，那么可以在一次尝试中将一系列数据变换应用于该子序列中，以增加所述子序列中的惟一字节值的频率。

[0022] 多个数据变换组成部分 40 被存储在直接存取存储器 25 中或者二级存储装置中。所述数据变换组成部分 40 可以包括多个随机产生的字节值的序列或者字节值的预定序列。该序列被存储为掩码结构 45。除了数据变换组成部分之外，可供选择地的或者优选地，还包括多个掩码公式 50，该掩码公式可用于产生另外的掩码结构 45。数据变换组成部分的应用将在下面进一步进行描述。

[0023] 所述系统还包括多个索引组成部分 60。在该数据文件 5 的子序列的处理期间，创建随后将被写入输出数据文件 35 中的索引 65。所述索引组成部分 60 还可以包括临时位置索引 70、位置索引 75 和置换索引 (permutation index)80。在某些情况下，位置索引 75 和置换索引 80 的内容将被添加到所述索引 65 中。下面将对各种索引组成部分 60 的操作进行进一步的描述。

[0024] 所述系统还可以包括在直接存取存储器 25 或者二级存储装置中存储的数据阵列 90。在数据阵列 90 的内容被写入输出数据文件 35 之前，该数据阵列 90 可用于存储各种索引组成部分 60 和正在被压缩的数据文件 5 的子序列的部分。

[0025] 图 2 至 4 举例说明本发明的优选形式的操作。二进制数据文件 5 优选地被分段成多个数据组。在本发明的一种优选形式中，各个数据组优选为 300 字节或者更少。然而，应当明白的是，正在被压缩的数据组的大小可以是超过五位的任何大小。该数据文件首先被检验 200，以确定数据文件的长度是否大于或等于预定长度。在本发明的一种优选形式中，初始的预定长度是 300 字节。在一种形式中，整个数据文件可以从二级存储装置检索到，并且将整个数据文件存储在随机存取存储器 25 的数据阵列 90 中。或者，数据文件 5 的部份可以从二级存储装置 10 中检索出作为数据流。

[0026] 数据组被计数 205，以便计算该数据组内惟一数据值的频率。将惟一数据值的频率与预定阈值相比较 210。在一种优选形式中，预定阈值为 256。如果在 300 字节子序列内存在少于 256 个惟一值，那么可以在一次尝试中将一个或多个数据变换应用于该子序列，以增加该子序列内惟一字节值的频率。

[0027] 如果 300 字节内的惟一字节值的频率降到 256 个值的预定阈值之下，那么测试 215 子序列，以识别数据变换“掩码”是否适用于所述子序列。在本发明的一种优选形式中，结构库被维持在比如直接存取存储器 25 之类的计算机存储器中。该库优选包括多个随机产生的数据集。这些数据集中的每个可以利用数据集标识符来识别，所述数据集标识符存储在计算机存储器中并且与各个随机产生的数据集相关联。

[0028] 在一种形式中，至少一个随机产生的数据集的长度基本上等于数据文件的子序列的长度。换句话说，所述子序列中的字节的数目与在变换数据集或者掩码中的字节数目相同。基于相应的字节值和所述检索到的变换数据集，通过将数据变换应用于所述子序列中的各个字节值，可以将这种掩码应用于所述子序列中。

[0029] 数据变换的一个实例的是模数加法。将所述子序列的第一字节值和所述数据集的第一字节值加在一起，然后对总数进行模 256 的计算。例如 :如果所述子序列的第一个二进制值是 168，而所述识别出的数据集的第一个二进制值是 203，那么合并后的总数是 371。由 371 进行 MOD256 的计算后的变换值是 115。然后利用所述数据集中的第二字节，以同样的方式变换所述序列中的第二字节。然后基于所述数据集中的第三字节来变换在所述子序列中的第三字节，等等。

[0030] 按照这种方式，将所述掩码应用 220 到所述子序列。

[0031] 在一种形式中，在计算机存储器中可以存储有 65,536 种掩码结构，各个掩码具有数据集标识符，该数据集标识符的形式为 0 和 65,536 之间的索引号。所述索引可以是指向相关数据集标识符的简单的 14 位段。

[0032] 该数据变换组成部分 40 可以包括掩码公式 (formula)，例如 :

[0033] • 300 字节或者更少字节的数据文件的在前 (preceding) 序列的标准误差。应当明白的是，由于数据文件的第一个序列没有在前子序列，因而该公式不可用于数据文件的第一个序列。

[0034] • 基于上述的子序列或者标准偏差来对所述子序列内的值进行反序 (Reversal)。

[0035] • 基于子序列的结构计算的适用结构。

[0036] • 基于文件结构随机产生的段，其被添加到所述相关子序列或者从相关子序列中减去。

[0037] 上述公式可以被预先应用来产生一系列掩码结构。或者，在数据变换期间，可以计算相关字节值。在一种形式中，512 个随机产生的结构或者掩码结构可以被存储在直接存取存储器 25 中。这些结构被应用于可以在 300 字节序列内具有 256 或更多个 0 值的数据文件的子序列。在形成许多软件应用程序的二进制文件部分的标题中，这是常见的。这些随机产生的结构还可以被应用于具有高重复级的其它格式中。

[0038] 在所述子序列上的数据变换之后，再次对所述子序列进行测试 210，以识别所述 300 字节内是否存在 256 个惟一值。如果不存在 256 个惟一值，并且没有另外的掩码应用于该子序列中，那么 300 字节的阈值被降低并且在较小的子序列上重复所述过程。在一个优选实施例中，阈值可以被临时地降低到 152 个 7 位值或者 77 个 6 位值以检查少于 300 个 8 位值（字节）。然后将阈值提升到 300 字节，以用于下一个子序列。以下将进行更为详细的描述。

[0039] 添加随机文件未必将在 256 字节段内创建 256 个惟一值，但是有大约 10％的可能。可以预料的是，一旦已经应用适当的随机文件结构，那么在不超过 300 字节的数据段内将有 256 个惟一值。在任何情况下，数据变换的目的是增加数据组中的惟一数据值的频率。

[0040] 本发明计算所述数据组内的 300 个数据值的索引。

[0041] 所述索引被优选存储在直接存取存储器 25 中的数据阵列 90 中。首先，利用两位来创建 300 个数据值的索引。如果在 300 字节数据组内已经识别出 256 个惟一值，那么位值“01”被写入 225 到所述索引中。

[0042] 在掩码已经被应用于所述子序列的情况下，所述掩码或者数据集标识符随后被写入 230 所述索引。该掩码标识符将优选为 16 位值，其识别在 0 和 65,536 之间的掩码值。所述掩码标识符中的值 0 表示这样的事实：即没有掩码或者零掩码已经被应用于所述子序列中。在零数据集被应用于所述子序列中的情况下，经过数据变换后的所述子序列基本上和数据变换之前的所述子序列相同。

[0043] 本发明的方法中的下一步是创建 235 临时位置索引。

[0044] 在正从 300 字节数据组中提取 256 个惟一值的情况下，临时位置索引创建方法开始于所述数据组中的第一字节，并且检查所述数据组中的随后字节，直到 256 个惟一值已经被识别出。如果被检查到的特定值是该数据值在该数据组或者前面数据组中的第一次出现，那么“1”位值被添加到临时索引中。另一方面，如果被检查到的数据值是较早数据值的重复，那么“0”位值被写入到所述索引中。一旦已经将 256 个“1”位写入到所述索引中，那么索引方法终止。

[0045] 所述临时索引便于将所述数据组中的每个数据值容易地放置在最终得到的压缩位流中且在最终得到的压缩位流中识别所述数据组中的每个数据值。所述索引中的“1”值

的数目表示使用了多少个位值。举例来说，如果临时索引中的 283 个条目 (entries) 后，在临时索引中出现了 256 个“1”值，那么这表示在所述子序列的 283 个字节内存在 256 个惟一字节值。

[0046] 如果在 300 字节数据组内存在 256 个或更多值，那么所述索引的开始两位将已经被设置为“01”。当所述临时索引可以被简单地添加到主索引时，存在存储该信息的更为有效的方法。在所述子序列中出现的“1”值的数目是已知的。如果不考虑它们出现的次序，那么仅仅需要记录惟一字节值的情形 (instances) 的数目。

[0047] 优选创建 240 位置索引并且将该位置索引写入到主索引中，而非记录临时索引本身。对于 300 字节子序列而言，在临时索引包括其后跟随 44 个“0”值的 256 个“1”值的情况下，可以为此分配位置索引“0 “。在 300 字节数据组内排列 44 个“0”值和 256 个“1”值的方式的数目是 ${}^{n}C_{r}$。这意味着，在存在 256 个“1”值和 44 个“0”值的情况下，在 300 个值内存在 300 ！ /256 ！ .44 ！ 种可能组合，该组合数等于 $1.34x10^{53}$。

[0048] 这种最大位置索引值 $1.34x10^{53}$ 小于值 2^{177}，该值需要 177 位来进行表示。

[0049] 这意味着，通过利用在临时索引中存在有至少 256 个“1”值的事实，位置索引可以被记录为 177 位或者 22.125 字节，而非存储 300 位的实际临时索引。

[0050] 记录数据组内的数据值的次序以启动压缩以及解压缩同样也是重要的。这是通过创建 245 置换索引并且通过将该置换索引写入主索引来实现的。

[0051] 置换索引计算是基于 256 个惟一值可被排序的方式的数目或者在不重复的情况下 256 个值的置换的数目的。对于第一个值，存在 256 种可能性，对于第二值，存在 255 种可能性，对于第三值，存在有 254 种可能性，等等。这被表示为 256 ！，称为“256 的阶乘”。256 个惟一值的可能置换的数目因此是 $8.57x10^{506}$。由于 2^{1684} 等于 $8.6x10^{506}$，其大于 $8.57x10^{506}$，所以该值可以由 1,684 位来表示。1,684 位等效于 210.5 个字节。

[0052] 序列 0,1,2,3,4…,254,255 将被表示为置换编号 1，而序列 255,254,253…3,2,1,0 将由置换编号 $8.57x10^{506}$ 表示。

[0053] 将所述置换索引写入主索引中。到现在为止，所述主索引将包括表示计算出的所述子序列内惟一位值的频率的数据值。这将是位值“01”，其后继之表示掩码已经被应用的 16 位，随后是表示位置索引的 177 位，随后是表示置换索引的 1,684 位。

[0054] 在到达这样的一点后 :在该点上，或者在数据文件中没有留下足够的位来获得足够长度的子序列，或者没有留下充足的唯一值，那么索引就被写入 250 输出文件中。

[0055] 该输出文件优选包括三个初始字节来标识文件类型。在所述文件类型标识符之后的另两个字节表示本发明的方法已经在特定数据文件上运行的次数，直到最大 65,536 次重复。

[0056] 在这五个字节之后，在数据阵列 90 中存储的索引被添加到输出文件。在所述索引之后增加的是这些索引中未用的、或者由于缺乏保持在数据文件中的足够位值的任何值或者唯一值的任意值。

[0057] 在大多数情况下，期望的是，存在五个标题字节，其后继之以主体，以及在输出文件末端以完整的未压缩的形式写入的 63 或者更少位值。所述输出文件的主体优选是被连续地写入的索引的集合，以便于以流的方式进行提取。

[0058] 如图 2 所述，存在这样的情形 :在本发明的方法的多次迭代之后，在所述数据文件

中不再剩余 300 字节，或者存在有 300 字节的子序列，在该子序列中，不存在 256 个惟一值并且没有另外的掩码适用。如在 260 中所示，在一种优选形式中，从所述数据文件检索到的子序列的大小可以减少。

[0059] 参考图 3，检验 305 数据文件以识别在所述数据文件中是否剩余至少 152 个字节。

[0060] 如果在所述数据文件中剩余至少 133 个字节，该 133 个字节包含 152 个 7 位值，那么在所述 152 个 7 位值内的惟一值的数目被计数 310。然后相对于例如 128 的阈值数目来检验 315 惟一值的数目。如果在 133 字节子序列中不存在足够的惟一值，那么识别适用的掩码（步骤 340）并且按类似方式将该适用的掩码应用于图 2 的步骤 215 和 220 中（步骤 345）。

[0061] 一旦在所述数据文件中的 152 个 7 位值中识别出惟一值的阈值数目，那么位序列“10”被写入 350 到所述索引中，并且该方法继续前进到图 2 中的 230 表示的步骤。

[0062] 如果在所述数据文件中不存在剩余来待处理的 152 个 7 值，或者 128 个惟一值不能位于 1527 位子序列内并且如 355 所示没有另外的掩码适用，那么该方法转到图 4 所示的步骤。如图 4 所示，处于检查中的数据文件中的位组的数目被减少到 77 个 6 位值。如果存在 405 剩余在数据文件中的 77 个 6 位值，那么在 77 个 6 位值中的惟一值的数目被计数 410。

[0063] 相对于阈值 64 来检验 415 惟一值的数目。如果在所述 77 个 6 位值中存在小于 64 个惟一值，那么该方法确定 420 掩码是否适用。如果掩码适用，那么应用 425 该掩码。这最后二个步骤 420 和 425 类似于图 2 的步骤 215 和 220 和图 3 的步骤 340 和 345。

[0064] 如果在 77 个 6 位子序列中存在 64 个惟一值，那么值“11”被写入 430 到所述索引中。然后控制返回到图 2 中的前面步骤 230。

[0065] 如果在数据文件中不存在剩余来待处理的 77 个 6 位值，或者在所述 77 个 6 位值序列内不存在 64 个惟一值，那么位值“00”被写入 435 到所述索引中，按照与图 2 所示的步骤 250 一样的方法将所述索引写入 450 到输出文件，并且所述数据文件中的剩余字节被写入 460 到所述输出文件中。

[0066] 应当明白的是，取决于处于检查中的字节的数目，如图 2 的步骤 245 所示，置换索引需要微小变化。当在 152 个 7 位组内存在 128 个惟一数据值的情况下，位置索引将为 152 ！ /128.24 ！，其等于 $5.48x10^{27}$。由于 $2^{93} = 9.9x10^{27}$，所以这可以由 93 位来表示。

[0067] 当在 77 个 6 位组内存在 64 个惟一值时，所述索引将为 77 ！ /64 ！ /.13 ！ 。$2^{48} = 2.81\times10^{14}$，其大于前者值 $1.84x10^{14}$，所以这可以由 42 位来表示。

[0068] 类似地，取决于处于检查中的字节的数目，如图 2 中的步骤 245 所示，置换索引需要微小变化。128 个值的置换为 128 ！ 或者 $3.86x10^{215}$。由于 $2^{717} = 6.89x10^{215}$，所以这需要 717 位来表示。

[0069] 64 个值的置换是 64 ！ 或者 $1.27x10^{89}$。由于 $2^{296} = 1.27x10^{89}$，所以这可以由 296 位来表示。

[0070] 图 5 举例说明关于数据组大小为 377 字节（8 位组）、350 字节、320 字节、300 字节、152 个 7 位组和 77 个 6 位组的预期结果的表格。该表格所包括的是所包含的变化的效果的指示。描述如下。

[0071] 解压缩仅仅是将上述过程反向。索引值表示从第一个到最后一个（第 256）的各

个值的范围。假设所述范围提供相关值。则索引可以被用于与所述标题一起重构。由于所有组成部分被一起打包，所以可以设想的是使用流。

[0072] 如果存在更为有效的方法，那么可以根据用于段的“0”和“1”值的字串来改变重复值的放置的索引。例如，如果存在仅仅一个或两个重复值，那么字节的数目将为 257 或者 258。已知的是，对于该段而言，所述第一个字节和最后一个字节是惟一的，而不是使用第 257 位和第 258 位。因此，在 257 个值的情况下，8 位将提供单个重复值的位置，而在 258 个字节段情况下，16 位将提供两个重复值的位置。

[0073] 所述方法可以被应用于所有文件类型和结构。对于利用诸如 PKWare 的 ZIP 产品之类的工具来相当的数量地压缩的文件类型或者结构，在一次完成 (single pass) 上本发明的方法不会达到相同的程度。然而，该方法可以被重复地应用于相同的文件，每次在大小上减小。所述次数或者重复次数取决于硬件处理和 / 或用户需求时间。

[0074] 由于所有组成部分是已知的，所以解压缩非常快。由于压缩需要对随机数据结构进行匹配时，所以解压缩比压缩更快。

[0075] 由于所有的索引被包含在实际数据本身中，那么可以同时地执行多个解压缩例程。

[0076] 其它的应用程序可以包括软件压缩，数据压缩，在掌机 (console) 之间的在线游戏以及语音 IP 和 / 或视频点播等，所述掌机诸如是索尼游戏站 2、微软公司 X-Box 等。该本发明具有以任何格式在任何地方存储、发送或者使用的数据或者二进制信息的应用程序。

[0077] 上述描述是基于在 300 字节代码段或者更小的代码段内的 256 个惟一值的。应当明白的是，这种选择的大小是仅仅用于示例性的目的。可以使用这方法重构 5 位或更多位的数据组，或者在 0 和 31 之间的值。减少随机产生的数据集的数目或者覆盖文件意味着还可以使用 3 位值和 4 位值。

[0078] 使用比所述 8 位 (256 值) 更大的位值可以实现更大的节省。例如，在 9 位值被压缩的情况下，与利用 8 位压缩达到的相比，其在压缩时有进一步的好处。

[0079] 节省或者压缩随着每值使用的位的数目的增加而增加。256 个值 (300 字节段) 的压缩不能与 512 个值 (600 字节段) 的压缩一样多。依次，512 个值数据的压缩不能与 1024 个值的压缩一样多。由于计算必须是基于文件大小的，因而不存在较高的确定级别。

[0080] 通过使用上述 300 字节的方法，这可以被扩展至 377 字节组。这意味着，所述有效范围是 256 至 377 字节组，其中对于在本说明书中描述的以及在图 5 举例说明的优选实施例而言，300 字节是最佳级别。

[0081] 针对 300 个 8 位组 (字节)，152 个位组以及 77 个位组的变化可以在压缩文件的标题中指示出。该变化可以包括两个 (2) 部分。它们是 :

[0082] 1. 每段大小的相关位组的数目的指示。对于 8 位组而言，大小范围是 256 至 377，其由 7 位来表示。对于 7 位组而言，范围可能以 5 位表示，而对于 6 位组而言，范围可以以 4 位表示。

[0083] 2. 可以在上述每一个的末端添加另一位，以表示在每一位组内是否出现变化。“0”表示 No(否)，而“1”表示 Yes(是)。

[0084] 因此，所述标题包含表示上述值的另外的 19 位。

[0085] 如果对于每个标题，变化值是允许的，那么可以将变化值逐组地写入到所述索引

中。

[0086] 例如，在一组 8 位组上的缺省可能是 300 个值，但是各个段可以在 256 和 377 个值之间变化，如所包括的变化值所表示的。

[0087] 本发明的另外实施例可以涉及多重复字节压缩增强，并且将参考图 6 和 7 对此进行描述。

[0088] 功能电子文件分成若干不同种类的字节结构。这些从简单的 2 颜色位图变化到使用任何目前可用的无损压缩算法压缩的文件。

[0089] 对于标题信息之后的 2 色位图而言，一个位值意味着黑色，另一个位值意味着白色。由于存在大量重复，所以以无损耗方式对这些文件进行压缩是简单的。

[0090] 移动到 24 位位图，图案的识别变得更加困难，由此使用现有算法进行的无损压缩比率不如针对更简单的位图结构进行的无损压缩比率一样大。

[0091] 这里描述的处理将更简单的图案引入到 24 位位图，对于标准照片类型图像，其允许使用任何目前可用的无损压缩算法，以显著地增加压缩量。

[0092] 为了实现这个目的，如图 6 中的步骤 610 所例示，原始图像被分解为 3 个组成部分，其中合并的尺寸明显地大于所述原始图像。

[0093] 然后，如 620 中所示，所有 3 字节 (24 位) 组按照升序的十进制数进行排列。例如，236，217，67 被重新排序为 67，217，236。使用 Huffman 结构，将在字节排列中的变化记录在所述索引中。

[0094] 由于仅仅有 6 种可能的原始结构，利用以下位索引来记录这些 ：

[0095] 00 ＝ 123

[0096] 01 ＝ 132

[0097] 100 ＝ 213

[0098] 101 ＝ 231

[0099] 110 ＝ 312

[0100] 111 ＝ 321

[0101] 上面数字中的每一个表示当与它们的排序后的放置相比时所述字节的原始位置。

[0102] 一旦图像已经被完全地扫描，这个索引被写入到文件 (文件 A) 中，如 625 中所示。

[0103] 然后，如 630 中所示，所有最低的或每个组的当前第一字节值被写入到一不同的文件 (文件 B) 中。

[0104] 如 635 中所示，由于字节值是有序的，第二字节值减第一字节值的值被写入到文件 (文件 C) 中，其后继之第三字节值减第二字节值的值。

[0105] 这样已经创建了 3 个文件 ：文件 A、文件 B 和文件 C。文件 B 和文件 C 的合并总数将与原始 24 位位图相同。由于文件 A 表示字节的索引，所以在尺寸上，文件 A 是额外开销 (overhead)。

[0106] 如果随后利用无损耗算法或如 WINZIP 之类的产品 640 将全部的三个文件 (A，B 和 C) 压缩成一个文件 650，那么得到的文件平均比通过简单地在未修改的图像文件上应用这些工具达到的要小 25％。

[0107] 测试已经表示了 2.5％下降的最坏情况，最好的情况是图像质量真实的 24 位位图的 82％。利用 JPEG 的无损压缩模式可以产生相同的效果。

[0108] 可以利用 3 个字节组将该过程应用于任何文件结构以保持数据。它可能同时扩大为覆盖 4,5,6,7,8 等等字节结构,以达到无损压缩的更高级别。

[0109] 由于位图文件被用于显示图象,Wave(.wav) 文件用于播放声音。现在将参考图 7,来描述涉及 Wave 格式文件的压缩增强处理的另一个例子。由于存在不同级别的位图文件(2 位、4 位、8 位、10 位、12 位、16 位、24 位和 30 位),其中每个位图提供更多颜色或者质量,对于 Wave 文件发生相同的情况。

[0110] 利用若干组成部分来创建 Wave 文件,这些组成部分是平均采样速率、采样速率、声音样本大小和通道数目。

[0111] 较低的采样速率意味着较小文件,但是具有较低质量。单声道文件同样小于立体声文件。

[0112] 在此被寻址的 Wave 格式,是将完整品质立体声音乐存储在商用 CD 上时使用的格式。这个格式是从 Wave 格式转换成 CD 格式。

[0113] 对于具有 176.4Kb/ 秒的平均数据速率、44.1kHz 的采样率、16 位声音样本大小和 2(立体声) 通道的 Wave 文件而言,可能适用以下情况。

[0114] 如果通过编号为 1 到 n 来表示所述文件中的全部的字节值,其中 n 是在文件中的最后字节 (对于正常的声音文件,这将是 50,000,000 的数量级),所有偶数位字节值被写入到一个文件 (文件 A),如在 725 所示,而所有奇数位字节值被写入到另一不同文件 (文件 B),如在 730 所示。例如 :

[0115]

字节值	255	167	33	0	0	24	24	167	167
顺序值	1	2	3	4	5	6	7	8	9
文件 1(奇数)	255	33	0	24	167				
文件 2(偶数)	167	0	24	167					

[0116] 如果随后利用无损耗算法或如 WINZIP 之类的产品将两个文件 (文件 1 和文件 2) 压缩成一个文件,再次如 640 所示,那么得到的文件 650 平均比通过简单地在未修改的图像文件上应用这些工具达到的要小 20%。

[0117] 测试已经暗示了 :在最坏情况下,已压缩文件的大小有 10% 额外的下降,在最好的情况下,在大小上有 43% 的下降。

[0118] 提取 / 解压缩是简单的,在使用相关的无损耗工具解压缩文件 1 和文件 2 之后,文件 2 的字节被重新插入到文件 1 的每个字节之间。

[0119] 以上描述了包括其优选形式的本发明。对于本领域技术人员而言,改变和修改是显而易见的,其趋于并入所附权利要求定义的范围内。

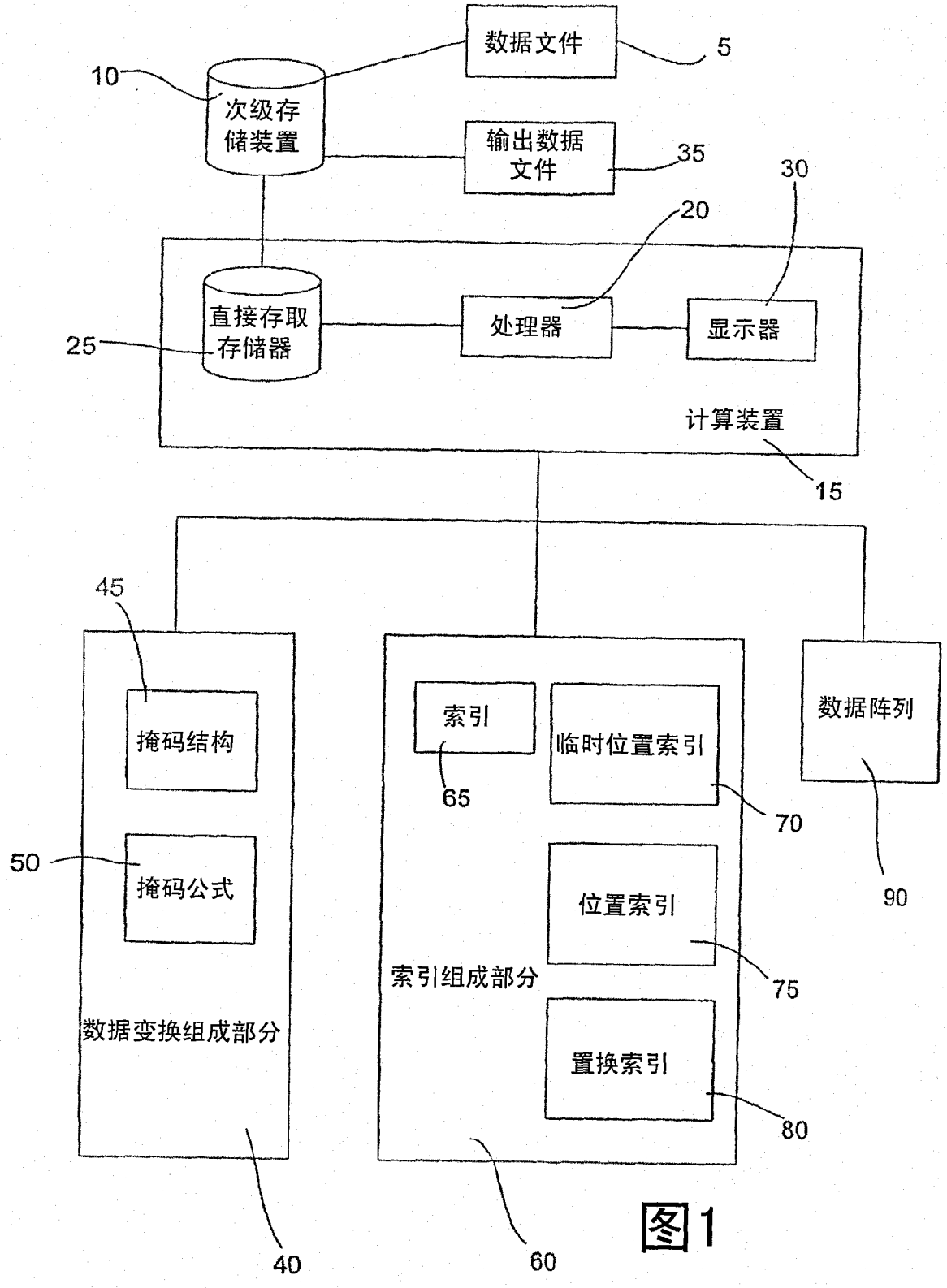

数据文件
5
10
次级存储装置
输出数据文件
35
30
20
直接存取存储器
25
处理器
显示器
计算装置
15
45
掩码结构
50
掩码公式
数据变换组成部分
40
索引
65
临时位置索引
70
数据阵列
90
位置索引
75
索引组成部分
置换索引
80
60

图1

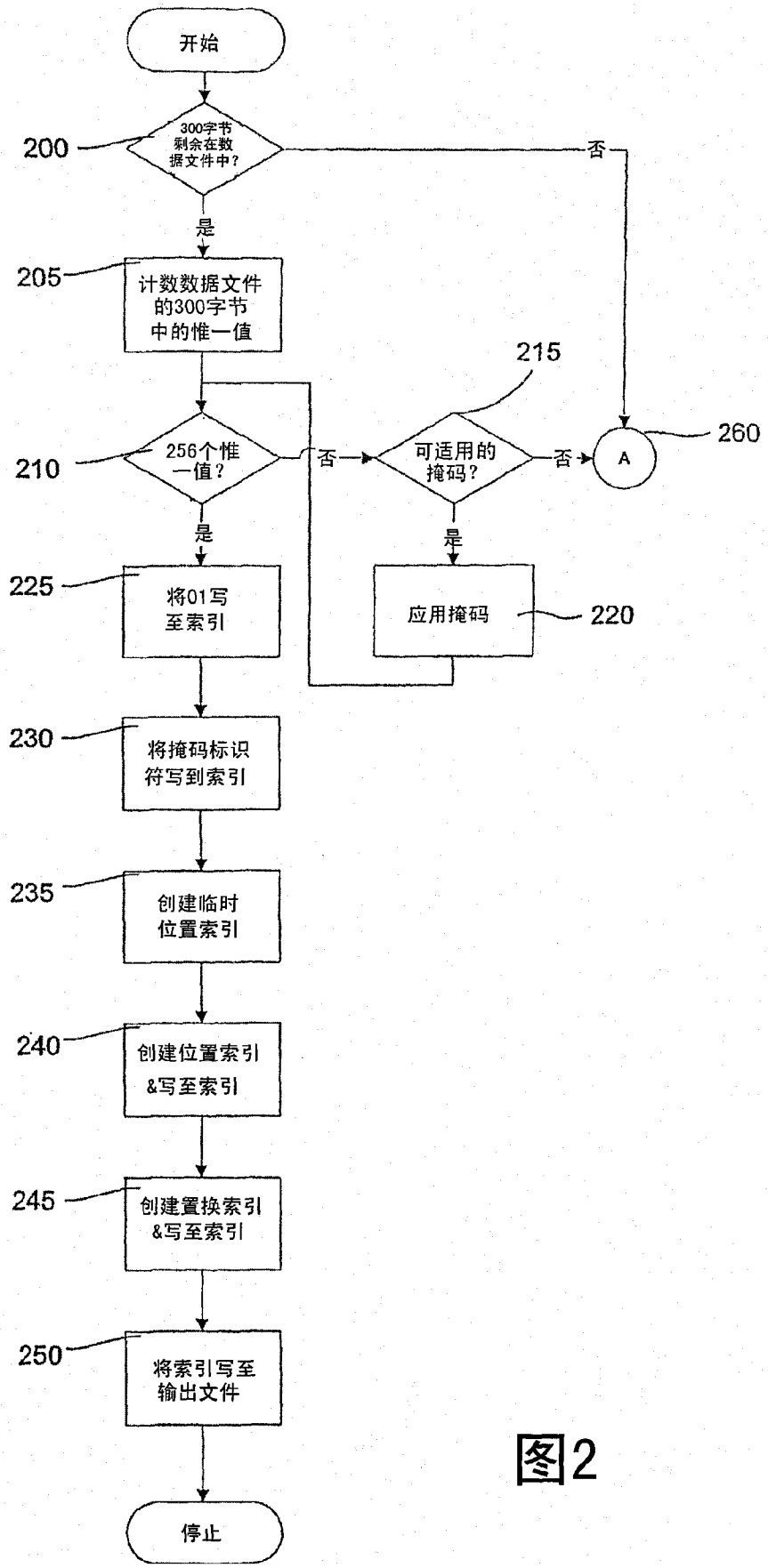

开始
300字节剩余在数据文件中?
200
否
是
205
计数数据文件的300字节中的惟一值
215
256个惟一值?
210
否
可适用的掩码?
否
A
260
是
是
225
将01写至索引
应用掩码
220
230
将掩码标识符写到索引
235
创建临时位置索引
240
创建位置索引&写至索引
245
创建置换索引&写至索引
250
将索引写至输出文件
停止

图2

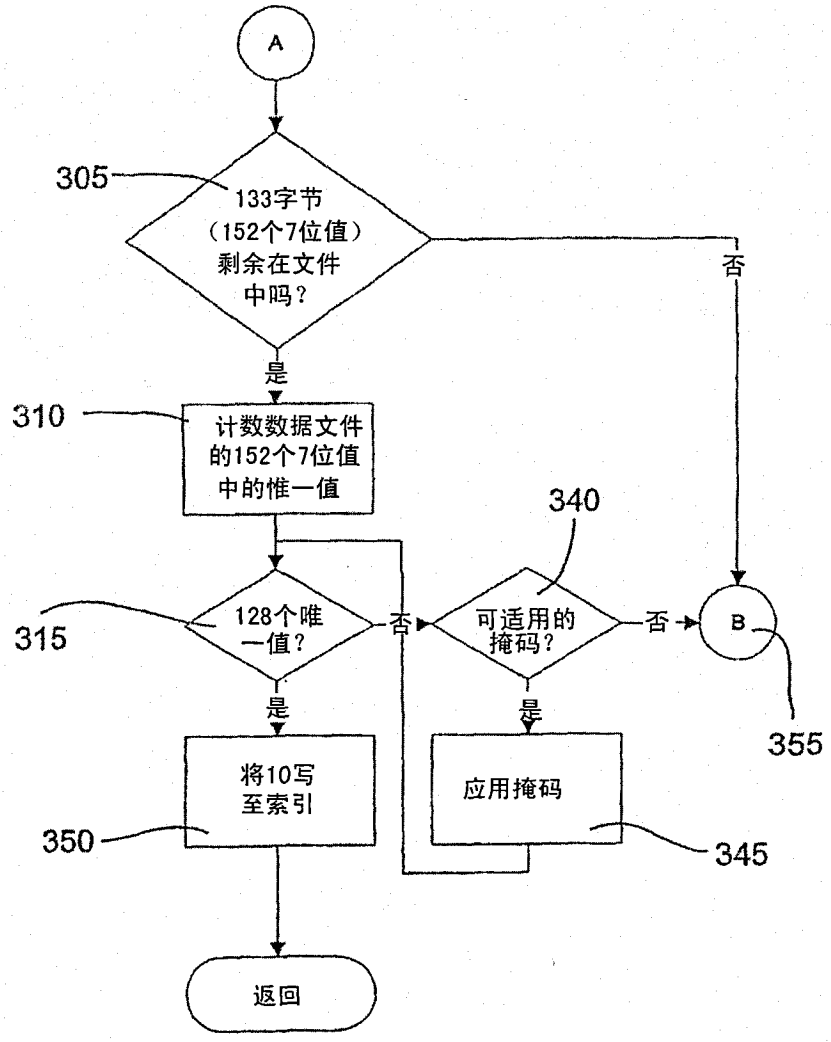

A
305
133字节
（152个7位值）
剩余在文件
中吗？
否
是
310
计数数据文件
的152个7位值
中的惟一值
340
315
128个唯
一值？
否
可适用的
掩码？
否
B
355
是
是
将10写
至索引
应用掩码
350
345
返回

图3

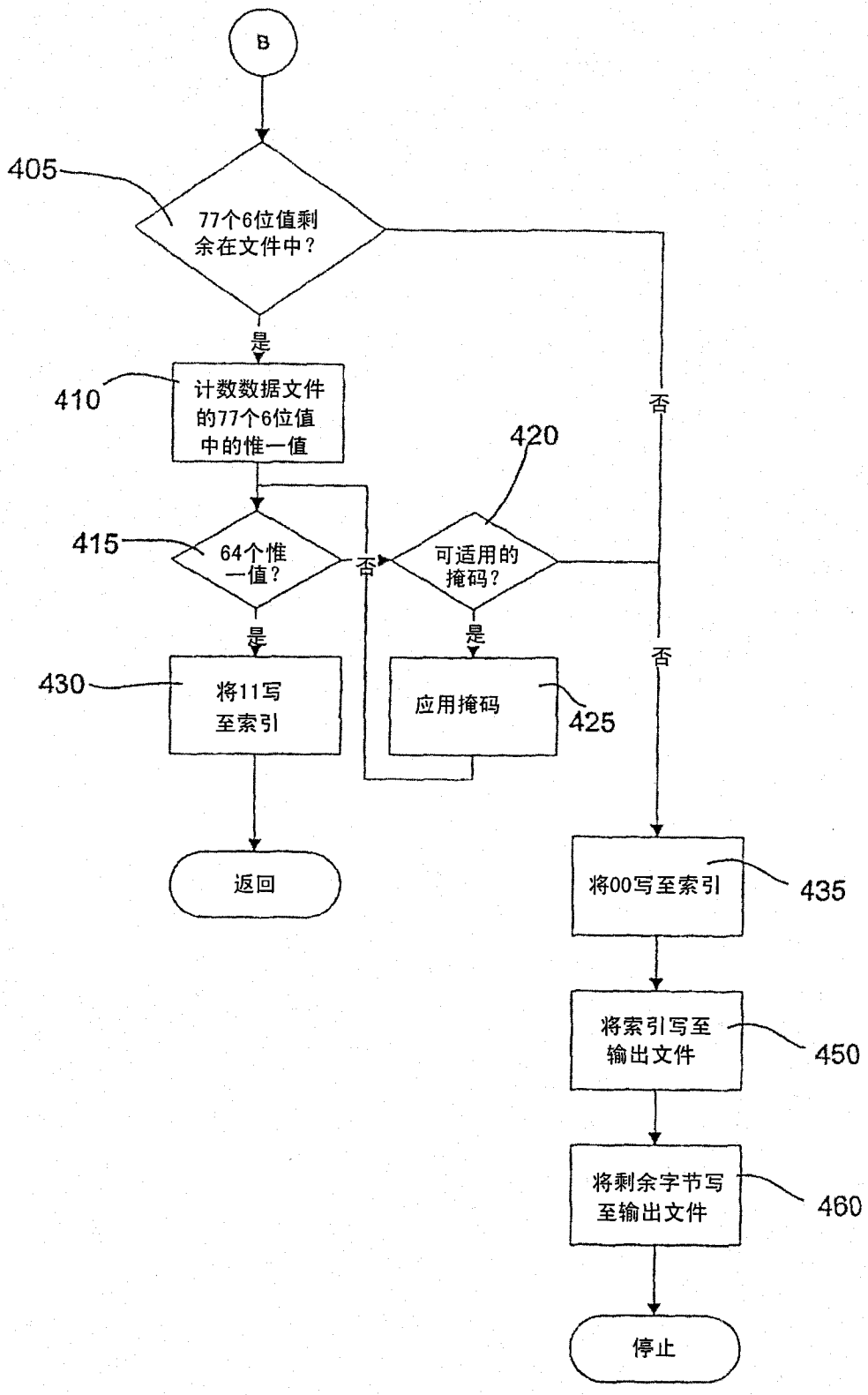

B
405
77个6位值剩余在文件中？
是
否
410
计数数据文件的77个6位值中的惟一值
420
415
64个惟一值？
否
可适用的掩码？
是
是
否
430
将11写至索引
应用掩码
425
返回
将00写至索引
435
将索引写至输出文件
450
将剩余字节写至输出文件
460
停止

图4

序列大小	字节（位）数	表示序列大小的索引位	表示根据需要变化的索引位	表示适用的掩码的位	需要来表示阶乘值的位	表示顺序组合的位	总数	保留
377	256 (2048)	2	7	16	1684	337	2046	0.10%
377	256 (2048)	2		16	1684	337	2039	0.44%
350	256 (2048)	2	7	16	1684	290	1999	2.39%
350	256 (2048)	2		16	1684	290	1992	2.73%
320	256 (2048)	2	7	16	1684	227	1936	5.47%
320	256 (2048)	2		16	1684	227	1929	5.81%
300	256 (2048)	2	7	16	1684	178	1886	7.91%
300	**256 (2048)**	**2**		**16**	**1684**	**178**	**1879**	**8.25%**
152	(892)	2		16	717	93	828	7.17%
77	(384)	2		16	296	48	362	5.73%

图5

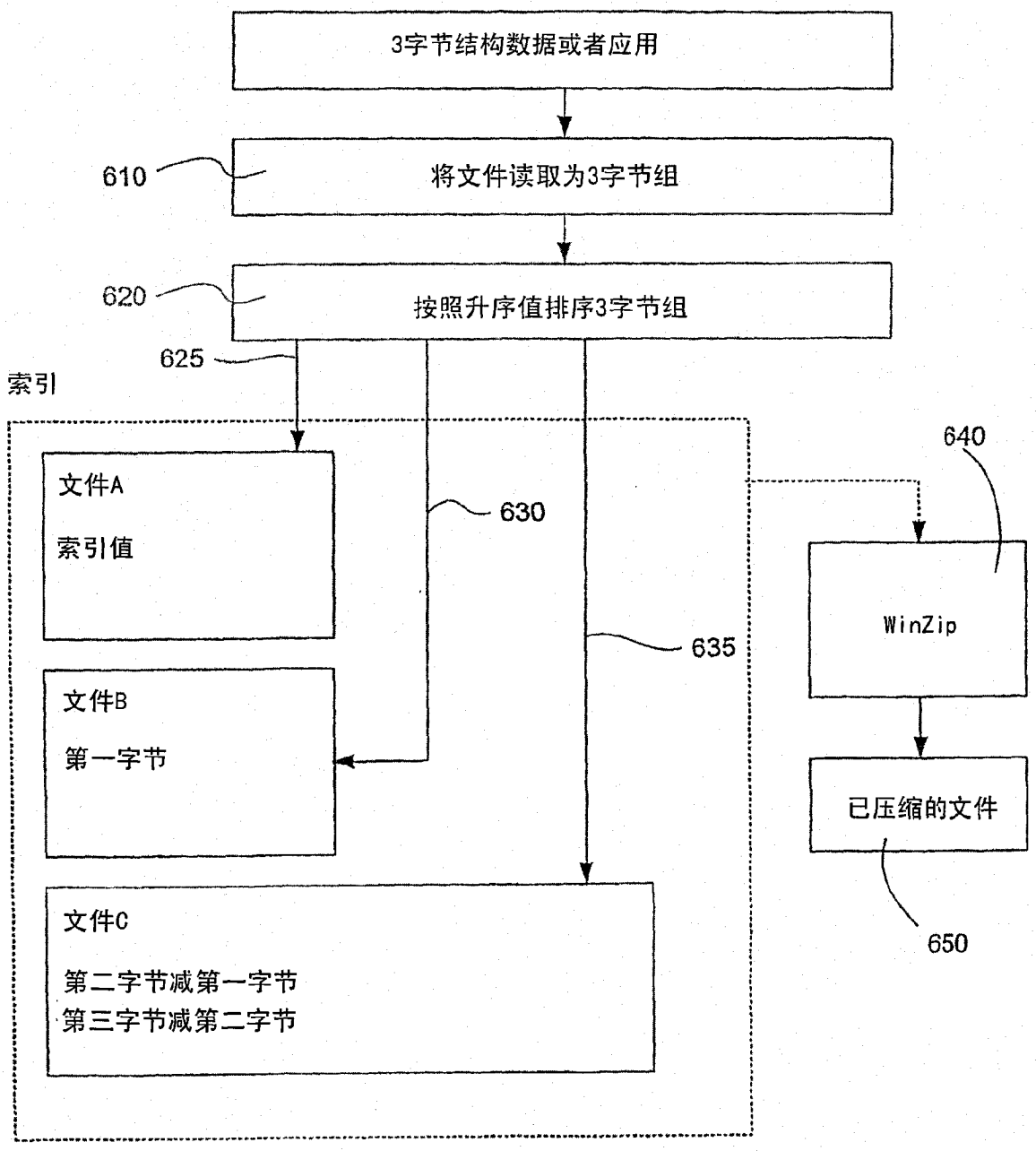

3字节结构数据或者应用
610
将文件读取为3字节组
620
按照升序值排序3字节组
625
索引
文件A
索引值
630
文件B
第一字节
635
文件C
第二字节减第一字节
第三字节减第二字节
640
WinZip
已压缩的文件
650

图6

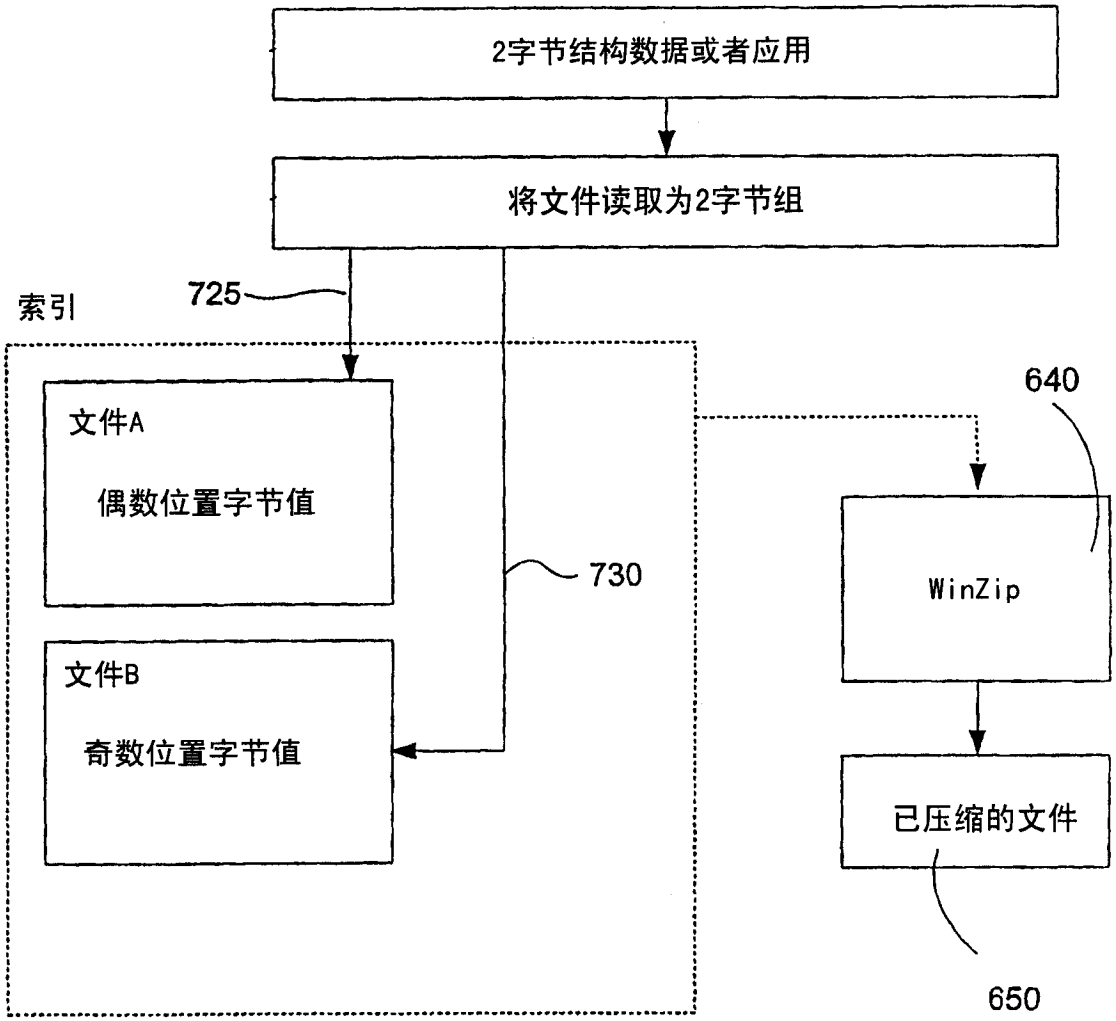

2字节结构数据或者应用
将文件读取为2字节组
索引
725
文件A
偶数位置字节值
730
文件B
奇数位置字节值
640
WinZip
已压缩的文件
650

图7