[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 12/00

[12] 发明专利说明书

[21] ZL 专利号 98801551.X

[45] 授权公告日 2004 年 5 月 5 日

[11] 授权公告号 CN 1148657C

[22] 申请日 1998.10.20 [21] 申请号 98801551.X
[30] 优先权
[32] 1997.10.21 [33] JP [31] 288559/1997
[86] 国际申请 PCT/JP1998/004736 1998.10.20
[87] 国际公布 WO99/021092 日 1999.4.29
[85] 进入国家阶段日期 1999.6.18
[71] 专利权人 富士通株式会社
地址 日本神奈川
[72] 发明人 片冈正弘 坪仓孝
审查员 石 岗

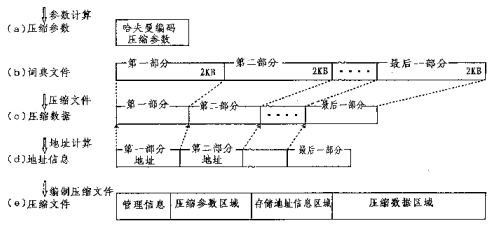
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 于 静

权利要求书 2 页 说明书 14 页 附图 9 页

[54] 发明名称 文件处理方法和数据处理装置

[57] 摘要

一种文件处理方法，一种数据处理装置和一种存储介质，把数据和关于该数据的索引数据划分成多个部分，并对各部分进行压缩以获得压缩文件，并把压缩文件和压缩后各部分的地址信息存储在存储介质中。



ISSN1008-4274

1. 一种文件处理方法，包括：

一个压缩步骤，该步骤把数据和对应于该数据并且在查找或搜索该数据时使用的索引数据划分成多个部分，并将该多个部分进行压缩以获得压缩文件；和

一个存储步骤，该步骤把压缩文件和压缩后该多个部分的地址信息存储在存储介质中。

2. 如权利要求1所述的文件处理方法，其中每一部分均为定长。

3. 如权利要求1所述的文件处理方法，其中每一部分均为变长，并且所述存储步骤还把地址信息在压缩数据之前存储在存储介质中。

4. 如权利要求1－3的任一项所述的文件处理方法，其中文件处理方法还包括：

一个恢复步骤，该步骤从存储介质中读取压缩文件并对每个部分进行扩展，以恢复数据和索引数据。

5. 如权利要求4所述的文件处理方法，该方法还包括：

一个用于把所恢复的数据和索引数据存储到一个辅助存储单元中的步骤。

6. 如权利要求4所述的文件处理方法，其中所述压缩步骤使用了对于每个部分的数据和索引数据所共同的压缩算法和压缩参数。

7. 如权利要求1所述的文件处理方法，其中所述压缩步骤根据第一位代码的出现频率在数据中选择预定数目的第一位代码，把余下的未被选中的第一位代码分解成第二位代码，根据对第二位代码进行选择的结果创建转换表，其中对第二位代码进行的选择是根据第二位代码的出现频率执行的，并且根据转换表实现数据压缩。

8. 如权利要求1所述的文件处理方法，其中数据包括词典数据。

9. 一种数据处理装置，包括：

压缩装置，该装置用于把数据和对应于该数据并且在查找或搜索该数据时使用的索引数据划分成多个部分，并将该多个部分进行压缩

以获得压缩文件；和

存储装置，该装置用于把压缩文件和压缩后该多个部分的地址信息存储在存储介质中。

10. 如权利要求 9 中所述的数据处理装置，其中每一部分均为定长。

11. 如权利要求 9 中所述的数据处理装置，其中每一部分均为变长，并且所述存储装置还在压缩数据之前把地址信息存储在存储介质中。

12. 如权利要求 9－11 的任一项所述的数据处理装置，该装置还包括：

恢复装置，用于从存储介质中读取压缩文件并对各部分进行扩展，以恢复数据和索引数据。

13. 如权利要求 12 中所述的数据处理装置，该装置还包括：

用于把所恢复的数据和索引数据存储在一个辅助存储单元中的装置。

14. 如权利要求 12 的所述的数据处理装置，其中所述压缩装置使用了各部分的数据和索引数据所共同的压缩算法和压缩参数。

15. 如权利要求 9 所述的数据处理装置，其中所述压缩装置根据第一位代码的出现频率在数据中选择预定数目的第一位代码，把余下的未被选中的第一位代码分解成第二位代码，根据对第二位代码进行选择的结果创建转换表，其中对第二位代码进行的选择是根据第二位代码的出现频率执行的，并且根据转换表实现数据压缩。

16. 如权利要求 9 所述的数据处理装置，其中数据包括词典数据。

文件处理方法和数据处理装置

技术领域

本发明一般涉及文件处理方法，数据处理装置和存储介质，尤其涉及一种文件处理方法和数据处理装置，用于压缩文件，例如压缩涉及一个或多个词典，百科全书等的词典文件，把压缩文件存储在存储介质中并从存储介质中读取所存储的文件，并把所读取的文件读到一种存储介质中，该存储介质中存储了文件，例如压缩的词典文件。

背景技术

近来，存在例如 CD-ROM 的存储介质，其中 CD-ROM 可以预先存储有关于词典，百科全书等的信息。通过从计算机访问这种 CD-ROM，可以读取并显示有关于词典，百科全书等的信息。结果是，有关于词典，百科全书等的大量信息可以存储在一张极其紧密的 CD-ROM 中。另外，当使用计算机时不必打开词典，百科全书等来获取必要的信息，可从 CD-ROM 中读取必要的信息，从而可大大的节省时间和减少麻烦。

常规的用于存储有关于词典，百科全书等的信息的 CD-ROM 中，词典文件由词典数据和有关于索引的数据（以后称为索引数据）组成。例如，在存储百科全书的情况下，词典数据包括有关于文本的数据（以后称为文本数据），其中文本用于解释单词的意思，有关于一个图象的数据（以后称为图象数据），其中如果单词描述的是一个动物则这一图象数据显示出一个动物，有关于一种声音的数据（以后称为声音数据），例如一种鸟的叫声，如果单词描述的是这种鸟的话，等。另一方面，索引用于从词典文件中检索所需的词典数据，并且根据词典数据提供索引。索引有时也指关键字。索引数据包括有关于标题的指针，有关于项的指针等。有关于标题的数据包括中心词。另外，有关于项的数据包括中心词，注释等。

目前，由于 CD-ROM 的存储能力是比较大的，因此文本数据和索引数据未经压缩就可存储在一张 CD-ROM 中。另一方面，声音数据尤其是图象数据中所包括的信息量非常大，因此在这些数据在被存储在 CD-ROM 中之前，根据适当的压缩技术分别对声音数据和图象数据进行压缩。

然而，如果每个词典或百科全书都需要一个 CD-ROM，则不能很好的利用词典数据。因此，需要把有关于多个词典，百科全书等的信息存储在一张 CD-ROM 中，但是由于即使词典数据已被压缩过，所需存储的信息量仍可能超过单张 CD-ROM 的存储量，因此就存在了问题。另外，甚至于在需要存储在 CD-ROM 中的词典文件只是有关于一个词典，百科全书等的情况下，当词典文件的信息量增加时，即使词典数据已被压缩过，所需存储的信息量仍可能超过单张 CD-ROM 的存储量，。

相应地，可考虑在把有关于词典，百科全书等的信息存储在 CD-ROM 中时，不仅需要对词典数据进行压缩，而且需要对整个词典文件包括索引数据进行压缩。但是到目前为止还没有提出一种可通过比较简单的技术来对整个词典文件进行有效的压缩，并且能够在短时间内对压缩的词典文件进行扩展的方法。特别是在词典，百科全书等的情况下，在这些情况下有关于索引数据的信息量非常大。因此，当对压缩的词典文件进行扩展时，如果花费很长的时间来执行索引数据的恢复处理，则访问所需要的索引数据或词典数据的时间将会变长，从而降低了词典，百科全书等的方便性。

而且，当把词典数据压缩成索引项的方式或定长的方式时，例如，由于有关于索引数据的信息量很大，特别是在词典，百科全书等的情况下，因此会花费很长的时间来执行对压缩词典文件的扩展处理，从而同样地减低了词典，百科全书的方便性，例如，日本公开专利申请 NO.9-26969 提出了一种电话号码簿检索系统，该系统使用了与上述方法类似的一种方法。然而，该系统所提出的方法并未对索引数据进行压缩。在电话号码簿的情况中，有关于索引数据的信息量与有关于相

应于词典数据的电话号码，姓，名，公司名和地址相比要小的多。因此，即使对电话号码簿的索引数据进行了压缩，整个信息压缩的效率也不会大大地提高。因此，即使把该系统所提出的方法应用于把有关于词典，百科全书等存储在存储介质中的存储中，也不会如所设想的那样大大地提高词典文件的整体的信息压缩效率。

相应地，在有关于索引数据的信息量非常大的情况下，甚至在这种信息量同有关于词典数据例如词典，百科全书等的信息量相比也是非常大的情况下，由于目前还不能有效地对词典文件进行压缩并把压缩文件存储在存储介质中，并且还不能通过比较简单的处理在短时间内访问压缩的词典文件，因此就存在有问题。

发明内容

因此，本发明的一个目的在于提供一种文件处理方法，一种数据处理装置和一种存储介质，利用这些方法，装置和介质可有效地对词典文件进行压缩并把它存储在存储介质中，并且可通过比较简单的处理在短时间内访问压缩的词典文件，甚至于在有关于索引数据的信息量与有关词典数据例如词典，百科全书等的信息量相比也是很大的情况下。

本发明提供一种文件处理方法，包括：一个压缩步骤，该步骤把数据和对应于该数据并且在查找或搜索该数据时使用的索引数据划分成多个部分，并将该多个部分进行压缩以获得压缩文件；和一个存储步骤，该步骤把压缩文件和压缩后该多个部分的地址信息存储在存储介质中。

本发明提供一种数据处理装置，包括：压缩装置，该装置用于把数据和对应于该数据并且在查找或搜索该数据时使用的索引数据划分成多个部分，并将该多个部分进行压缩以获得压缩文件；和存储装置，该装置用于把压缩文件和压缩后该多个部分的地址信息存储在存储介质中。

本发明的另一个目的在于提供一种文件处理方法，该方法包括一个压缩步骤和一个存储步骤，其中压缩步骤把数据和关于该数据的索

引数据划分成多个部分，并分别对这些部分进行压缩以获得压缩文件，其中存储步骤把压缩文件和经过压缩后的各部分的地址信息存储在存储介质中。根据本发明，可有效地对文件例如词典文件进行压缩并把它存储在存储介质中，其中词典文件由包括索引，每一项的文本等的数据组成。另外，可通过对每个部分的压缩文件进行扩展，从而可通过比较简单的处理快速地进行文件检索。

当每一部分都是定长时，在压缩文件中就不必在压缩数据之前包括地址信息，从而可提高数据压缩效率。另一方面，当每一部分均是变长时，则所述存储步骤还应在压缩数据之前存储地址信息，从而可以根据数据类型和部分把该部分设置为适当的长度，从而可快速的进行数据扩展。

当文件处理方法还包括一个恢复步骤时，可通过使用一个辅助存储单元来提高文件检索速度，其中恢复步骤从存储介质中读取压缩文件并对压缩文件的每一部分进行扩展以恢复数据和索引数据，其中辅助存储单元可进行快速数据访问并把恢复的数据和索引数据存储在辅助存储单元中。

当压缩步骤使用了每一部分的数据和索引数据所共同的压缩算法和压缩参数时，则可以在数据扩展时通过使用共同的压缩算法和压缩参数来简化数据压缩处理和数据扩展处理。尤其是，可以使用哈夫曼编码，通用编码等作为压缩算法。

本发明的另一个目的在于提供一种文件处理方法，该方法包括一个读取步骤和一个恢复步骤，其中读取步骤从存储介质中读取压缩文件和压缩后每部分的地址信息，对于每个部分来说，所述压缩文件是通过把数据和关于该数据的索引数据划分成多个部分并分别对各个部分进行压缩后得到的，其中恢复步骤对压缩文件进行扩展并恢复数据和索引数据。根据本发明，通过对每一部分执行压缩文件例如压缩的词典文件的扩展，可通过比较简单的处理来实现快速文件检索。

本发明的另一个目的在于提供一种数据处理装置，该装置包括压缩装置和存储装置，其中压缩装置用于把数据和关于该数据的索引数

据划分成多个部分，并分别对这些部分进行压缩以获得压缩文件，其中存储装置用于把压缩文件和压缩后各部分的地址信息存储在存储介质中。根据本发明，可有效地对由包括索引，各项的文本等的数据组成文件进行压缩并把压缩文件存储在存储介质中。另外，通过对压缩文件的每一部分进行扩展，可通过使用比较简单的处理来实现快速文件检索。

本发明的另一个目的在于提供一种数据处理装置，该装置包括读取装置和恢复装置，其中读取装置用于从存储介质中读取压缩文件和压缩后每部分的地址信息，其中所述压缩文件是把数据和关于该数据的索引数据划分成多个部分，并分别对这些部分进行压缩后获得的，其中恢复装置用于扩展压缩文件并恢复数据和索引数据。根据本发明，通过对压缩文件的每一部分进行扩展，可通过使用比较简单的处理来实现快速文件检索。

本发明的另一个目的在于提供一种用于存储计算机可读信息的存储介质，该存储介质包括读取装置和恢复装置，其中读取装置用于使得计算机从存储介质中读取压缩文件和压缩后每部分的地址信息，对于每一部分来讲，所述压缩文件是把数据和关于该数据的索引数据划分成多个部分，并对这些部分进行压缩后获得的，其中恢复装置用于使得计算机对压缩文件进行扩展并恢复数据和索引数据。根据本发明，通过对压缩文件的每一部分进行扩展，可通过使用比较简单的处理来实现快速文件检索。

本发明的另一个目的在于提供一种用于存储计算机可读信息的存储介质，其中该信息包括压缩后每部分的地址信息及压缩文件，对于每一部分来讲，所述压缩文件是通过把数据和关于该数据的索引数据划分成多个部分，并对这些部分进行压缩后得到的，其中是使用了每部分所共同的压缩算法和压缩参数对所述压缩文件进行压缩的。根据本发明，可有效地对文件进行压缩并把它存储在存储介质中。根据本发明，通过对压缩文件的每一部分进行扩展，可使用比较简单的处理来实现快速文件恢复。

本发明的另一个目的在于提供一种用于存储计算机可读信息的存储介质，包括计算机程序，该程序使得计算机执行压缩程序和存储程序，其中压缩程序用于根据词典数据把词典数据和数据划分成多个部分，并对各个部分进行压缩以获得压缩的词典文件，其中存储程序用于把压缩的词典文件和压缩后各部分的地址信息存储在存储介质中。根据本发明，可以通过执行比较简单的处理来快速地检索文件。

本发明的另一个目的在于提供一种用于存储压缩文件的计算机可读存储介质，其中存储介质包括一个压缩数据区，一个地址信息区和一个压缩参数区，其中压缩数据区用于存储通过把数据和关于该数据的索引数据划分成多个部分并对各个部分进行压缩后所获得的压缩数据，其中地址信息区用于存储压缩后各部分的地址信息，其中压缩参数区用于存储用于压缩的压缩参数。根据本发明，可通过执行比较简单的处理来地检索文件。

因此，根据本发明，即使在有关于索引数据的信息量比有关于词典数据例如词典，百科全书等的信息量还大的情况下，也可有效地对文件例如词典文件进行压缩并把它存储在存储介质中，并且可通过执行比较简单的处理在短时间内访问该文件，例如压缩的词典文件。

通过结合附图阅读下面的详细描述可清楚地了解到本发明的其他一些目的或特点。

附图说明

图 1 为一个表示使用一种文件处理方法的实施例的计算机系统的一般结构的系统框图；

图 2 为一个用于说明由 CPU 执行的压缩参数计算处理的流程图；

图 3 为一个表示压缩参数的数据结构的框图；

图 4 为一个用于说明由 CPU 执行的数据压缩处理和地址信息计算处理的流程图；

图 5 为一个用于说明由 CPU 执行的压缩文件编制处理和压缩文件存储处理的流程图；

图 6 为一个说明压缩文件的组成的框图；

图 7 为一个用于说明由 CPU 执行的索引读取处理的流程图；

图 8 为一个用于说明由 CPU 执行的数据扩展处理；和

图 9 为一个用于说明由 CPU 执行的文本读取处理的流程图。

具体实施方式

下面将描述根据本发明的一种文件处理方法的实施例和一种根据本发明的数据处理装置的实施例。这种文件处理方法的实施例和数据处理装置的实施例都使用了根据本发明的一种存储介质的实施例。在存储介质的实施例中，本发明应用于 CD-ROM。然而，除了 CD-ROM，本发明也适用于光信息存储介质，磁光存储介质例如磁光盘，磁性存储介质例如软盘，和多种半导体存储器件。

图 1 为一个系统框图，表示应用这种文件处理方法的实施例的计算机系统的一般结构，并且图 1 适用于数据处理装置的实施例。图 1 中所示的计算机系统一般包括一个中央处理器（CPU）1，一个由随机访问存储器（RAM）等组成的主存单元 2，一个由硬盘驱动器等组成的辅助存储单元，一个由键盘，鼠标等组成的输入设备 4，一个显示单元 5，和一个由 CD-ROM 驱动器等组成的 CD-ROM 输入/输出设备，这些组成部分通过总线 9 相互耦合。构成计算机系统的每一部分均可看作具有已知结构。

输入设备 4 用于把指令和数据输入到 CPU 1。CPU 1 根据所输入的指令和数据通过执行存储在辅助存储单元 3 中的程序来实现用户所要求的处理。存储在辅助存储单元 3 中的程序可能是预先安装的，或者是从 CD-ROM 6a 中装载的，其中 CD-ROM 6a 被装入到 CD-ROM 输入/输出设备 6 中。主存单元 2 用于暂时地存储由 CPU 1 执行的计算处理的中间结果等，操作所使用的数据等。显示单元 5 用于显示由 CPU 1 所执行的处理的结果，显示一条催促用户输入指令或数据的信息。也可以连接一个打印机（图中未示出）来替代显示单元 5，或除了显示单元 5 外再连接一个打印机，其中打印机用于打印由 CPU 1 等执行的处理的结果。

首先，将描述用于把词典，百科全书等的词典文件存储在

CD-ROM 6a 中的文件存储处理，其中 CD-ROM 6a 被装入到 CD-ROM 输入/输出设备 6。文件存储处理一般包括一个压缩参数计算处理，一个用于对索引，文本等进行压缩的数据压缩处理，一个地址信息计算处理，一个压缩文件编制处理和一个压缩文件存储处理。在该实施例中，为方便起见假设使得 CPU 1 执行文件存储处理的程序存储在 CD-ROM 6a 中，并且 CPU 1 使用一种已知的方法从 CD-ROM 6a 中读取该程序，并把该程序读到辅助存储单元 3。另外，为方便起见假设有关于词典，百科全书等的词典文件是从主机单元（图中未示出）传送过来并且通过总线 9 存储在辅助存储单元 3 中的，或者是通过 CD-ROM 输入/输出设备 6 从不同于 CD-ROM 6a 的其他 CD-ROM 中读取并通过总线 9 存储在辅助存储单元中的。

1a)压缩参数计算处理：

图 2 为一个流程图，用于说明由 CPU 1 执行的压缩参数计算处理。图 2 中，步骤 S1 访问辅助存储单元 3 并打开一个词典文件。步骤 S2 读取一个字符，也即一个 16 位代码，例如从该词典文件中读取。步骤 S3 通过使用 CPU 1 中的一个出现频率计数器来计算所读取的 16 位代码的出现频率。步骤 S4 判定是否已处理了该词典文件的最后一个字符，如果步骤 S4 中的判定结果是否，则处理返回步骤 S2。

另一方面，如果步骤 S4 中的判定结果是是，则步骤 S5 关闭这一词典文件。步骤 S6 根据各字符的出现频率对 16 位代码进行排序，并且步骤 S7 选择出 1024 个 16 位代码，例如，根据出现频率进行选择。步骤 S8 把余下未被选中的 16 位代码分解成 8 位代码，并计算出这些 8 位代码的出现频率。步骤 S9 根据 16 位代码的出现频率，通过把 8 位代码的出现频率设置为大概为 16 位代码的出现频率的一半来校正 8 位代码的出现频率。

步骤 S10 打开辅助存储单元 3 中的一个用于压缩参数的压缩参数保存文件。步骤 S11 把 1024 个 16 位代码及其各自的出现频率写入该压缩参数保存文件。另外，步骤 S12 把 256 个 8 位代码及其各自的出现频率写入该压缩参数保存文件。步骤 S13 关闭该压缩参数保存文件，

然后处理结束。

图 3 为一个表示压缩参数的数据结构的框图。例如，如图 3 所示，在使用哈夫曼编码进行压缩的情况下，所选中的 1024 个 16 位代码均有 256 种出现频率，对于所选中的 8 位代码也均有 256 种出现频率。这些出现频率成为用于生成哈夫曼树的数据。在使用通用代码进行压缩的情况下，压缩参数包括一个重试树或数据例如用于生成重试树的注册符号例子及其参数。

1b)数据压缩处理和地址信息计算处理：

图 4 为一流程图，用于说明由 CPU 1 执行的数据压缩处理和地址信息计算处理。图 4 中，步骤 S21 根据 8 位和 16 位代码的出现频率创建一个转换表，由于本实施例实行哈夫曼压缩，也即，创建一棵哈夫曼树。 步骤 S22 打开一个辅助存储单元 3 中的词典文件。步骤 S23 打开辅助存储单元 3 中的一个用于压缩数据的压缩数据保存文件和一个用于地址信息的地址信息保存文件。

步骤 S24 从该词典文件中读取 1 部分。该部分可能为定长或变长，但在该实施例中，为方便起见假设该部分为定长。有时该部分指一块。步骤 S25 使用哈夫曼树来计算 1 部分的压缩数据。步骤 S26 把一个结束代码添加到 1 部分的末尾。另外，步骤 S27 把压缩数据写入压缩数据保存文件。

步骤 S28 计算有关于地址的地址信息，其中地址中存储有上述 1 部分。例如，当上述 1 部分具有定长时，则根据分配给每个部分的部分编号来计算地址信息。步骤 S29 把地址信息写入地址信息保存文件。步骤 S30 判定是否处理了最后一部分，如果该步骤 S30 的判定结果是否，则处理返回到步骤 S24。例如，有可能根据添加到部分编号上或最后一部分上的最后部分代码来判定是否已处理了最后一部分。

另一方面,如果步骤 S30 的判定结果是是，则步骤 S31 关闭用于压缩数据的保存文件，并且关闭地址信息保存文件。另外，步骤 S32 关闭词典文件，然后处理结束。

1c)压缩文件编制处理和压缩文件存储处理：

图 5 为一个流程图，用于说明由 CPU 1 所执行的压缩文件编制处理和压缩文件存储处理。图 5 中，步骤 S41 打开一个辅助存储单元 3 中的压缩文件。步骤 S42 打开一个辅助存储单元中的压缩参数保存文件，并且步骤 S43 把压缩参数保存文件中的压缩参数复制到压缩文件中。步骤 S44 关闭压缩参数保存文件。

步骤 S45 打开一个辅助存储单元 3 中的地址信息保存文件，并且步骤 S46 把地址信息保存文件中的地址信息复制到压缩文件中。步骤 S47 关闭地址信息保存文件。而且，步骤 S48 打开一个辅助存储单元 3 中的压缩数据保存文件，步骤 S49 把压缩数据保存文件中的压缩数据复制到压缩文件中。步骤 S50 关闭压缩数据保存文件。步骤 S51 通过 CD-ROM 输入/输出设备 6 把压缩文件存储到 CD-ROM 6a 中。另外，步骤 S52 关闭压缩文件，然后结束处理。

图 6 为一个框图，用于根据上述 1a)压缩参数计算处理 1b)数据压缩处理和地址信息计算处理 1c) 压缩文件编制处理和压缩文件存储处理来说明压缩文件的编制。图 6 中(a)表示压缩参数。在该实施例中，压缩参数用于使用哈夫曼编码进行压缩。图 6 中，(b)表示词典文件的各部分。例如该实施例中，每一部分均由 2K 字节构成，且每一部分均由词典数据和索引数据组成。例如，在百科全书的情况下，词典数据包括有关于文本的文本数据，有关于图象的图象数据，有关于声音的声音数据等，其中文本用于解释单词的意思，如果单词描述的是一个动物，则图象用于表示这种动物，如果单词表示的是一个小鸟，则声音用于表示这种小鸟的叫声。另一方面，索引用于从词典文件中检索所需要的词典数据，并且根据词典数据来安排索引。索引数据包括有关于标题的指针，有关于项的指针等。有关于标题的数据包括一个中心词，注释等。

图 6 中，(c)表示压缩数据，其中每一部分均为定长或变长，且已被压缩。而且，图 6 中，(d)表示根据每一部分所计算出的地址信息，并且(e)表示压缩文件，其中压缩文件是通过对地址信息和压缩数据进行编制并在压缩文件的头部添加管理信息后得到的。管理信息包括检

索压缩文件时所用到的信息，例如词典文件名，词典文件类型，对词典文件所用的压缩类型。

接下来，将描述文件检索处理，这种文件检索处理通过读取存储在 CD-ROM 6a 中的压缩文件来检索所需要的数据，其中 CD-ROM 6a 被装入在 CD-ROM 输入/输出设备 6 中。文件检索处理一般包括一个索引读取处理和一个文本读取处理，并且文件索引处理是通过调用一个数据扩展处理来实现的。在该实施例中，为方便起见假设使得 CPU 1 执行文件检索处理的程序存储在 CD-ROM 6a 中，并且 CPU 1 通过已知方法从 CD-ROM 6a 中读取该程序和压缩文件，并把所读取的程序和压缩文件送到辅助存储单元 3。

2a)索引读取处理:

图 7 为一流程图，用于说明由 CPU 1 执行的索引读取处理。在图 7 中，步骤 S61 根据用户通过输入设备 4 所输入的索引数据，设置最重要索引的地址信息。步骤 S62 调用扩展处理，并从存储在辅助存储单元 3 中的程序中读取一个例程以执行扩展处理，并且步骤 S62 使得 CPU 1 执行文件检索处理，以对压缩文件中最重要索引的地址进行扩展。步骤 S63 根据索引数据获得重要索引的地址，也即，最重要索引的头字符。步骤 S64 调用扩展处理，并对压缩文件中重要索引的地址进行扩展。步骤 S65 根据索引数据获得下一层中次重要索引的地址。步骤 S66 调用扩展处理，并将位于压缩文件中的上述下一层的不太重要索引的地址进行扩展。步骤 S67 判定是否已对最不重要的索引的地址进行了扩展，如果步骤 S67 中的判定结果是否，则处理返回到步骤 S67。另一方面，如果如果步骤 S67 中的判定结果是是，则该处理结束。

2b)数据扩展处理:

图 8 为一个流程图，用于说明由 CPU 1 执行的数据扩展处理。数据扩展处理由索引读取处理和文本读取处理所调用。在图 8 中，步骤 S71 根据用户通过输入设备 4 所输入的索引数据把所请求的扩展地址，数据大小和存储区域存储到辅助存储单元 3 中，以根据扩展后数据大小在辅助存储单元 3 中预备一个足够大的存储区域。步骤 S72 判定从

CD-ROM 6a 中读取并且送到辅助存储单元 3 中的压缩文件是否已打开。如果步骤 S72 中的判定结果是否，则步骤 S73 打开辅助存储单元 3 中的压缩文件。步骤 S74 从压缩文件中读取压缩参数，并且读取压缩参数中的 8 位代码的出现频率，压缩参数中的 16 位代码及其 16 位代码的出现频率。步骤 S75 根据 8 位代码的出现频率和 16 位代码的出现频率创建哈夫曼树，然后处理进行到后面将描述到的步骤 S76。用于判定代码是8位代码还是16位代码的判定标识被添加到哈夫曼树的叶节点的数据上。

如果步骤 S72 中的判定结果是是或在步骤 S75 后，步骤 S76 从压缩文件中读取相应于所请求的扩展地址的地址信息。步骤 S77 根据地址信息，从压缩文件中读取相应的压缩数据的部分。步骤 S78 通过使用哈夫曼树来对压缩数据的部分进行扩展。步骤 S79 根据判定标识把扩展数据复制到上述存储区域，其中判定标识表示代码是 8 位代码还是 16 位代码。而且，步骤 S80 根据压缩文件判定是否完成了所请求数据大小的扩展。

如果步骤 S80 中的判定结果是否，则步骤 S81 从压缩文件读取相应于下一部分的扩展地址的地址信息。步骤 S82 根据相应于下一部分的扩展地址的地址信息，从压缩文件中读取相应的压缩数据的部分，然后处理返回到步骤 S78。另一方面，如果步骤 S80 中的判定结果是是，则处理结束。

2c)文本读取处理:

图 9 为一流程图，用于说明由 CPU 1 所执行的文本读取处理。图 9 中，步骤 S91 根据用户通过输入设备 4 所输入的索引数据，计算扩展数据中与索引相匹配的项数。步骤 S92 根据所输入的索引数据把索引的项指针的值指向地址。步骤 S93 调用扩展处理，并从存储在辅助存储单元 3 中的程序中读取一个例程以执行扩展处理，并使得 CPU 1 执行文件检索处理，以对压缩文件中由项指针所指示的文本进行扩展，也即，对 1 部分的词典数据进行扩展。

步骤 S94 判断进项指针所表示的词典数据是否已经结束。如果步

骤 S94 中的判定结果是否，则步骤 S95 设置下个 1 部分的地址。另外，步骤 S96 调用扩展处理，并对由压缩文件中的项指针所指示的下个 1 部分中的词典数据进行扩展，然后处理返回到步骤 S94。另一方面，如果步骤 S94 中的判定结果是是，则步骤 S97 根据所输入的索引数据判定是否对于所有项的处理均结束了。如果步骤 S97 中的判定结果是否，则处理返回到步骤 S92。另一方面，如果步骤 S97 中的判定结果是是，则步骤 S98 把扩展后词典数据显示在显示单元 5 上，其中词典数据的所有项均得到了扩展，然后处理结束。

可以在步骤 S97 之前执行步骤 S98。在这种情况下，每当对各项进行扩展后，步骤 S98 就把扩展后的词典数据显示在显示单元 5 上。

在上述实施例中，为方便起见假设每一部分均为定长。在这种情况下，数据压缩的效率是令人满意的，并且可以从压缩文件中恢复地址信息，而不必把地址信息存储在压缩文件中该部分的压缩数据之前。这是由于部分具有定长，并且对每一部分均添加了部分编号，因此可以计算出每一部分关于其他部分的相对位置。

另一方面，当部分具有变长时，则可以提高数据扩展速度。这是由于可根据数据和部分的类型来适当地设置部分的长度，从而不需对附加数据进行扩展。在部分具有变长的情况下，在压缩文件中需要把地址存储在该部分的压缩数据之前。相应地，可以根据是要提高数据压缩效率还是要提高数据扩展的速度，从而决定把部分设置成定长还是变长。

另外，CD-ROM 6a 中可存储一个或多个词典文件。当 CD-ROM 6a 中存储有有关于多个词典，百科全书等的多个词典文件时，可以通过使用图 6 中的管理信息（e）中的词典文件名或词典文件类型来指定所要检索的词典。

而且，虽然上述实施例中使用了哈夫曼编码进行数据压缩，但是除了哈夫曼编码技术，还可使用其他的编码技术，例如统一编码技术，只要所用的数据压缩技术能够通过使用每一部分所共同的压缩参数对词典数据有效地进行压缩。另外，进行压缩和扩展的数据并不仅限于

词典数据，还可包括包括索引和数据的数据库数据。

而且，上述实施例中，文件检索处理是通过把用于执行文件检索处理的程序和压缩文件复制到辅助存储单元 3 中实现的。然而，除了把程序和压缩文件复制到辅助存储单元 3 中外，还可以在主存单元 2 中展开程序和压缩文件，并执行类似于上述处理的处理。

通过使用上述实施例中的压缩算法，与使用 8 位哈夫曼编码的一般的数据压缩处理相比可提高数据压缩效率。结果是，可以减少存储在存储介质例如 CD-ROM 和用做辅助存储单元的硬盘中的压缩文件的区域。通过使用这种压缩算法，虽然可以提高数据压缩的效率，但是对压缩文件进行扩展所需要的处理时间与使用哈夫曼编码的普通数据压缩处理对压缩文件进行扩展所需要的处理时间相比并没有太大的改变。

实现数据检索处理所需要的时间由读取单元（驱动器）的搜索时间，读取压缩文件所需的读取时间，执行扩展处理所需的时间构成。

由于通过使用上述压缩算法可提高数据压缩效率，因此存储在存储介质中的压缩文件所减少的区域可减少文件检索处理的所需要的搜索时间。因此提高了文件检索速度。当硬件性能提高时，则文件检索速度提高所造成的效果会更显著。

而且，本发明并不仅限于这些实施例，但是可对此做出不同的修改和变更，而不背离本发明的范围。

图 1

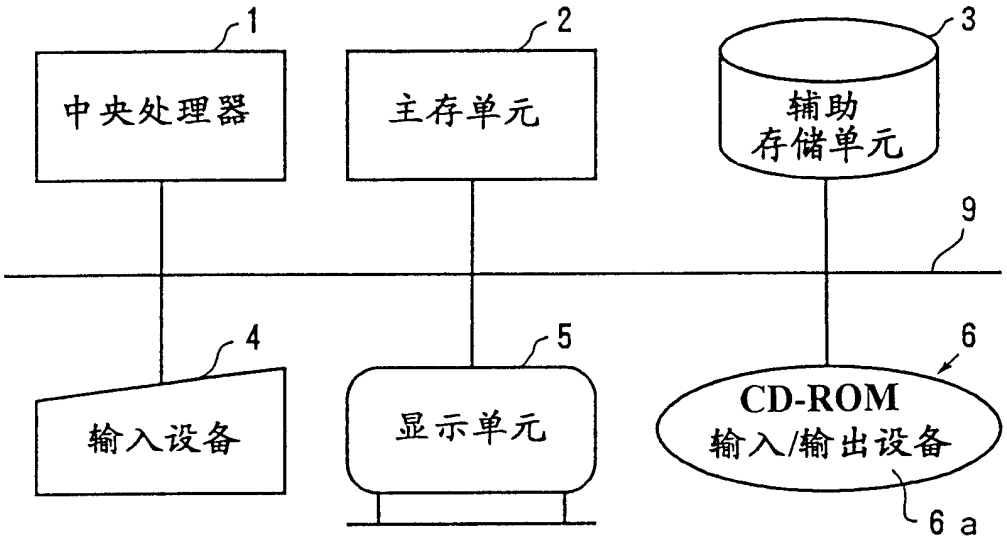


图 2

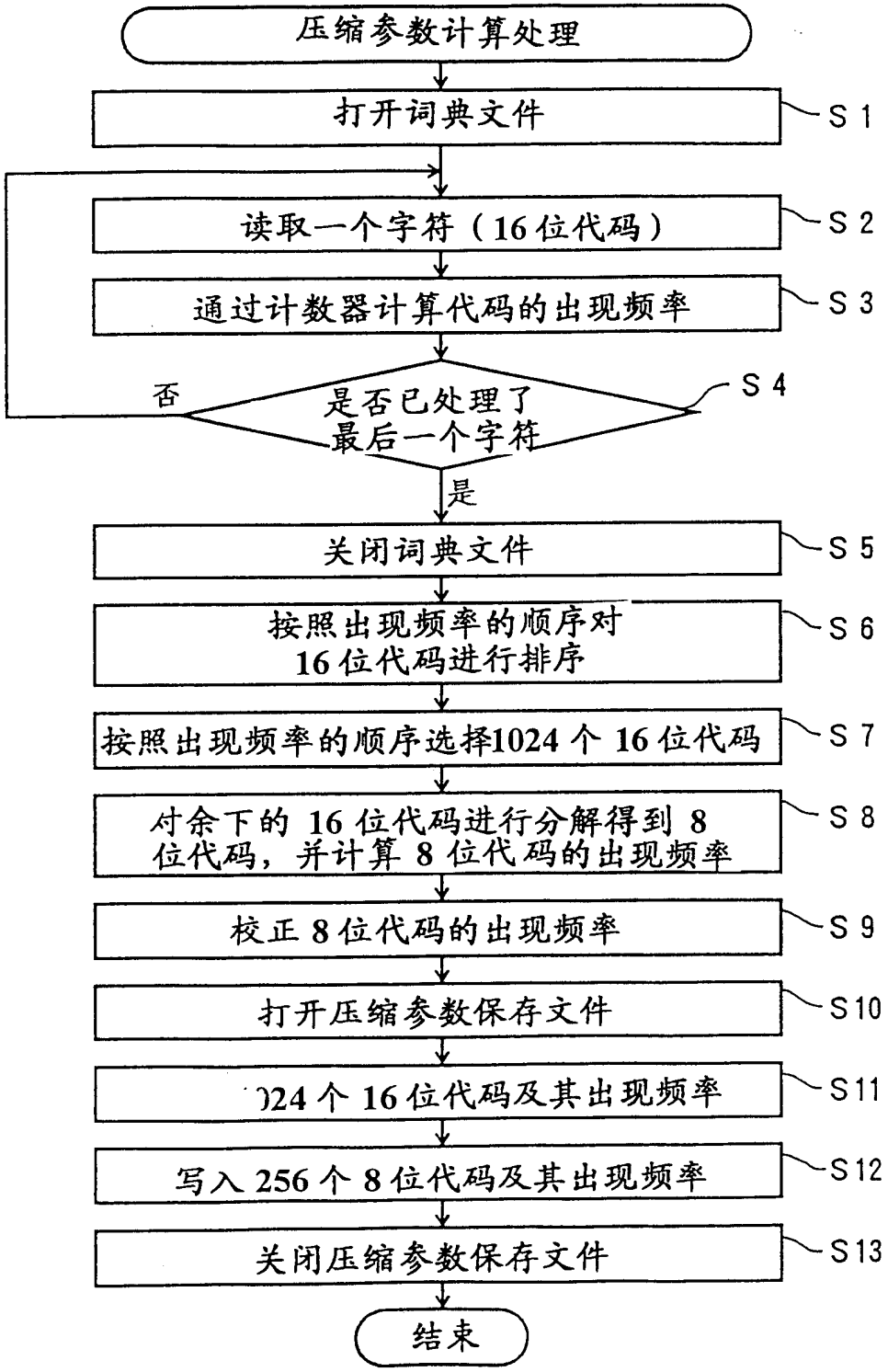

压缩参数计算处理
打开词典文件
S 1
读取一个字符（16 位代码）
S 2
通过计数器计算代码的出现频率
S 3
是否已处理了
最后一个字符
S 4
否
是
关闭词典文件
S 5
按照出现频率的顺序对
16 位代码进行排序
S 6
按照出现频率的顺序选择1024 个 16 位代码
S 7
对余下的 16 位代码进行分解得到 8
位代码，并计算 8 位代码的出现频率
S 8
校正 8 位代码的出现频率
S 9
打开压缩参数保存文件
S10
)24 个 16 位代码及其出现频率
S11
写入 256 个 8 位代码及其出现频率
S12
关闭压缩参数保存文件
S13
结束

图 3

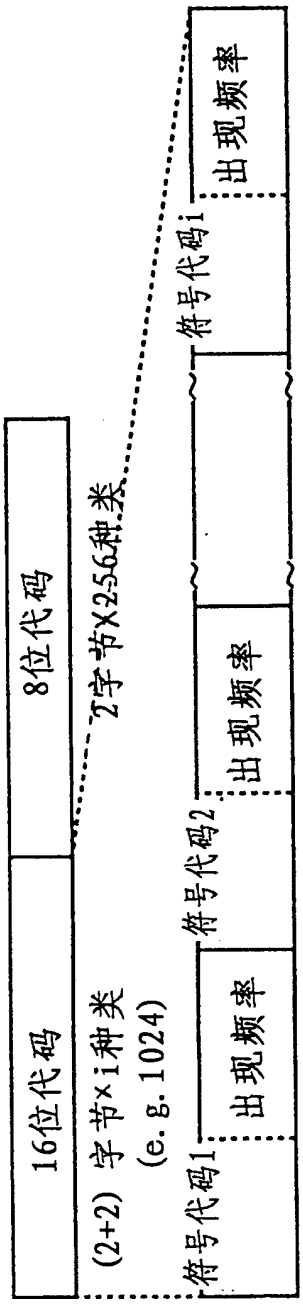

16位代码
8位代码
(2+2) 字节×i种类
(e.g. 1024)
2字节X256种类
符号代码1
出现频率
符号代码2
出现频率
符号代码i
出现频率

图 4

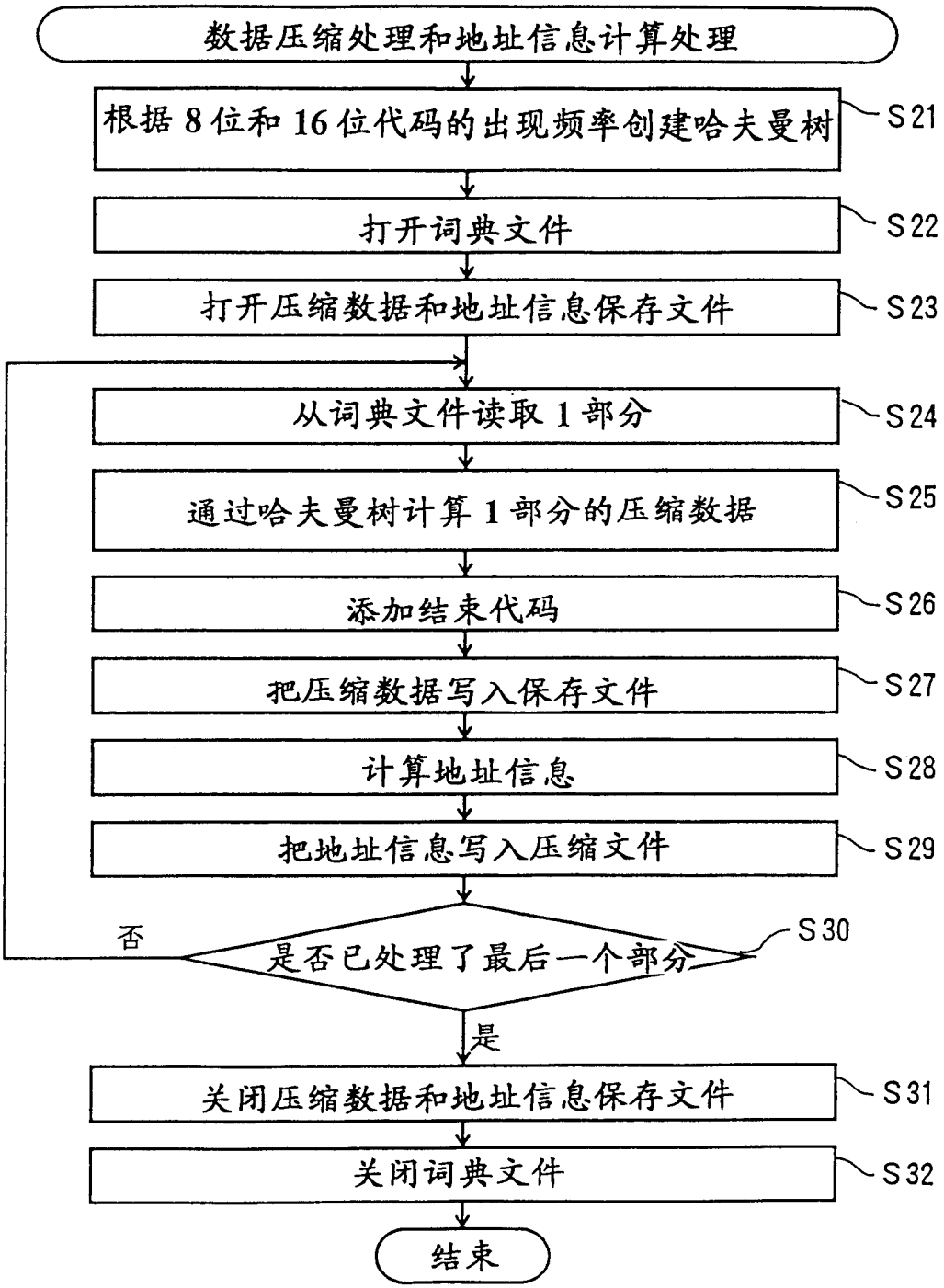

数据压缩处理和地址信息计算处理
根据 8 位和 16 位代码的出现频率创建哈夫曼树
S21
打开词典文件
S22
打开压缩数据和地址信息保存文件
S23
从词典文件读取 1 部分
S24
通过哈夫曼树计算 1 部分的压缩数据
S25
添加结束代码
S26
把压缩数据写入保存文件
S27
计算地址信息
S28
把地址信息写入压缩文件
S29
是否已处理了最后一个部分
S30
否
是
关闭压缩数据和地址信息保存文件
S31
关闭词典文件
S32
结束

图 5

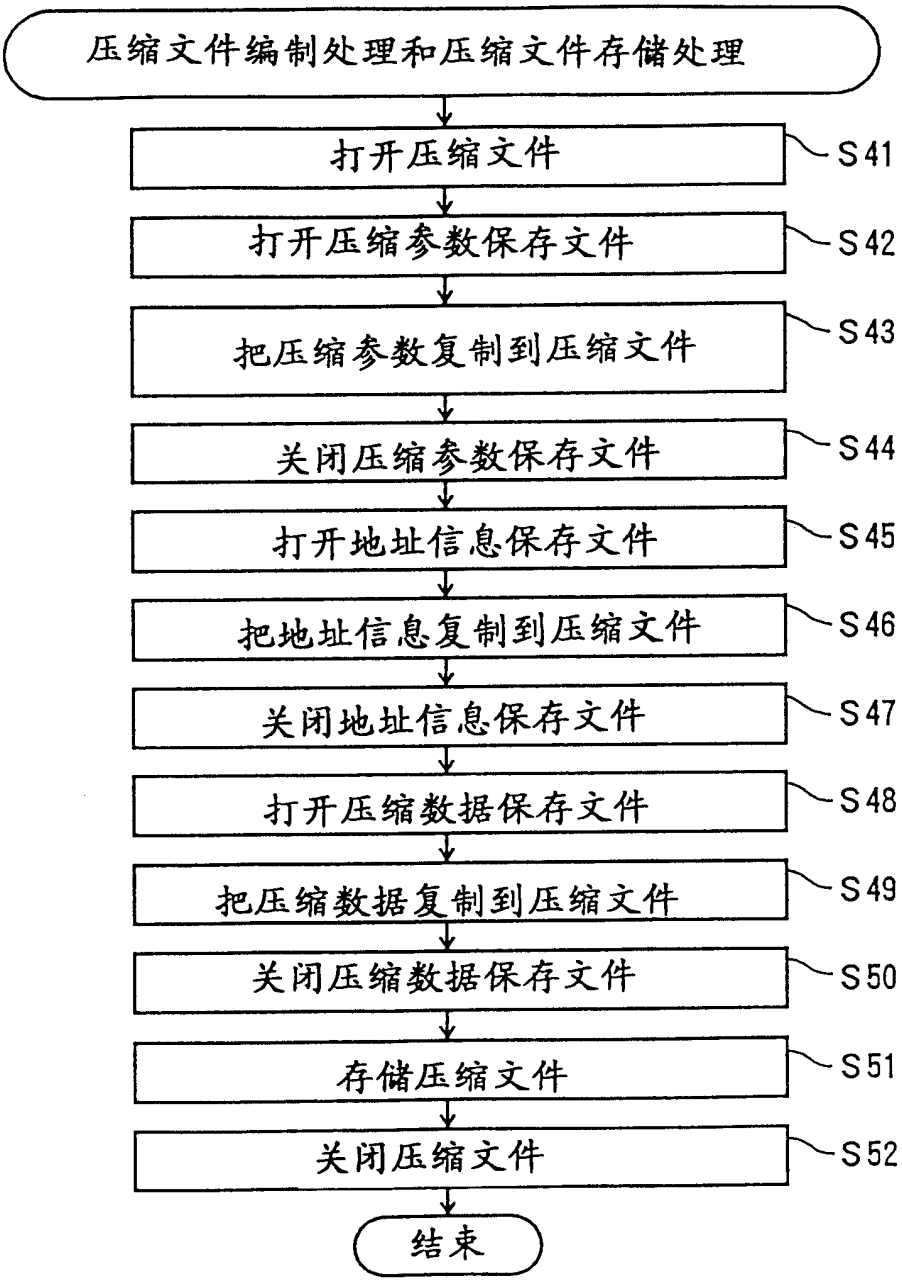

压缩文件编制处理和压缩文件存储处理
打开压缩文件
S41
打开压缩参数保存文件
S42
把压缩参数复制到压缩文件
S43
关闭压缩参数保存文件
S44
打开地址信息保存文件
S45
把地址信息复制到压缩文件
S46
关闭地址信息保存文件
S47
打开压缩数据保存文件
S48
把压缩数据复制到压缩文件
S49
关闭压缩数据保存文件
S50
存储压缩文件
S51
关闭压缩文件
S52
结束

图 6

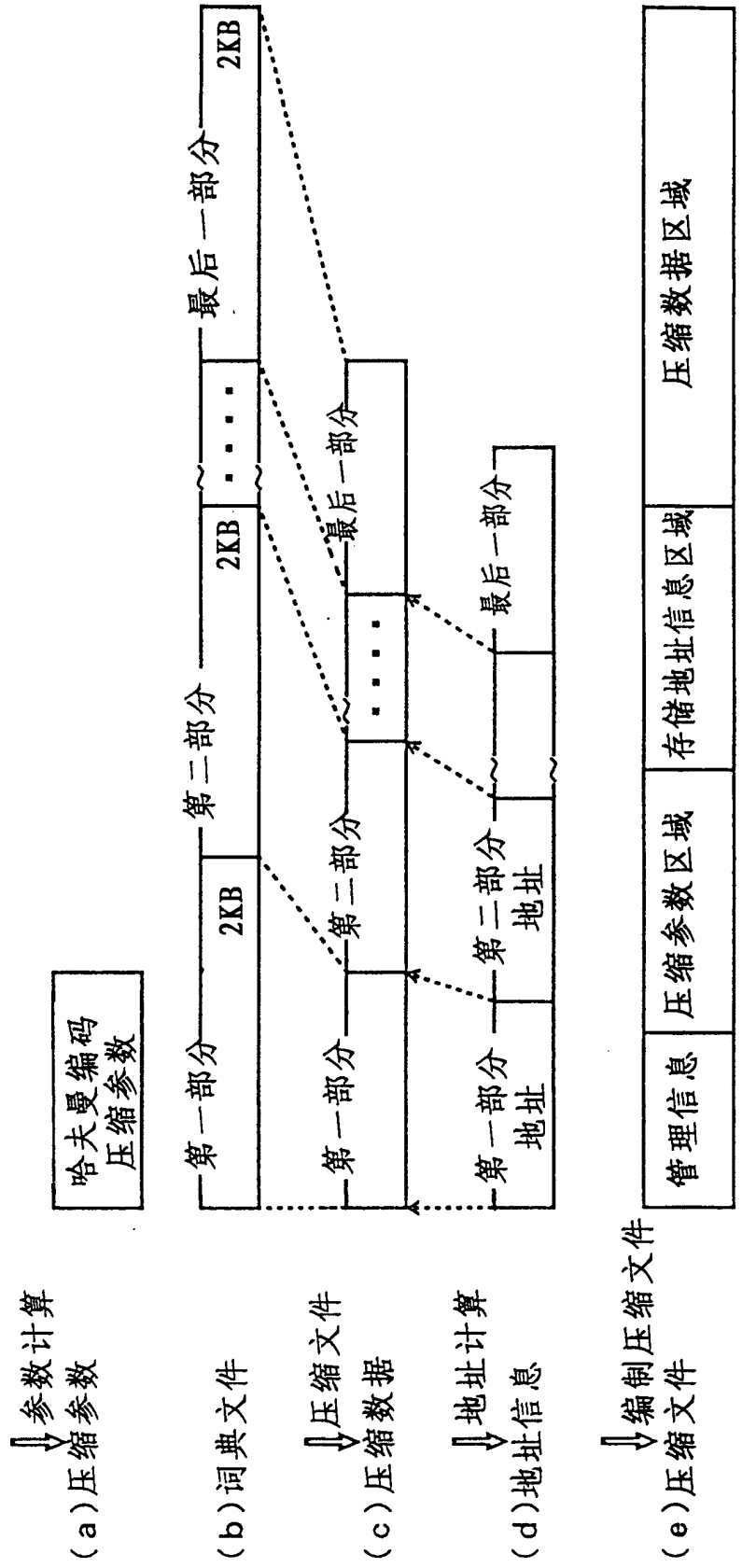

哈夫曼编码
压缩参数
参数计算
(a)压缩参数
第一部分
2KB
第二部分
2KB
最后一部分
2KB
(b)词典文件
第一部分
第二部分
最后一部分
压缩文件
(c)压缩数据
第一部分
地址
第二部分
地址
最后一部分
地址计算
(d)地址信息
管理信息
压缩参数区域
存储地址信息区域
压缩数据区域
编制压缩文件
(e)压缩文件

图 7

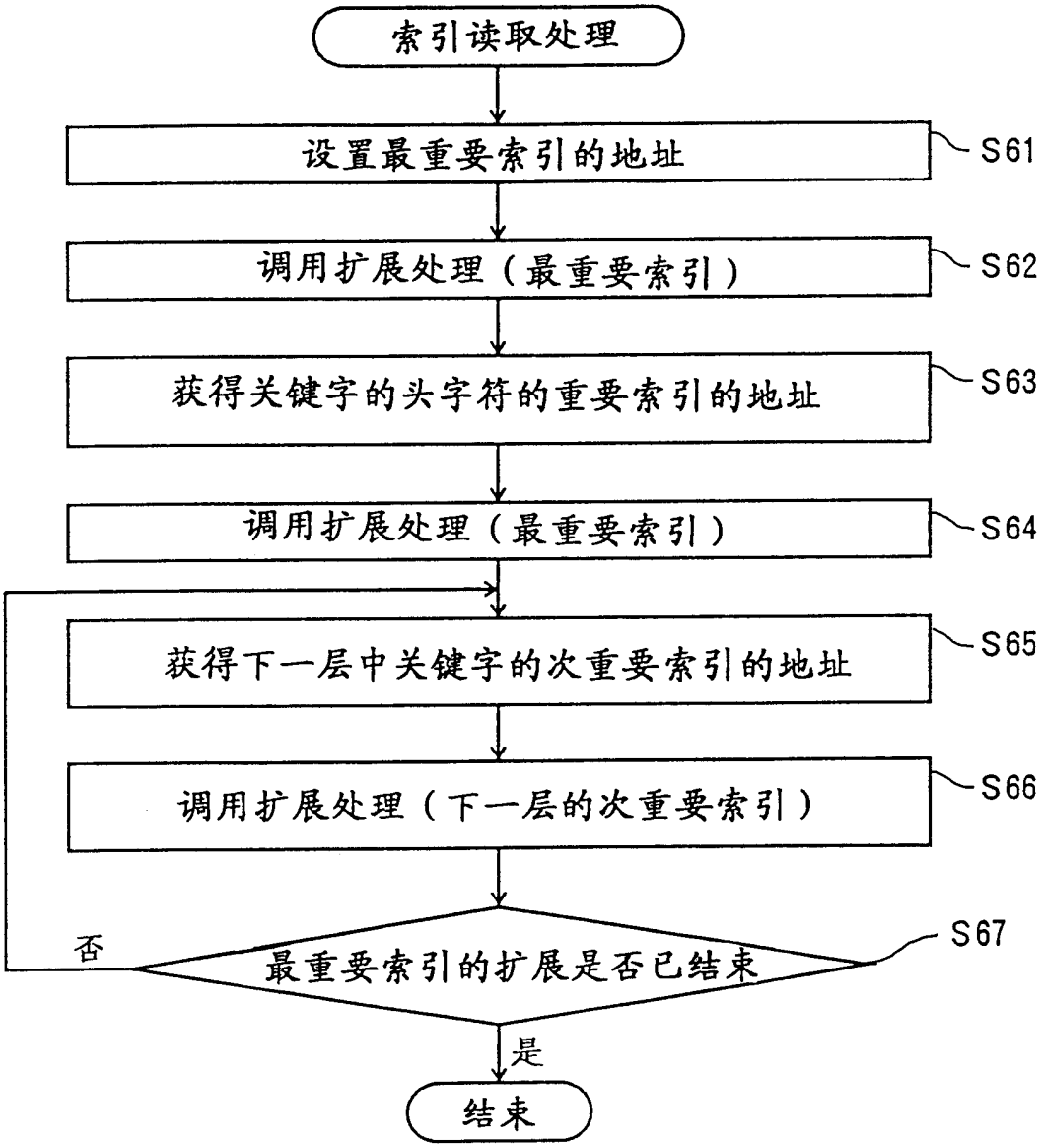

索引读取处理
设置最重要索引的地址
S61
调用扩展处理（最重要索引）
S62
获得关键字的头字符的重要索引的地址
S63
调用扩展处理（最重要索引）
S64
获得下一层中关键字的次重要索引的地址
S65
调用扩展处理（下一层的次重要索引）
S66
最重要索引的扩展是否已结束
S67
否
是
结束

图 8

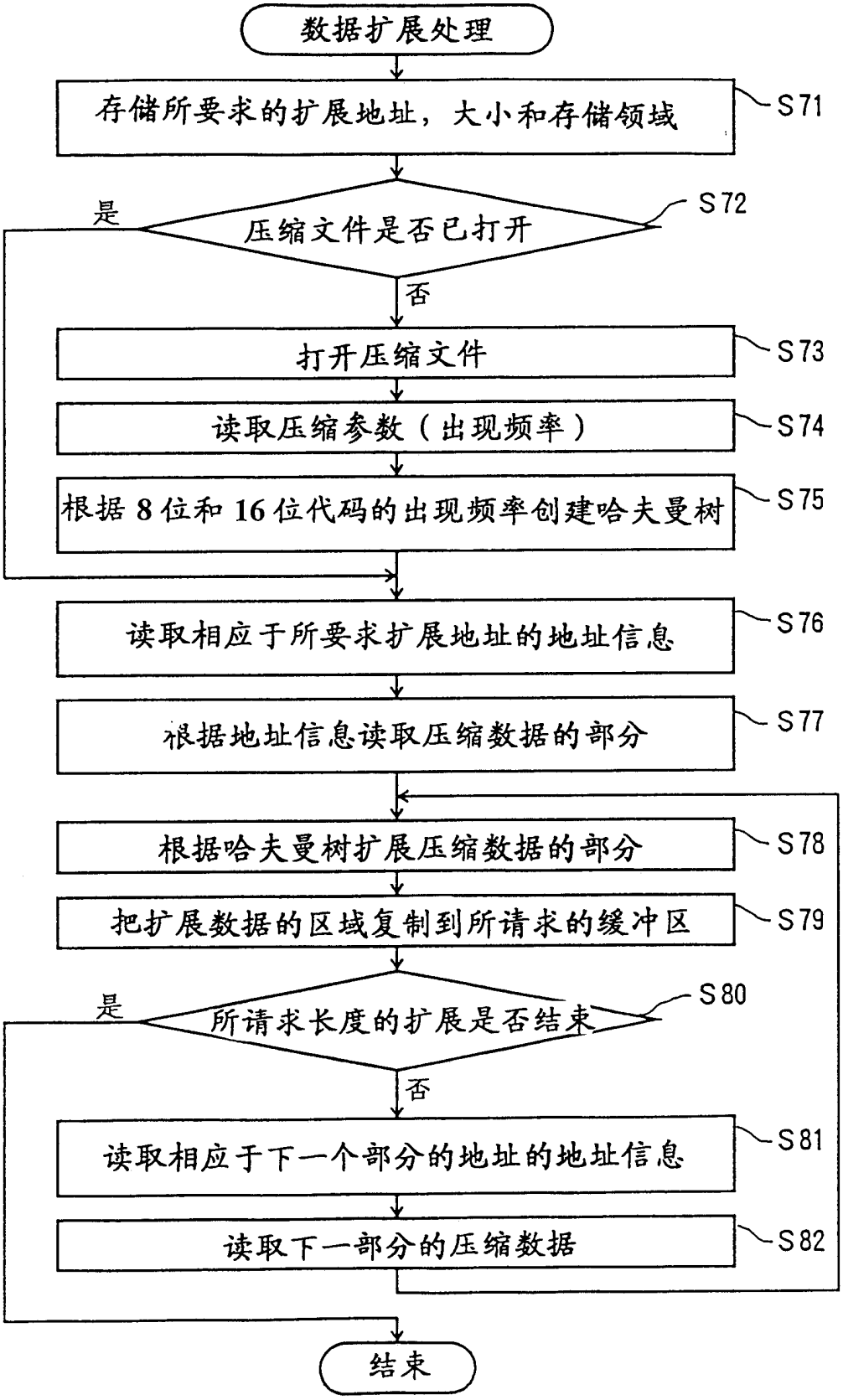
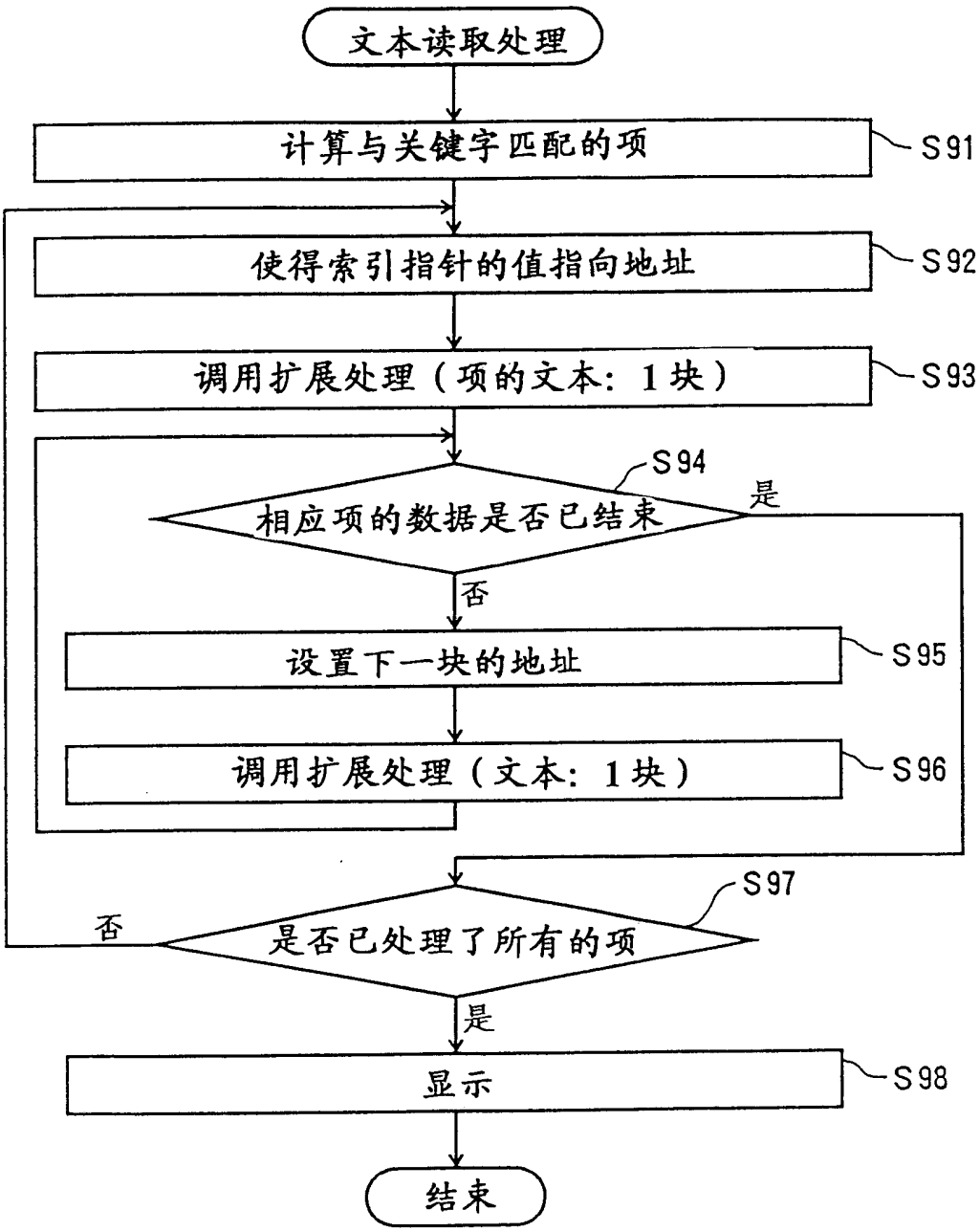

数据扩展处理
存储所要求的扩展地址，大小和存储领域
S71
是
压缩文件是否已打开
S72
否
打开压缩文件
S73
读取压缩参数（出现频率）
S74
根据 8 位和 16 位代码的出现频率创建哈夫曼树
S75
读取相应于所要求扩展地址的地址信息
S76
根据地址信息读取压缩数据的部分
S77
根据哈夫曼树扩展压缩数据的部分
S78
把扩展数据的区域复制到所请求的缓冲区
S79
是
所请求长度的扩展是否结束
S80
否
读取相应于下一个部分的地址的地址信息
S81
读取下一部分的压缩数据
S82
结束

图9


文本读取处理
计算与关键字匹配的项
S91
使得索引指针的值指向地址
S92
调用扩展处理（项的文本：1块）
S93
S94
相应项的数据是否已结束
是
否
设置下一块的地址
S95
调用扩展处理（文本：1块）
S96
S97
否
是否已处理了所有的项
是
显示
S98
结束