



[12] 发 明 专 利 说 明 书

[21] ZL 专利号 97101867.7

[45] 授权公告日 2005 年 6 月 8 日

[11] 授权公告号 CN 1205574C

[22] 申请日 1997.1.31 [21] 申请号 97101867.7

[30] 优先权

[32] 1996. 1. 31 [33] JP [31] 15012/1996

[71] 专利权人 株式会社日立制作所

地址 日本东京都

[72] 发明人 道明诚一 吉浦裕 服部芳明

天津丰 村上弘真

审查员 李 琼

[74] 专利代理机构 中国专利代理(香港)有限公司

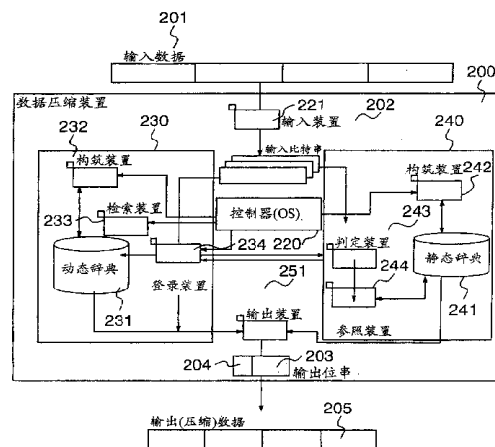
代理人 叶恺东 王忠忠

权利要求书 4 页 说明书 18 页 附图 16 页

[54] 发明名称 数据压缩、扩展方法和装置

[57] 摘要

一种使用词典的数据压缩方法，由于动态词典和静态词典并用，即使在用动态词典难以压缩的输入数据的开头部分也不降低压缩率，并且可以防止由于出现频度低的文字列占据词典而造成的压缩率的降低。由输入装置(221)输入应压缩的数据(201)，并把它分解为输入位串(202)；判定装置(243)和参照装置244把输入位串(202)作为检索关键词参照静态词典(241)；从把输入位串(202)与静态词典(241)的位串比较的结果和用检索装置(233)检索动态词典(231)的结果判定是否把输入位串(202)登录在动态词典(231)中，在登录的情况下，用登录装置(234)附加并登录索引码；输出装置(251)把输入位串(202)或与之一致的索引码作为压缩数据(205)输出。



1. 一种数据压缩方法，用第1辞典检索分解输入数据所构成的输入位串，所述第1辞典是对应位串和作为比该位串短的代码的索引码的码表数据；根据检索结果把所述输入位串登录在所述第1辞典上；把所述输入位串变换为该输入位串的代码或对应于所述第1辞典中的该输入位串的索引码之某一方，并作为压缩数据输出；其特征在于包含如下步骤：

对所述第1辞典，使用输入位串作检索，检索第1索引码；

对使位串和索引码固定地对应而构筑的第2辞典，使用输入位串，检索第2索引码；

与所述第2索引码的长度相比，所述第1索引码的长度较短时，使输入位串和第1索引码对应并登录到第1辞典中；

将指示第1辞典或第2辞典的代码，和第1索引码或第2索引码的任何一方，输出作为压缩数据。

2. 根据权利要求1的数据压缩方法，其特征在于用所述第2辞典检出与所述输入位串一致的位串的情况下，不把输入位串登录到所述第1辞典中。

3. 根据权利要求1的数据压缩方法，其特征在于比较所述输入位串和所述第2辞典的位串的步骤用所述第2辞典存储一致的第1索引码的长度；在把所述输入位串登录到所述第1辞典内的步骤中，计算用所述第1辞典登录的情况下的第2索引码的长度，与所述第1索引码相比，在所述第2索引码短的情况下进行登录。

4. 根据权利要求1的数据压缩方法，其特征在于在数据输入前定义构成输入候补的位串和索引码；在建立第2辞典的步骤中，把属性信息、出现概率、位串间的关联之中的至少一种附加信息加在所述位串和索引码上来进行定义；在所述输入位串和所述第2辞典的位串进行比较的步骤中，存储所述位串的附加信息；把所述输入位串登录在所述第1辞典上的步骤中，使用所述第2辞典的附加信息，而不登录输入位串。

5. 根据权利要求1的数据压缩方法，其特征在于在数据输入前定义构成输入候补的位串和索引码；在建立第2辞典的步骤中，把属性信息、出现概率、位串间的关联之中的至少一种附加信息加在所述位串和索引码上来进

行定义；在所述输入位串和所述第2辞典的位串进行比较的步骤中，存储所述位串的附加信息；在把所述输入位串登录在所述第1辞典上的步骤中，使用所述第2辞典的附加信息从辞典中删除至少一个位串。

5 6. 一种数据压缩装置，包括第1辞典，通过检索以分解输入数据所构成的输入位串，所述第1辞典具有对应位串和作为比该位串短的代码的索引码的码表数据；在所述第1辞典上登录有根据检索结果的所述输入位串；该数据压缩装置把所述输入位串变换为该输入位串的代码或对应于所述第1辞典中的该输入位串的索引码之某一方，并作为压缩数据输出；其特征在于，该装置具有：

10 通过预先固定定义的由候补输入构成的位串和索引码而建立的第2辞典；

 建立所述第2辞典的构筑装置；参照所述第2辞典内容的装置；和判定要不要对所述第1辞典进行位串登录的判定装置；并根据所述判定装置的判定结果把所述输入位串登录在所述第1辞典上。

15 7. 根据权利要求6的数据压缩装置，其特征在于所述第2辞典由英语或日语等的自然语言的一种以上的单词或其一部分、程序语言的一种以上的约定语言或其一部分、机械语言或一种以上的命令语言或其一部分之中的至少一种以上的文字列构成，并且结合了上述文字列的属性信息、上述文字列的出现概率、上述文字列动词的关联习惯之中的至少一种附加信息而构成。

20 8. 根据权利要求7的数据压缩装置，其特征在于所述第2辞典由自然语言的文字列或作为自然语言的正规显示的文字列构成；用所述属性信息定义词类、句型变化、词义，用所述文字列动词的关联习惯来定义语法。

25 9. 根据权利要求6的数据压缩装置，其特征在于所述第2辞典设置有加上属性信息、出现概率、位串间的关联之中的至少一种附加信息而构成并用所述第2辞典的附加信息从所述第1辞典中删除登录的位串的至少一个以上的删除装置；存储所述第1辞典的使用大小、登录位串数的上限值的至少一方，并且用所述第1辞典检索所述输入位串的检索装置包含检出输入位串的累计超过所述上限值的检出装置，和参照所述第2辞典的附加信息来删除第1辞典内的至少一个以上位串的所述删除装置。

10. 根据权利要求6的数据压缩装置，其特征在于：具有可以选择数据压缩的装置的设定画面，具有

多个登录装置，将输入到所述第1辞典的位串登录到辞典上；和

多个删除装置，将登录在所述第1辞典的位串删除；

5 特别是，包括有

选择多个登录装置之中的任意的登录装置的设定画面；和

选择多个删除装置之中的任意的删除装置的设定画面。

11. 根据权利要求6的数据压缩装置，其特征在于：具有可以选择数据压缩的装置的设定画面，具有

10 多个构筑装置，构筑所述第2辞典；

特别是，包括有选择多个构筑装置之中的任意的构筑装置的设定画面。

12. 一种数据扩展方法，将被压缩的数据扩展，所述被压缩数据由使用将成为候补的输入数据的位串和比该位串短的位串对应的多个码表，将压缩前的数据压缩变换了的、比该位串短的第1位串，和

15 将指示使用了所述多个码表中的哪一个的代码表示的第2位串的结合构成；

该数据扩展方法的特征在于：

具有从所述经压缩变换的数据中，抽出所述第2位串，并从所述多个码表，选择在扩展变换中使用的码表的规则。

20 13. 根据权利要求12的数据扩展方法，其特征在于压缩变换用的码表和扩展变换用的码表是同样形式的码表，压缩变换时利用该码表从构成为输入候补的输入数据的位串中得到短的位串，扩展变换时利用该码表从被压缩了的位串中得到作为候补的原数据的位串。

25 14. 根据权利要求12的数据扩展方法，其特征在于压缩变换时用的码表和扩展变换时用的码表是复制同一个代码的码表。

15. 一种数据扩展装置，用于将数据压缩装置所压缩变换了的位串扩展为原输入数据的位串，该数据压缩装置具有使用按照预定规则作成的第1码表，将输入数据变换为比该输入数据的位串还短的位串的第1数据压缩装置；

使用将成为输入候补的输入数据的位串，与比该位串短的位串预先固定对应的第2码表，将所述输入数据的位串变换为比该位串短的位串的第2数据压缩装置；和

5 使用所述第1，第2的任何一个数据压缩装置将输入数据分解了的输入位串压缩变换，并将压缩变换后的位串和表示使用了所述的哪一个数据压缩装置的代码一并输出的装置，所述数据扩展装置的特征在于，具有：

把所述压缩变换了的位串按照预定的规则扩展变换为原输入数据的位串的第1数据扩展装置；

10 用预先固定对应构成为输入候补的压缩变换了的位串和原输入数据的位串的码表，把所述压缩变换了的位串扩展为原输入数据的位串的第2数据扩展装置；以及

从所述压缩变换了的位串，将指示所述第1，第2码表的哪一个的码表的所述代码抽出，并选择在扩展变换中使用的码表的扩展控制装置。

15 16. 根据权利要求15的数据扩展装置，其特征在于第2数据压缩装置和第2数据扩展装置的码表是同样形式的码表，第2数据扩展装置利用码表从构成成为输入候补的输入数据的位串中得到短位串，第2数据扩展装置利用码表从被压缩了的短位串中得到作为输入候补的原输入数据的位串。

17. 根据权利要求15的数据扩展装置，其特征在于第2数据压缩装置和第2数据扩展装置是同一个码表。

数据压缩、扩展方法和装置

5

技术领域

本发明涉及一种把文字列等的数据变换为比该数据的比特数更短的位串的数据压缩、扩展方法和装置以及使用该方法和装置的数据处理设备及网络系统。

10

背景技术

数据压缩技术是削减硬盘等存储装置中的数据占有量和通信中的数据传输量，从而提高存储装置和通信线路的效率的技术。以往，提出了各种各样的数据压缩方法，作为有代表性的方法有数据压缩手册（トッパン、1994年）的第221～247页所述的LZ78及其改进型方法。

15

LZ78及其改进型方法是由下述的基本步骤构成的方法，这是一种按照预定的规则压缩数据的方法：

（1）存储代表输入数据的文字列。把这些存储的文字列的集合称之为动态辞典。

20

（2）当把与动态辞典中的文字列相同的文字列再现为输入数据的情况下，把代替该文字列的动态辞典中的该文字列的索引码（一般为正整数）作为输出数据输出。

（3）当所存储的文字列使动态辞典满额时，就不进行其超过部分的文字列的登录或删除登录完了的文字列。在删除登录完了的文字列的情况下，或是删除全部文字列，或是按顺序从更旧的文字列开始删除。

25

按照上述原来的方法，因为把由多个文字构成的文字列替换为一个索引码，所以能够压缩数据。而且，按上述的规则反向进行，可以容易地对用上述方法压缩的数据进行扩展变换。

在使用上述原来的动态辞典的数据压缩方法中，存在以下的问题：

(1) 在输入数据中最初出现的文字列还未登录到动态词典内，因此，并不替换为索引码，而是原样输出所输入的文字列。所以，输入数据的最初部分的压缩率特别低。

5 (2) 因为把输入数据中的文字列按顺序逐次登录在动态词典中，所以存在词典溢出的情况，作为这时的解决办法，例如可以采取以下的删除步骤：

(a) 在词典中不再新增登录文字列。

(b) 全部删除已登录的文字列，并把词典初始化。

(c) 保存更新的文字列，并删除旧的文字列。

10 不管在执行哪种删除步骤的情况下，存在于以后的动态词典内的文字列再次出现在输入数据中的可能性都未必很高。一般，由于输入数据中的文字列与动态词典中的文字列一致的频度小，所以压缩率低。

发明内容

本发明的目的在于提供能够实现更高压缩率的数据压缩方法和装置。

15 本发明的另一个目的在于提供即使在输入数据的最初部分也不降低压缩率的数据压缩方法和装置。

本发明的另一个目的在于提供防止由于出现频度低的文字列占据词典较多的部分而引起的压缩率降低的数据压缩方法和装置。

20 本发明的另外的目的在于提供把压缩了的数据扩展为原数据的数据扩展方法和装置。

本发明的另外的目的在于提供具备数据压缩功能和扩展功能的数据处理方法和装置。

本发明的另外的目的在于提供具备数据压缩功能和扩展功能的网络系统。

25 为实现上述的目的，本发明采取以下的构成。

第一，动态词典和静态词典并用，静态词典不变更登录内容。静态词典中登录着出现频度高而且比较长的文字列及其索引码。把静态词典的索引码加在动态词典的索引码上，作为压缩数据输出。

30 具体地讲，就是输出输入文字列（以下称“输入位串”）、动态词典的索引码、静态词典的索引码的某一个代码以及作为译码用的词典以用来选择

动态辞典、静态辞典的某一方的代码。这时，先存储输入位串代码的长度，在检索静态辞典时存储至少一个与输入位串一致的索引码的代码长度；在登录动态辞典时，同样存储一致的代码的长度，再把上述输入位串和索引码中更短的代码作为压缩数据输出。

- 5 在用静态辞典检出成为对象的位串阶段，也可以输出静态辞典的索引码，并采用静态辞典优先方式。有关使用动态辞典的索引码的不希望压缩的部分，即有关输入数据的最初的部分，例如开头的500字节部分，也可以采用所谓使用输入位串的代码或静态辞典的索引码进行编码的辞典切换方式。

无论以上的哪种构成，都能够解决以上的问题（1）。

- 10 第二，为解决上述的问题（2）可以采取以下①②的任一种构成：①不把利用率低的文字列登录在动态辞典中，②预见词典的溢出，并从动态词典中删除利用率低的文字列。

具体地对于第①点来说，在输入位串与静态辞典的内容一致的情况下，不把其输入的位串登录在动态辞典中。

- 15 而且，在事先调查把输入位串登录到动态辞典内时的索引码并作为输出数据编码的情况下，当动态辞典的索引码与静态辞典的索引码相比被判断为更短时，也可以把动态辞典的索引码登录下来。

另外，也可以使用静态辞典的属性信息、出现概率或位串之间的关联关系中的至少一种信息作为附加信息，而根据该附加信息不登录输入位串。

- 20 具体地对于第②点来说，把静态词典的各位串与属性信息、出现概率或位串之间的关联关系中的至少一种附加信息合起来定义，并利用附加信息把利用率低的至少一个位串从动态辞典中删除掉。

- 25 在存储动态辞典的使用大小的上限值或登录位串数的上限值的至少一方的同时，存储动态辞典的使用大小的当前值或登录位串数的当前值的至少一方，在检出当前值超过上限值的“溢出”的情况下，也可以参照静态辞典的位串的附加信息，由此来把动态辞典的位串的附加信息删除掉。

以上的任一种构成都能解决上述的问题（2）。

- 30 在实施按照本发明的数据压缩方法和装置的情况下，必须构成包含出现频度高的位串的静态辞典。在本发明优选装置的实施例中，静态辞典包含以下之中的至少一种，即：英语和日语等自然语言的一个以上的单元语或其一

部、程序语言的一个以上的预约语言或其一部、或机械语言的一个以上的命令语言或其一部。特别是,静态词典的内容是自然语言的文字列、或作为自然语言的正规显示的文字列,并把属性信息附加在其上,来定义词性、句形变化、语义、语法等。

5 在实施按照本发明的数据压缩方法和装置情况下,必须照顾到静态词典和动态词典的码表构成以及操作各词典的程序的選擇。特别是在本发明的优选装置的实施例中,是由具有最大 $2^{10} \sim 2^{15}$ (1K~32K)个索引码的动态词典和最大 $2^{12} \sim 2^{17}$ (4K~128K)个索引码的静态词典构成。而且选择构成静态词典的程序、把输入位串登录到动态词典的程序和删除动态词典的位串
10 的程序,对于各个程序设置设定数据压缩条件的初始设定画面。

 在实施按照本发明的数据压缩方法和装置情况下,必须共有压缩元和复原地址的静态词典。在本发明的一个适用对象中,有通过网络的文件传送时的数据压缩,对于此,可以采取以下的构成,在复原地址可以特定的情况下,预先在确认可以共有静态词典之后利用静态词典。或者也可以设置静态
15 词典的通用索引码,这种情况下,用压缩元把编码过的数据压缩为压缩装置固有的索引码中的通用索引码,并传送到复原地址。用复原地址把数据从通用索引码变换为译码装置固有的索引码后,把数据复原。

 如上所述,按照本发明的数据压缩方法与单独用动态词典的方法相比,能够使压缩率提高。

20 附图说明

 图1是表示在动态词典中并用静态词典的数据压缩步骤的流程图。

 图2是动态词典中并用静态词典的数据压缩装置的构成图。

 图3并用动态词典和静态词典,并可以选择位串长度更短的索引码的方框图。

25 图4是说明对输入位串、动态词典和静态词典的索引码之中的最短的数据进行编码,并作为压缩数据输出的方框图。

 图5是表示按照静态词典的优先方式的数据压缩方式的流程图。

 图6是表示按照词典切换方式的数据压缩方式的流程图。

 图7是说明在与静态词典一致的情况下不把位串登录在动态词典内的步骤
30 骤的流程图。

图8是表示计算静态辞典和动态辞典的索引码的长度，再在动态辞典的索引码短的情况下把位串登录到动态辞典内的步骤的流程图。

图9是由自然语言、程序语言、机械语言的至少一种以上的文字列及其附加信息构成的静态辞典的结构示例图。

5 图10是作为图9的上述附加信息，包含语言规则的静态辞典的结构示例图。

图11是附加信息的其它结构示例图。

图12是表示删除利用率低的位串并确保能够登录的区域的步骤的流程图。

10 图13是用静态辞典241的附加信息进行删除的构成图。

图14是表示初始设定程序的种类和压缩条件的构成图。

图15是表示静态辞典的结构的一个实施例的构成图。

图16是在分布系统上实施可以共有静态辞典的的情况的本发明的流程图。

15 图17是在分布系统上不能共有静态辞典的的情况的本发明的装置的构成图。

图18是本发明中的压缩数据结构和各索引码的关系图。

图19是在动态辞典中并用静态辞典的数据扩展装置的构成图。

图20是表示参照静态辞典和动态辞典的索引码进行扩展的步骤的流程图。

具体实施方式

20 以下来说明根据输入位串与静态辞典的位串进行比较的结果来把位串登录到动态辞典内的实施例，而本发明并不仅仅限定于该实施例。

首先用图1、2、7和8来说明本实施例的概况。

25 图2是把静态辞典并用于动态辞典的数据压缩装置的构成图。该数据压缩装置200是把输入数据201作为数据量小的压缩数据205输出的装置，是用作数据处理装置的计算机装置如个人计算机来实现的。或者也可以用连接在计算机装置和网络之间的专用装置来实现。数据压缩装置200具有动态辞典控制器230和静态辞典控制器240，还有控制整体处理的控制器220。控制器220例如由承载OS的计算机装置的CPU构成，控制器220还控制压缩数据的输入装置221和压缩数据的输出装置251。

动态辞典控制器230具有动态辞典231、构筑装置232、检索装置233和登录装置234。动态辞典231是存储位串和索引码的数据库，并且被构成在计算机装置的存储装置上，例如构成在装在计算机装置内的硬盘装置上。构筑装置232、检索装置233和登录装置234由操作动态辞典231的程序构成。

5 这些程序都存储在内装于计算机装置的硬盘装置上，并通过计算机装置的启动把它们转移到主存储装置上，再由CPU来执行，从而实现规定的功能。

静态辞典控制器240具有静态辞典241、构筑装置242、判定装置243和参照装置244。静态辞典241是存储位串和索引码的数据库，并且被构成在计算机装置的存储装置上，例如构成在装在计算机装置内的硬盘装置上。构筑装置242、判定装置243和参照装置244由操作静态辞典241的程序构成。

这些程序都存储在内装于计算机装置的硬盘装置上，并通过计算机装置的启动把它们转移到主存储装置上，再由CPU来执行，从而实现规定的功能。

15 输入装置221和输出装置251是分别进行计算机装置的硬盘装置那样的内装存储装置、经由网络的远距离文件系统或软盘和磁带那样的外部存储装置之间的数据输入和输出的程序（装置驱动器）。输入装置221输入输入数据201，并把它变换为输入位串202之后输出去。输出装置251把在作为来自动态辞典231或静态辞典241的索引码的输出位串204上附加了控制代码203的数据作为压缩数据205输出。

图1是表示把静态辞典并用于动态辞典的图2装置上的数据压缩步骤的流程图。在本实施例中的数据压缩步骤由构筑静态辞典241和动态辞典231的初始化处理（步骤100）和数据压缩的基本处理（步骤120）构成。

首先，来说明步骤100。当用户输入指定成为压缩对象的文件或数据的指令时，就开始进行处理（步骤101）。首先用构筑装置242来构筑静态辞典241（步骤102），例如在定义文字处理器和电子辞典的文字列的同时，定义位串和索引码。然后，用构筑装置232把动态辞典初始化，即：登录空的或既定的位串来构筑动态辞典（步骤103）。在动态辞典的初始化时，有两种情况，即：登录空的位串的情况和把既定的位串即ASCII码及JIS码内所包含的字母及片假名和平假名等代码预先作为基本位串进行登录的情况。根据检

索利用形态、计算机的存储容量或索引码时所使用的散列存储表等结构，可以采用两者的任意一种，进行过以上的步骤之后，结束初始化处理（步骤104）。

在本实施例中，在建立动态辞典231之前先建立静态辞典241，但是本发明并不被限定于此，也可以在静态辞典241的建立之前，先建立动态辞典231。另外，采用了ASCII码及JIS码内所包含的字母及片假名和平假名等代码作为在动态辞典的初始化时既定的位串，但是本发明并不被限定于此，也可以采用任何一种能够按照所规定的规则单一含义地指示文字的代码体系。

接下来说明步骤120。确认静态辞典241的建立后，开始数据压缩处理（步骤121）；首先输入应该压缩的数据（步骤122），例如：在用UNIX实施的情况下，输入指定了名称的文件或来自标准输入的数据；用输入装置221按数据块单位（例如一个数据块64比特）把输入数据201分解为输入位串202（步骤123）；然后，用判定装置243及其延长所启动的参照装置244把输入位串作为检索关键词，以参照静态辞典241（步骤124）；并且把输入位串202与静态辞典241的位串相比较（步骤125）。

根据比较的结果和由检索装置233得到的动态辞典231的位串检索的结果来判定是否把输入位串202登录在动态辞典231上（步骤126）；在登录的情况下，用登录装置234登录识别输入位串的索引码（步骤127）；在不登录的情况下，什么也不做。最后，用输出装置把输入位串或与输入位串一致的索引码作为压缩数据205输出（步骤128）；另外，在有输入数据的情况下，返回到步骤122，没有输入数据的情况下，结束数据压缩处理（步骤129、130）。

在本实施例中，一旦在步骤130结束了数据压缩处理步骤120，接着就在初始化步骤100对所输入的数据把静态辞典241初始化，但是本发明并不局限与此，可以半固定地建立静态辞典241，即使结束了数据压缩处理步骤120，也可以不把静态辞典241初始化，这样做的效果是可以缩短处理时间。

用图7和图8来详细地说明把输入位串和静态辞典的位串进行比较处理125的两个实施例。

图7是一个流程图，所表示的是在输入位串与静态辞典241中的位串一致的情况下不把该输入位串登录在动态辞典231中的实施例。先把输入位串作为关键词来检索静态辞典241（步骤701）；再判断输入位串是否与静态辞典

241中的位串一致（步骤702）；在一致的情况下，不把该输入位串登录在动态辞典231中（步骤703），而在不一致的情况下，就进行登录（步骤704）。按照本实施例，可以减少登录在动态辞典内的位串。

图8也是一个流程图，表示这样一个实施例，即：在输入位串与静态辞典241中的位串一致的情况下，计算静态辞典241和动态辞典231的索引码的长度，如果动态辞典231的索引码短，就把该输入位串登录在动态辞典231中。在图8中，步骤701~704与图7中的步骤是共同的。先把输入位串作为关键词来检索静态辞典241（步骤701）；再判断输入位串是否与静态辞典241中的位串一致（步骤702）；一致的情况下，提取出与该输入位串一致的静态辞典241中的位串的索引码，并存储其长度（步骤801），同时，把该输入位串作为关键词检索动态辞典231，并判断有没有与该输入位串一致的位串（步骤811）；如果没有一致的位串，在作为动态辞典231的位串进行登录时预测应该分配的索引码，并计算存储其长度（步骤812）；如果有一致的位串，就不进行登陆（步骤703）。当在步骤702判断为「一致」，而且在步骤811判断为「不」的情况下，比较在步骤801、812所存储的各个索引码的长度（步骤802）；如果静态辞典241的索引码短（步骤803），就不把该输入位串登录在动态辞典231中（步骤703），如果不是这样（步骤803），就进行登录（步骤704）。在步骤702判断为「不一致」的情况下，把该输入位串登录在动态辞典231中（步骤704）。

在本实施例中，比较动态辞典和静态辞典的索引码的长度之后，在下次出现相同位串时利用动态辞典可以期待高压缩率的情况下，进行向动态辞典的登录，因此，按照本实施例，就进行动态辞典与静态辞典并用的压缩率有限的处理，所以能够实现更高压缩率的数据压缩处理。

接下来用图3、4、5、6和18来详细说明动态辞典与静态辞典并用而输出某一方辞典的索引码作为压缩数据的处理128的一个实施例。

图3是表示由于动态辞典与静态辞典并用而使更短的比特长度的索引码可以选择的情况的方框图，与原来的动态辞典单独使用的方法相比较，压缩数据的结构更加复杂，所以提高了压缩率。

图3表示把普通文字301分解后的输入数据和在静态辞典中定义的文字列一致的情况。一般，把比特长度较长的文字列（例如由10个以上字母构成的

英语单词) 登录在静态辞典中, 并且在输出其索引码的长度短于文字列的长度的静态辞典的索引码时, 可以期待压缩率的提高。

作为本实施例的输出数据的一例, 有压缩数据304。压缩数据304成为在辞典的索引码或位串直接代码上附加了控制代码的数据列, 该控制代码204
5 由区别辞典索引码和位串直接代码的代码以及区别动态辞典231的代码和静态辞典241的代码的代码构成。虽然与原来的单独用动态辞典的压缩方法所得到的压缩数据305相比成为复杂的大的代码系, 但是由于静态辞典的索引码短, 所以作为一个整体来看, 压缩数据304的长度就比压缩数据305的长度短。

10 用图4和图18来说明实现输出最短索引码的步骤的方法。

图18所表示的是本实施例中的压缩数据的构造和各索引码的关系, 压缩数据1801、1811和1831分别代表把输入位串原样编码的情况、把动态辞典的索引码编码的情况以及把静态辞典的索引码编码的情况的压缩数据的构造。在本实施例中, 为简便起见, 以下的输入数据是英语的文章, 并且用8比特
15 码系(ASCII)进行输入。控制代码1803、1804、1813、1814、1823和1824是压缩数据代表输入文字列代码、动态辞典的索引码、静态辞典的索引码之哪一种的代码, 并且是数据本身所不使用的位串, 例如换码系列等。译码时, 先把所输入的压缩数据分解为规定的单位, 如每8比特为一个单位; 再识别这些控制代码; 然后把下一个控制代码到来之前的位串作为压缩数据的本体
20 进行处理。压缩数据1801采取把数据本体1802夹在控制代码1803和控制代码1804之间的构造, 数据本体1802由8比特固定长度的代码411构成, 压缩数据1811采取把附加了控制代码1815的数据本体1812夹在控制代码1813和控制代码1814之间的构造, 数据本体1812由包含在动态辞典231中的可变长索引码构成, 控制代码1815是代表构成数据本体1812的索引码长度的8比特代码, 例如
25 在图4的示例中, 由于用6比特代码来代表整个动态辞典231的索引码, 所以把6比特代码系代码设定在控制代码1815中。压缩数据1821采取把数据本体1822夹在控制代码1823和控制代码1824之间的构造, 数据本体1822由包含在静态辞典241中的可变长索引码构成。

图4是说明计算输入位串以及动态辞典和静态辞典的索引码的长度并采用数据长度最短的索引码的方框图。在本实施例中, 特别是把字数为4字以
30

上的英语单词登录在静态辞典上，关于这种英语单词的压缩，是按每一个字来对英文字进行编码，与变换为动态辞典的索引码相比，利用静态辞典的方法就非常有利。判定装置243从输入位串的各代码和辞典的各索引码中计算各自的比特长度，并进行比较判定。输入位串202由8比特代码系构成，即由8比特代码411和位串412构成。静态辞典241由最大长度为17比特的索引码421和位串422构成；动态辞典231由最大长度6比特的索引码431和位串432构成。控制代码取为8比特。例如：对普通文字401中的一部分单词“This”进行压缩，计算用输入位串202的代码411、动态辞典241的索引码431和静态辞典241的索引码421的表现方法来代表该单词的情况下所要求的比特数。在使用输入位串202的代码411的情况下，因为各代码是固定长8比特，在其上加上构成位串的字数4，再在前后加上控制代码8比特，这就成了48比特。在使用动态辞典231的索引码431的情况下，因为索引码长是最大6比特，在6比特上加上字数4，再加上代表6的8比特和前后控制代码8比特，这就成了48比特（最大值）。在使用静态辞典241的情况下，在该例中，因为对应于“This”的索引码是8比特代码“11111011”，加上前后8比特控制代码，就成了24比特。因此，在这种情况下，在判定装置243中，选择使用静态辞典的索引码421的编码。在该例中，与选择动态辞典231的索引码431和输入位串202的代码411相比，可以使压缩数据的长度做成为大约一半。

在以上的实施例中，说明了先计算辞典的索引码的长度，再选择静态辞典、动态辞典之某一种的方法。下面，作为其它实施例，首先来说明静态辞典优先方式，该方式是在输入位串与静态辞典中的位串一致的情况下，无条件地输出静态辞典的索引码。另外，作为其它的实施例，来说明辞典切换方式，该方式是设置门限值为从数据处理位置的数据开头算起的距离，然后切换在门限值前后使用的辞典。

图5是把与输入位串一致的静态辞典241的索引码作为压缩数据输出的步骤的流程图。本实施例是静态辞典优先方式的例子，是对图1的输入位串与静态辞典的位串进行比较的比较处理125的一个实施例。首先，全面检索静态辞典241（步骤501）；在位串一致的情况下（步骤502），判断为不要登录在动态辞典231上（步骤503），判定应该输出静态辞典241的索引码（步骤

504)；在位串不一致的情况下（步骤502），判断为必须登录在动态辞典231上（步骤505），并判定应该输出输入位串的代码（步骤506）。

图6是表示输出步骤的流程图，首先把门限值设置在从数据处理位置的数据开头开始的位置，并从数据开头到门限值的位置使用输入位串和静态辞典，如果超过门限值位置，就加上输入位串和静态辞典进行还使用动态辞典的编码，然后输出压缩数据。本实施例是辞典切换方式的例子，是图1和图4所示的数据压缩处理的其它实施例。这种方式是根据在文件的开头部分有效利用静态辞典的经验把文件的开头部分即从文件的开头到门限值位置的部分用静态辞典的索引码进行处理，其余部分也就是从门限值位置到文件末端的部分用动态辞典的索引码来处理。与图1的步骤125和126所示的实施例的不同点是不计算索引码的长度，而在规定的门限值之前一律用静态辞典。

本实施例是在辞典构筑的初始化处理步骤100之前把步骤分为设定门限值的初始步骤601~604和相当于作为数据压缩处理步骤120的内部处理的步骤123~128的处理的压缩主处理步骤611~617。

首先来说明初始步骤601~604，在初始步骤中，为了决定门限值的具体值，判别输入数据是输入时知道大小的文件还是像通信数据那样的终端未定的数据流数据（步骤601）；是文件的情况下，先调查文件大小，再用来自整个文件大小的比例计算门限值，或者根据文件的大小来选择不正规的门限值（步骤602）。例如：在把来自整个大小的比例为10%决定为门限值的情况下，如果整个大小是通常的大小，如10字节，那么最初的一字节就成为门限值。但是，在整个大小很大的情况下，例如为100字节的情况下，通常把第10字节取为门限值，并且不规则地把门限值设定在第2~3字节上。在数据流数据的情况下，利用固定的初始值（步骤603），如果决定了利用的初始值，就用数据压缩处理步骤把它设定到参照寄存器（步骤604）。

以下来说明压缩主处理步骤611~617，紧接数据的压缩处理步骤121~123之后，进行从输入位串的开头起始的比特数的累积计算（步骤611），并与在步骤604设定的门限值进行比较（步骤612）。在超过门限值的情况下，检索与输入位串一致的静态辞典中的位串（步骤613），对其索引码进行编码后输出（步骤615）。在步骤613没有一致的位串的情况下，对输入位串的代码进行编码后输出（步骤616），并把该位串登录在动态辞典中（步骤

127)。在步骤612, 在超过门限值的情况下, 检索与输入位串一致的动态辞典中的位串(步骤614), 并把其索引码编码输出(步骤617)。在步骤614, 不一致的情况下, 把处理步骤转移到步骤616、127。

在以上的处理结束并进一步存在下一个输入数据的情况下(步骤5 129), 返回到步骤122, 不存在下一个输入数据的情况下(步骤129), 就结束数据压缩处理(步骤130)。

按照本实施例, 切换辞典的控制代码的插入基本上为2次, 即: 完成是否超过门限值的代码的插入和在辞典中是否存在一致的位串的代码的插入。所以, 处理简单而且具有缩短所要时间的效果。与频繁地切换辞典的情况相比, 由于减少了控制代码数, 结果, 就具有提高数据压缩率的效果。10

作为最适宜的系统 and 实现的方法, 以下用图9和图10来说明上述的装置和处理步骤。具体地说, 所表示的是静态辞典的构成的实施例, 该静态辞典考虑用自然语言、程序语言或机械语言来记述输入数据201的实际情况, 把属性信息、出现频度或文字列的相互关联总合起来, 并利用语言上的特性来表示使压缩率提高的静态辞典的构成。15

图9所表示的是用程序语言上的一个以上的文字列及其附加信息构成的静态辞典的实施例, 特别是表示用C语言的控制句法中所使用的预约语及其附加信息构成的静态辞典900。本实施例的特征在于对登录在静态辞典内的文字列901, 把分别出现2次以上的概率902作为附加信息。例如: 在数据压缩的对象是C语言的源程序的情况下, 由于利用辞典900作为静态辞典, 所以可以期待提高静态辞典中的索引码的输出率, 并提高压缩率。由于事先用多个C语言的源程序进行了调查, 所以出现概率不依据输入数据的内容, 而可以作为期待值来加以利用。在登录和删除动态辞典中的位串的情况下, 通过把该出现概率的大小作为登录和删除的优先顺序, 也能使动态辞典中的位串和输入位串一致的概率提高, 当然也就可以期待压缩率的提高。另外, 在该25 实施例中, 由于把出现概率高的数据分配给比特长度更短的索引码903, 所以也可以期待检索速度的提高。

图10是作为附加信息包含了语言规则的静态辞典的构筑例。静态辞典1000扩展了静态辞典900, 并且其构成也包含语法。例如: 用C语言作为控制句法, 可以把这种句法结构作为是哪种图形来记述。使索引码1002对应于把30

多个预约语关联起来的语法1001。在该例中，由于是用3位整数来作为索引码1002，所以，由多个预约语构成的每一组语法的代码长度比使用索引码9003时的短，进一步还规定与句法结构一致的压缩数据的构造1003。

在本实施例中，虽然用程序语言的文字列及其附加信息来构成静态辞典，但是本发明并局限于此，也可以用自然语言、程序语言或机械语言之至少一种以上的文字列及其附加信息来构成。

以下用图11～图13来说明在用图9和图10所描述的静态辞典中根据输入位串和静态辞典的位串进行比较的结果不把位串登录在动态辞典中而是从其中删除的实施例。

图11是用别的附加信息的实施例，静态辞典241包含识别对应于各个代表登录位串1101的用语的领域的附加信息1102和索引码1103，在处理输入数据201之中，用判定装置243进行内容的分析，并判定是哪一个领域的文件，不登录在动态辞典231之中，或者用删除装置1301来删除领域以外的位串。

图13所示的构成例是先设定登录在动态辞典231中的上限值1302，再追加为在判定装置243中用静态辞典241的附加信息删除利用率低的位串的规则。图13和图2之不同点在于，具有在登录装置中检测到溢出时被调出的删除装置1301、设定规定动态辞典231可以登录的位串的上限值的寄存器1302以及静态辞典241的附加信息，并且还具有判定利用率是否低即判定是否是删除对象的规则1311。

图12是表示用图13的装置和图9的出现频度的附加信息的删除步骤的一个实施例的流程图。在把输入位串登录在动态辞典内时（步骤 127）检测溢出，并删除利用率低的位串，以确保可以登录的区域。首先，比较累积运算输入比特数和上限值，在检测到动态辞典溢出的情况下，启动删除装置1301，把处理步骤转移到步骤1202，否则就像原来那样进行输入位串的登录（步骤1206）。在本实施例的情况下，为说明简单起见，把具有比输入的位串的出现概率低的出现概率的位串取作删除对象。首先，参照作为输入位串的附加信息的出现概率902（步骤1202），把出现概率更低的位串提取出来，接下来检索动态辞典231的登录位串（步骤1203），存在利用率更低的位串的情况下（步骤1204），就把它从动态辞典中删除掉（步骤1205）。

图14是表示初始设定程序的种类和压缩条件的构成图。在本实施例中，因为在静态辞典中和动态辞典中分别有多个构筑装置242和多个登录装置234及多个删除装置，所以，其特征在于程序的种类和条件要由使用者来初始设定。1400是连接到压缩装置200的终端，1401是终端1400的设定画面，例如是用图表形式表示并选择条件的画面。数据压缩的使用者通过设定画面1401来选择构筑装置242、登录装置234或删除装置1301等的执行程序。与此同时，例如综合设定上述的门限值的设定以及对动态辞典的位串的登录・删除的基准等。

图15是静态辞典的构筑的一个实施例的构成图。在本实施例中，其特征是通过划分动态辞典和静态辞典的适用范围来发挥并用辞典的优点。因此，在静态辞典和动态辞典的索引码的最大值的设定中，以下所述的条件也是必要的。在静态辞典中多考虑比特长度比较长的文字列，而在动态辞典中则登录比特长度短但出现频度更高的文字列。特别是与动态辞典的情况相比，结合比特长度短的索引码进行登录，作为静态辞典构筑的一个实施例，采用约2万词的字处理器的辞典FEP1501、约10万词的英日辞典及汉日辞典1502、或每个专业领域的百科全书1503的至少一种以上的数据库，并用特定的基准抽出共同收录的词来构筑静态辞典241。在这种情况下，由具有最大 $2^{10} \sim 2^{15}$ (1K~32K)个索引码的动态辞典和最大 $2^{12} \sim 2^{17}$ (4K~128K)个索引码的静态辞典构成是最适宜的。

图16是表示在分布系统中可以共有压缩元和译码目的地的静态辞典的情况的实施例的流程图。在辞典的初始化处理步骤100之后，数据压缩步骤120之前，进行询问是否存在与译码地址相同的静态辞典的处理，采用这一点就成了图1的实施例。首先，作为进行压缩处理的前提，把连接在网络末端上的装置指定为译码目的地。在本实施例适用于文件传送（FTP）的情况下，例如把邮局名和IP地址指定为自变量。在确定了译码目的地的情况下（步骤1601），把压缩元和译码目的地之间连接起来（步骤1602），未确定的情况下（步骤1601），就结束。步骤1602的连接后，确认译码目的地的静态辞典的文件名，并调查静态辞典的版本和构成要素，根据其结果来判定译码目的地的静态辞典是否与压缩元的静态辞典相同（步骤1603），相同的情况下，采用动态辞典231和静态辞典241的并用方式（步骤1610）。不同的情况下，

进行修改并采用别的方式，如只利用动态辞典231进行压缩的原来的方式以及把压缩元的静态辞典241传送到译码目的地的方式等（步骤1604），并切断压缩元与译码目的地之间的连接（步骤1615）。在步骤1610之后，用压缩元压缩数据（步骤120），并发送数据（步骤1611）。在译码目的地在步骤
5 1611中采用双对的数据译码方式（步骤1612）。在译码目的地接收数据（步骤1613），并对该数据译码（步骤1614）。以上的处理结束之后，切断压缩元与译码目的地的连接（步骤1615）。

本实施例是按顺序进行步骤1610～1614的处理，但是本发明并不局限于此，也可以流水线式地执行各步骤，这样做的效果是可以提高处理速度。

10 图19和图20表示把用上述的压缩装置和步骤压缩的数据205进行扩展的扩展装置和方法的一个实施例。

图19是把静态辞典并用于动态辞典的数据扩展装置的构成图。在本实施例中，其特征在于采用与图2的压缩装置200成对的构成，并共有压缩装置200和构筑装置242。

15 数据扩展装置1900是把输入数据205作为数据量少的扩展数据1905来输出的装置，与前述的数据压缩装置200一样，用作为数据处理装置的计算机例如个人计算机来实现。或者，也可以连接在计算机装置和网络之间的专用装置来实现。该扩展装置1900具有动态辞典控制器1930和静态辞典控制器1940，还具有控制整个处理步骤的控制器1920。控制器1920由例如装载了OS
20 的计算机装置的CPU构成，控制器1920还控制扩展数据的输入装置1921和扩展了的数据的输出装置1951。

动态辞典控制器1930具有动态辞典1931、构筑装置1932、检索装置1933和登录装置1934。动态辞典1931是存储位串和索引码的数据库，构成在计算机装置的存储装置上，例如构成在装在计算机装置内的硬盘装置上。构筑装置1932、检索装置1933和登录装置1934由操作动态辞典1931的程序构成。
25

这些程序都存储在装在计算机内的硬盘装置上，并且由于计算机装置的启动而把它们转移到主存储装置上，由CPU执行的程序来实现规定的功能。

静态辞典控制器1940具有静态辞典241、构筑装置1942、判定装置1943和参照装置1944。静态辞典241是存储位串和索引码的数据库，构成在计算

机装置的存储装置上，例如构成在装在计算机装置内的硬盘装置上。构筑装置1942、判定装置1943和参照装置1944由操作静态辞典241的程序构成。

这些程序都存储在装在计算机内的硬盘装置上，并且由于计算机装置的启动而把它们转移到主存储装置上，由CPU执行的程序来实现规定的功能。

5 输入装置1921和输出装置1951是计算机硬盘那样的内装存储装置、经由网络的远距离文件系统或分别进行软盘和磁带之类的外存装置之间的数据输入和输出程序（装置驱动器）。在这里输入压缩数据205。

输入装置1921输入输入数据205，并把它变换为输入位串1902后输出。输出装置1951把对应于来自动态辞典1931或静态辞典241的输入位串1902的位串
10 作成为输出位串1903，并作为扩展数据1905输出。

图20是表示在图19中的参照静态辞典241和动态辞典1931的索引码同时扩展压缩数据的步骤的流程图。

判定装置1943输入输入位串1902，并经由参照静态辞典241的参照装置244，包含静态辞典的情况下（步骤2001）参照对应于输入位串1902的索引
15 码和位串（步骤2002），并输出位串（步骤2003）。

在步骤2001中，如果没有对应于静态辞典241的索引码，就用检索装置1933检索动态辞典1931（步骤2011），在包含于动态辞典内的情况下，参照对应于输入文字列1902的动态辞典1931的索引码和位串（步骤2012），并输出位串（步骤2013）。进一步把上次输出的输出位串和这次输出的输出位串的
20 开头位串（例如1字节）组合起来登录在动态辞典上（步骤2014）。

在步骤2011中，当动态辞典1931中未包含输入文字列1902时，检查不良的文字列1902，作为纠错处理结束（步骤2050）。

在以上的例子中，实际上，检索静态辞典和动态辞典，来判定是否包含输出位串1902，但是输入位串1902刚好具有输出位串203和索引码204的数据结构。预先设定索引码204的静态辞典（例如0～1023）和动态辞典（例如
25 1024～）的利用范围，当然也可以不检索辞典高速地进行判定。

计算机装置或个人计算机装置等的数据处理装置一般都设置有以上所说明的数据压缩装置200和数据扩展装置1900两种装置。而且，利用数据压缩装置200把所制成的数据存储在例如内装的硬盘装置内，当要利用该数据
30 时，就从硬盘装置中读出该压缩数据，并用数据扩展装置1900扩展为原数据

加以利用。在这种情况下，静态辞典241就可以共用这两个装置，动态辞典231也可以根据情况共用这两个装置。在数据处理装置用网络可以进行数据收发地相互连接的情况下，用数据压缩装置200从一方的数据处理装置把数据进行压缩传送，接收到该数据的数据处理装置就用数据扩展装置1900把它扩展为原数据，并利用它。

图17是在分布系统上不能共用静态辞典的情况的实施例。数据压缩装置200用动态辞典231或静态辞典241进行数据压缩，数据译码装置260用动态辞典261或静态辞典262进行数据译码。数据压缩装置200和图2中的数据压缩装置200是一样的，输入普通文字201，输出压缩数据。数据译码装置260是进行数据压缩装置200的逆数据变换的装置，输入压缩数据，而输出普通文字201。数据压缩装置200被结合在编码装置252内，编码装置252进一步被结合在数据传送装置270中。数据传送装置270通过网络与数据传送装置271连接，数据传送装置271被结合在解码装置253中，解码装置253进一步被结合在数据译码装置260内。编码装置252和解码装置253分别具有静态辞典代码变换表250和251。假如输入包含“COMPUTER”和“THIS”单词的普通文字数据201，由静态辞典241进行过压缩处理。“COMPUTER”和“THIS”在数据压缩装置200中分别被变换成为索引码“251”和“357”，并被作为压缩数据的一部分送到编码装置252。编码装置252接收这些数据，并参照静态辞典代码变换表250变换为通用索引码“2047”和“1023”。被变换为通用索引码的压缩数据经数据传送装置270、网络以及数据传送装置271发送到解码装置253。解码装置253接收该压缩数据并参照静态辞典代码变换表251把包含在其中的通用索引码“2047”和“1023”变换为静态辞典262中固有的索引码“34”和“143”。然后把包含索引码“34”和“143”的压缩数据传送到数据译码装置260。数据译码装置260接收该压缩数据，并用静态辞典262把索引码“34”和“143”分别变换为“COMPUTER”和“THIS”，然后把包含它们的文字作为原来的普通文字201输出。

按照本实施例，在编码装置252和解码装置253中设置有把静态辞典241和262的索引码变换为通用索引码的辞典代码变换表250和251，由于在传送部分定义标准的代码形式，所以具有可以保证不同结构格式的压缩装置・译码装置之间的数据的互换性的效果。

在本实施例中，虽然编码装置252和解码装置253仅进行索引码的代码变换，但是本发明并不仅仅局限于此，也可以分别组合进行数据的加密和译码。为此，可以提供数据的压缩加密装置，从而能够实现确保网络上的数据安全性的压缩数据的传送。

- 5 按照本发明，由于在应用动态辞典的数据压缩方法中并用静态辞典，所以与单独应用动态辞典的压缩方法相比可以提高压缩率。特别是由于在难以进行动态辞典的压缩的输入数据的最初的部分的压缩时采用静态辞典，所以能够进一步地提高压缩率。另外，由于把区别动态辞典和静态辞典的代码附加压缩数据的索引码上，所以，能够容易地进行译码时的辞典的切换处理。
- 10 而且，由于把输入位串的代码、静态辞典的索引码或动态辞典的索引码之中比特长度最短的代码进行编码，并作为压缩数据输出，所以，能够进一步提高压缩率。由于限于与静态辞典的索引码相比动态辞典的索引码为短的情况下才把输入位串登录在动态辞典中，所以，能够使动态辞典的大小最小。特别是由于把与静态辞典的内容一致的位串的索引码作为压缩数据输出，并不
- 15 把它追加到动态辞典上，所以，能够缩小动态辞典的规模，并且，因为索引码的长度也短，因此能够进一步提高压缩率。另外，当按顺序把输入位串登录在动态辞典中，并在动态辞典中的数据量达到规定的数量时，由于采用对此后的输入位串不进行登录的方法、或根据需从动态辞典中删除利用率低的位串的方法，或两种方法兼用，这样就能够解决位串溢出动态辞典的问题。
- 20

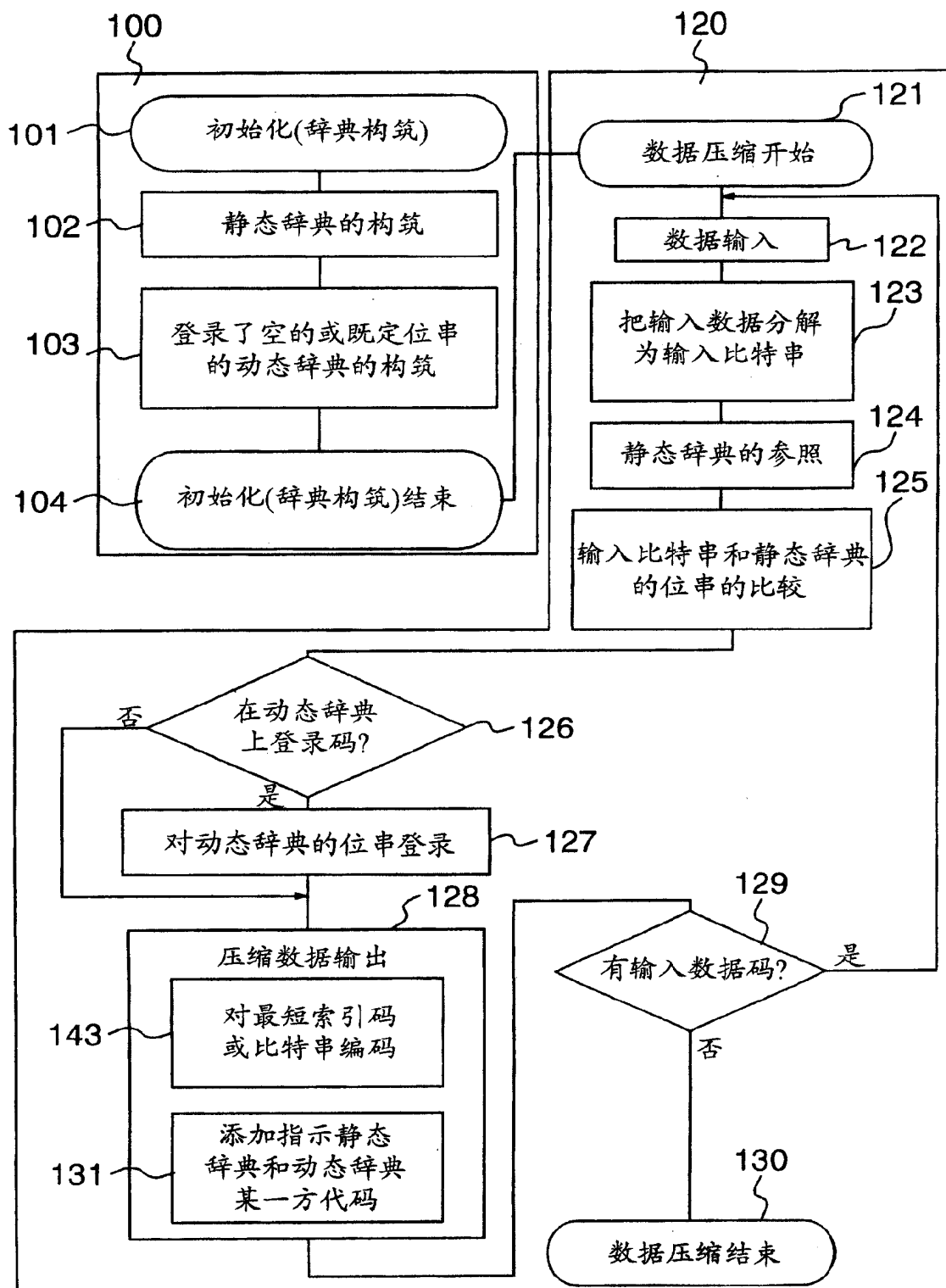


图 1

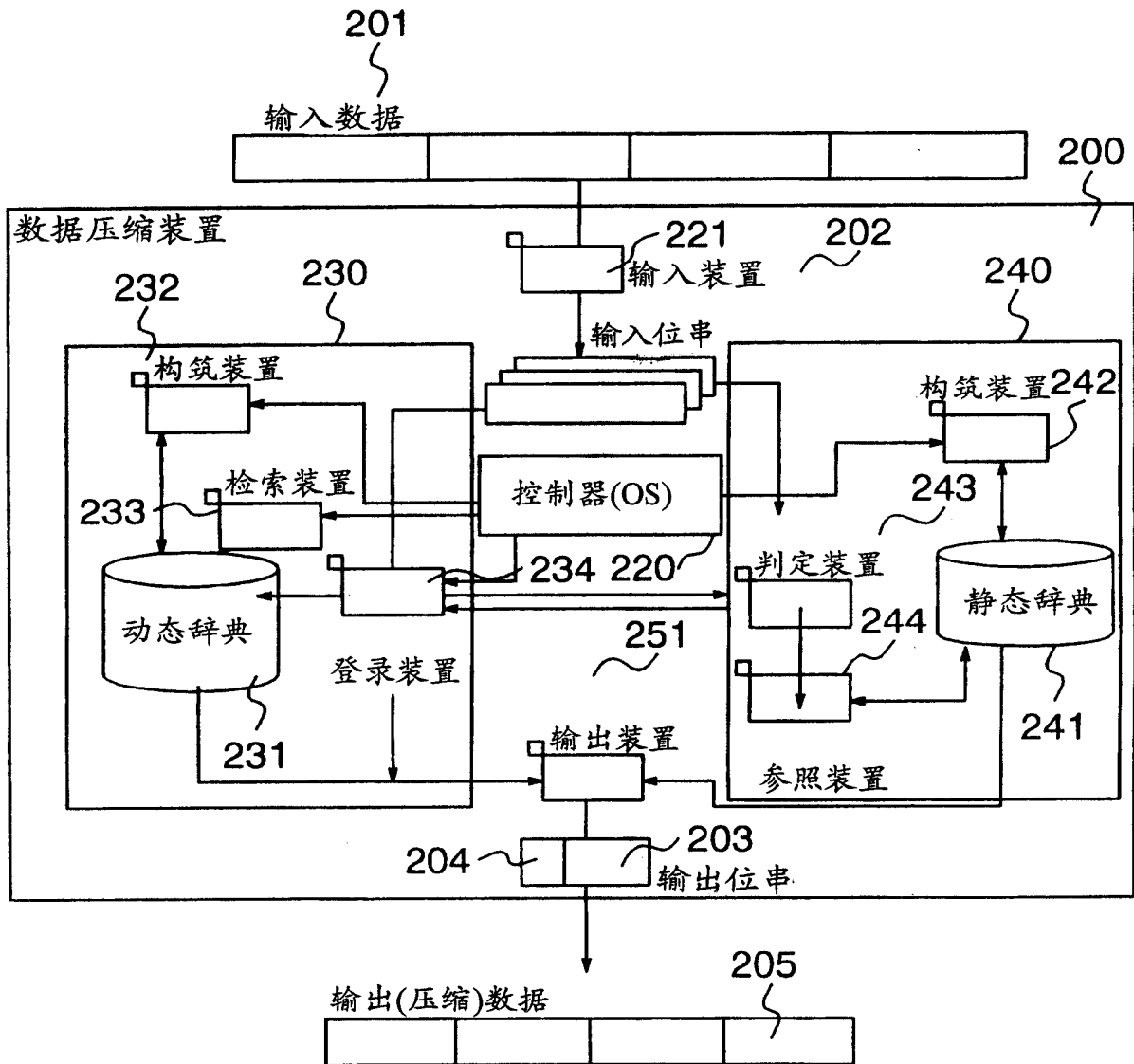


图 2

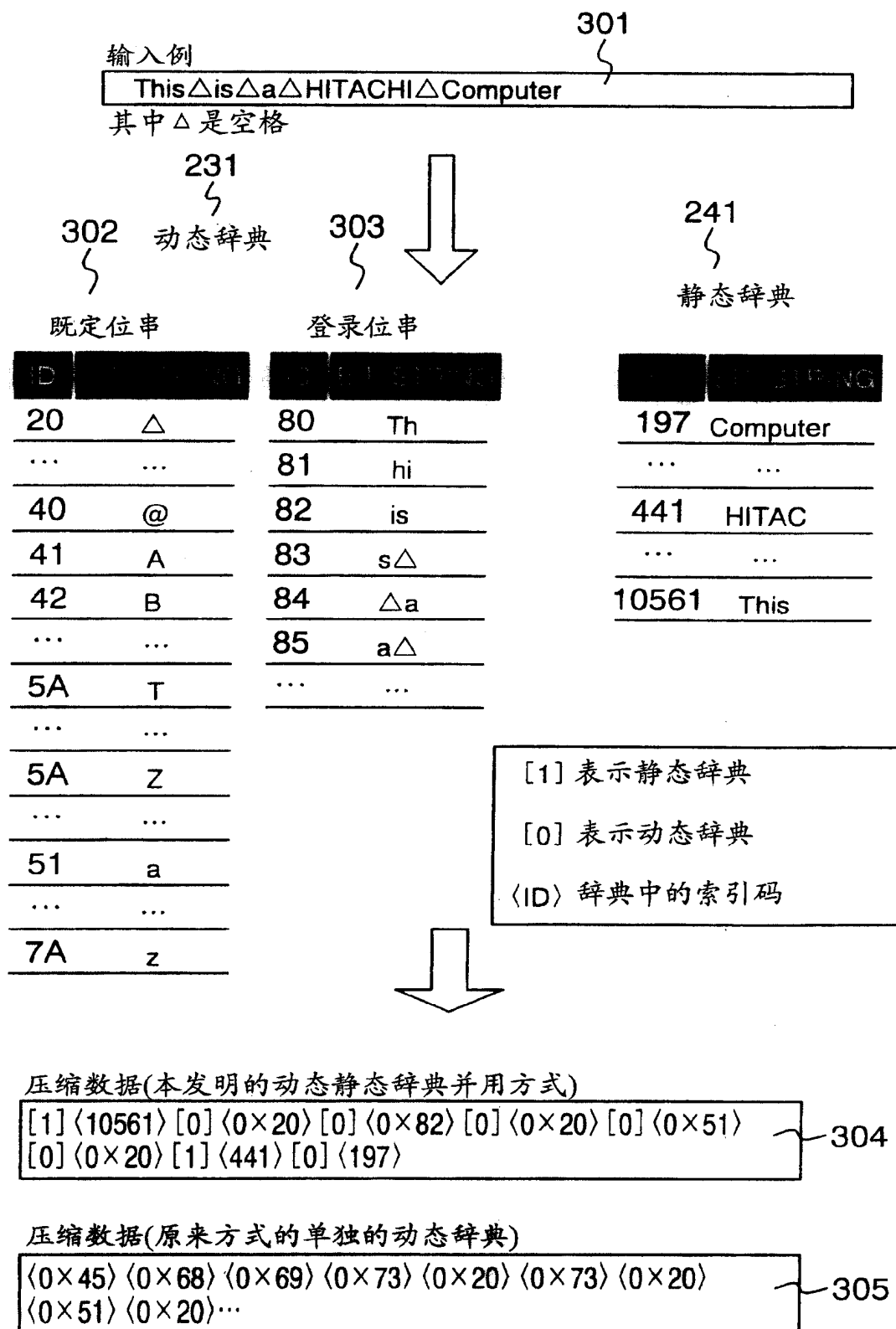


图 3

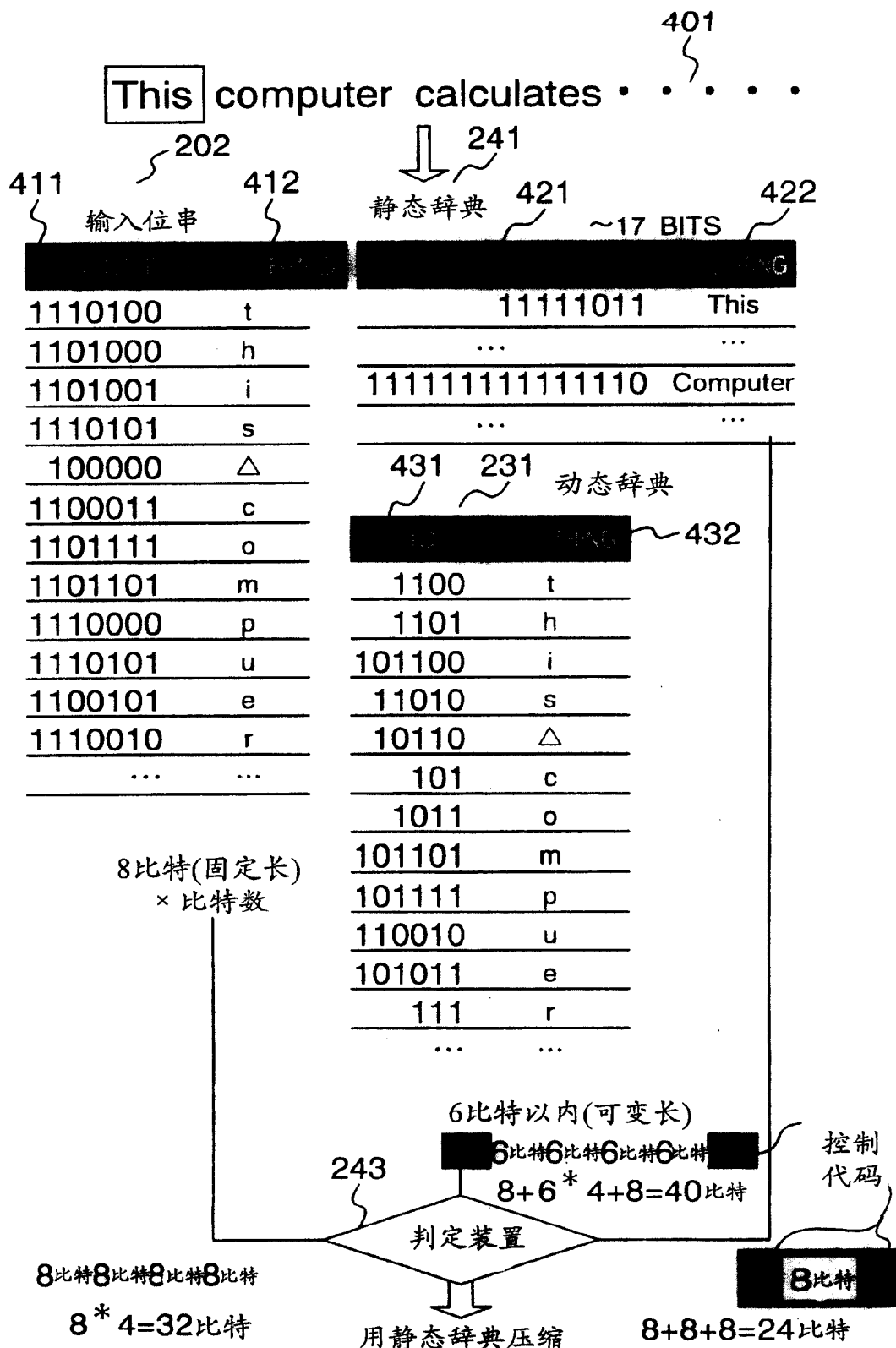


图 4

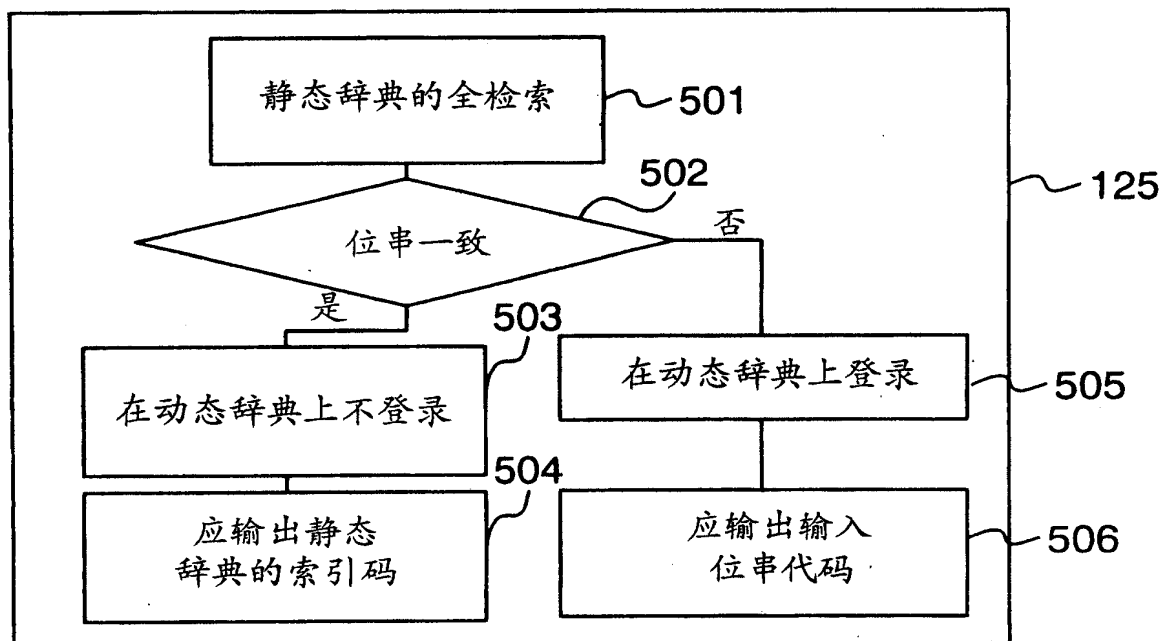


图 5

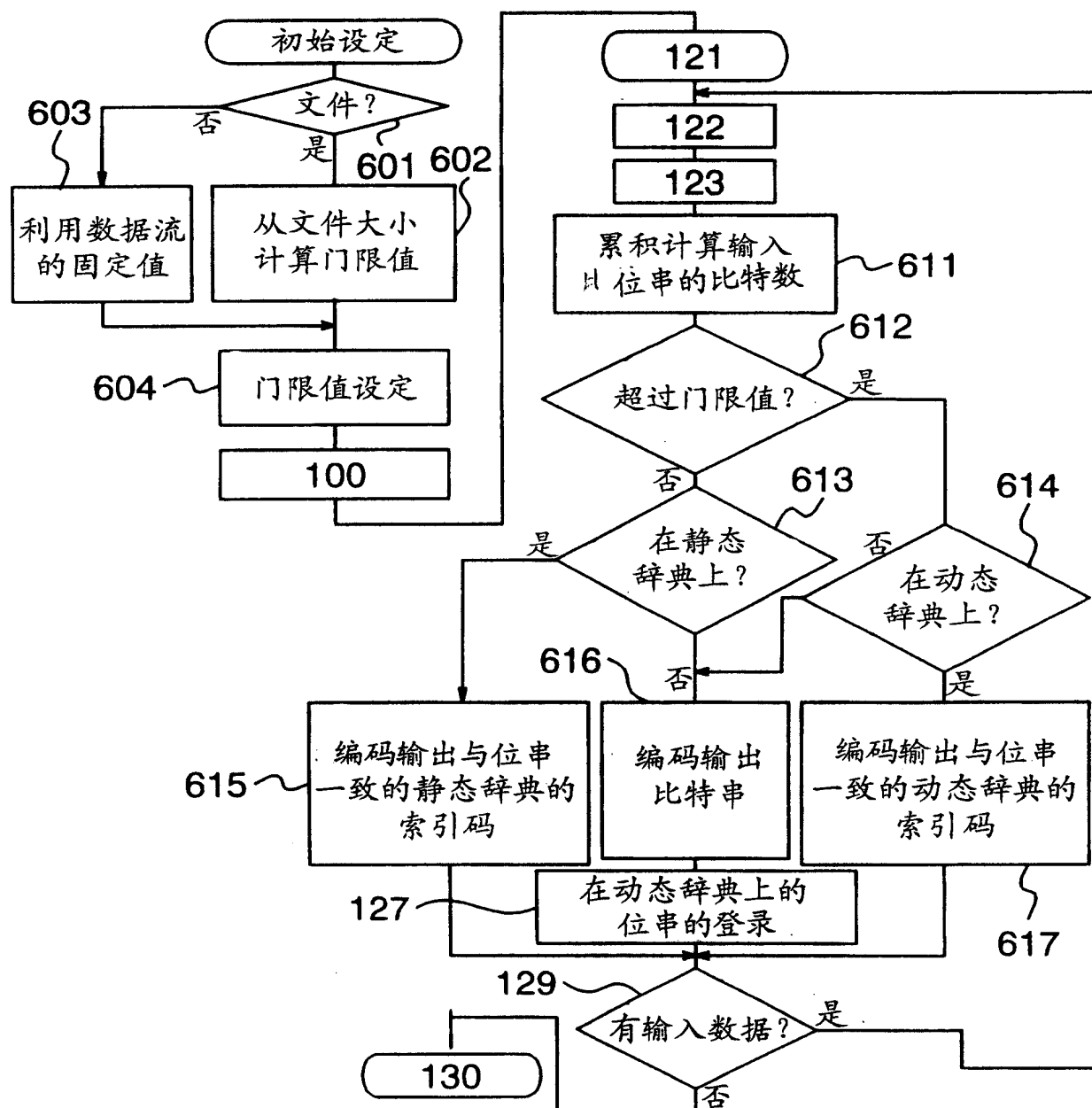


图 6

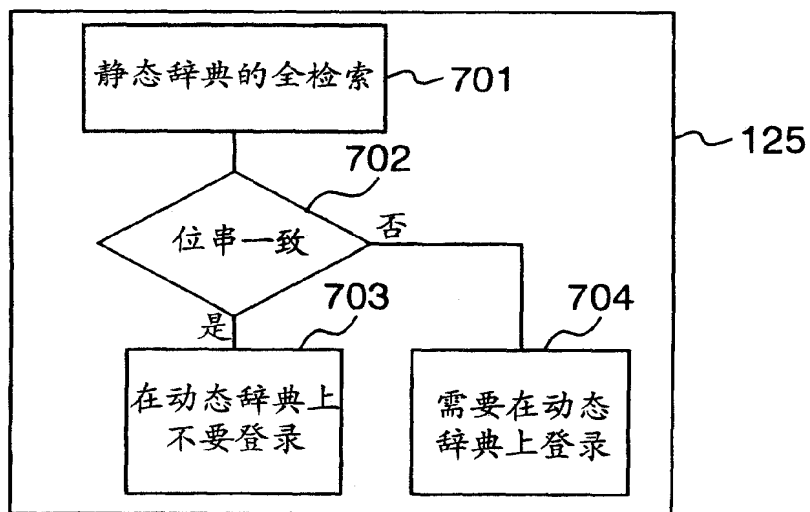


图 7

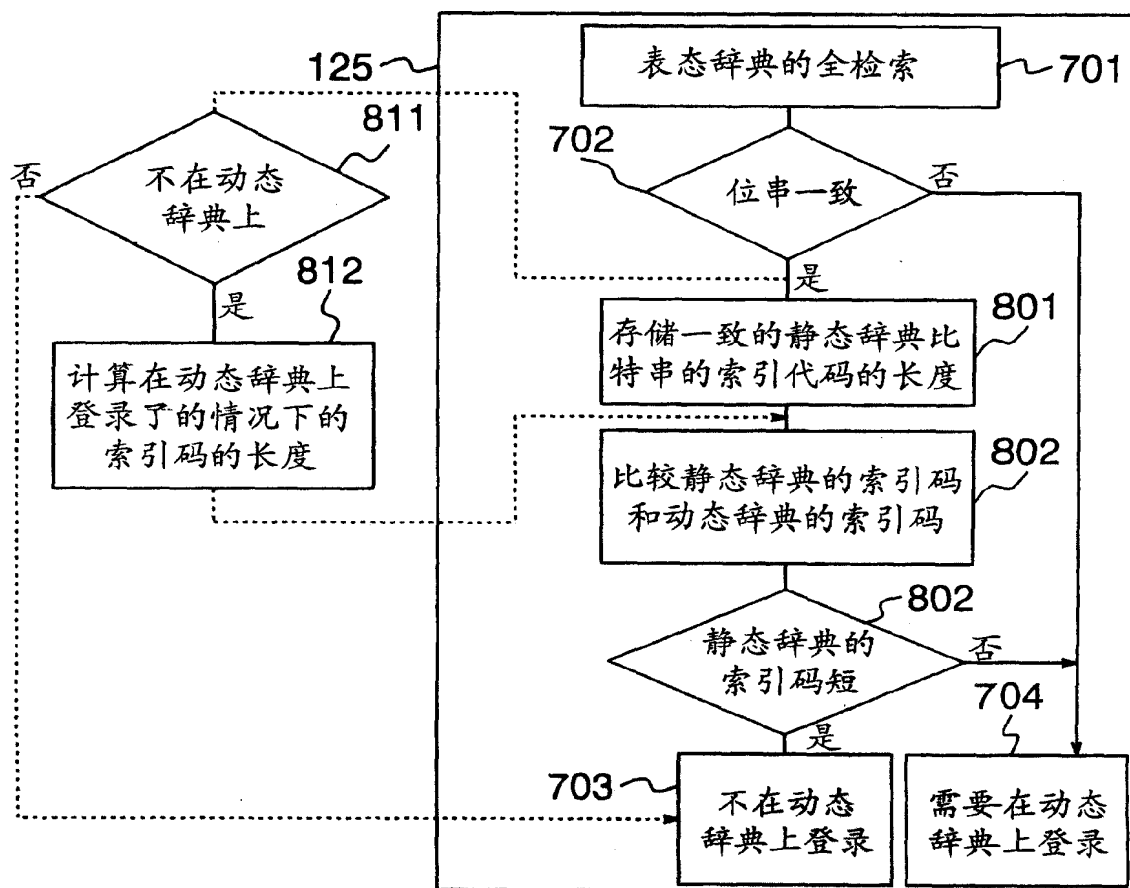
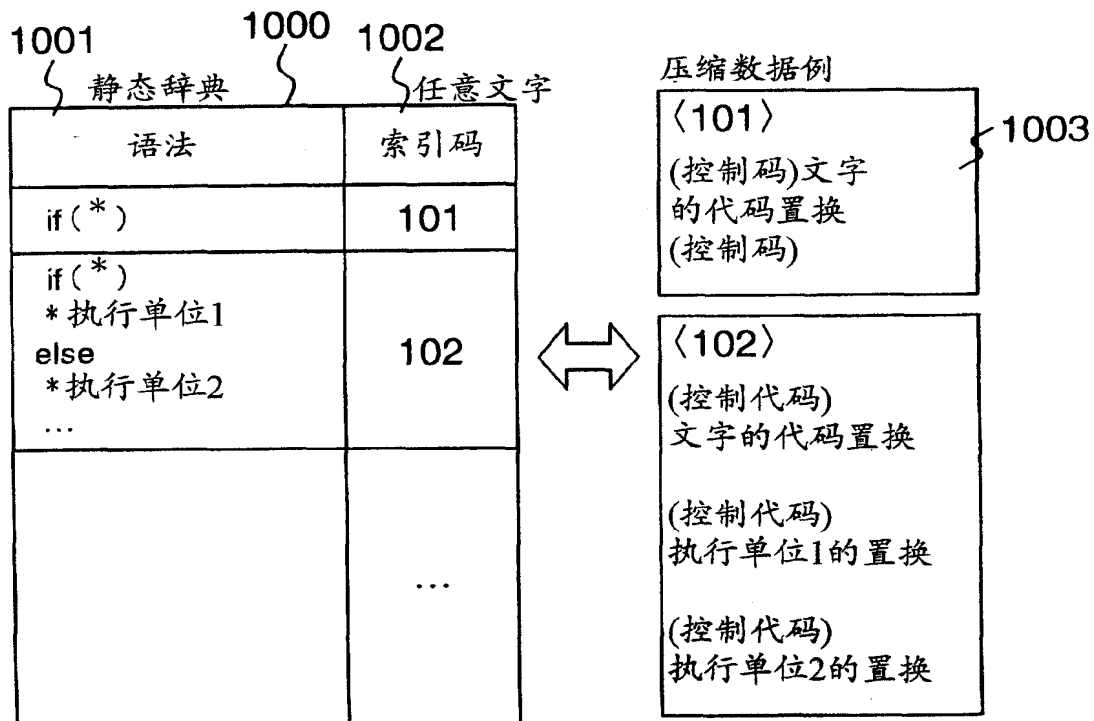


图 8

901 保留字	902 2次以上的 出现概率	903 索引码
if	0.089	2
else	0.075	3
swtich	0.043	5
case	0.043	6
while	0.051	4
for	0.095	1
do	0.040	8
goto	0.012	11
return	0.038	9
break	0.042	7
continue	0.03	10
...	...	...

900

图 9



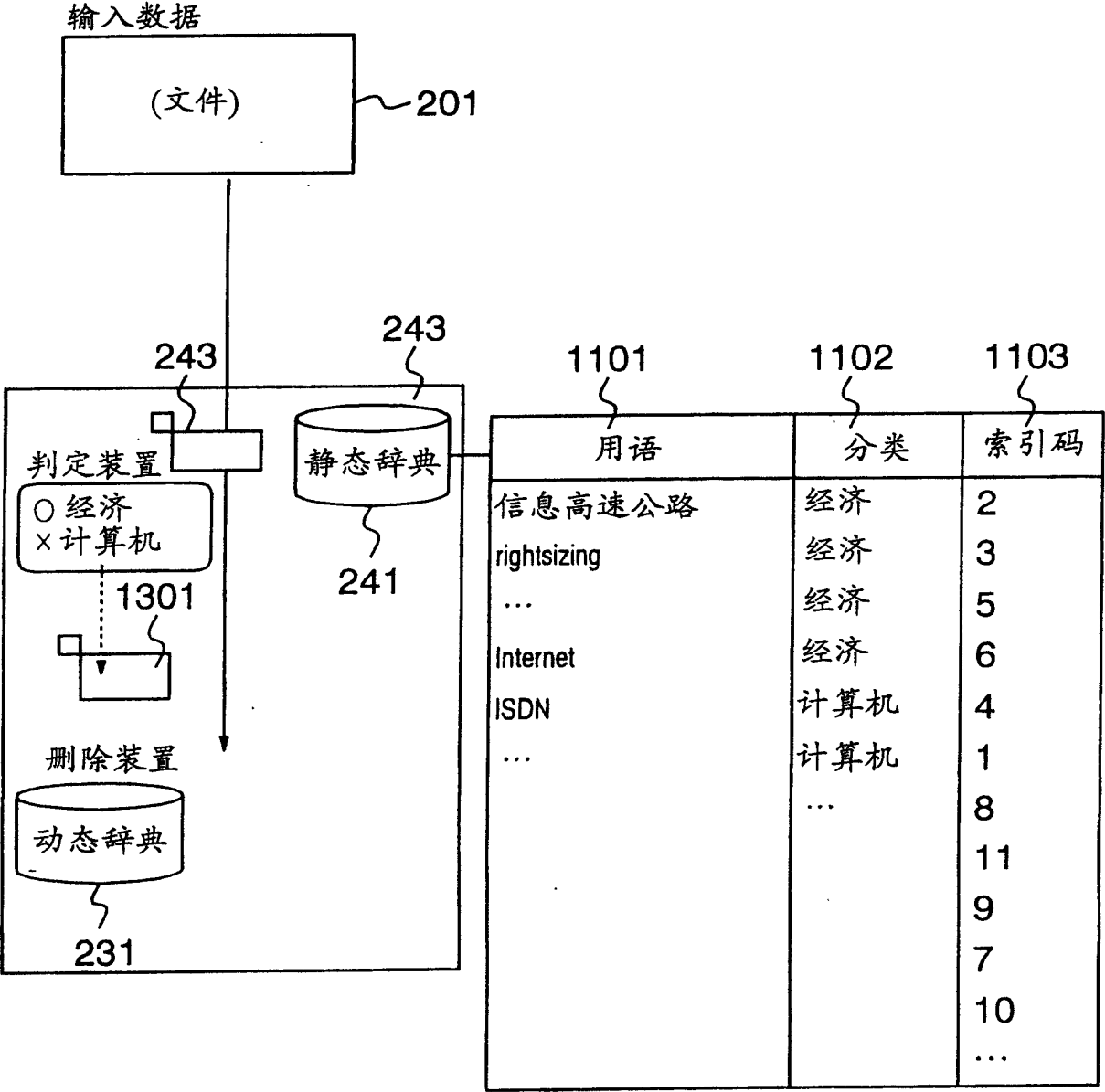


图 11

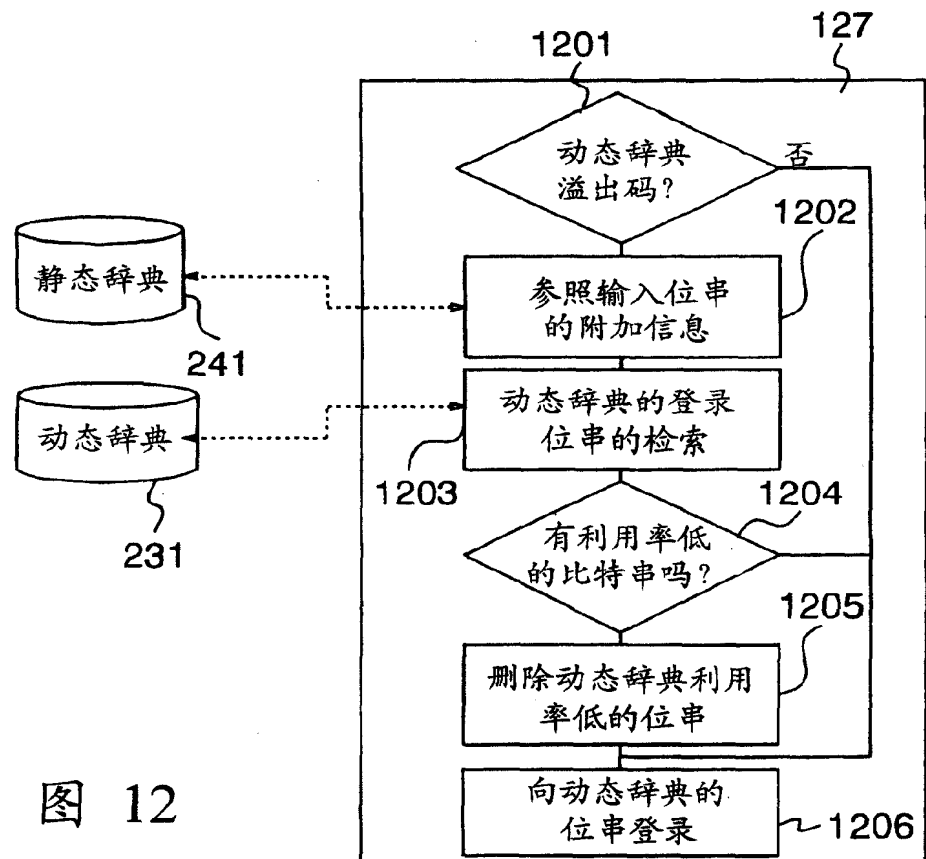


图 12

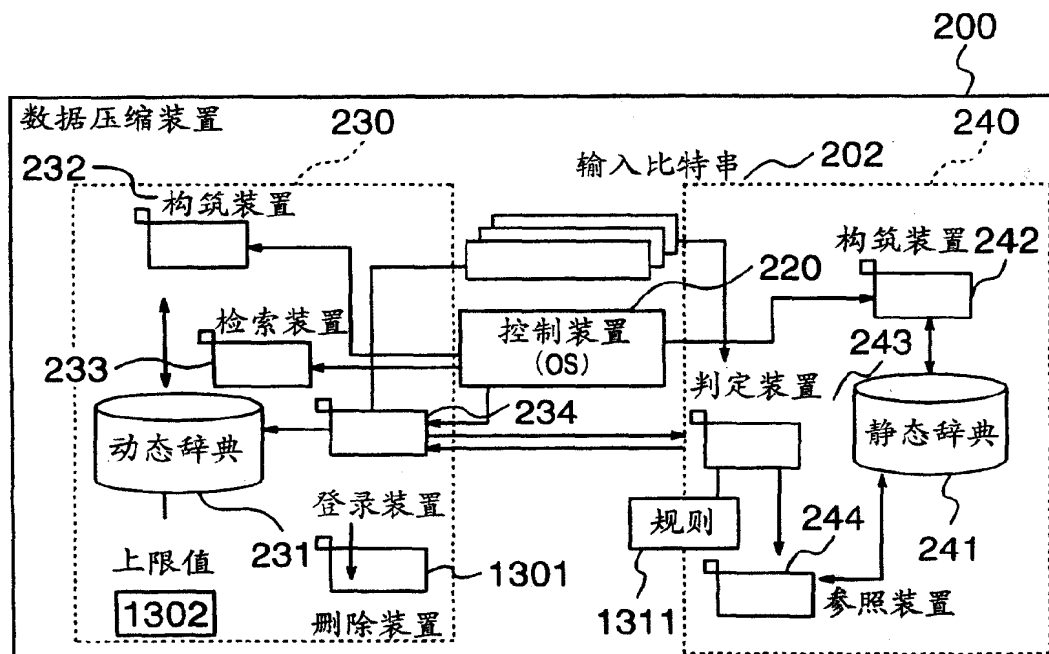


图 13

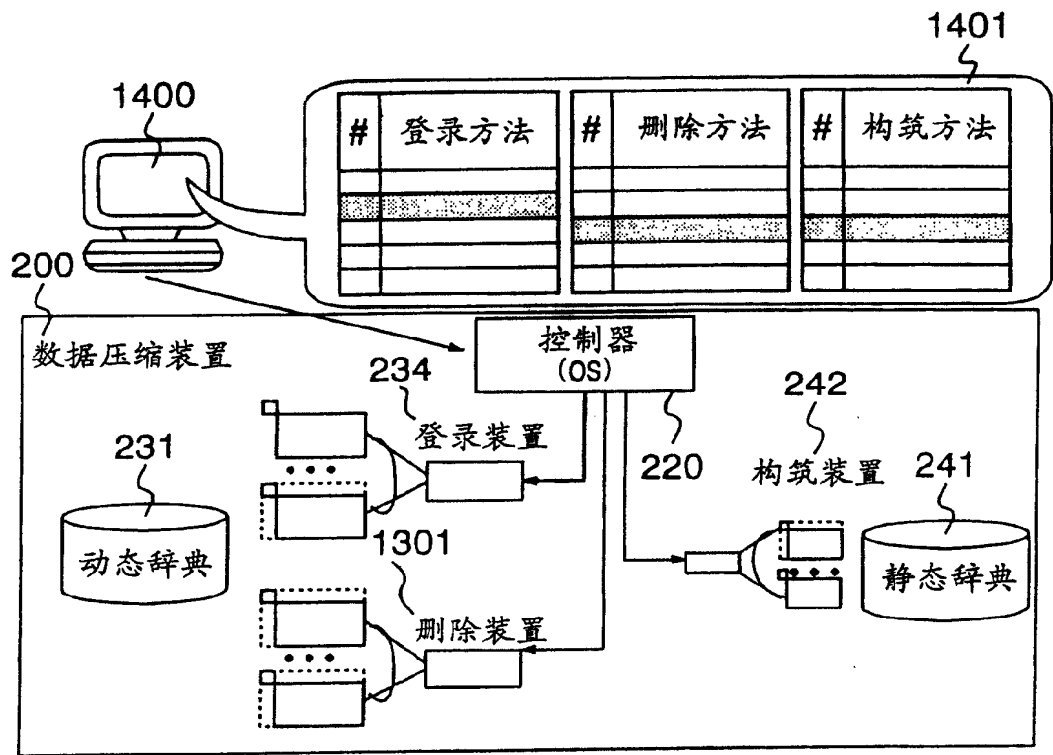


图 14

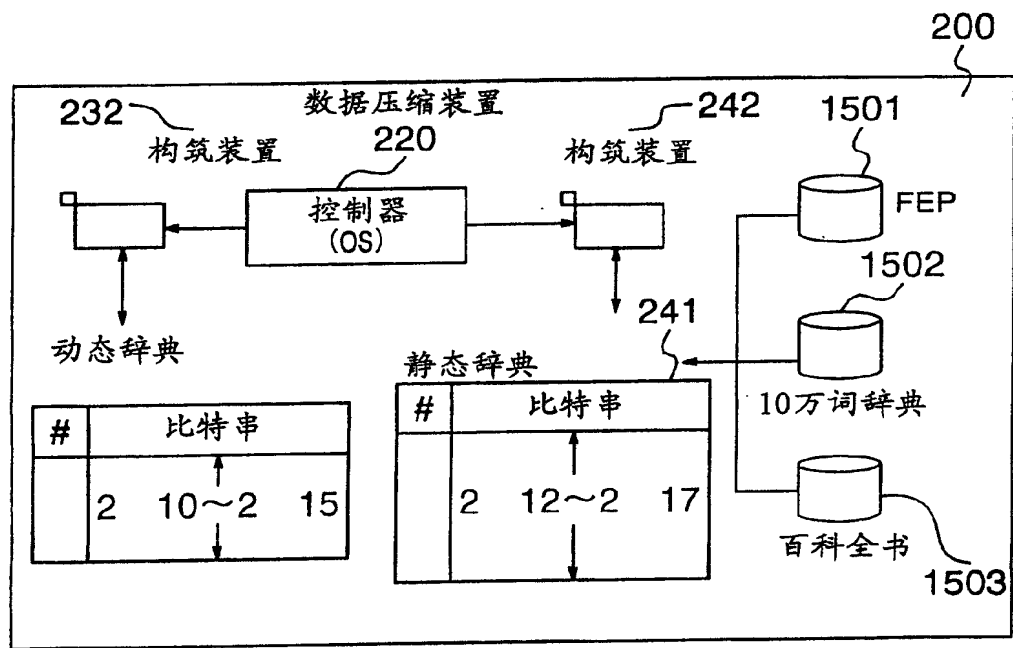


图 15

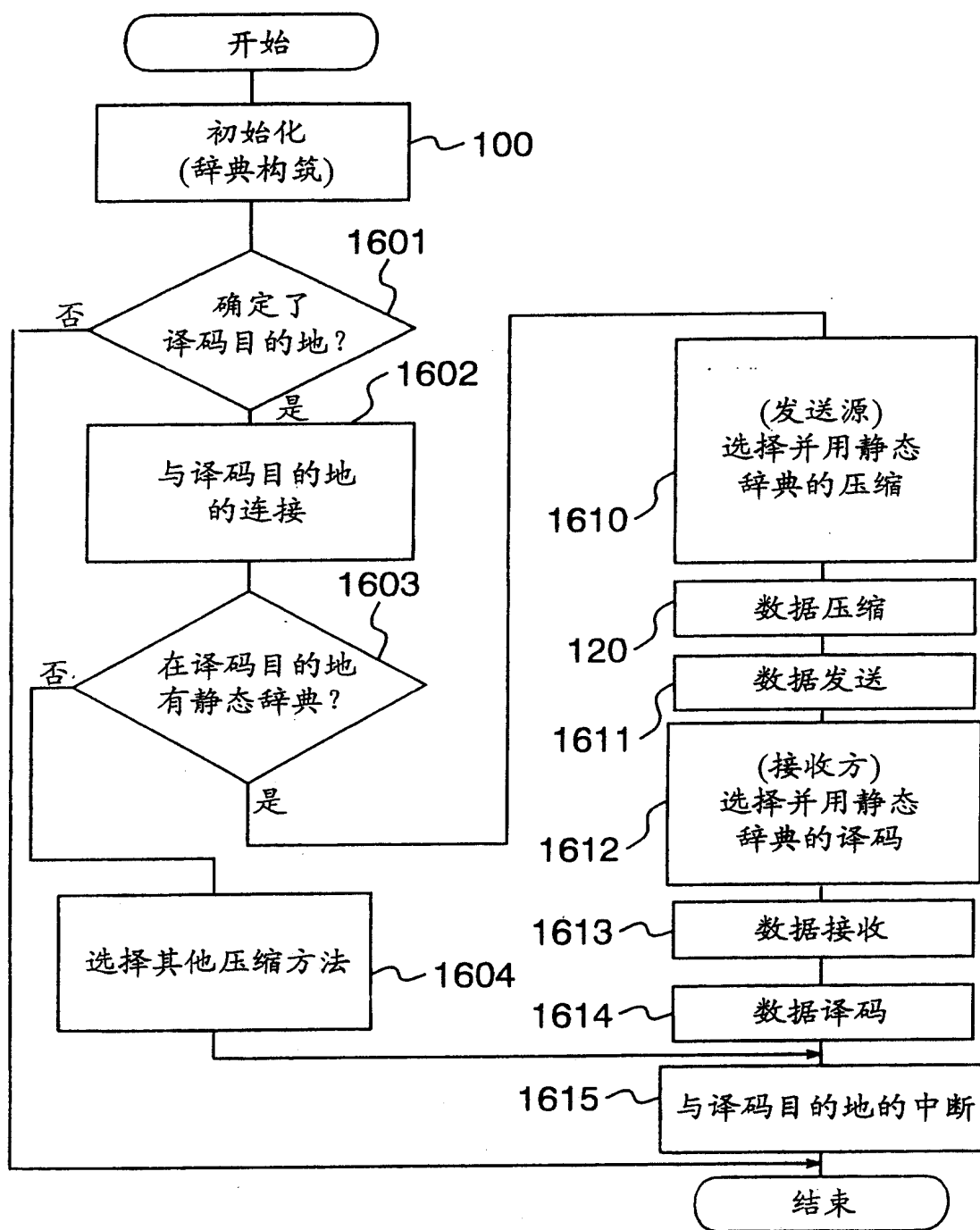


图 16

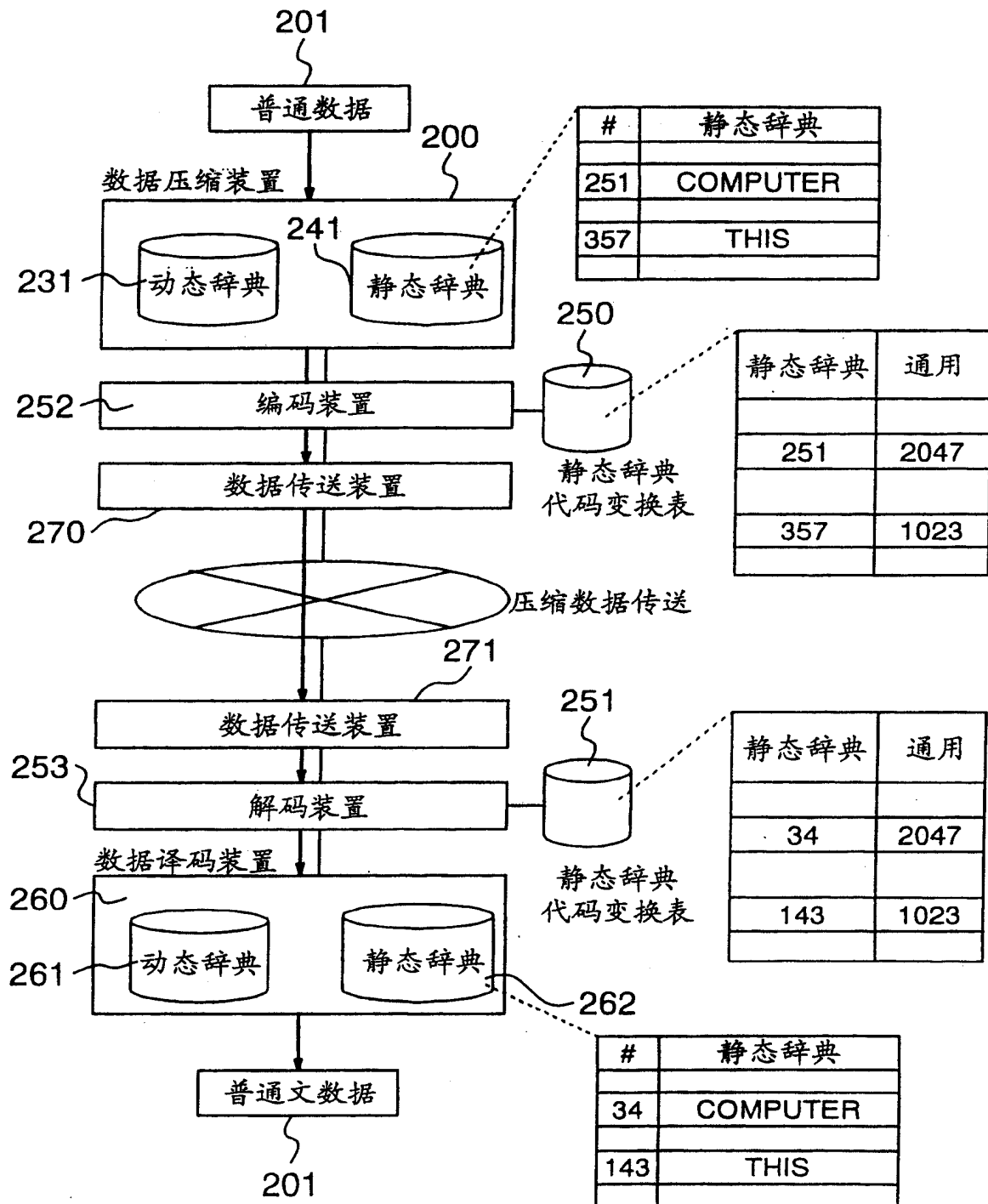


图 17

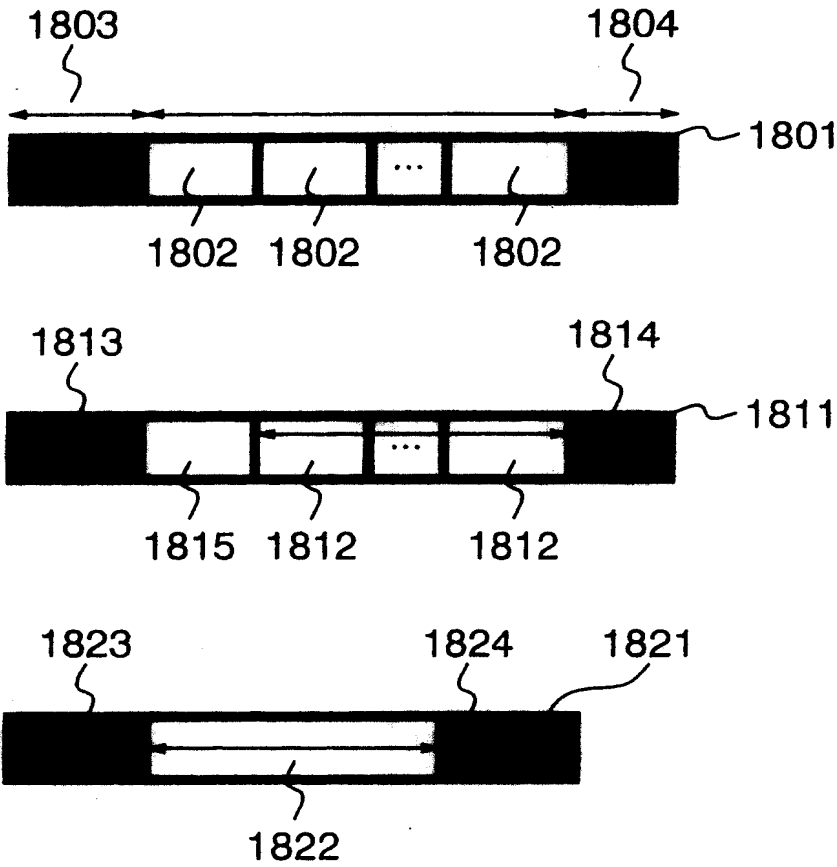


图 18

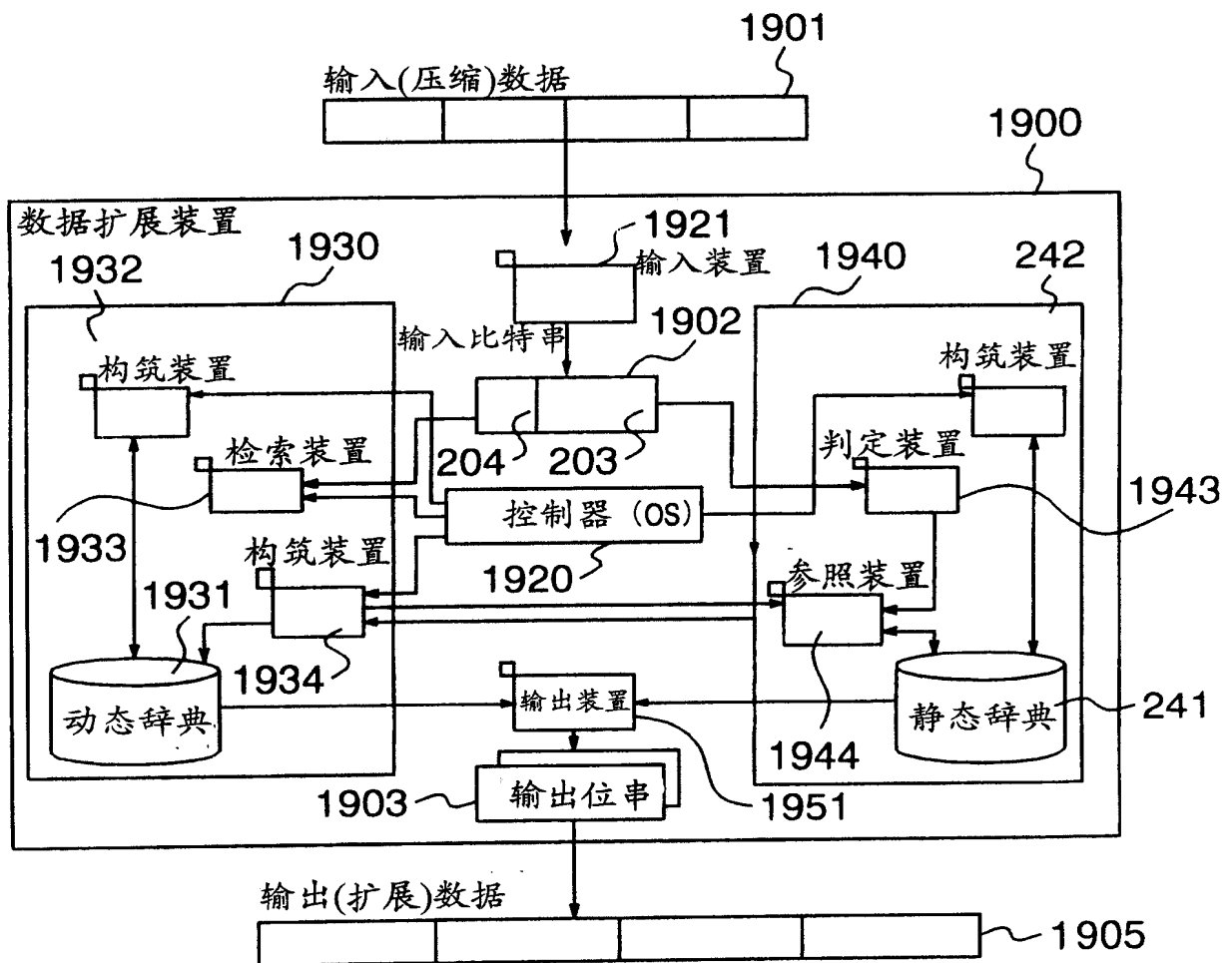


图 19

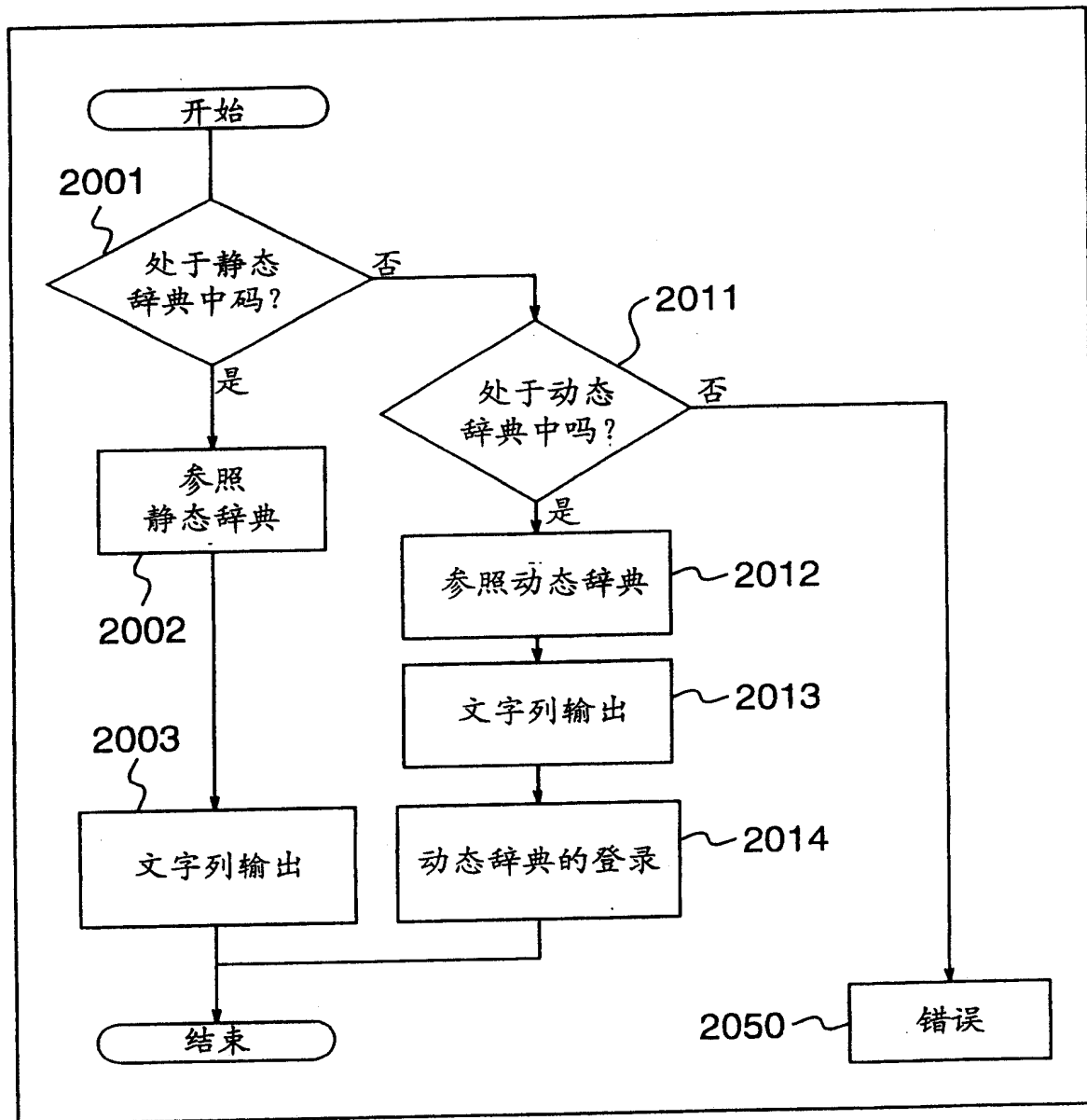


图 20