



[12] 发明专利申请公开说明书

[21] 申请号 03815759.4

[43] 公开日 2005 年 11 月 30 日

[11] 公开号 CN 1703692A

[22] 申请日 2003.7.3 [21] 申请号 03815759.4
[30] 优先权
[32] 2002. 7. 3 [33] US [31] 60/393,903
[32] 2002.11. 7 [33] US [31] 10/289,656
[32] 2002.11. 7 [33] CA [31] 2,411,227
[86] 国际申请 PCT/CA2003/001023 2003.7.3
[87] 国际公布 WO2004/006122 英 2004.1.15
[85] 进入国家阶段日期 2005.1.4
[71] 申请人 2012244 安大略公司
地址 加拿大安大略省
[72] 发明人 瓦迪姆·富克斯
迈克尔·G·埃利萨荣

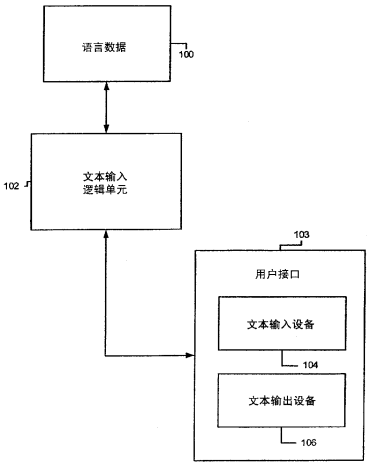
[74] 专利代理机构 中科专利商标代理有限责任公司
代理人 罗松梅

权利要求书 6 页 说明书 15 页 附图 9 页

[54] 发明名称 创建和使用紧凑语言数据的系统和方法

[57] 摘要

提出了一种创建和使用紧凑语言数据的系统和方法。计算出现在语料中的单词的频率。将单词中的每个惟一字符映射到字符索引，并以字符索引替代单词中的字符。将字符序列映射到置换索引，并以置换索引替代单词中的字符序列。按照公共前缀，将单词分组，并将每个前缀映射到针对以该前缀开头的单词组的位置信息。



ISSN 1008-4274

1、 一种创建紧凑语言数据的系统，包括：

5 语料；和

语言数据分析器，

其中所述语言数据分析器计算语料中单词的出现频率，将单词中
每个惟一的字符映射到字符索引，并以将所述字符映射到其的字符索
引替代所述单词中的每个字符，将出现在单词中的字符序列映射到置
10 换索引，并以将所述字符序列映射到其的置换索引替代每个单词中的
每个字符序列，将单词分成组，每个组包含以公共前缀开头的单词，
并将每个前缀映射到针对以所述前缀开头的单词组的位置信息；以及

其中所述紧凑语言数据包括所述惟一字符、所述字符索引、所述
置换索引、所述位置信息、所述单词组和所述单词的频率。

15 2、 根据权利要求1所述的系统，其特征在于还包括：

用户接口，包括：

文本输入设备；和

文本输出设备；以及

文本输入逻辑单元，

20 其中所述文本输入逻辑单元接收来自所述文本输入设备的文本
前缀，从所述紧凑语言数据中检索以所述前缀开头的多个预测单词，
并利用所述文本输出设备显示所述预测单词。

3、 根据权利要求2所述的系统，其特征在于所述文本输入设备
是键盘。

25 4、 根据权利要求3所述的系统，其特征在于所述键盘是简化键
盘。

5、 根据权利要求2所述的系统，其特征在于所述用户接口和所
述文本输入逻辑单元是在移动通信设备上实现的。

6、 一种针对多个单词的紧凑语言数据结构，其中将单词组织成
30 组，每个组包含具有公共前缀的单词，所述紧凑语言数据结构包括：

由所述单词中的每个惟一字符构成的字母表；

用于将所述字母表中的每个字符映射到字符索引的字符映射表；

用于将来自所述单词的字符序列映射到置换索引的置换表；

5 用于存储所述单词的多个单词定义表，每个单词定义表存储包括在所述组之一中的每个单词；以及

用于定位所述单词定义表的偏移量表，其中对于每个公共前缀，所述偏移量表包含存储了以所述公共前缀开头的单词的单词定义表的位置，

10 其中通过以由所述字符映射表将所述字符映射到其的所述字符索引代替所述单词中的每个字符，以及通过以由所述置换表将所述字符序列映射到其的置换索引代替出现在所述单词中的、来自所述置换表的每个字符序列，对所述单词定义表中的每个单词进行编码；以及其中去除每个单词定义表中的单词的所述公共前缀。

15 7、 根据权利要求6所述的紧凑语言数据结构，其特征在于还包括与所述单词相关联的多个频率，其中所述单词定义表存储所述频率。

8、 根据权利要求7所述的紧凑语言数据结构，其特征在于对所述频率进行标准化，从而使其在预定的范围内，并且针对每个单词组，以全精度存储该组中具有最大频率的单词的频率，并将该组中剩余单词的频率存储为所述最大频率的百分数。

20 9、 根据权利要求7所述的紧凑语言数据结构，其特征在于所述单词定义表中的至少一个包括：

用于存储由所述单词定义表存储的所述单词的多个本地单词定义表，每个本地单词定义表存储由所述单词定义表存储的、具有比所述单词定义表中的单词的公共前缀长的公共前缀的单词；

25 用于定位每个所述本地单词定义表的本地偏移量表；

用于将所述本地单词定义表中的所述单词中的每个字符映射到本地字符索引的本地字符映射表；

用于存储与包含在所述单词定义表中的组中的单词的最高频率相关联的单词在所述本地单词定义表中的位置的热词表；以及

30 用于存储与包含在所述单词定义表中的组中的单词的最高频率

相关联的单词的例外表，

其中通过以由所述本地字符映射表将所述字符映射到其的本地字符索引替代按此单词中的字符，以及通过以由所述置换表将所述字符序列映射到其的置换索引替代出现在单词中、来自所述置换表的每个字符序列，对所述本地单词定义表中的每个单词进行编码，以及
5 其中去除所述本地单词定义表中的单词的所述公共前缀。

10、根据权利要求6所述的紧凑语言数据结构，其特征在于还包括名称表，所述名称表包括用于识别所述单词的名称。

11、根据权利要求6所述的紧凑语言数据结构，其特征在于还包括用于存储单词的后缀的词形变化表，其中存储在所述单词定义表中的单词引用存储在所述词形变化表中的所述后缀。
10

12、根据权利要求6所述的紧凑语言数据结构，其特征在于包括在所述置换表中的字符序列具有不同的长度，其中，对于每个长度，所述置换表包含子表，将该长度的字符序列映射到置换索引，并且根据包含在所述子表中的序列的长度，按照降序组织所述子表。
15

13、根据权利要求6所述的紧凑语言数据结构，其特征在于所述偏移量表将来自所述字母表的每个两字符组合映射到所述偏移量表中等于所述组合中的第一字符在字母表中的位置乘以所述字母表中的字符数的积与所述组合中的第二字符在字母表中的位置的和的索引。

14、根据权利要求6所述的紧凑语言数据结构，其特征在于所述单词以具有最高有效位的比特序列表示，其中置位所述单词定义表中的每个单词中的最后一个字符中的最高有效位，并且根据除最高有效位之外的其他比特的比特值，确定所述单词中的每个字符的值。
20

15、根据权利要求7所述的紧凑语言数据结构，其特征在于还包括：
25

新单词表，用于存储利用所述字符映射表和所述置换表对其进行编码、且不包括在所述单词定义表中的新单词、与所述新单词相关联的多个频率、用于所述新单词的开头的索引、以及规定了所述新单词的字母排序次序的排序索引；以及

30 更新频率表，用于存储利用所述字符映射表和所述置换表对其进

行了编码的、存储在所述单词定义表中的单词、与存储在所述单词定义表中的单词相关联的已更新频率、用于所述更新频率表中的单词的开头的索引、以及规定了所述更新频率表中的单词的字母排序次序的排序索引。

5 16、根据权利要求 6 所述的紧凑语言数据结构，其特征在于所述字符映射表、所述置换表、所述单词定义表和所述偏移量表包含在计算机可读文件中。

17、根据权利要求 16 所述的紧凑语言数据结构，其特征在于所述计算机可读文件还包括头部，所述头部包括格式、版本和优先级信息。
10

18、根据权利要求 16 所述的紧凑语言数据结构，其特征在于还包括索引表，所述索引表将索引包含在所述计算机可读文件中，用于定位所述计算机可读文件中的所述表格。

19、根据权利要求 16 所述的紧凑语言数据结构，其特征在于在
15 计算机可读介质中实现所述计算机可读文件。

20、一种创建紧凑语言数据的方法，包括以下步骤：

创建包括最常出现在语料中的多个单词的单词列表；

按照字母表顺序对所述单词列表中的所述单词进行排序；

创建将所述单词中的字符映射到字符索引的字符映射表；

20 将所述单词列表中的单词分组，其中每个组中的单词具有公共前缀；

创建将所述单词中的字符序列映射到置换索引的置换表；

通过以由所述字符映射表将所述字符映射到其的字符索引替代所述单词中的字符，以及通过以由所述置换表将所述字符序列映射到
25 其的替换索引替代所述组中的所述单词中的字符序列，将单词编码为字节序列；

创建单词定义表并将所述编码单词存储在所述单词定义表中；

创建用于定位编码单词组的偏移量表；

存储所述字符映射表、所述置换表、所述单词定义表和所述偏移
30 量表。

21、 根据权利要求 20 所述的方法，其特征在于所述存储步骤包括将所述字符映射表、所述置换表、所述单词定义表和所述偏移量表写入计算机可读文件。

22、 根据权利要求 20 所述的方法，其特征在于：

5 所述创建单词列表的步骤包括以下步骤：

将所述语料过滤成多个已过滤单词；

计算与所述已过滤单词相关联的绝对频率；以及

将所述语料中最常出现的已过滤单词与所述最常出现的已过滤单词相关联的绝对频率存储在所述单词列表中；以及

10 所述方法还包括以下步骤：

通过将标准化函数应用于存储在所述单词列表中的绝对频率，创建多个相对频率；

创建与所述单词组相关联的频率集合；以及

将所述频率集合存储在所述单词定义表中，

15 其中对于每个组，所述关联频率集合包含该组中单词的最大相对频率，并且针对组中未与最大相对频率相关联的每个单词，包含与关联于所述单词的相对频率相对应的所述最大相对频率的百分数。

23、 根据权利要求 22 所述的方法，其特征在于所述计算与已过滤单词相关联的绝对频率的步骤包括以下步骤：

20 创建用于存储最常出现的已过滤单词及与所述最常出现的单词相关联的绝对频率的单词树，其中将与较高绝对频率相关联的单词存储在所述单词树的内节点中，而将与较低绝对频率相关联的单词存储在所述单词树的叶节点中；

25 将所述已过滤单词添加到所述单词树中，每个已过滤单词的添加步骤包括以下步骤：

确定所述已过滤单词是否在所述单词树中；

在所述已过滤单词不在所述单词树中的情况下，将所述已过滤单词及其关联绝对频率添加到所述单词树中；

30 在所述已过滤单词在所述单词树中的情况下，递增与所述已过滤单词相关联的绝对频率；

确定所述单词树中的节点数是否超过预定限制；以及

在所述单词树中的节点数超过所述预定限制的情况下，缩减所述单词树，其中所述缩减步骤包括删除包含与最低绝对频率相关联的单词的叶节点的步骤。

- 5 24、根据权利要求 22 所述的方法，其特征在于所述过滤语料的步骤包括以下步骤：

读取多个计算机文件，每个文件具有表示所述文件的类型的扩展名；

- 10 对包含在所述文件中的文本应用常规表述，以置换所述文本中的无效或不想要的字符；

获得针对每个文件的过滤器，所述过滤器对应于由所述扩展名所表示的类型；以及

将所述过滤器用于从所述文件中去除不是语料的一部分的文本。

- 15 25、根据权利要求 22 所述的方法，其特征在于还包括：

- 将用户从预测单词列表中选择的单词添加到学习单词列表中，其中所述预测列表包含从所述单词定义表和所述学习单词列表中检索出的单词；

获得与所述预测单词列表中的单词的最大频率相关联的单词；

- 20 确定与所述最大频率相关联的所述单词是否是从所述单词定义表中获得的；

在与所述最大频率相关联的所述单词是从所述单词定义表中获得的情况下，将所述选定单词与等于所述最大频率加一的频率相关联；

在与所述最大频率相关联的所述单词不是从所述单词定义表中获得的情况下，将所述选定单词与等于所述最大频率的频率相关联；

- 25 以及

在与所述最大频率相关联的所述单词不是从所述单词定义表中获得的情况下，从所述学习单词列表中删除曾经被分配给所述最大频率的单词。

创建和使用紧凑语言数据的系统和方法

5

相关申请的交叉引用

本申请要求2002年7月3日递交的美国临时申请S/N 60/393,903的优先权。该临时申请的全部公开，包括附图，在此一并作为本申请的参考。

10

技术领域

本发明大体上涉及语言数据，更具体地，涉及针对文本处理和文本输入的语言数据的存储和使用。

15

背景技术

移动设备和不同类型嵌入式系统越来越多的使用向这些设备的开发商和制造商提出了创建需要最小存储器使用量但执行良好的产品的问题。这些产品的关键元素是用户接口，典型地，使用户能够输入由该产品处理的文本。

20

语言数据的一个应用在于通过根据用户输入的单词的前几个字符预测单词完成来便于文本输入。给出从语言数据中检索到的一组预测，用户可以选择一个预测，从而不必输入单词中剩余的字符。

25

当包括在移动设备中时，用户输入的预测尤为有用，因为这种设备典型地具有包括键盘在内的尺寸上受到限制的输入设备。输入预测使在这种设备上输入单词所需的击键次数最小。

30

当利用简化键盘输入文本时，输入预测也是有用的。简化键盘具有比能够输入的字符更少的键，因此击键组合是不明确的。使用针对输入预测的语言数据的系统允许用户容易地解决这种不明确性。语言数据还可以用于消除使用简化键盘而输入的单个击键的歧义。

用于存储用于文本输入和处理的语言数据的现有解决方案典型

地依赖于散列表、树、语言数据库或无格式单词列表。由这些语言数据格式所覆盖的单词数受到已经存储了的单词的限制。

用在现有文本输入预测系统中的语言数据典型地来源于文本或言语等语言体，被称为语料（corpus）。语料具有以下用途：分析语言以建立其特征；分析人类在特定的环境下使用语言的行为；训练系统使其行为适应于特定的语言环境；经验验证与语言有关的理论；或者提供针对语言工程技术或应用的测试集合以建立其实际工作的优点。存在几亿个单词的国家语料，还存在针对特定目的而构建的语料。特定用途语料的示例是由汽车司机对识别口头命令的语音操作控制系统的仿真的谈话记录构成的语料。国家语料的示例是英语。

发明内容

提出了一种创建紧凑语言数据的系统。所述系统包括语料和语言数据分析器。所述语言数据分析器计算语料中单词的出现频率。所述语言数据分析器还将单词中每个惟一的字符映射到字符索引，并以将所述字符映射到其的字符索引替代所述单词中的每个字符。所述语言数据分析器还将出现在单词中的字符序列映射到置换索引，并以将所述字符序列映射到其的置换索引替代所述单词中的每个字符序列。所述语言数据分析器还将单词分成组，每个组包含以公共前缀开头的单词，并将每个前缀映射到针对以所述前缀开头的单词组的位置信息。所述紧凑语言数据包括所述惟一字符、所述字符索引、所述置换索引、所述位置信息、所述单词组和所述单词的频率。

还提出了一种针对多个单词的紧凑语言数据结构。将单词组织成组，每个组包含具有公共前缀的单词。所述紧凑语言数据结构包括由所述单词中的每个惟一字符构成的字母表、用于将所述字母表中的每个字符映射到字符索引的字符映射表、用于将来自所述单词的字符序列映射到置换索引的置换表、以及用于存储所述单词的多个单词定义表。每个单词定义表存储包括在所述组中之一个的每个单词。所述紧凑语言数据结构还包括用于定位所述单词定义表的偏移量表。对于每个公共前缀，所述偏移量表包含存储了以所述公共前缀开头的单词的单

词定义表的位置。通过以由所述字符映射表将所述字符映射到其的所述字符索引代替所述单词中的每个字符，以及通过以由所述置换表将所述字符序列映射到其的置换索引代替出现在所述单词中的、来自所述置换表的每个字符序列，对所述单词定义表中的每个单词进行编码。

5 去除每个单词定义表中的单词的所述公共前缀。

还提出了一种创建紧凑语言数据的方法。所述方法以创建包括在语料中最经常出现的多个单词的单词列表的步骤开始。所述方法继续按照字母表顺序对所述单词列表中的所述单词进行排序的步骤。所述方法继续创建字符映射表、以便通过以包含在字符映射表中的关联字符索引替代所述单词中的字符、对所述单词列表中的单词进行编码的步骤。所述方法继续将所述单词列表中的单词分组的步骤，其中每个组中的单词具有公共前缀。所述方法继续创建置换表、以便通过以由所述置换表将其映射到所述字符序列的替换索引替代所述组中的所述单词中的字符序列、对所述组中的单词进行编码的步骤。所述方法继续利用所述字符映射表和所述置换表、将所述组中的所述单词编码为字节序列的步骤。所述方法继续创建单词定义表并将所述编码单词存储在所述单词定义表中的步骤。所述方法继续创建用于定位编码单词组的偏移量表的步骤。所述方法以存储所述字符映射表、所述置换表、所述单词定义表和所述偏移量表的步骤结束。

20

附图说明

图 1 是其中将语言数据用于文本输入预测的系统的方框图；

图 2 是创建紧凑语言数据的系统的方框图；

图 3 是示出了过滤源文件的方法的流程图；

25 图 4 是示出了单词频率计算方法的流程图；

图 5 是示出了创建紧凑语言数据的方法的流程图；

图 6 是紧凑语言数据的格式的方框图；

图 7 是复杂单词定义表的方框图；

图 8 是示出了频率修改方法的流程图；以及

30 图 9 是示出了词形变化分析方法的流程图。

具体实施方式

提出了一种基于具有统计字符置换的单词前缀索引的、创建和使用紧凑语言数据的系统和方法。所述系统通过其存储语言数据的方法
5 需要最小的存储器使用率，并提供了对以特定前缀开头的单词及其关联频率的非常快速的存取。

图 1 是其中将语言数据用于文本输入预测的系统的方框图。系统包括语言数据 100、文本输入逻辑单元 102 和用户接口 103。可以在任何需要文本输入的计算设备上实现所述系统，尤其适用于具有较慢的
10 CPU 和重要的 RAM 和 ROM 限制的嵌入式设备，如移动通信设备等。

用户接口 103 包括文本输入设备 104，允许用户将文本输入系统。文本输入设备 104 是实现了文本输入的任何设备，如 QWERTY、AZERTY 或德佛札克键盘或简化键盘等。用户接口 103 还包括文本输出设备 106，向用户显示文本。文本输出设备 106 可以是显示在移动设备或计
15 算机的屏幕上的图形组件。

语言数据 100 基于具有统计字符置换的单词前缀索引，并将在下面进行详细的描述。

文本输入逻辑单元 102，例如，可以通过由包含在移动设备中的计算机处理器执行的计算机指令来实现。

20 文本输入逻辑单元 102 接收用户利用文本输入设备 104 输入的文本。然后，文本输入逻辑单元 102 使用文本输出设备 106 向用户显示对用户已经开始输入的单词的预测。预测是以用户作为文本输入的前缀开始的最可能的完整单词，并由文本输入逻辑单元 102 从语言数据 100 中进行检索。然后，用户可以使用文本输入设备 104 选择一个预
25 测。

在文本输入设备 104 是简化键盘的情况下，文本输入逻辑单元 102 还消除从简化键盘接收到的单独击键的歧义，根据语言数据 100 中的单词，向用户显示最可能的字符。

图 2 是创建紧凑语言数据的系统的方框图。语言数据分析器 202
30 通过分析自然语言的语料 200，如英语或法语等，创建语言数据 204，

稍后将对其进行描述。语言数据分析器 202 计算语料 200 中单词的出现频率，将单词中的每个惟一字符映射到字符索引，并以将字符映射到其的字符索引替代单词中的字符，将出现在单词中的字符序列映射到置换索引，并以将字符序列映射到其的置换索引替代单词中的字符序列，将单词分成组，每个组包含以公共前缀开头的单词，并将每个前缀映射到针对以所述前缀开头的单词组的位置信息。

语言数据分析器 202 对语料 200 的分析包括计算出现在语料 200 中的多个惟一单词的绝对频率。在图 3 和图 4 中描述了用于计算频率和创建单词列表的方法。一旦已经从语料 200 中得出单词列表，则将单词列表用于创建语言数据 204。语言数据 204 包括惟一字符、字符索引、置换索引、位置信息、单词组和单词频率。在图 5 中描述了用于创建语言数据 204 的方法。在图 6 中示出了由语言数据分析器 202 产生的语言数据 204。

可选地，可以通过将此组分为不同的文件并将自定义权重分配给此文件，修改在语料 200 中发现的特定单词组的绝对频率。此组可以由领域专用的单词构成，如地名或医学术语，而且根据用户的偏好，必须包含在所得到的单词列表中。结果，将利用分配给此组的权重，对此单词组的频率的绝对值进行修改，从而此单词组将具有不同于其否则将具有的频率的频率。

图 3 是示出了过滤源文件的方法的流程图。源文件包含包括了语料的文本。所述过滤方法的第一步在于计算语料中单词的频率。

所述方法从读取源文件的内容的步骤 300 开始。在读取源文件之后，所述方法继续步骤 302，根据可以存储在属性文件中的、与用户偏好相一致的文件，执行文本的置换。用户偏好规定了应用于文本的常规表述，以便置换无效或不想要的字符。例如，用户可能不想要包括在单词列表中的街道名称，或者意大利用户可能想要以 “è” 代替其后为非字母的 “e”，或者当可以预期最后一句只包含作者名时，用户可能想要跳过文本的最后一句。

然后，所述方法继续步骤 304，获得与由源文件的文件扩展名所指示的类型相对应的过滤器。例如，如果文件扩展名是 “.xml”，则假

设该文件包含可扩展标记语言（XML）文档，所以获得 XML 过滤器。类似地，如果文件扩展名为“.html”，则获得超文本标记语言（HTML）过滤器，以及如果文件扩展名为“.txt”，则获得文本过滤器。可以将其他文件扩展名映射到额外的过滤器。

- 5 然后，在步骤 306 应用在步骤 304 获得的过滤器，以便去除并非语料的一部分而只是格式定义的一部分的单词。例如，XML 过滤器将标记标签从自文件中读取出的文本中去除。

所述方法继续步骤 308，从自步骤 306 得到的数据中提取出单词，并在步骤 310，将所提取出的单词写入已过滤单词文件。

- 10 如果在步骤 312 确定存在更多的源文件要进行过滤，则所述方法继续步骤 300。否则，所述方法在步骤 314 结束。当所述方法结束时，已经对包括语料的所有源文件进行了过滤。

- 图 4 是示出了单词频率计算方法的流程图。该方法利用由图 3 所示的方法产生的已过滤单词文件。将来自已过滤单词文件的单词加载到单词树中。单词树是使用最小的存储器存储多个惟一单词及它们频率的有效结构。对树进行组织，从而使已过滤单词文件中频繁出现的单词位于树的内节点中，而较为不频繁出现的单词位于树的叶节点中。树的每个节点包含一个惟一单词和该单词的绝对频率。按照确保树保持平衡的方式，添加和从树中删除单词。
- 15

- 20 所述方法以读取已过滤单词文件的步骤 400 开始。所述方法继续步骤 402，从已过滤单词文件中读取单词，并且如果该单词并未存在于单词树中，则将其添加到单词树中。递增与树中的该单词相关联的频率。

- 所述方法在步骤 404 继续，确定树中的节点数是否已经超过了预定的限制，可以在属性文件中规定所述限制。如果单词树的大小未超过限制，则所述方法在步骤 408 继续。否则，所述方法在步骤 406 继续。
- 25

- 在步骤 406，缩减单词树，从而不再超过大小限制。通过从树中删除位于叶节点的、最不常使用的单词来缩减树。然后，所述方法在步骤 408 继续。
- 30

步骤 408 确定已过滤单词文件中是否还存在已过滤单词。如果存在，则所述方法在步骤 402 继续。如果不存在已过滤单词，则所述方法在步骤 410 继续。

步骤 410 确定是否存在任何剩余已过滤单词文件需要处理。如果存在，则所述方法在步骤 400 继续。否则，所述方法在步骤 412 继续。

在步骤 412，将存储了已经被添加到单词树中的单词及其频率的单词列表输出到输出文件。

图 4 所示的方法允许单独的计算机处理非常大的语料。所得到的单词列表最多包含预定限制数量的、语料中最常出现的单词以及与所述单词相关联的绝对频率。

图 5 是示出了创建紧凑语言数据的方法的流程图。所述方法利用包含单词频率信息的单词列表，产生紧凑语言数据，并包括单词前缀索引和统计字符置换。

所述方法在步骤 500 开始，从由如图 4 所示的方法等单词频率计算方法产生的输出文件中读取单词列表。然后，按照字母顺序排序单词列表中的单词。

所述方法继续步骤 501，对单词列表中的绝对频率进行标准化。以相对频率替代每个绝对频率。通过应用可以由用户规定的函数，将绝对频率映射到相对频率。可能的函数包括抛物线、高斯、双曲线或线性分布。

所述方法继续步骤 502，创建字符映射表。字符映射表用于在所述方法的随后步骤中对单词进行编码。当执行编码时，以字符映射表中这些字符的字符索引替代原始单词中的字符。由于字母语言的字母表大小比 256 小得多，单一的字节足以存储统一代码字符数据。例如，如果其位于字符映射表中的索引 10，则统一代码字符 0x3600 可以表示为 10。字符在字符映射表中的位置并不重要，并基于该字符出现在给定单词列表中的次序。

所述方法继续步骤 504，将单词列表中的单词分组。每组中的单词具有给定长度的公共前缀，并按照频率进行排序。最初按照两个字符长度的前缀对单词进行分组。如果存在 256 个以相同的两字符前缀

开头的单词，则将以更长的前缀进行额外的分组。例如，如果单词列表包含 520 个具有前缀“co”的单词，则将此组分为具有前缀“com”、“con”等的组。

所述方法继续步骤 506，产生每组单词的频率集合。为了减少存储频率信息所需的空间量，只以全精度保留每组中单词的最大频率。将每个其他单词的频率保留为其组中单词的最大频率的百分数。此技术引起了一些精度损失，但这对于文本输入预测的目的而言是可接受的，并导致了频率信息的较小存储需求。

所述方法继续步骤 508。为了减少存储单词列表中的单词所需的数据量，以置换索引替代单词中最常出现的字符序列。N 元置换，作为 n 个字符数序列，使其能够以单一字符表示多个字符。将此信息存储在置换表中。对置换表进行索引，从而将每个 n 元映射到置换索引。于是，可以通过当每次在单词中出现 n 元时，以置换表中的置换索引替代每个 n 元，使单词紧凑。

所述方法继续步骤 510，利用字符映射表和置换表，将单词组编码成字节序列，如上所述。从单词本身中去除用于将单词收集成组的前缀。结果，以给定其前缀的字节序列表示每个单词，其包括了找出原始单词所需的全部数据。

所述方法继续步骤 511，创建单词定义表。单词定义表存储在步骤 506 计算出的频率集合和在 510 产生的已编码单词。

所述方法继续步骤 512，创建偏移量表。偏移量表包含表示单词组的字节序列。此表实现了对表示特定单词组的字节序列的开始标识。偏移量表用于定位包括以公共前缀开头的特定组的已编码单词的字节序列。

所述方法以步骤 514 结束。在此步骤，已经将通过所述方法得到的语言数据存储在已经创建的表格中。将包括字符映射表、置换表、偏移量表和单词定义表的数据表格存储在输出文件中。

在步骤 514，可以可选地存储在创建紧凑语言数据的方法期间所收集的统计数据。所述统计数据包括存储在置换表中的 n 元在语言数据中的单词中出现的频率、语言数据中的单词数、单词列表和由其产

生单词列表的语料、以及语言数据、单词列表和语料中单词数之间的比例。

图 6 是紧凑语言数据的格式的方框图。该数据格式的主要目的在于，在最小化存储器使用和创建与解译数据所需的计算机指令数的同时，保持语言数据的解译的简单性。该格式的语言数据由语言数据分析器 202（图 2）产生，并且是如图 5 所示的方法的输出。

该格式允许对语言数据进行存储，具有或不具有单词频率信息。当语言数据包括频率信息时，可以实现稍后将进行描述的学习能力，并且可以将数据用于预测以简化键盘输入的输入。如果不包括频率信息，则不包括小于三个字符长的单词，由于其对预测用户输入没有用。

该格式定义了包含跟随有多个表格的头部 602 的计算机文件。

头部 602 包含包括了作为识别文件格式的幻数的签名的头部 602 还包含规定了文件中所包含的语言数据的版本和优先级的信息。优先级信息用于在文字输入逻辑单元使用包含语言数据的多个文件时将相对重要性分配给语言数据。头部 602 还指示文件是否包括频率信息。

索引表 604 跟随头部 602。索引表 604 包含文件中对以下所定义的剩余表格的索引，并允许添加额外的表格。利用在索引表 604 中的表格项中找到的索引信息，定位表格。

名称表 606 跟随索引表 604。名称表 606 包含识别单词列表的名称。

字符映射表 608 跟随名称表 606。字符映射表 608 包含用于此单词列表的字母表，并将字母表中的每个字符映射到字符索引。字母表由用在单词列表中的单词中的每个惟一字符组成。

置换表 610 跟随字符映射表 608。置换表 610 包含二元置换表，其后是针对所定义的每个高阶 n 元组的表格，如三元、四元等。通过置换表 610，将每个 n 元映射到置换索引。

偏移量表 612 跟随置换表 610。此表格用于根据要定位的单词定义表中的单词的公共前缀，定位单词定义表，稍后将对其进行描述。对于字母表中两个字符的每一组合，所述表格包括包含了以该字符组

合开头的单词的单词定义表在文件中的偏移量。对于空组，偏移量等于下一非空偏移量。每个偏移量还规定了定位在文件中的偏移量处的单词定义表是简单的或复杂的，如下所述。

假设两字符序列，通过以下公式定义位于偏移量表中的索引处的
5 偏移量： $((\text{字母表中第一字符的位置} \times \text{字母表中的字符数}) + \text{字母表中第二字符的位置})$ 。例如，如果字母表是英语，则字母表的大小为26，所以偏移量表中“ab”的索引为 $((0 \times 26) + 1)$ ，等于1。因此，偏移量表 612 的大小基于字母表的长度。

在语言数据中可以可选地包括并未示出的词形变化表。词形变化
10 表存储可能会被用在单词定义中的后缀。在图 9 中示出了词形变化分析方法。

语言数据还包含单词定义表 614。单词定义表存储来自单一单词组的单词和与所述单词相关联的频率，并且可以是简单或复杂的。简单表用于定义只以两字符前缀进行了分组的单词。复杂表用于定义以
15 更大长度的前缀进行了分组的单词。

利用字符映射表 608 和置换表 610，对定义表 614 中的单词进行编码。以来自字符映射表 608 的相应字符索引替代单词中的字符，以及以置换表 610 中的相应置换索引单词中处于置换表 610 中的 n 元。由于偏移量表 612 将字母表中的每个 2 元前缀惟一地映射到定义了以
20 该前缀开头的单词的文件中的位置，不需要保留前缀，并因而将其从单词定义中去除。

可以可选地以额外的特定字符标记大写单词。将所述特定字符存储在字符映射表 608 中，以并未用在单词列表中的单词语言中的额外字符扩展字母表。

25 简单单词定义表包含已编码单词组及与所述单词相关联的频率。通过应用对频率进行转换从而使其数值在预定范围内的标准函数，对所述频率进行标准化。在表格中，只以全精度存储组中单词的最大频率。按照频率存储已编码单词。但是，如果应用学习能力，如下所述，则原始排序不再有效，并可能需要对已编码单词进行重新排序。

30 如本领域普通技术人员所清楚的那样，在计算机系统中，以比特

序列表示字符。按照具有置位最高有效位的字符分离单词定义表 614 中的单词。如果字符使最高有效位置位 (MSB set)，则其是单词中的最后一个字符。然后，为了确定字符值，将该字符作为其最高有效位并未被置位，因而最高有效位并不影响字符值。

- 5 图 7 是复杂单词定义表的方框图。复杂单词定义表是递归的，其中包含本地单词定义表 708，每一个均为如上所述的简单或复杂单词定义表。

本地单词定义表 708 定义按照更高阶的 n 元前缀进行了分组的单词。每个本地单词定义表 708 存储由单词定义表存储的、具有共同前缀的单词，其中每个本地单词定义表 708 中的单词的公共前缀比单词定义表中的单词的公共前缀长。去除本地单词定义表 708 中的单词的公共前缀。

例如，如果单词组包括以前缀“co”开头的单词，并且存在多于 256 个以该前缀开头的单词，则针对“co”前缀单词的复杂单词定义表包含定义了以“com”、“con”等开头的单词的本地单词定义表 708。针对“com”前缀单词的表格可能是进一步包含针对以“comm”和“comp”开头的单词的本地单词定义表 708 的复杂单词定义表，而针对“con”前缀单词的表格可以是只包含以“con”开头的单词的简单单词定义表。

除了包含本地单词定义表 708 以外，每个单词定义表还包括本地偏移量表 706，用于定位本地单词定义表 708 中的每一个。每个偏移量还表示通过偏移量所引用的表格是复杂或简单单词定义表。

每个复杂单词定义表还包括本地字符映射表 704。此表格在功能上与字符映射表 608 (图 6) 相同，不同之处在于，其只包含被包括在位于本地单词定义表 708 中的单词中的字符。本地字符映射表 704 将本地单词定义表 708 中的单词中的每个字符映射到本地字符索引。通过以本地字符索引替代单词中的字符，对简单本地单词定义表中的单词进行编码。

复杂单词定义表还包含热词表 700 和例外表 702。热词是与复杂单词定义表中所包含的组中的最高频率相关联的单词。热词表 700 包含位于作为简单单词定义表的本地单词定义表 708 中的热词的索引。

例外表 702 存储位于作为复杂单词定义表的本地单词定义表 708 中的热词。可以利用热词表 700 和例外表 702 快速地检索热词，来代替执行对本地单词定义表 708 的搜索来找出热词。

上述语言数据的格式实现了利用最少量的存储器，对单词预测的非常快速的确定。当用户利用如 QWERTY 键盘等将字符映射到唯一键或键组合的文本输入设备输入单词前缀时，文本输入逻辑单元检索语言数据中、以该前缀开头、具有最高频率的单词，并向用户显示预测。当用户还是利用简化键盘键入单词时，单词前缀是不明确的，由于缩减键盘上的每个键被映射到多个字符。在这种情况下，文本输入逻辑单元从语言数据中检索以与由用户输入的前缀相对应的字符的任意组合开头的预测。

该格式还允许对单词频率的简易修改，以与单独用户的文本输入习惯相符合。文本输入逻辑单元学习通过用户在显示有单词预测备选时所做出的输入选择而进行确认的用户习惯，并将其存储在包括下述内容的表格中。

学习能力包括修改单词的频率信息和将单词添加到语言数据中。这些操作均基于与将单词及相应频率信息添加到单词列表中类似的处理。学习单词列表包括针对频率修改和针对新单词添加的表格。

图 8 是示出了频率修改方法的流程图。所述方法在作为在收集任何学习数据之前如上所述组成的语言数据的基本语言数据通常具有正确的频率信息的假设下进行。因此，所述方法允许对频率信息的有限修改。

所述方法以将用户选定单词 800 添加到学习单词列表中的步骤 802 开始。用户选定单词 800 是用户从所提供的、以用户输入的单词前缀开头的预测单词的列表中的选择的单词。用户利用文本输入设备选择预测。将选定的单词添加到学习单词列表中。

所述方法继续步骤 804，获得显示给用户的预测列表中的单词中、具有最大频率的单词。可以已经从基本语言数据中的单词定义表中或者从学习单词列表中获得了预测列表中的单词及其相应的频率。如果在步骤 806 确定从单词定义表获得了具有最大频率的单词，则所述方

法在步骤 808 继续，并向用户选定单词 800 分配等于最大频率加一的频率。

- 如果在步骤 806 确定未从单词定义表中获得具有最大频率的单词，而是从学习单词列表中获得，则所述方法在步骤 810 继续，并向
- 5 用户选定单词 800 分配等于最大频率的频率。然后，所述方法以从学习单词列表中删除在步骤 804 所获得的具有最大频率的单词的步骤 812 结束。

以下段落是图 8 所示的方法的示例。每个示例均假设用户输入三字符前缀。

- 10 给出三字符前缀“beg”，并且预测“began”，频率为 3024；“begin”，频率为 2950；“beginning”，频率为 2880，以及“begins”，频率为 2000，其中从基本语言数据中的单词定义表中获得上述全部单词，如果用户选择单词“begin”，则将单词“begin”以频率 3025 添加到学习单词列表中。

- 15 给出相同的三字符前缀“beg”，并且预测“begin”，频率为 3025；“began”，频率为 3024；“beginning”，频率为 2880；以及“begins”，频率为 2000，其中从学习单词列表中获得“begin”，如果用户选择“began”，则将单词“began”以频率 3025 添加到学习单词列表中，并从学习单词列表中删除单词“begin”。

- 20 以下是图 8 所示方法的示例，其中利用简化键盘输入三字符前缀。简化键盘包括用于输入“a”、“b”或“c”的键、用于输入“n”或“o”的键和用于输入“w”、“x”或“y”的键。在此示例中，假设用户通过按下“a/b/c”键、然后按下“n/o”键、最后按下“w/x/y”键输入三字符前缀。给出预测“any”，频率为 3024；“boy”，频率为 2950；“box”，
- 25 频率为 2880；“bow”，频率为 2000；“cow”，频率为 1890；以及“cox”，频率为 1002，其中从基本语言数据中的单词定义表中获得上述全部单词，如果用户选择“boy”，则将单词“boy”以频率 3025 添加到学习单词列表中。

- 学习单词列表包括：更新频率表，其包含具有已更新频率的单词；
- 30 以及新单词表，其包含新单词。这两个表格均包括利用与基本语言数

据所使用的相同的字符映射 608 (图 6) 和置换表 610 (图 6)、与基本语言数据中一样进行了编码的单词。每个学习单词列表还包括对表格中单词的开始的索引、与表格中的单词相关联的频率信息和规定了单词的字母排序次序的排序索引。每个表格还包括指示该表格是否包含
5 已更新频率或新单词的标记。学习单词列表一个接一个顺序相连,更新频率表排在最前面。

如果学习单词列表达到最大定义长度,则从表格中删除最早的单词,以便为表格中的新表格项留出空间。

通过创建表格更新表的字节序列并同时字节序列写入输出流
10 来执行将单词添加到学习单词列表中和从中删除单词。在完成更新时,重新读取已更新数据。在每次向学习单词列表中添加单词或从中删除单词时,发生写入输出流的处理。

为了向学习单词列表之一中添加单词或从中删除单词,如果其不包含出现在要添加的单词中的字符,则更新字符映射表 608 (图 6) 中的字母表。然后,利用字符映射表 608 (图 6) 和置换表 610 (图 6),
15 对要添加的单词进行编码,并将其插入到新单词表的开头。最后,更新学习单词列表的频率和排序索引。

图 9 是示出了词形变化分析方法的流程图。可选地,创建紧凑语言数据的系统和方法可以包括词形变化分析方法,其中创建具有高于
20 最小规定频率的频率的单词的列表和词形变化表。根据统计后缀分析,创建词形变化表,并封装语料的语言中用于单词创建的语言规则。词形变化表使其能够利用存储在词形变化表中的基本单词形式产生多于一个单词,确保语言数据覆盖更多的单词,同时基本单词列表保持紧凑。词形变化表可以可选地包括在如图 6 所示的语言数据格式中。

25 所述方法以根据单词的绝对频率找出单词列表中最常出现的单词的配置数的步骤 900 开始。

所述方法继续步骤 901,找出频繁出现的单词的后缀。所述后缀查找步骤基于递减长度的后缀的迭代搜索,以六字符长的后缀开始,以两字符长的后缀结束。这些后缀并不总是与给定语言的语法中的现有
30 有对应部分相匹配,而是后缀查找基于单词列表中后缀的出现次数。

所述方法继续步骤 902，以在前一步骤中找出的后缀更新词形变化表。第一次执行步骤 902 时，在对词形变化表进行更新之前，创建词形变化表。

在步骤 903，如果语言数据的大小小于配置最大大小，则所述方法在步骤 901 继续。否则，所述方法以创建单词列表中、不具有包含在词形变化表中的后缀的单词的列表的步骤 904 结束。

然后，可以如参照图 5 所述地那样，对词形变化表和不具有后缀的单词的列表进行编码。当使用词形变化分析方法时，如图 6 所示的、所得到的紧凑语言数据还包括词形变化表。于是，单词定义表 614 (图 6) 中的单词并不包括备包括在词形变化表中的后缀，而是包含对词形变化表中的后缀的引用。利用针对所存储的每个后缀的词形变化表所节省的空间为后缀出现的次数乘以后缀的长度。

上面的描述涉及本发明的一个示例。多种变体对本领域的普通技术人员而言都是显而易见的，而这些变体都在本申请的范围之内。

例如，尽管在大多数示例中所使用的语言是英语，所提出的系统和方法创建针对任意字母语言的紧凑语言数据。

此外，可以在个人计算机、PDA、蜂窝式电话、双向寻呼机、任何类型的可佩戴计算机、打印机、机顶盒和任何其他允许文字输入和显示的设备上，作为软件、固件或硬件或其组合，来实现创建和使用紧凑语言数据的系统和方法。

而且，如图 3、4、5、8 和 9 所示的方法可以包含比所示步骤更少、更多或不同的步骤。例如，尽管所述方法描述利用计算机文件存储所述方法的最终和中间结果，这些结果也可以被存储在如 RAM 或闪存存储器模块等计算机存储器中。

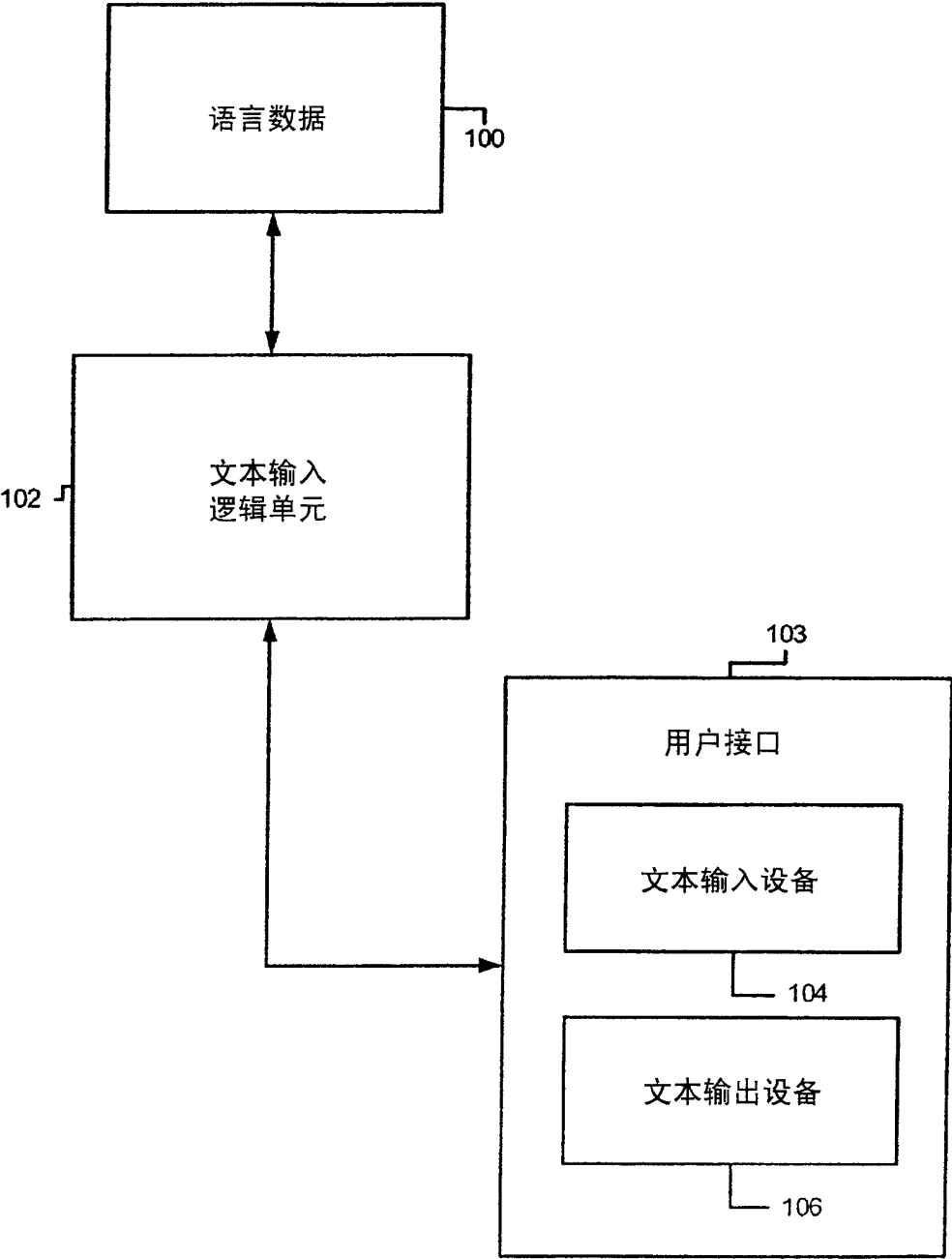


图 1

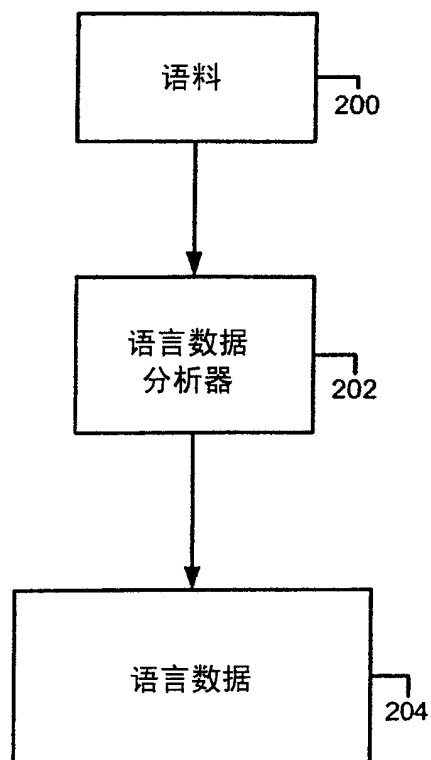


图 2

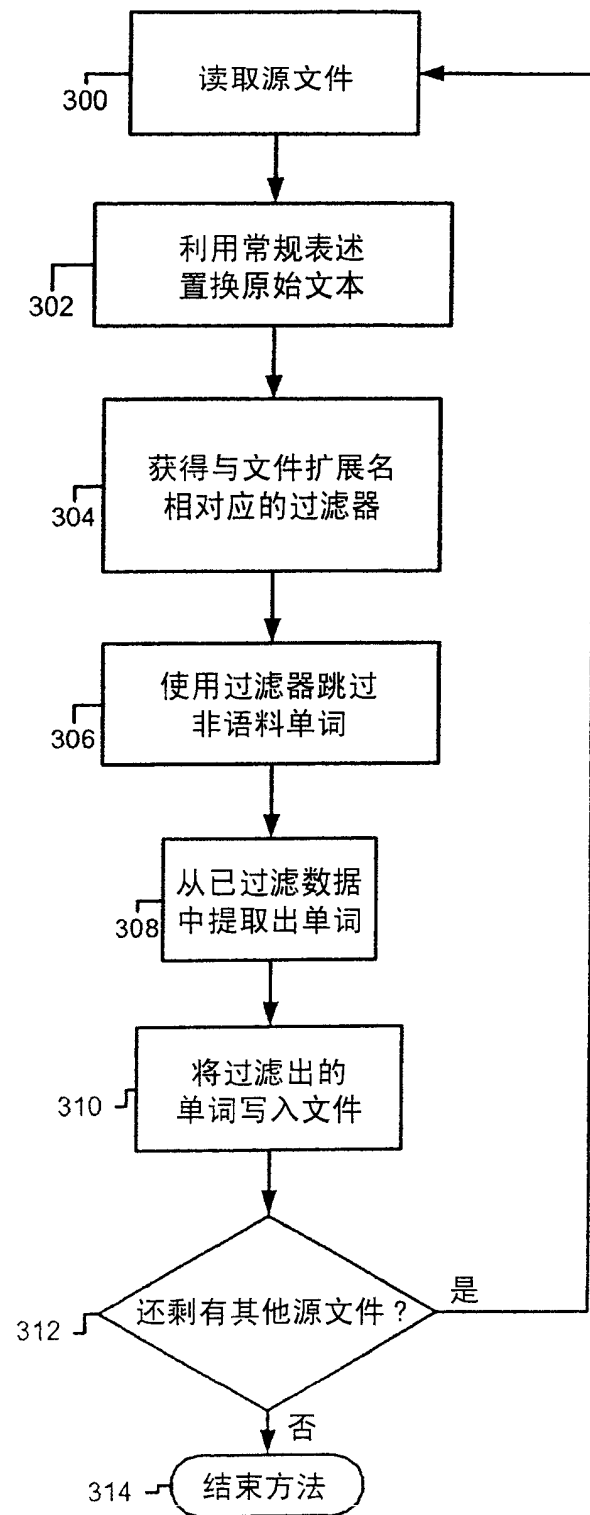


图 3

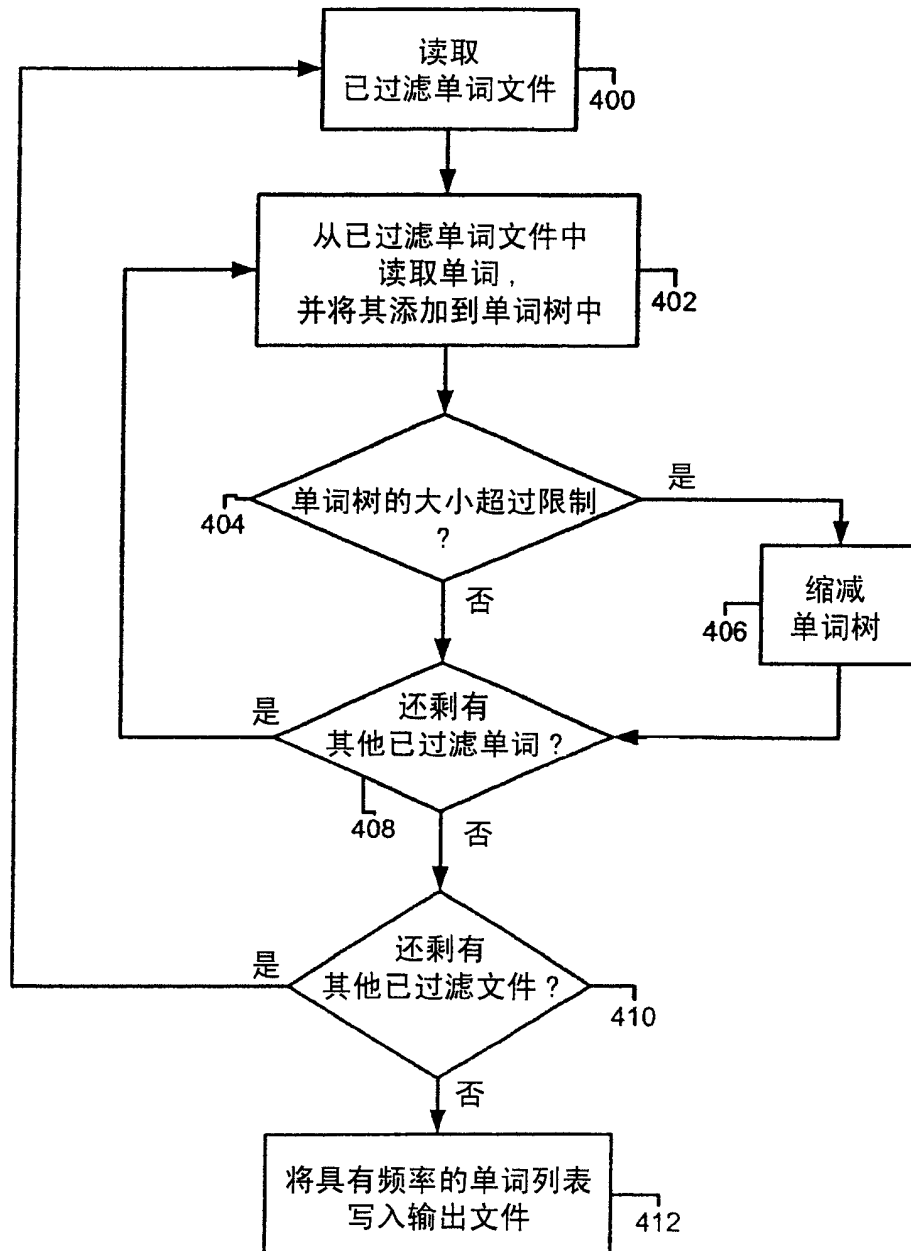


图 4

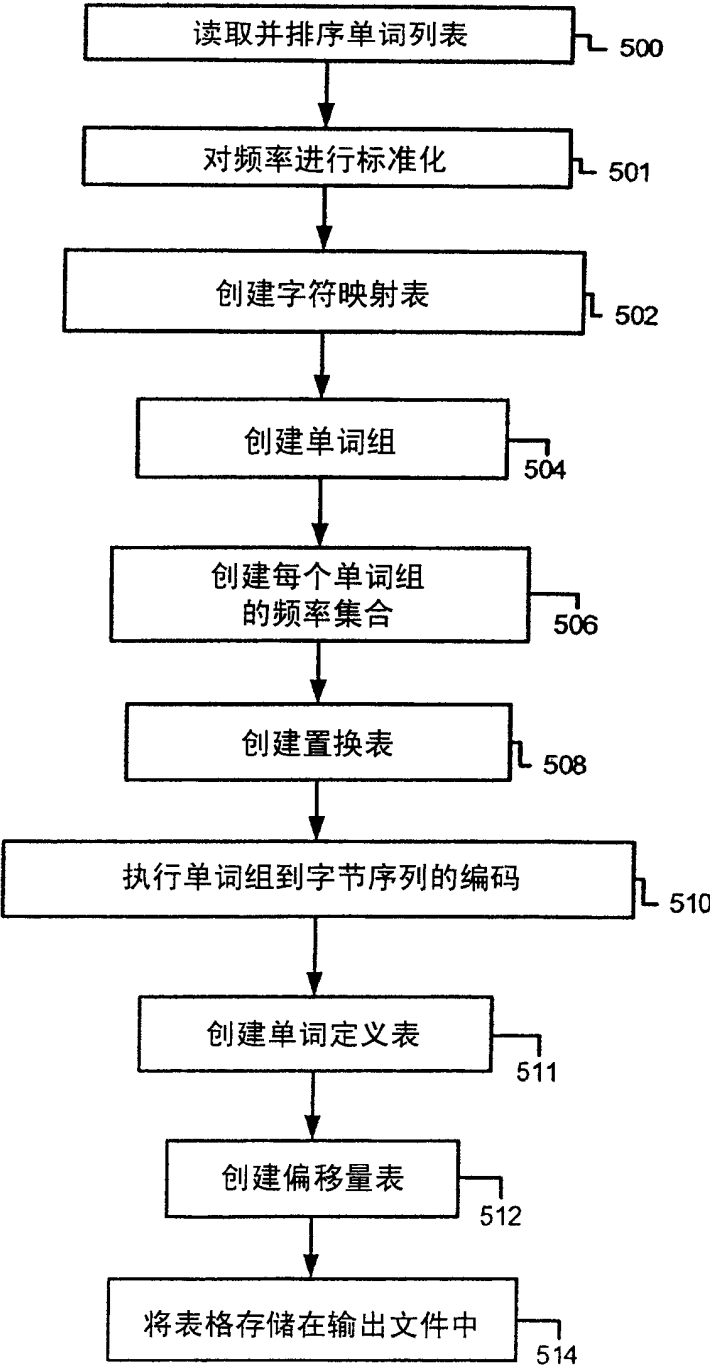


图 5

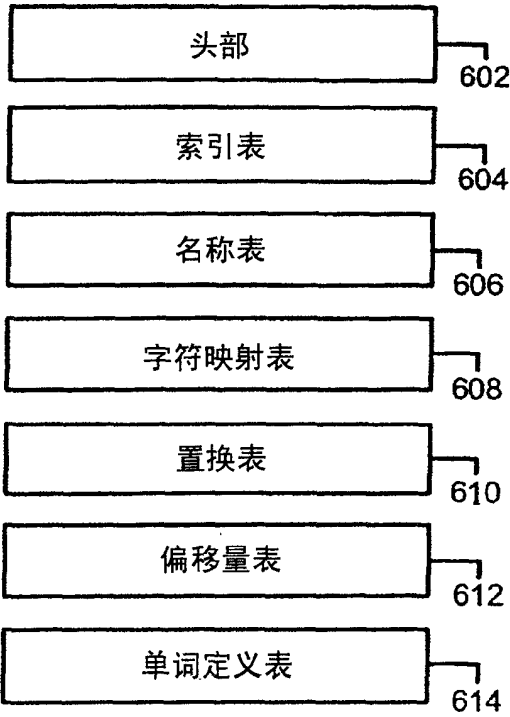


图 6

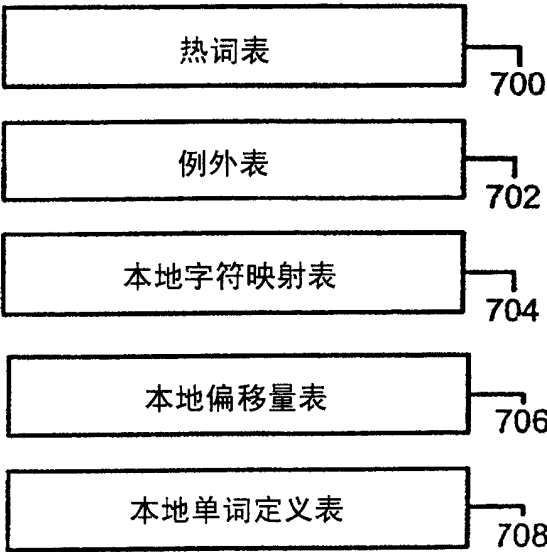


图 7

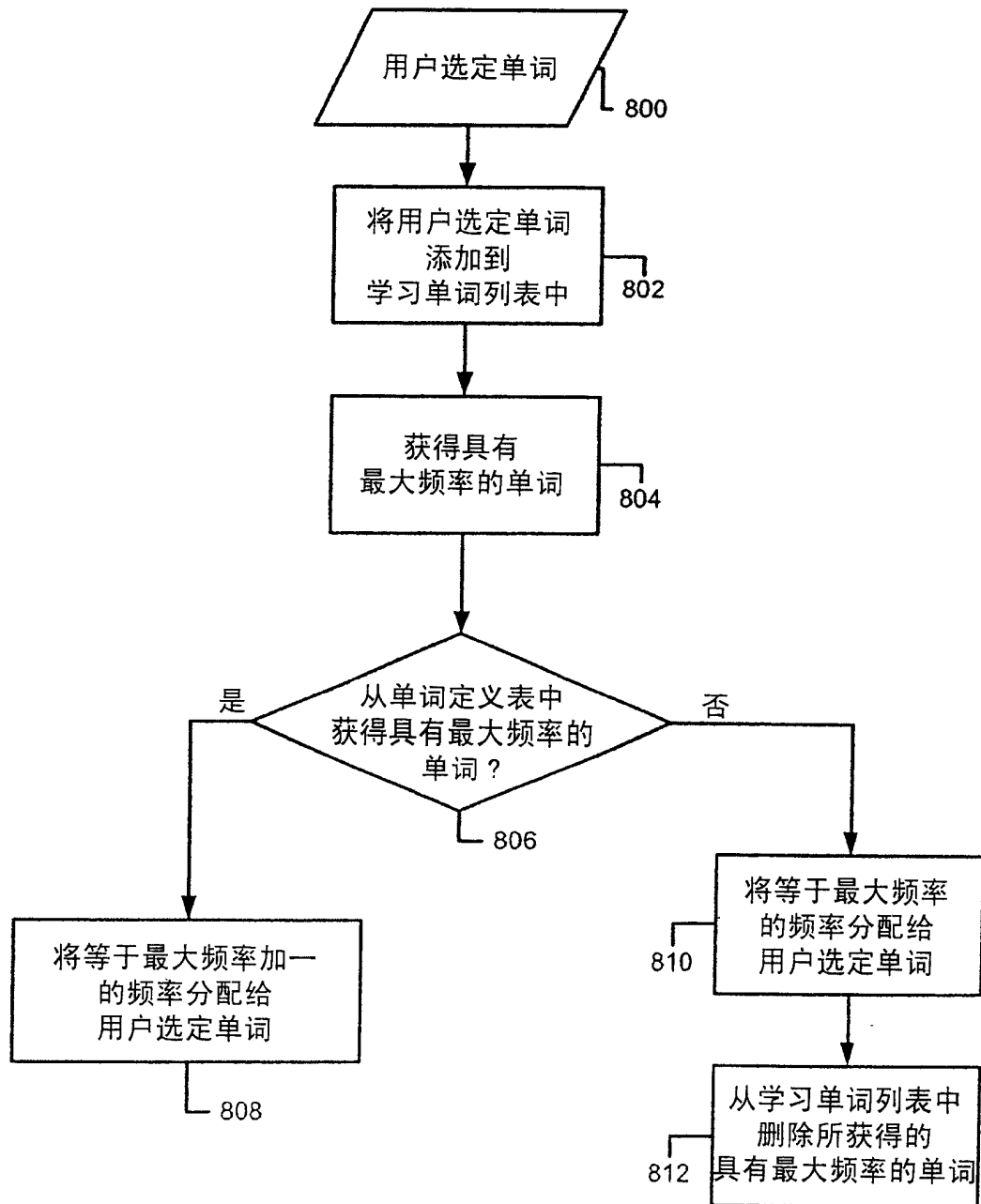


图 8

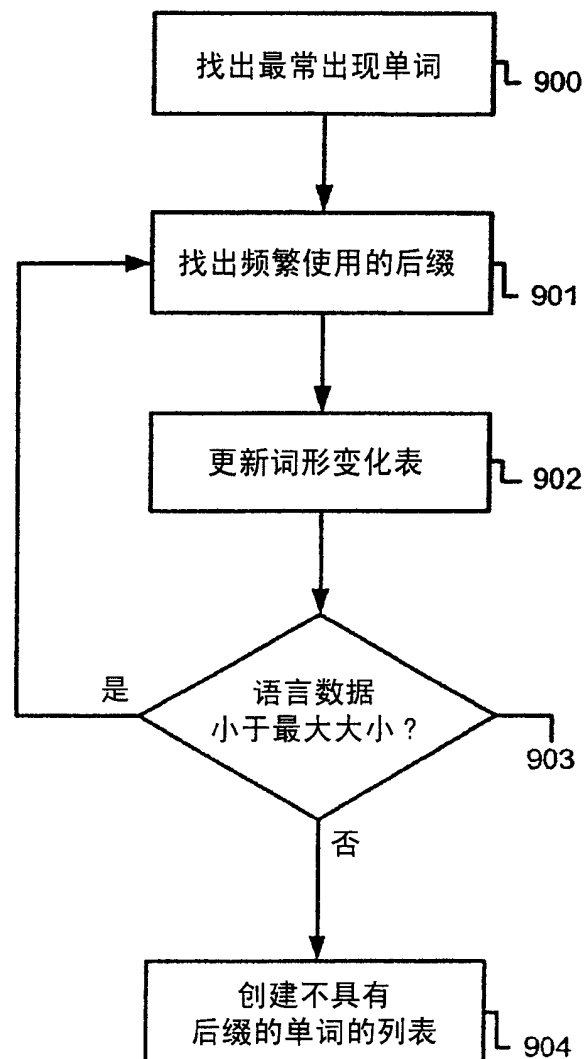


图 9