

说明书

发明名称：五笔输入系统及方法

技术领域

- [1] 本发明涉及输入法，尤其是涉及一种五笔输入系统及方法。

背景技术

- [2] 五笔字型输入法，简称五笔，是王永民教授发明的一种依照汉字字形进行编码的汉字输入方法，是目前中国以及一些东南亚的国家，最常用的一种汉字输入法之一。

- [3] 五笔的基本原理：汉字都是由笔划或部首组成的。为了输入这些汉字，我们把汉字拆成一些最常用的基本单位，叫做字根，字根可以是汉字的偏旁部首，也可以是部首的一部分，甚至是笔划。取出这些字根后，把它们按一定的规律分类；再把这些字根依据科学原理分配在键盘上，作为输入汉字的基本单位。五笔基本字根有130种，加上一些基本字根的变型，共有200个左右，这些字根分布在除Z之外的25个键上。当要输入汉字时，我们就按照汉字的书写顺序依次按键盘上与字根对应的键，组成一个编码；系统根据输入字根组成的编码，在五笔输入法的字库中检索出所要的字。

- [4] 五笔输入法以其重码率低而能够快速找到用户预期的词条，在熟悉该输入法的前提下可大大提高输入的速度，不过这需要用户能够熟练地拆解字词，同时一般需要三到四个五笔码才能快速确定所需要的词。在不熟练的情况下，用户只能通过一码或两码来获得大量的候选词条，并通过筛选来找到需要的词条，降低了输入速度。

对发明的公开

技术问题

- [5] 鉴于此，有必要针对传统的五笔输入法在一码或两码输入情况下重码率较高，影响输入速度的问题，提供一种能够提高用户输入速度的五笔输入系统及方法。

技术解决方案

- [6] 一种五笔输入系统，包括：缓存词库，存储一位码和两位码的常用词的词条信息及索引信息；核心词库，存储所有五笔编码的词条信息及索引信息；取词模块，用于当输入一位码或两位码时，根据所述缓存词库中的索引信息从缓存词库中取词并显示；当输入三位码或四位码时，根据所述核心词库中的索引信息从核心词库中取词并显示。
- [7] 优选地，所述缓存词库包括：缓存编码索引区，存储常用词的索引信息；缓存词条存储区，存储常用词的词条信息，所述常用词以五笔编码的前两码作为索引，并按照词频由高到低顺序存储。
- [8] 优选地，所述核心词库包括：核心编码索引区，存储所有五笔编码的词条信息的索引信息；核心词条存储区，存储所有五笔编码的词条信息，所有词条以其五笔编码的前三码作为索引按序存储，对前三码相同的词条按照词频由高到低顺序存储。
- [9] 优选地，所述取词模块包括：索引计算模块，根据输入的五笔编码得到索引信息；候选词输出模块，根据索引信息获得并显示词条。
- [10] 优选地，还包括判断模块，所述判断模块用于根据输入的一位码或二位码判断缓存词库中是否存在用户预期的词条。
- [11] 此外还提供一种五笔输入方法。
- [12] 一种五笔输入方法，包括如下步骤：接收五笔编码输入；当所输入的五笔编码为一位码或二位码时，向存储了一位码和两位码的常用词的词条信息及索引信息的缓存词库取词；当所输入的五笔编码为三位码或四位码时，向存储了所有五笔编码的词条信息及索引信息的核心词库取词。
- [13] 优选地，向所述缓存词库取词的步骤之后还包括：判断所述缓存词库是否包含用户预期的词条，若缓存词库未包含用户预期的词条，则向所述核心词库取词。
- [14] 优选地，所述向缓存词库取词的步骤具体是：将缓存词库中的词以五笔编码的前两码作为索引，并按照词频由高到低进行存储，将输入的五笔编码转换为索引信息，然后根据索引信息将词条按序读取并显示。
- [15] 优选地，所述向核心词库取词的步骤具体是：将核心词库中的词以五笔编码的

前三码作为索引顺序存储，并对前三码相同的词条按照词频由高到低顺序存储，如果输入的五笔编码为三位码，则将该三位码转换为索引信息，然后根据索引信息获得的词条按照词频由高到低进行排序显示；如果输入的五笔编码为四位码，则将由三码输入获得的词条中第四码与用户输入的第四码不匹配的所有词条过滤，获得与该四码输入相应的所有词条，将所获得的词条按照词频由高到低进行排序显示。

- [16] 优选地，所述向核心词库取词的步骤还包括：如果输入的五笔编码为一位码或二位码，则将该一位码或二位码转换为索引信息，然后根据索引信息将获得的词条按照词条在核心词库中的存储顺序依次读取并显示。

有益效果

- [17] 加入缓存词库后，根据用户输入可优先检索缓存词库，使用户在输入一位码或两位码时，显示常用词条，不用查找大量词条即可增大用户预期词条命中率，提高五笔输入速度。
- [18] 由于优先处理一位码或二位码，向缓存词库取词，因此根据用户输入可优先检索缓存词库，使用户在输入一码或两码时，显示常用词条，不用查找大量词条即可增大用户预期词条命中率，提高五笔输入速度。

附图说明

- [19] 图1为实施例1的五笔输入系统的结构示意图；
- [20] 图2为实施例1的五笔输入方法的流程图；
- [21] 图3为实施例2的五笔输入系统的结构示意图；
- [22] 图4为实施例2的五笔输入方法的流程图。

本发明的实施方式

- [23] 实施例1

- [24] 如图1所示，为本实施例的五笔输入系统结构示意图。该五笔输入系统包括取词模块100、核心词库200以及缓存词库300。核心词库200存储所有五笔编码的词条信息及索引信息，缓存词库300存储一位码和两位码的常用词的词条信息及索引信息。当输入一位码或两位码时，取词模块100根据缓存词库300中的索引信息从缓存词库300中取词；当输入三位码或四位码时，取词模块100根据核心

- 词库200中的索引|信息从核心词库200中取词。
- [25] 取词模块100包括索引|计算模块110和候选词输出模块120。其中索引|计算模块110根据用户输入将五笔编码转换为索引|信息。如将一位码或二位码转换为向缓存词库300取词的索引|信息，将三位码或四位码转换为向核心词库200取词的索引|信息。候选词输出模块120根据索引|信息获得词条并进行显示输出。
- [26] 核心词库200包括核心编码索引|区210和核心词条存储区220。核心编码索引|区210存储所有五笔编码的词条信息的索引|信息；核心词条存储区220存储所有五笔编码的词条信息，所有词条以其五笔编码的前三码作为索引|按序存储，对前三码相同的词条按照词频由高到低顺序存储。
- [27] 缓存词库300包括缓存编码索引|区310和缓存词条存储区320。缓存编码索引|区310存储常用词的索引|信息；缓存词条存储区320存储常用词的词条信息，常用词以五笔编码的前两码作为索引，并按照词频由高到低顺序存储。
- [28] 本实施例中核心编码索引|区210和缓存编码索引|区310都是一段连续的数组区域，数组的每个元素占4个字节，其中记录的是五笔编码所对应词条在核心词条存储区220或缓存词条存储区320中的起始位置。
- [29] 索引|信息即是指数组中存储的词条的起始位置，相应的，核心编码索引|区210中存储的索引|信息即是指核心词条存储区220中词条存储的起始位置；缓存编码索引|区310中存储的索引|信息即是指缓存词条存储区320中词条存储的起始位置。
- [30] 核心词条存储区220和缓存词条存储区320中保存的是具体词条信息，包括词条的五笔编码，Unicode文本，词频和其它一些附加信息。词条的五笔编码用于与用户输入对比确定是否匹配，Unicode文本用于显示词条，词频则可根据统计结果预先定义，也可在用户使用过程中实时更新，表示该词条使用的频率，因此词频较高的词条满足用户预期的可能性非常大。(Unicode是一种文本编码标准，每个字符用两个字节表示，是一种定长的2Byte多文种(语言)字符集编码，也属于现有技术)
- [31] 相应的五笔输入方法，如图2所示包括如下步骤：
- [32] S10：接收五笔编码输入。在键盘上a至y的25个按键上按照五笔输入法既定的

规则分布有字根，根据键盘字母的输入可得到字根所组合成的词条。本实施例的处理方法接收用户输入的一至四个由a至y的任意组合。

[33] S20: 判断五笔编码是几位码输入。如果是一位码或二位码，则转入步骤S30；如果是三位码或四位码，则转入步骤S50。

[34] S30: 向缓存词库300取词并显示。本步骤处理一位码或二位码输入，由于核心词库200包含了大量的词条，在一位码或两位码输入时，重码率较高，因此建立缓存词库300，收录较常用的词条，并且这些常用的词条是以不多于两位码的输入来进行索引的。

[35] 在缓存词库300中，所有词条是按照其前两码作为索引的，所以缓存编码索引区310的索引范围是从“a”至“yy”，因而数组包含 $25 + 25^2 = 650$ 个元素。

[36] 因此可建立一位码或两位码的五笔编码与缓存编码索引区310数组下标的联系。
。strCode代表用户输入的编码，长度从1到4，Index代表转换成的数组下标，则：

[37] $\text{Index} = (\text{StrCode}[0] - 'a') * (25 + 1) + 1;$

[38] If (编码长度 ≥ 2) $\text{Index} += (\text{StrCode}[1] - 'a') + 1。$

[39] 根据上述公式计算得出的结果如下：

[40] 编码：a 下标：1

[41] 编码：aa 下标：2

[42] 编码：ab 下标：3

[43]

[44] 编码：y 下标：625

[45] 编码：ya 下标：626

[46]

[47] 编码：yy 下标：650

[48] 根据上述公式可由五笔编码获得缓存编码索引区310数组下标，进而获得与该五笔编码在缓存词条存储区320中对应的词条的起始位置。由于缓存词条存储区320中的词条是按照两码索引，且以词频排序。

[49] 因此取词模块100从缓存词库300中取词的模式是：

- [50] 当用户输入一位码或二位码时，按照该一位码或二位码对应的数组下标获得词条的起始位置，并按照词条存储的顺序取词并显示。
- [51] 如在缓存词库300中与“aa”对应的词条仅以词频由高到低按序存储了“式”（aa）、“工作”（aawt）“工具”（aahw）、“工程”（aatk）、“工业”（aaog）、“工艺”（aaan）、“工资”（aaug）、“工厂”（aadg）、“工人”（aaww）以及“工”（aaa）十个词条。则在取词时就可以从“式”存储的起始位置开始顺序从缓存词库300中取词。
- [52] 若输入三位码以上，取词模块100不会向缓冲词库300取词。
- [53] 按照五笔用户的输入习惯，一般很少翻到两页以后去找候选词，因此在本实施例中，优选地，在缓存词库300中，每个五笔编码对应的索引至多存储10个词条。因此缓存词库300中至多存储 $650 \times 10 = 6500$ 条词条。
- [54] S50：向核心词库200取词并显示。本步骤处理三位码或四位码输入。当用户输入为三位码或四位码时，词条的重码率就已经很低了，因此可以直接进入核心词库200进行索引。
- [55] 在核心词库200中，所有词条是按照其前三码作为索引的，所以核心编码索引区210的索引范围是从“a”至“yyy”，因而数组包含 $25 + 25^2 + 25^3 = 16275$ 个元素。数组中每个元素的下标都与五笔编码建立一一对应的关系。
- [56] 如可通过如下方法建立五笔编码与核心编码索引区210数组下标联系：
- [57] strCode代表用户输入的编码，长度从1到4，Index代表转换成的数组下标，则：
- [58]
$$\text{Index} = (\text{StrCode}[0] - 'a') * (25^2 + 25 + 1) + 1 ;$$
- [59] If (编码长度 ≥ 2) Index += (StrCode[1] - 'a') * (25 + 1) + 1 ;
- [60] If (编码长度 ≥ 3) Indxe += (strCode[2] - 'a') + 1。
- [61] 根据上述公式计算得出的结果如下：
- [62] 编码： a 下标： 1
- [63] 编码： aa 下标： 2
- [64] 编码： aaa 下标： 3
- [65] 编码： aab 下标： 4

[66] 编码: aac 下标: 5

[67] 编码: aad 下标: 6

[68]

[69] 编码: y 下标: 15625

[70] 编码: ya 下标: 15626

[71]

[72] 编码: yad 下标: 15630

[73]

[74] 编码: yyy 下标: 16275

[75] 上述排序是典型的字典序, 根据上述对应关系可由五笔编码获得核心编码索引区210数组下标, 进而获得与该五笔编码在核心词条存储区220中对应的词条的起始位置。(是现有技术)

[76] 因此取词模块100从核心词库200中取词的模式是:

[77] 当用户输入三码时, 则将前三码相同的词条按照词频由高到低排序, 顺序取出并显示。如输入“fnt”, 若“fntj”对应的“专利”词频为1000、“fnta”对应的“专长”词频为 500、“fnnn”对应的“专书”词频为 200, 则在核心词库200中, “专利”、“专长”以及“专书”依次存储, 取词时依次取出并显示即可。

[78] 当用户输入四位码时, 将由三码输入获得的词条中第四码与用户输入的第四码不匹配的所有词条过滤, 剩余的词条即是与该四码输入相应的所有词条。

[79] 实施例2

[80] 由于五笔输入法本身的重码率较低, 在加入缓存词库300后, 将一位码或二位码输入的重码率也降低到一定程度, 提高了词条的命中率。一般来说, 利用二位码输入获得预期词条的几率很高, 或者说需要到核心词库中取词的几率很低, 可以满足大部分情形下快速取词。但是用户不可能熟记哪些字词在缓存词库300中有, 哪些没有, 因此仍然用户存在输入二位码后, 用户翻到最后一页也未找到预期词条的情况。按照上述实施例的处理方法, 若未在缓存词库300中找到预期词条, 则需要用户继续输入以构成三位码或四位码, 以从核心词库200中取词, 或者结束取词。因此本实施例在上述实施例的基础上加入判断模块400, 如

图3所示, 判断用户输入一位码或二位码后, 缓存词库300是否包含用户预期的词条, 如果用户翻到缓存词库300最后一页的时候还在翻页, 表示缓存词库300未包含用户预期的词条。

[81] 相应地, 如图4所示, 在上述实施例的基础上, 在步骤S30与S50之间插入步骤S40: 判断缓存词库300中是否包含用户预期的词条。如果否, 则转入步骤S50; 如果是, 则根据用户命令输出词条, 取词结束。

[82] 因此, 当用户输入一位码或二位码时, 若缓存词库300未包含用户预期的词条, 则很有可能该词条较为生僻, 用户可以选择继续翻页查找或者补充至三位码或四位码。

[83] 若选择继续翻页查找, 由于缓存词库300存储的词条有限, 需要转入核心词库200取词, 即步骤S30中还包括对于一位码或二位码输入的处理: 当用户输入一位码或二位码时, 由于词条先是按前三码索引排序的, 因此按照该一位码或二位码对应的数组下标获得词条的起始位置, 然后按照词条存储的顺序取词并显示。比如输入“aa”, 按照“aaa”、“aab”至“aay”的顺序依次取词显示。

[84] 不管用户如何选择, 因缓存词库300未包含预期词条, 因此有必要进入核心词库200进行索引。若找到词条, 则根据用户命令输出词条, 取词结束。

[85] 以上所述实施例仅表达了本发明的几种实施方式, 其描述较为具体和详细, 但并不能因此而理解为对本发明专利范围的限制。应当指出的是, 对于本领域的普通技术人员来说, 在不脱离本发明构思的前提下, 还可以做出若干变形和改进, 这些都属于本发明的保护范围。因此, 本发明的保护范围应以所附权利要求为准。

权利要求书

- [权利要求 1] 一种五笔输入系统，其特征在于，包括：
- 缓存词库，存储一位码和两位码的常用词的词条信息及索引信息；
- 核心词库，存储所有五笔编码的词条信息及索引信息；
- 取词模块，用于当输入一位码或两位码时，根据所述缓存词库中的索引信息从缓存词库中取词；当输入三位码或四位码时，根据所述核心词库中的索引信息从核心词库中取词。
- [权利要求 2] 根据权利要求1所述的五笔输入系统，其特征在于，所述缓存词库包括：
- 缓存编码索引区，存储常用词的索引信息；
- 缓存词条存储区，存储常用词的词条信息，所述常用词以五笔编码的前两码作为索引，并按照词频由高到低顺序存储。
- [权利要求 3] 根据权利要求1或2所述的五笔输入系统，其特征在于，所述核心词库包括：
- 核心编码索引区，存储所有五笔编码的词条信息的索引信息；
- 核心词条存储区，存储所有五笔编码的词条信息，所有词条以其五笔编码的前三码作为索引按序存储，对前三码相同的词条按照词频由高到低顺序存储。
- [权利要求 4] 根据权利要求1或2所述的五笔输入系统，其特征在于，所述取词模块包括：
- 索引计算模块，根据输入的五笔编码得到索引信息；
- 候选词输出模块，根据索引信息获得并显示词条。
- [权利要求 5] 根据权利要求1所述的五笔输入系统，其特征在于，还包括判断模块，所述判断模块用于根据输入的一位码或二位码判断缓存词库中是否存在用户预期的词条。
- [权利要求 6] 一种五笔输入方法，包括如下步骤：
- 接收五笔编码输入；

当所输入的五笔编码为一位码或二位码时，向存储了一位码和两位码的常用词的词条信息及索引信息的缓存词库取词；

当所输入的五笔编码为三位码或四位码时，向存储了所有五笔编码的词条信息及索引信息的核心词库取词。

[权利要求 7] 根据权利要求6所述的五笔输入方法，其特征在于，向所述缓存词库取词的步骤之后还包括：判断所述缓存词库是否包含用户预期的词条，若缓存词库未包含用户预期的词条，则向所述核心词库取词。

[权利要求 8] 根据权利要求6或7所述的五笔输入方法，其特征在于，所述向缓存词库取词的步骤具体是：将缓存词库中的词以五笔编码的前两码作为索引，并按照词频由高到低进行存储，将输入的五笔编码转换为索引信息，然后根据索引信息将词条按序读取并显示。

[权利要求 9] 根据权利要求6或7所述的五笔输入方法，其特征在于，所述向核心词库取词的步骤具体是：将核心词库中的词以五笔编码的前三码作为索引顺序存储，并对前三码相同的词条按照词频由高到低顺序存储，

如果输入的五笔编码为三位码，则将该三位码转换为索引信息，然后根据索引信息获得的词条按照词频由高到低进行排序显示；
如果输入的五笔编码为四位码，则将由三码输入获得的词条中第四码与用户输入的第四码不匹配的所有词条过滤，获得与该四码输入相应的所有词条，将所获得的词条按照词频由高到低进行排序显示。

[权利要求 10] 根据权利要求9所述的五笔输入方法，其特征在于，所述向核心词库取词的步骤还包括：如果输入的五笔编码为一位码或二位码，则将该一位码或二位码转换为索引信息，然后根据索引信息将获得的词条按照词条在核心词库中的存储顺序依次读取并显示。