(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106598920 A
(43)申请公布日 2017.04.26

(21)申请号 201611065190.0

(22)申请日 2016.11.28

(71)申请人 昆明理工大学
地址 650093 云南省昆明市五华区学府路253号

(72)发明人 邵玉斌 王逍翔

(51)Int.Cl.
G06F 17/22(2006.01)
G06K 9/62(2006.01)

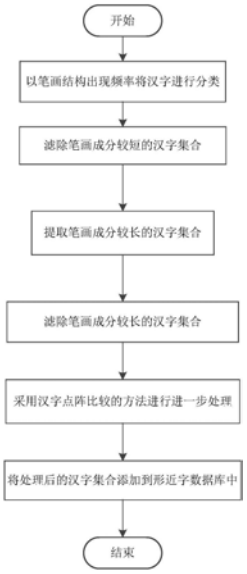
权利要求书1页 说明书7页 附图2页

(54)发明名称

一种笔画编码结合汉字点阵的形近字分类方法

(57)摘要

本发明提供了一种笔画编码结合汉字点阵的形近字分类方法，通过对汉字对应的笔画编码进行统计，以笔画结构出现频率将汉字进行分类生成数据表，每种笔画成分对应包含此成分的汉字集合；然后对集合进行筛选，滤除笔画成分较短和较长的集合，将后者添加到形近字数据库中；对过滤后的汉字集合采用汉字点阵比较的方法进行进一步处理，通过比较同一汉字集合内的汉字的点阵，将相似率较低的汉字滤除，将处理后的汉字集合添加到形近字数据库中；通过以上步骤，就得到了包含大部分汉字的形近字数据库，查询一个汉字的形近字只需要查询其所在的表就可得到它的形近字。本发明提高了形近字分类效率，节约了分类所消耗的时间，获得了较为准确的形近字数据。



CN 106598920 A

1.一种笔画编码结合汉字点阵的形近字分类方法,其特征在于,包括以下步骤:

步骤一、统计笔画编码表中所有笔画结构出现的频度,将出现次数低于10次的笔画结构滤除并排序,将包含相应笔画结构的汉字组成一个集合对应于此笔画结构,由此得到若干个汉字集合;

步骤二、对步骤一得到的汉字集合进行筛选,滤除编码长度小于4的笔画结构对应的汉字集合,将剩余的汉字集合添加到形近字数据库中,每个集合生成一张形近字表;

步骤三、对步骤二得到的汉字集合中的汉字进行点阵比较并计算平均相似度,求所有汉字的相似平均度的平均值,将低于平均值一定比例对应的汉字滤出,得到处理后的汉字集合,将其以步骤二中相同的方法保存到形近字数据库中。

2.根据权利要求1所述的笔画编码结合汉字点阵的形近字分类方法,其特征在于:所述步骤三种汉字点阵比较采用对位比较的方法,计算得到两个汉字的相似度,将点阵表示为由0,1表示的16×16矩阵,有汉字笔画的位置为1,反之为0,相似度计算公式如下:

$$X_{ij}=\sum_{m=0}^{16}\sum_{n=0}^{16}W_{imn}W_{jmn}(W_{imn},W_{jmn}\in\{0,1\})$$

其中,i,j为一个汉字集合中两个汉字的标号,X_{ij}表示两个汉字的相似度,m代表矩阵的行,n代表矩阵的列,W_{imn}代表汉字i的m行n列的值,W_{jmn}代表汉字j的m行n列的值。

3.根据权利要求1或2所述的笔画编码结合汉字点阵的形近字分类方法,其特征在于:所述步骤三中汉字集合中每个汉字的平均相似度计算如下:

$$S_n=\frac{\sum_{i=1}^{N}\frac{\min(L_i,L_n)}{\max(L_i,L_n)}X_{in}}{N}$$

其中,S_n代表字n的平均相似度,i代表字i,n代表字n,L_i代表字i的笔画编码长度,L_n代表字n的笔画编码长度,X_{in}代表两个字的相似度,N代表集合中除去字n的字的总数。

一种笔画编码结合汉字点阵的形近字分类方法

技术领域

[0001] 本发明属于语言处理领域，特别涉及一种汉字形近字分类方法。

背景技术

[0002] 汉字由简单的几种笔画组成，但由于它们在二维空间排列组合，便形成了种类繁多、结构复杂的汉字。构成汉字字形的各种特定的点和线，也是汉字的最小结构单位。根据楷书书写要求，从落笔到抬笔即为一笔，又叫一画，合称笔画，笔画的具体形状叫笔形。由此产生的各种字根形成了众多形态结构相似的汉字，被称为形近字。

[0003] 形近字的识别涉及字形识别。字形识别服务于生活的方方面面，如手写输入，从图像中获取汉字信息，纸质文本转录等，并且这项技术在生活中已经得到了广泛的应用。汉字的字形识别对于如今的技术而言已经不存在问题，对印刷体来说识别精度更高。获取汉字点阵是字形识别的第一步。在汉字的点阵字库中，每个字节的每个位都代表一个汉字的一个点，每个汉字都是由一个矩形的点阵组成，0代表没有，1代表有点，将0和1分别用不同颜色画出，就形成了一个汉字，譬如“我”字如图2所示。通过点阵的比较就可以发现字形结构之间的相关性。

[0004] 笔顺编码是为了记录汉字笔画的书写顺序而设定的具体的笔画的编号。其中，1代表横，2代表竖，3代表撇，4代表捺，5代表折，另外，提为横，点为捺，竖勾为竖，横折为折，竖提为竖，这样所有字就可用1、2、3、4、5这5个符号表示，如图2所示，“李”字的笔画编码是横、竖、撇、捺、折、竖、横，转换成编号是：1234521。笔顺编码包含了汉字的笔画顺序和结构信息，对汉字结构的对比识别有一定作用，但由于编码并未精确表示汉字的基础部件，所以编码所包含的汉字信息并不完整。

[0005] 识别形近字的意义在于，不仅可以帮助使用字形编码输入方式时，如五笔、郑码、手写等输入法，帮助用户提供易错参考，校验文本正确性，还可以应用于儿童识字教学作为参考实例，同时，对系统性的研究汉字结构特点有一定帮助。目前，形近字的识别多为人工收集方式，工作量大，费时费力。

发明内容

[0006] 为了解决上述问题，本发明提供了一种用于汉字形近字分类的方法，该方法实现了机器对形近字自动的分类，为人节约了大量时间和精力。

[0007] 本发明解决其技术问题采用的技术方案是：提供一种用于汉字形近字分类的语言处理方式，包括如下步骤：

[0008] 步骤一、统计笔画编码表中所有笔画结构出现的频度，将出现次数低于10次的笔画结构滤除并排序，将包含相应笔画结构的汉字组成一个集合对应于此笔画结构，由此得到若干个汉字集合；

[0009] 步骤二、对步骤一得到的汉字集合进行筛选，滤除编码长度小于4的笔画结构对应的汉字集合，将剩余的汉字集合添加到形近字数据库中，每个集合生成一张形近字表；

[0010] 步骤三、对步骤二得到的汉字集合中的汉字进行点阵比较并计算平均相似度，把平均相似度较低的汉字滤除，得到处理后的汉字集合，将其以步骤二中相同的方法保存到形近字数据库中。

[0011] 优选的，所述步骤三种汉字点阵比较采用对位比较的方法，计算得到两个汉字的相似度，将点阵表示为由0，1表示的16×16矩阵，有汉字笔画的位置为1，反之为0，相似度计算公式如下：

[0012] $$X_{ij}=\sum_{m=0}^{16}\sum_{n=0}^{16}W_{imn}W_{jmn}(W_{imn},W_{jmn}\in\{0,1\})$$

[0013] 其中，i，j为一个汉字集合中两个汉字的标号，X_{ij}表示两个汉字的相似度，m代表矩阵的行，n代表矩阵的列，W_{imn}代表汉字i的m行n列的值，W_{jmn}代表汉字j的m行n列的值。

[0014] 优选的，所述步骤三中汉字集合中每个汉字的平均相似度计算如下：

[0015] $$S_n=\frac{\sum_{i=1}^{N}\frac{\min(L_i,L_n)}{\max(L_i,L_n)}X_{in}}{N}$$

[0016] 其中，S_n代表字n的平均相似度，i代表字i，n代表字n，L_i代表字i的笔画编码长度，L_n代表字n的笔画编码长度，X_{in}代表两个字的相似度，N代表集合中除去字n的字的总数。

[0017] 本发明的有益效果在于：先对汉字通过分析其笔画编码的特征进行较粗略的形近字分类，节约了要对所有字进行分类所需大量的时间精力，效率大大提高；但由于笔画编码和笔画之间并不是一一对应，这样的分类还需要进一步优化，汉字点阵的比较的作用就在于此，它的使用可以滤除分类中不正确的成分，提高了结果的精确度；两种方法的结合，实现了由多到少，由粗到精的处理的过程，既保证了方法的效率，又达到所需的正确率。

附图说明

[0018] 图1是本发明的流程图；

[0019] 图2是笔画编码实例图；

[0020] 图3是包含相同笔画成分但字形不相似汉字实例图；

[0021] 图4是相似汉字点阵对比图。

具体实施方式

[0022] 下面结合附图和具体实施例对本发明的技术方案做具体阐述。

[0023] 如图1所示，本发明提供了一种用于形近字分类的语言处理方法分为以下三个步骤：

[0024] 一、从网上下载UNICODE汉字笔画编码表，是一个所有20902个汉字(U+4E00～U+9FA5)的笔画顺序表，部分如表1所示。

[0025] 表1部分UNICODE汉字笔画编码表

[0026]

汉字	序值	Unicode 编码	笔顺
一	00001	4E00	1
丨	00002	4E28	2
丿	00003	4E85	2
丿	00004	4E3F	3
丶	00005	4E36	4
乀	00006	4E40	4
乁	00007	4E41	4
乙	00008	4E59	5
乚	00009	4E5A	5
乛	000010	4E5B	5

[0027]

二	000011	4E8C	11
丁	000012	4E01	12

[0028] ……

[0029]

汉字	序值	Unicode编码	笔顺
求	01499	6C42	1241344
忎	01500	5FD1	1244544
孛	01501	5B5B	1245521
車	01502	8ECA	1251112
甫	01503	752B	1251124
匣	01504	5323	1251125
更	01505	66F4	1251134
亜	01506	4E9C	1251221
朿	01507	675F	1251234
吾	01508	543E	1251251
叟	01509	53D3	1251254
豆	01510	8C46	1251431
或	01511	6213	1251534
迓	01512	8FCA	1252454
两	01513	4E24	1253434

[0030] ……

[0031] 其中1表示“横”;2表示“竖”;3表示“撇”;4表示“捺”;5表示“折”统计所有笔画结构

出现的频度。笔画编码表汉字排列顺序由其笔画长度由短到长依次排列，每个汉字对应一个笔画编码，对其的分析过程为：由上到下依次分析每个字所包含的笔画成分，如果笔画成分之前未出现过，则将其保存为一类，其出现次数记为1，如果遇到出现过的笔画成分，则将其出现次数加一；将此表遍历后，就得到了所有笔画结构的出现次数，将出现次数低于10次的特例滤除并依次排序，就完成了笔画结构的统计。在两万个汉字中统计得到笔画和对应集合包含汉字数表，部分如下表2所示。然后，将以上步骤筛选得到的笔画结构即高频度笔画结构作为标志，如“511”(折横横)、“112”(横横竖)等，将包含相应笔画结构的汉字组成一个集合对应于此笔画结构，由此得到大量有交集的汉字集合，并保存到数据库中，生成待处理的数据表，部分如表2所示。

[0032] 表2笔画成分及对应的汉字数量分类表

[0033]

包含笔画成分	汉字数
1	20219
5	19309
2	19301
3	18297
4	17754
12	15343
25	13380
51	13364
11	13029
34	11191
251	10994
21	10065
13	9810

[0034] ……

[0035]

5411234	24
541254	24
541435	24
54444354	24
5444454	24
545231	24
5452312	24

[0036] ……

[0037] 二、对于已有的汉字集合表，显然笔画成分1、2、3、4、5出现会最多，它 们对应的表中汉字数也最多，但由于标志笔画成分太短，包含的形态信息有限，其集合中的汉字没有任何突出的共同特点，所以需要滤除较短笔画结构即编码长度小于4的笔画结构对应的汉字集合，其中也包含12(横竖)、25(竖折)等常见笔画成分，并结合其出现次数判断其是否具有

特征，包含汉字数过多的笔画成分则不具有特征；但就较长笔画结构，编码长度大于9的笔画结构来说，如3412515415，包含此结构的字有翕、翎、嗡、嶲等，它们已经具有较强的相似相似性，集合中也几乎不会含有不相似的成分，如表3所示，较长笔画结构对应的汉字数据表可直接复制添加到形近字数据库中，就可生成一张形近字表。形近字表添加完成后，为了后续处理需要将较长笔画结构对应的汉字集合也同样滤除。

[0038] 表3形近字表

[0039]

序号	汉字	Unicode编码	笔顺编码
09871	翕	7FD5	341251541541
09872	翎	7FD6	341251541541
14580	嗡	564F	251341251541541
14630	嶲	5D96	252341251541541
15207	滃	6F5D	441341251541541
15347	嬝	5B06	531341251541541
16224	歙	6B59	3412515415413534
16492	燊	71BB	4334341251541541
18765	蹯	8E79	2512121341251541541
19361	鬮	95DF	25112511341251541541

[0040] 三、通过以上步骤后，对于剩下汉字集合，可能会出现这样的情况，如图3所示，113533所代表的汉字集合中，会有“埸”、“啄”两个字，虽然都包含113533成分，但却不相似，这由于笔画编码的不精确引起的，笔画编码存在把横、横折钩、提表示为1的简化情况。所以需要把汉字集合中的汉字进行点阵比较，把那些特例，即平均相似度较低的汉字滤除，得到处理后的新的汉字表，将其以同步骤二中相同的方法保存到形近字数据库中，才能完成形近字的统计分类工作。

[0041] 对于同一集合内的汉字的点阵比较，其基本方法是将两个汉字的点阵对位相乘，如果有重叠部分则结果为1，反之为0，重叠越多其相似度越高，譬如，如图4所示，说和悦字，共有38个点重合，而“说”字一共63个点，重合点占了60％。形态上相似的字会有更多的重叠，其相似度也相应的高于不相似的组合。本发明采用汉字的16*16点阵，将两个汉字的相似度具体定义如下：

[0042] $$X_{ij}=\sum_{m=0}^{16}\sum_{n=0}^{16}W_{imn}W_{jmn}(W_{imn},W_{jmn}\in\{0,1\})$$

[0043] 其中，i，j为两个汉字的标号，X_{ij}表示两个汉字的相似度，m代表矩阵的行，n代表矩阵的列，W_{imn}代表汉字i的m行n列的值，W_{jmn}代表汉字j的m行n列的值。

[0044] 要在集合中滤除不具有集合内大多数汉字所具有特征的特例，就需要汉字在集合内平均相似度的计算，以此来滤除平均相似度低的汉字。因为笔画编码长度，即汉字总的笔画数的不同，汉字的结构就可能不同，如寸和付字，一个为独体结构，一个为左右结构，点阵的对比会出现误差，所以平均相似度的计算以加权方式进行，以减小这种情况的误差的影响，具体算法如下：

[0045] $$S_n = \frac{\sum_{i=1}^{N} \frac{\min(L_i, L_n)}{\max(L_i, L_n)} X_{in}}{N}$$

[0046] 其中，S_n代表字n的平均相似度，i代表字i，n代表字n，L_i代表字i的笔画编码长度，L_n代表字n的笔画编码长度，X_{in}代表两个字的相似度，N代表集合中除去字n的字的总数。笔画编码长度相差越大的汉字，计算此平均相似度中就会有更高的权值，以此减小误差。然后计算所有汉字集合中汉字的平均相似度的平均值，将低于平均水平一定百分比的汉字排除，根据实际需要可以设置70％、80％等不同的数值。

[0047] 以1.泮(44143112)、2.胖(351143112)、3.拌(12143112)、4.绊(55143112)、5.班(1121431121)五个字为例子说明：

[0048] 计算汉字两两的相似度，也就是代表两汉字点阵重叠的点数，计算得到以下相似度：

[0049] $X_{12}=24$ $X_{13}=43$ $X_{14}=14$ $X_{15}=23$

[0050] $X_{23}=25$ $X_{24}=34$ $X_{25}=23$

[0051] $X_{34}=18$ $X_{35}=14$

[0052] $X_{45}=21$

[0053] 其中，X_{12}显然等于X_{21}，所以没有列出后者，然后再利用前述平均相似度计算每个字在当前表中的平均相似度，五个字的笔画编码长度分别为$L_1=8$、$L_2=9$、$L_3=8$、$L_4=8$、$L_5=10$。

[0054] $$S_1 = \frac{\sum_{i=1}^{4} \frac{\min(L_i, L_n)}{\max(L_i, L_n)} X_{in}}{4}$$

[0055] $$S_1 = \frac{\frac{8}{9}*24 + \frac{8}{8}*43 + \frac{8}{8}*14 + \frac{8}{10}*23}{4}$$

[0056] $S_1=24.18$

[0057] 同理，算出S2＝23.62、S3＝23.60、S4＝19.75、S5＝16.78，汉字集合中所有汉字平均相似度的平均值为21.586，80％*21.586＝17.2688，5.班(1121431121)字低于这个值，所以滤除；至于从一组数据中筛选滤出低于平均水平的数据的方法有很多，在此不再赘述。

[0058] 由于一个汉字包含不止一种特征信息，所以根据不同的特征信息可能同时包含在多个形近字表中，例如，“斩”字和“折”、“近”等字会归为一类，也会和“轨”、“转”等字分为一类。在查询形近字的时候就需要查出所有包含此字的形近字表，将多张表中的重复结果滤除再合并，就可得到某个字完整的形近字表。

[0059] 通过以上实施方式，很好地实现了本发明的目的，本发明通过利用汉字笔顺编码所提供的特征信息对汉字形近字进行筛选，进一步通过汉字点阵比较提高了系统精度，很好的协调了效率和准确度，使人从繁重的手工统计中解放出来，使形近字的获取更加快捷方便。

[0060] 尽管本发明的实施方案已公开如上，但其并不仅仅限于说明书和实施方案中所列运用，它完全可以被适用于各种适合本发明的领域，对于熟悉本领域的人员而言，可容易地实现另外的修改，因此在不背离权利要求及等同范围所限定的一般概念下，本发明并不限

于特定的细节和这里示出与描述的图例。

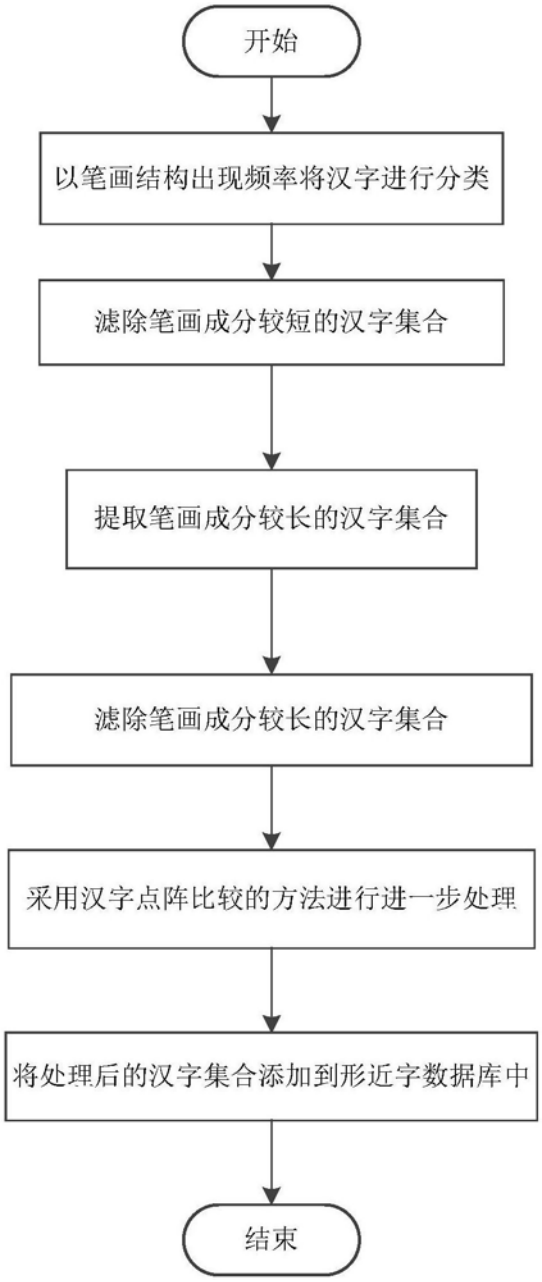

开始
以笔画结构出现频率将汉字进行分类
滤除笔画成分较短的汉字集合
提取笔画成分较长的汉字集合
滤除笔画成分较长的汉字集合
采用汉字点阵比较的方法进行进一步处理
将处理后的汉字集合添加到形近字数据库中
结束

图1

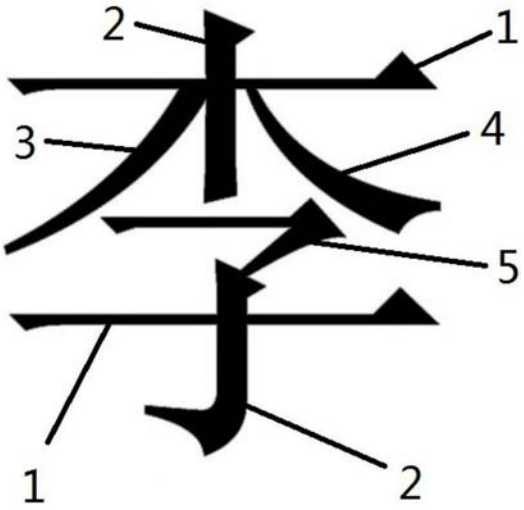

2
1
3
4
5
1
2

图2

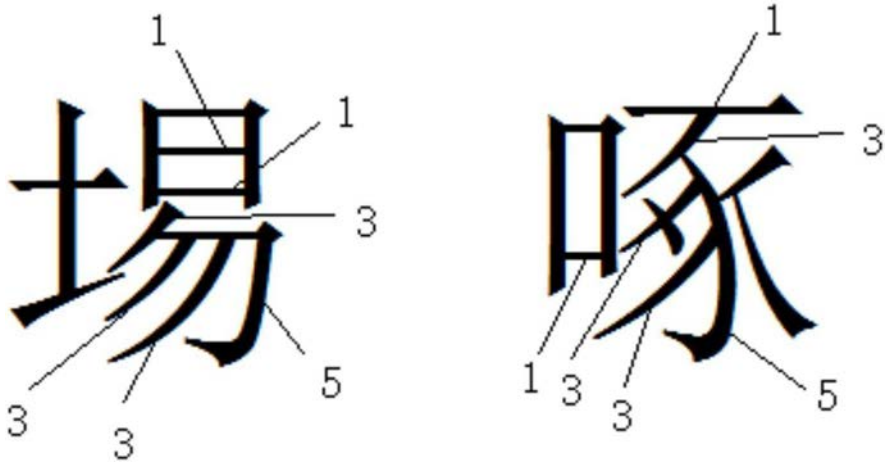


图3

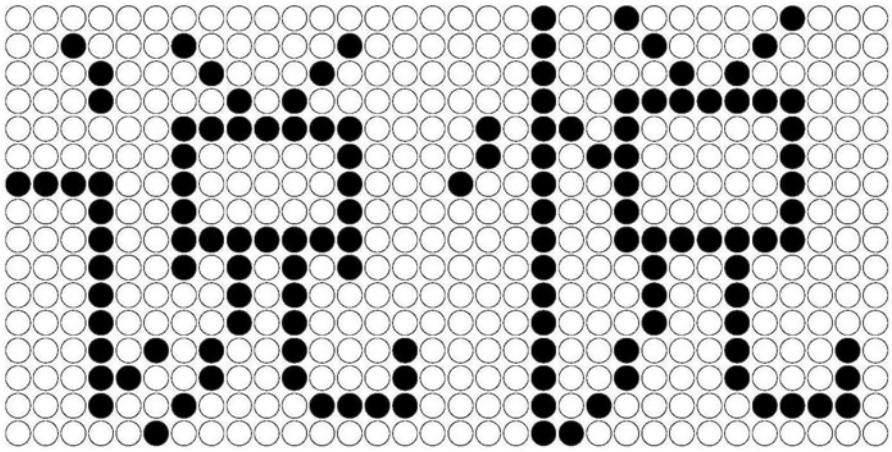


图4