基于编码的生僻汉字输入方法理论与测试研究

白　毅　易军凯*
(北京化工大学信息科学与技术学院, 北京　100029)

摘　要: 在 BNF 范式编码的基础上,深入讨论了生僻汉字数字化处理问题。根据对生僻汉字部件的统计和分析,本文提出了基于编码的生僻汉字输入方法,建立了相应的生僻汉字部件库,实现了生僻汉字的数字存储和显示。此外,应用测试用例自动生成的方法,对输入方法进行了测试,表明该方法具有造字速度快,应用范围广,与现有字体兼容性强等优点,从而为生僻汉字的数字化提出了一个新的解决方案。

关键词: 生僻汉字; 部件组合; 字库; TrueType 字体; 测试用例
中图分类号: TP391.1

引　言

古文物、古文献等文字材料数字化过程中遇到的一个最大问题就是缺字问题,即部分规范文字在现有的计算机系统字库没有编码,无法处理,这类文字称为生僻汉字或特殊字。在换发第二代身份证时,也有人因为名字中出现生僻字不能顺利领取身份证。生僻汉字主要出现在一些古代文献(比如文学、汉语、古代史、考古、中医等)和一些关于地理、人名、方言等名词性术语中[1]。如何解决特殊字输入、加工处理、显示打印等数字化工作环节中的问题,一直是困扰证件办理、文史研究以及古文献电子编辑出版的难题。因此,研究一种新的文字编码解决方案非常有必要。

本文提出了基于编码的生僻汉字输入方法,与通用汉字输入方法保持一致,把生僻汉字进行拆分,单独输入各部分的结构,再进行重新组合,最终显示为矢量字体。针对生僻汉字的构造,建立部件结构,填入各部分的内容,组成新矢量汉字。为了验证该编码规范的可行性和输出质量,本文应用测试用例自动生成的方法,对基于该编码规范的输入方法进行了实例测试。测试结果表明,该编码方法对生僻汉字的输入问题具有很强的解决能力,为生僻汉字的数字化提供了一种新的解决方案。

收稿日期: 2007-03-27
第一作者: 男,1983 年生,硕士生
*通讯联系人
E-mail: yijk@mail.buct.edu.cn

1　汉字编码技术

1.1　汉字编码原理

1.1.1　组字原理　汉字是一种音形结合的表意文字。从文字学的角度出发,汉字可由单字、笔划、部首三级构成。单字即一个结构完整的汉字,它是一个具有"音、形、义"的完整的文字书写符号组合块;笔划是指在书写汉字时按一定笔向连续写成的每一笔;部件是在汉字中反复出现的、能从字形中分隔出来的有固定形体的笔划[2]。

根据汉字的上述特点,可以看出解决生僻汉字的输入问题,实际上就是使得组成这些生僻汉字的单字、笔划和部首按照一定的顺序进行排列组合后输入;而且,一些生僻汉字也可以按照一定的排列方式组合成更为复杂的汉字。

因此,用户应该首先选择所要输入的生僻汉字的排列类型,然后用对应的单字、部首、笔划来填入相应的位置,需要的时候也可以嵌套上述过程以组成更为复杂的生僻字[3]。

1.1.2　编码规则　在对汉字字形特点进行分析的基础上,结合英文字母的象形意义,本文提出了汉字编码的符号定义规则,其由 BNF 范式表示如下

<汉字> ::= <汉字组字结构编码>
<汉字组字结构编码> ::= <左右结构汉字编码> | <上下结构汉字编码> | <内外结构汉字编码>
<左右结构汉字编码> ::= M(<构件序列>)
<上下结构汉字编码> ::= E(<构件序列>)
<内外结构汉字编码> ::= <全包围结构汉字

>| <半包围结构汉字>

<全包围结构汉字> ::= O(<基本构件序列>)

<半包围结构汉字> ::= <左上包围结构>| <左下包围结构>| <右上包围结构>| <上左下包围结构>| <左下右包围结构>| <左上右包围结构>

<左下包围结构> ::= C(<基本构件序列>)

<左上包围结构> ::= F(<基本构件序列>)

<右上包围结构> ::= D(<基本构件序列>)

<上左下包围结构> ::= G(<基本构件序列>)

<左上右包围结构> ::= N(<基本构件序列>)

<左下右包围结构> ::= U(<基本构件序列>)

<构件序列> ::= <基本构件序列>| A(<构件序列>)| V(<构件序列>)

<基本构件序列> ::= <基本构件><基本构件>| <基本构件><基本构件序列>| O(<基本构件序列>)| C(<基本构件序列>)| D(<基本构件序列>)| F(<基本构件序列>)| G(<基本构件序列>)| N(<基本构件序列>)| U(<基本构件序列>)

<基本构件> ::= BIG5 字符集中的汉字与部首等

汉字组字结构分 4 类:上下结构(E)、左右结构(M)、内外结构和附加结构。其中内外结构又分为全包围结构(O)和左上包围结构(F)、左下包围结构(C)、右上包围结构(D)、上左下包围结构(G)、左上右包围结构(N)、左下右包围结构(U)(注:结构编码字母基本按象形文字的习惯取字母形状,不区分大小写,下同),如图 1 所示。

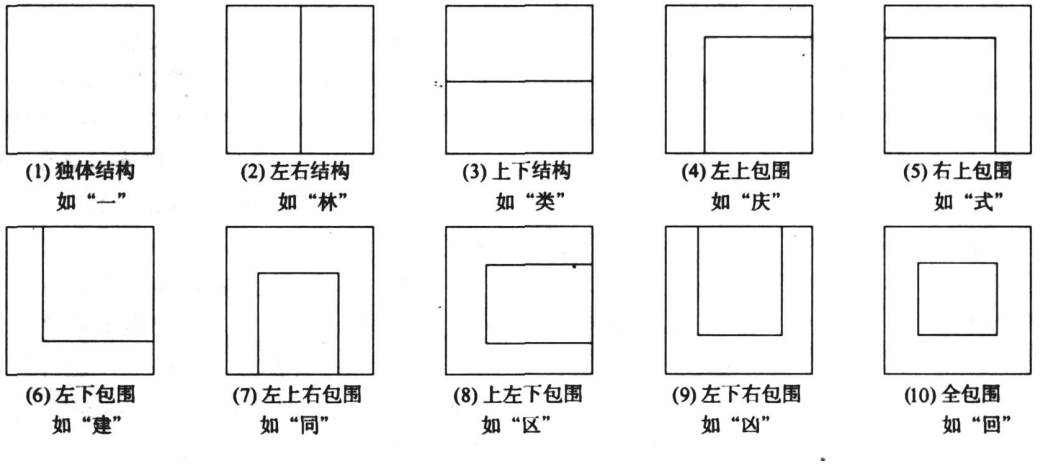

图 1 汉字结构示意图

Fig. 1 Drawing of Chinese characters' structure

1.2 编码与图形的转换

基于 1.1 节中定义的编码规则,对给定汉字的结构进行分析,不难得出该汉字的编码;根据编码,可以计算出该汉字每个部件的相对坐标值;根据这些坐标值将每个部件依次画出,即得到所需汉字,完成编码汉字的图形转化。

从根本上说,汉字部件相对坐标的计算规则取决于汉字的结构。但是,为了汉字的美观,需要对各种包围结构的部件进行缩放,具体来说,即是将被包围的中心部件区域扩大为原来的 150 %。

对于一个多层嵌套结构的生僻汉字来说,只有最底层嵌套的两个部件的相对坐标得以确定之后,整个汉字的相对坐标才能确定。由于汉字编码的读取是顺序逐个进行的,在此过程中,如果读取到的汉字部件不是处于最底层嵌套,就需要先将这些部件的相关信息存储起来,等最底层嵌套部件的相关信息处理完毕之后,再将其读取出来进行处理。

基于 1.1 节定义的汉字编码,实际上是一个中序遍历的二叉树(除去编码中的括号)。其中,每个基本构件的文本相当于叶子结点,结构编码相当于非叶结点。然而,计算每个汉字部件的坐标,是按照自下而上的顺序进行的,即先从最底层开始逐层计算,直到最高层为止,这个过程相当于对二叉树进行了一次后序遍历。因此,在计算坐标前,必须先将得

到的汉字编码由二叉树的中序遍历形式变为后序遍历形式。

汉字部件的存储是在对新得到的汉字编码进行顺序扫描的过程中同步进行的。在扫描的过程中,如果扫描到基本构件,则将其存入堆栈中,并继续扫描下一个;如果扫描到结构编码,则将堆栈中的最后两个元素取出,根据结构编码来计算并存储这两个元素坐标值,再将这两个基本构件看作一个新的汉字部件存入堆栈中,其中,部件的文本为上述两个基本构件与结构编码组成的一个新字符串,坐标值为两个基本构件区域并集的坐标值。如此循环,直到所有编码字符扫描完毕。

扫描完毕后,只需按照记录下来的各个汉字部件的坐标值,将部件文本逐个显示在其坐标值对应的区域,即将整个生僻汉字显示为图形。

1.3　图形与 TrueType 字体的转换

在字库中是用字的内码(1 个 unsigned short 值)来索引字形的,而造字程序及人们便于记忆的是字的编码,必须要有某种方法能实现两者之间的协调[4]。本方案的做法是使用哈希算法将字符形式的编码串映射为一个 unsigned short 值。由于哈希过程中会产生冲突,故需采用再哈希方法来解决冲突[3]。

1.4　生僻汉字部件的输入

由于本方案是基于汉字部件的,如何输入这些汉字部件及其结构信息成为一个重要问题。本方案提供两种方式来实现部件与结构的输入。

一种方式是对生僻汉字进行结构分析,得出其编码表达式(中序遍历形式),然后直接用键盘输入此编码表达式,由计算机对其进行处理后显示为图形并存入字库。

另一种方式则是先选择汉字的结构,由计算机绘制出结构框架,然后再向框架中输入对应的部件文本,然后显示为图形并存入字库。

比较之后不难发现,第二种方式要更好一些,因为大多数生僻汉字的结构都比较复杂,直接输入编码表达式显得比较繁琐。

2　输入方法测试结果与分析

2.1　生僻汉字用例生成方法

为检验基于本方案开发的输入方法的功能和性能,需要生成一些生僻汉字的用例来进行测试。根据 1.1 节中描述的汉字组字原理,将测试用例的选取方案制定为

(1)选取两个汉字部件作为测试用例。这个过程分两步进行:第一,两个部件都从笔划、部首和单字中任意选取一种;第二,两个部件之间的结构组合从图 1 所示的 10 种结构中选取一种。这样,两个部件的组字共需要选取的测试用例数量为

$$P_3^2 C_{10}^1 = 60$$

需要说明的是,作为测试用例的单字必须是 BIG5 字符集中的汉字。

(2)选取多个汉字部件作为测试用例,进行多层结构嵌套。这是由于图 1 的 10 种结构都是可以进行嵌套的。嵌套的层数不要超过 32。

(3)为了保证输入方法的通用性,所有的汉字部件的选取应该都是随机的。

2.2　输入方法测试结果的分析

根据 2.1 节中提出的测试用例生成方法,选取一定数量的汉字部件自动生成 200 个生僻汉字。统计得到的输出结果之后,对其进行速度、代表性以及相似程度等方面的分析。

2.2.1　速度分析　许多生僻汉字都是结构十分复杂的,嵌套的层数多在 5 层以上,因此造字速度就显得十分重要。由于本方案能够快速绘制出用户所要求的结构,再加上现有许多普通汉字输入法都能对 BIG5 集中字符进行快速输入,故而这些生僻汉字都能够在极短的时间内就输出在屏幕上,输出 200 个生僻汉字测试用例仅耗时 6.2 s,与主流的普通汉字输入法已经差别不大。因此,本方案完全满足用户对生僻汉字快速输入的要求。

2.2.2　代表性分析　2.1 节中提出的测试方案,实际上是用以偏概全的形式覆盖汉字的所有可能的排列结构。下面从 2.1 节中自动生成的 200 个生僻汉字中选取部分代表性较强的汉字进行分析。

刀笔划与笔划的组合,右上包围结构;编码为:D 刁丿。

乩笔划与单字的组合,左右结构;编码为:M 占乚。

亣部首与笔划的组合,左上包围结构;编码为:F 广乚。

囧部首与部首的组合,全包围结构;编码为:O 口厶。

匝部首与单字的组合,上左下包围结构;编码为:G 匚合。

岢单字与单字的组合,上下结构;编码为:E 山

可。

遵多层嵌套,包括左下包围、上下结构和左右结构;编码为:C辶(E(E穴(M糸(M言糸)))(E(M長(M馬長))心))。

由上述代表性很强的生僻汉字可以看出,本方案能够对1.1节中的生僻汉字组字原理完全实现,从而可以对几乎所有生僻汉字实现计算机输入。

在此基础上,参考黑龙江大学中文系教授范子烨博士编写的《中国繁难文字字库初编》,对其中的1.6万个左右的生僻汉字进行了测试,能够解决其中99.9%左右的生僻汉字的输入问题;参考东汉许慎编写的《说文解字》,则能够解决其中90%左右的汉字输入问题。这也再次证明本方案对生僻汉字输入问题具有很强的解决能力。

2.2.3 与实际字体的相容性 汉字是一种二维图形,空间结构上具有很强的美感,这就要求生僻汉字输出的图形必须具备不弱于已有实际字体的美观程度。通过对2.1节中自动生成的200个生僻汉字的输出结果进行统计分析,可以看出本方案生成的95%以上的字体图形与实际字体十分相似,很难看出差别。这说明本方案生成的字体在图形美感上也基本满足实际汉字的要求。

3 结束语

本文所提出的基于BNF范式编码的生僻汉字输入方法是一种全新的中文信息处理方法,可应用于很多领域。该方法不需参考字库,对现有汉字编码集之外的生僻汉字或虚构的汉字也可以自动生成字形,并可以将其嵌入到一些常用的开放的数据库管理系统、办公自动化系统中,从而弥补现有汉字内码包含信息不足等缺点,进一步实现中文信息按部件结构信息的索引、排序及查找等功能,使数据库管理系统、办公自动化系统等更科学、更全面、更系统地利用汉字结构知识。对本方案应用测试用例自动生成的方法,测试输入方法的功能和性能,证明其可行性和科学性,在理论上和实践上都有十分重要的意义。

参考文献:

[1] 阚映红. 地图数据库建立和应用过程中生僻汉字的处理[J]. 测绘学院学报,2000(1):42-45.

[2] 孙星明,殷建平,陈火旺. 汉字的数学表达式研究[J]. 计算机研究与发展,2002,39(6):707-711.

[3] XIAO Fanlin, XIAO Qingding, CHEN Ming. Adaptive confident transform based classifier combination for Chinese character recognition[J]. Research on handprinted Chinese character recognition, Ph.D, 1996,42(6): 1-89.

[4] 杨建红,刘蓉. TrueType字体在图形图像处理软件中的应用[J]. 武汉大学学报:工学版,2004(6):110-112,136.

[5] ZONGKER D E. Example-based hinting of true type fonts[J]. Proceedings of the ACM SIGGRAPH Conference on Computer Graphics, 2000, 5(3): 411-416.

Research and test on code-based rare Chinese character input method

BAI Yi YI Jun Kai

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: Based on BNF code, this paper investigates the problems concerning digital processing of rare Chinese characters. According to statistics and analysis of rare characters' components, the paper presented a code-based input method for rare characters and built a corresponding rare characters components library. The approach can realize digital storing and displaying of rare Chinese characters. Furthermore, the input method has been tested using automatic testcase generation approach. The result shows that the input method has lots of advantages, such as high speed, broad application, strong compatibility with current font and so on. The input method opens a new scheme in the filed of rare Chinese character digitalization.

Key words: rare Chinese character; components combination of Chinese character; font library; TrueType font; testcase