第 24 卷 第 3 期 华侨大学学报(自然科学版) Vol. 24 No. 3
2003 年 7 月 Journal of Huaqiao University (Natural Science) Jul. 2003

文章编号 1000-5013(2003)03-0331-04

古籍电子化中生僻汉字的处理

闫凡蕾 林仲湘 李 龙

(广西大学中文系,广西 南宁 530004)

摘要 “字库不够用”的问题是古籍电子化的一大障碍.文中就古籍电子化中遇到的生僻汉字问题进行了深入研究.在对生僻汉字进行分析统计的基础上,进而对其进行编码、造字.又根据生僻汉字的特点,设计了易学易用的输入法.从而,实现了古籍电子化中遇到的生僻汉字的存储、检索和显示,较好地解决了古籍电子化中遇到的生僻汉字问题.

关键词 汉字, 字库, 古籍, UNICODE

中图分类号 TP 391.43 文献标识码 A

曾经,饱览群书是做学问的基本条件.但是,任你学富五车,要穷究一个用典出处,就是上穷碧落下黄泉,也未必能即时找到.而今,随着科技的进步,古籍的电子化让你随时博古通今.文字在电脑和网络上能否正确地表达和显现,是古籍电子化的关键.汉字是表意文字,每个汉字必须用一个编码来表示.从古籍电子化伊始,困难就集中在操作平台的统一和建立古籍资源够用的大字符集上[1].由于 Windows 操作系统被广泛应用,因此基本上选用 Windows 操作平台,但“字库不够用”的问题仍是开发古籍产品的一大障碍[1].本文提出了基于 UNICODE 的古籍电子化中字库缺字问题的一种解决方案,在广西大学古籍整理研究所的科研项目中得到了应用,并且取得了良好的效果.另外需要说明的是,鉴于 Windows XP 操作系统应用越来越广泛,我们所讨论的解决方案是基于 Windows XP 操作系统的,对 Windows 9X 不支持.

1 国内研究现状

在探讨国内研究现状时,有必要回顾汉字编码发展的历史.从 1980 年开始,我国陆续制定、颁布了一系列的汉字编码标准.1981 年颁布的国家标准 GB 2312-1980《信息交换用汉字编码字符集——基本集》,收录了 6 763 个汉字.这对于绝大多数汉字信息处理系统来说,已能基本满足各种实用要求.但是,对于像古籍整理、辞典编纂、户籍管理等应用系统来讲,仍远不能满足使用要求.历代积累下来的汉字总字数(包括异体字)达 6 万之多[2].1993 年国际标准化组织颁布了 ISO/IEC 10646 标准(《信息交换——通用多八位编码字符集(UCS)》),它收录了 20 902 个 CJK 汉字.1996 年全国信息化标准化技术委员会颁布了《汉字扩展内码规范》,在该

收稿日期 2002-12-31

作者简介:闫凡蕾(1976-),男,硕士研究生,E-mail:fl_yan@163.com

规范中提出了一种与现行 GB 2312-1980 内码体系兼容的、能支持 ISO 10646 标准 CJK 汉字的两字节内码体系 GBK.

我们通常所说的 Unicode, 本身是英文 Universal Code 的缩略语, 顾名思义, 就是统一编码. 就内容而言, Unicode 和 ISO/IEC 10646 是一致的, 并行的. CJK 是"中日韩汉字"的简称. 自从 ISO/IEC 10646 实施中日韩汉字统一编码以来, CJK 便成为了"中日韩统一编码汉字" CJK Unified Ideographs 的简称.

从上述可知, UNICODE 编码空间越来越大, 编码汉字越来越多, 从而使古籍实现数字化成为可能. 在进行古籍的数字化进程中, 各科研机构采取了各种方法来处理字库中缺少冷僻汉字的问题. 目前, 互连网上的许多古籍文献采用的方法是用"□"代替偏僻字, 并在方括号内注释其汉字的结构.

北京国学时代文化传播有限公司与商务印书馆联合推出的"中国历代基本典籍库"使用 GBK 字库, 另附有专用图形字库. 国防科技大学研制的科研课题"汉字的数字表达式"项目, 能够让冷僻汉字根据所给出的数学表达式自动组合产生. 北京书同文公司开发的《文渊阁四库全书》, 形成了在国际标准框架内定义的、适合古籍整理与研究的 3 万汉字的大字符集. 宁夏大学开发的"西夏字处理系统", 虽然不属于古籍数字化的范畴. 但是, 其采用"位面"的设计思想却实现了西夏字、汉字、英文同平面共存的方法, 可为我们所借鉴. "该设计方法彻底突破了汉字码位紧张的瓶颈, 使得以后字库的升级、扩充变得简单易行……用户只需作简单的拷贝就可以获得新的、容量更大的字库"[3]. 广西大学中文系在开发《古今图书集成》(电子 1.0 版)时, 采用了外挂汉字平台的方法. 它用 Richwin 造字软件造字. 在录入人员输入数据、用户察看信息时, 先启动汉字平台 Richwin, 即可显示正确的汉字. 广西大学中文系在开发《多功能现代汉字辞典》(多媒体 1.0 版)时, 采用"伪字库"的方法. 即先用 Photoshop 或 Windows 自身带的画图软件, 制作每个冷僻汉字的图形文件. 然后, 在软件中需要显示该汉字的地方调用相应的图形文件.

以上方法, 虽然各有利弊, 但都是在目前条件下对"字库不够用"问题的积极探索. 它必将为 UNICODE 的发展做出贡献.

2 解决方案

目前我们正在开发《古今图书集成》(电子 2.0 版)和《多功能现代汉字辞典》(多媒体 2.0 版), 其工作是结合实际情况的. 比如, 为了解决资金短缺、设备简陋、人手不足、技术力量薄弱等问题, 我们借鉴了解放军信息工程大学测绘学院在数字制图中对生僻汉字的处理方法. 从中设计出一套方案.

2.1 生僻汉字的统计

我们在进行《古今图书集成》(电子 1.0 版)和《多功能现代汉字辞典》(多媒体 1.0 版)时, 已经积累了大量的实践经验, 统计出大约需要造的字有 3 500 个. 现在做升级版, 计算机的软硬件环境已经大大改善. 另外考虑到 Windows XP 操作系统已经推出 1 a 多, UNICODE 中汉字符号已经达到 27 000 多个, 所以我们重新进行了统计. 在 Windows XP 系统的造字程序中, 若选用 GBK 编码体系, 用户仅能造 1 894 个; 而选用 UNICODE 编码体系, 用户可以造 6 400 个字. 这么大的容量已经足够我们用了. 我们对生僻汉字进行统计的方法是, 先把原来统计出

来需要造的字, 在 Windows XP 系统中逐一排查, 把不能显示出来的另造. 在录入人员打字时, 把不能显示的字随时进行记录, 在计算机中把该字先用 4 位正整数来表示(不同的字所采用的数值应该不一样) . 最后, 对这些字统一编码造字, 在最后做成字库后, 再编写程序把这 4 位正整数统一转换成所造的字.

2. 2　制作生僻汉字图形

我们采用 Photoshop 7. 0 和 Windows 操作系统自身所带的画图软件进行造字. 这种方法相当于做出一个个汉字图形. 根据统计出来的生僻汉字, 取已有相同字体汉字的偏旁部首进行组合, 必要时要对汉字笔划进行拉伸、缩放、割断、移动、删除、拷贝和定位等操作. 对汉字的造字过程而言, 汉字笔划实质上是一个个多边形. 此时的汉字造字就是图形的编辑修改, 在对所造的汉字经过多次绘图检查和反复修改、确认所造汉字结构合理、大小适中以后, 形成所要的生僻汉字图形. 在基本保证使用的前提下, 我们选择了宋体字进行造字.

2. 3　**Windows** 生僻汉字资源字库

Windows 生僻汉字资源字库是指将上面造的汉字转换成 Windows 所支持的 ttf 或 tte 格式的字库, 这种格式的生僻汉字库在 Windows 平台上就能和 Windows 自身所带的字库混合使用. 为此, 先把造好的汉字图形分别剪贴到 Windows 的造字软件中, 剪贴的顺序同生僻汉字在空白区所排列的顺序相一致. 通过这种方法便得到 Windows 生僻汉字资源字库. 利用 Windows XP 系统中 TrueType 造字程序提供的字体连接技术, 使所造的生僻汉字资源字库和宋体字相连接. 生僻汉字便成为 UNICODE 的成员, 扩充了 UNICODE 字符集. 新的字库可以存储、管理、显示和打印, 和其他汉字一起使用. 所生成的字库能被安装到任何一台装有 Windows XP 操作系统的电脑中.

2. 4　生僻汉字输入法

汉字的处理过程是: 输入汉字编码后, 根据所选汉字对应的内码, 检索到汉字在汉字库中的地址码及相应的字模信息, 从而实现汉字的输出[4]. 在做生僻汉字输入法的时候, 我们采用 Windows 系统自身带的输入法编辑器. 输入法编辑器, 又称前段处理程序, 专门用于解决如何借助于标准的 101 键的键盘输入成千上万的东方象形文字. 在 Windows 环境中, 一个 IME 模块中包含一个转换引擎, 负责将键盘点击转换为拼音文字或象形文字, 以及一个含有常用象形字的码表[5]. 当 IME 程序被激活后, 它将捕捉所有的键盘事件, 经它解释并转换成用户真正希望的字符代码后, 再传递给应用程序.

2. 4. 1　*生僻汉字代码的编制*　根据生僻汉字的特点, 我们编写了笔划码. 笔划码是一种形码, 根据字形信息给汉字编码以实现汉字输入. 之所以选用形码, 是因为生僻汉字是表意的图形文字, 见字知其形, 具有相对的稳定性. 因此, 即使读不准音不知其义, 同样能够按其构形给出编码输入计算机. 笔划码由字母 1、字母 2、数字、字母 3 等 4 部分组成. 字母 1、字母 2、字母 3 分别是书写汉字的第一笔、第二笔及最后一笔笔划代码, 共分 6 种情况. 它们分别为横(h) 、竖(s) 、撇(p) 、点(d) 、捺(n) 、折(z) . 数字是指汉字的总笔划数.

2. 4. 2　*建立汉字代码与汉字内码的关联*　组织码表是制作输入法的核心工作之一. 汉字的输入实质上是以汉字外码位关键词对码表文件进行动态检索的过程. 汉字码表即编码字典是实现汉字输入不可缺少的数据文件, 它规定了汉字编码与汉字字词间的映射关系, 并按照特定的顺序进行排列[6]. 各种输入法所需的码表文件可从某种中文操作系统获得. 码表文件由 3 部分

组成,即输入规则字段、输入规则说明字段、汉字编码文本段. 输入规则字段和输入规则说明字段,可以根据系统的规定来加入,这里不再阐述. 汉字编码格式为

[汉字编码][中文词1][中文词2][中文词3]……[中文词n]

码表文件是包含目标输入法全部的特征信息和编码规则信息,是对目标输入法的完整描述. 用户只要提供标准格式的输入法码表原文件,就能够设计出 Windows 特性和功能的中文输入法,以便利于用户输入生僻汉字.

2.4.3 *生僻汉字的使用* 只要启动生僻汉字输入法后,就可以进行生僻汉字的输入、编辑、显示、打印等等.

3 结束语

本文所介绍的设计方案,在对生僻汉字进行分析统计的基础上,对它们进行编码、造字. 同时,根据生僻汉字的特点,设计了易学易用的输入法. 从而,实现了古籍数字化中遇到的生僻汉字的存储、检索和显示,较好的解决了古籍电子化中遇到的生僻汉字问题. 它不但对汉字库的研究具有一定的推动作用,同时也可以应用到其他少数民族语言的字体库及古文字库的建库方案上. 这对于拯救历史古迹文化、古文字及少数民族文字的研究,都具有一定的参考价值.

参 考 文 献

1 李肇翔. 古籍电子化应用之我见[J]. 古籍整理出版情况简报,2002,(11):2～4
2 郑恩培,陆汝占. 汉语词典编纂计算机化的若干问题[J]. 语言文字应用,1999,(2):93～94
3 柳长青,马希荣. 西夏字与汉字共存方案的实现[J]. 宁夏大学学报(自然科学版),2001,22(1):45～47
4 刘新贵,阚映红. 数字制图中生僻汉字信息处理[J]. 东北测绘,2000,23(3):6～7
5 吕肖庆. 超大字库输入法的实现[J]. 微型电脑应用,2001,17(2):44～47

Processing of Rarely Used Chinese Characters in the Electronization of Ancient Books

Yan Fanlei Lin Zhongxiang Li Long
(Dept. of Chinese Language & Literature, Guangxi Univ., 530004, Nanning, China)

Abstract In the electronization of ancient book, an obstacle lies in lacking enough characters in lexicon; another problem is the processing of rarely used Chinese character which is studied in-depth in this paper. Starting from analysis and statistics of rarely used Chinese characters, the authors let them be passed through coding and coinage as well as inputting by a method of designing according to the peculiarity of rarely used Chinese characters. Such probems as storage and retrieval and display in their processing are settled preferably. The plan saggeseing here is of reference value for studying the electronization of ancient Chinese books.

Keywords Chinese character, lexicon, ancient book, UNICODE

本期英文审校 李 坚