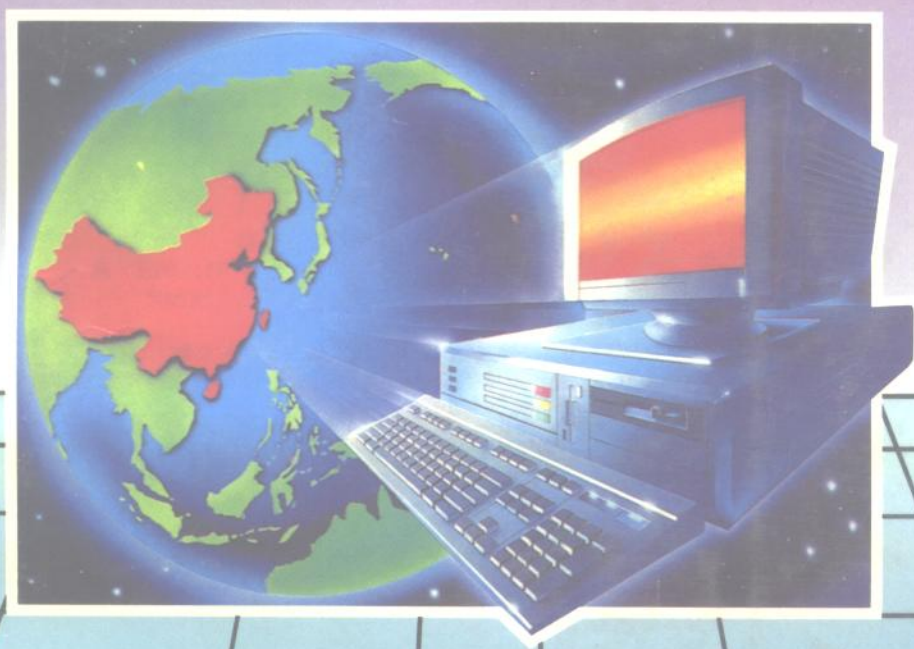


李亦何 编著

实用汉字操作系统大全



上海交通大学出版社

实用汉字操作系统大全

李亦何 编 著

上海交通大学出版社

内 容 简 介

本书介绍计算机汉字操作系统文字处理的应用、汉字系统的基本原理及基础知识;叙述了汉字操作系统的发展,以及目前国内较为优秀的汉字操作系统,如 CCDOS 2.13 系列和 SPDOS 等;还详细分析了“字符式”汉字系统的原理,首次公开了制表符智能识别的最佳算法;阐述了软件的汉化技术,尤其是在 Windows 下运行的应用程序的汉化技术;评析了流行的汉字输入法,介绍了新一代的汉字输入系统——万能码的数据结构和关键程序,为开发新一代的汉字操作系统提供了一些思路。全面介绍了目前国内应用面最广泛的几种桌面系统、轻印刷系统以及文字排版系统,如 WPS、《前景》、《科印》、《方正》和《华光》排版系统。本书可作为文字处理技术的研制、开发和使用部门的技术人员参考手册,可作为大中专院校和各类电脑操作培训班的培训教材,可供非计算机专业人员自学阅读参考,还可作专业排版人员的操作手册。

JS322/19

(沪)新登字 205 号

实用汉字操作系统大全

出版:上海交通大学出版社

(上海市华山路 1954 号 邮政编码:200030)

发行:新华书店上海发行所

印刷:立信常熟印刷联营厂

开本:787×1092(毫米)1/16

印张:24.25 字数:630000

版次:1995 年 10 月 第 1 版

印次:1995 年 10 月 第 1 次

印数:1-6000

ISBN 7-313-01448-1/TP·266 定价:23.90 元

前 言

电子计算机的发明,对于人类的科学文化生活具有划时代的意义。当代社会是充满信息的社会,由于信息量的日趋庞大,信息结构也越来越复杂,用计算机实现对信息的处理和管理,已势在必行。但是,计算机诞生在西方,其软件设备、操作系统和大量丰富的软件主要适用于西文的处理。目前外国公司虽然也在研制汉字系统,但汉字系统是我国的强项,不可能也不应该由外国公司来包办代替。西文拼音字的字符少、结构简单,处理起来较方便。而在我国所使用的文字是汉字,计算机所处理的信息大部分是汉字信息。由于汉字的数量很多,字形结构也较复杂,因此,用计算机处理汉字信息就受到一定的程度的限制。为使我国科学技术迅速发展,早日实现现代化,计算机的推广应用是十分必要的。由此,研究、推广和开发汉字系统处理技术,有重大的现实意义。

近年来,国内外的技术人员在计算机使用汉字方面作了大量的开发工作,克服了许多技术难点,推出了不少适应处理汉字的硬件设备和汉字操作系统以及许多的汉字应用软件,为我国计算机应用技术提供了良好的应用环境和开发平台。因此,在各行各业中,计算机的应用范围日益扩大,譬如书刊、文字处理、编辑排版系统、情报资料管理系统、宾馆、医疗卫生、金融、企业的信息管理。计算机的推广应用不仅是计算机事业发展的目的,也是科学技术水平发展的标志。

本书由浅入深地向读者介绍了汉字操作系统的原理,并向初学者提供了国内目前应用最为广泛的几种汉字系统的使用、操作等方面的知识。例如:CCDOS、2.13系列、SPDOS的使用以及常用排版软件《方正》、《科印》、《前景》的具体操作方法。对流行的汉字输入方法则实事求是地介绍了它们的优点、缺点和适用范围。另外,本书还首次公开了一些十分重要的在国内外处于领先地位的中文处理新技术的细节,如字符型汉字系统的原理、制表符的智能识别、通用中文词汇的数据结构及其关键程序,为计算机厂商和读者完善流行的汉字系统、进行软件的二次开发提供了一些思路 and 技巧。

在编写本书的过程中,张雯小姐和周宪先生给予了许多热情的帮助,在此谨表示衷心的感谢。由于作者才学浅陋,书中难免出现一些错误和不妥之处。敬请广大读者批评指教。

编者

1995年1月于上海

目 录

第一章 汉字操作系统概述	(1)
1.1 引言	(1)
1.2 汉字信息处理的有关问题	(1)
1.3 汉字处理系统的类型	(3)
1.4 汉字操作系统的发展	(5)
第二章 汉字操作系统基本原理	(7)
2.1 中西文软件的运行环境	(7)
2.2 汉字操作系统的结构	(7)
2.3 CC-BIOS 基本输入输出功能	(16)
第三章 2.13 系列汉字处理系统	(27)
3.1 2.13 汉字系统的特点	(27)
3.2 2.13 汉字系统的安装与启动	(28)
3.3 2.13 系统的应用	(41)
3.4 汉字的打印输出	(54)
3.5 2.13 系统造字功能	(67)
第四章 常用汉字系统操作指要	(70)
4.1 汉字系统的系统配置	(70)
4.2 常用汉字系统输入汉字的方法	(74)
第五章 WPS 文字处理系统	(77)
5.1 WPS 介绍	(77)
5.2 WPS 的使用方法	(81)
5.3 命令菜单的使用	(84)
5.4 编辑文本	(85)
5.5 文件操作	(90)
5.6 块操作	(95)
5.7 查找与替换文本	(99)
5.8 设置打印控制符	(102)
5.9 文本编辑格式化及其制表功能	(110)
5.10 模拟显示及输出	(115)
5.11 文件服务功能	(118)
第六章 《前景》排版系统	(120)
6.1 《前景》系统特点及功能	(120)
6.2 排版编辑功能	(125)
6.3 基本操作指令	(128)

6.4 报版预处理系统	(136)
6.5 书版预处理系统	(140)
6.6 《前景》表格预处理系统	(149)
6.7 图像扫描、造字、补字系统 PIWP	(154)
第七章 《科印》排版系统	(163)
7.1 系统构成	(163)
7.2 排版软件的主要功能	(164)
7.3 排版功能	(166)
7.4 排版操作	(168)
7.5 文艺理论版版式命令	(172)
7.6 数字排版命令	(196)
第八章 《方正》、《华光》排版系统	(210)
8.1 系统概述	(210)
8.2 系统软件的安装	(216)
8.3 小样的录入与编辑	(217)
8.4 交互式报版软件的使用	(220)
8.5 书版排版系统	(250)
8.6 科技排版系统	(277)
8.7 表格排版系统	(292)
8.8 图片、照片编辑系统	(297)
8.9 造字补字系统	(302)
第九章 软件的汉化技术	(306)
9.1 汉化软件的概述	(306)
9.2 软件汉化的步骤与方法	(310)
9.3 目标程序的跟踪技巧	(316)
9.4 屏幕模块显示汉字化	(319)
9.5 输入词法逻辑的汉化	(330)
9.6 屏幕显示重新布局	(333)
9.7 人机交互信息的汉化	(335)
9.8 中文 Windows 应用软件的汉化	(338)
第十章 字符式汉字操作系统	(352)
10.1 系统概述	(352)
10.2 字符式汉字系统的设计思想	(352)
10.3 字符式汉字系统的原理	(354)
10.4 显示模块的程序设计	(355)
10.5 扩展 ASCII 码字符的识别	(361)
第十一章 汉字键盘输入技术	(364)
11.1 汉字键盘输入方法评析	(364)
11.2 新一代汉字输入法——万能码(通用中文词库)	(376)
参考文献	(379)

第一章 汉字操作系统概述

1.1 引言

随着计算机系统功能的不断提高及应用领域的迅速发展,汉字信息处理的概念、涵义、作用和涉及的范围也大大扩展了。汉字操作系统属汉字信息处理技术的一个分支,从后来的基本平台发展到包括汉字的输入输出、各类办公自动化、文字处理以及书刊报纸的排版系统。实际上汉字处理技术正逐渐渗透到人们的生产和生活之中,科研、生产实践以至人类社会的一切活动都离不开它。以计算机为技术基础的汉字处理技术,正在促使人类的社会经济、科学技术和家庭生活发生日新月异的变革。

汉字是一种表意文字或像形文字,其字量多,字型复杂。在构成的汉字操作系统中,除了硬件设备尚须在底层配置相应的接口软件,使得计算机不但能处理西文,并且也能处理中文。通过科技工作者多年的努力,先后已推出不少优秀的汉字操作系统。为我国计算机应用的普及和推广作出了卓越的贡献。80年代初,就在PC以及兼容机上开发出CC-DOS以来,通过不断地优化、移植,逐渐发展成为兼容性强,更符合人们使用习惯的良好汉字平台。在此平台上软件开发人员又开发出了许多文字处理、桌面办公以及书刊排版系统。

在研制汉字系统过程中,随着汉字处理技术的理论不断地发展,汉字平台的显示技术、输入编码技术以及字型处理技术日趋成熟,不断有新的技术突破。譬如,显示技术包括兼容各类彩色显示器,支持彩色汉字显示,图文并茂等。输入技术包括近千种输入方法、手写识别、语音输入等。字型技术包括字库的压缩还原、点阵字库的平滑补偿、矢量字库以及曲线字库的算法等。实际上,汉字操作系统是一项综合性的技术,涉及许多方面基础理论,这些技术的发展,使我国的汉字处理发生了革命性的变化。我国的经济建设已离不开计算机汉字操作系统及汉字处理系统,这为我国计算机应用技术赶超世界先进水平打下了良好的基础。

1.2 汉字信息处理的有关问题

汉字的特点是字量大,字形复杂,因此要建立一个汉字系统,首先要解决汉字的输入、存储和输出问题。因此一个汉字信息处理系统需解决以下几个问题。

1. 汉字输入的编码问题

虽然可用许多种方法实现汉字编码,但要得到一种功能上最佳,并且适用面很广的汉字编码方法却非易事。

目前,汉字输入编码可粗略的分成三大类。一类称为整字编码法,实质上就是把汉字按某种规则排定先后次序,用其序号作为汉字代码。另一类是组合编码法,按照所采用的具体方法不同,组合编码可以分成许多种,例如采用字形特征编码的为形码;按照汉字发音特征编码的为音码;有形音结合的编码,也有采纳汉字其他特征的编码。还有一类就是目前国内外正研制

的手写识别或语音识别输入,这也许是一个方向,将来会突破汉字输入这一瓶颈的。

汉字整字编码输入方法的优点是直观,操作者容易学习和掌握,罕有重码问题。但它所用的键盘是通称的整字大键盘,其体积大、造价高、输入速度低。对于整字汉字键盘,除了盘内字以外,还要解决盘外字的输入问题。目前多数用户直接送入汉字代码,或者用汉字字根组合编码的方法实现外字的输入。大键盘只能在少数场合使用,无法普及推广。

当前,汉字编码输入方法很多,研究的汉字输入方案有几百种,实用的有拼音输入法,五笔字型输入法,区位码法,首尾码输入法,快速输入法以及自然码输入法等。由于用不同的输入法得出的汉字键盘码差别很大,故为便于不同的汉字信息处理系统相互交换汉字代码,需要确定一个统一的码制,称为标准汉字交换码。我国已颁布的作为国家标准的 GB2313《信息交换用汉字编码字符集——基本集》,就是一种标准交换码,在该基本集中,收集的一级常用汉字有 3755 个,二级常用汉字有 3008 个,共计 6763 个。除了有二级汉字外,还有几种外文字母、数字和符号,其总数为 7445 个。在国标交换码中,每个汉字、字母、数字和符号均用两个字节来表示。

2. 汉字字模的存储问题

计算机信息处理用的汉字字模,按用途可分精密型字模和通用型字模两种。精密型字模用于编辑排版系统,这种汉字字模,对字形,字体字号变倍等都有严格要求,必须适合出版行业所定的规格。通用型汉字字模适用于一般的汉字信息处理系统,应用面广。对于通用型的汉字字模,常用的点阵结构有以下几种:

简易型:15×16 点

精密型:256×256 矢量字库

普通型:24×24 点

精美型:TrueType 二次曲线字库

提高型:32×32 点,48×48 点

高精型:PostScript 三次曲线字库

在计算机的汉字信息处理系统中,存储汉字字模的存储体系叫汉字字模库或汉字字模发生器。其存储形式目前有三种,一种是存储整字字模信息的汉字字模库,这种字模库的特点是单位存储量的成本低,读取信息的速度快。第二种是存储压缩信息的汉字字模库,这种字模库采用的是信息压缩存储技术,其压缩原理可采用只收汉字字根的汉字字模库;利用向量组字法产生汉字字模库;用霍夫曼树型压缩法存储汉字信息等,目前用得较多的方法是矢量法和字根法两种。

矢量型或曲线型字库均是存储汉字的坐标信息以及提示信息再结合压缩技术,其存储率相当高,但还原速度没有点阵快,一般用于高精度的打印输出。

3. 汉字输出问题

汉字输出主要包括汉字显示和汉字打印。由于汉字字模的点阵密对显示器和打印机的技术要求比显示和打印西文的同类设备要高。因此,汉字的输出问题主要是汉字的显示技术和汉字的印刷打印技术。目前汉字显示用的显示器有 600×200 点阵发展到 640×480 点阵、800×600 点阵、1024×768 点阵等。汉字打印机也从原来的 9×9 点阵打印机发展到 16×16 点阵打印机、24×24 点阵打印机以及 300DPI、600DPI 甚至 1200DPI 的激光打印机,这些打印机所打印的汉字已经同印刷体所印的汉字没什么区别了。在汉字信息处理系统中,用户可以根据要求和不同的需要,选用合适的显示设备和打印输出设备。

4. 汉字终端技术

汉字终端技术是汉字信息处理中的重要设备,它同普通的显示设备不一样,这种终端具有

通信功能,采用标准的 RS232 串行接口,可以进行远程控制。目前汉字终端有两种,简易型汉字终端和智能型汉字终端。前者具有汉字的输入输出功能、汉字代码转换功能和汉字编辑功能,自身还带有汉字字模库;后者除了以上功能外,还有文件管理功能、表格处理功能和图形处理功能等。这种终端可以脱机使用,依靠本身的软件资源作为独立的工作站。

5. 汉字/西文兼容问题

计算机只能处理 0 和 1 两种信号,因此汉字信息处理也必须先把汉字量化,即数码化,使其转变成计算机能直接处理的信息。然而,计算机是首先在欧美等西文国家得到发展,因而西文字母先为计算机所接受,接着又用西文来替代机器码,发展成为汇编语言,进而又发展成各种接近自然语言的高级程序设计语言。这些长期积累的经验和技术都是针对西文系统,并用西文表达的,对于汉字信息处理系统的主要内容是汉字。在发展汉字信息处理时首先碰到的是实现中文信息的处理,特别是使用西文的高级程序设计语言也能适应处理汉字信息,目前通过软件手段已能顺利地解决汉字和西文的兼容问题,使实用的汉字信息处理系统既能处理汉字信息,又能处理西文信息。如目前流行的有 CCDOS 系列、Super DOS 系列、2.13 系列以及各类“字符式”汉字操作系统,如 UC DOS、天汇、高通汉卡等。

1.3 汉字处理系统的类型

1.3.1 汉字信息处理系统的构成

汉字信息处理系统的组成包括硬件和软件两大部分。

1. 硬件组成

汉字信息处理系统的硬件包括主处理机,常规外部设备和汉字外部设备。主处理机是通用的电子计算机。根据实际需要选择合适的主机类型。常规外部设备主要包括外存储器,例如磁盘机和磁带机。汉字外部设备包括汉字输入键盘,汉字打印机,汉字显示终端设备等。

在汉字信息处理系统中,汉字字模库和汉字显示终端是重要的组成部分。这些设备作为一个独立的外部设备,这样可以把字模库能多台设备所共享,特别是当汉字字模库价格较高时,这种设置方法有利于降低系统的造价,缺点是系统效率较低。第二种是字模库直接同打印机相连,这种系统只传送代码信息,可以提高系统的效率。第三种是字模设置在汉字显示终端上,目前大部分的汉字终端均配有字模库。

因此对于一个汉字信息处理系统,所选配的主机类型,字模库设置方式,汉字设备的要求,也决定了该系统的处理性能、工作方式以及成本价格。

2. 软件组成

汉字信息处理系统的软件包括系统软件和应用软件两大类。

(1) 系统软件

系统软件包括能兼容汉字和西文信息处理的操作系统。该操作系统除了保留通常西文系统全部功能的条件外,还包括各种汉字设备的驱动模块,能直接调用汉字输入输出管理程序以及汉字编辑程序。汉字输入输出管理程序也是系统软件的一部分,它包括汉字输入输出接口程序,以及汉字输入、换码,访问汉字字模库和汉字输出等程序;对于输入计算机的代码信息,当系统识别出它是汉字信息时,先把标准码变成汉字字模库中对应的地址码,然后读出字模信

息,以供显示或打印。除此以外,能直接处理汉字信息的各种高级程序设计语言都属于系统软件范围。

(2) 应用软件

应用软件的范围很广,可根据汉字信息处理系统的性质和用途,对各种西文系统的应用软件进行汉化,使其能直接处理汉字信息,这样便就成了汉字信息处理系统的应用软件了。例如,西文 Wordstar 文本编辑软件,经过汉化处理后,就成为汉字 Wordstar 的文本编辑软件。因此除了用户自己设计各种应用程序外,应尽量利用西文系统已有的一些应用程序包,这样便可迅速扩大汉字信息处理的应用范围。

1.3.2 汉字信息处理系统的分类

从汉字信息处理系统输出汉字文字的质量要求来看,目前可把汉字信息系统分成三大类:即精密型汉字编辑排版系统、通用型汉字处理系统以及汉字操作系统平台。第一种类型用于正式出版的书籍、报刊、报纸的编辑排版。第二种用于一般的汉字文件处理,统计报表以及文书处理等。第三种就是底层汉字操作系统平台,用于应用软件、管理软件和多用户网络系统中。汉字信息处理技术的推广应用取决于系统平台的发展。

1. 精密汉字编辑排版系统

这一系统最重要的关键技术就是高精度汉字字模和版面输出。对于精密汉字字模,不仅每个字的点阵信息量大,而且由于字量多,需要多种字体和字号,从而使总的字模信息很庞大,同时还要兼顾到适当的字模输出速度。这类系统的输出技术大部分采用激光扫描输出,在扫描的同时,输出版面在感光底片上记录成像,因而这种设备又称为照排设备。整个系统包括:排版用计算机,相应的外部设备,编辑、改错用的联机汉字显示终端,汉字数据采集用的汉字终端,校样印刷机,照排设备,字模自动制作设备,以及图片输入设备等。目前这类系统有:华光Ⅳ型书出版系统、北大方正系统、大屏幕报刊组版系统以及前景、科印书刊排版系统等。

2. 通用型汉字信息处理系统

这类系统亦称为桌面办公系统,其特点是作为汉字信息处理,其使用面广,系统的成本低,不需太讲究汉字字模的质量。通用型汉字系统的字模点阵目前常用的有 24×24 点; 32×32 点以及矢量字库等。汉字的输出以普通的 24×24 点阵针式打印机为主,有些系统也支持激光/喷墨打印机输出。目前这类系统主要用于现代办公自动化、家庭电脑等。现在市场上流行的有金山汉字(WPS)处理系统、巨人、瑞星等字处理系统。这类系统除了保持低价实用的优良性能,目前也在汉字输出质量方面做了大量的研究工作。汉字输出也可配置照排设备,以此同精密汉字编辑排版系统争夺市场,努力推广汉字信息处理系统在四化建设中的普及应用。

3. 汉字操作系统平台

这类系统的特点是兼容性强,占用系统资源少,在原西文系统下建立了一台汉字平台,通常用于数据处理、信息管理以及网络系统。譬如:CCDOS4.0、“字符式”汉字系统以及 P-Windows 图形汉字平台等。汉字平台的兼容性及其功能在不断优化和发展,国内已发展到第三代产品,并广泛地应用到各个领域。随着计算机硬件技术的迅速升级,汉字操作系统平台的性能也将愈来愈优越。

本书目的是系统介绍我国市场上流行的几种常用的汉字信息处理系统,促使我国各行各业的数据处理、办公文字处理、日常报表处理以及精密照排系统等工作早日实现计算机化。

1.4 汉字操作系统的发展

从1984年最早的CCDOS2.0开始到现在,中文系统的技术不管是显示、输入、打印都有了很大的发展。这里,我们把现在市场上已有的汉字操作系统分为三代。

第一代以全部修改BIOS为模式,DOS中涉及汉字输入、显示、打印部分的中断调用为基本技术,CCDOS2.0是其代表。在实现方法上,这一代汉字系统一般是全面修改INT5,INT10,INT17等。BIOS,DOS中断调用,使这些BIOS,DOS中断调用可以进行汉字信息的输入、显示和打印输出处理,限于当时的PC硬件环境和条件,CCDOS2.0仅仅支持11行中文显示。因此,几乎大部分实用的西文软件都要进行汉化处理,以适应11行的汉字显示支持。虽然有人开发出16行显示及单色25行显示的支持系统,但未形成大的市场。由于这些汉字系统BIOS,DOS中断调用的修改依靠具体DOS版本,也使得这时的汉字系统不能随DOS版本的升级而同时提高。

第二代是以修改BIOS,DOS中与中文显示、打印有关的部分为标志CCDOS4.0、2.13H、SPDOS、UCDOS 2.0等为其实现方法上对西文DOS的破坏较小,使得大部分西文软件只要经过简单的汉化就能在系统中正常运行。的研究,使得汉字系统不仅只在汉字显示上,还在中文打印输出,汉字输入上都有了长进。这些系统最主要的特点和第一代汉化系统的区别在于把显示、输入、打印均以模块化的方式提供给用户,使得用户根据不同的硬件配置和系统环境来选择安装,具有一定的灵活性。

这些进步使得第二代中文系统真正进入了实用阶段。从时间上讲正是第二代实用汉字系统的产生,才使得汉字操作系统形成一个独立的汉字系统市场。

第三代是以支持各种软件的直接写屏为标志,并且系统的模块结构更具优化。智能识别能力增强。如UCDOS 3.0~3.1、天汇、高通、倚天、新致等产品是较好的代表。第三代中文系统是近几年才出现的。由于这些产品完全支持大部分西文软件,同时又适合大多数流行的DOS版本,目前正广泛地在各行各业使用。

目前,中文系统的技术发展主要体现在下列方面:

- (1)与西文软件的最大兼容性。
- (2)无须特殊记忆,基于拼音的随意中文输入法。
- (3)英文制表符正确识别程度。
- (4)占用最少的PC资源,特别是内存使用的多样性。
- (5)中英文字型显示速度。
- (6)各种打印字体的设置。

这些技术是当前主要中文系统发展厂商争夺的焦点,也是未来几个月甚至几年内中文系统发展的驱动力。

那末,这些技术是否就是代表了汉字系统未来的发展方向,主宰汉字系统的技术呢?根据PC机软、硬件总体发展趋势,我们认为“开放式汉字系统”将是第四代汉字操作系统的主要特征。

“开放式汉字系统”就是系统将本身的系统资源(包括API)接口,大部分或全部开放给用户,供用户在实际使用中调用或按用户自己的需求修改,设置新的系统资源。同时,它也将全面

吸收当代各代中文系统的优点。

由于,目前我国使用的 PC 机和 CPU、RAM 容量、显示器、打印机种类繁多,PC 机装载的 DOS 操作系统版本多样,用户使用的应用软件层出不穷,应用软件的编制水平参差不齐,因此很难发挥用户 PC 机的最大效率,甚至有些还限制了一些应用软件的使用。

第四代以前汉字系统的封闭结构使用户不但要编写打印机的驱动程序,更要获得、掌握各种矢量字库(点阵字库)和它们的结构,不用说撰写字库驱动程序工作量之庞大,单就合法地取得这些字库的版权(使用权)也是一笔新的投资。

而现在,用户可以完全利用第四代汉字系统特点的一些表现。它的这一基本特性——开放性。使得这一代汉字系统适应性广、随意性强,不但能在 PC 机各种硬、软件条件下使用,更能让用户充分享受中文系统本身资源,随意修改或加入新的信息资源。可以毫不夸张地说,用户甚至可以编写出自己的特殊文种,特殊字型的系统。

第四代汉字操作系统从根本上改变以前的汉字系统封闭结构,能让用户改变自己 PC 机的软、硬件结构,以适应汉字系统的运行环境,或改变汉字系统软配置来适应用户的 PC 及兼容机,这使得用户的投资充分得到长期的使用和保护。

由于开放式汉字系统呈现出来的优良品质,使我们完全有理由认为,新一代汉字系统将是开放式汉字系统的天下。因此,汉字系统技术的发展在未来一段时间内将由开放式主宰,并形成汉字系统的新一代标准。谁最快掌握这一技术,谁就可能在巨大的汉字系统市场上取得主动权。

第二章 汉字操作系统基本原理

2.1 中西文软件的运行环境

由于计算机原有的西文操作系统不能直接处理汉字,这就需要在系统的底层开发适应汉字处理的接口。一般在汉字操作系统的环境中运行,汉字信息处理的结果和界面都是通过输入输出设备表现出来的。因此,了解和掌握计算机 I/O 接口的技术,对于使用汉字操作系统及其应用软件是很有必要的。如果对指令系统、汉字系统的功能调用和有关方法都有所了解的话,要想分析和移植某一指定的软件就比较容易了。

譬如,程序中与汉字信息处理有关的是键盘输入、屏幕显示、打印输出和信息通信等,纯西文系统与汉字系统在以上几个方面的处理过程是不尽相同的。我们所熟悉的 CCDOS 汉字操作系统则是在原计算机内基本输入输出系统(BIOS)的基础上开发的。实际上应该称为 CC-BIOS,是系统一级的环境软件,它为我国计算机的普及和使用打下了坚实的基础。CC-BIOS 对原有 ROM-BIOS 中的几个主要服务模块的功能进行了适当的扩充,诸如显示功能(INT 10H)、键盘功能(INT 16H)、打印功能(INT 17H)和屏幕拷贝(INT 5)等功能,都是以软中断的方式提供给调用者,使得调用程序在此环境下具备了处理汉字的条件。

有了上述条件,只能说 CC-BIOS 初步建立了一个汉字系统的基本运行环境,并不是所有的软件都可以处理汉字信息了。如果要想有效地处理汉字信息,则必须使得那些在 CC-BIOS 环境中运行的软件,毫无阻碍地调用 CC-BIOS 中那些扩充过的服务功能,才能确保正确有效地输入和输出汉字信息。

2.2 汉字操作系统的结构

2.2.1 系统层次与功能

CC-DOS 的设计目的是解决使 IBM 系列微机具有汉字信息处理环境的问题。汉字信息处理系统(Chinese Information Processing System)的关键在于计算机对汉字代码的数据处理使得人们在计算机上使用汉字和使用西文一样的方便和容易。在计算机系统内部,西文和中文只不过被人为地定义了不同的符号,用不同的信息表示而已。

从信息处理的角度来看,汉字信息处理和西文信息的处理并没有本质上的区别,两者均是非数值处理。这要求计算机系统具有特定的资源,并得到信息共享、数据库、数据通信和高级语言等功能的支持。由此,在 CC-DOS 的环境中就可以开发各类中西文兼容的软件,系统级的各类中西文兼容的软件。系统的各个层次如图 2-1 所示。

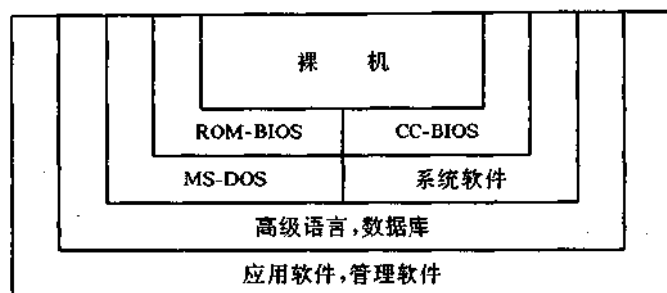


图 2-1 系统的层次

CC-BIOS 包括以下几个主要的模块和数据:

- (1)工作区和数据区;
- (2)显示器控制程序;
- (3)ASCII 字符字模数据;
- (4)键盘输入控制程序;
- (5)汉字输入码与内码对照表;
- (6)汉字字模数据。

汉字系统的结构是比较清晰的。汉字打印驱动程序是独立的模块,有些系统是根据不同的打印机型号装入不同的汉字打印驱动程序的。

工作区和数据区供系统程序运行中数据加工、传送和设置标志时使用。例如,汉字点阵的加工处理,外码和内码之间的转换等等。标志单元一般用于控制程序的走向,使程序完成预定的运行目标。

显示器控制程序的主要作用是:通过汉字内码,产生相应的汉字字型,根据 CRT 初始化后不同的显示器类型(如 CGA、EGA、VGA 等)和模式,通过对汉字字模点阵的检索,以图形方式在屏幕上指定的位置显示汉字的点阵信息。

字符数据分为 ASCII 码字模和汉字字模库。一个标准的 ASCII 码由 8×8 或 8×14 (EGA、VGA)点阵组成,由点位信息表示。常驻内存的汉字字模,每个汉字由 16×16 点阵 32 个字节表示,该字模库主要是提供给显示输出用的。

汉字输入码——机内码对照表,亦称为汉字扫描表,通常是指某种输入编码方案(如拼音法,五笔字形等)的码表。输入码与码表之间检索和比较的结果产生相应的机内码,尔后,根据这个机内码来定位该汉字点阵字模的首地址。

键盘输入模块可以根据功能键的切换分别同时或输入中文或西文信息,能把用户输入的外码进行有机地变换去检索码表,以达到汉字输入的目的。

汉字打印驱动程序主要解决汉字信息的打印输出。由不同的机内码检索到相应的汉字点阵字模,并在针式打印机或激光打印机上,根据不同的打印控制命令打印出大小不同、字型相异的汉字信息。

整个汉字系统的功能结构如图 2-2 所示。

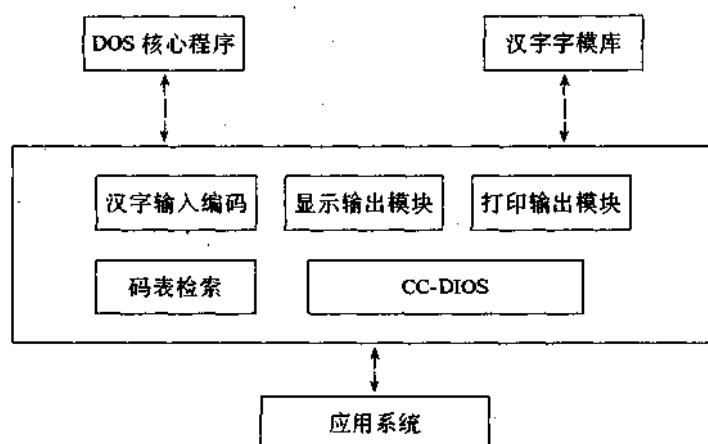


图 2-2 汉字系统功能结构

2.2.2 CC-BIOS 代码系统

作为一个完善的汉字操作系统,必须具备输入汉字、处理汉字和输出汉字三大功能。就汉字系统而言,无论是西文信息还是中文信息,其信息均是以代码的形式流通和传送的。在 CC-BIOS 汉字系统中,为了使各个功能模块处理信息更为方便,汉字的代码在各模块中的定义和名称均不相同,从而在汉字信息处理系统中存在着好几种汉字代码,这些代码构成了一个汉字代码体系。为了适应汉字系统内各组成部分对汉字信息进行处理的不同要求,在 CC-BIOS 内部各个功能模块之间,把这些代码进行了适当的转换和处理。如图 2-3 中的代码交流过程所示。

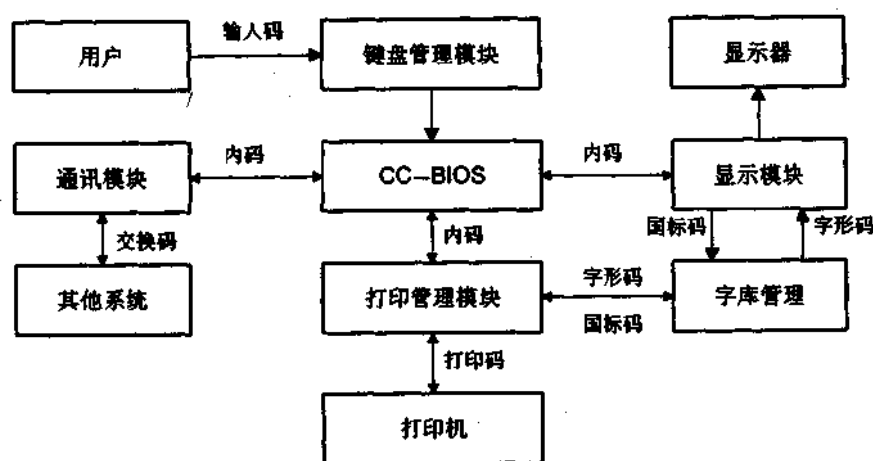


图 2-3 系统代码交换流程

1. 汉字输入码

汉字输入码是为了方便计算机用户把汉字信息输入到计算机内而定义的代码。这种代码必须通过操作员利用键盘等输入设备进行输入。因此说,输入码是位于人机界面之间的代码。相对计算机内部来说,又把输入码称为外码,所采用的输入外码方式,根据编码方案的不同而

各有差异。例如,拼音码是直接拼字母或数字作为外部编码,根据汉字的发音进行汉字的输入。如区位码、电报码等。现阶段已发明了近千种汉字输入法,如纯拼、首尾、大众、五笔、电报及钱码等等。众多的输入方法中,用户在输入一个汉字时需要敲键的个数及排列组合是不同的,一般需要2—4键。由于用户的层次不同,用途有别,故一个汉字系统中常配有数种汉字输入方式,以使其能广泛地适应各种用户。

2. 汉字机内码

汉字机内码(亦称内码)是计算机内部对汉字信息进行各种加工、处理的代码,不同的汉字系统,将选用不同的机内码。在一般情况下,不同输入方式,其汉字输入码的长度不等。要使计算机又能够接收,又便于存贮和传送,把多字节不等长的输入码按照一定算法变换为与国际码有一定对应关系的代码是十分必要的。CC-BIOS中采用的是以异形国标码作为其汉字机内码,有关内码的设计思想将在下一节中作介绍。

3. 汉字地址码

汉字地址码是指存贮汉字字模信息的逻辑地址,汉字是以点阵字节表示的数字信息,它可以存贮在计算机的只读存贮器(ROM, EPROM)中,如一些微机上配置的汉卡等,也可以存贮在磁盘上。当然,也可以装载到随机存贮器的RAM中。汉字字模库存贮的介质不同,汉字信息所在的逻辑地址也不同,因此,地址码的表示方式也就不尽相同。如果汉字装入内存RAM,则用内存地址数来表示。若汉字是在汉卡上(EPROM),则由选片控制来产生地址码;若汉字库放在磁盘上,其地址码通常由磁盘中磁盘某一扇区的数字码表示。由于CC-BIOS的汉字字模库中的汉字字形信息排列序列的规则,使得CC-BIOS的汉字地址码与汉字机内码之间形成一种简单的函数关系,所以很容易实现两者间的转换。

4. 汉字字形码

目前,方块汉字都是以点阵字节方式数字化表示。所以,汉字的字形码是指汉字字形点阵的数字代码。汉字字形大小和形状不同,其字形码亦不同。这些规格不同的字形码是根据不同的系统设置和输出要求而选定的。目前,我国已颁布了 16×16 , 24×24 , 32×32 和 48×48 点阵的字模标准。

5. 标准码

1981年,国家标准总局制定并颁布了GB2312信息交换用汉字编码字符集基本集,通称标准码或交换码。交换码是用于汉字信息处理系统之间或者与通信系统之间进行信息交换的汉字码。这要求汉字交换码必须采用统一的形式。目前,国内计算机系统所采用的标准信息处理交换码是根据有关国际标准制定的,即GB198《信息处理交换用的七位编码字符集》和相应的代码扩充标准GB2311《信息处理交换用的七位编码字符集的扩充方法》。标准中规定了汉字信息交换用的基本图形字符及其二进制编码,这些有标准的交换码,称为国标码。

国标码中共收集了7445个汉字和图形符号。其中汉字有6763个,分为两个级别:一级字库为常用汉字3755个,按汉语拼音字母顺序排列,对于同音字,再按笔划顺序排列;二级字库有3008个,按部首顺序排列。此外,收集有一般符号202个,包括间隔符号、标点符号、运算符号、单位符号和制表符号;序号有60个;数字有22个,英文字母大小写各有26个,共52个;日文假名169个,希腊字母大小写各有24个,共48个;俄文字母大小写各有33个,共66个;汉语拼音符号26个和汉字注音字母37个。

国标码的任何一个符号,汉字和图形都是用两个7位的字节来表示的。一般计算机中,每

个字节为 8 位。因此,最高位以 0 来表示。国标码有一张代码表,表中的纵坐标表示区,编号为 1~94,每个区有 94 位。表中的横位标表示位,编号也是 1~94,每个区位的焦点表示一个汉字的代码,例如汉字“啊”的区位码是 1601(16 区 01 位),其国标码为 3021H,把它展开为二进制码的话,是 00110000 和 00100001 两个字节。

2.2.3 汉字机内码的设计思想

汉字机内码是汉字系统内部处理和传送汉字使用的代码。汉字机内码的表示方式很多,有两字节法,三字节法和四字节法等。为了节省汉字代码所占的存贮空间和传送时间,CC-BIOS 汉字系统是采用两字节表示的。众所周知,西文字符的机内码均是采用一个字节表示的 ASCII 码。一般只使用了前 7 位来表示 128 个 ASCII 字符,而高位则是作为奇偶校验或不用。如图 2-4(a)所示。国标 GB2312 规定,一个汉字用两个字节表示,每个字节仅使用 7 位。

为保障汉字系统中西文兼容,在处理汉字机内码时,必须保障西文机内码(即 ASCII 码)的使用,但也要允许国标码的使用。显然,在一个系统中同时存在国标码和 ASCII,这将会产生二义性。因此,在汉字国标码的两个字节最高位分别置 1(如图 2-4(b)所示),用此方法来认识汉字,作为汉字的机内码。这种方法实际是变了形的同标码,称之为异形国标码。例如汉字“啊”的国标码为二进制的 00110000,00100001;高 8 位置 1 后所对应的机内码为二进制的 10110000,10100001,即十六进制 B0A1。这样处理,既解决了西文机内码与汉字机内码的二义性问题,又保证了汉字机内码与国标码之间非常简单的对应关系。CC-BIOS 就是采用这种类型的异形国标码作为机内码的。

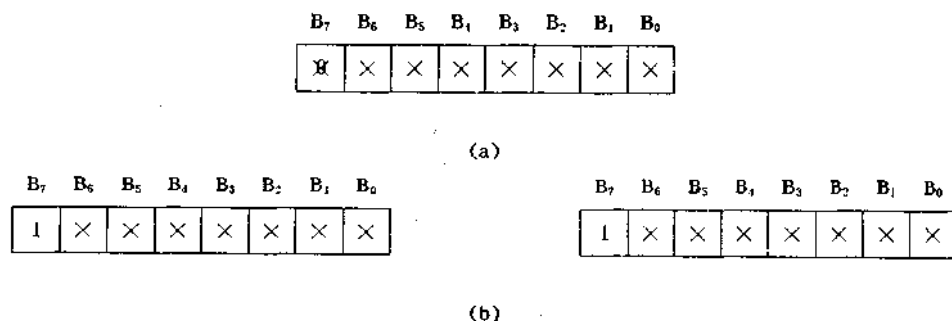


图 2-4 异形国标码的格式

(a)ASCII 的格式 (b)异形国标码的格式

无论采用哪种类型的机内码。其设计思想应满足以下几个原则:

- (1) 信息的位数,尽量减少信息的冗余度。
- (2) 与国标交换码有较简单的对应关系。
- (3) 排除汉字机内码与西文机内码间的二义性。
- (4) 便于汉字信息处理中的基本操作和运算。
- (5) 和原西文系统的兼容性要高。
- (6) 便于纳入各种程序设计语言中的字符类型中。
- (7) 尽量便于扩充字符集。