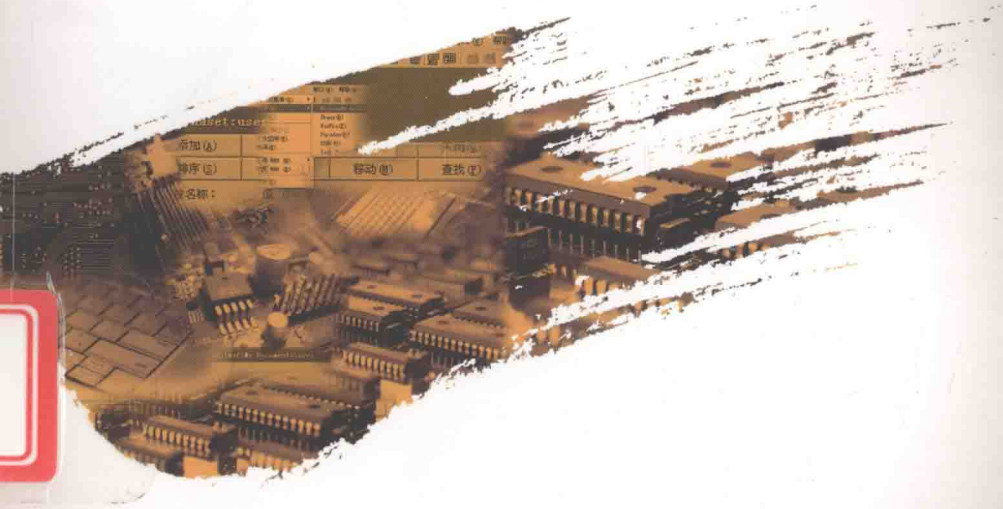
滇 | 西 | 学 | 术 | 文 | 丛

◎王善发 著

文字信息的计算机处理技术



云南大学出版社
Yunnan University Press

滇 | 西 | 学 | 术 | 文 | 丛

文字信息的计算机处理技术

◎王善发 著

云南大学出版社
Yunnan University Press

图书在版编目（CIP）数据

文字信息的计算机处理技术 / 王善发著. —昆明：云南大学出版社，2012
（滇西学术文丛）
ISBN 978-7-5482-0802-0

Ⅰ. ①文… Ⅱ. ①王… Ⅲ. ①文字处理 Ⅳ. ①TP391.1

中国版本图书馆CIP数据核字（2012）第018175号

策划编辑： 徐　曼
责任编辑： 石　可
　　徐　曼
封面设计： 刘　雨

文字信息的计算机处理技术　王善发　著

出版发行： 云南大学出版社
印　　装： 昆明研汇印刷有限责任公司
开　　本： 850mm×1168mm　1/32
印　　张： 6.375
字　　数： 170千
版　　次： 2012年4月第1版
印　　次： 2012年4月第1次印刷
书　　号： ISBN 978-7-5482-0802-0
定　　价： 18.00元

社　　址： 昆明市翠湖北路2号云南大学英华园内
邮　　编： 650091
电　　话： （0871）5033244　5031071
网　　址： http://www.ynup.com
E-mail： market@ynup.com

《滇西学术文丛》总序

蒋永文

保山学院的前身为保山师范高等专科学校，地处气候宜人、风景秀丽、历史悠久的滇西重镇保山，是一所建校已有30年，主要为拥有1 100万人口的滇西7个州市培养中小学师资的地方师范院校。长期以来，在艰苦的条件下，为该区域培养了上万名中小学教师和各行业建设者，为祖国西部边疆少数民族地区的教育发展作出了应有的贡献。2009年4月，学校被教育部批准为保山学院。这使我们站在了一个新的历史起点上，有了一个更为广阔的发展空间。

大学肩负着创造知识和传播知识的重任。学术是支撑大学的精髓，学科是构筑大学的基石，学者是大学精神的化身。教学与科研相统一是大学的基本理念。科研和教学是彼此促进的，在教学中，可以激发灵感，开阔思路，发现研究课题。而研究成果又可以丰富教学内容，促进教学质量的提高，二者相得益彰。为了给滇西地区提供更好的高等教育资源，保山学院必须建立一支热爱教育事业，业务过硬，高水平、高质量的教师队伍，为此，学校以重点学科建设为龙头，以形成科研特色，增强科研实力，提高效益为目标。学校近几年采取了资助科研立项、奖励科研成果，出版学术论文等措施，来不断提高广大教师的教学水平和科研水平，已收到了较好的效果。为更好地为广大教师提供出版学术论著的园地，学校决定继续出版《滇西学术文丛》，出版学术

水平较高的著作，相信《滇西学术文丛》的出版，一定会对保山学院科学研究的深入、学科建设和学科带头人、骨干教师的培养产生积极的影响。

辽阔的天空，允许大鹏展翅高飞，也允许小鸟上下蓬蒿。广袤的大地，允许参天大树生长，也允许无名小草成长。我们是小鸟，我们是小草，这套丛书，远非成熟完美，作者水平也还需要不断提高。我们期待着批评和指教。我们会做得越来越好。

2009 年 5 月

前　言

本书所撰写内容属于计算机科学技术的软件工程领域，任何汉字编码和汉字输入法都只是本书技术的实施对象。

钱学森在《电子计算机软件与新时期语言文字工作》一文中说："国防科工委系统工程研究所的汪成为同志提出'电子计算机软件也是语言文字工作'（见一九八六年五月十三日《光明日报》《语言文字》专刊）。这我很赞同。因为从文化发展的角度看，电子计算机的作用和影响，就同过去人类历史上语言的出现、文字的出现、造纸技术的出现以及印刷技术的出现一样，是文化建设中的一件大事。"

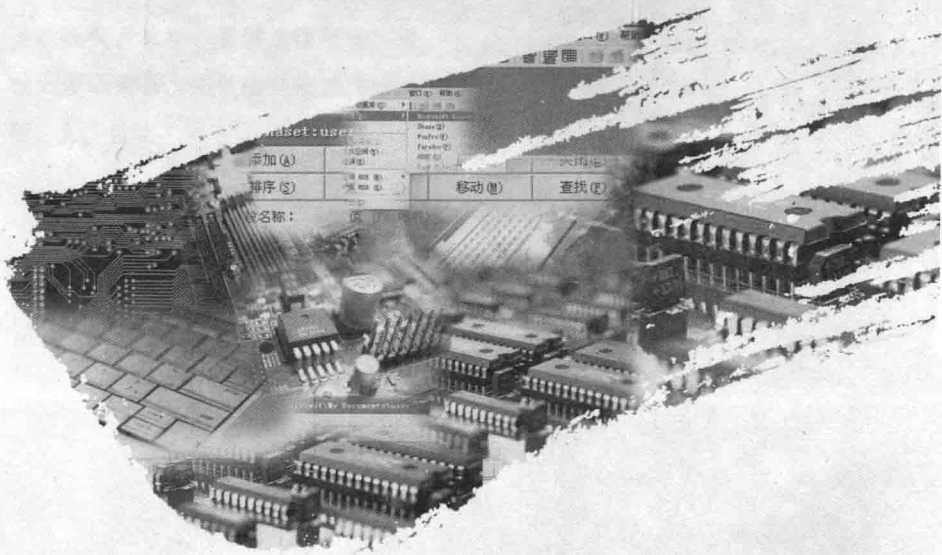
随着中国文化对世界的影响日益强大，把汉字信息输入计算机的需要越来越多，计算机的文字处理技术将是一个长期的过程。

为配合汉字编码人员做好相关编码研究和相关推广工作，作者在软件工程、软件开发平台、软件开发工具，汉字输入法的安装，汉字信息的获取、查询，当前使用汉字输入练习、测试、教程，汉字信息查询工具等领域进行研究、分析，参考很多资料上零碎的技术进行多年的艰苦研究，实现了获取 Unicode 字符集中的汉字和内码的方法，完成了"汉字输入法安装软件"、"输字学习软件"等软件的设计与实现，并将个人研究成果撰写成内容连贯，技术较全面的《文字信息的计算机处理技术》一书，便于文字信息工作者、汉字编码研究者、软件工程学习者、计算机软件开发者参考和借鉴。

王善发

2011 年 5 月

本书由“保山学院学术出版基金”资助出版



目　录

第 1 章　文字信息的计算机处理过程

本章从文字信息在计算机中的信息输入、信息加工和信息输出等几个方面讲述了文字信息的计算机处理过程。

1.1　文字信息的计算机处理

计算机处理文字的基本步骤包括信息输入、信息加工、信息输出，而文字输入、文字编辑、文本输出等步骤的前提是用 0、1 代码串表示文字符号，也就是编码问题。

以英文信息的计算机处理为例，英文字符的编码标准是 ASCII 码，即美国信息交换标准代码。这是七位的二进制代码，它是美国国家标准学会（ANSI）为计算机的信息交换提出的标准，后来由国际标准组织（ISO）确定为国际标准字符编码。为了和国际标准兼容，我国根据 ASCII 码制定了英文字符编码国家标准，即 GB1988。其中除了将货币符号置换为人民币符号外，其他都与 ASCII 码相同。

计算机的键盘原本就是为英文输入设计的，只要按照字母击键，就可以输入英文。键盘的译码电路按照所击的键产生英文字符的 ASCII 码，输入计算机的内存。

为了对输入的文字进行编辑加工，必须使用相关的应用软件，如 Word、WPS，或其他文字处理软件。经过编辑的文本仍然以 ASCII 码表示。

输出时，这些代码必须转换成字符字形的点阵，以便显示或打印。因此，计算机必须存储每个英文字符、数码以及标点符号

的点阵信息。这些点阵信息构成了“字模库”。字模库的点阵以有点或无点来表示文字和符号。文字、符号的点阵信息由显示器或打印机输出时，必须通过相应的驱动程序，将点阵信息转换为显示器、打印机的电子或机械的操作。文字信息的计算机处理过程如图 1－1 所示。

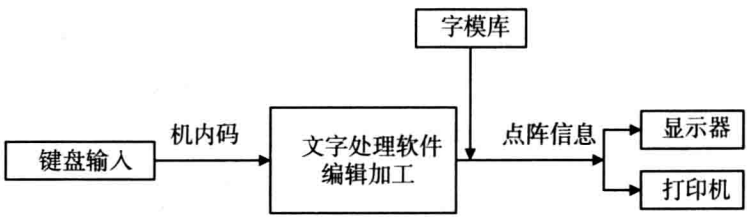


图 1－1　文字信息的计算机处理过程

1.2　文字信息的输入

要用计算机来处理文字，必须解决如何把文字输入计算机并在计算机中存储起来，即信息的输入问题。要实现信息输入首先要解决文字信息的编码问题，其次要解决如何按照编码标准将文字信息转换为计算机的机内码，将文字信息交给计算机处理的问题，即输入设备问题。

1.2.1　文字信息的编码

英文字符的编码标准是 ASCII 码，汉字的编码标准有国标码。英文字符和汉字的编码问题已经基本解决。如果要处理俄文、法文、德文、日文等其他国家的文字，也必须规定其编码的标准。

1.2.2　文字输入设备

1. 键盘输入

计算机的键盘一个键对应于一个字符或标点符号。只要按照字母击键，键盘的译码电路就会按照所击的键产生相应英文字符的 ASCII 码，并输入到计算机的内存中。其他字母文字如果其字符和英文对应，只要改变译码电路，也可以用英文键盘输入。

汉字的字符数目远远多于英文键盘按键的数目，因此要用几个键的组合来表示一个汉字。这种键的组合称为“汉字输入编码”。目前国内外提出的汉字输入编码方案不下 500 种。编码长度、规则的复杂程度、重码率等因素决定了不同编码方案的优劣。实际上流行的汉字编码输入方案只有十几种，它们分别对应于不同的输入法。

以汉字字型特征来编码的方案称“形码”。形码编码规则往往较复杂，与阅读文稿时大脑的思维习惯（读出的声音）不甚符合，要求用户熟悉汉字笔画、偏旁部首，且要经过较长时间的训练才能熟练使用。形码比较适合于以“看打”（边看文稿边输入）为主的专业录入人员。比较成功的形码有郑码、五笔字型码等。

以语音特征来编码的方案称“音码”。音码多数以汉语拼音为基础，编码规则简单，符合阅读的思维习惯，只要懂得汉语拼音，经过简单培训就可以使用，学习时间短。音码比较适合于边想边打的普通用户。音码首先实现了以词为单位、甚至以句子为单位的输入，实现了高频词优先、在线造词和词组等功能，使基于拼音的输入法做到得心应手、运用自如。比较成功的音码有微软拼音、智能 ABC、全拼输入法等。

不论哪一种输入方案，在具体实现时都要有软件的支持。汉字输入法软件按照汉字编码标准（国标码）将键盘输入的编码

转换为机内码，就可在计算机内存储和处理汉字。

汉字编码输入的研究目前还在继续。不过研究的重点已经从编码方案本身转向通过更好的软件技术和设计来做到重码少、适应面广、学习负担轻。

2. 其他输入设备

由语音转换成文字输入计算机的技术目前还不成熟。

通过光学字符阅读器（Optical Character Reader，OCR）可以将印刷体汉字作为图形点阵输入，然后进行字符识别，把汉字点阵转换成对应的机内码。这种方法已经达到实用阶段，但是设备较昂贵。

手写板输入基于计算机模式识别技术，专用的软件能够识别手写输入的文字、符号，将其转换为机内码。这种输入设备已经商品化，应用于微型机，适合不会键盘输入的用户使用。

1.3 文字信息的加工

为了对输入的文字进行编辑加工，必须使用相关的文字处理软件，如 Word、WPS 等。文字处理软件的工作主要有文本的增、删、改，字体、字号和版面布局设计等。文字信息的处理是由人与机器共同完成的。文字信息加工的结果是编辑完成的文本，它是输入的原始文本经过加工（变换）得到的。经过编辑的文本仍然以汉字机内码或者 ASCII 码表示。

1.4 文字信息的输出

输出编辑过的文本时，汉字机内码或者英文的 ASCII 码必须转换成字符字形的点阵，以便显示或打印。因此，计算机必须存储每个字符、数码以及标点符号的点阵信息。这些点阵信息构成了“字模库”。文字、符号的点阵信息由显示器或打印机输出时，还必须通过相应的驱动程序，将点阵信息转换为显示器、打

印机的电子或机械的操作。

1.4.1　文字字模库

文字输出时，不论显示或打印，都是把一个字符看成一个二维图形，并把笔画离散化，用点阵来表示。点阵的每个点位只有两种状态：有笔画上的点或无笔画上的点。这就可以用 0、1 代码来表示。取值为 1 表示“有点”，取值为 0 表示“无点”。那么，一个 0、1 代码串就可以表示点阵的一行。若干个代码串就表示整个字符的点阵信息。在具体实现时，点阵上取值为 1 就显示或打印一个“点”，否则不显示或不打印。例如，汉字“梅”就可用图 1 -2 所示的点阵图来表示。

图 1 -2　汉字的点阵表示

描述一个字符点阵信息的 0、1 代码串集合称为字符的“字模”，其作用跟铅字印刷所用的字模相当。所有汉字和各种符号的点阵信息组成汉字的“字模库”（简称字库）。显然，要实现近 8 000 个常用汉字和符号的显示和打印，字库要占用很庞大的存储空间。例如 16 × 16 点阵的汉字库（包括一级和二级汉字）就需要约 240kB 的存储空间。24 × 24 点阵的汉字库需要 580kB，精密字库所需的存储空间更大。我国两院院士王选提出“轮廓描述和参数描述”的字形压缩算法，发明了高分辨率字形的高倍率信息压缩技术和高速复原方法，使汉字信息压缩比达到 1:500。他设计出超大规模集成电路芯片，实现了高速度、高保

真的汉字字形复原、变倍、变形输出，突破了汉字激光照排中汉字的存储和汉字字形的还原输出两大瓶颈，达到世界先进水平，使我国的汉字激光照排系统占领了全球华文报刊的市场，因此获得了2001年度国家最高科技奖。

表1-1　汉字点阵类型和参数

点阵类型	点阵参数	每个汉字（行×列）占的字节数
简易型	16×16	32B
普及型	24×24	72B
提高型	32×32	128B
精密型	48×48	288B

字库可以存放在磁盘（软盘或硬盘）上，称为软字库。每次开机时，将字库从磁盘调入到计算机的内存中，供显示用。这样查找速度快，但要占用机器的内存空间。由于微型机的内存容量已经达到128M以上，装入软字库不成问题，因此软字库得到普遍使用。

另一种方法是把字库装在可擦除只读存储器（EPROM）或只读存储器（Mask-ROM）里，这就是所谓硬字库（汉卡）。把汉卡插到微机扩充槽内，作为机器的一个扩充ROM存储区使用。这种方法现在已经很少使用。

1.4.2　文字的显示输出

从键盘输入的字符经过键盘管理模块，变换成机内码。然后经字模检索程序，查到机内码对应的点阵信息在字模库的地址，从字库中检索出该字符点阵信息。利用显示驱动程序将这些信息送到显示卡的显示缓冲存储器中。显示器的控制器把点阵信息整屏顺次读出，并使每一位与屏幕的一个点位相对应，就可以将字