钱培德　陆建明　编著

DOS

汉字系统

高级技术

DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术
DOS 汉字系统高级技术

天津科学技术出版社

DOS 汉字系统高级技术

钱培德　陆建明　编著

天津科学技术出版社

津新登字(90)003号

责任编辑:徐　彤

DOS 汉字系统高级技术

钱培德　陆建明　编著

*

天津科学技术出版社出版

天津市张自忠路189号　邮编30002

天津市武清县振兴印刷厂印刷

新华书店天津发行所发行

*

开本 850×1168 毫米 1/32 印张 11 字数 271 000

1994 年 4 月第 1 版

1994 年 4 月第 1 次印刷

印数:1—3 000

ISBN 7-5308-1533-4

TP·48　定价:8.75元

内 容 简 介

目前使用的计算机汉字系统绝大部分是DOS汉字系统，本书全面和深入地介绍DOS汉字系统中的高级技术，这些技术能够有效地增强系统功能、提高系统效率和改善系统性能。

本书主要内容包括：系统概述和汉字代码体系，DOS汉字系统总体设计，DOS内存管理技术及其优化，词输入处理技术和多级词库结构，联想输入处理技术和智能型汉字输入技术，多级汉字库结构设计，字形变换技术，假脱机打印子系统设计，汉字处理数学模型，汉字造字程序设计，字模压缩和还原技术，汉字编码辅助技术和保护方式编程技术。

本书突出先进性、技术性和实用性。本书可供计算机开发和应用人员参考，亦可作为大专院校计算机专业的教学参考书。

前　言

DOS 操作系统是 PC 系列微型计算机的主操作系统。由于 PC 系列机的广泛应用，以及 DOS 自身的短小精悍，再加上 DOS 拥有极为丰富的应用程序，因而使 DOS 自 80 年代至今一直是拥有最多用户的操作系统。

在我国，DOS 同样获得了非常广泛地应用，是使用最广泛的操作系统。我国是使用汉语的国家，因此，汉字系统成为我国计算机应用的支撑基础。由于 DOS 使用的普遍性，所以我们使用的汉字系统基本上是基于 DOS 这一汉字操作系统的。最典型的汉字系统要数 CC-DOS，另外还有 UC-DOS、2.13 系统、SP-DOS 等，它们均是基于 DOS，在 DOS 基础上扩充了处理汉字的功能，所以我们把它们称为 DOS 的汉化版本，也称为 DOS 汉字系统。

CC-DOS 是一个典型的 DOS 汉字系统，它采用一系列技术实现 DOS 的汉化，我们称这些技术为 DOS 汉字系统的基本技术。CC-DOS 推出后，很快获得了用户的青睐，并且引起了越来越多的人对它的兴趣。随着对 CC-DOS 剖析的深入，DOS 汉字系统的基本技术已被越来越多的人掌握，目前在 DOS 用户中已相当普及。

近年来，我国的计算机汉字信息处理技术获得了飞速发展，DOS 汉字系统也有较大的发展。在此过程中，DOS 汉字系统采用了一系列新技术，这些技术是基本技术的升级，故称为 DOS 汉字系统高级技术。这些技术的应用，使得 DOS 汉字系统的性能更加优越，用户使用更加方便。

本书全面和深入地介绍 DOS 汉字系统高级技术，以帮助读者对它们有深入的了解，使这些技术得到普及，把我国的计算机汉字信息处理技术不断推向前进。全书共分四个部分，每个部分主要内

容如下：

第一部分（第一章至第二章）是基础部分，主要介绍汉字信息处理的概况，DOS 操作系统及其汉化原理，汉字代码体系和通用字符编码的国际新标准。

第二部分（第三章至第四章）是总体部分，主要介绍 DOS 汉字系统的总体设计，全汉字集处理系统的总体设计、内码设计、输入输出处理和系统实现方法，以及 DOS 内存管理技术。

第三部分（第五章至第七章）是核心部分，主要介绍词输入处理和联想输入处理技术，多级词库的结构与性能分析，输入系统的智能化，键盘模块的标准化，VGA 显示模块的设计，汉字库结构设计与性能分析，多级汉字库设计技术，字形变换技术，假脱机打印子系统设计，汉字处理数学模型的建立。

第四部分（第八章）是辅助技术部分，主要介绍汉字造字程序设计，字模压缩和还原技术，汉字编码辅助技术，保护方式程序设计技术。

本书注意突出先进性、技术性和实用性，并且力求作到概念清楚和通俗易懂。本书适合于广大从事计算机研究、应用和开发的科技人员参考，也可以作为大专院校计算机专业的教材和教学参考书。

本书由钱培德和陆建明执笔，并由钱培德主持编写和最后修改定稿。另外，杨季文、吕强、方奕、鲁言民等参加了本书编写过程中的部分工作。

在写作本书的过程中，我们始终得到天津科学技术出版社的支持，特别是徐彤编辑为本书的出版付出了大量的劳动。我们谨在此表示由衷的感谢。

最后，希望使用这本书的同行们能对本书提出宝贵意见和建议，以便今后进行修正和充实，我们对此表示谢意。

作者

1993.6

目 录

第一章 绪 论

第一节 汉字信息处理概况

一、汉字信息处理的重要性

信息的含义十分广泛，它是人们认识的一个基本要素，是对客观世界的直接描述，也是在人们之间进行传递的一些汉字知识。当今社会的一切活动都离不开信息的收集、组织、存贮、加工、传输、再生等等。它和物质、能量一起构成客观世界的三大要素。

信息处理是对信息的接收、存贮、转化、传递等操作。由于信息量的日趋庞大，使我们面临一个“信息爆炸”的社会。然而，人脑虽然在信息的识别、分析、综合、推理等高智能方面有着不可比拟的能力，但在存贮记忆、数值计算、检索能力及处理速度等方面却不够理想。因此，迫切需要人们用现代化的先进设备、技术来处理信息。

在信息处理中，对文字信息的处理称之为文字信息处理，其中对汉字信息的处理称之为汉字信息处理。我国是文明古国，是汉字的发源地，使用汉字已有数千年的历史，目前世界上使用汉字的人愈来愈多，汉语也做为联合国所采用的六种语言之一。由于汉字是一种表意文字，所以其字形复杂，而且字体繁多。传统的汉字信息处理基本上还是以手工抄写或机械处理，需消耗大量的时间和精力，因此远远不能满足现代社会的需要。

汉字信息处理的现代含义是用现代化的技术设备——计算机

去处理汉字信息，如存贮、分类、统计、检索、转换、传输等。主要处理汉字的文本信息、图像信息和语音信息。随着现代科学技术的迅速发展，用计算机对汉字信息进行处理已势在必行，如果不解决计算机的汉字信息处理，那么计算机的使用在我国将会受到极大的限制，影响了国内及国际间的信息交流，因此研究和开发汉字信息处理技术具有十分重要的意义。

二、汉字信息处理的原理

汉字信息处理是一门跨学科、多学科的综合性学科，包括自动检索、人机对话、自动翻译、语音识别、文字识别、自动分词、人工编码输入、自然语言处理等一系列课题的研究，其基本原理和西文信息处理的原理类似，它先将汉字转化为数字信息代码，而后经计算机进行代码处理，最后再在输出设备上将处理过的数字信息代码恢复为字形信息输出。

1. 汉字转化为数字信息代码

计算机无论是以数值计算为主，还是以非数值性的数据处理为主，都以代码信息的方式处理，目前，计算机一般均以 ASCII 码(USA Standard Code For Information Interchange)为内码，但也有的是以 EBCDIC 码(Extended Binary — Coded — Decimal Interchange Code)。由于 ASCII 码字符集较小，所以每个字符可全部集中在西文小键盘上表示，但是，当计算机处理汉字时，由于汉字的数量多，致使计算机处理汉字比处理西文字符更为复杂。从汉字输入的角度来说，说是要设法抽取方块汉字的特征信息，给每个汉字设计合理的信息编码输入计算机，这个编码称为汉字的输入码，它实际上是将汉字转化为数字信息代码。

例如，王永民先生的“五笔”输入法，它是以汉字的笔划和构件(部件、字根)为基础的字形编码。如“植”字，根据“植”字的特征信息，它的输入码为“SFHG”。

就目前而言，汉字编码方案不下六百多种，而且新的方案还在

不断涌现，大有“万码奔腾”之势，其主要原因在于汉字输入速度的问题，也可以说是重码率、击键次数以及用户记忆量之间的矛盾未彻底解决。所以，如何使汉字转化为合理的数字信息代码，即给每一个汉字编码，这是汉字编码方案研究中的一个重要课题。

2. 计算机内汉字代码的处理

任何用于处理汉字信息的计算机，均必须能同时处理西文字符信息，以保证中西文兼容，不然，许多先进的西文软件无法直接得到应用。为了确保计算机能自动区分西文代码和汉字代码，首要的问题是设计合理汉字代码。计算机内最重要的代码是汉字内码，有关汉字内码的详情将在第二章中作详细的论述。目前国内主要是采用与 GB2312-80 基本相吻合的变形国标码，用两个字节代表一个汉字，且每个字节的最高位为 1，这样作使用简便，易于区分中西文代码。但也有的汉字操作系统采用其它的汉字内码。

这样，计算机内汉字代码的处理首先是将键盘接收到的输入码转换为汉字内码，然后将汉字内码进行传输、存贮等处理转换。在输出时，根据汉字内码转换成汉字地址码，根据汉字地址码取出汉字字模供输出设备显示、打印。

3. 汉字的数字信息代码转换为汉字字形

将汉字的数字信息代码转换为汉字字形是汉字信息处理中的重要组成部分，是汉字输出的一个重要环节。一般汉字字形用点阵或矢量表示，对于汉字字形点阵而言，每个汉字都是由“0”、“1”按一定规则排列而成的信息代码，在输出时，“0”表示空白点、“1”表示汉字笔划上的点，这样将“0”、“1”组成的信息代码经一定转换，在屏幕或打印纸上，就可看到一个汉字的字形。对于用矢量表示的汉字字形而言，它的转换大体上分两步：第一步是将汉字字形的骨架和轮廓确定，第二步是对汉字字形的骨架和轮廓作必要的修饰，一般这两步在计算机内信息处理时完成。用户在屏幕或打印纸上所看到的是一个完整的汉字字形。用上述两种方式所表示的汉字，可有各种字体，如宋体、仿宋体、楷体、黑体等等。汉字字形的质量