微机中文信息处理技术

微机中文信息处理技术

主　编　蒋天发　陶　丹

副主编　肖荷生　郑克忠

武汉测绘科技大学出版社

(鄂)新登字 14 号

内 容 提 要

本书从实用的角度出发，介绍了微型计算机中文信息处理技术及方法。全书内容包括：中文信息处理概论；中文信息输入、输出设备及输出处理；汉字编码输入方法（五笔字型、自然码和表形码）；中文编辑软件 CWS 和 ZRED 的使用方法；WPS 桌面排版系统的使用及 SPT 的使用方法；软件汉化及中文软件设计基础；软件汉化的工具 PC TOOLS 和 DEBUG 的使用方法；软件汉化的原理及其方法；少数民族文字信息处理和国内应用较多的反病毒软件的使用方法。此外书中还给出了一些应用实例，可供读者借鉴和参考。

本书可作为大、中专院校理、工、农、医和经济等各类专业计算机应用课程的教材，也可供各级、各类培训班师生、录入员、程序员和书刊编辑人员作为工作手册或自学参考书。

图书在版编目(CIP)数据

微机中文信息处理技术/蒋天发，陶丹主编.
武汉：武汉测绘科技大学出版社. 1994,6.
ISBN 7-81030-336-8/T・54

Ⅰ.微…
Ⅱ.①蒋… ②陶…
Ⅲ.汉字处理-汉字信息系统-微型计算机
Ⅳ.TP391

微机中文信息处理技术

主 编 蒋天发 陶 丹
责任编辑 张立福

武汉测绘科技大学出版社出版发行
（武汉市珞喻路 39 号 邮编 430070）
武汉测绘科技大学出版社丹江印刷厂印刷

*

开本：787×1092 1/16 印张：17.5 字数：448 千字
1994 年 6 月第一版 1994 年 6 月第 1 次印刷

印数：1～4 000 册 定价：14.00 元

序

涉及微机中文信息处理技术的书已经不少，但蒋天发、陶丹同志主编的这本书有它自身鲜明的特点：除汉语的信息处理外，还有朝鲜文、藏文、蒙文、满文、维吾尔文和彝文等的信息处理。中华民族是个多民族和睦相处的大家庭，各兄弟民族在历史发展的长河中都形成了自己的民族文化，同时又对中华民族优秀文化传统做出了卓越的贡献。文字的计算机信息处理是继承和发展民族文化的重要手段，也是发展国民经济和科学技术的需要。这本书就适应了这样一种需要。

此外，这本书在汉语信息处理技术、软件汉化技术和工具方面也都有深入浅出的阐述。书中还包含有桌面排版系统和抗病毒方面的内容。

总之，我认为这是一本难得的中文——包括汉文和多种少数民族文种——微机信息处理的工具书。我相信它一定会受到广大读者——包括汉族和许多兄弟民族读者的欢迎。

王振宇

九九年五月

王振宇教授系中国计算机学会软件分会委员、中国软件行业协会理事、湖北省计算机学会常务理事、学术委员会主任。

——编者注

前　言

现代社会是充满信息的社会，对信息用现代化的工具进行处理和管理是现代社会的需要。随着计算机及计算机科学技术的发展，电子计算机，特别是微型计算机的应用已深入到现代社会的各个方面。计算机文字信息处理及计算机事务管理是现代社会信息处理的两大主要方面。

针对我国，要进行处理的信息主要是中文信息。中文信息处理，狭义上特指汉语汉字信息处理；广义上则指中国民族语言文字信息处理。通常，中文信息处理，除汉字信息处理外，还应包括朝、藏、蒙、维、哈、柯、彝等几十种民族语言文字的信息处理。显然，中文信息处理实质上是多文种信息处理。面向世界诸多文种的民族文字支撑就形成了计算机软件及系统的国际化。民族文字化和国际化正是同一个大潮的两个侧面，从个别国家、民族、地区的视角观察是民族文字化，从综观国际信息界的视角观察则是国际化或多文种化。汉化正是这种历史必然发展的一环，是大潮中的一股潮流。

我国在文字信息处理方面，现在由计算机产生的公文、报纸、杂志、书籍……比比皆是。计算机的文字信息处理与排版系统正在迅速取代传统的机械打字机和铅字印刷行业；文秘人员、作家、新闻工作者、编辑等正在开始或早已开始用计算机这一现代化工具从事自己的日常工作；写字不用笔，办公不用纸已并非笑话。事务管理方面亦是如此，在许多单位的人事、工资、设备、计划、统计、财务、经营、销售等部门都已开始或正在打算用计算机进行管理；银行、电信、飞机场、火车站等早已普遍应用计算机开展业务；以前很多手工做的事情，现在逐步变成了计算机处理，有的已发展到了离不开计算机。这一趋势正在迅速发展，并远远没有达到高潮。因此，现在很多单位在用人、招工时，在录用条件上往往有一条是懂计算机、会计算机操作者优先。面对这一情况，学计算机知识，掌握微机中文信息处理技术已成了长期持续不退的热潮。

因此，为满足社会和教学的需要，笔者根据多年讲授《中文信息系统》课的体会和经验，在原讲稿的基础上，编写了《微机中文信息处理技术》一书。全书共分十一章，第一、七、八、九、十章由蒋天发编写，第二、三、四章由陶丹编写，第六章由肖荷生编写，第五、十一章由郑克忠编写。全书由蒋天发统稿。

在写作过程中，得到了武汉大学计算机科学系黄俊杰教授和中南民族学院计算机科学系主任张群教授的指导和帮助；中南民族学院民族学系谢志民教授对本书少数民族文字信息处理等部分章节进行了认真审阅；中国船舶工业总公司第七〇九所副总工程师王振宇教授为本书作了序；谨在此向他们表示衷心的感谢。

为了体现先进性和技术性，我们广泛参阅了现有的有关资料，从中得到了许多启发，吸收了不少营养，部分内容借用了这些资料的图表，这些均以参考文献的形式列出，并致以谢意。

由于中文信息处理是计算机科学中的一门新科学，发展很快，加上我们水平有限，书中不妥之处在所难免，深望本书读者批评指正。

编　者

1994年5月于武昌

目 录

第一章　中文信息处理概论

§1.1　信息与信息处理的基本概念

信息是一个正在不断发展和变化的概念，至今还没有一个公认的定义进行描述。但是，人们也试图用各种理解去进行解释，例如：信息是具有新内容、新知识的消息；信息是事先不知道其结果的消息；信息是对客观世界现象通过直接观察、或对讯号的语义解释领会而得到的知识；信息是反映客观世界中各种事物的特征和变化的组合，是一种有用的知识；……我们认为，信息是自然环境和人类的一切活动所产生的各种状态和消息的总称。信息这一概念人们很早就知道，从定性的意义上讲，人们在得知某个消息后，若他在事前认为消息中所包含的事件发生的可能性愈小，则认为这个消息给他带来的信息量就愈大。可见信息的量值与事件的随机性或不定度有关。信息在人类社会活动的各个方面无不显示出其极大的重要性。信息的价值体现在信息的准确性、及时性和适用性，对于任何一个决策者来说，只要失去其中之一，该信息就将变得毫无价值。例如，棉花增产对于纺织行业和棉农来说是有用的消息，但对建筑行业并无价值，因为不适用。

现实世界的信息可分为两大类：一类是自然信息；另一类是社会信息。

自然信息是由于自然环境的变化而发送的信息。例如候鸟的迁移，鱼类的回游和爬虫的冬眠等就是它们接受了自然信息后所产生的反应。这种信息的流程是被动的，即：信息发送→传输→选择与接收→达到适应环境的目的。

社会信息是人类群体生活中产生和交换的各种频繁和复杂的信息。人类对外来信息不像低等动物那样只能进行简单的适应，而是具有记忆和辨别的能力，能进行逻辑推理和形象思维，建立新的概念，发现新的规律，以便把客观环境改造为适合人类自身生活所需要的环境。所以，社会信息的流程是：信息发送→传输和交换→选择与接收→记忆和辨别→处理和加工→达到改造环境的目的。社会信息与自然信息的最本质区别在于社会信息可以由人类进行各种加工处理，成为改造世界和能够不断发明创造的有用知识。

任何信息都需要载体。自然信息的载体是未经加工的自然物，而社会信息的载体却和人类社会的发展和进步密切相关。在远古时代，人类用五官、表情、手势、语言等作为载体传递和交换信息。以后发明了文字，使信息可以长期积累和保存，人类文化可以直接遗留给子孙后代。在科技不甚发达的时代，信息的作用及其利用价值被限制在较低的程度。例如，信息技术的一种手段为传递，在电信技术发明以前，人们只能用人工通信，或者用其它简单的表示方式或各种约定来传递信息。随着电气通信技术的问世与发展，人类开始用电波作为信息的载体。社会信息以光速进行传播与交换，使信息的传递速率大大提高，效能也大为改善，但目前还仅限于传递信息。信息技术的另一种手段为处理技术。20 世纪 40 年代发明了电子计算机，大大提高了人类对信息处理、存贮、传播与交换的速度与能力，为实现生产过程、办公和家庭自动化，为向信息化社会过渡创造了必要的条件。对信息进行加工处理离不开计算机技术，所以信息处理这一术语就和计算机技术联系在了一起。用计算机处理或加工信息，扩大了信息的利用范围，使

信息的利用价值也大为提高。由于这一意义深远的科技成果的应用，使信息愈益成为现代社会的科技进步、经济发展、人类文明进程所不可缺少的社会财富。它和物质、能源被列为同等重要的地位并被看作为现代人类社会生存和发展的三大要素。科技先进的国家，已经建立起强大的信息产业，并仍在以很高的速度向前发展，在整个国民经济生产中占有的比重愈来愈大。信息处理技术在人类文明和科学技术现代化的进程中正在发挥其重要的作用。广义的信息涉及多种范畴。例如：人类社会活动所产生的各种信息；科学技术和生产活动产生的各种信息和自然现象所包含的各种信息。在这些含义丰富的信息中，信息的表示形式又是多样性的。例如有声音、图像、文字、图形等多种表示形式，这称为信息的多元化表示。信息的多种物理表示形式，成为信息的多种载体或媒质，表现为媒质的信息。

我们用计算机处理多元化信息，是信息处理技术的范畴。传统的信息处理技术在近十多年来有了很大的发展。这是由于微电子技术和计算机技术的飞速发展。微电子技术的进步体现在制作超大规模集成电路的技术水平日益提高以及各种大容量存贮器芯片和具有复杂逻辑运算功能的集成电路芯片的不断问世，并且迅速推广应用。计算机技术的进步体现在计算机硬件性能价格比的大幅度提高，微型机和以微型机技术为基础的各种终端设备的日益普及应用。这些因素大大推进了信息处理技术的实用化进程。另一方面，计算机软件技术也有很大的进步，如软件工程、第四代程序设计语言和各种先进的软件工具的实用化、数据库管理系统等各种公共支持软件技术的进步与普及应用；人工智能软件技术的发展以及各种应用软件的开发和利用，不仅使数据和文字信息处理技术更加完善，获得了更为广泛的应用，而且开拓了信息处理技术的更新的应用领域，如图像信息处理、模式识别、语音识别和语音合成、自然语言处理、语言的翻译等高技术领域。

众所周知，计算机也具有通信功能，即利用数据通信技术实现计算机网络通信。传统的通信技术以传输模拟信号为主。自从发展了数据通信技术后，经计算机存储和处理的信息可以在两台或多台计算机或数据处理设备之间互相传输，更加增强了信息处理技术的效能，并扩展了信息处理技术的内容，这称为广义的信息处理技术。若从另一个角度来看，把传统的通信技术的内容加以扩展，现代化通信技术的概念把信息传输和信息处理两种功能结合起来，称为计算机与通信技术。现代化通信技术的信息载体是综合的多元化信息，即包括数字、文字、语音、图形、图像。传统的信息处理只指狭义的信息处理，如信息的存储和检索；传统的通信技术只是完成信息的传输或转移，而现代化的通信技术（即广义的信息处理技术）则兼有信息处理和信息传输的功能。60年代初期，开始发展计算机与计算机之间，或计算机和终端设备之间的直接互连通信。60年代末期又发展了以报文分组交换为特点的公共数据通信网络，使计算机远程通信技术较快地走向通用化和标准化，为这项技术迅速地推广应用创造了条件。由于远程计算机通信技术的较快发展，使实现电子邮件技术、远程情报资料检索、数据库检索以及远程批处理等技术成为可能，从而有可能大大扩展信息处理技术的距离范围。由于通信网络技术的发展，给分布式信息处理和资源共享等新技术的发展创造了条件。70年代中期又发展了局域网络通信技术，在1～10km的范围内，能以较高的通信速度实现信息通信和资源共享，并能在这个距离范围内实现电子邮件传送。这为目前正发展的办公自动化技术提供了很大的方便，从而扩展了计算机信息处理技术的应用范围。局域网络和远程通信网络的连接，使办公自动化技术的作用范围可以跨越城市、国界甚至洲界的距离，有可能使地球上任何距离的办公室之间可以实现同时办公通信。

§1.2 文字信息处理

信息的表示形式是多样的，在多元化的信息中，文字信息是一种最通用、最普遍的表示形式。无论是公文、文件、信函、报表以及各种印刷出版物等，绝大多数都使用文字的形式来记录。文字是一个国家或民族文化的象征，在社会和历史的发展中有着特殊的地位。计算机的应用已从数值计算发展到非数值以及文字信息处理，这是计算机发展史上的一个重要转变，因为它大大地开拓了计算机的应用领域，使计算机渗透到各行各业。文字信息处理的应用范围非常广泛，从编辑文稿、建立文件档案资料、排版印刷，到行政管理、办公自动化，凡是需要用文字表达信息的应用场所，都可以利用文字信息处理技术。随着个人计算机应用的普及，以微机为基础构成的文字处理机目前已有很大的发展。文字处理机根据其应用的不同要求，可设计成不同的档次。使用最为普遍的是便携式的文字处理机，又称为电子打字机，其使用的范围也在不断扩大。这和传统的机械式打字机相比，具有编辑功能丰富、灵活的独特优点，并且可以提供一定数量的文件存档，其价格也在逐渐降低，今后有可能逐步取代机械式打字机。高档次的文字处理更具有传统的机械式打字机所无法比拟的优点。随着微型机性能和软件技术水平的不断提高，文字处理机的功能也会不断扩展。如高档文字处理机利用计算机人工智能，在字、词处理的基础上有可能增加句法和语法处理、书面自然语言处理等新功能。

计算机能具备高速运算和处理能力，是因为它利用了电子技术处理或执行二进制数运算这一法则；其中的运算器利用半导体器件的二个状态(即通与断)的变化，代表二进制数字串中的一个二进制数位上的"1"或"0"的变化，即能高速地执行二进制数的数值或者逻辑运算。实际上，无论计算机作数值的或其它任何种类信息的运算或者处理，最基本的运算操作，都是这种二进制数的演算。由此可见，文字信息处理的实质，是先把文字信息数字化，也就是用一个固定的数码代表一个字母或文字。比如在英文信息中，以 26 个字母作为文字信息处理的单位。因此要对 26 个字母逐个地确定代替它的数码。在汉字的情况下，一般是以一个整字作为文字信息处理的单位，所以要对每一个整字确定唯一地代表它的数码。这一数码，统称为代码。在计算机内部处理文字信息时，就像处理数据一样对待。处理完毕后，再把替代的数码还原成相应的字母或文字。利用计算机能够高速处理数据的性能，使文字信息处理能够分享计算机技术的这一独特优点，从而实现文字信息处理的高效能化。

§1.3 中文信息处理的必要性

世界上的文字可分为表意文字、音节文字和音素文字三种。表意文字单独为一个体系，音节文字和音素文字合称为拼音文字体系。全世界所有文字符号都可以包括在这两大体系中。拼音文字体系字母总数少，笔划简单，词形清楚，无论多少万个词都能由几十个字母线性组合而成。也就是说，在拼音文字中，单词和词组都是字母的线性序列。比如，英文的单词与词组都是由 26 个英文字母线性组合而成。用计算机来处理拼音文字，其输入输出方法和信息存储都能很方便地解决。

表意文字是文字发展史上最早出现的一种类型，例如古埃及文字、古玛雅文字和方块汉字都属于这一类型。表意文字与语言没有直接的关系，因此可以用来表达不同的民族语言。我国是个民族众多的国家，汉族因居住区域广，有着各种各样的方言。表意文字恰好适合于方言分

歧的民族。汉字是世界上许多种古代的表意文字中唯一能够巩固和流传下来的文字体系。由于汉字是用各个符号表示个别的完整的词或它的独立部分，用无数独立的符号来记录语言，用不同的符号来表达各个不同概念的词。因此，随着语言词汇的丰富，使它在四千多年的历史演变中，使用符号的数量随语言的发展而增长，以至达到数万种之多。汉字是把形、音、意三个方面结合起来的独特文字。每个文字都有其特定的形体，都有一定的读音，表示一定的意义，它们是不可分割的统一整体，因此，汉字的形、音、意称之为文字的三要素。

中文是我国的通用文字，严格地讲，中文应包括我国各民族所使用的文字。由于汉字在我国使用得非常广泛，汉族人口众多，汉字和汉语自然就成了我国的特定通用文字和语言。由此可见，汉字在中国具有特别重要的地位，因此，在不少场合，中文信息处理就是指汉字信息处理。汉字不仅是我国使用最广的文字，而且是联合国五种通用文字之一。据统计，全世界约有1/3的人使用汉字，而我国有12亿人口，使用汉字的人口约占80%。因此，我国不仅在对汉字的形成、发展与改革中起了主导的作用，而且在汉字信息现代化处理的变革中，更应该做出主要的贡献；所以，加速中文信息处理技术的研究以及推广应用具有重要的意义。

现代社会是充满信息的社会，对信息进行处理和管理是社会的需要。由于社会信息日趋庞大和复杂，如果仍用传统的人工方法来实现对信息的存储、传递和处理，则要花费大量的劳动，而且由于信息繁多和人脑工作的固有特点，往往使这些工作不能达到令人满意的效果。目前我国微型计算机大量普及，使利用计算机进行信息处理已成为现实，又由于计算机所具有的优点，使得其完全能够胜任这种工作。所以，利用计算机进行社会信息的处理已经势在必行。随着计算机技术的不断发展，计算机系统的功能也不断增强，计算机的应用领域也在不断拓宽。中文信息处理的函义与涉及的范围也大大扩展了，现在已包括情报资料和图书的自动编目与检索；书刊和报纸的自动编辑与排版；事务处理和企业管理；办公自动化与数据通信等。因此，解决计算机的中文信息处理问题，已到了刻不容缓的时候了。由于我国是汉字的发源地，对于汉字的研究最深入，因此，对汉字结构的特性及使用情况最熟悉。同时，我国对发展计算机中文信息处理技术的要求最为迫切，得到的收益也最大。所以，我国理应在计算机中文信息处理领域中走在世界的最前列。

§1.4 中文文字信息处理的特点

一般来说，中文文字是指在中国广泛使用的汉字。中文的基本组成单位是汉字，汉字的字种多，字形复杂且形、音、义缺乏有机联系，用计算机处理起来较为困难。因此，要用计算机处理中文信息，必须了解汉字的特点。

汉字的主要特点是它属于象形文字，字量大，字形复杂，和西方国家广泛使用的拼音文字有显著的区别。西文的特点是字形信息和语言信息的关系密切，用少数结构简单的字母图形，自左而右在一维空间里依线性关系顺序排列成单词。汉字不仅构成的笔画多，而且它是一种二维结构的图形，比西文单词的线性排列结构要复杂得多。由于这些特点，在汉字编码方法输入计算机的问题上造成不少困难，国内外有不少学者从研究汉字结构与汉字编码的角度出发，致力于把汉字拆分成基本笔画、字根或字元，希望从这些分析中找出汉字结构的规律性，从而归纳出一套简明而容易掌握的组字规则或编码规则。这些工作虽然已取得了一些成果，但还未能达到满意的程度。汉字的字量大，据统计，中国的汉字总数超出六万个。而且，不同的汉字在不同的历史时期，不同的专业领域中使用时，其频度的差别是很大的。根据1974年对国内使用的

现代汉字综合使用频度的统计，要求覆盖率达到99.99%的情况时，所需要的汉字量约在七千个左右。所以在1979年国家制订颁布的《信息交换用汉字编码字符集(基本集)》(国家标准代号为GB2312-80)中，共收容了6 763个汉字。在6 763个汉字中又分为两级，第一级为常用汉字，共3 755个；第二级为次常用汉字，共3 008个。对6 763个汉字用计算机技术加以分区，按最小信息冗余度的原则，需要用13位二进制信息($2^{13}=8\ 192$)。实际上是用二个字节(16位二进制信息)表示一个汉字信息交换码，或简称汉字交换码。

计算机进行汉字信息处理的全过程大致包括三个环节：

其一是文字信息的输入。通常是通过键盘把组成英文词汇的各个英文字母逐个地输入。这一过程中键盘的作用是把输入的每个字母、数字或各种符号转换成它们所对应的代码，供下一步信息处理用。键盘同时也是使用或操作计算机的人和机器系统之间的界面。因此，键盘要设计成方便人们的使用和操作，以提供良好的人机界面。由于汉字的字量大、字形复杂的特点，使汉字输入技术成为中文信息处理上的一个主要难题。汉字输入计算机的主要方法目前仍是利用键盘，通过汉字编码方法输入。汉字编码输入方法有两大类，一类是汉字整字编码法，对于六千多个汉字，采用某些规则排出它们的流水号，顺次把它们排列在键盘上。使用整字编码的键盘，是一种专门设计的汉字(整字)键盘，造价较高，因此，这种输入方法不易推广。另一类是按汉字的字形，或发音特征，或利用汉字的形、音特征相结合的编码方法。由于把汉字拆分成笔画、字根或字元，把按发音的音、韵、调等作为编码的依据，所使用的码元较少(和汉字的字数相比)，因此这类编码方法绝大多数就利用英文字符系统的通用字符键盘作为输入工具，这种键盘不仅造价低，而且和字符系统在输入设备上的通用性好。因此，这种编码方法目前得到广泛的应用。现有汉字编码方法的种类很多，仅国内提出的汉字编码方案就有五百种之多。可真正得到用户接受并能推广应用的尚不到其中的1/10。汉字编码输入方法是一个主要的人机界面，所以要经过认真考查、评测，优选出技术指标较高，并且能为广大用户接受的汉字编码输入方法。

其二是文字信息的处理。文字信息处理包括多种不同的处理要求。如在文稿的编辑操作中，有对文字(或文字中包含的字母)的增、删、改的操作；有对若干个字、整个句子或整段的增、删、改的操作。在对文字串的处理中，有分类、合并、比较、排序、检索以及对齐等的操作。这些种类的操作都可以预先编制成相应的处理程序来实现。对于汉字信息的处理，因字量大，字形复杂，所以对汉字字形的存贮器的容量提出了较高的要求。

其三是文字信息的输出。文字信息处理完毕后，要把处理结果的代码信息转换成文字的形式输出，输出方式包括显示和打印。为此，在计算机系统中要存储有关文字的字形信息。计算机中存储的文字字形，是以点阵式字形的形式表示的。通常英文字符信息用5×7或7×9的点阵表示。这样的字形点阵信息和计算机中二进制数的存储相对应，即有笔画的点用二进制数1表示，无笔画的点用二进制数0表示。因此，在计算机中存储的字形信息实际上也是一串二进制数。在英文信息处理系统中，字形信息的存储问题比较容易解决。因为只需存储大、小写的52个字母，10个阿拉伯数字，加上一些图形符号，总共94个符号。用容量不大的存贮器芯片，即可解决全部字符点阵信息的存储。而汉字是图形文字(象形文字)，是笔画的二维结构。它的字形复杂，笔画繁简不一，少至一笔，多至三十余笔，笔画的方向及形状的变化也较复杂。汉字的字形，可以用两种方式来表示，一种是将单个汉字的字形离散成网点，每点用一个二进位表示，一个字的所有网点数据构成了该字的点阵式字模。点阵式汉字可用于汉字的显示和打印。另一种是向量式表示法，它将汉字的笔画表示成一组二维线段，并将线段的端点数据保存在汉

字向量字库之中。向量式汉字可用于图形显示器和绘图仪，字形可任意进行比例和旋转变换。在中文信息处理系统中，为了显示或打印汉字就必须建立汉字库，汉字库是用来存储汉字字形信息的。一般来讲，大多计算机中存储的汉字字形，都是用点阵方式来表示的。和结构简单的英文字符相比，点阵式汉字字模要求用较高的点阵密度来表示。最少的汉字字模点阵表示要求15×16点，字形质量稍好些的要24×24点阵。这样的点阵密度，一个汉字字模便要占用较大的存储量，总数为六七千个汉字要求有大量的字模库存储容量。在发展汉字信息处理技术的早期，因为当时集成电路存贮器芯片的容量小，价格贵，汉字字模的存储曾经是中文信息处理技术的一个棘手的问题，当时也曾设法采用过存储字根或字元，用软件方法来组成完整汉字的方法，以节省汉字库的存储容量；也曾一度广泛使用磁盘等用软字库方法存储汉字。这些方法虽然局部地解决了节省存储空间的问题，但在汉字字形质量和汉字输出速度等方面都受到影响和限制。在近几年内，由于半导体超大规模集成电路存贮器芯片的存储容量迅速提高，单位存储容量的价格下降，使汉字字形信息的存储问题得到基本解决。例如用于存储汉字字形信息的只读存贮器ROM，目前常用的有1兆位、2兆位、4兆位等几种。对于15×16点阵的汉字收存全部因素标准基本集(GB2312-80)两级汉字只需一片2兆位的ROM芯片。这样的汉字字模库不仅成本低，容易制作，而且体积小，使用、安装方便，容易普及应用。

对于不同的使用条件，汉字字模的质量规格也有不同的要求。如15×16、24×24点阵的汉字属于目前常用的针式打印机(分辨率为7～9点/毫米)印出的较低质量的字模规格。若使用较高分辨率的印字机，印出同样大小尺寸的汉字，则点阵规格必须相应地提高。因此，需要设计出32×32、48×48等点阵规格的字模。此外，若考虑要求印出大小尺寸不同的汉字，则对于一种分辨率规格的印字机，也要配备几种不同点阵规格的字模。对于通用型的汉字字模，主要用于印制一般的中文文件、报表。而精密型汉字字模的用途是利用计算机技术的印刷排版。两种字模的主要差别在于它们所用的点阵规格。通用型字模要求的分辨率一般在7.08～11.8点/毫米的范围；而精密型字模的分辨率则要求在27.4～40点/毫米的范围，两者差别很大。对于通用型字模，目前一般采用逐点存储的方法；而精密型字模，由于其信息量太大，即使目前存贮器芯片的应用已较普及，但是仍有必要采用压缩信息的技术以减少字模信息所需的存储量。

一个中文信息处理系统必须具备汉字输入、汉字存储、汉字显示、汉字打印和汉字传输的基本功能。对于这五项功能，每个汉字都有对应的五种表示法：一是内码，即系统内部处理和存储汉字使用的统一编码。代码定长。选取机内码一般遵循四个原则：(1)能与机内的ASCⅡ基本码区分开来不产生歧义；也不允许汉字之间有重码。(2)应便于检索，与字库地址间的关系也应简单，以便于字库管理的简单化。(3)应与GB2312-80规定的国标码有简单的对应关系，便于中文信息在不同系统中的交换。(4)码长在能区分8 000至10 000字的前提下，应尽量地短。二是输入码(键盘码)，即用户从键盘输入汉字所使用的汉字编码，又称外码。代码不定长或定长。用字母、数字串代表汉字的输入编码应具有易记忆或不需记忆编码规则，使各类用户不加训练或稍加训练就能学会使用；编码长度应尽可能地短，以便加快汉字的输入速度；编码与汉字对应性要好，以便减少重码。一个系统可以选用多种输入码方案。无论采用哪种输入方案，输入的汉字编码都将由键盘驱动程序转换成内码，以便保存或进行显示、打印和传输操作。三是显示字模码(字形码)，即用于显示汉字字形；16×16点阵汉字为32字节码，24×24点阵汉字为72字节码。显示处理程序的功能是将汉字的内码转换为显示字模码，并将表示汉字字形的显示字模码在显示器上显示出来。内码到字模码的转换是通过查找汉字库来完成的。汉字库按其存放介质的不同分成软字库和硬字库两类；软字库在系统启动时装入内存，需占用用户

存贮器；硬字库则固化在 EPROM 或 PROM 中，启动时不必装入，也不占用用户的 RAM 区。四是打印字码，即用于图形打印机打印汉字；该码也可由显示字模码加以变换而成。由于点阵打印机的输出单位为行，所以必须将一行汉字的内码全部转换为打印字模码后，才能打印。转换过程中需要使用的字模库可以是专用的打印字库，也可以利用显示字库。显示字库中取出的字模码必须经过转置操作，才能变成打印字模码。在打印过程中还可以选择打印字号；不同大小的字是整倍放大而得到的。放大的字用 2 个以上的点表示原字模码中的一个点，例如：24×24 点阵的汉字可以放大为 48×48(大号)、24×48(扁体)、48×24(长体)等多种字号。五是传输码(交换码)，即终端与主机或主机与主机通讯时使用的汉字编码，也就是说它是一种用于系统间或计算机通信用的汉字信息交换码。它是中文信息处理技术的基础标准。

现在我国所使用的微机绝大部分是中英文兼容的，这是什么原因呢？这是因为不论是英文字符，还是中文的汉字信息，在计算机内部都已转换成二进制的代码表示。唯一的差别在于英文字符是用一个字节代表一个字母、数字或图形符号；而汉字是用二个字节代表一个汉字信息。因此，凡是英文字符能实现的信息处理功能，汉字信息也能实现。但是，由于历史原因，中文信息处理系统不宜单独地自成系统，而必须在国际通用的英文字符系统的基础上开发。这是由于不论是系统硬件和软件，通用的英文计算机系统已有了相当的基础。若撇开原来英文字符系统的硬、软件环境基础，独立地开发中文计算机系统，在技术上并非不能实现，但是这样做，工作的起点就很低了。大量已成熟的、国际上通用的各种软件资源就不能加以利用，限制了系统功能的发展。而且，也不利于和国际上的标准技术相兼容。因此，开发中文信息处理技术，必须走和国际上通用技术相兼容的道路。同时，这样做也可以站在较高的起点上开发中文信息处理系统，收到事半功倍的效果。这项技术称之为中英文兼容技术。它的出发点是完全保留并利用原来英文计算机系统的一切硬件和软件功能。在此基础上，再增加中文信息处理能力，把中文信息和英文、数字信息的处理功能兼容于同一系统中，并不损失原英文系统的功能，使系统能方便地处理中、英文混合的信息流。

在原英文系统的基础上扩充中文信息处理功能，在设计上会受到一定约束。例如，为了达到中、英文信息兼容的目的，汉字信息交换码要遵守英文、数字系统字符代码体系的数据格式。同时，要利用计算机原有的系统软件兼容中、英文两种代码，又要求系统能明确地区分两种代码，以便在信息输出时，系统能对两类信息在逻辑上区分开，并分别处理。以上第一点要求是容易达到的，因为汉字信息交换码的设计是根据标准字符代码(即 ASC Ⅱ)扩充而来的。ASC Ⅱ 共包括 94 个字符，用二个 ASC Ⅱ 交叉组合成汉字信息交换码，共 94×94=8 836 个，汉字基本集实际使用了其中的 6 763 个。它们都是七位二进制信息表示的代码，仅有的区别是，英文字符用单字节表示，而汉字则用双字节表示。数据格式相同，可以为系统所接受。第二点要求是中文信息处理所特有的条件。因为无论单字节的字符代码和双字节的汉字代码，都是七位二进制信息进入系统后，若不加其它的标识信息，则将无法对二者加以区分。因此，汉字信息进入系统后，应对汉字代码添加相应的标识信息。加上标识信息后的汉字交换码，称为汉字内部码。比如，在微型机系统中，目前常用的汉字内部码的表示方法，就是对每个汉字交换码的双字节中，每个字节的最高位(原来未使用)置 1，作为汉字代码的标识。这是一种最简便易行的添加汉字代码标识信息的方法。

综合上述情况，可以归纳出中文信息处理系统技术的要求以及特点，主要有以下四个方面：

(1)要解决使计算机系统能输入和输出汉字信息。

(2)要解决信息量很大的汉字字形在系统内的存储。

(3)系统技术上,要解决中英文信息的兼容问题;要求系统能处理中、英文混合信息流。

(4)中文信息处理系统技术必须走和国际标准相兼容的道路,以便中文信息处理能共享原英文系统所开发的各种硬件和软件资源。

§1.5 中文信息处理技术发展概况

1."748"工程的前后

我国中文信息研究和计算机相结合的历史并不短,几乎和我国计算机的历史一样长。早在50年代末期,我国就研制了第一台104大型计算机,科技人员就在机器上进行俄汉机器翻译工作。从这时起,汉字开始与计算机结下了不解之缘。这就是我国计算机中文信息处理研究的开端。限于当时的技术条件,计算机的存储量有限,只能容纳少量汉字信息。打印汉字的技术也还没有解决,只能将中文译文用拼音字母打印输出。

到了60年代后期,我国开始对汉字信息处理技术进行进一步的探索和研究,并成功地研制出了汉字电报译码机。这种机器能以点阵方式在纸上输出汉字字形,为以后大量使用的汉字点阵式打印机提供了基础。集成电路的出现及其集成度和运行速度的不断提高,价格的不断下降,加上磁盘容量的迅速提高,为中文信息处理的现代化,为汉字信息进入计算机提供了坚实的物质基础。

从70年代开始,我国开始系统地研究和开发汉字信息处理技术,1974年8月原第四机械工业部、原第一机械工业部、中国科学院、新华通讯社和国家出版事业管理局五个单位联合向国务院和国家计划委员会提交了"关于研制汉字信息处理系统工程"的请示报告,拟名"748工程"。同年9月24日国家计委批复,同意把研制汉字信息处理工程列入1975年国家科学技术发展规划,正式列为国家重点工程,即著名的"748工程"。这项工程项目包括三个研制任务,即精密型汉字编辑排版系统、汉字情报检索系统、汉字通信系统与汉字终端设备。这三项任务均取得了重大成果,从而把我国的汉字信息处理水平提高了一大步。

进入80年代以来,我国的汉字信息处理技术更加蓬勃发展。这方面的学术研究和学术交流更加活跃,各种学术团体和组织纷纷成立。1981年6月成立了中国中文信息研究会,由著名的教授钱伟长担任理事长,下面设立了基础理论、汉字信息处理系统、汉字编码、汉字信息处理专用设备、自然语言处理和汉字字形等专业委员会。中国计算机学会也设立了中文信息处理技术专业委员会。这些专业学术团体组织了大量的国内和国际学术交流活动,有力地推动了我国的中文信息处理技术的发展。

2.汉字输入技术的发展

汉字输入技术在上一节已有提及,对于汉字编码输入方案在第四章中要较详细地介绍。今后的重点工作是优化汉字编码输入方案,分开几个技术层次积极推广应用。要开发智能化的汉字输入方法,如光学汉字识别技术,汉语语音识别技术。智能化的汉字输入方法并不能完全取代键盘输入方法,而是相互补充,使汉字输入方法多样化,满足不同应用目的用户的需要。

3.系统软件技术的进展

中文信息处理的系统软件技术经历了一段时间的演变过程。在70年代中期所用的汉字信息处理软件方法比较简单,在系统软件以外,编制一个"汉字输入输出管理程序",又称之为"汉字驱动程序",置于用户级。当用户程序运行到要输出汉字表示的结果时,由用户程序调用这一

程序模块，执行输入、输出汉字信息的任务。由于数据处理和汉字信息处理是分别进行的，未经操作系统软件统一调度运行，因此其运行效率低。此外，用户的编程工作也比较繁琐。

80年代初期，微型机技术在国内开始推广应用。微型机的硬件和软件的结构比较简单，容易对它进行二次开发。例如早期8位微型机的CP/M操作系统，经扩充汉字功能后，把它改造成中英文兼容的操作系统，使汉字的输入输出操作能和英文信息处理一样，直接由操作系统来管理和操作运行。稍后IBM PC机在国内推广时，对它所用的MSDOS操作系统进行了类似的功能扩充，经扩充汉字功能后的MSDOS称为CCDOS，在国内应用很广。它适用于IBM PC各档机器和它的多种牌号兼容机，对微型机汉字系统在国内的推广应用起了不小的作用。近几年来，对微型机上运行的Unix操作系统的汉字功能扩充也作了有成效的工作。

今后的发展方向是自行研究和设计能同时处理单字节和双字节代码信息具备汉字信息处理功能的操作系统软件。设计这样的系统软件要考虑和通用的国际标准相兼容。

操作系统汉字化的功能应达到如下几点要求：

(1)兼容性好，系统能输入输出和处理中英文混合的信息流；汉化的操作系统应与原系统保持向上兼容，以充分利用原系统的软件资源。

(2)可用性好，即具有良好的用户接口，便于采用多种输入方法并提供交互式操作。

(3)在源程中可以包含汉字字符串或汉字常量，还可以使用汉字与字母混合书写的注解。

(4)扩充性好，即允许用户增加输入方式或更换设备。

(5)直接用操作系统或程序设计语言的输入输出语句，实现汉字或字母数字的输入与输出操作，并可用汉字或字母数字表示文件名。

其它像程序设计语言的汉化问题，一些实用程序的汉化问题，国内也已做了不少的工作，并取得了一些成果。

4. 中文设备

中文设备是指一些中文信息的输入输出功能的设备。输入设备能实现把中文信息变换成中文信息在系统内部的表示形式，送入主机内。目前常用的输入设备有键盘、语音识别器、字形扫描识别器等。输出设备能实现把中文信息在系统内部的表示形式变换成中文信息的字形或语音。目前常用的输出设备有显示器、打印机或语音合成输出器等。

80年代以来，随着微型机技术和集成电路技术的不断发展，中文设备的研制和生产发展很快。下面简介几种设备的情况：

(1)中文键盘。目前主要发展和应用以字符键盘作为输入工具的汉字编码输入技术。我国字符键盘的生产目前已达年产20多万台，其中主要用于配备国产的各种牌号的微型计算机系统和终端设备，其中半数以上具备中文处理功能。其它像用作整字输入的笔触式中文键盘，以及某些特殊规格的整字输入键盘，我国也有小批量的生产。

(2)中文信息存贮芯片和汉卡。根据国家标准的几种规格的点阵式汉字字模，近几年来陆续生产了多批汉字字模存贮芯片。如15×16、24×24、32×32、40×40、48×48点阵，并根据用户使用要求，分别制作了宋体、仿宋、黑体、楷体等几种字体的字模芯片。汉卡是近几年来研制、生产和应用很活跃的一种中文信息设备，它按照某种汉字编码输入方案，把相应的处理和译码程序固化，并把固化后的集成电路芯片安装在印制电路板上，可以方便地在微型机的总线槽中安装或拆除，方便用户的选择和使用。有些汉卡把汉字字模的存贮芯片也安装在插件板上，这样的汉卡兼有汉字输入和输出的功能。目前我国已研制了多种用于某些汉字编码输入方案的汉卡，如联想式汉卡、五笔字型汉卡等，年产量总数达到十万块以上。

(3)中文印字机。中文印字技术是计算机中文信息处理技术的重要部分。印字质量的好坏直接影响到中文信息处理系统的使用效果。目前所用的中文印字机，按所采用的记录方式分成两大类：一类是撞击式记录，另一类是非撞击式记录。前者适于 100 点字/秒以下的低速印字，价格比较低，适合作微型机和终端印字；后者主要用于数百至数千行/分的中高速印字。中档的可作终端设备，高档的一般用于脱机印字设备。目前国产的中文印字机以针式中文打印机和热感式中文印字机为主，都用于印制 24×24 点阵的汉字。智能式的中文印字机并带有汉字库。这类印字机的年产量达 5 千台以上。简易激光印字机国内的产量尚较少，主要用进口散件装配。当今，为满足中文信息技术应用不断发展的需求，进一步发展中文印字机的研制与生产是一项重要的任务，

(4)中文终端。它是一项重要的中文信息处理设备，其用途非常广泛，同时也是国内目前产量最大，国产化程度最高的一项设备。由于中文终端基于微型机技术构成，这项技术近年来在中国发展很快，并掌握了不同的技术档次，可以设计生产不同性能规格要求的中文终端设备，如简易型、基本型、智能型以及具有多用户分时功能的工作站终端等，能够形成一定的型谱系列。

(5)中文字处理机。它是在微型机的基础上设计、研制成的一个汉字处理设备。近几年来，由于政府机关、工厂企业、事业单位打印文稿的需要，利用文字处理机取代老式的中文打字机，需求量大，因此生产和应用发展很快。据统计，目前国内生产的中文字处理机有 15 种之多。其中有相当一部分是和外商合作开发生产的，也有一部分是国内自行研制和生产的。年产量达到 6 万台以上。由于生产成本在不断下降，而在技术上不断有所创新，这项产品有很好的发展前景。对中文信息处理技术的普及应用可产生不小的推动作用。

5. 中文信息处理技术的标准化

任何一项技术的标准化工作都十分重要，它关系到这项技术能否得到迅速推广应用和今后长远的发展。在中文信息处理技术发展到一定阶段时，要及时地制订出有关的技术标准。

进入 80 年代以来，国家对中文信息处理技术极为重视，先后颁布了一系列的中文信息处理标准，有力地支持和推动了这项技术的发展。国家已颁布的中文信息处理标准如下：

(1)GB1988-80，“信息交换用七位编码字符集”。

(2)GB2311-80，“信息处理交换用七位编码字符集的扩充方法”。

(3)GB2312-80，“信息交换用汉字编码字符集(基本集)”，共包含 6 763 个汉字。

(4)GB3453-82，“数据通讯基本型控制规程”。

(5)GB3454-82，“数据终端设备(DTE)和数据电路终端设备(DCE)之间的接口电路定义表。”

(6)GB5199.1～5199.2-85，“信息交换用汉字 16×16 点阵字模集及数据集”。

(7)GB5007.1～5007.2-85，“信息交换用汉字 24×24 点阵字模集及数据集”。

(8)GB6345.1～6345.2-86，“信息交换用汉字 32×32 点阵字模集及数据集”。

(9)GB5261-86，“文字和符号图形设备的增补控制功能”。

(10)GB7589-87，“信息交换用汉字编码字符集第二辅助集”，共包含 8 836 个汉字。

(11)GB7590-87，“信息交换用汉字编码字符集第四辅助集”，共包含 8 836 个汉字。

第一、三、五辅助集分别为基本集，第二、四辅助集为繁体字版本。

目前正在研究制订的有汉字内部码标准与汉字内部码的数据格式；还有双八位汉字编码标准。