中央广播电视大学
继续教育教材

中文信息处理技术

赵珀璋　张淞芝　主编

宇航出版社

中央广播电视大学

继续教育教材

中文信息处理技术

赵珀璋　张淞芝　张炘中
姚天顺　董振东　陈秀群
编　著

宇航出版社

内　容　简　介

本书全面通俗地介绍了中文信息处理技术的基础知识、基本原理、基本设计要领及其使用方法。全书与电视录相配套，也可自学。其内容分别介绍了中文信息处理概念、汉字编码输入方法、中文信息代码概念、汉字字形存储及信息压缩、汉字设备、多文种信息处理概论、中西文兼容信息处理、中文信息网络通信、汉字识别、汉语语音识别及语音合成、汉语自然语言理解、机器翻译、中国少数民族语言文字处理、中文信息处理应用系统。在附录中附有国际标准ISO646、ISO2022中译本节录，还有中国国家标准GB2312—80、GB5007.1～5007.2—85、GB5199.1～5199.2—85、GB6345.1～6345.2—86的有关节录。

这是第一本中文信息处理技术的电视录相教材，可供高等院校、大中专院校师生作教材，也可供从事计算机和中文信息处理技术的研制、生产、开发和使用部门的技术人员及管理人员使用或参考，还可供管理部门的在职干部及技术人员使用与参考。对普及与提高中文信息处理技术，对中文文字改革及语言研究、对与世界各国有关中文信息处理技术的研究、开发及交流，将会起促进作用。

中文信息处理技术
编著：赵珀璋　张淞芝
张炘中　姚天顺
董振东　陈群秀
责任编辑：廖寿琪
•
宇航出版社出版
北京和平里滨河路1号
邮政编码：100013
新华书店北京发行所发行
各地新华书店经销
常州武进凹凸印刷厂印刷
•
开本：787×1092　1/16印张：36　字数：818千字
1990年5月第1版第1次印刷　印数：0001—5000
ISBN—80034—344—8/Tp·022　定价：16.20元

序　言

中文在我国已沿用了几千年，对中华民族的统一与文明发展有着不可磨灭的贡献。中文信息处理技术是近十几年来发展起来的新学科，也同样会影响到中华民族的未来发展。

自电子计算机问世以来，对人类社会、科学、文化起着巨大影响，尤其是在语言文字信息处理领域影响更大，从而对人类社会的其它领域都发生链锁式的影响。计算机起源于西方，当时曾有人说，计算机只能处理西文而不能处理中文。从七十年代末开始，我国从事计算机信息处理的科技工作者与语言文字学界配合，逐步探索研究并开发了一系列中文信息处理系统（包括少数民族语言文字信息处理系统），从汉字编码输入技术、汉字识别技术、语音处理技术等方面，到中文文字及语词处理系统、中文通讯系统、机器翻译系统、中西文兼容处理系统等，一直到各种应用处理系统：电子排版印刷系统、办公自动化系统等，已从一种专门技术逐步形成一门各学科交叉的和正在完善的专门科学技术。这门科学技术是以计算机科学、语言文字学、历史学、心理学、逻辑学、生物学、数学、物理数、电子学、声学、光电学、信息处理技术与通信技术等各种基础学科为基础，而综合发展起来的，而且还在发展中不断完善。这门科学将推动我国的信息处理产业的进一步发展，这是不容忽视的。

有人说：新能源、新材料、信息是构成当今物质社会三大要素，而信息则是这三大要素的基础与先导。那么中文信息处理科学技术在我国应是信息科学的基础与先导。

中国科学技术协会、中国中文信息学会（咨询服务中心）、中央广播电视大学、清华大学电教中心、东北理工大学继续教育中心联合录制的这套《中文信息处理技术》电视讲座将有利于中文信管处理科学技术的推广与普及，这部电视教材深入浅出，它将在教学中不断修改完善，希望能逐步形成一套较正规的教材。

钱伟长

一九八九年十一月十七日

前　言

计算机中文信息处理、包括字形信息处理、语音信息处理、词汇信息处理、语义信息处理及语言信息处理（如自然语言理解、机器翻译等）等学科，它是信息处理科学的一个重要领域。

中文信息处理技术，不仅涉及到计算机体系结构、操作系统、程序设计语言、数据库和网络通信技术，还涉及到语言文字学、语音学、词汇学以及应用科学技术，比如文字识别和生成、语音识别和合成、排版印刷技术等，它是一门高度综合性的科学和技术。

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，还应包括蒙、藏、维、哈、朝、彝、壮、傣、苗等几十种民族语言文字信息处理。显然，中文信息处理实质上是多文种信息处理。从信息处理编码技术（字符集编码、字形集编码、语音集编码、词汇集编码、语法集编码和语义集编码等），到系统软硬件、应用软硬件以及输入输出设备和技术的设计研制，都不能只考虑汉字信息处理，而应考虑两种或两种以上文种信息处理，即多文种信息处理。

中文信息处理，既有许多理论问题，又有许多技术问题。比如中文信息处理编码体系结构、操作系统、程序设计语言和数据库的多文种兼容问题，中文输入方法（编码输入、文字识别输入和语音识别输入等），中文输出方法（编码输出、字形显示和印刷输出以及语音合成输出等），中文网络通信（字符通信、字形通信以及语音通信等），机器翻译（中翻外、外翻中和多语种互译等），自然语言理解（词汇切分、语法、语义直至语用形式化等）、系统设备和应用设备等设计、制造和应用问题。这些问题大都没有很好解决。

为此，中国中文信息学会受中国科协之托，邀请中国中文信息学会常务理事张淞芝高级工程师、常务理事赵珀璋高级工程师、理事姚天顺教授、理事张炘中副教授、理事董振东研究员以及清华大学计算机系陈群秀讲师等六位专家讲授《中文信息处理技术》电视讲座，并编写了本教材。本教材共分十四章，一、二、五、十四章由张淞芝编写；三、六、七、八、十、十三章由赵珀璋编写；四、九章由张炘中编写；第十一章由姚天顺、陈群秀编写；第十二章由董振东编写。

中国中文信息学会理事长钱伟长教授为本书写了序。

本讲座自开始筹备以来，得到了社会各方面的大力支持，在电视录相与书面教材编写的过程，尤其得到中国中文信息学会咨询服务中心姜德存副主任的全力支持，并进行大量编审工作，才促成了该项工作的顺利进展。同时也得到了宇航出版社的大力支持。廖寿琪副编审对本书提出了宝贵的建议并做了大量编审工作。还得到了刘润松同志的大力支持。没有他们的支持和帮助，本讲座是不可能实现的。在此一并表示感谢。

该讲座的编写过程中是几位作者分头编写，因此全书连贯性较差，不太紧凑，风格不太一致，我们谨向读者致歉。由于时间紧和水平所限，书中错误在所难免，敬请广大读者指正，不胜感激之至。

编　者

一九九〇年三月

目　录

第一章　中文信息处理概论

1.1　信息和信息处理概念

1.1.1　广义的信息和信息处理

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。人们在很早就已知道信息这一概念。从定性的意义上说，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一切方面都有着很大的重要性。但是，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上。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。电气通信技术的发展，从电话电报开始，直到传真、电视，从有线通信发展到无线通信，直到微波、光纤通信、卫星通信，信息的传递速率大大提高，效能也大为改善，但只限于传递信息。信息技术的另一种手段为处理技术。本世纪40年代发明了电子计算机，开始时只是利用它处理数值运算。但是很快就意识到利用数据代表广义的信息，发展了数据信息处理这一意义深远的应用技术。利用计算机处理数据信息，不只是作单纯的信息传输，而主要是对信息按某种规律或作某种意义的加工，使它适应某种特定目的的需要。例如，气象预报中的信息处理，结合信息传感技术，对采集到的原始信息按预先设计的数学模型进行处理，得出的结果可以作为气象预报的资料。对信息进行的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，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上发展，在整个国民经济生产中占有愈益增大的份额。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重要的作用。

广义的信息涉及多种范畴。例如，一些自然现象所包含的各种信息；人类社会活动，如政治、经济、军事、文化、商业等活动所产生的各种信息；科学技术和生产活动，如揭示自然和物质结构的奥秘，从事地质研究、探矿等产生的各种信息。它们涉及到人们生存的环境，和从事科研、生产、生活等活动的一切方面。在这些含意丰富的信息中，信息的表示形式又是多样性的。例如，信息可以有数据、文字、声音、图形、图象等多种形式，这称为信息的多元化表示。或者说，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。

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。这要归功于微电子技术和计算机技术的飞速进步。微电子技术的进步体现在超大规模集成电路的技术水平日益提高，各种大容量存储器芯片和具有复杂逻辑运算功能的集成电路芯片日益增强，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进步，例如，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和普及应用；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，如图象信息处理、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。

以上简单介绍了计算机在传统的信息处理技术方面的进展情况。另一方面，计算机也具有通信功能，这就是利用数据通信技术实现的计算机网络通信。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内容，可以称为广义的信息处理技术。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，称为计算机和通信技术。

现代化通信技术的信息载体是综合的多元化信息，即包括数据、文字、语音、图形、图象（图象又包括静止的和活动的图象）。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。

图1—1示现代化通信技术所包含的实质内容。

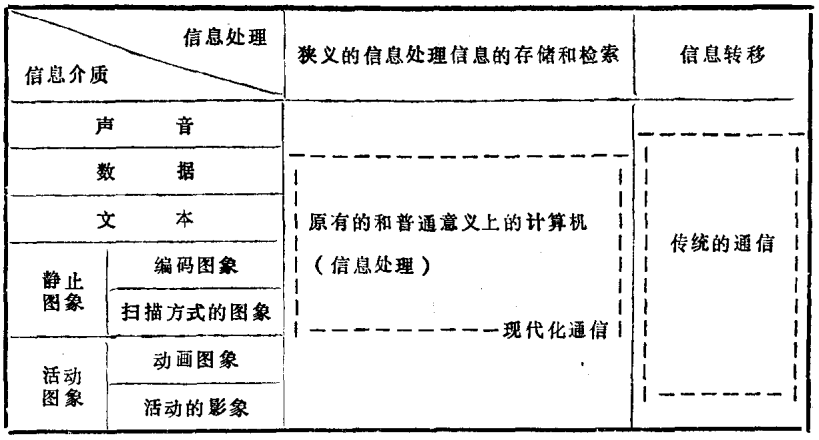

图1—1 现代化通信技术的内涵

近20多年来，数据通信技术有了很大的发展，这是和计算机通信技术的发展分不开的。60年代初期，开始发展计算机和计算机之间，或计算机和终端设备之间的直接互连通信；60年代中期发展了由通信子网（由通信处理机互相连接而成）和用户资源子网（由用户计算机构成）构成的计算机网络通信体系；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趋向通用化和标准化，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，大大扩展了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，即在同一个或相邻的多个建筑物的距离内，可以较高的通信速度实现信息通信和资源共享，也可以在上述距离范围内实现电子邮件传送。这对于目前正在发展的办公自动化技术提供了很大的方便，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。

现代化通信技术在近几年和未来若干年内又有新的发展。由于数字化通信技术的优越性，并且可以充分地利用计算机信息处理技术所具有的优点，今后不论何种类型的信息，要利用脉码调制（PCM)技术，把模拟信号转换成数字信号后进行传输和处理。并利用正在发展中的综合服务数字网络（ISDN）技术，建立一体化的通信体系。在这一目标下，可以把原来为适用于各种通信业务而建设的电报通信网，公共电话网，专线电话通信网，数据通信网等，合并成一个ISDN通信体系。在交换转接技术上，充分利用分组交换技术的优点，给多元化信息的混合传输提供了方便。在通信媒体方面，可以实现多种传输媒体的交换和转接。例如电束，微波信道，光束，以及卫星通信等传输媒体的综合利用（见图1—2）。从而把现代化通信技术的功能提高到一个新的水平。

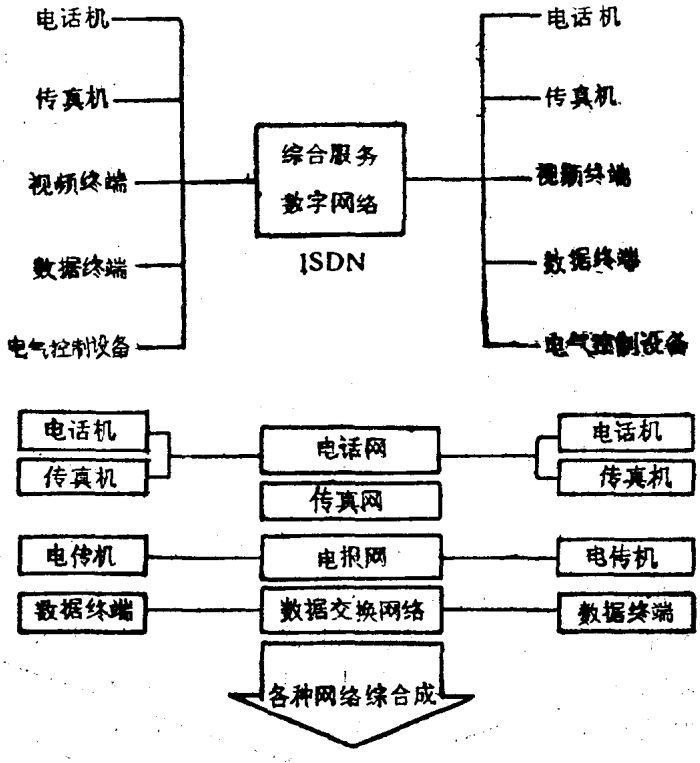

图1—2示综合服务数字网络通信技术的示意。

1.1.2 文字信息处理

前面已经提到，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、各种印刷出版物等，绝大多数都使用文字的形式来记录。文字也是一个国家或民族文化的象征，在社会和历史的发展中它有着特殊的地位。计算机从处理数据发展到处理文字信息，代表应用技术上的一个重大进展，否则计算机的应用将局限在一个较狭小的范围内。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，办公室自动化，凡是需要用文字表达信息的应用场合，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，以这类计算机为基础构成的文字处理机目前已有了很大的发展。文字处理机依据其应用的不同要求，可以设计成不同的档次。使用最为普遍的一种是便携式的文字处理机，或称为电子打字机，使用范围正在日益扩大。和传统的机械式打字机相比，电子式打字机具有编辑功能丰富、灵活的独特优点，并且可以提供一定数量的文件存档，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。如高级的文字处理机利用计算机人工智能，在字、词处理的基础上，有可能增添语法和句法处理，书面自然语言处理等新功能。随着高技术的开发和工业生产的发展，文字处理技术的推广应用前景是很乐观的。

文字信息处理的实质，是先把文字信息数字化，即用一个固定的数码代表一个字母或文字。例如，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，一般是以一个整字作为文字信息处理的单位，因此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(Code)。在计算机内部处理文字信息时，就象处理数据一样对待。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。

计算机所以能有高的运算和处理能力，是由于它利用了电子技术处理或执行二进制数运算这一法则。计算机中的运算器，利用半导体器件的二个状态（通和断）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。这样就能够高速地执行二进制数的数值或逻辑运算。实际上，计算机无论作数值的或任何种类信息的运算或处理，最基本的运算操作，就是这种二进制数的演算。

在本节中先讨论英文信息的处理。英文信息处理技术中，要考虑以下各种字母、数字和一些必须用的符号，它们是：

A，B，C，…，X，Y，Z，共26个字母，包括大写和小写形式，共52个。

0，1，2，…，9，共10阿拉伯数字。

+，－，×，÷，＝，＞，＜，…，！，？，＊，〔，（，｛，…，共32个图形符号。

用于计算机动作控制的控制符号，共34个。

以上共计128个字母、数字、符号的总和，统称为字符。对于这些字符，应制定统一的字符代码标准，以便各种不同型号的计算机系统都遵守这一标准，从而使各个计算机系统之间能够互相交换信息。对于字符代码的标准，在60年代已由美国国家标准局制订了美国国家标准信息交换码（英文缩写为ASCII．这是一种用七位二进制数表示的代码。七位二进制数共可作出128种编码（$2^7=128$），正好分配给总数为128个字符）。实际上对于每个字符使用一个字节（BYTE）的信息量。一个字节包含八位二进制数，实际使用其中的七位，尚留出一位，作为每个字符信息的奇偶校验用。

美国标准的ASCII码，国际标准组织（ISO）规定依据它制定作为英文字符编码的国际标准，即ISO646。这对于世界各国的计算机产业界从事计算机设备的工业生产，信息处理技术的国际化、通用化提供了依据。中国在1975年由当时第四机械工业部颁布

依据ISO646制定的七单位字符的编码标准（代号为GB 1988），其中除了个别货币符号有了改动外，其余内容完全相同。

文字信息处理的全过程大致包含如下三个环节：

（1）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。

（2）文字信息的处理。文字信息处理包括多种不同的处理要求。例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、改操作，有对若干个字、整个句子、或整段的增、删、改操作。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。

（3）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，如图1—3所示。这样的字形点阵信息和计算机中二进制数的存储相对应：即有笔画经过的点用二进制数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，加一些图形符号，总数在94个字符。用容量不大的存储器芯片，即可解决全部字符点阵信息的存储。计算机输出处理结果时，根据每个字符的代码，容易计算出字形信息在器中的存储地址，按照这一地址读出字符的点阵信息，供显示器或印字机输出。

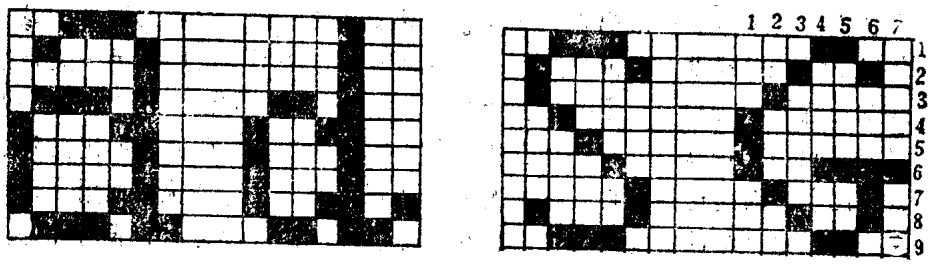

图1—3　英文字符的点阵化表示

显示器和印字机是用于输出信息处理结果的设备，它们也是一种人机界面，输出的结果应符合人的需要和习惯。对于字符显示器，标准的显示格式是每帧显示24行，每行80个字符。印字机的种类较多，目前使用较普遍的有针式打印机和简易激光印字机。针式打印机是一种普及型的印字机，简易激光印字机具有较高的印字质量。关于显示器和印字机的技术特性在第五章信息处理设备中作详细介绍。

1.1.3　中文文字信息处理的特点

一般情况下，中文文字指在中国广泛使用的汉字。要用计算机处理汉字信息，必须从汉字的特点说起。

一、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。西文的特点是用少数结构简单的字母用线性规则排列组成单词。

汉字不仅构成的笔画多，而且它是一种二维结构的图形，比起西文单词的线性排列结构来要复杂得多。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和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。这些工作虽然已得出一些成果，但未能达到满意的程度。也就是说，由于汉字独特的字形结构，不容易把它们拆分成一些由基本笔画、字根或字元的简单（线性）组合。因而，也不易得到简明的编码规则。这就是汉字输入编码问题难度大的根源所在。汉字的字量大，据统计，中国的汉字总数超出六万个。但是，不同的汉字在不同历史时期，不同的专业领域中使用时，其频度的差别是很大的。按照中国在1974年，对国内使用的现代汉字综合使用频度的统计，要求复盖率达到99.99%的情况，所需要的汉字量约在六七千个左右。中国在1979年制订颁布的《信息交换用汉字编码字符集　基本集》（国家标准代号为GB 2312-80）中，共收容了6763个汉字。这个数量的汉字，就是根据上述对现代汉字综合使用频度的统计资料作为依据而订定的。6763个汉字中，又分成两级，第一级为常用汉字，共3755个；第二级为次常用汉字，共3008个。对6763个汉字用计算机技术加以区分，按最小信息冗余度的原则，需要用13位二进制信息（$2^{13}=8192$）。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。

二、汉字输入技术

由于汉字字量大，字形复杂的主要特点，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，不仅这种键盘的造价低，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。

目前，汉字编码方法的种类很多，仅是国内提出的汉字编码方案就有五百种之多。然而，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。

利用字符键盘通过汉字编码的输入方法，不论编码方案的技术指标有多高，一般来说，其输入速度和计算机的信息处理速度相比，总是很低的。因此，用键盘输入汉字的环节，过去对它有“瓶颈”之称。除了利用键盘输入汉字的方法外，近几年来，由于计算机硬、软件技术的进步，若干种智能化的输入方法开始得出研究成果，有的已开始走向实用化。例如，联机手写汉字识别输入，在图形输入板上写入汉字，可以不按严格的笔顺次序，计算机可以对输入的汉字加以识别，给出它的标准代码。但这种输入方法的速度决定于手写汉字的快慢，并且不能潦草，因此速度并不高。另一种智能化的汉字输

入方法是光学汉字识别，目前主要是对印刷体汉字，原稿上的印刷体汉字经光学扫描后，经二值化处理送入计算机，由程序把送入计算机的字模信息和原先存在计算机中的标准字模信息进行比较，判定和识别输入的汉字，这种方法的识别速度较高。例如使用16位的微型计算机，识别速度可达每秒8～15个汉字。识准率在95%以上，初步达到实用阶段。另一种智能化的输入方法是汉语语音识别输入。用标准普通话的汉字发音，结合词汇输入，经计算机识别后，给出相应汉字的代码。目前普通微型机能识别的汉语词汇量达1000个以上。经改进可望在若干年后实现如声控打字机，能接受汉语输入的汉字终端等功能。因而，在中文信息的输入技术上，可以有多种选择，相互配合形成一种较完整的输入体系。

三、汉字字形的存储

前面已经指出，汉字结构不仅笔画多，而且它是基本笔画或字根的二维空间组合，除了对汉字编码造成困难外，也对汉字字形的存储提出较高的要求。计算机中存储汉字字形，也是用点阵方式来表示。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(详见第三章附录)。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的字模库存储容量。在发展汉字信息处理技术的早期（70年代中、后期），由于当时集成电路存储器芯片的容量较小，价格也贵，汉字字模的存储曾经是中文信息处理技术的一个棘手问题，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。80年代以来，特别是近几年内，由于半导体超大规模集成电路存储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ROM（只读存储器）芯片，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，收存全部国家标准基本集（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。上述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，须要设计32×32、40×40、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。

以上介绍的是通用型的汉字字模，主要用于印制一般的中文文件、报表。除了通用型的汉字字模外，尚须要考虑有很高文字质量的精密型汉字字模，它们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，即使目前存储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。

四、汉字的输出技术

和字符的输出要求相比，输出汉字字形要求输出设备的分辨率较高。对于汉字显示规格，目前最常用的是15×16点阵的汉字字模。为了和英文字符的显示格式相兼容，每行显示40个汉字，一帧24行汉字，若加上1～2行提示信息，一帧的总行数为25～26行。因此，要求显示屏的分辨率为640×420以上。若要显24×24点阵的字模，则显示屏的分辨率必须达到1000×700点。

汉字印字设备，常用的分辨率有7.1、9.4、11.8、15.7点／毫米(180、240、300、400点／时）。通常7.1、9.4点／毫米属于低档印字机品种，如针式汉字打印机，热感式汉字印字机；11.8、15.7点／毫米属中、高档印字机，如简易激光印字机，液晶开关式汉字印字机等。

五、中英文兼容技术

在计算机系统技术方面，要考虑系统能输入、输出，并能处理中文信息。从原则上来说，可以独立地设计一个专用于处理中文信息的计算机系统。这是因为，不论是英文字符，或是中文的汉字信息，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；而汉字则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设备和软件，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，我们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，可称为中英文兼容技术。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、软件功能。在此基础上，再增加中文信息处理功能，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，汉字的代码(即汉字信息交换码)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作分别的处理。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ASCII）扩充而来。ASCII共包括94个字符，用二个ASCII交叉组合成汉字信息交换码，共94×94＝8836个，汉字基本集实际使用了其中的6763个。它们都是七位二进制信息表示的代码，所有的区别是，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，则对二者便无法加以区分。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