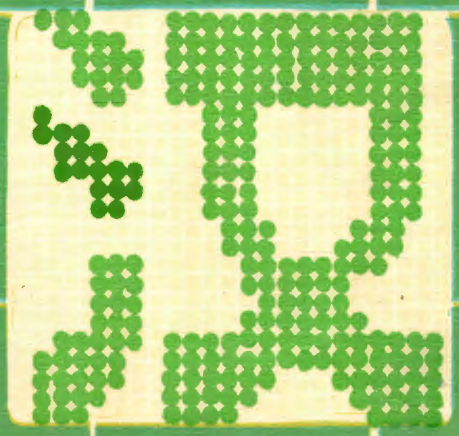
郭平欣　张淞芝　主编

汉字信息处理技术



国防工业出版社

装帧设计：
杨庆英
正文设计：
郭　林

科技新书目　110-113
统一书号　15034·2973
定　价　7.30　元

目　　录

汉字信息处理技术

郭平欣　张淞芝　主编

国防工业出版社

内 容 简 介

电子计算机是当今新技术革命的先导技术。而在我国，要实现电子计算机的普及应用，还必须解决计算机输入输出的汉字化问题，也就是说，要妥善解决汉字信息的计算机处理问题。只有这样，才便于使计算机技术被我国大众熟悉和掌握。

本书系统而全面地阐述了汉字信息处理技术。全书共分十六章，其内容大致可为四大部分：汉字信息处理技术基础；汉字输入输出设备和技术；汉字信息处理系统的基本软件和系统软件；汉字信息处理技术的应用。本书是由电子工业部计算机工业管理局组织我国有关专家编著的，力图做到理论与实际结合，尽可能反映国内当前水平。

本书既可作为从事计算机科研、设计、生产、使用与维护的工作人员的参考手册，也可作为大专院校师生的教科书。对于企、事业单位的领导干部、管理人员以及想涉足计算机应用领域的业余爱好者，本书也可起到一定程度的响导作用。

汉字信息处理技术

郭平欣　张淞芝　主编

责任编辑　张均武

*

国防工业出版社出版

新华书店北京发行所发行　　各地新华书店经售

国防工业出版社印刷厂印装

*

787×1092 1/16　印张34 3/8　插页2　794千字

1985年12月第一版　1985年12月第一次印刷　印数：0,001—8,500册

统一书号：15034·2973　　定价：7.30元

主　编

郭平欣　张淞芝

执　笔　第一章　张淞芝

第二章　郑易里

第三章　扶良文　胡宣华　王绪龙

第四章　陈跃星　向维良

第五章　蒋攒平　陈瑞源　华一满　胡春年　王绪龙

第六章　毛德行

第七章　朱子龙　周根林

第八章　毛德行

第九章　周根林　傅大林　单启成　雍殿书

第十章　陈胜凡

第十一章　郭瑞枫

第十二章　朱进业　陈瑞源

第十三章　王　选　陈堃铓

第十四、十五、十六章　陆大绚

审　稿　刘　源　周根林　吴克忠　龚滨良　石运程　李约瑟

责任编辑　张均武

前　言

信息是存在于客观世界的一种事物形象。凡是物质的形态、性能在时间或空间上的变化，以及人类社会的各种活动都产生信息。千万年来，人类用自己的感觉器官从客观世界获取信息，藉以认识和改造客观世界。人类对信息的运用需要有载体。信息科学认为，语言就是一种信息载体，有了语言，人们就能描述和解释直接或间接的感觉、知识等各种信息，并进行人与人之间的信息交换。但是在古代，这种以声音作为信息符号的载体，只能在同一时间、同一地点进行信息交换。在信息用语言表达过之后，除了可能在人类的记忆系统中有所存储以外，立即会消失。因此，人类在语言的基础上又创造了文字这种交换信息的载体。这种载体的显著优点是可以反复阅读，不受时间、空间的限制，并且可以作出优化，经多次传递而不易发生差错。因此，文字是促进人类文明、加速社会向高级阶段发展的有力工具，它是人类社会历史发展过程中创造出来的精神财富。

自从科学技术发展以后，人类可以借助各种工具、仪器获得用直观方法不能得到的客观世界的各种信息。比如用望远镜获得宇宙间宏观世界的信息；用各种显微镜、原子能仪器获得微观世界的信息。特别是无线电通信和电子技术的发展，产生了可以运载声音、文字和图象的新的信息传送技术和载体，人类就能够以更高的效率获得更为丰富的自然信息和社会信息。

第二次世界大战结束后，科学技术以更快的速度发展，从而促进了经济的飞速发展。据悉，现代的科学技术知识有百分之九十是从本世纪五十年代以后积累起来的。过去人类对于信息的处理（包括采集、存储、回忆、思维、判断和决策）只能依靠大脑进行，处理的效率很低，可能提供的信息量很少。自从有了电子计算机技术，并把它用于信息处理以来，电子计算机便成为强有力的工具，使信息处理从人工时代进入了自动化时代。所谓信息处理自动化是指对于科学研究、技术开发、工业制造、农业开发、邮电交通、商业、财政银行、文教卫生和行政管理等部门日常产生和积累的数据、图表、文件、报告等信息，借助计算机自动地加以分类、编目、存储，检索、分析、综合、打印输出和提供咨询服务。现代社会中的一切领域，若不能及时掌握信息，就无法有效地进行工作。由于计算机技术的高度发展，通信系统的日趋完善，信息自动化已逐步变为现实。计算机的广泛应用，特别是数据库和计算机通信网络的普遍建立，使社会进入信息化时代。信息的重要意义和作用已逐渐为人们所认识。信息正在成为除农、林、牧、渔等可再生资源和矿业等非再生资源以外的人类社会的第三种重要资源。有目的地开发和利用这种资源，人类社会便会加快新的技术革命。

用计算机处理的信息，包括数据、文字、图形、语音等，其中主要的是文字信息。我国是个多民族的国家，大多数人民属于汉族，因此我国使用的文字主要是汉字。新中国成立后，承认全国各民族一律平等，反对民族沙文主义和狭隘民族主义，因此有必要区分“汉字”和“中文”的提法，不能把它们混同起来。中文应包括少数民族的文字。

我国要推广计算机应用，必须使各民族的文字都能在计算机上实现输入和输出。这些民族的文字中，只有汉字是一种象形文字，其它的都是拼音文字。拼音文字的字符最多不超过数十种，而现代我国使用的汉字仅常用部分就有六、七千字，而且字形复杂，这就使计算机实现汉字信息处理存在较多困难。如果解决了汉字信息处理的技术难题，对于我国其它民族的文字处理就比较容易解决。

目前，我国计算机的应用范围日益扩展，除了上述的情报资料自动化管理和数据库系统外，还有书刊、报纸的自动编辑排版，文字处理，公用事业的咨询服务，旅馆服务，医疗诊断，企业管理，工业自动控制，计算机辅助设计，以及办公室自动化等，在我国开展这些应用项目是很迫切的。社会需要是技术进步的强大动力，计算机的推广应用是计算机事业发展的目的。要在我国推广计算机的应用，必须要妥善地发展汉字信息处理技术。

我国最早有目标地开展“汉字信息处理技术”的研究是在一九七四年八月。当时，由电子工业部、新华通讯社、文化部出版局等几个单位发起，得到国家计划委员会的大力支持，并正式批准开展这一课题（定名为“七四八工程”）的研究。当时，根据现实条件和国内的实际需要，确定汉字信息处理的研究课题包含以下三个项目：

（一）精密汉字编辑排版系统

这是一个用于书刊、报纸编辑排版的专用系统。把计算机技术用于出版印刷行业，实现“无纸编辑”（借助荧光屏实现汉字文稿的加工和校对），发展“冷排”技术（逐步取代铅字的人工检排），这对于减轻编辑排版的劳动强度，改善排字车间的劳动条件，提高排版速度和效率，缩短出书周期，加速出版印刷行业的技术革新，将产生深远的影响。

（二）汉字情报检索系统

国家的经济发展依赖于科学技术的进步，而科学技术情报是促进科技发展的重要手段。长期以来，我国的科技情报管理是很落后的。由于完全靠人工方式，不仅所能提供的信息量很少，而且很不及时，因此，远不能满足使用者需要。此外，还会造成大量的图书和情报资料积压，利用率很低。如果利用计算机技术及时地加以收集、加工、编目、评价和存储，并按照使用者的要求，提供检索查询的方便，那么就可以大大提高科技情报资料的管理水平，提高情报资料的利用率。这项技术的实现和在科技情报资料管理部门的推广应用，对我国科学技术的发展具有重大的意义。

（三）汉字通信系统和汉字终端设备

这项课题的目标是要建立汉字通信网络并研制发送和接收汉字信息的汉字输入输出终端。汉字信息处理技术发展到一定程度后，通信网络技术无疑是很重要的，只有把计算机技术和通信技术结合起来才能更好地发挥汉字信息处理技术的功能和效用，提高信息源和设备的利用率。

上述三项研制任务，在不太长的时间内，都已在不同程度上取得了成果。特别是在七十年代中、后期，在国内器材条件很困难的情况下，能取得较好成果是令人欣喜的。汉字精密照排系统，在某些技术指标上达到了国际先进水平，形成了我国自己的特色。这一事例说明了我国有能力在开展汉字信息处理系统的研制工作中，发挥我国对汉字特点和使用规律掌握得最彻底的优势。从七十年代后期至今，汉字信息处理技术在国内得到

了迅速的发展，如汉字编码和输入技术、汉字字模存储技术、汉字输入输出设备、汉字终端技术、汉字西文兼容的软件技术和汉字系统应用等，在研制水平上都有很大的提高。同时在器材和设备条件方面，也有了明显的进步。例如在汉字系统中广泛地采用中、大规模集成电路，微处理机技术，以及软盘和温式磁盘等外存储器，从而提高了系统的性能价格比，缩小了体积，这就向实用化方向前进了一大步。特别是微型机汉字信息处理系统和用微型机控制的汉字显示终端设备，取得的研制成果更为明显。这也符合我国发展电子计算机以微、小型为主，以微型机的推广应用为重点的方针。可以预期，在今后一段时期内，在我国具有汉字功能的各类业务处理系统、管理系统、办公用计算机、咨询服务系统、数据库系统、局部网络系统、文字处理机，各类汉字终端设备等的应用会更迅速地发展。

汉字信息处理技术，除了计算机硬、软件技术外，还涉及其它的研究领域。例如：对汉字属性的研究，剖析汉字的字形、字音特性，对于优化汉字编码技术是重要的依据；对汉字使用频度的统计，确定常用字、次常用字、稀用字、罕用字，并确定各级字的收容范围，可供系统装备汉字字模时选择；进行汉字组词特性的研究和作词频统计，这有利于掌握组词的特点及其使用规律，从而可以提高汉字信息处理的效率。此外，对汉语自然语言、语法和语义的研究，可使汉字信息处理技术向深度发展。

在汉字信息处理技术的开发过程中，必须重视建立标准化的工作。过去几年内，在这方面已取得了一些成绩。例如已颁布了《信息交换用汉字编码字符集——基本集》。目前正在进一步制订辅助集和汉字系统控制功能码标准。此外，也在制订数字化汉字字模标准。这些标准的建立和实施，对于这项技术的进一步研究开发、设备研制和推广应用是很重要的。除了这些基础的标准外，也要重视汉字设备生产的工业标准，设计系列化产品和模块化结构，以利于用户的选择使用。使汉字设备和系统的研制生产成为我国计算机工业生产体系中的一个重要组成部分。

面临着信息化时代的到来，汉字信息处理技术要适应当前形势，在研究工作上不断深入，积极开展推广应用，使它在我国的社会主义建设事业中取得巨大的技术和经济效益。

为了适应当前国内汉字信息处理技术飞速发展的形势，为了满足广大读者对全面而系统地介绍汉字信息处理技术专著的迫切要求，电子工业部计算机管理局特组织和约请了有关专家编写了此书，力图使本书能对有关人员有一定的参考价值，对推广我国的汉字信息处理技术起某种程度的推波助浪的作用。

在本书的组织和编写过程中，得到了有关部门和单位的大力支持，得到了华北终端设备公司和无锡电子计算机厂的热情资助。在此，一并表示衷心的感谢。

由于汉字信息处理技术涉及面广、包含的专业门类多，因此参加该书编写和审阅的人员也较多。这样就给统编工作带来了困难。尽管我们在确定编写大纲、审定内容和统编全稿等阶段作了努力，但由于时间仓促和水平所限，书中一定会有不少缺点和错误，敬希读者批评指正。

国际信息学会常务理事　郭平欣

一九八四年八月

目　录

第一章　概　论

第二章　汉字属性

第三章　汉字输入编码方法

第四章　信息处理交换用的汉字代码

第五章 汉字输入方法和设备

第六章 汉字字形发生器

第七章 汉字输出设备

第八章 汉字显示终端

第九章 汉字信息处理系统的配置及基本汉字处理软件

第十章 汉字数据处理的系统软件

第十一章 汉字情报检索系统

第十二章 汉字联机通信网络

第十三章 精密汉字编辑排版系统

第十四章 汉字企业管理系统

第十五章 中医诊疗系统

第十六章 其他应用

第一章 概 论

1.1 汉字信息处理的意义和任务

1.1.1 什么叫汉字信息处理

对于“信息”一词，目前尚有多种定义。其中的一种定义是：信息（information）是各种事物所发出的消息、情报、指令、数据和信号中所包含的表征该事物的内容。随着人们对客观世界认识的日益深化，确认信息和物质、能量三者，是构成客观世界的三大要素。信息对于人类社会的重要性，表现在除了可再生资源（如动、植物）和非再生资源（如矿物）以外，信息是维持人类生产活动、经济活动和社会活动的第三种资源。信息具有多种性能，例如可传输性，可转换性，可存储性，可处理性，以及可再生性等。随着科学技术的发展，信息的传输效能愈益增强，其作用范围也愈益宽广。例如，由于电子通信技术（特别是光纤通信和卫星通信技术）的发展，信息传输技术从电话、电报发展到传真、电视，包括声音、文字、图形和图象的传送，从而大大增强了通信效能。但是，这种技术的发展还只限于实现信息的传输。约在本世纪六十年代，电子计算机这项重大的科学技术成果在非数值计算领域内得到推广应用。由于电子计算机不仅能存储和控制信息，更重要的是由于它能加工或处理信息，因此相应产生了信息处理（information processing）这一新的概念。这里所指的计算机是电子数字计算机，所处理的信息是量化信息，即相应于二进制数码“0”和“1”的各种组合所代表的数字信息。因此，有人把用于信息处理的电子计算机称作信息处理机（information processor），这是比较确切的。本书中所指的信息是人类所特有的信息，这就是文字或代表这种文字的语音所包含的信息。随着计算机系统功能的不断提高，应用领域的迅速扩展，信息处理的概念、涵义、作用和涉及的范围也大大扩展了。特别是文字信息处理所包含的内容更加丰富了，例如：情报资料和图书的自动编目和检索；书刊和报纸的自动编辑和排版；事务处理；企业管理；办公室自动化；文字处理；文字翻译；医疗诊断；公用咨询服务；数据通信等。实际上，文字信息处理技术已逐渐渗透到人类思维、生产和生活等活动的一切方面。文字信息处理技术同科研、生产实践、社会活动、生活环境等的联系日益密切，以至人类社会的一切活动几乎都有它的用武之地。以电子计算机为基本手段的现代信息处理技术，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。这项技术的发展速度和应用水平已成为人类进入信息化社会、国家走向现代化的一个重要标志。

信息处理技术中，对文字信息的处理称为文字信息处理。本书的主题就是文字信息处理中的汉字信息处理。

事实上，计算机系统只能处理数据，而数据所表示的意义就是信息。因此，本书中讨论的信息处理，体现为对数据的处理。表示文字信息或符号信息的数码，称为代码（code）。例如，在对西文字母以及符号的处理中，对应于26个字母（分为大写和小写体）

和一些常用符号，按照某种规律和约定，编成一组数码，这组数码称为字符代码（如我国国家标准GB1988，或国际标准ISO646七位代码，EBCDIC八位代码等）。因此，对文字信息的加工，就是对代码数据的加工。概括起来，可以把文字信息的处理过程分成三个阶段：

（1）信息的输入。通过输入设备把文字信息转换成代码，并送入计算机。

（2）信息的加工或处理。根据各类不同的应用，借助预先设计好的程序对输入的信息进行加工和处理，从而得出结果信息。

（3）信息的输出。通过输出设备把以数据代码形式表示的结果信息，复原成文字。

科技发达的西方国家，目前已相当成熟地应用文字信息处理技术。究其原因，除了这些国家对计算机系统的设备、技术开发得较早，掌握得较早以外，还有另一个重要原因，这就是这些国家采用拼音文字，拼音文字的字母数量少，字形简单，从而容易实现对文字信息的处理。我国所用的文字主要是汉字，汉字是一种表意文字，字量多，字形复杂。从以后的讨论中可以知道，这两个特点使汉字输入方法和建立汉字字模库的工作遇到不少困难。为了构成一种汉字信息处理系统，在硬设备方面，除了需要一些通用的设备（例如通用电子计算机及其外部设备）外，还需配备汉字输入输出设备。在软件方面，要使系统软件具有适应对汉字处理和西文处理两者兼容的能力。因此，这个课题难度较大。目前我国在汉字信息处理技术领域内已经取得了可喜的成绩。我国是汉字的发源地，我国人民最了解汉字的使用特点，因此，在研究和开发汉字信息处理技术方面，理应作出更大的贡献，使其在我国的四化建设中显示出强大的生命力。

1.1.2 汉字信息处理技术涉及的范围

一、汉字属性（attribute）有关的内容

汉字信息处理技术是一项综合性的技术，其核心是计算机技术。因为计算机处理的对象是汉字信息，因此，为了合理地制定一些计算机处理汉字的技术规则，先要研究有关汉字的一些基本特性（又称为汉字属性）。它大体上包括以下几个方面：

（一）汉字字量

汉字是表意文字，或称象形文字（ideographic character），它的每个字有其特有的形状和构造，这是不同于各种拼音文字的一大特色。在使用中，所用汉字字量的多少是一个重要问题。我国的汉字字量多至五、六万。目前实际应用的汉字，据1981年颁布的我国《信息交换用汉字编码字符集——基本集》（即GB2312）中所收字量，一级字有3755个；二级字有3008个，共计6763个。根据需要，今后尚需制定扩充集。一个汉字信息处理系统中究竟应该收容多少汉字字量，这应根据实际的使用要求来确定。

（二）字形分解

汉字字形是汉字属性中的一个重要项目。汉字分解后，其基本组成部分有部首、字首、字根、笔画、位点。可以把位点看成是组成字的最小单位。分解字形是为了找出汉字的结构规律，以便为汉字信息处理技术在字形信息存储方面提供依据。例如在建立汉字库，特别是在建立用字根合成（又称向量组字法）的汉字库的过程中，字形分解显得更为重要。通过对字形的分析研究，可以选取最少数量的字根，合理地组成所需的汉字，从而可以改善经济性，提高效能。此外，在基于字形特征的汉字编码方法中，为了得到

高性能的编码方案，更需注重字形分解的研究工作。

（三）汉字字体（style）

在用于印刷排版的汉字处理系统中，对字体种类的要求较高。一般而言，汉字字体至少可以分为宋体、仿宋体、楷体、黑体等四种，而每一种又有方体、长体、扁体的区别。一般的汉字信息处理系统对字体种类并无特殊要求，甚至只需具备一种（例如宋体字）即可。

（四）使用频度（frequency）

对于不同的汉字，其使用频度的差别是很大的。同一汉字在不同专业领域中使用时，其频度也有差异。因此，对于不同专业的字频，要分别进行统计。一种综合频度是在若干有代表性的专业领域中统计出各自所用汉字的频度后，求取其平均值得到的。根据汉字的使用频度，可以把汉字分为常用字、次常用字、稀用字、罕用字等几个等级。在建立不同种类的汉字处理系统中，必须根据使用频度来选用字库中所收容的汉字。

（五）汉字发音（pronounciation）

每个汉字有它的标准（普通话）发音。目前国内主要推广拉丁化的汉语拼音，作为它的发音属性。汉字发音特性的研究，对于在计算机系统中按音序检索汉字，或以发音特性作为汉字编码的研究工作是很重要的。

（六）汉字索引（indexing）

可以从不同角度检索汉字，例如以笔画、偏旁或部首来检索，也可以用汉字发音的音序来检索，或以国标交换码区/位号检索，还可能有其他索引方法，其目的都是要以简捷的法则准确地查得某个汉字或它的标准编码。

（七）汉字排序（sequencing）

与西文文字相比，汉字排序是一个较复杂的问题。汉字可以用笔画的多少排序，也可以用汉字拼音排序，或者以汉字的综合使用频度排序。无论用哪一种方法作出的汉字排序表，记忆和掌握起来都较困难。在国标《信息交换用汉字编码字符集——基本集》中，汉字的排序是综合部首、音序和频度三者确定的。汉字的序数对于汉字信息处理技术来说是很重要的。例如，目前国内研制的大多数汉字信息处理系统，都是采用国标码的汉字序数作为内部码的。在汉字情报检索系统或汉字数据库系统中，文献名或属性项的排序，是以构成这一文献名或属性项的汉字序数作为依据的。

（八）汉字标准交换码（Chinese standard exchange code）

1980年，我国颁布了《信息交换用汉字编码字符集——基本集》（即GB-2312）。GB-2312是和GB1988兼容的。它用两个七位码代表一个汉字，用以作为计算机系统之间汉字信息交换用的标准代码。自从公布了这一标准码后，它便成了汉字的一项属性。

以上列举了一些主要的汉字属性项目。在实际应用中，可以根据需要加以增删。对汉字属性进行较彻底的研究，掌握汉字的基本特性和应用规律，才有可能合理地设计出各种类型的汉字信息处理系统。

二、对汉字词组及文句结构的研究

除了汉字属性外，为了更有效地研究汉字信息，需要对组成的字或字组（称为词）进行研究。所谓词是指经常使用并有特定含义的单个汉字、或多个汉字的组合。词的属性包括词的种类、组词字数、词的使用频度、词的含义、排序特性等。在汉字信息处理

技术中，对词的研究是很重要的。在汉字输入方案中，对于使用频度特别高的词，可以用软件方法设定，用一个键位代表一个词，也可以根据需要改变某个键位所代表的词。在汉字情报检索和数据库系统的应用中，需要使用主题词或关键词进行存储或检索。在这类系统中，要求把词排序组成词典，存入系统的存储器中，以便建立索引或进行查找。

汉字信息处理属于中文信息处理的范畴。严格地说，中文信息不仅包括汉字信息，也包括国内各少数民族语言和文字信息。而实际上，有时这两个名称互相混用了。若从意义和信息处理功能上分析，汉字信息处理包含对汉字本身的处理，例如汉字在系统中的输入、输出，以及汉字的编辑；中文信息处理的范围除了对汉字信息处理外，还应包含对少数民族语言、文字的处理，即使只考虑汉字，尚需包括对汉字文件中的句法和上、下文结构的处理。显然，后者比前者在含义上深刻些，所涉及的范围也要宽广些。例如，中文信息处理要涉及信息处理的更高级形式，这就是对自然语言（natural language）的处理，要使计算机能够理解语言的内容。而语言的内容又有两个方面，即语法和语义。语法是指语言的结构格式；语义是指语言代表的信息内容。计算机必须对这两者具有分析和综合能力。当计算机达到能处理中文自然语言时，其应用前景就更为广阔了。这样一来，就可以实现诸如计算机翻译（外文译为中文，或相反）系统，情报、文献的自动管理系统，用途广泛的咨询系统等。

三、汉字信息处理和计算机技术有关方面

信息处理离不开计算机技术。对于用作汉字信息处理的计算机系统，所用的硬设备除了通用电子计算机和通用的外部设备外，还需要增加汉字输入输出设备，如汉字键盘、汉字字模库、汉字显示终端、汉字印刷机等。通用外部设备中，作为外存用的磁盘子系统、磁带子系统，这些和通用计算机系统中所用的外存设备并无区别。根据汉字系统规模的大小和应用特点，可配用各种适当的外存设备。

系统所配备的软件，除了通常的操作系统，高级语言编译程序外，还需要有汉字服务程序（Chinese service program）（或称汉字管理程序），它是汉字信息处理系统中软件的一个重要组成部分。此外，还要配备面向汉字信息处理系统作业任务的应用程序。实际上，西文信息处理系统所用到的一些技术，在原理上都适用于汉字信息处理系统，只是汉字信息处理系统所要处理的字符范围扩大了，除了西文字符外，又增加了数量大得多的汉字。并且，汉字信息处理系统要利用西文信息处理系统原有的如高级程序设计语言等软件资源，使技术复杂程度增加了，这给汉字信息处理技术的实现和完善造成了一定困难。

除了系统方面的技术外，汉字输入方法是汉字信息处理技术中的一个重要课题。目前，输入汉字信息主要用键盘方法，完全依靠手工操作，故其效率低，速度上远不能和计算机的运行速度相适应。利用计算机直接识别汉字字模或汉字语音的方法，目前国内尚处于研究阶段，离实用尚有相当距离。汉字识别（Chinese character recognition）又分为光学汉字识别（optical Chinese character recognition）和联机手写体（on-line hand described）汉字识别两种，它们有着不同的用途。汉字语音识别（Chinese speech recognition）是利用计算机直接辨别汉语语音。目前国内已有的成果只能识别分离的汉字、字组或词的语音，尚不能识别连续的汉字语音。这些都是今后必须努力解

决的课题。它们对于实现汉字输入方法的多样化，并从根本上改变汉字输入方式和提高输入速度和效率有着深远的意义。

汉字信息处理系统除了输出打印或显示结果的汉字外，还可以用汉字语音的方式输出，这就需要研究汉字语音合成（Chinese speech synthesis）和输出方法。

综上所述，汉字信息处理技术所涉及的范围是很广的，必须分别解决各方面的课题，才能使汉字信息处理技术水平不断提高。

1.1.3 汉字信息处理技术要解决的问题

由于汉字的特点是字量大，字形复杂，因此，要建立一个汉字系统，就需要解决汉字的输入、存储和输出等问题，这在实现上要比实现西文信息处理系统更困难。此外，因为西文信息处理技术已开发了不少成熟的软件，因此，要把这些软件用于汉字信息处理系统，并使计算机能兼容西文和汉字两种文字的信息处理功能，则还有许多软件方面的工作要做。以下就分别对汉字信息处理技术要解决的问题作些说明。

一、汉字输入编码和汉字键盘

虽然可以用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。

对于汉字编码输入方法，可以粗略地将其分为两大类。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，例如按照字形特征编码的，称为形符法；按照汉字发音特征编码的，称为音符法；有形音结合编码的；也有按照汉字的其他特性编码的。这些种类繁多的方法，简称为汉字编码输入法。

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；没有重码问题。但它所用的键盘是通称的整字键盘，其体积大、造价高、输入速率低。整字键盘有两种主要形式。一种是早期使用的主辅键式〔或称移位键(shift key)式〕汉字键盘，盘面上布置有约400个键，每个键上收容9～12个汉字，双手操作，用辅键来确定字键上9～12个汉字中的某一个字。由于这种键盘体积大，造价高，故不易推广使用。另一种目前流行的整字键盘，是一种所谓笔触式（pen touch）汉字字盘，盘面上可收容3000～4000个键位，每位一字。这种整字键盘的体积可以做得较小，其造价低于前一种。因此，它在不少应用领域内有推广应用价值。对于整字汉字键盘，除了盘内字以外，还要解决外字（相应于二级字或非常用字）的输入问题。目前，多数用户采用直接送入汉字代码，或者用汉字字根组合编码的方法实现盘外字的输入。

汉字编码输入方法很多，以下仅就汉字形符编码法、音符编码法、形音结合符编码法，以及联想式编码法作简要介绍。

（一）汉字形符编码输入法

现仅举出以下两种作简要介绍：

1．笔画（stroke）编码法　它是按组成汉字的基本笔画（五种或八种）编码的。在输入组成一个汉字的笔画时，按照特定的先后次序进行。这种方法的优点是：输入一个汉字就象书写这个字一样容易学习操作。整个汉字的输入码是组成该字基本笔画代码的组合。这类编码方法的输入码较长，而且是不等长的。

2．偏旁、部首〔或称字根（radical)〕编码输入法　这类方法按所选定的汉字偏旁、部首的种类不同，又可分成多种。例如，采用传统的汉字偏旁、部首作为编码输入的单元（其总数可达200多个)。要实现这种汉字编码的输入，需要设计专用的字根键盘。为了利用普通字母数字键盘作为汉字编码输入的工具，已发展多种方法。这些方法所选用的字根是特殊的，其数目大大减少，使能收容在普通字母数字键盘的键位上。例如仓颉汉字编码，选用了特殊的24个汉字字根。其特点是：对输入键盘要求低，但学习掌握这类编码输入方法需要熟记一套编码规则，难度较大，从而一般只适合于专业操作人员。

（二）汉字音符编码法

它是以汉字发音为基础的编码输入方法。这种方法也包括很多种类。例如，其中之一是按照较早流行的汉字四角号码发音规则来编码的。又如，国内近几年来发展的声、韵双拼编码法，是以汉字普通话发音的声母、韵母为基础的。在双拼方案中，分别以相应的字母代表各个声母和韵母来作为音符。这类编码输入法都可以用字母数字键盘作为输入工具。汉字音符编码方案的一个特点是由于汉字同音字多，因此重码率一般较高。如何区分重码字，这是这类方案的一个重要问题。

（三）形符、音符混合编码法

这种混合编码法利用了汉字形、音两种属性的各自特点，从而在区分汉字上有可能节省信息或得到其他好处。有的混合编码方案中，除了利用了形、音特征外，还添加了其他信息，例如汉字的字义。

（四）联想式汉字编码（conceptualized Chinese character encoding）输入方法

这是一种联机操作的汉字输入方法。这种方法事先把经常在一起使用的汉字代码链接起来。在输入操作时，只要键入为首的一个汉字的编码，或者用某一串汉字所共有的起始笔画作为索引，就可在荧光屏上显现相关的一串汉字。然后，再借助光笔或键盘按所需次序加以挑选，就完成了在系统中输入这些汉字的目的。这种借助人-机对话方式的汉字输入方法，较易学习掌握，有利于普及应用。

对于上述各种汉字编码输入方法，除了有的用常规的偏旁、部首作为字根的编码方案，需用专门设计的字根键盘输入汉字外，其他一些编码方案都可以利用普通的字母数字键盘来输入汉字，这样就可以和西文信息处理系统的输入键盘完全兼容。使用字母数字键盘输入汉字的好处是，操作人员在熟悉了某种汉字编码输入方法之后，可以实现盲打，从而可以得到较高的输入速度。这对于专业操作人员是比较适合的。它们共同的缺点是：存在重码问题；此外，由于用编码方法输入汉字时，按键次数多，输入码较长，信息的冗余度较大，因而在输入机器后需要转换成码长为两个字节的汉字内部码。

由于用不同的输入方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，此称为标准汉字交换码。我国已经颁布了作为国家标准的GB2312《信息交换用汉字编码字符集——基本集》。目前国内研制的大多数汉字信息处理系统都采用国际码作为内部码。在基本集中，除了有两级（包括常用和非常用）汉字外，还有几种外文字母、数字和符号，其总数为7445个。所以，国标交换码中每个汉字或字母、数字、符号用两个字节(实际上只用了14位)来表示。

二、汉字字模的存储

（一）汉字字模的点阵规格

计算机信息处理用的汉字字模，按用途可以分成精密型字模和通用型字模两大类。精密型字模用于精密汉字编辑排版系统，这种汉字字模，对字形、字体、字号变倍等都有严格要求，必须适合出版印刷行业所定的规格，要求分辨率达到30线/毫米。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。目前，国内正对通用型汉字点阵字模制订国家标准，在字模存储设备和输出印字设备能达到的技术条件下，要求有较高的文字质量。通用型的汉字字模，常用的点阵（dot matrix）结构有以下几种：

简易型：15×16点；

普通型：24×24点；

提高型：32×32点。

（二）存储整字字模信息的汉字字模库

一个汉字系统必须设置一个存储汉字字形信息（Chinese ideographic information）的存储体系，此称为汉字字模库(Chinese font library)或汉字字模发生器（Chinese font generator）。

对于汉字字模存储媒体的要求主要有两点：一是单位存储量的成本要低；二是读取信息的速度要快。在普遍采用大规模集成电路存储器（如EPROM，ROM或RAM）以前，主要采用磁心存储器和磁盘存储器来存储汉字字模。磁心存储器虽然可以提供较快的读取速度，但是由于包括译码和驱动线路在内，其单位存储量的成本高，而且体积大，消耗电流大，工作时产生的电磁干扰也大，故不能推广应用。用磁盘存储器（包括硬盘和软盘）存储汉字字模，虽然单位存储量的成本较低，但由于从磁盘中随机读取汉字字模信息时，平均等待时间较长，故字模输出速度低。近几年来，大规模集成电路存储器由于价格不断降低，故已成为一种较为理想的汉字字模信息的存储媒体。但由于目前它的单位存储量的成本仍高于大容量磁盘存储器，所以一般做法是把常用字（例如国标码中的一级汉字，或其中的一部分）存入大规模集成电路存储器，而把非常用字（如国标码中的二级汉字）存入磁盘存储器。这是一种两级存储体制。因为使用频度高的字模是从高速集成电路存储器中取出的，所以平均来说，仍可得到较高的字模输出速度，而整个汉字字模库的成本又不致于过高。对于少数要求汉字输出速度特别高的实时系统，例如作为高速电传通信用的汉字终端，可以考虑全部用集成电路存储器存储汉字字模。

（三）存储压缩信息的汉字字模库

对于通用型汉字字模，除了采用上述整字存储的方式外，还可采用信息压缩(information compression）的存储方式。汉字信息压缩的原理和技术有多种，例如：采用只收存汉字字根的汉字字模库；利用向量组字法产生汉字字模；用霍夫曼树型压缩法存储汉字信息等。由于用向量组字法和字根字模库相结合的汉字字模库技术，可以得到较大的信息压缩比，故目前已得到较广泛的应用。利用这种汉字字模库技术来组成和输出某个汉字时，只要用地址链的方法把组成该汉字的有关字根（这些字根是用坐标、斜率、长度等原始信息记录的）从存储器中取出，经过信息还原、比例尺寸选择、字根间相对位置的确定后，就可得出相应的汉字字模。这样的汉字字模库，可以较多地节省存储成本。特别是在要求收容的汉字字数较多时，可以不把汉字分级，不要用磁盘存储器作辅助存储，只需用数量不大的ROM或EPROM组件（一般在40～60K字节），就可以产生7000～20000个汉字。这种汉字字模库的缺点是：所产生的汉字字模质量不如整字存储

的汉字质量好。此外，由于需要用软件方法组成汉字，所以输出速度较低。

今后，随着ROM器件的集成度迅速提高和价格的不断下降，有可能普及完全用ROM器件作存储媒体的汉字整字字模库。在字形和点阵规格标准制定后，以较大数量成批制作这样的汉字字模库更可降低成本。上述字根式汉字字模库，还可进一步改善字形质量，提高软件功能和组字速度。这种汉字字模库在某些应用场合，仍然有其使用价值。

三、汉字输出技术和设备

汉字输出技术主要包括汉字显示和汉字打印。由于汉字字模的点阵较西文字符的点阵密，因此，对于显示器和印刷机的技术要求，自然要比显示和打印字符数字的同类设备高些。

（一）汉字显示技术

汉字显示目前主要采用荧光屏显示技术。对于汉字显示器的技术指标，主要是一帧画面能显示的汉字数，这和荧光屏的分辨率（resolution）、显示器的视频通带等指标有关。目前，一般都采用光栅扫描的显示方式。通常，一帧画面的字数在500～1000范围内，字数愈多，要求荧光屏的分辨率愈高。但是，这是有一定限制的。例如，对于显示12行×40字的情况，若显示的汉字点阵为15×16，则要求荧光屏的分辨率[1]为320×640点；如果汉字点阵为24×24的情况，那么，即使每屏的显示字数降为360字（30字×12行），仍要求分辨率为480×800点以上。若要求再增多显示字数，或要求使显示器兼有图形显示功能，则需要再提高显示器的分辨率，但同时也会提高显示器的成本。此外，视频范围同一帧显示的汉字数、汉字点阵多少，以及显示器的帧频等成比例增加，但有一定的上限值。汉字显示器的刷新存储（refresh memory）方式有两种：一种是缓冲存储汉字在字形发生器中地址的刷新方式；另一种是对画面信息逐点缓冲存储器的刷新方式。后者可以兼顾对图形的显示要求。

（二）汉字印刷机

对于汉字印刷机，主要的技术指标是打印分辨率和印字速度、用纸要求等。目前对通用型汉字字模的印刷机，其印字分辨率在4～12线/毫米的范围内。对于根据不同的印字原理和机制所制作的各种类型的汉字印刷机，其印字速度可以在每秒数十字到数千字的很宽范围内变化。按照印字速度，可以把它们分成低速、中速、高速三档。

低速档：其印字速度为每秒数十到上百字。属于这档的印刷机有针式汉字打印机、热感式汉字印刷机、喷墨式汉字印刷机等。其中，针式汉字打印机是目前应用得最多的一种，其分辨率为4～6线/毫米。

中速档：其印字速度为每秒数百到一千字。它包括简易型激光扫描汉字印刷机、发光二极管静电转印式汉字印刷机等，其分辨率为5～8线/毫米。

高速档：其印字速度为每秒一千字以上，最高可达每秒上万字。它包括高精度激光扫描汉字印刷机、发光二极管（LED）或光纤管（OFT）转印汉字印刷机等，其分辨率为8～12线/毫米。

---

[1] 荧光屏的分辨率系指满屏所有的扫描线数。通常认为相邻扫描线之间是连续的，因此满屏的线数指连续的扫描线数目。但在文字点阵的显示中，为了要分清相邻的两个位点，以一个亮点间隔一个暗点来计算扫描线，所以通常说的分辨率就相当于满屏扫描线数目的一半。纵向扫描线和横向扫描线数按荧光屏的宽度和高度比例计算，例如320×640指横向扫描线有320条，纵向有640条。此外，荧光屏中心部位同四个角上的分辨率是不相同的，而上述数字一般是指中心部位的。

在上述各种类型的汉字印刷机中，针式打印机虽然其技术指标较低，但是由于结构简单、成本低、使用维护方便、维持费用低，所以被目前绝大多数汉字信息处理系统(特别是微型机、小型机汉字信息处理系统)所采用。对于目前流行的汉字针式打印机，可根据汉字字模点阵规格，设计成16针头和24针头的，将其交叉排成两纵列。其打印速度为40～60/秒。另一种所谓梳齿状汉字针式打印机，它的多个针头间隔成一定距离并排成横列，每个针头可以在一段距离的几个汉字的区间内移动。而纵向点阵则由走纸动作形成。这种针式打印机的输出速度可达100字/秒以上。

某些微型机汉字信息处理系统，也采用普通7针或9针的字符打印机。它用较少的针头往返打印两次或三次来形成一行汉字。

激光扫描汉字印刷机是一种较新发展的机种，它在印字速度和分辨率方面都可以达到较高指标。这种印刷机使用普通纸，纸幅可以在较大范围内变化。在降低造价和提高工作可靠性的情况下，可望广泛配置在小型或中、大型机汉字信息处理系统中。

四、汉字系统在软件方面要解决的问题

对于一个完善的汉字系统，不仅要求它具有汉字的输入、输出以及汉字文件的编辑等功能，而且还要求它兼备字母数字处理功能。也就是说，应要求它把汉字处理和数据处理结合起来。理论上说，汉字处理和字母数字处理这二者是不会发生矛盾的，但是事实上，要使它们能很好地协调，并非一件易事。下面来分析一下原因。

(一) 汉字、西文兼容性的要求

应该指出，电子计算机的基本语言并不是西文，而它只能接受0和1两种信号，计算机的运行总是要受这种信号的组合〔称为机器码(machine code)〕的控制。由于计算机技术首先在欧美国家得到发展，因此西文字母、数字、符号先为计算机所接受，接着又用缩写的西文字来替代机器码，发展成汇编语言 (assembly language)，然后进一步发展成各种较接近人类自然语言（西文）的高级程序设计语言。这些都是软件工具方面的进展。经过较长时间的积累，目前已有大量同系统关系密切的软件，特别是各种高级程序设计语言，都是针对西文系统设计，并用西文表达的。我国在发展汉字信息处理系统软件时，有人主张应从操作系统开始，设计出把各种汉字设备资源都考虑在内的操作系统；在程序设计语言方面，主张应该遵照汉语规则，程序语言要完全中文化，因此也计划设计出中文程序设计语言的编译系统。这样做当然在原理上完全可以实现，但却要花费大量的人力、物力和时间。更为重要的是，这样的中文计算机软件系统同国外迅速发展的软件技术不相适应，不能共享国际通用的软件资源，其应用也就有很大的局限性，这条路径看来是走不通的。另一条道路是国内目前多数人所主张的，这就是，鉴于上述历史情况，在发展汉字信息处理系统软件时，应当尽量继承和利用已有的西文系统软件，达到汉字系统和西文系统兼容 (compatible) 的目的。本书在这项技术方面的观点，是围绕后一见解阐述的。

但是，对于数量很大的汉字（更确切地说是汉字的国标码)，如果将其加到已配有西文软件体系的计算机环境中，并同原有的西文字符系统协调一致实现信息处理，那就要妥善解决使系统能分辨字符代码和汉字代码，使高级程序设计语言能适应处理汉字信息等一系列问题，也就是说，要解决汉字和西文在一个计算机中兼容的问题。这主要通过软件手段来解决。

（二）汉字信息处理软件技术的发展过程

以下简要介绍国内目前在计算机系统中实现汉字、西文信息兼容的解决方法。

（1）早期(七十年代中期到七十年代末期)所用的汉字处理软件方法比较简单。这就是借助于操作系统软件以外的“汉字输入输出管理程序（Chinese character I/O management program)”（有人称它为“汉字输入输出驱动模块)，来完成汉字输入输出的功能。当用户程序运行到要输入输出汉字信息时，可调用这一软件模块来实现所需的操作。由于这一软件模块未纳入操作系统，所以只能在用户程序中调用这一模块。但由于数据处理和汉字信息处理是分别进行的，因此其运行效率较低。此外，由于在这种汉字信息处理系统中，要间接地调用汉字信息处理软件，因此，编写用户程序的工作较为繁琐。

（2）八十年代初期，开始研究直接利用高级程序设计语言处理汉字信息问题。这一时期，由于开发了微型机汉字信息处理系统，而且许多微型机都配有BASIC语言，所以对采用BASIC语言扩充汉字信息处理功能的工作做得较多。此外，也有采用COBOL和FORTRAN语言来实现这一功能的。为了实现这一功能，主要是在高级语言中增加汉字输入输出语句，把原来的语言改造成“有汉字功能的程序设计语言”。这种在语言级上扩充汉字信息处理功能的方法，可以提高运行效率，简化编制用户程序的工作。

（3）一个系统中可以包含多种高级程序设计语言，但必须使各种语言都具备处理汉字信息功能，从而必须对多种语言逐个地加以修改和进行功能扩充。因为各种高级语言都由操作系统支援，因此可以通过修改操作系统来使它具有处理汉字信息的功能。但是，这并非一件简单易行的工作。特别是对于一些中、大型计算机上的操作系统而言，由于其规模大、结构复杂，故对其修改和扩充起来比较困难。并且，由于受原系统软件设计上的一些限制和各个不同系统固有特点的约束，故没有统一的修改和扩充操作系统的方法。由于微型机的操作系统一般比较简单，因此，近几年来，对微型机操作系统（例如CP/M、MS-DOS或UNIX）的扩充工作进展较快，而且已取得了令人满意的效果。通过扩充操作系统功能，把原来的汉字输入输出管理程序的功能同原操作系统中西文字符的输入输出驱动模块的功能结合起来，使系统处理汉字信息像处理西文字符一样方便，较妥善地解决了汉字西文兼容问题。

（三）区分汉字和西文字符的标识问题

不论是通过扩充高级程序设计语言，还是扩充操作系统的功能来实现西文和汉字信息处理的兼容性，都要使系统具备识别字母数字和汉字代码的能力。按照GB1988码，字母数字用7位码表示，占一个字节；按照国标交换码GB2312，汉字用两个7位码表示，占两个字节。为了能区别它们，就要使这两种代码有各自的标识。目前，在实际应用中，采用了多种区分字母数字和汉字代码的方法。

第一类方法是利用汉字国标交换码中的第8位。若两个字节中的第8位都为1，则标志这两个字节组成的是一个汉字代码；若该位为0，则所代表的是一个字母或数字代码。这种方法把汉字代码纳入字符代码的数据类型，可看成是字符代码的扩充。虽然这种方法有一定的优点，但是在某些情况下也存在一定的限制。例如，在使用汉字终端和主机系统通信的场合，由于每个字节的第8位必须用作奇偶校验位，故不能再把它作为汉字代码的标识了。此外，在使用EBCDIC 8位码的系统，无法再利用第8位来标识。

第二类方法是在一串汉字代码的前和后，加上起始符和结束符，这种符号称为标识符（identifying code）或引导符（leading code）。这种符号可以利用某个不常用的字符代码或功能码（function code）的组合来代表。这种方法的优点是可以用每个字节的第8位作为代码信息，从而使可表示的汉字数增多。但是，这又带来了下述一些问题：用了这一标识符后，在系统中可能引起意义上的混淆（即二义性），例如所选用的标识符可能和汉字代码数据的某种组合发生重码；标识符本身需占用存储单元（特别是在汉字和字母数字交替出现频繁的情况下）；在对汉字字符串作合并、分离等运算操作时，会造成一些困难。这类方法中，由于所用引导符的方法不同，又可以分成很多种。

以上所举的两类标识汉字和字母数字的方法，分别适用于不同类型和不同用途的汉字设备或系统中。一般地说，用最高位的0或1来区分，这对汉字信息的加工运算来说是比较方便的。目前国内多数微型机汉字系统或汉字终端都采用这类区分方法。但是，它的应用有一定的局限性。例如在计算机联机网络中，为了实现信息的传输，需要加进奇偶校验，即最高位要作为奇偶校验用，在这种情形下，就不能再用它来标识汉字或字母数字了。通常可以在联机工作时，经过转换采用第二类标识方法，而进入主机或终端设备后，再转换到采用第一类标识方法。

（四）发展接插兼容汉字终端技术

前面列举的是通过扩充高级程序设计语言或操作系统来实现汉字和西文信息处理兼容的软件方法。事实证明，对于微型机系统来说，特别是通过修改和扩充其操作系统达到这一目的是有效的。对于一个大的计算机系统来说，即使能改动和扩充它的操作系统，但因为其工作量很大，而且进口的各类中、大型主机的品种很多，需要逐个地修改它们的系统软件，因此困难很大。为了解决中、大型机系统处理汉字信息的能力，近年来提出了一种汉字信息处理系统插接兼容（plug compatibile）技术，其中包括有关插接兼容汉字终端或插接兼容汉字印刷机的技术。

以插接兼容汉字终端为例，由于这类终端是由微型机构成的，由微型机的操作系统容易扩充汉字功能，因此这一终端本身就是一个汉字、西文兼容的系统。终端系统的硬设备包括汉字键盘、汉字字模库、显示器和汉字印刷机。输入功能方面，汉字终端具有汉字码转换（从键盘码转换成国标码或内部码）功能；对汉字和西文字符添加标识的功能；能实现汉字、西文字符的混合编辑和屏幕显示。把加有标识的汉字和西文字符送入中、大型计算机系统，经加工处理后的信息送回这类终端，根据不同的标识对汉字和西文字符作出适当的处理后供显示或印刷输出。这样，对于中、大型主机系统，只要不致引起代码的二义性，不需要辨别汉字或西文字符，允许能在用各种语言书写的程序中写入汉字常量和注解，在文件系统中可以用汉字字符串书写文件。也就是通过接插兼容的汉字终端能使原来没有汉字功能的中、大型机系统进行汉字信息处理。由于终端和主机联机工作，因此终端尚需有主机和终端间控制码的转换功能，以及和计算机通信有关的数据链路功能。上述终端所具备的各种处理功能，主要用软件方法实现，但其中能固化的则尽可能装备成固件（firmware）的形式。发展这类汉字终端设备，就可以使更多的计算机系统为汉字信息处理服务。

接插兼容的汉字印刷机的基本原理同接插兼容汉字终端相似。

（五）汉字信息处理软件技术要达到的目标

具有汉字信息处理功能的数据处理（Data processing）要达到的目的，可以列举以下几个方面：

（1）在源程序中，可以包含汉字字符串或汉字常量；

（2）在源程序中，使用汉字或汉字、字母混合写出的注解；

（3）直接用操作系统或程序设计语言中的输入输出语句实现汉字或字母数字的输入输出操作。

如果再提出进一步的要求，例如要求在程序中使用以汉字定义的变量名，那么，这就要涉及到所用的变量名和输入的汉字字符串是否能在系统中明确区分，以及怎样才能避免混淆的问题。此外，对于高级语言中所用的一些保留字，能否用汉字替代，这是一个有争议的问题。如果能做到这一点，就能完全用汉字写出源程序，从而会使汉字信息处理技术更向前推进一步，即程序语句汉字化。当然有不少人认为，象保留字等的汉字化，并非必需。在一些微型机系统，对某些结构较简单的高级程序语言（例如 BASIC），国内已实现了这点。但并非普遍情况都能既实现语句的汉字化，又不损失原语言的全部功能。这些工作有待于进一步的研究，使汉字信息处理的软件技术不断达到完美的程度。

五、汉字终端技术

目前一般所指的终端（terminal）大多数是指微型机控制的显示终端。汉字显示终端对于汉字系统的配置来说是一项很重要的设备，许多大、中、小型计算机系统都少不了配用汉字显示终端。而多数情况下，汉字显示终端本身就是一个基于微型机构成的系统。

按照汉字终端的功能等级，可以把它分成简易型和智能型两种：

（一）简易型汉字终端

它具有汉字输入输出功能、汉字代码转换功能和汉字编辑功能，本身还带有汉字字模库。对于外部设备，为了降低成本，除了配备显示器外，只配用汉字输入键盘。它相当于西文字符系统灵巧终端的功能。它所用的微处理机可以是国内优选的 8 位或16位微处理机。

（二）智能型汉字终端

除了汉字输入输出，汉字编辑功能外，智能型汉字终端还应具有某些处理功能，例如文件管理功能、表格处理功能和图形处理功能等。它的硬件，除了配备简易型汉字终端所有的设备外，还应配备汉字印刷机、软盘机或温式磁盘机。软件方面，在这种终端中应配有操作系统、文件管理系统，一至两种高级程序设计语言。这类终端也可以脱机使用，依靠本身的软件资源作为独立的汉字工作站(Chinese character work station)。这类终端既可以用 8 位微处理机，也可以用16位的微处理机。

作为终端设备，都应具备通信功能。对于通信接口，目前大多采用 RS232C 串行接口标准，或其它相应的接口标准。考虑到远程通信的需要，应配备远程通信控制软件。

以上两种类型的汉字终端可看作是两种基本的类型。在此基础上，可以形成专用性质的汉字终端，例如：编辑排版系统用的校改汉字终端；情报或数据库系统用的联机查询终端；业务处理用的窗口系统服务终端；工业控制等实时系统用的测控终端；电传通信网络用的收发终端；数据采集用的汉字终端等。

终端设备应采用模块化、积木化结构，这样可以在基本类型的终端设备基础上，添加有关的模块即可得到所需的功能。

1.2 汉字信息处理系统的构成和分类

1.2.1 汉字信息处理系统的构成

汉字信息处理系统应包括硬件和软件两大部分。它和通常的计算机系统的组成情况相似。

一、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。主处理机是通用电子计算机。根据所要求的处理能力（包括容量、速度和信息吞吐量）和工作方式（包括脱机成批处理和联机工作方式等），可以选用适当的主机，它可以是大、中、小型计算机，也可以选用微型机。常规外部设备主要包括外存储器，例如各种类型和规格的磁盘机和磁带机。汉字外部设备包括汉字输入键盘，汉字印刷机，汉字显示终端设备等。

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。对它们的设置和连接决定了该系统的汉字部分的工作方式。

通常，汉字字模库可以设置在系统的三种不同的部位。

（1）汉字字模库作为一个独立的外部设备。例如，采用大规模集成电路存储器存放汉字字模，作为外部存储器，或者采用硬磁盘或软磁盘存放汉字字模。这些外存储器连接在系统的总线上。这种方法的优点是汉字字模库可以为多台设备所共享，特别是当汉字字模库的价格较高时，这种设置方法有利于降低系统造价。其缺点是，当输出字模信息时，要占用总线传送时间，从而会影响系统效率。对于一个较大的汉字系统，在要求经济性较好，而对工作速度等指标要求不高的情况下，这种设置方案是可取的。

（2）汉字字模库直接和汉字印刷机连接。如果系统中只配一台汉字印刷机，而这台印刷机的利用率又很高，就应该直接把汉字字模库和印刷机相连，并且可以构成智能印刷机。在这种情况下，系统总线只传送代码信息，而对于信息量大的字模信息则只在汉字字模库和印刷机之间的局部线路上传送，从而可以得到较高的效率。这样设置的汉字字模库，多数采用 ROM 存储器。

（3）汉字字模库设置在汉字显示终端上。目前，一般小型机以上规模的汉字系统，都配有汉字显示终端。主机系统在其本身未配备汉字字模库或汉字印刷机的情况下，可以通过与其相连的汉字终端输出汉字。

汉字终端和主处理机系统之间一般有并行和串行两种连接方式。

1．并行连接　这种连接方式又称外部设备型连接。如果汉字终端采用标准的 8 位微处理机，除了地址线和控制线外，另用 8 条数据线和主处理机系统相连，每次交换 1 字节信息。对于设在主处理机房内或距主处理机房近的汉字终端，通常是采用这种连接方法。

2．串行连接　这种连接方式又称为通信连接方式。这里，由于信息是逐位传输的，所需传输线的数目少，故适合于长距离或远程通信，其接口一般采用 RS232-C 标准或 RS422 标准。大多数汉字终端和主处理机系统之间，都是用串行方式连接的。当用于计

算机通信网络时，也采用这种连接方式。

二、汉字信息处理系统的软件

和通常的计算机系统相似，汉字信息处理系统的软件包括系统软件和应用软件两类。

（一）系统软件

其系统软件包括以下项目：

（1）能兼容汉字和西文信息处理的操作系统。它在保留通常西文系统全部功能的条件下，还包含各种汉字设备的驱动模块，并且，它还能直接调用汉字输入输出管理程序和汉字编辑程序。此外，系统所支持的各种高级程序设计语言也能识别和处理汉字信息。

（2）汉字输入输出管理程序。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，便根据输入码的不同编码方式，将其转换成标准汉字代码，供加工处理。当输出汉字时，先把标准码变换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。

（3）汉字文本编辑程序。它和西文编辑功能相似，汉字文本编辑程序除了具有对单个或多个汉字以及整句或整段的增、删、改的功能外，还具有换行和换段功能。对于功能强的编辑程序，还具有行首、行末禁则处理和行对齐，以及自动成页等功能。

（4）高级程序设计语言。汉字信息处理的应用程序可以用高级程序设计语言来编写。它们在操作系统支持下，可以调用相应语言的编译程序。这些高级程序语言在保持原有功能的情况下，还能识别出应用程序中的汉字字符串和注释，直接处理汉字信息。

（二）应用软件

根据汉字信息处理系统的性质和用途，各种应用项目都要有相应的应用程序。对于一些典型的应用程序，应提供商品化的应用程序包。由于应用项目种类繁多，设计各种应用程序是一项工作量很大的任务，因此，要尽量利用西文系统已有的一些应用程序包（例如数据库应用程序），使其在经过必要的改动后，能够适合在汉字工作方式下应用。这样便可迅速扩大汉字信息处理的应用范围。

（三）汉字文件系统

对于许多应用项目，建立文件系统是十分重要的。以汉字字符串写成的文件，包括如汉字情报检索系统或汉字数据库系统中必须建立的各种文档，汉字编辑排版系统中设计的作为排版格式依据的编排文件，翻译系统中用作文法和句法规则的文件等。

1.2.2 汉字信息处理系统的分类

从对系统功能和输出文字质量的要求上来区分，可以把汉字信息处理系统分成两种类型：即精密型汉字编辑排版系统（Chinese editing and typesetting system）和通用型汉字系统。前一种类型用于正式出版的书、刊、报纸的编辑排版；后一种类型用于汉字文件处理，统计报表，数值和数据处理等，这一类型的系统使用范围是很广的，汉字信息处理技术的推广应用很大程度上取决于这类汉字系统的发展。

一、精密汉字编辑排版系统

这一系统最重要的技术关键是高精度汉字字模的存储和版面输出。该系统的特点对汉字字模的点阵密度要求很高。如果要求达到分辨率为30线/毫米以上，那么，对于一个五号字（尺寸为3.675毫米²），就要求其点阵密度达96×96点。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量很庞大。此外，还要兼顾有适当的字模输出速度。因此，需要解决一种高倍率压缩信息的课题。这类系统的版面输出，可以采用两类技术，即电子束扫描（cathode ray scanning）输出和激光扫描（laser scanning）输出。在扫描的同时，输出版面在感光底片上记录成象。这种输出设备又称为照排设备（phototypesetting eguipment）。整个系统除了包括上述设备外，还包括：排版用计算机（一般用现代的小型计算机或16位微型计算机即可）；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（prove printer）；字模自动制作设备；图片输入设备；等等。

汉字编辑排版系统要配备大量专用的软件，它们包括：专用的操作系统；编辑排版专用语言及其编译系统；汉字文件系统；书、刊、报纸等各种版式的排版应用程序；图片处理软件等。

二、通用型汉字信息处理系统

它的特点是：主要用来实现数据处理或一般的汉字信息处理，其使用面广；力求系统的成本低；不需太讲究汉字字模的质量。通用型汉字系统字模点阵规格目前流行的主要有两种：15×16点；24×24点。对低于15×16点阵的字模，因质量太差，故目前已很少采用。这类通用型系统的汉字印刷机，目前以针式打印机为主，今后也可能推广采用简易型激光扫描印刷机。这类系统对于字号尺寸的变化，不象对精密型字模那样要求严格，故通常采用软件变倍或调节针头间距的方法便可解决。采用这些方法，只能变出很少几种尺寸，而且不一定符合出版印刷业定出的字号尺寸规范。

以下列举几种通用型汉字信息处理系统的例子：

（一）汉字情报检索系统（Chinese information retrieval system）

用于书、刊、情报资料的存储和检索等的自动管理系统，进一步可发展为汉字数据库（Chinese data base）检索系统。这类系统的特点是需配备容量很大的外存储器（如磁盘机和磁带机），以收容尽可能多的情报资料。一种类型是配置多台问答式联机汉字显示终端，供用户以询问方式向系统索取情报资料。这样的系统称为联机（on line）汉字情报检索系统。另一种是采用集中提问输入方式的系统，此称为批处理（batch processing）汉字情报检索系统。系统的响应或回答可以由荧光屏显示中间结果，或由汉字印刷机印出检索结果。

（二）企业管理系统（factory management system）

它用于大型工矿企业的生产管理、计划调度、行政管理、人事工资管理、设计图纸和工艺资料管理、产品管理、合同管理，以及供销计划管理等。工矿企业应用了计算机管理后，不仅可以减轻和节省人力，而且能大大提高管理效能，便于实行现代化的企业管理体制。

（三）事务处理系统（business processing system）

它用于如下一些业务：政府机关的计划拟订、公文管理、档案管理、统计报表制作；银行、金融机构的金融情报管理、财务管理，日常各类业务处理、报表统计；学校、教育机构的数学行政管理，学生的学籍档案和成绩记录管理；大型医院的医疗事务和病历管理；大型旅店的业务管理；公安机关的户籍管理和案犯档案管理等。总之，它主要面向各种不同的事务或业务管理。

（四）办公用计算机（office computer）

它不仅用来实现办公室范围内的文件和书信印制，还可用作简易的文件档案管理。此外，可以结合办公室业务的特点，专门设计一些应用程序进行其他业务方面的处理。对于文字处理方面的应用中一些常用的公文、书信格式、文摘等，可以预先将其收存在系统中，以便起草文件、书信时随时调用。对于编辑好的文件、书信等，可在系统中加以存档，以便需要时随时打印。这种系统又称为文字处理机（word processor）。办公用计算机系统大多采用微型机构成。这类设备也可和通信线路相连接，用作汉字电传的通信终端设备。

（五）汉字通信系统（Chinese character communication system）

汉字信息也可以实现有线或无线传送。不过，在线路上传送的是汉字代码〔因此，它属于数据通信（data comunication）〕。在接收端，再把汉字代码转换成字形输出。数据系统的信息交换中心是一个计算机系统，它用来控制数据流的传送（包括确定通路，选择信道，存储转发等功能）。在发送和接收终端可以采取加密、解密措施和纠错措施。这类通信既可以是终端和终端之间、终端和计算机之间、计算机和计算机之间的通信，也可以是远距离的数据通信，或是计算机局部网内的通信。有了数据通信技术，各类计算机信息处理系统可以发挥它们更高的效用，例如在建立了全国性的情报通信网络后，科技情报检索系统便可使边远地区的读者方便地向设在大城市的图书、情报中心查询所需的情报资料；各省、市的图书、情报中心可以相互间或向国家图书情报中心交换情报资料。

（六）窗口系统（cashier system）

它可供许多公用事业机关或各种服务行业等进行各类日常的业务处理。窗口系统的例子有：飞机、车船等订票用的业务处理终端；商店、批发部门制作发票、发货单据用的终端；银行存款、取款用的出纳终端；公用事业收费记帐用的终端，以及大型商店、超级市场用的售货终端等。

（七）文字自动翻译系统（word automatic translation system）

通过计算汉字信息处理系统把不同种类的外文译成中文。目前对于文法结构比较简单的外文资料（如一般的科技文献资料等），已可以部分实现用计算机自动翻译。对于语法和修辞等复杂的文章结构，计算机翻译系统需要具备语法和句法分析，上下文分析，词汇选择等大量复杂的程序，这属于中文信息处理自然语言研究课题的范围，尚需研究探索。

（八）其他智能系统（intelligent system）

利用计算机逻辑判断和数据处理能力，还可以构成除上述七类系统以外的其他应用系统，例如医疗诊断系统、服务行业咨询系统（consultant system）、计算机教育系统、电话查号系统、军事指挥系统，以及法律诉讼的推理分析系统等。

上述各类应用系统，有些是为了节省人力，提高工作效率；有的是为了提高准确性和可靠性；有的是为了完成无法用人工方法实现的难度很高或极为烦琐的工作。总之，汉字信息处理系统的应用，能够大大促进在我国建立有效的管理体制和手段。

1.2.3 汉字信息处理技术标准化问题

一、标准化工作的重要性

汉字信息处理在目前还是一门新技术，它虽和传统计算机技术的关系极为密切，但却有许多独特的技术课题有待进一步去探索，例如：有关汉字属性方面的研究工作；汉字信息处理设备的研制和生产；汉字信息处理系统软件的开发等。一项新技术在推广应用前，必须先确立技术标准，这是工业技术发展必须遵循的途径。如果没有或缺少技术规范或标准，各行其是，必然会影响这项技术的进一步发展。但是技术标准也不可能是凭空产生的，它必须建立在一定的研制成果生产和使用经验的基础上，并从基础的标准开始，逐步形成一个比较完整的技术标准体系。

二、标准化工作内容

汉字信息处理技术标准化工作的内容是较多的，可以列举出如下项目。

（一）有关汉字属性的标准

1．汉字数字化字模标准　这主要指通用型的汉字字模。先应确定字体和字形。对于点阵密度可选择两种规格，即 15×16 点，24×24 点。以后根据需要，可以再扩展 32×32 点的标准。要求在这样的点阵密度条件下，得到质量较高的字模图形。此外，制订汉字标准字根，对以字形为主的汉字编码输入方案设计和字根汉字字模库设计工作是很有意义的。

2．汉字交换码标准　为便于系统之间的汉字通信，应定出交换码标准。这项标准要求先定出计算机处理汉字的字量。1981 年我国已颁布了 GB2312《信息交换用汉字编码字符集——基本集》国家标准，共收集汉字 6763 个。目前正在制订辅助（扩充）字符集及其交换码标准。

3．汉字索引（indexing）和排序(sequencing）标准　象通常从字典中查找(检索）汉字的方法相同，汉字信息处理系统中要建立汉字属性字典，以便在系统中查找汉字或由它组成的词。汉字的检索方法和拼音文字的检索方法相比，困难要大得多。这是因为它的字量大，不易记住国标汉字字符集中汉字的排列顺序。为此，有必要建立汉字索引标准。通常检索汉字，可以按部首（字根），笔画数，或汉字拼音所用的拉丁字母的顺序等。

（二）汉字系统所用的控制功能标准

控制功能是用来控制计算机系统中各项设备的特定动作，包括控制功能的种类、符号和含义。因为它直接关系到各类外部设备的控制动作和软件，所以制订控制功能标准对于汉字系统的配置和应用是很重要的。

（三）汉字编码和输入方法标准

编码和输入方法有密切联系。目前各种类别的汉字编码方案很多，其输入方法也各不相同。制订汉字编码方案的标准需在优选汉字编码方案的前提下进行。为此，需要首先拟订汉字编码评测标准。制订汉字输入方法标准时，要注意分开层次等级、提高型和普及型并重的方针，以利于切合实际，推广应用。在制订输入方法标准的同时，也应包

括制订笔触式汉字字盘盘外字的输入方法的标准。

（四）汉字设备方面的标准

1．汉字键盘标准　除了字母数字键盘外，字根式汉字键盘、笔触式汉字键盘都需有技术标准（如汉字键盘总技术条件，汉字在盘面上的排列标准等）。

2．汉字字模库标准　对于成批生产的 ROM 汉字字模库，应有关于 ROM 器件的集成度、收容字数、编址方法、接口技术等的技术标准。对于字根式汉字字模库，也要有相应的技术标准，以利于建立汉字终端等设备设计和生产的标准体制。

3．汉字印刷机　对于汉字印刷机，需要对常用机型（例如汉字针式打印机、简易激光扫描印刷机）制定技术标准（其中包括控制功能码和汉字字符点阵码标准）和建立型谱系列标准。

4．汉字显示终端　对于汉字显示器，应该订出显示管尺寸系列、分辨率等级、所显示的字模点阵和满屏的字数标准。对于显示终端，需要对汉字字模库的设置，扫描刷新方式，接口技术等订出标准。同时还应制订汉字显示终端的功能等级和型谱系列标准，并作出显示终端模块化结构的规定。

5．其他硬设备方面的技术标准　对于汉字光学字模识别技术和设备，联机手写体汉字识别技术和设备，汉字语音输出设备等，都需根据研制工作的进程，制订相应的技术标准。

（五）汉字软件技术标准

不仅应制定出汉字信息处理系统软件应具备的基本功能和操作系统扩充汉字功能的技术标准，而且还要制定出各种程序设计语言处理汉字信息的技术标准，汉字内部码标准，系统汉字和西文字符的标识符标准等。

1.3　汉字信息处理技术的现状和展望

1.3.1　国内汉字信息处理技术的现状

我国六十年代末就已开始对汉字信息处理技术进行探索和实践。邮电部门在 1968 年研制成的汉字电报译码机，是我国汉字信息输出设备的最早型式。七十年代中期开始系统地研究和开发这项技术，明确地提出“汉字信息处理系统”（七四八工程）的研制课题，并列为以电子计算机技术为中心的重点系统工程项目。在实施的最初几年内，由于受当时器材和设备条件的限制，进展比较缓慢，但也创造了一些技术条件和研制经验。1978 年以来，由于我国开始广泛应用大规模集成电路存储器和成套的微处理机芯片，因而在很大程度上促进了汉字信息处理技术的发展，不仅使原有的一些技术得到更新，而且研制成了一些新型的汉字输入输出设备。在技术指标、可靠性、实用性和经济价值方面，都有很大程度的提高，已能用国内研制的汉字设备和计算机配置成多种应用系统，特别是以微处理机为基础的汉字信息处理系统或用微处理机控制的汉字终端发展更为迅速。目前，国内在继续进行汉字基础理论研究的同时，已制订了汉字信息处理设备和系统的研制和生产规划。汉字信息处理系统的推广应用工作愈益得到政府部门和各类业务部门的重视。可以预期，在今后的几年内，汉字信息处理技术将会以更快的速度向前发展。

以下简述国内在汉字信息处理技术，汉字设备的研制、生产和汉字系统配置方面已

取得的成果。

一、汉字信息处理技术的基础工作

在对我国使用的汉字频度统计的基础上，结合我国习惯使用的汉字部首索引分类法和我国汉语普通话拼音规则等制定的汉字字符集（基本集）及其标准编码，在1981年公布后，对我国汉字信息处理技术的发展起了很大的促进作用。目前这项工作尚在继续，正在进一步制订辅助的汉字字符集及其标准编码。并且还在制订能适用于汉字信息处理的控制功能码（或称控制功能）的标准。

为了使汉字信息处理技术水平推向更高的阶段，正在开展对汉语词汇的研究和进行词频统计工作，开展对汉语自然语言处理的研究。近年来，在机器翻译的理论探讨和工程实践方面，也有较大进展。

国内目前有不少单位在研究汉字自动识别的课题，其中包括光学汉字识别和联机手写体汉字识别。虽然这项技术目前已得到的成果离开实际应用尚有不小距离，但是有信心通过努力，使这项技术早日达到实用阶段。

近来国内对汉语语音识别已获得一些进展。对于上百个不连续的汉语语音（字组或词的发音），经过“训练”的计算机系统能够加以自动识别，准识率达到95%以上。而且，由于这样的识别设备是在普通的微型机上由软件实现的，故可用于一些特定的领域。对于不限定词汇以及连续汉语语音的识别问题，尚需进行细致深入的研究工作。

可以预期，上述两项研究工作在取得实际应用的成果后，将大大改善我国目前的汉字输入方式。

对于汉字语音合成输出，国内目前已掌握这项技术，而且字数不限。这是通过在微型机上附加一种专门研制的声音频率合成电路装置的方法实现的。

对于汉字编码输入方法而言，除了采用整字输入方法外，目前主要是用编码输入法。据不完全统计，国内已写成书面资料的各类汉字编码方案已超出400种，并且还在继续发展。其中技术指标较高并得到国内多数用户实际应用的也有十多种。有关汉字编码技术的学术组织正在致力于制订评测标准，力图吸取一些技术指标较好的编码方案的优点，制定出一种具有综合性指标能为大多数用户所接受的方案。这在我国是一项有深远意义的工作。

二、汉字设备的研制和生产

（一）汉字键盘

除了字母数字键盘外，对字根式汉字键盘和笔触式汉字字盘也都开展了不少研制工作，除了字根式键盘因使用不普遍而未计划投产外，对笔触式汉字字盘已组织小批量投产，同时，为了提高其可靠性和降低造价，正力图改进设计，提高工艺水平。

（二）汉字字模库

汉字整字字模库发展方向是采用集成度高的ROM器件，大量制作整字字模库。这项工作应在确定数字化字模的点阵标准和汉字字模库设计规范后进行。国内目前已制作用256KROM器件实现的15×16和24×24点阵汉字字模库。今后将用集成度更高的（例如用1兆位的ROM芯片）ROM器件制作点阵密度更高的字模库。

目前正在进行的字根汉字字模库的研制工作，应在信息压缩比、字形质量、汉字输出速度等几个主要指标方面达到更高的水平，这类字模库在需要提供更多的汉字字模（例

如要求包括辅助集汉字字模）时有较大的应用价值。

（三）汉字印刷机

国内目前已研制成多种类型的汉字印刷机，并已配置在国产汉字信息处理系统中。例如：16针、24针的低速针式打印机，其分辨率和打印速度已接近国外产品水平。热感式汉字印刷机已有小批量投产。中速的简易激光扫描汉字印刷机，已研制成样机，需要进一步提高可靠性，向小型化方向发展，组织批量生产，降低成本。根据实际需要，还要研制一些适合于我国使用的新型汉字印刷机。今后随着微型计算机的普及应用，在汉字印刷机方面，估计会有较大的发展。

（四）汉字显示终端

近几年来，国内微处理机控制的汉字显示终端设备的研制和生产发展很快。这类终端有的是在微处理机芯片的基础上自行设计研制成的，也有的用进口的个人计算机经改制而成。在功能方面，有简易的汉字输入输出终端，也有汉字智能终端。所用微处理机的类型也较多，大多数是标准8位的微处理机，也已开始研制16位的微处理机汉字终端。目前有些型号的汉字终端（例如ZD2000、HZ8401型汉字终端）已组织批量生产。今后汉字显示终端的需要用量很大，在制定有关的技术标准后，需进一步确定型谱系列，完善品种，扩大生产。

三、汉字信息处理系统的配置

对于汉字信息处理系统的配置，除了提供必要的汉字设备及其接口外，重要的是软件的配置。在系统软件扩充汉字功能方面，目前主要在微型机上做的工作较多。例如，在微型机使用较普遍的CP/M操作系统中，加入汉字输入输出管理模块。扩充功能后的操作系统（有些系统称它为EC-DOS，即西文汉字兼容的磁盘操作系统），可以解决它所支持的多种高级程序设计语言处理汉字的问题。对于PDP-11系列的小型计算机系统，也已完成了把它的RSX-11M操作系统扩充为具有汉字处理功能的工作，这一成果适用于PDP系列小型机的各档机种。

汉字终端同各类小、中、大型机系统连接的工作，已有了不少实例，其中，所连接的有进口的各种机型和国产的小型系列机。对于国产的系列机，实现这项工作是很有意义的，目前已对1000系列、2000系列，S16等机型完成了这项工作，使它们能通过终端实现汉字输入输出。一些单位已开始研制前述的接插兼容的汉字终端，使各种类型的主机系统通过汉字终端实现汉字处理功能。

对于国产系列机本身的汉字化工作，目前也正在小型系列机（例如1000系列）上进行，今后将扩展到国产的其他系列机种。

近几年来，国内完成了不少专用汉字信息处理系统，这些系统不少是用国产的计算机和汉字设备组成的。在精密汉字编辑排版系统中，其主机用的是国产1053机。其中，采用了高倍率汉字信息压缩技术。用位片式微处理机高速还原成汉字，采用高精度激光扫描技术输出版面，能提供多种汉字字体、字号。排版应用软件能排出各种类型的版式。整个系统的技术已达到国际先进水平，并能供实际使用。在小型汉字情报检索系统中，其主机是1053机，配有大容量磁盘外存储器、联机汉字终端等设备，能供科技情报单位和小型图书馆使用。用国产小型机构成的企业管理系统，已在试点应用。此外，微型机汉字信息处理系统的各类应用更为普遍，象小型汉字数据库系统（包括汉字化的

dBASE 2 和dBASE 3 数据库系统)、办公用汉字信息处理系统和一些业务处理系统等。

1.3.2 国外汉字信息处理技术状况

一、日本汉字信息处理技术的发展

日本是研究汉字信息处理技术较早的国家，因为日文中包括许多汉字（常用汉字就接近 2000 个),要解决用计算机处理日文，就先要解决汉字处理技术。

日本也是目前对汉字设备研制和生产较多的国家，在日文信息处理系统的应用中积累了较多的经验。

对于汉字输入方法，经过多年来的探索和实践，目前主要推行三类方法：整字输入法；汉字假名变换输入法；二次击键联想式输入法。其中，整字输入法同我国所用的没有多大的区别，只是汉字数目和盘面文字排列有些不同。目前整字输入法大多用笔触式字盘,也有仍在使用移位键式大键盘的。汉字假名变换输入法是利用与汉字相应的假名读音。但是,由于有些字的读音由多个假名音节组成,故不能得到较高的输入速度。它的优点是,除了不需特殊操作训练外,可以用日本流行的假名键盘输入汉字。二次击键联想输入法是日本近年来发展的一种高速汉字输入方法。这种方法固定用两个假名来代表每个汉字，从而可以组合成 2500 个汉字。在操作训练时，因为必须强记这种对应关系，故训练难度较大。但是，只要熟练掌握了这一方法，就可以得到每分钟输入 250 个汉字的高速度，适合于专业操作人员使用。

在日本，鉴于生产计算机和汉字设备的厂商较多,各厂家所定的汉字代码都不相同,为了各厂商生产的设备具有通用性，便于通信交换，于 1979 年 1 月公布了汉字交换码的日本工业标准（JISC6226），以便各厂商生产的设备有一个共通的代码标准。

在汉字字模的存储技术方面,日本汉字点阵规格和我国的相似,主要有两种：15×16 点；24×24 点。所采用的字体称为明体（相当于我国的宋体)。日本较高采用了中、大规模集成电路来实现字模存储。例如，日本昭和情报机器公司 1973 年研制的脱机批处理汉字系统（T4100）中，就已使用中规模集成移位寄存器来存储汉字。近年来，用大规模集成电路构成的汉字字模已商品化。例如，富士通公司的 MB8364 16×16 点阵汉字字模库，收容 2965 个汉字；冲电气公司的 MSM38128 24×24 点阵汉字字模库，收容 3418 个汉字。这些字模库采用了 128K、256K 位的组件。最近，又研制用 1 M 位的超大规模集成电路制作的汉字字模库。这样，只需用 2 ～ 3 块芯片就可解决上万个汉字的存储，所以汉字字模库技术更变得简便可行。

在汉字印刷机方面，日本生产的品种很多，例如有 16 针、24 针和梳齿状的汉字针式打印机、模拟式汉字书写机、喷墨式汉字印刷机、激光扫描汉字印刷机、发光二极管硒鼓转印汉字印刷机、光纤管转印汉字印刷机等多种。其中，以针式汉字打印机的使用较普遍，生产量最大。近年来，简易型激光扫描汉字印刷机的应用面也逐渐扩大。针式汉字打印机和喷墨式汉字印刷机已能印出彩色的文字或图形画面。

汉字显示终端和微型机汉字信息处理系统技术近年来发展很快。在日本有代表性的产品是日本电气公司 1979 年生产的个人计算机 PC-8000，已在 1981 年将其扩充成具有汉字功能的微机系统（把汉字和图形技术结合起来），即 PC-8800 系统。最近，该系统又改用 16 位的 Intel 8086(和 IBM 公司的个人计算机兼容)，从而推出NEC PC9800

系统。

在小、中、大型机方面，日本几家大的计算机公司都在他们所生产的一些系列机上扩充了汉字功能，并相应设计了西文和日文兼容的操作系统，在日文工作方式下使用时，能利用系统所有的一切软硬件资源，和西文工作方式一样方便。

日本在推行汉字信息处理技术的同时，很重视建立技术标准的工作，对于汉字输入输出，汉字存储、各种汉字设备、汉字软件（包括汉字代码和数据类型、高级程序语言扩充汉字处理功能、汉字文件格式等），有些已制订出标准，有些正在积极准备中。其中有国家的工业标准（即 JIS），也有各厂商自己的标准。

日本国内汉字系统的应用已相当普遍。1980 年以来，在解决汉字信息处理技术的基础上，又将其和日文假名、西文的处理结合起来，开发日文信息处理技术，从而使其应用范围日益扩大。目前，许多部门都已应用了日文计算机系统。随之，个人计算机、文字处理机和办公室计算机系统的得到更为迅速的发展。

在计算机汉字编辑排版技术方面，日本已有近二十年的研究和应用历史。六十年代中期，日本发展由计算机控制的光机式汉字照排系统（通称二代机），用于书、刊的照相排版。七十年代中期，发展全电子式的汉字编辑排版系统（通称三代机），用高分辨率 CRT 输出版面，在排版功能和输出速度上比二代机大为提高。如朝日新闻、产经新闻、每日新闻、经济新闻等全国最大的四家报纸都已采用全电子式日文照排系统。八十年代初期，日本开始研制用激光扫描输出的汉字照排系统（称为四代机）。

一些新的研制项目，如光学汉字识别（OCR）技术，日本已有较好成果。例如，东芝公司的汉字 OCR 装置，能自动识别 2000～3000 汉字，识别速度可达 100字/秒。联机手写体汉字识别装置，也已接近实际应用。汉字声音（假名发音）识别的研究，目前正在积极开展。

为了促进日文信息处理技术的发展，日文信息处理学会（Information Processing Society of Japan）定期召集国内和国际会议，交流和促进这项技术的发展。

二、汉字信息处理技术在欧美

在美、加、澳等国家，研究汉字信息处理技术主要是一些外籍华人。特别是在美国，其研究工作具有一定规模。例如美国的中文计算机学会（Chinese Language Computer Society，USA）组织，每隔一、两年定期举行学术会议，交流汉字信息处理的理论和方法。对汉字输入方法的研究中，大多倾向于用字母数字键盘按编码方法输入汉字。较多的编码方案是基于汉字的拉丁字母拼音。也有按字形的编码方案，例如，澳大利亚的墨尔本大学有人提出用笔画输入的方法，美国王安公司采用三角号码法。此外，在联机汉字手写体识别，光学汉字识别，汉字语音输入以及程序设计语言的汉字化方面，也取得了一些成果。

在汉字信息处理系统研制生产方面，近几年来美国的一些大公司都在作这方面的努力，主要是在系统软件上扩充汉字功能，如 IBM，Honeywell，CDC，Univac，HP和王安公司等都在他们所生产的大、中型机系统上进行这项工作。在微型机的汉字化工作上，也开展了不少工作。例如星茂公司的 Sigma-10 汉字微型机系统，IPX 的通信用汉字终端，ASIAGRAPHICS公司的“自来字”汉字系统，Xerox 公司的微机汉字终端等。

英、法、西德、意大利等国家，也在某些用途的汉字信息处理系统和设备方面进行

研制工作。例如西德西门子赫尔公司早在七十年代前期就已研制全电子式的汉字照相排版系统。英国 Monotype 公司在一九七九年研制成激光汉字编辑排版系统。此外，西欧一些国家也在研制和生产几种汉字设备，例如法国巴黎大学和法国自动化设备公司合作研制成联机手写体汉字识别设备。英国 Sindex 公司正在研究汉字语音识别技术，意大利等国家正在研制和生产像热感式汉字印刷机和伺服笔式汉字书写机等。

三、台湾和香港地区对汉字信息处理技术的研制

我国台湾省的汉字信息处理技术研究工作，在六十年代后期已经开始，近年来汉字信息处理系统的研制和实际应用发展也很快。

汉字输入方法，整字输入法和以字形为基础的输入方法应用较普遍。如三角号码法，仓颉编码法、汉通编码法（首、末笔画加字形特征）等编码方案，都利用字母数字键盘作为输入工具。

仓颉编码法是一种在当地较有影响的编码输入方法，其特点是采用 24 个基本形符和辅助键，这样可以组成 22000 个繁体汉字。每个汉字由 1～5 个字节组成，平均每字 3.8 字节。这些字中有 107 对同码字，可用软件方法确认。这种编码方法的缺点是它的 24 个形符和传统的偏旁、部首相比差别较大，从而掌握起来比较困难。其优点是可以用普通的字母数字键盘，并可以输入较多汉字，在台湾省被评为最优的汉字编码方案。

近年来，台湾省的微型机汉字系统和汉字终端技术发展较快。较有代表性的有天龙微型机系统和长城 75 型汉字终端。天龙 8002/256 采用 16 位微处理机 Z8000 作为 CPU，主存容量为 256K 字节，配用 *Unix* 操作系统，可带 4 台长城 75 型汉字终端（用Z80 微处理机控制）。该系统还带有一台 20 兆字节的温式磁盘机，24 针的针式汉字打印机。汉字输入采用仓颉编码法。汉字存储采用字根和矢量组字结合的方法，字根和汉字编码所用的形符一致(称为仓颉造字法或称 CCG 技术)。汉字点阵为 14×15 点；22000 个繁体字使用 60K字节的存储器；字模产生速度约 40 字/秒。此外，该系统还配有关系数据库结构的数据库管理系统。

仓颉造字法技术目前尚在几个主要技术项目上加以改进，目前已发展到第四代 CCGⅣ。主要目标为从软件角度改进汉字字形质量，提高字模点阵密度，提高输出速度，并进一步压缩字模信息的存储容量。

汉字系统的实际应用方面，诸如办公用计算机、文字处理机、数据库管理系统和汉字通信网络系统等，都已有成果。主要的报刊出版正拟用汉字编辑排版系统，已开展版面格式自动设计、小样自动产生、字形压缩技术、字形变倍变体技术、排版、显示、激光拷贝、缩微输出等一系列研究工作，并已开展省内和海外的直接汉字新闻通信。

香港地区近年来很重视对汉字信息处理技术的研究，一些大学等学术机构正从事这项技术的理论和方法研究，对于汉字编码和输入方法很重视。香港中文计算机学会经常定期交流这方面的技术。该地区的一些计算机公司也重视在各类计算机系统上扩充汉字功能，以便于向东南亚等使用汉字的国家和地区推销产品。汉字系统大都采用微型机，例如用 Apple Ⅱ微型机实现。采用仓颉字根和矢量组字技术，制成一种名叫“汉卡”的插件板，插入 Apple Ⅱ机后，就能实现汉字输入输出功能。由于其价格低，故推广使用的面广。现正设法把这一“汉卡”用到其他类型的微型机系统上。

在应用方面，香港地区的一些办公室和商店等，已开始使用办公用汉字计算机或文字处理机。今后，汉字信息处理技术将在该地区得到更广泛的应用。

1.3.3 汉字信息处理技术的发展前景

到目前为止，汉字信息处理技术的研究开发已取得一定成果，除继续探索和开发一些新的领域外，应使它在我国的计算机推广应用方面发挥更大作用。

一、技术开发

为了不断提高汉字信息处理技术并扩大使用范围，应做好以下几方面的工作

（1）加强汉字信息处理技术基础理论方面的研究工作，使这项技术不断向深度和广度方面发展。

（2）重视汉字信息处理技术标准化的研究和标准的制定，逐步完善标准体系，以利于汉字设备的工业生产和推广应用。

（3）对于各种汉字设备，应注意优选机型，扩大批量生产，不断提高质量。加强型谱系列化工作，加速研制开发新品种，建立我国完整的信息工业体系。

（4）确定汉字系统体制，加强汉字系统软件的研制工作，做好国产计算机系统的汉字化工作。

（5）加强汉字系统应用软件的研制工作，移植西文系统的应用软件包，扩大汉字系统的应用范围。

二、推广应用

有重点地装备和推广一些效果显著和影响面大的汉字信息处理系统。

在推广汉字系统的实际应用过程中，应重点推行一些应用系统，例如：

（1）事务管理系统、企业管理系统、办公用计算机、文字处理机和局部网络系统；

（2）科技情报检索系统和数据库系统；

（3）汉字编辑排版系统（报纸书刊自动编印系统）；

（4）咨询服务系统；

（5）公用事业窗口系统；

（6）计算机教育系统；

（7）计算机通信网络系统；

（8）办公用系统、文字处理机等。

三、汉字信息处理技术应用发展的前景

实现了以上各类汉字系统的推广应用后，会使我国的计算机应用展现令人鼓午的前景。例如：政府机关利用计算机进行行政管理，编造统计报表，通过通信网络汇总全国工农业生产数字和收支平衡情况，从而取代大量繁琐的人工劳动；工厂企业和物资部门利用计算机进行物资管理，建立现代化生产管理体制；科技情报管理系统配合通信网络，把全国各省、市的主要图书馆、科技情报研究单位联系起来，提高图书、情报资料的利用率，使它们更好地为科研、生产和教学服务；利用汉字编辑排版系统，在全国有计划地集中建立计算机排版中心，各出版部门可把记录原稿信息的软盘等媒体送到排版中心制版，从而可大大提高排版效率，实现书刊出版工作的全面技术革新。各大报社、新闻

通信社利用新闻通信网络和汉字编辑照排系统把新闻采集、通信、发稿、报纸编排等组成一体化作业，加快新闻宣传报导的时效；公用事业窗口系统可以改进服务质量，节省人力。现代化教育系统可以提高教育质量，改进教育方法，提高教育效率。咨询服务系统，可以提高信息的价值和利用率，把信息资源提高到可以和物质资源相提并论的高度，为进入信息化社会创造条件。总之，这些系统的广泛应用，将对我国的四化建设产生深远的影响。

为了促进我国汉字信息处理技术的发展，全国应团结一致，大力协作。1980年 6 月，我国成立了中国中文信息研究会，并且定期组织开展这一技术领域内的学术活动，现已初步见到成效。目前正开展和海外学术界的联系，组织学术交流，以便共同促进这项技术的不断完善。

第二章　汉字属性

汉字是汉语的书写符号。它是历史悠久、影响深远的文字之一。本章从汉字的演变出发，对汉字信息处理与汉字属性之间的关系作了系统的阐述。

2.1　汉字演变概况

汉族最初创造的文字，已无实物可考。但可估计，黄帝以后，已经陆续出现了汉族的文字了。1880 年河南安阳小屯村发现刻有文字的龟甲骨片后，人们便开始亲眼见到了我国古代汉字的实物。这些汉字是公元前 1300 至 1028 年间的产物，上距黄帝时代约一千多年，下距现在约三千多年。这种已搜集到的汉字约有两三千个，其中有象形字、表意字、形声字和假借字。可以想见，当时造字，已逐步像宋代郑樵所说的“形不可象则属诸事，事不可指则属诸意，意不可会则属诸声，声则无不谐矣”那样，日益往形声字方面发展了。那时造成的形声字还不多（只占 20％左右），而且作为形旁的那个义符，还无一定写法，还很混乱。例如甲骨文“牝牡”这两个字，取作形旁时，那个义符不一定都用“牛”，用“羊、豕、犬、马、鹿、虎”等情况都有，直到秦代搞“书同文”才作出规范化，统一用“牛”。至于形声字的声旁，则无论是甲骨文或是后来的汉字，都一直没有统一过。同一个音常有两种以上的符号；不同的读音，又因地区方言不同，而常用同一音符来表示。以致形声字的声符渐渐失去其标音作用。又字形变化，同音借用，多数形声字的义符已经失去其表示类属的作用。例如原为“者”声的“赌、睹、煮、暑、箸、锗、楮、猪、诸、屠、都、奢”等字，“者”已成了多种读音的一个音符；“笑”字、“骗”字已难识别“笑”与“竹”义、“骗”与“马”义有何关系。由此可见，形成形声字的情况既繁且乱，而且机动多变。以致形声字大量产生，使汉字单字的数量日益增多。

2.2　汉字字量

2.2.1　汉字的累积字量

社会不断演进，新事物不断发生、发展，旧事物不断老化，淘汰。作为汉语记录工具的汉字也不例外。

现在有实物可考的最早的汉字是甲骨文字，约有三千个。其后经过一、二千年才有汉代编辑的仓颉篇，共 3300 字（当时社会中所实有的字量可能不只此数）。其后经过三百多年，东汉许慎著《说文解字》有 9353 字，加上异体重文 1163 字，共为 10516 个汉字。

其后经过四百多年，有南朝顾野王著《玉篇》(公元 543 年)，收字 22726 个。宋代司马光等著《类篇》(1066 年)，收字 31319 个；明代梅膺祚著《字汇》，收字 33179 个；清代张玉书等著《康熙字典》(1716 年)，收字 42174 个；民国时代中华书局编《中华大字典》(1915 年)，收字 44908 个。据估计，累计到现在，古今汉字总数大约已有六万个左右。由《说文解字》到《中华大字典》共经过 1815 年，字量增多 34392 个，平均

每年约增多 18.95 个字。若按目前所估计的六万字计算，则时间经过是 1983－100＝1883 年，共增多 49484 个字，平均每年增多 494.84 个字。这说明从东汉到现在共经过一千八百多年，汉字逐年有增无减。解放以前增加的，可以说绝大多数都是形声字，只有少数是异体字和简体字。解放后增加的，几乎百分之百是简化字，同时淘汰了大批繁体字和生僻字，保留下来的通用汉字不过六、七千个。其他五万多个生僻字，虽然受到淘汰，但还原样保存在古籍中。在当代口语化了的书面语言中，日益新生滋长的已不是单字（尤其不是从前那种以生僻自炫的单字），而是能用少数单字灵活组成的新词。目前，汉字造字这条老路已被堵死，只有造词这条新路畅通无阻。

2.2.2 汉字实用处理（一）

汉字字量过大，既不利于学习、记忆和应用，更不利于信息处理。所以在实际应用中宜有区分，作相应的处理。

实用处理之一是按照字频分级。一般分作三级：（1）常用字，（2）次常用字，（3）生僻字。用得多的字优先处理，用得少的字和生僻字另作安排，一切为高频度字让路。

据北京新华印刷厂 1977 年版《汉字频度表》的《综合频度表》对两千一百六十二万多字中不同汉字总字量使用频度的统计结果，并参照郑林曦等所编《按字音查汉字频度表》，可以把 6373 个单字分作如下三个等级：

1．一级字　由频度最高的原编号码为 1 号的“的”字到 560 号的“河”字共 560 个单字，和频度次高的原编号码为 561 号的“帝”字到 1367 号的“震”字共 807 个单字所组成。这 1367 个字累计出现 20553572 次（占 95.03%），称为“常用字”。

2．二级字　由原编号码为 1368 号的“岁”字到 2400 号的“茫”字，共 1033 个单字，累计出现 859951 次（占 3.97%）。这些字称为“次常用”字。

3．三级字　由原编号码为 2401 号的“筹”字到 6376 号的“腴”字共 3976 个字，这些字累计出现 216192 次（占 1%），称为“生僻字”。

上述统计的资料包括工业、农业、军事、科技、政治、经济、文学、艺术、教育、体育、卫生、天文、地理、自然、化学、考古、文字改革等方面。统计范围之广和总字量之多，是空前的。但由于所查书刊的内容和时间的特点，也受到一定的限制，除出现次数特多的字有其普遍性外，也反映了一定的特殊性。那时正逢宣传“批林批孔”，所以有些字的出现次数受时代背景和被统计资料的干扰，就显得不正常地增多。例如“彪”字本不是常用字，可是在《综合频度表》中却出现 9422 次。“孔”字在教育部《常用字表》中尚列为次等常用字，但在《综合频度表》中则出现 25585 次，序码是 215 号，一跃而为最常用字。当然，绝大多数汉字的使用频度还是能够正确反应出来的。

按照上述三级处理时，总累积频度占 99%的 2400 个单字（其中一级字有 1367 个，二级字有 1033 个）可作为常用字，可按最方便最有利的条件来处理。而总累积频度只占 1%的其余 3976 个单字，则可作为备用字处理。

2.2.3 汉字实用处理（二）

汉字实用处理之二，是要分清下述两类情况：（1）高频度汉字和基本汉字；（2）专业汉字与通用汉字。

高频度汉字是当前书刊中出现次数较多的前一、二级汉字，也就是目前一般人们文化生活中用得最多的那些汉字，总量约 2400 个单字。基本汉字，最初是由洪深在 1935 年从所谓的实物字、动作字和形容字中选定的 1100 个汉字。继经舒新城和吴廉铭补充后共为 1600 个。分为人、天地、鸟兽、草木、亲友、衣、食、住、行、物具、色味、部位、数量、次序、形象、状态、质料、感觉、品格、（口的、手的、足的、眼耳的、心理的）动作词等，共 24 类。打算实现：（1）用基本汉字编一本字典，各字释义一律限用基本汉字；（2）日常写作和书信也一律限用基本汉字；（3）教材和教学也限用基本汉字。这个尝试终于失败了。因为以 1600 个基本汉字作为认识汉字的最低限度是可以的，但不能以此为限，作茧自缚。在汉字信息处理中，必要时，把高频度汉字 2400 个字中的前 1600 个字作为基本汉字而把其余 800 个汉字作为常用字，也是可以的。

专业汉字是该专业常用、而其他专业罕用以至不用的那一类汉字。通用汉字是各行各业都通用的那一类字。二者在各专业中出现的频度是不同的。例如，在通用文学中，“我”、“你”、“他”、“她”是常用字，而“羧”、“酐”、“嘌”、“呤”则几乎用不到；但在生物化学中，“羧”、“酐”、“嘌”、“呤”是常用字，“我”、“你”、“他”、“她”则几乎一次也用不上。为了提高各专业的工效，宜在汉字信息处理技术中对二者所选字数各有不同安排才好。

2.2.4 汉字实用处理（三）

汉语单音词很多，汉字中的单音词更多。口语说“黑马”已很好，汉字又特别造成一个“驪”字，简化后又有“骊”字。如此“青白马”又造“驄”字，“千里马”造“驥”字，“公猪”造“豟”字，“小猪”造“豚”字等等。据统计，在《新华字典》所收 8500 个字中，除开“囫囵”、“葡萄”等不能拆开单用的“联绵词”（共有 578 个，占 6.8%）外，单音成义的单字竟达 7922 字（占 93.2%）之多，可以说，这是世界语文中的一个奇迹。英语词典中的单音词就很少很少，是无法和汉语单音词相比的。

由于汉语单音词特别多，所以组词能力特强。只要两个单音汉字，颠来倒去，便能组成绝大多数词汇。例如：“兄”、“弟”可以组成“兄弟”、“弟兄”、“弟弟”；“青”、“年”可以组成“青年”、“年青”、“年年”、“青青”。至于三个或四个单音汉字，其组词性能就更复杂、更灵活了。汉字的这种性能使汉字成为世界上极其难能可贵而异彩独放的一种文字。这是汉字的突出优点。

有优点同时就有缺点。汉字的突出缺点是单字字量特多，在信息处理技术中难办。所以，除了分级处理和专业处理之外，还有必要进行汉字词频统计的研究。掌握了汉字词频的科学数据以后，我们便可在汉字信息处理技术中进一步提高工效，促进四化建设。

2.3 汉字字形

2.3.1 汉字的形体结构及其分解

汉字字形是汉字形体结构的图象。篆书图象与楷书图象不同，分解方法也不同。《说文解字》把所收篆书按照构形特征，分为 540 部。每部各有一个共通字根，叫做“部首”。这是对汉字处理的最早的科研成果。其后，汉字的字形、字音和字义互有变迁，到楷书

成为规范字体后，有不少汉字已难于归部，有不少部首也已失去部首作用，所以有必要加以归并简化。许慎所用“据形系联”的分类方法也同时改为按笔画画数分类。发展到现今的计算机时代，部首分类法和笔画画数处理法又已远远跟不上时代要求，因此又出现“字根”、“字首”、“字式”、“字型”、“笔形”、“位点”等等的分解处理概念。只有彻底理解这些新概念，我们才能在全新的基础上掌握汉字科学，提高汉字信息的处理技术水平。

2.3.2 汉字图象的细胞——位点

从前，汉字应用范围有限，由部首分解到笔画就算尽顶了。现在，所联系到的专业、科学技术和文化工具等等日益复杂，汉字应用范围日益广泛，笔画已显然不是汉字结构的最小单位。在一些粗疏的照相版印件上显然可见，每一笔都是由许许多多网状微点（简称“网点”或“位点”）组合而成。由此可知，笔画的最小单位是“位点”。汉字的笔画一般都不少（每字平均约 11～13 笔画），组成一个汉字的点阵的最少点数为 15×16＝240 点。

在坐标图的同一面积内，网点只有各不相同的位置上的区别，没有长短、大小等的不同，所以叫做“位点”。每一位点都是X-Y坐标的一个交点。它没有“方向性”，只有共同的“位性”。分散孤立的一个位点，毫无意义。但要注意，这里所说的“点”，并不是汉字笔画中一般所说的“点”。汉字楷书笔画起码要两个互相邻接着、形成一定方向性的位点才能表现出来。两个位点只能形成极短的一、丨、丿、㇏；三个位点则能形成一丨丿 ㇏㇕㇄中的任一种笔画。每一种笔画，无论其位点数有多少，都是一种抽象量，彼此同为一画，互无差异。每一种笔形，即使仅仅是由两个表现有一定方向性的位点形成的，但彼此都赋有具体的不同质的表现，所以又是互有区别的。

2.3.3 笔画

一、单向笔画和复向笔画

最原始的汉字本来是一种图画文字，是一笔一笔画成的，所以下笔后按一定笔向连续画成的每一笔，就叫作“笔画”。

篆书（尤其是古篆）还没有完全从图画文字中蜕化出来，笔势婉转曲折，构形机动多变。所以汉代许慎依据篆书分解时，虽然提出“据形系联”的编排标准，制定 540 个部首，按字形分为 14 组；但牵强附会，缺乏科学规律。

后来演变成楷书，圆弧形笔画没有了，笔势走向一定。由汉到宋，郑樵（约 1161 年）写作《六书略》时，才开始对汉字的基本笔画，从具体的笔形方面进行分解研究。他说：

“衡为一，纵为丨，邪为丿，反丿为㇏，至㇏而穷。”

“折一为㇕，反㇕为厂；转厂为㇗，反㇗为㇄，至㇄而穷。”

“折一为厂者侧也，有侧有正，正折为∧，转∧为∨；侧∨为＜，反＜为＞，至＞而穷。”

“一再折为冂，转冂为凵；侧凵为匚、反匚为コ，至コ而穷。”

“引一而绕合之，方则为□，圆则为○；至○则环转异势，一之道尽矣。”

"丨与一偶，一能生，丨不能生，以不可屈曲又不可引，引则成丨；然丨与一偶，一能生而丨不能生，天地之道，阴阳之理也"（见《通志》第35卷"六书略"）。

"衡"即"横"；"邪"即"斜"（现在一般叫"撇"）；"㇏"(反丿)现在一般叫"捺"。

"㇇"就是我们所说的"弯笔"，"𠃊"就是我们所说的"拐笔"。"𠃌"（反㇆）和"㇅"（反𠃊），现在楷书书写时已不连作一笔而是分作两笔书写了。

"∧∨〈 〉"，现在除"巢"等极少数字中还有"〈"笔外，"∧∨>"这三种笔形已转化变形，不复存在。

"冂凵匚コ"这四种结构，在楷书书写习惯上应依次作"丨㇆、𠃊丨、一𠃊、㇆一"两笔看待。

楷书中已无纯圆结构"○"，只有方形结构"□"，但不是一笔，而是"丨㇆一"三笔。

篆书的"丨"，不偏不斜，楷书已不存在这种笔形。楷书点笔，书写习惯上常常带有一定偏斜度。落笔偏左者，事实上就是短撇，落笔偏右者，事实上就是捺笔)。此外的㇀笔（由左下落笔于右上），事实上仍然是撇笔。

可见，在楷书早已成为正式字体的宋代，篆书在一般思想认识上仍有很大的残余影响。但郑樵对汉字基本笔画的分解研究，十分符合我们今天所要求的"一分为二"的观点，是最早最难能可贵的、对汉字基本笔画的辩证分解研究成果。若按楷书标准，去粗取精，去杂取纯，得到的正是楷书汉字的"一丨丿㇏㇆𠃊"这六种基本笔画。

在笔数方面，古代图画文字，只要画成其物，多一笔少一笔，都没关系；楷书则不然，笔数全有一定。在笔形方面，楷书的横、竖、撇、捺、弯、拐，区别显然，各笔有各笔的一定笔形和笔向。笔画的画数是抽象的，非本质的表现，笔形则是具体的、本质的表现，所以我们着重分析的便是笔形。

楷书的"一丨丿㇏㇆𠃊"这六种笔形，按其方向性区分，不外如下两种：

1．单向笔画　横笔（"一"）是从左向右横书而成的一笔，竖笔（"丨"）是从上向下直书而成的一笔，撇笔（"丿"）是从右上向左下斜书而成的一笔，捺笔（"㇏"）是从左上向右下斜书而成的一笔。

2．复向笔画　弯笔（"㇆㇇>"）是从左到右按90°角度弯向正下方或按45°角度弯向左下方书写而成的一笔，拐笔（"𠃊∠く"）是从上到下按90°角度拐向正右方或按45°角度拐向右上方书写而成的一笔。某一些字中弯后带拐和拐后带弯、一气连续写成的复杂笔画也是一笔。例如："乃弓"的"㇌㇉"，同为一笔，分类时按其第一个弯（或拐）归属弯笔（或拐笔）。

二、笔画的两个方面——笔形和画数

笔画形体和笔画数量，同为"笔画"的两个方面，既矛盾又统一。形成字体结构时，二者有如下不同表现。

1．单笔结构　笔数同为抽象的"1"，笔形则有"一丨丿㇏㇆𠃊"六种各不相同的具体形象。

2．二笔结构　笔数同为抽象的"2"，但笔与笔间形成的具体形象，则多达数十种，其中有表现在接触关系方面的，也有表现在位置关系方面的。接触关系的有：

散离结构　二刂〃儿八丷氵⺀巛等；

连接结构　丁厂𠂆匚丂丄卜　⺆冂𠂉亻　人勹冖ㄷ　亠㇋工ユ　刀凵刂　亼夕等；

交叉结构　十𠂇匕七犭乂力九七也等。

位置关系的有：

独体结构（各笔画互不离散，或连接，或交叉，囫囵一体）人丁乂𠂇；

左右离散结构　川八犭丷巛；

上下离散结构　二彡犭灬。

3．三笔结构　笔数同为抽象的“3”，具体结构形象则互有差异。例如：

独体结构　（连笔结构）工丌兀千上下万口个夕久饣习阝尸巳弓山乜等，（交笔结构）艹毛尹毛大丈扌女犭等；

左右离散结构　川丷巛；

上下离散结构　三彡习。

4．多笔结构　笔数同为抽象量，但具体结构形象则互有差异。例如：

独体结构　（连笔结构）扌卡凸凹臣丐丐勿日月欠爿占龟等；（交笔结构）丰韦车韦事秉中聿等；

左右离散结构　灬州洲；

上下离散结构　竺。

由上列各例可见：

（1）笔数上的抽象表现各只一种，笔形及其结构关系上的具体表现则繁多而复杂。

（2）离散关系的结构功能很有限。笔画愈多，离散关系愈不利于区别。所以散式二笔结构还比较多，三笔以至多笔结构就依次减少，甚至全无。

（3）连、交关系有利于使复数笔画紧密结合而形成一个特征突出、区别明显的独体结构，其中以连、交式二、三笔结构最多，在字体中出现的频度最大。连、交式多笔结构，频度不大，个数也较少。

（4）在复笔结构中，同种笔形的结构功能很差而异种笔形的结构功能则很强。

（5）在复笔结构中，表现为左右、上下关系的结构极少，表现为独体的结构特多，但在字体结构中则是左右结构很多，上下结构较少，独体结构最少。

2.3.4　字根

一、字根的意义和个数

位点组成笔画，笔画组成字根。字根是汉字形体的基本结构单元。

字根可以在通用的二百多个部首的基础上，根据应用情况，按照一定标准，决定选择个数。过去人们对“字根”这一概念还比较模糊，无统一名称（有中文字母、形母、组件、部件、构件、元件、基本符号等等），也无选择标准，以至所选数量多少不一（少的只十个，二、三十个，多的多到一、二百个以至四、五百个甚至一千多个），十分混乱。现在我们初步认为应以那些不宜分拆的原为部首的基本结构单元在单字总数中的出现频度，作为选择标准。我们曾以4356个常用汉字分解统计，字根的出现频度依次如下：

口732，木496，氵433，人430，扌360，土339，日317，门312，十298，艹284，亻278，丶229，勹220，八217，又209，大201，忄182，㐅168，广167，冂165，阝160，月159，小155，𠂉152，刀149，二145，贝141，讠137，女131，纟130，

宀129，二127，田121，广112，卜105，冖105，讠103，山102，彳101，七99，金98，匕97，三93，廿93，匚90，王88，火87，文85，禾80，尸75，虫75，目73，厂72，广68，礻66，灬63，子62，罒58，白56，车55，犭55，米54，大53，石53，罒52，巾51，口51，工50，足47，止47，⺮44，疒43，弓42，习41，马41，牛41，礻40，门39，羊38，酉37，幺35，臼34，雨34，钅32，方30，户29，穴29，矢28，……(以下从略)

这个统计，虽因所用单字总数（只4365个汉字）不多而不够精密，但却已足够作为一个重要的参考了。

汉字字典检索中以及一般汉字信息处理中所需要的字根个数，以控制在一百个左右最为适宜（因为易学、易记、易用）。特种专业所用字根个数以200～500个为宜（因可减少击键次数）。

二、单结构字根和复结构字根

只一笔就形成一个独立结构的字根，叫做“单结构字根”。二笔或多笔形成一个独立结构的字根，叫做“复结构字根”例如：“一、灭、归、么、升、主、乙、司、断、乱”等字结构中的“一、丿、乀、乙、乚”便是单结构字根，其中的“火日口𠂉艹王米舌”等则是复结构字根。现在在单结构字根中，仅“一乙”是有意义的单字，其他都是无意义的单笔结构。

三、字根结构的形态特征和笔画数特征

字根结构形态特征之一是笔画相互间的“单、散、连、交”式的关系表现。在所选定的下列字根中，只有单笔关系的字根是“一丨丿乀冂乚”；只有散笔关系的字根是“二三八冖丷艹讠习”；只有连笔关系的字根是“工厂匚歹卜止冂月口足日目四，广亻白𠂇夕饣勹冖广宀户彐尸已弓刀⺕阝匕凵山𠂉”；只有交笔关系的字根是“十艹扌子七车女牛乂犭力九又”；有散、连两种关系的字根是“贝亻灬疒门穴讠礻衤之纟”；有散、交两种关系的字根是“忄小米”；有连、交两种关系的字根是“耳王土西廿木大田巾虫角毛禾火子”；散、连、交三种关系都有的字根是“雨酉舟鱼”。

这种笔画与笔画之间的关系（单笔、散笔、连笔或交笔关系），从小学认字、习字时开始就已有所认识。但在很长一个时期内，一直是一种感性认识，还没有上升到理性认识阶段。“感觉到了的东西，我们不能立刻理解它。”千百年来，人们能在不知不觉中从速度缓慢的认字、习字开始，自然而然地进到速度飞快的看书认字，便是由于人们早已熟识笔画结构的这种单、散、连、交关系，使这种结构形态特征，能在一瞬间之内作为一个综合整体由视觉反映给大脑。但是自从篆书演变为楷书以来，一直未能成为一种科学总结，使其在汉字检索方面有所应用，以致人们一直忘了笔画结构形态这一最具体、最重要的特征，也一直忘了视觉这一闪电式的敏捷功能。以致明代梅膺祚著《字汇》(1615年前后）时开始“按笔画数”这一抽象数量关系编排了三万多个汉字。在汉字应用过程中，由直观的看书认字发展到按笔画数量这一抽象规律来查字，也可以说是认识提高了，进了一步了。但汉字的笔画数量一般都不少，不能一看就辨别清楚，一定要一笔一笔，合计到最后一笔才行。加之同一笔数的字也非常普遍。不要说几千几万个字如此，就是简化到如上所列的100左右个字根来说，一画字根有“一丨丿乀冂乚”，共6个；二画字根有“二厂匚十𠂇七卜冂八𠂉亻𠂆勹刈讠丷亠氵刀⺕力九又匕凵山𠂉”，共27个；三

画字根有“三工土艹扌 大女口口巾彳 饣犭 灬氵 广门宀之忄 小习⺀尸已弓子阝 马山”，共30个；四画字根有“歹王廿木车止月贝日罒钅牛 毛户礻 火纟”，共17个；五画字根有“石目四田白禾疒礻 穴”，共9个；六画字根有“耳西虫⺮舟米”，共6个；七画字根有“酉角足”，共3个；八画字根有“雨鱼”，共2个。可见，在100个字根中，笔画数特征有八种。每一种必须一画一画，数完最后一画，才能精确地知道它一共有多少画。一般都不能一见字就立刻把它的画数精确地判断出来。

但是，在同样的100个字根中，结构形态特征同样只有“单、散、连、交、散连、散交、连交和散连交”等八种，其中四种各只有一种形态，三种各只有两种形态，仅一种有三种形态特征。每一字根的形态特征平均为：

$$(1+2+3)\div 8 = 6 \div 8 = 0.75$$

但每一字根的笔画特征则平均为：

$$(1+2+3+4+5+6+7+8)\div 8 = 36 \div 8 = 4.5$$

可见，每一字根的笔画数平均特征比其结构形态平均特征大六倍。所以结构形态特征能一见就立刻精确地判断出来。笔画数特征就不能如此。

由此可知，“一丨丿 丶乛乚” 6种笔形及其“单、散、连、交”四种关系，是实现快速检索工作的最主要的条件。

2.3.5 部首和字首

一、部首的来历

“部”有“统属”、“门类”、“部类”等意义。远在春秋战国时代，已有六书学说，对当时汉字初步有所分解研究。东汉时代，许慎开始在其所著的《说文解字》中说：“此十四篇五百四十部九千三百五十三文。其建首也，立一为耑，方以类聚，物以群分，同条牵属，共理相贯，杂而不越，据形系联，引而申之，以究万原”。他把9353个汉字分类为540部。每一部类中的汉字都含有各字所同有的一个表意结构。列为第一个部类的是“一”部；“一”部中又把“一”字立为类首，排列在该部各字的最前头。后来就把这个带队的表意汉字叫做“部首”。虽然作为部首的某些汉字，后来已演变成全无字义的纯粹符号（例如手部的“扌”，水部的“氵”等），但“部首”的意义以及“部首”在汉字分类应用上的功能却一直保持未变。

分部（或分类）是向科学进军的重要途径之一。“科学研究的区分，就是根据科学对象所具有的特殊的矛盾性，因此，对于某一现象的领域所特有的某一种矛盾的研究，就构成某一门科学的对象。”汉字所特有的矛盾性由“形、音、义”三者体现出来。《说文解字》依据篆书分部，标准约有如下四种：

1．按义旁分部　例如“攴”（攵）部的“攸败牧”，“木”部的“某朱休”等，便全是由表义结构合成的“会意字”。

2．按形旁分部　例如“木氵土”等部的形声字以及“樂巢（木部）雷（雨部）胃（肉部）母（女部）禽离（内部）”等合体象形字便是。

3．按声旁分部　例如“殺”部的“弒”、“喜”部的“憙”便是。

4．按反义反形分部　例如“⼯”部的“⼞”（反⼯）、“彳”部的“亍”（反彳）、“正”部的“乏”（反正）便是。但“匕”（反人）又作独立部首，不归人部。

许书按汉字结构中一个共通义符（形旁）选作部首的字，有：

1．纯象形字　日月雲雨气火山厂阜氏水泉川井大土田㠯臣民弟人子女首目自口牙耳亢而手广又吕止力幺工尸鬼鸟乌佳燕於羽牛羊马犬豕鹿象虎豸兔龟鼠易虫鱼它巴已黽角皮革毛肉来禾米韭瓜艸竹木甲系絲巾鼎舟方车豆壺勺匕酉几且琴弓矢刀戈矛勿册玉丹贝斗斤网率曲缶也广宀門户瓦等。

2．合体象形字　彡虍屮匚臼石京包谷足巫血束才鬯弋齿金龍衣身能等。

3．指事字　一二三四五六七八九上爪行攵克矛鬥乃卜爻交厶凶出乙齐耑飛西冓口冂永午日只欠甘示豐干豕天交凵面尺寸亦采大久勹不至高等。

4．会意字　六十士半吉走音用鼻骨青老赤黑幸此从舛牧休㗊品㗊㗊㤅劦多雔玨等。

能“画成其物”的象形字不可能很多（许书9353字中仅364字，占3.89%，所以又依据“视而可识，察而可见”的现象创制成为数更少的指事字（又叫“象事字”，仅125字，占1.33%）。进而依据象形字和指事字，“比类合谊”（“谊”即“义”），按意义互相组合，如此所能造成的会意字（又叫“象意字”，仅116字，占1.24%）也很少。由于依据物象、事象和意象表达成字的可能性十分有限，所以又进而以语音为主，作为音符，并以其形体类属作为意符，两相配合，创制成为数最多的形声字（在许书总字数9353字中多达8748字，占93.54%）。可见，绝大多数汉字（包括会意字和形声字）全是由象形字和指事字组合成的合体字。可以说，象形字和指事字是汉字演化增多的根源、基础，所以现代人们就把它叫做“字根”。

有许多象形字并不单象一物，如“爨能雷母樂巢”等便是。有的指事字也不单指一事，如孕乎秀”等便是。许书对这一类象形字和指事字，有的作独立部首处理，有的则归入他部处理。例如：“雷”归“雨”部，“樂”归“木”部，“孕”归“子”部，“秀”归“禾”部等。而“爨”则作独立部首，其下所属仅二字。“易”也作独立部首，下无所属，一部仅部首本身一字。可见，许书选部归部的情况，虽有一定科学意义，但存在下述一些问题。由于（1）折衷并列，混杂不一。（2）选定部首和分部归类，若不一一熟知六书学说，则有很多构造复杂的字（如“喜”、“爨”等）以及那些已不象形（或象事）的篆书（如“易、母”等），就很难判断它是不是部首。而且字中部首部位不一（上、下、右、左，中全有），此外还常有省略部首，从而这一类字也很难正确分部归类。（3）对于选作部首的字，象形、指事、会意（如“喜”等）皆有，标准不一；加之，“殺”部的“殺”字虽是一个“形声字”（“杀”声，“殳”形），但其下所属的“弒”字竟然是二声（“杀”是声，“式”也是声）合成的纯声字，这已超越六书学说之外，无法解释。（4）部首数目定得过多（共达五百四十部），其中又突出地存在着“字多部少，字少部多”的矛盾现象，因而难学，难熟，难用。鉴于上述原因，许书的编排方式是很值得商讨的。

篆书演变为楷书后，汉字结构，面目全非。字音和字义，变异也很多很大。许慎以后约经过一千五百年，直到明代梅膺祚著《字汇》时，才大刀阔斧地把许书540部简化为214部。这些简化后的部首，虽然已有很大改进而且一直通用到现在，但因不能彻底摆脱许说影响，以致未能明确选部理由，更不能进一步统一规定字中部首所在部位，不能很好地贯彻简化办法，所以，在检索上一直存在着难学、难熟、难用的老问题。

二、选部、归部的巨大演变及其规律

许慎是按篆书和六书学说选定部首的。当时的篆书，构形上已和古代汉字有一定差

异。例如：像蜥易的“易”字已完全不像蜥易，而许慎固守六书学说，选作部首，部内别无它字，形成一个光棍部首。这种一部仅一个字或二、三个字的部首还有：

毛、耑、丐、能、𠂤、燕等部（一部仅部首本身一个字）；

孱、哭、只、古、共、異、熊、𠂇、左等部（一部仅两个字）；

半、是、品、叢、隶、色、包、卤等部(一部仅三个字)。

如此等等，所选定的部首竟多达540部，未免烦琐。直到明代梅膺祚，在字体结构变异更多更大的楷书条件下，大胆归并简化，选定了214个部首，这是一个巨大的改进。但是对于这214个部首是依据什么标准选定的，各字分类归部的标准是什么，几百年来，这两个问题一直未能得到明确答复。

直到1963年，《辞海》(未定稿)在其“部首查字法查字说明”中才有比较明确的如下两条规定：

（1）依据字形定部。一般采取字的上、下、左、右、外等部位作部首；其次是中坐和左上角。按照以上七种部位都无法确定部首的，列入余类。

（2）部首共250个，按笔画数排列，同笔画数的按一（横）丨（竖）丿（撇）丶（点）𠃌（折）五种笔形顺序排列。另有余类，排在最后。

这个选定部首的标准，开始明确了“依据字形定部”。这在“部首法”中是一个难能可贵的标准，使人们开始能够果断地彻底地摆脱六书学说的拘束，按照现代通用的楷书汉字结构和现代要求来考虑问题。

然而，由于部首选定标准早已越出六书学说之外，概念十分混乱。《辞海》（试行本）虽说依据字形取部，但部首所在部位又无一定标准，有上、下、左、右、外、中坐和左上角七种部位以及其例外的余类。这在汉字结构复杂多变的情况下，仍不免岐异多端，从而难于顺利处理。

可见，由六书规律发展到纯字形规律，虽是汉字应用科学中选部、归部的一大跃进，但因规律杂乱，标准不一，所以部首查字法还是不能很好地得到发展和应用。

三、简化部首的标准

在楷书条件下，明确了“依据字形定部”以后，第二步要明确的便是“部首的意义和作用”。

（一）部首和字首的意义

可以说，“楷书部首”应该是某一类汉字形体中共同的一个表意结构，这好象只要对现有一万个左右楷书汉字字形进行一次分析统计，就会明确下来的。实际上，这件事并不那么容易。因为有不少楷书汉字中的那个表意结构已无法辨认，辨认部首的标准便只能依靠字形了。但这也不是那么简单的。例如：对于“一二元未”四字，由于其共通结构显然可以是“一”，所以这四个字都可以归入“一”部，而“二元未”这三个字又都可以归入“二”部。加之，“元”又可分解为“一、二、兀、儿”这四种结构，可能选归的部首有“一”部、“二”部、“兀”部、或“儿”部。因为同有“二”的字还有“示亍云井夫未”等，同有“兀”的字还有“髡鬀虺光先完”等，同有“儿”的字还有“兒兄克兑见鬼允充兆”等等。由此可见，单凭“共通结构”这一标准，已不能顺利无阻地辨认部首，部首已失去其分类和检索的积极作用。为此特提出“字首”(即字体起笔处的那个共通字根)作为楷书汉字分类、检索的新标准。这样，由于“部首部位在字的上、下、左、右、外、

中”等多岐办法所形成的多岐情况就不存在了。这样，可以说，在汉字早已楷书化了的今天，字首也就是广义的部首。二者名异实同。

（二）简化标准

若仅按“起笔处的共通结构”这一标准来选定字首，则字首个数可能很多。例如：“垂甀郵舌甜刮舔舐天喬蚕呑斤丘乒乓兵岳血蠛衃衇衅衂”等字，既可通通按起笔选作“丿”部，又可分别选作“垂”、“舌”、“夭”、“斤”、“血”等部。由于今天主要是按楷书字形（不是按六书学说）选定字形中作为字首的那个共通结构，所以还必须考虑到那个共通结构在检索中究竟有多大作用。假若（1）它所属字数不多，（2）当作字体结构中一个字根看待时，它出现的频度不高，（3）它所应用的范围只限于一般检索，那就尽可能简化为好。若以上述这三个方面作为简化标准，那么就有如下情况：

第一，起笔是一横的“豕”（14字）、“至”（16字）、“而”（7字）等部，其字数都不多，故可并入“一”部。起笔是二横的“韦”部（仅14字），几乎全是生僻字，所以可以并入所属字数也不多的“二”部。起笔是三横的“髟”部虽有34个字，但绝大多数也是生僻字或简化后的废用字，所以可并入所属字数同样不多的“三部”。其他如“革”部，虽有44字，而绝大多数也是生僻字或简化后的废用字，所以可以并入所属字数不多的“廿”部。

第二，有一些不出现或很少出现在起笔部位的字根（如“文寸鸟欠心……”等）可以不选作字首。有的字根（如“羊”）虽可选作字首，但考虑到以其起笔构件（如“羊”的“丷”）作字首时，可以包括较多字数，所以也就可以不选作独立字首了。

第三，一般字典或索引，字首过多，实无必要，因既费脑力，又费工效。即使在汉字信息编码问题中，这也是重要的考虑因素之一。这是因为，它牵涉到键盘键数的多少问题。

由此可见，从实用角度出发，把几百个部首简化为100个左右字首，是较为合适的。

2.3.6 单字

一、单字的意义

位点构成笔画，笔画构成字根，字根构成单字。单字是代表一定“形、音、义”的、结构完整的一个意义单位。现在选定的所有字根，原来都是单字。发展到现阶段，有一些字根已失去其“音、义”，已不是单字，而成为一种纯粹的符号单位。例如在100个简化字根中，这种符号单位就有：

单笔字根　“丨丿 ㇏㇕𠃊” 5个

散笔字根　“爫丷氵冫冫” 5个

连笔字根　“匚冂囗𠂉亻亠讠礻衤冖⺈勹龴阝凵”15个

交笔字根　“扌𠂇乜㐅犭忄” 6个

在这31个符号单位中，对于“乛爫氵冫讠礻衤冖扌犭忄”等，现在还不能把它们原样算作单字，必须分别将其改写为“乙爪冰水言示衣幂手犬心”，才是单字。其余“丨丿……丷⺌亠匚……”等20个字根，目前已无音和义，纯粹是一种符号单位。

二、单字构形的量变特征

在100个简化字根中，能成为单字的单笔字根，只有“一”和“乙”。除了符号字根20个和“爫氵冫⺌”等11个要改写后才能单独成为单字的字根外，其他能成为单字的单独字根一共是66个。可见，在几千个单字中，绝大多数单字都是由两个（或多个）字根

4．内外型　字体内一个内根被一个外根全部（或局部）包围着的散式构型便是“内外型”。例如：图式为  的“囚困”等，图式为  的“冈罔区”等，图式为  的“这历司”等。内外型的单字构型虽多，但为数较少，故一律作上下型处理，这样最简便，最适宜。

 型字和  型字全是由古代会意字和形声字演变而来的合体字。

字型是汉字信息的重要指标之一，分类过分繁琐，反而于事无补。

独体型包括单式、连式、交式三种，而左右、上下、内外三型则全属散式一种。在一万个左右的汉字中，独体字约占10%，上下字（包括内外字）约占30%，左右字约占60%。

字体结构中的连、交式独体结构，虽可分拆为两个（或多个）字根，但仍应和其它字根一样作为一个结构单元看待。所以按字体结构中所含单元数来说，汉字可以分为：

一根字：三、月、鱼（单式独体型单字）和天、中、弗（连、交式独体型单字）等；

二根字：明、蚕、费、磐等（散式左右型或上下型单字）；

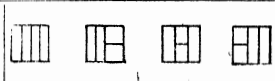
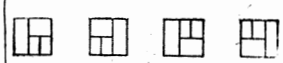
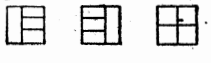
三根字：彻、曼、盟、敏等；

多根字：疑（四根字），赢（五根字），赣（六根字）等。

汉字结构中所含单元的多少，是量变，是抽象表现；字式和字型是质变，是具体表现。二者互相配合，演化发展，成为千差万别的成千上万个汉字单字。其中一根字，即独体字，浑然一体，只一型。二根字，由两个独体字组成的合体字，独体和独体间显然有隙可辨，共两型。三根字，是由一个独体和一个二根合体组成的合体字，共六型。四元字，是由一个独体和一个三根合体或由两个二根合体组成的合体字，共二十二型。四根以上的合体字极少，可舍弃其冗余部分，概括作四根字来处理。

表2-1列出了汉字的各种类型。

表2-1　汉字类型总表

	一根字	二根字	三根字	四根字	合计
独体型	 田月由中天母且 甫疋等				1种
左右型		佃 仲 肌	 湘 沼 邵 概 海 动 娜 塘 郁	 州 褓 湖 鹕  碗 鸪 溜 剔  漫 鹭 疑	15种

组合而成的。

按照单字中所含字根数量的多少，可以把单字分为以下四类：

（1）单根单字：一女十木口日马又力等。

（2）二根单字：从劝权旦早杏另等。

（3）三根单字：树查曼驾萌盟等。

（4）四根（或多根）单字：楂碳疑爵壹恣戆等。

在八、九千个汉字中，单根字和四根以上的多根字都不多，字数最多的是二根、三根字。由此可见，按字根数来区分单字的这一方案，要远远优于按笔画数来区分单字。

三、单字构形的质变特征

字体内各字根相互关系间的表现形式，是单字结构的最重要的质变特征。它有两个方面：一是“结构方式”，即“字式”；二是“结构类型”，即“字型”。

（一）字式的种类

字式是字体结构内各字根间体现在接触关系上的一种结构方式，有如下四种。

1．单式　字体内部构件浑然一体的、不能分拆的、单独存在的一个原始字根，它所体现的结构方式便是“单式”。例如：日月木火鱼雨等。

2．散式　字体内字根与字根间只有离散关系的那种结构方式便是散式。例如：“盟”字的“日、月、皿”三者间间隙显然，所以是散式结构。

3．连式　字体内一单笔根与一复笔根（或一复笔根与一单笔根）相连，这样的结构方式便是连式。例如：

单复相连：天（一大）正下千夭  耳（𠂇耳）

复单相连：丕(不一）韭业少尺（尸丿）久（夕丿）。

起笔（或末笔）是一个捺点的字（主永良义术凡兔等）也常作为连式处理。

4．交式　字体内字根笔画互相交叉。这种结构方式便是“交式”。例如：未(“二小”相交)，末(“一木”相交)，兆(“  儿”相交)，  (“林爻”相交)，臾(“臼人”相交)，等。

字式是单字结构的一种具体形态特征，无论是出现在字体结构的开始、字中、或字末，都极为突出明显，识别速度异常快，在检索应用上要远远优于笔画数那样的抽象数量特征。

（二）字型的种类及其辩证演进规律

字型是字体内各字根相互间的一种结构类型，有如下四种：

1．独体型（图式：  ）　独体字是由古代象形字和指事字演变而来的单式（或连式、或交式）单字，结构紧密，独自一体。它的构型便叫“独体型”。例如，

单式独体型：三石鱼米山等（单根结构）

连式独体型：天下千少尺等（复根连笔结构）。

交式独体型：夫丈事秉半坐等（复根交笔结构）。

2．左右型(图式： )　字体内的左根与右根间互有一定间隙的散式构型便是“左右型”。例如：相鸠邢炳锉等。

3．上下型（图式：  ）　字体内上根与下根间互有一定间隙的散式构型便是“上下型”。例如：杏荚蚕杂岩等。

（续）

	一根字	二根字	三根字	四根字	合计
上下型（内外型）		宙男卷笑贯	曼　茄　盟 毒　藕　墅 粪　宛　盥	蔓　荔　葬　巽 露　鉴　辙　照 鼐　樊　琵	15种
合　计	1　种	2　种	6　种	22　种	31种

2.4 汉字字音

2.4.1 汉语和汉字

语言的表现形式是“音”，其潜在内容是“义”。文字的表现形式是“形”，其潜在内容是“音、义”。二者同出一源，同含一义；但在演进中，互相影响，互相淘汰，各有新生。情况十分复杂。

汉字的“音”，源自汉语的“音”。语音来自人们的发音器官（唇、齿、颚、舌、鼻、喉、口腔、声带等）。各个主要的发音部位，叫做“音位”。来自各个不同音位的语音的最小单位，叫做“音素”。汉语普通话共有元音音素（主要的）6个，辅音音素22个。一个元音音素可以单独成为一个音节。一个辅音音素则必须配合上一个（或二、三个）元音音素，才能形成一个音节。每一个汉字就是一个音节，其中读音纯为元音音素的汉字不多，多数汉字每字读音都含有一个辅音和一个（或二、三个）元音（其中辅音部分叫“声母”，元音部分叫“韵母”）。但每一个汉字的读音并非互不相同，一音（一个音节）多字的现象十分普遍。例如《新华字典》收编的8500个字，纯元音字只有a，ai，an，ang，ao，e，ê，ei，en，eng，o，ou，ya，yan，yang，yao，ye，yi，yin，ying，yo，yong，you，yu，yuan，yue，yun等单音，包含一声母和一韵母的字则有lea，lai，……ca，……zuo等共397个单音。其中除ê，ei读音各仅一字外，其他各读音，一音最少由三、四个字、十几个字，几十个字，多到一百二、三十个字的都有。可见，无论是汉语或汉字，同音现象十分突出。但在社会实践应用中，除信息处理技术外，由于语言环境和上下文等多种客观因素，同音现象所产生的矛盾并不突出，它表现灵活、清楚，应用婉转自如，不愧是世界语言、文字中的一个奇迹。

2.4.2 汉字反切法和拉丁字母式双拼法

由于汉字不是拼音文字，因而标注每字读音时仍然只能采用另外的汉字。办法有两种：

一、“直音”

这是一种用一个汉字标注另一个读音相同的汉字的方法。由于汉字读音多到四百多

种，即令每种选定一字作注音标准，也难记难用。再加上汉语方言众多，变异复杂。此地所选，不合彼地所用；此时所选，异时异地又常有音变的可能。问题不少。

二、“切音”

“切音”又叫“反切”。这是我国用两个汉字标注一个汉字读音的传统方法，实质上也就是一种“双拼法”。例如：“马”，“莫下”切。意即把前字“莫”的声母(相当于m-)和后字“下”的韵母(相当于-a)分切出来拼读，发音就是“ma”。没有声母仅有韵母的字，则用一个没有声母的字代表“零声母”。例如：“乌”，“哀都”切（“哀”声母为“零”，“都”韵母相当于-u音)；“亦”，“羊益”切（零声母是“羊”），等等。

我国远在公元前200年前后就已正式应用反切。但由于下述三个原因，这种反切法就逐渐淘汰了：

（1）同一个音常常不用同一个汉字表示；即使是同一个音用同一个汉字表示，但因汉字读音多达四百多种，故难于记忆和应用。例如，声母相当于b-音的字及其反切有：包（“布交”切），崩（“北滕”切），逼（“彼力”切），等等；韵母相当于-a音的字及其反切有：八（“博拔”切），插（“楚洽”切），乏（“房法”切），妈（“莫下”切），拏（“女加”切），等等。

（2）用作声母、韵母的汉字又常常由于方言或语音变异等关系而枝节丛生，难于通用。例如：“但”、“徒旱”切(“徒”的声母现在相当于t-，而“但”的声母是d-)；“私”，“息夷”切（“私”的声母是s-而“息”的声母则是x-)；“悲”，“府(bu)眉”切；“飙”，“甫(bu)遥”切；等等。

（3）汉字结构本来就不像拼音字母那样简便，在切音（即拼音）应用中又甩不了它那无用部分（即反切时所不用的“韵母”、“零声母”，或“声母”），故十分繁冗难用。

最近几十年以来，这种传统的用汉字拼汉字的反切法已不用了。然而作为一种“双拼注音法”来说，它和汉字同样，却是几千年以来世界文化史中永难凋谢而更将异彩独放的奇花。反切法使每一个拉丁字母也像每一个作反切用的汉字那样，既可作声母又可兼作韵母。因此，每一个汉字便可一律用两个预先规定了读音的拉丁字母来注音。近十多年内发展起来的以古代“反切法”为基础的“拉丁式双拼法”，作为汉字编码的表音部分，已经逐渐形成信息处理领域中的一个不可忽视的流派。

另外拉丁式双拼文字用于信息处理技术，有相当的经济价值。

2.5 汉字字义

汉字信息处理中用得最多的是字音。因为汉字虽然普遍存在着一音多字的同音现象，但按照1957年版的《新华字典》中8500个汉字来说，单音不过四百多个。而这四百多个单音，又只要26个拉丁字母中的若干个字母便可拼写出来，既容易，又方便。当然由于存在同音字的问题，只将拼音键入计算机是不能准确无误地打印出所需的汉字来的。此外，对那些不知其读音的汉字也是不能随意键入的，而且，各种不同方言地区的人也常常会拼错字音。由此可见，拼音法虽简捷，却也还存在不少问题。

字义方面，问题更为复杂。因为有8500个汉字就有8500种基本定义。一般人不知字义的字可能还普遍多于不知字音的字。即使各字定义都已熟识，但一个字大都不只一

个意义，一般常有2～5个意义，有的多达6～9个意义；有少数字，一字多达十几种意义。假设平均一字有四种意义，那么8500个字便有8500 × 4 ＝34000种意义。这在汉字编码技术中是一个很伤脑筋的问题，我们就不深入研究了。

2.6 汉字排序

2.6.1 汉字排序的意义

汉字是在历史长流中创制出来的。它的发生和发展，有先有后，自然是一种有序过程。我们说位点组成笔画，笔画组成字根，字根组成单字，在字体结构的演进过程中，各有一定内在规律，这也是一种有序过程。我们写字，由起笔到落笔，也总是有一定顺序的。这是书法规律。我们读字正音，用字正义，也有一定规律。这些规律，不外都是一种有序表现。汉字字音繁多，演变复杂，所见常常因人而异。但现在汉字要和电子计算机发生联系，必须把汉字的这些有序特征统一作出科学总结，逐一编定人们便于应用而机器又便于处理的序码，以便我国各行各业都能得心应手地利用电子计算机，使其在四化建设中发挥最高经济效益。

可见汉字排序的意义是十分重大的。

2.6.2 汉字排序法

汉字排序法很多，约有如下几种：

一、千字文式排序法

在科学技术极不发达的旧时代，一般常作序码用的文章是《千字文》、《百家姓》等。《千字文》的一千个汉字就相当于一千个序码。人们常说“天字第一号”，但“某字第几百几十号”就从来没有听人说过了。可见在一千个汉字中只有开始几个汉字易于顺口溜，能作序号代码。序号越大越难记，越不便实用。

二、流水式排序法

这是一种随机排序法。例如医院挂号，病人依次取号，按序就诊。又如电报号码，收编多少字便依次有多少号。这种排序法的好处是无虚号，便于统计，比较经济。但只便于按码叫号，不利于检字取码，因而除专职人员外，其他人员难以应用。

三、拼音字母式排序法

汉语拼音所用拉丁字母是26个，一律按ABC顺序排列，极便于应用。但由于有一音多母的二十多个复韵母，以致一字代码有短有长，长者多达五、六个字母，故在信息处理上很不经济。

双拼式同样应用26个拉丁字母，按照汉字反切办法，每一汉字一律由两个字母拼成，前面一个字母代表声母（其中包括zh，ch，sh），后面一个字母代表韵母（其中包括ai，ao，en，eng，uan，ueng等复韵母）。

但是汉字双拼法和汉语拼音法有同样情况，同音字太多，不便单独应用。但若结合字体结构，另用一个或两个字母代表字形，便可能形成一种音形结合的汉字编码方案。

四、拆字定码排序法

最初的汉字是一笔笔画成的象形字或象事字，每字浑然一体，故名独体字。后来用二个（以至三、四个）独体字互相配合，便创制成为数最多的一个个合体字。在约一万

个汉字中，合体字约占 94％以上。合体字可以分解为独体字，而独体字也就是合体字结构中的字根，字根可以分解为笔画。首先，笔画可以排队成序，按序定码。其次，作为字根的独体字也可以在种种特征量的统查下排队成序，按序定码。最后，便可把合体字的结构类型以及各型中的字根顺序总结出来，排队成序，按序定码。如此，汉字结构的内在规律便逐一体现为序码，实现“见字识码”，像西文打字那样使汉字在电子计算机上显示和打印出来。现分述如下：

（一）笔形的种类和顺序

汉字笔形，在楷书已成为规范字体而印刷技术又有了巨大创新的宋代，已有郑樵(1161年)依据笔形演变的辩证规律，作出了很有价值的论述。经过明清两代，演进到现在，楷书笔画各有定形，笔画种数则一直没有定论（最少的说有 4 种，最多的说有 72 种）。解放以后，汉字基本笔画，其概念日趋明确。笔画的名称及其先后顺序，可依次总结为“横、竖、撇、捺、弯、拐”六种。极短极短的“撇”或“捺”，就是一般所说的“点”。这一科学总结所依据的标准约有如下五个方面：

（1）宋代郑樵《六书略》中的理论性总结。这是对笔画辩证演化规律的权威性理论总结。现在去粗取精，笔数和笔序便依次是“横竖撇捺弯（𠃌）拐（乚）”六种。

（2）《康熙字典》以及《辞源》、《辞海》等字书对汉字笔画数的区分法。《康熙字典》一画字有“一丨丶丿乙亅乚𠂉…”等，二画字有“儿冂七乃马…”等，三画字有𠃌弓（フ-ㄣ）丸阝…”等，四画字有“𠂆丐丐兮及殳弔……”等，五画字有“甘瓦癶世凸(13-1-)凹(1217-)目…”等。可见复折笔画有两种（乙𠃌和勹々），各为一画。

（3）笔画频度。我们曾按“一丨丿(𠂇)㇏(丶)𠃌乚”这六种笔画对九千个汉字的起笔和末笔初步作过统计（见表）：起笔频度的高低依次是“一㇏丿丨乚𠃌”，末笔频度的高低依次是“丿𠃌一㇏乚丨”。起笔出现次数最多的是“一㇏(丶)”，末笔是“丿(𠂇)𠃌”。可见宜以“一”笔排列第一而以其矛盾对立面的“丨”笔次之，以“丿”笔排列第三，而以其矛盾对立面的“㇏”笔次之，以“𠃌”笔排列第五，而以其矛盾对立面的“乚”笔次之。六笔笔序宜为“一丨丿㇏𠃌乚”。表 2-2 列出了这六个笔形的起笔、末笔使用频度统计数字。

表2-2　汉字起笔、末笔频度

笔形	起笔		末笔		首末笔合计	
	字数	%	字数	%	总字数	%
一	2753	24.9	1961	21.80	4714	26.20
丨	1270	24.33	871	9.60	2161	11.97
丿(𠂇)	2087	23.19	2041	22.70	4128	22.94
㇏(丶)	2240	24.90	1104	12.30	3344	18.60
𠃌	194	2.15	1962	21.80	2156	11.97
乚	436	4.84	1061	11.80	1497	8.32
合计	9000	100	9000	100	18000	100

（4）坐标图中的笔画。平面上的任何一笔都可表示为直角坐标轴$x'x$和$y'y$上的数量x和y。按直角坐标规则，x的正规方向是从左向右，y是从下向上。任何一个位点都可以用它的坐标（x、y）的两个值来表示。由点集合而组成汉字笔划。横笔的习惯书

法是从左向右，竖笔的习惯书法是从上而下，两笔交叉，习惯书法是先一后丨（如：十），先丿后㇏（如：乂），先㇆后㇄（如：乜）。

（5）点阵图中的笔画。由于汉字笔画一般都不少，因而高质量的文字最少要32×32的点阵才能把汉字笔画的风格表现出来。一般微型机容量不大，只用15×16或16×18的点阵，使表现出的每一个汉字清瘦到似乎仅有骨架结构。这种点阵中的笔画，不宜按毛笔书写体区分，可概括地分为一丨丿㇏㇆㇄六大类较为合适。

综合以上五个方面来说，汉字笔画数目及其顺序宜是（1）一，（2）丨，（3）丿，（4）㇏，（5）㇆，（6）㇄。

（二）字根的数目和顺序

首先，我们曾对4356个汉字分解后，按其出现频度选出约一百个字根作为排序样本。

第二，各字根先按起笔的笔形，依次分为一丨丿㇏㇆㇄六类；其次按各根内部结构的质变关系，分为单笔根、散笔根、连笔根和交笔根四类，然后再按其次笔形排列，于是，各根便依次成序，如表2-3所列。

表2-3　各种笔形的组合排序结构

	单笔根	散笔根	连笔根	交笔根
横类	一	二 三	王工耳，雨西酉，厂石歹，匚	十土艹木寸，大，七车女
竖类	丨		卜止，口口口足，日月目贝田四	巾 虫
撇类	丿	亻八𠂉	𠂉毛牛(牜)禾竹，亻白舟，人金食，𠂉勹夕角鱼尸	犭 乂
捺类	㇏	丷灬氵礻	讠广疒，门宀穴户，讠礻衤之	米小忄火
弯类	㇆	习	彐己弓，子，刀，𠂆，阝，马	力九又
拐类	㇄		匕，凵山，𠂇，纟	

第三，要明确字根在字体结构中的顺序。我们知道，百分之九十以上的汉字都是由那些为数不多的、可以作为一整个字根看待的独体结构合成的合体字。合体字有左右型和上下型（包括内外型）两类。在字体结构内，左和右，上和下，其间间隙显然。例如：


或

，各为1、2两个独体合成。

与

，

与

，

与

，各为一个独体与一个合体合成。

与

各为两个独体与一个合体合成。结构中每一个互相邻接着的双联体都是一个合体。结构中的三联合体，以左(或上)一单元为独体，其余两个单元为合体。任何汉字皆可如此分解，根序显然不紊。例如“懿”，首先可以看作是一个

型。左为

型，上为一独体（即土），下为一合体（即𠂇与豆）。右为

型，上为一合体（即亻与欠），

下为一独体（即心）。如此，按照“先左后右，先上后下，先左右后上下”这一习惯规律，“懿”字各根的顺序是  ，“夔”是  （其中3与4，5与6各为一双联独体）。一切内外型汉字一律并作上下型处理最简便。例如：因→吴，匠→  ，近→，司→，等。

以七千个左右汉字来说，五、六个根的合体字为数不过二、三十个。各字结构一目了然，不费思索，处理便捷。

（三）单字的数目和顺序

古今汉字约有六万多个。以前是按部首和笔画数排序。同部首同画数的字很普遍。1980年公布的国家标准汉字信息交换用字符集——基本集收容6763个。其中一级字3755个，按汉语拼音字母顺序逐字实排定码；二级字3008个，按部首逐字实排定码。这是计算机内应用的交换码。人手键入时，只有一字一键的整字键盘或笔触式字盘能沿用此码，从而可在不知码情况下找字按键输入。字母数字键盘一键以至多键一字，可如下文所说，各按不同字形特征顺序编码输入，然后，由机器查表转换成交换码。

1．整字按笔画结构排序法　先总结一个字中的笔画种数及其先后顺序，然后按照一定笔顺规律逐笔编码排序。这种方法的优点是简单易学。但由于汉字笔画普遍较多（每字平均约十二画），同画数单字普遍不少，一字平均码长不可能很短，若要略笔减码，则有二义性情况等等。这一切都是需要踏踏实实进行研究的。

2．字根排序法　字根来自单字，故必须确定下述一些内容：（1）作为采根对象的汉字字量（例如国标一、二级共6763个汉字）；（2）字根的构形特征（例如：散笔构形、连笔构形、交笔构形等）；（3）分解原则（例如断笔不断笔等）；（4）字中字根的位置关系（例如左右关系、上下关系等）；（5）字中各根的顺序（例如按书法惯例与按顺时针方向等）；（6）表列各根的顺序（例如笔数顺、笔形顺、音顺、义顺等）等等。目前，这一方面的研究成果很多，不难从中做出科学总结。字根的数目和顺序明确以后，便可以见字拆根成序。今后，人们便可以像西文打字那样利用标准字母数字键盘输入汉字。

2.7　汉字信息处理与汉字属性

研究汉字信息处理，首先需要研究汉字的输入编码问题。对输入编码的研究同掌握汉字的属性密切相关。此外，在设计字根式汉字库工作中，对汉字字形属性的分析研究也是十分重要的。汉字的每一个字，以至每一个字根都是形、音、义三种属性的一个统一体。它与西文不同，既可以见形知义，也可以由音知义。因而，汉字与拼音文字的字母并不相同。实际上，有些汉字相当于拼音文字的词，有些汉字相当于拼音文字的词素。而有些汉字本身既是词，却又可以作为另一个词的构词词素。此外，汉字所存在的一字多音（或称同形异音），以及一音多字（或称同音异形），却给输入编码带来了很大麻烦。所以认真研究汉字属性，尽可能做到使汉字编码简单易学，操作方便，重码率低，输入迅速，这对汉字信息处理技术的发展具有重大意义。反过来，汉字信息处理技术的发展也推动了对汉字属性更广泛深入的研究工作，并提供了现代化的研究手段，对我国古老的汉字总结出更科学的结构规律，使他更丰富多采，并使他有更强的生命力。

第三章　汉字输入编码方法

3.1　汉字输入编码概述

在汉字信息处理技术中，汉字输入问题是一个一直受到重视、并有待妥善解决的问题。

汉字是记录汉语的图形符号，利用光学扫描方法将汉字的图形信息直接读入计算机，本应是汉字输入的一种较好的方法。但是，由于汉字的字量极大，字形繁杂，因此，向计算机直接输入汉字的图形信息，并使它能够识别和处理，其难度较大，成本较高。目前，要使汉字识别技术成为普遍的实际输入手段，还有一定距离。

利用声音识别技术，将汉字的字音直接读入计算机，也是汉字的一种输入方法。但是，目前它还处于研究阶段。

目前，键盘输入设备是应用最广的设备，是进行人机联系的基本工具，在一定时期内，它仍然是汉字信息输入的主要手段。

对于拼音文字，由于其字母的数量不多，故采用键盘来输入拼音文字信息较为方便。但是，如何利用键盘来输入汉字，却是一个比较复杂的问题。

有人曾研制过一种一字一键式的汉字键盘。但是，由于汉字字量太大，使得键盘的键数也多，体积大，成本高，不便操作，不切实用。为了减少键数，后来发展了一种一键多字式键盘，键盘上另加了选字键（又称移位键），用以选定一键上某一个汉字。这种主辅键式的整字输入键盘，在日本曾大量使用。但是，由于我国的汉字量比日本汉字多得多，使用一键多字也很难包括我国通常使用的所有汉字。而且这种设备的体积仍较大，成本也较高，操作时需双手并用，因此，其适用范围受到一定限制，不利于普及推广。从缩小键盘尺寸着眼，近年来发展了笔触式（pen touch）字盘，这种设备体积较小，造价也较低，初学者容易掌握，适合于在某些专业性系统中应用。但是，由于它不能实现“盲打”，操作时需较多时间注视盘面，不但影响输入速度，也容易出错，输入速度也较低，因此，也难于普遍采用。

为了提高汉字的输入效率，人们认为选用普遍使用的字母数字键盘是比较合适的，而且也是有可能的。为此，要对汉字进行编码，利用输入代码的方法，来取代直接输入汉字的方法。这就是我们要讨论汉字输入编码方法的原因。

汉字输入编码与汉字字典的查字法是一脉相承的。任何一本汉字字典都有一种或几种对汉字进行分类排序的查字法。就其实质而言，查字法也可以说是一种汉字编码法，只不过用于信息处理中的汉字编码法与用于字典中的查字法，在要求上有所差异而已。字典查字法是供查阅字典用的一种汉字分类方法，其主要的要求是见字即能定其类属，只要同一类属中的汉字不是很多，一般都可使用。例如，东汉时的许慎，按照汉字的形声特征，以形为主，对汉字进行分部归类，形成了沿用至今的按部首查字的方法。这种分类方法符合汉字的客观属性，所以，两千年来，部首查字法一直是汉字分类检索的基本

方法之一。按照部首对汉字分类，各部所属字数的分布是不均匀的，一个部首多的可有数百个汉字，少的则仅有几个汉字，从而使查字很不方便，因此，需要考虑对同部首的字进一步分类的问题。传统的作法是按笔画数的多少进行再分类。这种方法对用字典查字来说，虽不理想，但却是可用的。对汉字编码法来说，由于同部首、同笔画数的汉字常常不是一个，仅依部首和笔画数来进行编码时，将出现若干个汉字对应于同一个代码的重码现象，而且这种情况较为普遍，因此，一般是不适用的。

一百年前，在我国创办汉字电报通信中，将汉字按照《康熙字典》的214个部首分类，然后再按笔画数从少到多顺序排列，从而列成了电报通信用汉字字表。表中每一汉字用四位数目字顺序编上号码，故被称为“四码电报”。这是汉字代码的一个最早应用实例。四码电报虽然用了汉字的部首特征和笔画数特征，但因重码较多，在码型结构中，放弃了对部首特征和笔画数特征进行明确描述的设想，而采用了按流水线编号的编码方法。这种编码法对一般人而言，使用起来是相当困难的，只有经过训练的专职人员才能熟练使用。因此，这种电报编码法也不便于在汉字信息处理技术中普遍采用。

在以计算机技术为基础的信息处理技术迅猛发展的今天，为了寻求能被普遍采用的汉字输入编码方法，国内外许多研究人员作了大量工作，据统计，目前汉字输入编码方案已逾四百个，其中有不少方案已在一些汉字信息处理系统中试用，并有了可喜的效果。但是，总的说来，汉字输入编码技术仍不能适应实用的需要，尚需作更加艰苦细致的探索工作。

为了帮助读者对汉字输入编码方法有个基本了解，本章将从一些基本概念出发，进而结合汉字特点对汉字输入编码方法作一般描述，最后，对汉字输入代码的设计方法和评测方法加以讨论。

3.2 汉字集及其划分

3.2.1 按汉字天然属性划分的子集

从集合论的观点出发，我们按照汉字的字音、字形、字义等天然属性将其划分成若干子集，分别进行讨论。

一、按字音划分的子集

汉字虽然复杂，其字音却简单而有规律。除极少量的字(如千瓦、英寸等)在正式用语中已被废弃外，一般都是单音字节。现代汉语普通话中共有四百一十多个音节（未计儿化音节），各音节都有音调变化，带调音节共一千二百八十余个。一般按音序排列的字典或音序检字表，都已将所考察的汉字集，按字音划分成了四百多个同音字子集或一千二百多个同音同调字子集。利用字音属性设计汉字输入用代码，需要研究解决的问题有二：一是如何区分同音字；二是如何描述字音，即如何给各个同音子集命名。

二、按字形划分的子集

汉字的字形结构是有一定规律的。

汉字是由笔画组成的，按照汉字笔画特征，可以将汉字集划分成不同的子集。以笔画数目为特征，汉字集可划分成数十个大小不等的子集，其中，笔画数目相同的汉字同属一个子集。一般字典中的笔画数检字表便是例子。麻烦之处在于，要对笔画数进行计数，各子集汉字数目也有较大差别。汉字的起笔笔形是很醒目的，有些小字典便是

按起笔类型来划分子集的。这种方法简便易行，但子集数较少，且各子集所容字数的分布颇不均匀。若把汉字起笔笔形分为横、竖、撇、点、折五类，则根据9000汉字的统计，起笔为横的子集字数为2753个，占30.6%；起笔为竖的子集字数为1290个，占14.3%；起笔为撇的子集字数为2078个，占23.2%；起笔为点的子集字数为2240个，占25%；起笔为折的子集字数为630个，占6.9%。类似地，也可按末笔笔形划分子集，不过这时各子集中的字数将更不均匀。例如，在同样统计中，末笔为撇的子集字数为99字，而末笔为折的子集字数则拥有3023字。子集字数差别的悬殊，对汉字输入代码的设计是很不方便的。

利用和起笔、末笔相关联的笔形组合这种特征，也可把汉字划分成不同的子集，这就是所谓按“上下形”来划分子集的方法。由于笔形组合有一定的自由度，故可以根据统计规律和需要来调整和确定，从而有可能使子集类型和子集容量设计得较为适中。这样，凡“上形”和“下形”相同者便被列入同一子集。例如，上形为“艹”下形为“灬”的字有{蒸燕薰蕉蘸燕……}等，它们属于相同的子集。为区别子集中不同的字，可用追加补充信息的方法解决。例如，为了指出“燕”字，则只要追加“是子集中的第二字”这一补充信息，便可唯一确定了。由于这种方法忽略了上下形以外的汉字信息，因此，“重码”的现象较为普遍。为了避免“重码”现象，一般要依赖系统的提示来追加补充信息，因此，这在使用上受到一定限制。

也有按照与四角号码查字法中相类似的“角形”划分子集，以及按照与传统的偏旁部首相类似的“字根”划分子集的。由于角形的取法以及字根的取法较为多样，故使得这方面的内容相当丰富。基于这些子集划分方法，产生了许多以字形为主的汉字编码方案。

在按汉字的字形划分子集时，要研究解决究竟选定哪些字形属性特征来划分子集，以及如何给这些子集命名等问题。

三、按字义划分的子集

汉字是一种望文可以辨义的表意文字。虽然经过几千年的演变与发展，汉字字义已变得十分丰富。但是，归根究底，引伸或转化后的字义与该字的本义总还存在着一定的渊源关系和逻辑联系。所以，按字义划分子集时，宜按字的本义确定其归属。由于汉字的部首是汉字本义的明显表征，因此，按字义划分子集的方法，同部首检字法的分部归类法颇为相似。在汉字输入代码中充分利用部首信息，这对某些类型的汉字编码法是很有意义的。不过，有如下问题需要考虑：

（1）给定哪些部首较为合适？

（2）如何给部首命名？

（3）有些汉字的部首已演变得不很明显（如“衣”部的衰、衷、衮等）；有的甚至不易确认（如“衣”部的表、袁等）；有的可同属两个部首（如“鸿”本应属“鸟”部，“杲”本应属“日”部，但某些字典分别将它们列入了“氵”部和“木”部）；还有些字则是难以定出部首的，等等。对于上述这些情况，都要采取相应的解决措施。

（4）按部首划分的子集，其子集容量悬殊，少则几个或几十个字，多则有数百乃至上千字，这也应妥善处理。

3.2.2 字母集上的有序组

我们将两个按一定次序排列的容体 a 和 b 组成的有序序列（a，b）叫做有序组，a 和 b 分别称作第一和第二容体。两个有序组（a，b）和（c，d），只有当 $a=c$ 且 $b=d$ 时，才认为（a，b）=（c，d）。

有序组中容体的个数称为有序组的维数。例如 (a_1, a_2)、(a_1, a_2, a_3)、$(a_1, a_2, \cdots a_n)$ 等有序组分别称为二维、三维、n 维有序组。对 n 维有序组 $(a_1, a_2, \cdots, a_n)$ 和 $(b_1, b_2, \cdots, b_n)$，只有当 $a_i=b_i$（$i=1, 2, \cdots, n$）时，才可以说两有序组相等，并记为 $(a_1, a_2, \cdots, a_n)=(b_1, b_2, \cdots, b_n)$。

对于各种拼音文字，实质上乃是拼音字母集上的不等长有序组的集合。如英文中的 a，ab，aba，abb，等等，就是英文字母集上的有序组。用有序组的概念来描述字母集上的拼音文字是较为方便的。

汉字不是拼音文字，但若能按照汉字特有的规律，建立一个与拼音字母集相类似的汉字字母集（character set），不也就可以通过有序组来描述和处理汉字了吗？

先看一个简单的例子。若将汉字中常用的构字元件（字根），依照书写顺序来排列，则汉字便是字根集上的有序组。如：

设 a = 口，b = 木，则有（a，a）= 吕

（a，a，a）= 品

（b，b）= 林

（b，b，b）= 森

（a，b）= 呆

（b，a）= 杏

（a，b，b）= 啉

（a，a，a，b）= 桑

（a，a，a，a，b）= 噪

（b，a，a，a）= 椞

扩大字根的数目，将能得到表示更多汉字信息的有序组。例如，再增三个字根：

c = 艹（草字头）

d = 氵（三点水）

e = 亻（立人旁）

则可增加能表示汉字信息的有序组有：

（e，b）= 休

（e，a，a）= 侣

（e，a，b）= 保

（c，a，a）= 莒

（c，e，a，b）= 葆

（c，d，a，a，a，b）= 藻

（d，b）= 沐

（d，b，b）= 淋

$(d,\ a,\ a,\ a,\ b)$=澡

如果把这种字根集上的有序组看作是相应汉字的代码，那么在将汉字用代码输入时，便和拼音文字的输入十分相似了。

n 维有序组中的容体性质和数量可以是不相同的，上例中的容体都是字根。下面再举一例：人们常用“弓长张”、“立早章”、“耳东陈”、“禾呈程”这样的叙述方式来描述汉字。这是一种用三维有序组描述汉字信息的方法。在这种描述模式中，第一容体描述汉字图形前半部分称谓，第二容体是后半部分的称谓，第三容体是汉字整体图形的称谓。此时便是

(Gong, Chang, Zhang)=张

(Li, Zao, Zhang)=章

(Er, Dong, Chen)=陈

(He, Cheng, Cheng)=程

或略写为

(G, C, Z)=张

(L, Z, Z)=章

(E, D, C)=陈

(H, C, C)=程

这样的有序组也可看作是一种汉字编码方法。

一般说来，任何一种能够表示汉字信息的有序组，都可作为一种汉字输入编码方法的模型。但是，由于汉字的复杂性，在具体实现时都遇到不同程度的困难。例如，对于上述的字根模式，要能表示出数以万计的汉字，所需字根数往往不是几十个，而是数百个甚至更多。对于“弓长张”模式，要所有汉字分解出“半边字”并给出称谓来，这是十分困难的，当仅取缩写字头时，还会增加更多的重码现象。

3.2.3 汉字的笛卡尔积集

当有序组 $(a_1,\ a_2,\cdots,a_n)$ 的容体 $a_1,\ a_2,\ \cdots,a_n$ 分别是集合 $A_1,\ A_2,\ \cdots,\ A_n$ 的元素，即 $a_i\in A_i$ $(i=1,\ 2,\ \cdots,\ n)$ 时，有序组的全体将组成一个新的集合，我们把它称作 $A_1,\ A_2,\ \cdots,\ A_n$ 的 n 维笛卡尔积集，并记作

$A_1\times A_2\times\cdots\times A_n=\{(a_1,\ a_2,\ \cdots,\ a_n)|a_i\in A_i,\ i=1,\ 2,\ \cdots n\}$，同时，我们称 A_i 为 a_i 的属性集。

我们仍以“弓长张”模式来考察“张、章、陈、程”四个汉字的集合。此时，$n=3$，且有

A_1={Gong, Li, Er, He}或缩写成 A_1={G, L, E, H}

A_2={Chang, Zao, Dong, Cheng}或缩写成 A_2={C, Z, D}

A_3={Zhang, Chen, Cheng}，或缩写成 A_3={Z, C}，

则 $A_1\times A_2\times A_3$={(Gong, Chang, Zhang), (Li, Zao, Zhang), (Er, Dong, Chen), (He, Cheng, Cheng), (Li, Chang, Zheng), …, (Er, Cheng, Cheng)}

或缩写成 $A_1\times A_2\times A_3$={(G, C, Z), (L, Z, Z), (E, D, C), (H, C, C),

(L, C, Z), …, (E, C, C)}

这里，第一维容体长度为 4，第二维容体长度也为 4，第三维容体长度为 3，笛卡尔积集 $A_1 \times A_2 \times A_3$ 的体积为 $4 \times 4 \times 3 = 48$。而缩写后，第一、维容体长度仍为 4，第二维容体长度由于原容体 Chang 与 Cheng 混同为C，故减少为 3。类似地，第三维容体长度减少为 2，笛卡尔积集的体积则减少为 $4 \times 3 \times 2 = 24$。

如果考察的对象不只是上述四个汉字，而是更多的汉字，那么情形将是怎样呢？显然，此时汉字的属性集 A_1，A_2，A_3都将扩展，而其笛卡尔积集的体积扩展得更快。实际上，对汉字来说，部首属性集 A_1 的元素约二百个上下，偏旁属性集元素约一千个，字音属性集元素为四百多个（不计声调，不计儿化），或一千二百多个（计声调，不计儿化）。它们的三维笛卡尔积集元素将在亿个左右。如果采用首字母缩写法，这些属性元素都将映照到以 26 个字母为元素的字母集或其中某一子集内，这时，笛卡尔积集中的元素为 26^3个。

汉字的笛卡尔积集，可以看成是汉字编码研究中的基本对象。

3.2.4 汉字代码集

我们称笛卡尔积集 $A_1 \times A_2 \times \cdots \times A_n$ 的一个子集 R 为由集合 A_1，A_2，…，A_n 所确定的一个 n 维关系（Relationship），显然，这时 R 是一些 n 维有序组的集合。

仍以上面考察的 $A_1 \times A_2 \times A_3$ 为例，它有一个子集 R 如下：

R = {(Gong, Chang, Zhang), (Li, Zao, Zhang), (Er, Dong, Chen),
(He, Cheng, Cheng)}

它有四个元素（element），每个元素都是一个三维有序组，它们分别描述“张、章、陈、程”四个汉字信息的一组代码。由这个元素所组成的集合 R 可以代表“张、章，陈、程”四个汉字，我们特地把它称作这四个汉字的代码集。代码集 R 关于积集 $A_1 \times A_2 \times A_3$ 的补集（Complement Set）$\sim R$ 为

$\sim R$ = {(Li, Chang, Zhang), (Er, Chang, Zhang), …, (Er, Cheng, Cheng)}

它共有 $48 - 4 = 44$ 元素。这些元素所描述的是一些形似汉字而实非汉字的图形，不属于我们的考察对象。

我们感兴趣的是由所选定的诸汉字属性集 A_1，A_2…，A_n 所构成的汉字属性笛卡尔积集 $A_1 \times A_2 \times \cdots \times A_n$，以及汉字在这个积集中的映象 R，即汉字代码集（Chinese code set）。

3.2.5 汉字输入编码的简单模型

由上可知，n 维笛卡尔积集 $A_1 \times A_2 \times \cdots \times A_n$ 的一个子集，即关系 R，可以用来对汉字进行描述。当 $n = 2$ 时，取

A_1 = {x | x 是汉字的部首}

A_2 = {y | y 是汉字的偏旁}

R = {(x, y) | $x \in A_1$，$y \in A_2$，存在一个有意义的汉字，其部首为 x，偏旁为 y}

显然，当所考察的字表中，所有汉字的部首都属于 A_1，所有汉字的偏旁都属于 A_2 时，R 就是该字表中所有汉字的一种描述。如表 3-1 所列。

表3-1　字根组字表

A1 \ A2	金	木	水	火	土	口	日	月	目	各	佳	容	者	少	亢	戋	甘	吾	丁	尧
金				钬	钍	钔	钼	钥	钼	铬	锥	镕	锗	钞	钪	钱	钳	锢	钉	铙
木		林			杜	杏	杳		相	格	椎	榕	楮	杪	杭	栈	柑	梧	杠	桡
水	淦	沐					汨		泪	洛	淮	溶	渚	沙	沆	浅	泔	语	汀	浇
火				炎	灶					烙		熔		炒	炕			焐	灯	烧
土					圭						堆		堵		坑		坩			
口	唸	呆			吐					咯	唯			吵	吭		咁	唔	叮	
日		杲		炅			昭	明												
月								朋		胳					肮					
目											睢			眇					盯	
各											雒									
佳																				
容																				
者																				
少																				
亢																				
戋																				
甘																				
吾																				
丁																				
尧																				

当取　A_1＝{金，木，…，土}

　　A_2＝{金，木，…，丁}

时，二维笛卡尔积集的关系

$$R=\{\text{金，木，…，坩}\}$$

便描述了字表中相应部分的各有意义的汉字。

如果取$A=A_1\cup A_2$，并考虑二维笛卡尔积集$A^2=A_1\times A_2$上的一个子集R'，此时，$R'\subseteq A^2$，$R'=\{x\mid x$是有意义的汉字$\}$，则将有

$$R'-R=\{\text{唸，呆，吐，…，盯，雒}\},$$

即R'比R多出了“唸，呆，吐，…，雒”共22个汉字，这是由于A_1扩展了的缘故。这一扩展使积集中的元素由原来的5×20＝100，扩展到了20×20＝400，即增加了三倍，但关系R仅由原来的62扩展到84。用于编码时，表明了码空间的利用率由原来的62÷100＝62%下降到84÷400＝21%。虽然如此，代码集R'所能覆盖的字数还是有了显著的增加。

为进一步增加代码集 R' 所能覆盖的字量，可以扩充属性集 A_i 的基数。例如，为了把汉字“钯、把、吧、肥”等，按表 3.1 扩充进代码集中来，此时只需增加一个属性元素“巴”即可。如果按照“有序”的含义，这时汉字“吧”和“色”，都扩充进代码集 R' 中了，并且对应于同样的有序组（口，巴）。这就是说，在这种情况下，“吧邑”二字成了重码字。

原则上，可以通过扩充 A_i 而使代码集 R 能完全覆盖任何一个汉字字表。不过，随着 A_i 的扩充，编码效率将逐渐下降，重码字组也将逐渐增多。为此，不少编码方法都附加了某些约束来对重码字进行分化。A_i 扩充到何种地步，要保持怎样的编码效率，附加何种约束来分化重码字等，应根据系统的总体考虑作出决择。

上述的这种汉字代码二维模型很易推广成 n 维的情况（$n \geqslant 2$）：若汉字的属性集有几种类型 X_1，X_2，…，X_n，第 i 种类型 X_i 有 m_i 个状态（$i = 1, 2, \cdots, n$），即

$X_i = \{x_{ij} | x_{ij}$ 是汉字的第 i 类属性集的第 j 种状态（$j = 1, 2, \cdots, na_i$）$\}$

（$i = 1, 2, \cdots, n$）

我们可以使某汉字与其属性集的有序组（$x_1, x_2 \cdots, x_n$）对应起来，这里 $x_i \epsilon z_i$（$i = 1, 2, \cdots, n$）。并通过这种对应关系在汉字集 $H = \{h / h$ 是一个汉字$\}$ 与笛卡尔集 $X = X_1 \times X_2 \times \cdots \times X_n$ 的有序组之间建立起一种联系 f，于是，函数 f 便把 H 中的任一元素（即汉字）h 与 X 中的某一元素（即代码有序组）$x = (x_1, x_2, \cdots, x_n)$ 联系了起来，记为

$$H \xrightarrow{f} X$$

或　$f: H \longrightarrow X$，

即　$x = f(h)$，其中 $h \epsilon H$，$x \epsilon X$。

这种从汉字集 H 到笛卡尔积集 X 的对应联系是映照关系，映照函数（mapping function）f 便是编码规则，汉字集 H 是函数 f 的定义域，代码集 R 是函数 f 的值域。

这里对汉字属性集 X_i 只给出了抽象的涵义，并未要求具体内容，所以，这种关于汉字代码的简单模型带有普遍意义，不同类型的汉字编码方案都可通过它来进行讨论。

3.3 汉字输入编码的设计

一般地说，我们可以通过汉字输入代码的简单模型来对编码方法进行设计。为了达到预期的效果，模型中涉及到的汉字集 H、汉字属性集 X_i、汉字笛卡尔积集 X、X 中的特定子集——代码集 R、编码规则、编码效率、重码分化等等，都需要作大量统计计量工作。它与计量语言学、工程心理学、数学和文字学等都有着密切的关系，因而它是一门综合性的技术。为不致使篇幅过于庞杂，以下仅就有关的主要内容予以讨论。

3.3.1 字种的确定

汉字的字量很大，每个字的使用频度相差悬殊，这是人所共知的。一个汉字信息处理系统通常无须具有处理全部汉字字种的能力，这是因为在大约六万个汉字中，多数现在已不使用了的缘故。只是在较为特别的情况下，才有可能用到较多的汉字。

一个汉字信息处理系统所能处理的字量称作系统内字量，内字量随使用目标而定。

表3-2 前十号汉字的概率及其累计和

编号	政治		文艺		新闻		科技		综合	
	字	概率	字	概率	字	概率	字	概率	字	概率
1	的	0.0536	的	0.0324	的	0.0375	的	0.0320	的	0.0384
2	是	0.0165	一	0.0218	一	0.0132	一	0.0097	一	0.0125
3	一	0.0136	了	0.0196	了	0.0120	在	0.0092	是	0.0098
4	在	0.0115	不	0.0165	和	0.0086	用	0.0079	在	0.0095
5	这	0.0109	是	0.0141	在	0.0086	有	0.0073	了	0.0082
6	主	0.0108	说	0.0130	人	0.0083	是	0.0070	不	0.0081
7	不	0.0101	他	0.0130	大	0.0083	不	0.0069	和	0.0075
8	和	0.0098	这	0.0119	主	0.0083	中	0.0066	有	0.0069
9	人	0.0087	着	0.0107	是	0.0078	大	0.0064	大	0.0069
10	们	0.0087	个	0.0097	们	0.0065	时	0.0063	这	0.0064
累计		0.1544		0.1627		0.1189		0.0922		0.1141

表3-3 常用汉字集中情况

概率累计和	汉字序号				
	政治	文艺	新闻	科技	综合
0.05	102	96	132	169	163
0.90	650	860	780	900	950
0.99	1790	2180	2080	2250	2400
0.999	2966	3204	3402	3719	3804
0.9999	3917	3808	4575	5116	5265
1.0000	4356	3965	5084	5711	6359

为了较为客观地选择内字量，原则上应结合系统的使用范围，收集足够多的可能作为处理对象的汉字文字资料进行统计。这种统计可以给出被统计资料中的字量，以及各个汉字的使用频度（即各个汉字在统计资料中所出现的次数）。统计资料愈丰富，则愈接近于客观用字情况。不同汉字的个数及其使用频度是确定系统内字的主要依据，如果再辅以各字构词能力等语言学方面的考虑，或考虑使用环境的特点，那么便可确定出较为符合实际需要的字量来。这些汉字便构成了上述的汉字集H。

这种统计工作是十分浩繁的。表 3-2、表 3-3 是我国有关单位公布的统计资料的一部分，总共统计了政治、文艺、新闻、科技四大领域中两千一百多万字的文字资料，得到的不同汉字的数量为 6359 个。这项统计为开发我国汉字信息处理事业作出了重要贡献。

根据各种统计资料的分析研究，1981 年我国公布了《信息交换用汉字编码字符集基本集》的国家标准（即 GB-2312），它是汉字信息交换用的代码依据。不过，一般说来，这个标准所收容的字种数量为 6763 个。目前它已成了我国在汉字输入编码设计时选择汉字集的重要依据和最低要求。

3.3.2 汉字的熵值

汉字的熵值H（汉字）是完全确定系统中一个汉字所需平均信息量（以“位”为单位）的最小值，它可由下式算出：

$$H\ (\text{汉字}) = -\sum_{i=1}^{n} P_i \lg P_i$$

其中，n 是汉字集中的字量；P_i 是第 i 个汉字的相对使用频度。

汉字熵值是汉字总体的一个统计特性，从汉字编码来说，它给出了代码信息量的理论最小平均值。在对同一个汉字集进行编码时，不同的编码方法，其代码信息量是不同的。如果把理论值和实际值之比叫做编码效率的话，那么，不同编码方法的编码效率便可计算出来。在编码设计中，应该尽可能地提高编码效率。

计算汉字熵值是很困难的，目前大多停留在零阶熵的计算上。在某种程度上，它只是把使用情况集中的汉字当作一种没有实际含义的元素来处理的（尽管使用频度可反映出某些含义）。这与汉字是作为语言元素而存在的实际情况不相符 合。因 此，引起了人们研究高阶熵的兴趣。

3.3.3 选择键盘类型

不同类型键盘的键位数是各不相同的。一般整字方式的汉字输入键盘的键位有数百个或数千个，直接采用字根输入汉字的字根键盘的键位有数百个，作为一般计算机输入用的字母数字键盘的键位为数十个。键位的多少对键入速率有着明显的影响。据日本实务用字研究协会的统计，典型键盘的键入速率如下：

字母打字机（26 键）	450 次/分钟
假名打字机（50 键）	250 次/分钟
假名汉字混合打字机（2000 键）	50 次/分钟

键位数与击键反应时间，一般可用海曼（Hyman）公式来描述：

$$T = a + b \lg k$$

其中 T 为各键位以相等速率工作时每击一次键所需时间，k 为键位数，a 可看作为 $k = 1$ 时的简单反应时间，是因人而异的常数，b 是寻找键位时所需的选择时间，也是因人而异的常数。若各键位使用频率不等，式中应以

$$H(k) = -\sum_{i=1}^{k} p_i \lg p_i$$

来取代 $\lg k$，$H(k)$ 是键位的熵值，p_i 是第 i 个键位的使用频度。

由此可见，对输入键盘类型的选择是很有意义的，这随系统的实际使用环境作配置上的考虑。从采用编码输入汉字的立场看，目前在我国较为普遍的是希望选用字母数字键盘来实现汉字的输入，那么，它的基本键位数便是在 0 ～ 9 共10个数字键和 A ～ Z 共26个字母键这样的范围内。

3.3.4 选择汉字属性类型

汉字的属性集合有字音、字形、字义等等类型，每一类型中还有多种元素状态。当确定了属性类型 X_i（$i = 1, 2, \cdots, n$），便可得到笛卡尔积集

$$X = X_1 \times X_2 \times \cdots \times X_n$$

根据特定的规则，X_i 中元素的有序组应能覆盖系统的全部汉字，这些有序组便 构 成 了

汉字的代码集。

如何选择属性集及属性集中的元素状态，是编码设计中的一项基础性决择。它基本上代表了汉字输入代码的类型，也在很大程度上反映了编码方法自身的素质。所谓以形为主，以音为主、音形结合等的汉字代码类型，主要就是以所选择的汉字属性类型来分类的。

恰当地选择汉字属性类型及其中的元素数量和状态，应较全面地掌握有关汉字的大量统计材料，如：汉字字量、汉字使用频度分布，汉字构词能力的统计分析、构成汉字的字根使用频度分布、汉字字音分布规律、音形相关性的统计分析、字根间的组合关系和分解关系的统计分析，等等。全面系统地获得这些材料是相当困难的，在建立了各种类型的汉字属性数据库后，这一问题可望得到较好的解决。

目前的多数做法是，从某个侧面出发，以某些属性类型为基础，然后根据特定的编码规则试编，当发现用既定属性中元素，用既定的规则，会产生较多歧义甚至不能覆盖汉字集合时，除对编码规则作出调整外，再逐步吸收属性中更多的属性元素，乃至增加属性类型。例如，当确定选用字形属性的一定数量的字根元素后，给出这些字根的组字规则，当它们仍不能组合成汉字集的代码集的话，或者出现了大量重码时，就有必要调整组字规则，或者扩充字根元素的数量，乃至增加另外的属性类型。

汉字属性的选择，是一个较长的设计过程，往往需要经过反复实践和调整才能完成。

3.3.5 汉字属性元素在键位上的配置

确定了键盘类型和汉字属性类型之后，键位数和属性元素也就确定了。当键位数多于属性元素时，每一属性元素一般均可单一地配置在相应键位上。在相反的情况下，由于属性元素多于键位数，此时每一键位上配置的属性元素将多于一个，同一键位上的不同元素应避免在对应于有意义的汉字的有序组（即代码集）中作为相同的容体被采用，否则将会引起岐义。以字形属性为例，设属性元素有“才，木，氵，肖，…”，它们按照字根有序组构成汉字，当需要在一个键位上配置多个元素时，那么对于“才，木，氵”三个元素不宜配置在同一键位上，因为它们在相应的有序组（才，肖）、（木，肖）、（氵，肖）中都作为相同的容体被采用，否则将使汉字“捎、梢、消”产生混淆。

另外，从工程心理学观点来看，键位配置应尽可能符合操作人员的运指规律和能充分发挥各手指的功能。对于一般的字母数字键盘，常有三排键位，中排是手指的常驻键位，熟练后，可根据各键位与常驻键位之间的相对位置关系，以“音打”方式击键，提高键入效率。一系列的击键实验还表明：

（1）同指连击最慢；

（2）同手越排连击较慢；

（3）双手交替击键最为协调轻快；

（4）食指灵活，中指次之，小指较弱，无名指较笨。

因此，在考虑键位配置时，宜将各属性元素尽可能分布在三排键位上，进而把它们再按使用频度的高低分别配置在各相应手指所触打的范围内。这种配置也需要以大量统计数据作为基础。

3.3.6 重码数量预测

对于某一具体编码方法，在未完成代码表的情况下，准确预测重码数量是很困难的。但若在代码表完成之后发现重码过多，则必须加以调整。这种调整往往涉及到汉字集中的许多汉字，消除了一些重码字组，可能出现新的重码字组，使编码工作遇到困难。为此，以下给出一种估算重码数量的方法。

设某一编码方法选择了属性类型X_i，它们的元素为x_{ij}（其中$i=1, 2, \cdots, n$，$j=1, 2, \cdots, m_i$），在汉字集H中，元素x_{ij}所领有的汉字子集为(Z_{ij})，(Z_{ij})中汉字的数量记作$|Z_{ij}|$（不需十分准确），选择较大的$|z_{ij}|$，按下列公式

$$C_{ij}=\frac{|Z_{ij}|\cdot(|Z_{ij}|-1)\cdot m_i}{\prod_{i=1}^{n} m_i}$$

可以估算出该子集重码字对数量C_{ij}，它是与属性类型X_i有关的汉字子集中出现的重码字对数目。算出的重码字对，若为两字一组的重码以一对计，三字一组的重码以三对计，四字一组的重码以六对计，q字一组的重码以$q\cdot(q-1)/2$对计。

按此估算出的重码字对，如果在数量上可被接受，即重码字对数量在允许范围内，便可选择较大的(Z_{ij})进行实编试验。实验结果若与估算值接近，表明这种编码方法在区分重码字方面是有效的。一般实编结果均较大于估算值，若实编了几个子集，它们都与估算值较为接近，其平均值为估算值的E倍，那么可期望系统汉字集H中的重码字对数量C约为

$$C=E\cdot\frac{|H|^2}{\prod_{i=1}^{n}|X_i|}$$

式中，$|H|$为汉字集H的字种数；$|X_i|$为属性集X_i的元素基数；n为属性类型数。

3.3.7 代码表的编制

代码表（code table）的编制就是给出汉字集H中各元素（即汉字）在汉字笛卡尔积集中的各有序组——代码集。它是属性集中各元素按照一定规则排列起来的，每个汉字原则上对应于一个有序组——代码（假定忽略掉一字数码的多码情况和数字一码的重码情况）。通常的做法是，将每一汉字分离成若干个必须包含在属性X_i中的元素x_{ij}，然后根据建立有序组的规则，把x_{ij}配置在相应容体的位置上，由此组成的有序组就是该汉字的代码，建立了H中所有汉字的有序组，代码表的编制工作便算完成。例如，当X_1仅选择字根属性类型（此时$i=1$），而它的元素为（讠（言），讠，五，口，士（士），……），建立汉字有序组的主要规则是：按通常书写习惯排列各容体。于是，“语”字的有序组（即代码）为（讠，五，口），“洁”字为（讠，士，口），“诘”字为（讠，士，口），“吉”字为（士，口），“吾”字为（五，口），等等。对于非代码集中的有序组，如（讠，口）…等，因为不能组成有意义的汉字，它们不属于应考察的对象。完成了所有汉字的有序组，代码表便编制完成。

有序组中容体的数量（即代码长度）有固定的和不固定的两类，应事先做出规定。上述例子中的代码长度属于不固定类型。

编制代码表的过程中，常常仍会发生所用的编码规则或属性元素等选择欠妥，需不断进行修改和调整。目前的许多编码方法，大都经过多次反复实践才逐渐完成的。

为了提高输入效率，许多编码方法还制定了一定数量的“简码”或“词汇码”。一般它们应是上述基本码的简约形式，并与基本码在机内兼容，因而在系统中可以混合使用。简码或词汇码主要用于使用频度极高的汉字或某些常用的专门词汇术语等，对提高输入效率颇有裨益。

代码表是提供给用户的必备文件，供使用中参考，同时也是将输入代码通过相应键盘设备转换成机器内部码的设计依据。

3.4 汉字输入代码的类型

3.4.1 概述

1978年12月，我国召开《第一次全国汉字编码学术交流会》，会上提出了各种类型的汉字输入编码方案约40个。接着，汉字输入编码方法的研究，吸引了许多志士仁人，经过四年半的时间，于1983年5月所作的统计，各种编码方法已逾400个。1978年前仅有个别方法在计算机系统上实践过，到了1983年，则有40余个方法已在各种类型的汉字信息处理系统中获得了应用，为我国的汉字信息处理事业起了很大的推动作用。

初期的编码方法所需的键盘输入设备，大体上有主键-辅键方式和笔触字表方式等整字键盘、多种类型的字根键盘和字母数字键盘，目前则较为集中在笔触式整字键盘和字母数字键盘两个方面，大大减少了对输入键盘设备在种类上的要求。在汉字属性的选用方面，也由原来依赖字音属性偏多逐步过渡到依赖字形属性偏多的状况。这些都是汉字输入编码方法研究不断深化的结果。

通常，汉字输入代码的类型可以按照在编码方法中所使用的主要汉字属性来进行划分。目前较多被使用的汉字属性有字形、字音、字义、字频等属性，有的编码方法中仅使用某单一汉字属性，有的则混合使用多种类型的汉字属性。在汉字字形属性中，一般根据某种规则，把构成汉字字形的基本结构单位统称为字根。在讨论输入代码类型时，为清晰起见，我们还将使用某些其他名称，如笔形、角形、字元等等。对于整字输入方法的有关内容，请参阅本书第五章。这里仅就主要用字母数字键盘输入汉字的各种编码方法加以概括的叙述，并且为简明起见，并不对特定的输入代码作具体罗列，读者对某种特定输入代码有兴趣时，可自行查阅有关文献。

3.4.2 字根代码类

由于拼音文字是由线性排列的字母串组成的，故在对拼音文字进行输入时，就是顺序输入相应的字母串。每一文字的字母串，可以看成是该字的有序组，即代码。通常，方块汉字不能象拼音文字那样按照一维的符号序列来书写，它是一种二维的图形符号。但是，通过对汉字的分析发现，这种二维图形可以分解成若干基本单元，它在数量上远比汉字总数要少，如果按照一定的顺序规则，把汉字的基本单元排列起来，而成为一维

的基本单元串，那么它就是相应汉字的有序组，即代码了。这样，在输入汉字时，只要把组成该字的基本单元——字元顺序输入即可。

字根（radical），也称作字元，它的代码类型可作如下描述：

设有字元集X，即

$$X=\{x_i|x_i\text{是字元}，\ i=1，2，\cdots，\ k\}，$$

作X的n维笛卡尔积集Y，

$$Y=X^n$$

将汉字按某种确定的顺序（例如，可以按照通常的书写顺序），分解成字元的有序组（x_1，x_2，$\cdots$，x_m），其中$x_i\in X$。当$m\leqslant n$时，使（x_1，x_2，$\cdots$，x_m）与Y中元素（x_1，x_2，$\cdots$，x_m，0，0，$\cdots$，0）相对应，这里“0”可在逻辑上看成零。当$m>n$时，则按某种确定的约定弃去（$m-n$）个字元。例如从第（$m-n+1$）个字元起弃去（$m-n$）个字元，使（x_1，x_2，$\cdots$，x_{n-1}，x_n，x_{n+1}，$\cdots$，x_m）映照到（x_1，x_2，$\cdots$，x_{m-1}，x_m），通过这一映照而得到汉字集H在Y上的象集R，R便是H的代码集。

利用汉字的字元属性为汉字编码时，必须考虑的是如何选择字元以及字元的数量k，把汉字分解成字元后，按何种顺序排列成有序组以及有序组容体数量——代码长度n，各个字元如何命名以及在键位上如何配置等。

统计表明，当选定$n=2$，则k约为1800才能自然地覆盖约18000个左右的汉字。$n=4$时，为能自然地覆盖《现代汉语词典》中所收的11000多个汉字，k约需800左右。如果n不限，则用496～504个字元，可以自然地覆盖几乎全部通用汉字。由此看来，汉字的字元数量远比拼音文字的字母数量为多。同时，汉字的输入代码是面向操作人员的，对这些字元的命名和摄取方法，往往直接影响输入效果。随着对字元及字元数量的选择、字元的命名、字元摄取方法等的差异，形成了字元代码类型的多种多样的具体设计方法。

例如，有一种方法，总共选择588个字元，把汉字由字元组成的结构形式分为29种，每个汉字由1～4个字元组成，字元的摄取方法按通常书写习惯，由上而下，由左而右，由外而里的步骤进行，每个字元用两个字母来命名。这样，每个汉字的输入代码便可由2～8个字母组成，另加一个汉字结构形式的字母对不同结构形式加以区别，代码共由3～9个字母组成。如果字元和字母的对应关系为：立⇒EG，日⇒YY，刀⇒DI，口⇒YU，……，而“韶”字是由“立、日、刀、口”字元组成，则“韶”字的代码便是“EGYYDIYUG”，最后的“G”是“韶”字的结构代码。

另有一种方法，力图把组成汉字的字元选择成与《信息处理交换用的七位编码字符集》（GB1988）中的字符形象非常相似的单元，总共选择约90个基本字元，每一基本字元与GB1988中的某一字符相对应。每一基本字元可以派生出数个变形字元，它和基本字元对应于同一个GB1988的字符，这样便使数百个字元仅与GB1988中的约90个字符相对应。字元的摄取规则，根据汉字的“方块”形特点，按顺时针方向，从右上角开始，尔后为右下角、左下角，至左上角止。码长自然地规定为1～4码，多余的字元代码被舍弃。这个方法的字元与传统的偏旁部首有许多不一致的地方，但根据字母数字键盘原有键位上的字符，能引起操作人员的联想，在某种程度上可减少操作员的记忆负担。如果字元和GB1988中字符的对应关系有：厶⇒A，口⇒o，冂⇒n，二⇒=，日⇒Q，

人⇒r，乙⇒2，米⇒米，一⇒一，卩⇒p，……，则“命”字的代码为“Apo”，“明”字的代码为“n＝Q”，“氣”字的代码为“r2米一”，等等。

还有一种代码，通过对汉字字形结构的分析，选择了24个主字元，这些字元的代码，为便于记忆和联想，根据它们的起笔笔形（如横、竖、撇、捺、折等）分为六类，每个字元用一个字母或两个数字来表示。然后，按照字形结构的特征，对约400个付字元或从属字元，以主字元的编码规则，推定其代码。主字元及其代码可直接配置在键位上。取码时，按照习惯书写顺序，每字取1～4码，即1～4个字母，或2～8个数字。如果字元和代码的对应关系有：一⇒‖(S)，丿⇒31(O)，木⇒16(G)，目⇒24(A)，……，则“厢”字的代码应为“11，31，16，24”，或“SOGA”，等等。

在利用汉字的字元属性对汉字进行编码时，目前的基本做法是把构字能力低的字元尽可能地加以归拼或简化，或者将字元形象相似的加以合併，或者直接取它们的首末笔形来作为表征等等。这样，将有数百个字元被“同化”掉，使得在过去由于字元过多，不得不采用数百键位的字根键盘，有可能用字母数字键盘来实现汉字的输入。不过，在归并字元时，应注意在同化为一个字元中的各字元尽量减少码同字不同的重码情况，否则有可能使重码指标显著恶化。例如，“氵”和“忄”，习惯名称和形象都较为相似，自然地会引起将两者合并的念头，但实际上它们将产生大量重码字对：准，淮；冶，治；冼，洗；冷，泠；…；等，使“忄”部的39个字中与“氵”部的字生成22组的重码字。再如，习惯上有一种说法，即“水火不相容”，那么“水”和“火”是否可合并为一个字元呢？实际表明，这样合并也会产生大量重码字对：泡，炮；浇、烧；洪，烘；沙，炒；沪，炉；澡，燥；…；等，总共约80余对（虽然其中有些字是不常用的）。在给字元命名时也会遇到类似的问题，如字元“木”和“目”，当用它们的共同字音“Mu”命名时，那么有可能出现重码字对：棰，睡；枕，眈；根，眼；梱，睏；…；等近30对（虽然其中有些字是不常用的）。可见对字元的处理是十分重要的。

利用字元属性对汉字编码，直观性较强，不受方言影响，不认识的字并不影响代码的编制，由于它和汉字字形联系紧密，有可能与汉字的字形识别技术中的笔画——子图形——汉字的分层识别技术结合起来。

3.4.3 角形代码类

角形代码（cornor code）是利用汉字四个角上的字形结构特征来描述汉字的。假设选定的角形属性集合 X_i，它表示为

$$X_i=\{x_j|x_j \text{是汉字的角形属性状态，} \quad j=1,\ 2,\ \cdots,\ m_i\}$$
$$(i=1,\ 2,\ \cdots,\ n)$$

式中，m_i是第 i 个角位的属性状态数；n 是所选用的边角数。它的笛卡尔积集为

$$Y=X_1\times X_2\times\cdots\times X_n,$$

按照约定的顺序取汉字的角形属性元素作有序组 $(x_1,\ x_2,\ \cdots,\ x_n)$，并通过有序组相等的关系将汉字集$H$映照到$Y$中去，所得到的象集$R$，便是汉字的代码集。

当取 $n=4$，$m_i=10$（$i=1,2,3,4$）时，可以得到一般字典中常用的“四角号码查字法”的字码对照表（假定它的角形属性元素及取角顺序和四角号码查字法是一致的）。

当取 $n=2$，$m_1=45$，$m_2=24$，且取的“上下形”元素与《当代汉英大词典》所用的“上下形”元素一致时，便得到与该词典所用查字法相同的字码对照表。

取 $n=1$，$m=4$ 或 5 或 6 时，得到的将是与某些字典中所用的“起笔笔形查字法”相类似的汉字检字表。

取 $n=4$，$m=16$时，得到的将是曾在俄汉机器翻译系统中使用过的“新四角号码汉字编码”表。

取 $n=3$，$m_1=m_2=m_3=100$ 时，得到的是一种所谓的“汉字三角号码法”检字表。

常用的十进制四角号码查字方法，角形元素少，规则简单，易学易用，但同号字较多，大多数字均可能出现重码现象，有的重码字组甚至可达数十字或更多。如《现代汉语词典》中的4422号字就有 59 个。加上附号，用五位号码编排的《康熙字典》，其中的 44227 号字仍有 197 个。十六进制的四角号码编号方法，由于角形状态数量的增加，重码指标有了一定的改善，不过不加其他规则的约束，直接用于汉字输入亦是很不方便的。一百进制的三角号码法，大量增加了角形元素，在通用汉字的范围内重码可降至 3% 以下，因而在一些系统中得到了应用。不过这种三角号码法，除了 100 个主角形元素外，还有许多辅助角形元素，从而使角形元素共达 300 个左右，虽然这些角形的编号方法有某些联想规律可供记忆参考，但本质上仍是强制性记忆，易于混淆，非专职操作人员是较难熟练掌握的。

还有一种与字元编码相类似的“几何编码法”，它把汉字看成某些基本几何图形的组合结构，按约定的几何位置顺序对汉字依图赋码，这些基本几何图形包含在34个汉字之中，直接配置在34个键位上，顺序按键，便得到汉字的代码，操作人员不需记忆代码，由系统自行赋码，并且它还可以兼容“三角号码法”，在一定范围内获得了应用。

3.4.4 笔形代码类

笔形代码（stroke code）的编码方法，力图完全使用汉字的笔画形状或辅以少量笔形结构较为固定，且使用频率甚高的组合笔形来表示汉字的代码。这种情况下，由于笔画元素较少，因此汉字的笔形属性集合 X 也比较简单：

$$X=\{x_i|x_i \text{ 是汉字的笔形元素，} i=1, 2, \cdots, m\}$$

它的 n 维笛卡尔积集为

$$Y=X^n$$

根据由笔形构成汉字的有序规则，将汉字集 H 映照到 Y 中去，构成各个汉字在 Y 中的有序组，便得到汉字的代码集 R。

随着 m，n 选择的不同，形成了多种的汉字笔形编码方法，兹略举数例如下。

取 $m=8$，$n=3\times3$，即取 8 种笔形，基本码长为 9，建立有序组的规则为：先将汉字分成三块（不足三块时可只分作两块或一块，多于三块的依次取首次末三块，其余弃去），然后按笔形起点位置的高低或左右，以先高后低、先左后右为序取其三个笔形（不足三笔时可少取，多余的笔形则弃去，唯不足三块时则对末块连续摄取笔形），直至 $n=9$ 时为止。为提高输入效率，对常用字可用简码输入，简码是基本码的前部，码长不定，以能与其他汉字的代码相区别为原则，故在取简码时，其码长原则上应作强制性

记忆。

取$m=5$，$n\leqslant 13$。有序组按习惯书写顺序排列，当笔形超过13时被弃去多余部分。对高组字率的复合笔形，如“口，土，…，等”可另用指定的键位整体输入，也可按笔形序列顺序输入，两种方式在系统中可兼容并用，以提高输入速度。通常组成汉字的笔形数量较多，因此这种方法的平均码长偏长，然可按习惯书写顺序编码，较为自然。不过，相同的笔形有序组往往会产生多义性，如（一，一，丿，㇏）可以是“天”，也可能是“夫”，同样（一，丿，㇏，丶）可以是“太”，也可能是“犬”，因此形成的重码现象也较多，在使用上还需增加较多的区分重码的措施。

取$n=5$，$m=5\times 5$。将笔形按习惯分作横、竖、撇、点、折五种，以书写顺序为序，每两个笔形合为一键，将$5\times 5=25$种复合笔形配置在字母数字键盘的25个键位上，这样便可象写字一样按笔画顺序击键输入，两个笔形一键输入，也对键入速率有利。但笔顺不是完全规范化的，加之操作人员不同的笔顺习惯，往往产生岐义现象，为此应加强消除岐义的措施。例如，将较易产生笔顺岐义的常用字元，如“王、女、火、山、忄、阝、门、艹、鸟、可、方、龙……”等直接配置在键位上，作为习惯复笔笔形直接输入，在一定程度上可避免岐义现象。对未配置在键位上的其余易于发生笔顺岐义的字元，还可按多种笔顺习惯输入，使其最终生成等价的效果。

3.4.5 字音代码类

我国汉字几乎都是单音节字。根据1958年2月全国人民代表大会通过的《汉语拼音方案》，我国汉语普通话字音，不分声调时有410余个，分声调时有1280余个，一律采用26个拉丁字母来拼写。利用汉字的字音属性来对汉字进行编码时，实质上就是把汉字的拼音作为汉字的代码来使用。输入时，原则上可象其他拼音文字一样，直接按字母输入即可。即便附上声调标志，汉字的全部字音仅1280余个，而汉字则数以万计，因此同音字（在代码表中表现为重码）很多。如何区分同音字是按字音属性编码的一个非常重要的问题。

一般按字音属性编码时，都利用了汉字音节是由声母和韵母两部分构成的这一特征。声母有约22个，韵母约35个，由它们构成410多个音节（加四声时为1280多个）。此时，声母属性集合X_1和韵母属性集合X_2（计声调时还有声调属性集合X_3）为：

$$X_1=\{x_i|x_i\text{是声母},\ i=1,\ 2,\ \cdots,\ 22\}$$

$$X_2=\{x_i|x_i\text{是韵母},\ i=1,\ 2,\ \cdots,\ 35\}$$

$$(X_3=\{x_i|x_i\text{是声调},\ i=1,\ 2,\ 3,\ 4\})$$

它们构成的笛卡尔积集

$$Y=X_1\times X_2$$

$$(\text{或}Y=X_1\times X_2\times X_3)$$

构成汉字字音的有序组规则一般为声母在前，韵母继后（计声调时，再加上声调），把汉字集H中各汉字有序组映照到Y中去，便得到按字音属性编码的汉字代码集R。

由于声母、韵母、声调的总数量仅约60个，可把它们直接配置在字母数字键盘的键位上，操作人员只需知道汉字的拼音，便可按规定的顺序击键，从而完成汉字的输入工作。

在字音代码（phonetic code）类型中，有直接按《汉语拼音方案》的音素制方法，一个汉字的字音由1～6个字母组成，此法码长虽不固定，但与汉字的拼音形式完全相同。也有把汉语音素加以简并处理，每个声母由一个字母表示，每个韵母也由一个字母表示（由于韵母数比字母数为多，有必要使某些字母同时代表两个韵母），这是一种称为“双拼”的方法。在有序组中，声母、韵母分别占据固定的容体位置。这样做可使汉字的码长规整划一，不过必须对字母和声母、字母和韵母的替代关系应作周密考虑，避免由此而产生许多新的重码字组，也要便于操作人员的记忆。当需要辅以声调属性时，不论是音素制的方法或者双拼的方法，都应在有序组中附加表示声调属性的字母。

普遍认为，完全采用汉字字音属性为汉字编码，重码过多，使用中很易产生混乱。为了区分同音字，目前常采取的措施有：

（1）借助于系统的提示，进行二次选择。按字音属性有序组输入汉字后，如果它有同音字，则借助系统专门提供的功能，把同音字全部显示在荧光屏上，操作人员根据需要，再从这些同音字中选择出某个特定的汉字。

（2）采用词汇编码的方式。确切地说，汉字书面语言是以词汇为单位的、按分词连写的规则写出汉字词汇的字音，在通用词汇的范围内，同音词组比同音字组明显减少，仍有的同音词组可按类似于第一种方法处理，或者尽可能地扩充系统的智能，使常用词汇中的绝大部分同音词组，由系统根据语言学规律来自动作出选择，在某种程度上，它和日语输入中的假名汉字变换方法较为相似。

（3）除汉字的字音属性外，增加字形、字义、字频等其他一些属性，这正是下面将要讨论的内容。

3.4.6 音、形等相结合的代码类

这类代码中大致有：

（1）主要属性来自字音，但为了区分重码字（即同音字），增加字义属性（具体地说，大都选用汉字的部首，它也可看成是字形属性的一部分），字形中的笔形属性（例如起笔或末笔的笔形），或者根据字频统计材料附以字频属性等。在这种情况下的字音属性一般不采用音素制的拼写方法，而采用不同的简并处理（例如某种形式的“双拼”等），以减少代码长度。这样，声母、韵母、声调将各用一个字母来替代（不同的设计者确立的替代关系并不一样），为了规则起见，其他属性也分别用字母（或数字）来表示。例如把《新华字典》中的183个部首，按定义划分成数类，每类中各部首均由字母来代替，不同类别中的部首再用笔形字母区分。字频属性也一样，把声母、韵母、声调均相同的同音字，按它们的使用频度高低顺序用字母A，B，…（或数字1，2，…）来加以区分，等等。这种代码的码长，除简码外，一般是固定的，例如有序组大多由四个容体所组成。随着所用各属性在有序组容体中的对应关系的不同，其中有（部首、声母、韵母、声调）、（声母、韵母、声调、字频）、（声母、韵母、部首、笔形）等多种形式的有序组，它们都将由操作人员按照替代规则变汉字为四个字母来表征，通过字母数字键盘输入。不论哪种替代规则，都应便于操作人员的掌握，才能达到较好的效果。

（2）主要属性来自字形分解成的字根，并给各个字根命名，再按照字根组成汉字的顺序，以字根的字首属性元素构成汉字的代码。由于字根数量远比汉字数量为少，当给

字根命名后，每个字根也就有了确切的字音属性。有的为了避免音素制的字母偏多或类似“双拼”方法的声母、韵母的替代，字根的字音属性直接取自字根字音的第一个字母，这样，汉字便可由组成该字的各字根字音的第一个字母来表示。例如，假设“韶”字可分解为如下字根“立、日、刀、口，它们的确切字音为：Li、Ri、Dao、Kou，则有序组（LRDK）可作为“韶”字的代码而被用于输入。从分解字根来说，它和字根代码类的方法相类似，但它直观地取用了字根字音的第一字母来作为汉字的表征，对操作人员降低了汉语拼音的要求，不过在字根分解和字根摄取规则方面应作周密考虑，以减少重码字组和便于人们的记忆。

此外，还有许多不同的方法，在此不再列举了。描述上也都可以按照 3.2 中的简单数学模型来加以概括。

3.5 汉字输入编码方法的计算机辅助设计

计算机辅助设计，在各种工程设计中已获得广泛的应用。在汉字输入编码方法研究中引入机助设计的手段是近年来才出现的，目前还只能在一定程度上优化具体方法中的某些评测指标方面应用，尚不能将一个本质上缺陷很多的方法，通过机助设计手段来达到优化的目的。在特定的输入编码方法中，属性元素的调整；属性元素在键位上的配置是否合理；在已确定属性元素集合的情况下如何使重码字组最少；在确定重码指标的前提下如何合理选择属性元素的集合；对有关评测指标如何给出确切的论证等等，都是非常繁杂的。人工处理这些问题费时极多，甚至是人力所不能及的。因此，计算机辅助设计方法被引入了这一领域。

所谓汉字输入编码方法的计算机辅助设计，是以计算机为工具，把汉字的属性元素（字音属性或字形属性等）以一定的格式送入计算机，通过分析选择或者调整比较，寻求一种优化的结果。这是在编码方法已基本确定的情况下，一种优化设计的方法。只有在充分了解汉字的特点，把握汉字内在的规律，合理选择属性特征的条件下，利用机助设计方法才能使编码方法优化，充分体现出一个编码方法的特点。当然，不是所有编码方法都有可能或必要采用机助设计手段的，对于那种不存在调整或选择可能性的编码方法，就不需要进行机助设计；对于只能在很小范围调整的编码方法，用人工方法也是可以奏效的。机助设计的效益随编码方法的可调整范围扩大而明显地增加，编码方法的可调整性事实上是由一个代码有序组和多个属性元素之间的对应关系而形成的。因此，机助设计的目的可归结为：如何确定有序组和属性元素间的映照关系，而最终得到一个优化的汉字代码集。

机助设计的流程如图 3-1 所示。根据设计指标的要求，首先应收集汉字及汉字的有关各种属性资料，然后初步制订出编码方案，并将它们以某种数学模式来描述，在选择恰当的算法后，通过程序，计算机可按初步确定的编码规则编制出相应的汉字代码表，并及时地计算出有关各项性能指标与设计要求相比较，当不满足设计要求时，由人工干预或由机器自动地在可调整范围内，再次重复上述过程，直至满足设计要求后，给出最终的汉字代码表。

目前汉字输入编码方法的计算机辅助设计可分为两种类型。（1）以综合调整为主要手段的方法。这种方法是以有序组的容体和属性元素的某种映照关系，形成最初的汉字

代码集。然后以此代码集为基础，模拟人工调整的过程，进行综合分析求得优化途径。它是通过以反复进行局部分析为基础来实现综合调整，也是一个人机交互作用的过程：计算机综合结果而人进行分析处理。由于计算机的高速和逻辑判断能力，使得用人工完成需要数月或数年的工作量，用计算机在短时间内就可完成。（2）以分析为主要手段的方法。它的特点是全面分析特定编码方法中全部汉字的各种代码组合的可能性，以供设计人员选择。其中如果存在着满足设计要求的某种代码集合，计算机可以直接给出答案。当然，全面地分析全部汉字的各种代码组合，其计算量是十分惊人的，只有经过科学地约束和简化处理后，才有可能进行。即使这样，仍然对算法有着较高的要求。

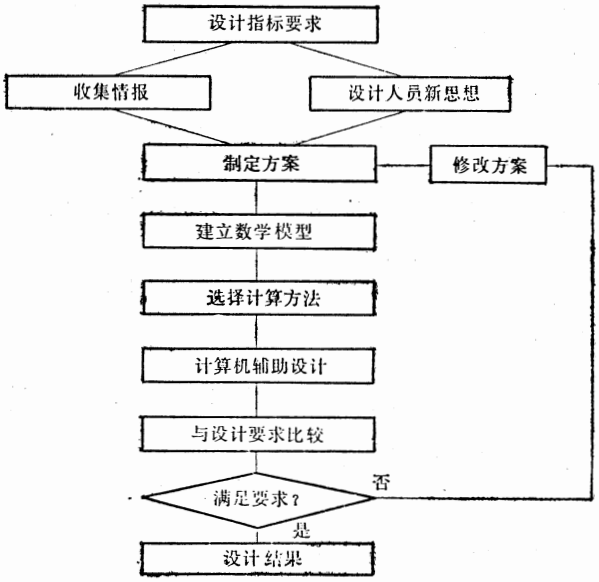


图3-1　机助设计流程

机助设计的主要困难在于：全面地、客观地、科学地总结出汉字的各类属性特征及其规律，并把它们以相应的数学模式来描述，建立可靠而有效的算法。由于这一工作的复杂性，是至今的绝大多数编码方法未能引用机助设计手段的主要原因。目前，机助设计还只在少数编码方法中，通过某些约束和简化处理，取得了较为优化的结果。详细地讨论机助设计方法，将涉及到更多的集合论和算法论方面的内容，有兴趣的读者可参阅有关文献。

3.6　汉字输入编码方法的评测

一般地说，在汉字集和汉字输入代码集之间建立映照关系的可能性是非常多的，这就是说有可能出现各种各样的汉字输入编码方法。短短数年内出现了数百种编码方法便是例证。随着研究的不断深入，新的编码方法还将出现。不过，汉字输入代码是一种人机界面上的代码，是通过操作人员的击键操作来输入机器的，它和人、机器的关系非常密切，并非任何编码方法都会被接受，在经过一段大发展之后，输入代码的数量上的增加将会明显减慢，并且随着使用环境的不同，将有一定数量的输入代码在相当一段时间内同时被采用。

随着汉字信息处理系统的普及推广应用，如何对为数众多的汉字输入编码方法从理论上和实践上作出评价和测试，已成为一个普遍关心的问题。由于评测方法本身，目前还处于研究阶段，标准的评测方法正待逐步确立，以下就一些基本内容作一介绍。

目前对汉字输入编码方法进行评测的指标中，大致可分为两类，一类是根据某些统计或计算可以得到的定量指标；另一类是属于尚不便统计或计算的定性指标。同时，对某些影响因素较多的综合性指标，有时还有必要采取抽样实测的评估方法。

属于定量评测指标的内容一般有：

1．汉字集H的容量　指的是可给出代码的汉字的数量，一般应满足GB2312基本集的需要，还应考虑将来辅助集中的汉字。

2．码元数量　即代码中所用汉字属性元素的符号数，其中用于区分出一个个汉字的空格也应包括在内。

3．码元熵值　暂不考虑码元的相关性，只考虑码元的概率分布，通过下列公式计算所得的平均每个码元的信息量。

$$H(k)=-\sum_{i=1}^{k} p_{mi}\log_2 p_{mi}$$

式中$H(k)$是码元熵值，单位为位，k为码元数量，p_{mi}为第i个码元的使用频度。

4．汉字熵值　完全确定一个汉字平均所用的最低信息量，它由下列公式计算出

$$H(\text{汉字})=-\sum_{i=1}^{n} p_i\log_2 p_i$$

式中，H（汉字）是汉字熵值，单位为位；n为汉字数；p_i为第i个汉字的使用频度。

5．平均码长　平均每个汉字所用的码元位数，使用空格区分汉字时，空格应包括在码长之内，一般用$\overline{L}$来表示。

6．汉字编码效率　指的是理论码长的下限（即汉字熵值）与实际平均码长之比，若用η表示，定义式为：

$$\eta=\frac{H(\text{汉字})}{\overline{L}\cdot\log_2 k}$$

7．键入速率　指的是单位时间内击键输入汉字的字数，通常用的单位为字数/分钟，若以符号S_j表示，则可沿用下式计算

$$S_j=\frac{60}{[a+bH(k)]\cdot\overline{L}}\ \text{字/分钟}$$

式中，当$H(k)$尚未求得时，可近似取$\log_2 k$代替$H(k)$，a、b两系数根据试验的具体条件决定，通常因人而异。

8．重码数　若以C表示，其定义为

重码数C＝重码字数－重码组数。

注意，这是代码表完成后的实际统计结果，与3.3.6中对重码进行预测时所作的讨论不完全。

9．重码率　若以P_c表示，定义式为

$$P_c=\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij}$$

式中P_{ij}为各重码字组中第i组第j字的使用频度，j按频度高低排序。

10．非常规代码数　所谓常规代码就是本章所讨论的基本代码，有时为了区别过多的重码字或者按照基本规则无法给出某些汉字的代码时，不得不附加某些规则或特殊定义才可，对于这样的汉字代码就称作非常规代码。

11．非常规代码出现率　用下式定义

$$P_f = \sum_{i=1}^{F} P_i$$

式中 P_f 即非常规代码出现率，F 为非常规代码数，P_i 为第 i 个非常规代码的使用频度。

12. 多码数　这是指一个汉字与多个代码相对应的数量，通常简码也按多码计算，若以 D 表示，则定义式为

$$D = \text{总代码数} - \text{总字数} + \text{重码数}。$$

13. 错码率　错码次数占全部处理字数的百分比，一般按特定条件由实测结果判断。

14. 学习曲线　这是特定条件下，在学习过程中根据实测结果绘制的键入速率-时间曲线和错码率-时间曲线。

15. 编码学习期　操作员从学习编码开始到不用码本进行编码的错码率稳定下降至1%所需要的时间，常用小时作单位。

16. 编码熟练期　随着键盘类型的不同，操作员键入速度达到 30～50字/分，同时错码率不高于 1%所需要的时间，以小时计。

17. 外字数量　指汉字代码集以外的汉字。属于定性评估方面的内容一般有：

（1）编码方法的论证是否合理和充分。

（2）在编码规则方面有：编码规则的数量多少；代码和属性元素对应规则的逻辑性和规律性是否简明扼要，是否前后一贯；编码规则对用户要求的高低等。

（3）在记忆量大小方面有：编码规则是否容易记忆；非常规编码的字数和规律性；附加规则或规定是否繁琐等。

（4）在检索和兼容方面有：对于有疑难的集内字是否有检索的手段及其难易程度；编码的兼容性及用户进行选择的自由度；对集外字有无处理能力及处理方法的难易等。

（5）在与机器的联系方面有：编码方法的设计是否考虑了与计算机的特点相结合；可使用的计算机类型，是否需要增加其他设备或采取其他措施；最大码长 L_{max}；软件程序所占存储器容量的大小；是否需要人机相互作业及人机相互依赖的程度等。

上述各项内容，不论在定量或定性方面，目前大抵是独立地进行各单项指标的评测，有些计算内容，如汉字熵值、汉字键入速率等，还有待进一步充实和论证。随着评测方法研究的不断深入，评测内容还将不断更新和完善，逐步建立起标准的评测方法来，用户可根据评测结果，对编码方法作出适合本身需要的选择。

汉字输入编码方法的评测工作是一项相当浩繁的工作，为对某种编码方法做出全面的、较为可靠的评测，往往需要耗费大量人力。因此，很自然地应将尽可能多的评测内容通过计算机来完成。下面是按部分评测内容，对角形代码类型中的所谓“三角号码法”进行计算和统计而事先编制的处理程序的一部分：

```
The tested results of 3CC code system
    K = 11
    H(K) = 3.251683036455
    Average length of code = 7
    Coding efficiency = 0.3989C^1603189
    Number of Chinese Character = 10558
```

```
    Number of duplicate code = 328
    Probability of duplicate code = 6.34442486E-07
    Number of multi-code = 0
    Rate of Keyin = 32.64259671395
Elements of code
'0'  '1'  ' '  '2'  '3'  '4'  '5'  '6'  '7'  '8'  '9'
The list of duplicate code
```

在上述例子中，涉及到的评测结果有：

码元数量 $k=11$；

码元熵值 $H(k)=3.251,683,036,455$；

平均码长 $\overline{L}=7$；

编码效率 $\eta=0.3,989,094,603,189$；

汉字集中的字数 $N=10,558$；

重码数 $C=328$；

重码出现率 $p_c=6.34,442,486\times10^{-7}$；

多码数 $D=0$；

键入速率 $S_j=32.64,259,671,395$；

给出的具体码元为 0，1，…，9，空格，共 11 个；

最后列出的是全部重码字组。

随着对汉字输入编码方法的评测工作不断开展，更多的评测项目或对多种编码方法进行综合分析等，都可在通过适当准备后，用计算机予以完成。

第四章　信息处理交换用的汉字代码

4.1　概　　述

用电子计算机处理汉字时，必须先将汉字代码化，即对汉字进行编码（encoding）。我国最早可用于计算机处理汉字的代码是在十九世纪末编制的电报四码——用四个阿拉伯数字表示一个汉字，从0000～9999可编出一万个汉字代码。五十年代，有人采用该代码进行过俄汉机器翻译。但是，当时的电子计算机技术正处于前期发展阶段，它主要用于数值计算方面，只有少数科研单位进行汉字信息处理的试验性工作，所以也只有少数人对汉字编码进行局部的研究与应用，汉字的输入输出问题还不突出。但是，随着电子计算机技术的飞速发展，特别是七十年代，集成电路技术的发展与进步，普遍使用计算机处理汉字信息才成为可能，随之而来的汉字输入输出问题也就逐渐突出，特别是如何解决汉字的输入编码，引起了许多专业和业余爱好者的兴趣和广泛研究。为了交流经验，1978年在青岛召开了第一次全国汉字编码学术交流会，并成立了中国汉字编码研究会。1980年在杭州召开了第二次会议。1981年6月在天津成立了中国中文信息研究会。上述的中国汉字编码研究会成为该会下设的汉字编码专业委员会，专门组织和协调汉字输入编码的研究。到1983年四月为止，登记在案的汉字输入码方案已达四百多个，并还有继续增长的趋势。目前汉字编码专业委员会正在拟制汉字输入码的评测规则，以便从中优化出若干种方案在全国推广使用。

随着计算机系统和网络技术的发展，我国在七十年代中后期，先后开展了对汉字信息交换码和控制码的研究工作，并于1981年颁布了国家标准GB2312——《信息交换用汉字编码字符集——基本集》。目前正在制订后续的辅助集和与汉字交换码配套使用的控制功能码标准。

汉字输入码的广泛研究和汉字标准交换码的制订，促进了对汉字信息处理系统内部码的研究。这是一个如何充分利用计算机现有资源处理汉字信息的问题。它说明我国汉字信息处理技术的发展，已从汉字输入编码的研究和应用，深入到汉字信息处理系统的研究和应用。汉字编码技术是汉字信息处理技术的一个重要组成部分。

4.2　汉字代码种类

从汉字代码的角度看，一个汉字信息处理系统，就是一个进行各种汉字代码的转换系统。这些汉字代码，除上述提到的汉字输入码、汉字交换码、控制功能码、汉字内部码外，还有汉字地址码、汉字字形码以及汉字扩充码等。这些代码的名称叫法可能不一，但是它们所表示的含义和具有的职能却是明确的。各种代码之间的关系，以及它们在系统中的位置，可以用图4-1作简单表示。

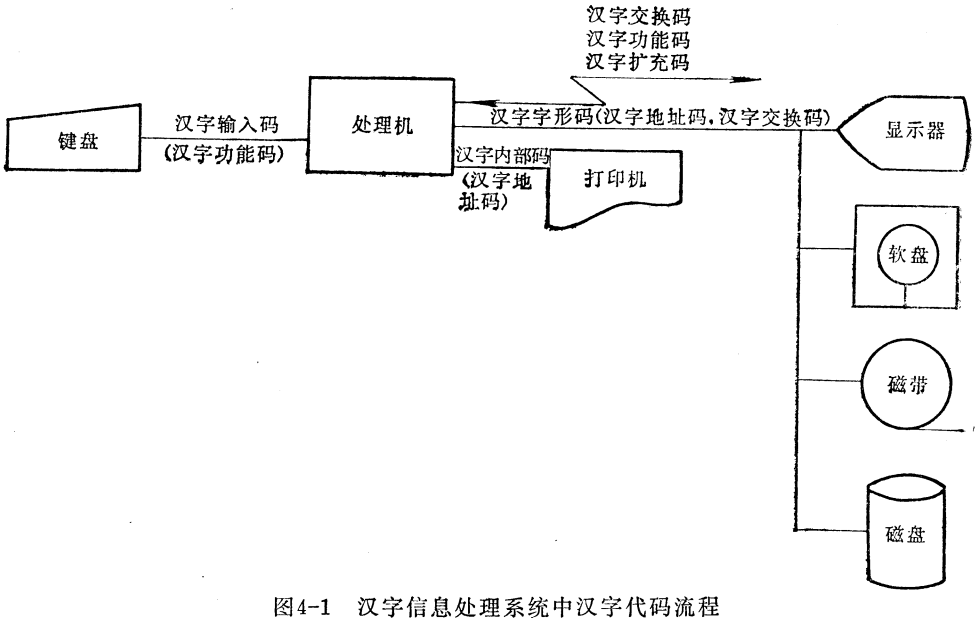


图4-1　汉字信息处理系统中汉字代码流程

当然，由于各个汉字系统的结构和功能不同，因此，各个汉字系统不都同时具有上述所有的汉字代码，某些汉字代码在系统中的位置也不是固定不变的。下面逐一介绍各种汉字代码。

4.2.1　汉字输入码

汉字输入码是为了将汉字输入计算机而编制的代码，它位于人机界面之间。目前汉字主要由人工通过汉字输入设备输入，所以，为便于人们熟悉和容易掌握，常常要求输入码规则简单、易于记忆、操作方便、编码容量大、码位短、输入速度高和重码率低等性能。为了达到这些要求，人们根据汉字的各种属性提出了数百种汉字输入编码方案。究竟哪种编码方案最好，这有待实践检验。一般来说，由于用户不同，用途有别，故对采用什么输入编码方法和输入方式不能强求统一。不过应该制订一个评测汉字输入编码的规则。对各种输入编码方案进行测试，并公布各项测试结果，以便提供给用户和研制单位选择使用和研制汉字输入设备。如果能通过对各种方案的评测，优选出几种输入编码方案，这对汉字信息处理系统的推广应用无疑是有积极意义的。

到现在为止，几乎所有的汉字输入编码都是由人工进行的，而汉字输入编码又是一项极其繁琐的工作，这种繁琐性必然地影响编码质量。如何利用计算机进行辅助编码，将繁琐的输入码编制工作从沉重的手工劳动中解脱出来，这是今后必须重视的工作。

4.2.2　汉字内部码

通常绝大部分的汉字输入码都要超过两个字节，有的并不是等长码。因此，为了节省内、外存储空间和处理方便，将多字节的输入码转换成信息量较少而且便于处理的系统内部码（internal code），这是必须经过的步骤。

汉字内部码的设计往往和具体的系统及使用要求密切相关，它没有统一的格式。所

以，目前汉字内部码的形式有多种。有的采用输入码的序号；有的采用交换码；有的采用交换码中的区、位号；有的将交换码中的两个七位字节改为八位字节，分别在每个字节的高位或仅在第一字节高位加一个汉字标识码；有的采用三个字节等等。但是从目前来看，汉字内部码究竟采用哪种形式，一般需要考虑如下几点。

（1）码位尽可能短，而所表示的汉字数要尽量多。

（2）便于操作运算，码值有序且连续。

（3）与标准交换码兼容，即与交换码有尽可能简单明确的对应关系。

（4）便于纳入各种高级语言的字符类型。

下面介绍目前国内三种汉字内部码的形式。

1．国标 GB2312 汉字交换码作为汉字内部码　国内有些COBOL、FORTRAN汉字支援系统，采用国标汉字交换码或它的区、位号作为汉字内部码。但在汉字（地址码）和GB1988 的字符汉字混合字符串中，分别加有标识汉字（地址码）开始和 GB1988 字符开始的标识符。例如，用“&&”表示汉字（地址码）开始、用“¥¥”表示汉字（地址码）结束和 GB1988 字符开始。例如 1053 计算机，要写成¥¥1053 && 计算机。

2．用三个字母字符表示一个汉字内部码　本方法的汉字输入输出转换可用图 4-2 表示：

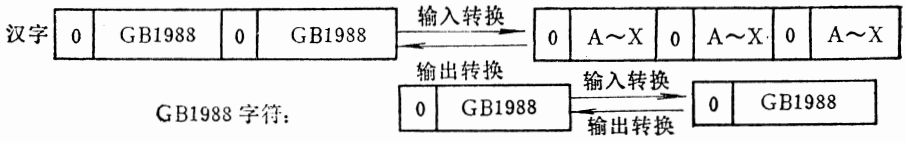


图4-2　汉字输入输出转换

3．加有字标识位的汉字内部码　本方法是用两个八位字节表示一个汉字内部码，但是第一个字节高位恒为 1，记作 B11，第二个字节高位可以为 1，也可以为 0，记作 B21 或 B20。B11、B21 为汉字基本集的内部码，取值与标准交换码一一对应，去掉高位即为交换码。B11，B20 为汉字扩充集的内部码，留待汉字集进一步扩充用。

上述第 1 种内部码直接采用汉字交换码，汉字的存储和输出都是两个字节，格式一致。但是对高级语言的语法影响较大，需要设置区别汉字和 GB1988 字符的特殊字符或控制码。

第 2 种内部码在对高级语言语法的影响上和在汉字输出转换时，对汉字内部码和 GB1988 字符的识别方面要优于第 1 种内部码形式，并且给实现变量名，文件名的汉字化带来方便，但是本方案的汉字文件存储空间增大 50%。在汉字检索识别上，仍然需解决汉字内部码同英文关键字及用户定义的英文符号之间的混淆问题。而且，在汉字的输入输出的码制转换上开销也较大些。

第 3 种内部码可以将数据纳入高级语言的字符类型，不需要特殊字符或控制代码作为汉字标识符去区别是汉字字符还是 GB1988 字符。此外，只要视两个字节为一个汉字，就可直接进行字符串的运算操作。并且由于存储与输出都占用两个字节位置，故保持了存储与输出格式的一致性。但是，这种内部码由于占用了八位中的一位，所以损失了48

K的容量，即（$2^{16}-2^{14}$）个两个字节的内部码。

4.2.3 汉字地址码

汉字地址码（address code）是指汉字字模库(这里主要指整字形的点阵式字模库)中存储汉字字形信息的逻辑地址码。目前作为汉字字模库的存储媒体主要有两大类。一类是半导体存储器，如只读存储器 ROM，可编程只读存储器 PROM，随机存取存储器 RAM 等。另一类是磁性存储器，如软磁盘、硬磁盘、磁带、磁鼓等。构成汉字字模库的存储媒体不同，汉字地址码的物理表示法也不一样。例如半导体存储器汉字字模库的汉字地址码通常由插件地址、存储模块地址、存储片地址、数据地址等组成。软磁盘字模库的汉字地址码通常由盘片地址、区段地址、磁道地址、数据地址等组成。磁带存储器字模库的汉字地址码则由带地址（磁带机号），记录块地址、数据字段地址等组成。不论哪种汉字字模库、汉字字形信息都是按一定顺序（大多数按标准汉字交换码中汉字的排列顺序）连续存放在存储媒体上。所以汉字地址码也大多是连续有序的，而且与汉字内部码间有着一定的对应关系。为了简化汉字内部码到汉字地址码的转换，这种对应关系通常应尽量简单。因此，有些汉字地址码的形式与汉字内部码是完全一致的。

4.2.4 汉字交换码

汉字交换码是一种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代码。汉字交换码位于一台机器的出口和另一台机器（包括输出设备与记录设备）的入口之间。为了要达到系统设备之间或记录媒体之间信息交换或互换的目的，汉字交换码必须采取统一的形式。为此汉字交换码的编制应考虑以下有关问题。

（一）同现行计算机系统所采用的标准信息处理交换代码（指字母、数字和符号)的兼容性

从“信息”本身看，“汉字信息”与“非汉字信息”没有本质上的差异。原则上，凡是能处理非汉字信息的计算机系统都能处理汉字信息，所以无论从过去、现在还是从将来的发展趋势看，几乎所有的汉字信息处理系统都是在原有非汉字信息处理系统的基础上扩充了一定的软、硬件功能形成的。为了最大限度地利用原有计算机系统的资源为处理汉字信息服务，显然汉字交换码必须与现有计算机系统所采用的标准信息处理交换码相兼容。这种兼容性带来的另外一个好处是，汉字信息处理技术本身的任何发展成果也可以方便地为大家所享用。

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 GB1988《信息处理交换用的七位编码字符集》；还制定了相应的代码扩充标准，即 GB2311《信息处理交换用七位编码字符集的扩充方法》。因此汉字交换码应与 GB1988 兼容，并根据 GB2311 所规定的扩充方法进行编制。由于汉字数量远远大于七位编码所能表示的容量，所以根据 GB 2311，一个汉字必须用两个或两个以上的七位编码表示。

（二）汉字选择的实用性、通用性及编码效率

汉字的特点是字量大，字形复杂，同音字多，异体字多。据估计，目前汉字约有六万，但是每个字的使用频度相差很大，有个别字如“的”字的使用频度很高，覆盖率达 4 %左右，但有些字使用频度却很低，在万分之一以下。从总的情况来看，绝大部分字

都是现代汉语中很少用到的“古字”。所以对任何一个汉字信息处理系统，首先就面临着一个选多少字，选哪些字的问题。选多了造成系统软、硬件负担过重，浪费资源，降低了系统的存储、处理、传输及使用效率；选少了又不能满足应用的要求。因此过多过少，都达不到实用的目的。

此外，汉字的使用是相当复杂的，各行各业使用的汉字又不完全相同，我们使用的字和我国台湾省、香港地区使用的字差别更大，再加上一些少数民族用的汉字，更增加了汉字使用的复杂性。

如何使所选用的汉字具有广泛的通用性，这也是编制汉字交换码所必须考虑的问题。根据上述要求，我国首先对六千多个常用汉字制订了交换码的国家标准，即GB2312《信息交换用汉字编码字符集——基本集》，其中每个汉字用对应于GB1988的两个七位码来表示。

目前正在对另外不常用的一万六千余字编制第一个和第二个辅助集，每个汉字也是用两个七位编码表示，但要用不同的控制字符去调用。此外，对于六万个汉字是用两个字节表示还是三个字节表示，这涉及到编码的效率。若用两个字节表示，那么，根据GB2311，一个双字节集合只能编码8836个汉字。若用三个字节表示，则容量足够，但是由于增加了汉字信息处理时对存储和传输的时间、空间开销，降低了整个系统的效率。如何利用汉字使用的特点，提高编码效率，但又有足够的编码容量，这才是编制汉字交换码时另一个不可忽视的问题。

4.2.5 汉字控制功能码

以上我们讨论了汉字输入码，汉字内部码，汉字地址码和汉字交换码。这些汉字代码实际上都是与一个具体的汉字或其他图形字符相联系的，即它们都表示一个汉字数据，我们称之为“图形字符码”或“汉字数据码”。在一个汉字信息处理系统中，除了表示汉字数据的“图形字符码”以外，还有一种并不表示一个具体的汉字数据，但是却影响汉字数据的格式处理，传送控制或解释执行等的代码，这就是汉字控制功能码。(control function code)。它和汉字数据代码组成汉字数据流，它贯穿于汉字输入、机内处理、汉字输出等整个汉字系统的处理流程，汉字控制功能码的设计和选择将直接影响整个汉字系统的效率和性能。

汉字控制功能码的设计包括两个方面。一是控制功能的编码表示；二是控制功能含义的确定。控制功能编码表示的确定要考虑以下两点：

（1）应与原计算机系统控制码的编码表示兼容。特别对于通用型的汉字信息处理系统更应如此。否则将会增加处理的复杂性，降低系统的效率，并丧失系统的通用性。

（2）编码要简单明确，效率要高。即应以尽可能短的码位表示尽可能多的控制功能。

以上两点在实际应用中往往是互相制约的。对通用型的汉字信息处理系统尤为突出。这是因为作为一个通用型的汉字信息处理系统，其汉字控制功能码的表示必须按照有关的国家标准或国际标准进行设计。自行选择的余地小。

功能码的控制功能的确定取决于系统本身的用途。对于不同的汉字信息处理系统可选择不同的功能码。例如对于通用型汉字信息处理系统，应该具备汉字信息处理和交换

的基本控制功能。对于专用的汉字信息处理系统（例如汉字编辑、排版系统），可以在基本功能码的基础上再扩充一些专用的控制功能码，以满足专用领域中的汉字信息处理的需要。

目前我国正在参照有关国际标准和国家标准，制订能处理汉字的控制功能码的国家标准，即《文字和符号图形设备的增补控制功能》。其中，包括控制串定义符、引导符、格式控制符、文稿组版设备用的增补格式控制符，修改可见数据用的编辑功能、移动操作位置用的编辑功能，区的限定，模式设置及其他控制功能等共有一百多个功能码。控制功能码的编码表示从两个字节到四个字节不等。

4.2.6 汉字扩充码

这里讨论的汉字扩充码是指按照国家标准 GB2311 进行扩充的编码，因此上述的汉字交换码和汉字控制功能码，事实上都是一种扩充码。

作为一个非汉字的西文计算机系统，处理对象主要是字母、数字及有限的几个图形符号。通常只要一个七位或八位编码字符集就足够了。但是对于一个汉字信息处理系统，作为处理对象的汉字，其数量远远超过一个七位或八位编码字符集的容量，控制功能码虽然没有那么多，但是也已过百，所以必须寻求一种扩充代码的办法，只有将代码加以扩充，才能表示数以万计的汉字和大量控制功能。

扩充代码的方法很多，扩充码的形式也多种多样，例如可以用增加位数的办法扩充编码容量，对六万汉字，用 2^{16} 就可表示，也可用三个五单位字节表示，甚至也可以用连续的十进制数字表示。但是根据以上所述，对于一个通用的汉字信息处理系统，无论是汉字交换码，还是汉字控制功能码，其编码表示必须与现有计算机系统所使用的信息代码，特别是应与标准码兼容，同时应该尽可能提高编码的效率。

由于我国已经制订了国家标准 GB1988 及 GB2311，所以汉字代码的扩充方法实际上只能按照 GB2311 进行。GB2312 就是一个根据 GB2311 扩充而来的汉字交换码标准。

4.2.7 汉字字形码

目前汉字信息处理系统中产生汉字字形的方式，大多是数字式的，即以点阵的方式形成汉字，所以这里讨论汉字字形码，也就是指确定一个汉字字形点阵的代码。

汉字字形通常分为通用型和精密型两类，通用型汉字字形点阵分成三种：

简易型　15×16 点阵

普通型　24×24 点阵

提高型　32×32 点阵（日本称为轻印刷）

精密型汉字字形用于常规的印刷排版，由于信息量较大（字形点阵一般在 96×96 点阵以上），通常都采用信息压缩存储技术。

不论是通用型还是精密型的汉字字模，字形信息一般均以字节的形式存储在半导体或磁性存储媒体上。通用型字模由于信息量少，通常存储整字信息，这些整字信息就是汉字字形码。精密型字模通常只存储经压缩后的字形信息。

现以简易型 15×16 点汉字点阵为例，“寸”字的字形点阵如图 4-3 所示。

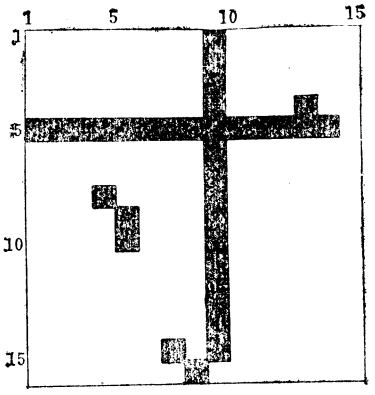


图4-3 “寸”字的字形点阵图

假设存取字形的单位为字节（通常为8位）存取次序为由第1行→第16行，同一行左边8位为第一字节，右边7位为第二字节。对于任何15×16点的汉字字形码，显然都有16（行）×(16位/8位)≈32字节，所以寸字的字形码可表示如下：

① 00 00 00 00 $(00)_H$（H表示括号内为16进制数）
② 10 00 00 0 $(80)_H$
③ 00 00 00 00 $(00)_H$
④ 10 00 00 0 $(80)_H$
⑤ 00 00 00 00 $(00)_H$
⑥ 10 00 00 0 $(80)_H$
⑦ 00 00 00 00 $(86)_H$
⋮ ⋮ ⋮
㉙ 00 00 00 10 $(02)_H$
㉚ 10 00 00 0 $(4.0)_H$
㉛ 00 00 00 01 $(01)_H$
㉜ 00 00 00 0 $(0.0)_H$

上述字形存取的顺序也可改为由第一列→第16列，同一列上面八位为第一字节，下面八位为第二字节，则“寸”字的字形码与上述完全不同，读者可以自己写出。

即使对于相同点阵结构的同样一个字，并取相同的存取顺序，由于不同的字形设计，其字形码也不同，特别对于那些因受点阵结构限制需要简化的字更是各不一样，这样会造成字形信息的混乱，各单位又都自造字模点阵，重复工作造成浪费，并且不利于工业化大批量生产。另外汉字字形码的设计不仅是一个计算机技术问题，同时又涉及文字学和书法艺术，所以这是一个技术与艺术相结合的问题，有相当的难度，不是任何人都能设计好的。所以有必要统一汉字字形码，目前我国正在制定通用型汉字字形码的国家标准。有关情况将在下节介绍。

4.3 汉字代码的标准化

上节我们讨论了汉字信息处理中各种汉字代码的基本概念、设计思想及应用。由于

这些汉字代码直接影响到汉字信息处理系统硬件与软件的研制，所以应该尽量使之标准化。一旦实现标准化，就必须按照标准执行。只有这样，才能有助于迅速发展我国的汉字信息处理技术。

下面分别介绍用于我国汉字信息处理技术领域的已经制定或正在制订的有关汉字代码标准。

4.3.1 GB1988《信息处理交换用的七位编码字符集》

GB1988《信息处理交换用的七位编码字符集》不是汉字代码标准，而是非汉字代码标准，但是由于许多汉字代码标准是在该标准基础上扩充而来的，而且绝大部分汉字信息处理系统都是既处理汉字信息，又用这一标准处理西文或数字信息，所以有必要对该标准作一介绍。

“GB1988”是国家标准代号。其中“GB”是“国标（GOOBIAO）”汉语拼音的首字母，“1988”为标准序号，该标准简称为“GB1988”。

GB1988是我国计算机专业基础标准，它是根据国际标准化组织（International Standards Organization，简称ISO）的标准ISO646《信息处理交换用的七位编码字符集》制定的。

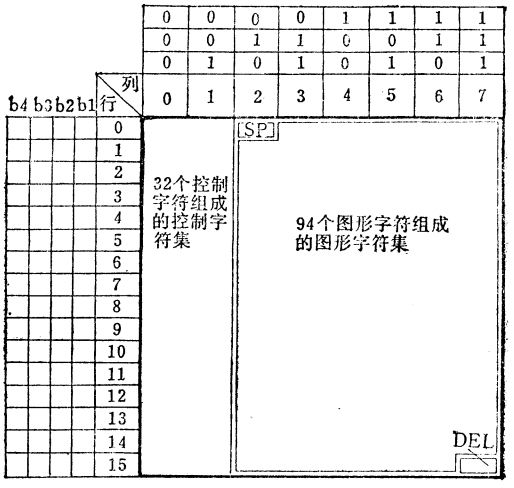


图4-4 GB1988代码的构成

一、GB1988代码构成

GB1988七位编码字符集规定了信息处理交换用的128个字符，每个字符都是用$b_7b_6b_5b_4b_3b_2b_1$标识，b_7为最高位，b_1为最低位。也可以用它在代码表中的位置（列号/行号）表示，列号与$b_7b_6b_5$三个二进制位对应从0到7列，共有8列。行号与$b_4b_3b_2b_1$四个二进制位对应，从0行到15行，共16行。这128个字符按其在代码表（见图4-4）中的位置和功能分成下列两个部分。

（一）控制字符集

0列和1列的32个字符为控制字符，它们组成一个控制字符集（属于C0集）。这32个控制字符按照它们的功能含义分成五类。

1. 传输类控制字符　用于各种数据终端设备或系统之间的数据传输控制。这类字符共有以下10个：

- SOH（0/1）标题开始字符
- STX（0/2）正文开始字符
- ETX（0/3）正文结束字符
- EOT（0/4）传输结束字符
- ENQ（0/5）询问字符
- ACK（0/6）承认字符

• DLE(1/0) 数据链转义字符
• NAK(1/5) 否认字符
• SYN(1/6) 同步空转字符
• ETB(1/7) 组传输结束字符

2. 格式类控制字符　用于控制所要打印、显示或记录数据的位置。这类字符共有以下6个：

• BS(0/8) 退格字符
• HT(0/9) 横向制表字符
• LF(0/10) 换行字符
• VT(0/11) 纵向制表字符
• FF(0/12) 换页字符
• CR(0/13) 回车字符

3. 设备类控制字符　用于控制同数据处理系统或数据通信系统相联系的辅助设备，而不能用于控制电信系统。控制电信系统的是数据传输类控制字符。设备类控制字符有以下4个：

• DC_1(1/1) 设备控制字符1
• DC_2(1/2) 设备控制字符2
• DC_3(1/3) 设备控制字符3
• DC_4(1/4) 设备控制字符4

4. 信息分隔类控制字符　它们都是在数据的层次排列中，用于从逻辑上分隔或限定数据。这类字符有以下4个：

• US(1/15) 单元分隔字符
• RS(1/14) 记录分隔字符
• GS(1/13) 群分隔字符
• FS(1/12) 文卷分隔字符

5. 其它控制字符　包括以下8个：

• NUL(0/0) 空白字符
• BEL(0/7) 告警字符
• SO(0/14) 移出字符
• SI(0/15) 移入字符
• CAN(1/8) 作废字符
• EM(1/9) 媒体结束字符
• SUB(1/10) 取代字符
• ESC(1/11) 转义字符

此外，在图形字符集的首尾还有两个字符也可归入控制字符范围内：

(1) 间隔字符SP(2/0)。它用来使打印或显示位置在同一行内前进一个字符位置，通常在一串字符中突出一个位置来表示一个符号。间隔字符不是一个控制字符。但是它具有格式控制字符和信息分隔控制字符的功能。

(2) 抹消字符DEL(7/15)。它用于清除错误或不要的字符，在穿孔纸带上该字符

由每个穿孔位置上都有孔的代码组成。它也可用于实现媒体填空或时间占空。从字符串中去掉或插入该字符，均不影响字符串的含意。但是它对格式或设备可以起控制作用。

（二）图形字符集2列至7列（不包括位置2/0和7/15的字符）的94个字符为图形字符。它们组成一个图形字符集（属于G0集）。

上述34个控制字符的定义及94个图形字符的名称及在代码表中位置见GB1988文本。

二、GB1988标准代码表

GB1988有两张标准代码表。即国内通用代码表和国际通用代码表（见附录表），两者主要差异是位置（2/4）的货币符号，前者用货币元符号“¥”，后者用国际通用货币符号“¤”。

三、GB1988的应用

GB1988的所有控制字符都可用于汉字信息处理系统有关软件和硬件的设计。例如SOH等10个传输类控制字符就完全可用于汉字数据通信的软件设计；格式类控制字符可用于汉字印刷机、汉字显示器等终端设备的设计制造；信息分隔类控制字符可用于汉字数据文件的书写；SO，SI移位字符及ESC转义字符可用来表示各种汉字扩充码。同样GB1988的图形字符也完全适用于汉字信息处理。由下节可见，每个汉字交换码就是用两个GB1988中的图形字符表示的。我国于1981年将GB1988与GB2312向国际标准化组织（ISO）提出登记申请，已获批准，并给定了这两个标准的转义序列。ISO正式通知我国和其他各成员国，从1982年6月30日起开始使用。在GB1988中，国内通用的基本代码表的转义序列为：

G0集　ESC　2/8　5/4
G1集　ESC　2/9　5/4
G2集　ESC　2/10　5/4
G3集　ESC　2/11　5/4

GB2312的转义序列为：

G0集　ESC　2/4　4/1
G1集　ESC　2/4　2/9　4/1
G2集　ESC　2/4　2/10　4/1
G3集　ESC　2/4　2/11　4/1

这些转义序列分别用于指明我国的两个标准字符集，它由我国唯一占有并可在国际通行。

4.3.2　GB2311《信息处理交换用七位编码字符集的扩充方法》

上述的GB1988由于仅规定了128个字符及其编码表示，它只能满足西文系统信息处理的需要。但是作为汉字信息处理系统，通常需要处理成千上万个汉字，而且还需要一些特殊的控制功能，所以必须制定一套编码扩充方法。用这一方法扩充汉字交换码和控制功能码，使它既能满足各种系统的需要，又能与GB1988规定的七位编码字符集兼容，以便进行国内外的汉字信息交换。为此，我国在1980年根据国际标准ISO2022《七位与八位编码字符集的扩充方法》制定了国家标准GB2311《信息处理交换用七位

编码字符集的扩充方法》。GB2311 是以七位编码字符集为基础进行代码扩充的。

一、七位编码字符集的构成

七位编码字符集的构成与 GB1988 的相同，均由下述的控制字符与图形字符的有序集合区域构成。

（1） 0 列与 1 列为 32 个控制字符集的区域，属于 C0 集；

（2）位置 1/0 的“间隔”字符（SP），属于单个控制字符；

（3） 2 列～7 列为 94 个图形字符集的区域，属于 G0 集；

（4）位置 7/15 的“抹消”字符（DEL），属于单个控制字符。

二、代码扩充方法

（一）替换法

替换法就是在不破坏 GB1988 代码结构的基础上将 GB1988 中的某些字符用新的字符来代替，代替后的字符集构成新的编码字符集（控制字符集或图形字符集）。该字符集需用新的转义序列指明（和调用）。

（二）增加字符法

这种方法是在七位码结构提供的 128 个字符以外，增加以下若干字符（见图 4-5）。

（1）单个扩充的控制字符：CS。

（2）由 32 个控制字符组成的增补控制功能集：C0 集或 C1 集。

（3）由 94 个图形字符组成的多字节图形字符集：多字节 G0 集，G1 集，G2 集和G3 集。

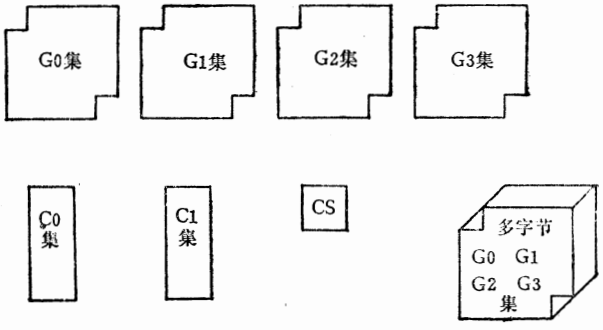


图4-5 增加字符法扩充代码示意

用于指明和调用上述扩充的控制字符集或图形字符集需要下面一些控制字符：

（1）转义字符 ESC。其位置在 GB1988：1/11。

（2）移出字符 SO。其位置在 GB1988：0/14。

（3）移入字符 SI。其位置在 GB1988：0/15。

（4）单移位字符 SS2、SS3。其编码由专门的标准规定。

（5）锁定移位字符 LS2、LS3。其编码也由专门的标准规定。

SO、SI、SS2、SS3、LS2 和 LS3，用于调用图形字符集，其中 SO 和 SI 分别调用 G1 集和 G0 集，调用后，系统将改变原来的状态；而 SS2 和 SS3 分别调用 G2 集和 G3 集中的一个字符，调用后仍然回到系统原有的状态。

（三）转义序列

对于一个使用多个扩充控制字符集或图形字符集的系统，为了便于信息交换，必须对每个扩充集用其相应的转义序列指明和调用。

转义序列的一般表示式为：

$$ESC \cdot I \cdot F$$

其中，ESC 是转义字符；I 是中间字符，它取 GB1988中 3 列内的码位；F 是终止字符，

它取 GB1988中 3 列～ 7 列码位（码位 7 /15除外）。

根据有无中间字符或中间字符的多少，转义序列一般有如下三类。

1．二字符转义序列 ESC · F　没有中间字符的二字符转义序列用于控制功能 的 扩充。根据终止字符 F 的不同码位，又分以下三种：

（1）F 取 6 列和 7 列（7 /15除外）码位。供扩充单个控制功能用，记作：ESC·Fs。

（2）F 取 4 列和 5 列码位。供扩充控制字符集 C1 集用，记作：ESC · Fe。

（3）F 取 3 列码位。供专用含义，由使用者自己定义。记作：ESC · FP。

2．三字符转义序列ESC · I · F　三字符转义序列的含义又由中间字符“I”决定，其具体意义如下：

（1）I 是 2 / 0 的三字符转义序列（ESC 2 / 0 F）为宣布序列，用以宣布后续图 形字符集的使用形式。

（2）I 是 2 / 1 的三字符转义序列（ESC 2 / 1 F）作控制字符集（G0 集）扩充用。

（3）I是2/2的三字符转义序列（ESC2/2F）作增补的控制功能集(C1集)扩充用。

（4）I 是 2 / 3 的三字符转义序列（ESC 2 / 3 F）作单个控制功能扩充用。

（5）I 是 2 / 4 的三字符转义序列（ESC 2 / 4 F）作指明多字节 G0 集用。

（6）I 是 2 / 5 的三字符转义序列（ESC 2 / 5 F）作完整代码扩充用。

（7）I 是 2 / 8 和 2 /12的三字符转义序列（ESC 2 / 8 F 和 ESC 2 /12F）作指明单字节 G0 集用。

（8）I 是 2 / 9 和 2 /13的三字符转义序列（ESC 2 / 9 F 和 ESC 2 /13F）作指明单字节 G1 集用。

（9）I 是 2 /10和 2 /14的三字符转义序列（ESC 2 /10F 和 ESC 2 /14F）作指明单字节 G2 用。

（10）I 是 2 /11和 2 /15的三字符转义序列（ESC 2 /11F 和 ESC 2 /15F）作指明单字节 G3 集用。

（11）I 是 2 / 6 和 2 / 7 的三字符转义序列目前尚未规定确切的含义，留作以后实施标准化使用。

3．四字符转义序列 ESC·I·I·F　四字符转义序列有两个中间字符“I”，现已制定确切的含义的有下面四种。

（1）指明多字节 G0 集用的 ESC 2 / 4，2 /12F；

（2）指明多字节 G1 集用的 ESC 2 / 4，2 / 9 F 和 ESC 2 / 4，2 /13F；

（3）指明多字节 G2 集用的 ESC 2 / 4，2 /10F 和 ESC 2 / 4，2 /14F；

（4）指明多字节 G3 集用的 ESC 2 / 4，2 /11F 和 ESC 2 / 4，2 /15F。

由上可见，由于具有标准含意的终止字符 F 共有63个，即七位代码结构 中 4 列～7 列（不包括位置 7 / 5）。所以，可供扩充的控制功能和图形字符是非常多的。例如若用二个七位编码，则每个图形字符集能对 $(2^7-32)(2^7-32)=8836$ 个的汉字进行编码。如果要对六万汉字进行编码，则只需要七、八个图形字符集。而上述的多字节图形字符集可以用 G0 集、G1 集、G2 集和 G3 集。每种图形字符集按其终止字符不同，各有126个（63× 2）具有标准含意的图形字符集，所以对六万汉字编码是毫无困难的。即使 再 多也没有问题。再如对控制功能码的扩充，由上述可知，转义序列 ESC 2 / 2 F 可用于 增

补的控制功能集（C1集）的扩充。显然，可以扩充63个C1集，每个C1集有32个控制功能，通过二字符转义序列ESC · Fe调用。

现举例说明调用字符和转义序列的用法。

如果一个汉字信息处理系统同时使用GB 1988和GB2312(这时可以将GB1988作为G0集，GB2312作为G1集)，则系统可用SI调用GB1988，用SO调用GB2312。如果GB2312不是作为G1集，而是作为G0集，则因一个系统同时存在两个G0集，故须用转义序列指明和调用哪个G0集。(这时调用字符可以省略)。但如果使用多个G0集和多个G1集，则必须先用转义序列指明使用的是哪个图形字符集，然后再用SI或SO调用。显然，对同时使用四个以下图形字符集的汉字信息处理系统，为了省掉指明的转义序列，分别将它们作为G0，G1，G2与G3集处理是最为简捷的。对于控制功能，则直接使用转义序列指明与调用。

以上简要地介绍了GB2311的内容及其使用方法，读者如需进一步了解，则应查阅GB2311文本及有关专门文章。

4.3.3 GB2312《信息交换用汉字编码字符集——基本集》

国家标准GB2312《信息交换用汉字编码字符集——基本集》是根据GB2311的代码扩充方法制定的汉字交换码标准。它是汉字信息处理系统中的基础性代码标准之一，由于它与GB1988兼容，所以能使通用计算机系统方便地扩充汉字处理功能并进行汉字信息的交换。

一、GB2312代码的构成

GB2312规定了进行一般汉字信息处理交换用的6763个汉字和682个非汉字图形字符的代码。每个汉字（包括非汉字图形字符）用两个字节表示，每个字节为七位二进制码。这七位二进制码分别与GB1988图形字符集G0集的94个七位代码，即0100001(2/10)～11 11 11 0(7/14)相对应。6763个汉字及682个非汉字图形字符就排列在这94×94个编码位置所组成的代码表中。

这个代码表纵向分成94个区，由第一字节标识，横向将每个区分成94个位置，由第二字节标识。因此代码表中的每个汉字或非汉字图形字符也可用它在代码表中所在位置的区号每位号来标识。显然这个代码表最多可收8836(94×94)个汉字与非汉字图形字符，构成一个双字节图形字符集。国际标准化组织对基本集给定的转义序列见前节。GB2312图形字符代码表中的空白位置留作将来实施标准化用。

二、汉字的选择、分级和排列

基本集中的6763个汉字的选择是在中华人民共和国文化部与中国文字改革委员会于1965年联合发布的《印刷通用汉字字形表》(6196字)的基础上，根据汉字信息处理的实际需要增加了五百多个科技、地名和姓名用字确定的。所以对我国绝大多数汉字信息处理系统来说，只要具有这基本集内的六千多汉字就能基本满足各种使用要求，而不必预备几万个汉字，这既有利于降低汉字信息处理系统的成本、缩短常用字的码长和提高汉字编码的效率，又有利于汉字信息处理技术的推广和应用。

根据我国1975年进行的查频统计，基本集内6763字的使用覆盖率可达99.99%左右，但是它们的使用频度却相差很大，根据二十年代至七十年代的各种汉字使用频度统计，

三千至四千个常用字覆盖率达99.9%左右。因此，实际上只要具备这三四千个字就能大体满足一般应用的需要。这样，无论从便于使用，或是从便于设备的分档制造，都有必要将这三、四千字作为一级常用字区分开。在综合考虑了汉字频度的高低、构词能力的强弱，实际用途的大小等情况，选择了一些具有代表性的常用字表，再进行重合字数统计，最后选出3755个字作为一级常用字，其余3008个字作为二级次常用字。

为便于人们检索代码表中的汉字，由于第一级汉字都是“常用字”，一般都知道读音，故按汉语拼音字母顺序排列。多音字取它的常用音，同音调字以起笔笔形横、竖、撇、捺、点、折为序。若起笔相同，则按第二笔，依次类推。第二级字大部分较生僻，不容易掌握读音，所以按部首排列较易检索。部首与一般通用字典相同，略有合并，部首顺序按笔画数排列，变形部首排在正部首之后，同部首字按除去部首以外的画数排列，同画数的字也按起笔笔形顺序排列。

三、汉字的字体与字形

基本集中汉字的字体以中国文字改革委员会1964年编印的《简化字总表》以及中华人民共和国文化部和中国文字改革委员会联合公布的《第一批异体字整理表》为准。字形一律以《印刷通用汉字字形表》为准。

四、GB2312的应用

GB2312目前广泛用于我国通用汉字系统的信息交换及硬、软件设计工作中。例如汉字字模库的设计目前都以GB2312为准。汉字整字输入键盘盘面文字的选择、汉字输入码的转换，以及汉字输出设备的汉字地址码，也都遵照GB2312来设计。此外，目前绝大部分处理汉字的高级语言、汉字数据库系统，以及汉字情报检索系统等软件的设计，也都采用GB2312或以GB2312为基础进行设计。

其次，由于GB2312是汉字信息处理技术领域内的基础标准，故许多标准都与它密切相关。例如：汉字点阵字形标准，磁带和磁盘等格式标准，各种汉字输入输出设备标准的制订等等，这些都应贯彻执行GB2312标准。

4.3.4 汉字点阵字模的设计与标准化

在汉字信息处理系统中，以汉字字模的形式直接进行输入输出，对其字形是有一定要求的。在输入技术方面，利用光学字符识别（OCR）技术，将印刷体或手写体的汉字直接输入计算机，在第五章中有详细的叙述。这里只着重介绍输出用汉字点阵字模的设计及其标准化。当然，在设计中应尽可能考虑到光学字模识别对汉字字模的一些要求。

一、标准化的目的和对象

汉字用于表达信息，它可以记录下来进行信息交换。为了便于人和机器的识别，汉字的字形、笔画等必须统一。由于有些复杂汉字的笔画繁多，故为了字迹清晰，有时需要减少笔画。这种简化应全国统一，否则会造成人们交换信息的困难。统一就是标准化的任务，汉字点阵字模标准化的目的在于：

（1）统一字形和笔画，防止异体字的产生；

（2）降低汉字字形发生器的造价；

（3）使汉字字模易于进行信息交换，并有利于采用光学字模识别技术。

标准化的对象和内容，从总体上和长远观点来看，或者从满足所有各方面的使用要

求来看，总数将近六万个的汉字都应该是标准化的对象。但从当前和满足一般广大用户的要求来看，有基本集和第一辅助集及第二辅助集的字就够了。因此点阵字形的标准化，首先应以 GB2312《信息交换用汉字编码字符集——基本集》中的6763个汉字和682个一般字符为对象，制订我国的汉字点阵字模国家标准。

二、点阵的系列化

若任各种汉字点阵的结构形式自由发展，必然会造成种类繁多（目前国内就已有十来种，字形更是各异），这种状况不利于工业生产和推广使用。因此，按照标准化的原则，根据技术经济合理的要求，考虑到生产和使用的实际情况，确定选择汉字点阵结构的种类和系列是：15×16；24×24；32×32；96×96。该系列的品种基本上能满足近期内生产和使用方面的要求。对于32×32以下的点阵，普遍用于针式打印机和显示器；在32×32及其以上的点阵，多用于激光印刷机和汉字照排系统输出设备上。

三、设计汉字点阵字模所用参照字体的确定

我国汉字有多种书写体和印刷体，它们各有自己的独特风格。汉字信息处理系统中应选哪一种作为它的基本字体呢？普遍认为选用宋体比较合适。这是因为，宋体汉字在书籍和报刊上用得最为广泛，人们见得最多，最为熟悉，从而容易被人识别和接受。宋体字具有横笔画细、竖笔画粗的特点，这对汉字点阵字模的设计尤其有利，因为汉字大多数是横笔画多，易于设计和造字。通过几年来的研究和实践证明，对于用计算机处理的汉字字形，以点阵的形式来表示有下述许多突出的优点：汉字能以数字化的形式表示，易于处理；汉字的设计造型和修改都较容易；设备制造的价格便宜。这种点阵式的汉字字形，目前已广泛采用在打印机和显示器上。

对于汉字点阵，点数的多少直接影响到汉字造型的难易和字形的质量。点数越多，汉字的点阵字模越接近于人们熟悉的参照字体，字形的质量也越高。例如，32×32比24×24点阵表现汉字的能力强。在32×32网络上设计出的汉字，基本上可做到与参照字体一样，它能用于印刷一般公文和其他资料。

四、汉字点阵字模设计的具体要求

本节中所描述的点阵是以字面为限度来表示的，即：横向点数×纵向点数。整套字都在规定的某种点阵范围内设计，但并不是每个字都占满格，只有一部分字需要占满格。因此，它是以实际字面的最多点数为限度来计算的。每个汉字的点阵字模，由置于小方格内的许多小点确定。例如：汉字“跑”的点阵字模可以由图 4-6 表示。

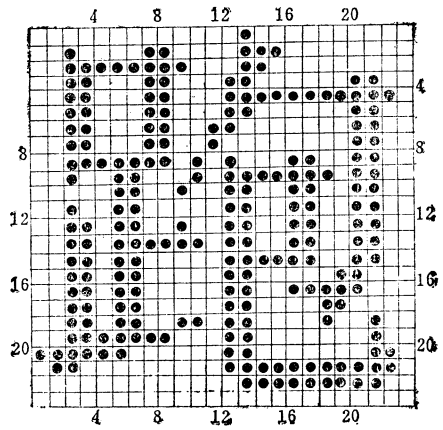


图4-6　24×24点阵的汉字字模“跑”

设计好一套汉字字模需解决设计中的许多具体问题。这就必须遵循下述一些设计原则：

（1）字形正确，符合标准。例如：国家标准 GB2312 中的“启”字，不能用“啓”来表示。

（2）字的笔画数要符合查字法。如：“长”是四画，而不是六画的“长”。

（3）字要符合书写规范。不能用短横或长横笔画代替点。

（4）从总体上看，要使一整套字的大小匹配适宜。有的字设计宜大些，如："命令"这两个字，横向要占满格，否则看起来字显得小。而"田"字设计宜小，不应占满格，否则与别的字排列在一起就显得字体大。

（5）字的结构端正，重心要稳，上下左右要协调。

（6）字的一笔一画应尽可能的舒展自然，撇、捺势顺，不要使字显得僵硬。

（7）分清字构件的主次，粗细、黑白处理要适当。

除了遵照上述原则外，在实际设计中，还要根据汉字的形态、笔画的多少、传统的结构、字面的大小、字间间隔等灵活地处理字面，并进行必要的验证和反复的修改，以达到汉字的形神兼顾和均匀美观的目的。

若要考虑汉字的光学字模识别，对形体相近的汉字，应使它们之间的差异在5％以上，这是一条总的原则。由于汉字的数量相当大，要实现这条要求是有不少困难的。

为了使印出的汉字笔画连续，字形美观，在某些设备（例如：针式打印机）上实现时，点的大小的确定和距离的选择与相邻点的重合度有关，即是与相邻点覆盖的多少有关，一般取其重合率在30％左右为宜。

4.3.5 汉字交换码辅助集的标准化

GB2312颁布以来，国内外已有许多厂家和机构按照这一标准生产与研制了各种汉字信息处理设备与系统。根据两年多来的使用情况来看，基本集的六千七百余字，已基本满足绝大部分用户的使用要求。但是，对于某些仍然使用繁体字的地区及使用字数特别多的用户（例如台湾省、北京图书馆、大城市的户籍处理系统等），则感到只有基本集的六千多汉字还不够用。为此，我们根据一些实际使用的字表、辞典，通过对各字在一些典型辞典中义项数的统计，从五万余字中（不包括基本集的六千七百余字）筛选一万六千余字，分配在两个集合中，作为信息交换用汉字编码字符集的第一辅助集和第二辅助集。

一、编码结构

辅助集的编码结构与基本集完全相同，即各个辅助集中的每个汉字也用两个七位编码表示。第一区至第四区为保留标准区。区字排列在第五区至第94区（见图4-7）。

区号＼位号	1 2 3 4 5 …… 93 94
1 2 3 4	保留标准区
5 6 …… 93 94	汉字图形字符区

图4-7 辅助集编码结构

二、汉字的选择与分级

根据我国1975年进行汉字查频统计，基本集以外文字的使用频度都很低（在万分之一以下）。所以我们认为，辅助集中的字不宜采用阶梯式查频统计的选择方法，而应根据一些实际使用的汉字表与一些典型辞典，参照各字在这些辞典中的字义项数(构词能力)和实际用处等进行选择和粗略分级。

具体方法简述如下：

以《辞海》、《中华大字典》、《中文大字典》中所收字为主要汉字，并作为统计、筛选的基础，同时增加《汉语大字典》(正在编纂中）的部分新增字。

用卡片分别登记各个单字在上述三部辞典中的音项数，义项数，例句数，书证数，所带的条数以及该字的异体、伪体等情况，并进行分类统计。根据各集应收字数，第一辅助集为各大类中的《辞海》卡；非《辞海》卡中四个义项以上的各类中有词条卡，有例句卡。第二辅助集为两个义项和三个义项两类中有词条卡和有例句卡；三个义项以上各类中有书证卡和无词条、例句、书证卡。以上两个义项一类中有书证卡和无词条、例句、书证卡。

上述选择是基于以下这样一些事实。

对于大部分单字，义项数越多，使用频度就越高。

其次，单字的构词能力与该字的使用频度有关：单字的构词能力越强，使用频度就越高（只有少数例外）。

第三，仅有书证或甚至没有书证的，因其使用频度已无可查考，即使其义项数较高也不归入第一辅助集。

第四，考虑到《辞海》是一部界于《新华》字典、《现代汉语词典》与《中文大字典》、《中华大字典》之间具有较为广泛实用性的辞书，所以，对于《辞海》所收单字，

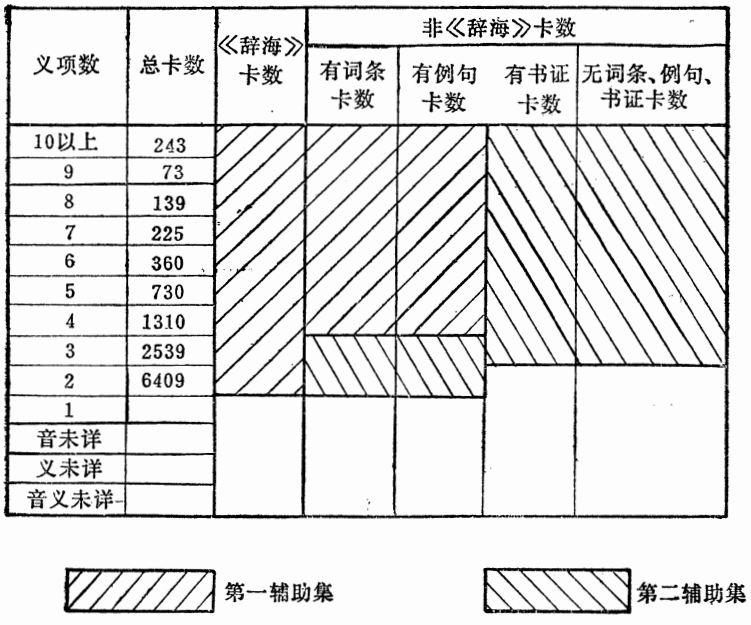


图4-8　第一、二辅助集汉字表提取汉字情况

即使义项数较低，也基本收入第一辅助集，而将一些不常用的异体字则收入第二辅助集。

经如上规则的处理，第一、二辅助集字表提取情况将如图 4-8 所示。

图 4-8 中的空白部分为第二辅助集以后各集的收字范围。

需要指出的是，上述根据字义项数的分类只是一种粗略的分类。有少数字虽然义项数很少，但仍然用到。所以，我们认为，根据实际情况应将未收入上述第一、二辅助集字表中的北京图书馆《中文图书目录检字表》、邮电部《标准电码表》以及《新华字典》、《现代汉语词典》中的 565 字补入第一辅助集。

综上所述，我们建议第一辅助集中的八千余字是继基本集之后的，包括《新华字典》、《现代汉语词典》、北京图书馆《中检字文目录表》、邮电部《标准电码本》中几乎全部的单字，以及《辞海》中的绝大部分单字；第二辅助集的八千余字中为少数《辞海》字（不常用异体字），以及五万七千余字中未收入基本集、第一辅助集中的具有二个义项以上的单字。

三、字体与字形

如上所述，第一、二辅助集的编制，主要是为了便于目前尚未推广简化字地区的汉字信息处理与交换，以及一些用字量较大的用户需要。这些字不是在使用上有一定的地区局限性，就是带有较强的“专用”性，所以我们建议，一律采用旧字体和旧字形，而不再进行类推简化。只是为了利用现成的字模，便于将来标准的印刷，第一辅助集中的字形暂采用新字形。

四、字的排列

由于辅助集的字绝大部分都是生僻字，为便于检索，我们建议一律按部首、笔画数排列。其方法同基本集。

五、使用说明

鉴于信息处理交换用汉字编码字符集将有三个集合：基本集；第一辅助集；第二辅助集。在实际应用中，汉字编码字符集组成的数据结构一般有如下几种方式。

（一）只用基本集

当系统只用基本集 GB2312，则无须使用转义序列和调用字符。

（二）同时使用 GB1988 和 GB2312

当系统同时使用 GB1988 与 GB2312 时，可将 GB1988 作为 G0 集，GB2312 作为 G1 集，用移出字符 SO 调用 GB2312，用移入字符 SI 返回 GB1988(见图4-9)。

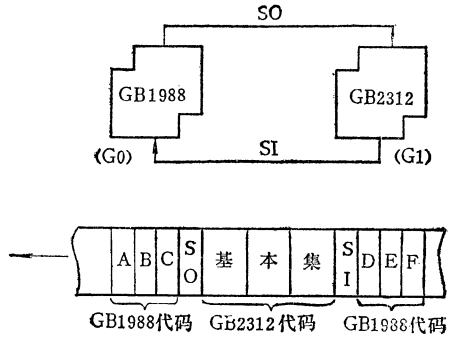


图4-9　同时使用GB1988和GB2312时的调用方法

（三）使用 GB1988、GB2312 及一个或两个辅助集

这时可分别将 GB1988、GB2312 及一个或两个辅助集作为 G0 集、G1 集、G2 集或G3 集。用移出字符 SO 调用 G1 集（基本集），用单移字符 SS2 或 SS3（只调用 G2 集或 G3 集中一个汉字，然后自动返回原有状态）；或用锁定字符 LS2 或 LS3 调用G2 集(第一辅助集）或 G3 集（第二辅助集），

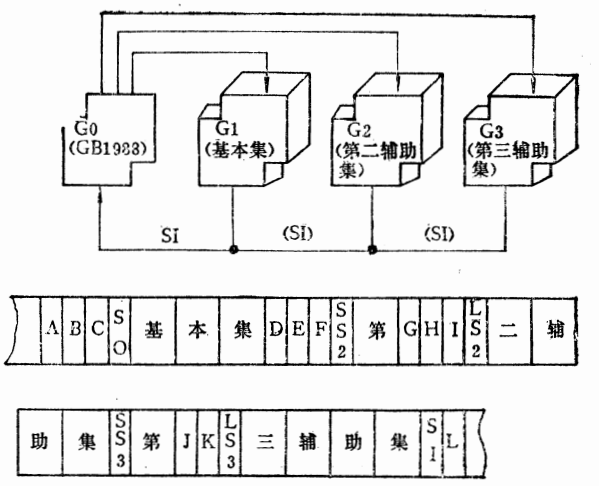


图4-10　同时使用GB1988、GB2312及一个或两个辅助集时的调用方法

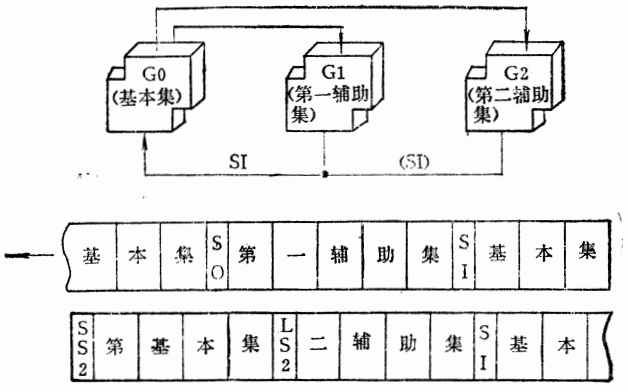


图4-11　同时使用GB2312及第一和第二辅助集时的调用方法

用移入字符 SI 返回 G0 集（见图4-10）。

（四）使用基本集和两个辅助集

这时可将基本集、第一、二辅助集分别作为 G0 集、G1 集和 G2 集（见图4-11）。用移出字符 SO 调用 G1 集（第一辅助集）；用移入字符 SI 返回 G0 集（基本集）；用单移字符 SS2 或锁定字符 LS2 调用 G2 集（第二辅助集）；用移入字符使其返回 G0 集（基本集）。

（五）超越字界处理

无论是基本集还是辅助集，每个集合最大容量为8836字。当使用字数超过此数时，就要同时使用两个或两个以上的集合。这时，因增加调用字符而使汉字代码不等长，故引起内部处理复杂化。为了系统内部处理方便，在系统内部每个汉字可用两个八位编码、三个七位编码或其它等长码表示。

4.3.6　文字图形设备增补控制功能的标准化

目前我国即将颁布用于汉字设备的汉字控制功能码国家标准，即“文字和符号图形

设备的增补控制功能”，以下简称“汉字功能码标准”。

汉字控制功能码是用于汉字数据处理、传送或解释执行的代码。GB1988C0 集中的32个控制字符是进行汉字数据处理的基本控制功能。

ISO6429 字符图形产生设备用的增补控制功能，是设计汉字数据处理用的附加控制功能的基础，共有87个。但是必须指出，由于 GB1988 和 ISO6429 主要用于处理西 文、数字等非汉字字符，未充分考虑汉字等表意文字的特点，所以都不能完全满足汉字数据处理的需要。

我国根据汉字的特点，为便于汉字数据的处理，对 ISO6429 进行了扩充性修改，在保持与其兼容的基础上，又增加了一些用于汉字数据处理的功能和模式，它是既适用于汉字数据处理，又适用于西文数据处理的汉字控制功能码标准。

一、汉字功能码的分类

汉字控制功能码标准共规定了 101 个控制功能，按用途分为以下十类：

（1）控制串定义符：

APC	应用程序命令
DCS	设备控制串
OSC	操作系统命令
PM	保密消息
ST	串终止符

（2）引导符：

CSI	控制序列引导符
ICSI	汉字控制序列引导符

（3）转移功能：

SS2	单移 2
SS3	单移 3

（4）格式控制符：

HPA	（*n*）	字向位置绝对移动
HPB	（*n*）	字向位置反向移动
HPR	（*n*）	字向位置正向移动
HTJ		对齐的字向制表
HTS		置字向制表
HTSA	（*n* …）	置字向制表绝对位置
HVP	（*n*，*m*）	置字向和行向位置
IND		移行
INL	（*n*）	汉字新行
NEL		下一行
PLP		行部分下移
PLU		行部分上移
RI		反向移行
SGR	（*s* …）	图形显示选择

TBC (*s* …) 清制表
VPA (*n*) 行向位置绝对移动
VPB (*n*) 行向位置反向移动
VPR (*n*) 行向位置正向移动
VPS 置行向制表

(5) 文稿组版备用的增补格式控制符:

ECR 字符旋转结束
ESN 注音结束
FNT (*s*; *t*) 字体选择
GSM (*n*; *M*) 字符图形大小修正
GSS (*n*) 字符图形大小选择
HSS (*n*) 半字指定
JFY (*s* …) 整版
LLS (*n*) 行长指定
QUAD (*s* …) 单行对齐
SCR 字符旋转开始
SFD (*s*) 字符图形书写方向选择
SPI (*n*, *m*) 间隔指定
SSN 注音开始
SSU (*s* …) 单位大小选择
TSS (*n*) 间隔增量

(6) 修改可见数据用的编辑功能:

DCH (*n*) 删字
DL (*n*) 删行
EA (*s* …) 消区
ECH (*n*) 消字
ED (*s* …) 消页
EF (*s* …) 消段
EL (*s* …) 消行
ICH (*n*) 插字
IL (*n*) 插行

(7) 移动操作位置用的编辑功能:

CBT (*n*) 位标反向制表
CHA (*n*) 位标字向绝对移动
CHT (*n*) 位标字向制表
CNL (*n*) 位标正向移行
CPL (*n*) 位标反向移行
CTC (*s* …) 位标制表控制
CUB (*n*) 位标反向移动

CUD	(n)	位标行向正移
CUF	(n)	位标正向移动
CUP	(n；m)	置位标
CUU	(n)	位标行向反移
CVT	(n)	位标行向制表
NP	(n)	正向换负
PP	(n)	反向换负
PPA	(n)	页面位置绝对移动
PPB	(n)	页面位置反向移动
PPR	(n)	页面位置正向移动
SD	(n)	下滚
SL	(n)	左滚
SR	(n)	右滚
SU	(n)	上滚

(8) 区的限定：

DAQ	(s …)	限定区定义符
EPA		防保区结束
ESA		选择区结束
SPA		防保区开始
SSA		选择区开始

(9) 模式设置：

RM	(s …)	清模式
SM	(s …)	置模式

(10) 其它控制功能：

CCH		字符作废
CPR	(n；m)	位标位置报告
CPT	(s)	字形点阵传送
DA	(s …)	设备类型
DMI	(Fs)	禁止人工输入
DSR	(s …)	设备状态报告
ECC		字符合成结束
EMI		允许人工输入
EXC	(s …)	外字代码指示
IDCS	(s)	设备控制串标识
INT	(Fs)	中断
MC	(s …)	复制
MM		消息等待
PU1		专用 1
PU2		专用 2

REP	(*n*)	重复
RIS	(*Fs*)	恢复初始状态
SCC		字符合成开始
SCP	(*s* …)	字形点阵构成方式选择
SEC	(*s*; *t*)	外字代码构成方式选择
SEE	(*s* …)	编辑范围选择
STS		置发送状态

注：缩写词后括号内字母含义：

a(*n*) ：带一个数值参数的控制序列；

b(*n*; *m*)：带两个数值参数的控制序列；

c(*n* …) ：数值的参数的个数可变的控制序列；

d(*s* …) ：选择参数的个数可变的控制序列；

e(*s*; *t*)：带两个选择参数的控制序列；

f(*Fs*) ：ESC Fs序列

限于篇幅，本章中不能详细介绍每个控制功能符的定义，读者实际使用时将来可参阅该标准文本。

二、功能码的编码表示

汉字功能码的表示方法有三类。

(一) 属于C1集的控制功能

在七位编码中，用二字符转义序列表示，其形式是ESC Fe，其中Fe是4列或5列的位组（见表4-1）。

表4-1 C1集中控制功能的位组分配

行号	列号		行号	列号	
	A	B		A	B
0	—	DCS	8	HTS	—
1	—	PU1	9	HTJ	—
2	—	PU2	10	VTS	—
3	ICSI	STS	11	PLD	CSI
4	IND	CCH	12	PLU	ST
5	NEL	MW	13	RI	OSC
6	SSA	SPA	14	SS2	PM
7	ESA	EPA	15	SS3	APC

(二) 用控制序列表示

控制序列由“控制序列引导(CSI)”和“汉字控制序列引导符(ICSI)”的编码表示及跟在它后面的一个或几个确定控制功能的位组及表示控制功能参数的位组等组成。控制功能CSI和ICSI均是C1集中的一个控制功能。

CSI和ICSI构成控制序列的格式是一样的。其通用格式表示为：

$$CSI\ \ P_1\cdots P_n\ \ I_1\cdots I_mF$$

其中：

(1) CSI为CSI或ICSI。其七位编码表示为ESC 5/11和ESC 4/3。

（2）$P_1 \cdots P_n$ 是3列中的位组，它表示参数值。若该控制功能没有参数，或采用隐含值，则这些位组都可以省略。

参数有两类：数值参数和选择参数。数值参数表示一个数，而选择参数则表示一字符串，其意义取决于控制功能本身。

（3）$I_1 \cdots I_m$ 是2列中的位组，它与终止字符F一起确定控制功能，如果该控制功能仅由终止字符F确定，则这些位组就省略。

（4）F是4列、5列、6列或7列（7/15除外）中的位组，它使控制序列终止。若有中间位组，则与中间位组 $I_1 \cdots I_m$ 一起确定控制功能（见表4-2、表4-3以表4-4）。

表4-2 无中间位组的CSI控制序列的终止位组分配

行号	列号		
	4	5	6
0	ICH	DCH	HPA
1	CUU	SEE	HPR
2	CUD	CPR	REP
3	CUF	SU	PA
4	CUB	SD	VPA
5	CNL	NP	VPR
6	CPL	PP	HVP
7	CHA	CTC	TBC
8	CUP	ECH	SM
9	CHT	CVT	MC
10	ED	CBT	HPB
11	EL	—	VPB
12	IL	—	RM
13	DL	—	SGR
14	EF	—	DSR
15	EA	SFD	DAQ

表4-3 有一个中间位组2/0的CSI控制序列的终止位组分配

行号	列号		
	4	5	6
0	SL	PPA	
1	SR	PPR	
2	GSM	PPB	
3	GSS		
4	FNT		
5	TSS		
6	JFX		
7	SPI		
8	QUAP		
9	SSU		
10	—		
11	—		
12	—		
13	—		
14	HTSA		
15	IDCS		

表4-4 ICSI控制序列的终止位组分配

行号	列号		
	4	5	6
0	ECC	HSS	SEC
1	SCC	LLS	
2	SSN		SCP
3	FXC		CPT
4	SCR		
5	ESN		
6	ECR		
7			
8			
9			
10		INL	
11			
12			
13			
14			
15			

（三）用 ESC Fs 序列表示

用 ESC Fs 序列表示的控制功能按 GB2311 的规定，这些控制功能的编码表示是以 ESC Fs 形式表示的二字符转义序列，其中 Fs 是 6/0 至 7/14 中的一个位组（见表 4-5）。

表4-5 ESC Fs序列

缩写形式	名称	编码
DMI	禁止人工输入	ESC6/0
EMI	允许人工输入	ESC6/2
INT	中断	ESC6/1
RIS	恢复初始状态	ESC6/3

三、汉字功能码标准的应用

汉字功能码标准适用于各种汉字与字符图形设备及各种汉字数据的交换体系。对于一台具体的设备，并不要求也不可能具备该标准所规定的所有功能。但是，它可以选择其中的部分控制功能、控制功能的参数和模式，或者使用别的控制功能和模式。一台与国标汉字功能码标准兼容、或有限兼容的设备应该符合以下各点：

（1）设备能用指定的“编码表示”和意义实现本标准中规定的控制功能子集、控制功能的参数和模式；

（2）如果实现的控制功能子集包含有本标准中规定的隐含参数值的控制序列，则当这个隐含值不论是明示或暗示时，该设备应能接收该控制序列，并能进行正确的解释；

（3）标准中用于指定控制功能的任何编码表示，不应该再用于表示其他的控制功能；

（4）标准中留作今后标准化用的任何一个编码表示不应被占用；

（5）设备可以实现标准指定的模式以外的模式。但是这些专用模式的一个状态应该用标准规定的编码表示和意义来实现。

汉字功能码标准不规定控制功能的具体实现方法，控制功能码的具体实现取决于设备本身的设计或通信各方的规定。

四、控制功能使用举例

（一）字符合成

一般说来，合成简单的图形字符可以利用格式控制符“退格(BS)”。

例如：

○BS★→⍟

但是，当合成较为复杂的汉字或其它图形字符时，使用“字符合成开始(SCC)”与“字符合成结束(ECC)”较为方便。

例如：

〔数据〕　　〔表示〕

SCC 田又又又又 ECC→㬎

产生合成字符“㬎”的具体方法，取决于执行过程。

（二）注音处理

当给汉字加注汉语拼音时，可以使用格式控制符“注音开始(SSN)”和“注音结束

(ESN)"。但是被注音的字与注音字母间应用GB1988中的"单元分隔字符(US)"分隔开。

例如:〔数据〕 SSN 畧USzhvō ESN

〔表示〕 zhvō 畧或畧 (zhuō)

注音字母标注在被注文字的上部还是在后面,这取决于执行过程。

(三) 汉字新行

当书写汉字文稿需要分段或另外起行时,可以使用"格式控制符(INL)"。其数值参数的隐含值为3,意即将操作位置移到下一行的第三个字符位置(汉字文稿的段落起行通常都是从第三个字符位置开始)。

例如:〔数据〕INL 七绝 SP SP 黄鹤楼 INL8 李白 INL 故人西辞黄鹤楼,INL 烟花三月下杨州,INL 孤帆远影碧空尽,INL 唯见长江天际流。

〔表示〕

七绝　黄鹤楼

李白

故人西辞黄鹤楼,
烟花三月下杨州,
孤帆远影碧空尽,
唯见长江天际流。

(四) 外字处理

外字处理是汉字数据处理中一种特有的处理技术。对于我国目前已经制定的图形字符代码标准,GB1988和GB2312以外的图形字符,均作为外字处理。

外字代码用控制功能"外字代码指示(EXC)"标识,外字代码的构成方式由"外字代码构成方式选择(SEC)"确定。

例如:EXC 5 7 6 SEC 1 3

意思是外字代码是用GB1988中的0～9(3/0～3/9)表示,共三个字节,具体代码为"576"。

当然,SEC显示置于EXC之前,即SEC 1 3 EXC 5 7 6。

如用GB2312中的0～9(2/3 3/0～2/3 3/9)表示上述外字代码,则为:

EXC 5 7 6 SEC 2 6

五、字形点阵的传送

字形点阵的构成方式用"字形点阵构成方式选择(SCP)"指定。根据SCP指定的字形点阵结构,用"字形点阵传送(CPT)"传送该字形点阵所表示的图形字符的代码及其字形点阵。

例如:SCP 15 16 0 0

表示:字形点阵为15(宽)×16(高),并按宽方向顺序扫描(见图4-12)。

用此模式传送上例中外字代码为"576"(共3个字节)的字形点阵时,用如下方式:

CPT 10 a_1 a_2 a_3 a_4 a_5 …… a_{67}

10 外字　a_1 a_2 a_3 外字代码　a_4 a_5 …… a_{67} 字形点阵(64字节)

选择参数“10”意为“外字”，$a_1a_2a_3$表示该外字的代码，a_4～a_{67}共64个字节表示该外字的字形点阵。这里，a_4～a_{67}用GB1988列3位组的下4位（b_4～b_1）对应字形点阵。

例如：a_4～a_7的下4位同点阵的对应关系如下：

	b_4	b_3	b_2	b_1
a_3	①	②	③	④
a_4	⑤	⑥	⑦	⑧
a_5	⑨	⑩	⑪	⑫
a_6	⑬	⑭	⑮	

由于指定为“非连续装配”传送，故a_6的最后位（b_1）不对应字符点阵。而a_7的下4位对应点阵⑯⑰⑱⑲。下同。这样每个字形点阵将一共浪费16个“b_1”，即16位（bit）或16点。

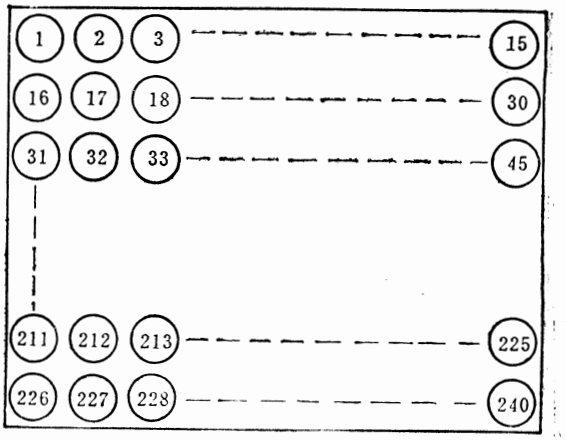


图4-12　用控制功能码构成和传送15×16图形字符的点阵

当模式指定为：

SCP 15 16 0 1时，则由于指定为“连续装配”传送，故上述外字字形点阵传送为：

CPT 10 a_1 a_2 a_3 a_4 a_5 …… a_{67}

外字　外字代码　字形点阵（60字节）

这时，同样的字形点阵只用60个字节，这是因为要避免上述非连续装配传送时每扫一行就要浪费一个点的缘故。即上述a_6的最后一位（b_1）对应字形点阵第⑯点，而a_7的下4位对应点阵⑰⑱⑲⑳。下同。

显然，在传送大量图形字符的字形点阵时，采用“连续装配”传送可提高传送效率，并节省有关资源，但处理过程比“非连续装配”传送要复杂些，故通常传送较少或个别字形点阵时，采用“非连续装配”传送；而传送大量字形点阵时，则采用“连续装配”传送。

以上介绍了汉字信息处理领域中我国已经制定和正在制订的几个代码标准。其中GB1988是最基本的标准，几乎所有的信息处理设备和系统都离不开这一标准。GB2311提供了一个几乎是可以无限地扩充对应于GB1988的各个元素，而又保持与其兼容的编码扩充方法。根据这种扩充方法制定了GB2312和汉字功能码标准。有了GB2312就可以制订点阵字模标准，信息交换用磁带、磁盘等记录格式标准，以及传输系统内一系列有关标准。所以，本节介绍的都是汉字信息处理系统中极为重要的基础标准。它们直接影响汉字信息处理系统的硬件和软件的设计周期、产品质量以及产品的推广应用。因此，作为一个汉字信息处理系统的设计师必须了解和熟悉这些标准，并在具体设计应用中贯彻执行这些标准，否则很难想象能设计出一个质量高、成本低、通用性好，适于在我国批量生产和普及推广的汉字信息处理系统。

第五章　汉字输入方法和设备

在汉字信息处理系统中，首先遇到的是如何输入汉字的问题。然而，要完善地解决这一问题是相当困难的。其基本原因是汉字的字量多达 6 万，常用字即有 3000 字以上。因此，不可能象拼音文字那样，用普通的字母数字键盘来方便地解决输入问题，此外，由于用户的需要不同，使用的文字范围也不同。因此，较难设计出一种标准的汉字输入设备。

目前所使用的汉字输入方法大致可以分为两种类型：一种是用人工操作某种物理设备输入；另一种是用人工智能方式直接对文字或语音进行识别输入。典型的物理输入设备包括各种典型的键盘及字盘；人工智能输入设备主要是汉字识别及语音识别系统或装置。

各种键盘和字盘是目前国内外使用最广泛的汉字输入设备。虽然它仍处在不断地发展和完善的阶段，但已在各种实用的汉字信息处理系统中成功地解决了汉字的输入问题。汉字识别目前还处于研究阶段，印刷体及手写体汉字的识别技术已取得了一定的进展，某些系统的识准率达 98%以上，拒识率为 1.3%，误识率为 0.3%。语音识别，目前限于特定的发音者，并主要用于单个汉字或字组等某些特定的场合，作为汉语输入用的一般语音识别系统，还有待进一步探索。

5.1　汉字键盘输入方法

目前在各种汉字信息处理系统中所使用的汉字键盘是多种多样的。但是，按照操作方式，基本上可分为直接输入方式、间接输入方式和人机对话输入方式三种。直接输入方式，是在键盘或字盘上选择所需的汉字键或接触字盘直接输入汉字代码，其典型代表是汉字整字键盘和各种笔触式字盘。间接输入方式则是利用输入汉字的编码来输入汉字。各种字根式汉字键盘和字母数字键盘属于这种方式。对话输入方式则是利用终端设备的处理功能通过人机对话来输入汉字由于使用的目的不同，在不同的汉字信息处理系统中，往往采用不同的汉字输入键盘。

汉字键盘的主要分类如表 5-1 所示。

表5-1　汉字键盘分类

键盘输入方式	直接输入方式	汉字整字键盘	全键式整字键盘 多段移位式整字键盘	
		笔触式汉字字盘	静电耦合式字盘 电磁感应式字盘 光电式字盘	磁致伸缩式字盘 压感式字盘
		中文打字机式汉字键盘	电磁感应板式 活字代码式	条型码式 全息照相编码式
	间接输入方式	汉字字根键盘	字根式汉字键盘 非字根式汉字键盘	
		标准字母数字键盘	汉字拼音输入键盘 汉字字根或笔形编码输入键盘	
		联想式人机对话汉字输入		
	人机对话输入方式	标准 字母数字键盘	汉字上形、下形的联想检索方式 专业常用字的联想检索方式	

5.1.1 汉字整字键盘

汉字整字键盘是使用最早的一种输入设备，它具有直观易学的特点。一般盘面收容字量在 2,000 以上。汉字整字键盘主要有全键式、多段移位式两种。

一、全键式汉字整字键盘

全键式汉字整字键盘是采用一字一键的方式，键盘上的每个键都与一特定的汉字相对应。它的基本操作就是选字和击键，一次输入一个汉字。键可以用机械触点式，也可以用无触点式（如霍尔效应键）。全键式键盘的主要问题是键盘太大，收容字量受到键盘尺寸的限制，输入速度较低。这种键盘已不多见。

二、多段移位式汉字键盘

多段移位式汉字键盘是在全键式汉字键盘的基础上，采取了类似于字母数字键盘用移位键（*shift key*）来区分大、小写字母的方法，即在一个文字键上定义多个汉 字，用相应数目的移位键来区别文字键上的各个汉字。

实用的多段移位键盘的移位键数，通常有 4 、 8 、 9 、12、 13、 14、 15、 30等多种。移位键数越少，文字键数就要求越多。目前用的最多的是 9 ～15段移位键盘。典型的文字键数为 2000～2700 个，多的可达 5000 字。

图 5-1 为一种多段移位式汉字整字键盘。

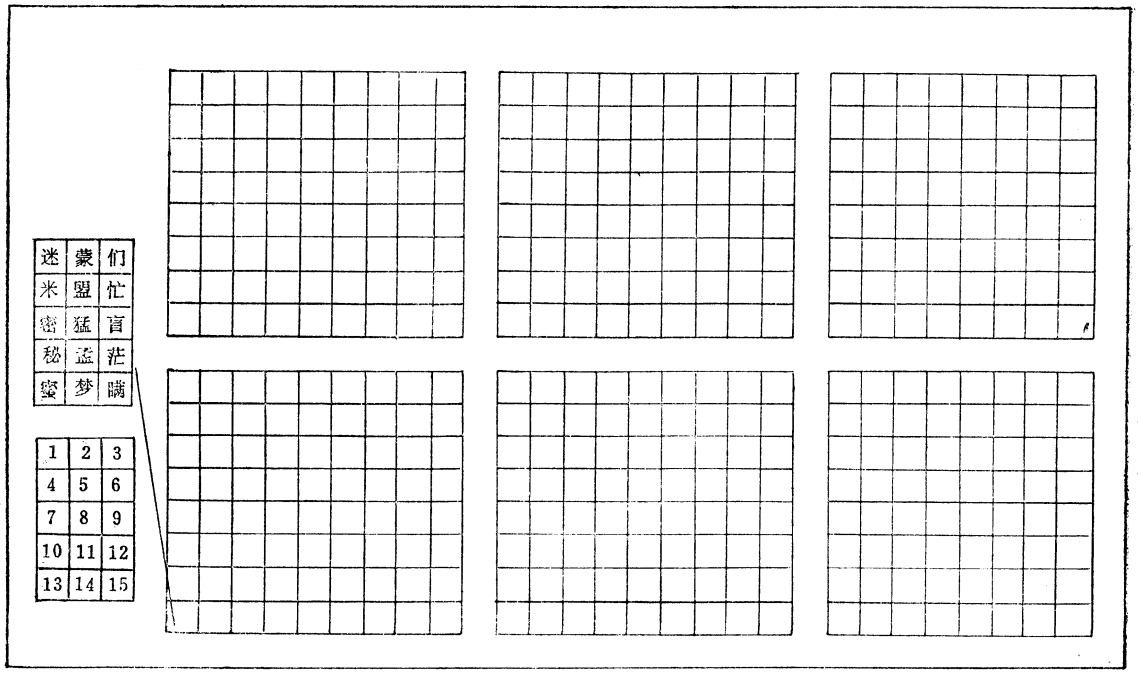


图5-1 多段移位式汉字整字键盘

汉字整字键盘输入方式的特点是：直观、操作简单，一般人都可以使用。据称经过专门训练，输入速度可以达到每分钟 100 字。而且，由于操作员比较容易掌握键的大小和间隔，因此具有一定经验的操作员可以盲打移位键（即选择键），甚至可以盲打频度高

的汉字的文字键。一般适用于输入数量很大的报社及和书刊的编辑排版部门。

它的主要缺点是设备比较庞大，价格比较贵，因此，不利于普及应用。

各种汉字整字键盘盘面设计与笔触式汉字字盘的盘面设计是类似的，将在后面介绍。

5.1.2 笔触式汉字字盘

笔触式汉字字盘(pen touch tablet)输入方式是近几年来发展很迅速、并且也已经被大量使用的输入方式。

笔触式汉字字盘输入方式和一般图形输入方式很相似，一般由坐标盘、字盘、接触笔、控制器等四个主要部分组成。字盘上印有按矩阵排列的汉字，将字盘覆盖在坐标盘上，使字盘上的汉字与坐标盘上的X、Y扫描线的交叉点一一对应。这些交叉点就是文字位置检测点。当用输入笔触及字盘上的汉字时，坐标盘就输出它所对应的X、Y坐标值。控制器的作用是对坐标盘进行驱动、检测（扫描）出X、Y坐标值，并把它转换成对应的编码。通过接口部分与终端设备或计算机相连接。其组成框图如图 5-2 所示。

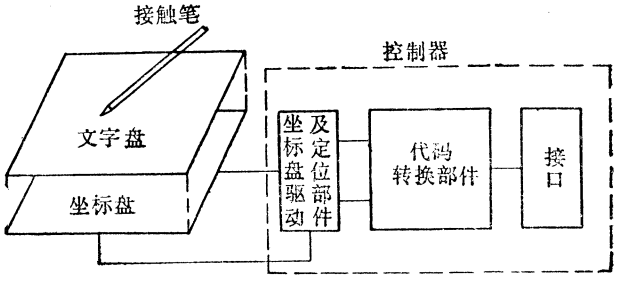


图5-2 笔触式字盘组成框图

笔触式字盘是一种目视检字输入装置，它也是采用整字排列，一一对应的输入方式。文字盘是一个印有文字的字板，在这种文字盘面上找字与在字典上按笔画或音顺检字一样。因此，操作很直观和方便。对于非专业的操作人员，它是一种比较容易接受的输入装置。但是，由于只能单手操作，而且必须用输入笔逐字进行输入，所以，速度不会很高。即使是熟练的操作员，要想达到高于60字/分钟的输入速度也是比较困难的。而且，用笔在很小的格子里点字，时间长了就很容易疲劳。

笔触式汉字字盘输入装置，按其坐标盘的工作原理不同，可以分为静电式、电磁式、光电式、磁致伸缩式及压感式等。输入笔一般分为有线式及无线式两种。某些压感式字盘不用笔触，而是直接用手指触摸输入。

一、静电耦合式字盘

静电耦合式字盘的结构形式尽管多种多样，但基本上都是由屏蔽网络、X扫描导线组、Y扫描导线组和基板等几个部分组成，经精确定位后把几部分叠加在一起，形成一个完整的坐标定位盘。

图 5-3 为我国生产的一种静电耦合式字盘的结构。共分三层，从下向上第一层为基板。它是一块 1.5 毫米厚的双面复铜板。该层的下面为铜箔，作为屏蔽层。主要是屏蔽控制器及空间的高频信号的辐射对字盘产生的干扰。该层的上面布有若干根沿X轴方向相互平行的X扫描导线。X扫描信号脉冲依次加在各条X扫描导线上。这些X扫描导线

被扫描信号驱动时，它就成为传送某根导线在坐标盘的X方向的脉冲信号发射体。

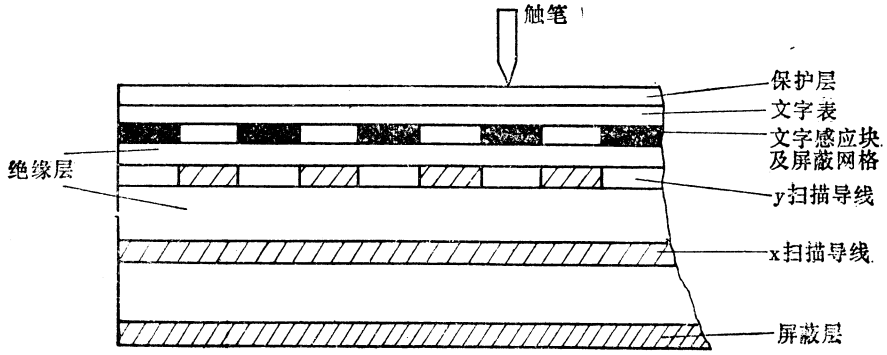


图5-3　静电耦合式字盘结构图

第二层为绝缘层，它是一块厚度约 0.2 毫米的玻璃纤维布（通常称为工艺填料），起第一层与第三层之间的粘合和层间的绝缘作用。

第三层是一块 0.3 毫米厚的双面复铜板，该层的下面布有多根沿Y轴方向相互平行的Y扫描导线，其基本作用与X扫描导线相类似，只是发射Y方向扫描脉冲信号。该层的上面是坐标点位置感应块和屏蔽网格。每一个文字感应块都处在屏蔽网格的中心，并与下面的X、Y扫描导线的交叉点相对应。而字盘上的每一个汉字与每个位置感应块又一一对应。位置感应块是一块金属导体，它所感应到的X、Y扫描信号在位置感应块平面上的任何一点都是等电位的，而每个汉字的字形正好覆盖在一个位置感应块内，当触笔触接在字位的中间区域或者四周边缘时，其探测效果都是一样的，都能可靠地进行位置或相应的汉字检测。

在使用时，当触笔触及字盘上所选定的某一个汉字时，X、Y扫描导线组在扫描计数器的选通下依次加上扫描信号，由于触笔的笔尖部分与X、Y扫描导体之间分别构成了静电容，对应于触笔的X、Y扫描信号便由电容耦合到触笔内，经放大整形后，送给控制电路，并反馈到计数器，立即读出扫描计数器在这一瞬间的计数值。该值就是所选的文字在字盘上的X、Y坐标位置代码。

典型的静电耦合式坐标盘的指标如下：

文字数：3072字（64字×48字）

盘面字尺寸：5 × 5 （毫米^2）

文字间格：6 毫米

接触笔重量：40克

盘面尺寸：502×388×47（毫米^3）

二、电磁感应式字盘

电磁感应式字盘分有线的和无线的两种。

（一）有线电磁感应式字盘

有线电磁感应式字盘的工作原理如图 5-4 所示。当输入笔尖的线圈通过电流时就产生磁场，通过坐标盘内的读取线将磁场检出，从而得到笔的位置信号。在一个读取回路

中感应到的电动势的相位随着笔尖与回路的相对位置而改变。即笔尖在回路内侧（图5-4 a）和外侧（图5-4 b）所感应的电动势的方向是不同的（图5-4 c）。检出这种相位差，就可以知道笔尖与回路的相对位置。将线圈配置在X、Y两个方向，就可获得X、Y两坐标上的信号，经数字化后，就可得到笔的位置的平面坐标值。回路的排列使得所产生的相邻坐标位置代码只有一位不同（格雷编码），这样就可以准确检出笔的位置（见图5-5）。盘面上的文字与文字之间设置一个对触笔不敏感的间隔，以防止错误输入。即使产生这一区域的信号也不作处理。

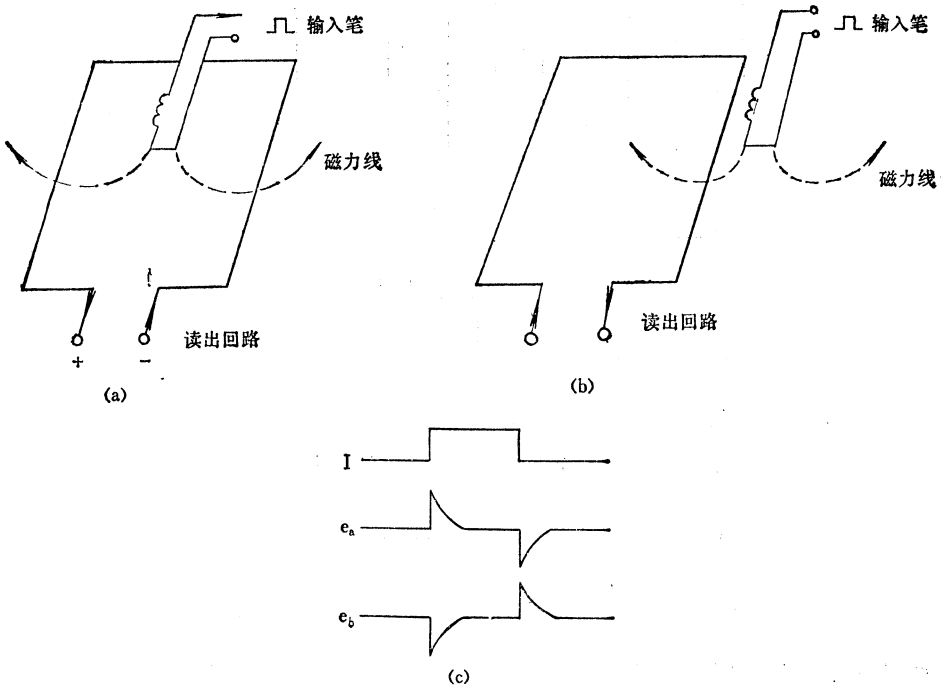
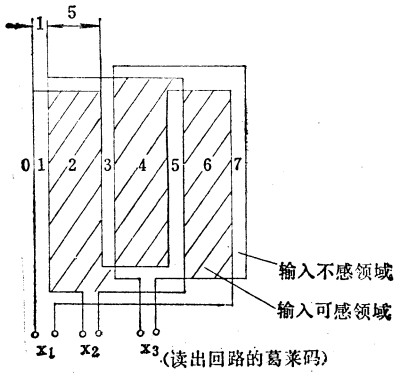


图5-4　笔的位置检测原理

（a） 笔尖在回路内侧；（b） 笔尖在回路外侧；（c） 波形。



区域 / 回路	0	1	2	3	4	5	6	7
x_1	0	1	1	0	0	1	1	0
x_2	0	0	1	1	1	1	0	0
x_3	0	0	0	0	1	1	1	1

图5-5　有线电磁感应式字盘原理

（二）无线电磁感应式字盘

上节所述的字盘，坐标盘与检测部分是分离的。因此，使用一根导线来传递检测信号。无线电磁感应式字盘则不同，在这种坐标盘的内部，有互相正交地排列成矩阵状的一组驱动回路和一组检测回路，如图 5-6 所示。

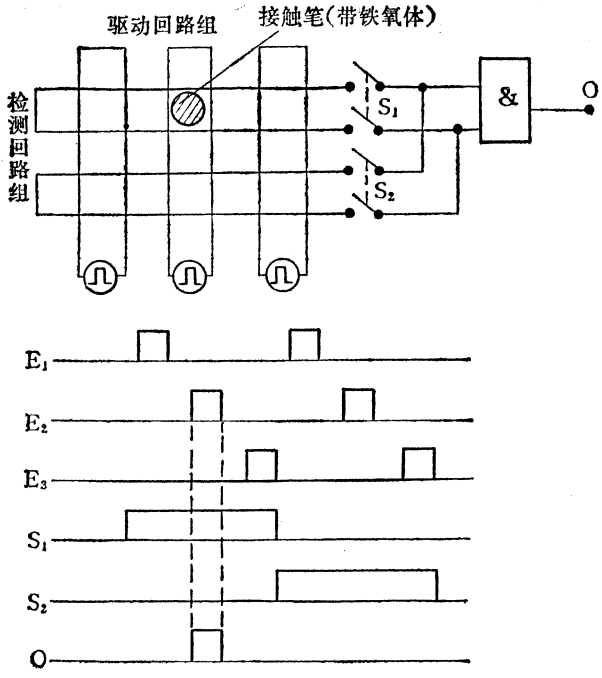


图5-6　无线电磁感应式字盘

驱动回路中加有时序脉冲驱动信号，对检测回路进行时序扫描，接触笔内装有铁氧体磁芯，当笔接触盘面时，笔所在位置的驱动回路与检测回路交叉部分的互感量增加，在检测回路一侧感应出电压，得到检测输出。这种坐标盘由于检测和驱动部分都在盘面上，因此，接触笔可以是无线的。通常，为了提高信噪比，在笔的铁氧体磁芯的外侧，装有带短路线圈的开关，在开关呈短路状态时，铁氧体的电磁效应被屏蔽，在笔动作时开关呈开路状态。

三、光电式字盘

在汉字字盘里安装一个与文字位置相对应的发光二极管矩阵。用光笔接触字盘，使光笔的碰障开关接通，即启动$X-Y$扫描计数器工作，对全部文字进行高速扫描。当扫到触笔所指的对应文字位置时，触笔内的光敏元件接收光信号，经光电变换后反馈到控制电路，使扫描停止，这时$X-Y$计数器的内容就可作为文字代码输出。因为检出的文字发光，所以眼睛还可以确认，起到校验输入的作用。

在光电式字盘中有一种新型的设备采用等离子显示技术。它是用等离子显示板代替了以往的发光二极管元件作为光源。它和发光二极管比较，制作容易，价格便宜。其构造见图 5-7 。

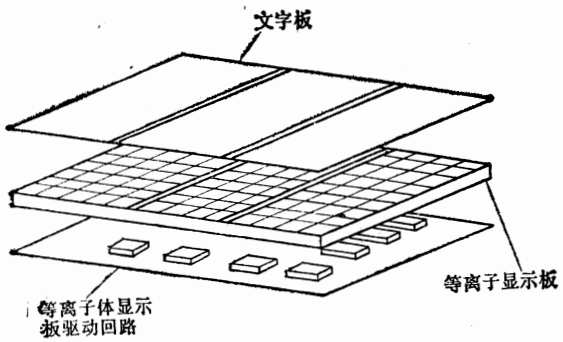


图5-7　等离子体显示盘式汉字字盘的构造

四、磁致伸缩● 式字盘

它的基本工作原理是将一张磁致伸缩板（例如电镀Fe-Co合金薄板）与一张印有等间隔导线的印刷板重叠，构成输入平面。如图 5-8 所示。

将磁致伸缩板固定在印制板上，在励磁线圈上加上脉冲，使发生磁致伸缩振动。所产生的磁致伸缩振动波以一定的速度向另一端传播，这时在印制板的导体下端上感应起相应电压，该电压的周期与导体折叠间隔 d 对应，把该电压波形放大、整形、微分后，得到了所示的与磁致伸缩板上的间隔 d 相对应的基准脉冲序列。另外，在接触笔的内部带有检测线圈，可以检测出传播脉冲。如果对励磁脉冲的加入时间与接触笔检出信号的时间间隔计数，就可以知道接触笔指示点的区域（图 5-8 的 0-8），再将计数值转成相对应的坐标编码。为了确定汉字在盘面上的平面坐标，还有一组 Y 方向的磁致伸缩板及激励线圈以及 Y 方向折叠的印制线。典型的磁致伸缩坐标盘的原理如图 5-9 所示。

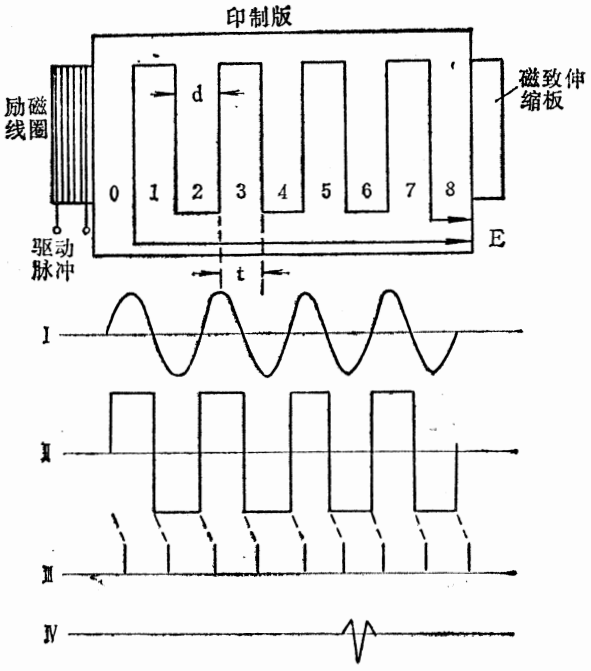


图5-8　磁致伸缩式字盘工作原理图

当在坐标盘的 X、Y 励磁线圈上分别加上激励脉冲时，用检测笔在矩阵网格上检测出 X、Y 方向的传播脉冲。在励磁脉冲发生的同时，启动二进制计数器，对时钟脉冲计数，检测出脉冲时计数停止，由内部的主控制器（微处理机或逻辑电路）处理，转换成相应的平面位置坐标的编码，经接口部分输出。

● 铁磁物质在外加同向磁场作用下，其外形发生变化的现象叫磁致伸缩效应。常见的铁磁物质有Fe、Co、Ni及其合金。

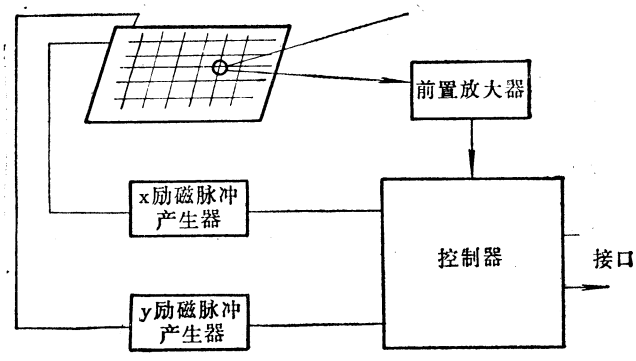


图5-9　磁致伸缩坐标盘的原理图

磁致伸缩式字盘在使用上要注意防止冲击振动，及避免靠近强的磁场环境工作。

五、压感式字盘

它的工作原理如下。在X-Y正交的电极板上，重叠上一层压感导电胶，盘上再加上一层保护板，最上面是文字盘。当接触笔在文字盘面上按下时，压感导电胶受到机械压力，按下部分呈导电状态。这部分导电胶经贯穿孔使X、Y电极相应部也呈短路状态。当在各个X电极上加上顺序电压时，对Y电极顺序检测，就可以检出触笔位置。由于指示笔的作用只是施加压力，故不需要特殊加工，而且坐标盘的构造也比较简单。主要的问题是选择比较合适的导电胶，能够具有比较好的压力与电阻值的变化特征。这样，不致发生要求输入压力过大，或过于灵敏而容易造成误输入等情况。其基本构成如图5-10所示。压感式字盘的特点是：键结构简单可靠，抗干扰力强，不需专门的检测笔，可用普通圆珠笔或手指触压输入。

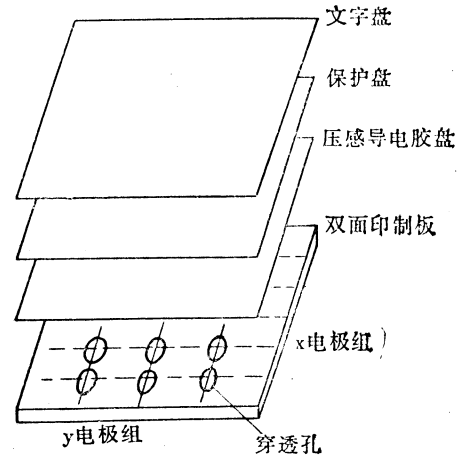


图5-10　压感式字盘示意图

我国生产的一种压感式字盘输入装置的主要技术指标如下：

文字数　3968个（62×64矩阵排列）。

盘面尺寸　650×540×55/135毫米^3。

文字大小　5×5毫米^2，14×7毫米^2，14×14毫米^2。

文字间隔　1毫米，2毫米。

接触压力　约85克。

六、书页式字盘

书页式字盘是在普通的笔触式字盘的基础上发展起来的，它的主要目的在于克服单个字盘收容字数的限制。如果把一般笔触式字盘换成多页式结构的文字盘。再加上对页的检出部分就可实现。页检出部分可以用光学式，也可以用与文字盘类似的页输入部分。

典型的书页式字盘如图5-11所示。

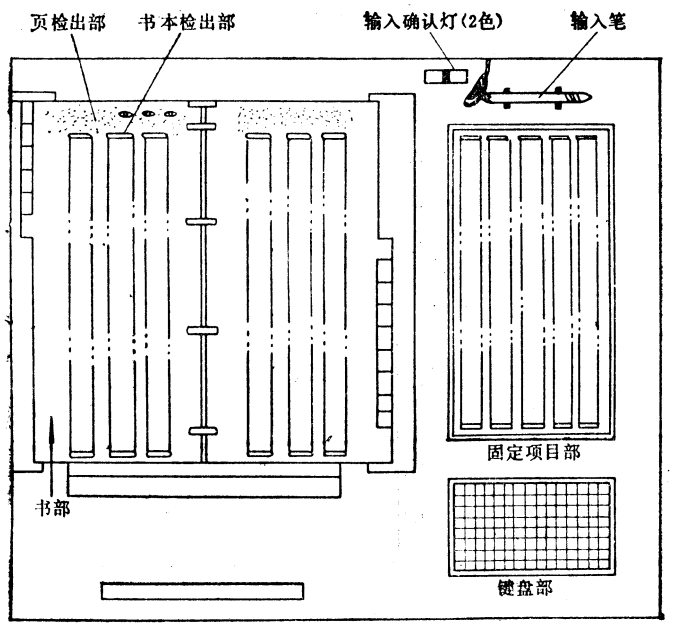


图5-11　书页式汉字字盘

该字盘上有 160 个（10×16矩阵）字符输入键和15个移位键，可产生总数为2400个字符编码。键的上面是按矩阵排列的字表。为扩大盘面收容字数，备有若干页字表，操作时只要按压相应的页号键，或在翻动页面时，自动形成的页面编码，就能很容易地选到所需要的汉字。这样 8 个表能产生 2400× 8 ＝19200 个字符代码。

书页式汉字字盘的特点是收容字数多，可达上万字或更多的字，而且可以根据需要，输入一些特定的词组或项目。

七、字盘主要参数的确定

字盘主要参数的设计包括两方面的内容，一是坐标盘主要物理及电气参数的确定，第二是字盘的文字盘面的设计。前者由于各种坐标盘所依据的物理效应不同，需要针对具体的对象进行讨论。一般说来，最主要的参数包括：扫描导线的线宽、间距、感应块面积；工作灵敏度；信噪比等等。由于篇幅所限，这里从略。而文字盘面的设计，由于它直接影响到字盘的使用性能，因此，从使用的角度来看，它是笔触式字盘最重要的设计问题。其中主要的设计参数包括：盘面尺寸；盘面字数；盘面排列；盘面表示方式；盘外字的输入方法等。当然这两方面的问题并不是截然分开的，它们有着相互联系和相互制约的关系。

（一）盘面尺寸

笔触式字盘的基本操作是找字和动笔，因此盘面的大小就直接影响到视力活动范围和手的活动范围，从而直接影响输入速度，因此，盘面不宜过大；但盘面过小，如果收容字量一定，文字尺寸就小，笔的定位就困难，影响输入速度，其间应有一个比较合适的选择范围，从当前实际使用的各种字盘来看，比较合适的盘面尺寸在 500×400 毫米^2

至 400×250 毫米²左右。

（二）盘面字量

盘面字量的确定主要是在找字时间和盘外字出现频度之间取得合理的折衷。盘面字越多，盘面字的排列就会越复杂。一方面操作员记忆和熟悉盘面排列的训练时间要加长，同时，前面已提到，盘面尺寸一定时，盘面字越多，字的尺寸越小，定位越困难，输入速度自然要下降。因此，减少盘面字量，输入速度就会提高，这一点对于非专业操作人员尤为明显。但是，当盘面文字字量减少时，盘外字（指键盘上未收容的汉字）的出现频度就会增高，要输入一个盘外字，需要几次击键动作，这样，就使总的输入时间增加。因此，在两者之间应有一个比较合理的选择，从目前使用的一般情况来看字盘总字量在3000字左右是比较合适的。从我国报刊出现的文字频度统计表明，3000个字约占累计总频度的 99.7%，这时盘外字的出现概率只有 0.3%。盘面收容字数与外字出现概率曲线如图5-12所示。

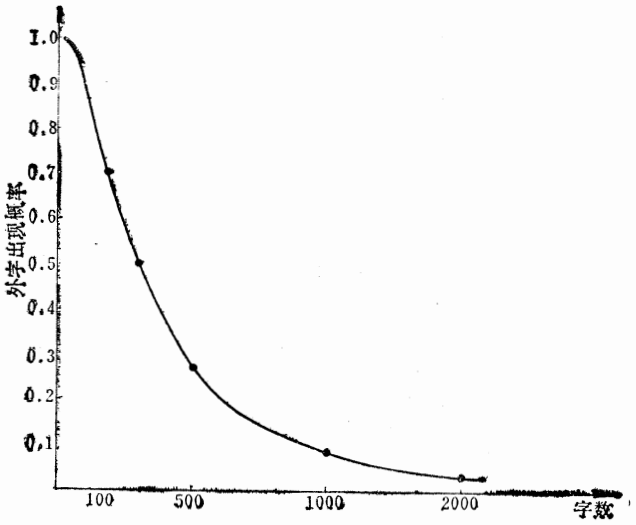


图5-12　盘面收容字量与盘外字出现概率曲线

（三）盘面排列

确定文字排列的两条基本原则是：（1）便于记忆；（2）笔的平均移动距离要小。目的也在于减少找字和笔的定位时间。目前，我国常用的字盘排列方式有：按汉语拼音字母音序排列；按中文打字机标准字表排列；按汉字的出现频度分级分区排列；按笔形笔顺排列等数种。在分级分区排列字盘的每个分区内又有按音序或笔形笔顺排列两种方式。但归结起来，可大致分成顺序排列和分级排列两大类。顺序排列方式适于非专业操作员，分级排列方式适于专业操作员。对于专业操作员，如果记住了各个字所在的分区，而且在分区内部把常用的搭配字安排在一起，会使输入速度显著提高。

（四）盘面表示方式

除了选择适当的字体外，最好用醒目的标志或颜色来标明每一分区汉字的属性，如按音序的首字母、不同的频度级别来区分等等。总之，应力求清晰、界限分明，但整个盘面又显得统一调和。

还应该指出，在选择或设计一种字盘输入装置时，除了上述的文字盘面的各种性能之外，还必须注意到它的词组键、功能键以及字符键，因为这对于简化输入操作，提高使用上的灵活性是非常重要的。我国设计的一种笔触式字盘盘面上键位的具体分配如下：

功能键　　64个

词组键

固定词组键	96个
活动词组键	32个
文字键	
常用汉字	648个
次常用汉字	1784个
西文、符号	176个
用户自定义字	128个
拚字部件	144个
总计	3072个

八、盘外字输入

上一节已经指出，由于键盘物理尺寸的限制，及其它操作上的原因，盘面收容字数一般不过3000字左右，它虽然能保证盘外字的出现概率足够低，但仍然是不能忽视的。特别是对于某些专业领域，盘外字输入可能是相当突出的问题。因此，在设计字盘式输入装置时，采取一种比较方便灵活的外字输入方法是十分必要的。目前比较常用的盘外字输入方法有：

1．汉字电报码输入法　利用数字键，按汉字电报码（四位数字），击键四次输入一个汉字。

2．字根拚字输入法　上面提到的一种笔触式字盘，在盘面上专门设置了144个组字“部件”，它们一般都不是字，而是组字的“部件”(或称字根)，例如“氵”、“纟”、“亻”等。如要输入盘外字“潼”，则需输入“氵”和“童”。这样就可以方便地解决盘外字的输入，大部分盘外字可以是二拚，少数是三拚，极少数是四拼、五拼。在拼字部件输入后，用分级检索，对分法查表，或利用杂凑寻址法等很容易查到对应的国标码。

3．汉字交换码输入法　用4个“0～F”十六进制代码键输入，输入一个汉字击键四次。另有清除键，在未到四次时发现击键错误可以用清除键清除后重新输入。

4．其它　如按照字典（或称词书）所给出的唯一性汉字编码输入。其过程是：查字典，找到相应字的代码，并据此代码击键输入。

上述几种外字输入方法中，第二种方法，在使用上得到了较好的效果。

5.1.3　中文打字机式汉字键盘

中文打字机在我国已有很长的历史。目前，已经普及到各个部门，使用量也很多。所以，设法利用这种装置来进行汉字输入，是有一定意义的。在中文打字机上加装发生代码信号的机构，则用同样的操作，就可以同时进行打字和纸带穿孔。这就成为中文打字机式的汉字输入装置。

中文打字机式汉字键盘的输入方式，其最大的特点是原来使用中文打字机的打字员可以不经特殊训练就成为键盘操作员。在国内中文打字机的生产和使用台数是很大的，远远超过其它汉字输入装置的数量，另外国内熟悉中文打字机的人员也很多，这些正是中文打字机式汉字输入装置的重要优点。

但是，中文打字机的机械活动部分特别多，所以可靠性差，而且输入速度不会太高。

根据所加装的代码发生器的不同，可分为几种不同的类型。

一、电磁感应板式

这种方式的原理同上述笔触式汉字字盘方式输入装置的原理相似。对应于字盘式输入装置的笔的部分是铅字选择联动机构。一按下印字手柄，就能够一面印出文字，一面从字盘联动的电磁感应式平板上检出位置坐标的信号。

二、全息照相编码式

将微全息照相排列在矩阵式的全息照相代码记录胶片上。在各微全息照相上都记录有文字代码。如图5-13所示。用指示器指定中文打字机的文字盘，并借缩放仪的作用，把激光引向全息照相记录胶片上的相应位置，透过全息照相的一部分激光，可以显示出这一代码，再经光检测器的变换，把它转成代码的电气信号。全息照相的优点是代码容易变更。这只要简单重写全息照片即可。

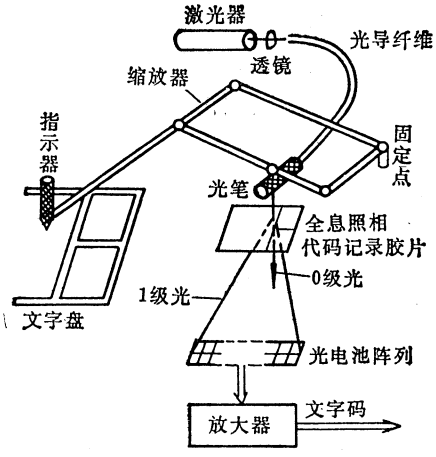


图5-13　全息照相编码式原理图

三、其它

在早期还曾使用过铅字代码式、条型码式及坐标输入式汉字输入装置，但目前已不多见，这里从略。

5.1.4　汉字字根键盘

上述汉字整字键盘，虽然具有直观和操作简便的优点，但是，尺寸一般较大，价格也较高。普及使用比较困难。因此，有人研制中等尺寸的用字根组合输入汉字键盘，简称字根式汉字键盘。

这种键盘盘面上的字根（包括部分整字）虽然不及汉字整字键盘的多，但数量仍多至近百个甚至上千个。它的基本思想是：最常用字一次输入，其它字则用字根组合方式输入。

字根键的数量视所用的拼字方法而定。如果一个汉字用较多的字根来拚写，总的字根键数就少，在极限情况下可以少到只用24种基本笔形。但是，很多汉字不仅笔划数多，而且纵横交错很难分辨。因此，对每个汉字分解的部分不能太多。否则，在分辨组合时会发生困难。此外，分解部分太多，编码就长，从而增加了输入一个汉字的按键次数，使输入速度降低，也降低了编码传输的效率。

我国研制的一种汉字字根键盘，规定每个汉字最多只能分成两部分。也就是说，输入一个汉字最多只能用两个字根，一次组合拼成。据统计，能输入6603个汉字的键盘，实用字根数为1411个。其中单个整字占20%（约1300字）再拼一次的字占72.5%，复字的占7.5%。这种键盘没有脱离整字键盘直观的概念，也保留了编码比较短的优点。

这种键盘设计，把汉字形体的拓扑图分成如下几种类型：

1．单字体：

□ 例字：王

对这类字一律不拆，整字一次输入。

2．合体字：

（1）左右字。有五种类型：

例字：明

可拆成：“日”“月”字

例字：湘

可拆成：“氵”“相”字

例字：绮

可拆成：“纟”、“奇”字

例字：邵

可拆成“召”、“阝”字

例子：韶

可拆成“音”、“召”字

规定：所有带边旁部首的字都把边旁部首作为一字根，其余为一个字根。还有一些字拆不开的，如班、辨等，这类字一律不拆。

（2）上下字。有三种类型：

例字：星

可拆成“日”、“生”字

例字：菩

可拆成“艹”、“音”字

例字：碧

可拆成“珀”、“石”字

规定：对于上下或左右字，若是能形成两个字根或两个字元的，则可以拆，否则不拆。

（3）包孕字。有七种类型：

例子：园

可拆成“口”、“元”字。


例字：同

不拆。


例字：凶

不拆。


例字：匡

不拆。


例字：反

这类字中只有“麻”和“病”为头部的两种可拆，其余的不拆。


例字：司

不拆。


例字：近

可拆成“斤”、“辶”字

有一种方案，将每个汉字用4个字根来组合，则字根数可降到六百多个。不过拼字结构的形式也相应增多（约30多种）。

基本字根650个（其中单体字500个），组成汉字时最多要用四个字根。补充字根234个（其中单体字134个），加用这些补充字根，每个汉字就可以不超过三个字根，只有36个字需要用4个字根。难拼字61个，共计945个，其中常用字为700个。如果再把繁体字、异体字根（200个）及日语汉字字根（45个）加入则可解决繁（异）体字和日语汉字的输入。

上面介绍的两种字根式汉字键盘，字根键数分别为1400和600个左右。为了进一步减少字根键数，一般采用字根合成法或某些特殊编码输入法。三点定字编码方案所用的键盘就是一种例子。

现行汉字通用部首为214个，因而可以认为汉字就是由这214个字根按照一定的规则编排、集“形”为字。在组成文字时，都必须遵循严格的位置规定而不允许位置发生混乱。统计表明，上述汉字的214个字根并不是完全相容的。大量的字根只会和某一定的字根相结合成字而不可能和另一些字根相结合。例如“巛”和“才”“水”和“钅”等等。这些互不相容的字根在编码上就可以合并而不会发生重码字的问题。三点定字编码方案将这214个字根归并为64个字根组合，平均约三个字根共用一个键，其中还可以包含64个常用整字。

三点定字编码方案字根键盘见图5-14。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
6			1	2	3	4	5	6	7	8	9	10			
5		74 0.0047	17 0.0073	56 0.0096	07 0.0128	35 0.0134		73 0.0035	36 0.0142	24 0.0120	06 0.0090	15 0.0041	61 0.0040		0.0946
4	50 0.0048	36 0.0049	51 0.0084	52 0.0113	13 0.0162	21 0.0178	16 0.0040	42 0.0034	30 0.0450	46 0.0207	40 0.0202	31 0.0098	45 0.0069	41 0.0013	0.1747
3	43 0.0049	63 0.0066	03 0.0089	75 0.0118	67 0.028	44 0.0473	14 0.0059	64 0.0057	00 0.0469	10 0.0305	60 0.0111	47 0.0074	33 0.0044	37 0.0052	0.2246
2	27 0.0020	34 0.0049	53 0.0096	01 0.0109	25 0.0243	04 0.0172	66 0.0038	20 0.0054	65 0.0194	12 0.0243	76 0.0078	55 0.0059	62 0.0053	71 0.0017	0.1425
1		70 0.0042	22 0.0077	05 0.0096	02 0.0151	02 0.0151	54 0.09995	54 0.09995	11 0.0158	72 0.0153	23 0.0082	57 0.0054	32 0.0029		0.0993 + 0.1999
	0.0117	0.00253	0.0419	0.0532	0.0964	0.1108	0.11365	0.11795	0.1413	0.1038	0.0562	0.0326	0.0235	0.0082	

图5-14　三点定字编码方案字根键盘示意图（图中每一键上部数字为该键名称或编码，下部数字为其出现概率）

有人设计了一种通用汉字键盘，能兼容多种编码方案。键盘具有90个字符键、8个定位键、10个十进制数字键、35个编辑功能键、2个定位备用键、11个备用功能字符扩充键、1个汉字分隔符号键和1个清除键，共158个。这种键盘是一个独立的输入设备，使用时不需要附加任何其它设备。

通用汉字键盘示意图见图5-15。

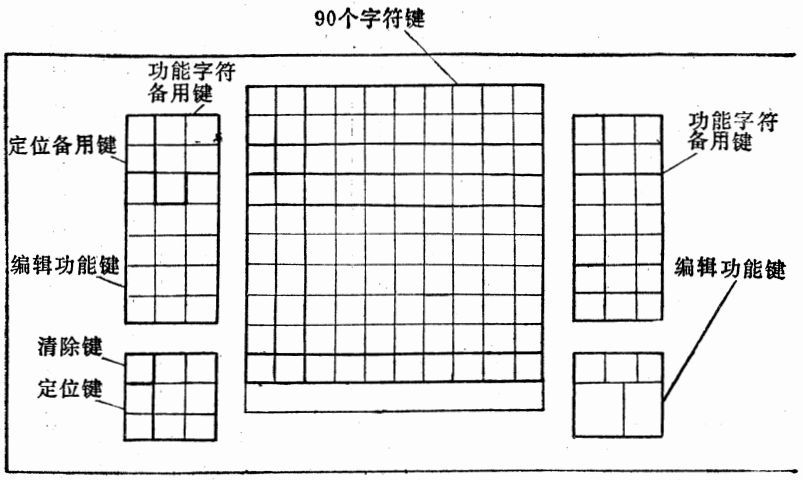


图5-15　通用汉字键盘面板布置图

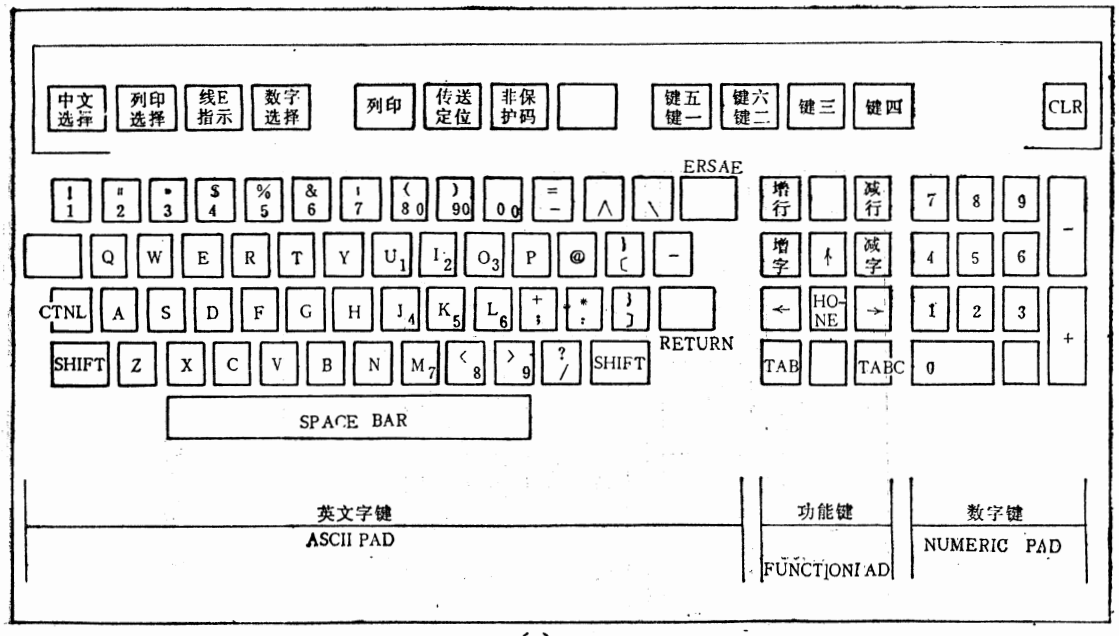
5.1.5　标准字母数字键盘

标准字母数字键盘指国际上通用QWERTY键盘。国内提出的汉字输入编码方案绝大多数都使用于这种键盘。这种键盘最主要的意义就在于它具有通用性，不要另外设计专门的汉字输入键盘，只要利用计算机本身配备的标准字母数字键盘就可以解决汉字的输入问题。此外，标准字母数字键盘还具有轻便、键位少、适于盲打、输入速度快等一系列优点。

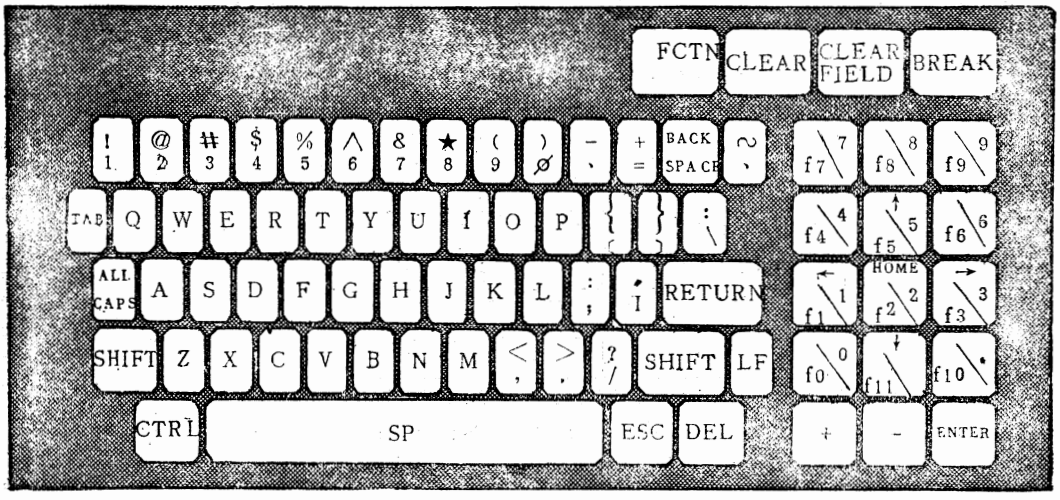
我国对于汉字编码的研究无论在方案的多样化和完善性方面都处于世界的领先地位。使用标准字母数字键盘，正好使这种优势得以充分的发挥。

随着微处理机及其应用的迅速发展，今后我国各个部门一定会越来越多地使用微型机系统。字母数字键盘作为一种小型的普及型汉字输入键盘，非常适合这方面的需要。

各种标准字母数字键盘都能用于汉字输入。但作为汉字信息处理系统配套用的键盘输入设备，比较适宜的是扩充型的标准字母数字键盘。图5-16是两种典型的扩充型字母数字键盘。它由三部分组成：（1）字母键；（2）数字键；（3）功能键。所谓扩充型就是指扩充了专门的数字键和功能键。使用标准字母数字键盘输入汉字可以采用多种方法，以下分别介绍几种较常用的方法所使用的标准字母数字键盘。



(a)



(b)

图5-16 扩充型字母数字键盘

(a) 之一；
(b) 之二；

TAB 制表键；
ALLCAPS 全大写；
SHIFT换挡键；
CTRL 控制键；
BACK SPACE 退格键；
RETURN 回车键；
DEL 删除字符键；
LF 换行键；
ESC 换码符；
FCTN 功能键；
CLEAR 清除键；
CLEAR FIELD 清除字段键；
SPACE 间隔键；
BREAK 中断键；
HOME 归O键；
ENTER 输入键；
fo～f11 12个功能键。

一、汉语拼音键盘

图 5-17 为一种标准 QWERTY 键盘经重新定义键名而成的汉语拼音输入键盘。

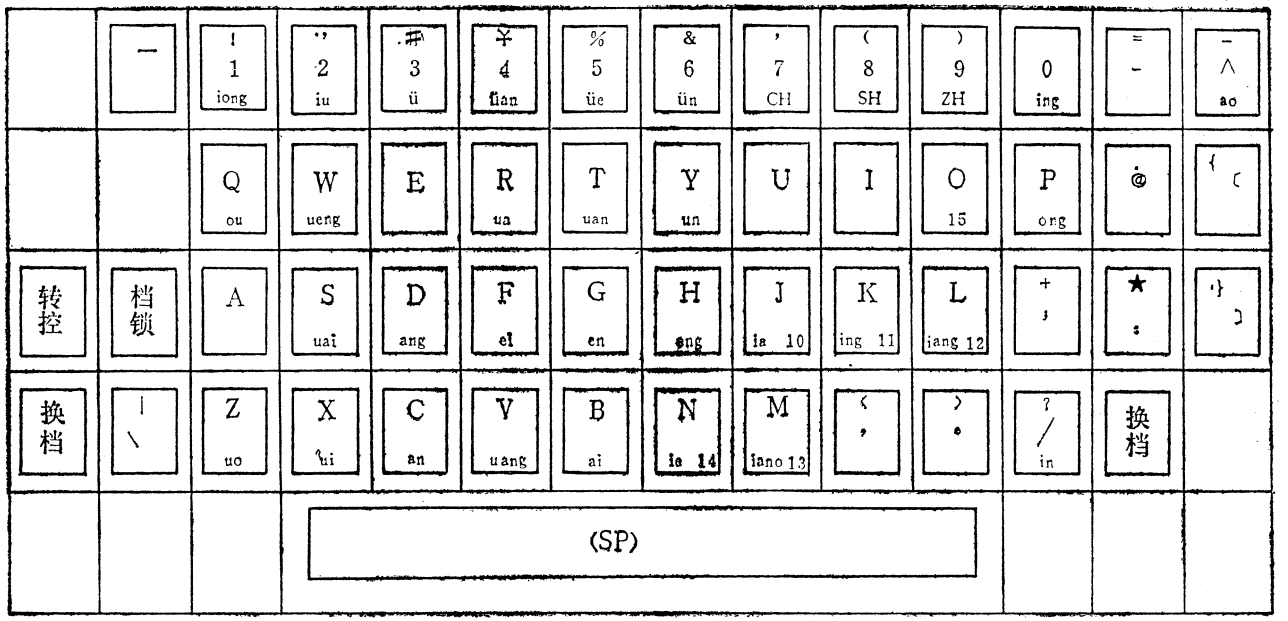


图5-17　标准字母数字键盘经重新定义键名后而成的汉语拼音输入键盘

有人设计一种汉语拼音键盘，以键盘的中心线为界把声母键和韵母键分别定义在两边，左手操作声母键，右手操作韵母键。

还有一种汉语拼音键盘（包括扩充型标准字母数字键盘）不经任何更动或重新定义，当用作拼音编码输入汉字时，虽然操作击键次数是增加了，但是完全按字母数字键盘击键指法操作，可以达到很高的输入速度。

音韵双拼编码方案是以汉语拼音为基础，只是把汉语拼音的声母和韵母各用一个字母表示，两个字母拼一个字音。击键时，对标准键盘不作任何更动，靠操作人员按编码规则直接输入。

二、汉字字根或笔形编码键盘

字根编码方案中，如果能把字根数压缩到用48个字母数字键能够表示时（一键也可定义若干个字根），就可以用标准字母数字键盘输入汉字。使用这种方法的标准字母数字键盘就叫做汉字字根或笔形编码键盘。

图5-18图5-19，是二种典型的字根式字母数字键盘。

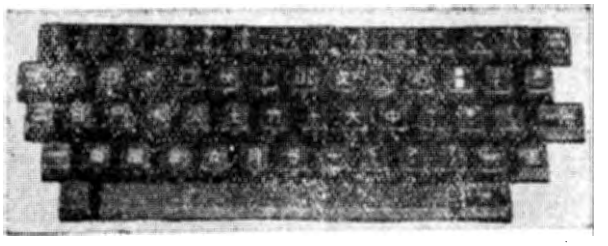


图5-18　字根式字母数字键盘（一）

石 12Q	白 34W	虫 25E	大 18R	口 23T	日 21Y	月 22U	女 19I	人 31O	言 42P
四 24A	一 11S	的 33D	手 17F	木 16G	火 44H	山 61J	王 13K	心 41L	
		了	金	土	水	草			
18Z	71X	51C	32Y	14B	43N	15W			
			间隔键 (代号"␣")						

图5-19　字根式字母数字键盘（二）

用字母数字键盘来实现笔形编码的输入方案，主要使用其中的数字键，另外对一些高频度的笔画组合需要定义一些多画键。

5.1.6 联想式人机对话汉字输入方法

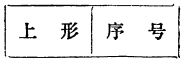
一、人机对话汉字输入方法

人机对话式汉字输入方法的基本原理是：用按键输入一个汉字的编码，但这种编码并非唯一性的，机器把具有这种编码的所有汉字，一次或分批地显示出来，由操作员进行挑选。这种方法的优点是编码规则简单明确，可确保一码一字，输入和校对同时进行，便于发现和纠正错误。它必须要借助于显示器、汉字库和实现人-机对话的控制设备，一般微型机汉字终端上具备这些功能。

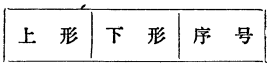
我国研制的联想式人机对话汉字输入方法在实现人机对话输入时将汉字分成为四类进行分级处理、各类汉字可采用如下所示的编码形式：

（1）最常用字（32个）

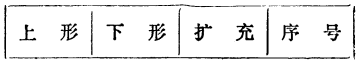
（2）基本字（600个）



（3）常用字（2500～3500个）



（4）扩充字



输入时可用光笔或键盘进行操作，机器则通过显示屏进行问答。整个显示屏从上到下分为三个区。上面三分之二是文件区，下面三分之一是选择区和上下形键盘区，它们各占两行。使用键盘操作时，屏幕上最下面两行的上下形键盘可以不要。

文件区显示输入、输出的内容，可用手动控制，将其内容往上、下、左、右各个方向任意移动。

键盘区开始显示上形，选取了上形后，键盘自动转换为下形；选取了下形或选择区的内容后，又自动转换为上形。

选择区经常显示的是32个最常用字，用光笔在所需的字上一点，或用键盘按序号键，就可将该字输入，它的字形显示在文件区中光标指示的位置。

如果所需的字不在最常用的32个字以内，则用光笔或键盘选取该字的上形，选择区将出现具有该上形的一组字（基本字），从它们中间选取所需字的方法与选取最常用字一样。该字输入后，选择区恢复为最常用字。

若按上形键仍未找到所需要的字，则需要按该字的下形键（或用光笔点），这时，显示区显示出具有指定上、下形的一组字（不包括扩充字），每组平均10个字左右，最多的20多字，选取它们的方式同上。

按了上、下形键后仍未出现所需要的字，则可按“扩充字键”，机器将把扩充字中具有指定上、下形的字调出来，显示在选择区，供人挑选。

二、联想记忆

在设计编码时，有意识地使编码形式适应汉字出现频度不等的规律，这样有效按键率会有较大提高。

进而还可以发现，在现代汉语中，大多数汉字不是孤立地出现的。它们一般是组成某种固定的文法结构。因此，汉字的出现概率是有条件的，即随前面已经出现的字的不同而具有不同的条件概率。我们可以进一步利用语言中的这种性质提高编码效率。

通过对汉语词组的分析可以看出，以某一个字为词头的词组，多的有几十个，少的只有几个。这意味着，一个汉字出现后，经常跟着它出现的字并不多。这少数字占了在这种条件下汉字出现概率的很大部分。“联想记忆”的方法正是利用这一特性，在输入一个字后机器会自动地“联想”起和它相关的字（称为联想字），并把它们显示在选择区，可以不考虑它们的上、下形就直接选取了。经过一次联想选中了所需的字后，又可由它引起另一组联想字。这种逐次联想的办法，可以使大多数常用词组，甚至于一句话，都一连串地被引出来。因此，可以迅速地输入常用的词组和句子，在整个联想式输入过程中，输入每一个字都只需按键一次。

联想记忆的编码方法，可以看作一种“自适应”编码。即根据前面出现的字，不断地自动调整最常用字（在选择区显示这种在具体条件下出现的最常用字）以达到最佳的编码效率。联想结构本身也可以在使用过程中，根据对实用信息的统计分析，自动地进行修改，使系统在一定程度上具有“自组织”的功能。

除了词组等汉语的固有结构外，讨论某个具体问题的文章，有关的字和词出现的次数往往较高。而文章作者个人惯用的字和词也会经常出现。怎样使系统也能自动适应这种变化呢？

分析一下显示屏的内容就可以发现，文件区显示的汉字正好反映了每个人和每篇文章的特殊性。当前要输入的字或词，往往前面已经输入过。并且很可能还留在文件区中。在用光笔操作时，可以将电路设计得使得文件区和下面的选择区毫无差别。光笔点到哪个字，就把哪个字输入机器，并同时在文件区光标指出的位置上显示出来，有时十个、八个字，用光笔一画就完成了输入操作。例如要输入“我为人人，人人为我”这样一句话，输入了前一半，后一半完全可以利用文件区进行输入。这将使输入速度进一步提高。

5.1.7 键盘控制器

一个完整的键盘控制器至少应由三部分组成，即一、扫描控制电路，二、译码电路，三、接口。下面分别加以说明。

一、键盘扫描控制原理

键盘扫描的基本目的在于确定被按下键的位置，下面以字母数字键为例来说明。

一般终端设备所使用的字母数字键盘有两种基本类型，即编码键盘和非编码键盘。

（一）编码键盘

它具有必要的硬件（或 LSI 芯片），能检测出键盘矩阵中被按下的键，并输出该键所对应的代码，同时，它还能产生相应的选通脉冲，使所连接的微处理机产生外设中断，启动相应的中断处理程序。我国生产的一种字母数字键盘采用 SMC 公司的 KR2376 扫描控制芯片，它是一种 MOS 电路芯片，具有 40 针双列直插结构，其原理如图 5-20 所示。

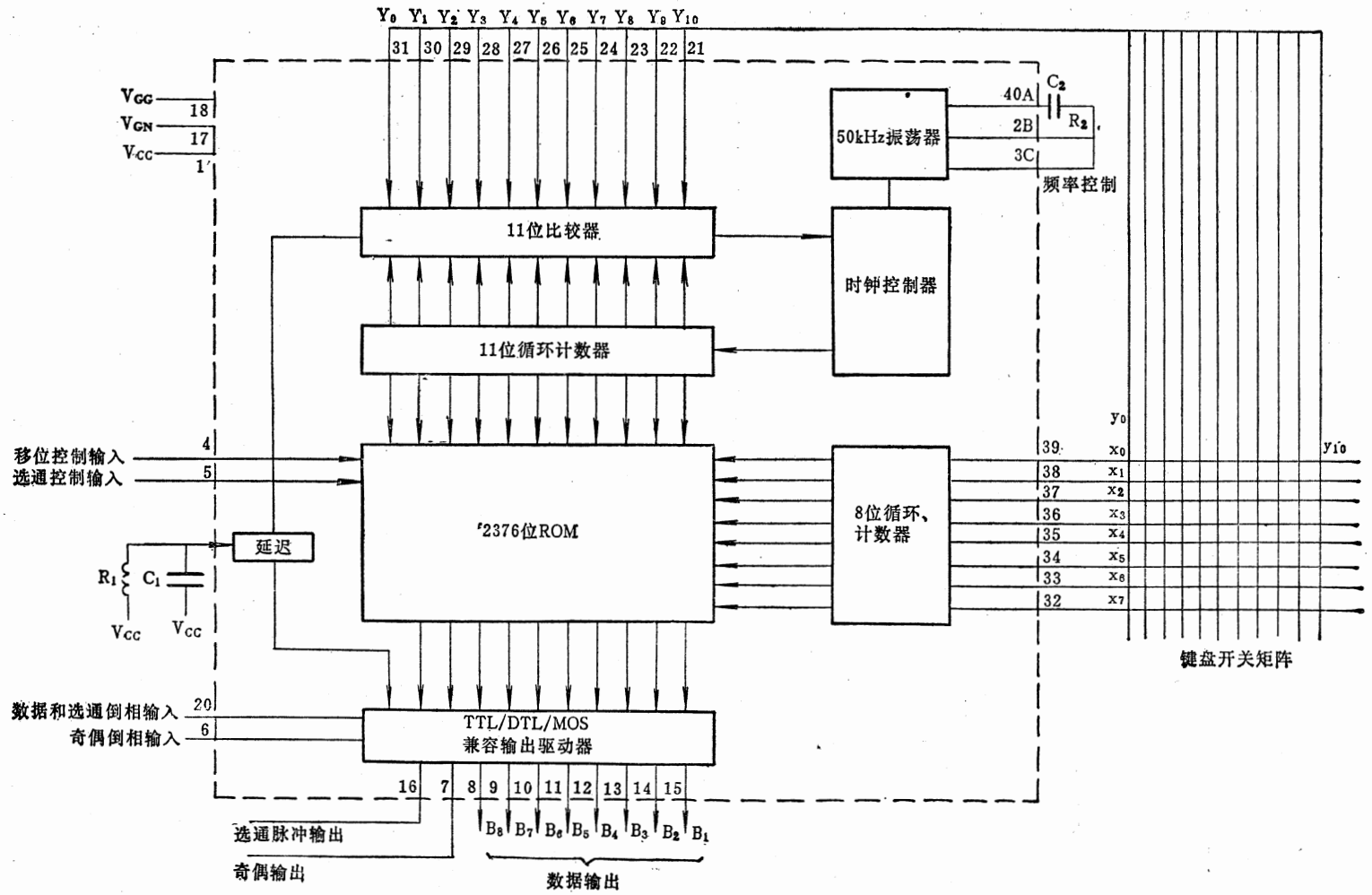


图5-20　KR2376键盘编码器原理框图

其工作原理如下：8级环形计数器依次对8根X方向扫描线发选通信号，X与Y扫描线之间的交叉部位有按键开关，当某个键闭合时，在相应的Y线上便产生输出，加到11位比较器的输入端，与Y方向的11级扫描计数器的输出比较，当两者符合时，即确定了相应的X、Y按键矩阵的闭合点，此时产生选通信号，并输出ROM编码器产生的8位键盘编码信号。这种键盘编码芯片适用于有88个键的键盘，加上移位控制功能共可产生256种不同的代码。

此外，比较常用的键盘编码器还有Intel公司的8279等。

当然，也可以用TTL集成电路及EPROM来实现键盘编码，其工作原理与上述的KR2376十分相近，这里就不再赘述。

（二）非编码键盘

它只提供键盘定位所必需的行列矩阵，扫描控制必须由用户附加的硬件及软件来完成。通常的做法是采取并行接口芯片如PIO（Zilog公司），PIA（Motorola公司）、8255（Intel公司）与键盘的行列线连接，用软件编程技术来实现扫描，常用的编程方法有行扫描法和线路反转法两种。

1．行扫描法

下面用4 × 4键盘矩阵为例来说明（见图5-21）。并行接口芯片输出依次对各行扫描，每扫一行，检测各列线是否有输出，若有，即可确定某行某列的键被按下。

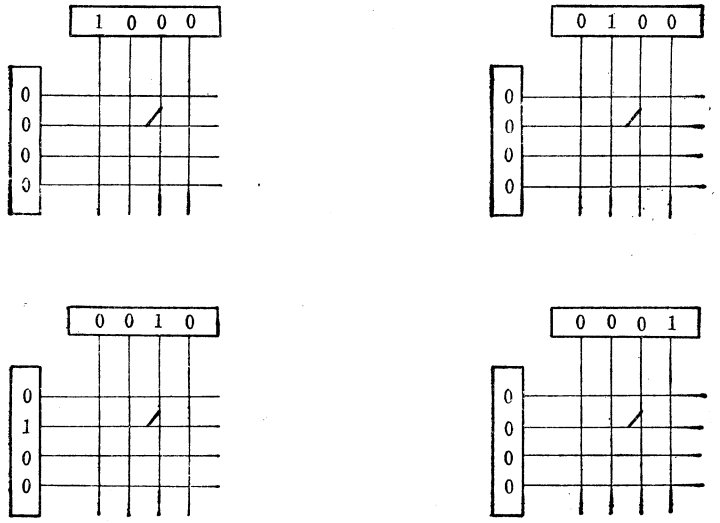


图5-21　行扫描法原理图

2．线路反转法

图5-22所示为线路反转法原理图。它仍以4 × 4键盘矩阵为例来说明。

由程序定义并行接口芯片中的一个通道（通常是8根线）内的4根线为输入线，4根线为输出线。如定义D_0～D_3为输入线，D_4～D_7为输出线，这只要在芯片的数据方向寄存器中写入“000011 11”即可实现。如果把芯片中的数据寄存器的初值全置为“0”，则D_4～D_7的输出为“0”。当按下某个键时，相应的行线也变成“0”。图中“0”在D_2线上出现，D_0、D_1、D_3因无键闭合，仍“1”，即数据寄存器的高四位的值为“1011”

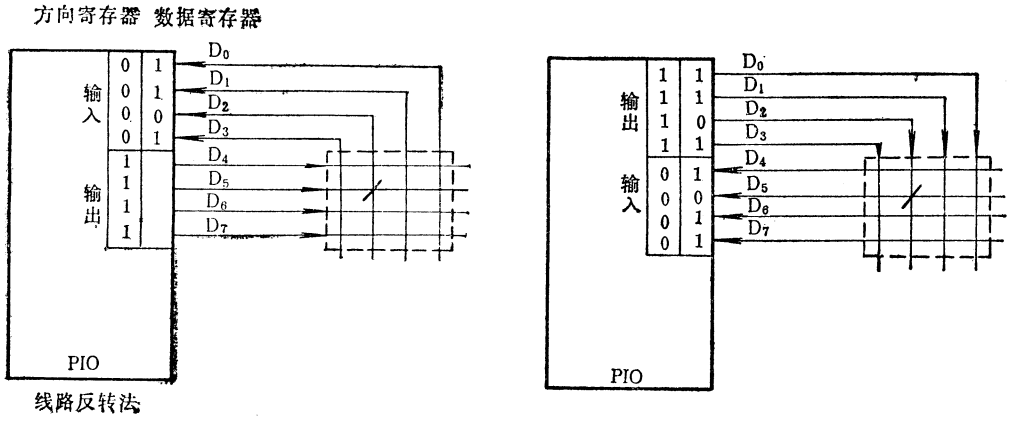


图5-22 线路反转法原理图

只是确定了键所在的列，要确定键的行位置，可以通过重写数据方面寄存器来实现。这时在数据方向寄存器中写入“1111 0000”，并由 D_4～D_7 作为输入线，则数据寄存器低四位的值为“1101”。结果数据寄存器中最后得到的值是“1101 1011”，由此，可以确定键盘被按下的键位于第 3 行第 2 列再经过简单的转换，即可得到与该键对应的代码，类似的原理也适用于更多行列的键盘扫描。

在汉字整字键盘中，通常采用X、Y两度平面坐标译码来实现多键扫描。这种译码方式对 4096 键矩阵的对外接口线需要 64＋64＝128 线。为了减少接口线的数目，有人设计了X、Y、Z三度译码技术，这样 4096 键矩阵的对外接口线只需 要 16＋16＋16＝48 线，从而使电器简化。

二、编码电路

键盘编码器的基本作用是确认被按下的键的位置，并提供与该键位置相对应的代码。目前使用的主要有两种编码器：静态编码器和扫描编码器。静态编码器单纯产生与键对应的代码。例如，一个具有 64 键的线性键盘，如每个键按下时接通相应的导线，因此，某个键被按下时，对应的线上就出现脉冲信号，并将该键的位置简单地变换成相应的 8 位编码（实际并未用足 8 位信息）。这就意味着 64 根单独的输入线 产 生 64 个 8 位代码中的一个代码。为了简化电路，大多数键盘是按矩阵方式排列。这时 8 × 8 键矩阵只需要 16 根线，这时需要用对矩阵的行列扫描来确定按键位置。即使用扫描编码器。

大多数键盘编码器都是用 ROM 或 EPROM 芯片来实现的。一般字母数字键盘的扫描控制芯片（如前面提到的KR2376）中便含有 ROM 编码器。

三、接口

目前，不少的字母数字键盘除了含有扫描控制芯片外，还配有微处理机，可向外部提供标准的串行或并行接口，并附有字符校验等功能。

当前所使用的键盘接口电路并没有完全标准化，在使用时，接口电路的设计要针对具体的情况来处理。但一般说来接口不外乎两种方式，一种是并行方式，一种是串行方式。

并行接口　编码型键盘（或字盘）通常具有并行接口，接口信号至少应包含选通信号和数据信号，此外，还包含有数据校验信号线、电源线、地线等。这种类型的接口一般可以用并行接口芯片来实现。用 PIO 芯片实现的用 KR2376 芯片作为键盘编 码 器接

口的典型程序框图如图 5-23 所示。

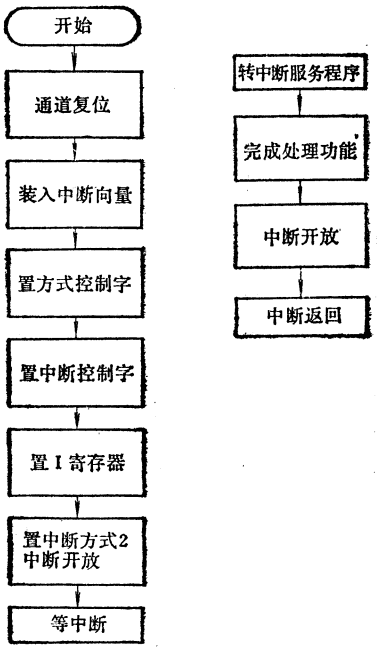
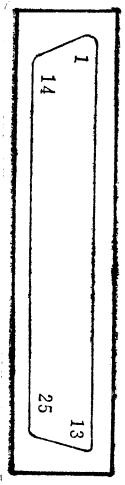


图5-23　字母数字键盘并行接口程序框图

串行接口：目前使用的串行接口，大都是 RS-232-C 接口。一种笔触式汉字字盘的串行接口线如图 5-24 所示。



针号	信号	针号	信号
1	GND		
2	T_xD	14	
3	R_xD	15	
4	RTS	16	
5		17	
6		18	
7	GND	19	
8		20	DTR
9		21	
10		22	
11	T_xD(TTL)	23	
12		24	
13		25	

图5-24　一种笔触式键盘的串行接口线（RS-232-C接口线）

实现串行接口可以使用各种串行接口芯片，如 8251（Intel 公司）、SIO、DART（Zilog 公司）、ACIA（Motorola 公司）等。汉字终端采用 SIO 芯片与键盘实现串行接口的典型程序框图如图 5-25 所示。

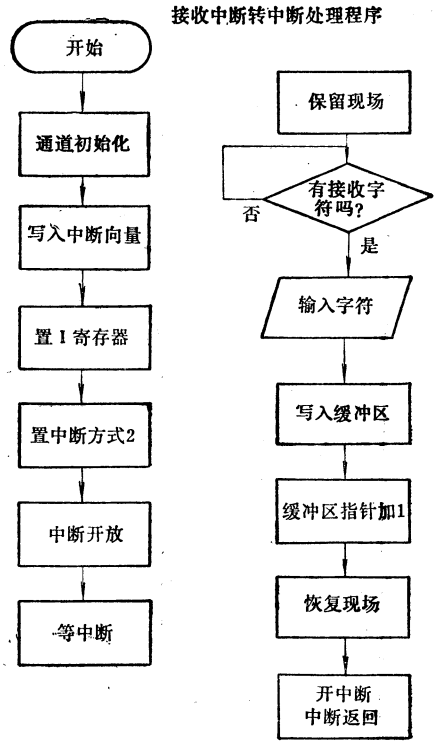


图5-25　串行接口程序框图

四、笔触式字盘控制器

上面所说的键盘控制器的基本原理，对于笔触式字盘是完全适用的。为了使读者对笔触式字盘的控制器的作用原理有一个完整的概念，下面用一种国产的静电耦合式字盘的控制器作为一个实例来说明。控制器的原理如图 5-26 所示。

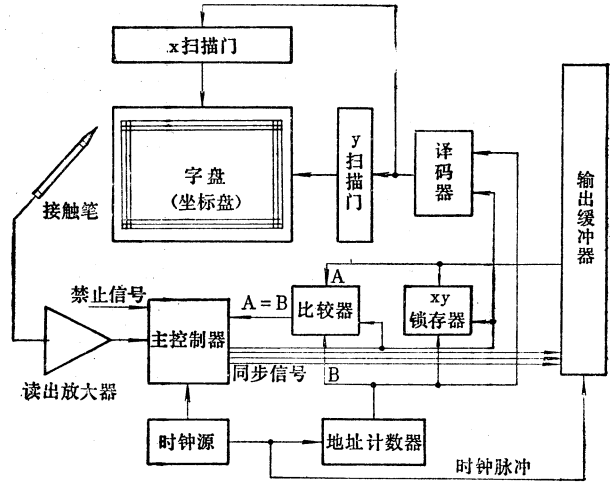


图5-26　字盘控制器框图

正交的X、Y扫描线在字盘上形成 3072 个字位检测点，当接触笔接触这些点上的文字时，启动笔内开关，扫描计数器开始计数，通过译码器译码，依次选通X扫描门和Y扫描门，使X扫描线和Y扫描线都分别从头到尾扫两次。当扫到对应于接触笔位置的X、Y扫描线时，便通过接触笔与扫描线的耦合关系把扫描线上的信号回授给主控制器。在整个扫描过程中，主控制器将先后得到两个X耦合信号和两个Y耦合信号。每次得到这些信号时，扫描计数器状态即为接触笔所接触文字的位置代码。在得到第一个X耦合信号时，主控制器将扫描计数器状态送到X锁存器中锁存，以便和出现第二个X耦合信号

时的扫描计数器状态相比较，比较符合则表明所需文字的X位置代码正确，否则将发出告警信号；对Y方向扫描也作同样的比较，以确定所需文字的Y方向位置代码的正确性。此后，设备发出数据选通脉冲，将锁存器中的X、Y位置代码作为文字数据送至汉字终端系统。

某些汉字字盘控制器，还可以将X、Y位置代码转换成对应的汉字代码(如国标码)，这通常是用译码电路来实现的。图5-27(*a*)示出了一种键盘译码及接口电路框图。

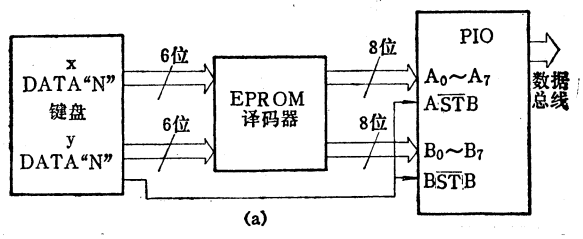


图5-27(a) 键盘译码及接口电路框图

图中的EPROM组成译码电路，它将键盘的X、Y坐标位置码直接译成所定义的汉字国标码。键盘与汉字终端的连接，是通过一个可编程的并行接口芯片来完成的。

某些键盘内部用微处理机作控制器，但其基本原理与上述情况是十分相近的，一种实用的微处理机控制的键盘编码及接口电路如图5-27(*b*)所示。

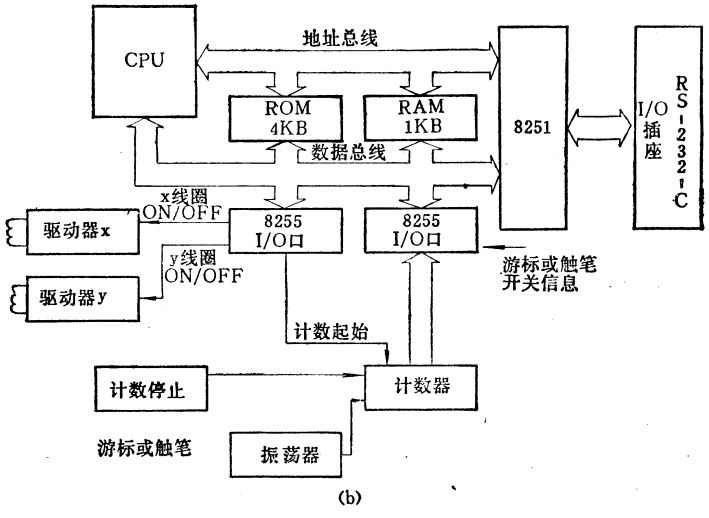


图5-27(b) 微处理机键盘控制器框图

5.1.8 键盘式汉字输入设备的操作性能

定量地评定各种键盘式汉字输入设备的操作性能是非常必要的。通常用输入速度和学习曲线作为评定的标准。前者，反映出不同输入设备的输入速度的统计特性，后者反映出操作人员的操作水平随时间的变化。

一、输入速度和学习曲线

表5-2是对10个操作人员在各种键盘上进行输入操作的实验成绩。包括在各种键

盘上的实验操作次数和最后一次实验的平均输入时间。它反映各种键盘的输入速度。图 5-28 是各种键盘式输入设备的学习曲线。

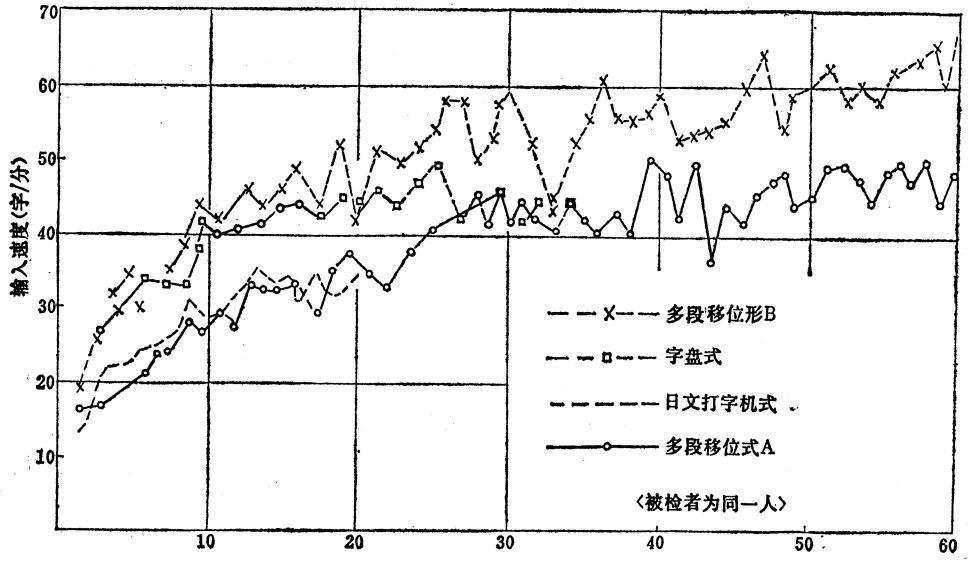


图5-28　各种键盘式输入设备的学习曲线

从表 5-2 和图 5-28 可以得出以下结论：

（1）以盲打为基础的标准字母数字键盘输入速度最高。

（2）多段移位式 B 的输入速度仅次于标准字母数字键盘的输入速度。多段移位式 A 不仅熟练操作需要时间长，而且最高输入速度也比 B 低，不过，比其它输入设备还是快一些。多段移位式 B 的盘面字不是以汉字频度顺序排列的；多段移位式 A 的盘面字是按汉字频度顺序排列的。

（3）笔触式字盘的熟练操作需要时间短，但是输入速度比多段移位式的低。

（4）中文打字机的输入速度最低。

表5-2　实验次数和输入时间（秒/字）

被验者	多段移位式A		多段移位式B		笔触式(字盘式)		字母数字键盘	
	次数	汉字	次数	汉字	次数	汉字	次数	汉字
A	74	1.8	60	1.3	49	1.7	145	0.4
B	67	2.7			53	2.0	180	0.5
C	74	1.6	60	1.1	45	1.9	145	0.5
D	74	1.8	30	1.4	42	2.1	75	0.6
E	76	1.7	60	1.0	48	1.8	180	0.3
F	32	2.7			53	2.5		
G					7	3.8	72	0.7
H					32	2.0	27	0.9
I	61	1.4			45	2.0		
J					33	3.0	55	0.8

二、对各种输入装置的操作性能简要分析

在笔触式汉字字盘上的输入操作，很象拿笔在稿纸上写字一样，但是，要达到熟练

并有高的输入速度，是不太容易的。被试验者用笔触式字盘输入设备，每分钟可输入约45字，比多段移位式装置要低。

一般被验者都认为使用字盘式输入装置容易疲劳。为了进一步验证被验者的这种印象，进行了CFF（Critical Flicker Frquency——临界闪烁频率）测验。即让被验者看一亮一灭的光点来测定被验者判别灯光闪亮的临界频率。频率降低就说明疲劳度增大。

图5-29所示的是一个被验者的CFF测定结果。由此，也可以说明字盘式装置比多段移位式装置更容易使操作员疲劳。

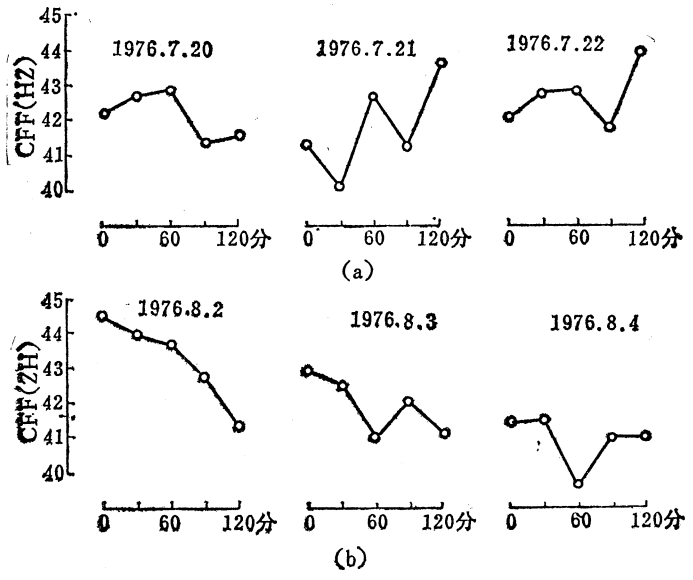


图5-29 输入操作的CFF（闪烁频率）试验

（a） 多段移位式输入装置；（b） 下盘式输入装置。

标准字母数字键盘最大输入速度每分钟近200字符。但是，要熟练地掌握这种键盘的输入方法要比别的装置难得多。一般要进行较长时间的专门训练，才能运用自如地进行操作。而对于汉字整字键盘，很多人只经过较短操作试验便可达到相当于职业操作员的水平。其输入速度不会比标准字母数字键盘低一倍。

操作标准字母数字键盘时有严格的指法，手指按照一定的节奏动作，而其它装置没有这种节奏感。

同时，由于标准字母数字键盘比各种汉字整字键盘尺寸要小，操作员手腕的移动范围也小，因此，在熟练掌握后可以得到高的输入速度。

5.2 汉字语音输入方法

在汉字信息处理系统中，除了汉字的键盘输入外，还有语音输入，它主要是利用产生声音的物理模型，通过语音分析手段，预先将一些语音的特征参量提取出来并储存在处理系统中。当语音信号输入时，处理系统根据对该信号所提取的特征参量和所储存的参考特征参量进行对比，通过逻辑判断方法或“距离”测量方法，对语音进行识别辨认。

这一节首先根据声音产生过程来说明语音特征参量的物理意义，然后介绍一下语音识别过程和几个语音识别系统。

5.2.1 声音产生的基本物理原理

图5-30为发音器官简图。声带由粘膜和肌肉组成，有一定厚度，左右各一片。声带之间的缝隙称为声门。在正常呼吸不发音时，声带处于平衡位置，声门是开的。声道由咽腔和口腔组成，它可以看作一端是口唇，另一端为声带的声管。在发音期间，由于发音器官（如舌，下巴颚和口唇等部位）的运动使声道形状连续变化。通常情况（即非鼻音）下，声音由口唇辐射，没有声音从鼻孔辐射出去。但在发鼻辅音时，声道的前端是封闭的，软腭使鼻腔与声道耦合，声音从鼻孔辐射出来。

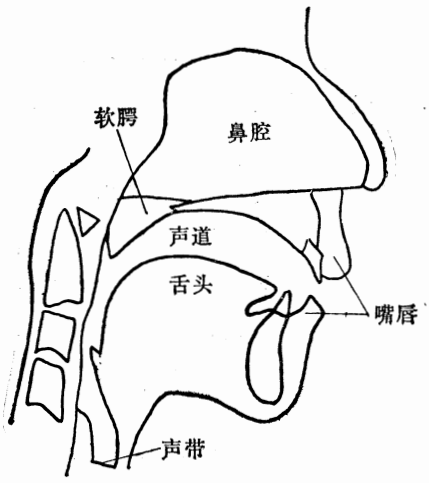


图5-30 发音器官简图

按激励方式可以把声音分成三类。

（一）浊音

浊音从肺部送出一股压缩空气迫使声带振动，产生一系列准周期空气脉冲去激励声道，从而产生浊音。浊音包括元音、半元音、浊辅音。声带振动的基频由声带的质量、张力以及通过声门的大气压决定。在发音期间，声带张力和声门大气压都是变化的。一般成年人说话的基频范围约60～400赫。

（二）清音

在声道某处形成一个收窄点，当高速空气通过这个收窄点时便产生干扰，形成一个宽带噪声源激励声道，从而产生清音。

（三）爆破音

先在封闭口腔中建立一定的大气压，然后突然释放从而产生爆破音。例如塞辅音。

在发音期间，由于激励源的不同以及声道形状和面积的不断变化，形成了时变语言信号。同时，声道的共振在语音的动态频谱上出现一系列共振峰，它们的位置和频谱的包络形状是由声道的大小和形状所决定的。

上述发音的生理过程可以用图5-31的语言产生简化模型来描述。图中，线性滤波器代表声道及口唇辐射；周期脉冲源或白噪声源为激励源；G是增益参数。显然滤波器的输出就是语音信号$S(t)$，它等于线性时变滤波器脉冲响应和激励函数的卷积，即

$$S(t)=\int_{-\infty}^{t} u(\tau)\cdot v(t,\ \tau)d\tau \tag{5.1}$$

上式中$u(\tau)$为激励函数；$v(t,\ \tau)$是线性滤波器的脉冲响应。通常声道形状的变化是缓慢的，因此，语音的产生过程可看作为准稳态过程，在10～20毫秒时间内，激励源和线性滤波器基本上保持不变。对于浊音，图5-31中的激励源采用周期脉冲源，其基频为声带振动的周期。对于清音，激励源采用无规白噪声源，清音频谱完全由声道

响应决定。通常高频能量较大，它们的频谱见图5-32。

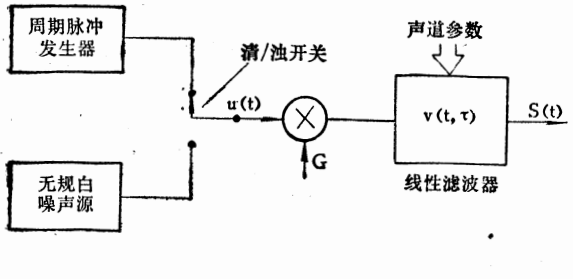


图5-31　语音产生的简化模型

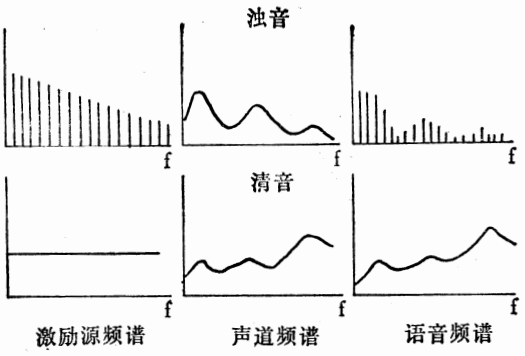


图5-32　清音和浊音的频谱

图5-31的线性滤波器是由声道的面积和形状决定的，它的数学表达式能从管中声传播理论推导出。为便于数学描述，声道可简单地看作由M个长度相等、面积不同的声管联接而成，每一节都有均匀的面积，其尺寸比波长小得多（见图5-33）。声波以平面波方式在管中传播，在某些频率出现共振，根据管中声传播的理论推导可知，声管等效于一个全极点滤波器，声管共振相当于系统传输函数的极点，因而，图5-31的线性滤波器可用全极点滤波器表示，它的转移函数$V(z)$有下面的形式：

$$V(z)=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{5.2}$$

式中，复变量$z=\gamma e^{j2\pi f}$；f为频率；p为数字滤波器的阶；a_k和G分别为数字滤波器的系数和增益，它们的数值随着语音的不同而变化，由于语音信号是准稳态过程，一般取20毫秒～30毫秒作为分析帧，因而每帧内a_k和G都是恒定的。这些参数都能通过信号处理技术直接从语音信号中逐帧求得。图5-33为七个等长管的级联图。

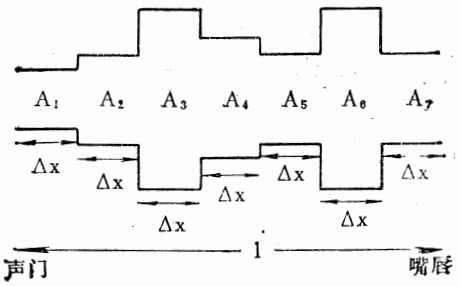


图5-33　七个等长声管的级联图

5.2.2　语音识别的特征参量

目前的语音识别系统都有许多不同的规定。其中，有连续语音的识别和单字的声音识别，以及指定讲话人的识别和不限定讲话人的识别等等。但不管何种语音识别系统，都必须首先用数字信号处理技术提取语音的特征参量。识别所采用的特征参量很多，但都是以上面描述的产生声音的物理模型为基础的。这一节将简要介绍语音识别中常用的几个特征参量。

一、频谱参量

语音频谱中包含着语音的重要信息，它显示了激励源的性质和声道的频率响应，如图5-34所示。从它的动态频谱中使人一目了然地看出共振峰、基频随时间的变化。因此，

语音的频谱(即语图),一直是研究语音的重要工具,也是最早用于识别语音的特征参量。它已成功地用于某些限定词、句、指定说话人的语音识别系统中。

语音频谱和语音时间信号一样是时变的,但在短时间内可以看作恒定不变,所以它能从短时信号的傅里叶分析中计算出。根据数字信号处理的理论,要使离散傅氏分析得到正确的结果,就得满足如下条件:其一,满足抽样定理,即抽样频率 f_s 必须是信号的最高频率分量的两倍以上;其二,短时信号不能从语音信号中简单截取,即截取函数不能采用矩形傅氏函数。因为时域截断的效果就相当于频域上的卷积,矩形函数在频域中的旁瓣很大,使傅氏分析产生较大误差,因而必须采用旁瓣很小的傅氏函数例如汉明商,它的数学表示式如下:

$$W(m)=\begin{cases}0.54-0.46\cos(2\pi m/N_0-1) & 0\leqslant m\leqslant N_0-1\\ 0 & 0>m,\ m>N_0-1\end{cases}\tag{5.3}$$

假设语音信号为 $S(n)$,每一分析帧取 N_0 个样点,则第 n 帧的信号表示为 $S(m+n)$,$m=0,\ 1,\ \cdots,\ N_0-1$,它的短时频谱可用下面的富氏分析计算。

$$S_n(e^{j\omega})=\sum_{m=0}^{N_0-1}W(m)S(m+n)e^{j\omega m}\tag{5.4}$$

如对连续语音信号截取不同时间的信号作频谱分析,就可以得到频谱随时间变化的信息。上式的运算量是很大的,但在识别系统中,语音的频谱分析已做成专用硬件,占用内存较少,只需要微型计算机及一些模拟外围电路就能做到实时识别。

二、零交叉率

零交叉率是从语音时间信号中提取的参量。在离散信号中,如相继的样点有不同的代数符号就存在过零点,零交叉率是信号频率成分的简便测量方法,这对窄带信号来说更为确切。例如频率为 F_0 的正弦信号,当取样频率为 F_s 时,每周正弦波有 F_s/F_0 个样点,且存在两个零交叉点,所以平均零交叉率为 $Z=2F_0/F_s$,此式能正确估计正弦波的频率,而语音信号是宽带信号,平均零交叉率与频谱没有定量的关系,可是根据短时平均零交叉率能大概估计语音频谱的性质,下面列出短时平均零交叉率 Z_n 的表示式,

$$Z_n=\frac{1}{2N}\sum_{m=0}^{N-1}|\mathrm{sgn}[S(m+n)]-\mathrm{sgn}[S(m+n-1)]|\tag{5.5}$$

$$\mathrm{sgn}[S(m)]=\begin{cases}1 & S(m)\geqslant 0\\ -1 & S(m)<0\end{cases}\tag{5.6}$$

$S(m+n)$ 为第 n 帧的语音信号,零交叉率与信号的频谱有很大关系,例如频谱主要是高频成分,则零交叉率较高,反之如主要是低频能量,则零交叉率就较低。上节已提到浊音能量集中在低频(低于3千赫),而清音大部分能量集中在较高的频率,所以从零交叉率的数值能判别语音是清音,还是浊音。图5-34是清音和浊音零交叉率的分布曲线。从图中可以看出,两种音的零交叉率分布有很大差异,然而也有小

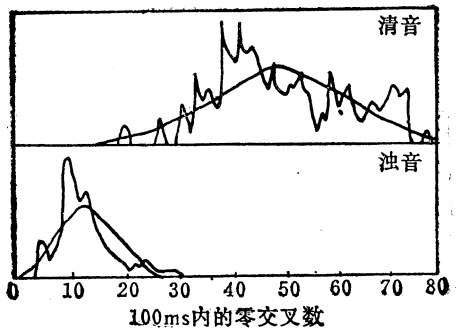


图5-34 清音和浊音零交叉率的分布

部分是重叠的，故虽不能做到绝对正确地用零交叉率判别清音和浊音，但是零交叉率具有运算简单的特点，在识别技术中是常用的参量，它与其他特征参量配合能正确识别语音。

三、线性预测（LP）参量

线性预测分析方法是一种重要的语音分析技术，它可以有效而精确地确定声道转移函数$V(z)$中的常数$\{a_k\}$，图5-37声管模型中各节的反射系数（又称PARCOR系数），面积函数也都是根据线性预测方法提取的。

线性预测的基本概念是对一个语音样点能用前面若干个语音样点的线性组合来近似。设前面p个语音样点的线性叠加为$\tilde{S}(n)$，即

$$\tilde{S}(n)=\sum_{k=1}^{p} b_k S(n-k) \tag{5.7}$$

其中，b_k称为线性预测系数，$\tilde{S}(n)$和语音样点$S(n)$之间的差值，称为预测误差$e(n)$，即

$$e(n)=S(n)-\tilde{S}(n)=S(n)-\sum_{k=1}^{p} b_k S(n-k) \tag{5.8}$$

m个点的预测误差信号的能量为E_n，则

$$E_n=\sum_{n=1}^{m} e^2(n) \tag{5.9}$$

线性预测原理就是找出一组预测系数$\{b_k\}$，使能量E_n为最小值。此预测系数值就等于线性滤波器$V(z)$中的常数$\{a_k\}$，为求E_n最小时的预测系数a_k，令$\frac{2E_n}{2b_i}=0$，$i=1$，2，…，p，得到下面的方程组：

$$\sum_{k=1}^{p} a_k R_n(|i-k|)=R_n(i) \qquad i=1,\ 2,\ \cdots p \tag{5.10}$$

式中

$$R_n(k)=\sum_{j=0}^{H-k} S_n(j)S_n(j+k) \qquad k=0,\ 1,\ 2,\ \cdots p \tag{5.11}$$

$R_n(k)$是短时内自相关函数，为求预测系数$\{a_k\}$，首先要求出（$p+1$）个自相关值，然后再解式（5.10)的p个线性方程组，从而便可得出图5-35中线性滤波器的频率响应，可以通过下式得到：

$$V(e^{j\omega})=\frac{G_k}{1-\sum_{k=1}^{p} a_k e^{-j\omega k}} \tag{5.12}$$

这里必须指出，计算预测系数G_k的过程中必须保证系统$V(z)$是稳定的，即极点在z平面的单位圆之内。例如，在计算自相关函数时精度不够也会影响系统的稳定性。此外，可以证明，$V(z)$与极点数p有关。当p足够大时，全极点线性系统的功率谱$|V(e^{j\omega})|^2$

近似等于语音功率谱$|S_n(e^{j\omega})|^2$，为了使线性滤波器仅代表声道和辐射的性质，即它的频谱相当于语音的频谱包络，必须适当地选取p值。根据声道长度的大致估计以及实验的证实，声道极点数是取样频率的函数，约等于取样频率f_s（以千赫为单位），再加上2～3个声音辐射的极点数，所以在取样频率10千赫时，通常p选取12～14。

利用线性预测方法除了能从语音信号中直接提取预测系数外，还能导出其他一些特征参量。下面简要介绍它们的物理意义以及计算公式。

（一）预测误差

把预测系数$\{a_k\}$的值代入式（5.8）和（5.9），可得到预测误差的能量E_n的最小值，它的归一化值为E_p，E_p是随p值增加而减小的函数。见图5-35。从图可看出，p增加时E_p减小，但p值达到12后，E_p值减小很缓慢，最后当p足够大时，E_p趋近极小值E_{min}。极小值E_{min}完全与信号谱的形状有关，频谱越平坦，E_{min}值越大，反之E_{min}值就越小。如图中曲线所示，清音的误差值E_p较高，而浊音的E_p值较小，这个性质很重要，它可用于识别清音和浊音。

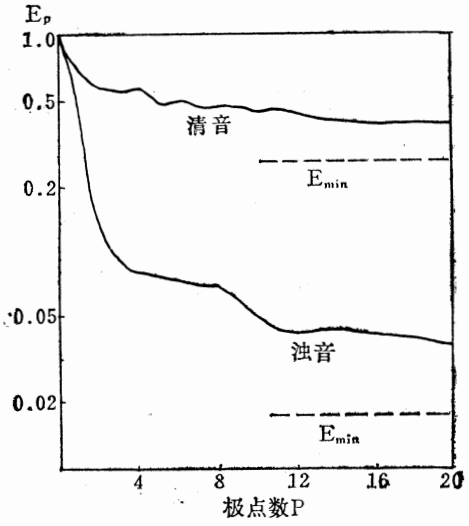


图5-35 归一化误差曲线

（二）反射系数（PARCOR系数）和对数面积比系数

前面已简单介绍了声管模型（见图5-33）声道可简单地看作声管，因而声道的作用可用声管中声音传播的物理过程来描述。在声管模型中，各节声管之间的反射系数及它们的面积比是声管的重要参量，因此它们与预测系数相比有更明确的物理意义。PARCOR系数，即反射系数k_i能从线性预测系数$\{a_p\}$用递推方法求得，即

$$k_i=a_i^{(i)}$$

$$a_j^{(i-1)}=\frac{a_j^{(i)}+a_i^{(i)}a_{i-j}^{(i)}}{1-k_i^2} \qquad 1\leqslant j\leqslant i-1 \tag{5.13}$$

当$i=p$时，$a_j^{(i)}$表示第j个线性预测系数。上式求解时，先使$i=p$，再等于$p-1$，最后等于1，这样经p次运算可求出p个PARCOR系数。

对数面积比系数g_i可用下式从RARCOR系数推导出：

$$g_i=\lg\left|\frac{1-k_i}{1+k_i}\right| \qquad 1\leqslant i\leqslant p \tag{5.14}$$

（三）二极点线性预测参数

其本语音单元分成两大类元音和辅音，元音又分成前元音、中元音和后元音，辅音也能再分成清辅音和浊辅音。在计算机识别时，利用两极点线性预测参数能自动对语音分类。

在线性预测中，如极点数p值选用2，则式（5.12）成为

$$V(e^{j\omega})=\frac{G}{1-\sum_{K=1}^{2}a_k e^{j\omega k}}$$

此两极点系统的频响代表语音频谱的特征。它有一个复数共轭极点或者两个实数极点，频响$V(e^{j\omega})$的极大值频率就是语音频谱中能量的集中部分，因此能用二极点线性预测系数来辨别不同类型的语音，因为清音的能量相对集中在高频域，而鼻音和元音的能量相对集中在低频域。

二极点频谱$V(e^{j\omega})$的峰值频率总是在语音的第一共振峰频率F_1和第二共振峰频率F_2之间，由于语音随时间的变化过程中，F_2常有较大移动而F_1移动较小，因此二极点频谱峰值频率的移动表示第二共振峰F_2的移动，而F_2的移动特性在辨别某些音时是很重要的。

从上面可看出计算二极点频谱的极大值频率能确定语音频谱中能量集中部分的位置和F_2的动态特性。此外，两极点线性预测误差最小值还包含了频谱能量集中程度的重要信息。可以证明，频谱中能量越集中，预测误差越低，在三类元音中，后元音预测误差最低，而前元音预测误差最高。如鼻音的特征是预测误差较低，且两极点频谱的峰值频率等于0赫，而对于清音来说，其规一化误差，峰值频率及零交叉率都比较高。

上面介绍了语音识别中常用的几个特征参量。各种识别系统必须根据具体要求提取适用的特征参量，而特征参量之间的正确配合能得到较好的识别效果。

5.2.3 语音识别方法

前面已谈到，可以用作语音识别的参量很多，例如：零交叉率；能量；基调；共振峰F_1、F_2、F_3；频谱包络；线性预测系数a_1，a_2，…；PARCOR系数；声道断面积函数；LPC剩余误差等等。这些参数有些是独立的，有些是相互有关的，究竟选用哪几个参量，应由被识别语言的特征：是单字还是连续词汇；词汇量大小；参量分析处理实时性能；发音场合有无噪音等来决定。一般要经过下列几个处理步骤：

（1）语言信号起始点和结束点的确定；

（2）语音参量的提取；

（3）时间的非线性“扭曲”处理；

（4）根据特征参量进行逻辑判断，或进行距离测量，求出距离最小者。

其框图如图5-36所示，以下解释各个步骤的作用。

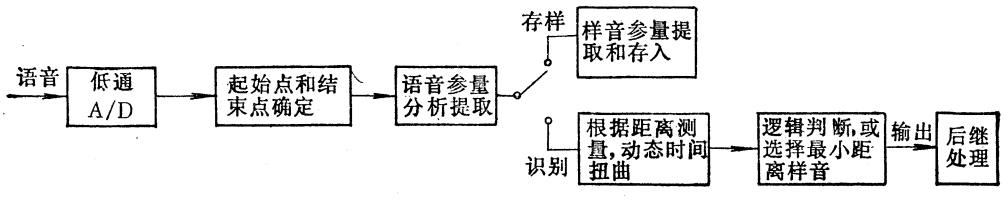


图5-36 语音识别的处理流程框图

（一）低通滤波器和A/D模数转换

一般通过传声器或录音机输入的语音信号都是模拟量，要转换成计算机能运算的数字量，必须经过A/D模数转换，A/D精度一般在10位以上，有些处理系统也有用8位的，采样率一般每秒10^4次，也有低到6.7×10^3次的，A/D的位数和采样频率越高，当然转换质量也越高，但这样会大大增加计算机的运算量。A/D前的低通滤波器是由采样标准所要求的，它的截止频率应小于A/D采样频率的一半，以消除折叠效应的影响，滤波器的阻带衰减应大于60分贝/倍频程，这个要求是较容易满足的。

（二）起始点和结束点的测定

为了使待测的一个语音（或语句）能与样音参量进行很好比较，必须精确的测定语音在时间上的起始点和结束点，这对语音识别是非常重要的，倘若发音现场很安静，背景噪声能量低于语音中清辅音能量，那末，只要测量语音能量就能确定语音的起始点和结束点。但一般场合并非如此，需用能量和零交叉率两个参量才能确定端点。其步骤如下：能量和零交叉率的测量都是以10毫秒为一帧来计算的。若采样频率为10^4次/秒，则一帧内有100个样点。用绝对值i的和代替帧内各点平方和进行能量的计算，可得

$$E(n)=\sum_{i=-50}^{50}|S(n+i)| \tag{5.15}$$

这样，可提高运算速度，并可得到较光滑的能量曲线。在测定时，首先让发音者在指定时间间隔中发音，并保证在此时间间隔前的100毫秒内无语音存在，那末就可利用此100毫秒时间测量“寂静”时平均能量$I_{\min}$和零交叉率统计特性：零交叉率平均值$\overline{I_{zc}}$和标准偏差σ_{Izc}。就可选定零交叉率阈值为：

$$I_{zCT}=MIn(I_F,\ \overline{I_{zc}}+z\sigma_{Izc}) \tag{5.16}$$

式中，I_F为常数，I_F=25次/10毫秒，然后计算整个时间间隔的能量函数，由能量函数的极大、极小值就可决定能量的高低两阈值I_{TU}和I_{TL}，如图5-37所示。

$$\left.\begin{aligned}&I_{TL}=\min(I_1,\ I_2)\\&I_{TU}=5^*I_{TL}\\&I_1=0.03\cdot(I_{\max}-I_{\min})+I_{\min}\\&I_2=4\cdot I_{\min}\end{aligned}\right\} \tag{5.17}$$

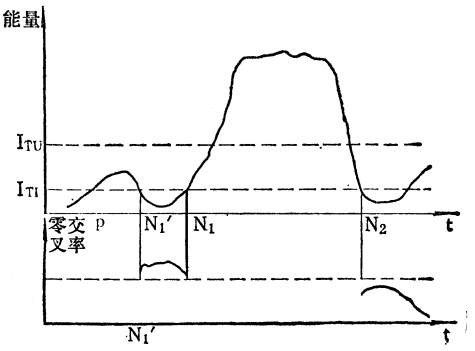


图5-37 语音能量和零交叉率对时间t的关系曲线

接着，就根据I_{TU}和I_{TL}在时间轴上找出能量超过I_{TL}的第一点，并且在这点以后，能量在超过I_{TU}之前不再低于I_{TL}，在图中对这点标为N_1。在PN_1间隔内由于能量小于I_{TL}，所以语言起点初步定在N_1，而不定在p点。用同样方法，可定出结束点N_2。从能量观点定出的起始点和结束点是偏于保守的，即语音起始点和结束点肯定不会落在N_1N_2区间内，但是是否就是N_1，N_2点，还需要检查零交叉率数值的大小。倘若在N_1点到（N_1-25）点区间内，语音信号零交叉率低于零交叉率阈值I_{ZCT}，那末N_1点就定为起始点；倘若在此区间内零交叉率大于阈值I_{zCT}有三次以上，那末起始点还应向前修正到超过I_{zCT}的第一点$\hat{N}_1$处，同样，检查N_2到（N_2+25）区间内零交叉率有无超过I_{zCT}达三次以上，决定是否需要对N_2位置进行修正，整个方法流程见图5-38。

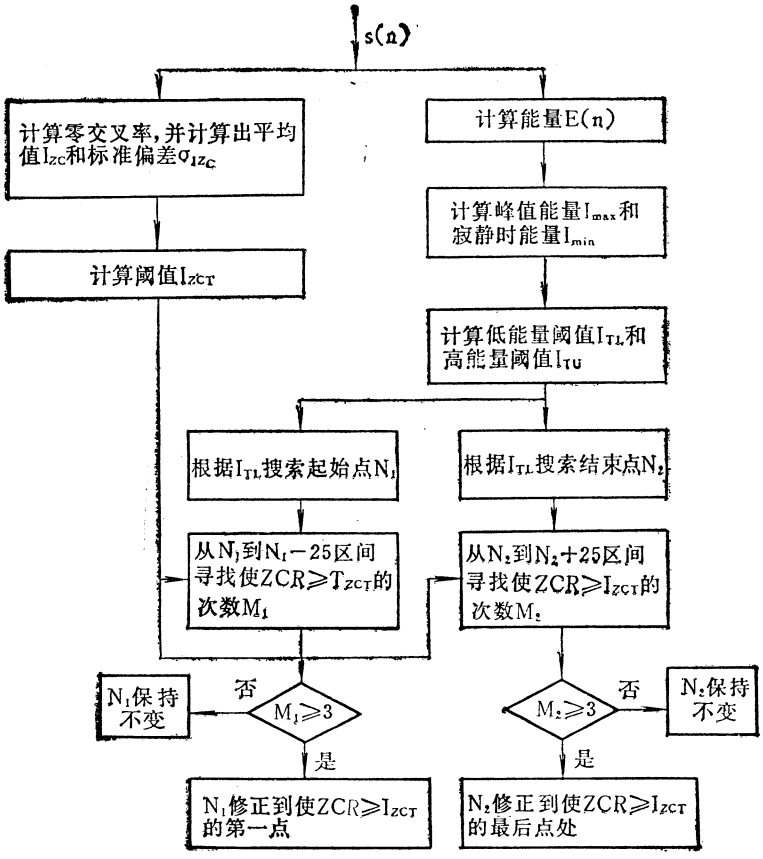


图5-38　确定语音起始点和结束点的流程框图

（三）时间的扭曲处理

在正确地测定了语音的起始点和结束点，并测出该区间的有关语音参量后，还不能与存储的样音参量进行直接比较。这是由于，即使同一个人重复发相同的音，语音的各帧参量也不可能完全重复，因此还需要进行时间的“扭曲”处理，才能使被测定的语音和样音很好地对齐，这一措施是极为重要的。决定扭曲函数的方法有几种，从其原理来看，大致相同，这里仅介绍一种。考虑一对样音和被测语音的强度包络曲线（一般选用语音强度作时间扭曲处理），从语音的起点到结束点的取样点分别为

$$n=1,\ 2,\ \cdots,\ N \qquad m=1,\ 2,\ \cdots,\ M$$

设时间的扭曲函数为$W(n)$，则$m=W(n)$

函数的边界条件是

$$W(1)=1 \qquad \text{起始点}$$

$$W(N)=M \qquad \text{结束点}$$

倘若时间的扭曲是线性函数，则$W(n)$应有如下形式：

$$W(n)=\left[\frac{M-1}{N-1}(n-1)+1\right] \tag{5.18}$$

但实际上往往选用非线性扭曲的效果更好些。

考虑到被检测语音和参考语音两者瞬时速度之比不会超过一倍以上，因此，对非线性扭曲函数$W(n)$可作如下约束：

$$\left.\begin{aligned} W(n+1)-W(n)&=0,\ 1,\ 2 \quad \text{当} W(n)\neq W(n-1)\\ &=1,\ 2, \quad \text{当} W(n)=W(n-1)\end{aligned}\right\} \tag{5.19}$$

然后，测量在n点和m点处被测语音和样音的强度值，求出它们之间的距离。根据使总距离为最小的原则，决定在（n，m）点的扭曲函数究竟取式（5.19）中哪个值。扭曲函数的非线性扭曲情况，如图5-39所示。图中，$N=20$，$M=15$。图5-40表示一组测试语音和样音强度曲线经过非线性扭曲处理前后的情况，可看出经非线性扭曲处理后，两者的吻合程度有了改进。其中实线表示被测语言强度曲线；虚线表示样音强度曲线。

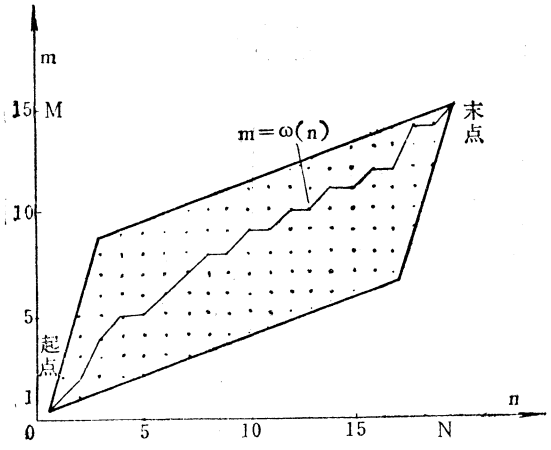


图5-39　n和m点扭曲函数的取值情况

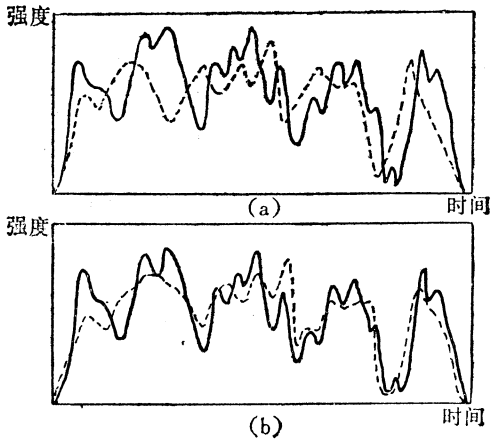


图5-40　被测语音和样音的强度曲线在非线性扭曲处理前后的符合情况

（a）非线性扭曲处理前；（b）非线性扭曲处理后。

（四）距离的测量

当输入的被测语音经过端点的确定，并经过时间的非线性扭曲处理后，就转入识别处理。所谓识别也就是从存储的样音组中找出一个样音，使它的语音参量和被测语音参量最为近似，或者说两者距离最小，那末这组样音就被选中。因此识别过程就是求距离的过程。距离的计算方法很多，最简单方法是把语音参量的N个取样值（扭曲处理后的值）作为N维空间，然后测量多维空间规正后的距离：

$$d_j=\sum_i \left[(a_{js}(i)-a_{jr}(i))/\sigma_{aj}(i)\right]^2 \tag{5.20}$$

式中，$a_{js}(i)$是第j个被测参量（或第j次测量）在时间i的值；$a_{jr}(i)$是第j个样音参量（或第j次测量）在时间i的值；$\sigma_{aj}(i)$是第j个参量在时间i的标准偏差。总距离D为

$$D=\sum_j W_j d_j \tag{5.21}$$

W_j是根据第j个参量（或第j次测量）的重要性所决定的加权系数。

有时为了提高识别正确率，需要采用更完善的距离测量方法，ATAL 从多维向量的高斯概率密度分布函数 $g(x)$ 出发

$$g(x)=(2\pi)^{-1/2}|W_i|^{-1/2}\mathrm{EXP}\left[-\frac{1}{2}(x-m_i)^T W_i^{-1}(x-m_i)\right] \tag{5.22}$$

求得距离表达式为：

$$d_i=(x-m_i)^T W_i^{-1}(x-m_i) \tag{5.23}$$

d_i 是被测向量与第 i 个样音向量的距离；x 是被测 L 维参量向量。式中，m_i 是代表第 i 个样音的平均向量：

$$m_i=\frac{1}{N_i}\sum_{n=1}^{N_n} x_i(n) \tag{5.24}$$

W_i 是相应的第 i 个样音的协方差矩阵

$$W_i=\frac{1}{N_i}\sum_{n=1}^{N_i} x_i(m)x_i^T(n)-m_i m_i^T \tag{5.25}$$

N_i 是存储第 i 个样音参量时，输入的样本语音次数，x_i 代表第 i 个样音的 L 维参量向量。

5.2.4 语音识别系统实例

汉字输入系统，应能识别几千个以上单字，并且字的长短组合是任意的，这在目前还较难做到，需进一步研究。目前只能在词汇量，读音方式等等作一定限制的条件下识别语音。这样的系统很多，这里例举几个，就它们的工作原理，分析的参量等作一简单介绍，至于更详细的了解，请参阅有关专著。

一、单字识别系统

系统对使用对象不加限制，不需专人发音，可在具有噪声的机房内使用，但识别字数较少，如0～9十个数字，系统的框图如图5-41所示，它是根据对各个字的语音参量之间的差异，进行仔细的逻辑判断来识别的。

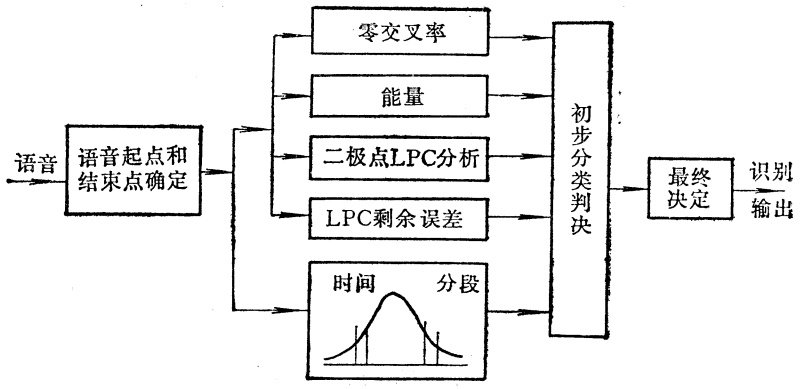


图5-41 单字语音的识别系统框图

在确定输入语音的起始点和结束点后，分段平行计算语音的零交叉率、能量、二极点线性预测和预测误差，例如图 5-42 和图 5-43 是英语中 NINE 和 SIX 两个字的参量随时间分布情况。从图中可看出，NINE 的鼻音部分规一化后的预测误差较小，两个极

点频率为零赫，而SIX的摩擦辅音部分具有高的预测误差，极点频率和零交叉率值也很高。对其他数字也可类似地进行分析，得到各自的参量特征，利用这些特征就可以对语音进行分类，并可建立一定的逻辑判别法测出输入的语音。

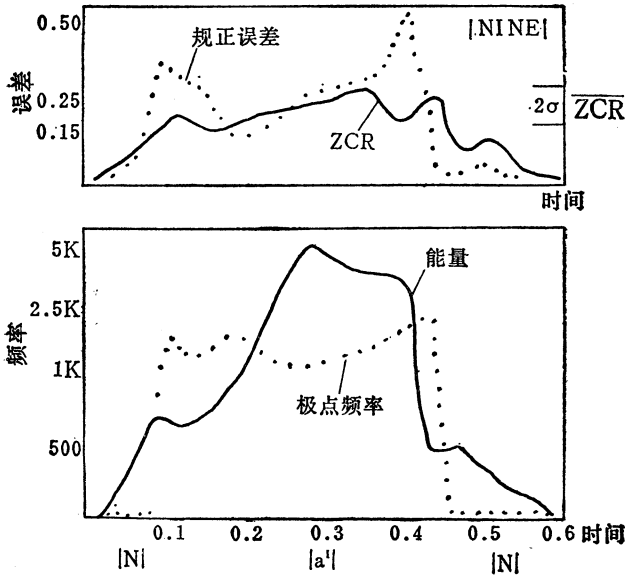


图5-42　英语NINE语音的各参量随时间变化的曲线

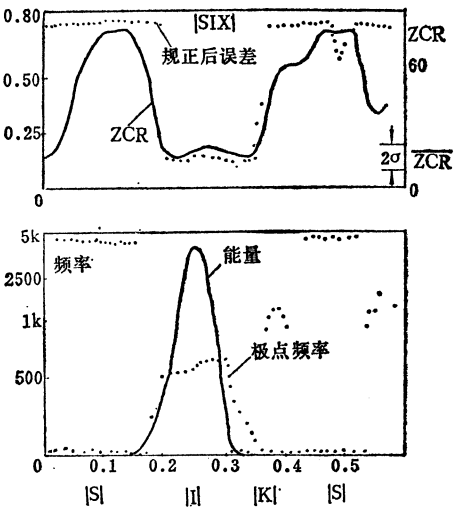


图5-43　英语six语音的各参量随时间变化的曲线

二、词汇量较大的识别系统

此系统特点是：

（1）采用单字输入格式，连续字之间发音要求有停顿。

（2）词汇量为100～500单字。

（3）对讲话环境没有太多限制，可采用电话传输系统。

（4）对讲话者性别年龄没有限制，但必须固定专人发音，若要更换人发音，必须经过训练，对词汇表中每个字发音一次或多次。

系统的结构框图见图5-44它是利用自相关系数来进行识别处理的。

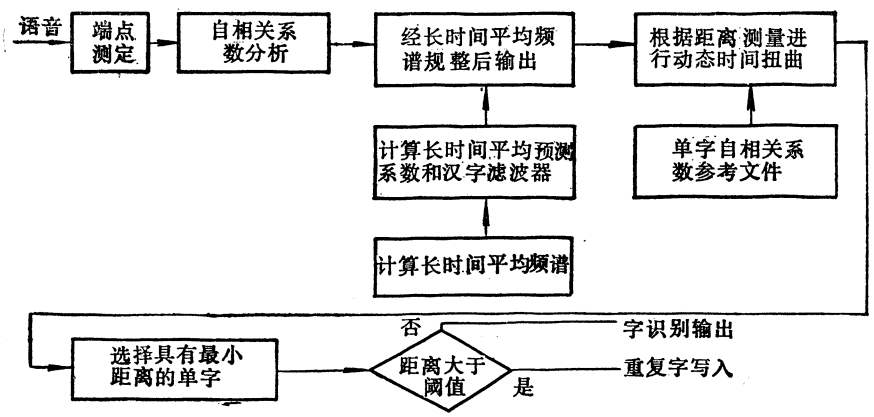


图5-44　词汇量较大的语音识别系统框图

系统的A/D采样率为6.67×10^{3}次/秒，语音带宽3千赫，输入语音先经过端点检测，然后每秒钟重复67次，计算自相关系数的前八个系数。为了补偿电话传输系统的频响不均匀，系统对自相关系数一帧帧地进行平均计算处理。由平均自相关系数算出二极点LPC系数和相应于平均谱的R滤波器，各帧的自相关系数经过R滤波器规整后输出，再经过时间的扭曲处理，与原先存储的样音自相关系数进行比较，找出与之距离最小的样音，倘若这最小距离小于预先规定的某阈值（由实验确定），则这个样音就被选中。否则，不作判断。

三、通用实时语音识别系统

近年来，我国在语言识别方面工作也进展得很快，具有代表性的是中国科学院声学所首创的通用实时语言识别系统，该系统适用于由任意字组成的汉字短语，但组成短语的字的个数应是确定的，下面简单地介绍一下该系统的工作原理。

该系统与上面介绍的两个系统不同，它是利用频谱参量进行识别处理，由于人耳对语音信息中位相部分不敏感，因此，只需要幅度谱（一般称频谱）就可有效地识别连续语音。其处理框图如图5-45所示。

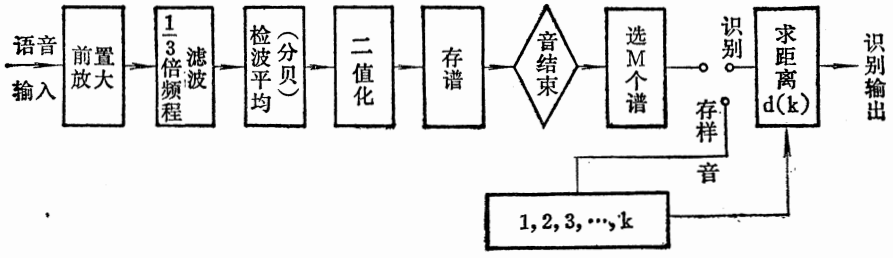


图5-45　我国研制的汉字语音识别系统工作流程

考虑到语音信号在10～20毫秒时间内可认为是准稳定的，因此在识别处理时可把待识别语音按10～20毫秒为一帧分为N帧，然后对每帧按$\frac{1}{3}$倍频程进行频谱分析，得到N帧频谱数值：

$$\overrightarrow{BP}=\begin{cases}A_{11}\quad A_{12}\cdots A_{1L}\\A_{21}\quad A_{22}\cdots A_{2L}\\\qquad\quad A_{ij}\\A_{N1}\quad A_{N2}\cdots A_{NL}\end{cases}\quad\begin{matrix}i=1,2,\cdots N\\j=1,2,\cdots L\end{matrix}\qquad(5.26)$$

其中元素A_{ij}的左脚标i指时间序号，右脚标j指滤波通道序号，为了压缩内存，加速识别过程，考虑到频谱变动部分对识别有较大作用，因此可删去频谱变化不大的相邻帧，从N帧中选出M帧，进行识别处理，通过实验

$$M=(2\sim4)\times\text{字节数}$$

选帧流程如图5-46所示，为压缩内存，减少运算量，对选出的M帧频谱需要进行二值化处理，下面对此加以说明。假定被选中的第P个频谱原始数据是：

$$A_{P1},\ A_{P2}\cdots,\ A_{PL}$$

右脚标是指滤波通道序号，由于A_{P1}，A_{P2}，…，A_{PL}是以分贝表示的数据，在L个A_{Pj}中往往有好几个是零，为了克服音量变化时零的个数的影响，采用非零数据的平均值作为二值化处理的阈值，假设有s个值不为零，则可用下式求出频谱平均值：

$$\overline{A_P}=\sum_{j=1}^{L}A_{Pj}/\,s$$

然后按下式进行二值化处理：

$$a_{Pj}=\begin{cases}1 & \text{当 } A_{Pj}\geqslant\overline{A_P}\text{时}\\ 0 & \text{当 } A_{Pj}<\overline{A_P}\text{时}\end{cases}\qquad j=1,2,\cdots,L$$

对于每一个选中的频谱都进行这样的二值化处理，由于所用的阈值随频谱而变，是浮动的，这有助于克服音量变化所引起的问题。经过二值化处理后，信息存储量大大压缩，一个频谱分量只要一个比特就可以了，并且距离的测量也可大大简化。假定一帧频谱用L个通道数据来表示，将二帧二值化频谱进行比较时，只需进行一次“异或”运算，然后查一下有几个位是1，于是，1的个数就可作为这二帧频谱间的距离。假定每句话选用M个二值频谱，字表大小为B，则第K个参考音数据为：

$$\begin{matrix}a_{11}^{(k)} & a_{12}^{(k)}\cdots a_{1L}^{(k)}\\ a_{21}^{(k)} & a_{22}^{(k)}\cdots a_{2L}^{(k)}\\ \vdots & \\ a_{M1}^{(k)} & a_{M2}^{(k)}\cdots a_{ML}^{(k)}\end{matrix}\qquad k=1,2,\cdots,B \tag{5.27}$$

其中，$a_{ij}^{(k)}=0$或1，$i=1,2,\cdots,M$，$j=1,2,\cdots,L$

若待识别语言的二值谱数据为：

$$[X]=\begin{cases}b_{11} & b_{12}\cdots b_{1L}\\ b_{21} & b_{22}\cdots b_{2L}\\ \vdots & \\ b_{M1} & b_{M2}\cdots b_{ML}\end{cases} \tag{5.28}$$

其中，$b_{ij}=0$或1，$i=1,2,\cdots,M$，$j=1,2,\cdots,L$。

这个待识别的语音和第k个样音之间的距离可用下式计算：

$$D(k)=\sum_{i=1}^{M}\sum_{j=1}^{L}[[a_{ij}^{(k)}+b_{ij}]]\quad k=1,2,\cdots,B \tag{5.29}$$

其中[[·+·]]指进行“异或”运算。从这B个$D(k)$中选出最小者，例如$D(k)=\min D(k)$，$k=1,2,\cdots,B$，则识别结果便是字表中的第k个语音。

语言识别系统有很多，以上仅介绍了几个例子。单字识别系统目前已有广泛的应用，但它要求使用者间断地发每一个字的音，这就影响了声音的自然度和人机对话的效率。为解决这个问题，必须进行连续语音识别的研究，近几年来，这方面的工作有很大进展，在某些实验室已制成连续语言识别系统。连续语言的识别除了需要语音的特征参量外，还需附加复计的句法、语法等条件。

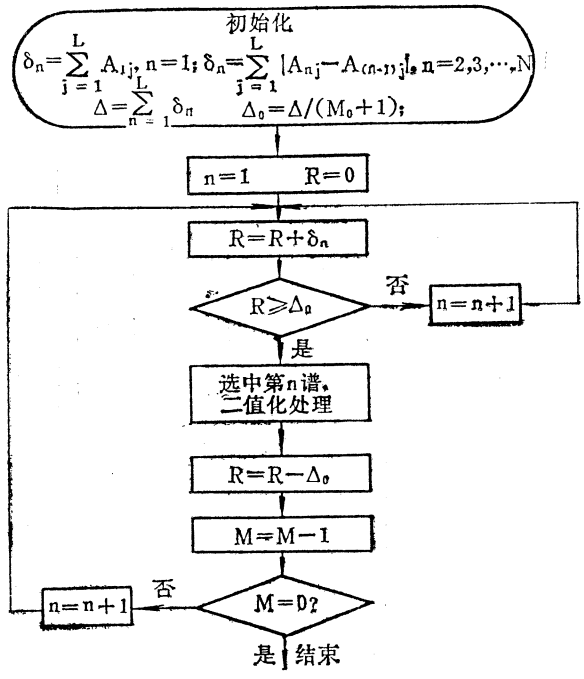


图5-46 汉字语音识别的选帧流程图

5.3 汉字字形输入方法

5.3.1 概述

在汉字输入方法中，除了汉字的键盘输入方法和汉字的声音输入方法外，汉字的字形输入方法，也就是汉字字形识别方法，一直受到普遍的重视。1955年印刷体阿拉伯数字的光学字符阅读机（OCR：Optical Character Reader）商品化以后，便大力开发对数字、拉丁字母以及日语假名等的OCR装置。到了七十年代，印刷体及手写体的数字、字母以及日语假名等的OCR已达到商品化。这种小型OCR的普及十分迅速，同时也进一步推进了汉字OCR的研究和发展。

汉字字形识别方法的研究，还在六十年代初，美国的R.卡赛（Cacey）和G.纳杰（Nagy）便已开展了许多工作。其后，在使用汉字国家之一的日本，把这类研究列为通产省重点资助的研究项目，取得了迅速的进展。1977年底，日本东芝公司首先发表了针对2000个印刷体汉字、输入速度为100字/秒的汉字光学识别装置样机，基本上达到了实用水平。接着，日立公司等也相继研制成同类样机。对手写体汉字识别方法的研究，特别是近几年来，进展也很迅速，根据已发表的资料分析，不少成果已达到或接近实用水平。

在我国，虽然汉语拼音方案的推广已有了一定的基础，但用拼音文字来取代方块汉字并非短期内所能解决。随着我国向信息化社会的进程，各种汉字文献资料迅速增加，当用计算机处理这些大量的汉字信息时，仅用传统的键盘输入方法已远远不能适应，需要探索更高速、更有效的汉字输入方法。因此，对汉字字形识别技术的开发日益迫切，

并已引起有关部门的重视。目前，许多单位正开展着积极的研究，并取得了明显的进步。

汉字字形识别的流程一般如图5-47所示。

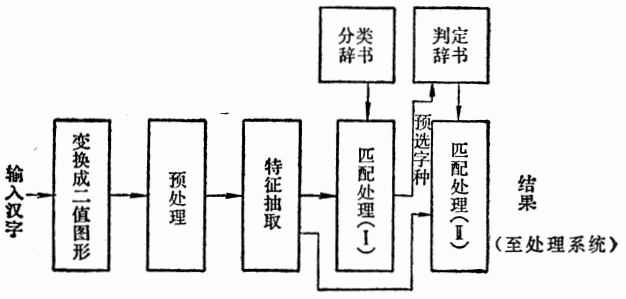


图5-47　汉字识别流程

首先，用光学的方法，对输入原稿上的文字进行光学扫描，并以8～10点/毫米取样，取样点信息用“1”或“0”赋值。我们可以假定：文字笔画经过的点的信息为1，背景部分的点为0，或者相反。这样便将输入的汉字字形变换成二值化的图式。此后，在预处理阶段，除去污染和杂音，使文字正规化。在特征抽取时，是对经过预处理的汉字图形抽出它的图式特征，例如，笔画的长度、角度、端点、交叉点、笔画分布、四周特征或背景特征等等，这些特征一般都表示成多维向量的形式。作为识别标准的学习图形，以多维向量形式，存放在分类辞书和判定辞书中。最后对输入汉字的图式与辞书中的标准图式进行比较，与输入图式最接近的标准图式（也称作两者相匹配），便作为判定结果输出，送到处理系统中。

这一领域的研究开发工作，正在广泛深入开展之中，识别过程各环节的处理方法多种多样，新理论、新技术不断涌现。但在本节中，我们将着眼于基础部分的讨论，较系统地介绍单一字模印刷体汉字识别技术，而不讨论多字模印刷体汉字识别技术。对手写体汉字识别技术主要讨论具有总体性质的特征抽取方式。

5.3.2　印刷体汉字的字形输入

印刷体汉字字形识别的研究，是在汉字字形识别研究中最早开展的领域。在已发表的研究成果中，字量在2000字，大都已达到误识率为10^{-4}，拒识率为10^{-3}，识别速度为50～100字/秒的水平，基本满足实用要求。

一、汉字字形的取样

对原稿上印刷体汉字字形的取样，通常是利用光学的方法，由光电变换系统，对纸面上的文字进行逐点扫描，以一定的时间间隔，取出扫描范围内的信息，再以一定的阈值，把它们变换成二值化的电信号。目前常见的扫描机构有如下几种形式。

（一）阴极射线飞点扫描方法

以飞点扫描管为光源，在纸面上顺序扫描。用光电倍增管接收扫描信号，并转换成较强的电信号输出。其原理如图5-48所示。

这种扫描管的光点是直接受偏转电路控制的，因而，它的扫描形式可以在逻辑上予以控制。在字母、数字识别系统中，它也是一种较常采用的方法。

飞点扫描管的扫描速度，主要受荧光体材料余辉时间的限制。如果余辉时间为0.1微秒，则其速度为1000万位/秒。为判断一个方块汉字所需的二值化点阵字形信息，一般不超过100×100点，或者只在40×40点左右，因此它的扫描速度可以达到每秒钟一千字以上。

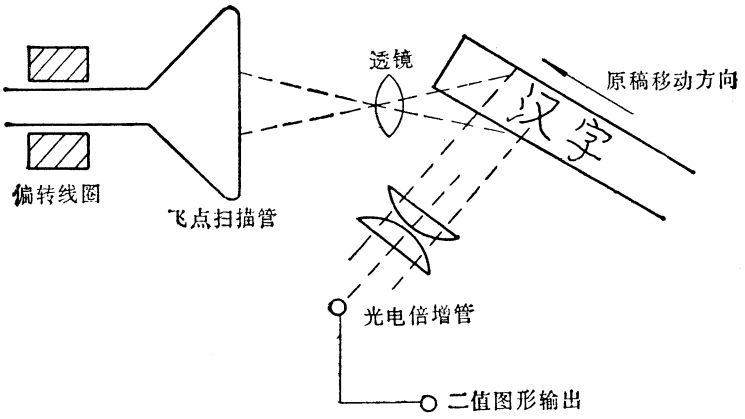


图5-48　飞点扫描法原理

扫描所得的信号是从纸面上的字形轨迹取出的，纸面上的污染，或者字形轨迹周围的污染点，将使接收信号的噪音增加。外界光的影响，也会降低信噪比，所以，这种方法通常需要遮蔽外界光线的措施。在外围电路上，扫描管要有一套高压电源。为避免荧光体的不均匀而招致的荧光面各点上发光强度不一致影响，一般也要采用电路上的措施予以控制。从结构上看，这种方法较为复杂。

（二）激光扫描方法

它的原理如图 5-49 所示。这种方法用激光束为光源，光源强度高，方向性强，光点小。与飞点扫描方法相比，分辨能力大大提高。飞点扫描时，为提高分辨率，要选用光点尺寸小的扫描管，光学系统中的透镜组，一般要用经过各种误差校正的高性能透镜组成。由于激光束定向性能强，可以只使用透镜的中心部分，对光学系统的要求有所降低，并可提高激光束的利用效率。同时，因入射光强大，有利于提高信噪比，也可减少对外界光的屏蔽要求。结构较为简单，成本有所降低。

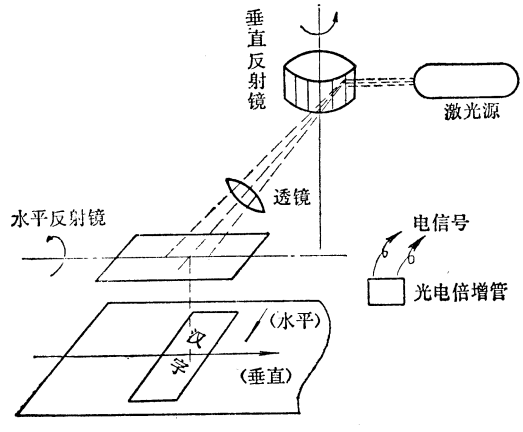


图5-49　激光扫描法原理

激光扫描方法，大多采用机械式光点偏转机构，对激光束的控制不够灵活，因而它的速度较低。

（三）光敏元件方法

利用光敏元件的光电变换性能，将经光源照射的纸面上的反射光，用光学透镜加以扩大，投影到光敏元件上，这样，便可在光敏元件上得到输出的电信号。

一种最简单的结构形式其原理如图 5-50 所示。将光敏晶体管元件排成一列，使其高度略高于文字高度，以允许文字有一定范围的上下偏差。纸面上的反射光在经过透镜扩

大后，投影到光敏元件列上。随着纸面的移动，投影象不断变化，从而完成对纸面的扫描过程。也有用较多的光敏元件排成一个阵列，使其与纸面上一个文字的范围相对应，这样，在扫描一个字时，纸面便可以不必移动。此外，还有采用集成光敏元件的方法，而在光路上更多地采用光导纤维作引导等先进技术。

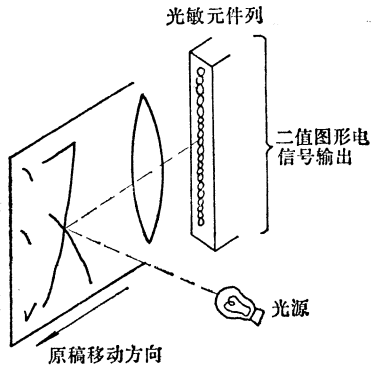


图5-50　光敏元件取样法原理

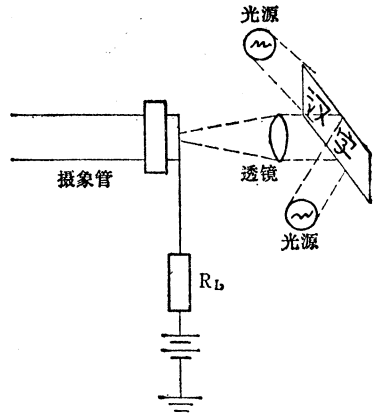


图5-51　光导摄象管方法原理

采用光敏元件取样的方法，元件的寿命较长，稳定性能较好，结构较为简单。但是元件数量较多，需要配以适当的电路，对元件参数间的偏差予以修正。

（四）光导摄象管方法

将光导电物质蒸发在透明的导电膜上作为光靶。字模通过透镜成象后，在光靶上由电荷积累形成图象，用电子束对光靶进行扫描时，在负载 RL 上就有代表字模图象信号的电流输出。如图 5-51。这种方法一般需对摄象管的残象特性进行修正，速度也受到一定的限制。

除上述种种方法外，目前还有采用生成平行光的方法束对文字进行扫描，以及在摄象管原理的基础上用普通的电视摄象机稍加改装来获取汉字点阵字形信息的方法等，都取得了较好的效果。

二、输入字形的预处理

通过光学系统对纸面进行扫描而取样得出的字形信息是点阵式的字形信息。由于印刷质量和纸张的差别，往往会带有各种不同程度的污染，由于光学系统的转换，或者由于机械运动的误差，字形信息中也会有各种噪音和变形等。这些都对正确地判别出输入文字有很大影响。通常为去除这些污染或噪音等，应在预处理阶段予以完成。预处理阶段的具体方法，有赖于整个识别系统的考虑，特别与匹配阶段选取什么样的特征信息等有直接关系。

一般在预处理阶段应当去除文字本身的杂音，并使文字规格化。

文字的规格化，是对输入文字的大小、位置等施以特定的标准化处理，系统中对文字的大小和位置作了规定。对于输入的文字，当大小不符时，可把输入文字作为平面上的二维图形，按纵横方向分别乘以适当的比例因子，便可进行放大或缩小。位置的规格化，通常是把文字区域的中心对应到特定的位置上，以校正文字的偏移。必要时可围绕中心施以适当角度的旋转，以校正输入文字的倾斜。这是一种很直观的方法，原理和相

应的算法都很简单，用软件或硬件的处理都很易于实现。当对预处理阶段要求较高时，则往往要把输入文字作为一个空间曲面图形来处理，这时的规格化过程将略显复杂。

所谓污染或杂音，是指检测出来的字形点阵图式中，不该有黑点的地方有了黑点，或者该是黑点的区域中却窜进了白点。这里我们假定：在一个文字区域中，笔画通过的各点为黑点，其余为白点。为去除因杂音而引起的黑点或白点，以下用一个例子来简单地说明它的方法。

例如，在点阵字形中，假设我们所关心的点$P_{i,j}$为白点（或为黑点），而它周围的8点$P_{i-1,\ j-1}$，$P_{i-1,j}$，……$P_{i+1,j+1}$全部与其相反（可参考图5-52），为黑点（或为白点），这时我们便认为$P_{i,j}$是一个孤立的白点（或黑点）。如果把孤立的白点看作是缺漏，孤立的黑点看作是污染，那么，遇到这种情况，就应将孤立的白点填黑。若是孤立的黑点就将它除去，使之成为白点。若黑点以“1”表示，白点以“0”表示，“∧”为逻辑乘符号，$\overline{P_{i,j}}$为$P_{i,j}$的“非”，即$P_{i,j}$为1时，$\overline{P_{i,j}}$为0。那么，它的处理算式便可列出如下：

$P_{i-1,\ j+1}$	$P_{i,\ j+1}$	$P_{i+1,\ j+1}$
$P_{i-1,\ j}$	$P_{i,\ j}$	$P_{i+1,\ j}$
$P_{i-1,\ j-1}$	$P_{i,\ j+1}$	$P_{i+1,\ j-1}$

图5-52　去除杂音示意图

当

$$P_{i-1,j-1}\wedge P_{i-1,j}\wedge P_{i-1,j+1}\wedge P_{i,j-1}\wedge P_{i,j+1}\wedge P_{i+1,j-1}\wedge P_{i+1,j}\wedge P_{i+1,j+1}=1$$

成立时，便使$P_{i,j}=1$；

当

$$\overline{P}_{i-1,j-1}\wedge\overline{P}_{i-1,j}\wedge\overline{P}_{i-1,j+1}\wedge\overline{P}_{i,j-1}\wedge\overline{P}_{i,j+1}\wedge\overline{P}_{i+1,j-1}\wedge\overline{P}_{i+1,j}\wedge\overline{P}_{i+1,j+1}=1$$

成立时，便使$P_{i,j}=0$；

如此进行下去，便可除去文字的杂音或污染。另外，还有笔画的粗细及浓度等规格化工作，这些在本节稍后的内容中，我们还将有机会提到。经过这样预处理的字形，便可用来与标准字形作相应的匹配处理了。

三、汉字字形的分类

在数字、字母，甚至包括日语假名的场合，作为输入对象的字量，大约为100个上下，匹配处理一次便可完成。但是，由于汉字的字量大，字形复杂，因此，若直接用一次匹配的方法，则费时过多，要占用大量高速存储空间，是不切实用的。通常，对汉字的匹配处理要分两次进行。在匹配处理（Ⅰ）的阶段，先对输入字形作出粗略的分类，每类中的预选字数限制在10～100个范围内。在匹配处理（Ⅱ）的阶段，再将输入文字与所在类别的预选字进行匹配，并最后判定出结果来。

分类时，在输入字形图式中抽取出适当的信息，以此作为输入文字所属类别的表征。以下介绍几种较为典型的分类方法。

（一）复杂指数法

复杂指数指的是文字的线段密度，即单位扩展量中文字笔画的长度。如果将文字看

成平面上的二维图形 $f(x,y)$，那么，便可计算出它在不同方向笔画线段上的长度之和。另外，对图形 $f(x,y)$ x 方向的 i 次、y 方向的 j 次的矩 m_{ij} 可作如下定义：

$$m_{ij}=\iint x^i y^j f(x,y)dXdy$$

特别是，x 方向重心 $M_{10}=m_{10}/m_{00}$

y 方向重心 $M_{01}=m_{01}/m_{00}$

进一步再给出：

x 方向的二阶矩 $M_{20}=(m_{20}-m_{10}^2/m_{00})/m_{00}$

y 方向的二阶矩 $M_{02}=(m_{02}-m_{01}^2/m_{00})/m_{00}$

我们把文字的纵向线段长度之和记为 L_y，横向二阶矩（即扩展量）的平方根记作 σ_x，则定义横向复杂指数 C_x 为：

$$C_x=L_y/\sigma_x \qquad \sigma_x=\sqrt{M_{20}}$$

类似地，纵向复杂指数 C_y 为：

$$C_y=L_x/\sigma_y \qquad \sigma_y=\sqrt{M_{02}}$$

这样，可以给出文字的整体复杂指数 C 为：

$$C=(L_x+L_y)\sqrt{\sigma_x^2+\sigma_y^2}$$

由此，对每个字便可用复杂指数作为表征该字的特征量而进行分类。

为对这种方法有个较为直观的了解，可参阅图5-53。这是日本东芝研究所对日语汉字（也包括数字、字母及日语假名等）所作的复杂指数分布图的一部分。

由图看出，上方文字的纵向复杂指数 C_y 大，如曇、量、震等；右方文字的横向复杂指数 C_x 大，如剛、綱等；分布在 45° 线附近的文字，如粉、姬等，其 C_x 与 C_y 比较接近。另外，数字、字母、假名以及常用符号等的 C_y 都比较小，可见，用 C_y 特征量便能将它们与汉字作一次分类。同时，整体复杂指数 C 与文字的笔画数有着线性关系，笔画数愈多，C 愈大。因此，用复杂指数来定义文字的复杂性，便和人们的直观感觉相一致了。

由于复杂指数不因文字位置而变化，因此对位置偏移有很强的抗干扰能力。用它对文字分类，在约2000字数的日语汉字中，每类平均字数 190 个，对于复杂的汉字，每类字数往往可降低到 50 个左右。当识别的对象字数增加时，这种方法的分类效果也会有所降低。

（二）四边码法

这种方法着眼于文字四周的笔画分布情况，如图5-54所示。将文字四周的笔画浓度分布予以数值化，例如，分成 0、1、2 三个浓度等级。从上边起，按顺时针方向分别为第一、二、三、四边。根据文字笔画在四边中所占的比例，选择确定的阈值，分别赋以浓度等级 0、1、2 的值，按四边顺序写成的浓度值，便是该字分类特征量——四边码。图中的“中”字四边码为 0101，“国”字为 2222。类似地，“子”字应为1000，“用”字应为 2202，“因”、“囚”的四边码均为2222 等。

四边码法对文字的断线抗干扰能力很强。在与上述方法类似的统计中，用四边码作分类特征量，每类的预选字数约 270 个。如果和复杂指数特征量同时使用，则每类字数

图5-53　复杂指数分布例

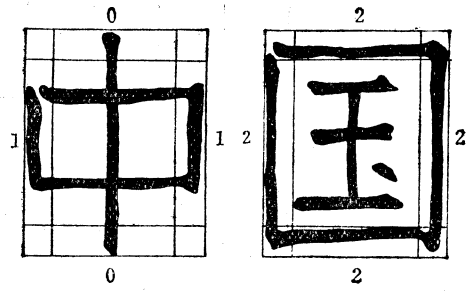


图5-54　四边码法例

中：0101　国：2222

约 98 个。对于复杂的汉字，这两种特征量的独立性很强，都可使每类字数减少到 50 个左右。

（三）带图形法

如果把文字的笔画设想为一个导体，并向周围发出等电位线，那么在稍离笔画轨迹的点观测到的等电位线，就不大受文字变形的影响。带图形法就是抽出这种对文字断线极不敏感的特征来进行分类的。代替等电位线，这种方法是将文字分阶段粗化，在各阶段中表示四周黑部比例的图形便是带图形，如图5-55所示。图中把文字的四边，从上边起，按顺时针方向依次作为第一、二、三、四边。为清楚起见，这里仅示出了第二边的图形，各边中取 4 点宽度而成带状，每边的带状图形再以 5 点一组，分成 8 个小区域，根据各小区域中笔画所占部分（即黑的部分）的比例，分别赋以“0”或“1”的值。这里的 0、1 的值，是按确定的阈值赋给的。这样，每边便有一个 8 位的二进制代码与之对应，如图中的第二边代码为01000000，将它记作 T_{02}。经三次粗化，可分别得到T_{12}、

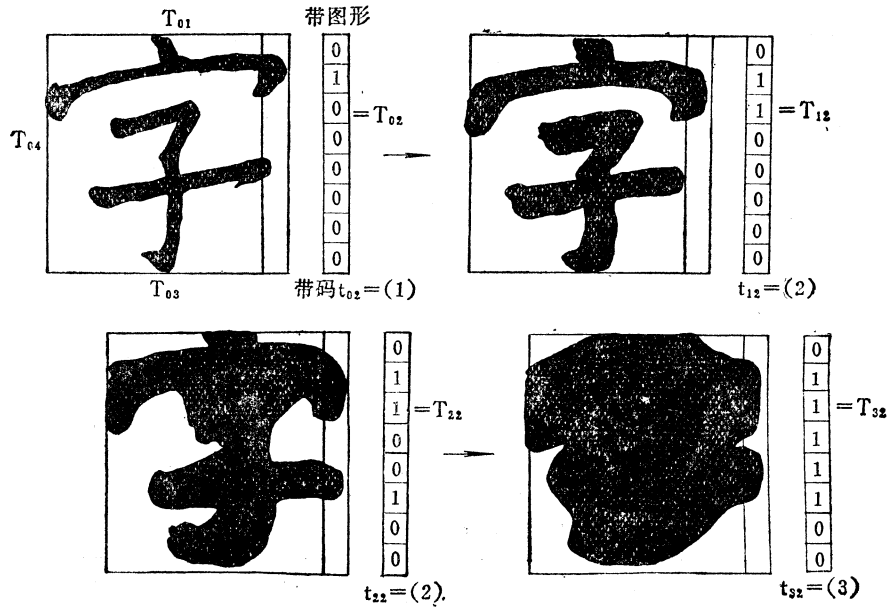


图5-55　经三次粗化的带图形及带代码例（仅取第二边）

T_{22}和T_{32}。从各边求出的T_{mn}($0\leqslant m\leqslant 3$，$1\leqslant n\leqslant 4$)，便是该字的带图形。于是，每字可用 8 × 4 × 4 ＝128 点来表征，比原图形的 40×40＝1600点大大减少了。

接着，再定义带状码。按 T_{mn} 的 8 位码中“1”所占的比例，再以一定阈值分别赋以 1 ～ 3 的值作为带状码。把这个值记作 t_{mn}，则 $1\leqslant t_{mn}\leqslant 3$ ($0\leqslant m\leqslant 3$，$1\leqslant n\leqslant n$)，图中的 $t_{02}=1$，$t_{12}=t_{22}=2$，$t_{32}=3$。固定 m 时，用 t_{mn} 的组合，即用 (t_{m1}，t_{m2}，t_{m3}，t_{m4}) 来定全体码，参照图5-56所示代码表。

全体码T_m	带代码	带码3的位置	(例)
1	四边全为1		
2	一边为2		
3	二边为2		
4	三边为2		
5	四边全为2		
6	一边为3		院
7			士
8			利
9			天
10	二边为3		
11			向
12			
13			
14			工
15			司
16	三边为3		凶
17			区
18			冈
19			
20	四边全为3		国

图5-56　代码表

例如，对于（1，1，1，1），$T_m=1$；对于（3，*，3，*）●，$T_m=14$；等等。T_m 是对粗化各阶段的四边黑部比例进行的宏观描述。T_m的最小值是1，最大值对应于(3，3，3，3)，其值为20，即表中最末一行。把$T_0T_1T_2T_3$ 的顺序排列作为全体码链，分类中便使用全体码链这一特征。按全体码链分类的总类别数较多，如果把总类别数据限制为64，那么就1740个汉字的试验而言，每类字数约40个，而每字平均可能属于的类别数为 2。

● “*”表示 2 以下的任意值。

带图形法用的带图形、带状码、全体码等特征量，可以在扫描字形时，直接由硬件抽取出来，因此，它所需的处理时间是很少的。

（四）分层匹配法

这个方法的概念，可用图5-57来作简单说明。将字形以32×32点、灰度为2位的图形抽取出来，作为第四层。对该层作模糊处理，采样成16×16点、灰度为4位的第三层图形。再以类似方法作出第二层和第一层。标准字形也在事先这样做好。可以看出，不同汉字经过模糊化处理后，可能得到相同的图形，把它们都看成是同一类的文字。第一层同一类别的字最多，例如为m_1个。经过逐次的比较，每一类别的字逐渐减少为m_2、m_3、m_4个。一般用32×32点、灰度为2位的点阵字形信息，就可对汉字作出区别，因此经过第四层的比较以后，便可得到最终的判定结果。完成匹配处理时，这种方法以相同的特征参数，在各层作同样的处理，各层都可用同一的算法。

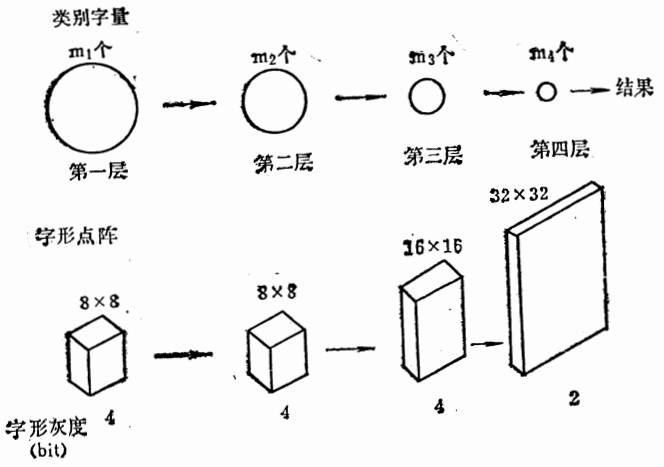


图5-57　分层匹配概念图

实际试验中，以2000汉字为对象，第一层是8×8点、灰度为4位；第二层是40×40点、灰度为1位的图形。经第一层匹配后，每一类别的字数平均约10个，再经第二层的匹配处理便可得到结果。

四、汉字字形的判定

在匹配处理（Ⅰ）的阶段给出了输入文字所属的类别，在匹配处理（Ⅱ）的阶段，再将它与同一类别的文字作进一步的比较，便能判定出结果。在匹配处理（Ⅱ）的阶段常用的方法有

（一）单纯相似度法

如前所述，表征文字的特征量通常是以多维向量来描述的。如果输入文字的字形向量记作X，标准文字的字形向量记为X_0，则单纯相似度$S(X_0, X)$用下式定义：

$$S=(X_0, X)/|X_0|\cdot|X|$$

式中，(X_0, X)为X_0和X两向量的内积，$|X_0|$和$|X|$分别为各自长度的模。

可以看出，S是X_0和X两向量间夹角的余弦。当夹角为零时，S值为1，表明两向量重合，即两个文字是相同的。当两个文字不同时，S的绝对值小于1。因此，S值的大小，可作为两个文字相似程度的量度。判定时，S值最接近于1的X_0便作为对输入文字X的判定结果。这在字母、数字等的识别系统中是一种有效的方法，在汉字识别系统中也被用作一种最基本方法。

（二）复合相似度法

这种方法的理论导出较为复杂，有兴趣的读者可参阅有关文献。如果把每一类别的

文字字形函数展开成正交函数列ϕ_j，第k类的正交函数列用k_{ϕ_j}来表示，则输入文字X与第k类函数的复合相似度S_k定义如下：

$$S_k=\left[\sum_{j=0}^{m}\frac{k\lambda_j}{k\lambda_0}\cdot\frac{(k_{\phi_j},X)^2}{|X|^2}\right]^{1/2}$$

S_k被看成是文字X分别在第k类正交函数k_{ϕ_0}，k_{ϕ_1}，…，k_{ϕ_m}的函数轴方向的投影，用$k\lambda_j$对其投影值的二次方进行加权取和。

如果特别选择第k类正交函数列k_{ϕ_0}、k_{ϕ_1}和k_{ϕ_2}（如图5-58所示），对位置的偏移很稳定，此时$\lambda_0=\lambda_1=\lambda_2$，则复合相似度$S_k$便可简化为

$$S_k=[(k_{\phi_0},X)^2+(k_{\phi_1},X)^2+(k_{\phi_2},X)^2]^{1/2}/|X|$$

采用复合相似度方法能较好地识别窜有噪音的输入文字，对位置偏移也有较强的抗干扰能力，但对非常相似的文字则判别能力较弱。

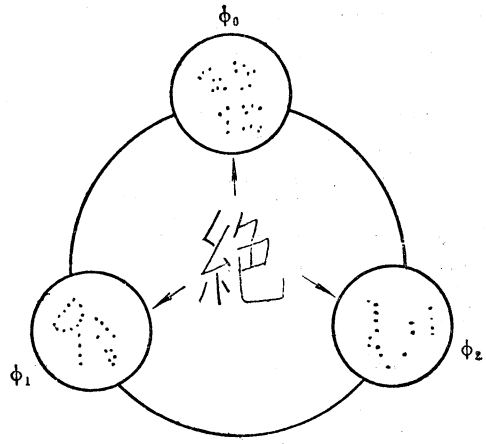


图5-58　正交函数例

（三）混合相似度法

汉字的特点之一是，有些字之间是非常相似的，如王-玉，大-太-犬，哀-衷等。混合相似度便是在复合相似度法的基础上，再进一步强调相似文字间的差别。设第k类中非常相似的文字的标准图形为f，用下式定义ψ：

$$\psi=\left[f-\sum_{j}^{m}(f,k_{\phi_j})k_{\phi_j}\right]\Big/\left[|f|^2-\sum_{j}^{m}(f,k_{\phi_j})^2\right]^{1/2}$$

ψ和ϕ_j正交，$(\psi,\phi_j)=0$。则混合相似度S_k^*为

$$S_k^*=\left[\sum_{j=0}^{m}\frac{k\lambda_j}{k\lambda_0}\cdot\frac{(k_{\phi_j},X)^2-\mu(\psi,X)^2}{|X|^2}\right]^{1/2}$$

式中μ为常数。当输入字形X与f非常相似时，$\mu(\psi,X)^2$项起作用。

这种方法对于有若干个相似文字的情况，也可作出最终的判定。

（四）计算距离的方法

这种方法是对输入文字向量X和标准文字向量X_0之间的距离进行计算，把文字间距离最小的X_0作为是X的判定结果。

设标准文字的特征向量X_0为

$$X_0=(X_{01},X_{02},\cdots,X_{0n})$$

输入文字的特征向量X为

$$X=(X_1,X_2,\cdots,X_n)$$

则文字间距离定义为

$$D(X,X_0)=\sum_{i=1}^{n}|X_i-X_{0i}|$$

除上述几种方法以外，在强调相似文字的匹配处理时，有时也采用二次比较的方法。例如，文字时“间-问”，“困-因”，“刀-刃”等很相似，那么在第一次比较时，先找出相同的部分，然后把不同的部分取出来，再次进行比较，即先去掉“门”、“口”、“刀”，剩下的部分再作比较。这种比较一般可用单纯相似度方法进行。也有另一种方法是，事先将相似的文字组集中起来，给出它们间的差别，特别强调出差别的特征，用软件模拟的方法或硬件抽取的方法，把它们作为判定辞书的一部分保存起来，遇到相似文字时，便可直接比较而得出结果。

在匹配处理（Ⅱ）的阶段所选择的匹配方法，一般并非独立的，常常有赖于匹配处理（Ⅰ）的阶段，即分类阶段所选用的参量。整个匹配过程决定了系统的处理方式。此外，尽管各阶段的匹配方法大致相当，但是由于细节的不同，效果也常有较大的差异。

五、例子

作为例子，我们举日本富士通研究所发表的一个试验装置作简要介绍。它是用带图形法进行分类的，处理流程如图 5-59 所示。将输入文字作成带图形，再作成全体码，按全体码链分类。判定时，采用单纯相似度法作匹配处理。为区别相似文字，用软件方法事先做出了强调文字间细微差别的模块。系统结构如图 5-60 所示。

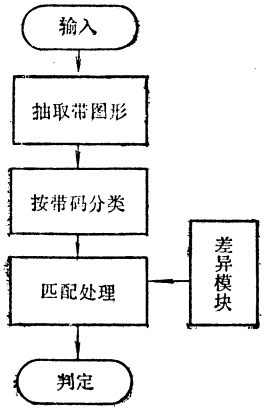


图5-59　处理流程

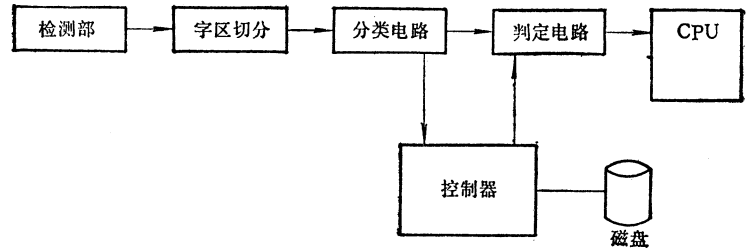


图5-60　试验系统结构图

输入文字的带图形、带状码、全体码是扫描纸面时由硬件抽取的。每一类别的标准字形放置在磁盘的同一磁道上，磁盘转动周期为 20 毫秒，转动一周，一个字的判定便结束，因此识别速度为 50 字/秒。分类电路用全体码指出磁盘的地址，控制器把指定地址的标准字形从磁盘送到判定电路。判定电路对位置偏移进行校正，再以确定阈值作出判定结果。图中字区切分部分是针对汉字特点设置的。汉字中有许多字是由各自分离的部分组成的，如“刻、北、川、非”字等。这一部分用以切分出输入文字的字区，其流程如图 5-61 所示。先读取横向的文字列，由横向连接状态检出线段。该线段可能属于一个字，也可能是一个字的一部分。选定适当的阈值W，如果线段长度大于W，便视为正常字，去作下一步处理。否则，应判断输入文字是狭窄文字还是分离文字。存储器中备有狭窄文字的标准图形，如“目、占、月、贞”等。与狭窄文字作匹配处理，若不一致，

便认为是分离文字的一部分，再检测出下一线段，返回操作。

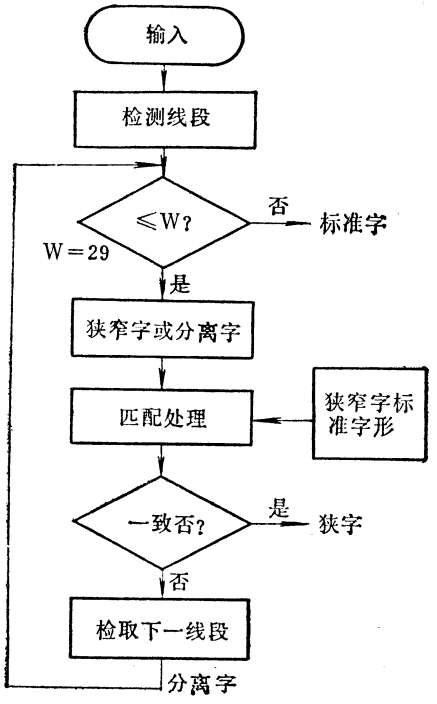


图5-61　字区切分流程

试验中输入字形为 40×40 点的二值整量黑白图形，W值选为 29 点。据报导，试验系统的拒识率为 0.15%，误识率为 0.05%。识别的汉字包括 1840 个明体汉字（打印字体），加上英、数字、假名等总字数约 2000 个，文字尺寸为 3.5×4.0 毫米。

5.3.3　手写体汉字的字形输入

七十年代后期以来，手写体汉字识别技术的研究发展很快。以工整的手写体汉字的1000～2000字数为对象，实验室的成果，最高识别率已达 97～98%。它的识别过程也如光学汉字识别过程相似。字形检测部分可采用与印刷体汉字相类似的光学系统来组成，抽取文字的分类特征，使每一类别的预选字数为 10～100 个，再和预选字作进一步匹配处理，以完成对输入文字的判定。

手写体字形是变化多端的，用没有约束的手写汉字作为识别对象几乎是不可能的。实际上，即使在字母、数字等字量很少的情况下，对手写体字不加约束也是不行的。然而约束过多，将丧失其实用价值。因此，一般只要求输入文字的大小应有一定的范围，目前较多选择在 8～16 毫米的方框内，书写工具大多是最常使用的铅笔或圆珠笔，字体要工整，即笔画数不错，无连笔现象。实验中凡草写的或连笔的字一般都不能识别。

目前这项研究正是纷纷出成果的阶段，尤其对识别过程具有总体性质的特征抽取方式的研究很为活跃。另一方面，各种方法的研究大体上是独立进行的，由于基础条件等的差别很大，对各种方法作出评价尚为时过早。以下以特征抽取方式为重点，对一些典型方法进行讨论。

对于文字的特征，大体有从文字笔道（即文字笔画轨迹）抽取，和从文字背景抽取两种方式，如图 5-62 所示。

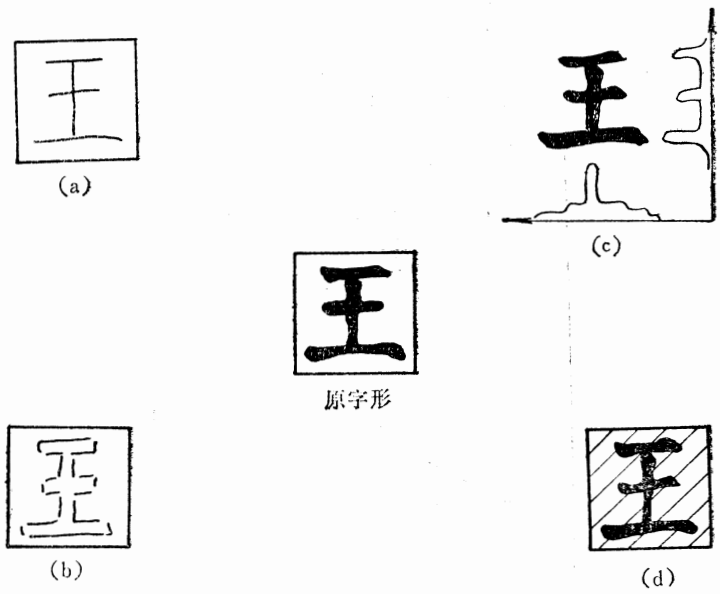


图5-62　抽取字形特征的比较

（a） 笔道特征；（b） 轮廓特征；（c） 笔道分布特征；（d） 背景特征。

图 5-62（a）是把原字形二值化以后，细化成笔道为一点的细线字形，取出笔道，以笔道方向和位置作为特征。图 5-62（b）不进行细化，从二值字形或原字形求出轮廓，将轮廓线分成线段，采用与笔道相同的特征。汉字笔道的曲线很少，几乎都可看成是直线组成的。笔道或轮廓线段常可表示出原字形的构造。图 5-62（c）是把文字笔道的各点按纵向、横向、有的还包括斜向进行分类，将其在各方向上的分布作为特征。图 5-62（a）、（b）、（c）都是从文字笔道抽取特征的方式。图 5-62（a）和（b）保留有笔道、轮廓的位置和角度信息，较好地反映出原字形结构，常称作结构分析方法。图5-62（c）用了特定方向的笔道分布函数，称为笔道分布方式。图 5-62（d）是将文字背景部分的白点相对于文字笔道（黑点）的位置来作为特征的。从已有资料的分析，图5-62（c）、（d）的特征较多用于分类阶段；（a）、（b）的特征则较多用于判定阶段。

一、结构分析法

（一）笔道特征的抽取

把输入文字的二值黑白图形予以细化，从而取出笔道。我们用一种称作井罩法的细化方法来说明细化原理。

如图 5-63，选用一个井字形外罩（图中的 a），它由 9 个网点组成，以其中心点与二值字形中的某点 p_i 重合，来检测它周围 8 点的状况。设笔道通过的点为黑，记作 1，否则为白，记作 0。用 $\gamma(k)$ 表示周围 8 点的值，$k=1\sim8$。p_i 点的值 $C(p_i)$ 用下式定义：

$$C(p_i)=\sum_{k=1}^{8}|\gamma(k+1)-\gamma(k)|$$

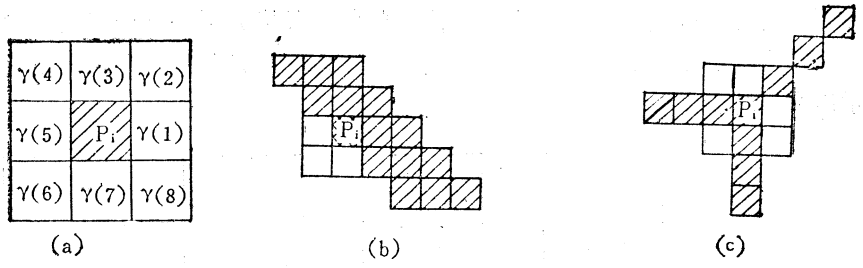


图5-63　井罩法的应用

（a） 井字罩；（b） 细化；（c） 求特征点。

且设 $\gamma(9)=\gamma(1)$

当满足下述条件时，删去 p_i 点：

（1） $C(p_i)=0$ 或 2

（2） $\sum_{k=1}^{\infty}\gamma(k)\neq 1$

（3） $\gamma(1)\wedge\gamma(3)\wedge\gamma(5)=0$　　$\wedge$ 表示逻辑乘

（4） $\gamma(1)\wedge\gamma(3)\wedge\gamma(7)=0$

由于图 5-63（b）中的 p_i 点满足上述四个条件，因此，该黑点 p_i 应被删除。如此反复进行，便可得到文字笔道仅为一点的细化字形。井罩法是一种较常使用的细化方法，不过细化过程还应附加某些规则，以避免可能产生的不同细化结果。

用井罩法还可以检测出笔道的端点、三叉点、四叉点等特征点。如图 5-63（c）那样，已被细化的笔道，$C(p_i)=2$ 时，为笔道端点，$C(p_i)=6$ 时为三叉点，$C(p_i)=8$ 时为四叉点。$C(p_i)=0$ 时，p_i 是孤立点；$C(p_i)=4$ 时，p_i 是笔道上的点，它不是特征点。设有特征点 p_i 和 p_j（$i<j$），检测两点间的 p_{i+1}，p_{i+2}，…，p_{j-1}，当其中某点 p_R 在边界上以一定角度转折时，则将 p_R 作为转折点，它可看作是特征点。

在上述四种特征点中，连接相邻两特征点间的点列，便是笔道线段，再把夹着分叉点背向延伸的笔道线段连结起来，如此得到的线段就是笔道，一直下去可以取出全部笔道（它们在多数情况下与通常笔画相一致）。

图 5-64 是笔道抽取的例子。图 5-64（a）是二值点阵字形，经细化处理得到图 5-64（b），其中的 * 号是特征点，图 5-64（c）是对每一笔道再附上号码。

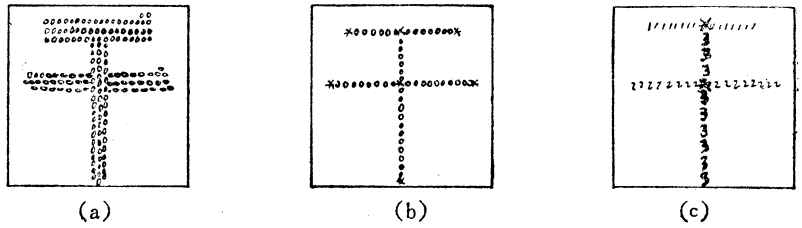


图5-64　笔道抽取例

（a） 二值字形；（b） 细化字形；（c） 取出笔道并编号。

接着可对文字各笔道给出相应的特征。例如可按图 5-65（a）所示的 0～3 四个方

向，作为笔道的方向代码。再以适当的阈值，将笔道长度分成0～2三种，如图5-65（b）。将笔道方向和笔道长度的特征量赋给笔道中点，于是具有这些值的中点的位置分布便表示出文字特征。

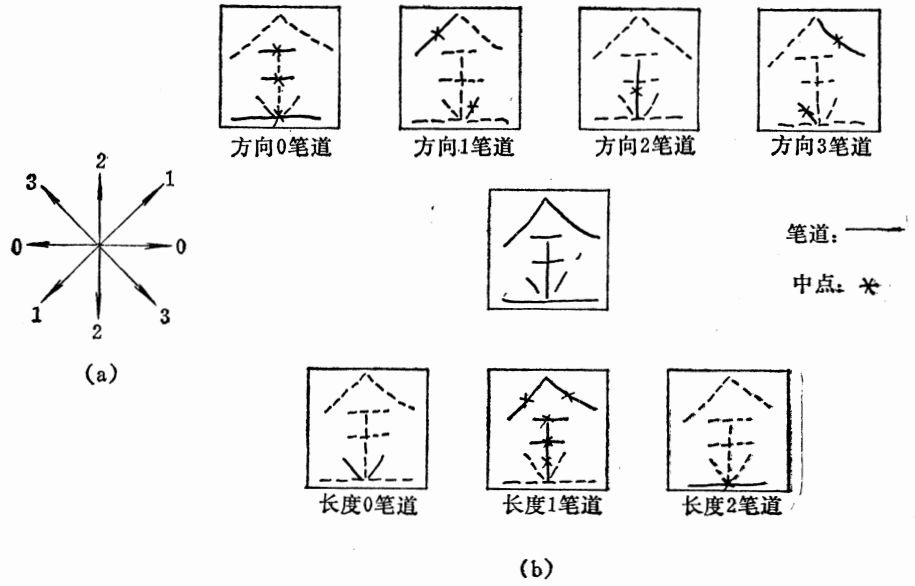


图5-65　笔道特征例：方向按0～3分类；长度按0～2分类。
（a）方向定义；（b）“金”字的笔道特征。

把从输入字形抽取出来的笔道特征，和以类似方法事先作出的标准字形进行匹配处理，对应于相似度最大的标准文字便是识别的结果。匹配过程大多用单纯相似度的方法，也有沿用对印刷体汉字有较好效果的复合相似度方法。

（二）轮廓特征的抽取

这种情形可不作细化处理，用多边形近似笔道轮廓，把多边形的边的始点和终点坐标，以及线段的角度和长度作为特征。这种情况下用得较多的是一种称作relaxation的方法。此法较复杂，我们把它简化成图5-66来说明它的原理。

图5-66　relaxation方法
（a）输入字形轮廓；（b）标准字形轮廓。

图5-66（a）是输入文字的轮廓线；图5-66（b）是标准文字的轮廓线，轮廓线段的方向是按顺时针指向的。不难看出，有可能存在方向相同，而坐标、角度、长度又相近似的线段，如l_1有可能对应于L_1和L_3。因此，在作相似度计算时，应同时再考虑与l_1相

邻的 l_0 和 l_2，先找出 l_0 和 L_0、l_2 和 L_2 的对应关系，才能给出 l_1 和 L_1 匹配较好的结论。重复这一过程，可以求出各线段间的相似度。取各线段间相似度的最大值，再计算出文字间的相似度，从而得到判定结果。

（三）笔道-子图形-汉字的逐层匹配方法

上述两种方法，主要着眼于文字笔道这一最基本的成分，把它们真实地抽取出来，用它们之间的联结关系和位置关系对文字进行识别。在考察汉字的几何结构时，常常还注意到汉字的子图形。不同研究者所给出的子图形虽有差异，但历史上沿袭下来的偏旁部首中的多数，大都列为子图形的基本成分。这些子图形是按照上下、左右、内外等结构形式来组成汉字的，当子图形的数量和形状选择恰当时，GB 2312 中六千多汉字的绝大多数可由 1～4 个子图形组成，只有少数例外。在不同层次上对笔道、子图形和汉字进行分析，使结构分析方法的内容十分丰富，特别在我国，正作为一种主要的识别方法受到广大研究人员的重视。

笔道-子图形-汉字的识别过程可作如下描述：首先，抽取出文字笔道。接着不直接与标准字形进行匹配处理，而按照子图形划分的规则，分析出各个子图形，并对各个子图形按照它们在构成汉字时所在的上、下、左、右、内、外等的位置关系，给出一个形心参量，形心参量便反映出子图形在汉字中位置的特征。最后，根据子图形相对位置关系，可以判断它是一个什么样的汉字。

图 5-67 是一个划分子图形的例子。这里规定把没有笔道相连的、组成汉字的一个个相对独立的部分称作子图形。例如："形"是由"开丿丿丿"四个子图形组成；"字"是由"宀、子"两个子图形组成；"何"由"亻口丁"三个子图形组成等等。

汉　字	子　图　形	子图形数量
形	开丿丿丿	4
字	宀子	2
何	亻口丁	3
大	大	1

图5-67　子图形划分实例

这种方法，在识别用的辞书中，不必给出过于庞大的字形信息量，只需在不同层次上给出较少的特征信息：在笔道匹配时，提供笔道的长度（始点、终点）和方向特征。在子图形辞书中，提供数百个子图形特征，它们可由笔道和笔道数量等来表示。在最后的判定辞书中，提供汉字中各子图形所在位置的特征，它们可由子图形的形心位置来表示。这样一来，大大减少了辞书空间，这不仅减少了系统的负担，也可提高识别效率。

二、笔道分布法

笔道分布函数也称作笔道密度函数，它可以看成是由文字边框的两边各自向着对边扫描时和笔道相交的次数给出的。如图 5-62 的（c），便是文字在两个方向上的笔道分布函数。当进一步考察汉字字形外部特征时，人们注意到汉字的笔道几乎都是由直线组成的，同时笔道的走向呈"米"字形规律，即笔道方向可用 8 个方向来表征，如图5-68所示。它也正是历来的书法家所重视的"米字八法"。

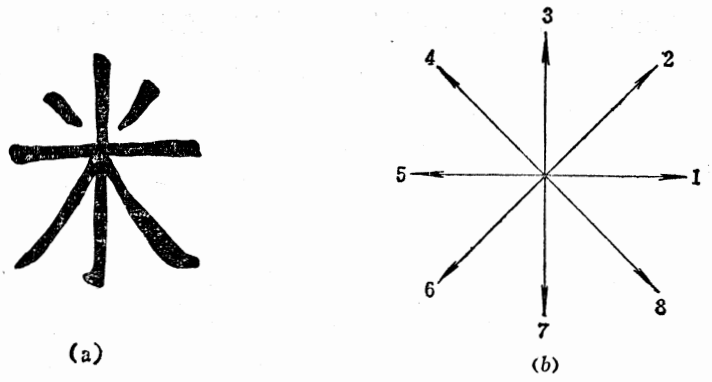


图5-68　汉字笔画米字形规律

（a）“米”字；（b） 八个方向。

如果将文字图形沿着米字形方向，即八个方向投影，这时需要四个投影轴。取文字笔道对投影轴垂直的成分投影，便可求出各笔道在不同方向上的分布曲线，如图5-69所示。把这样得到的文字笔道分布特性，用作对文字分类，也是目前重要实验内容之一。

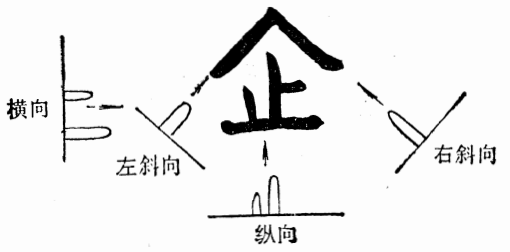


图5-69　笔道分布曲线

下面的方法是目前试验较多的一种方法。如图 5-70（a），在四个方向上找出文字笔道的数目，每个方向的取样点数为32，于是可得到 $32\times 4=128$ 维的特征向量。这种方法的特点是抗污染（线间有黑点）能力较强，但对文字的倾斜和偏移较为敏感。

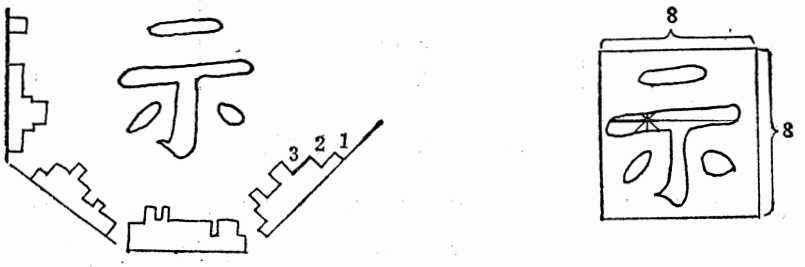


图5-70　对笔道计数例子

（a） 按四个方向对笔道计数；（b） 由笔道内点沿八个方向求至轮廓线的距离。

如图 5-70（b），找出文字笔画线内的黑点沿八个方向伸展到轮廓线的距离，沿 i（$i=1\sim 8$）方向的距离记作 l_i，定义 d_j（$j=1\sim 4$）如下：

$$d_j=(l_j+l_{j+4})\Big/\left[\sum_{i=1}^{4}(l_i+l_{i+4})^2\right]^{1/2}$$

用 d 表示 $(d_1,\ d_2,\ d_3,\ d_4)$，d 便是局部笔道向量。如果把全字分成 8×8 的小块，用每小块全黑点的 d 的平均值作为该块特征，这样每个字便可以表示成 $4\times 8\times 8=256$

维的特征向量。此法的抗污染能力较弱，而对文字的倾斜和偏移并不敏感。

联合使用上述两种方法时，可以得到较精确的识别效果，不仅可用于分类，也可作出判定。不过这种方法相当复杂，向量维数高达512维。

三、背景特征

利用文字背景部分的白点抽取文字特征的方法，曾在印刷体文字识别研究中得到应用，目前正进行应用在手写体汉字识别中的可能性的试验。我们用图5-71来说明在抽取背景特征中的某些不同方法。

（1）是把背景部分（白点）中的一点记作P，由点P向上下左右四个方向作直线，它和文字笔道相交的次数，按各自方向计数，则P点的特征可用一个四维向量表示。如果文字的范围是16×16毫米，并分解成128×128点的点阵字形，则信息量太多。于是将一个字的范围再分成16×16个小块，每小块为8×8点。小块中全白点的特征按方向相加，表示各小块特征。这样全字的特征为16×16×4＝1024维向量，仍然很多。不过，实际试验中，只使用对角线上的32个小块，即文字特征向量仅有32×4＝128维。

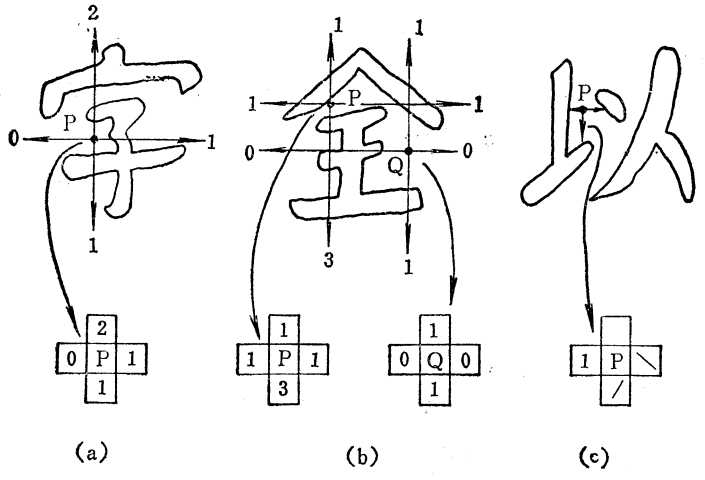


图5-71　背景特征抽取例

（a）由背景点P沿四个方向对笔道计数；（b）对四方向近似垂直的笔道计数，近似45°时，纵横皆计数；（c）抽取倾斜代码特征。

（2）也是由P点向四个方向作直线，对它和笔道相交的次数计数。但只当直线和笔道近似垂直相交时才计数，当近似45°时，则纵横两个方向同时计数。

（3）不对直线和笔道相交的次数计数，而是用由P点向四个方向延伸的直线和文字笔道相交时的角度作为倾斜代码，求出其方向。倾斜代码规定为六种，即：垂直、平行、＋45°、－45°、无笔道、其他。一个文字范围分成3×3＝9个小块，用各小块中白点的倾斜代码分布来作为文字中某个部分的特征。

5.3.4　联机手写体汉字的识别

在汉字的键盘输入方法中，我们曾经介绍过一种笔触式汉字字盘设备，输入时不用打键，而是检测笔在字表上的位置。由于每字在字表上的坐标位置是单一的，因此，它的位置坐标便可作为该字的代码而被输入。在坐标型平板字表输入原理的基础上，如果

把它的X、Y坐标的分辨率做到4～10点/毫米，这样的装置便可用作图形或文字输入。联机手写体汉字的字型检测设备常采用这样的平板型图形输入板。

这一课题的研究是七十年代中期广泛开展的。联机状态下，由于系统的实时处理功能的支持，在输入板上工整地书写汉字时，可即时地抽取如下信息：a）笔头在模板上移动轨迹的坐标按产生的次序作为时间序列抽出；b）每一文字是由许多笔画组合而成，根据笔头接触和离开模板的信息可分离出各个笔画；c）书写时，笔画是有序发生的，这种笔顺信息同时被抽出；d）每个文字是限定在模板一定范围内的，对每个字的区分很方便。这样，便可以即时地得到如文字的笔画、一个字的笔画数以及它们的笔顺等信息。

如果用一个有序的基本笔画的组合来作为文字的表征，那么它即为文字的分类特征。假定在分析汉字结构特性的基础上，确定了如图5-72所示的笔画为基本笔画，每个汉字就是由它们组合而成，笔画形状、笔顺和笔画数就是汉字的特征。如“大”字是由三个笔画按照“一丿㇏”的笔顺组合而成，“学”字是由七个笔画依照“丶丶㇀一了一”的笔顺组合而成，等等。标准字形特征亦是这样做出，经过和标准字形的比较，便可对输入文字作出判定。自然，这里存在着特征相同但文字相异的情况，如“天”和“夫”同是以四个笔画按“一一丿㇏”笔顺组成，“太”和“犬”同是以四个笔画按“一丿㇀丶”笔顺组成等，这就需要作进一步的判别。

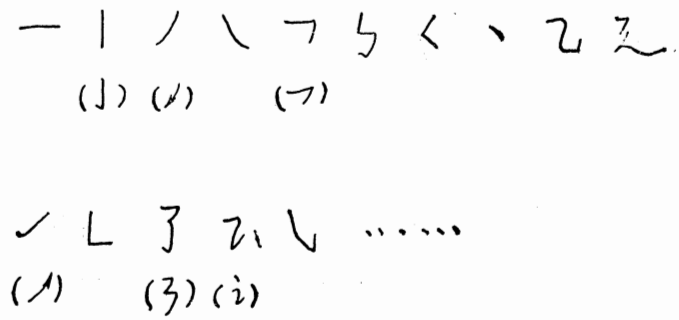


图5-72　汉字基本笔画示例

另有一种方法是尽量压缩基本笔画的种类，如只取点、横、竖、撇、捺五种笔画，并对它们在平面坐标轴上的投影值的界限作出规定，所谓“点”，在X、Y轴上的投影值应小于规定的阈值；“横”在X轴投影大于某值，在Y轴投影小于某值；“撇”在X、Y轴投影均大于某值等等。这样的笔画，计算机在接收书写信息时，只需作简单的判断，便可依次识别出来。进一步是判别子图形，子图形是较少基本笔画的有序组合，而汉字是由子图形按左右、上下、内外…，等关系组成的。这里我们不妨把子图形理解成通常的偏旁部首。若以左右偏旁为例，左偏旁应是最先写出的，输入时计算机及时判别笔画，记录笔画数，与偏旁辞书作比较，对偏旁作出相应的判断。除偏旁外，其余部分再抽取笔画数及其有序组合，这时由于特定偏旁包含的相同笔画数及其有序组合的字是很少的，与特定偏旁的辞书作比较，便可对输入文字作出判定。这种方法也是在笔道和汉字间增加了判别子图形这一层次，并且可以得到系统实时处理的支持，因此较为易于实现。目前的试验中，我国的许多研究人员对此方法作了大量实验，并取得了较为满意的结果。

如上所述，汉字的书写笔顺是作为特征取出的，但是汉字笔顺不能认为是规范化的，由于人们长期的书写习惯，笔顺差异也是实际存在的。为排除笔顺的影响，对不依赖于笔顺信息的联机手写体汉字识别系统的研究，进展也很快。日本电电公社武藏野通信研究所曾发表一个研究结果，它是在商品化的文字处理机上附以专用的识别硬件，相应软件和字表型输入板组成的。对1～23画的文字，识别所需的时间为6.1～475毫秒/字，通常用笔书写的速度为1.5～2.5秒/字，因此，系统能实时识别。在包括片假名等共2057个字的识别中，识准率达99.5%。它的识别方法如下：在字表式书写模板上的15×15毫米的方框内书写汉字，每10毫秒，接收分解成4点/毫米的X-Y坐标序列，再以适当的时间间隔取样，进行规格化处理。然后，把各笔画的起点、终点坐标以及它们之间的点数N作为特征取出。在文字笔画总数为6以下时取$N=6$，笔画总数为7以上时取$N=3$。最后，用计算距离的方法，计算标准字形各笔画和输入字形各笔画间的距离，把笔画间的距离总和作为文字间距离，取文字间距离最小的标准字形作为判定结果。这种方法没有考虑笔顺，不同的笔顺，并不影响结果。该公司目前还希望在笔画数不准确的情况下，也能得到满意的结果。分析表明，笔画数不准确的情况几乎只有三种，即：笔画数少一画（两笔连写成一笔）的占74%，多出一画的占14%，少2画的占10%。这一系统的发表，表明了联机手写体汉字识别系统已达到实用化水平，正在显示出作为文字处理机的辅助输入手段而被商品化的可能性。

5.3.5 小结

汉字字形识别方法的研究，是汉字信息处理技术向深度和广度发展的标志之一，也是这一领域力图突破汉字输入难题的一大重要类型，它在国内外普遍受到人工智能、语言文字、数学、工程学、计算机科学等各学科领域内的专家们高度重视。目前就单一字模印刷体汉字而言，技术问题已大体解决。对多字模印刷体汉字的研究也时有报告发表。手写体汉字识别技术的研究，近年来非常活跃，仅日本发表的对识别技术具有总体意义的特征抽取方式就达数十种，虽然还有许多问题需要解决，但一般预测，八十年代将至少达到实用水平。也许在某种程度上由于这一原因，以致有必要考虑是否立即把印刷体汉字光学识别设备向商品化推进。联机手写体汉字识别，在精度方面已基本达到实用化水平，它有可能在普及化的文字处理机和办公用计算机系统中较早地获得应用。

我国这方面的工作虽然开展较晚，但进展很快，特别在手写体和联手写体汉字识别方面都取得了较好的实验成果。不过，要推进到实用化、商品化的程度，还面临许多课题。例如，（1）与识别技术有关的各种外围装置技术的研究和开发；（2）建立汉字识别技术坚实的理论基础；（3）建立标准字形和不同等级标准的汉字辞书；（4）对低品位的输入文字（例如由印刷质量、纸质、手写变形等引起）如何辅之以语法和文脉信息进行识别；（5）大力开发汉字字模识别用的应用软件；（6）降低设备成本以利于这项技术的推广应用等。

第六章　汉字字形发生器

汉字信息处理系统除了需要具有汉字信息的输入功能和对这些信息进行处理的能力以外，还必须能够输出汉字字形信息。为了输出汉字字形信息，根据不同的要求有各种形式，粗略地可以把它们分为两类。一种是“精密汉字字形输出”。例如，由计算机控制输出的可供印刷制版用的精密汉字字模输出形式，不仅要求有高质量的字形，并能输出各种字体和不同的汉字尺寸，而且还往往要求能组成复杂的版面，有的甚至要求能兼有图形、照片的输出功能；另一类是由计算机控制在各种印刷机上印出汉字，或在荧光屏上显示汉字。这一类的汉字字形信息输出形式，使用范围十分广泛，适用于统计制表、情报检索、文件或档案管理、事务处理、企业管理，以及数字通信等各种场合。它的特点是对字形的要求不太高，而主要是要求字形清晰，在清晰的前提下力求美观。它实现起来比较容易，是一种经济实用的输出形式。

本章所讨论的汉字字形发生器属于后一种应用目的，也称为“通用型汉字字形发生器”，简称“汉字字形发生器”。

汉字信息的输出同一般计算机的字符信息输出相比较，在工作原理方面，无论是软件还是硬件，都有许多共同的或相似的地方。但是，前者有它自己的特点。有必要研究汉字信息输出技术中对汉字字模的特殊要求和特殊处理方法。其中就包括本章所要讨论的汉字字形发生器。习惯上也常把它称为“汉字字模存储器”，或者称为“汉字字模库”、“字模库”等。它是汉字信息处理技术中一个比较重要、也是比较特殊的问题。

6.1　汉字字形的数字化表示

6.1.1　汉字字形的特点

国家标准《信息交换用汉字编码字符集——基本集》共收汉字6763个。其中第一级汉字3755个，第二级汉字3008个。这一标准中所收的汉字都是“正字”，也就是说，没有包括繁体、异体、旧体等“非正字”。即使不考虑非正字，基本集所收的6763个汉字也是不完全的，还有许多字没有收集进来。有一些生僻字使用的频率很低，但只要用到它，就一定要有这个汉字的字形。因为按照我国使用汉字的习惯，是不能用符号或其他信息代替汉字的，所以在实际使用时一定要设法“备足”所需要的汉字字形。这也是汉字字量大的原因之一。

除了字量大这一特点外，汉字的字形也复杂。可以从各种不同的角度来说明汉字字形的复杂性。

（1）尽管每个汉字都是由横、竖、撇、点、折、捺、提等“笔画”构成的，但这些笔画的位置、大小（长短）、方向等往往错综复杂，较难找出统一的规则。

（2）近几年来，汉字字形的研究人员对汉字字形的部件构成规律作了大量的研究。大多数汉字可以看成由若干“组字部件”构成。例如“购”字可以看成由“贝”和“勾”

构成，“赔”字可以看作“贝”和“音”构成。这里所谓“组字部件”就是指贝、勾、音等部件，近似于常说的偏旁、部首，但又不全是偏旁、部首。为了和偏旁、部首相区别，这里采用“组字部件”这样一个直观的名词。有人将其称为“字素”、“字根”、“字元”等，其意思都是一样的。假如每个汉字都可由某些组字部件拼合起来，那末汉字信息处理中的输入输出问题就比较容易解决了。遗憾的是这样的拼字规律较难得到。其原因如下。其一，这样的部件本身就很多，而且有些字的部件很难分离。举例来说，“里”字既可以看成“田”和“土”的合成字，也可以看成一个不可分的“部件”。类似这种情况的字有好几百个，而且往往是常用字。其二，组字部件的使用频度相差很大。有的部件使用频度很高（象“口”、“木”、“亻”、“氵”等），也有的部件使用频度很低（例如“𠂇”、“凸”等），我们无法按使用频度来取舍。这是因为，有的部件即使只有在极少数情况下使用，也很难舍去这些部件。其三，组字部件的位置、大小也相当复杂。例如“篇”、“遍”、“匾”、“编”这几个字，都是有由两个部件拼起来的，可以看出，作为拼字的组字部件在这几个字中的大小、形状、位置是不相同的。其四，还有相当数量的多拼字更增加了拼字的困难。多拼字占全体汉字的20％。有许多字很难确定它究竟是多拼字还是两拼字。例如“屣”字，可以分解为“尸”“彳”“走”，也可分解为“尸”、“徒”两字相拼。从这些例子中可以看到，用组字部件来拼成汉字是相当复杂的。由于上述这些原因，故至今还不能找到一种非常理想的组字规则。因为文字和语言一样，是逐步形成的，它具有社会性和历史性。我们很难用形式上的几条“法则”或者“规律”来强行统一，只能承认它的复杂性。

以上讨论了汉字字形的复杂性。另一方面我们说汉字字形也还有一些处理上的方便特点。最突出的一点就是汉字是“方块字”。所谓方块字就是指每一个汉字都一样大小，（指在同一种字号尺寸下）无论笔画多少都可以放在一个固定大小的方格之中。所以说，汉字的字形很复杂，但是它的“大小划一”。每一个汉字在计算机中可以看成是一个同样大小的“图形”。

6.1.2 汉字字形的数字化

因为不论汉字字形的笔画多少，都可以写在同样大小的方块中，从而可以把这一方块划分为许多小方格，组成一个“点阵”。每一个小方格是点阵中的一个点（即组成字模笔画的最小单位位点）。例如，对于一个16×16的点阵，是把一个方块横向分成16格，纵向也分成16格，从而有256个小方格，也就是说，该矩阵有256个“点”。点阵中每个点可以有“黑”、“白”两种颜色。用这样的点阵就可以描出汉字的字形，此称为“汉字点阵字形”。

我们很容易用二进制数字来表示点阵。如果用二进制的“1”来表示黑点，用二进制的“0”来表示白点，那么一个16×16的点阵字形，就可以用一串二进制数（256位二进制数）来表示了。这种方法，我们称为“点阵的数字化”。

例如，对于图6-1中的“次”字点阵就可以用256位二进制数来表示，由左向右，从上到下逐点地记录、组成一个二进制数串。因为二进制数写起来太长，使用不方便，故常采用“八进制数”或“十六进制数”来代替二进制数。以十六进制数为例，一位十六进制数实际上就是四位二进制数。所以在点阵中，并列的四个点就可以用一位十六进制数来表示。

表 6-1 表示十进制数、二进制数、十六进制数和点阵图象的关系。

用十六进制数表示一个汉字的16×16点阵只需要用64个数字就可以了。仍以上述“次”字的点阵为例。它的点阵数字化信息可以用如下一串十六进制数表示：00, 80, 00, 80, 20, 80, 10, 80, 11, FE, 05, 02, 09, 44, 0A, 48, 10, 40, 10, 40, 60, A0, 20, A0, 21, 10, 21, 08, 22, 04, 0C, 03。

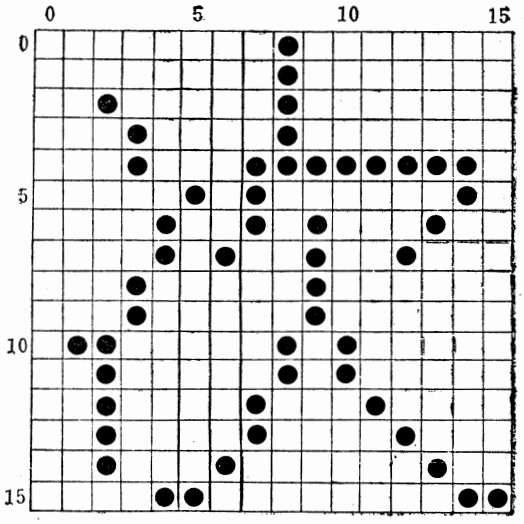


图6-1 汉字点阵字形

在计算机技术中常用“字节”(byte)这样一个名词，一个字节通常是指八位二进制数。所以，一个16×16的字形点阵需要用256位二进制数表示，也可以说由32个字节的数字来表示。

一个字节的内容可以用八位二进制数来表示，也可以用两位十六进制数表示。例如，若某一字节的内容是“01011101”，则可以写为“5D”。为了特别表示这个数为十六进制数，常常用一个括号加一个小的注脚“16”表示，例如十六制数5D可写成$(5D)_{16}$。这就是说，下列关系式成立：$(01011101)_2=(5D)_{16}$。在不至于混淆的情况下，经常不必特别指明这是十六进制数。

表6-1 十进制数、二进制数、十六进制数和点阵图象的关系

十进制数	二进制数	十六进制数	点阵图象
0	0 0 0 0	0	○○○○
1	0 0 0 1	1	○○○●
2	0 0 1 0	2	○○●○
3	0 0 1 1	3	○○●●
4	0 1 0 0	4	○●○○
5	0 1 0 1	5	○●○●
6	0 1 1 0	6	○●●○
7	0 1 1 1	7	○●●●
8	1 0 0 0	8	●○○○
9	1 0 0 1	9	●○○●
10	1 0 1 0	A	●○●○
11	1 0 1 1	B	●○●●
12	1 1 0 0	C	●●○○
13	1 1 0 1	D	●●○●
14	1 1 1 0	E	●●●○
15	1 1 1 1	F	●●●●

汉字字形数字化以后，就可以把字形转化为一串数字。这一串数字称为“汉字字形的数字化信息”，简称为“字形信息”。

一个24×24点阵的汉字共有 576 个点。一个32×32点阵有1024个点，其信息量分别比16×16点阵的信息量大2.25倍和 4 倍。显然，点阵的信息量越多就能把字形表示得越精确。为了和计算机普遍采用八位二进制信息为一个字节这样一个约定相符合，字形信息多数采用16×16，24×24，32×32这样一些 8 的倍数点阵。也有少数采用16×18，20×20等非规格化字形点阵的。

从实际使用上来看，16×16是最简单的汉字字形点阵，除少数笔画特别复杂的汉字需要做些“变通”以外，它基本上能表示所有的汉字字形。24×24点阵一般可以做到横细竖粗，带笔锋，从而逼近于汉字宋体字形。

在16×16字形点阵中，一般总是在点阵的左侧或者右侧留出一列空白，做成15×16的点阵。总信息量仍是16×16，但是在边上有一列信息全为 0 。这是因为，输出信息时需要有一定的“字间隔”（横排汉字的每个字和字之间需要有一些空隙）。而“行间距”一般不在字形发生器中留空白，而由输出设备另外用程序或硬件方法添加。

6.1.3 汉字点阵字模的制作

为了获得美观清晰的汉字输出，一件很重要的工作是如何制作汉字点阵字模。也就是说怎样把汉字字形数字化。

我国目前有少数单位研制用计算机自动制作汉字字模的设备，但大多数还是用手工制作。手工制作的过程大致可分为两步。先是在印好的方格纸上由人工照字形“描点”，这一步也可以说是在纸上先把字形数字化。然后，第二步通过各种方法把字模信息输入给计算机。可以用光笔输入、笔触式字盘输入、光标移动或用擦、写键输入等多种办法。然后，由计算机把输入的字模信息按一定的方式存储起来。

显然，字模制作是一件相当繁重的工作。要把几千字的字形一个一个地“制作”，不是件容易的事。但是，要特别强调的是，这种制作是一次性的，也就是说全部字模只需要做一次，就可以永久性地存储在字形发生器中，没有必要每种汉字系统都需独自设计、制作一套汉字字模。目前我国正在制定16×16、24×24汉字字模点阵的国家标准，待正式颁布后，一般说来都不必重复进行汉字字模的制作工作了。

初看起来，字模制作是一件不太难的事。其实，这项工作也很有讲究。字形美观与否全在对字模描点这一道工序上。图 6-2 表示同一汉字的点阵化字模可以有不同结果。显然，有的字形好看，有的字形很差。其中，图6-2(a) 是一半放两点水，一半存放“欠”字，其结果是字形很不匀称。图6-2(b) 是写两点水太简单(随手一画)，其结果很不理想。图6-2(c) 的字形较好。此外，即使能点好某些字，但是否能点好所有的字，使整套字模都有统一的风格，也还是一个问题。所以，多数有经验的制作者建议最好能

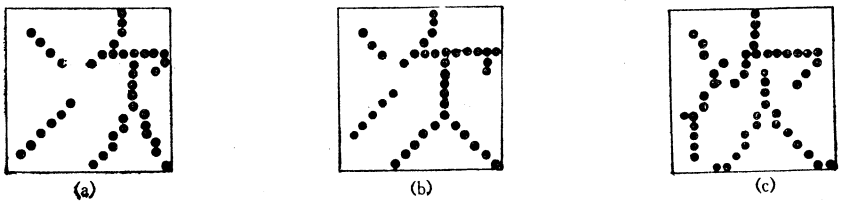


图6-2 同一汉字的不同数字化字模质量

参考一种标准字体。我国印刷出版行业用的“字模”因为已经过长期使用，并被我国广大读者所熟悉，故应该采用它做为数字化字的“楷模”，尤其是用印刷行业的“书版宋体”字模稿作为通用型数字化字形的字稿更为合适。基本上照字模稿点字，局部地方要做些调整，应使数字化点阵所表现的汉字字形保持书版宋体字所具有的整齐、美观的特色。即使是16×16点阵的字模，也应参照字模稿来描字。这是因为，尽管15×16的规格点阵的位点少，无法表现横细竖粗的笔画，每笔难以带笔锋，但是“基本参照字模稿、局部地方作调整”仍然是制作此数字化字模的原则。汉字字形的美观是由汉字本身的结构特点所决定的。每个笔画，每个部件的大小、高低、宽窄都有一定的规则。事实证明：凡是按照印刷行业所用字模稿制作的数字化点阵字模，一般都经得起推敲，字形的美观和准确性都比较有保证。图 6-3 所示是书版宋体汉字字模稿。

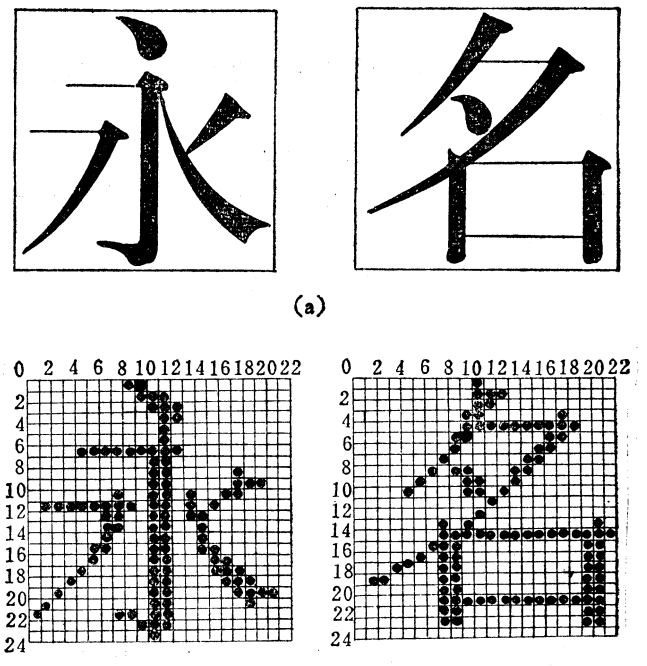


图6-3　书版宋体汉字字模稿及其数字化字模

（a）书版宋体汉字字模稿；（b）用这种字模稿制作的数字化字模

6.2　汉字字形发生器

汉字字形经过数字化以后，就可以存放到存储器中构成汉字字形发生器，以便向各种输出设备提供字形信息。这种字形信息有的不必进行“加工”，可以直接供输出设备使用；有的还要进行必要的加工处理才能使用。此外，各种不同的输出设备要求提供的字形信息速度差别也很大。因此，实际上可采用的字形发生器的组成方案也是多种多样的。

6.2.1 存储器概述

从计算机的角度来考虑汉字字形发生器是比较简单的。它只是一个存储字形信息的存储器。对于计算机的存储器来说，要考虑的问题无非是采用什么样的存储元件、存储容量有多大、存取时间要求多快、采用何种存储方法，以及如何组织存储器等。

计算机的主存储器在五、六十年代大多数采用磁心作为存储元件。七十年代以后随着半导体MOS技术的逐步成熟，大规模集成电路以它的可靠性高、功耗低、体积小，以及便于维护等优点取得了优势，目前它几乎完全取代了磁心存储器。

在计算机的外部存储器方面，磁盘存储技术发展快，应用广。它是较理想的大容量后援存储器。

下面我们介绍几种大规模集成电路（LSI）存储器芯片的特性。

（一）随机存取存储器RAM（Random Access Memory）

RAM可在很大的地址范围内随机地“存”“取”信息，使用方便，存取速度快，所以往往是理想的计算机内存或快速存储器。在字形发生器中常用RAM组成“缓冲存储器”或者用作临时性的“转存”存储器。RAM一般又可以粗分为两大类。一类称为“静态RAM”（Static RAM）；另一类称为“动态RAM”（Dynamic RAM）。在静态RAM中存入信息以后可以一直保留，直至下次被改写或断电后才消失。动态RAM在写入一次信息以后，经过若干时间（约几毫秒至几十毫秒）如不再访问这一存储单元，在这个单元中所存放的信息就可能被丢失。为了让动态RAM中所存放的信息不致丢失，需要不断地“再生”。因为读一下就能起再生作用，所以实际的做法是不断地去读它。用动态RAM所组成的存储器需要再生电路。在动态RAM存储器中，每个存储单元通常是由一个MOS管和一个电容组成，工艺也较简单，所以容量大，成本低。虽然需要有再生措施，但仍被广泛地采用。

表6-2列出了几种典型的RAM芯片的参数表。

表6-2　几种典型的RAM芯片参数表

类型	型　　号	存储规格 地址数×位数	存取时间 （毫微秒）	功　耗 （瓦）	引脚数
静态	2114	1024×4	450	500m	18
	2114L-1	1024×4	300	357m	18
	2114-2	1024×4	200	475m	18
	2147	1024×4	70	1.0	18
	2141-3	4096×1	150	1.2	18
	HM6116-4	2048×8	200	1.0	24
	HM6116-3	2048×8	150	1.0	24
	HM6116-2	2048×8	120	1.0	24
动态	TMS4027	4096×1	250	470m	16
	4116-300	16384×1	300	450m	16
	4116-250	16384×1	250	460m	16
	MK4816	2048×8	300	150m	24
	4164-200	65536×1	200	250m	16
准静态	Z-6132	4096×8	300	250m	28

目前，还有一种叫做准静态的RAM芯片（quasi static RAM)。它实际上是动态RAM，也是单管存储，但是在片子中能自动再生，使用者可以不必考虑它的再生措施，就同用静态RAM一样，故称“准静态”RAM器件。

（二）只读存储器ROM（Read Only Memory）

所谓只读存储器是指这样一种存储器，在一般情况下，它只能读出不能写入，一旦写入后便可长期使用，即使断电，也不会破坏已写入的内容。这种只读存储器适于用作字形发生器。这是因为，汉字字形发生器中的字形信息往往就是“一次写入，长期使用”的。只读存储器也有多种类型。

通常所称的ROM也称为“掩模ROM（mask ROM)”。这种存储器一般由工厂先做好一种特殊的掩模，一次做定，出厂时它所存储的信息内容已经固定。它适合于大批量生产，而且成本很低。因此它是目前存储容量最大的ROM器件，每片已可做到一百万位，从而特别适合制作字形发生器。在我国实现汉字字形标准化以后，采用掩模ROM做成标准的字形发生器是很有前途的。

PROM称为可编程的ROM（Programmable ROM)，产品出厂时，存储器内所有的存储单元内容都是全“1”。用户可以用程序决定片子中每一个单元中应存放的内容，写入一次，就可永远使用。每次写入的单元数可多可少。但每个地址只能写入一次，不能修改。因为是一次性的“写”入，所以编程写入时要特别小心。

EPROM可擦可编程只读存储器(Erasable Programmable Read Only Memory)，这是一种最常用的只读存储器，因为它可以编程，即出厂时是“干净”的（全部信息内容为全1)，可以由程序写入，写入一次后可长期保存（约十年）不被破坏。它和PROM不同的是，若有必要，可以把已写入的内容擦去，重新再写。擦写次数可达上百次。所以使用十分方便。编程时可分多次写入，每次可以只写一个单元，也可以整片都编程。擦除的方法也比较容易，用紫外线灯照射就可以擦除。这种擦除方法简单方便、但擦除后整片存储器的内容全为1，不能局部改写。

由于EPROM使用方便，又可擦可写。所以对这类片子的需求量也较大。从技术上来说，这类片子也是各厂家重点开发的对象。近几年来，每块EPROM片子的存储容量以8K位，16K位，32K位，64K位，128K位……很快地增长。对于当前每片64K位的片子，每K位的价格已比以前的16K位和32K位片子相应的价格更便宜。目前我国许多汉字字形发生器也都采用这类芯片。

一些典型的PROM，EPROM器件的参数如表6-3所列。

E^2PROM电可擦可编程的只读存储器（Electrically-Erasable Programmable Read Only Memory)，是一种可擦可写的只读存储器。但和EPROM有一点不同，它不是用紫外线照射的办法来擦除的，而是利用电（流）来擦除的。这有两个优点，（1）不必备有专门的紫外线擦除灯（Eraser)，擦除时也不必拔下片子。（2）因为是电擦除，故允许局部擦除，局部修改而不象EPROM那样，紫外线一照，全片都被清除。就这方面来说，性能比EPROM优越。目前这种片子的品种较少、每片的容量不太大，价格也比较高。估计今后会得到广泛的应用。用这种片子在汉字字形发生器中可以用来存放一些临时性的、需要随时更改的字形信息，故有较大的灵活性。

以上只是简单的介绍了存储器芯片的一些概况，显然详细论述它们的特性和工作原

表6-3 典型的PROM，EPROM器件的特性

类型	型　　号	规　　格 地址数×位数	存取时间 (毫微秒)	功　耗 (瓦)	引脚数
PROM	MCM7642	1024×4	70	700m	18
	MSM3770	1024×8	600	745	24
	MCM7680	1024×8	70	750	24
	2136	2048×8	450	1.0	24
	2332	4096×8	300	1.0	24
EPROM	2708	1024×8	450	430m	24
	2716	2048×8	450	520m	24
	2732	4096×8	450	750m	24
	2732-250	4096×8	250	750m	24
	2764	8192×8	450	525m	28
	MCM68764	8192×8	450	800m	24
	27128	16384×8	450		28

理不是本书讨论的范围。存储器芯片的种类很多，就其工作方式来分有静态RAM、动态RAM、准静态RAM、ROM、PROM、EPROM、E^2PROM等。存储器的容量也有许多种，每片4K位，8K位，16K位，32K位，64K位等芯片的读出速度一般有450毫微秒，300毫微秒，200毫微秒，150毫微秒等，有些类型的芯片存取速度更高，可小于100毫微秒。大规模集成电路技术的飞速发展使得存储器芯片的集成度（每片的存储容量）、可靠性、速度和价格等各方面的指标不断上升，总的趋势是每片的容量不断增加，成本逐年下降。

6.2.2 汉字字形发生器的存储容量

汉字字形发生器的存储容量是一个值得讨论的问题。决定汉字字形发生器的存储容量有两个因素，一是该汉字字形发生器所收的汉字字形总数；另一因素是每一个汉字的点阵规格。就我国目前使用汉字的情况来说，一般的汉字信息处理系统备有四千个左右的汉字字形就已经基本够用了。根据汉字使用频度的统计，国标一级汉字3755个已占使用汉字的99.9%以上。如果字形发生器备有6000～8000个汉字字形，就能满足绝大部分汉字信息处理系统的使用要求。只有极少数系统需要用到大量的生僻字、繁体字和异体字。这种特殊的系统可能需备二万以上汉字字形，但这是属于较少的使用要求。

常用汉字字形发生器存储容量的计算结果如表6-4所列。

表6-4 几种汉字数量和点阵规格不同的汉字字形发生器所需的存储容量

汉字字形的点阵	每个汉字字形信息的存储量（字节）	1024个汉字字形信息的存储量（K字节）	4096个汉字字形信息的存储量（K字节）	8192个汉字字形信息的存储量（字节）
16×16	32	32	128	256K
24×24	72	72	288	576K
32×32	128	128	512	1M

就目前的技术水平来说，汉字字形发生器仍是一个存储量比较大的存储器。以16×16点阵的4096个汉字字形发生器为例，如果采用每片存储容量为16K位的芯片，则共需

用64片，即使采用每片存储容量是64K位的片子，也要16片。当然，假如现在已经有百万位的存储芯片，那就只要一片就够了。所以说存储容量大小的讨论，不能绝对化，要根据具体的技术指标来讨论。

6.2.3 汉字字形发生器的缓冲存储器

在各种汉字信息处理系统的总体设计要求下，汉字字形发生器可以有不同的结构。除了用简单的直接存取方法以外，还可以采取另外的一些构成方法。主要的方法有：

一、多级存储方案

对许多汉字信息处理系统来说常用字三、四千字已足够应用。但是备用字的数量往往很多。有时还需要临时制作字模。如果把所有备用字也存入字形发生器内，字形发生器的存储容量就需很大。而备用字的使用频度相当低，有的甚至是备而不用的，因此，若都将它们存储在一个大的字形发生器里，则不太合理。多级存储方案是一种有效的解决办法。

多级存储方案典型的模式是由一个ROM存储器存放若干常用汉字的字形，另有一个后援存储器（一般采用磁盘存储器），其容量较大，可以存放较多的备用字，也便于存放需要临时制作字模的生僻字，这样的结构有较大的灵活性。

二、字形发生器的共享方案

因为字形发生器存储量较大，需花一定的硬件成本。但是，一般说来，访问字形发生器的速度并不要求很高。为了有效地发挥字形存储器的作用，降低整个系统的造价，可以采用共享的方案。即一个字形发生器可以同时为若干台设备、或若干个终端所使用。

以上两种设计方案比较适合我国使用汉字的实际情况，所以在国产的汉字信息处理系统中被广泛使用。在采用多级存储方案或者共享字形发生器的方案中，都需要一个缓冲存储器。在直接存取的方案中也常常采用缓冲存储器。图 6-4 表示了几种汉字字形发生器结构的框图。

汉字字形发生器中缓冲存储器有重要的作用，一般它是由RAM存储器芯片组成。它

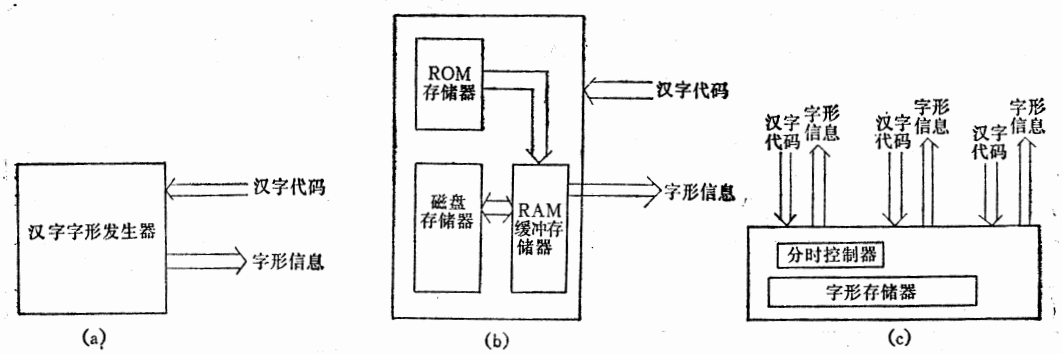


图6-4 几种汉字字形发生器的构造框图

（a） 直接存取的汉字字形发生器；（b） 多级存储方案的汉字字形发生器；

（c） 多个用户共享的汉字字形发生器。

是直接面向使用者的，这里所说的使用者就是指打印机或CRT显示器等输出设备。缓冲存储器的容量可以用得较小，一般只存放256或512个汉字字形。由于它是RAM存储器，所以可以不断地更新，把当前需要使用的汉字字形调进来以供输出。在兼有图形显示或图形打印功能的系统中，这个缓冲存储器也可以存放图形的点阵信息。

缓冲存储器是汉字字形发生器的ROM存储器或后援存储器同使用者之间的纽带和桥梁。此外，它的存取时间较短，便于控制和变换。所以，使用缓冲存储器能使整个系统各部分的配合有较强的适应性。

以下讨论如何决定缓冲存储器中存放的内容？什么时候，在什么条件来更新缓冲存储器的内容？

我们通过几个具体的例子来说明缓冲存储器的更新问题。

〔**例 1**〕 一种汉字终端，采用软盘存储器存放汉字字形信息。约有8000个16×16点阵的汉字字形点阵，全部存放在一片200毫米（8英寸）的软磁盘上。同时有一个64K字节的缓冲存储器，直接面向使用者——CRT显示器。这个64K字节的RAM存储器，同时也就是显示器的再生存储器，存放需要显示的汉字字形信息。此外，在终端的主存中有一个2K字节～3K字节的显示文件区，用来存放当前需要显示的汉字代码，它是屏幕编辑的依据。显示文件中一个代码同缓冲存储器中的一个点阵（16×16）信息是一一对应的。它的更新方法比较简单。

（1）清除状态，也是初始状态。64K字节内容全部清除为0，显示文件区也一律清除为0。

（2）需要显示一个汉字时，要先查一下显示文件，看文件中是否已经有过这个字了。如果有，则表示点阵信息一定在RAM存储器中，可以不必访问磁盘存储器，但要在缓冲存储器中“拷贝”一次（将先存入的这一字形点阵传输到新的位置上去），并保留原有的点阵信息。同时，在显示文件中登记这一汉字的代码。如果在显示文件中查不到这个字的代码，则表示缓冲存储器中还没有用过这一汉字的字形，需要从软磁盘中调入。调入以后，也要在显示文件中登记这个汉字的代码。

（3）要在屏幕上删去某字，先删去显示文件区的代码，同时也删去缓冲存储器中的点阵信息。

总之，这种方法的特点是，永远保持显示文件区和汉字点阵的一一对应关系。重复使用的汉字不需要再次访问磁盘，而且采用缓冲存储器内部传输的方法。不用的字，可随时删去。显示文件是随机的（即每次要显示什么汉字是随机的），所以说这里的查找只能采用逐个字查找的办法。在显示文件区中每次要从头开始对汉字内部码作逐个比较。在显示文件不太长时，查找比较快，愈往后愈费时间。该终端的屏幕上可以显示1300多个汉字，如果要在最后显示一个汉字，这个字又和已经显示的一千多字都不一样，则要逐个比较一千多次，最后仍需要访问一次磁盘存储器，很费时间。当然，这是一个比较极端的情况。一般地说，如果显示文件已有k个字，比较$k/2$次能查到重复的字（这是一种平均估计的概率），那么，显示N个字的文件，平均就需要比较$N(N+1)/4$次。假定有一个1000字的文件，其中有200个不同汉字，有800个字是重复字，则显示这一文件需要访问200次磁盘、查表约需比较25万次，缓冲存储器内部传输800次。折合所花费的时间，访问磁盘约需60秒，查表需要用10～20秒，缓存器内部传输时间约1秒，总

的响应时间大约是一分多钟。由此可见，时间主要花费在访问磁盘上了。从这一估计中可以看到，由于采用了缓存器内部传输的方法，可以减少访盘次数。所花费的查表和拷贝的时间是值得的。如果没有这一措施，显示一个一千字的文件，响应时间则要五分多钟。

〔**例 2**〕 某汉字终端的字形发生器，4096个汉字，每个汉字为16×16点阵，存放在EPROM存储器内。另外备用汉字 3000～4000 字的字形信息存放在软磁盘上。有一个可容纳 256 个汉字的缓冲存储器（8K字节的 RAM 存储器）为输出设备服务。在缓冲存储器中的汉字字形是没有重复放置的。

为了便于 RAM 存储器的更新，在终端的主存中除了显示文件区以外，另有一个缓冲存储器内所存储的汉字的登记表（256×2＝512字节），称为使用汉字登记表。登记表也就是当前使用的 256 个汉字代码的目录。它包含两项内容，一是当前所使用汉字的代码（例如国标码），另一项是相应的汉字字形信息在缓冲存储器中的地址编码(0～255)，也就是该汉字在汉字登记表中的单元地址。它的更新法则是：

（1）清除显示器时，显示文件清除为 0，使用汉字登记表也清除为 0。实际做法是把一个登记箭头指向登记表的始端就可以了。缓冲存储器的内容不必清除。

（2）显示一个汉字时，先查登记表，从始端逐一查找，一直查到登记箭头所指的位置。若查到这一汉字代码，则说明已经登记过了，表示该字的字形点阵已在缓冲存储器中，只要把“字形地址码”替换汉字代码，并存入显示文件区即可。这里所说的“字形

使用汉字登记表

	国标码	序
汉	3A3A	0
字	5756	1
系	4F35	2
统	4D33	3
的	3544	4
形	504E	5
发	3722	6
生	497A	7
器	4677	8
		9
		A

主存显示文件区

显示信息

汉	0
字	1
系	2
统	3
的	4
汉	0
字	1
字	1
形	5
发	6
生	7
器	8

图6-5 〔例 2〕的汉字字形发生器工作状况示意图

地址码”就是使用汉字登记表中的“序号”，从0到255。它也是临时的显示文件信息，用来控制显示器的再生。如在登记表中查不到这个汉字，则要从ROM库或软盘上把该汉字的字形信息调入缓冲存储器中，并在使用汉字登记表中登记，登记箭头下移一格。用新定义的字形地址码替换汉字代码，并存入显示文件区。

（3）显示器要删除某一汉字时，只要在显示文件中删除就可以了，并不删去汉字的字形信息，也不修改字形信息登记表。

图6-5是这一更新过程的示意图。

这一方案的特点是，增加一个使用汉字登记表，保证重复使用时可以不重复调入字形信息，而无需在缓冲存储器中拷贝字形信息。这样就提高了缓冲存储器的使用效率，缩短了查表时间，加快了响应速度。其缺点是删除时不能删去字形信息，所以，除了“重新开始”以外，没有其它更新的办法。这也会给使用带来一些不便。

〔**例3**〕 这是上述〔例2〕的一个改进，主要是改进它的更新办法，使得在显示器需要新的字形信息时，可以随时将其调入缓冲存储器，同时挤掉那些当前用不着的字形，保证使用者当前要用的字形信息都在缓冲存储器中。这就使RAM存储器常用常新。具体的做法很简单，只要改造一下原来的使用汉字登记表就可以达到目的。若在原来的使用汉字登记表中再添加一项内容，即把该汉字的使用情况也记录下来，此称为“使用情况登记表”，简称U表。凡是进入显示器的，在查到该字后应在U表上加“1”。退出显示器时，要在该字的U表中减1。若某汉字在U表中记录的值为0，则表示当前无用，可以被更新。若在U表的值为1，则表示在显示器中仅在一处使用。如果显示器当前使用这个字N次，则对应的U表的内容登记N。

更新法则：

（1）加电总清时，清除缓冲存储器全部内容以及使用登记表的全部内容，U表全部清为0。

（2）屏幕总清时，只要把U表全部清为0就可以了。

（3）显示一个汉字时，先查使用汉字登记表，要从头查起，一直查到末尾。如果查到这个字，则表示缓冲存储器中已有这个字的字形信息，用字形地址码替换汉字代码并存入显示文件区。同时，在该字的U表中加1。若从头到尾查不到，则表示缓冲存储器中没有这个汉字的字形。这时，要查阅U表，看哪里有“0”的项，查到0以后，就可以顶替，用新的字形信息替换掉没有用的字形。新的字形调入缓冲存储器以后，将该字的汉字代码填入使用汉字登记表，U表加1，并把该字的字形地址码替换汉字代码存入显示文件区。

为了有秩序地查询U表，可设一个U表查询箭头，每次查询U表，箭头便下移，下次查询U表时，应从上次结束处开始，到表尾以后，再回到顶端。这样可以保持大致的先进先出的秩序，否则每次都从头查U表，可能会使前面几个字变动频繁。

（4）退出屏幕的字，每个都要相应地改正U表的内容。退出一次，U表减1。

图6-6是这一更新法则的示意图。

利用使用情况登记表来管理缓冲存储器的进入和退出，掌握当前缓冲存储器的使用情况，就可以有目的地更新缓冲存储器的内容。由于使用汉字是随机的，所以查询使用汉字登记表只能逐个比较，查表时间较长。根据平均概率计算一篇1000字的文件，若要

使用汉字登记表

	国标码	字形地址码	U
你	4463	0	0
的	3544	1	1
们	4347	2	0
生	497A	3	1
年	446A	4	0
汉	3A3A	5	2
系	4F35	6	1
统	4D33	7	1
形	504E	8	1
发	3722	9	1
器	4677	A	1
日	4855	B	0
英	5322	C	0
了	414B	D	0
字	5756	E	B
海	3A23	F	0

←U表查询指针

显示文件区

显示信息

汉	5
字	E
系	6
统	7
的	1
汉	5
字	E
字	E
形	8
发	9
生	3
器	A

图6-6 〔例3〕的缓冲存储器内容更新示意图

调入200个新的汉字点阵，有800个字是重复的，则需要的比较次数共有200×256+800×128=153600次。

一种实用的改进办法是把登记表分成几块。例如可按汉字代码的最后两位是00，01，10，11来把登记表分成四块，这样，每次查表只要四分之一长度就可以了，一般说这种分法的溢出机会并不比单独用一张表的更大，而查表时的比较次数可以减少很多。

从程序设计的角度来看，实现缓冲存储器的更新，很类似于虚拟存储器的管理办法，但又不完全相同。所以，在借鉴虚拟存储管理技术时，必须结合具体情况加以改进。另外，从理论上来说，完全可以找到一种方法来建立汉字代码和字形信息的地址码之间的对应关系，从而可以实现以国标码为“关键词”的快速查表。但在实用上来说，这种做法可能还不如分块逐一查找更切合实用。

除了上述〔例3〕对〔例2〕的这一种改进方法以外，还可以有多种设想。例如，可以设计一种“优先链”的控制方法，来管理缓冲存储器的更新。

总之，汉字字形发生器的构成，既可以是简单的直接读取方式，只用一个只读存储器；也可以根据系统设计的要求，采取多级存储方案，或者采取多个设备共享一个字形

发生器的方案。同时要考虑缓冲存储器的结构，包括考虑它的更新方案等，使它成为一个利用率高，既经济又有效的汉字字形发生器。

6.3 汉字字形的压缩存储方法

汉字的字形信息可以有两种存储方法，一种是整字存储法，就是把汉字字形的点阵信息逐个字节地全部存放在字形信息存储器中，需要使用时可直接读出。这种存储方法原理简单，使用方便，响应时间快，也可以保证字形质量。另一种称为压缩信息存储方法，不是直接将字形信息存储起来，而是采用信息压缩的办法，存储器中只存文字的压缩信息，使用时需要将压缩信息“还原”成字形。这样做的目的是为了减少字形发生器的存储容量。

字形信息的压缩技术很多。我们只能介绍目前使用较普遍的几种方法。

6.3.1 部件组字法

前面曾经叙述过汉字的字形可以看作由“组字部件”拼起来的。据粗略的统计，在国家标准汉字基本集的一级字（3755个）中，由左右两部分拼起来的汉字有2021个，约占 53.8%；由左、中、右三部分拼起来的有 177 个字，约占 4.7%，这两种字合在一起占总数的58%以上。二级字（3008个）由左右两部分或左、中、右三部分拼起来的字有2082个，占69%。在国标一级字和二级字总共6763个汉字中，左右两拼或左、中、右三拼的汉字共有4103个，占总数的61%。特别值得一提的是：有几个组字部件的使用频度很高，或者说组字能力很强。以下列出在左、右两拼和左、中、右三拼的汉字中使用得最多的20个组字部件，及它们在6763个汉字中使用的次数：

氵	352	亻	213	虫	126	阝	83
扌	264	讠	142	阝	121	王	80
口	261	纟	139	土	107	火	73
木	237	月	137	女	102	犭	69
全	214	忄	127	石	93	鱼	64

除了左、右拼的字以外，还有7～8%的上、下两部分拼合的字，其部件分布也很集中，在6763个汉字中使用次数最多的前十个部件是：

艹	336	竹	111	宀	68	心	67	亠	51
曰	42	木	34	穴	32	雨	29	山	28

除上述两种最普遍的字形结构方法以外，还有其它一些值得考虑的字形结构。如在6763个汉字中走之旁的字就有 104 个；病字头有98个；门字部有45个等等。

既然在许多汉字中都有同样的组字部件，就可以设法只存储一个部件信息，相同的部件信息不再重复存储。需要输出字形时再由软件“生成”（或者说“组成”）汉字字形。这就是用部件组字法压缩信息的汉字字形发生器的基本工作原理。在这种字形发生器中存储的信息有两部分：一部分称为“部件字形信息”，它是存放各种不同尺寸的部件的点阵；另一部分称为“组字信息”，每一个汉字应该有一个组字信息。它应包含这一汉字的拼法和所需组字部件的点阵信息的地址。有了部件点阵信息和组字信息以后，才能生成整个汉字的点阵信息。

部件字形信息应包括各种不同“尺寸”的部件信息，一般采用密集存储的方法。为了节省存储量和便于读取，可以按字形信息的长度（存储容量）来分类。凡存储长度一样的就归为一类。这样，只要有一个类首址表就可以方便地找到某一类中某一个部件信息了。

在整个部件点阵信息中有一些字形仍要以整字的形式存放，这些字往往是一些常用字或者是不便拆开的字。象年、月、日、我、了、千…，这些字估计有500～700个。有些部件应该有几种不同尺寸的部件字形信息。例如单人旁“亻”，它在“倒、侧、例”等字中是窄长的单人旁，但是在“仍、仅、仇、代”等字中要宽一些，为了保存字形的美观，有可能要收存几种不同尺寸的单人旁。

据统计，汉字的组字部件，如果不考虑其尺寸大小，总共约一千个左右，其中还有一半左右的字本身就是独立的字。但在考虑了组字部件的尺寸以后，部件的数目有四千多个。有的设计者考虑用程序方法实现字形的“变倍”，也就是说只存一种尺寸的部件点阵信息，需要时再由程序把它压扁或变窄，以期减少信息的存储量。但是用程序方法来变换点阵的尺寸有一定的限制和困难，程序开销较大，字形的正确性也要受到影响。所以在多数的情况下只得收集全部各种大小尺寸的部件点阵信息。

用部件拼出来的字，有一个缺点，就是字形不美观。最主要的“破坏性”的原因是两部分拼起来的衔接部分太生硬。这是因为，实际上许多汉字的各个部件不是截然分开的，而是各部分之间互相“渗透”的。例如一个“妙”字，如果把左半部分放一个“女”字，右半部分放一个“少”字，并排放在一起是不好看的。必须使“女”字和“少”字互相有些交叉才好。这种渗透现象相当普遍，假如一律不考虑渗透，字形必然变差。反之，只要设法改善一下，字形就可以变得自然、美观。其具体实现方法并不难，只要使两个部件都放宽一些，在组字的时候可使衔接部分的信息故意有些重叠，就能起到较好的效果。

以上只是粗略地讨论了用部件组字法压缩字形信息存储容量的基本原理。具体实现起来会有各种不同的方案。这一方法的特点是，若能设计得当，就有可能保持字形的美观，生成过程不太复杂，但是压缩倍数不会很高。可以说是为了保持有一定字形质量的一个折衷的压缩方法。如果用24×24或32×32的点阵，就有可能使得输出字形相当美观，而存储容量则减少一半。对16×16点阵的汉字字形，由于组字信息所占存储容量的比例较大，故单纯用部件组字来压缩存储信息的容量，其效果并不理想。

6.3.2 向量存储法

将一个汉字的字形看做由许多直线笔画所组成，这些直线笔画可以有各种不同的方向和长度。这样，就可以用平面上的一系列向量来描述这些笔画。这就是汉字字形的向量表示法。用向量来表示图形是一种典型的办法，它在图形处理的领域内有极广泛的应用。汉字字形也是一种图形，因此，同样可以用向量来表示汉字的字形。

“向量”是一个数学的概念，一般是指在坐标空间中(在我们这里是一个固定的坐标平面)，由坐标原点指向空间中任意一点的一个有方向和长度的量。例如坐标平面上有一个点A，它的坐标是(x, y)，我们可以记作$A(x, y)$。则坐标原点指向点A的向量可以记作$\boldsymbol{A}$，它的坐标分量是x和y，所以，向量$\boldsymbol{A}$也可以写成$\boldsymbol{A}(x, y)$。

若空间中有两点$A(x_1, y_1)$和$B(x_2, y_2)$，就有向量$\boldsymbol{A}$和向量$\boldsymbol{B}$。由A点指向B点也有一个向量，我们用$\boldsymbol{AB}$来表示，向量$\boldsymbol{AB}$可以表示成：$\boldsymbol{AB}=\boldsymbol{B}-\boldsymbol{A}$。向量$\boldsymbol{AB}$的分量是$x_2-x_1$和$y_2-y_1$，同样，由$B$点指向$A$点的向量记作$\boldsymbol{BA}$，它的分量是$x_1-x_2$和$y_1-y_2$。

用向量加减法可以表示空间中任意两点之间的一个向量。但是，要注意向量$\boldsymbol{AB}$只给出了这个量的大小和方向，并没有规定起始点在什么地方。数学上的向量是设有起始点的，或者规定坐标原点为起始点。为了表示平面上的一个直线段，需要确定一个起始点和一个向量。或者给出两点，一个起始点，一个终止点。起点到终点的这个直线段的方向和大小（长度）可以用一个向量来表示。

用平面上的两个点来表示一个直线段，称为直线段的两点表示法。用平面上的一个点，和一个向量来表示一个直线段，则称为直线段的向量表示法或增量表示法。实质上这两种表示法是一样的。因为有了两个点的坐标，就可以把一点作为起始点，由两点的坐标差来确定增量。所以这两种方法一般都可以称为直线段的向量表示法。

现在我们来讨论用向量表示法描述一个汉字字形的例子。

首先，我们假定平面上有一个坐标系，左上角为坐标原点。数学上常常习惯把左下角取作坐标原点。因为汉字的书写习惯是从左到右，自上而下，所以我们不妨把坐标原点取在左上角。又假定x方向有十六个单位，y方向也有16个单位。平面上有一个点A，它的坐标是x，y，记作$A(x, y)$，和数学上不同的是，这里的x，y一定是整数。都是大于等于0，小于等于15的正整数，A是16×16点阵中的一个“点”。

在这种16×16的点阵坐标系中的一个向量也是由若干个点组成的。图6-7描出了几个向量的图形。其中，x增量为正，表示自左向右；x增量为负，表示向量自右向左。y增量为正，表示向量自上向下；y增量为负，表示向量是由下向上。图中的向量都没有考虑起始点的位置。

其中$\overrightarrow{12}$表示从“1”点到“2”点的向量。它的$\triangle x=5$，$\triangle y=2$，表示为$\overrightarrow{12}(5, 2)$，另外的三个向量可以表示为$\overrightarrow{34}(14, 3)$；$\overrightarrow{56}(0, -5)$；$\overrightarrow{67}(-4, 1)$。

从这些例子中可以看出，用向量表示的直线段，可以有各种不同的方向和长度，其

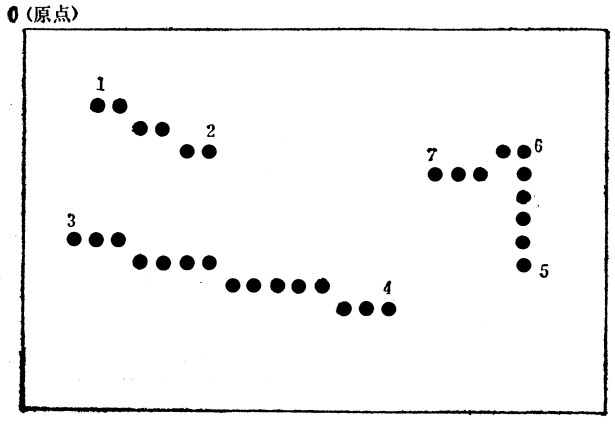


图6-7　点阵坐标平面中的向量

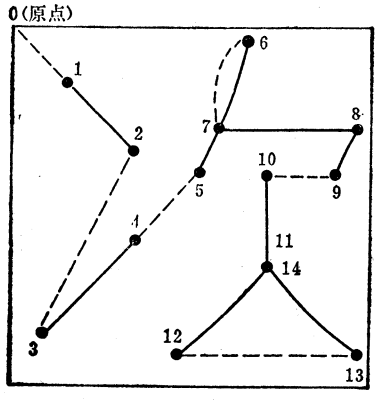


图6-8　“次”字的向量表示

表现能力是比较强的。但是，一个直线段中间不能“转弯”，不能改变方向。所以需要转弯的折线段要用几个连接的向量来描述。

在这个基础上，就可用向量来表示一个汉字字形了。一个汉字的字形可以看成许多直线段组成的。我们称这些直线段为“笔画”，这里的所谓笔画和一般说的汉字笔画不是同一的含义，而是指一段直线。图 6-8 是“次”字的笔画图。起笔是在坐标原点。第一个向量$\overrightarrow{01}$，这是一个“虚向量”，也可称为“空笔画”。它的作用是为下一笔确定起始位置。$\overrightarrow{01}$向量的终点就是$\overrightarrow{12}$向量的起点，$\overrightarrow{12}$是一个实向量。$\overrightarrow{23}$又是一个空笔画。$\overrightarrow{34}$是一个实笔画。…，依次类推。

把这些向量按先后次序列表可以得到如下结果：

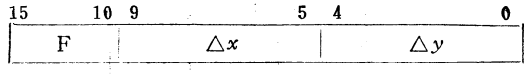
序	$\triangle x$	$\triangle y$	实/虚
01	2	2	0
12	3	3	1
23	-4	8	0
34	4	-4	1
45	3	-3	0
56	2	-6	1
67	-1	4	0
78	6	0	1
89	1	2	1
910	-3	0	0
1011	0	4	1
1112	-4	4	1
1213	8	0	0
1314	-4	-4	1

最后一笔称为末笔，其中共十四画，八画是实笔画，六画是虚笔画，末笔是实笔画。

我们规定的每一个汉字的第一画起点一定是坐标原点。第一笔的终点一定是第二笔的起点，第二笔的终点是第三笔的起点，依次类推。每一笔都有一个$\triangle x$和$\triangle y$。此外，需要指出是，究竟为实笔画还是为虚笔画，这可用一位信息表示：1 表示为实笔画；0 表示为虚笔画。这样一组有序的笔画信息构成汉字的字形信息。用向量组成的汉字字形信息来代替点阵信息可以使信息量得到压缩。在输出的时候，再把向量信息还原成点阵。以下进一步说明信息压缩的原理。

（一）字形信息的数字描述

用两个字节的数字表示一个向量，如



其中F称为笔形码，共有六个二进制位。用“000000”表示空笔画；用“000001”表示实笔画。当F＝“111111”时，表示实的末笔。如有必要，则可用“111110”表示空的末笔。

$\triangle x$，$\triangle y$各用 5 位二进制数，其中包括一位符号位。$\triangle x$，$\triangle y$可以表示－15 至 15 之间的任何整数，它们是向量的分量。

一个汉字的向量表示可以由这样一组有序的笔画信息组成。

（二）笔形码的扩展

上述向量的数字表示中有一个笔形码F，共有六位，除能表示空笔画、实笔画、空末笔和实末笔以外，并没有其它用处。实际上，可以利用这一信息来扩展向量表示法的功能。例如在汉字中“口”字用得很多。我们就可以定义一种专门表示“口”字的笔形码。

0 0 0 0 1 0	$\triangle x$	$\triangle y$

F＝000010表示一个口字的笔画，这个口字的宽度是$\triangle x$，高度是$\triangle y$。实际上它可以看作是由四个实向量组合而成的。这四个向量是：

0 0 0 0 0 1	0	$\triangle y$
0 0 0 0 0 1	$\triangle x$	0
0 0 0 0 0 1	0	$-\triangle y$
0 0 0 0 0 1	$-\triangle x$	0

先是左边一竖，然后是下面一横，再是右边一竖，最后是上面一横，回到出发点。这样，一个笔画信息就不再是一个向量了，而是一串向量。

再例如，我们可以定义“口”字下面开口的“冂”。

0 0 0 0 1 1	$\triangle x$	$\triangle y$

F＝000011 表示下开口的“冂”字。它也由四个向量组合而成。

0 0 0 0 0 1	0	$\triangle y$
0 0 0 0 0 0	$\triangle x$	0
0 0 0 0 0 1	0	$-\triangle y$
0 0 0 0 0 1	$-\triangle x$	0

和“口”字相似，只是第二笔是虚笔画。这样，我们在写“秃宝盖”时，只需一个笔画信息就够了。

F＝000100 表示一个“日”字。宽度是$\triangle x$，高是$\triangle y$，另外在$\triangle y/2$处再加一横，最后仍回到出发点。

F＝000101 表示一个“丅”型笔画。先画一横，再回到$\triangle x/2$处，画一长为$\triangle y$的一竖，最后笔画停留在竖的下端。

F＝000110表示一个“丄”型笔画，起点在一竖的上端，竖的下端有一横，宽度为$\triangle x$，终点规定在下横的中点。

这样，我们就可以定义许多种“组合笔画”，这种组合笔画扩展了向量信息。由于每一种笔形码后面仍有$\triangle x$和$\triangle y$作为笔形的参数，故它的“表现能力”是很强的。例如“口”字笔形信息可以表示各种大、小、长、扁的“口”字，而且只要把起点设置好，就可以放到任何位置上。由于笔形码共有六位，故我们可以定义几十种常用的组合笔画（称为笔形组合信息），这样就能进一步压缩汉字字形的信息。

（三）字形信息的调用

在笔画信息扩展以后，可以压缩一部分汉字的字形信息，但还是有局限性的。因为

这种笔形信息仍然受到 $\triangle x$ 和 $\triangle y$ 的限制，太复杂的笔形无法使用这一手段。这里介绍的字形调用，可进一步压缩字形信息。

我们可以定义一种“调用命令”。

例如规定 $F=100000$ 为调用命令：

100000	D	

这里的 $\triangle x$，Δy 已经不再是向量的增量，或者是组合向量的参数了，而是一个“地址”，或者说是一个字形的编号。而 $F=100000$ 表示一种“命令”。D 共有十位二进制，可以有 1024 种笔形。用这一方法我们就可以把一些部件固定化。例如 氵、亻、扌、禾、西、鱼、马、乌、焦、革、气、阝 等都用一些向量信息和笔形信息来固化，每一个部件都做成标准部件。起点可以一律定在左上角，部件的终点一律定在右下角。每一个部件的末笔也有末笔标志。这样，在许多字中可以直接调用部件信息。也就是说，可以吸收部件拼字法的许多优点并进一步压缩字形信息。

结合采用向量信息、笔形信息和部件信息以后，字形信息的压缩倍数是很高的。

（四）字形信息的还原

下面讨论一下字形信息的还原问题。从向量信息再转换成点阵信息，这就是所说的信息还原。

假定程序已经把字形信息中的部件调用信息和笔形组合信息都转换为标准的向量信息，那么，这时每一个字形都是一组只有实笔画和虚笔画的向量组了。现在的问题是如何把它们还原。

向量信息的还原遵照以下规则：

（1）凡是虚笔画一律不描点，只是改变起点的位置。

（2）实笔画都要“走步”。每描一个点称为走一步。一个向量有两个增量 $\triangle x$ 和 $\triangle y$。如果 $|\triangle x|\geqslant|\triangle y|$，则这个向量应该是走 $|\triangle x|+1$ 步。若 $|\triangle x|<|\triangle y|$，则应走 $|\triangle y|+1$ 步。例如，$\triangle x=5$，$\triangle y=-3$，那么这个向量应该走 6 步。第一步一定在起点上点出一点，就算走了一步。余下一共还要走 5 步。

（3）已知 $\triangle x$ 和 $\triangle y$ 如何决定向量的走法。可以有一个计算法则，这就是直线方程：

$$y=\frac{\triangle y}{\triangle x}\cdot x$$

直线方程中 $\triangle y/\triangle x$ 称为斜率。x 是自变量，取值 1，2，3，4，5。若 $\triangle x$ 是正数，x 每走一步都向右移动；若 $\triangle x$ 是负数，x 每走一步都向左移动。对应每一个 x，都可计算一个 y，四舍五入取 y 的整数值，就决定了 y 该走的位置。例如 $\triangle x=5$，$\triangle y=-3$。

计算函数如下：

$x=1$	2	3	4	5
$y=-\frac{3}{5}$	$-\frac{6}{5}$	$-\frac{9}{5}$	$-\frac{12}{5}$	$-\frac{15}{5}$
y 走步 -1	-1	-2	-2	-3

根据这种算法，可以保证走步误差最多不大于 1/2 格。即描出来的点子和直线方程最接近。可以很容易用程序方法来计算这种还原方法。也可以一次把 $\triangle x$ 和 $\triangle y$ 的各种

情况下的走法计算好，然后用查表的方法来实现还原。

向量存储法是一种典型的图形处理方法。结合汉字字形的特点，可以做到字形信息的压缩。整个16×16点阵的汉字字形发生器，共8000个汉字，总存储量可以压到50～60K字节，有的甚至可以更少一些。所以，这是一种高压缩倍数的存储方法。其缺点是和整字存储的汉字字形发生器相比，字形质量较差。

除上面重点介绍的两种比较典型的压缩存储方法外，当前有关汉字字形压缩存储的实施方案有许多种。例如笔画函数法、黑段白段法、哈夫曼树形压缩法等，由于这些方法的信息压缩倍数都不大，实际上在通用型汉字字形发生器中很少使用。

压缩存储是否有利，采用什么样的方案好？这个问题应该结合实际环境来讨论。压缩存储可以减少存储容量，但是字形质量受到影响，字模生成速度较低。有的方案虽然能保证字形质量，但压缩倍数太低。总之，要对存储容量、字形质量、响应时间三项因素作统一考虑。特别是随着存储器价格的不断下降，存储容量大的矛盾会逐步得到解决。直接存储的优势会愈来愈明显。这一趋势很可能使整字存储的汉字字形发生器得到更为普遍的应用。

第七章　汉字输出设备

7.1　汉字印刷技术概述

7.1.1　汉字印刷输出设备的功能

汉字印刷输出设备是汉字信息处理系统的重要组成部分。各类汉字信息处理作业所形成的文件、表格需用它来实现输出。为了适应多种用途的需要，对汉字印刷机性能方面的要求越来越高。

汉字印刷输出设备和一般的字符输出设备不同之处，是它必须照顾到汉字的固有特性。如前面章节所述，这些固有特性是：汉字字量比西文多百倍以上；字形复杂，组成汉字的点阵位点远较字符的点阵位点多；通常还要求不同尺寸的汉字混合使用；要求既能横写也能竖写；汉字中可以夹杂西文字符；带汉字的表格处理也较为复杂，它包括输出格式的表头、表旁、格线和数据的输出处理等。所有这些都说明，汉字印刷技术的难度大，设备成本也高。

要使印刷输出的汉字清晰度高，就要求组成汉字字形的点阵位点数目增加，这样，字形信息量也随之增加。由于汉字字量大（即使对国标一级、二级字而言也是如此），故要求的总信息存储量相当可观。因此，必须根据实际的需要安排适当的字形点阵，以便使印刷机结构更为合理。表7-1列出了不同应用领域所使用的文字尺寸；表7-2列出了文字字形清晰度、文字尺寸与字形点阵之间的关系。

表7-1　不同领域使用的文字尺寸（磅为0.35毫米）

使用领域	本文的文字大小	标题文字的大小
新　闻	约8磅×6.8磅	10～24～48磅
杂　志	8～9	10～20磅
式样印刷	10.5	12～14磅
一般帐票	字母数字9磅×7.2磅	
	汉字9～12磅	12～16磅 6～9磅（表头小文字）

从上述表的内容可知，若选取相同的点阵规格，则字号越小，其清晰度就越好，但字号不能太小。若相同字号点阵的位点数增大，清晰度就增强。字形结构简单，其清晰度也好，因此对结构复杂的字进行简化可以提高清晰度。此外，印刷机性能、墨色、纸张质量等都与清晰度有关。据此，必须根据不同的应用对象，选用不同的字号和点阵。一般说来，对于印刷普通文件，其要求可低些。对印刷出版物，其要求就很高。

综上所述，汉字印刷机的最基本的功能，就是把所需的汉字按规定的质量印刷到纸的既定位置上。为实现这个功能，发展了各种类型的汉字印刷机。

表7-2 字形清晰度、文字尺寸、字形点阵的关系

文字尺寸＼清晰度（字形点阵）	4线/毫米	6线/毫米	8线/毫米	10线/毫米	15线/毫米	20线/毫米	30线/毫米
7.2磅	10	15	20	25	38	50	76
9磅	13	19	25	32	48	64	96
12磅	17	25	33	42	63	84	126
应用	一般用			小型印刷		照排	
印刷机性能	串行打印机						
		喷墨方式					
			电子照相式				
					阴极射线管、激光照排装置		

7.1.2 汉字印刷输出设备的类型

用于计算机的印刷设备，按其印刷工作原理、印刷方式、印刷字符、图形及汉字、印刷质量、印刷速度、行宽、印刷色调、用途等因素可划分成多种类别。这里，按汉字印刷输出设备的工作原理，可把它们分成击打式和非击打式两大类。

击打式汉字印刷机以针式点阵打印机为代表机种。其输出形式较为灵活，成本较低，所以使用很普遍。非击打式汉字印刷机可把激光印刷机作为代表机种。由于充分利用微电子技术、激光技术及电子照相技术的成果，这类印刷机的性能较击打式设备有很大的提高，当然，它的价格目前是较高的。

这两类印刷机都是属于非字型式印刷机。即汉字是以点阵（点或线）组合来表现的，汉字字形信息存储在汉字字形发生器中，这样，使打印记录的灵活性（文字种类、字形尺寸变化灵活并适用于图形输出）好，输出速度高；字模点阵位点多时，其印刷质量好；但一般说来这类印刷机的印字质量不如采用活字的字型方式印刷机。

表7-3 印刷机的分类

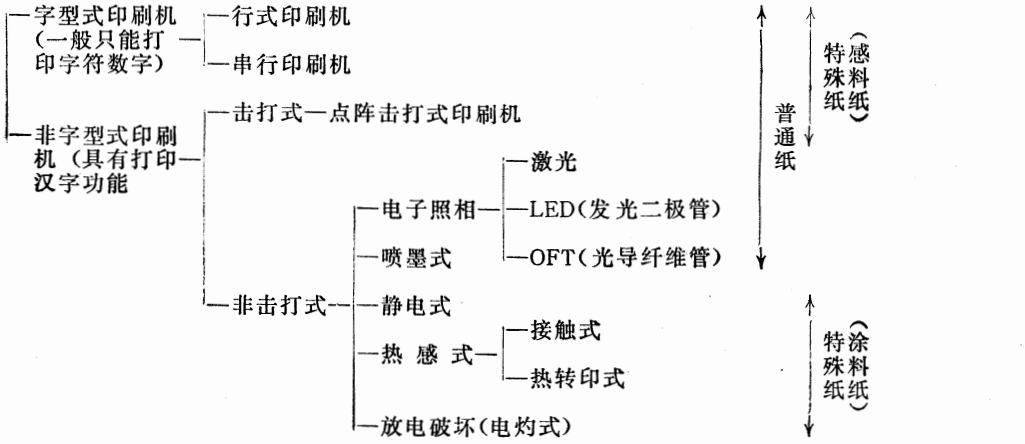


表 7-3 列出了各类印刷机的分类情况。其中主要的一些印刷机类型、工作原理、结构特点和技术性能等将在以下各节中作进一步介绍。

7.1.3 汉字印刷机的发展概况

汉字印刷设备是在传统的计算机西文打印设备的基础上发展起来的，可视作西文字符印刷设备功能的扩展和延续。

在六十年代，主要是使用字型方式印刷机，汉字印刷机尚处于实验阶段，因价格过高而很少使用。七十年代由于非字型方式印刷机如针式打印机、静电印刷机、激光印刷机相继问世，汉字印刷设备也开始了向实用化发展，到七十年代末，汉字印刷设备的性能已较完善，并开始实际应用，出现了以针打、喷墨、激光印刷机为主流的多种形式的汉字印刷设备。

表 7-4 是有关印刷机的发展情况，随着汉字信息处理领域的不断扩大，进一步促使汉字印刷设备向着提高印刷质量、提高印字速度、减小噪声及改善操作性能的方向发展，并继续在高可靠性、低价格及多功能方面改进。

表7-4 主要印刷机的发展过程

		1961	1965	1970	1975	1980
印刷机	击打式串式打印机	整体字模型 15字/秒	整体字模型 20字/秒	针式点阵型 40～60字/秒	字模型80字/秒 汉字针式点阵型 针式点阵型	汉字字模型10～20字/秒 汉字针式点阵型 80字/秒
	非击打式串式印刷机			热感式 30～120字/秒	汉字热感式 80字/秒 喷墨型50字/秒	随机喷墨式 30字/秒
	击打式行式打印机	链　式 600行/分	鼓　式 1250行/分	列车式 2000行/分	鼓式，带式 2000行/分	带　式 3800行/分
	非击打式行式印刷机		静电型 2万字/秒	喷墨式 8000行/分	激光印刷机12000行/分 光导纤维管印刷机4000行/分	汉字激光印刷机 15000行/分

7.2 针式汉字打印机

7.2.1 针式汉字打印机的性能特点

击打式汉字印刷机中的针式汉字打印机占据主流位置。它作为终端打印设备，在事务处理等应用领域得到了广泛应用。针式汉字打印机是在西文字符针式打印机的基础上发展而来的，由于用点阵来组成文字、图形，并不会因字种增多而增加选字时间，因此易于实现多字种的汉字打印。针式打印机由于可以使用普通纸，故在进行大量的汉字信息处理的情况下，有利于降低维持费用。

一般来说，对针式汉字打印机的性能要求是：

（1）为保证印字质量，打印头应有较高的分辨率，即打印针的排列应细密；
（2）印字速度应适应计算机信息通信的要求；
（3）印字头不需调整，结构上具有可维护性；
（4）运行可靠，便于操作和故障修复；
（5）运行噪声要小，达到能在办公室里使用的水平；
（6）具有扩大文字、竖线印刷等多种功能；
（7）体积小、重量轻、价格低。

7.2.2 针式汉字打印机的工作原理

针式汉字打印机与针式字符打印机的工作原理是相同的。其机械部件的主要区别在于，汉字针式打印机结构较字符针式打印机复杂。打印机构是印刷汉字的执行部件，它由打印头和驱动器两部分组成。打印头是打印机的关键部件，它直接关系到打印质量、输出速度和可靠性。驱动器是控制打印针高速运动的部件。图 7-1 示出了印字机构的原理图。

印字头安装在托架上，在压板和印字头之间安放色带和印字用纸，纸由导孔引导前进。印字时，装着打印头的托架向右移动，与时序栅产生的时序脉冲同步，由控制器送来的印字信息，驱动被选打印针击打压板（或压辊），以实现印字动作。印完一行后，印字头返回左端，纸前进一行。

印字头由打印针、针管、导板和驱动电磁铁四部分组成。其结构如图 7-2(b)所示。有两种类型的印字头。图中左边所示的结构采用衔铁在外的C型电磁铁，依靠电脉冲的激励，提供足够的驱动力，并将其产生的机械动作传递到打印针上。在打印针穿过的导孔中设置宝石轴承，用以导引针的动作。由于这类方式的象点直径不能太小，因此印字质量不高。右边的结构采用针和衔铁分离的方式。打印针在静止状态时，由复位弹簧使它处于平衡位置。当衔铁动作时，针即开始打印动作（见图中上方的打印针），衔铁停止后针还在继续动作，向着纸和色带击打。这种方式容易实现高速印字，并且打印点阵较细密，现已被广泛采用。

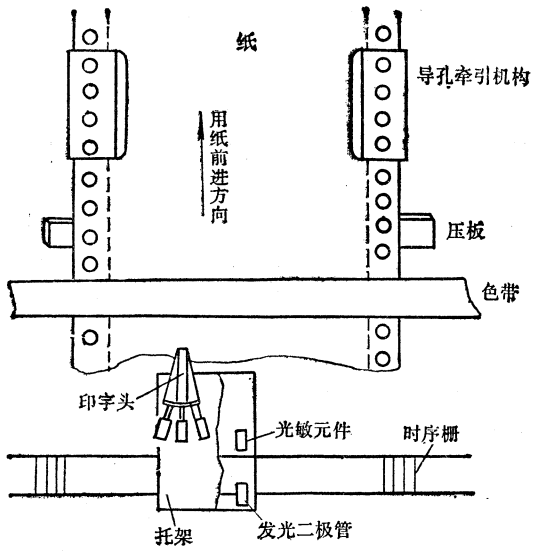


图7-1 针式打印机印字机构原理图

把这样的多个打印针机构组合起来，装配成圆形结构的印字头（见图 7-2 a）。针的前端集中后形成 1 列或交错的 2 列，由于汉字打印机针头数目多，通常排成 2 列，印字头沿水平方向作等速运动，一旦同步，针便高速击打形成文字。这种印字方式如图 7-3 所示。

打印针的质量参数直接关系到印刷质量和印字头的寿命，在结构设计中应考虑各种因素：

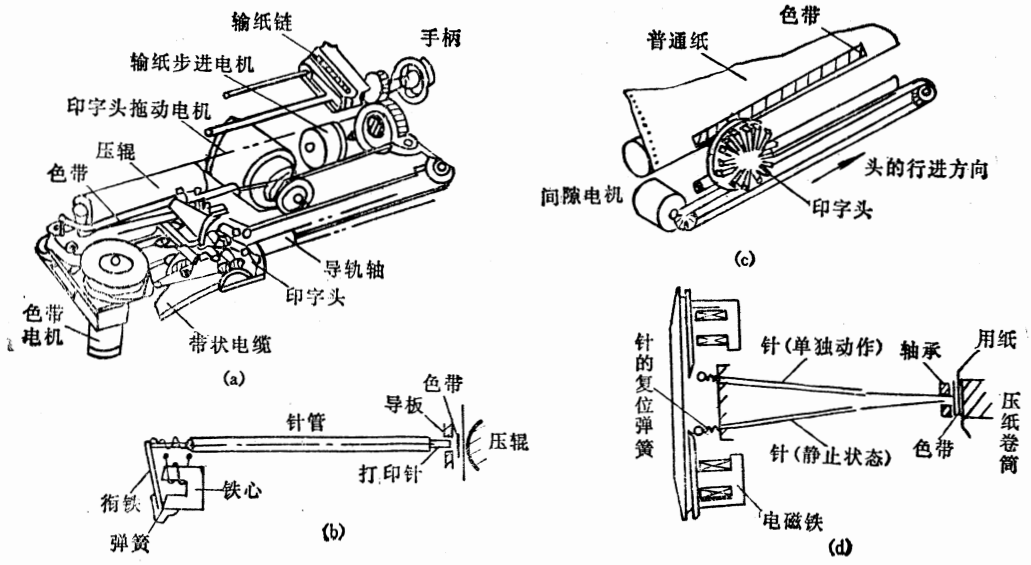


图7-2　打印机结构和打印头结构

（a） 针式打印机结构；（b） 印字头结构；（c） 印字机构；（d） 印字头的断面。

（1）打印针的配置。这是决定印刷质量的关键。针的配置与文字尺寸、点阵数直接有关，由此确定出打印针的点距、针的直径及针点的重叠量，在图 7.4 中表示出这些量之间的相互关系。其中，H为文字尺寸(高度)，q为空白量，m为斜向重叠量，p为针距，n为点阵数，d为针的直径，D为位点直径，e为墨水扩大量。用公式来表示时，它们之间的相互关系为：

$$H=(n+1)p-q \tag{7.1}$$

$$q=2p-(d+2c) \tag{7.2}$$

$$m=\sqrt{2}\,p-(d+2c) \tag{7.3}$$

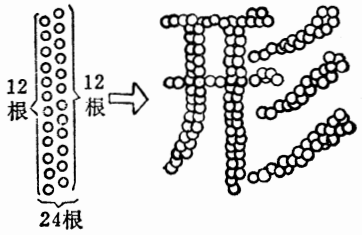


图7-3　印字方式示意

为了取得好的印刷质量，在增加N的同时，必须尽可能减小点距p，减小针端面积和控制墨水的扩大量。例如，当打印汉字的点阵数为24×24时，取针头直径为 0.2 毫米（一般字符打印机的针头直径为0.35毫米）；允许打印点的直径为针头直径的 1.5 倍，$D=d+2e=0.3$毫米；点距取 1.5 毫米；墨水扩大量取 0.1 毫米。此时打印点阵的密度为6.7点/毫米，可获得较高的印刷质量。

（2）打印针的寿命，应保证在 1.5 亿次以上（相当于打印 3 千万个汉字）。

（3）针的动作特性。必须考虑针与压纸滚筒之间的最小间隙、印字压力的设定范围等因素。

（4）为使打印针驱动机构高速化，要设计好电磁铁的动作条件和驱动特性。

图 7-4 示出了打印针位点的配置。

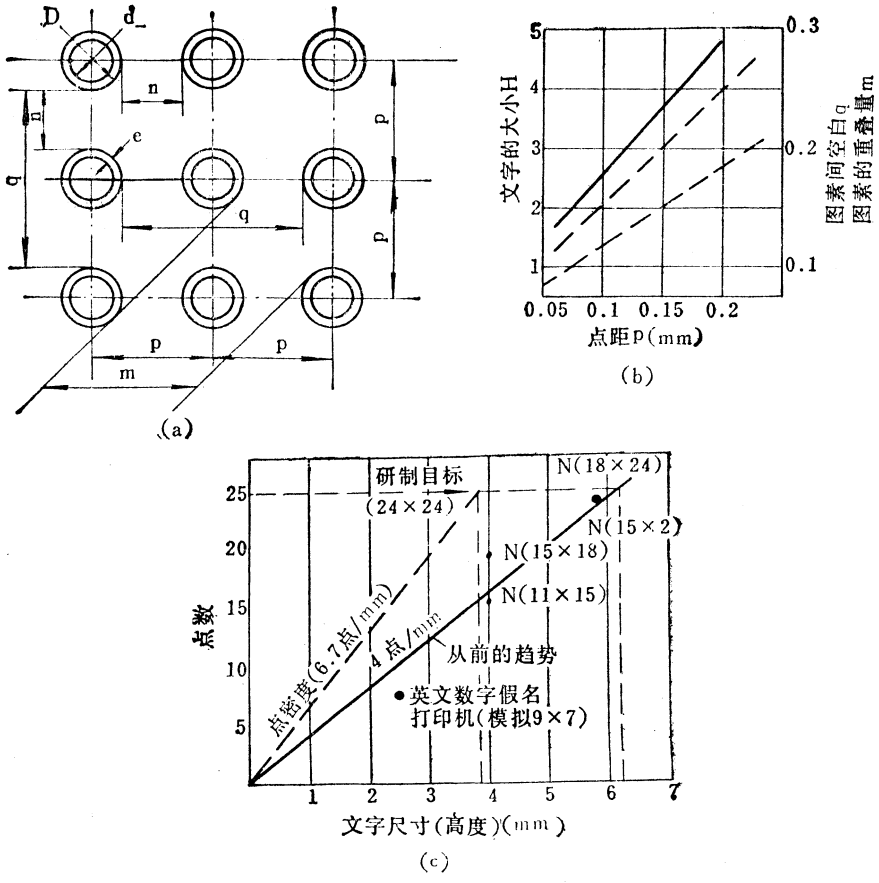


图7-4　打印针位点的配置

（a） 点阵排列；（b） H和p，q的关系；（c） H和N的关系。

7.2.3　针式汉字打印机的组成与动作

针式汉字打印机通常由打印机械、驱动电路和打印机控制器三个部分组成。各部分协调动作，完成从计算机送来的汉字信息的印字操作。图 7.5 是打印机的组成简框。

一、打印机械

由印字头、托架传动机构、送纸机构和色带传送机械四个主要部分组成。印字头包括打印针和驱动电磁铁部件。托架传动机构带动印字头实现按行的印字，其驱动源常采用四相步进电机，这样可进行两个方向的印字，而且可以随着所打印的文字形式不同（字母数字或和汉字形式）而改变托架的传送速度，不需要检验点、构造简单、控制也较简单。图 7-5 示出了针式汉字打印机的组成简框。

送纸机构和托架传动相似，驱动源也采用四相步进电机，可以实现以一排点，一行文字或一页纸为行距的送纸动作。为了适应汉字处理中办公室业务的需要，可利用单页纸插入机构，实现单页纸的打印。

色带传送机构通常不带驱动源，只在托架下用一根张紧的钢丝绳缠绕在色带传输滚

筒上。随着托架移动，色带便自动传送。

二、驱动电路

这是打印机的电气控制部分。其功能包括：产生印字头的各个电磁铁驱动信号；产生托架驱动步进电机和走纸步进电机的各相励磁信号；根据复位微动开关和换行微动开关的状态控制所需的各种驱动信号的发送等。此外，针式打印机中一些辅助开关(如用纸检测等)和指示灯都由本部分产生驱动信号。各部分组成如图 7-6 所示。

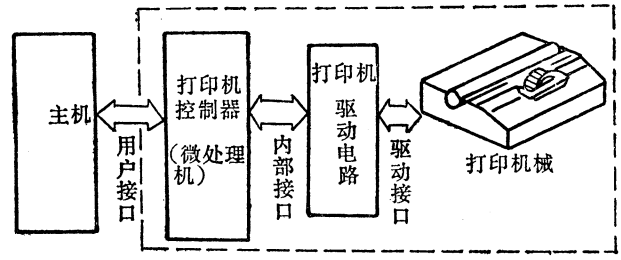


图7-5　针式汉字打印机的组成简框

图 7-6 示出了针式汉字打印机的驱动电路框图。为适应针式打印机的动作特性，驱动电路所产生的各类驱动信号必须遵循严格的定时关系。此外，为满足针式打印机驱动的需要，驱动电流脉冲波形也满足一定要求，这部分的定时信号关系如图 7-7 所示。

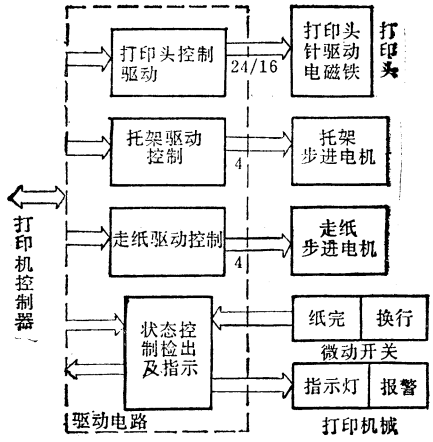


图7-6　针式汉字打印机的驱动电路

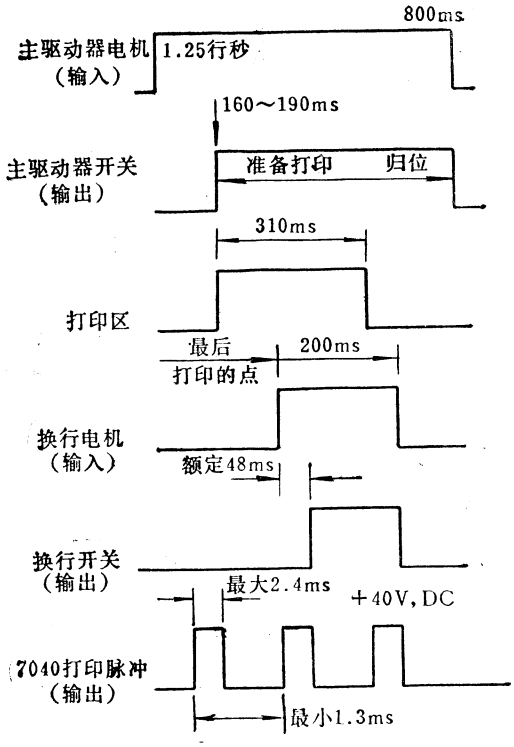


图7-7　针式汉字打印机驱动电路的定时

三、打印机控制器

打印机控制器是针式汉字打印机与主机相连接的接口部件。打印机控制器接受主机送来的各类打印控制信息和代码信息，在控制器内转换成对针式打印机驱动电路的控制信息，再由它去驱动打印机械的动作，完成全部打印操作。

随着针式汉字打印机功能的不断完善，打印机控制器已普遍采用微处理机作为控制器。这样，汉字文件打印所需的格式编排、印字速度切换，文字的扩大及缩小、竖向及横向打印、文字和图形打印等就很容易由微型机实现控制。

图 7-8 所示为针式汉字打印机控制器的框图。CPU 为 8 位微处理机（8080或 Z 80），打印监控程序固化于 ROM 中，实现对打印机全部控制任务。作为打印信息的缓冲器，总线上接有 8K 字节的 RAM。微处理机的外围输入输出电路用来实现主机（即使用这台针式打印机的计算机）与控制器之间，以及控制器与驱动电路之间的连接，信息的收发由打印程序来规定。由于复杂的打印操作可由软件来实现控制，故可简化打印机控制器结构，提高可靠性，增强功能。

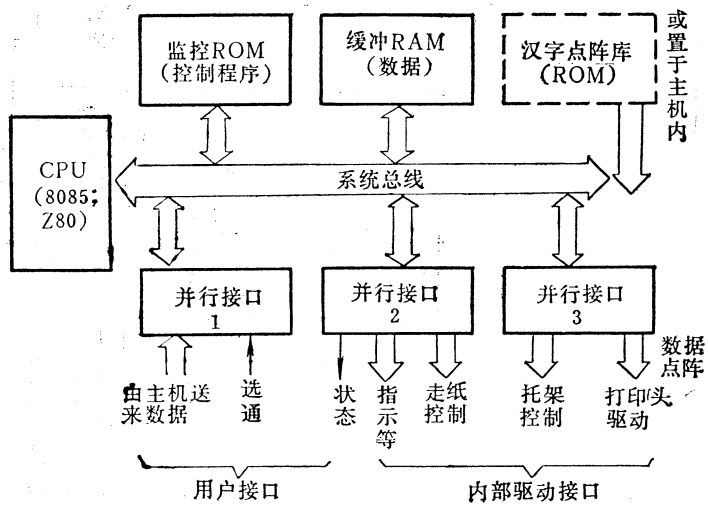


图7-8　针式打印机控制器的结构

驱动针式打印机所包含的软件可分成两个部分。一部分称为缓冲器管理程序，它主要用于处理针式打印机与主机的接口信息。主机送来的待打印文字被存储在控制器的缓冲存储器中。程序中设置两个指针，以供实现对缓冲存储器的存取。其中，输入指针总是指向最后存入缓存器的那个文字；输出指针则指向要打印的下一个文字。这样就可实现缓存器与主机以及打印驱动部分之间的联系。

软件的另一部分称为打印机服务程序。它用来完成待打印文字的打印操作控制。根据输出指针所指示的缓存位置取出要打印的文字信息。服务程序利用定时器保持对印字头当前位置的跟踪，并向印字头的电磁铁发送驱动信号，以实现打印。在服务程序的控制下，从缓存器取出的文字点阵信息按列发送到打印机驱动电路的缓冲寄存器，以驱动打印针动作。针式打印机可实现的各种功能，都配备有相应的服务子程序。图 7-9 示打印程序的结构。

最后，说明一下汉字字模库的设置。针式汉字打印机本身可以不带汉字字模库。这种情况下的汉字的点阵信息是由主机发送到打印机控制器的缓存器的。由于点阵信息量大，故针式打印机工作时要大量占用主机时间，从而使系统效率较低。随着 ROM 价格的不断下降，本身带有汉字字模库的针式汉字打印机已逐渐增多。这样，打印机只需从主机接收待打印的汉字代码，由打印机控制器在机内转换成对应汉字字形的点阵，并送给缓冲存储器。由于这样可以大大减轻主机负担，提高系统效率，因此这是今后针式汉字打印机发展的趋势。

为了进一步弄清针式汉字打印机的工作参数，以下用国产 CYD-901 型24 针针式汉字打印机为例作详细的剖析。

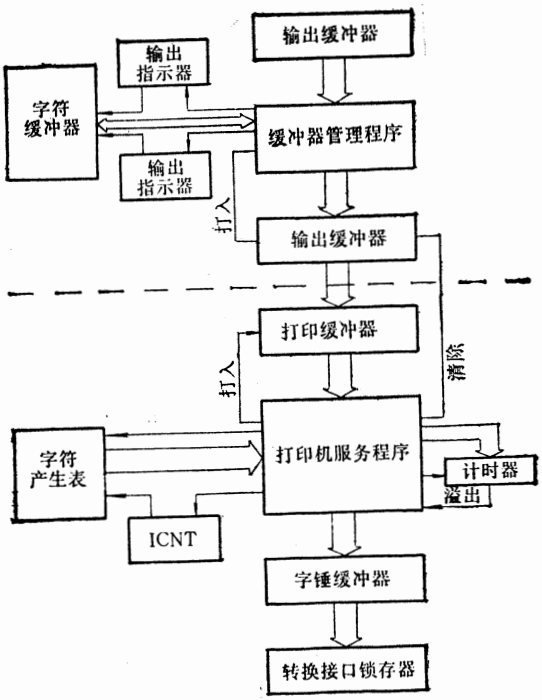


图7-9　打印程序的结构图

7.2.4　针式汉字打印机的接口及其控制信息

一、CYD-901型针式汉字打印机的技术性能

CYD-901型针式汉字打印机在技术性能上具有以下一些特点：

（1）印字头用24根打印针，采取两列交错排列。纵向覆盖率为20%；横向移动单位为1/120英寸，其覆盖率为20.6%。可打印出质量较高的24×24点阵、4毫米见方的汉字。

（2）采用步进电机驱动机构，具有双向最短距离印字机能。该机构可按1/24英寸间距精细地传送印字纸，走纸距离可以自由设定，行间能首尾相连，便于打印表格和图形。

（3）具有打印汉字35字/秒和打印字符105字/秒两种印字速度，并具有水平、垂直制表机能。

（4）采用主-从微处理机作打印机控制器，其功能强、可靠性高，并可脱机自检，从而维护方便。

打印机的主要技术规格列于表7-5。打印机由印字机械、驱动电路及控制器几个部分组成，各部分关系见图7-10。下面介绍的接口信息及控制信息都以此为基础叙述。

二、针式汉字打印机的接口信息

（一）针式汉字打印机控制器与驱动电路之间的接口（即驱动器接口信息）

这些接口信息包含有：

（1）CRA_0、CRB_0、CRC_0和CRD_0——托架传送步进电机的A～D极的励磁信号；

表7-5 CYD-901针式汉字打印机的主要技术规格

印字方式：点阵击打式；汉字	字母、数字
印字速度：35字/秒	105字/秒
位　　数：90字/行	136字/行
文字构形：22×24点	9×11点
文字大小：5.13×3.88毫米	2.76×1.79毫米
文字节距：3.81毫米	2.54毫米

改行间隔：1.05毫米的整数倍

用　　纸：两端链轮间12.7-38.1厘米宽

接　　口：TTL电平，8位并行。

（2）LFA_0、LFB_0、LFC_0、LFD_0——走纸步进电机的A～D相的励磁信号；

（3）PD010～PD240——相应于24根打印针的驱动信号；

（4）THM_0——给印字头加电压的信号；

（5）$FIRE_0$——激励打印针的信号，根据这个信号，若PD010～PD024有了数据，就激励相应的打印针；

（6）$LFMH_0$——加在走纸步进电机上的电压（+24V）；

（7）$LMGSW_0$——用于印字头初始置位和超出运行的检测工作开关接点信号；

（8）$LMGSW_1$——接在字盘左侧的开关触点信号；

（9）PENDSW——用纸检测开关的接点信号；

（10）$ONEC_0$——是原来位置检测器的信号。

该接口信息的线路实现如图7-11所示。

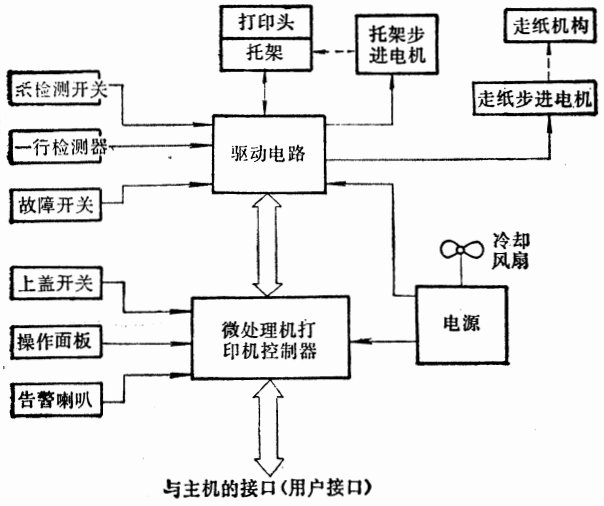


图7-10 CYD-901型汉字针式打印机的组成

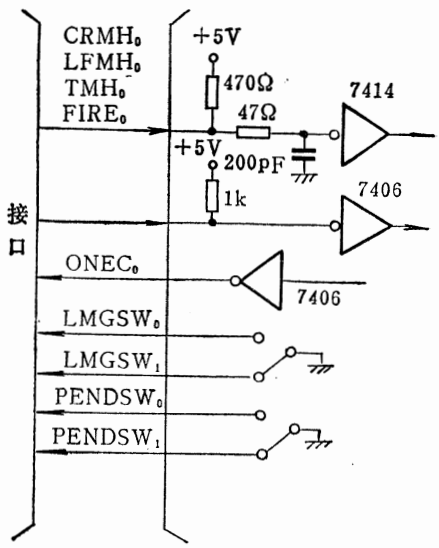


图7-11 驱动器接口线路

（二）针式汉字打印机接口信息

打印机接口信息通常是指针式汉字打印机与主机间的接口信号。所采用的8位并行接口，如表7-6所列。用户将针式汉字打印机连接到主机时，必须对打印机接口信息的

安排有清晰的了解。虽然各类针式汉字打印机的接口表面上很不相同，但原理上是相近的。打印机的并行接口大都遵循 Centronics 的接口规程。简单地说，这一规程是用选通脉冲同步来进行数据的读入，用 BUSY（忙）及 ACKNOWLEDGE（认可）来实现信息交换（handshaking）方式的传送。图7-12是这种方式中主要信息的时间关系。

表7-6　针式打印机与用户交换信号表

信　　号	来往方向	线　　数	信　号　名　称
DATA11-DATA81	用户⟶打印机	8	数据DATA1-8DATA11为低位
MODE10	⟶	1	打印模式 1
MODE20	⟶	1	打印模式 2
DSTBO	⟶	1	数据选通
PRIMEO	⟶	1	初始化
ACKO	⟵	1	回答
BUSY1	⟵	1	忙
CMDR1	⟵	1	命令请求
SLCT1	⟵	1	选择
ERRO	⟵	1	错误
PE1	⟵	1	纸完
OSC	⟵	1	振荡器输出1.536MHZ
+5V	⟵	1	+5V（<200mA）

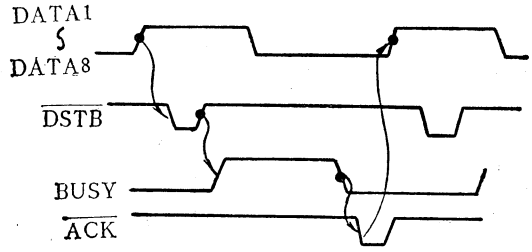


图7-12　针式汉字打印机并行接口的定时图

以下结合上述规程来说明表 7.6 中各信号的作用：

（1）$DATA_1$～$DATA_8$——8 位并行输入数据。

（2）SLCT——通知主机打印机是否可使用的信号，即设备是否被选中的信号（SELECT）。

（3）DSTB——data Strobe 为读入数据的选通脉冲，主机发送此信号给打印机，使打印机接收由主机送来的 8 位并行数据 $DATA_1$～$DATA_8$。

（4）ACK——设备（打印机）在接收了数据后，向主机发回的应答信号，据此，可再传送新的数据。

（5）BUSY——打印机通知主机是否现处于工作中的状态信号，若正处于工作中（忙），就不能向打印机送新的数据。

以上是基本接口信息。

此外，不同的针式汉字打印机还配有各类辅助接口信息。本例的针式汉字打印机的辅助接口信息有：

（1）PRIM——打印机接收到此信号时，便处于准备状态，使打印机初始化。

（2）MODE——用此信息改变数据总线 $DATA_1$～$DATA_8$ 上信息的含义。高电平时总线上信息为字符代码，低电平时总线上信息是打印命令或汉字点阵的信号。

（3）CMDR——在汉字方式下按序定义数据。当其为高电平时，总线上送来的数据是命令信息；然后，CMDR变为低电平，打印机接收汉字点阵或图形信息；最后，在该数据序列全部接收完后，CMDR再次变成高电平，以便接收下一命令数据 。在字符方式下，该信号始终为高电平。

（4）ERR——打印机产生故障时发出的信号。

（5）PE——指示打印机的打印纸已不变的信号。

其他诸如：OSC——是打印机向外送出的 1.536 兆赫时钟脉冲；打印机向外界提供 +5V 电源（一般不设置）；等等。

并行接口是打印机实现高速打印动作的必备接口。有的针式打印机还同时设置串行接口，利用 RS-232C 通信规程实现 打印机与主机的联系。由于 连接线简单，故在打印速度要求不高时，使用串行接口是很方便的。

三、针式汉字打印机的数据格式和控制信息。

有了接口信息，就可将主机和针式汉字打印机连接起来。下一步就是用来组织由主机向打印机发送的信息。它包括汉字数据信息和控制信息两个部分。

（一）汉字数据信息的组织

首先要弄清针式汉字打印机一行可打印的最大点数，通常称为针式打印机的“打印位置”。对于 CYD-902 针式汉字打印机，一行横向最大点数为 2176 点，纵向为 24点。于是，在用汉字方式打印时，一行最多可打印 90 个（24×24 点阵）汉字；在 用字符方式打印时，为每行 136 个（16×11点阵）字符。在组织打印数据时，不能超越这个范围。

汉字的点阵为24×24，但接口中只安排了 8 位数据总线。所以，一个汉字点阵数据按每纵列 3 个字节（3 × 8 点＝24点），共24个纵列来组成。图7-13为“漢”这个字的点阵信息，由22×24点阵组成。图中，E/0 的表示相当于用十六进制书写的 8 位数 E0，或二进制数11100000。

主机要实现打印一行汉字时，则按图7-14来组织打印信息。第一字节H ‘01’ 为打印汉字命令（H为十六进制数的标志）；第二、三字节以 15 位信息填入一行 要打印 的距离 N；例如在一行打印 13 个汉字时：N＝13×24点＝312 点，换算成十六进 制为 0138，其中，将 01 填入第二字节，将 38 填入第三字节；从第四字节开始，用来安放汉字的点阵数据，例如：第四、五、六字节为一行汉字的第一个字的第一纵列的点阵数据。依次类推，直至一行汉字点阵全部容纳下为止。

（二）针式汉字打印机的控制信息

针式汉字打印机的控制码比较多，用它来实现汉字文件的各种格式的打印功能。可以把控制码分成以下几类来说明。

1．设备控制　实现由主机对打印机进行联机/脱机的指定控制，主要有：

（1）DC1——(H ‘11’)　用此码使打印机转为选择状态。

（2）DC3——(H ‘13’)　用此码使打印机转为非选择状态。

2．位置控制　用以移动打印位置。在沿文字打印方向，可指定的位置 移 动单位为

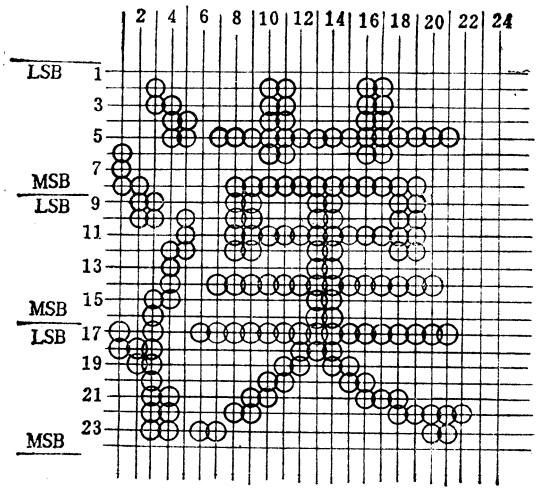


图7-13 “漢”字的点阵信息

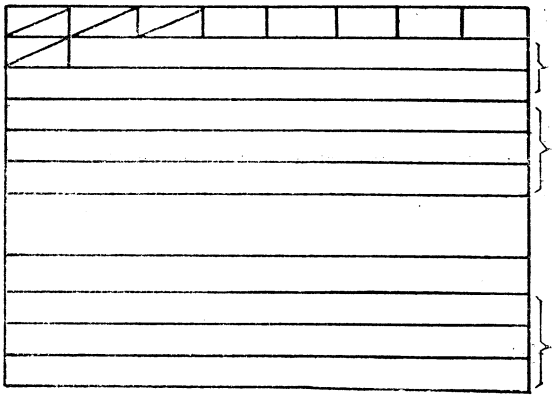


图7-14 打印一行汉字的信息组成

1/160 英寸（1 点）；沿纵向的行移动单位为 1/24。其控制码有：

（1）HT——水平制表。即缓存中打印位置进到下一个制表位置。

（2）LF——改行，或记为 NL（换行）。即打印位置移到下一行的 行首或第一个制表设定位置。

（3）VT——垂直列表（也可不用）。

（4）FF——换页。把打印位置按同一打印位数推进到下一页的最初一行。通常一页设定为 66 行。

（5）CR——回车。打印位置返回到同一行的行首或第一个制表位置。

（6）SP——间隔。使打印位置在同一行中推进 16 点。

（7）KSP——汉字间隔。在同一行中推进 24 点。

3．制表控制　在水平方向，以 1/160 英寸为一个点单位，在整个打印位置范围内进行列表设置。纵向，以 1/24 英寸为单位，在 6 页（396行）范围内进行列表设置。这类控制码有：

（1）SCT——水平方向列表设置。

（2）RCT——取消水平方向列表设置。

（3）SLT——垂直方向（纵向）列表设置。

（4）RLT——取消垂直方向列表设置。

（5）STF——列表位置的指定，可带数个距离参数N_2，…，N_n。

4．格式控制　它用于指定文字的传送量和走纸量。控制码后接连续的参数，因此，适当地改变这些参数，就可以指定文字的数量、间隔和改行的位置等。常用的控制码有：

（1）CP——字符间距。用以确定字符间点距。

（2）LP——行间距。控制换行时的行距。

（3）LC——指定不同行间距的组合走纸。

5．扩充的控制　它是打印机进行各种特殊处理所用的控制码。一般使用 ESC 码作

为先导，后面再接识别字符组成复合的控制命令，通称为ESC控制码。由于命令为2字节，故使打印机的控制能力大幅度地增强。以下举出有代表性的扩充控制码来说明。

（1）加下线——指定用ESC+ I；解除用ESC+ X。

（2）扩大文字——用ESC+SO。可将文字在横向扩大一倍后印刷。有了这个功能，就不需要自行编制扩大文字的打印程序了。

（3）纵写汉字——指定用FS+ J；解除用FS+ K。

FS的功能与ESC相同，也是扩充控制码标志。

（4）单页纸装入，卸出——用IN及OUT命令来执行。有的打印机用DC2+ O及DC2+ P来执行。

（5）响铃——BEL。

（6）字符增强打印、字体类型更换等。

总之，扩充的控制码所提供的功能，为编排实用的汉字打印文件提供了很大的方便。因此，在使用针式汉字打印机时，必须对其有详细的了解，充分利用这些扩充的控制码，使打印程序有效而简捷。

必须指出，各类打印机的基本控制码如CR、LF、VT、FF、BEL、DC1、DC3等的定义是相同的，扩充控制码却很不相同。此外，有的控制码，在字符方式下和汉字方式下都可以使用。有的却不然，不能乱用。

如上所述，在掌握针式汉字打印机的接口信息、数据格式、控制码以后，就可以把汉字打印功能组织到用户程序中去。

7.2.5 针式汉字打印机的选用

近年来，随着汉字信息处理技术的发展和推广应用，用于各个领域的针式汉字打印机也相继问世。尤其是在微机上配置了汉字信息处理功能后，对价格低而功能完善的针式汉字打印机的需要，已日益迫切，这就促使针式汉字打印机迅速发展。针式汉字打印机品种规格繁多，各具特色，故必须作出比较后再选用。所比较的内容主要是印刷质量、印字速度、噪声、行宽、机械可靠性及扩充功能的强弱等。下面就对几种常用的针式汉字打印机作些介绍。

一、24针针式汉字打印机

目前，这是针式汉字打印机中印字质量最好、功能较强的机种。它用来打印24×24点阵的汉字，可实现仿宋体汉字的输出。因此，虽然其价格较高，但在需要打印质量较高的文件等场合，都必须配置这一档的针式汉字打印机。若仅利用24针打印机中的16针，则也可实现15×16汉字点阵的打印。但是，这样做的结果是，输出字形质量差，且尺寸过小。为此，通常只有用扩大文字的方法，执行扁体字（30×16）的打印输出，才能满足实际要求。

二、16针针式汉字打印机

带有汉字处理功能的微型机系统，限于屏幕显示汉字的能力及汉字字模库的容量等因素，大多采用15×16点阵的汉字作为处理对象。16针针式汉字打印机就可以作为这类系统的汉字文件输出设备，用来输出15×16点阵的汉字信息。打印机本身可不带汉字字模库，而直接利用微型机内用于显示汉字的字模库，从而可降低系统价格。16针汉字打

印机的字形质量不如24针式的打印机。一般来说，用纵向16点阵表示笔画多的汉字是比较困难的，字体只能采用细线体，不够美观。对于笔画较少的汉字，则15×16点阵也可勉强表示出仿宋体字样，这对于要求不高的应用来说，是可以接受的。此外，为增加打印底线等功能，也已制造出18针式的汉字打印机。

与24针的相比，16针汉字打印机的点数少，打印速度较高，噪声较小，价格较低，扩充的控制功能较齐全。因此，它是易于普及的机型。

三、用9针针式字符打印机打印汉字

9针式打印机通常配置在西文信息处理计算机系统中用来打印字符。尤其是广泛配置在微型机系统中用来实现文件输出。当在原有的用于西文处理的微型机系统上扩充汉字功能时，往往可以不再增添汉字打印机，而是把原配置的9针打印机兼作汉字打印机来使用，从而达到整个系统汉字、西文兼容的目的。

用9针打印机输出的汉字点阵通常为15×16点式，必须分两行来打印，每纵列的16点信息要分散到两行中，每行分打8点信息。显然，要使汉字点阵的上、下两半连接无缝，则要控制行距量为零。这在技术上是比较困难的。同样，要使两行汉字能相互区别开，则要安排一定行距。这样，9针打印机在打印汉字文件时行距是变化的，打印奇数行后的行距同打印偶数行后的行距设置应有区别。此外，在汉字点阵信息的组织和传送上，也要充分考虑到汉字点阵需拼接的特点，这需要根据打印机所提供的扩充控制功能来组织打印程序。

9针打印机印出的汉字呈矩形，字形质量较差，有明显分离点感觉。只适合用于要求不高的汉字文件打印。但由于这种打印机使用普遍，且价格较低，故为用户所乐于使用。

用户在选用各类汉字针式打印机时，除字形质量外，还必须了解打印速度，每行最多可以打印的文字个数（即行宽）等。表7-7列出了各类针式汉字打印机的各项性能的比较。

表7-7　针式汉字打印机性能表

机　型	高　档	中　档	中　档	低　档	简易型
国外参考型号	PC-PR201	NM-9100	SM-16P	FP-100K	GP-550E
打印头针数	24	18	16	9	2
汉字点阵	24×24	16×16	16×16	16×16二次打	16×16多次
打印速度	40汉字/秒	60汉字/秒	60汉字/秒	24汉字/秒	17汉字/秒
适用范围	微型机	同左	同左	同左	兼图形打印

四、其他类型的针式打印机

各类针式字符打印机都可以作为汉字打印机来使用。与9针打印机相似，还可以使用8针、7针的打印机，甚至可利用单针或针点交错的双针针式打印机来打印汉字。显然，它必须往复打印数行才能拼打成汉字，双针针式打印机由于针点重叠一部分，故可以打印出质量较好的汉字。此外，由于它的噪声很小，造价更低，故是一种简易的针式汉字打印机。

针式汉字打印机由于是串行工作的，故打印速度不易提高。为此，出现了一种梳齿

状并行针式打印机。打印头呈梳齿状，并行排列60根针，用每根针打印一个汉字。其工作方式是，60根针同时打印完一点后，打印头便移动到相邻点的位置，再同时打印第二点。如此重复下去。当打印完每个汉字的横向点阵行后，便走纸相当于一排点的距离，再重复上述打印周期，如此将一行汉字全部打完。这样做，由于多针并行工作，故可减少打印头横向行程，提高打印速度。此外，由于它的打印方式与显示器的扫描方式一致，控制程序简单，字形也较好，故是一种颇有特点的针式汉字打印机。

7.3 激光汉字印刷机

7.3.1 激光汉字印刷机的工作原理

激光汉字印刷机是汉字印字输出设备中结构精密、功能完备的印刷装置，适于配置在规模较大的汉字信息处理系统中。它可以成批输出有高质量汉字的文本或报表。

激光汉字印刷机的印字原理及机构如图 7-15 所示。

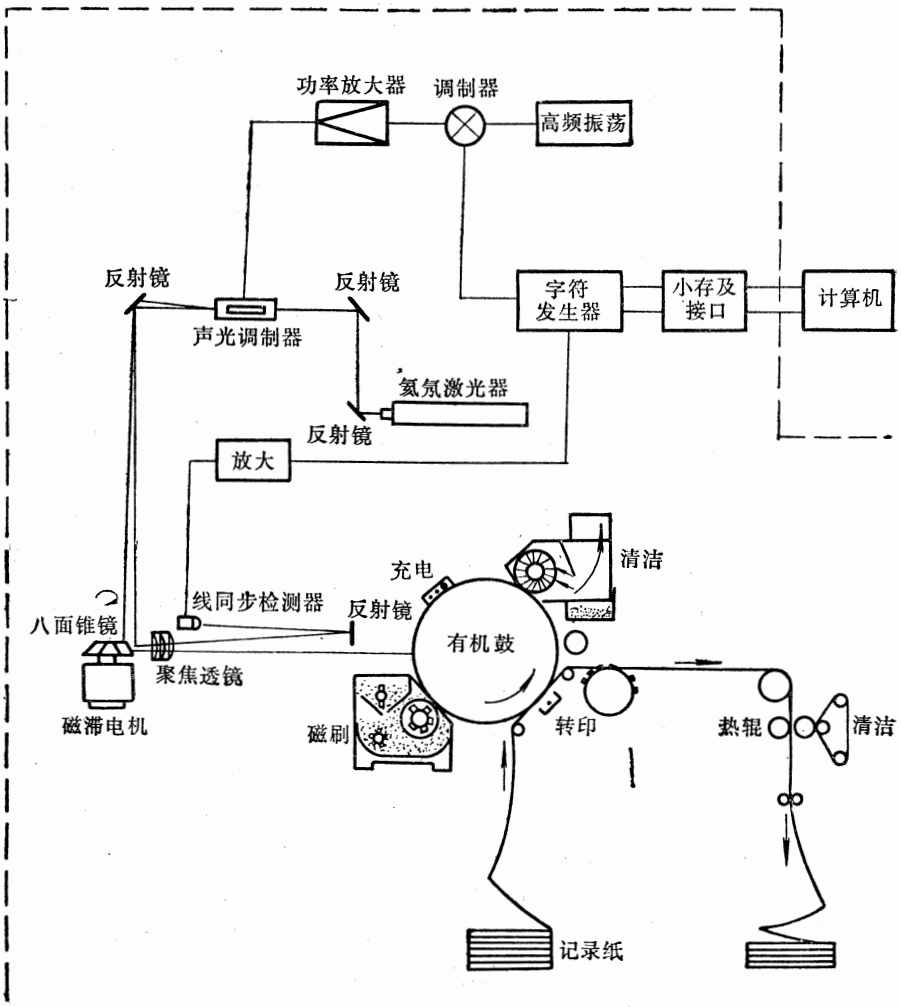


图7-15 激光汉字印刷机的工作原理

激光印刷机的核心部件是激光器（通常采用氦氖激光器）。它输出的单色激光束经反射镜投向声光调制器作信息调制。然后载有汉字信息的第一级光束，通过反射镜投射到旋转的多面锥镜上，由于锥镜的旋转将激光束展开成水平方向的扫描线段，再经聚焦透镜的会聚，将一条载有汉字信息的激光束投射到旋转的记录鼓面上。硒鼓上预先充有均匀的电荷，当鼓面经载有汉字信息的激光束曝光后，鼓面上被照部分失去电荷，形成静电潜象。

鼓面潜象转到显影区时，由磁刷进行干式显影，于是鼓面上呈现带有墨粉的汉字图象，该图象继续转到转印区，由于转印电极对记录纸充电，使纸面也带有一定的电位，在静电场的作用下，将鼓面上和汉字图象相应的墨粉分布吸附到记录纸上。转印后的汉字图象是不能长久保存的，极易擦掉，尚需定影处理。当记录纸移到定影区后，经加热辊加热，使墨粉内的树脂溶解，使字符图象牢固地与记录纸结合，完成汉字文件的印刷。

转印后的鼓面，尚留有相应于汉字图象的残余墨粉。为便于清除，经消电灯去除鼓面电荷，墨粉基本成中性而浮在鼓面。然后，经清洁区，由旋转毛刷的擦刷，将残余墨粉全部清除。擦刷下的墨粉由吸粉管道回收到墨粉匣中。经清洁后的记录鼓表面继而重复上述充电、曝光、显影等一系列印刷记录过程。

声光调制器对激光束的调制，由激光印刷机的控制器来执行，控制器接收主机送来的汉字印字信息，经控制器转换为汉字字形信息而去调制激光束。

激光束经声光调制器调制后的零级光束，通过反射镜投向线同步检测器检测。所产生的控制信号送入控制器，用以控制扫描的起始点。这样就能确保锥镜在每个镜面在扫描出水平线段时，始终有一个一致的起点位置，以保证印刷质量。

激光束的扫描方式有上述单束扫描方式和多束同时扫描方式两类。对激光束的调制方式，更是因调制器的类型不同（例如有声光调制和电光调制两大类）而有差别。

7.3.2 激光汉字印刷机的组成

激光汉字印刷机由光-机-电结合的印刷机构和控制电路两部分组成。如图 7-16 所示。

一、光-机-电结合的印刷机构

如上所述，激光印刷机的印刷机构综合了激光，电子照相，机电控制等多方面技术。整个机构由激光部分、印字部分、转写部分、显影部分、接纸部分、定影部分、清扫部分组成。

这些部分使用了大量控制电机，其驱动和停止全部由微处理机来控制。因此，动力系统结构清晰，控制灵活。这些电机有：

棱镜电机（反射激光束）；硒鼓电机（印字）；磁辊电机；墨粉搅动电机；墨粉浓度检测器传动电机（显影）；加热辊电机；硅油毡输送电机（定影）；走纸链驱动电机；转写器顶纸板驱动电机；拖纸辊电机；偏歪走纸调整电机；吸纸器抽风电机；吸纸器内滑块移动电机（输纸部分）；接纸台升降电机；前折纸板转动电机；后折纸板转动电机（接纸台部分）；清扫辊电机；清扫辊外罩移动电机；墨粉收集抽风电机（清扫部分）等。

这些电机是保证高速、高质量进行汉字印刷所必需的。当然，对于中、低速的激光汉字印刷机，其动力系统可大为简化。

为了使微处理机能准确、及时地控制各部分动作，必须对系统在印字过程中各部分的状态（电流的过流、欠流；电机转速；机械构件的位置；压力；墨粉浓度、温度；打印纸存量等）作出监控，为此系统使用了下述一些传感器：

（1）激光部分——棱镜电机转速检测传感器。

（2）印写部分——带电器、转写器电压检查、锥电灯电源检查、硒鼓电机转速检出等传感器。

（3）显影部分——磁辊电机转速检出、墨粉浓度检出、墨粉箱空了检出，墨粉加入口打开位置检出、放显影体拉杆打开位置检出传感器。

（4）定影部分——预热板温度检出，加热辊温度检出、压力辊电机位置检出，硅油毡用完检出、硅油用完检出、硅油毡后边的门打开状态检出传感器。

（5）输纸部分——纸将用完检出、用纸宽度检出、导纸孔是否准确检出、缓冲臂位置检出、走纸电机位置检出、吸纸器滑块位置检出、走纸偏歪检出、转写器顶纸板位置检出、拖纸辊处理卡纸检出、纸未进入接纸台检出、走纸支架位置不适当检出等传感器。

（6）接纸台部分——接纸台卡住、接纸台位置高低检出、接纸台纸满检出、接纸台中折纸板转矩检出等传感器。

（7）清扫部分——清扫辊电机转速检出、清扫辊外罩移动电机状态检出、墨粉收集抽风电机压力检出、收集墨粉的桶打开状态检出等传感器。

此外，还有各驱动电机的温度过高状态输出及电流异常检出，各种动作时间间隔监视等传感器。

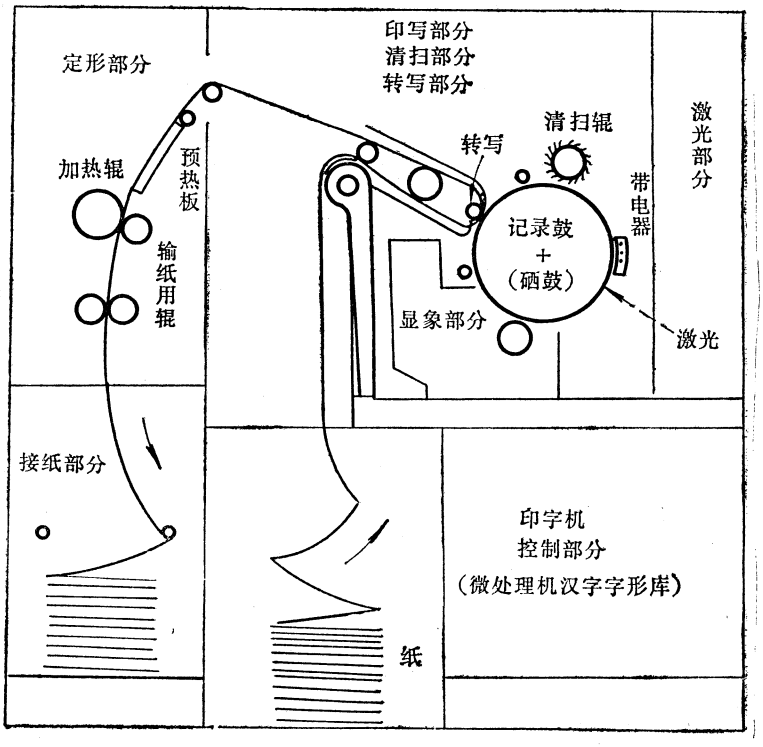


图7-16　激光汉字印字机的组成

二、激光印刷机控制电路

印刷机控制电路用来控制印字机构的动作，同时，又负担印字机与主机的信息交换。这样的控制电路由多台微型计算机组成，它用来实现汉字印字信息处理与对激光印字机构的控制。其构成如图 7-17 所示。

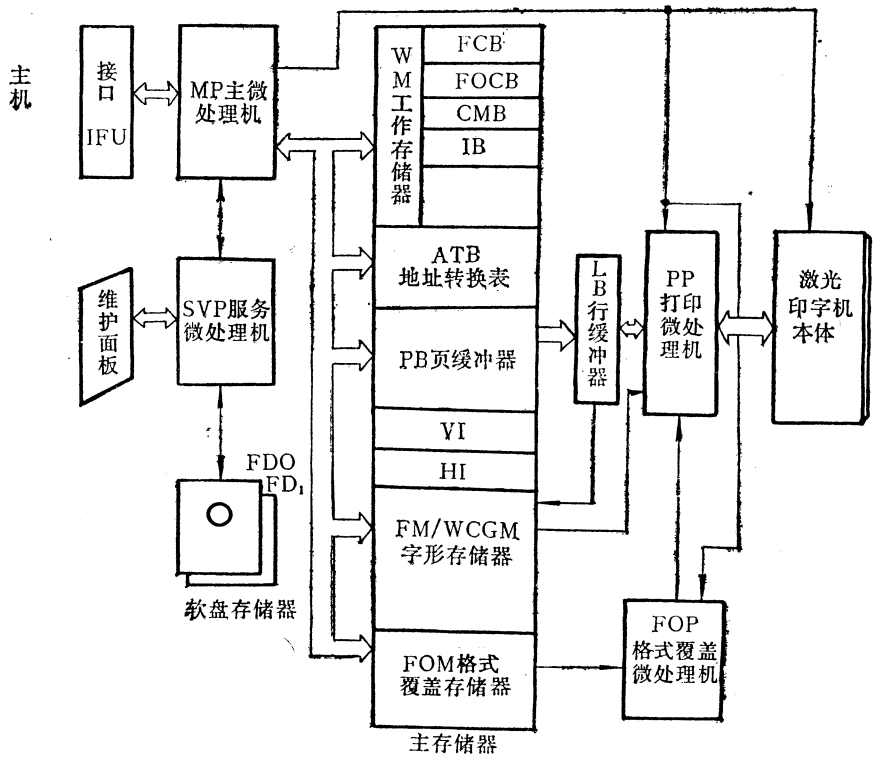


图7-17　激光汉字印刷机的控制电路框图

整个控制器内包含四台微型机。图中，存储器各部分的功能为：

（1）WM(工作存储器)——它包括印字信息控制所需要的格式控制缓冲器(FCB)、格式覆盖控制缓冲器（FOCB)、拷贝修饰缓冲器（CMB)、中间缓冲器（IB）等，其容量为 64 K字节。

（2）FM、WCGM(字形存储器)——它用以存放汉字字形的点阵库，其容量为1M字节以上。

（3）ATB——地址转换表，64K字节。

（4）PB——页缓冲器，256K字节。

（5）VI——垂直方向扩大缩小指示，64K字节。

（6）HI——水平方向扩大缩小指示，64K字节。

（7）FOM——格式覆盖存储器，192K字节。

（8）LB——行缓冲寄存器，2K字节。

控制电路中各个微处理机的功能为：

（1）主微处理机（MP)——它是一种总控处理机，负担激光汉字印刷机与主机间

的各种接口信息变换。

（2）服务微处理机（SVP）——它用来控制印刷机所附的软盘驱动器动作。同时，分担在维护印刷机时的各种操作。

（3）印字微处理机（PP）——用来控制各类工作存储器（PB、VI、HI、FM、LB），组成供KPR使用的激光调制信号，提供对印字机械的执行控制信息。

图7-18所示为印字微处理机系统的组成。其中，控制存储器用来存放各项控制程序（固化于24K字节的EPROM中）；数据存储器为4K字节；诊断程序存储器用来存放印字机械自检时的诊断程序（固化于16K字节的EPROM中）。各种子系统控制包含：管理程序控制、数据存储地址控制、中断控制、时钟控制、错误检测控制等。其他部分则是与印字机构一一对应设置的控制和检测部件。

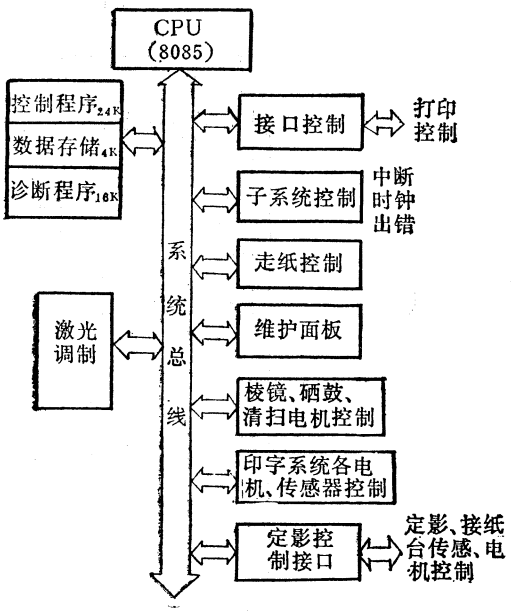


图7-18　控制电路的印字微处理机系统的组成

（4）格式覆盖微处理机（FOP）——它专用于控制从FOM中输出格式覆盖的信息，与LB的信息重叠，送印字机构印出。

7.3.3　激光汉字印刷机的动作与性能

一、激光印刷机的动作过程

完整的印刷机动作分为三个过程（参见图7-19）：

（一）数据传送过程

这是从主机通过输入输出通道将以行为单位的汉字数据送至汉字印刷机的页缓冲器（PB）的过程。在这个过程中，首先把由通道送来的以行为单位的数据存放在中间缓冲器（IB）中，接着由地址转换表部件（ATB）将IB中的数据代码（汉字编码）变换成汉字字模存储器（FM）中相应字形点阵的地址，然后，存放于页缓冲器PB中。最后，由主机送来走纸信息存放于格式控制缓冲器（FCB）中。

（二）数据记录过程

当满一页的数据存入页缓冲器后，记录过程便开始。这时，将存在页缓冲器的数据一行一行地移入行缓冲器（LB），行缓冲器的数据访问字模存储器FM，读出相应汉字的字形点阵，从而把一行信息变换成点阵。接着，用此点阵信息去调制激光束，在记录鼓面上引起汉字行潜象。这样，记录完一页的信息后，便把下一页的信息送入页缓冲器，开始一次新的记录过程。

（三）转印、定印过程

这是前面已介绍过的激光印刷机的动作过程，目的是完成汉字文件的印字作业。

二、激光汉字印刷机的处理功能

激光印刷机的处理能力远较针式打印机强。除了一般的打印机的控制功能外，它还

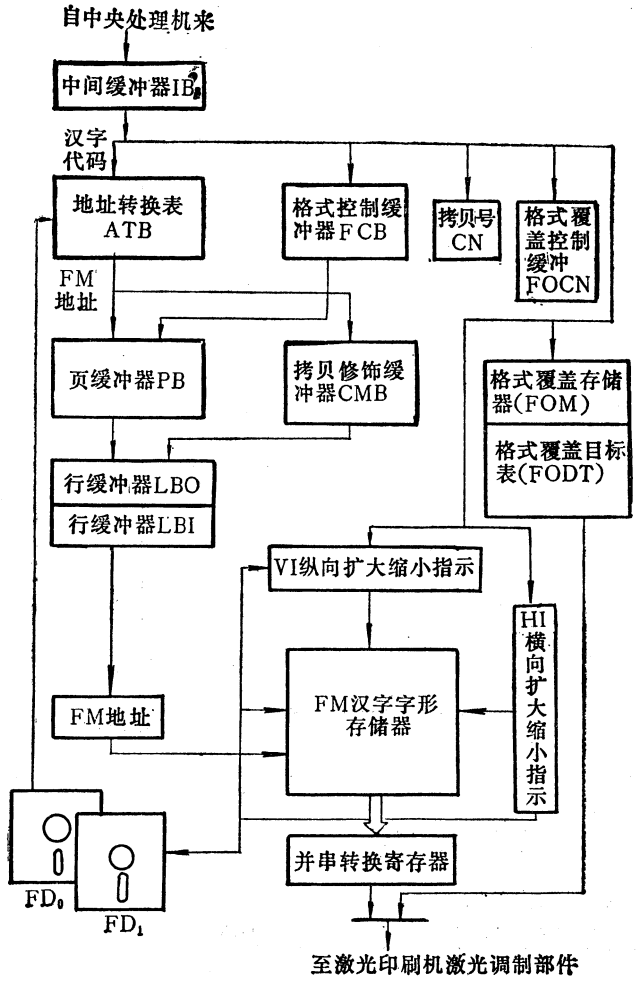


图7-19　激光汉字印刷机的动作过程

有不少特殊的处理功能，如格式覆盖、字形尺寸变换、拷贝修饰、字形修改、外字处理、自检诊断等。包括硬件和软件的控制命令在内，其命令不下90余种。表 7-8 中列出了几种典型的主机送来的通道命令，各种控制功能由这些命令来配合成，以下就几种主要功能作原理性的介绍：

（一）格式覆盖功能

汉字文件中常使用表格（包括划线、标题、栏目、项目等）。同样的表格内可填入不同的汉字和数据。为了减少主机与汉字印刷机之间的信息传送量，可把描述表格的程序独立出来，编上号。使用时，一次向印刷机内装入。之后，当需要使用某个表格时，主机只要再送该表格的编号和需填入的汉字数据，即可得到带有汉字数据的表格来。这就是格式覆盖功能。

格式覆盖功能的执行过程如图 7-20 所示。首先，用格式覆盖记录来描述要输入的格式，经过格式覆盖生成程序，作成格式覆盖模块。然后，将格式覆盖模块中的格式覆盖程序存入格式覆盖存储器（FOM），并将存入时的起始地址和程序的标识符存入格式

表7-8　控制激光汉字印刷机的通道命令

• 走纸格式命令	(1)	空走纸M（M = 1～3）行
	(2)	跳到N（N = 1～12）所指定的行
• 写命令	(3)	把要印的字写到页缓冲器中，不改行
	(4)	把要印的字写到页缓冲器中，空M（M = 1～15）行
	(5)	把要印的字写到页缓冲器中，跳N（N = 1～12）行
• 装填命令	(6)	装填走纸格式
	(7)	装填拷贝某些页要少量修改的数据
	(8)	装填进行拷贝的起始页号
	(9)	装填格式覆盖程序和它的目录表
	(10)	装填控制格式覆盖程序如何使用的数据
	(11)	装填字形点阵到字形存储器(FM)
	(12)	变更写命令所指示的数据打印位置
• 控制命令	(13)	空操作
	(14)	打印初始化(打印PB剩余内容后置初始状态)
	(15)	清除打印(打印PB剩余内容)
	(16)	块数据检查
	(17)	传送结束
	(18)	非汉字方式转成汉字方式工作
	(19)	汉字方式转成非汉字方式工作
	(20)	测试输入输出设备状态
	(21)	读出判断字节(记录设备向主机报告的信息)
	(22)	读出传送给印刷机的字符中的无效字符
	(23)	读出记录错误的寄存器内容
	(24)	读出中间缓冲存储器的内容

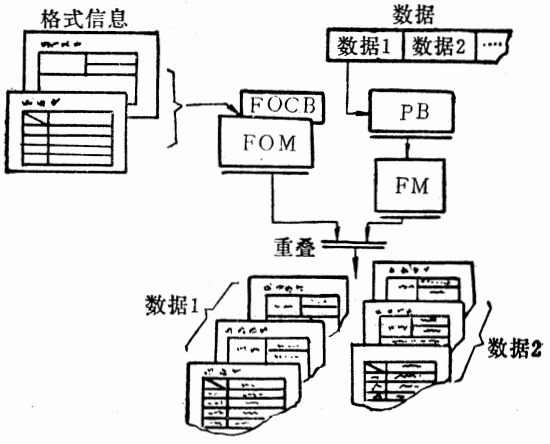


图7-20　格式覆盖功能的执行过程

覆盖目录表（FODT）。印字时，将数据和汉字经过输入编辑程序、变换成要印出的结果并放在输出缓冲区。接着，把这页数据要使用的格式覆盖程序的标识符和要印出的张数装填到格式覆盖控制缓冲器（FOCB）。然后，把输出缓冲区的数据送页缓冲器(FB)，再将此数据和表格重叠后送到印刷机印出。

以上过程需用通道命令来组织程序。这里就省略了。

（二）拷贝修饰功能

激光汉字印刷机不能像针式汉字打印机那样，一次打印数分复本。但可以用重复印刷页缓冲器的内容来得到文件的复制。如果需要使要输出的每一页面都有少量的不同（例如，通知书中收件人的姓名）或者需要对已印过的页面作少量修改等，都可以利用拷贝修饰功能来处理。这样做，不仅可减少机内信息的传输量，而且可大大减少编制程序的工作量。

图 7-21 为拷贝修饰功能的流程图。其执行过程为：首先，利用“加载拷贝修饰命令”，从主机送来拷贝修饰数据。数据格式中前面 6 个字节存放要修饰的位置，后面最大 255 个字节存放修改用汉字的编码；然后，经地址转换表（ATB）将编码变换成字模存储器 FM 的地址，并连同其他数据写入拷贝修饰缓冲器（CMB）。在指定了从第几页开始印刷和使用哪个格式的命令后，便把要印的数据写入页缓冲器PB。接着，按CMB中的位置参数比较符合的页编号、行号和文字位置。符合时，将修改用字送到行缓冲器LB的相应位置，冲掉原有的文字。这样就可实现页文件的局部修改。图 7-22 就是使用拷贝修饰功能执行文件输出的实例。这在事务处理范围内是很有用的。

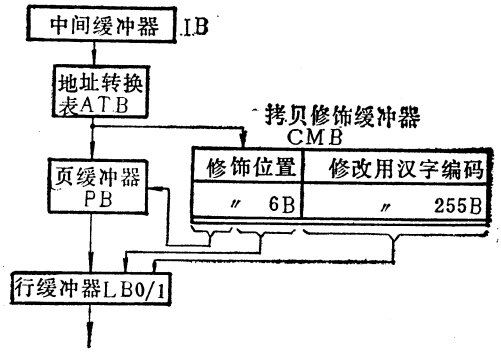


图7-21 拷贝修饰功能的执行原理

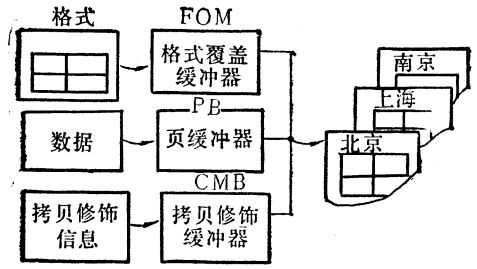


图7-22 拷贝修饰功能的使用实例

除了用程序方法组织常用表格外，激光汉字印刷机还可选配胶片覆盖功能。此功能是将常用表格格式图用照相方法作成胶片，使用时，把胶片上的表格图直接投影到记录鼓上，并与控制器送到激光扫描器的数据重叠，从而实现印出操作。显然，对图素复杂的表格，用这种方法是很简便而有效的。

（三）字形变倍功能

一般激光汉字印字机使用32×32的汉字点阵，即每个汉字要用 128 字节来存放点阵信息。8192 个汉字需备 1M字节的字模存储器（FM）。在汉字处理作业中经常需要在文件中加入小号字（如注解用字）和大号字（如标题）。为这些字体再组织字模库显然是

极不经济的。设置字形变倍功能即可用32×32字形点阵为基础，获得扩大和缩小的汉字点阵。以下介绍它的实现方法：

如果把32×32点的字形，纵向看作32行光栅，横向看作32列，则只要实现每4行抽掉一行，每4列抽掉一列，即可变成字形相近的24×24点阵字形，这就可用于小号字印刷；若实现每四行增加二行，每四列增加二列，则可形成字体相近的48×48点阵字形，这就可用于大号字印刷。因此，若用增加一位信息来标志该行（或列）变倍，则整个字扩大和缩小标志（又称HI、VI信息）反用16字节，这样，8192字共需增加128K字节。扩大和缩小标志先在主机上用程序作成（按一定规则，确定一个汉字点阵中本行或列是否可去掉或重复），然后再经过人工修正，最后和字形一起填到FM、HI和VI的存储区，并存入软盘。

硬件在读FM时，如果要印小号字，则缩小标志为1的行和列禁止输出；若要印大号字，则扩大标志为1的行和列重复输出。

对于64×64点阵汉字，128×128点阵汉字，只要让每行、每列重复印两遍、四遍即可。

硬件结构可参见图7-18。

（四）字形修改和外字处理功能

激光汉字印刷机的汉字字模库先是存放于软盘中的，开机后，便加到FM中去。在使用过程中，由于汉字字形使用的变化（如简化字、字体更动等），有时需要作修改，此外，用户往往需要一些特殊的汉字或图形符号。文件中要用到国标一、二级字以外的汉字时，就需要在汉字字模库中增添新的汉字，这个功能称为外字处理，这里，把加电后从软盘自动装填到FM中的字模称为内字。

字形修改的过程如图7-23所示，其执行过程为：

（1）清除字形。将地址转换表ATB中各单元的V位（为“1”时，标志FM对应单元中存入字形有效；为“0”时，无效）全部置“0”。把指示FM地址单元使用界限的FMA指针置于08(00～07为硬件固定使用的字模)。

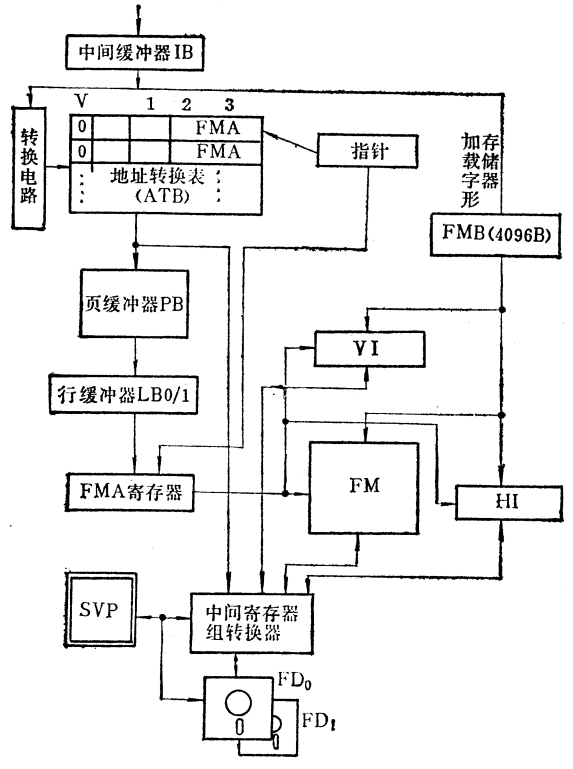


图7-23 汉字字形修改的流程图

（2）装填字形点阵。按规定的数据格式把汉字字形数据及HI、VI数据先存入缓冲器FMB，然后由指针所指出的地址，写入FM、HI、VI。并将相应的ATB表中V位置成“1”，这个过程不断重复，直到新的汉字字形全部装完。

（3）存入软盘。即在SVP处理机的控制下，读出FM、HI、VI中所存数据并记入

软盘 FD0，读出 ATB 记入软盘 FD1。

图 7-24 所示是一个汉字字形数据的结构。其中，L 为字符级别(ATB 中登录的汉字分为四级)；A 用来区分汉字和非汉字；B～E 用来区分字体、字种；F 用来指示后面数据的有效长度；HIC为水平方向缩小信息；HIE 为水平方向扩大信息；VIC和VIE分别为纵向缩小和扩大信息。

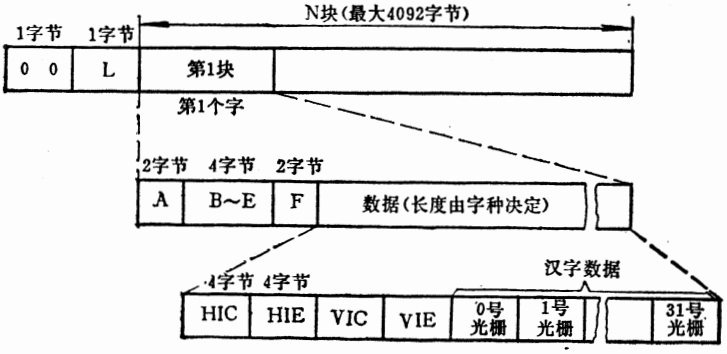


图7-24　汉字字形数据的结构

在外字处理时，其情况就稍为复杂一些。由于 FM 的容量有限，只能装入 0 级字(ATB中登录的级别)。还有大量的罕用汉字作为 1 、2 级字装在软盘中，作为“预装填外字”。此外，还有在使用中需临时装填汉字字形点阵，称为“请求装填外字”，ATB 将其作为 3 级汉字看待。

预装填外字的处理过程如下。先装填字形点阵；根据要装填的汉字代码在 ATB中找到一个单元，记入FMA 指针，并将字形数据存入 FMB；然后，待PB中的内容输出完，就将 FMB 的内容写入指针指出的 FM、HI、VI存储单元中，并将ATB中相应的V位置成“1”。此后，在写入命令中就可使用预装填的外字。

请求装填外字的处理过程如下。在写入命令中使用了字模库中没有的汉字代码时，在ATB上检出为失效字符。待一页写完后，向主机报告检测结果，主机便发出读判断字节命令，将印刷机记录中是否存在失效字符及失效字符的个数读入主机。然后，主机发读出失效字符命令，将失效字符的代码取入主机。由主机完成该外字的字形数据装配，再使用装填字形点阵命令存入有关信息；字模数据存入 FMB，并在 ATB中登录此汉字代码（作为 3 级字)。再按指针指示写入 FMA。最后发打印命令，将该页印出。

7.3.4　激光汉字印刷机的选用

激光汉字印刷机的优点是：印字质量高——由于激光束聚焦很细，故印字图象清晰；印字速度快——由于激光功率强，从而可缩短曝光时间（在高速印刷时可达 2 万行/分)；可用普通纸印刷；等等。此外，利用强功率激光可直接制作胶版，其效率很高。当然，激光印刷机是很复杂的设备，成本很高，只宜在规模较大的计算机系统中使用。为适应汉字信息处理技术的发展需要，近年来也在研制和生产各种简易型激光汉字印刷机，供小型汉字系统配用。下面介绍几种常见的激光汉字印刷机。

一、高精度激光汉字印刷机

它通常称为行式激光汉字印刷机，适宜于连续高速输出汉字文件。如果增加分页部件，也可实现高速的以单页纸为单位的文件输出，并可以实现自动分页操作，使输出文件自动按页整理好。行式激光汉字印刷机除了前面所介绍的各个组成部分外，还可以根据需要增添外围部件，例如带显示器的控制台，文件记录用磁带存储器，磁盘存储器等，从而组成规模较大的独立汉字印刷系统。除联机作业外，还可以完成大量的脱机印字作业。

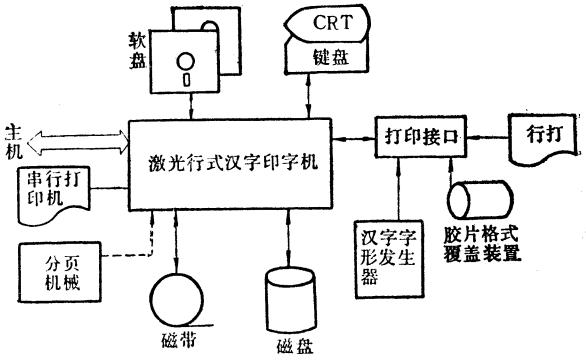


图7-25　行式激光印刷机的扩充组成

图 7-25 是行式激光汉字印刷机的基本系统与扩充系统的构成。大型计算机系统可配置扩充系统，中、小型计算机系统常只配置基本系统。由于在性能和价格上，激光汉字印刷机的各个机种差别甚大，故在选用时必须充分调查。表 7-9 列出各类激光印刷机的型号、性能及用途，作为选用参考。

表7-9　激光汉字印刷机的性能一览表

机　型	高速行式	中速行式	中、低速页式
参考型号	IBM3800	IBM6670	LBP-10
印字速度	20040行/分	1800行/分	500～700行/分
分辨率	水平　5.7～8点/毫米 垂直　7.1点/毫米	4～8点/毫米	水平　18.9点/毫米 垂直　9.45点/毫米
应用领域	汉字、数据处理	数据处理、印刷文件、报表	文件处理电传通信

二、简易型（页式）激光汉字印刷机

这是在静电复印机基础上发展起来的激光印刷机。由于它的价格低、性能优良，故深受用户欢迎。静电复印机的记录、显影、定影等功能与上述激光印刷机的相应部分是近似的，只要再加上激光、调制部分即可组成完整的激光印刷机。若舍去走纸系统，改用静电复印机原来具备的页式送纸功能，则成为一台页式激光印刷机。

若页式激光印刷机配上比较简单的控制器（其规模与针式汉字打印机相近），就可以与主机相连，成为页式激光汉字印刷机。

依据静电印刷机的机构及控制器的规模不同，页式激光汉字印刷机品种极为丰富。由于价格低，在小型计算机系统中可普及应用，高档微型机亦可望使用低档的页式激光汉字印刷机。

这类印刷机的代表性能可参阅表7-9。

7.4　喷墨、热感、静电、光纤管等式样的汉字印刷机

针式及激光汉字印刷机已成为目前汉字印刷机发展中的主流。但对于其他类型的印刷机，由于各自的特点，也在某些场合被用作汉字印刷机。本节将对它们的原理、特点

及性能作简单的介绍。

7.4.1 喷墨式汉字印刷机

一、工作原理

喷墨式汉字印刷机是使从喷嘴喷射出的墨滴发生偏转而形成字形的印字记录设备。按墨滴产生的性质，可将其分为连续式及冲击式两类。连续式是喷嘴连续喷射墨滴，并在墨滴运动过程中控制其运动方向，使它落到纸面的预定位置而印出字形；冲击式是控制墨滴本身的喷射(仅在需要时才进行喷射)，而喷出墨滴的运动方向是固定的。后一方式虽技术难度较大，但结构简单，因而是有希望发展的方式。若按喷墨印刷机的墨滴偏转方法来分，则又可分为电场控制型及电荷控制型两类。以下介绍使用较普遍的电荷控制型喷墨印字原理（见图7-26)。

如图7-26所示，墨水容器中的墨水经泵加压后送至墨水喷管头。由压电振子加上一定频率（如115千赫）的振动，墨水与之同步，形成墨水微粒连续地从喷嘴射出。喷出的墨水微粒通过带电的电极时，由于电极电压受汉字字形信息的调制，墨滴上就分别地带上了与字形相应的电荷。接着，带有电荷信息的墨滴通过带高压的垂直方向偏转板使墨滴产生静电偏转（偏转量大小与锯齿波电压的瞬时值有关，使墨滴在到达记录纸时形成文字上的一点。不带电荷的墨滴则保持水平方向飞行，由墨滴收集器回收，送到墨水容器中再使用。偏转板上的偏转电压每变化一个周期就形成汉字文字的一个纵列，控制喷墨头作横向移动就可形成一行完整的文字。如果增加横向偏转的静电偏转板，则也可直接扫描完整的文字。

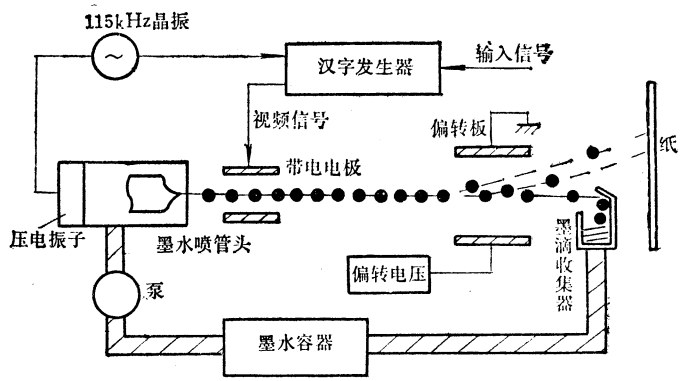


图7-26 电荷控制式喷墨印刷机工作原理

二、喷墨印刷机的组成及控制

从上可知，由于电荷式喷墨印字操作与针式汉字打印机的情况很相似，因此，在结构上也很相近。图7-27是喷墨印刷机的组成框图。其组成部分是：

（1）喷墨机构——由墨水系统、喷射头及偏转系统所组成。

（2）电气及驱动电路——它是直接控制喷墨印刷机构动作的电路。它包括喷射头传动电机驱动电路、走纸电机驱动电路，墨水泵控制电路，高压偏转电压电路以及使墨滴带电荷的控制电路等。

（3）控制器——这是用微处理机构成的全机控制核心，用来控制电气驱动电路的动作，并担负喷墨印刷机与主机的接口。与针式汉字打印机一样，喷墨印刷机控制器可自带汉字字模库，也可以由主机送汉字点阵信息到印刷机中。控制器的组成与针式打印机相似。其接口也趋于采用标准化并行接口。

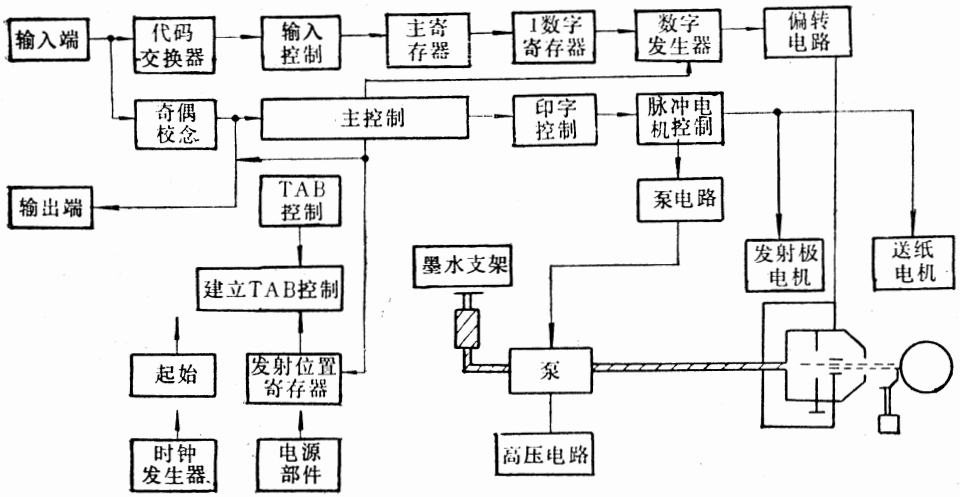


图7-27　喷墨印刷机的组成电路框图

三、喷墨汉字印刷机的性能特点

喷墨汉字印刷机作为一种非击打式印刷机具有以下一些特色：

（1）可使用普通纸印字，运行成本低。

（2）具有较高的印字质量。由于喷射头的喷嘴直径仅数微米，墨滴极细小，很易达到8～12点/毫米的分辨率。可与铅字印刷比美。普及型机种常采用24×24及15×16点阵的汉字字形。

（3）印字速度快，可达90～120字/秒。

（4）噪声小，宜于办公室环境使用。

（5）易实现彩色印字。在喷射机构中安置三个喷射头，分存三种颜色的墨水，即可实现彩色文字和图形的印刷。这是彩色喷墨印刷机目前很受用户欢迎的重要原因。

目前，在国外对于能作汉字及图形输出的彩色喷墨印刷机，在其体积及价格方面，都已与针式汉字打印机不相上下。因此，它是一种很有希望发展的印刷机。表7-10列出了几种喷墨印刷机的性能参数。

表7-10　喷墨式汉字印刷机的性能表

机　型	行　宽	速　度	汉字点阵	适用范围	国外参考型号
喷墨式	45汉字	40汉字/秒	16×16	微型机	
	90汉字	50～90汉字/秒	16×16 24×24	微型机	JD-1000
彩色喷墨	132字符 90汉字	50汉字/秒	24×24	彩色硬拷贝	SHARP500

必须指出，喷墨式印刷机要求保持有清洁的环境，以免灰尘堵塞喷嘴，同时保证满足墨水特殊的性能要求。

7.4.2 热感式汉字印刷机

一、热感印字原理

热感印字是将热感印字头同受热便改变颜色的热感纸相接触，热感纸接触发热点（计算机输出的印字信息用来控制印字头上的发热元件瞬间发热），发热点发生化学变化而变色，形成所需的点阵文字字形。

热感印字头有多种制作方式。常见的有薄膜电阻方式、厚膜电阻方式及半导体扩散电阻三类。薄膜型印字头由于它具有可形成较精细的图形及热惯性小等优点，故应用较广。这种印字头用薄膜技术在陶瓷基片上制作出一个文字点阵数的发热单元，呈矩阵形式排列。发热单元为点状，由电阻材料（如 Ta_2N）制成。当电流流过电阻体时，由于电热效应而产生热量。电阻发热单元所组成的矩阵的一端为公共接地端，另一端则与字模信息线相连，由控制器控制。

热感纸是种特殊加工的纸张。常用的是有机染料发色型。它是用无色染料和酚类化合物当作发色材料，将它们研磨成微小颗粒，然后与粘合剂、填料均匀搅拌在一起，最后喷涂在纸基上而制成的。纸受热就会因化学反应而变色，反应速度与纸上染料，酚类化合物的含量有关。一般当印字头的发热单元电阻为 100 欧姆时，加上宽度为 4 毫秒的脉冲，即可在热感纸上形成色点。图7-28所示为热感印字头和热感纸的结构。

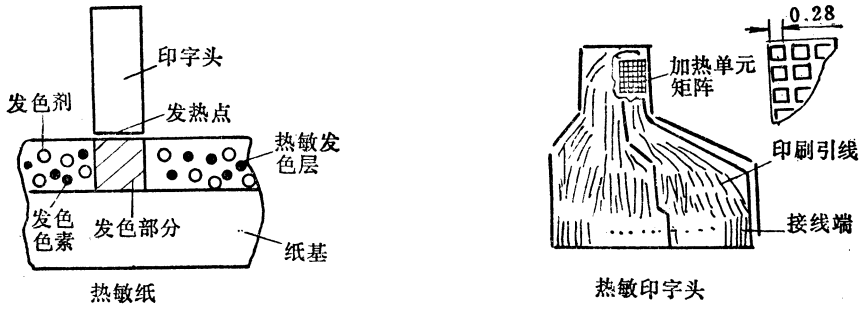


图7-28　热感(敏)印字头与热感(敏)纸

二、热感印刷机的组成

热感汉字印刷机的组成与针式打印机相类似，由于热感印刷的机理特性使热感印刷机可省去色带及其传送机构，故整机结构较为简单、紧凑。

（一）热感印字机构

它是热感印字的执行部件，如图7-29所示。热感印字机构是接触式的，其印字头沿纸面作横向运动，以完成一行文字的印字。换行时，为提高返回速度和印字头使用寿命，必须使印字头与纸张脱离接触，在砧板和印字头之间留有 1 毫米的间隙，然后返回左端位置。这是由电磁铁拉动让位摆杆实现的。印字时，使印字头与纸接触，并维持一定压力。砧板上填有硬橡皮以保证接触良好。印字头传动与走纸传动系统都使用步进电机。

（二）印字驱动电路

由热感印字头的发热单元矩阵的驱动电路及印刷机的各种传动系统的步进电机驱动电路所组成。根据热感印字头的性能要求，驱动电路产生所需的加热脉冲波。

（三）控制电路

与针式打印机一样，也是采用微处理机来实现整机的控制，负责与主机的接口和控制印字驱动电路工作。汉字字模库可设在印刷机控制器内，也可以不设而由主机提供字形信息。图7-30中列出了一台热感汉字印刷机的控制器框图。其结构与针式打印机是相似的，也采用标准接口。

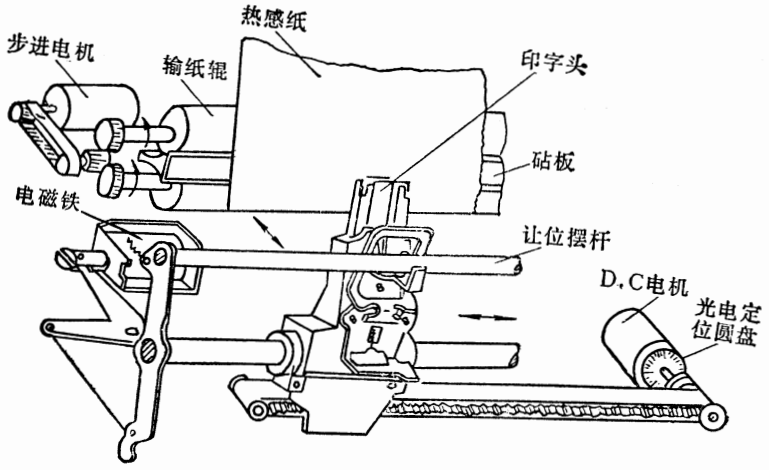


图7-29　热感印刷机结构示意图

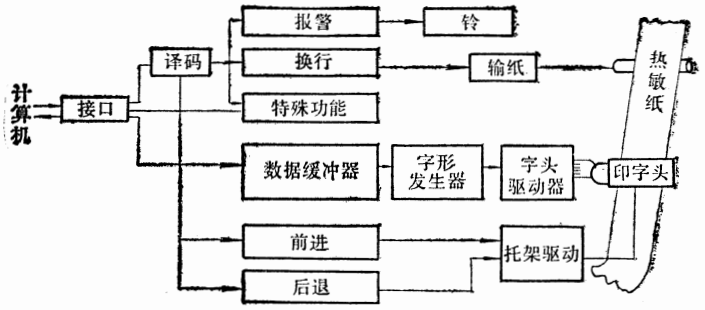


图7-30　热感汉字印刷机的控制器

三、热感汉字印刷机的性能特点

热感式汉字印刷机也是非击打式印刷机的一种，它的特点是：

（1）机械传动结构简单，从而成本低，可靠性高。

（2）低噪声，易小型化，可做成便携式。

（3）可实现中、低速印字（约 30 汉字/秒）。这与印字头热响应特性不易高速化有关。

（4）需使用热感纸，在特殊加工的记录纸中，热感纸的价格较低，但高于普通纸。印好的文件不宜作长时间保存。

（5）可实现高分辨率，印出 15×16 及 24×24 点阵的汉字字形，也可用于图形印刷。

热感式汉字印刷机作为低价格、小型化的汉字印字设备是很实用的，其主要缺点在

于不易实现高速度且必须使用特殊纸。然而，随着热感印字技术的发展，已在研究使用普通纸来实现热感式印字技术，并开始使其实用化，这就是热转印式的热感汉字印刷机。

四、热转印式热感汉字印刷机

以上所介绍的热感印刷机属于接触式，与之相比，热转印式汉字印刷机有很多技术上的突破，从而使热感印刷机开拓了新的应用前景。

热转印式印刷机的原理，简言之，就是用带有热感墨水的热感转印色带代替热感纸，在热感印字头与普通记录纸之间通过热感转印色带，热感印字头上相应的发热单元，加热时，热感转印色带上加热位置处的墨色便会转印到与其接触的普通记录纸上，从而产生字形。其结构很象前述的针式打印机，不同之处在于，热感转印色带长度比普通色带为长，这是因为色带上墨的需要，有的色带只能用一次。此外，若将转印色带换置成热感纸，则照样可以作接触式热感印字之用。

已实用化的热转印式热感汉字印刷机具有良好的性能，印字头上发热单元可做到每毫米 8 点，所印出的 24×24 点阵汉字质量优于针式打印机，而且噪声小。但是，这种印刷机的印字速度尚较低。热转印式印刷机除可使用普通纸印刷外，还可具有彩色印字功能。只要配置三色热敏转印色带即可实现彩色转印，这些特点使热转印式热感印刷机得到迅速发展。

表7-11　各类热感式印刷机性能表

机　型	行　宽	速　度	汉字点阵	适用范围	国外参考型号
热感式	45汉字	20～30汉字/秒	16×16	微型机	STT-201
	90汉字	20～30汉字/秒	24×24	微型机	PU-6000
热转印式	90汉字	15汉字/秒	24×24	微型机	PC-8824(NEC)
彩色热转印	45汉字	6 页/分	8×8 点/毫米^2	七色彩色文件	TN-5000(东芝)

各种热感式汉字印刷机的性能指标列于表 7-11 中，可供选用参考。

7.4.3　静电式汉字印刷机

一、静电式印字原理

静电式印刷机原理与广为使用的静电复印机很相似。首先在作为纪录纸的电介质材料上直接加以高压，获得文字的静电潜象，通过静电力吸附显色剂形成可见图象（即显象）。然后再经过定影，即得印字文件。

图 7-31 为静电印字原理图。当在记录电极和控制电极上施加一定电压时，在电极和记录纸间隙内，由于气体放电产生静电荷保留在纸面上，形成文字潜影。为了使记录纸能使静电荷保留的时间较长些，需在纸面上附着一层合成树脂材料的记录层，约 5 微米。该记录层具有高于 10^{12} 欧姆/厘米的体电阻，这种纸称为静电记录纸。图 7-32中示出静电记录纸的结构和记录电极的印字结构。图 7.32(a)，(c) 所示是采用两层结构的记录纸，适用于记录电极和控制电极分置纸两侧的印字方式。图 7-32(b)、(d) 采用三层结构的记录纸，适用于记录电极和控制电极都在一侧的印字方式，后者能在记录纸厚度变化时，不会影响印字质量，所以使用得较多。电极常采用镍铬丝、钼丝、铜丝等制成，

呈针状，在纸上印出一个圆点，点的密度取决于电极排列密度，一般每毫米可排列 4～8 个电极，即每毫米可印刷 4～8 点。记录电压在 500～1500 伏特，施加电压时间为数微秒。为实现高速印字，多采用横排印字方式，即记录电极按印刷机可印字的行宽横向排成一行，记录针达上千根。

显影过程分干式及湿式两种，图 7-33 为这些方式的示意图。干式有用色粉和载体(铁粉和玻璃粉）混合物向静电潜影加上显象剂；湿式是将色粉掺在以高纯度汽油作为高电阻溶剂之中制成的显影材料。印字时，采用湿式喷射法或用液压装置送到已产生静电潜影的纪录纸上。此称为湿式滚压法。

定影是通过加热，使色粉周围的树脂层融化，将色粉固化在记录纸表面上，最终形成图象。

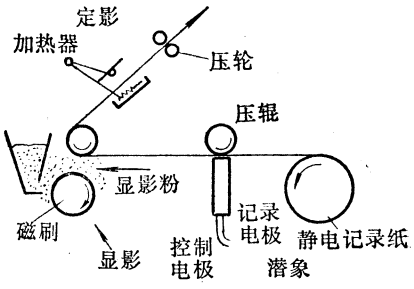


图7-31　静电印字原理

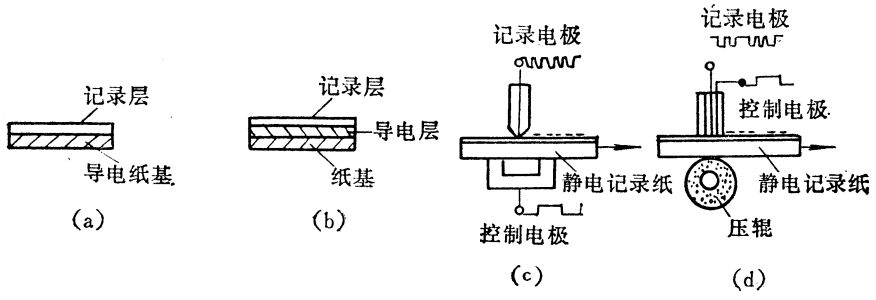


图7-32　静电记录纸和记录电极结构

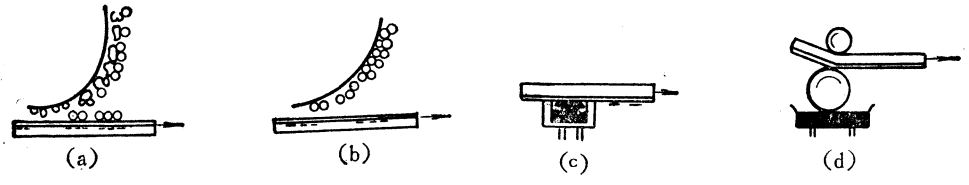


图7-33　常用显影方法示意图

（a） 干式色粉和载体混合法；（b） 干式磁性色粉法；（c） 温式喷射法；（d） 温式液压法。

二、静电式汉字印刷机的组成

静电式汉字印刷机是一种高速行式汉字印字设备，属非击打式印刷机。其组成与简易式激光印刷机相似。其驱动电路部分主要是数量庞大的印字针的高压驱动器，通常是采用编组并串行驱动方式来组织的。其他部分与一般印刷机相似。

控制器也使用微处理机实现，由于汉字字形点阵的处理与激光印刷机类似，这里从略。

三、静电式汉字印刷机的性能特点

与前述的各种类型汉字印刷机相比较，静电式汉字印刷机的主要特点是可实现高速行式印字，输出速度达每分钟数千至 1 万行。而系统结构较激光行式印刷机简单，价格低，工作可靠，能兼作图形印刷输出，印字针直径小，分辨率达每毫米 8 点，印字质量

优良，印字时噪声也小，所以作为中、小型计算机系统的行式汉字印字设备是有前途的。国外已开始出现与微型机配用的简易型静电印刷机。

当然，需要采用特殊的静电纪录纸这一点，是其主要缺点。但由于纸质好，记录可长期保存，作为重要文件的高速输出设备还是可取的。此外，必须把电灼式放电印刷机与静电式印刷机区别开来，前者广泛应用于传真设备，也有作成文字、图形印刷机的。电灼式印字采用特殊的电灼纪录纸，这种纸的纸质差，不易保存，放电记录时有臭味。

表 7-12 列出了静电式汉字印刷机的性能参数。

表7-12 静电式汉字印刷机性能表

机　型	行　宽	速　度	字形点阵	适用范围	国外参考机型
窄行(高速)	45汉字	18000行/分	无限元	大、中、小型机	PPSII/E型
宽行(高速)	90汉字	12000行/分	无限元	大、中、小型机	PPSO500型
中速小型	45汉字	1000行/分	16×16	小、微型机	V-80(XEROX)

7.4.4 光纤管转印汉字印刷机

光导纤维管（Optical Fiber Tube，简写为 OFT）转印汉字印刷机的工作原理如图 7-34 所示。

纪录纸采用连续供给的氧化锌纸，先使纸通过充电器，在高压电晕放电下使之带上负电荷，汉字字形信息送至光导纤维管的控制栅上，当涂有氧化锌的带负电的纸表面受到由汉字字形信息所控制的光照射时，被照射部分的氧化锌层便导通而放电，残余电荷几乎为零。未被光照射部分的氧化锌纸面则继续带有负电，这样在纸的表面便形成了汉字潜象。

接着，带汉字潜象的纸通过显象箱。箱内的显象液是一种墨水，色粒直径在 1 毫米左右，并带有负电，于是纸上带负电的部分排斥墨水，仅潜象部分吸引墨水，使汉字在纸上成象。再经过烘干，进行热定影，送至裁剪部，按规定的页标记将印字纸裁成一定规格的文件页，并在积纸部收集。

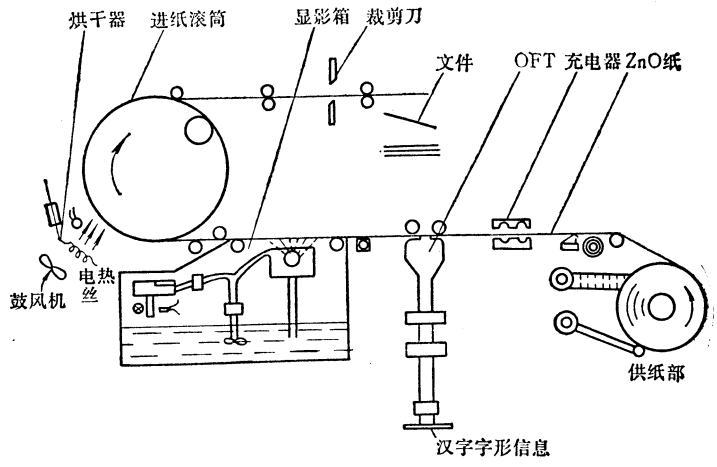


图7-34 光纤管转印汉字印刷机原理图

光纤管是一种用光导纤维作屏面的阴极射线管。它直接在记录纸上曝光成象，可得很好的印字质量，分辨率可达 10～50 线/毫米，采用行扫描中短余辉管头，屏面为 210×10毫米，呈窄带形，工作时，一次显现一行字符。

光纤管转印印刷机的驱动系统十分复杂。为防止失真，保证光点在全屏面能获得良好的聚焦，需要复杂的偏转补偿电路。为了保证高速工作，必须使用静电偏转系统，其

造价很高。此外，必须用特殊加工的氧化锌纸作记录纸，其维持价格亦高。因此，目前只在少数中、大型计算机系统中配用，以充分发挥其高分辨率的字形输出及每秒数千汉字的高速印字能力。国外较新生产的光纤管印刷机已可以使用普通纸印字，这是一个重大改进。

总之，汉字印刷机随着汉字信息处理技术的飞速发展，已成为计算机外部设备中一类重要设备。由于它品种多，性能差别大，价格相差悬殊，故在选用这些设备时，必须根据实际需要和可能，加以正确选择，以保证整个计算机系统在性能、价格上的合理性，使之在应用中发挥其应有的效能。

7.5 汉字语音输出技术和设备

汉语语音输出是汉字信息处理技术用的一种输出手段，它有重要的使用价值。

这里所说的汉语语音输出是指利用语音的数字信息，采用语音合成（speech synthesis）的方法由数字信息还原成模拟量而输出人耳能听到的语音。这种输出方法又称为“语音合成”。

汉语语音的输出可以采用以音节为单位的输出方法，因为汉语的特点是一个字一个音节，字是独立的发音单位，这比一般采用拼音文字的外国语言（如英语、法语、俄语等）要简单得多。汉语的音节共有一千二百多种，比较单纯。外国语言虽然音素不多（美国英语只有 42 个音素），但是音节种类繁多，而且多音节的词很难完全按单个音节的发音来组合。正由于汉语发音的特点，使汉语语音全部参数的存储量可以较小，而且发音的控制也较简单。

下面我们来介绍一个实用的汉语语音输出装置。

7.5.1 汉语语音合成装置的组成

图 7-35 是汉语语音合成装置框图。其中，CPU 是一个简单的八位微处理机，M是存储器，存放 CPU 的工作程序等。有一个通信接口 SIO。通过通信接口可以接收到系统发来的命令和所要求发音的汉字代码。

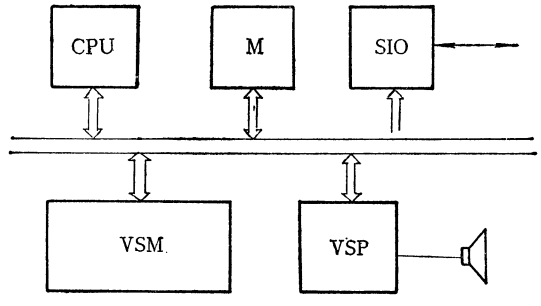


图7-35 汉语语音合成装置框图

其中，VSP（Voice Synthesis Processor）称为语音合成处理器；VSM（Voice Synthesis Memory）称为语音合成存储器。

VSP 是一个关键的部件。这里介绍的一种 TMS 5200 芯片，是美国德克萨斯仪器公司的单片语音合成器产品。

TMS 5200 的逻辑框图如图 7-36 所示。

对 CPU 来说，VSP 可以看作是一种低速的存储元件。实际上，VSP内部的处理还是比较复杂的，但它以较低的速度（工作频率较低）获取外界的语音参数。

VSP 在面向 CPU 这一侧一共有十二条信号线，它们是D_0～D_7八条数据总线以及读

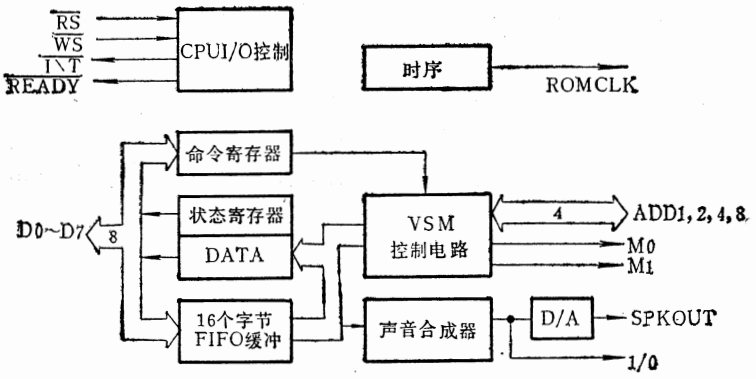


图7-36 TMS5200的逻辑框图

（$\overline{RS}$）、写 （$\overline{WS}$）、中断请求（$\overline{INT}$）和准备就绪（$\overline{READY}$）等。

$\overline{READY}$ 线是为了和 CPU 的速度相匹配而设置的。图 7-37 示出了 VSP 的读写周期。

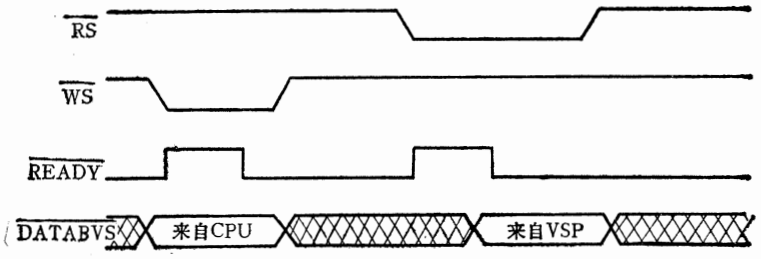


图7-37 VSP读写周期

当 VSP 接到 CPU 的读（$\overline{RS}$）信号或者写（$\overline{WS}$）信号时，$\overline{READY}$ 信号立即变成高电平，告诉 CPU 信号读、写操作正在进行，但未完成。当 VSP 接收了 CPU 信号（写的情况），或者已经准备好了数据可以被 CPU 所读出（读的情况）时，$\overline{READY}$ 信号才变成低电平。

TMS 5200 在发声的过程中要求提供发声的参数。有两种给 TMS 5200 提供参数的模式。一种模式称为外部发声的模式（SPEAK EXTERNAL），这种模式，是由CPU 通过 DATABUS 从语音合成存储器中取出参数并送给 VSP 的，其传送方式是 8 位并行传送，由中断信号及程序进行控制。另一种模式是 SPEAK 模式，VSP 直接从 VSM 获得参数，不需要程序控制〔只需 CPU 给 VSP 置地址（LOAD ADDRESS）命令和 SPEAK 命令就可以进行〕。为此，德克萨斯仪器公司生产一种 VSM 的专用 PROM 存储器芯片 TMS 6100，它是一种 128 K × 1 位的 PROM。一个 VSP可以接 16 片 TMS 6100（共有256K字节的存储容量，可供说话半小时），而不需要外加硬件电路，它通过

VSP 的 ADD 1、ADD 2、ADD 4、ADD 8/DATA、MO、MI 和 ROMCLK 等七条线与 VSM 相连接，以串行方式传输参数信息。这两种模式都可以在同一片 VSP 上使用，并可以对 TMS 6100 进行编程（写入）。在图 7-36 所示的结构中，只表示了外部参数发声模式的情形。这时，VSP 中的数据寄存器及VSM控制器不起作用。VSP面向 CPU 的寄存器，这时只有命令寄存器（COMMAND REGISTER）、状态寄存器（STATUS REGISTER）和一组先进先出缓冲寄存器（FIFO BUFFER）等。

FIFO BUFFER 共有 16 个 8 位的寄存器。CPU 送给 VSP 的发声参数都送到这十六个寄存器中。

状态寄存器只有 3 位（D 0＝TS，D 1＝BL，D 2＝BE）。其中，TS 称为“说话状态”（TALK STATUS），若这位是“1”，则表示正在说话（发声）；BL 称为缓冲寄存器低态（BUFFER LOW），表示缓冲寄存器已被使用一半以上（八个字节）所剩参数不多了，需要“补充”；BE 表示缓冲存储器空（满）。当接收到外部发声命令后，缓冲存储器没有数据，这时称为空，BL＝1，BE＝1；当说话进行中如果缓冲存储器已经接收满了，则 BE也将表示为高电平（即BE＝1，BL＝0），这时称为满。

$\overline{\text{INT}}$是中断请求信号，当下列情况之一出现时，VSP产生中断请求：

若 TS 由 1 变到 0，则说话状态结束。这时，可能有两种情况：其一是接收到一个“结束发声”的参数；另一种情况是 FIFO 缓冲寄存器已使用完了，又没有新的参数送入，则自动停止发声，TS 由 1 变到 0。

CPU 给 VSP 的命令形式是非常简单的。如表 7-13 所列。

表7-13　CPU送给VSP的命令

DATA BUS的命令码	命　令　内　容
×0 0 0××××	空命令
×0 0 1××××	读一个字节
×0 1 0××××	空命令
×0 1 1××××	读并转一地址
×1 0 0AAAA	置地址
×1 0 1××××	SPEAK
×1 1 0××××	SPEAK EXTERNAL
×1 1 1××××	总清

在外部参数发声的情况下，只要用两种命令，即 SPEAK EXTERNAL 和总清（RESET）。

当发出外部参数发声命令以后，将不能再发总清命令。因为这时 VSP 分不清是总清命令还是发声参数，所以一旦发出外部参数发声命令，就一直发声，直到检测到一个停止发声参数或者 FIFO 空了以后为止（此时 TS 由 1 变 0）才算结束。此外，只要一加电，就一定能总清。

7.5.2　语音的合成

图 7-38 是 TMS 5200 中语音合成部分的细框图。

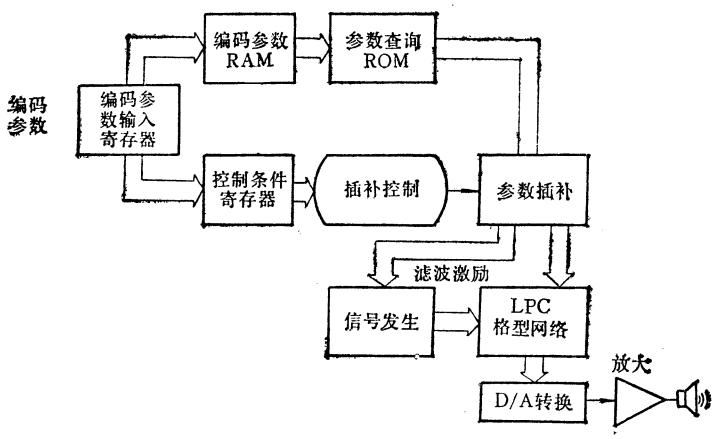


图7--38　TMS5200语音合成部分的框图

一、线性预测分析法

送给 TMS 5200 的参数称为“基音激励线性预测码”，这是一种专门的表示语音参数的编码。线性预测码是目前应用最广泛的语音参数表示。

线性预测分析法是最有效的语音分析技术之一。这种方法其所以重要是因为它能够极为精确地估计语音参数，而且它的相应计算速度比较快。在 TMS 5200 中，采用了软件固化、插补计算的方法，使得处理速度快，语音质量也好。

线性预测分析所包含的基本概念是，一个语音抽样能够用过去若干个语音抽样的线性组合来逼近。通过使实际语音抽样和线性预测抽样之间的差值的平方和（在一个有限的间隔上）达到最小值，能够决定唯一的一组预测系数。

可以用几种不同的方法由线性预测分析参数来合成语音。最简单的方法是采用一个系统，它和分析系统（产生语音参数的声码器）有着相同表示式的合成器。图7-39是这种语音合成器的方块图。

合成器所需的时变控制参数为基音周期、浊音/清音开关、增益以及预测系数等。浊音语音激励源的冲激发生器，在每一个基音周期的起始处产生一个单位幅度的脉冲。用于清音语音激励的是白噪声器，它产生不相关的、均匀分布的随机抽样。其标准方差为 1，平均值为 0。由清音、浊音开关对这两种激励源进行选择。增益G（这里用能量ENERGY）决定整个激励信号的幅度。

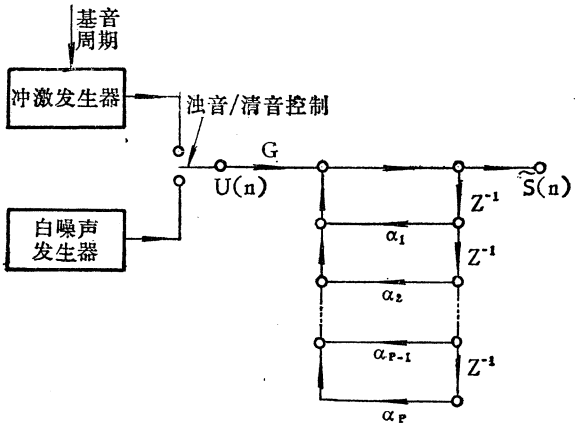


图7-39　线性预测合成器原理图

TMS 5200 的线性预测码共有 12 个参数（见表 7-14），它们是能量（ENERGY）、基音（PITCH）、十个反射系数和一个重复标志。

表7-14　TMS5200线性预测编码表

参　数　名　称	级　数	编　码　位　数
能量 (ENERGY)	15	4
基音 (PITCH)	64	6
K 1	32	5
K 2	32	5
K 3	16	4
K 4	16	4
K 5	16	4
K 6	16	4
K 7	16	4
K 8	8	3
K 9	8	3
K10	8	3
总计12	247	49 + REPEAT = 50位

语音参数就是按这种编码方案存储在 VSM 中的。

特别是，这里，PITCH= 0，表示清音输出，应该由白噪声发生器来激励，PITCH ≠ 0，表示浊音输出，并由 PITCH 参数给出基音周期激励声道。此外，这里用 ENERGY= 0 来表示不发声（语音之间的间隔）等。另外还加了一位重复位（REPEAT)。因为语音过程很慢，而且不是每一帧声音都是有很大变化，实际上很可能没有什么改变，故用了 REPEAT，以减少参数的存储量。当能量参数为 1111 时，表示结束发声标志。

在发清音时，声道参数一般都较少。数字滤波器所需的反射系数只须用 K1～K4 四个参数。整个参数的结构如图 7-40 所示。

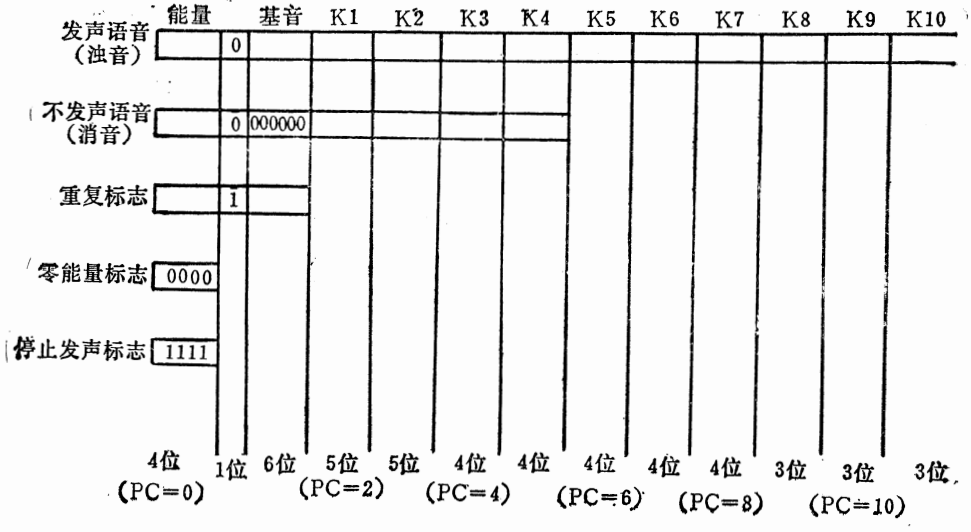


图7-40　线性预测参数结构示意图

二、语音合成

语音合成过程是这样的，当VSP接到发声命令后，BL、BE由0变到1，VSP开始接收CPU发送来的参数，逐个存入FIFO缓冲寄存器。当BL由1变到0（即CPU给VSP已经发送了九个字节的数据）时，发声就开始，TS由0变到1。发声开始以后，VSP就从FIFO中取数据。同时继续接收CPU发来的数据。在不断的发声过程中，也不断地接收CPU送来的参数。如果CPU送数足够快，则产生BE中断，表示FIFO缓冲寄存器已满；如果CPU发送参数不快，FIFO已用一半以上，则BL发出中断。如此，发声过程一直继续着，直到接收一个“停止发声”的参数或者将FIFO缓冲寄存器取空（BL＝1 BE＝1）为止。发声停止后，TS＝1。通知CPU已停止发声。

从FIFO缓冲器取出数据后，首先要检测这组参数是什么性质的。一共有五种可能，它们是清音、浊音、重复、零能量（不发声）和停止发声。这五种情况决定合成过程的不同控制方式。

从FIFO缓冲器取出的参数数据，并不直接被使用。除了首先判别五种性质以外，这些参数只是实际参数的一个“索引”。在线性预测参数的分析中证明，线性预测码为了保证合成的稳定性和语音输出的质量，要求参数的精度要高。尽管在VSM中取来的参数只有3～6位，但在内部一般都用10位参数。这些参数存放在VSP芯片的一个参数查询ROM（LOOK-UP ROM）中。TMS 5200的线性预测码每次取8～64（用3～6位二进制码表示）级可能的参数。而LOOK-UP ROM读出的数据精度要高（级数相等）。由ROM读出的数据再交给插补计算器。

三、TMS 5200的时序

TMS 5200时序比较简单，它采用一个独立的时钟(CLOCK)。可以外接晶体振荡器，也可以接一个电阻来组成多谐振荡器。其主频可以是640～800千赫。在TMS 5200设计指标中，声音的抽样频率是10千赫。25次抽样周期称为一个插补周期，8个插补周期（相当于200个抽样周期）称为一帧。

当抽样频率是10千赫时，帧周期为20毫秒，即每秒钟要50帧，每帧的参数是50位。所以，语音合成器TMS 5200每秒需要的参数是2500位。如果主振频率为640千赫，抽样频率就是

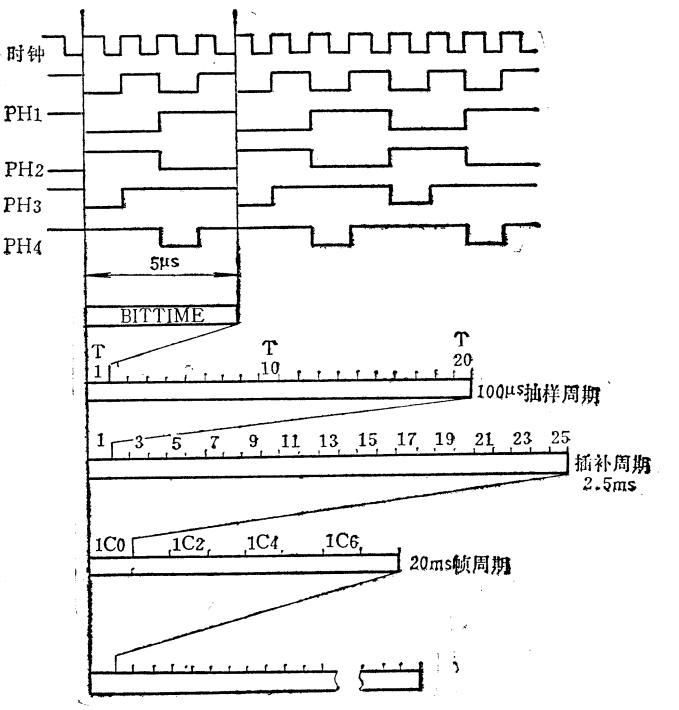


图7-41 TMS5200时间关系图

8 千赫，每秒的帧数就是 40 帧。这样数据需求率是 2000 位/秒。

实际上，由于使用了缩短的清音参数、REPEAT 和 0 能量参数，其实际 数据率还要低一些，估计在 1200～1500 位/秒之间。

TMS 5200 的时间关系如图 7-41 所示。

图中参数是以主振频率为 800 千赫为例来标明的。

四、TMS5200的 D/A转换及输出电路

TMS 5200，内部有一个 8 位的数模转换器。分辨频率为最低位的 1 / 2，每 100 微秒(抽样时间)从格型网络的输出端输出一个 14 位的数。取其高十位，进行D/A转换(高十位数据中，又只用符号位和其中的低七位去转换，余下两位只用来做逻辑判断，以决定是全开还是全关)。详见表 7-15。

表7-15 TMS5200的D/A转换及电流输出表

NO	Y 输 出				D/A输入	模拟输出 (微秒)
	Y13	Y12	Y11	Y10—Y4		
	0	1	1	×	11111111	0
＞+127	0	1	0	×	11111111	0
	0	0	1	×	11111111	0
127	0	0	0	1111111	11111111	0
126	0	0	0	1111110	11111110	5.86
			—			
			—			
			—			
+1	0	0	0	0000001	1(1)0(1)001	738
0	0	0	0	0000000	10000000	744
*-1	1	1	1	1111111	01111111	750
-2	1	1	1	1111110	01111110	755.8
			—			
			—			
			—			
-128	1	1	1	0000000	00000000	1500
＜-128	1	1	0	×	00000000	1500
	1	0	1	×	00000000	1500
	1	0	0	×	00000000	1500

TMS 5200D/A转换器的输出是一个电流源。输出电流从 0～1.5 毫安，分 辨 率是 5.9 微安。它可以用来驱动音频放大器，也可以直接接到扬声器上。

当Y寄存器所存的数大于 +127 时， 输出为 0 伏； 当Y所 存 的 数 小于 −128 时，线路的输出为 3 伏 (电流是 1.5 毫安)。当没有发声时， Y的 内容 是− 1，电流是 750 毫安。

以上简略地介绍了语言声音合成器 TMS 5200 的工作原理和主要性能。

采用 TMS 5200 作语音合成电路，再配以一般的 ROM 存储器，就可以让计算机作汉语语音输出，ROM中的语音参数需要预先“录”好。

平均每个汉字约用 600～1000 位的参数。这样，一千二百多个汉语语音参数总计需要的存储容量约为 120 千字节。

第八章 汉字显示终端

8.1 汉字显示器

目前计算机系统中所用显示器的绝大多数是利用阴极射线管(Cathode Ray Tube)显示，它的常用符号为 CRT，是一种应用十分普及的电子器件。家用电视机的显象管就是 CRT 的一种。CRT 是一种电-光转换器件，可以把电信号转换为光信号。作为计算机的外部设备，广泛使用各种 CRT 的字符显示器和图形显示器。近年来虽有液晶显示，等离子显示，发光二极管显示等新技术的出现，但尚未成熟，价格也高，不及 CRT 使用普遍。

用 CRT 做成的显示器，又称监视器（MONITOR)。它有许多优点：

（1）由于全部采用电子元件，没有机械部件，无磨损，从而提高了可靠性和耐用性，维护也比较容易。

（2）没有噪声。

（3）成本低廉。

（4）响应速度快。

正因为有这些优点，现代计算机终端几乎都采用这种显示器了。有人称 CRT 显示器是“无噪声的打印机”，“玻璃上的打印机”等。尤其是七十年代以来，半导体工艺、大规模集成电路和微型机技术，以及 CRT 显示技术等各方面都有了突飞猛进的发展，使得 CRT 显示器提高了性能，缩小了体积。特别是微型机技术的发展，使得 CRT 显示终端的功能有了更多的扩充。如彩色显示、图象显示、光笔的使用，以及各种“智能”的开发等。

能显示汉字字形的 CRT 显示器是汉字信息处理系统中不可缺少的设备。汉字信息的输入，文件的修改、校对以及各种汉字信息的输出，都离不开汉字显示器。

从技术上来说，要让 CRT 显示汉字并不十分困难，因为都是一些成熟的技术。然而它和普通的字符显示器相比，有某些特殊的要求。下面介绍 CRT 汉字显示器的工作和结构原理。

8.1.1 CRT 显示器的扫描方式

用于计算机系统中显示字符和图象的显象管称为“字符显象管”，它和普通电视机用的显象管很相象，都属于磁场偏转式的。但有几条非本质的区别这就是：

（1）普通电视机用显象管的光点（spot）比较粗，而字符显示要求光点细一些。这倒不完全是普通电视机的光点细不了，而是人们有意让它“散焦”，以便电视图象收看的效果更好一些。一般用“分辨率”（resolution）这个指标来衡量光点的粗细。一般电视显象管的分辨率比字符显象管的要低一些。字符显象管要做到 600 线是不困难的。分辨率是 600 线，是指屏幕中间部分能清楚地分辨出每行 600 个点，即在屏幕的中间一行显示 300 个亮点，300 个暗点，亮暗相间如能分辨清楚，就说这只管子的分辨率为 600

线。有些场合要求高分辨率的字符显象管。现在已经有分辨率为1000线，2000线，甚至4000线的高分辨率的CRT显象管。此外，在屏幕的边缘和四个角上分辨率要低一些。

（2）电视显象管的余辉时间比较短。而字符显象管往往采用中、长余辉的为好。余辉时间长一些，再生周期也可以长一些，这对降低视频信号的频率，减少屏幕显示的闪烁是有好处的。尤其是对高分辨率的显示器更为重要。

（3）字符显象管的光点一般采用绿色，而家用电视则是白色（黑/白）的。绿色光点对人眼的刺激较小，适应较长时间工作的耐疲劳性，更受使用者的欢迎。

现在各种彩色显象管应用也较普遍。特别用彩色显象管来显示图形有更好的效果。为了集中讨论“汉字的显示”这一题目，本节的内容仍以讨论单色显示器为主。

一、CRT电子束光栅扫描原理

CRT的电子枪只射出一束电子，怎样才能使整个荧光屏显示出图象呢？这就要靠“扫描”来实现。字符显示器最常见的扫描方法称为“光栅扫描法”。

如果没有扫描，那么荧光屏上永远只有一个亮点。如果在横向偏转线圈中通过一个锯齿形变化的电流（如图8-1），那么电子束就会在偏转磁场的影响下作周期性的往复运动，结果便在荧光屏上扫出一条水平的直线来。

如果在偏转线圈中的电流波形不是锯齿形的，而是其他变化形状，电子束也能扫出一条直线。锯齿形的优点是电子束从左到右扫描的时间是匀速前进的。称为“扫描线性”好。从左到右的扫描称为扫描的“正向行程”，简称“正程”。从右到左的扫描，就是锯齿电流下降的那一部分，称为“反向回程”，或者称为“逆程”。逆程扫描速度较快，线性不好。经常不用逆程扫描而只让正程起作用。在回程的时间内只要在控制栅极上加上一个负电压，就能抑止电子束的发射。这种做法称为“回程消隐”。

如果在正程扫描的同时，在控制栅极上加上一个同步的脉冲信号，那么在荧光屏上就可以看到有亮点和暗点相间的一条直线。控制栅极电压高时，对应屏幕上的亮点；电压低时，抑制了电子束对荧光屏的轰击，就成了暗点了。加到栅极上的信号称为“视频信号（video signal）”。

有了横向的扫描就能在荧光屏上获得一条直线的图象。如果在纵向也让它扫描，那就能在屏幕上获得完整的图象。

横向扫描称为“行扫描”，纵向扫描称为“帧扫描”。相应的扫描线圈分别称为“行扫描线圈”和“帧扫描线圈”。

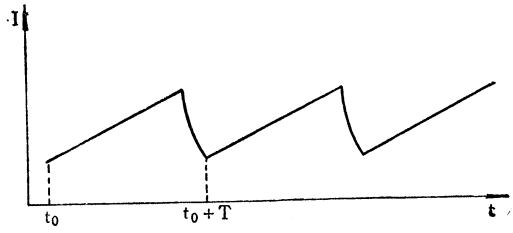


图8-1　显示器的扫描电流

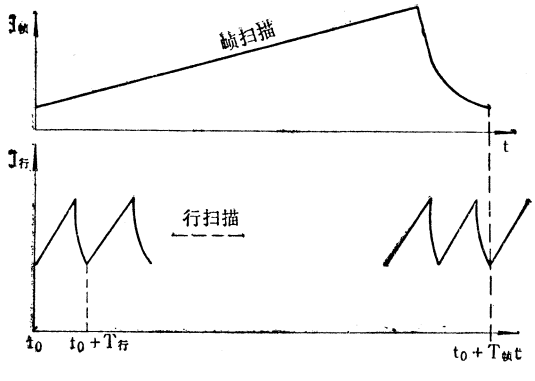


图8-2　行扫描与帧扫描的时间关系

帧扫描也是采用锯齿波电流，不过帧扫描的周期很长，是行扫描周期的几百倍，如图8-2所示。

帧扫描电流也有正程和逆程之分。帧扫描和行扫描的关系除了它们的频率有固定的比例外，还要求它们是“同步的”。

有了帧扫描以后，电子束就能从左到右，从上到下一行一行地扫描了，扫过整个屏幕。这种扫描方法称为“光栅扫描（raster scanning）法”。如图 8-3 所示。

在扫描电路的控制下，电子束从上到下，从左到右扫过整行屏幕，如果在控制栅极上再加上视频信号，就能获得相应的图象。要在屏幕上看到固定的画面必须不断地扫描，扫过一帧，再重复扫一帧，因为光点扫过荧光屏亮一下（余辉）以后就熄灭。实际上扫到下面时上面的点就熄灭了。所以必须连续不断地扫描。这种不断地重复扫描称为“刷新”（refresh）过程。

如果每次刷新都能在同一地点亮或暗，那么就能在屏幕上得到一个稳定的图象。要保证每次刷新都在同一地点呈现亮或暗，这要靠视频信号与扫描信号的同步来得到。

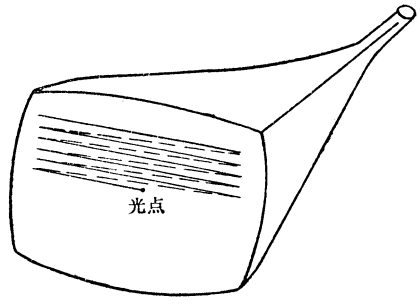


图8-3　电子束光栅扫描示意图

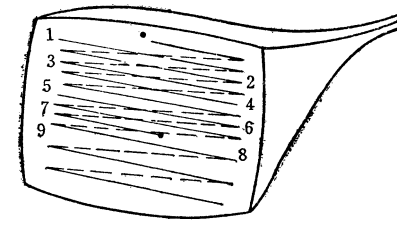


图8-4　隔行扫描示意图

二、逐行扫描与隔行扫描

人眼的特点之一是：如果一秒钟内亮了又灭，灭了又亮三十次以上，就觉察不到“闪烁”了。所以，对于无间隔的行扫描，每秒要大于三十帧，这种方法称为“逐行扫描”。但是如果采用隔行扫描的办法，那么，每秒钟即使扫描二十五帧，仍有相当于五十帧的效果。家用电视均采用“隔行光栅扫描”体制。

所谓隔行扫描就是：第一“场”先扫第一行，第三行，第五行……，称为“奇数场”；下一场是偶数场，偶数场只扫第二行，第四行，第六行……如图 8-4 所示。

我国的电视扫描制式规定为：

（1）每秒钟共扫描 25 帧，50 场（25 场奇数场，25 场偶数场）。帧扫描的周期是 20毫秒。

（2）每帧 625 行，每场是 312.5 行。

（3）按照这个规定计算，每秒钟扫描 15625 行。每行扫描时间是：

$$1\text{秒}/15625=64\text{微秒}$$

（4）规定每行 64 微秒中，52 微秒是“行正程”；12 微秒是“行逆程”时间。

（5）规定帧扫描过程中逆程时间是 25 个行扫描时间，即 1.6 毫秒，这样，帧正程时间是 18.4 毫秒。

国外的电视扫描体制也有采用每秒 30 帧，60 场，每帧 525 行，每场 262.5 行的。

一秒钟共扫描 15750 行。

隔行扫描每一帧的行数是奇数，也就是说，每一场一定有一个“半行”，这样才能把奇数场和偶数场的图象在荧光屏上叉开。

字符、图象显示的扫描体制既可以是逐行扫描的，也可以是隔行扫描的。

字符、图象显示的扫描制式可以灵活的设计。这与电视接收机有些区别。

普通电视机的“行同步”、“帧同步”信号必须和电视台的信号一致，是一种已规定好的制式，实际上它们是从接收的电视信号中“分离”出来的。但是，字符显示器就不同了，它自己有一个独立的扫描控制电路。

此外，字符显示和普通电视机还有一个区别。电视机的视频信号是一个连续变化的电压，只有这样才能在荧光屏上看到不同层次的黑白影象（称为辉度）的连续变化。但字符显示则采用数字脉冲信号，它在屏幕上显示一个“点阵”，每个点的位置在屏幕上是固定的，每个点不是亮就是暗，只有两种辉度。

三、逐行扫描的控制电路

字符、图形显示器扫描控制电路的核心是一个脉冲分频电路。图 8-5 是一个逐行扫描控制电路的框图。

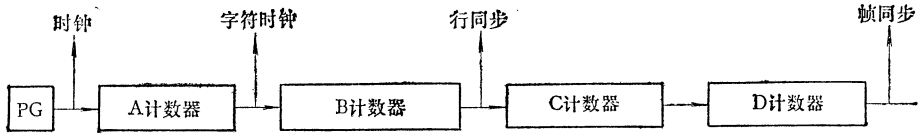


图8-5　逐行扫描控制电路

其中A、B、C、D是四个计数器。PG 是一个时钟（clock）脉冲源。时钟脉冲决定视频信号的脉冲频率，称为“点时钟”。

A计数器称为“点计数器”，它表示每个字符横向要有几个点。如果显示一个西文字符采用的是 7 × 9 的点阵，那么每一个字符宽度可用 8 个点或者 9 个点，因为还要留出一、二个点作为字符与字符之间的间隔。如果采用 8 个点，那么A计数器是一个从 0 ～7 的计数器。也可以说是一个三位计数器，或者说是一个 8 分频的计数器。A计数器的输出称为“字符时钟”（character clock）。

如果点时钟的频率是 12.5 兆赫，那么字符时钟的频率就是 1.5625 兆赫。点时钟的重复周期是 80 毫微秒，字符时钟的周期就是 640 毫微秒。

B计数器称为“字符计数器”，用来记录每扫描一行有几个字符。需要特别说明的是，这个字符个数还应包含逆程时间的字符脉冲个数。例如每行显示 80 个字符，回程时间规定是 20 个字符（20个字符时钟的周期），那么B计数器应是从 0 ～99的一个计数器，或者说是一个 100 分频的计数器。按这样的安排，这个例子中行扫描周期是 64 微秒。正程是0.64×80＝51.2(微秒)，逆程是0.64×20＝12.8(微秒)。

B计数器应输出行同步脉冲。这个例子中行同步脉冲的周期是 64 微秒。

C计数器称为“行计数器”，它的计数个数由每一个字符行应有几行来决定。例如要显示 7 × 9 点阵的西文字符，它的每一字符行应有 16 个扫描行，其中 9 线显示字符点阵，还有 7 线是行和行之间的间隔。那么C计数器应该是一个四位计数器，因为计数

状态是0～15，故也可以称为16分频的计数器。C的输出送到D计数器。

D计数器称为“字符行计数器”。它的内容由一帧图象应包含几个字符行来决定。同样，它应包括帧逆程所需要的行数在内。如果设计在屏幕上有24行西文字符，帧逆程的时间规定为2个字符行的时间，那么D计数器应该是一个26分频的计数器。D计数器的输出是帧同步信号。在这个例子中，字符行的时间是16×64＝1024微秒＝1.024毫秒。帧同步周期26×1.024＝26.624毫秒，即每帧扫描的时间就是26.624毫秒。一秒钟扫描37.5帧。帧扫描的正程时间是24×1.024＝24.576毫秒，帧逆程的时间是2×1.024＝2.048毫秒。

正如前面说过的那样，字符显示器中A、B、C、D计数器的计数值和时间的分配情况是可以根据需要设计的。在上述例子中也给出了设计计算的一些方法。当然，在设计时必须考虑到点脉冲的频率（即视频信号的频率）、行同步信号的频率、帧同步信号的频率、字符点阵的多少、逆程时间的安排等多种因素，这样才能设计得合理。

汉字显示的设计方法实际上和字符显示没有原则上的差别。我们仍用这个电路作为例子。假定汉字的点阵是15×16，每帧显示16行，每行显示40个汉字。主脉冲仍是12.5兆赫，那么，这时只要把A计数器改为0～15，B计数器改为0～49，C计数器改为0～23（每行汉字包含24扫描行，其中16行显示汉字字符，8行是行间距），D计数器改为0～17。

由此可以计算出以下一些数据：

字符时钟的周期为80（毫微秒）×16＝1.28微秒；

行同步周期为50×1.28＝64微秒；

每一帧的行数为18×24＝432行；

每帧扫描时间为64×432＝27648微秒＝27.648毫秒；

帧逆程时间为48×64＝3072微秒＝3.072毫秒；

行逆程时间为10×1.28＝12.8微秒；

每秒扫描帧数为1000÷27.648≈36.5帧。

从这一组参数来看，汉字显示和西文显示在扫描控制电路上没有太大的区别。

从这个例子中，可以计算出整个荧光屏上点阵的点数。每一行有640个点，每一帧显示384行。这里所指的640点和384行都是指实际显示的点阵，不算逆程时间的净显示的点阵。这一点阵也称为“屏幕点阵”，或者说“全点阵”。在这种扫描电路的控制下，可以让显示器显示一个384×640点阵的图形。

四、隔行扫描的控制电路

图8-6是一个隔行扫描控制电路的框图，它和逐行扫描的控制电路基本上类似。所不同的是：（1）B计数器分为B和B′，B′是单独一个触发器，B计数器的输出是半行同步脉冲。（2）C计数器是原来计数值的一半（少一位）（3）另外要加一个E计数器。E计数器所计的数是“半行数”，E计数器也可以称为“场计数器”，它输出场同步信号（也称为帧同步信号）。E计数器计满后输出一个清除信号，用来清除C计数器和D计数器内容，同时把B′的状态传送给C′。C′是奇数场与偶数场的标记。C′和C合起来决定当前扫描的行数。C′＝1表示奇数行；C′＝0表示偶数行。

如前所述E计数器一定是一个奇数分频的计数器，那么奇数场和偶数场每行起始的

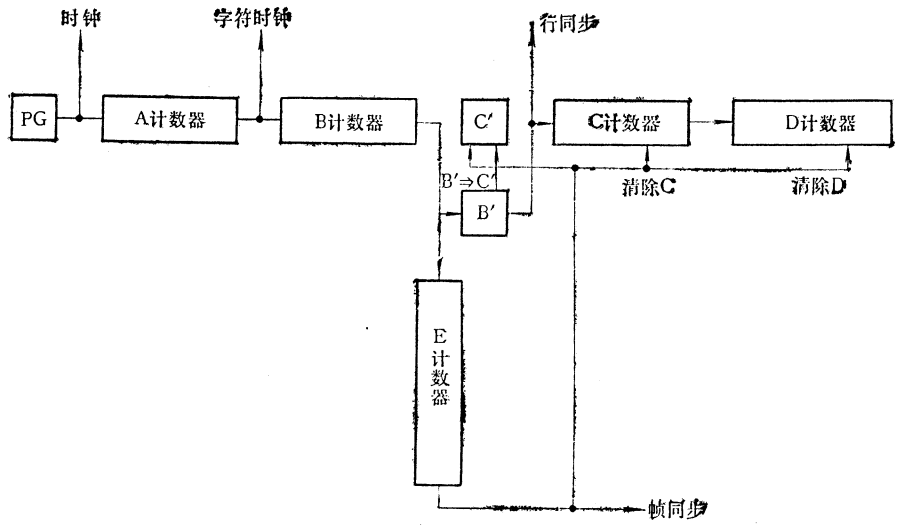


图8-6　隔行扫描控制电路

地方正好叉开。图 8-7 是隔行扫描控制电路的时序图。

以上简要地讨论了在字符显示或图形显示时经常采用的扫描方法——光栅扫描法的原理。通过数字分频电路来实现扫描控制是一种简单而可靠的方法。利用计数器的状态可以组合成各种控制信号，主要有行同步脉冲、帧同步脉冲、行逆程消隐信号、帧逆程消隐信号等。

8.1.2　显示器的刷新方法

有了扫描控制电路，CRT 的电子束就可以循环往复地扫过整个荧光屏。荧光屏上显示的是一个点阵。但是在显示过程中还必须控制点阵中每一个点是亮点还是暗点。只有这样才能显示一个点阵图形。控制点阵中每一个点的亮和暗的信号就是“视频信号”，点阵显示的视频信号是一种数字脉冲信号。

在扫描过程中，视频信号也要循环往复地加到 CRT 的栅极上去，也就是要作屏幕刷新。当然，为了保证亮暗清晰，还必须使扫描控制信号和视频信号同步。由于我们采用数字式分频扫描，同步问题比较容易解决。所以下面将重点讨论刷新问题。

刷新的方法基本上可以分为两种不同的类型。一种是“整屏幕点阵信息缓冲存储刷新”；另一种方法称为“字符信息控制刷新”。

一、整屏幕点阵信息缓冲存储刷新方法

整屏幕点阵信息的刷新，也称为“屏幕刷新”，“全点阵刷新”等，它主要依靠一个缓冲存储器，或者称“刷新存储器”。

屏幕上的每一个点与刷新存储器中每一位存储信息一一对应。存储信息为“1”代表屏幕上的一个亮点，存储信息为“0”代表屏幕上的一个暗点（也可以反过来规定）。顺便提一下，若是彩色显示，则要加大屏幕存储器的容量。彩色显示一般采用 RGB（红、绿、蓝）三色控制法。这时，屏幕上的一个点，要对应存储器的若干位信息。若一个点对应三位存储信息，就能获得八种颜色。若一个点对应四位存储信息，就可组合出十六

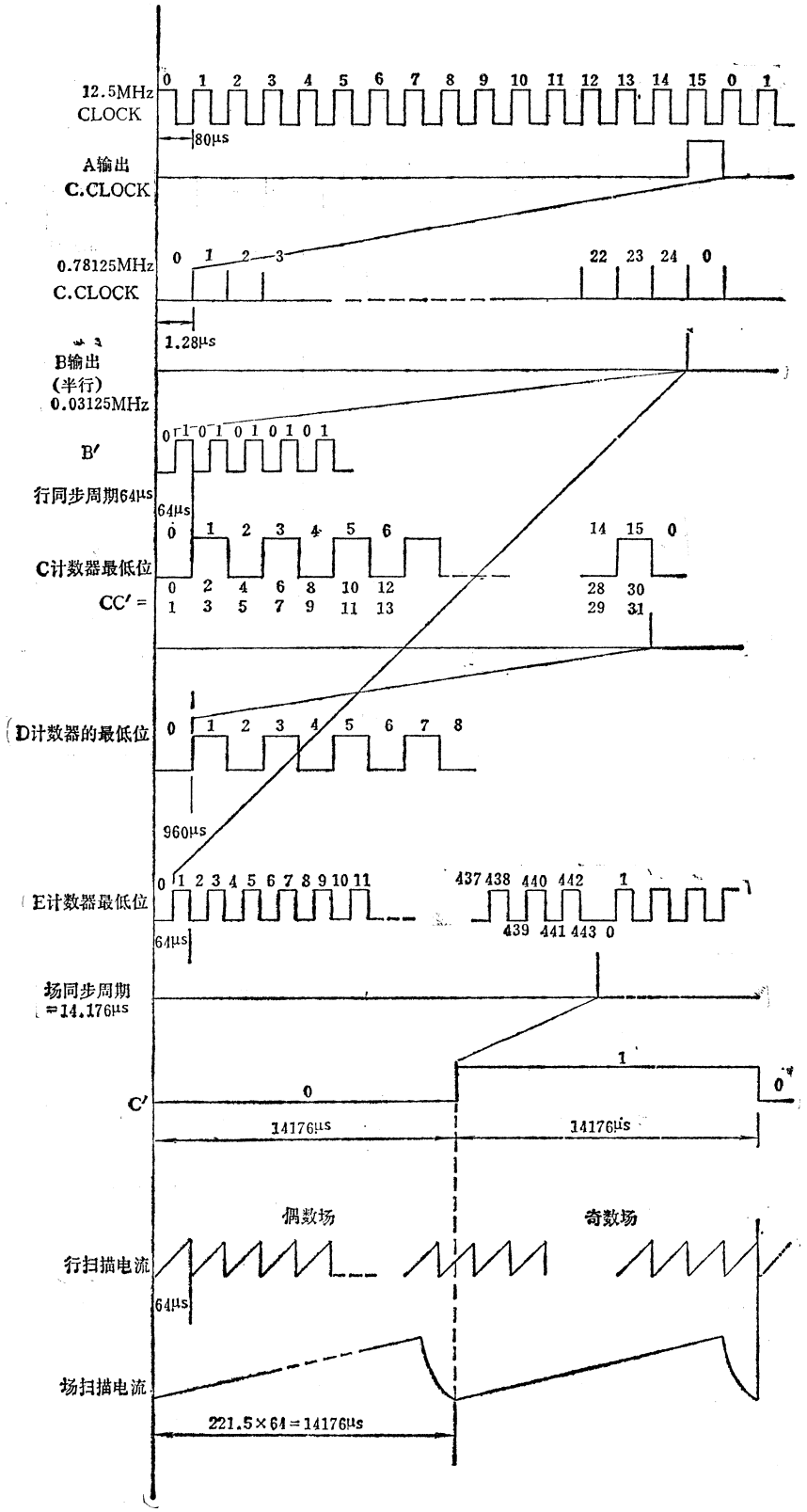


图8-7　隔行扫描的时序图

种彩色。

就单色显示刷新而言，屏幕上每一点的位置与刷新存储器的一位相对应，所以可以把屏幕点阵的位置与刷新存储器的地址对应起来。我们仍以640×384的屏幕点阵为例来说明。扫描线共有384条。每行扫描有640点。即每行扫描的内容应有80个字节的存储信息来提供刷新的视频信号。存储器的总容量是384×80字节。80个字节的地址称为“列地址”，384行的次序称为“行地址”。

这样，如果用一个地址计数器，就可以实现对整个屏幕的刷新。实际上，上节中介绍的扫描控制电路中的计数器正好是屏幕刷新存储器的地址计数器。图8-8画出了扫描控制与屏幕刷新的逻辑关系图。

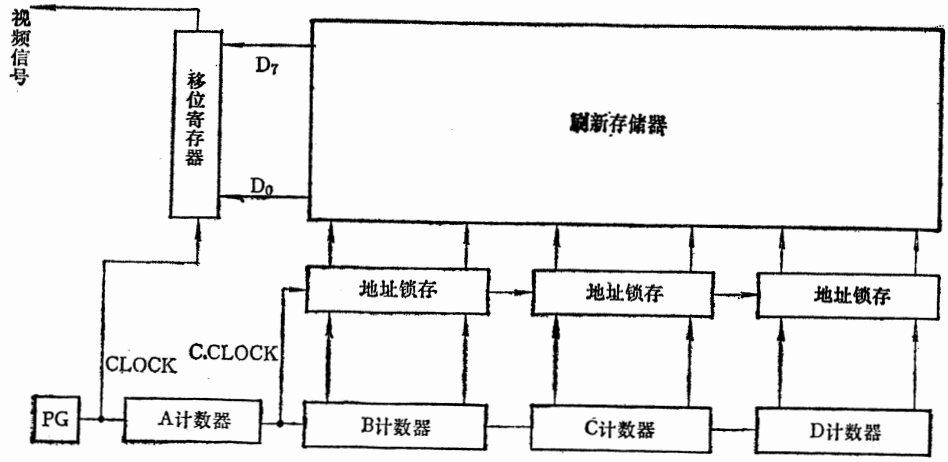


图8-8　扫描控制电路与屏幕刷新的工作原理

其中，计数器A是点计数器，每计满八个“点脉冲”，便发出一个“字符时钟”脉冲，用来控制B计数器，同时它也是“读”一个字节的控制信号。读脉冲也就是把刷新存储器的内容读出来“置入”移位寄存器。这时，移位寄存器的输出就是视频信号。B计数器在扫描控制电路中称为“字符计数器”，从屏幕存储器的角度来看，它正好就是列地址计数器。同样，C、D两者合起来作为刷新存储器中的行地址计数器。

图8-9所示为屏幕点阵信息与刷新存储器地址的关系。

在扫描过程中，计数器不断地计数，正好与刷新的时序相同，而且是同步的，两者的配合就可完成屏幕信息的正确显示。有几点需要说明如下：

（1）B、D两个计数器的计数值比实际要刷新的地址更大一些，因为在扫描回程时，不需要刷新，但仍要计数。这一点并不影响刷新工作的准确性，因为这时无论有什么视频信号，屏幕上是不显示的（这是因为有回程消隐控制）。

（2）回程消隐信号要用字符脉冲同步并延迟一个周期。这是因为：地址寄存器为0时读出信号尚未置入移位寄存器；当地址寄存器由0变为1时，才将0号地址的内容置入移位寄存器，再输出视频信号。所以，实际显示的点比计数器落后一个周期。为此，回程消隐信号也要让它落后一个周期。其处理方法是用一D型触发器来实现延迟。

（3）如果B、C两个计数器的分频数不是2的某次幂，那么存储器的地址就不连

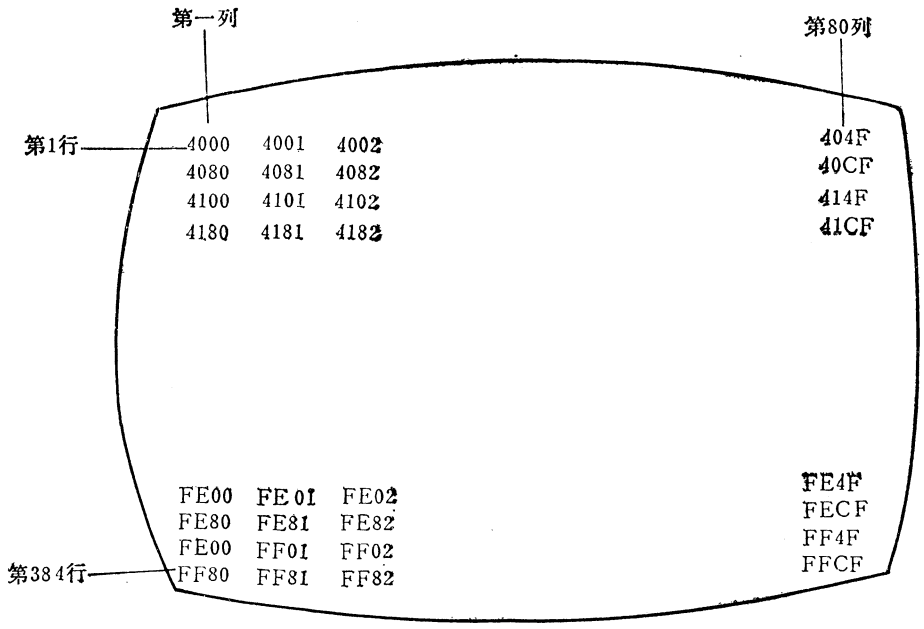


图8-9　屏幕点阵信息与刷新存储器地址的关系

续了。例如B计数器是从0计数到99，实际使用的是0～79，共80个地址需要刷新。而在刷新时实际上使用的地址要128个。其中有48个地址没有利用。

举例说，若刷新存储器的第一行的首地址是$(4000)_{16}$，那么第一行存储器刷新地址是$(4000)_{16}$～$(404F)_{16}$，而第二行的存储器刷新地址是$(4080)_{16}$～$(40CF)_{16}$，第三行的存储器刷新地址是$(4100)_{16}$～$(414F)_{16}$，…，其中4050～407F，40D0～40FF，4150～417F，…，没有利用。见图8-9所示。

（4）在具体电路设计时还可以充分利用计数器芯片的功能组成回程消隐信号和行（帧）同步信号，使行（帧）同步信号在回程的中间，这样可以获得更好的时间关系。

图8-10给出了一个字符计数电路的实例。

U_1，U_2，U_3是三片四位同步计数器电路74LS163。时钟脉冲12.5兆赫统一供给三个计数器。U_1是一个点时钟脉冲计数器，这里采用8分频。它的输出信号CARRY使它本身置成“8”，由“8”计数到“15”，再置为“8”，如此往复，成为8分频计数器。CARRY的输出加给U_2进位，U_2的进位通过CARRY给U_3进位。U_2和U_3组成B计数器（字符计数器）。用“与非”门组合成X、L、Y等三个信号。

X信号在U_2-U_3为0～79状态时为高电平，其余状态时为低电平，可用于回程消隐。

L只有在状态88时有一负脉冲输出作为U_2-U_3的$\overline{\text{Load}}$信号，置U_2-U_3为245（即二进制的11110101），所以B计数器（U_2-U_3）从0计数到88，然后“跳”到245，再计数到255并回到0，共100个状态。输出信号Y当B计数器的状态为84～87时有一个负脉冲。它可以用来作为行同步脉冲，也可用它作为下一级C计数器的计数输入脉冲。

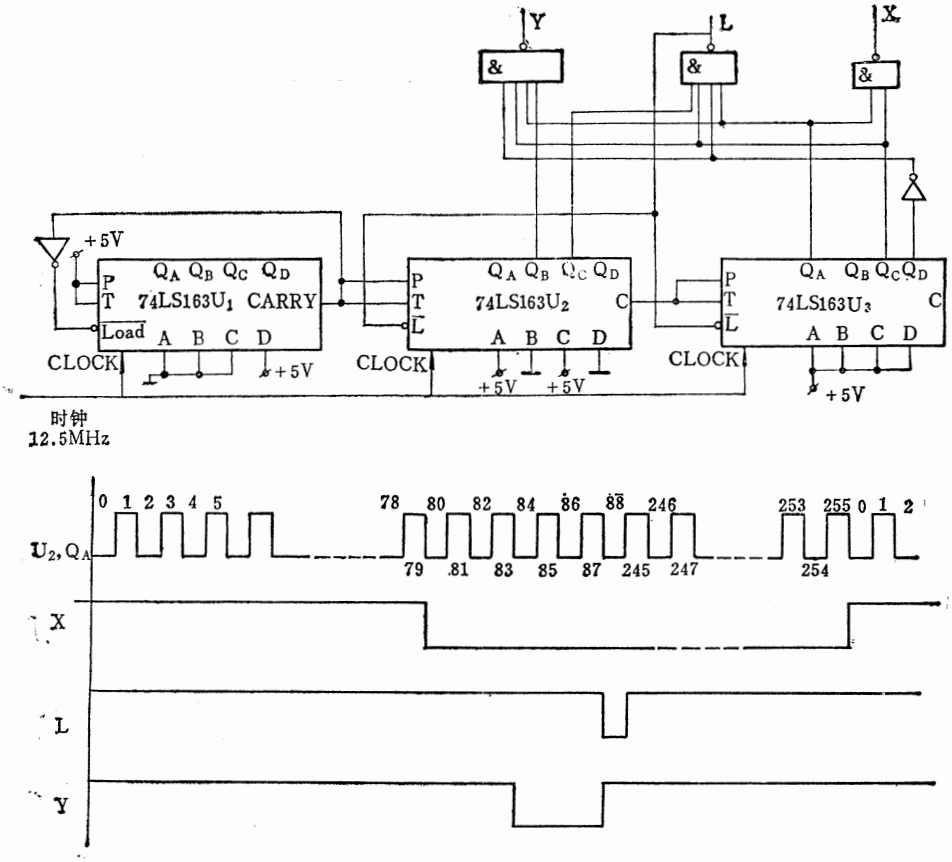


图8-10　字符计数器实例

以上所述是整屏幕点阵的刷新方法，无论对逐行扫描还是隔行扫描都是适用的。隔行扫描只要用C′和C合起来作为地址寄存器就可以实现。

二、字符信息控制刷新方法

整屏幕信息刷新可以显示整个屏幕点阵，这种方法不但可以显示字符、汉字，也可以显示图形。但是，需要容量较大的刷新存储器。此外，当显示字符时，如果屏幕上要有些改动（例如增一个字符；删去一个字符；字符行的滚动等），往往要使刷新存储器的内容“移位”，这种移位传输的信息量有时是很大的。而在多数情况下，我们只要求显示字符，特别是一些西文字符显示器，它能显示的字符一共只有几十种，若把每个字符的点阵都在刷新存储器中存起来，就有相当多的重复。为此，可以采用字符信息控制刷新的办法。图 8-11 是字符信息控制刷新的逻辑电路框图。

计数器B和D组成字符列和字符行地址。存储器M称为字符信息存储器，或简称信息存储器（message memory），它的内容是字符信息（即字符的内部码或地址码），每个字符一个字节（实际上只需要七位）。这里的G称为字形发生器。对于西文字符来说，这种字形发生器十分简单，只有几十种 5 × 7 或 7 × 9 的点阵图形，总存储容量不大于 2 K 字节（只需用一片 ROM 组件）。同时，也要把C计数器的内容送给G，指示当前显示字符行中的那一“线”。所以C计数器也可称为“线计数器”，C中的内容称作“线地址”。

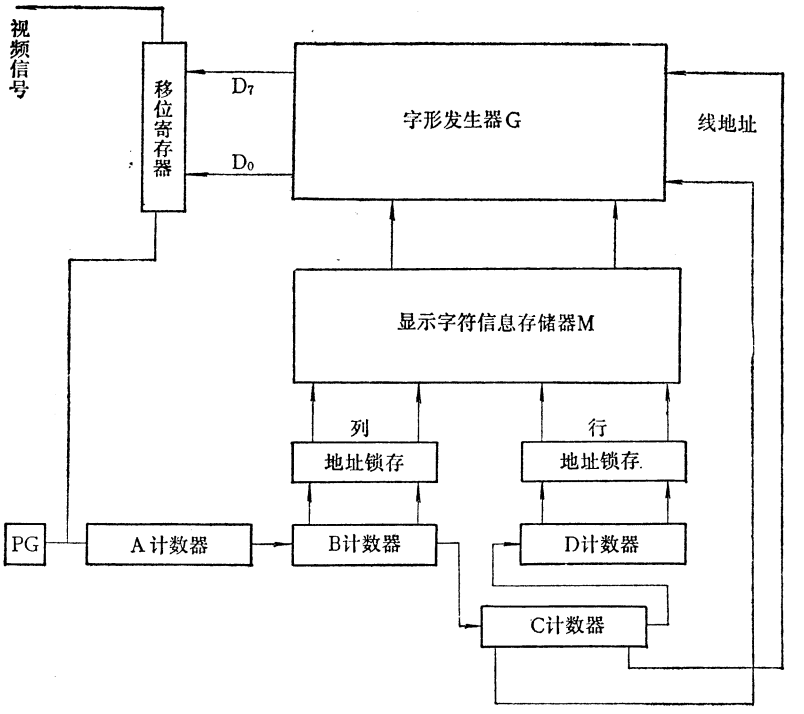


图8-11　字符信息控制和刷新逻辑电路原理图

除了存储器的结构不同外，其他如回程消隐信号，同步电路等都和全屏幕刷新的情况一样。

字符信息控制刷新有两个优点。第一，节省存储器。字符信息存储器中存储字符的内部码或地址码，每个字符一个字节。字符发生器也十分简单；此外还易于修改屏幕信息。

在汉字显示的例子中，也可以采用字符信息控制刷新的方法。

若采用一个兼顾显示 16×16 汉字点阵和 8 ×16 西文字符点阵的显示器，则扫描刷新控制电路的框图如图 8.12 所示。

这里，扫描制式为每行 80 个字符，或 40 个汉字。可以在一行中同时显示汉字或西文字符。显示的每一行汉字（字符）共 24 线，其中有 8 线是行间距。每屏幕共显示 16 行汉字。显示器显示点阵为 640×384。

采用逐行扫描的方法（隔行扫描也类似）。A计数器为 8 分频，B计数器仍是 100分频。C计数器是 24 分频，D计数器因考虑帧回扫时间，故为 18 分频。

为了能兼有显示汉字和西文字符的功能，M存储器的容量要略大一些。也就是说，M存储器中存储的信息，每一个字符（汉字）只用一个字节表示就不够了，故采用每一个西文字符要用两个字节，每一个汉字信息用四个字节表示，具体地说，M存储器存储信息的模式如下：

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
西文字符	0	0	0	0	0	0	0	0	0	×	×	×	×	×	×	×
汉字左半边	1	0	×	×	×	×	×	×	×	×	×	×	×	×	×	×
汉字右半边	1	1	×	×	×	×	×	×	×	×	×	×	×	×	×	×

整个M存储器对应 80 列、16 行的显示位置，实际上 80 列要用 128 个地址。所以，总的存储量是：

$$16\times128\times\ 2\ =4096字节$$

M存储器的结构如图 8-12 所示。

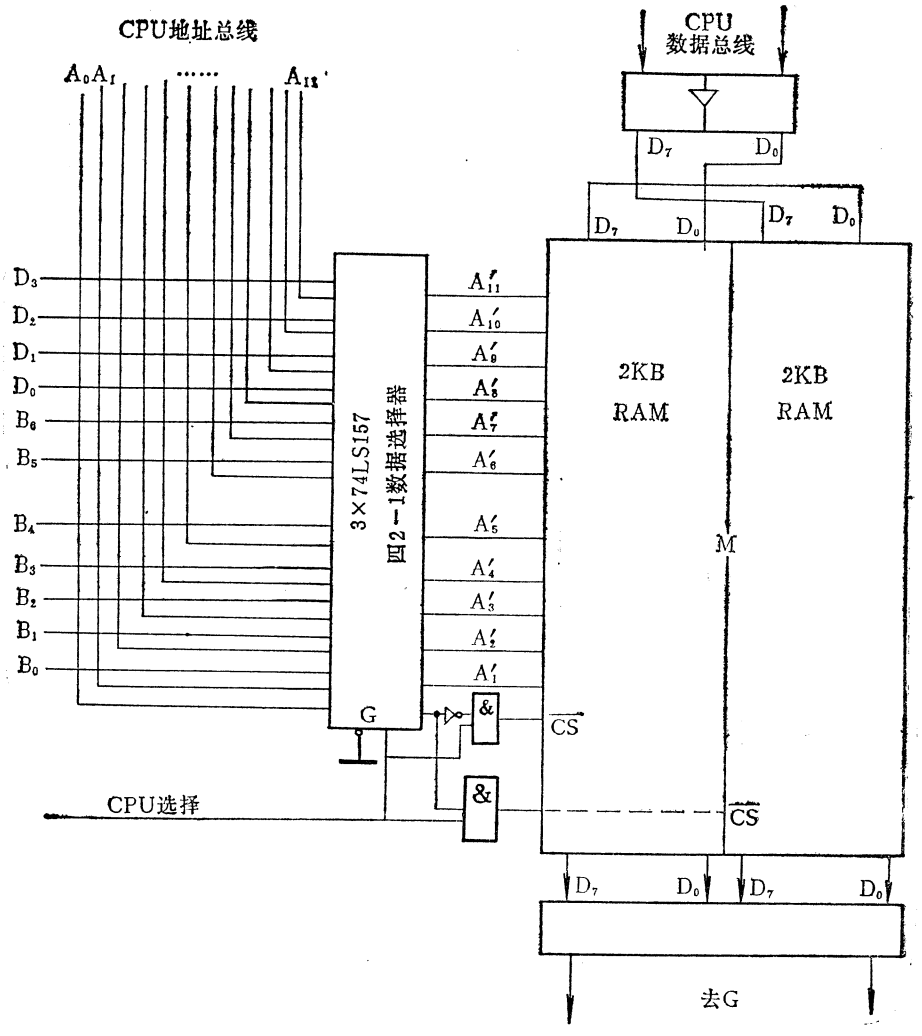


图8-12　兼有显示汉字和西文的显示器字符信息存储器的结构图

CPU对M只有写的功能，除了CPU向M写入的情况以外，都是显示器从M存储器中读出。每次写时，只写8位，而读时则要读16位。

在工作情况下，M存储器在每一个字符脉冲时读出一次。由它输出的数据作为字形发生器G的地址加到G上，并读出字形信息。每次读出8位，经移位输出即为视频信号。G的地址除了M读出的信息外还需加上C计数器指示的线地址。

这样组织的M存储器信息，汉字的编码最多可达14位（相当于16384个汉字），一般可不用那么多位。M的输出加到G，作为G的地址码。这里的G是汉字和西文字符的字形发生器。

G的规模若以512个汉字字形缓冲存储为例，它应有16K字节的容量，外加一个2K字节的西文字符发生器。这时M存储器实际上只需用11位就够了，汉字编码用9位。从以上叙述可知，汉字显示技术与西文字符显示以及图形显示很相似，只要稍加改动就可以显示汉字，而且也不难做到使西文、汉字、图形等都可在一个显示器上显示。

现在，有一些大规模集成电路，已经能在一个芯片上集中所有扫描和刷新控制电路了。例如MOTOROLA公司的MC6845，它的功能和我们所介绍的图8-11电路的功能相似，只是更完善、更灵巧。MC6845称为可编程的CRT控制器（programmable CRT controller），所谓可编程主要是指可以由程序来确定下述内容：是逐行扫描还是隔行扫描；整屏能显示多少字符行；每字符行内含几条扫描线；每个字符行中有几个字符；行回程的时间含几个字符脉冲；帧回程时间含多少字符行及多少扫描行，等等。由于这些参数都是通过程序给定的，所以使用很方便，也适合用于各种不同技术要求的显示器。

此外，MC6845的输出信号是M存储器的地址，相当于图8-11中D、B计数器的输出端，共有十四条地址线，编址空间可多达16K地址。此外它与B、D的输出有一点不同，即不论B的计数分频数是多少，输出地址线总是连续的。也就是说，M存储器中信息可以一行一行地连续存放，除十四条M存储器地址线以外，还有五条“线地址”输出线，它们相当于图8.11中C计数器的输出线。此外还有各种控制信号输出，详见生产厂家的技术资料。

用MC6845来控制汉字显示，无疑也是可以的。就象前面所介绍的方法（图8-12）那样，只要把M存储器的信息扩大成两个字节为一个单元就可以了。采用扩充的M存储器，把M存储器的输出端接到汉字字形发生器G（或者汉字字形发生器的缓冲存储器），则G的输出信号经过移位就是视频信号了。

Intel公司的8275，也是一种可编程的CRTC，它和MC6845一样，扫描控制电路中的各种参数都是可编程序的。但是8275只能控制逐行扫描，而不能控制隔行扫描。

在刷新方法上，Intel的8275与MC6845有较大的区别。8275在芯片内部有两个行缓冲存储器（缓冲寄存器），每个行寄存器最多可以容纳80个字节。这两个行缓冲器交替地工作，当一个行缓冲寄存器在进行显示刷新时，另一个便向M存储器取下一行要显示的信息。在当前这一行显示完成后，这两个缓冲寄存器的位置作了一次交换。如此交替循环地工作。8275在取显示信息时是通过DMA（直接存储器存取）方式向M存储器获取信息的。所以，8275还必须与DMA控制器连用。

8275的输出是七条线。它的输出信息不是地址而是要显示的字符信息，可以直接接

到字形发生器G上去。8275同样有四条线地址输出线（每个字符行最多可有十六条扫描线）。

8275因要和DMA连用，故要付出一定的代价，但输出端可直接产生视频信号。此外，它有较多的控制信号输出，例如：加亮、反显示、闪烁、加低线、光标等控制信息；还有两个用户可自己定义的控制信号，以及光笔输入的控制信号。因此，8275的功能比较齐全。

8275也可用作对汉字显示的控制，这主要是要设法解决七条输出线太少的问题。一个简单的方法是用锁存器来锁存两次输出信息而当作一次用，这样显示信息的代码就可以有十四位，从而解决了汉字显示的问题。

8.2 汉字显示终端

8.2.1 概　述

“终端”是电子计算机技术的一个术语，我们所说的终端是用来与计算机系统进行通信的一种输入输出设备，它是人和计算机进行“对话”的工具。

一台大的计算机系统往往有很多台终端。每台终端对计算机来说可以看成一个“用户”。“多终端系统”也可称为“多用户系统”，这是指某计算机系统带有多台终端。当然，也有单用户的计算机系统，这里它只带有一台终端。

现代计算机系统对各个部分的调度和控制是由操作系统来统一实现的。首先由用户向操作系统提出申请，然后，操作系统根据情况分配给每一个用户一定的机器“资源”（例如存储空间等）。这样，用户就可以通过终端使用计算机。这时操作系统将允许终端向它发出“请求命令”。根据这些命令，操作系统能使计算机完成各种操作，直至用户发出“撤离命令”为止。另外人们往往可以通过终端输入某些原始的程序和数据，并可对某些程序或数据进行编辑、修改。也可以通过终端输出某些计算机处理的结果。上述这些功能通称为“人机对话”。

早期的终端设备比较简单，称为“哑终端（dumb terminal）”象电传打字机、控制台打字机等就是这一类典型的设备。人们只要在键盘上按一个键，就向操作系统输入一个信息（一个字节）。输出时，由操作系统向终端逐个地发送字符，终端接收以后，就“照本宣科”地在打印机上打印出来。终端本身完全是被动的，只起一个收收发发的作用。就连最简单的命令分析也要靠操作系统来完成。这种类型的终端现在还在许多计算机系统上使用，不过往往不用打印机而用CRT显示器来替代，或者既有CRT显示器，又有打印机。

近年来，随着微处理机技术的发展，终端的功能也得到了迅速发展。除了反应速度快、性能可靠、没有噪音等特点外，终端的处理能力也有了很大的提高，出现了所谓“灵巧终端（smart terminal）”、“智能终端（intelligent terminal）”等。灵巧终端已经可独立于操作系统进行“行编辑”，甚至能进行“屏幕编辑”，还能对请求命令进行简单的语法分析。有时也可以实现自行编页输出。总之，原来某些要由操作系统进行处理的事情，可以“下放”到终端一级来完成了。这样，既可使操作系统更有效地工作，又可给使用者提供方便。

智能终端所包含的意义就更广泛了。例如：图形处理终端具有某种类型的图形输入、

图形识别、图形修改和输出等功能；语音对话终端具有语音（人的自然语言）识别和声音输入输出的功能；还有一些智能终端具有自动管理通信、自动检测、自动控制等功能。智能终端的特点除了具有一般计算机终端的功能以外，往往还有一种或几种“智能”处理能力。

汉字终端也应属于智能终端的范畴。所谓汉字终端是指具有汉字输入、汉字显示、汉字打印以及汉字屏幕编辑、文件管理等功能的终端设备。

所有的终端和主机之间（或者说和操作系统之间）都应该有一个通信接口。这里所说的通信接口应该包含两重意思。一是硬件上有一个通信接口，就是指终端和主机之间有一条信息传输的通路。一般都采用“半双工”或者“全双工”的串行通信接口，并有近程、远程之分；另一重意思是指终端和主机操作系统之间的软件通信接口。这是指软件规定的通信方式，如中断方式和询问方式等，此外软件还要规定一些通信的控制代码，这些都是必不可少的。汉字终端无疑也应有一套完整的通信方式。

应该强调一点，终端和微型计算机是两个不同的概念。终端仅是具有人机对话功能的输入输出设备。尽管它在硬件结构上和微型计算机没有多大的区别，但是它仍不能称为微型计算机。两者在本质上的区别在于微型计算机应有独立的处理能力，它应有完整的、属于本身的操作系统和各种系统软件，而终端一般来说是不需要的。终端的所有软件都是为实现人机对话和各种智能而设计的，而它是在主机的操作系统控制下运行的。当然，目前国内外也确实有不少微型机可兼作终端用，特别是可以在微型机上利用其比较强的软件、硬件的开发能力来开发一些智能。也有把微型计算机的操作系统作些改变来增强原有的功能（例如增加汉字输入输出的能力），变成某种具有特殊智能的微型计算机，而不失去操作系统对各个系统软件的支持。例如近年来国内某些性能较好的汉字微型计算机确实做到了这一点。但是，我们认为汉字终端仍然是一个不同于汉字微型机的概念。某些具有汉字处理能力的微型计算机系统可以看作是一个汉字智能终端和一台微型计算机的结合体，它们之间实质上仍然是各自独立的。

此外，我们常听说有“联机终端”和“脱机终端”这样的说法。其实，这种说法并不十分确切。实际上，所有的终端都应该是联机的。终端应该是整个计算机系统的一个组成部分。所谓脱机终端的确切名称应该称为“数据工作（或收集）站（data workstation）”。这里所说的数据是一个广义的名词，它包括汉字、字符、数字等各种信息。

在某些情况下，需要输入大量的数据，如果通过终端把数据逐个地输入就显得太慢，而且要占用主机较多的时间。为此可以采用数据工作站预先把数据准备好，这就是说，在脱机的情况下，按照一定的格式把大量的数据存放到某一种“存储媒体”上。曾经大量使用过的穿孔纸带、穿孔卡片等就是两种老式的“存储媒体”。往往把纸带穿孔机生动地称为“从键到纸带”的数据收集装置。卡片穿孔机称为“从键到卡片”的数据收集装置。现在，因为微型计算机配用软磁盘比较普遍，所以多数采用“从键到软盘”的数据采集装置。一个微型机系统当然也很容易兼有某些数据工作站的能力。

如上所述，终端、数据工作站和微型计算机三者从结构上来说是相似的。但其功能是不一样的。为了突出本章的重点，我们将着重讨论终端，特别是汉字终端的有关问题。

因为终端在计算机系统中是直接面向使用者的，且与操作系统紧密相连，近年来又赋予了许多智能，所以它具有重要的意义。

在汉字信息处理系统中，汉字终端也是关键性的设备之一。它应具有下述两方面的功能。

（一）汉字输入输出功能

使用者通过汉字终端向计算机输入汉字信息。计算机系统通过汉字终端输出汉字。汉字的输入输出功能应是汉字终端本身的功能。它能把“有形的汉字”转化为“无形的汉字”(即汉字代码信息)，再输入给计算机。计算机给汉字终端送出无形的汉字（汉字代码信息）之后，由终端转化为有形的汉字而实现输出。

除此以外，汉字终端应该具有汉字和西文字符混合输入输出的功能。

（二）通信功能

汉字终端是和计算机的操作系统相连接的。它们之间必然要有一个通信接口。除了硬件有一个通信接口外，还要求有软件通信接口，这样才能使汉字终端输入的信息被操作系统所认识。操作系统和各种系统软件（包括各种高级语言的编译系统、数据库管理系统等）如果不认识汉字信息，或者汉字信息与系统软件不能相容是不行的。计算机系统要和汉字终端在软件上兼容，这是一条基本原则，也是汉字终端必须要有的功能。

8.2.2 汉字显示终端的硬件结构

由于目前广泛采用微处理机的成熟技术来实现汉字的输入输出功能，也就是说，汉字终端的硬件结构是在微处理机的基础上建立的，并没有很多特殊的技术。

一、一种典型的汉字显示终端结构

图 8-13 是一个典型的汉字显示终端构成框图。

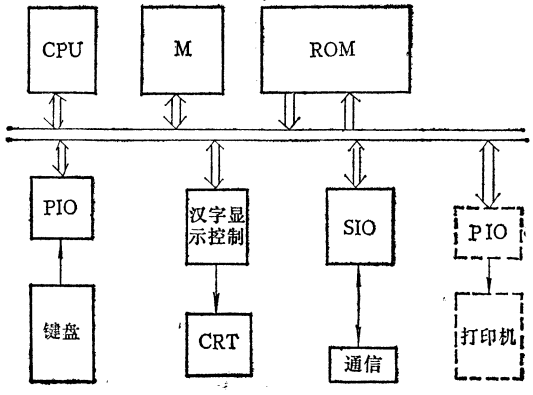


图8-13 简易汉字显示终端构成框图

汉字显示终端由下列几个主要部分组成：

1．中央处理机 中央处理机CPU是一个8位（也可以用16位）的微处理机，作为汉字终端的控制部件。它用来控制一个内部总线（包括地址总线、数据总线和控制总线）。

2．汉字字形发生器(字模库) 汉字字形固化在ROM存储器中，以四千汉字字形计算，15×16的点阵共需ROM存储器128K字节。国内已有成套的掩模ROM的汉字字模库。当然也可以采用其它方法构成字形发生器。

3．汉字输入键盘 它可以是一个标准的字母数字键盘，采用某种编码方案在此键盘上输入汉字，也可以采用笔触式汉字字盘来输入汉字。汉字输入键盘一般通过一个标准的并行接口与内部总线相连。如果采用标准字母数字键盘，则可以是并行接口，也可以是串行接口与内部总线相连。

4．汉字显示器 汉字显示器包括一个CRT扫描、刷新控制电路。为了能显示汉字，扫描控制电路可以采用中小规模集成电路，也可以采用大规模集成电路。显示器一般采用标准的单色的字符显示器；如有特殊需要，也可采用彩色显示器。

5．通信接口　汉字终端必须要有一个与主机相连接的通信接口。一般采用标准的串行异步接口，通信速率应是可选的，通常选用的通信速率是每秒 19200，9600，4800，2400，1200，600，300 位等。每秒多少位称为波特率，常用“bps”来表示。

6．汉字印刷机　最常用的是选用 9 针或 16 针、24 针的针式打印机。这种打印机一般都能打印图形，所以也称为图形打印机（graphic printer）它不同于只能打印字符的那种针式打印机。这种图形打印机要求把打印的每一点信息都传输给它。它也是通过一个标准的并行接口和内部总线相连的。

7．存储器　除上述汉字字形发生器需要占用存储空间外（汉字字形发生器的存储地址一般采用扩展存储地址），还要存放汉字键盘码与汉字内部码或国际交换码的转换表；终端工作的管理程序；显示文件存储区；打印信息缓冲区等。一般说除显示文件存储区及打印信息缓冲区及某些工作单元要用 RAM 存储器外，所有转换表及管理程序等都是固化的（采用 ROM 存储器）。

这个方案的特点是结构紧凑、简单、成本较低。其缺点是不够灵活，特别是要求联机通信软件也要固化，这样一来，如果要适应和各种系统相连接，则显得很不方便。

二、带有软盘存储器的汉字显示终端结构

图 8-14 是一种带有软盘存储器的汉字终端构成方案。它和前一种方案的不同之处是采用了软磁盘存储器。除了 CRT 控制部分为了能显示汉字需作一些特殊的考虑外，它和一般的微型计算机在硬件结构上没有多大差别。这样做有两个优点：

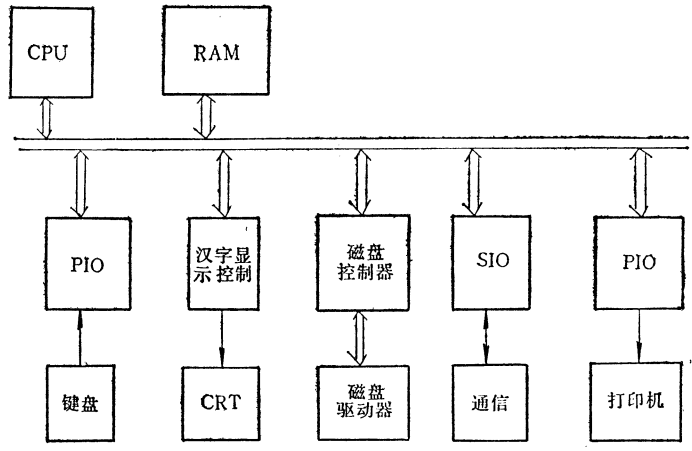


图8-14　带有软盘存储器的汉字终端结构框图
PIO—并行输入/输出接口；SIO—串行输入/输出接口。

（1）一机多用，既可以当作微型计算机用，也可以作汉字终端用。在装入了系统软件后，便可当作微型计算机使用。在装入了终端管理程序后，它便是一个汉字终端了。同时，也可以把它作为汉字数据工作站来使用。

（2）由于有了软盘存储器，故功能的开发比较容易。例如：可以实现多种输入方案；可以连接各种不同型号的印刷机；可以增加某些屏幕编辑功能；如有需要，也可以临时增加一些不常用的字模等。

尤其是可以比较容易地通过软件把终端的汉字信息代码转换成各种主机的机内代

码。进一步还可以仿真西文终端的联接软件，实现插接兼容。

以上我们从原理上介绍了两种汉字终端的结构。由于各生产厂家的原始设计不会相同，用户也会有各种不同的要求，所以其具体的硬件构成方案会各有特色。我们就不再一一列举了。

三、汉字显示终端结构的模块化问题

现在我们来讨论一下硬件的模块化结构问题。在计算机硬件设计中，模块化、标准化、系列化是一个重要的问题。

模块化、标准化、系列化的好处可以归结为:(1)可以根据不同的需要构成各种产品;(2)产品标准化以后，一个工厂生产的产品(某一模块)可以直接用到另一个工厂的产品中去。这样，既可避免重复设计，缩短试制周期，也有利于工厂生产的专业化;(3)便于设备的更新换代，组成系列产品;(4)易于实现软件设计。

模块是把某种能独立完成一定功能的部件做成一块板或者一台设备。象微型机中常见的扩充存储器模块、软磁盘存储器控制器模块、温式(Winchester)磁盘控制器模块、模拟-数字转换器模块等。此外，针式打印机和标准字母数字键盘等设备，也都可以看成是一种模块。

模块和外界的联系，常见的有三种方式:

(1)标准的总线方式;

(2)标准的并行接口连接方式;

(3)标准的串行接口连接方式。

有的模块在设计时就考虑到可同时兼有几种连接方式。用户可以自由选择。

下面介绍两种模块化汉字设备的例子。

(一)汉字字形发生器

如果把汉字字模库做在终端内部，就要扩充主存，这既不便于使用，也难于更新换代。如果做成模块化的设备就有许多方便之处。把汉字字模库做成一个外部设备，要什么汉字的字形，就向它发送一个汉字代码，然后汉字字形发生器送出这个汉字的字形点阵信息。

要增强汉字字模库的功能，可以考虑在字模库内设置一个微处理机，用以完成字模库的管理工作。这种微处理机当然也可以采用Intel8080、8085、Z-80、MC6800等8位微处理机，但是更合理的做法是采用单片机，例如Intel8048、8748;MC6801,MC6803和Z 8861等。单片微处理机的特点是把CPU、PIO、SIO，甚至还把RAM,EPROM都集中在一块芯片上，这样既可以降低设备成本，也可以缩小体积。用微处理机来管理汉字字模库，就有可能适应汉字字形的变化等工作。例如，可以在汉字字模库模块中完成汉字字形压缩信息的还原，以及实现字模信息的旋转。因为普通针式打印机要的字形信息与显示器所要的字形信息是不一样的。一般CRT的字模信息是横向一行一行地显示。而针式打印机要求竖向一列一列的信息。这就需要把字形旋转90°。而模块化结构的汉字字模库就可以承担这个任务。

由于采用了标准化接口，故字模库也可以有多种产品。例如:15×16点阵，采用压缩存储办法，每秒钟能提供40个汉字点阵的模块;15×16点阵，采用全点阵存储，每秒可提供800个汉字点阵的模块;24×24点阵，每秒可提供100个汉字字形的字模

库等。可以把多种产品配置在各种不同的设备上。

还可以把汉字字模库做成多用户分时共享的模块。它可以同时为多台终端（或打印机）服务。

这种多终端共享的汉字字模库，因为完全由字模库本身的分时方法所控制，故用户（终端）并不觉得有任何不便。

当然，把汉字字模库做成独立的设备也会有些缺点，带来些新的矛盾。例如：多了一对接口；反应速度必须与终端匹配等。但是，这同它带来的好处相比，却是次要的。

（二）汉字输入键盘模块

无论是采用汉字整字键盘，还是采用标准字母数字键盘，输入汉字时都有一个代码转换问题。因此，设计汉字终端时必须考虑适应不同用户的要求。而不同的编码方案和代码转换方法会给终端输入程序带来一些问题，例如查表的方法不同，所占的存储空间也不同。如果把输入键盘连同编码方案做成标准化的模块，就可以方便地解决这个问题。在键盘里设置一片 CPU（可以是单片微处理机），主要处理代码转换操作。键盘和终端的连接是一个标准的 PIO 接口或 SIO 接口。这样一来，所有的汉字编码方案的代码转换查表工作都可由键盘模块独立完成。键盘的输出信息是汉字代码，或是西文字符代码、控制码等。不同的编码方案可以改变键盘模块内部的存储器内容和固化程序来解决。模块化输入键盘便于更换，终端软件几乎可以适应多种不同的汉字输入编码方案。

随着新技术的发展，象语音识别输入、文字字形的识别输入、语音输出等设备都可以做成模块化结构，从而可以比较方便地扩展汉字终端的功能。

第九章　汉字信息处理系统的配置及基本汉字处理软件

9.1　汉字信息处理系统的配置

用户应按其应用环境的差异，对所需的汉字信息处理系统作不同的配置，以取得系统整体性能与价格的合理平衡。

9.1.1　系统配置的基本考虑

汉字信息处理系统的配置规模，按用户的使用环境、待处理的业务量及性质的不同，而有多种不同的方案。这里，把系统配置按规模分为三类，它们是：

1）以微型计算机为基础的微机汉字系统。这种系统是各类汉字系统的基本构成单元。按系统的性能又可分为如下几种类型：

（1）独立型汉字系统，通常为单用户工作；

（2）多功能工作站，通常为多用户工作；

（3）汉字文字处理机，这是一种专用系统。

2）以小型计算机为主机，配以各类汉字终端设备的小型机汉字系统。

3）以大、中型计算机为主机，配以各类远、近程汉字终端及工作站的中、大型机汉字系统。

对于具体的系统而言，用户根据自己的特殊需要，在系统配置上的不同考虑，主要反映在汉字外部设备及终端的类型、数量上的差异。这些设备的配置显然要考虑到系统功能与价格的平衡，用过小的系统去从事大的作业或用大的系统去从事少量作业的方案都是不合理的。此外，系统配置还应考虑以下的因素：

1．兼容性　系指系统在具有汉字处理功能的同时还应保留西文处理功能。这样，原西文处理技术中所积累的丰富的软件资源（如高级语言、文件管理、数据库管理及实用程序等）不仅能照样使用，而且还可以得到汉字功能的支持，从而可扩大系统的通用性；

2．可扩充性　鉴于汉字处理技术的不断发展，已建立的汉字信息处理系统应具有可扩充新增汉字功能的能力；

3．经济性　汉字系统的建立可从硬件及软件两个方面考虑，系统的配置要权衡硬件实现技术与软件开销的合理性，以实现有效而经济的系统。

下面以通用型系统为例，介绍各类汉字信息处理系统的配置。

9.1.2　微型机汉字信息处理系统的配置

以微型计算机为基础的系统是发展汉字信息处理技术比较方便、并且容易普及的类型。随着微型计算机技术的飞速发展，微型机汉字信息处理系统的性能也日趋完善，其

机种也不断增多。这里，仅介绍几类典型系统的配置。

一、独立型微机汉字系统

这通常指带有汉字信息处理功能的微型机系统。系统本身应具有汉字输入、汉字显示、汉字打印、汉字文件存储等基本功能。这种系统规模虽小但性能较齐全，容易设置，应用面广。

独立型微型机汉字系统的配置，有两种主要途径：一种是在已开发的西文处理用的微型计算机上作汉字功能的扩充，这包括硬件的改造和软件的开发；第二种是在设计微型机系统时，同时考虑实现汉字功能，从而开发出本身就具有汉字处理功能的微型计算机。前者主要基于对已有系统的改造；后者着眼于新的微型机系统的开发。随着微型机技术的发展，尤其是16位微型机的推广，一开始设计就考虑实现具有汉字功能的微型机系统配置，这应是发展微型机汉字系统的重点。

独立型微型机汉字信息处理系统的配置如图9-1所示。其中，可分为主机及外部设备两个部分。

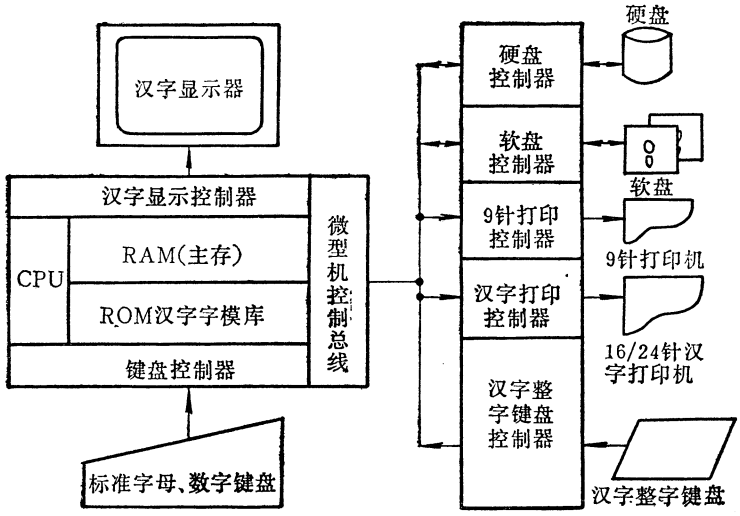


图9-1　微型机汉字信息处理系统（独立型）的配置

（一）主机部分

带有汉字信息处理功能的微型机系统，首先要求有足够的主存空间，以提供汉字输入码的转换、汉字文本编辑、汉字外部设备控制及汉字处理的服务程序等的存储能力。和西文处理系统相比，它应该有更大的存储空间，通常必须保证有64K字节的主存容量。此外，主机内应设置汉字字模库，为此，可用ROM芯片固化（仅国标一级汉字就需要128K字节以上的存储容量），这在存储器芯片价格不太高的情况下已经可以做到。与此相应，对于8位的微型机系统，由于其CPU芯片寻址能力的限制，必须采用页面方式或块方式来扩充对存储器的寻址范围；而对于16位微型机系统，由于其CPU寻址能力强，加上能直接处理2字节的汉字代码，因此它是比较理想的汉字系统构件。16位微型机汉字信息处理系统的结构简单，处理效率高，软件开发也比较方便。

汉字显示用控制器（CRTC）的配置可在设计主机内部结构中统一考虑。鉴于汉字本身是一种字符，因此，采用字符型显示控制器可获得较高的汉字显示效率。但是，也可把汉字看作图形，使用图形方式显示控制器来实现汉字显示。后者的控制方法比较简单，易于实现。汉字显示的水平是评价微型机汉字系统的重要因素，因此，在设计中应尽量采用先进的CRTC器件，以简化结构，提高性能。从实用角度来看，汉字显示屏的满屏显示汉字数应大于400字，高档机可达1,000字，这时汉字文字处理来说是必需的。尤其是在处理带表格的汉字文件时，满屏800字以上才能取得较好的效果。

标准字母数字键盘控制器也属主机配置的考虑范围。使用各类汉字编码方式进行汉字输入，相对西文输入来说，它对键盘的功能有更高的要求。因此在键盘控制器的设计中，越来越多地使用独立的单片微型计算机来实现对输入代码的转换控制。这样一来，由于它分担了主机的处理任务，故可提高主机的处理效率。由于单片机便于使用，适应性好，价格也不高，故应推广应用于汉字系统的输入控制器。

（二）外部设备

汉字印刷设备是取得汉字文件硬拷贝的主要设备。微型机汉字系统所配置的汉字印刷机通常是低速针式打印机。这种机器类型很多，其中，16针的汉字打印机可以输出汉字字模点阵为15×16的汉字，24针的汉字打印机可以打印24×24的高质量汉字，也可兼作15×16点阵汉字的输出设备。这些印刷机本身正随着微型机智能化的进展而强化其功能。例如，自带汉字字模库的汉字印刷机，可以较大地提高汉字输出的效率。不少系统，为实现汉字西文兼容，直接使用原配置的字符打印机来输出汉字。例如，对9针打印机而言，一行汉字可分二次输出。这样，尽管其输出速度低，字形质量低，但可降低系统造价。

汉字文件存储器是一种用来保存汉字文件、系统程序、高级语言的编译程序、数据库管理程序以及非常用汉字的汉字字模等的存储器，它是微型机汉字系统的重要组成部分。一般系统中都配置软盘存储器（包括软盘驱动器和软盘控制器两部分）。为区分系统用盘和用户作业用盘，常配置两台软盘驱动器。130毫米（$5\frac{1}{4}$英寸）[1]软盘容量虽不及200毫米（8英寸）[1]软盘，但因其价格低、保存和携带方便，故使用很普遍。对于每盘片容量低于200 K字节的软盘存储器，因其容量小，故不宜用于汉字信息处理系统。此外，随着汉字系统应用的深入，用户对大容量文件存储器的需求愈益迫切。密封型温式硬盘存储器由于具有体积小、易维护、价格低、并且容量大等特点，在微型机汉字系统中已越来越多地采用。130毫米温式硬盘的存储容量可达5至15兆字节，200毫米的达到40兆字节，有了这样大的存储空间，系统在处理汉字作业时提高了处理能力，应用范围也随之扩大。

由于汉字整字键盘可以直接输入汉字，故有一定的应用价值。独立型汉字系统一般应带有连接汉字整字键盘的接口。由于整字键盘的智能化发展，键盘控制器可以很简单，且易于和主机相连接。

上述独立型系统的硬件配置中，文件存储器、汉字打印机、汉字整字键盘等通常是与微型机系统的扩充总线相连接的，这样做易于扩充新的汉字设备；其他部分则是和微

[1] 按照国家标准，英制单位应改为公制单位。8英寸约相当于200毫米，$5\frac{1}{4}$英寸约相当于130毫米。

型机内部系统总线相连，以简化结构、提高效率。

随着微型机汉字系统的发展，通信功能已受到重视，通信接口的性能提高，使微型机汉字系统可以作为更高一级系统的汉字终端设备来使用，也可以由多台独立型系统组成局部网络，实现汉字处理功能的扩展。

微型机汉字系统的软件配置如图 9-2 所示。由于系统规模小，很多汉字功能可以借助软件来实现，所以软件配置是很重要的。

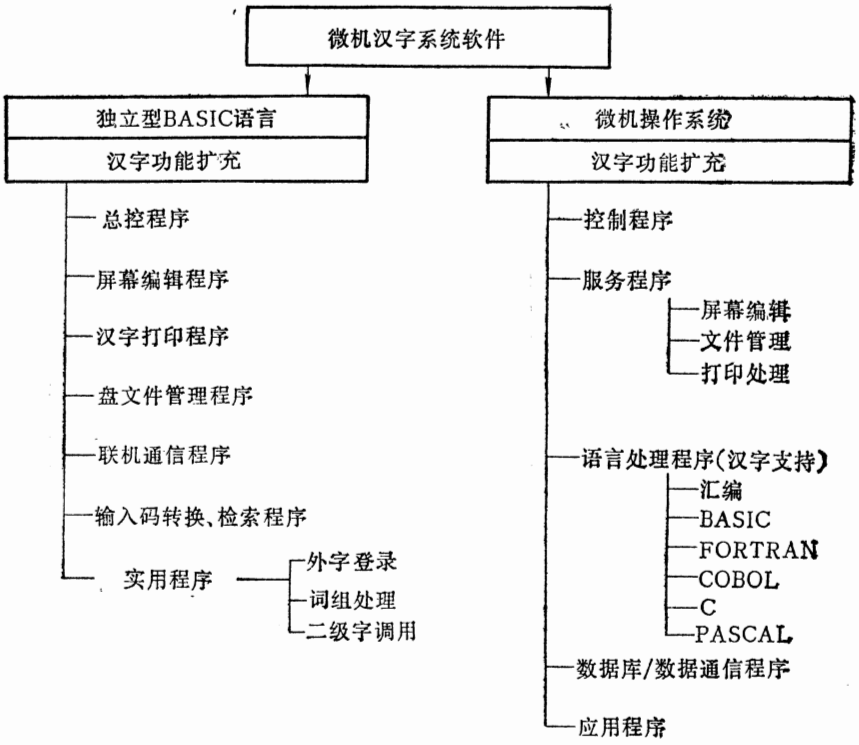


图9-2　独立型微型机汉字系统的软件配置

可以用微型机汉字系统的软件来实现汉字的输入和显示、具有汉字内容的程序的运行、汉字文件的编辑和存取管理，以及汉字的输出、通信等任务。按照汉字系统本身的能力可有多种实施方法。

一种是利用带有汉字功能的可扩充独立型 BASIC 语言来实现汉字信息处理。独立型 BASIC 语言通常固化于 ROM 中，本身带有简单的磁盘文件管理功能。因为它使用很简便，故在低档的微型机汉字系统中广为使用。

比较完善的方法是使操作系统带有汉字处理能力。这样，在操作系统支持下，汉字磁盘文件管理、高级语言、各类实用程序（如汉字屏幕编辑、汉字文字处理等）及通信软件等都可以得到汉字功能的支持，这样，可以增强系统的汉字处理能力。随着操作系统下各类软件的开发，很容易实现汉字处理功能的扩充。

综上所述，独立型微型机汉字系统由于规模小、容易建立、使用方便、功能扩充性好、系统适用性强等特点，已成为汉字信息处理中最为广泛应用的系统。

二、多用户汉字工作站

这是指以高档微型计算机为中心，连接多台汉字终端设备组成的系统。

多用户汉字工作站的硬件配置如图 9-3 所示。以下分主工作站与终端两部分来说明。

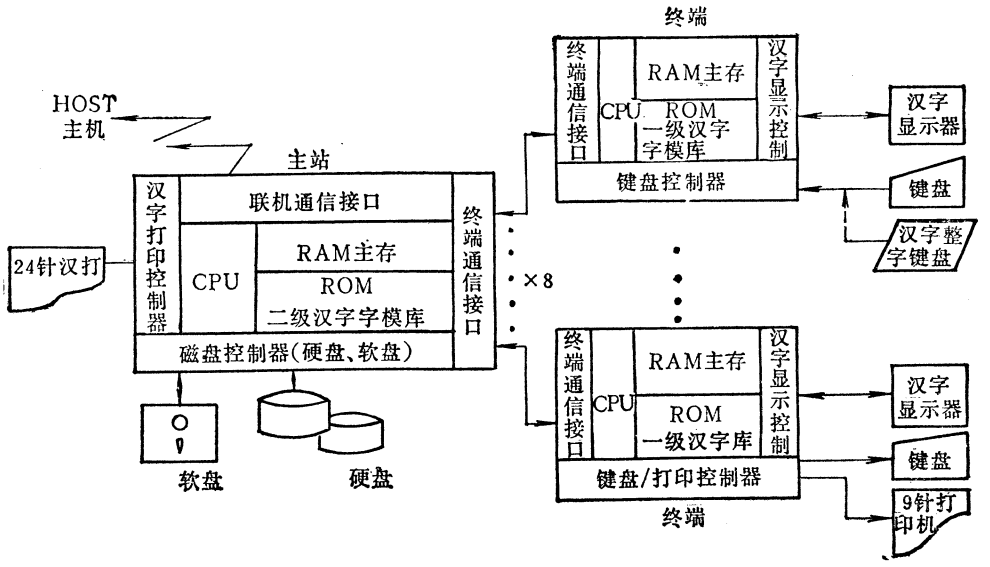


图9-3　多用户汉字工作站配置

（一）主工作站

这是一台功能较强的微型计算机，通常为 16 位微型机，主存容量为 256/512K字节，并带有汉字字模库（包括常用字和非常用字），供多台终端共用。汉字系统中价格较高的设备，如大容量硬盘存储器、软盘存储器、16/24 针汉字打印机等，由主站统一管理，为所有终端共同享用。

（二）终端部分

这是前述微型机汉字系统的简化设备，通常只带有汉字输入及汉字显示功能，并附加与主站的通信功能。

终端的配置比较简单，主存可小于 64K字节，但也应带有固化的常用汉字字模库，以提高系统工作效率。汉字显示部分的考虑和独立型微型机汉字系统相同。为适应多种汉字输入方式的需要，除使用标准字母数字键盘外，还应能加配或改配汉字整字键盘。考虑到终端设置的场所条件，为了作业的方便，终端也可加配简易的打印设备，能兼作汉字及西文打印。

终端与主站的连接，一般采用较低层的通信协议，接口简单，以近程方式工作。一台主站可连接达 8 台终端，同时进行汉字信息处理作业。

多用户汉字工作站的软件配置较独立型系统复杂。通常，由主站上配置多用户操作系统（如 UNIX，MP/M等）来统一管理，显然，操作系统应作汉字功能扩充。这样做，各终端在进行汉字处理作业时，如同在使用独立型微型机汉字系统时一样。

就工作站系统整体来说，性能价格比高于独立型系统。终端结构虽然简单，但主机配置的功能较强，可完成较复杂的汉字信息处理作业。多用户汉字工作站在配置了高一级通信协议的通信软件及接口后，就可以作为大、中型汉字系统的一个子系统（或称为

终端工作站)。因此，这样的系统对于小型的汉字信息处理应用是较为理想的，系统的扩充性能也比较好，适合于小型企业单位使用。

三、汉字文字处理机

这是一种专用的微型机汉字系统，用于汉字文稿的输入、编辑、存储、印刷及传送。随着办公室自动化的发展，文字处理机的需求量也迅速增大。

汉字文字处理机的典型配置如图 9-4 。可以看到，其结构与微机汉字系统是类似的。不同之处在于文字处理机对汉字显示要求较高，满屏汉字数要求在 800 字以上，对汉字打印的要求也较高，通常要求汉字字模点阵为 24×24，以满足正式公文文件规格的需要。文字处理机所带的软盘存储器容量要大，即使是 130 毫米软盘也要求每盘片容量在 500K 字节以上，以容纳更多的汉字文稿。

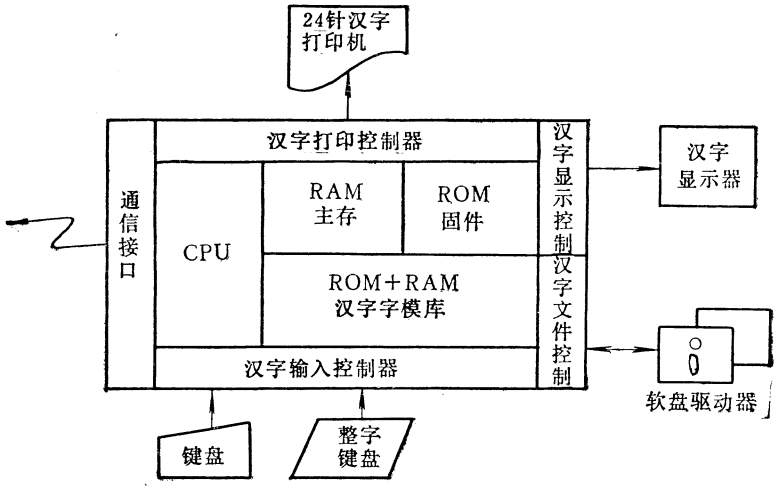


图9-4　汉字文字处理机配置

汉字文字处理机的软件是专用软件，要考虑到操作人员是普通的办公室人员，不具备专门的计算机使用知识。文字处理机并不提供传统的高级语言处理程序，而是使用简明的交互式操作方式，用带汉字提示信息的菜单方式，指示用户进行作业。为了提高输入效率，机内常带有各种汉字属性典及常用汉字词典，以便于汉字的键入和索引查找。为了满足财务处理上的需要，文字处理机还提供键盘级的计算功能，如同普通计算器一样操作简便。

汉字文字处理机为普通人员从事汉字信息处理提供了有效的工具。它问世以来，得到迅速的发展。文字处理机在加进了通信功能后，也可连接成网，组成功能齐全的办公室自动化系统。

9.1.3　小型机汉字信息处理系统的配置

小型机汉字信息处理系统使用通用小型计算机作为主体，配置以各类汉字外部设备和数量众多的汉字终端设备。与微型机汉字系统相比，它有更强的汉字处理能力。这类系统在配置上的特点是：

（1）系统本身是通用型系统，西文字符处理仍是这类系统的主要功能。扩充汉字

功能后，整个系统便成为汉字、西文兼容的系统。

（2）由于主机功能很强，故可以连接很多台汉字终端，一般可连接32～128台。终端的类型也较多，以近程终端为主，也可通过远程通信控制器连接远程终端。除汉字终端外，还可以配备较多的字符终端。

（3）小型计算机所提供的终端连接通道因机种而异，接口信息的规定差别也较大，因此，为了符合主机的要求必须精心设计终端与主机的通信规程，也可以通过仿真技术同时从软件上解决连接上的差异。

（4）小型计算机汉字系统的主机可以配置各类高档汉字外部设备（如简易激光汉字印刷机等），以提高汉字信息处理效率。主机由于资源丰富，存储容量大，可在机内建立汉字属性典、常用词典等，以提高汉字输入、检索能力。

图9-5所示是单总线结构的小型计算机汉字系统的典型配置。国产2000系列机就可按此方式配置汉字系统。可以看出，由于主机能力强，这类汉字系统的配置齐全，较微型机汉字系统有更强的处理功能。

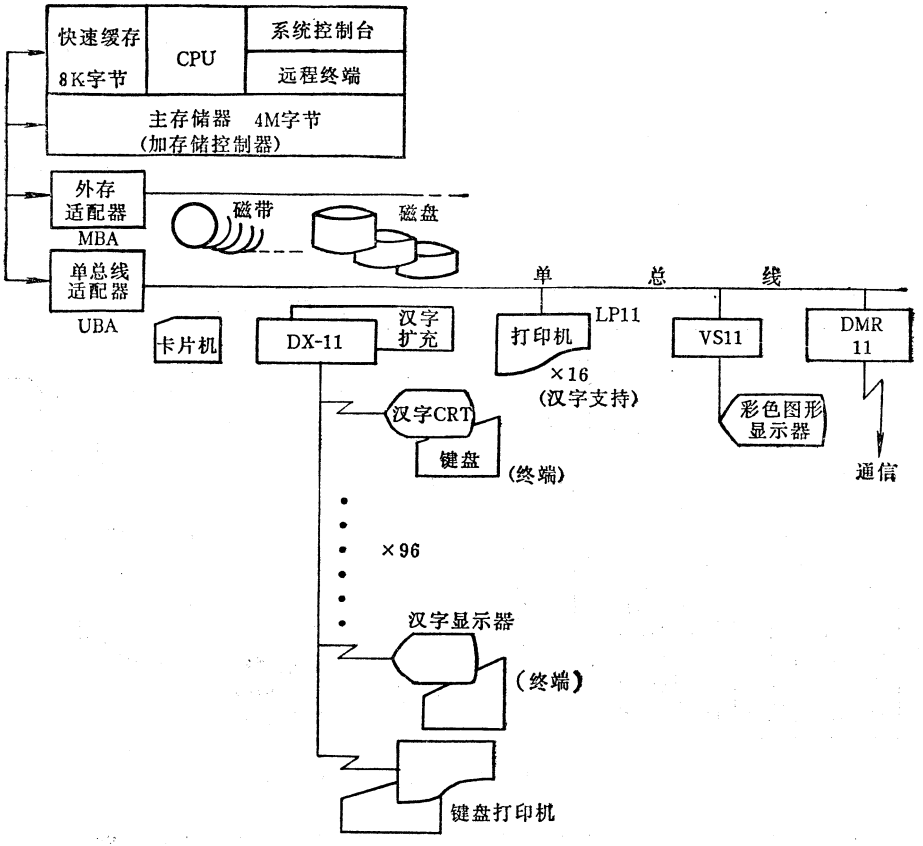


图9-5 小型机汉字系统的典型配置

小型机汉字系统的软件配置与系统的汉字功能建立方式有关。由于小型机的操作系统远较微型机复杂，故用修改操作系统的方法使之具有汉字处理功能工作量很大。因此，一种较为方便的办法是原则上不对小型机的操作系统作修改，而是利用接插兼容的汉字智能终端来实现用户的汉字作业。主机层主要担任汉字文件管理等任务。但是，对

主机而言，是不需区分汉字或字符的，汉字只是作为普通字符来处理。汉字外部设备也接在智能终端上，完全由终端来担任汉字信息处理任务。这样就使原有的小型机系统很容易成为能处理汉字信息的系统，而要做的软件工作量却很小。因此，这是一种很简便的方法。

但是，小型机汉字系统的根本建立方法还应从操作系统着手。目标是使操作系统能识别汉字代码或西文代码，建立汉字外部设备的驱动程序模块。这样就使小型机操作系统（实时、多用户）所支持的丰富的软件资源都可直接为汉字作业所利用，高档的汉字外部设备也可由主机来管理，供各终端用户分享。这样一来，整个汉字系统便可获得更高的性能价格比，从而可充分发挥小型机的功能。

小型机汉字系统由于功能强，故可以组成多种应用系统，如小型情报检索系统、小型数据库管理系统，中、小型企业的企业管理系统等。这类系统的应用覆盖面广，价格适中，与微型机汉字系统相配合可满足较大范围用户的需要。

9.1.4 中、大型机汉字信息处理系统配置

这是以通用中、大型计算机为核心，配置上大量的远、近程汉字终端、汉字工作站（终端计算机）以及各类高档汉字外部设备的汉字信息处理系统。

可以从图 9-6 所示的这类系统的典型硬件配置看出它的特点：

（1）通过终端控制器、远程通信控制器及通信设备实现主机与上百台汉字终端设备（包括部分字符终端）的连接。

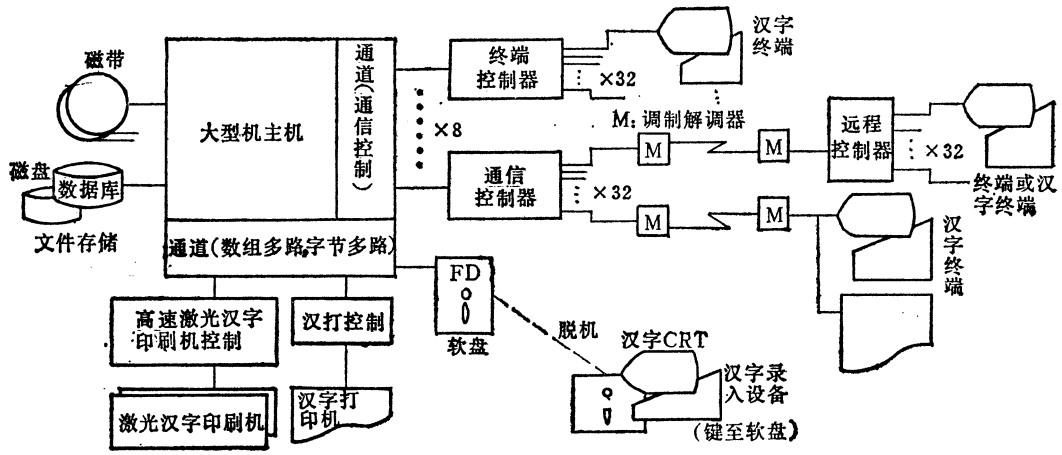


图9-6 中、大型机汉字信息处理系统的配置

（2）在终端设置方面，远程终端占相当比重。根据设置场所的实际需要，终端类型可以多种多样。有只能联机工作的简易型及灵巧型汉字终端；也有可以脱机工作的独立型汉字智能终端；也可配置规模更大的汉字工作站，由它再连接一批终端。所以，在大型机汉字系统中，其通信接口部件是至关重要的。其通信方式较前述系统复杂，通信协议也需用较高的规格，以提高联机工作效率。显然，对于不同的机种必须对所需的通信接口作特别的设计。目前，为使汉字终端能适应几种大型机的连接需要，通信仿真技

术也已成为主要的研究课题。

（3）汉字外部设备的种类繁多。大型机汉字系统应配置高速的汉字输出设备。例如配用激光行式印刷机（每分钟可输出数千行以上的汉字），或者配用价格稍低的页式激光印字机（如同复印机一样）。为了根本改变汉字输入的速度，在大、中型机汉字系统中也可以使用汉字光学字符阅读设备（汉字 OCR）。

由于大型系统信息处理能力强，汉字信息输入速度远不能适应要求，故大型机汉字系统还需配备脱机的汉字数据采集装置。以软盘为信息载体的汉字数据采集装置，在大型机汉字系统中可以得到广泛的应用。

（4）大型机系统由于其主存及文件存储设备容量极大，故可获得更高水平的汉字信息处理功能。以汉字字模库来说，所收容的汉字字数可达数万。可建立不同字体的字模库，可以采用 32×32 点阵的高质量汉字字模。可以建立比较完整的汉字属性库及汉字词库。由于可以配备高性能的汉字设备和软件，能充分扩展系统的汉字信息处理能力，以满足大型汉字信息处理系统用户的使用要求。

中、大型通用计算机是以数据处理为主要对象的系统，汉字处理只是该系统功能的一个部分，所以必须具备汉字、西文处理的兼容能力。这类系统的软件配置与小型机系统类似，也有两种考虑方案：

一种是由于大型机的实时/虚拟存储操作系统十分庞大和复杂，故原则上不对这样的操作系统作扩充，而是使用接插兼容技术，使汉字终端及汉字外部设备通过仿真程序及接口同主机标准通道的控制器相连接，汉字信息处理在接插兼容汉字终端或接插兼容汉字印刷机上实现，主机并不区分汉字与字符。这样做的优点是易于实施，工作量

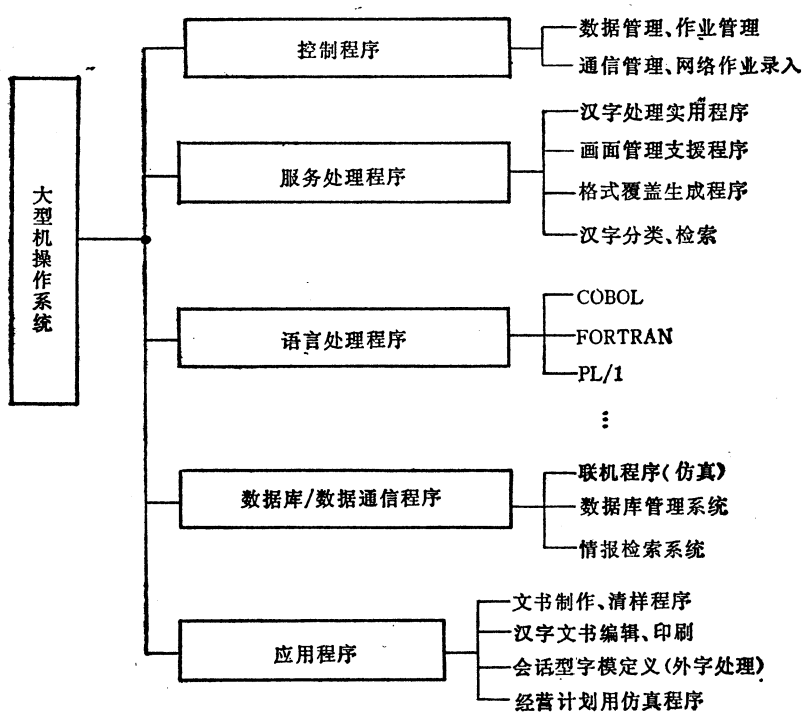


图9-7　大型机汉字系统的软件结构

小。因此，作为实现大型汉字系统的第一步，这是一种有使用价值的方案。

另一根本性的途径是对大型机操作系统进行汉字功能扩充，或者设计能兼容汉字和西文处理的操作系统。虽然这种扩充和改造还会涉及到其所支持的大量系统软件，从而工作量很大，但这样做，可以使系统原有的各种西文信息处理方式（如批处理、交互型作业方式等）都能得到汉字功能的支持；相应的高级语言、文件管理系统、数据库管理系统及公用程序，都可以具有汉字处理能力。因此这是一个十分可观的软件资源。这样做，才能把大型计算机的处理能力和汉字信息处理功能更紧密地结合起来，从而可以实现规模更大的汉字信息处理系统。

图 9-7 是大型机汉字系统的软件组成。它分为控制程序、语言处理程序、服务程序、数据库和数据通信程序，以及各类应用程序。这是一个庞大的软件系统。开发这些汉字信息处理软件对建立大型机汉字信息系统有着重要的意义。

上面介绍了各种规模的汉字信息处理系统的基本配置。用户在建立自己特定的汉字应用系统时，应根据实际的应用环境来考虑配置相应的汉字信息处理系统。汉字外部设备、汉字终端的种类繁多，系统主机的规模差别也很大，因此必须精心选配，以组成合理的系统。同时，也应着重考虑软件的配置，才能使系统发挥应有的效能。

以下，就汉字信息处理系统所应具备的基本处理软件作进一步的介绍。

9.2 汉字输入处理

汉字输入方法是人（使用汉字系统的人）机联系的关键问题。就输入设备来说，有汉字整字键盘、汉字字根键盘、标准字母数字键盘等各种汉字输入设备。不同的用户有着各种不同的要求。为了满足各类用户的输入要求，有效地管理各类汉字输入设备，并将输入的汉字信息转换成机内的标准形式。这就需要设计专门的汉字输入程序。

9.2.1 汉字输入方式

一、主控制台输入方式

这种输入方式比较简单，也较直观。但由于主控制台输入方式是人工按键，所以很费机时，很不经济。但作为联机键入少量汉字数据，使用这种方式还是必要的。

对于通用型计算机来说，一般均采用标准字母数字键盘（有的也可配上汉字整字键盘）作为主控制台。主控制台输入期间，汉字输入程序的部分工作可由操作系统的输入控制功能来完成。在输入控制将输入的汉字信息原封不动地放入内存区后，汉字输入程序根据用户程序提供的信息加工参数，进行必要的加工处理。

从主控制台输入的汉字信息都转换成相应的数据。根据汉字的某些具体特征，用户采用主控制台方式输入汉字数据时，应采用某种输入方法，而这些输入方法取决于汉字输入编码方案，计算机接收了这些汉字数据后，再进行一些后继处理。

例如，若每个汉字输入编码都是固定长的，就可以采用直接按键的方法来输入它们的编码。汉字输入程序可根据汉字编码的固定长度把汉字一个个分割开来，直接将键入的数据转换成机内码，并存入主存储区；若每个汉字的输入编码是可变长度的，则可根据某种约定补足固定长度后再键入下一个汉字，或在各个汉字的编码之间添加一个分隔符。总之，不论采用上述那种方法，都是为了加工处理时能把相继输入的汉字信息区分

开来。假定要输入“计算机”三字，且这三字的编码分别为

计　JIIK

算　SP

机　JI

它们中最长的编码为4个字符，于是，在“算”字的编码“SP”后添两个“⊗”凑成4个字符“S P ⊗⊗”；同样也可在“机”字的编码“J I”后添两个“⊗”凑成4个字符“J I ⊗⊗”。这样即得到“计算机”三字的汉字编码为“J I I K S P ⊗⊗ J I ⊗⊗”，以达到输入“计算机”三个汉字的目的。当然，也可以在每个汉字编码之间添加一个特殊字符“；”，把一个个汉字人为地隔开。这样得到“计算机”三字的输入数据串“J I I K ; S P ; J I ;”，而后通过主控制台依次键入以上的汉字信息。

从上面的讨论可知，主控制台输入作为一种联机输入方式，在整个输入过程中始终占用了主机时间，因此很不经济。这种方式仅使用于少量汉字数据的输入，而大量的汉字数据输入主要依赖于媒体记录输入方式。

二、媒体记录输入方式

媒体记录输入方式适合于大量汉字数据的输入。这种方式是首先用脱机方式将汉字数据记录在纸带、卡片、软盘、盒式磁带这类信息媒体上，接着将记录有汉字数据的媒体安置在联机的汉字输入设备上，然后在汉字输入程序控制下，将记录的汉字数据高速读入主机。常用的输入设备有“键到软盘”，“键到磁带”等脱机数据采集设备。

显然，采用这种方式必须着重考虑以下两点：首先考虑主机系统是否具备可读出相应汉字数据媒体上信息的输入设备。如果将汉字数据记录在盒式磁带上，则主机系统应具备盒式磁带机；其次，必须考虑汉字数据格式的相容性。如果将汉字数据记录在软盘上，而主机系统的软盘机不能识别该软盘上的信息，则同样是不可行的。

下面以软盘输入方式为例作详细说明。该方式首先使用键到软盘输入设备将汉字数据记录在软盘上，再把记录有汉字信息的软盘置于联机的软盘机上。在输入程序控制下，将记录在软盘上的汉字数据读入内存，并加工输入信息，使之转换成后继处理所需的数据格式。假定采用4位十六进制数组成一个汉字的输入编码，且软盘上的数据形式为：“1 A B 3 … 4 A B C ⊗ 1 A C … A F”。在读出软盘上记录的信息时，汉字输入程序首先调用操作系统的输入控制，将汉字数据原封不动地读入内存缓冲区。根据用户要求，把上述汉字数据加工成下述的数据格式：

地　址	内　容
1 ⊗ ⊗ ⊗	1 A B 3
⋮	⋮
1 5 ⊗ ⊗	4 A B C
1 5 ⊗ 1	⊗ 1 A C
⋮	⋮
	A F

软盘输入方式作为一种媒体记录输入方式，解决了主控制台输入时人工按键速度慢、过多地占用主机的矛盾，并且软盘具有便于修改、体积小、价格低等优点。因此，将汉字数据记录在软盘上，通过联机的软盘机将软盘上记录的汉字信息读入主机，这是一种较为理想的媒体互换方式。

三、联机终端输入方式

联机终端输入方式较为常用。它是通过通信接口将汉字信息送至主机处理加工的。

汉字终端设备有智能型、灵巧型、简易型等多种，它们的配置除键盘和显示器之外，根据需要还可加上汉字整字键盘、软盘存储器，汉字印刷机等；它们的功能差别也很大，有的可以不设置汉字字模库，汉字功能由主机承担；有的却相当于一台具有汉字功能的微型计算机。在汉字输入手段上，仍不外乎采用标准字母数字键盘按编码输入，以及采用整字键盘的直接键入等类方法。

主机系统分时或实时地响应各台终端的汉字输入操作，并将各台终端输入的汉字信息分别存入它们各自的输入缓冲区。

主机除了响应各台汉字终端输入信息外，若有空余时间，还可运行其他程序。因此，这种方式具有主控制台输入方式的优点，既可随时输入汉字信息，也可随时增加、删除、修改这些信息。这样，主机时间也能得到充分合理的利用。

图 9-8 表示上述三种输入方式的示意图。

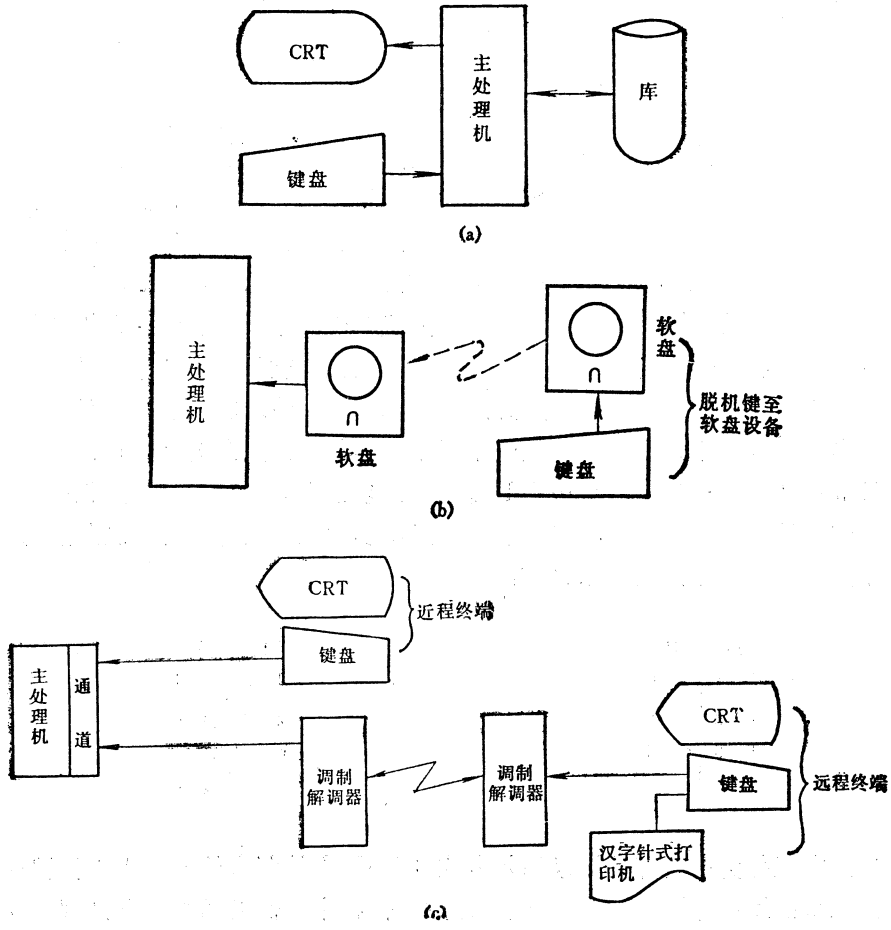


图9-8　三种输入方式示意图

（a）主控制台方式；（b）载体记录输入方式；（c）联机终端输入方式。

9.2.2 汉字输入程序

汉字输入程序是汉字信息处理系统的基本程序之一。它是解决汉字信息进入机内的关键。下面从四个方面加以叙述。

一、汉字输入程序与操作系统输入控制的关系

目前大多数具有汉字信息处理功能的计算机系统是在通用型操作系统支持下，由“汉字信息处理程序模块”来实现汉字信息处理的。因此，在此环境中，汉字输入程序是在该程序模块的管理下运行的。当输入汉字数据时，汉字输入程序首先向操作系统提出调用输入控制的请求，操作系统的输入控制在被汉字输入程序调用时，必须赋于存放输入信息的内存起始地址等输入参数。然后，汉字输入程序根据输入要求、汉字的一些基本特征以及汉字数据类型等因素进行必要的加工处理。

鉴于汉字输入与西文字符的输入有某些相似之处，又有某些特点，为了兼顾操作系统的兼容性和原有软件的可利用性，绝大多数汉字处理系统都把输入控制和汉字输入程序分开，但汉字输入程序也作为系统软件资源。当出现输入汉字请求时，若有部分过程可由操作系统的输入控制实现，则汉字输入程序可以通过系统调度模块调用输入控制。若在输入过程中既有汉字输入又有西文字符输入，则操作系统的输入控制也可通过系统调度模块调用汉字输入程序。

二、汉字输入程序的设计

汉字输入程序是汉字信息输入计算机所必需的。由于汉字输入设备的种类不同，且性能各异，输入方法也有多种。所以，汉字输入程序要面向不同的输入设备和各种输入方法来设计，以满足各类用户输入汉字数据的需要。

当用户程序请求输入汉字数据时，汉字输入程序首先将用户程序提供的输入方式、存放当前输入的汉字数据的缓冲区起始地址、缓冲区长度等必需参数赋于操作系统的输入控制，输入控制在汉字数据（二进制形式）输入过程中不断地测试汉字输入设备所处的状态，协调主机与输入设备在时间上的差异，同时把输入设备送至计算机的信息转换成主机相容的格式。输入控制返回时，送出汉字输入程序所需的结果参数。然后，汉字输入程序按照汉字的基本特征（例如；一个汉字的编码不能拆开；有的方案不允许汉字编码大于某个确定的数等），检查缓冲区中读入的汉字信息的正确性。若有错误，则将错误信息反馈给用户。在某些系统中，由于汉字输入编码同机内码，机内码同字模库中的地址码存在某种映照关系，因此，汉字输入程序还必须查找输入编码与机内码的索引表，把输入的汉字信息转换成标准的机内处理用的汉字信息。汉字输入程序并根据用户的要求，汉字数据类型等因素，对转换后的信息进行必要的加工。此后，该输入程序将必需的信息赋于后继模块，或置入约定的单元。

三、汉字输入程序的工作流程

上述的汉字输入程序模块的功能，可以用图 9-9 的工作流程图表示。

四、调用汉字输入程序的方法

由前面叙述的汉字输入程序的功能及实现方法可知，调用该输入程序的方法按照它和操作系统的关系不同而异。

若汉字输入程序是系统软件资源，在操作系统开工后，汉字输入程序即处于准备状

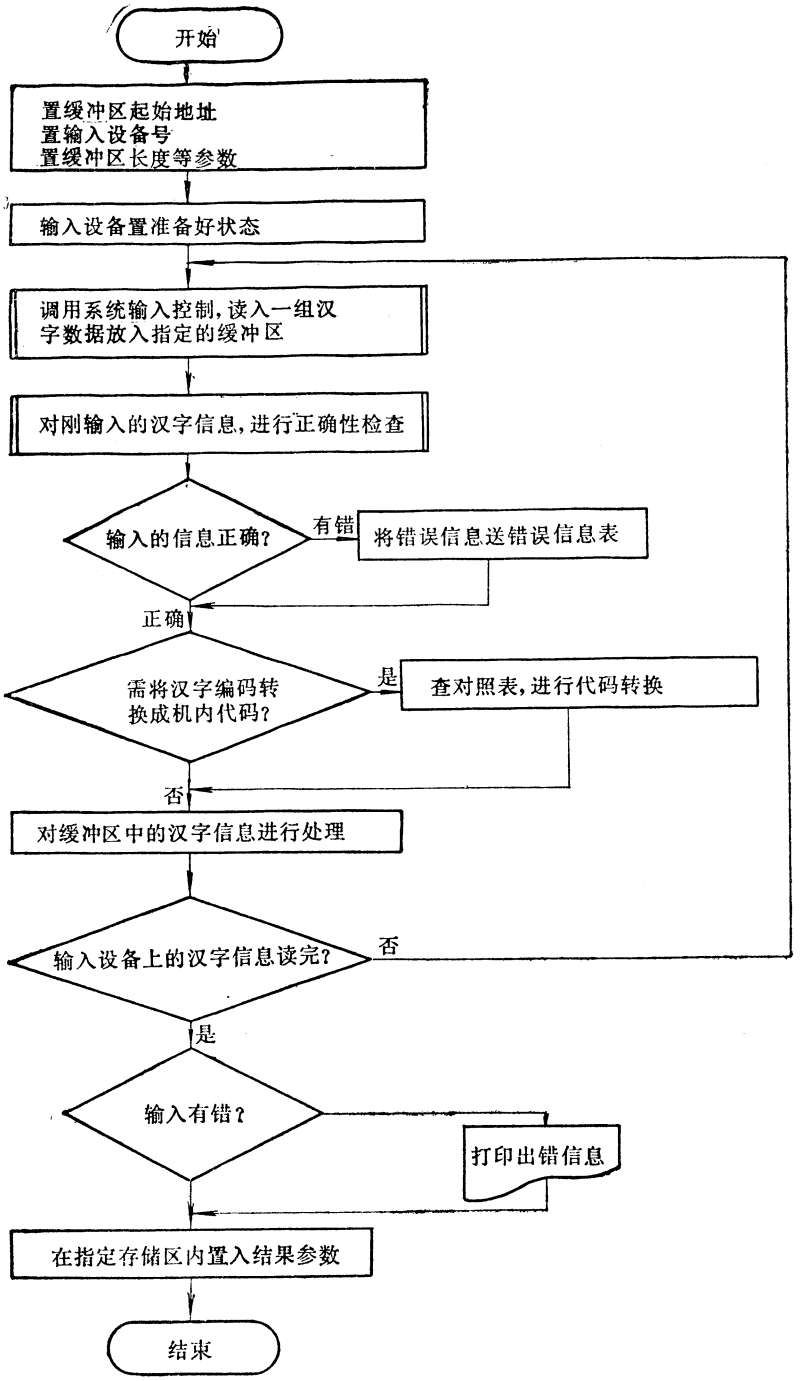


图9-9　汉字输入程序工作流程图

态。若需调用某台汉字输入设备的输入程序运行，只要在特定的存储单元中设置输入设备号、内存缓冲区起始地址、输入方式、数据区长度等参数。根据该程序名，使用操作系统提供的广义指令即可调用。输入程序返回时，在约定的存储区内设置结果参数或调用出错信息。

若汉字输入程序模块不是系统资源，而是把汉字输入设备作为用户设备定义的，则调用汉字输入程序运行有两种方法。第一种方法如下。在操作系统启动后，启动汉字信息处理管理程序，并在特定的缓冲区中记入汉字输入程序名、输入设备号和输入方式等所需信息，管理程序根据缓冲区中的信息转去调用某台汉字输入设备的输入程序，该程序运行结束时，将结果信息馈给汉字信息处理管理程序，以供后继处理用。第二种方法如下。若没有管理程序，则在按程序名调用之前，必须将该输入程序运行中所需的信息置入它的工作区，而后键入输入程序名，即可调用该程序运行。调用返回时，输入程序按照约定方式将结果信息通知用户程序。

9.2.3 汉字输入编码转换程序

一、汉字输入编码转换的意义

在汉字信息处理系统中，经常使用多种不同形式的汉字输入编码。这种编码进入计算机后，必须经过转换处理，才能完成检索汉字、显示、打印等信息加工的任务。

将汉字的输入编码和要转换的汉字代码之间建立一个对照表，再根据表的结构采用不同的查找方法，将输入的汉字键盘码转换为所需要的汉字代码。这就是汉字编码转换的基本思想。例如，汉字采用声韵部形编码，通过字母数字键盘输入，其键盘码是四个西文字母组成的代码，执行转换程序后就可以把它转换成国标交换码或其它代码，再转换为字模库的地址码就能检索并输出汉字了（见图9-10）。

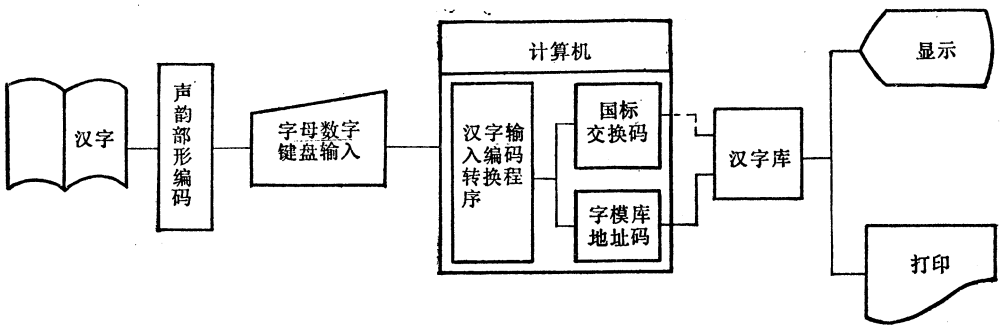


图9-10 汉字输入编码的转换示意图

汉字编码的转换对联机汉字系统是很重要的。因为转换技术的好坏直接影响到机器的响应时间，而机器的响应时间则是评价一个联机系统优劣的重要因素。所以，编制好汉字代码对照表，优选转换方法，是一项重要的技术。

二、汉字输入编码转换的方法

汉字在字模库中是按照地址码递增的顺序存储的。对于随机输入的汉字，按照编码规则确定的代码不可能是连续的，要迅速地从汉字库中找到输入的汉字，关键是要找到随机的汉字代码所对应的汉字存储地址，然后由该地址码找到相应的汉字。

将汉字的输入编码转换为字模库中的地址码，通常是将输入的汉字编码（例如一级字、二级字共6763个）和存储在字模库中的汉字地址码之间建立一个对照表。

要编制好汉字代码对照表，首先要确定系统选定的汉字输入编码和转换的结果代码之间的结构关系，然后选择一种适合于这种结构的算法。例如，一一对应的关系或者由输入编码直接计算出相应的存储地址等。如果对选定编码的结构特性不太了解，最后就作一定数量的实验，以便根据统计结果，找出它的规律性。

例如，英文字典是按照单词的字母在字母表中的顺序编排的，所以，查阅英文单词时，可以根据单词中每个字母在字母表中的位置能很快地找到它。

查找对照表的算法直接影响到计算机的使用效率，快速的查找方法能够大大提高程序的运行速度。

查找汉字代码对照表，实现汉字代码转换的方法有很多种，常用的包括顺序法、对分法、索引法和散列法四种。

现在举例加以说明：

（一）顺序法

设有七个采用声韵部形方式的汉字编码以及相应的汉字在字模库中的地址码，其排列情况如表9-1所示。

表9-1　声韵部形编码和字库地址码对照表

汉　字	编　码	地　址　码
动	DWLK	8341
办	BJLN	8061
财	CRBF	8211
爱	AIJN	800A
目	MUML	8E48
老	LXLU	8D4F
朋	PYYC	9049

当“爱”字的声韵部形编码 AIJN 输入计算机后，采用顺序法转换为地址码，先从表头的“动”字开始，逐个与“爱”字的输入编码 AIJN 进行比较，若在表中找到与“爱”字相同的输入编码 AIJN，便转换为与 AIJN 相应的字模库地址码 800A；如果查遍整个表而没有找到与“爱”字相同的输入编码时，便给出相应的信息或采取其它的措施进行处理。

顺序法转换的效率主要是根据汉字的输入编码在对照表中的位置所进行的查找次数来评价的。

对于较长的代码对照表，可以在表尾赋一特殊记号，以避免每一步查找都要判断整个表是否结束。这样做几乎可以节省一半的机器时间。

设 n 为对照表中被查汉字输入编码的数量，则用顺序法转换的平均查找次数为：

$$M=\frac{1}{n}\sum_{i=1}^{n} i=\frac{1}{n}(1+2+\cdots+n)=\frac{n+1}{2}$$

（二）对分法

设有十一个用声韵部形方案的汉字编码和国标交换码，其对照关系如表 9-2 所列。

表9-2　汉字输入编码和国标交换码对照表

序　号	汉　字	输 入 编 码	国 标 交 换 码
1	啊	AAKO	16-01
2	岸	AJMA	16-22
3	兵	BJBW	17-88
4	本	BLMF	17-30
5	当	DQXY	21-17
6	电	DSDE	21-71
7	出	IUUY	19-86
8	借	JKRK	27-72
9	宽	KPRP	31-77
10	秋	QHHI	39-79
11	英	YGCG	51-02

这是一个汉字编码按递增的顺序排列的表。因此，可以采用对分法将输入汉字编码转换为国标交换码。

假如我们输入“秋”字的汉字代码 QHHI，要求使用对分法进行查找，并把它转换为国标交换码。其查找过程如下：

先将十一个汉字依次编号，然后执行以下查找过程。

第一步，取$H=\frac{1+11}{2}=6$，即取位置 6 的“电”字输入编码 DSDE 与 QHHI 比较，因为 DSDE＜QHHI，所以可确定在表的后半部查找；

第二步，再取$H=\frac{6+11}{2}=9$（四舍五入），取位置 9 的“宽”字的输入编码 KPRP 与 QHHI 比较，因为仍然有 KPRP＜QHHI，所以仍在表的后半部查找。

第三步，再取$H=\frac{9+11}{2}=10$，取位置 10 的“秋”字的输入编码与 QHHI比较，这时两者相等，就将QHHI 转换为国标交换码3979。

用对分法进行代码转换，要求汉字编码对照表是一个有序排列，它的平均查表次数为

$$N=\frac{\sum_{i-1}^{n} i\cdot 2^{i-1}}{2^n}=n-1+\frac{1}{2^n}\approx\log_2\cdot n-1$$

其中 n 为对照表中被查汉字输入编码总数，N 为对分法查表次数。

可以证明，这是一种平均查表次数最快的线性查表方法。

（三）索引法

先用对分法查找索引表，再在所确定的区域使用对分法或顺序法，查找与输入汉字相同的代码进行转换。它的速度介于对分法和顺序法之间。

（四）散列（HASH）法

我们知道，对分法和顺序法都是通过一系列的比较，才能确定与输入汉字相同的代码。所以，可以统称为对汉字输入编码进行比较的转换方法。

若对汉字输入编码采用某种算法，就能直接确定相应的位置，这就是散列地址查找法的基本思想。

设汉字输入编码表A是一个长度为n（例如$n=5$）的表，$a_i(1\leqslant i\leqslant 5)$为输入编码的地址，$X_i(1\leqslant i\leqslant 5)$是汉字输入编码，则汉字输入编码$X_i$和地址$a_i$之间存在着一定的函数关系（用$H$表示），使得等式$H(X_i)=A(a_i)$（$i=1,2,\cdots,n$）成立。

表9-3是汉字输入编码表A。

表9-3　汉字输入编码表A

地　址	汉　字	输入编码
a_1	的	X_1
a_2	大	X_2
a_3	了	X_3
a_4	人	X_4
a_5	上	X_5

这里，$A(a_i)$是汉字输入编码在表中的地址，$H(X_i)$称为散列函数（Hash function），通过散列函数就可以把汉字输入编码集合中的编码映照到一个地址集合中去。

散列函数是一个压缩函数。一般情况下，汉字输入编码集合比地址集合大得多，所以，不同的汉字输入编码有可能映照到同一个散列地址（即所谓共址代码）中去。

因此，构造散列函数必须解决如下两个问题：（1）对给定的一组汉字输入编码，选择一种算法简便、均匀分布的散列函数，尽量减少共址代码的发生；（2）制定解决共址代码的方法。

1．构造散列函数的步骤及方法

由汉字输入编码求存储地址的转换算法常分为如下几步：

（1）如果汉字输入编码非数字，则在不丢失其有用信息的情况下，先转换成能够推算的数码。

（2）对数码选择适当的算法，将数码转换到所允许的地址编号范围内，尽可能地得到均匀分布。

（3）把结果编号乘上一个压缩常数，压缩到准确的地址范围里。

上述三步中，第二步的算法有多种多样，这里只取常用的除算法加以说明。

所谓除算法，即用一个小于但又接近于可用地址的质数来除汉字输入编码的数值，所得的余数作为该汉字输入编码的相对存储地址。

$$\frac{\text{汉字输入编码}}{\text{可用地址}}=\text{商}\cdots\text{余数}$$

如果取有效存储地址为余数加1，可用地址为1000，汉字采用某种数字编码，则使用除算法计算散列地址的例子见图9-11所示。

一般地说，用除算法计算存储地址，只要选择的质数不与组成汉字输入编码的基数直接相关，就能得到较均匀的地址分布。

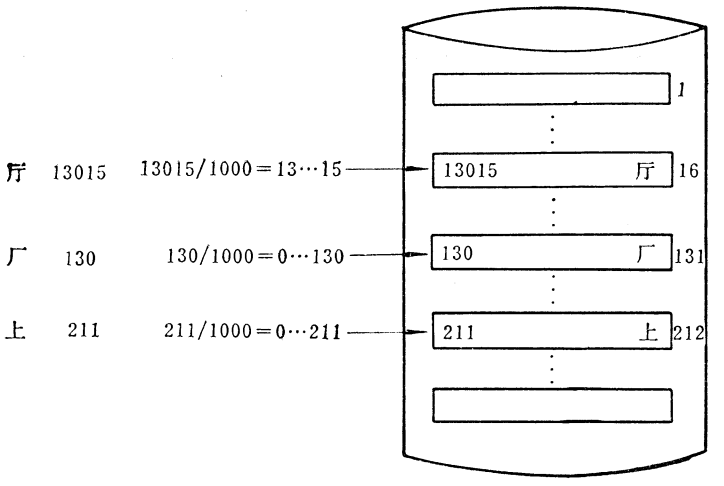


图9-11 用除算法计算散列地址的例子

2．解决共址汉字代码的方法

如果两个不同的汉字输入编码经过计算转换后得到同一地址，则必须设法把它们区分开。

例如，设一汉字的输入编码为48325，另一汉字的输入编码为15325，则使用除算法得到的存储地址如下：

$$\frac{48325}{1000}=48\cdots325 \qquad \text{存储地址为326}$$

$$\frac{15325}{1000}=15\cdots325 \qquad \text{存储地址为326}$$

得出相同的地址码。

解决共址汉字输入编码的常用方法是链地址法，即用链表来表示溢出表（其中存放共址汉字码）的结构。

构造链表的方法如下：将某一汉字输入编码，经过散列函数转换得到的每个散列地址建立一个链表。若无共址映照，则在基本表的指针域中填 0；否则，在链域中存储同一散列地址在溢出链表的头指针（head pointer），见图 9-12 所示。

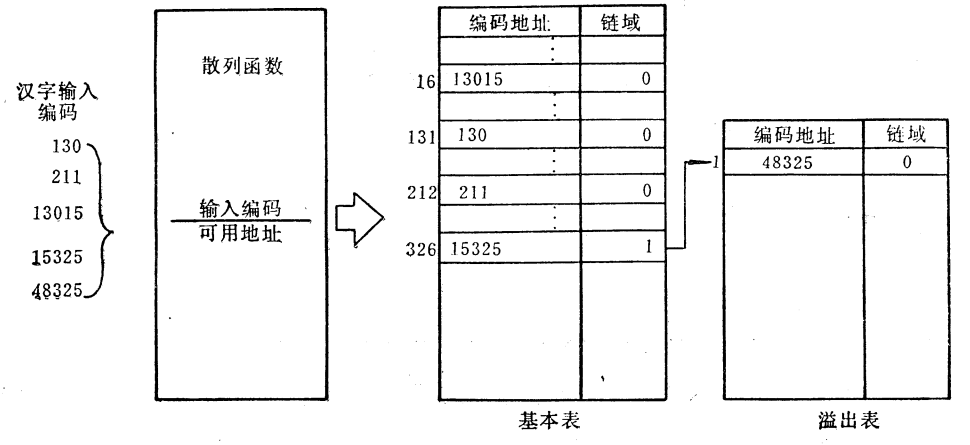


图9-12 用链表法解决共址代码的散列表

查找用散列法构成的表，先查基本表，若未找到输入汉字编码，再沿着链域查找溢出表。若溢出表中也无该汉字输入编码，则将此汉字输入编码存于溢出链表中。

散列法是一种直接计算存储地址的方法。当选择的散列函数能得到均匀地分布时，由于不必经过比较判断，因而比顺序法、对分法、索引法的效率要高。实际上，由于难免发生共址情况，所以查找效率又与解决共址问题的方法有关。

三、汉字输入编码转换程序流程

汉字输入编码和转换代码对照表的类型主要分为单表和复表两种。

(一) 单表

如果一个汉字信息处理系统只允许一种汉字编码输入（例如声韵部形编码），则要转换为国标交换码。这时对照表中只有两项，前一项为声韵部形编码，后一项为国标交换码，这种对照表称为单表，例如表 9-2 就是这种类型。

(二) 复表

如果一个汉字信息处理系统的使用允许两种以上的汉字编码输入方案，例如，既可以使用电报码，又可以使用声韵部形编码输入，要转换为国标交换码，为了节省内存单元，可将两个单表合并为一个复表，其中前两项为电报码和声韵部形码，后一项为国标交换码，这种复表不仅可以实现由电报码、声韵部形码转换为国标交换码，需要时还可以实现电报码转换为声韵部形码（见表9-4）。

表9-4 复表例子

汉字	电报码	声韵部形码	图标交换码
碧	4310	BISA	17-44
部	6752	BUEN	16-40
的	6104	D	21-36
导	1418	DXJU	21-28
发	4099	FAYY	23-02
服	2591	FUYB	23-94
纲	4854	GQSQ	24-57

汉字输入编码的转换过程实际上就是直接查找对照表、检索汉字代码的过程。它可以简单地用图9-13所示的程序流程图来说明。

9.3 汉字输出处理

就输出设备来说，有汉字显示器、针式打印机、激光印刷机，以及OFT（光学纤维管）汉字印刷机等。为了满足汉字字形的输出，有效地管理各类汉字输出设备，以便用户获得满意的输出效果，则需设计专门的汉字输出程序。

9.3.1 汉字字形输出

汉字字形输出包括汉字整字字形输出和以字根组合方式构成的汉字字形输出。

汉字字形的输出是将存储在汉字字模库中的相应字形信息取出，送到所指定的汉字输出设备上输出。

如前所述，汉字信息在机内加工完后，按机内码形式存放在缓冲区内。当用户需要

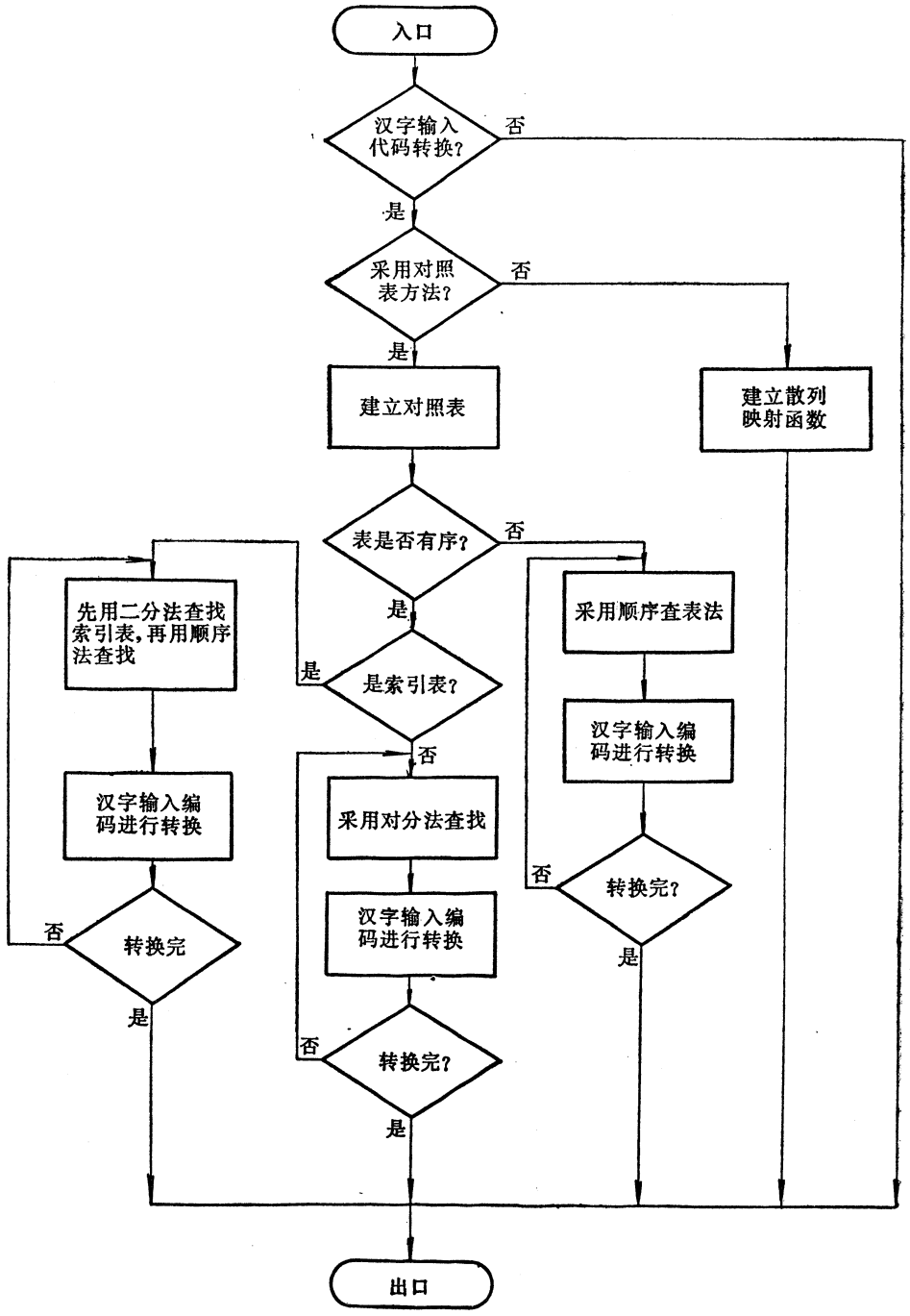


图9-13　汉字输入编码转换流程图

在指定的输出设备上输出汉字字形时，首先将机内码转换成汉字字模库中该汉字所对应的地址码，然后根据地址码从字模库中取出汉字字形信息并存入字形信息缓冲区。汉字输出程序根据输出设备的特点，将从字形信息缓冲区中取出的汉字字形信息进行某种特定的映照变换，并将变换后的数据送至相应输出设备的输出缓冲区，以便输出汉字。

9.3.2 汉字输出程序

汉字输出程序是汉字信息处理软件的必要组成部分。由于汉字字量大，不能象西文信息处理那样，设计几十个字锤就可解决西文字符的输出问题。因此，无论从输出设备和输出程序方面来看，汉字输出问题在技术上要比西文字符输出的难度大些。下面简单阐述汉字输出程序与操作系统输出控制的关系。

一、汉字输出程序与操作系统输出控制的关系

如同汉字输入程序那样，汉字输出程序也依其它所处的系统环境的不同，和操作系统输出控制的关系也不同。当汉字输出程序是操作系统资源时，汉字输出设备是作为系统设备定义的，它在操作系统调度模块管理下工作。按照系统设计原则和系统结构，它作为系统资源，被安置在某一确定的层上。按照约定，汉字输出程序和操作系统其它模块通过传递信息实现通信。当系统开工后，即可调用该程序运行。

若为了处理汉字信息而配置了相应的汉字输出设备，则视技术力量及各种因素和条件而定，在较全面地分析现行操作系统的情况下，可以把汉字输出设备定义为系统设备，并将自编的汉字输出程序作为系统模块插入操作系统的结构内。这样，汉字输出程序与操作系统的关系即如前所述；若把汉字输出设备定义为用户设备，则汉字输出程序不是操作系统的资源，且整个汉字信息的输出处理是在用户区内进行。如果汉字输出程序需要调用操作系统输出控制时，也可利用系统调用命令调用。

二、汉字输出程序的设计

由于汉字输出设备的种类繁多，汉字输出程序的计划方法也因不同的输出设备而异。对于不同类型的汉字信息输出，设计方法上也有所不同。本段着重介绍汉字字形的输出。

当用户程序请求输出汉字时，输出程序首先将用户程序提供的存放待输出的汉字数据的内存起始地址，内存区长度，输出数据类型等参数置入特定的存储单元中，然后输出程序转去查找汉字机内码与汉字地址码索引表，将指定内存区中的汉字机内码转换成相应汉字的地址码。转换完毕，输出程序根据汉字地址码及查找汉字字模库方法，转去访问汉字字模库，取出与汉字地址码相对应的汉字字形信息，同时将它存入相应的缓冲区。

至此，输出程序完成了汉字机内码到汉字字形数据的转换，并将转换后的数据置入相应的缓冲区。进而根据输出设备的具体特性，将缓冲区中的汉字字形数据转换成输出设备所规定的数据格式，以适应相应输出设备的特点。在输出过程中，汉字输出程序不断地测试输出设备所处的状态，协调主机和输出设备在时间上的差异。

输出完毕，汉字输出程序可将输出控制馈出的结果参数及其它有关信息置入约定的存储单元。

这里有一点必须提及，在大量输出汉字字形信息时，汉字输出程序为了解决内存紧张的矛盾，往往将汉字信息分组，一组一组的输出，而不是采用将内存区中的汉字机内

码全部转换成汉字地址码，再查找汉字字模库、取全部汉字地址码相对应的汉字字形信息的方法。

三、汉字输出程序流程图

图 9-14 所示为执行汉字输出程序的流程。

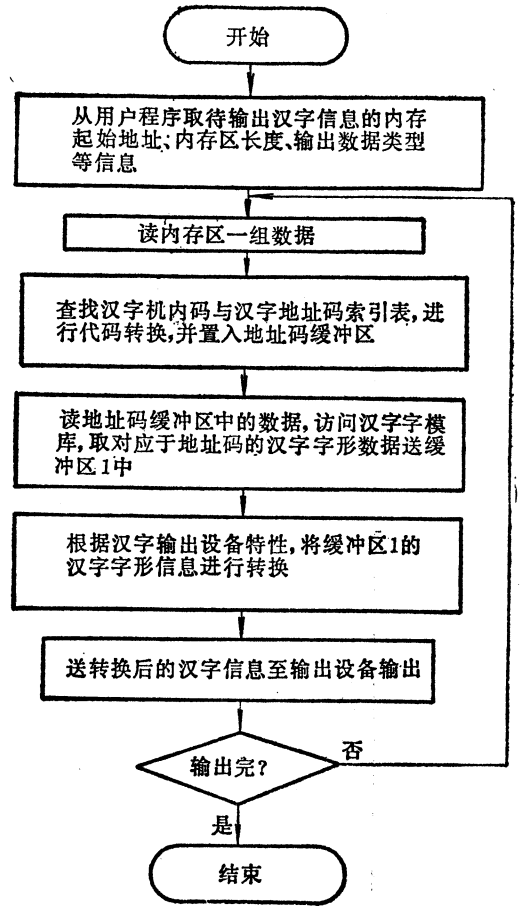


图9-14　汉字输出程序流程图

四、调用汉字输出程序的方法

由前节的论述可知，输出程序是汉字信息处理的主要程序之一，而且由于操作系统和该程序的关系不同，调用的方法也各有差别。

若输出程序是操作系统资源，则当系统启动时，该程序同时启动。调用该输出程序运行前，首先必须在指定的存储单元中置待输出汉字信息的内存区起始地址，内存区长度、输出汉字数据类型等参数。根据输出程序名，使用操作系统的相应广义指令即可调用该程序运行。调用返回时，该程序馈出结果参数。

若输出程序是操作系统支持下的应用程序，则操作系统启动后，可按输出程序名调用该程序运行。调用前必须在指定的存储区内记入该程序运行时所必需的参数信息。运行完毕，可根据用户程序要求或汉字信息处理管理程序的约定，输出程序在约定的存储区内写入结果参数。

五、介绍几种汉字输出程序

虽然各种不同汉字输出设备的输出程序在设计原则上并无多大差异，但由于各输出设备的特性、输出方式及其对输出数据格式的要求不同，对于某种输出设备必须配置适合的输出程序。以下叙述汉字字形的几种输出程序。

为叙述方便，假定汉字字模库中汉字是 24×24 点阵，并按点阵从上到下的顺序存放汉字字形数据。由于下面将详细介绍访问汉字字模库的方法，所以在本段中只从简叙述访问汉字字模库的过程。下面依次介绍汉字显示器输出程序，针式打印机输出程序、激光印刷机输出程序的基本设计思想。

(一) 汉字显示器输出程序

汉字显示器是最为常用的汉字输出设备之一，它的输出程序的主要功能是将缓冲区中的汉字字形数据转换成汉字显示器所需的数据格式，并将转换后的数据按地址送入刷新存储器，以达到显示汉字的目的。

由汉字显示器的硬件设计原理可知，它的显示缓冲区（即刷新存储器）的数据是按点阵行排列的。因而，本程序的主要任务是将汉字字形数据转换成按点阵行排列的格式。

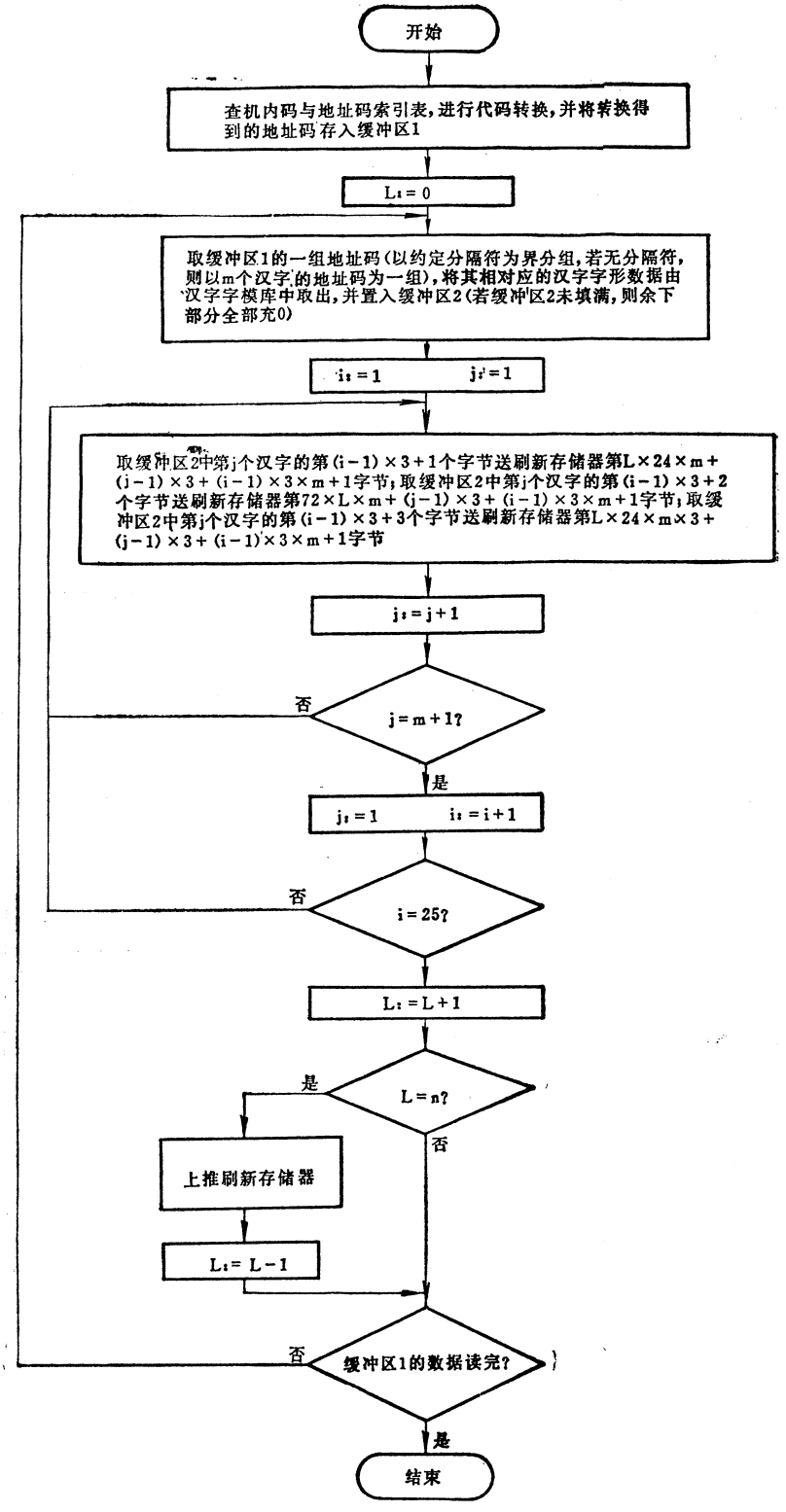


图9-15　汉字显示器输出程序流程图

为叙述方便，假定显示器一行所显示的汉字个数为m，一帧可显示n行汉字。

该汉字显示器输出程序的工作流程如图 9-15 所示。

（二）汉字针式打印机输出程序

由一般针式打印机的硬件设计原理可知，它用一列钢针按照机内提供的脉冲信息，逐列打印汉字的各列点阵，以构成汉字字形。因此，它与显示器输出的情况不同，信息是以点阵列方式组织的。该输出程序的主要任务是将取自汉字字模库中的字形数据转换成针式打印机所相容的数据格式，并输出到打印机缓冲区。在输出过程中，不断测试打印机所处的状态，协调主机与打印机在时间上的差异。

通常的针式打印机为了提高打印速度，都采用双向打印的方法，其程序的设计原理是相似的。由于篇幅限制，本段只作原理性介绍。

输出程序的流程可叙述如下：

（1）首先查找汉字机内码与地址码的索引表，将机内待输出缓冲区内的汉字机内码转换成相对应的汉字地址码，并置入缓冲区 1。

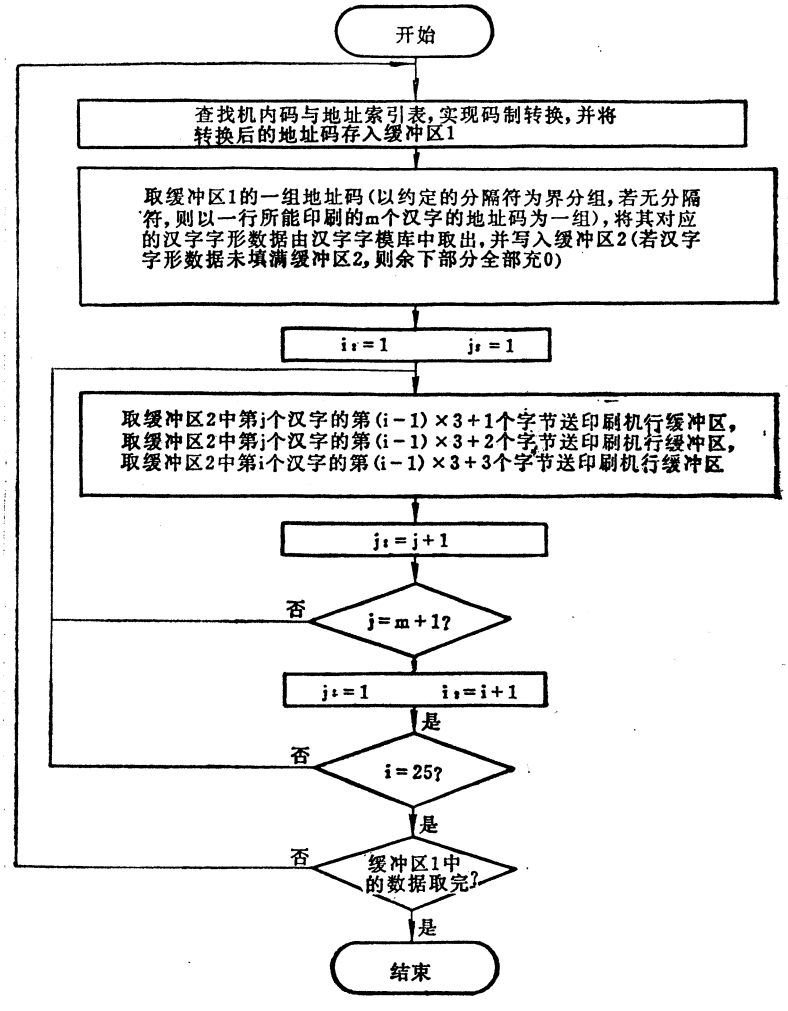


图9-16 激光印刷机输出程序流程图

（2）读一组汉字地址码（以约定分隔符为界分组，若无分隔符，则以一行所能打印的汉字个数为限），将其对应的汉字数据从汉字字模库中取出并置入缓冲区2。

（3）取缓冲区2中的每一个汉字字形数据按点阵列从左到右的顺序进行变换，并将变换后的数据置入打印机缓冲区，以供打印机打印出一行汉字。

（4）倘若待输出的汉字地址码已取完，则结束；否则转2。

（三）激光印刷机的输出程序

激光印刷机是一种印刷质量较好、输出速度高的汉字输出设备，特别适用于成批印制汉字文稿及报表。

由激光印刷机的一般设计原理可知，它是以行方式印字的。因此，激光印刷机输出程序除了实现上面两个输出程序所作的把机内码转换为地址码、并且取一组地址码所对应的字形数据存入缓冲区2外，还必须将缓冲区2中的数据转换成印刷机所需的特殊数据格式，并将转换后的汉字信息送入激光印刷机的行缓冲区。在整个输出过程中，该输出程序不断测试印刷机所处的状态。

其工作流程如图9-16所示。

9.3.3 访问汉字字模库程序

汉字字模库是存放汉字字形点阵信息的设备，也叫字形发生器。它是根据某个汉字的一组特定代码，产生该汉字的字形码点，供汉字印刷机印出汉字或由汉字显示器显示汉字，或进行其它汉字信息处理用。在汉字信息处理系统中，汉字输出程序模块要经常访问它。汉字字模库是汉字信息处理系统中一个必要的组成部分。汉字字模库中存储汉字的数据结构、类型，对于汉字的检索、系统的运行效率有着直接的影响。

一、介绍汉字字模库的几种存储方式

目前，汉字字形信息的存储，绝大多数采用数字存储方式，并可利用计算机存储器的先进存储技术。下面介绍几种以整字形式存放的汉字字模库存储技术。

根据节省查找汉字字模库的时间和节约存储空间的原则考虑，数字式汉字字形存储器的构成方式包括：1）只读方式（ROM方式）或可改写的只读方式（EPROM方式）；2）随机存储存取方式（RAM方式）。

采用ROM方式或EPROM方式，字形已固化，一般不能再更改，这种方式结构简单、方便，成本低，找字速度快。适用于字模库中常用字(国标一级字)的存放。随机存取方式，将汉字点阵信息存放在某种外存设备上，如磁带、磁盘、软盘等。在系统开始运行时，从外存储设备上将需用的汉字字形信息存放到随机存取存储器—RAM里，以备应用。这一操作，称为汉字字模信息的装填。这种方式具有后援存储容量大，字形容易修改，增删灵活的优点。但在遇到所需的字模不在RAM中，要临时访问外存时，会影响汉字输出速度。我国目前多数汉字系统采用两种方式结合的汉字库结构，以下讨论多级存储的汉字字模库。

在一个大的汉字系统中，如果要充分提高汉字字模库的利用率，则可以设置多级（例如三级）汉字字模库。最常用的汉字存放在ROM中，它的使用频度最高，各台终端都备有一级字模库。次常用字存于RAM，可根据使用情况改变存放内容，可设置在主机内，由各台终端所共享。第三级字存放在主机系统的磁盘存储器中，这样可以节省对

ROM、RAM组件的使用，使整个系统的成本降低。由于存放在外存储器中的汉字字模使用频度很低，虽然输出速度低，但对整个处理过程的影响并不大。

二、汉字字模库的地址编码

汉字字模库的地址编码是给出汉字字形信息存放在汉字字模库里的顺序号，因此，要求汉字字模库的地址编码是连续的。目前国内汉字输入编码方案已有上百种。大多数方案的编码值是不连续的，所以汉字输入编码一般不能直接做为汉字字模库的地址编码。

可以依据汉字的某些属性或借助于汉字输入编码构造一种函数，算出汉字在汉字字模库中的顺序号。例如，把汉字按其电报码的数值大小顺序排序，排序的序号就做为对应的汉字的地址编码号。

汉字	电报码	地址编码号
中	0022	172B
国	0948	053E
	……	

我们还可以将汉字按其在汉字使用频度表中出现的先后次序，按频度高低的顺序，给出汉字的顺序号。另外也可以按组成汉字的偏旁部首归类计算出地址编码等等。

总之按汉字的某些属性，找出一种计算方法或映照函数，算出它在汉字字模库中的顺序编号，作为汉字在字模库中的地址编码值。用H*表示汉字的某种属性值或汉字输入编码值，$F(X)$为映象函数，ND为汉字字模库中汉字的地址编码号。则有

$$F(H^*)=ND$$

这里要求$F(X)$为单值函数，ND为一组连续的自然数。

固化于ROM或EPROM中的汉字点阵信息的地址编码，是根据国标GB2312的次序排列的，从而使访问地址的计算与国标编码能直接对应。寻找比较方便。

三、访问汉字字模库的基本步骤

汉字是以机内码形式存储在计算机中的，从汉字机内码到检索汉字字形信息，只需把机内码用计算公式转换成存放汉字字形信息的实际始地址。而后从始地址开始取出汉字字形码点信息，送到输出缓冲区。如果是用汉字输入编码去访问汉字字模库，则必须查找各种索引目录表，先将输入编码转换成机内码，再用上述方法去访问汉字字模库。

以三级存储的汉字字模库为例，说明访问步骤。各汉字终端备有一级汉字字模库，用ROM固化，存放512个最常用汉字。终端内的随机汉字字模库用RAM存储器，可存放512个汉字。在主机上备有可存放1024个汉字的二级字模库。系统将使用的其余汉字存放在磁盘上。对于每一级字模库都建立汉字输入编码与机内码（或汉字顺序号）的目录索引表。

各级字模库地址编码序号分配如下：

0000～0511　ROM　存储器固定字模库的汉字顺序号
0512～1023　RAM　随机存储器字模库的汉字顺序号
} 一级字模库

1024～2047　二级字模库的汉字顺序号。

2048～N　　三级字模库的汉字顺序号。

汉字文件中如果其汉字机内码在0000～0511之间，则从ROM字模库中去取出汉字字形。如果机内码在0512～1023之间，则从RAM字模库中取出汉字字形。同样，若机

内码在1024～2047之间，则去访问二级字模库。若在2048～N之间，则去访问三级字模库，并取出相应的汉字字形信息。

在建立汉字文件时，对输入的汉字输入码首先查找ROM字模库的目录表。若有，则给出相应汉字的顺序号；若无，则查找RAM字模库目录表。若仍未找到，则查找二级字模库目录表；若再找不到，则最后查找三级字模库目录表，以取得汉字的相应顺序号。需要输出汉字字形时，按取得的汉字顺序号推算出存放该汉字字形信息的实际地址，取出字形码点，供显示或打印汉字用。访问汉字字模库的流程如图9-17所示。

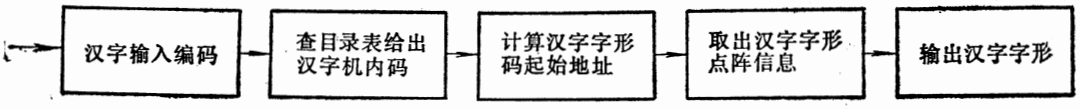


图9-17　访问汉字字模库流程图

四、查找汉字字模库的方法

前面我们谈到汉字字模库都包括有目录表部分。目录表由若干目录项组成。每个目录项包括汉字的输入编码和对应的汉字顺序号。从某种意义上讲，查找汉字字模库就是查找目录表。这里仅举两种不同的查找法。

（一）对分查找法

采用对分查找法，可以提高检索速度。这种查找方法和前述把汉字输入编码转换成国标码或内部码的方法相同，不再重复介绍。

采用这种查找方法，其速度比采用顺序查找法快得多。例如查找一个含有八千汉字的字模库，查找次数不大于13次。

当汉字字模库增删汉字、汉字输入编码、机内码更新时，目录表必须重新排序组织。

（二）分级查找法

按汉字输入编码的数字字母大小顺序建立总目录表，并划分为若干区（设为L个区）。取出每个区中的第一个目录项，组成第二级目录表。我们称为B_2表，共有L个目录项。再将总目录表划分出的若干分区分成若干小区（设有k个）。依上述方法，分别取出k个小区的第一个目录项，组成L个第三级目录表。每个目录表含有k个目录项。这些目录表用B_3^i表示（$i=1,2,\cdots,L$）。仿此，我们可以建立第4级、第5级，……目录表。总目录表分级的个数依字模库的大小和需要而定。

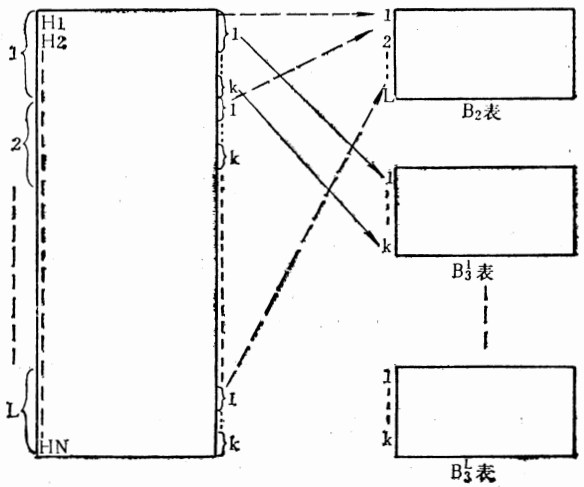


图9-18　目录表分级示意图

目录表的分级如图9-18所示。

为了查找和制表处理的方便，对于B_3^k表，若其目录项少于k个，则用空目录项使表填满。

检索多级目录表的方法是：首先在第二级目录表（B_2）中查找。若在B_2中就查到所需查找的检索项，则取出此目录项中的后一部分的汉字顺序号。若不等，则判定检索的码值究竟将落在目录表的哪两个目录项之间，以确定下一步要在第几个三级目录表中查找。例如：检索项的码值大于第二级目录表的第j个目录项的输入编码值，小于第j＋1个目录项的输入编码值，则确定下次将查找目录表B_3^j。若大于最后一个目录项的输入编码值，则去查找目录表B_3^k。按照上述查找方法一直可查到最高一级目录表，而后再据查表结果确定顺序查找总目录表的某一小区域，直至查到含有与检索项同码的目录项。若不在这个小区域内，则表示此汉字不在此汉字字模库内。

五、举例

下面我们给出一个综合使用上节所介绍方法的例子。

假设汉字字模库为二级存储。汉字点阵为24×24。汉字总数为7000个。其中最常用的512个汉字存放在固定字模库（ROM存储器）中。RAM为随机字模库，存放512个汉字字形信息。其余汉字字模存放在软盘中。汉字输入编码用hd表示。hd_i（$i=1,2,\cdots,7000$）为汉字字模库里第i个汉字的输入编码。并且设：

$$hd_1<hd_2<\cdots<hd_i<hd_i+1<\cdots<hd7000$$

ROM汉字字模库中汉字的顺序号为0000～0511；RAM汉字字模库中汉字的顺序号为0512～1023。二级汉字字模库的汉字顺序号为0000～7000。二级汉字字模库里的汉字在计算机里实际顺序号为1024＋N。N为二级汉字字模库的汉字顺序号。

为了说明前面介绍的各种方法。我们设ROM字模库存放的512个汉字按其在汉字频度表中出现的先后次序排序。即顺序号为0000的汉字使用频度最高，0511号汉字使用频度最低（仅对这512个汉字而言）。查找ROM汉字字模库用查表法。RAM字模库采用随机方式建立，查找方法为顺序查找法。ROM字模库选用的512个汉字输入编码为：

$hd'_1, hd'_2, \cdots, hd'_{512}$。$hd'_j \in \{hd_i\}_{j=1,2,\cdots,512}^{i=1,2,\cdots,7000}$

并且有　$hd'_j<hd'_{j+1}$　　$j=1,\cdots,511$

首先建立目录表B1，如图9-19所示。

目录表地址	汉字输入编码	顺序号
k_0+1	hd'_1	n'_0
k_0+2	hd'_2	n'_1
— —	— — — — —	— — —
k_0+512	hd'_{512}	n'_{511}

$0\leqslant n'_i\leqslant 511; i=0,1,\cdots,511$

$n'_i\neq n'_j$ 当$i\neq j$时

$hd'_i<hd'_{i+1} \; i=0,1,\cdots,510$

图9-19　ROM汉字字模库总目录表（B_1表）

将B_1平分为八个区，并取出每一个区的第一个目录项，建立二级目录表B_2。如图9-20所示。

目录表地址	汉字输入编码	顺序号
kk_0+1	hd'_1	n'_0
kk_0+2	hd'_{65}	n'_{64}
—— —— ——	—— —— —— —— ——	—— —— ——
kk_0+8	hd'_{499}	n'_{448}

图9-20　ROM汉字字模库二级目录表（B_2表）

分别把已划分的八个区再细划分为八个小区。取出各个小区的第一个目录项，以及此项在总目录表中的存放始地址，并建立三级目录表，用B_3^j表示（j＝1，2，…，8）。如图9-21所示。

汉字输入编码	顺序号	在总目录表中始地址
hd'_1	n'_0	k+1
hd'_8	n'_8	k+9
hd'_{57}	n'_{56}	k+57

（B_3^1表）

汉字输入编码	顺序号	在总目录表中始地址

（B_3^8表）

图9-21　第三级目录表

随机汉字字模库（RAM存储器）因为其内容经常更新，不宜采用查表法和对分法，所以采用顺序查找法。并建立随机汉字字模库目录表，如图9-22所示。

汉字输入编码	顺序号
hd''_1	n'_{512}
hd''_2	n'_{513}
hd''_{512}	n'_{1023}

$hd''_j \in \{hd_i\}$
$i=1,2,\cdots,7000$
$j=1,2,\cdots,512$

图9-22　随机汉字字模库目录表

汉字输入编码	顺序号
hd_1	0000
hd_2	0001
hd_{7000}	6999

$hd_j < hd_{j+1}$
$1 \leqslant j \leqslant 6998$

图9-23　二级字模库目录表

二级汉字字模库也按汉字输入编码的数字字母顺序排列，排列的序号作为对应的汉字顺序号。按顺序号从小到大把相应的汉字字形信息依次存放在软盘上。二级汉字字模库目录表如图9-23所示。

查找二级汉字字模库，我们采用对分法查找。下面具体讨论查找汉字字模库的过程。

1）从键盘输入汉字输入编码 hd。第一步查找固定字模库（ROM 字模库）。

（1）取 hd 与B_2表中的目录项的前一部分比较。若相等（$Hd=hd'_j$），则从 B_2 表中取出相应的汉字顺序号（n'_j-1）；若不等，则判定 Hd 的值界于B_2表中哪两项的值之间。若 $hd'_{64(i-1)+1}<hd'_{64i+1}$，$i=1,2,\cdots,7$，则确定下一步查找 B_3^i 表。若 $hd>hd'_{64\times7+1}$，则查B_3^8表。

（2）查找B_3^i表方法与查找 B_2 表的方法一样。若 hd 等于表中的某一项的前一部分，取出相应汉字顺序号。若不等，则通过 hd 与B_3^i表中前后两连续项的前一部分的比较，确定 Hd 的值将落在哪个小区内。再从目录表项目后一部分取出小区在总目录表的始地址，从此始地址开始，逐项取出小区内的目录项的前一部分与Hd比较。如果小区内含有与Hd等值的项，则取出相应汉字顺序号，若查不到，则转去查找随机字模库。

由汉字顺序号n'_j按下列公式算出存放汉字字形信息的实际始地址。

$$\text{实际始地址}=D^*+72\times n'_j \tag{9.1}$$

式中，D*为 ROM 字模库在内存中编址的始地址号。

最后从算出的实际始地址开始取出汉字字形信息。

2）顺序查找随机汉字字模库目录表。如果某汉字在这一字模库内，给出相应顺序号n'_k，则按式（9.2）计算出存放汉字字形信息的实际始地址。如果不在此字模库，则转去访问二级字模库。

$$\text{实际始地址}=D^{**}+72\times(n'_k-512) \tag{9.2}$$

式中，D**为随机字模库在计算机内的始地址编号。

按实际始地址取出汉字字形信息。

3）查找二级字模库目录表用对分查找法。对分查找法的主要计算公式如下：

$$X=\left[\frac{\text{低项码}+\text{高项码}}{2}\right] \tag{9.3}$$

符号X代表对分点码值，为低项码与高项码之和的平均值的整数部分。X_0为目录表中对应于第X个目录项的前一部分。即对应有顺序号X的汉字的输入编码。

先以0000作为低项码值，6999为高项码值代入公式。比较 Hd 与 X_0 的值。若 $X_0=Hd$，则给出 Hd 的汉字顺序号 $n=X_0$ 若 $X_0\neq Hd$，则可能有下述两种情况：（1）若 $X_0>Hd$，则以X做为高项码、而低项码不变进行计算，得出新的对分点码值 X′，重新比较 X'_0 与Hd。（2）若 $X<Hd$，以 X作为低项码，而高项码不变进行计算，算出新的对分点码值 X′，并得到相应的 X'_0，比较 X'_0与 Hd。这样反复地进行计算和比较，直至某一个 $X_0^i=Hd$，或进行若干次计算和比较后，有 $X^i=$新的低项码值（或 $X^i=$新的高项码值），但对应的 $X_0^i\neq Hd$，表示字模库中无此汉字。

由汉字顺序号N，按下列公式计算存放汉字顺序号为N的汉字实际盘地址。

$$(N-1)\times72/128=N_1\cdots\cdots\text{余数 }1 \tag{9.4}$$

$$N_1/52=N_2\cdots\cdots\text{余数 }2 \tag{9.5}$$

N_2+6+1 为所求磁道号，余数 2 + 1 为扇区号，余数 1 + 1 为字节数。所以，存

放此汉字字形信息的实际盘始地址为：第N_2＋6＋1磁道，第（余数2）＋1扇段，第（余数1）＋1个单元（字节）。

这里作为二级汉字字模库用的软盘，其规格是8英寸单面双密度盘，其具体规格为：128字节/扇段，52扇段/磁道，77道/片。0～6磁道用来存放各种索引目录表。汉字字模从第7道，扇段1开始存放。

9.4 扩充的汉字信息处理程序

9.4.1 编制汉字信息典程序

在信息检索系统或处理中心，汉字库中最好能包含汉字的属性信息。例如，要求系统能将各种汉字编码转换为电报码或国标信息交换码；要求系统能接收所选择的输入某种编码方案的汉字码；要求计算机能帮助人工输入汉字（提供汉字索引信息），从一个汉字联想到另外的汉字等等。这些功能不仅要用到汉字的字形信息，而且要用到汉字的属性信息，现介绍如下。

一、汉字信息典的类型

汉字的信息典主要应该包括下面这几方面的内容：

（1）国家标准汉字信息交换码和汉字输入编码（例如电报码，四角号码等）的对照表。

（2）汉字的读音（例如汉语拼音）。

（3）汉字的字形特征（例如：偏旁部首；字形；文字代码；部首外的画数；总画数；笔顺；结构类型等）。

（4）其它如综合频度、排序特性等。

利用这些属性信息，可以进行汉字编码和各种字典的辅助设计；可以帮助人们检索和输入汉字，从而提高汉字的输入速度；可以从各个不同的方面检索汉字。例如：按总笔画顺序检索；按部首笔画顺序检索；按四角号码检索；按汉语拼音检索等等。这样就可加强汉字信息库的利用率。

二、汉字属性在计算机内的存储方法

如果把汉字库的概念加以扩大，即不但存储汉字的字形数据，而且存储有关各汉字间的相互关系，则形成汉字信息库。例如，主题词库就可以进一步增强机器的扩检和缩检功能，更加充分地发挥机器的效率。

汉字的主题词典展示了词间的同义概念、上位概念、下位概念和相关概念等词间关系。现举例说明如下：

（一）同义概念关系（用：Y；代：D）

这就是词与词之间为同义的关系。例如："电子计算机"与"电脑"为同义主题词，若规定"电子计算机"为标准主题词，则

用（Y）：电子计算机代（D）：电脑

（二）上位概念关系（属：S）

这就是对某一词为其上位概念的关系。例如："电子计算机"为"数字计算机"的上位主题词，则

数字计算机　　属（S）：电子计算机

（三）下位概念关系（分：F）

这就是对某词为其下位概念的关系。例如："数字计算机"，"模拟计算机"为"电子计算机"的下位主题词，则

电子计算机　　分（F）：数字计算机

　　分（F）：模拟计算机

（四）参照概念关系（参：C）

这就是词与词之间为参照的关系。例如："数据检索"与"情报检索"为参照关系主题词，则

数据检索　　参（C）：情报检索

在电子计算机中判别一个数据，必须有类型（TA）、长度（LNG）、位置（POS）三个要素。若将此记为目录（T），则$T=$(TA, LNG, POS)，目录与数据（D）构成记

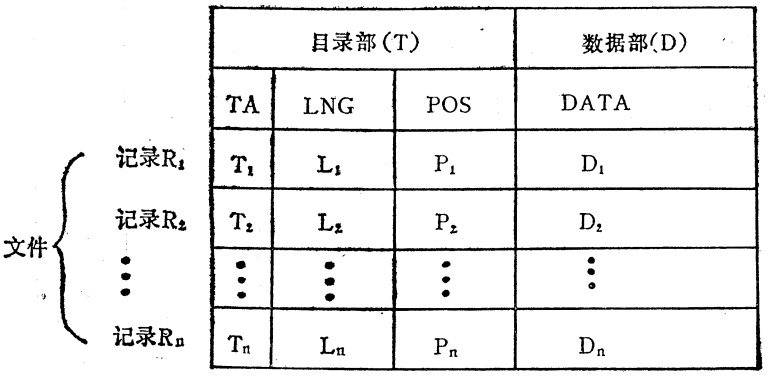


图9-24　文件及记录的描述

录（R），即$R=(T, D)$，若干记录的集合则构成文件（图9-24）。

现在，我们来研究主题词典在计算机内的存储方法。

设"电脑"和"电子计算机"为同义概念关系，"电子工程"是它们的上位概念关系，"磁盘"为"电脑"的下位概念关系，"模拟计算机"和"数字计算机"为"电子计算机"的下位概念关系，它们之间为同义概念关系，则可用树型结构表示。如图9-25所示。

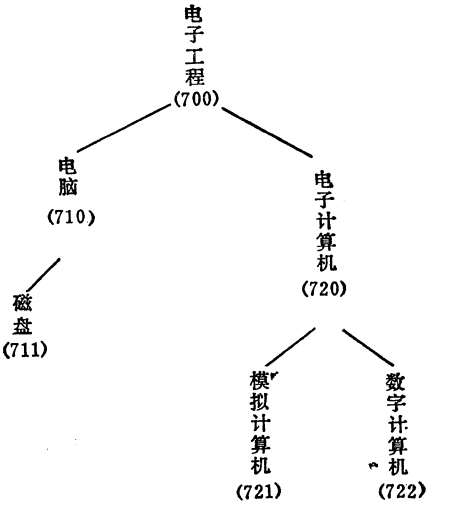


图9-25　主题词间的树型结构表示

为了使词间的关系便于在计算机内存储，我们可以把目录部扩展为两部分。一部分用以记述词号、长度、词的始地址；另一部分用以记述词间关系（上位、同位、下位）的始地址。若地址用三位数表示，数据部分的汉字用2个字节的代码表示，字间用链指针构成词，则词或短语可以用如下方式

记录

目录(扩展)

目录地址	数据目录表			词间的关系位置		
	TA	LNG	POS	POS(上位)	POS(同位)	POS(下位)
001	700	08	104	000	000	018
018	710	04	104	001	035	052
035	720	10	104	001	018	069
052	711	04	225	018	000	000
069	721	10	258	035	086	000
086	722	10	291	035	069	000

数据

数据地址	汉字	第一链	第二链
104	电	115	159
115	子	126	181
126	工	137	
137	程	*	
148	700		
159	脑	*	
170	710		
181	计	192	
192	算	203	
203	机	*	
214	720		
225	磁	236	
236	盘	*	
247	711		
258	模	269	
269	拟	181	*
280	721		
291	数	302	
302	字	181	*
313	722		

图9-26　主题词在计算机内的存储

存储：

汉字始地址	汉字代码	下一字地址	次下一字地址

若同一汉字可以连成几个词，则可以在链域中用几个指针指示。在相应的链域出现“*␣␣”时，则表示该词到此为止，从而给出相应的词号。由于主题词的长短不一，因而记录为可变长，所以，可用处理可变长的方法来处理。图9.25所示的主题词的存储结构如图9-26所示。

有了主题词典，可以提高汉字情报检索的查全率和查准率；利用主题词之间的关系(用、代、属、分、参)，还可以进行扩检和缩检。

三、建立汉字信息典的软件实现

汉字信息典收录项目应该由国家统一制订，这里以某个汉字信息处理系统为例，说明汉字信息典的结构内容：

（1）汉字的字模信息（128字节）；

（2）电报码（4字节）；

（3）四角号码；

用五位数字表示一个汉字，前四位数字表示汉字四角的笔画特征，最后一位数字用来区别同码字。对于无同码的字，末位数字记0；对于有同码的汉字，按频度高低，第一个同码字末位数字记作1，第2个同码字末位数字记作2，以此类推。

（4）汉语拼音（7字节）；

其中声母（2字节），韵母（4字节），声调（1字节）

（5）部首代码（2字节）；

（6）总笔画代码（2字节）；

（7）常用编码（4字节）；

（8）结构类型（1字节）。

汉字分独体字（I）和合体字两类。在合体字中，又分上下结构（V）、左右结构（H）、包含结构（S）三种。

于是，对每个汉字可采用如下记录格式：

汉字字模信息	电报码	四角号码	汉语拼音	部首代码	总笔画	常用编码	结构类型

有了这些汉字属性信息，可以建立汉字信息库，使用参数指定，用计算机系统的分类排序语句，可以很容易地编制各种索引表（图9-27）。

如果我们根据卡片或控制台指定的分类键按增序或减序来分类，则编制各种分类索引表的处理流程如图9-28所示。

9.4.2 汉字字形旋转程序

编制汉字字形旋转程序可从坐标系的旋转算法中得出。由于这个程序较复杂、执行时间也较长，因此，除非特殊需要的场合，一般是不配备这种程序的。

然而，作为字形旋转的特例，即字形旋转90°，形成汉字竖向排列，则在汉字系统中

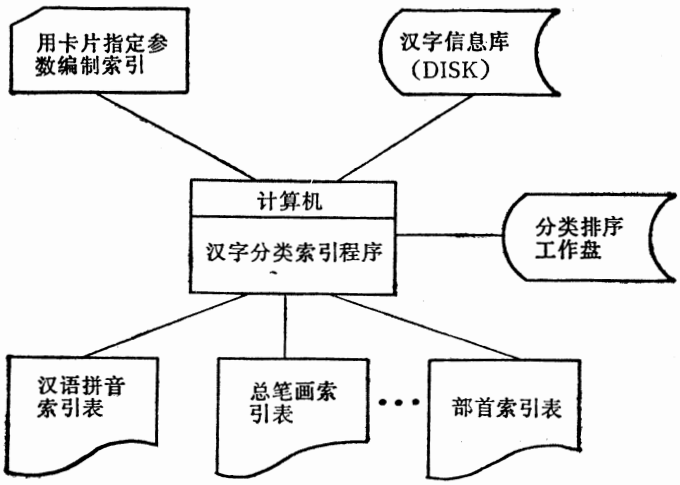


图9-27　根据汉字属性信息编制各种索引表

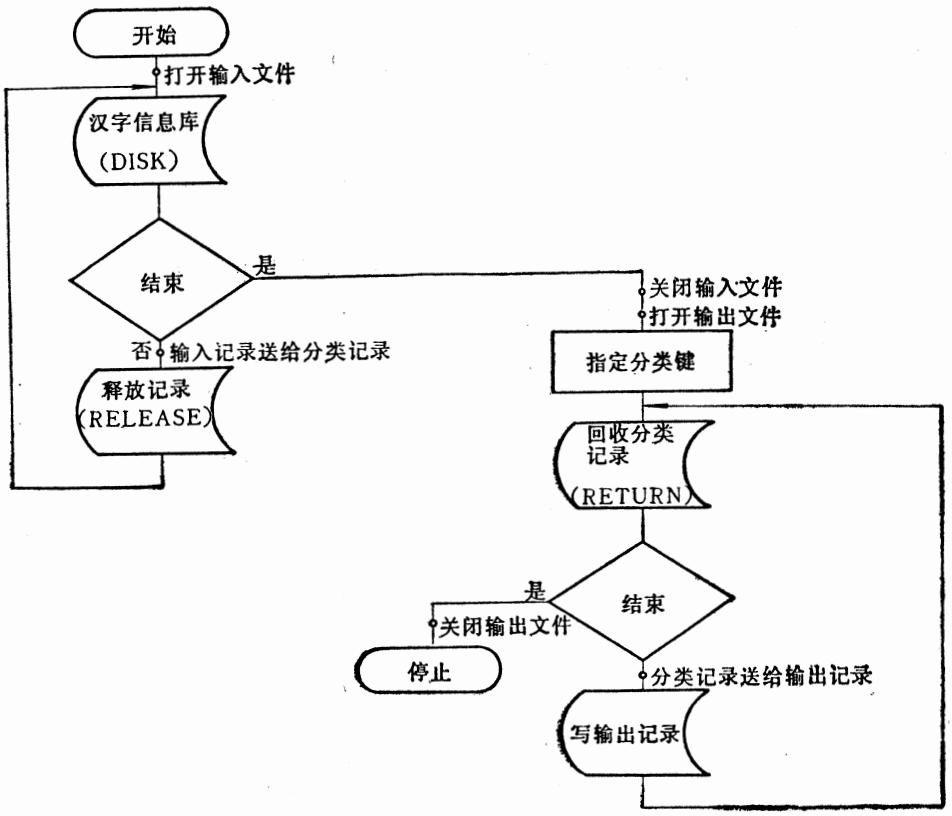


图9-28　编制汉字分类索引程序的流程图

却颇为重要。

从汉字字形库中汉字点阵的结构来看，只需要简单地更换点阵输出的次序，即可实现汉字从横向打印转换成竖向打印（图9-29）。

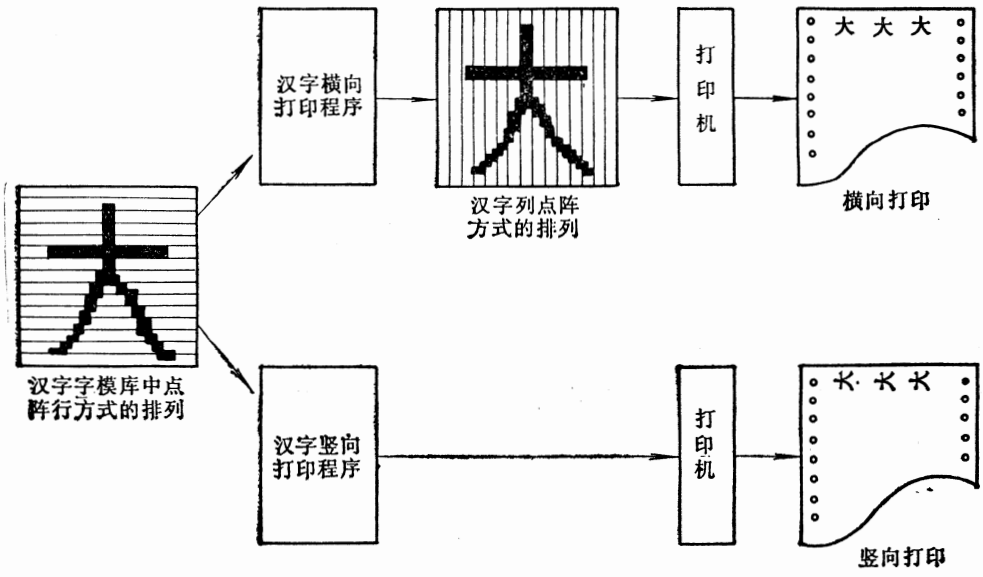


图9-29　汉字横向打印及竖向打印的原理

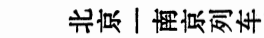
由于实现方法比较简单，程序就不在此列出。此外，随着印刷机智能化的进展，一些汉字印刷机本身具有横向、竖向两种打印方式的切换能力。主机的打印输出程序可仍按横向方式取点阵，由印刷机的控制器转换成竖向方式的点阵，即可实现竖向打印。

必须指出，在竖向打印中，若对全部文字作竖向处理，就会在某些文句的打印中，导致不正常现象。例如：

横向打印：

北京——南京列车时刻表

若用此法作竖向打印时，就会出现：



这里的连接号也随之作了竖向变换，出现了与应用要求不符合的结果，对此，必须在程序中作出相应的处理。

9.4.3　汉字尺寸变倍程序

一、问题的提出

在汉字印刷业务中，为了使文章层次分明，对文章的各层标题等，常采用比正文字号大的汉字；而对脚注以及提供给读者参考的部分，常采用比正文字号小的汉字。为了满足这种要求，对同一个汉字必须准备许多不同字号的铅字模。由于汉字的数量很大，每个汉字又有多种不同的字号，因而需要准备的铅字模的数量就特别多。

在以电子计算机为中心的通用型汉字信息处理系统中，汉字字模常采用数字式存储。由于字模点阵中点的数量和位置都是固定的，要表示比基本字模点阵大（或小）的文字，

必须准备相应的汉字字模点阵以及使用相应的硬设备。

利用存储设备中存储的某种规格的字模点阵（称之为基本字模），通过输出设备输出几种不同字号的汉字，这是汉字信息处理业务中需要解决的一个特殊问题。对某些输出设备来说，改变一下输出的扫描行数就可以解决汉字的变倍问题，而对另一些输出设备（例如，汉字点阵显示器，汉字针式打印机）来说，则必须改变字形点阵中的位点数目、才能解决汉字字形的放大或缩小问题。

这就是说，实现汉字字形的放大或缩小，可采用硬件和软件两种方法。由于硬件方法对汉字点阵中的位点进行变换，需要复杂的电路，故成本较大；而采用软件方法，则可以克服这个缺点。

本节试图介绍利用一种基本字模，可以产生任意规格点阵字模的软件方法。重点介绍几种软件放大的方法，供读者参考。而对缩小算法感兴趣的读者可以去看有关的参考书。

二、汉字字形的放大原理

汉字的字形放大，是利用汉字图形的相似变换方法。汉字字形的放大原理，可简单地用图9-30来表示。

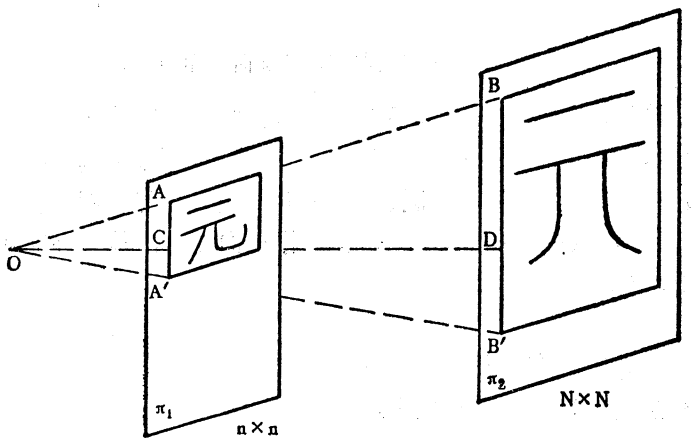


图9-30　汉字字形的放大原理

如图所示，过O点作平面 π_1 和 π_2 的垂线，垂足为C、D。则显然有

$\triangle ODB \approx \triangle OCA$，于是 $BD/AC=OD/OC$

设 $OD/OC=\alpha$，则 $BD=AC\times\alpha$

同理 $DB'=CA'\times\alpha$

所以 $BD+DB'=(AC+CA')\times\alpha$，即 $BB'=AA'\times\alpha$

或 $BB'/AA'=\alpha$。

于是，α 可理解为放大系数。

三、汉字尺寸的变倍方法

（一）逻辑方程插入法

设基本字模的点阵A为 $n\times n$ 阶方阵，变换后的字形点阵B为 $N\times N$ 阶方阵，取放

大系数$\alpha=N/n$，则由方阵A变换为方阵B，其下标变换是将A的下标乘以α再取整而得到B的下标，其数值（0，1）变换是根据A中的点及其相邻点的值，用一组简单的逻辑方程插入，而得到B中相应点的值。在$1<\alpha\leqslant 2$的情况下，可以使用此法对字形放大。

先把$n\times n$内的点$(k,\ l)$和$N\times N$内的点$([\alpha k],\ [\alpha l])$相对应（其中方括号[ ]是取整记号）。对于$N\times N$内和$n\times n$内那些不相对应的点，最简单的方法是取和邻近点（例如“列”的下邻，“行”的右邻）相同的值。但这样做会使汉字中的某些笔画变粗，或使倾斜笔画的联系错乱，因而使放大后的字形失真。

因此，要考虑这一点周围的信息，根据其周围点的值来决定该点的值（1或0）。

例如，取基本字形点阵A为24×24阶方阵，放大变换后的字形点阵B为32×32阶方阵，放大系数$\alpha=32/24=4/3$，由A变换为B，先把24×24点阵中的各点$(k,\ l)$，变换为32×32点阵中相对应的点$\left(\left[\frac{4}{3}k\right],\ \left[\frac{4}{3}l\right]\right)$。设$32\times 32$点阵中与$24\times 24$点阵中不相对应的点为（i，j）（见图9-31）。其值（0或1）用A（i，j）表示。

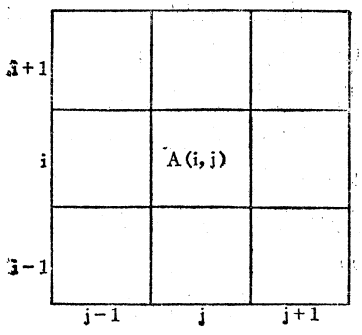


图9-31　点$(i,\ j)$与周围点的关系

为了保证汉字笔画之间的连结，若以下八个式子中有一个以上的等式成立，则取A(i，j)=1。

$$A(i+1,\ j)\cdot A(i-1,\ j)=1 \quad (9.6)$$

$$A(i+1,\ j-1)\cdot A(i-1,\ j)=1 \quad (9.7)$$

$$A(i+1,\ j+1)\cdot A(i-1,\ j)=1 \quad (9.8)$$

$$A(i,\ j+1)\cdot A(i,\ j-1)=1 \quad (9.9)$$

$$A(i-1,\ j+1)\cdot A(i,\ j-1)=1 \quad (9.10)$$

$$A(i+1,\ j+1)\cdot A(i,\ j-1)=1 \quad (9.11)$$

$$A(i-1,\ j-1)\cdot A(i+1,\ j+1)=1 \quad (9.12)$$

$$A(i-1,\ j+1)\cdot A(i+1,\ j-1)=1 \quad (9.13)$$

式（9.6）按照通常矩阵行与列的指定，考虑点$(i,\ j)$的上邻和下邻。若两者都为黑（即值为1），就把点$(i,\ j)$定为黑，取$A(i,\ j)=1$否则，取$A(i,\ j)=0$。

式(9.7)和(9.8)是关于点$(i,\ j)$的下邻和上邻的左右点定为黑的条件（纵向连结）；式(9.9)、(9.10)和(9.11)是关于点$(i,\ j)$的左邻和右邻的上、下点定为黑的条件（横向连结）；若点的上下及左右都不存在与$n\times n$内相对应的点，就采用式(9.12)和(9.13)斜组合的条件。

上述运算操作可以从行或列的单侧（例如编号小的一方）进行，也可以从行和列两个方向同时进行。

（二）曲面插入法

使用2变量内插（即曲面插入）的方法，可以进行矩阵的行、列元素数的任意倍数的变换。

设基本字模点阵的阶数为$I\times J$，$z=f(x,\ y)$是给出基本字模的函数，则用

x，y，z三维空间（其中z对应着0或1）内的数据集合F：

$$F=\{(x,\ y,\ f(x,y))|f(x,\ y)=0,\ 1;\ x=1,\ 2,\ \cdots,\ I;\ y=1,\ 2,\ \cdots,\ J\} \tag{9.14}$$

若把基本字模F变换成阶数为$K\times L$的字模（数据集合）G，则用下式表示：

$$G=\{(x_i,\ y_j,\ f(x_i,\ y_j))|x_i=x_1,\ x_2,\ \cdots,\ x_K;\ y_j=y_1,\ y_2,\ \cdots,\ y_L\} \tag{9.15}$$

式中

$$x_i=1+\frac{I-1}{K-1}(i-1)(i=1,\ 2,\ \cdots,\ K)$$

$$y_j=1+\frac{J-1}{L-1}(j-1)(j=1,\ 2,\ \cdots,\ L)$$

字模点阵的阶数变换是由数据集合F生成数据集合G而实现的。

由式（9.14）表示的$I\times J$个数据点，$1\leqslant x_i\leqslant I$，$1\leqslant y_j\leqslant J$，需要计算取样点$(x_i,\ y_j)$的函数值$f\ (x_i,\ x_j)$，换句话说，必须把式（9.15）的离散值化为模拟值，为此，可采用简单的z变量内插（曲面插入）的方法来实现。

取$k_i=[x_i]$，$l_j=[y_j]$（方括号[ ]是取整记号），则包围任一取样点$(x_i,\ y_j)$的四个数点是：$(k_i,\ l_j)$，$(k_i,\ l_{j+1})$，$(k_{i+1},\ l_j)$，$(k_{i+1},\ l_{j+1})$。由式（9.14）得到各点的函数值为$z_1=f(k_i,\ l_j)$，$z_2=f(k_i,\ l_{j+1})$，$z_3=f(k_{i+1},l_j)$，$z_4=f(k_{i+1},\ l_{j+1})$。把上述四点用三维坐标表示如下：$(k_i,\ l_j,\ z_1)$，$(k_i,\ l_{j+1},\ z_2)$，$(k_{i+1},\ l_j,\ z_3)$，$(k_{i+1},\ l_{j+1},z_4)$。

经过解这四个数据点的最低次曲面方程式，可用式（9.16）来内插近似计算取样点z的函数值。

$$f(x_i,\ y_j)=C_1x_iy_j+C_2x_i+C_3y_j+C_4 \tag{9.16}$$

其中

$$C_1=z_1-z_2-z_3+z_4$$

$$C_2=l_j(z_2-z_4)+(l_{j+1})(z_3-z_1)$$

$$C_3=k_i(z_3-z_4)+(k_{i+1})(z_2-z_1)$$

$$C_4=z_1(l_{j+1})(k_{i+1})-z_2l_j(k_{i+1})-z_3(l_{j+1})k_i+z_4l_jk_i$$

由式（9.16）得到的$f\ (x_i,\ y_j)$的值是区间〔0，1〕上的实数，所以根据阈值τ（$0<\tau<1$）可以进行二值化处理，即：

若$f(x_i,\ y_j)\geqslant\tau$，　　则$f(x_i,\ y_j)=1$

若$f(x_i,\ y_j)<\tau$，　　则$f(x_i,\ y_j)=0$

根据τ的选择方法，可以得到字模的线幅（笔画）变化。

根据式（9.16）得到的阶数变换后的字模G，其文字质量已可供实用。为了提供更高质量的字模，对于直线段的端部（纵线段和横线段）以及交叉处产生的变形，可采用与值变换技术进行整形处理。

（三）公式变换法

设基本字模点阵A的次数是一个$n\times n$阶的矩阵，$A\ (i,\ j)$是给出基本字模的函数，则用i，j，$A(i,\ j)$空间内的数据集合A表示如下：

$$A=\{(i,\ j,\ A(i,\ j))|A(i,\ j)=0,\ 1;\ i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ n\}$$

若把基本字模A变换为阶数为$m\times m$阶的方阵B，则可用下式表示：

$$B=\{(K,\ L,\ B(K,\ L))\mid B(K,\ L)=0,\ 1;\ K=1,\ 2,\ \cdots,\ m;$$
$$L=1,\ 2,\ \cdots,\ m\}$$

此外，若用$〔x〕$表示x的整数部分，$\langle x\rangle y$表示x/y的余数部分，则对放大变换算法可作如下讨论：

1．放大算法（$N<M$）

令$〔NK/M〕=I$

$〔NL/M〕=J$

$A(I,\ J)=A_1$

$A(I+1,\ J)=A_2$

$A(I,\ J+1)=A_3$,

$A(I+1,\ J+1)=A_4$

$FK=MIN(N,\ M-\langle NK\rangle M)$

$FL=MIN(N,\ M-\langle NL\rangle M)$

使

$$B(K,\ L)=FK\cdot FL\cdot A_1+(N-FK)\cdot FL\cdot A_2+FK\cdot(N-FL)\cdot A_3$$
$$+(N-FK)(N-FL)\cdot A_4 \tag{9.17}$$

于是得到了变换后的字形点阵$B(K,\ L)$相应小方格的值。由于$A_i(i=1,\ 2,\ 3,\ 4)=0$或1，显然有$0\leqslant B(K,\ L)\leqslant N^2$成立，即$0\leqslant B(K,\ L)/N^2\leqslant 1$成立。这样，$m\times m$阶矩阵$\{B(K,\ L)/N^2\}$的所有元素值都在0与1之间，选取适当的阈值$\tau$（$0\leqslant\tau\leqslant 1$），就得到了放大后的字形矩阵$\{B(K,\ L)\}$。

2．整形判别及修正量的计算

在简易文字处理系统中，直接使用上述变换方法所得到的字形，已可供实用。但在对字形质量要求较高的情况下，由于直线段的端部和交叉处有可能产生变形，所以仍需进行整形处理。

这种变形是指在直线段的端部本来应该为“1”的元素变成了“0”，而在直线段的交叉处本来应该为“0”的元素变成了“1”（参见图9-32）。

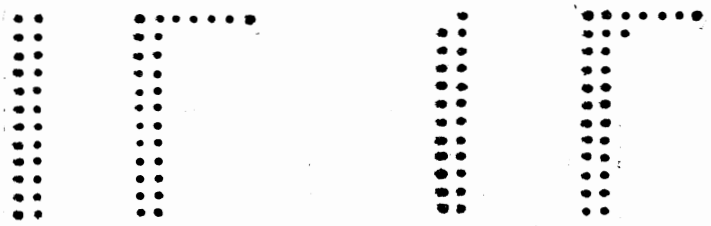


图9-32　由变换产生的直线段端部和交叉处的变形

（a）变换前；（b）变换后。

现在讨论进行整形处理。它的方法是当满足一定的逻辑条件时，使$B(K,\ L)$的值加上一个修正量，消除在直线段端部和交叉处产生的变形。

当$\langle NK\rangle_M>M-N$并且$\langle NL\rangle_M>M-N$时

若 $\begin{cases} A_1\cdot\overline{A}_2\cdot\overline{A}_3\cdot\overline{A}_4\cdot\overline{A}(I-1,\ J+1)\cdot\overline{A}(I+1,\ J-1)=1 \\ 或\ \overline{A}_1\cdot\overline{A}_2\cdot\overline{A}_3\cdot A_4\cdot\overline{A}(I,\ J+2)\cdot\overline{A}(I+2,\ J)=1 \end{cases}$

则 $\Delta_1=\min[(N-M+\langle NK\rangle_M)\cdot(M-\langle NL\rangle_M),\ (M-\langle NK\rangle_M)\cdot(N-M+\langle NL\rangle_M)]$

若 $\begin{cases} \overline{A}_1\cdot\overline{A}_2\cdot\overline{A}_3\cdot\overline{A}_4\cdot\overline{A}(I,\ J-1)\cdot\overline{A}(I+2,\ J+1)=1 \\ 或\overline{A}_1\cdot\overline{A}_2\cdot A_3\cdot\overline{A}_4\cdot\overline{A}(I-1,\ J)\cdot\overline{A}(I+1,\ J+2)=1 \end{cases}$

则$\Delta_2=\min[(N+M+\langle NK\rangle_M)\cdot(N-M+\langle NL\rangle_M),(M-\langle NK\rangle_M)\cdot(M-\langle NL\rangle_M)]$

修正项 Δ_1、Δ_2 将使与直线段端部相应的那些$B(K,\ L)$的值增加，从而阻止本来应该为“1”的元素变成“0”。

若 $\begin{cases} \overline{A}_1\cdot A_2\cdot A_3\cdot A_4\cdot[A(I-1,\ J+1)\cdot A(I+1,\ J-1)+A(I-1,\ J+1)\cdot\overline{A}(J+2,\ J)+A(I+1,\ J-1)\cdot\overline{A}(I,\ J+2)]=1 \\ 或A_1\cdot A_2\cdot A_3\cdot\overline{A}_4[A(I,\ J+2)\cdot A(I+2,\ J)+A(I,\ J+2)\cdot\overline{A}(I+1,\ J-1)+A(I+2,\ J)\cdot\overline{A}(I-1,\ J+1)]=1 \end{cases}$

则 $\Delta_3=-\Delta_1$

若 $A_1\cdot\overline{A}_2\cdot A_3\cdot A_4[A(I,\ J-1)\cdot A(I+2,\ J+1)+A(I,\ J-1)\cdot\overline{A}(I+1,\ J+2)+A(I+2,\ J+1)\cdot\overline{A}(I-1,\ J)]=1$

或 $A_1\cdot A_2\cdot\overline{A}_3\cdot A_4[A(I-1,\ J)\cdot A(I+1,\ J+2)+A(I-1,\ J)\cdot\overline{A}(I+2,\ J+1)+A(I+1,\ J+2)\cdot\overline{A}(I,\ J-1)]=1$

则 $\Delta_4=-\Delta_2$。

修正项 Δ_3、Δ_4 将使与直线段交叉处相应的那些$B(K,\ L)$的值减小，从而阻止本来应该为“0”的元素变成“1”。

3．文字线幅和二值化阈值τ的关系

使用本方法所产生的文字线幅，由于τ的选择，有一部分为一线（位）幅，另一部分为2线（位）幅，这看来好象是存在的一个问题，但在实用上并不是大的障碍。若有必要，可用1×4（或4×1）的逻辑屏蔽把线幅凑成1线（位）或2线（位）幅。

关于文字线幅和二值化阈值τ的关系是，τ若大，则线幅变细，τ若小，则线幅粗大；线幅细，τ大就可能断线；线幅粗大，τ小就可能损坏整个字形。因此，τ的选择要特别注意。最好通过实验，根据各文字对于一个点阵阶数可得到最好字模的τ值而选择决定。

4．汉字尺寸变倍程序的处理流程

对汉字尺寸变倍，如果采用公式变换法放大，不但能从方阵到方阵，而且可以实现方阵到矩阵的放大处理。其放大子程序见图9-33。

（四）直接放大法

近年来，在以微型机为中心的汉字信息处理系统中，对汉字库中存储的字形常采用直接放大的方法，提供几种放大的字形供用户选用。

这种直接放大法的基本思想是根据汉字库中存储的一种基本字形以及要放大的倍数（通常是整数倍），直接映照出放大后的字形数据，然后供显示或打印用。

使用直接法放大汉字，又分横向放大、竖向放大、横竖向同时放大三种情况。

1．横向放大　设基本点阵字模为15×16，只在横的方向上放大一倍，即为30×16，

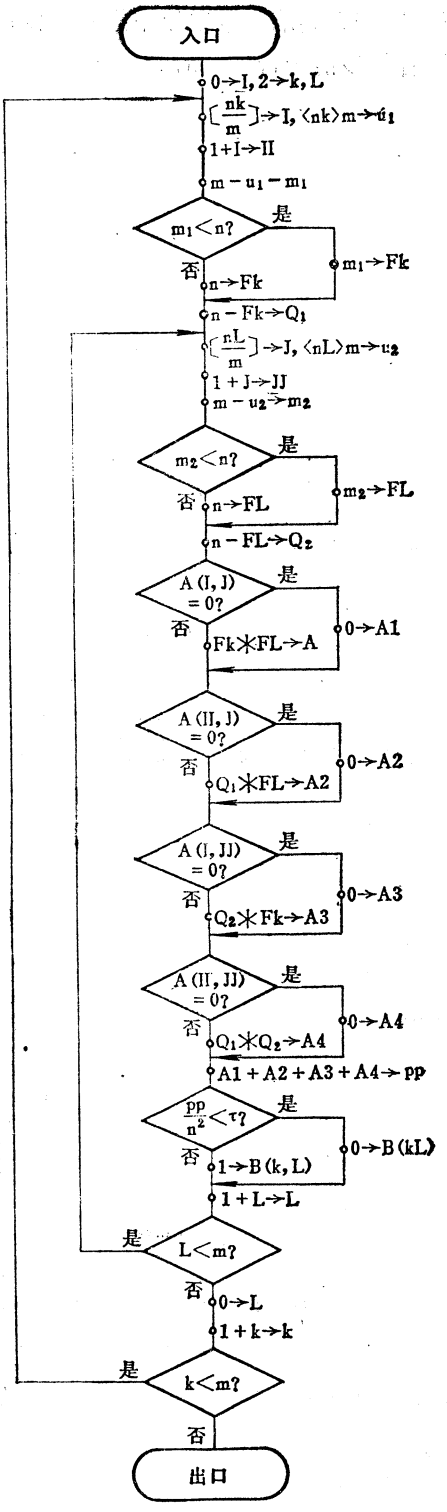


图9-33　放大子程序流程图

这就是说，放大后的汉字为基本汉字的双倍宽。

横向放大的算法思想是：按列读取基本字模中的数据依次写到放大缓冲区的奇数列，而放大缓冲区的偶数列接着复抄前一奇数列的数据（见图9-34）。

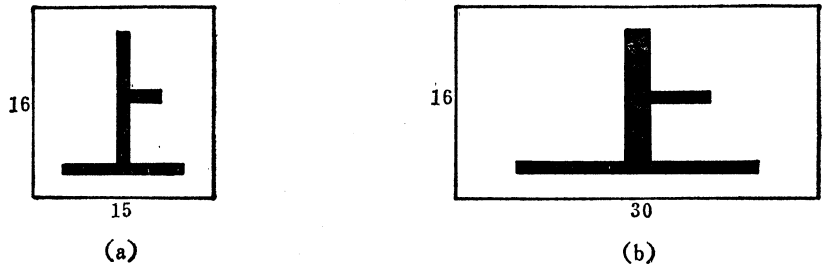


图9-34　横向放大字例

（a）基本字模；（b）横向放大后的字模。

由于印刷机智能化的进展，这项工作可由印刷机控制器完成，主机只需发放大命令即可。

2．竖向放大　设基本点阵字模为15×16，在竖的方向上放大一倍，即15×32。这就是说，放大后汉字的高度为基本字模的两倍。

竖向放大的算法思想是：按行读取基本字模的数据依次写到放大缓冲区的奇数行，而放大缓冲区的偶数行接着复抄前一奇数行的数据（见图9-35）。

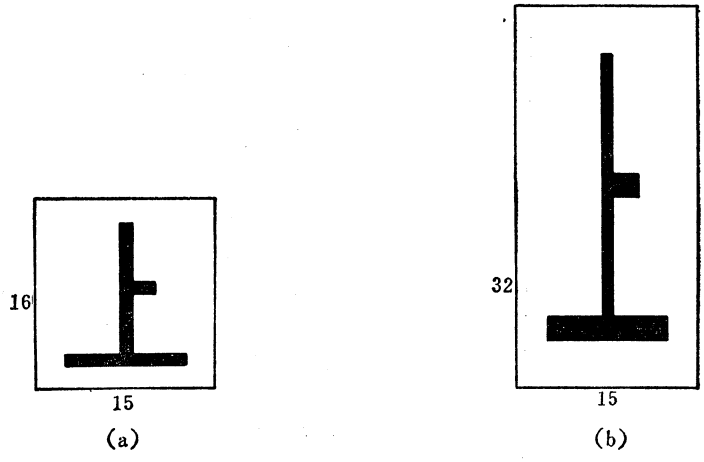


图9-35　竖向放大字例

（a）基本字模；（b）竖向放大后的字模。

3．横、竖向同时放大　设基本点阵字模为15×16，在横、竖方向上同时放大一倍，就是说，放大后的汉字为30×32，它是基本字模的四倍。

放大时，先在内存中开辟15×16和30×32两个缓冲区，前一个用来存放基本点阵字模，后一个用来存放放大后的点阵字模，按行（或列）读取基本字模中的数据，放大后依次写到放大缓冲区的奇数行（或列），而放大缓冲区的偶数行（或列）接着复抄前一奇

数行（或列）的数据。

对于一行中的十六进制数据（09，$A\sim F$），的放大操作，举例如下：

1	○○○●	○○○○○○●●	0 3
⋮	⋮	⋮	⋮
9	●○○●	●●○○○○●●	C 3
A	●○●○	●●○○●●○○	C C
⋮	⋮	⋮	⋮
F	●●●●	●●●●●●●●	F F

将一行中的点阵字模数据（0～9，$A\sim F$），分别转换为相应放大后的数据，写入放大缓冲区内。

使用直接法放大汉字的字例见图9-36。

用直接法放大汉字的处理流程见图9-37。放大子程序见图9-38。

综上所述，软件人员必须根据需要以及具体的设备条件选择适当的变倍算法，才能编制出高质量的切合实用的汉字变倍程序，提高汉字信息处理系统的效能。

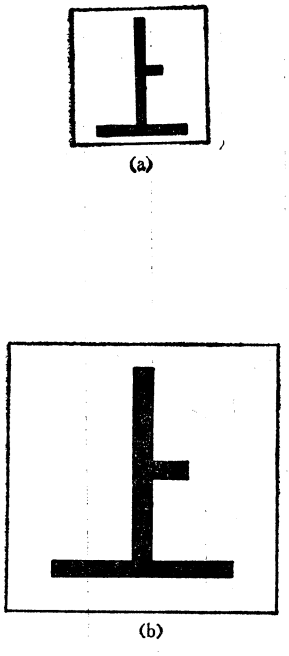


图9-36　横向、竖向同时放大两倍的字例

（a）基本字模；（b）放大四倍的字模。

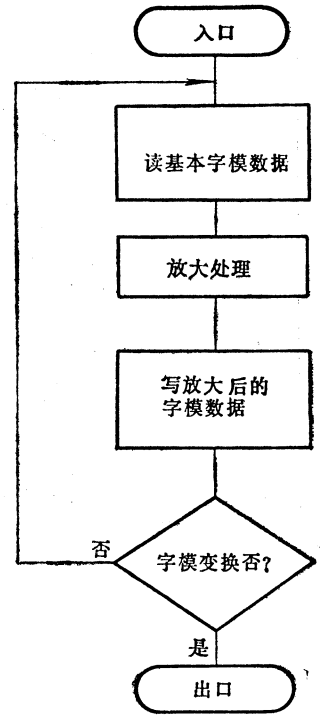


图9-37　直接法放大汉字的流程

9.4.4　汉字文本编辑程序

汉字编辑程序是对输入的汉字文本（text）按指定的格式编排，从而实现汉字文件的显示、打印等功能的程序模块。

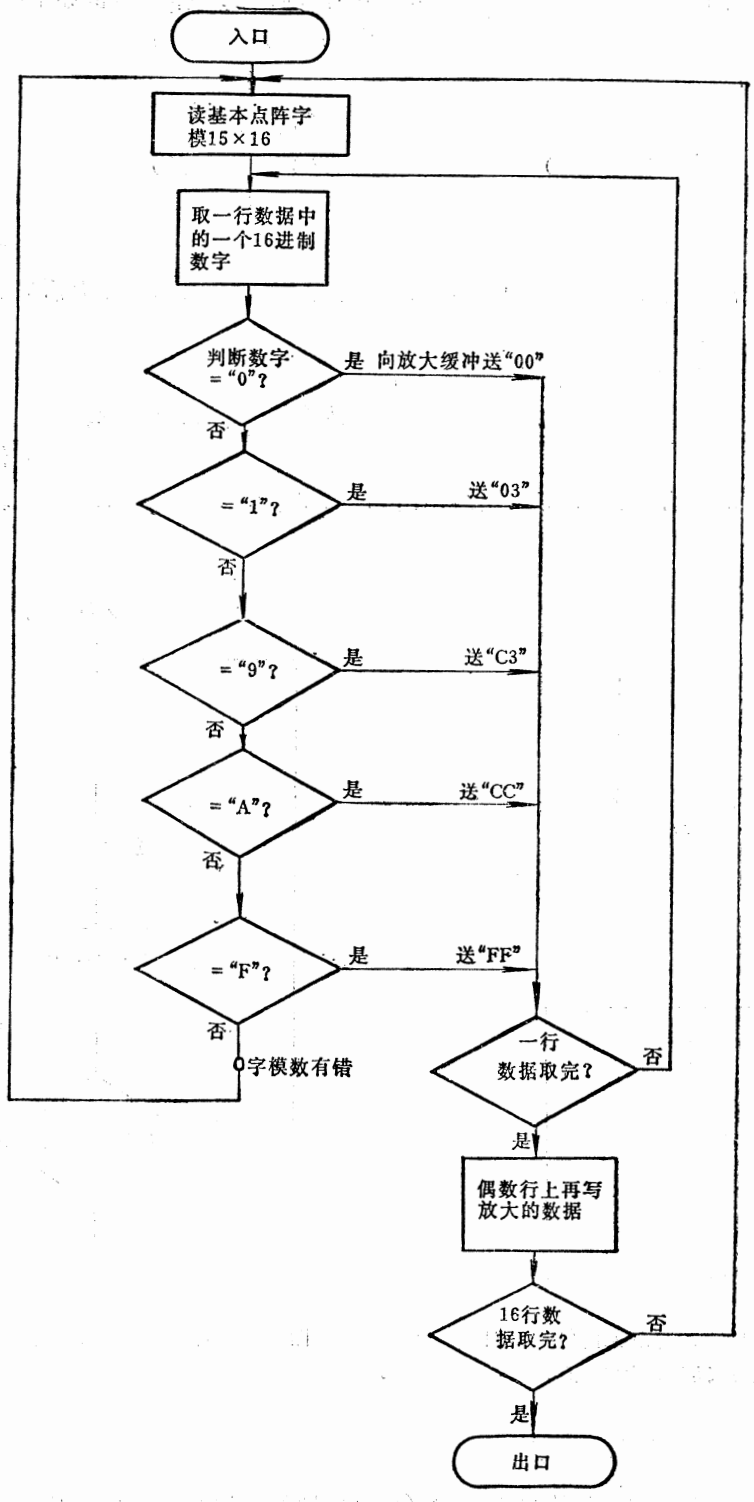


图9-38 直接放大四倍的子程序

一、汉字编辑程序模块的功能

在汉字信息处理过程中，要求把计算机加工的汉字文件或中间结果在屏幕上显示，或用印刷机打印出来。汉字编辑程序模块能使计算机处理的汉字文件满足一定的编排格式，以及对中间结果文件进行校正和更改。具体的功能有：游标（cursor）定位；增、删、改单个汉字；增、删、改一行汉字；行对齐，选择页面格式等。

二、汉字编辑的实现方法

汉字编辑程序由若干子程序模块组成，当用户程序进入汉字编辑程序模块后，由控制台或用户程序提供编辑命令或控制码，而后转入相应的子程序模块完成对单个汉字及汉字文件的编辑。

（一）游标定位

汉字在计算机内部的存储、加工、判断比较等工作都是用机内码形式进行的。存储在内存中的汉字文件由一组汉字机内码组成。每个汉字机内码占两个字节，其他非汉字字符占一个字节。

游标定位（cursor positioning）是指用程序方法将游标显示在屏幕上一个指定的物理位置上。这里游标是这个物理位置的指示标记。

游标定位的物理表现形式有游标左移、游标右移、游标上移、游标下移，以及游标回原点（屏幕上第一行第一列的汉字位置）。对应于显示屏幕的每一个汉字的物理位置，我们用表示物理屏幕位图的矩阵D的元素d_{ij}表示。这里，$1 \leqslant i \leqslant n$；$1 \leqslant j \leqslant m$。$n$为物理屏幕每帧显示汉字的最大行数，$m$为每行显示汉字的个数。对应于$D$在内存设置一个屏幕映照代码区。这个区也称为显示存储器，用来存储汉字机内码。其大小按照系统允许逻辑文件的长度而定。有的系统一个逻辑文件对应于一个物理文件；有的则一个逻辑文件可用若干个物理文件来表示。对于包含多个物理文件的逻辑文件的显示，加工处理可以用替换屏幕显示的方法。因此，对于物理屏幕的游标控制相当于对逻辑文件的映照代码区矩阵的行和列的指定。

游标右移实际上就是当前游标所在位置，其对应的映照代码区的列数j增加1；游标左移则为列数j减少1；游标下移是游标当前所在行的行计数i增加1，列计数不变；游标上移为游标所在行的行计数i减少1，列计数不变；游标回原点是游标返回左上角显示，即行计数i置1，列计数j置1。

游标定位程序模块流程如图9-39所示。

（二）添加汉字

在汉字信息处理过程中，人们必须对输入的汉字文件信息和编辑的汉字文件进行检验查错。发现有漏字情况，调用添加汉字子程序模块完成汉字的插入。

实现汉字的添加方法是先用游标定位程序确定被添加汉字在汉字文件中的位置，依次将指定位置以后的汉字代码或符号的代码向后移动。空出游标所指定的那个位置，以备添加被插入的那个汉字。

添加汉字的处理流程是：

输入汉字输入编码（包括用汉字整字键盘输入汉字）。

将汉字输入编码转换成汉字机内码，填入到被插入的位置上。

若添加汉字是在屏幕显示情况下进行，除在映照代码区填入汉字机内码外，还必须

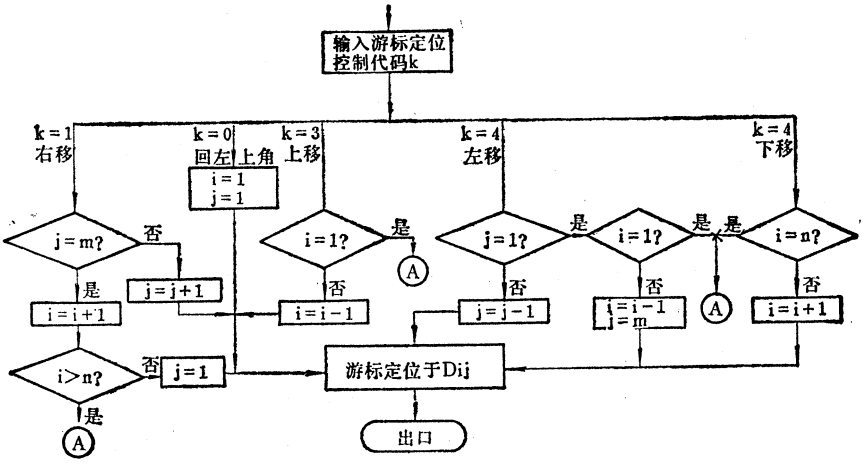


图9-39　游标定位程序流程图

图 9-39 的注解：*A*为游标越界处理，通常采用两种方法。其一是显示出错信息，并将游标显示于屏幕左上角；其二，在游标产生越界时（即列数值大于*m*或小于 1，行数值大于*n*或小于 1），产生屏幕滚动。向上滚动是产生在游标上移，而现行行 $i=1$ 时；或现行行 $i=1$，列数 $j=1$ 需要游标左移的情形。这时将映象代码区里对应于现行行的前一行（若现行行已是映象代码区的第一行则将映象代码区的最末行）显示于屏幕的第一行位置上，原屏幕上的行相应下滚一行，末行去掉。并把游标显示在当前行的相应位置上。向下移动游标位置产生越界处理方法与此类似。

根据汉字机内码访问字模库，取出汉字字形信息，显示于屏幕上。

要想在存放汉字文件代码的映照代码区中空出一个汉字机内码位置。必须把原来存放在这个位置上的代码信息向后移位，而且又不能使汉字文件原有的信息丢失，所以首先要判定要移位的代码信息的最后一个移动位置。通常是含有空格符的汉字文件段的末行。移位从最后一个需后移的代码信息开始，这样，文件段的末行末尾的首空格位置被其前一位置的代码信息覆盖，此文件段以后的所有位上的信息不变。

〔**例子**〕

“目前北京、上海、广州、南京、杭州、大连、青岛、天津、成都、沈阳、福州，南宁、西安、昆明等地均办理国际及港澳地区的用电报业务。

……。”

我们需向第三行的“用”字后插入一个“户”字。用游标定位于第三行的“电”字上，判定下行即为被添加汉字所在文件段末行，而且末行含有空格符，所以确定移位从第四行第四个位置开始。将“·”移至第五位，“名”字移至第四位，“业”字移至第三位，“报”字移至第二位，前行“电”字移至第一位，空出第三行最后一个位置。如下所示：

“目前北京、上海、广州、南京、杭州、大连、青岛、天津、成都、沈阳、福州、南宁、西安、昆明等地均办理国际及港澳地区的用户电报业务。

……。”

将“户”字插入到“用”字后，即得到正确的结果。

添加汉字子程序流程如图9-40所示。

说明：若处理的汉字文件在被插入的汉字所在文件段末行无空格，则指定*k*为文件段末行行号数。并将（Dkm）转送到一个缓冲区。而后执行上述流程图的过程，最后将

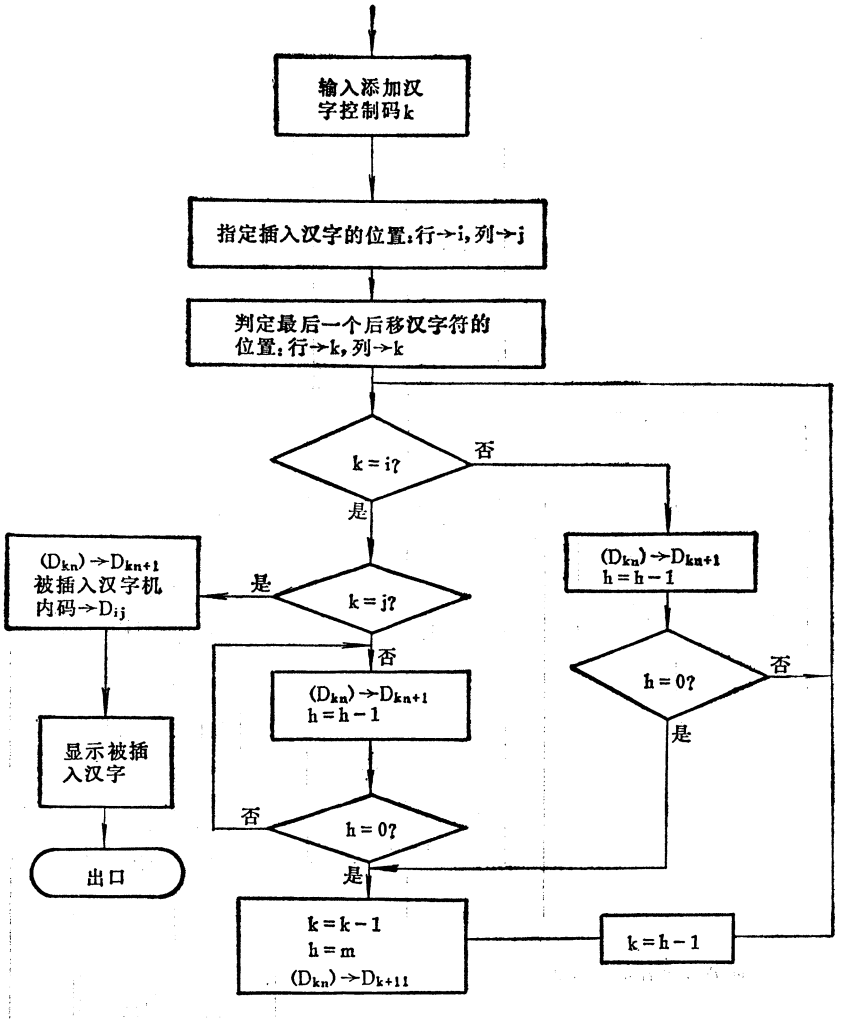


图9-40　添加汉字子程序流程图

缓冲区内容作为新的一行插入到文件段末行之后。如果系统不允许增加新的文字行，则插入行工作仍执行，实际上是将原汉字文件最末行内容删去。

（三）删除汉字

删除汉字的工作原理和处理流程同添加汉字情况相同。其步骤如下：

调用游标定位模块，指定被删除汉字在汉字文件中的位置。

自游标指定所在位置开始依次将后一位的代码填入前一位置，直至被删除汉字所在的文件段末行最后一个非空字符填入前一位为止。

用空白符号替换原文件段结束所在行的最末一个非空符号。

检查是否产生空白行，若有则调用删去一行的子程序模块，删去空白行。否则返回删除汉字子程序模块出口。

若被删除的符号是一个“信息处理交换用七位编码字符集”(即 GB1988-80字符集)的符号，则用“GB1988-80”字符的空白符“*b*”代入被删除的符号，并重排GB1988-

80字符串。使“b”出现在被删除的GB1988-80字符所在字符串末尾。若末尾出现两个“b”,则作为一个汉字空白符被删去。若为一个则保留，返回子程序模块出口。

删除汉字子程序流程如图9-41所示。

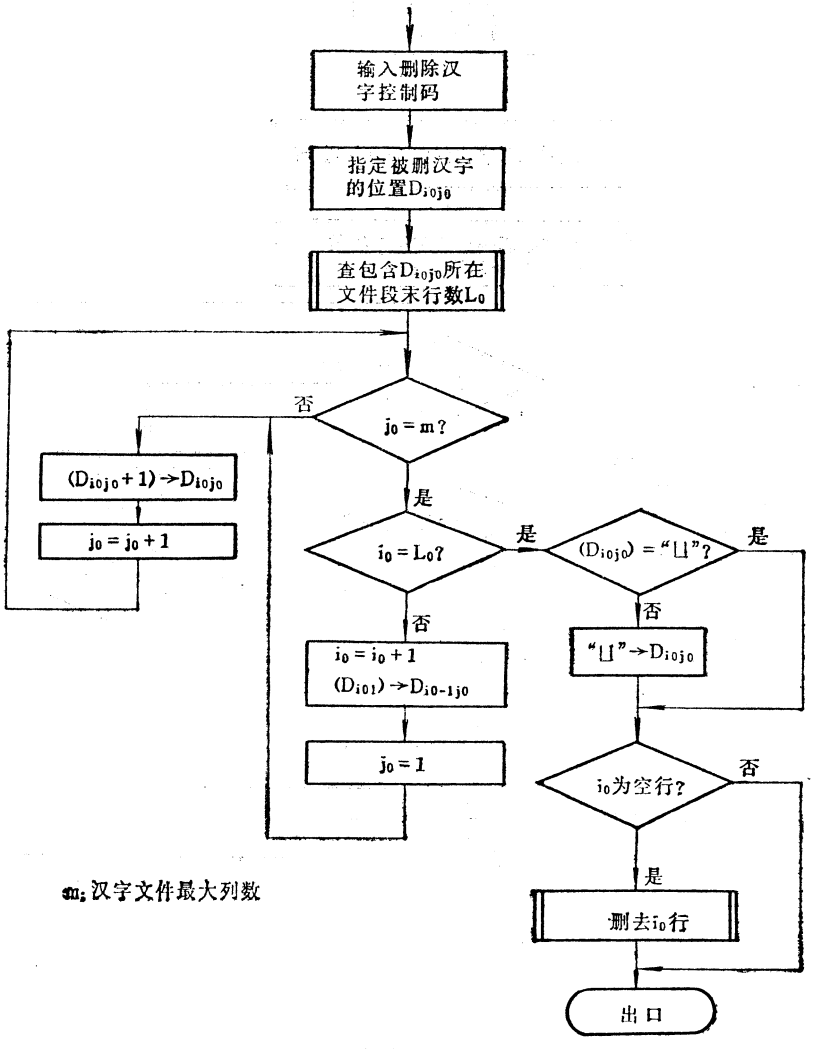


图9-41　删除汉字子程序流程图

（四）更改汉字

在汉字文件编辑过程中，若发现有错字，则可调用更改汉字模块进行修改。这个子程序模块处理过程较为简单。先用游标定位程序指出被更改的汉字所在位置。然后，由汉字输入设备输入正确的汉字输入编码，通过转码程序将其转换成汉字机内码，用新的汉字机内码替换错误的汉字机内码，从而完成汉字文件的单个汉字更改工作。

更改汉字子程序流程图省略。

更改一段汉字文件处理方法与更改单个汉字类似。用游标定位程序指定被更改的汉字段首址，并输入更改区间长度。然后，输入正确的汉字代码串。用正确的汉字字符串覆盖被修改的汉字代码串。这里有下面几种情形发生。

（1）正确的代码串（新输入的字符串）长度同被更改的区间长度相等。用新的代码串内容覆盖错误的串。

（2）正确的代码串长度大于被更改的区间长度，截去新输入的代码串，截去后长度等于被更改的区间长度，并用截去后的代码串内容覆盖被更改区间的串，截去后剩余部分调用添加汉字程序将其插入。

（3）若新输入的代码串长度小于被更改的区间长度，用新输入的代码串覆盖被更改区间的内容，覆盖从指定区间的最左部开始，区间内未被覆盖的部分用删除汉字程序将其删去。

（五）页面尺寸格式

汉字文件在计算机内的表现形式是一组汉字机内码。页面尺寸的指定实际上是在这一组汉字机内码符号中间，按照用户要求插入一些回车换行符号和页文件结束符。这样

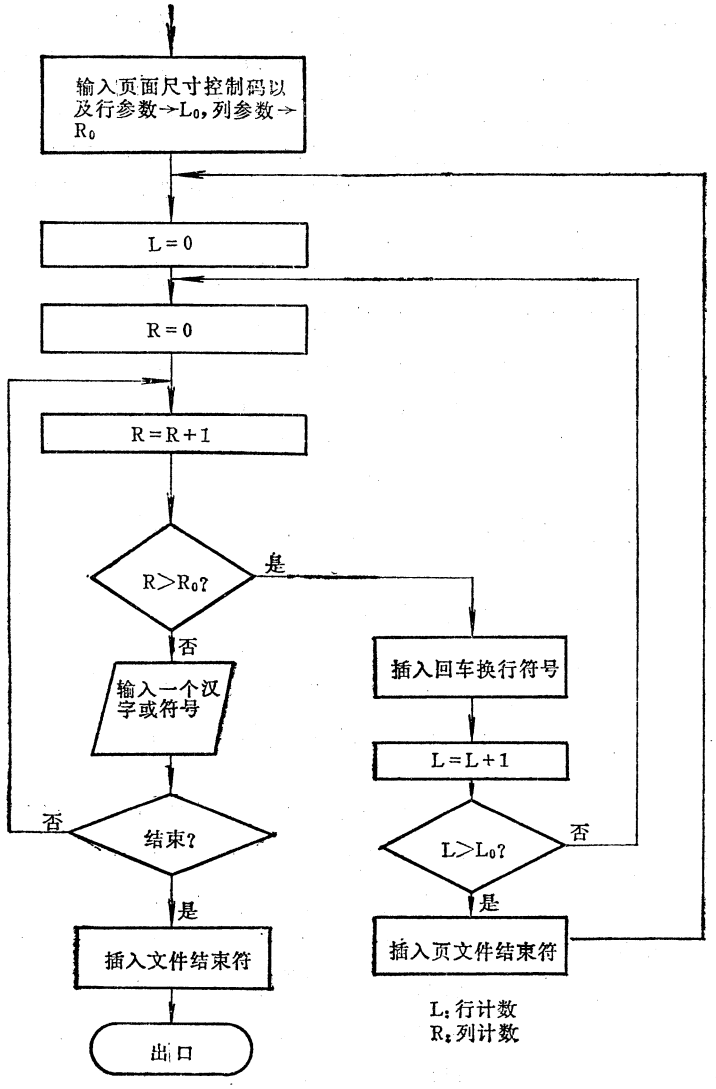


图9-42　页面尺寸格式子程序流程图

在汉字文件输出时即可得到用户所需要的页文件格式。

页面尺寸格式子程序流程如图9-42所示。

（六）添加一行汉字，删除一行汉字

对于计算机已经建立的汉字文件，有时除添加单个汉字或删除单个汉字以外，常常还要对其整行的汉字进行删除或添加。下面分别对删除和添加一行汉字进行说明。

（1）删除一行汉字　当需要删除汉字文件的某一行时，由控制台输入删除一行文字的控制码和被删除的行的行号数。程序从给出的行号开始将以后的每一行内容顺序地填入前一行，直至汉字文件的结束行为止。更改原汉字文件逻辑行参数L。即$L=L-1$。

删除一行汉字的子程序流程如图9-43所示。

（2）添加一行汉字　添加一行汉字的步骤为：

输入添加一行汉字的控制码，指定插入行的位置。

（1）检查原汉字文件是否有可压缩的空白行。例如，汉字文件的末行是空白行，或标题行与正文行之间有可压缩的空白行。当末行是空白行时，把原汉字文件的总行数送入行计数器L_0，通过移行将指定的插入行区空出。若在标题行与正文行之间有可压缩的空白行，则L_0等于正文行的首行行号数，然后将L_0行的内容送入L_0-1行，将L_0+1行内容送入L_0行。用(L_0)表示第L_0行的内容，即$(L_0)\to L_0-1$，$(L_0+1)\to L_0$，直至被插入行原有内容$(L+1)$送入前一行L。

（2）输入添加行的汉字输入编码，调用译码程序，将其转换成汉字机内码，并填入被指定的插入行位置。

（3）若汉字文件无可压缩的空白行，则修改原汉字文件的行参数L_0。使$L=L+1$，$L_0=L-1$。而后将$(L_0)\to L_0+1$。即将原汉字文件的最末行填入新的汉字文件的末行，依次将汉字文件上一行的内容填入下行，直到被指定插入行的内容填入下行为止。进行步骤（2）完成在汉字文件中输入插入行的汉字内部码。

添加一行汉字的子程序流程如图9-44所示。

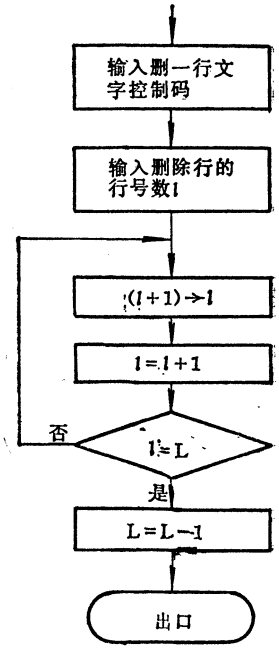


图9-43　删除一行汉字子程序流程图

$(l+1)\to e$表示将$l+1$行的内容送入l行；L—汉字文件的行数。

（七）行首和行末对齐

为了使汉字文件的编辑满足一定的格式要求，系统配有行首和行末对齐子程序模块，行首、行末、单字不成行、单行不成页等禁则子程序模块。

调用行首、行末对齐子程序模块，使得汉字文件编辑输出时，汉字文件正文的上下行起始位置和结束位置都是整齐一致的（文件段的开头行行首和文件段的末行行末例外），即每行正文前空格个数，正文行后的空格个数相同。如果汉字文件印出时是双面印刷（一张纸的正、反两面都要印刷文字），则必须使得正文在正反两面上的位置对称。设汉字文件在正面上每行汉字正文首字开始前（包括因装订而空出的位置）空k个空格，行末后有k个空格，则其反面的一页，正文行首字前应空k个空格，行末后有k

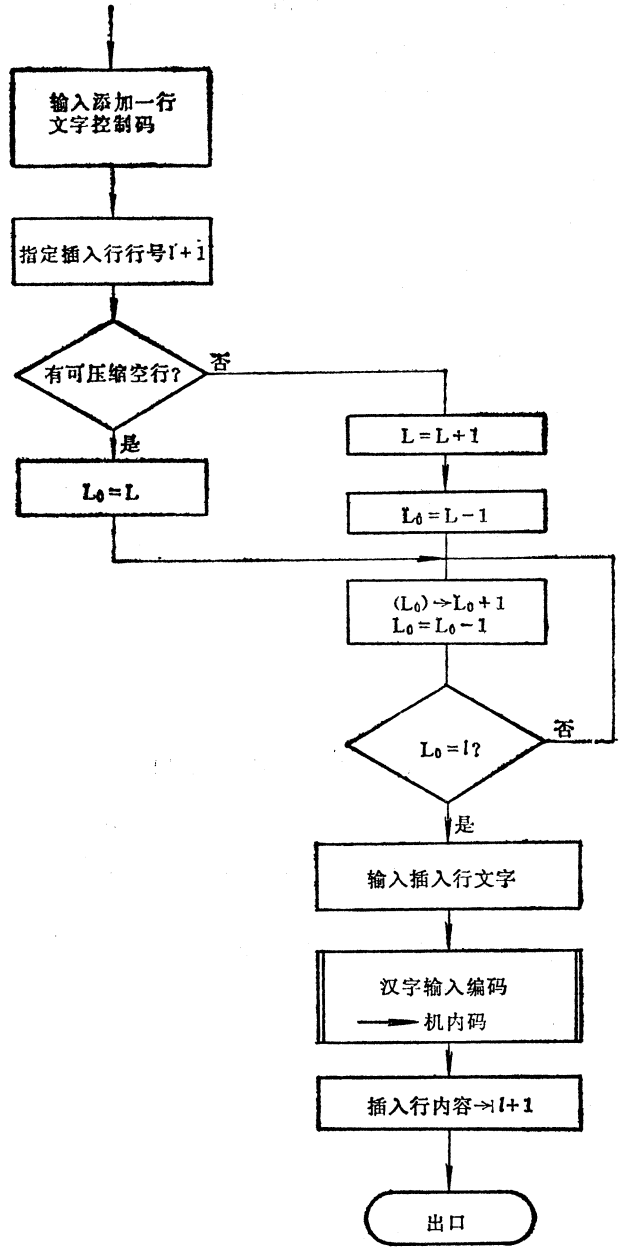


图9-44　添加一行汉字的子程序流程图

L_0—行计数器；(L_0)—L_0行的内容；L—汉字文件的行数；$l+1$—插入的行号数。

个空格。

行首行末对齐情况如图9-45所示。

如果不需留出装订用的位置，则要使得汉字文本印刷在纸上的位置左右对称，上下行一致。

计算行首、行末空白个数的公式是：

$$k_{反}=k_{正}=[(R_*-R_0-k_0)/2]$$

$$k_{反}=k_{正}=〔(R_*-R_0-k_0+1)/2〕+k_0$$

其中，R_* 为汉字输出设备每行文字个数；

R_0 为用户指定的汉字文件正文每行文字的字数（标点符号和 GB1988-80 字符，占汉字位置的一半）。

〔x〕表示取数值 x 的整数部分；

k_0 装定用的空格个数。

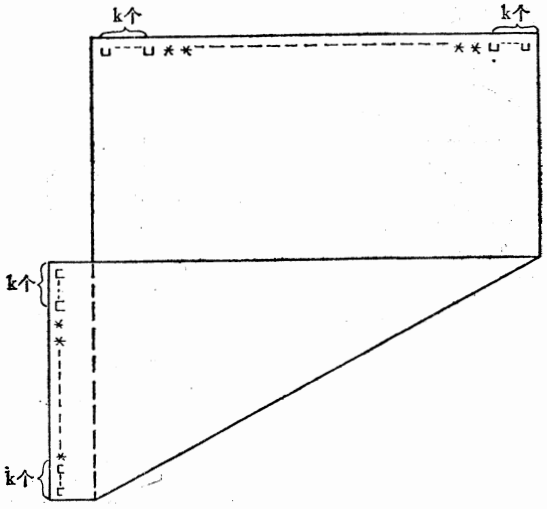


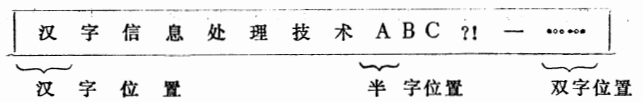
图9-45 行首、行末对齐示意图

上面第二式中的〔 〕内加 1 是处理当汉字印刷机允许印出总字数减去正文所要求字数为奇数时，多余的一个空格放在装订线一边。不需要留装订空格数时$k_0=0$。

（八）排版禁则处理

编辑排版要求标点符号，右括号，右引号等不能出现在文字行的开头；破折号，省略号不能一半在行末，一半在行首；左引号，左括号不能出现在一行的末尾；单个汉字不能成一行；单独一行不能组成一页等禁排规则。由于受技术的限制，目前计算机还很难实现对GB1988-80字符串回行规则及化学分子式、数学公式等的回行规则要求。但可以通过显示屏幕和其他汉字印刷机，完成对汉字文本的编辑排版面的某些要求。这里叙述用软件方法实现某些文件禁排规则的处理。

前面已经说过，一个汉字在计算机内用两个字节的代码表示，括号、引号、标点符号和GB1988-80字符，用一个字节的代码表示(即非汉字符号的代码占一个汉字位置的一半)。破折号和省略号占两个汉字位置。汉字、标点符号和破折号等所占位置如图所示。



在外部设备上输出的汉字文件，每行显示或印出的汉字个数是一个整数。因此，文件行中若含有GB1988-80字符或标点符号则应有偶数个。若为奇数个，则要添加占半个汉字位置的半字空白符。

计算机实现行首、行末、单字不成行、单行不成页禁则的方法，类似于手工铅字排版时使用的方法。就是压缩或插入一些占半个汉字位置的半字空白符或空格。

行首、行末、单字不成行、单行不成页禁则模块流程图如图9-46（a）、（b）所示。

（九）回车换行

回车换行对于不同的汉字输出设备采用不同的特殊符号。输出汉字文件时遇到回车换行符号，硬件自动产生回车和换行动作。若硬件上没有自动回车换行功能，则采用软件的方法来实现。即当遇到回车换行符号时，将回车换行符号去掉，并在本行包括回车换行符号所在位以后的部分用空白符填满至行末。这样硬件在输出汉字文件时，输满一行即产生换行动作。

（十）屏幕的上滚、下滚、左移、换幕

前面说过，一个汉字逻辑文件可用多个物理文件表示。通常称一个物理文件为一个页文件。调用屏幕滚动子程序模块可以使一个逻辑文件连续的在屏幕上显示。其外部表现形式是逻辑文件不断上滚。逻辑文件的开始行接在结束行后不断地在屏幕上显示，直至用户键入结束上滚控制码为止。

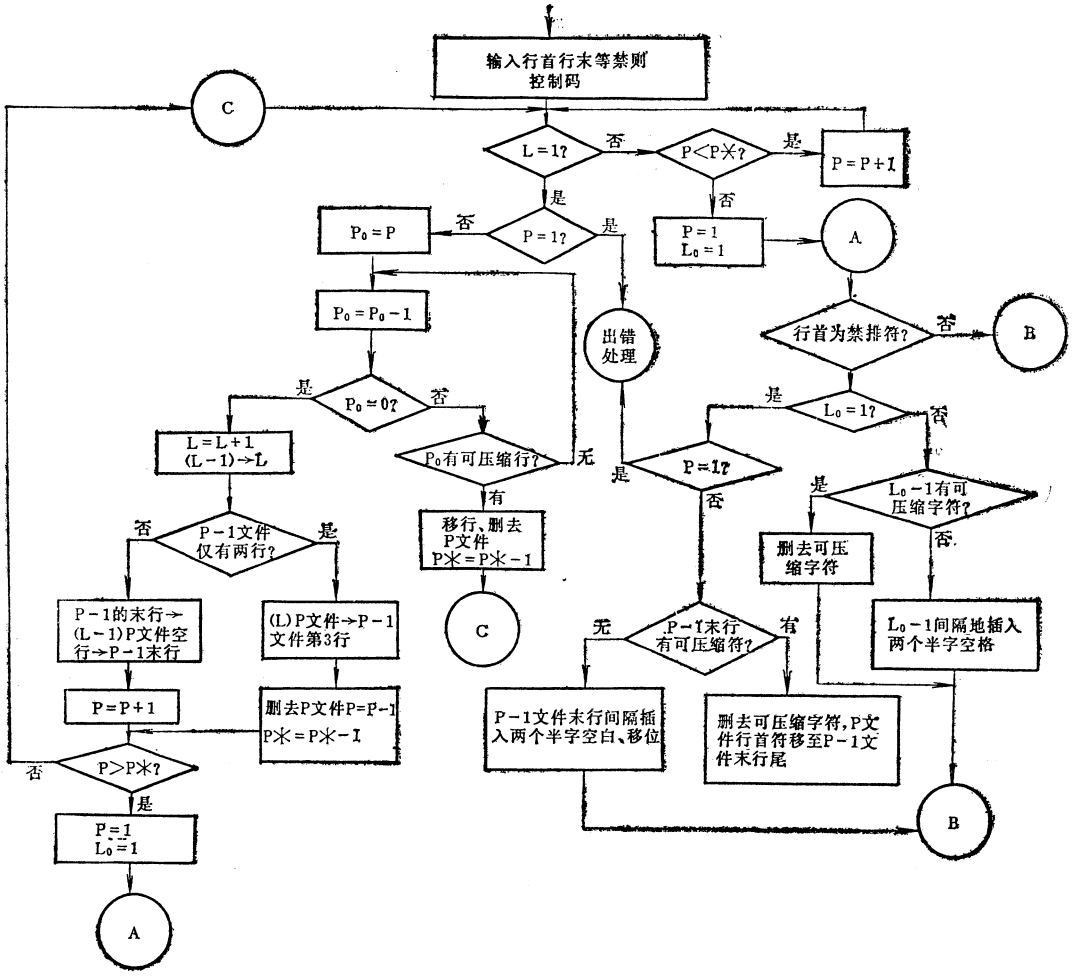


图9-46（a） 禁排规则处理流程图之一

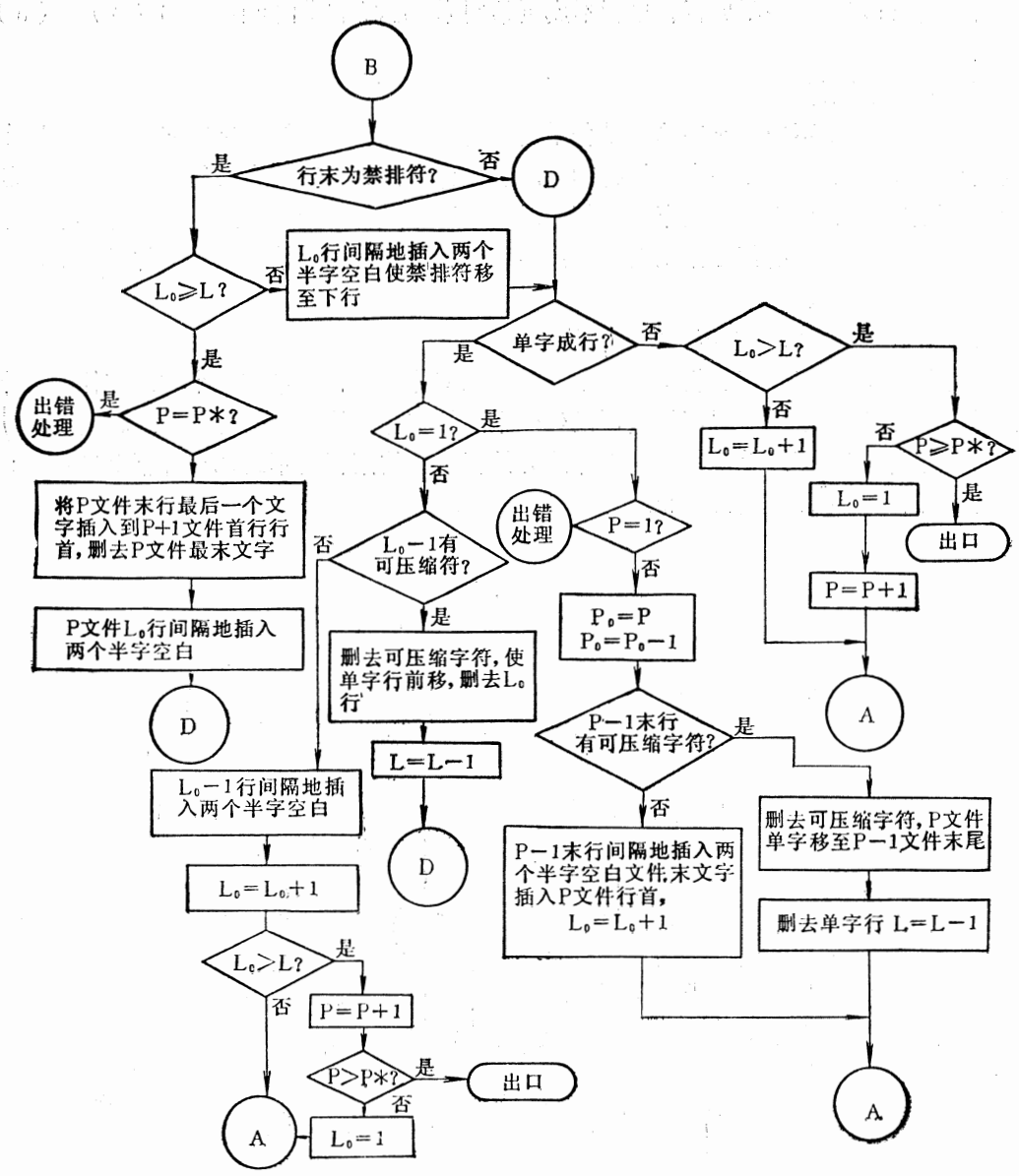


图9-46(b)　禁排规则处理流程图之二

形象地说，一个逻辑文件就好象首尾连接起来的一卷文件，显示屏幕好比是一个窗口，用户通过窗口可以看见逻辑文件中某一段的内容。

其内部表现形式为：设一个逻辑文件含有 n 个页文件。其数据结构是：

页文件 1	页文件 2	……	页文件 n

1．屏幕的上滚　我们假设屏幕上滚从第一个页文件开始，处理流程为：

（1）在屏幕上显示页文件 1；

（2）将页文件 2 的第一行内容存入滚动行缓冲区；

（3）将页文件 2 的第二行填入第一行，直至页文件 2 的末行填入前一行；

（4）页文件3的第一行填入页文件2的末行，第二行填入第一行，……，依次将页文件3的后一行内容填入前行；

（5）类似（3）、（4）处理页文件4至页文件 n；

（6）页文件1的第1行内容填入页文件 n 的末行；

（7）页文件1的第2行内容填入第1行，第3行内容填入第2行，……，末行内容填入前行；

（8）缓冲区内容填入页文件1末行。

在屏幕上显示页文件1的内容，再执行（1）至（8）的工作。如此循环，直至要求结束时为止。

此外，我们还可以采用另一种方法。页文件信息不变，另开辟一显示屏幕缓冲区(设有 m 行)供显示用。其步骤如下：

（1）取页文件1送入显示缓冲区，把缓冲区内容显示于屏幕。

（2）以页文件1第2行起至页文件2的首行将其内容送入显示缓冲区，并在屏幕上显示。

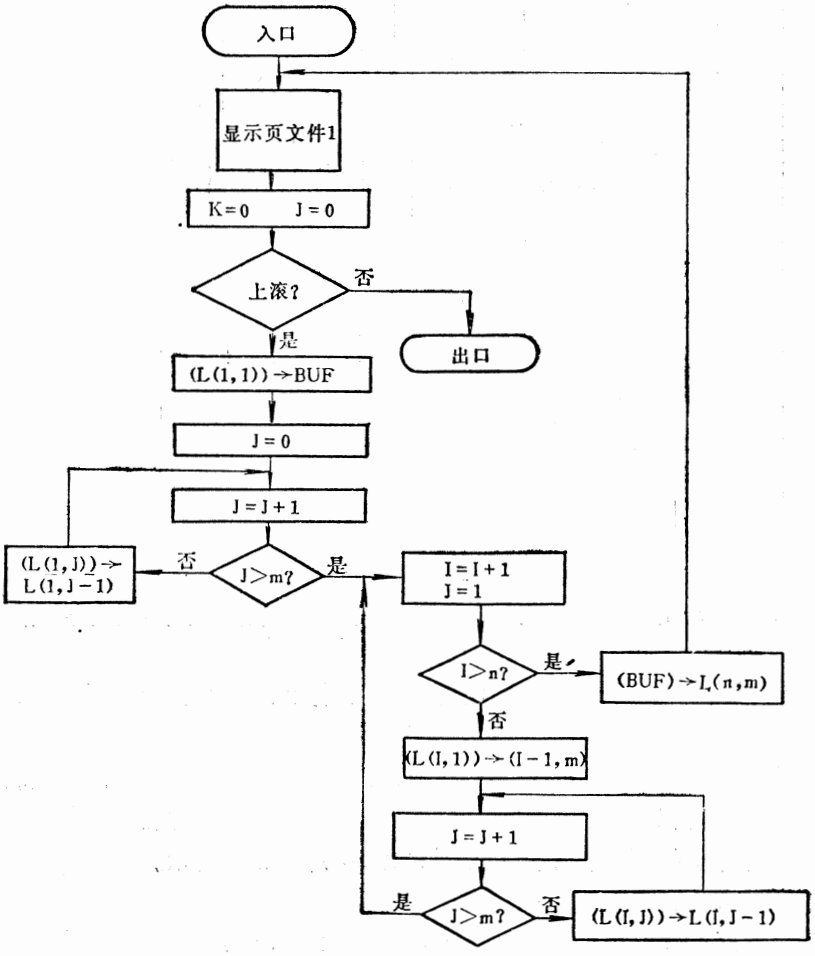


图9-47(a) 屏幕上滚子程序流程图之一

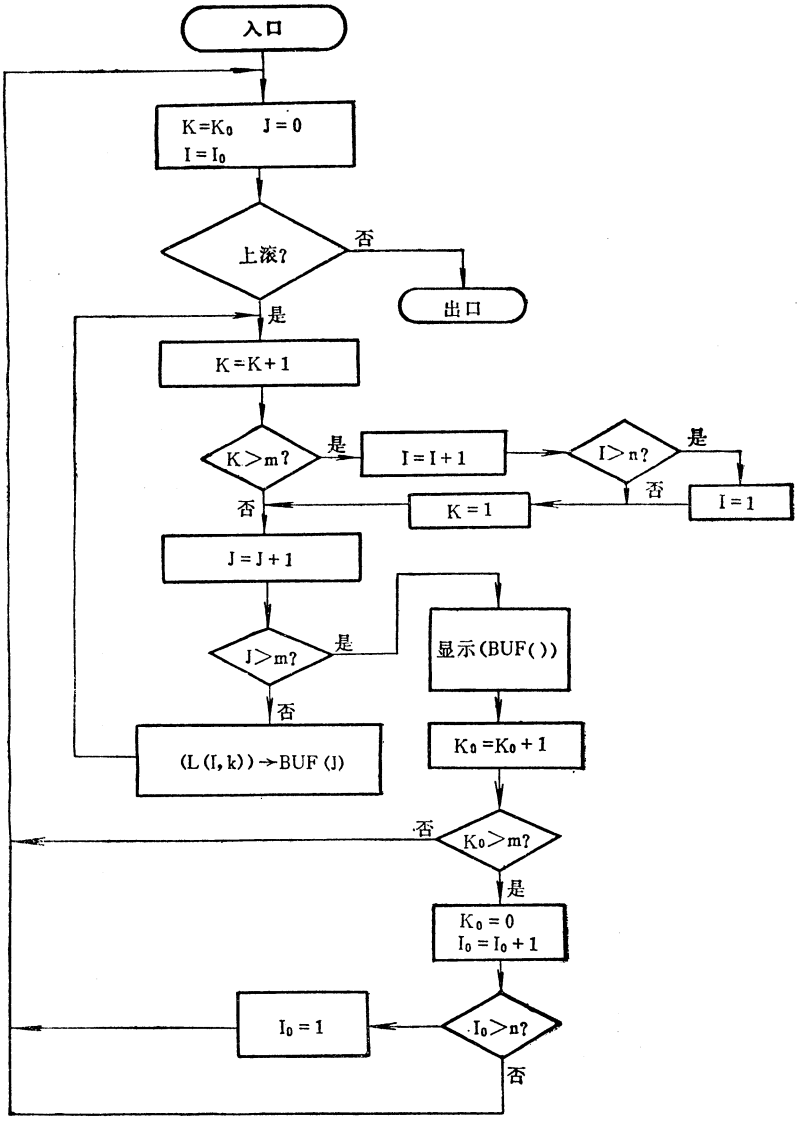


图9-47(b) 屏幕上滚子程序流程图之二

BUF( )—BUF(1), …, BUF(m) 的全体; I_0—上滚前页文件计数; K_0—上滚前页文件起始行。

(3) 依次类推，从页文件1第3行起至页文件2的第2行止，送入显示缓冲区，显示于屏幕上。

这种方法处理速度较高、简单，每次显示的内容都是从显示缓冲区获得，但多占内存。系统可以根据所花时间和空间的开销，选择不同的处理方法。

用上述两种方法处理的程序流程如图9-47(a)、(b)所示。

对于第二种方法再作如下说明:

设显示缓冲区为BUF(1), BUF(2), …, BUF(m);

页文件存储区为$L(1, 1)$, …, $L(n, m)$;

J为行计数器，I为页文件计数器；

K为滚动行计数器。

设滚动从第I_0个页文件的第k_0行开始，则$I=I_0$，$K=K_0$，$J=0$。

判断是否要求上滚。若不要求转出口，则往下执行。

自页文件I_0的第$K+1$行开始，依次将各行内容送入显示缓冲区（BUF( )）的相应行，直到取页文件I_0+1的若干行，且使得显示缓冲区的各行填满内容为止。实际上，将在页文件I_0+1取k_0行内容送入显示缓冲区，并在屏幕上显示缓冲区内容。如果还需要上滚，则修改滚动起始参数。即$K_0=K_0+1$，若K_0大于页文件最大行号数，则$K_0=0$，页文件计数加1；若页文件计数大于n，则置页文件计数$I=1$，执行以上的过程。

BUF( )表示BUF(1)，BUF(2)，…，BUF(m)的全体，即整个显示缓冲区。

用第一种方法，上滚后改变了原有页文件存储区的内容。如果要想恢复上滚前页文件存储区的状态，则必须反复执行上滚，直到屏幕上显示的全部内容为第一个页文件的内容时为止。

2．屏幕的下滚 （其处理的方法与上滚时采用的处理方法相似（见图9-48）。我们

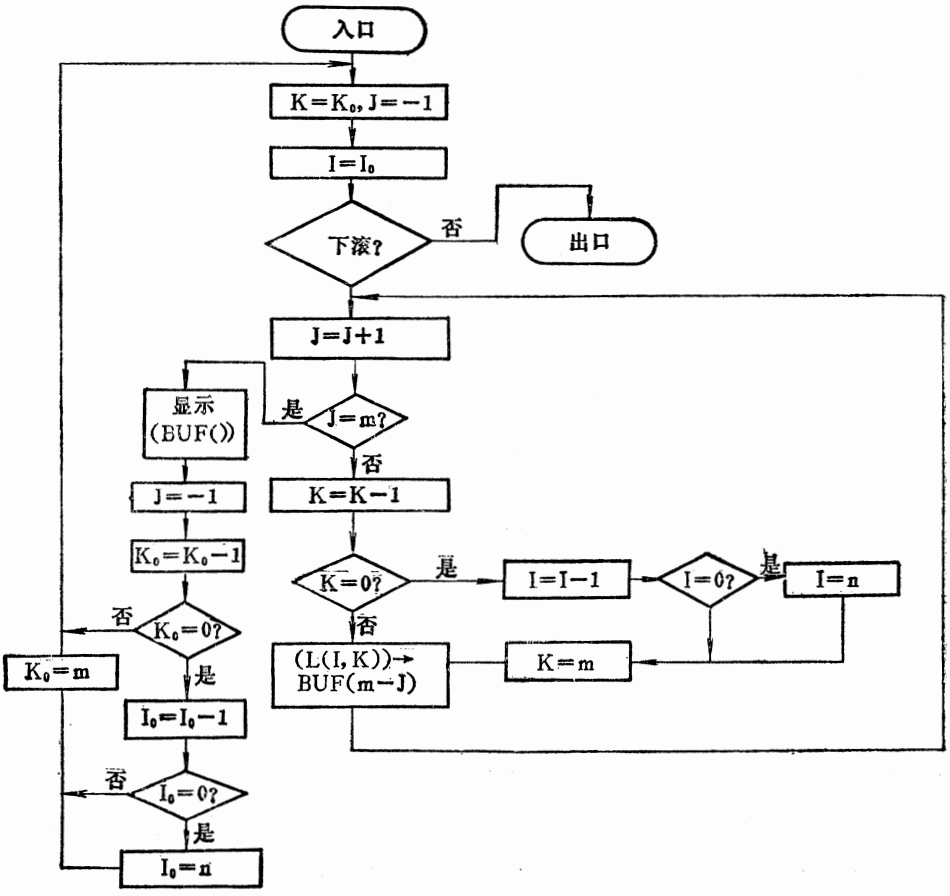


图9-48 屏幕下滚子程序流程图

K_0—下滚前屏幕显示页文件最大行号数；I_0—下滚前页文件计数。

以采用上述的第二种方法为例来说明。

设下滚从当前屏幕所显示的第 I 个页文件、第 K_0 行开始。即当前屏幕显示的内容包括第 I 个页文件的前 K_0 行，第 $I-1$ 个页文件的 $m-K_0$ 行。

顺序将第 I 个页文件的 K_0-1 行内容，K_0-2 行内容，…，第 2 行，第 1 行内容送入显示缓冲区，再将第 $I-1$ 个页文件的后 $n-(K_0-1)$ 行内容，从末行开始依次送入显示缓冲区，并在屏幕上显示缓冲区内容。如果还要下滚，则修改 K_0 值，使 $K_0=K_0-1$。若 $K_0=0$，则令 $K_0=m$，再修改 I 的值，$I=I-1$。若 $I=0$，则令 $I=n$，继续上面的过程。屏幕上将逻辑文件不断滚动显示出来，直至输入滚动结束为止。

3. 换幕显示　换幕显示分为换上幕和换下幕两种。换幕显示程序处理方法较简单。换上幕处理程序是将当前在屏幕上显示的页文件的前一个页文件从存储区中取出，并在屏幕上显示。若当前处理的页文件是第 1 个页文件，则取出的页文件是第 n 个。换下幕处理程序是将当前显示于屏幕的页文件的后一个页文件从页文件存储区中取出，并在屏幕上显示。若当前处理的页文件是第 n 个，则调出的是第 1 个页文件。它们的处理流程省略。

4. 屏幕的左移和右移　屏幕的左移、右移分为屏幕的内容向左移动一列和向右移动一列，以及左移一幕和右移一幕。在编辑过程中，特别是对汉字表格文件的处理，往往逻辑文件的行宽大于物理文件的行宽，为了直观地将逻辑文件的一行或逻辑文件的全部内容直观地表现于屏幕上，我们设制了左移、右移子程序模块。

为了能做到屏幕的左移、右移，逻辑文件在存储区内的存放必须满足相应的要求。以逻辑文件的宽度为三个屏幕为例，其存储结构如图9-49所示。

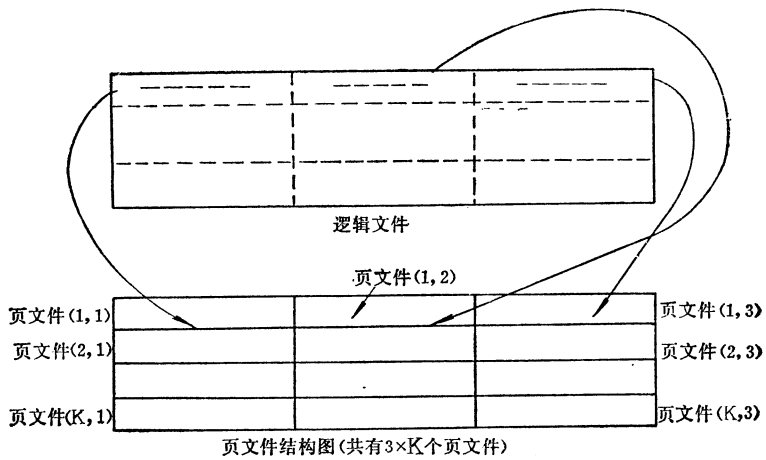


图9-49　多页文件结构示意图

若当前显示的屏幕是页文件（K_0，I_0），执行左移一幕显示时，取出页文件（K_0，I_0-1）在屏幕上显示。如果 $I_0=1$，则显示的将是页文件（K_0，3）的内容。同理，若 $I_0=3$，则在右移一页时，得到页文件（K_0，I_0+1）应为页文件（K_0，1）。在向下移动一页时，页文件（K_0，I_0）变为页文件（K_0+1，I_0）。若 $K_0=K$，则（K_0+1，I_0）改为页文件（1，I_0）。为了不失一般性，通常的情况是：$I_0\neq 3$，$K_0\neq K$。

左移、右移一列的处理方法和上、下滚动的方法类似，只是这里处理的对象是列。读者可以自行思考。

9.4.5 汉字文件加密、解密程序

随着计算机科学的迅速发展，多任务、多用户和多终端的出现，计算机的应用范围越来越广。在此状况下，若一个系统没有必要的保密措施，用户就可任意存取系统文件，修改系统软件模块，从而会造成严重后果，也不利于用户文件的保密。文件的保密情况可以分成几种等级。例如，一个人事档案系统，按照通常规定及设计原则仅允许档案管理人员读写和修改档案文件；某些赋于一定权限的人可读取档案目录，而不允许了解档案的内容；某些赋于特殊权限的人可读取档案文件的部分内容，而不允许他修改这些内容，更不允许他得到他不应该知道的档案材料。因此，为了确保文件的安全及保密，必须设计汉字文件加密、解密程序。

一、文件加密和解密的一般概念

计算机系统中，文件的内容可以有多种类型，一个源程序，一篇文章，一类报表，一批数据等都可组成一个文件。按用途，文件大致可分成三类：

（1）系统文件。除了系统程序外，各种系统信息所组成的文件。这类文件对用户不开放，只有系统才可调用。

（2）程序库文件。例如汉字信息处理系统中的一些公用程序（utility program），访问汉字字模库程序，汉字编辑程序，输入和输出服务程序等信息组成的这类文件，允许用户调用，但不允许修改或写入信息。

（3）用户文件。用户信息组成的文件属于此类（例如，源程序文件、用户数据文件）。通常，这类文件的所有者或授权者可对其进行存取或修改。

为了安全可靠，每个文件常被规定保密级别。文件按保密级别可分为四类：

（1）执行文件。用户可将该文件当作程序执行，但不能修改。

（2）只读文件。允许文件所有者或授权者读出或执行，但不准写入。

（3）读写文件。限定文件所有者或授权者可以读写。

（4）不保密文件。不作限制，可读可写，可修改或执行。

为了实现不同的保密级别，根据规定的访问性质（读或写），规定的使用权限（是否有权撤消等），可以采用相应的各种加密措施。

为了实现对文件不同深度的加密，即对文件的某个记录，甚至某个字段的保密，可以对记录和字段加密。

保密的级别有多种，加密的手段也有多种。有的由硬件设备提供加密措施；有的由操作系统提供加密措施；某些语言还提供了加密语句。有了加密手段及具体的实施办法，就必然会有解密的具体方法。例如，对某些用户是开放的，而对其他用户是保密的文件，则在该文件加密后，被授权使用的用户必须首先对文件实现解密，然后使用该文件，而未授权者就不知如何解密。

二、文件加密和解密的实现方法

文件加密和解密的方法有很多种，如修改文件名法、硬件加密和解密法、口令法、存取控制表、多级树形目录结构法和伪随机密码法等。

本节着重介绍几种主要的软件加密和解密方法。

（一）口令法

对每个汉字文件配置一组口令。如果允许某个用户访问这个文件时，就将口令通知他。当用户提出访问该文件的请求时，加密和解密程序首先核对口令。如果口令符合，则该用户可根据他的访问权限和使用权限，访问和使用该文件。这样做开销较小，实现方便。当文件经过一定时间，要回收某个用户的使用权时，必须更改口令；而更改后的新口令必须通知系统所允许使用的用户。为了简单明了起见，以程序A为例加以说明。

程序A：

```
          ⋮
110       B$ = "ABC"
120       DIM  A$
130       INPUT  A$
140       IF B$< >A$ THEN GO TO 500
150       OPEN FILE (1, 2)  "MAD"
          ⋮
500     END
```

程序A在运行过程中，首先于标号130处要求用户键入口令。在用户键完口令后，于标号140处测试用户回答的口令是否为“ABC”，若不是“ABC”，则不作任何响应；若回答正确，则该用户可以打开文件名为“MAD”的文件，从而可以处理加工该文件。

（二）伪随机机密码法

这种加密和解密法是采用一个“混合器”（加密过程）。输入到“混合器”的信息有两种。一种是需要保密的信息，即被加工的信息；而另一种是“伪随机密码”，它可以是一段易记的文章、诗、词或歌谱。根据不同的使用者，可以自行约定。解密法是用同一个“伪随机密码”经过“分离器”（解密过程）进行解密，以得到原来的信息。

此法不易泄密，混合后的信息量并不增加。然而，加密和解密过程的系统开销较大。当更换“伪随机密码”时，必须将新的“伪随机密码”通知它所授权使用的用户。

（三）存取控制表法

它是由汉字信息处理系统的全部用户和该系统所具有的全部文件组成的一个二维矩阵：

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ & \cdots\cdots & & \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}$$

矩阵中的每一个元素a_{ij}就规定了第i个用户对第j个文件的访问和使用权限。

例如规定：

$$a_{ij}=\begin{cases} 0 & \text{表示不可访问} \\ 1 & \text{表示可读} \\ 2 & \text{表示可写} \\ 3 & \text{表示可读可写} \\ 4 & \text{表示可进行包括撤销文件在内的一切操作} \end{cases}$$

当用户需使用某个文件时，文件解密程序首先根据用户名、文件名，得出它是该系

统的第几个用户，第几个文件。其后查二维矩阵，得出该用户使用和访问该文件的权限。但是，如果用户和文件较多，这个矩阵就很大，实现起来开销也较大。

（四）多级树型目录结构法

文件是“按名存取”的，而“按名存取”主要是用文件目录实现。一般说来，文件目录可放在内存区，用来登记文件和这些文件在文件存储器中的位置。文件目录不但能实现由文件名转换成物理地址的功能，而且还能提供对文件存取的控制和保密措施，以防止未经授权的用户及文件的所有者或授权者错误地使用文件。通常文件目录设置诸如下述的一些内容：

文件名；

文件所有者存取权限；

文件授权者存取权限；

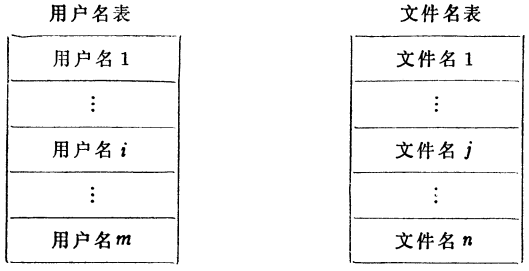
文件的物理位置。

凡获得某级目录表存取权限的用户，都能自动地获得此目录所属的全部目录或文件的存取权。这种方式实现简单，开销也不太大。为了明瞭起见，我们将在第四段用一个二级目录结构来说明这一方法。

以上叙述的各种文件加密和解密的基本思想及实现方法同样适合于对记录、字段的加密和解密。这里就不一一叙述了。

三、文件加密和解密的流程图

以上阐述了几种实现文件加密和解密的主要方法。但由于文件加密和解密的方法很多，往往一个系统仅仅提供一两种文件加密和解密的方法。为了叙述方便起见，仅以存取控制表法作为选用的方法来举例说明。为此，假定系统已建立了一个用户名表和一个文件名表，且该系统有m个用户，n个文件。

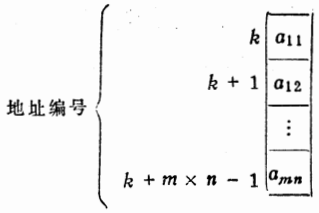


采用存取控制表加密就是在用户名表．文件名表的基础上，根据各用户对每一个文件的使用和访问权限，建立一个二维矩阵（存取控制表）来实现对文件的加密。其加密的具体流程如图9-50所示。

在加密程序运行后，则在存储区内建立起文件存取控制表（即二维矩阵）：

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & \cdots & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

其存储形式为：



当用户需要访问和使用文件时，解密程序首先询问用户姓名，并根据用户回答的姓名，在用户名表中查找有无此名的用户。若无此用户名，则拒绝该用户访问文件库；若有此用户名，则取出该用户名在用户名表中的编号，同时询问他需要访问的文件名。在用户回答文件名后，解密程序在文件名表中查找有无此文件名。若无此文件名，这说明文件库中无此文件；若有此文件名，则取该文件名在文件名表中的编号。其后，根据用户名编号和文件名编号，取出存取控制表中相应的数值，从而决定该用户访问和使用此

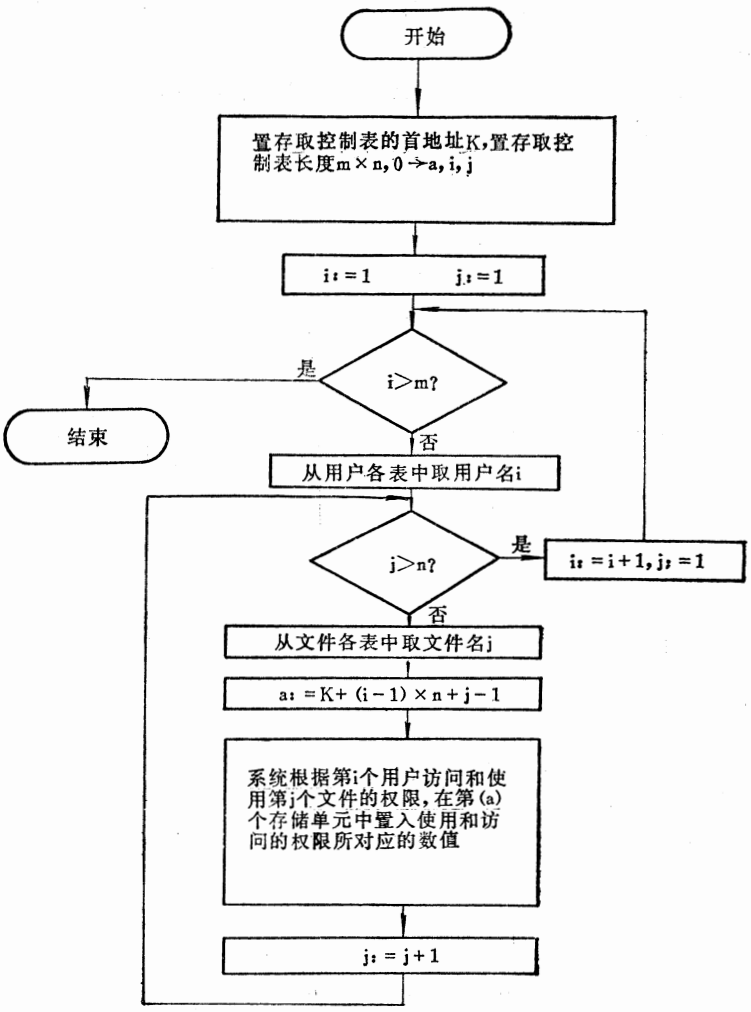


图9-50　建立存取控制表的流程图

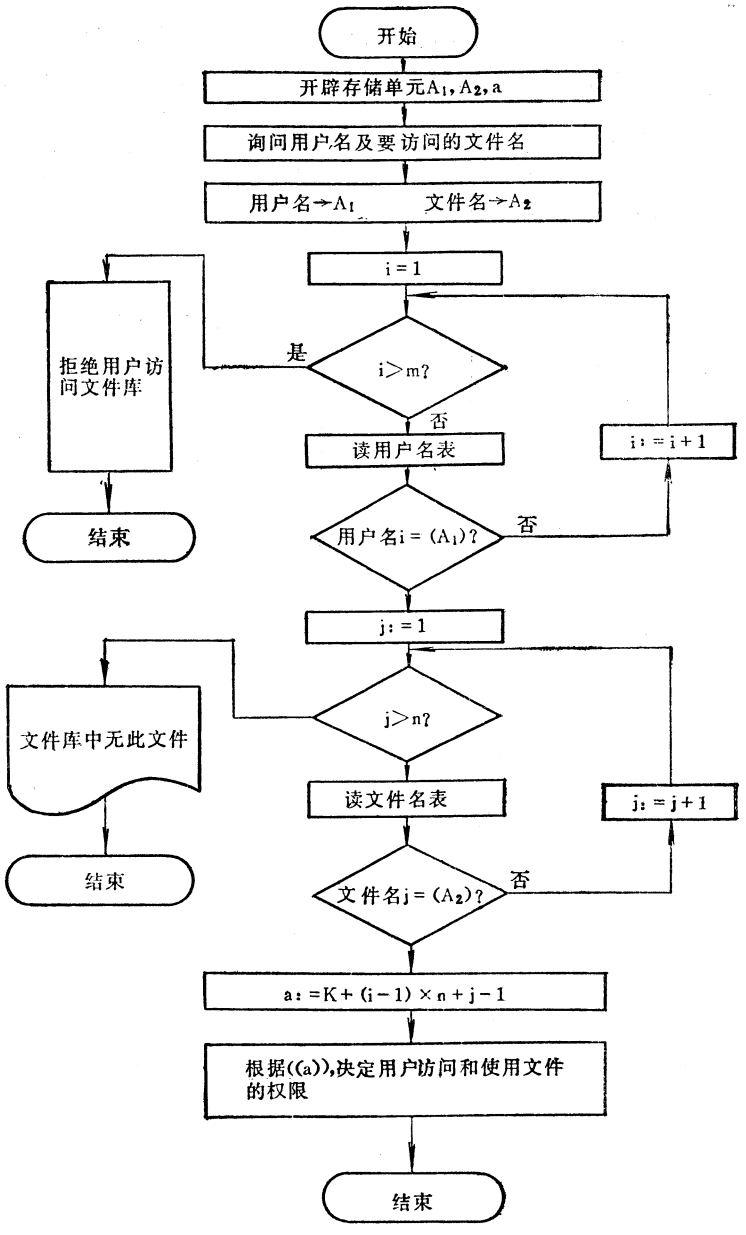


图9-51　解密程序的流程图

文件的权限。其解密程序的流程如图9-51所示。

四、文件加密和解密的系统方法以及应用程序员方法的实施过程与举例

尽管文件加密和解密的手段及方法有多种，但无论哪个系统都不可能提供很多种文件加密和解密的方法。一般是根据系统特点、用户要求、加密和解密的系统开销等因素来决定采取何种方法。因此，本段拟以叙述一般系统常采用的多级树形目录结构，并以二级文件目录为例加以说明。

为了实现对文件的加密和解密，系统必须建立一个主文件目录和一个用户文件目录。

如图9-52所示。

在用户需要建立文件时，首先判断该用户是不是系统允许的用户。若不是系统允许的用户，则拒绝建立文件；若是系统允许的用户，则在存储器中写入文件后，文件加密程序首先在用户文件目录中登录刚写入文件的文件名以及它在存储器中的物理位置，同时在主文件目录中登录用户名和文件名在用户文件目录中的地址。至此，系统登录了该文件的档案，即实现了对此文件的加密。

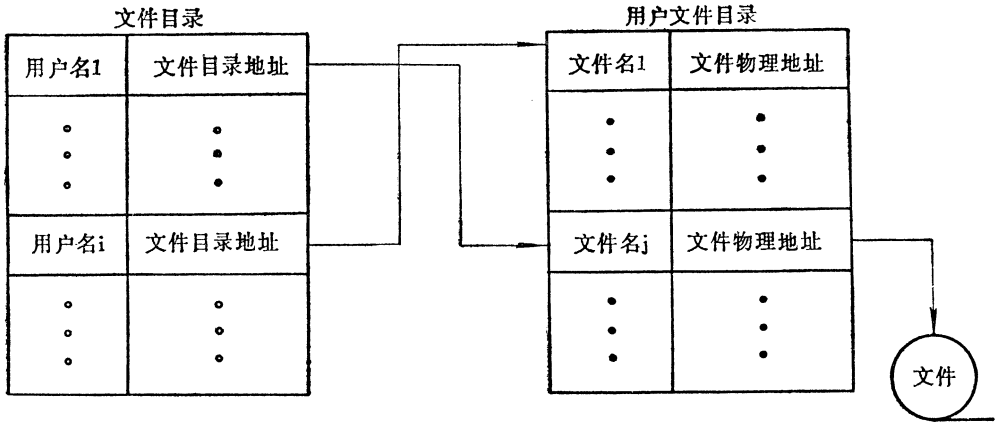


图9-52　主文件目录与用户文件目录

当用户提出访问文件的要求时，解密程序首先询问用户名以及该用户所要访问的文件名，由用户名查找主文件目录。若主文件目录中无此用户名，则拒绝该用户访问任何文件；若主文件目录中有此用户名，则取附在该用户名后面的该用户有权访问的那些文件的文件名在用户文件目录中的地址，再根据用户要访问的那个文件的名字，同刚才得出的那些地址中的文件名相比较。若有相同的文件名，则根据所指的文件物理位置去访问文件；若无相同的文件名，则说明该用户无权访问该文件，或该文件根本就不存在。

因为任何文件的存取都通过主文件目录。所以，采用二级文件目录对文件加密和解密，不但可以避免一个用户存取另一个用户的文件，使用户文件的独占性得到保证。而且不同用户具有同一文件名时也不会导致混乱。

除了系统提供对文件加密和解密的措施外，通常应用程序员也可以利用系统提供的保密功能，或利用高级语言中提供的保密语句对文件加密和解密。总之，这种加密和解密方法一般说来都较简单，较容易实现。

由于文件加密和解密的应用程序员方法如同系统加密和解密的方法一样有很多种，故不可能一一列举，仅以采用COBOL 语言中的某些语句实现对文件加密和解密的全过程来阐述应用程序员方法。为此，我们引用以下程序（程序名B）：

⋮

数据部分：

```
FD   SAMPLE-FILE
     RECORDING  MODE  IS  F,
     BLOCK  CONTAIN  5  RECORD,
```

```
     RECORD CONTAIN 100 CHARACTER,
     LABEL RECORD IS SAMPLE-LABEL,
     DATA RECORD IS SAMPLE-RECORD,
01   SAMPLE-LABEL
  02 LABEL-ID PICTURE X(4)
  02 LABEL-INFO PICTURE X(76)
01   SAMPLE-RECORD PICTURE X(100)
     ⋮
PROCEDURE DIVISION.
DECLARATIVES.
L  SECTION. USE BEFORE STANDARD FILE LABEL
   PROCEDURE ON SAMPLE-FILE.
   ⋮
END DECLARATIVES
M  SECTION
   OPEN OUTPUT SAMPLE-FILE
   ⋮
```

程序*B*在文件描述体中使用“LABEL RECORD子句”说明SAMPLE-FILE文件，采用用户自己建立的标号，而不是系统的标准标号。因为任何文件在使用前必须打开文件，使用后必须关闭文件，所以，在打开SAMPLE-FILE时，系统自动地转入申明节中“USE子句”所引出的用户标号处理过程的执行，以建立具有数据名为SAMPLE-LABEL的文件头标。因为任何文件的读取，首先是打开文件核对文件头标，而在此程序运行后，用户自己建立了具有非标准文件头标的文件，未授权者并不知此文件头标内容，所以，只能由该用户或授权者读取此文件。显然，“USE子句”引出的用户标号处理过程即为文件的加密过程。

倘若将程序*B*中“OPEN OUTPUT SAMPLE-FILE”改为“OPEN INPUT SAMPLE-FILE”（省略号所隐含的那部分语句改动除外）得程序*C*，那么，当执行打开已建立的SAMPLE-FILE文件时，系统自动地转“USE子句”所引出的标号处理过程，核对标号。若标号有错（即标号对不上），则转入出错处理，拒绝用户使用此文件；若标号无误，则回到打开例程，从而可以读出该文件。由此可见，程序*C*中的“USE子句”引出的标号处理过程实际上是文件解密过程。

第十章　汉字数据处理的系统软件

10.1　什么是系统软件

现以 IBM370 计算机的操作系统为例，简单地说明一下什么是系统软件。

操作系统由两大部分组成，即控制程序和处理程序。如图10-1所示。

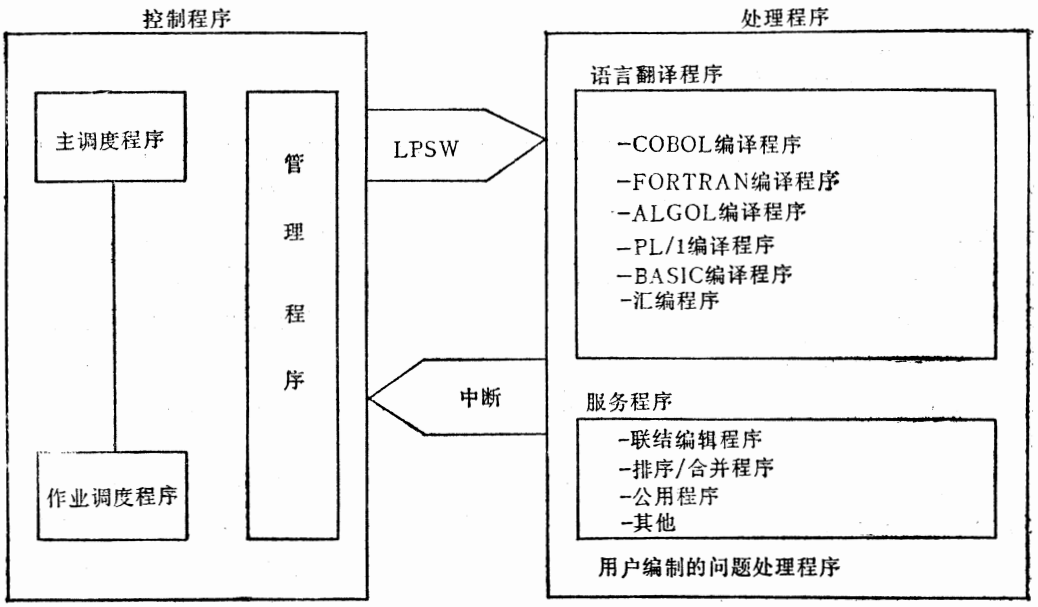


图10-1　操作系统的主要构成部分

控制程序是非生产性的程序，是监督、管理处理程序的程序。管理程序、主调度程序和作业调度程序等都属于控制程序。

管理程序是操作系统的核心部分。它是由各种中断分析程序、中断处理程序组成。其中包括数据管理程序和任务管理程序。当然也包括输入输出管理程序。因此，本书第九章所涉及的内容，从操作系统的角度来看，都属于管理程序部分。管理程序属于系统程序。我们把一些内容列入第九章，是为了让读者对汉字的输入输出有一个感性的认识，从内容上说，它当然也是属于汉字数据处理的系统程序。

主调度程序是主管操作员和系统之间的通信的，通常又把这个程序叫做命令解释程序。主调度程序是系统和操作员之间联系的桥梁。操作员通过控制台键盘打入各种命令，对系统进行控制和管理，反过来系统将运行的情况通过该程序在控制台键盘上输出。

作业调度程序把用户编制的作业按某种优先次序一个一个地进行处理。关于主调度程序和作业调度程序我们只是作了如上感性的介绍。

处理程序是生产性的程序。

程序设计语言可分为四种类型：面向过程的语言，面向问题的语言，面向机器的语言和面向系统的语言。FORTRAN 语言和 ALGOL 语言是用于科学计算的主要语言。PL/1 语言既能用于数据处理，也能用于科学计算。它们都是面向过程的语言。

RPG 语言和BASIC语言是面向问题的语言。RPG语言用于报表等数据处理；BASIC语言常用于远程终端，在微型机上也使用得较为广泛。

汇编语言可分为简单汇编语言和宏汇编语言。

简单汇编语言是面向机器的语言。

宏汇编语言中的宏指令分为系统定义的宏指令和用户定义的宏指令。我们在这里不谈用户定义的宏指令。常用的系统定义的宏指令如 OPEN，CLOSE，GET，PUT，READ，WRITE等。

宏汇编语言是面向系统的语言。具体地说是面向一个具体的操作系统的语言。每一个操作系统对系统宏指令的功能都有具体的规定，所以一个操作系统功能的强弱完全取决于该系统宏指令功能的强弱。系统宏指令对了解操作系统的结构和使用操作系统都有重要的作用。

如图10-1所示，在IBM370计算机系统中处理程序和控制程序（特别是管理程序)之间的联系是通过中断来实现的。由管理程序进入处理程序要通过一条 LPSW 指令。

控制程序好比是管理部门，处理程序好比是生产部门。生产部门通过“中断”把生产和处理的情况及时上报管理部门——控制程序；而管理部门又及时把指示下达（通过LPSW指令）给生产部门——处理程序。如此往复循环，以达到提高生产效率（在计算机系统中叫吞吐量）的目的。

那么，究竟什么叫系统程序或系统软件呢？

所有的控制程序、所有的语言翻译程序、所有的服务程序，包括联结编辑程序，排序/合并程序，公用程序等都属于系统程序。

用户编制的问题处理程序则不算系统程序。

但是，系统程序和非系统程序之间并没有一条严格的界限。甚至可以这样说，一切由制造者提供的具有较为普遍使用价值的程序都可算系统程序。用户编写的程序一般不属系统程序，但是一旦用户编写了一个通用性很强的程序，具有普遍使用价值，可以提供给其他用户使用，这样的程序就完全可以列入公用程序，因此也可算是系统程序。

10.2 汉字数据处理问题的提出

近年来我国汉字数据处理工作取得了很大的进展，在不同的计算机上建立起来的具有各种特性的汉字处理系统，为计算机的推广应用起了很好的作用。这些汉字数据处理系统有的建立在应用程序这一层，有的建立在数据库的基础上，也有的直接建立在操作系统上。涉及的范围有：汉字报表格式的编排和打印；汉字文本编辑；汉字数据的存储、检索和查询等。但是，如何加速汉字数据处理技术的推广和应用，除在理论上要进一步研究之外，在实践上还有许多工作要做。

目前，至少有两类关于如何建立汉字数据处理系统的思想。

一些人认为：要建立汉字操作系统，定义汉字数据类型，建立和使用对应于各种使用面较广的程序设计语言的汉字程序设计语言，使用汉字命令等等。

另一些人提出一种向上兼容扩大汉字功能的设想。他们认为，汉字的输入输出只不过是一个汉字的用户界面。此外，汉字在计算机内部也一定有一种代码表示，不管用什么程序设计语言，都能处理这种代码。因此，只要把国际上普遍行之有效的操作系统加以扩充，引进汉字输入输出和处理功能，使各种程序设计语言都能用来调用和处理汉字就行了。

其所以提倡采用国际上普遍使用的操作系统，是因为这种操作系统的信息资源很丰富，如果能把这些资源全部或大部分搬过来为汉字数据处理服务，将大大节省人力和物力。不仅如此，这种操作系统的潜在资源更丰富，世界各国的程序员都在不断地为之增添新的软件资源。一旦把汉字功能扩充到这类操作系统，那么这种不断增加着的潜在软件资源也会为汉字数据处理服务了。

所以，后一种设计思想更能切合实际需要，易为研制部门和广大用户所接受。

10.2.1 汉字与西文兼容问题

汉字数据处理要解决的问题实质上就是一般的西文（实际上常指英文）数据处理所要解决的问题。汉字，作为一种数据类型在计算机内部的表示，它同字母数字（也是一种数据类型）在计算机内部的表示没有什么本质上的区别。汉字数据在磁盘媒体、磁带媒体中的表示，在通信媒体中的表示同西文字母数字在相应媒体中的表示都没有什么本质上的差别。下一节将说明汉字数据和西文数据在计算机内部的代码表示的各种方式。汉字和一般的字母数字的最大差别在于用户界面——键盘输入、显示输出，印字输出。这些问题在前面几章中已经详细叙述了。

本章主要涉及汉字数据处理问题，要强调的是汉字数据和字母数据的共性。目前各种程序设计语言几乎都有“字符的”或“字符串”的数据类型。借助于这些数据类型，都可用来描述和定义汉字数据。这样我们就可参照用西文写的各种程序设计语言的文本来书写汉字数据处理的程序。此外，对于原有的各种程序设计语言的编译程序也无须作太多改动，就可适于编译用西文程序设计语言写的、用来处理包含汉字信息的源程序。

以下我们要举例说明如何用西文的程序设计语言来编写处理包含汉字信息的源程序的问题。先讨论汉字和一般的字母数字的共性问题——在计算机内部的代码表示。

10.2.2 汉字和字母数字在计算机内部的表示

严格说来，计算机既不能直接处理汉字，也不能直接处理字母数字。所谓处理字母数字数据，实际上是处理字母数字代码——GB1988码（与ASCII码兼容）、ISO码或者EBCDIC码。同样，处理汉字数据，在计算机内也就是处理汉字代码。

在计算机内部不能直接表示A, B, C, D, …。

这些字母，表示成GB1988码，写成十六进制的形式，在计算机内部的表示则是：

41，42，43，44，…。

同样，若把这些字母表示成EBCDIC码，写成十六进制的形式，在计算机内部的表示则是：

C1，C2，C3，C4，…。

如何在计算机内部表示汉字代码是首先要解决的一个课题。GB1988中的字母，只

有 26 个。而汉字的数目有几万个，因此，用代码表示汉字要困难得多。由于我国已颁布了通信用汉字字符集（基本集）及其交换码国家标准 GB2312，因此，汉字的机内代码也可以参照 GB2312 码来设置。

使用GB 2312 交换码，可以将汉字“啊，阿，埃，挨，…”表示成计算机可以识别的十六进制代码，其代码为：

3021，3022，3023，3024，…

每个汉字的代码，用了两个字节。如果直接用 GB2312 码来表示汉字，则会和 GB1988 码相混淆。因为把上述四个汉字的代码，对应于 GB1988 码的字符则为：

O!　O′′　O#　O¥

为了要把GB1988代码或EBCDIC代码和汉字代码区分开，就要想一些办法。下面举例说明汉字代码的表示方法。这里所说的汉字代码既可以用于汉字的机内表示（在内存中的表示），也可用于汉字在磁记录媒体（硬盘、软盘、磁带）中的表示，还可用于汉字的数据传输形式。因为各种计算机系统都有自己的个性，数据传输方式也各有其特点，因此，即使在同一计算机系统中汉字的机内代码和汉字的传输代码也可以相同，并且要和 GB2312 交换码有简单明确的一一对应关系，使得将来不同的系统之间进行汉字数据交换时有章可循。

下面列举一些汉字的代码表示方式，无非是给读者以感性的认识，并不是说就不能设置新的汉字代码了。

一、自名表示法

如图10-2所示，1 个字节有 8 位，可有 256 个代码，但GB1988码只使用了前面 7 位，共128个代码，即最高位是冗余位。这样我们就可用第 8 位为 0 的 1 个字节表示1个GB1988字符；用第 8 位为 1 的 2 个字节来表示 1 个汉字。如果表示汉字的 2 个 7 位码采用了GB2312代码，则这种自名表示法与GB2312 码建立了最简单的一一对应关系，而且也决不会和GB1988 码混淆。

例如上面提到的汉字“啊，阿，埃，挨”，若用上面的方法表示成汉字代码，则是：

B0A1，　B0A2，　B0A3，　B0A4。

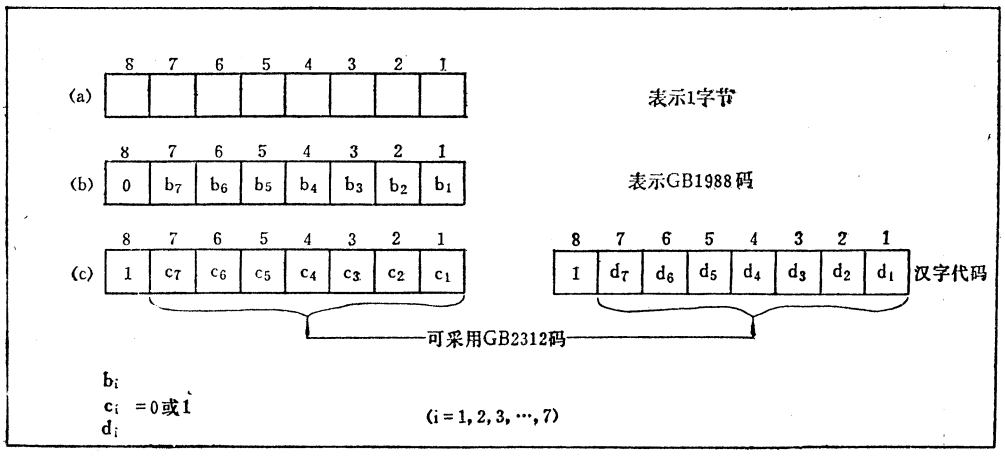


图10-2　GB1988码和汉字代码

显然，这跟 GB1988 码不会相重。

因为是用同一字节中的某一位来区分 GB1988 码或汉字代码，不必另加区分符，所以叫自名表示法。由于不必另加区分符，就能区分汉字数据和字母数字数据，故在数据处理的字符串合并、查找、比较等运算时，会带来很多方便。

这种表示法的不足是由于第 8 位要置 1，因此两个字节的信息最多只能表示：

$$128\times128=16384$$

个汉字。然而，GB2312 代码在 7 位代码的128个代码中只用了 94 个。为了使汉字代码和 GB2312 码建立最简单的一一对应关系，这种表示法一般只能表示：

$$94\times94=8836$$

个汉字。一般说来，这已经够用了，因为这个数目比 GB2312 所包含的一级汉字和二级汉字总数还要多。

现在，再讨论用 COBOL 语言是如何表示这些数据的。现实情况中，我们常常遇到一种字母和汉字混合的数据，如“学生A”，由 2 个汉字和 1 个字母组成，共占用 5 个字节，用COBOL语言来描述这个数据，其形式为：

```
01  XY.
  02  ANAME  PIC  X(5)  VALUE  IS "学生A".
  ⋮
```

现在的微型机系统中，终端和处理机的连接多半采用通用的 RS232 接口，而终端驱动模块往往采用 7 位的输入输出码（第 8 位是作为奇偶校验用的）。在这种情况下，如果打算不修改终端驱动模块而接收和传输汉字代码，则就不能采用上述方法了。

二、控制字符区分法

如图10-3所示，用 控制字符 a 开头，后跟GB1988 码（或EBCDIC码）表示字母数字字符（或 EBCDIC 码字符）序列。用 控制字符 c 开头，后跟 2 个字节一组的代码序列表示汉字序列。

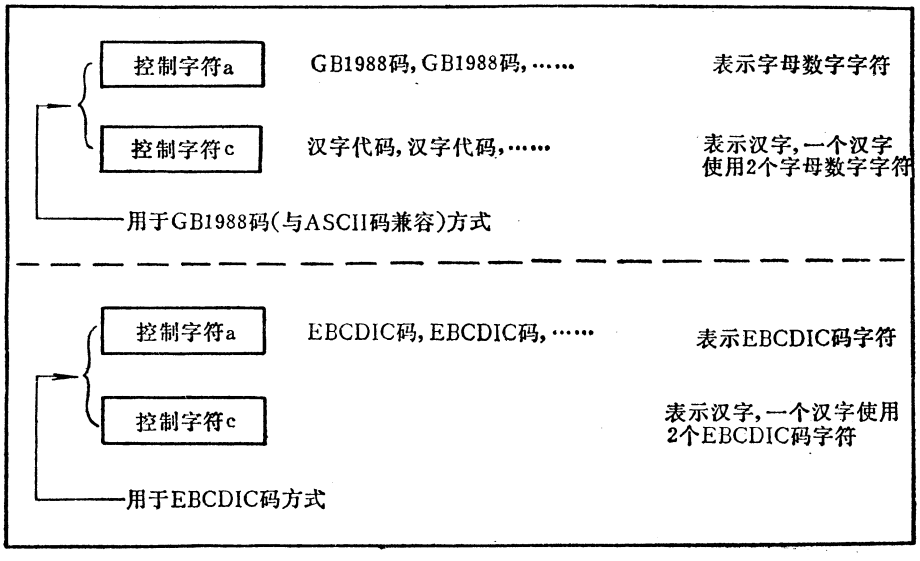


图10-3　用控制字符区分汉字和字母数字

现在再来看“学生A”这个数据。若采用以上汉字代码，则该数据在计算机内的表示为：

“ 控制字符c 学生 控制字符a A”

这个数据共使用了7个字节。用COBOL语言来描述该数据，其源程序的形式为：

```
01  XY.
  02  ANAME  PIC X(7)  VALUE  IS “学生A”
  ⋮
```

由于控制字符不能印出，也不能显示，因此它在源程序中的形式仍为“学生A”。但是这个数据在计算机内部占用了7个字节。而且，在计算机内部的表示也不是唯一的，也可以表示为：

“ 控制字符c 学 控制字符c 生 控制字符a A”

这样在计算机内部就占用了8个字节，其输出形式仍然是“学生A”。

这种表示法的优点是表示的汉字个数比前一种方法要多。若采用GB1988码，则由于要去掉32个控制字符、空格符和DEL符，第8位可以是1也可以是0，因此可以表示：

$$(256-68)^2=188^2=35344$$

这种表示法的缺点是一个汉字要用2个以上的字节（把控制字符也计算在内）才能表示。此外，由于采用了控制字符区分，同一汉字数据在计算机内部的表示也就不唯一了。这样，会给数据运算带来不少的麻烦。

三、控制字符作为括号的方法

如图10-4所示，用一对控制字符作为括号将汉字代码序列括在里面，表示汉字序列。而GB1988码序列或EBCDIC码系列则不需要任何控制字符来区分，也不需要一对控制字符作括号。

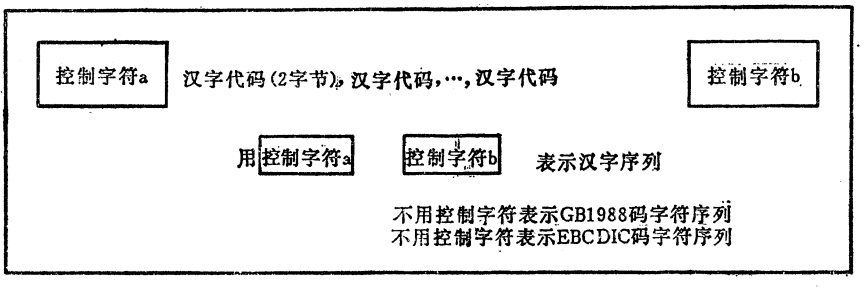
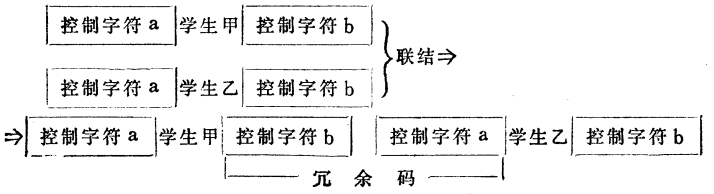


图10-4　用控制字符作为括号的汉字表示法

这种表示方法的优缺点大体上和第2种表示方法相同。尤其在汉字数据作联结运算时，结果显得不太自然。例如将“学生甲”和“学生乙”作联结运算：



上述结果中的两个控制字符是冗余码，若不把这2个控制码去掉，则显得累赘；而要把这2个码去掉，则需要在联结运算时增加特殊的措施。

四、用一对图形字符作为括号的汉字表示法

用一对图形字符作括号的方法，如图10-5所示。

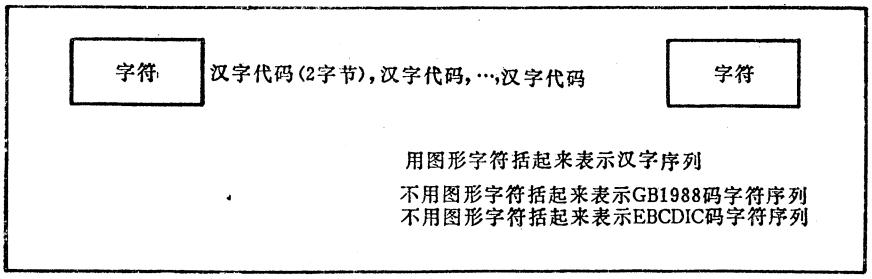


图10-5　用一对图形字符作为括号的汉字表示法

这种表示方法类似于第三种，所不同的是这里采用了图形字符作为括号。这种方法的缺点是源程序的可读性比前述的三种要差。这是因为，每当汉字出现在源程序中时，都有一对字符来定界。而且用作括号的字符最好不再作别的用处了，否则会有二义性。

但是，这种方法比起第三种方法却有一个优点。我们知道，不同的操作系统中，各种终端驱动模块对于控制字符都有各自的具体的解释和用法。因此，在使用第三种方法时，把什么控制字符作为括号，要特别小心，以免混淆。但是采用这种方法，则不会出现这个问题，因为没有用控制字符作括号。

五、用4个图形字符表示一个汉字

用4个图形字符表示一个汉字，其中第一个字符恒为一固定的不常用的图形字符。这种表示方法的缺点是用4个字节表示一个汉字太浪费。然而它也有不少优点，在汉字数据运算时，如在查找、合并、比较时，一般不会有二义性。此外，因为只用了7位码，故在一般的微型机多用户系统中，可顺利地通过终端驱动模块。

六、用3个或4个大写字母和数字表示一个汉字

因为有的小型机甚至某些大型机在连接终端设备时，系统只允许从终端进入大写字母和数字，因此这种方法也是有实际意义的。

这种方法的缺点是要3个以上字节才能表示一个汉字，开销很大。但是它也有第五种方法所具有的同样的优点，因为它不需要区分符，在汉字数据运算时一般不会有二义性。

上面我们简单地列举了一些汉字数据在计算机内部的表示，即汉字代码的若干种形式，究竟采用哪种形式为好，这取决于究竟是在什么样的计算机系统的环境中使用。例如有的计算机系统只能接收7位代码，输出7位代码，因此，在数据输入和输出时，不能采用第一种表示方法。但是即使是这样的计算机系统，在其内部的汉字代码表示中，仍可采用第一种表示方法。也就是说，在同一系统里，在输入输出时，在内部运算时以及在数据通信时，汉字数据可以采用不同的代码表示形式。

那么，在一个具体的计算机环境中，选择何种代码形式为好，有哪些基本原则可以

遵守，我们提出了以下几点，供读者参考：

（1）要以尽可能少的字节数表示尽可能多的不相同的汉字；

（2）和GB2312码有简单的一一对应关系；

（3）在汉字数据（包括字母汉字混合数据）运算（查找、合并、截断、比较、替换、传输等）时，不容易产生二义性和不确定性。

10.3 汉字数据处理系统软件建立的方法

近几年发展起来的汉字数据处理系统有各种各样的支持环境：有的是在不增加外部设备的条件下，纯粹用软件的办法建立汉字处理系统；有的在系统中增加了汉字输入输出模块，然后通过某一高级语言来调用它；有的把汉字输入输出模块放在操作系统的输入输出管理程序中，并使之融合成一个整体。

此外，这些汉字数据处理系统也有建立在微型机上的，也有建立在小型或中型机上的。

首先，我们从汉字系统软件的发展阶段，来看一看汉字系统软件建立的方法。

10.3.1 汉字数据处理系统软件建立的发展阶段

一、从裸机上进行汉字输入输出试验

在七十年代中后期，国内就有些单位在国产计算机上尝试进行汉字的输入输出。这个阶段有两个主要特征：一是以研究和试验各种汉字输入方案（尤其是编码方案）为主；二是在汉字处理软件方面，通常是另外编制一套独立于原有操作系统的“汉字管理程序”来实现汉字的输入、输出和各种汉字文件报表的打印；在这类系统中，对汉字的处理一般只能使用汇编或机器语言，而计算机系统所配的各种高级语言还不能直接调用和处理汉字。

八十年代以前，建立汉字数据处理系统的工作进展是较缓慢的。直到微型计算机引入国内，情况才起了很大的变化，因为随着微型机和大规模集成电路的引入，汉字存储库和汉字显示器都容易解决了。于是，过去在中、小型机上进行的汉字输入输出试验，完全可以移到微型机上来进行，而且在微型机上做试验更容易。当然，这时候的工作不仅是软件的工作，还有不少硬件的改造工作，例如改变显示器控制器的线路，使之能显示汉字等等。上述工作大体上是八十年代初开始的。

由于多数微型计算机都带有BASIC语言，因此用BASIC语言直接调用汉字的输入输出（I/O）模块，一开始就成为国内、国外所关心的课题。

二、高级语言直接调用汉字I/O模块

作为例子，以下说明BASIC语言如何调用汉字的I/O模块。这样的系统需要具备哪些最少的硬件条件？如图10-6所示的为具有汉字I/O功能的微型机系统配置情况。图中，

CPU：主频为4兆赫的Z-80A

内存：64K字节静态MOS存储器

CRT：全点阵图象CRT，横向640光点×纵向800光点。刚好与64K字节RAM的刷新存储器相对应。

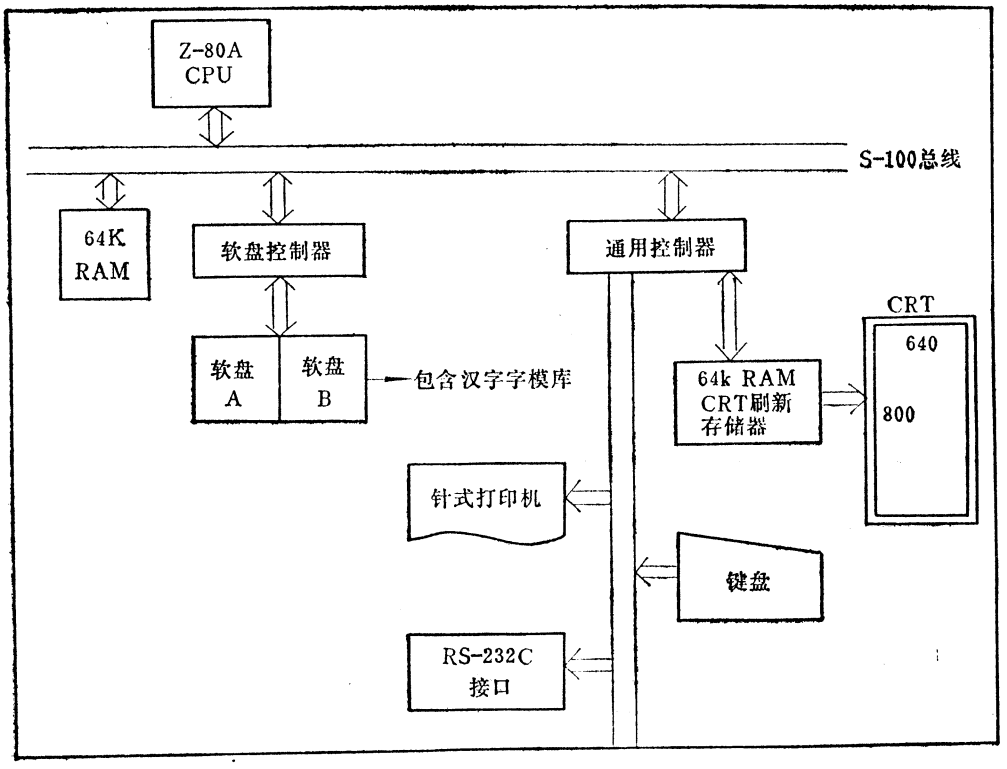


图10-6　具有汉字 I / O 功能的微型计算机配置图

键盘：普通的字母数字键盘。

软盘：200毫米（8英寸）单面双密度。128字节/扇区，52扇区/磁道，77磁道/软盘。

打印机：24针式汉字打印机。

先讨论内存、刷新存储器，和CRT之间的对应关系(图10-7)。CRT中的光点可分为16页，每页640×50个光点，与刷新存储器的16页中的1页相对应，每页4K字节。

4K字节共有32768位。

CRT每页共有32000点。

因此，刷新存储器中每页中余下的96字节除用作控制字节外，其他的是冗余字节。另外，内存中的最高4K字节与刷新存储器的某1页相对应，究竟对应哪一页由指令控制。

每个汉字所需的点阵为25×16，每个字母数字的点阵为25×8。实际上有7点为行之间的空白，因此实际汉字点阵为18×16，字母数字字符点阵18×8。CRT中1页可显示2行，每行40个汉字或80个字母数字。

字母数字点阵存放在西文汉字显示印字驱动模块中；而汉字字模点阵则存放在软盘B中。1个汉字的点阵需要36个字节来表示，因此，每个扇区可放3个汉字点阵。

用全点阵图形方式缓存汉字显示点阵和字母数字点阵、用软盘来作汉字字模库是陈旧的方法。我们是为了说明其中的原理而这样做的。这样做还有一个好处，即说明汉字输出和字母数字输出，并无本质上的区别，因为都可以看成由点阵构成的图形。

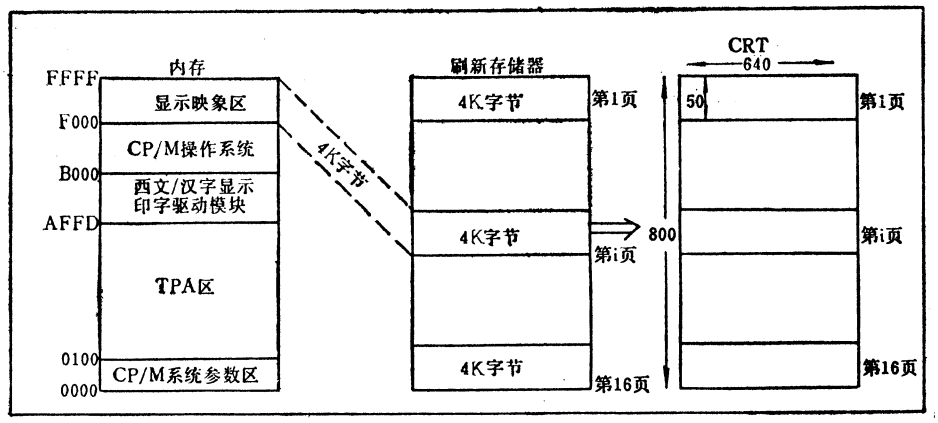


图10-7　内存、刷新存储器CRT之间的对应关系

图10-7中的TPA的全文是 transient program area，即临时性的程序区。编译程序、BASIC 解释程序以及用户程序都是在TPA上执行的。

图 10-8 是 BASIC 程序显示输出一个汉字的例子。下面我们逐句解释它们的作用。

```
100  A¥ =: "127164"
110  CALL& B000 , VARADR(B¥), VARADR(A¥)
120  CALL& B024
130  CALL& B006, VARADR(C¥), VARADR(B¥)
140  CALL& B00F, 2, 1
150  CALL& B01B, VARADR(C¥)
160  CALL& B01E
170  CALL& B021
180  END
```

图10-8　BASIC程序显示输出一个汉字的例子

100 句：将“电”的键盘输入码作为字符串赋给A¥，其中A¥的长度为 6 字节(本输入方案是用汉字三角号码法、每个汉字占 6 字节)。

110句：将A¥中代码转换为B¥中的字库地址码。其中B¥的长度为 2 字节。用 16 进制表示的B000为程序模块入口地址，参照图 10-7 可知，该地址属于西文汉字显示印字驱动模块。

120句：将屏幕上的光标消去。

130句：根据 B¥中字模库地址找出字形送 C¥。该语句涉及到从软盘B上取字形，C¥占36字节。

140句：将显示位置定于第二行第一列。

150句：在当前显示位置显示一个“电”字。

160句：将显示指针位置推移到下一位置。

170句：在当前显示位置上显示光标。

上面介绍了如何用 BASIC 语句来直接调用汉字 I/O 模块的例子。上述办法有两个很大的缺点：通过CALL语句来进行汉字的输入输出显得很不自然，也不方便；上面我们已经说到汉字输出和字母数字输出在本质上是没有什么区别的，那么为什么对字母

数字的输出是用PRINT语句，而汉字的输出却要用 CALL 语句呢？

是否可以把汉字的输入输出和字母数字的输入输出都通过原来的程序设计语言的I/O语句，如INPUT、PRINT、READ、WRITE、ACCEPT和DISPLAY来实现呢？这是完全可能的。

为了说清楚这个问题，先要分析一下BASIC语言中的PRINT "A" 是如何在显示屏上输出一个"A"的。其基本的步骤如下（请参阅图10-9和图10-7）：

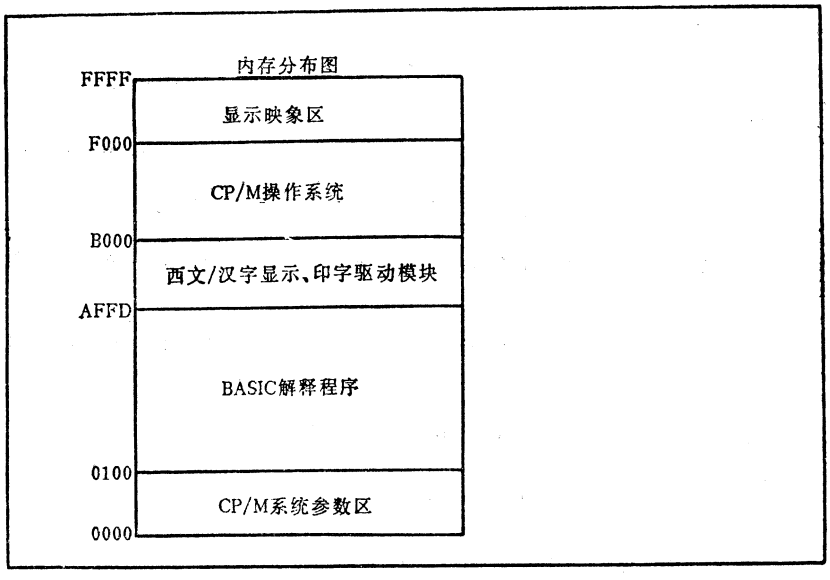


图10-9 执行BASIC程序时的内存分布

（1）BASIC解释程序将PRINT"A"语句加以解释，知道是要在显示器上输出一个"A"字，于是将这件工作委托给CP/M操作系统；

（2）CP/M操作系统将英文字 "A" 的输出任务委托给西文/汉字显示驱动模块；

（3）西文/汉字显示驱动模块将 "A" 字的字形（18×8的点阵）送到显示映象区。硬件机构又将 "A" 字点阵送入刷新存储器，于是就在显示器上显示一个 "A" 字；

（4）CPU的控制权从西文/汉字显示驱动模块返回CP/M操作系统，再从后者返回BASIC解释程序；

（5）BASIC 解释程序解释执行下一语句。

弄清了 BASIC 程序输出一个字母的内部过程后，再来看看图10-8例子中的BASIC程序是如何输出一个汉字的。我们只考虑第 150 句。BASIC 解释程序将输出 "电" 字的工作经解释后直接委托给西文/汉字显示驱动模块的某一子模块，该子模块的入口地址是 16 进制的 B01B 。以后的过程就和输出西文字母类同。

从上面的分析可以看出，在上述系统中输出西文和输出汉字唯一的不同点是后者的输出是不经过CP/M操作系统的。因此将上述系统稍加改进，不用 CALL 语句来调用和输入输出汉字，而是利用原有的输入输出语句，完全有可能通过CP/M 操作系统的输入输出模块来处理汉字。但是要注意以下两点：

（1）要将汉字显示、印字驱动模块和CP/M操作系统中原有的输入输出模块有机

地结合起来。其结合方式参照西文显示、印字驱动模块和CP/M操作系统中的输入输出模块的联结方式；

（2）要将字母的代码跟汉字的代码明显地区分开来，以便让西文/汉字显示、印字驱动模块很容易识别和区分它们。

下面一节介绍如何将汉字输入输出模块纳入CP/M操作系统的输入输出模块，使各种高级的、低级的程序设计语言都能使用原有的输入输出语句来输入输出汉字和字母数字。汉字和字母数字的输入和输出，所使用的是同一语句，汉字和字母数字的差别仅仅在于代码不同。

三、汉字I/O模块纳入操作系统的I/O管理程序

为了把操作系统上建立起来的各种软件资源，包括各种程序设计语言，公用程序(utility)和数据库系统(data base system)，都能用来为汉字的数据处理服务，最好的办法是把汉字的输入输出驱动模块纳入操作系统中的输入输出管理程序。我们以微型机上的CP/M操作系统为例来说明这个方法(图10-10)，它是具有普遍性的。

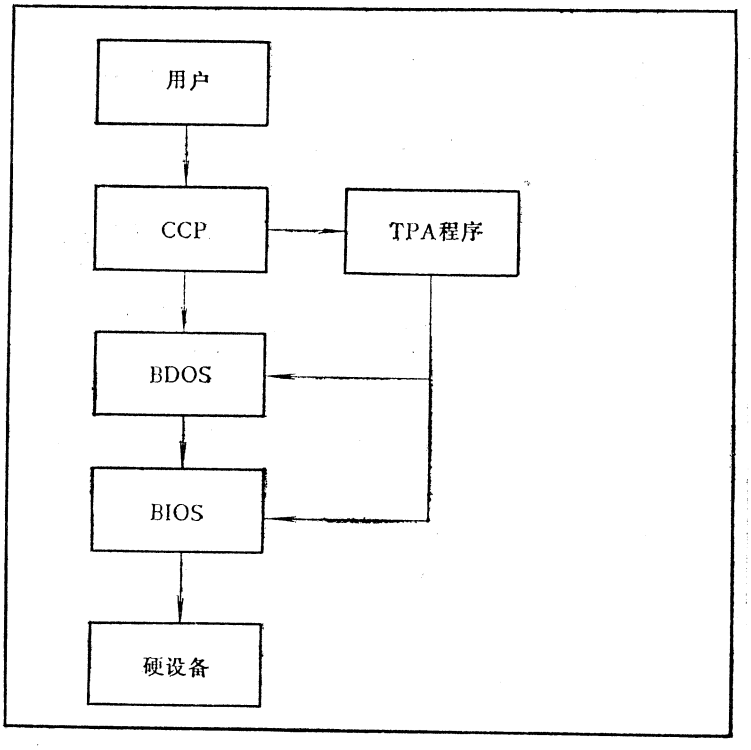


图10-10　CP/M操作系统结构图

在CP/M操作系统中，模块采用层次结构，即整个操作系统分成若干个基本程序模块，这些程序模块排成若干层，各层之间只能单方向依赖，不构成循环。因而具有比较好的可靠性、易懂性和适应性。整个系统的正确性可以由各层模块的正确性来保证。对整个系统的了解变成对各个局部模块的了解。图中的CCP为命令解释程序；TPA程序就是用户程序，也包括各种编译程序；BDOS是CP/M操作系统的基本控制部分，后面要提到的系统调用就是由BDOS来处理的。BIOS是CP/M操作系统的I/O管理程序，

用来直接驱动各种外部设备。各个模块之间的单箭头的具体含义是：用CALL指令进入，一定用RETURN指令返回。不能转入别的模块，以防止造成层次的混乱。

为了要弄清楚怎样把汉字 I/O 模块加入到 I/O 管理程序中去才能达到我们的目的，首先应当了解CP/M操作系统中的数据流。

在CP/M操作系统中，用户程序的每一个输入输出要求，都是通过系统调用来实现的，BDOS模块分析这些系统要求，然后命令BIOS模块来完成这些要求。再由BIOS模块直接启动外部设备。此后，BIOS 模块将控制返回 BDOS 模块；当用户提出的系统调用全部完成后，再将控制返回用户程序。

因此，在CP/M操作系统中，BIOS模块也就是一般所说的 I/O 管理程序。它直接承担着测试和启动外部设备的任务。

要使该系统具有汉字信息处理能力和汉字输入输出能力，首先要在系统中增加如下的设备：具有汉字显示能力的显示器；具有汉字输出能力的印刷机；汉字字模库。图10-6给出的系统配置是符合这一要求的。

如果我们使用普通的字母数字键盘进行汉字输入，而不是通过笔触式汉字字盘进行汉字输入，则上述三类设备就够了。

于是就需要在CP/M操作系统的BIOS模块上增加三个子模块：

（1）汉字印刷输出子模块；

（2）汉字输出并显示（在CRT上）子模块；

（3）汉字键盘输入并显示（在CRT上）子模块。

以上三个模块总称为汉字驱动模块（Chinese character drive module）。

第（2）个模块和第（3）个模块是不一样的，第（3）个模块要调用第（2）个模块。前两个模块比较简单，纯粹是输出。第（3）个模块比较复杂，是交互式的模块，既有输入又有输出，不仅如此，还涉及汉字输入编码（键盘输入汉字时采用的代码）到汉字代码（机内表示方式）的转换。

把这三个模块加到 CP/M 操作系统的 BIOS 模块中去，并把汉字的输入输出和西文的输入输出统一起来考虑，把相应的两类程序模块联系起来就构成了图 10-11 所示的汉字CP/M操作系统了。

从操作系统的结构上看，如果把 BIOS 和汉字驱动模块作为两个模块，那么汉字CP/M操作系统已经不是层次结构的操作系统了，因为BIOS和汉字驱动模块之间没有单向依赖关系。但是如果把这两个模块加在一起看成扩充的 BIOS(即XBIOS)，那么汉字CP/M操作系统仍然是层次结构的操作系统。

由于在汉字驱动模块中包含了汉字的输入编码到机内代码的转换（甚至可以包含多种编码方案），故XBIOS 要比 BIOS 庞大得多。而 XBIOS 可以看成字母和汉字混合数据的输入输出管理程序。

在汉字 CP/M操作系统控制下，原来的各种程序设计语言都可以调用和处理汉字了。一种程序设计语言能调用、处理汉字、其标准有如下四点：

（1）在注解中能使用汉字、字母数字以及字母和汉字混合字；

（2）文字常量或非数值常量可使用汉字、字母数字，以及字母和汉字的混合字；

（3）原有的输入输出语句可输入输出汉字、字母数字以及字母和汉字的混合字；

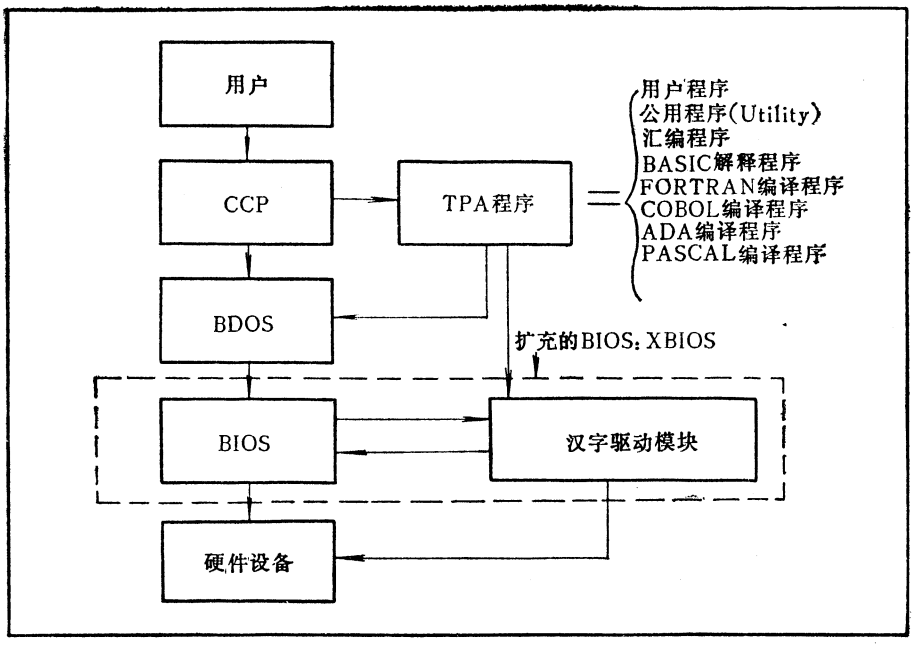


图10-11　汉字CP/M操作系统结构框图

（4）不用 CALL 语句或其它类似语句来输入输出汉字、字母以及同汉字的混合。

这样，就做到了汉字和字母数字兼容，并使得原有的 CP/M操作系统上的全部软件，乃至将要开发出来的软件，都可以为汉字数据处理服务。

图 10-12、图 10-13 和图 10-14 分别为汇编源程序、BASIC 源程序和 FORTARN 源程序的例子。也就是说，在这样的系统环境中，人们可以做到，处理汉字、编辑汉字、输入输出汉字，就好象处理字母数字、编译字母数字，输入输出字母数字一样方便。

各种编译程序编制出来的目标程序与编译程序本身都能调用和处理汉字，是因为它们都是TPA程序，它们对汉字和字母的输入输出要求都是以系统调用的形式向 BDOS 模块提出的。这里再对系统调用的手段和形式补充说明几句。

CP/M操作系统共有三十多个系统调用，都有统一的调用方式。

在 5 、 6 、 7 单元中放了一条转移指令转向 BDOS 模块在内存中的起始地址。不同的系统由于硬件配置不一样，这个起始地址也各不一样。但是不论什么配置下的 CP/M 操作系统中的任何一个 TPA 程序（用户程序）只要调用 5 号单元就可转入 BDOS 模块的入口地址。此外，还有如下约定：

调用序号存入寄存器C；

调用参数存入寄存器D和E；

返回参数存入寄存器A和B。

当 CP/M操作系统的版本不断更新时，系统调用的序号在增加，但是原有序号的意义不变，因此能向上兼容。

这里不把全部系统调用列举出来，只举第 10 号系统调用-输出字符串的例子。

系统调用序号：（C）=09。

（a）汇编源程序

```
; THIS IS 汉字 ASSEMBLE PROGRAM
          ORG    0100H
          JMP    BEGIN
;
DATA1     DB     ，中国软件技术公司，
DATA2     DB     ，汉字输出输入，
;
BEGIN     LXI    D, DATA1
          MVI    C, 09H
          CALL   0005H
;
          LXI    D, DATA2
          MVI    C, 09H
          CALL   0005H
;
          JMP    0000
          END    0100H
```

（b）通过汇编编译后的结果

```
                    ; THIS IS 汉字 ASSEMBLE PROGRAM
0100                      ORG    0100H
0100  C31F01              JMP    BEGIN
                    ;
0103  8E818781AEDATA1     DB     ，中国软件技术公司，
0113  8785DF80CBDATA2     DB     ，汉字输出输入，
                    ;
011F  110301        BEGIN LXI    D, DATA1
0122  0E09                MVI    C, 09H
0124  CD0500              CALL   0005H
                    ;
0127  111301              LXI    D, DATA2
012A  0E09                MVI    C, 09H
012C  CD0500              CALL   0005H
                    ;
012F  C30000              JMP    0000
0132                      END    0100H
```

图10-12　汇编源程序

开始列表

```
10  GOTO100
20  RESET:RESAVE''TESTKAN
30  END
100  REM            计算机汉字直接输出，直接列表演示程序
110  REM      在本程序中，汉字能够直接出现在源程序的注解，字符串常量
120  REM  和字符串数据中，能够被赋值，读取，并能象普通 ABC 字符串一样进
130  REM  行加减和置换运算，例：
140  A￥="汉字"                    : REM  赋值
150  B￥="输出"                    : REM  赋值
160  PRINT# 2, A￥                 : REM  输出
170  PRINT# 2, B￥                 : REM  输出
180  C￥=A￥+B￥                    : REM  加运算
```

```
190  PRINT# 2, C￥            : REM  输出
200  MID￥<C￥, 22, 7>="入"   : REM  置换运算
210  PRINT# 2, C￥            : REM  输出
220  DATA 程序, 列表          : REM  数据
230  READ A￥, B￥            : REM  读取
240  C￥=A￥+B￥              : REM  加运算
250  D￥=LEFT￥<C￥, 21>      : REM  减运算
260  PRINT#2, A￥             : REM  输出A￥
270  PRINT#2, B￥             : REM  输出B￥
280  PRINT#2, C￥             : REM  输出C￥
290  PRINT#2, D￥             : REM  输出D￥
300  PRINT#2                  : REM  走纸三行
310  PRINT#2
320  PRINT#2
330  PRINT#2, "开始列表"      : REM  输出
340  LIST# 2                  : REM  列表
350  END                      : REM  程序结束
```

图10-13 BASIC源程序

(a) FORTRAN 源程序

```
C     这是一个汉字 FORTRAN 源程序
C     在本程序中，汉字可以出现在格式语句中
C     在处理时如同普通 ASCII 代码 FORTRAN 源程序一样
C
C     ***************************************** **
C     这个程序求解一个一元一次方程
C     **************************************** ***
C
      WRITE<5, 100>
100   FORMAT<1X, '这个方程的形式为：AX+B=C'>
      WRITE<5, 200>
200   FORMAT<1X, '请输入A, B, C的值'>
      READ<5, 50>NA, NB, NC
 50   FORMAT<3<15>>
      ND=<NC-NB>/NA
      WRITE<5, 10>NA, NB, NC
 10   FORMAT<1X, '输入的值是：', 3<15>>
      WRITE<5, 20>ND
 20   FORMAT<1X, '您需要的解是：', 15>
      WRITE<5, 30>
 30   FORMAT<1X, '谢谢您'>
      STOP
      END
```

(b) 执行结果

```
这个方程的形式为：AX+B=C
请输入A, B, C的值 2, 14, 4
输入的值是： 2  14  4
您需要的解是： -5
谢谢您 STOP
```

图10-14 FORTRON 程序和执行结果

入口参数：(DE)＝输出字符串首地址。

返回参数：无。

功能：在显示器上将￥前面的字符串显示出来。

由于汉字数据和字母数据在计算机内部都是以代码表示的，即都表示成字符串的形式，因此 BDOS 模块对它们的处理是一视同仁的，只是到了 XBIOS 模块，由于后者能识别和区分汉字代码和字母数字代码，因此分别输出汉字和字母数字。

根据系统调用的个数，在 BDOS 模块内部也划分了相应个数的子模块，各自处理对应的系统调用。

当用户程序要输入一串字符，例如为“啊ABC”，那么用户程序中要用一个相应 的系统调用，发向 BDOS 模块，后者责成 XBIOS 执行输入一串字符的操作。用户输入这个字符串的过程如下：

（1）按（Control, B）键〔不妨认为这是约定的汉字的引导符〕，然后是“啊”字的输入编码（例如采用声韵部形码），再按“SPACE”键。

（2）在显示屏上出现“啊”字。

（3）按（Control, A）键（不妨认为这是约定的字母数字的引导符）。

（4）按A键，在显示屏上显示“A”。

（5）按B键，在显示屏上显示“B”。

（6）按C键，在显示屏上显示“C”。

（7）按“RETURN”键。

于是这一串字符“啊 ABC”的代码就送入用户程序区，由用户程序加以处理。在用户输入这一串字符的过程中，XBIOS 模块的工作过程大致如下：

（1）由（Control, B）码知道将输入汉字，置上汉字输入标识位。

（2）由编码转换表找出“啊”字字形地址，在显示屏上显示“啊”字。并根据 GB2312 交换码，将“啊”字的代码表示为 16 进制的 2 个字节的国标码 BOA1，并送入缓冲区。

（3）由（Control, A）码知道将输入字母数字，置上字母数字输入标识位。

（4）显示“A”，并将 41 码送入缓冲区。

（5）显示“B”，并将 42 码送入缓冲区。

（6）显示“C”，并将 43 码送入缓冲区。

（7）用户按 RETURN 键，表示本次输入结束。

将汉字输入输出驱动模块纳入操作系统的输入输出管理程序，是一个很好的办法，可以使操作系统所有的更多的软件资源为汉字信息处理和汉字输入输出服务。但是上述四条标准是比较高的，并非建立汉字数据处理系统非要遵照这些标准不可，但可把它作为一个努力的目标，以便使得汉字、西文程序设计语言一致化。

10.3.2 几种汉字数据处理系统软件的建立方法

目前我国汉字系统软件的建立，多半是在微型机上和小型机上进行的。下面将分别加以说明。

一、在CP/M操作系统上扩充汉字功能的方法

七十年代后期，国内少数用户是在裸机上进行汉字试验。在这样的系统里只有汉字的输入输出功能而无处理能力。到八十年代初，才发展成用高级程序设计语言、特别是BASIC语言编制的程序来调用和处理汉字。在图10-6到图10-9中，我们给出了详细的例子。在这些例子中，可以看到，用BASIC语言的CALL语句来调用和处理汉字在使用上是非常不便的。另外，从体系结构上来说，这样的结构也是不合理的。请参见图10-7中的内存分布图，其中的字母数字的显示、字母数字的打印驱动模块已纳入CP/M操作系统的BIOS部分，所以用CP/M的系统调用就能输入和输出字母数字。然而，汉字的显示和汉字打印驱动模块却和CP/M操作系统毫不相干，以致要使用CALL语句才能调用汉字显示、汉字打印驱动模块，显得很不方便。因此，我们要把汉字驱动模块纳入BIOS模块，即前节所说的扩充为XBIOS模块。这就是将汉字输入输出程序纳入操作系统的输入输出管理程序——BIOS模块的办法。经过这样一番改造，原来的CP/M操作系统就改造成为汉字CP/M操作系统了。

如果CP/M操作系统原来所支持的设备中并无汉字设备，那么就要把这些汉字设备加上去，并编写相应的驱动模块，同时把这些驱动模块同原有的BIOS模块有机地结合起来，使汉字的输入输出和字母数字的输入输出统一到新的XBIOS模块中。这就是在CP/M操作系统上扩充汉字功能的办法。新的系统是汉字和字母数字兼容的系统。经过试验，我们证实不但原有的各种程序设计语言都能处理汉字，而且CP/M操作系统上的原有的数据库系统DBASⅡ也能处理汉字。

现在再回过头来谈谈汉字数据及其表示的问题。操作系统输入输出管理程序（经修改的CP/M操作系统中的XBIOS）中的汉字印刷机驱动模块和汉字显示器驱动模块之所以能够识别、区分汉字和字母数字，是因为它们在计算机内部的代码表示各不相同。假如我们采用10.2.2节中的“自名表示法”，那么就可以如同图10-2中的（b）表示一个GB1988码字符，而（c）则表示一个汉字，它们彼此是可以识别不会混淆的。但是不论汉字或是字母数字，都可以看成各种程序设计语言中的“字符的”或“字符串的”数据类型。这样我们就有可能利用原有的程序设计语言来表示汉字、处理汉字和输入输出汉字。这时，既不需要另搞一套程序设计语言，也不需要修改和扩充原有的程序设计语言，而使原有的各种程序设计语言都能处理汉字、字母、以及字母和汉字的混合字（请参阅图10-12到图10-14）。

只要程序员明白汉字数据在计算机内部的表示法，他就不难做出这样的程序设计。

汉字数据和字母数据在计算机内部的表示是不相同的，因此也有人认为要把汉字数据作为一种新的数据类型加到原有的程序设计语言中去，如扩充到COBOL中去，或FORTRAN中去。那样，新的COBOL程序语言文本，新的FORTRAN程序语言文本就要同国际标准的COBOL程序语言文本和国际标准的FORTRAN程序语言文本不一致。当然，还需要另搞一套编译程序。不过在国外，这是一种颇为流行的观点。

采用前一种观点和前一种做法，不把汉字作为新的数据加到原有的程序设计语言文本中去，这样作有如下三个理由：

（1）可把国际标准的程序设计语言或国际上普遍使用的程序设计语言作为我国国家标准的程序设计语言。

（2）可以使操作系统中原有的程序设计语言表示汉字、处理汉字和输入输出汉字。

（3）我国是多民族的国家。假如为了要把汉字数据类型加到COBOL程序语言文本中去，创建中国的汉字COBOL程序语言文本，那么还需要创建我国多种少数民族的COBOL程序语言文本，类似的工作就需要没完没了地做下去。反过来，只要解决了用原有的程序设计语言来处理汉字的问题，就可用大体上相同的方法，来解决我国其他少数民族文字的数据处理问题，而又和国际标准的程序设计语言或国际通用的程序设计语言保持一致。

上面是以CP/M操作系统为例，说明了在操作系统的核心部分——输入输出管理程序中扩充汉字外部设备的驱动模块的方法，以及这种方法在理论上和实践上所带来的好处。

下面举例说明在小型机中，在多用户环境下的操作系统的核心部分——I/O管理程序中如何扩充汉字外部设备的驱动模块，以及这种做法在实际上所收到的效果。

二、PDP-11系列机上汉字系统的建立方法

在PDP-11系列机的RSX-11M操作系统上，可以建立汉字数据处理系统。PDP-11系列有十来种操作系统，其中尤以RSX-11M使用最广。RSX-11M原有多达成百万条指令的系统软件，经过许多软件公司、厂家和用户的开发，又在其上增加了新的系统软件和应用软件，它们象滚雪球一样越滚越大。如果能把这些系统软件和应用软件都移植过来为汉字数据处理服务，其效果是十分可观的。这就是为什么首先在RSX-11M上建立汉字系统的指导思想。

RSX-11M是一个实时、分时操作系统，它的I/O结构如图10-15所示。

不论是用户程序还是公用程序，若要进行输入输出，则都要明显地或蕴含地通过QIO宏指令才能进入管理程序。管理程序中的输入输出子程序接受用户程序的委托，对QIO宏指令中表述的输入输出要求作大体如下的处理：

（1）对QIO宏指令进行预处理。根据QIO宏指令中提供的参数，来识别是哪个设备的哪种要求。把这种要求填写到专门的“I/O要求的报文”中，并把报文发送给相应的设备驱动程序，在那里排队挂号。

（2）把控制转给相应的设备驱动程序。

（3）当相应的设备驱动程序得到控制权时，取出最早排队挂号的“有输入输出要求的报文”进行处理，完成用户所需的I/O要求。在这个过程中，对某些设备驱动程序，也许要处理若干次的输入输出中断，才能完成某一用户的输入输出要求。输入输出要求完成后，控制又转向原先的管理程序的输入输出子程序；

（4）管理程序的输入输出子程序，对用户的I/O要求进行后处理，把完成的情况（成功或失败）告诉用户程序，并把控制转向用户程序中的QIO宏指令的下一指令。

这就是RSX-11M操作系统中的输入输出操作的大致过程。要在这上面建立汉字数据处理系统，就是要把用户提供的汉字设备加到系统配置中去，把用户提供的汉字设备驱动程序加到操作系统中去，并与原系统保持一致。

用户提供的外部设备不外乎下面两类：

（1）汉字印刷机；

（2）汉字显示器（或汉字终端）。

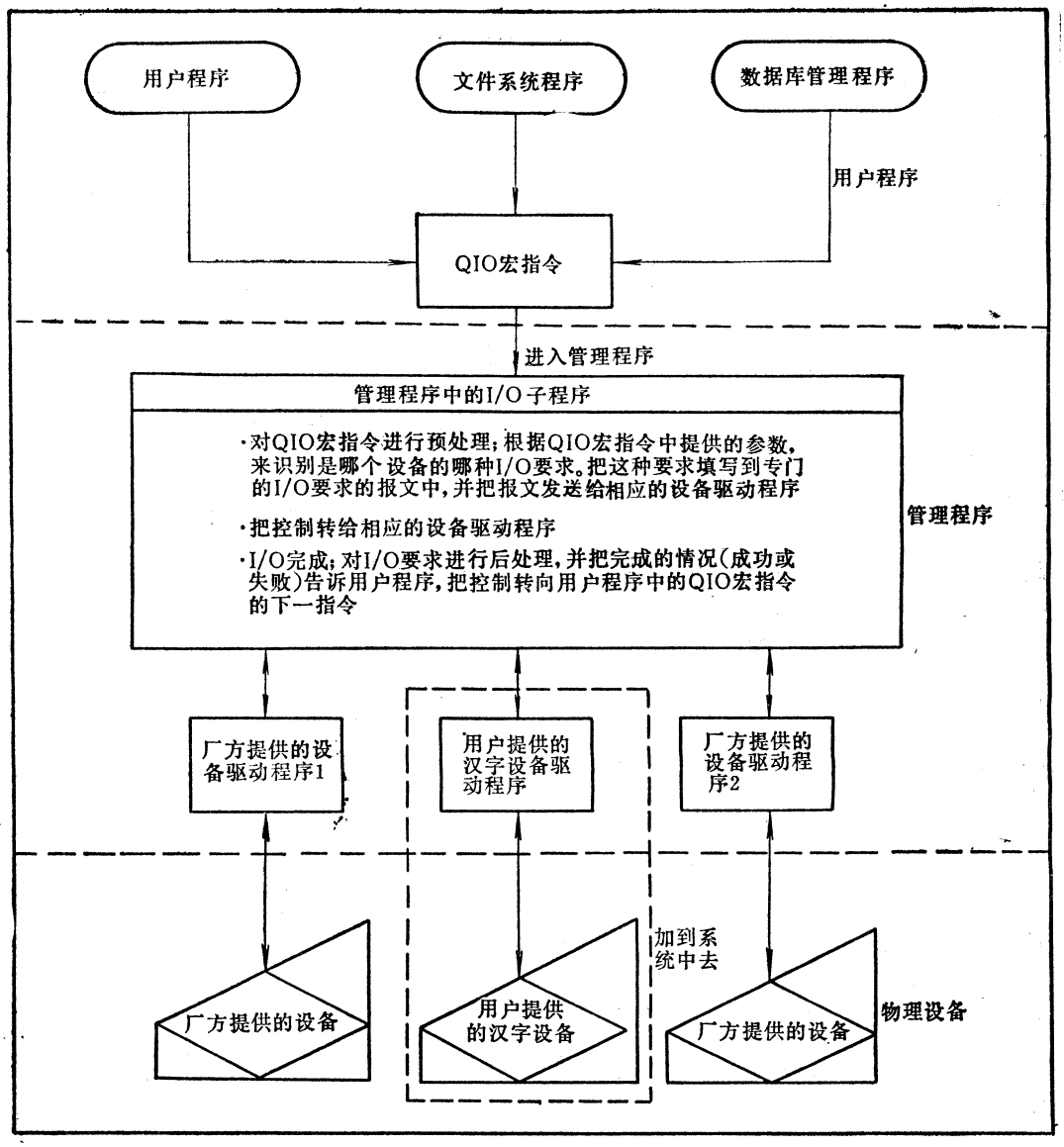


图10-15　RSX-11M操作系统的I/O结构

最简单的汉字终端是不具备汉字字模库的，字模库ROM可设置在PDP-11的总线上，作为PDP-11的一个外部设备。对于这样的终端，终端驱动模块中要有输入编码到机内代码的转换表。前面说过，机内代码最好同我国的GB2312交换码有简单的一一对应关系，而且汉字字模库中的字模地址最好也按国标码的次序存放，使终端驱动模块占尽可能小的内存空间，获得尽可能快的处理速度。

在上述情况下，用户输入汉字的过程大致如下：

1）用户程序发出一个QIO宏指令。该宏指令的意思是：从汉字终端输入一串汉字，并把这一串汉字的代码存放在用户程序指定的地址（QIO宏指令的具体功能是由该宏指令的参数表示的）。

2）管理程序的I/O子程序将上述要求转达给汉字终端驱动模块。

3）汉字终端驱动模块启动汉字终端，进行如下一系列的工作：

（1）用户键入汉字编码。

（2）终端驱动模块将汉字编码转换为汉字内部码，并根据内部码从汉字字模库中找出相应的汉字，把这个汉字的字模信息按字节逐个地发送到汉字显示器中，于是显示出用户所需的汉字。

（3）用户继续键入汉字编码。

（4）终端驱动模块重复（2）的工作。

（5）用户按RETURN键，表示汉字字符串输入完毕。

（6）终端驱动模块完成了用户委托的I/O要求，把控制转向管理程序中的I/O子程序。

4）管理程序中的I/O子程序将一串汉字代码转送到用户程序指定的地址，并将控制转向用户程序的QIO指令的下一指令。

以上是最简单的汉字终端的汉字输入过程。如果汉字字模是由15×16的点阵构成的，那么就要输出30个字节才能在显示器上形成一个汉字，使输入输出的量大大增加，从而会减低系统的效率。

因为RSX-11M是具有分时功能的操作系统，汉字字模库做在总线上可为多用户共享，因此成本低。其缺点是系统开销太大。

另一种汉字终端是把字模库做在终端中，而把汉字编码到内部码的转换表设置在终端驱动模块。这样做的优点是：可由用户选择自己所习惯的编码方式（因为在终端驱动模块中可以有几个不同的汉字输入编码到内部码的转换表）；而且发送2个字节就可显示一个汉字，系统开销也不大。

如果把汉字字模库和汉字编码到内部码的转换表都设置在终端内，那么就可以大大节省系统开销。

由于RSX-11M操作系统允许将用户提供的外部设备驱动模块加到操作系统中去，因此不论什么样的汉字终端、不论什么样的汉字印刷机都可加入系统。

另外，该系统有较好的“与设备无关”的性能。由于系统程序和公用程序中所用到的输入输出设备不是在程序中预先规定的，因而可以在程序执行时或程序联结（task building）时指定设备。由于有了这个功能，就可以用原操作系统中的所有的各种编译程序和公用程序来处理汉字。也就是说，用汉字终端驱动模块代替原有的字母数字终端的驱动模块。

前一节中已经说过，因为汉字代码和字母数字代码在计算机内部表示是不同的，所以汉字终端驱动模块很容易区分汉字代码和字母数字代码，从而分别显示出汉字和字母数字。也能把汉字印刷机做到这样，由于汉字印刷机驱动模块能识别汉字代码或字母数字代码，故可分别输出汉字和西文。

在这个扩充了汉字功能的系统里，各种原有的程序设计语言都能处理汉字，都能用原来的程序设计语言编写处理汉字数据的程序。这对于用户来说是方便的，他不需要学习新的程序设计语言就可以进行汉字数据处理了。

以上就是PDP-11上建立汉字系统的基本工作原理、基本方法和基本功能。

10.3.3 从软、硬件的角度讨论汉字数据处理系统的建立方法

一、不增加汉字设备的方法——纯软件的方法

在一个虚拟机（virtual computer）的环境中建立一个汉字数据处理系统并不存在一种统一的方法，也没有一成不变的模式。何况用户的计算机也是各式各样的，有国产的，也有国外引进的。即使是同类机种，也可以有各种不同的配置和各种各样的操作系统。同一机种的不同操作系统可以看成不同的虚拟机器。不同的虚拟机器有不同的特点，针对不同的特点人们可建立各式各样的汉字数据处理系统，使之满足不同用户的各种需要。

在各种方法中用纯软件的方法来建立汉字数据处理系统，是一个简易可行的方法，因为它利用的是厂方提供的设备，不需要购买专门的外部设备，而仅仅需要投入人力。

有人在HP3000的MPE操作系统上建立的汉字处理系统就是纯粹软件工作的一个例子。

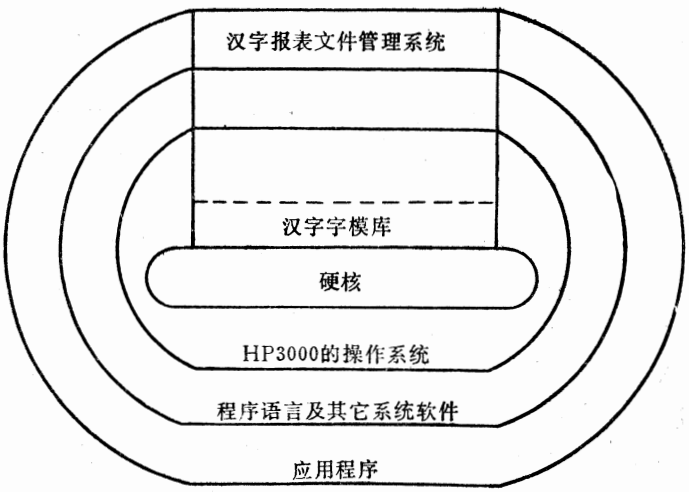


图10-16 HP3000汉字处理系统

图 10-16 给出了这个汉字处理系统的三个基本部分：汉字字模库（包括维护程序）、汉字处理程序库和汉字报表文件管理系统。

（一）汉字字模库

这个汉字处理系统的汉字字模库放在操作系统的层次中。可以把它看成为操作系统的组成模块。这样汉字字模库可成为常驻内存的系统模块。这对加快汉字处理速度、提高系统效率、减少系统开销是有好处的。

（二）汉字处理程序库

汉字处理程序库存放着已成为 HP3000 操作系统的内部过程（intrinsic process）的各个汉字处理程序。这些程序可以看成是操作系统的组成模块。它们可以为各种高级程序语言 COBOL、FORTRAN、BASIC 和 SPL（HP3000 的面向机器和系统结构的系统程序设计语言）所调用，使这些高级语言有处理汉字数据的能力。不论是从实现的结果还是从实现的方法上看，它们都是这个汉字处理系统的特点。

在 HP3000 上具有汉字输入或输出能力的设备有点阵式的行式打印机，显示终端及与终端连在一起的点阵式打印机。汉字处理系统的汉字输入输出操作就是在这些设备上实现的。这些设备是系统本身所具备的（由厂方提供），用户并未提供新的外部设备。

（三）汉字报表文件管理系统

汉字报表文件系统具有模块式结构，这些模块不纳入 HP3000 的操作系统。这是一

个面向以报表文件为基本组织形式和输出形式的应用系统。

如上所说，这个系统的汉字处理模块是放在某一系统库（system library）中的，因此各种程序设计语言可以通过CALL语句或类似的方法来调用这些汉字处理模块，以达到汉字输入输出和汉字处理的目的。于是，这些系统有相当的向上扩充的余地。用户可以自己编制和逐渐积累起各种汉字处理的公用程序，以满足用户自己特殊的需要。

但是这个系统的不足之处也是很明显的，由于汉字输入输出模块是建立在系统程序库上的，不是建立在I/O管理程序中（当然，汉字输入输出模块是要借助于I/O管理程序的），因此，汉字的输入输出和原有字母、数字的输入输出不能统一于同一语句之中。这样，原有的系统程序和公用程序，如原有的字母数字的编辑程序等就不能为汉字编辑服务。有关汉字信息处理的工作都要系统设计者自己做或让用户来做，MPE操作系统上的某些软件资源，包括用户在其上开发的应用软件一般都不能为汉字处理服务。

此外，也有在PDP-11/70上配置汉字信息处理系统，也是立足于软件的。因为他们的系统中配有P6000型点阵印字机和4010型图形显示器，可以用来输出和显示汉字。

二、增加汉字设备的方法——软件同硬件结合的方法

目前国内在微型机上实现的汉字处理系统，大都是增添了用户提供的汉字输入输出设备，或者是对某些设备进行了一些改造来实现的。于是相应地必须对原操作系统，或对某些程序设计语言的编译程序进行一些改造，才能建立汉字处理系统。其中，对CP/M操作系统的改造最为成功，使其上的原有系统软件都能为汉字数据处理服务。这些我们在前面已经说明过了。这一类方法在系统配置上说是增添了汉字设备，从操作系统上说是在输入输出管理程序上增添了汉字设备的驱动模块。然而，却没有改变该操作系统和用户程序间的界面。因此，人们使用CP/M操作系统时的方法、手段，同使用具有汉字功能的CP/M操作系统时的方法、手段是完全一致的；使用该操作系统中的程序设计语言对西文数据进行处理的方法，与使用同一程序设计语言对汉字数据进行处理的方法是一样的。

此外，在前一节中介绍的PDP-11系列的硬件组成虽然在配置上增添了汉字设备，在RSX-11M操作系统上增添了汉字设备的驱动模块，然而用户程序和操作系统的界面都是不变的（用户程序仍是通过QIO宏指令来委托操作系统进行输入输出），即用户使用RSX-11M操作系统的方法同使用具有汉字功能的RSX-11M操作系统的方法是一样的；用其上的程序设计语言来进行西文数据处理的方法与用同一程序设计语言来进行汉字数据处理的方法也是一样的。

从日本和美国引进的不少微型机系统也具有汉字数据处理功能，但是对汉字数据进行输入输出的方法跟对字母数字数据进行输入输出的方法不同，不是采用同一语句实现，看来这并不是一种好的方法。

三、接插兼容的汉字设备——硬件的方法

除了用纯软件的方法和用软硬件相结合的方法来建立汉字系统以外，也可采用硬件的方法来建立汉字系统。

在10.3.1节和10.3.2节介绍了在CP/M操作系统上和在RSX-11M操作系统上建

立汉字处理系统的基本方法和原理。概括起来有两句话：用汉字设备代替字母数字设备；用汉字设备的驱动模块代替字母数字设备驱动模块。

我们所说的汉字设备是指能输入输出字母数字、汉字和字母汉字混合字的设备；我们所说的汉字设备的驱动模块是指既能输入输出字母数字，又能输入输出汉字、也能输入输出字母数字和汉字混合字的设备驱动模块。

计算机发展的历史告诉我们，软件往往是硬件的先导。工业上常常把软件上行之有效的算法固化而实现成批生产。随着工业上 ROM 价格的不断下降，容易实现把汉字设备的驱动模块，例如汉字终端的驱动模块(包括汉字字模库和汉字输入编码到机内码的转换表)固化到 ROM 上，用到汉字终端设备上去，我们把这样的设备称为接插兼容的汉字终端。对于和这种汉字终端所连接的主机来说，汉字和西文的区别仅仅是代码的不同，从而所说的操作系统中的汉字设备的驱动模块又可以还原到原来的字母数字设备的驱动模块。

这样，就可以不修改操作系统，也不扩大操作系统，利用原来的字母数字终端的驱动模块来驱动接插兼容的汉字终端。同样的道理，可以用原来的字母数字印刷机的驱动模块来驱动接插兼容的汉字印刷机。

这就是用硬件的方法建立汉字系统的手段。

用软件的方法，还是用软硬件结合的方法，或是用硬件的方法来建立汉字系统，这要取决于用户的要求，取决于人力、物力等条件。

用软件的方法建立汉字系统，其投资较少，软硬件结合的方法投资略多一些，而用硬件的方法则投资要更多一些，因为每个汉字设备上都要有汉字字模库，以及其它一些经固化的软件。

然而，从系统的效率看，硬件的方法效率最高，软硬件结合的方法次之，软件的方法效率最低。

因此，从长远的观点来看，硬件的方法最有前途，而且这是发展的趋势。“接插兼容的汉字设备”虽然是硬设备，但是它的设计思想是从软件来的，它的硬件组织中包括了固化了的软件功能。

10.4 大、中、小和微型机系统扩充汉字功能的考虑

汉字数据处理系统的目标就是要实现西文数据处理系统的全部功能。现在国内使用的计算机有的是从国外引进的；有的是国内生产的。国外引进的计算机大都是优选的有代表性的计算机；国内生产的计算机大体上也和优选的计算机兼容。建立汉字数据处理系统实质上就是在已有的大型机、中型机、小型机和微型机上扩大汉字数据处理功能。因此，只有当优选系列的操作系统上的全部软件——包括各种程序设计语言，公用程序，文件系统，数据库系统，各种应用软件，乃至西文的编辑程序都能为汉字数据处理服务时，我们才能说我们上述的目标真正实现了。这样，我们不需要重复国外已经做过的而且对我们也是行之有效的工作，而可以在已有系统资源的基础上建立各种适合于我国应用的各种汉字数据处理系统或应用系统。

10.4.1 在微型机上扩充汉字功能

在微型机上扩充汉字功能，现在已经做了很多工作，目前大多是采用软硬件结合的

方法，而且做得比较成功。典型的做法是将原有的CRT控制器改成具有汉字显示功能，此外，如果原有的印刷机是针式打印机，那就不必加以改造，便能成为汉字打印机。否则，就需要在微型机系统中增加汉字打印机。因为单用户的微型机操作系统比较简单，故只要对它进行适当的改造就可实现汉字的输入输出，工作量并不大。在改造的过程中要注意不要改变原有的系统调用（即用户程序对操作系统的界面），这样就可以使原有的系统软件为汉字处理和汉字输入输出服务。同时在汉字终端驱动模块中应妥善考虑汉字代码（汉字的机内表示）在汉字显示中所占的空间以及字母数字代码在字母数字显示中所占的空间的协调。只要协调得好，原有微机系统中的西文编辑程序也可以用作汉字的编辑程序。以后我们还要提到，作为接插兼容的汉字终端，上述协调性也是非常重要的，否则字母数字的编辑程序就不能用于汉字的编辑了，而需要另外设计汉字编辑程序。这样一来，原有的系统资源就不能充分加以利用。

以上的讨论主要是对单用户的微型机而言的，对于具有分时功能的多用户微型机，问题更复杂些。

目前的微型机，除了那种小型机微型化以外，一般都是没有系统生成功能的。我们把具有系统生成功能的微型机作为小型机看待，在下一节讨论。在微型机系统中把一个行之有效的操作系统推广到有相同CPU、而有不同配置的微型机系统中，这叫做可移植性（portability）。移植的主要工作是重写输入输出管理程序，以适应不同的配置。但是不改变用户程序对操作系统的调用界面，因此原系统上的全部软件都可搬到新的配置上。对于用户来说是在使用着同一操作系统。

在一个微型机系统中扩大汉字功能，实质上就是将同一操作系统移植到有相同CPU的并配用汉字外部设备的配置中。当然，也不能改变用户程序对操作系统的调用界面。

微型机系统一般是没有系统生成功能的，它通过移植操作系统来扩大它的应用范围。因此，对于一个好的微型机操作系统，其移植一定是非常方便的，否则它只能固定在某一种配置之下，从而不利于这种操作系统的推广。因此，一个好的微型机操作系统也易于扩大汉字功能，使该系统的软件全部或大部都能为汉字数据处理服务。在扩大汉字功能时，要注意以下两件事：

（1）选择一种（或同时用若干种）适当的标识汉字代码的方法。

（2）在用汉字外部设备来代替字母数字外部设备的同时，用汉字设备驱动模块来代替字母数字设备的驱动模块，千万不要改变用户程序对操作系统的调用界面。

10.4.2 在大、中、小型机上扩充汉字功能

一、汉字系统的设计准则

要在小型机以上的机型上（包括中型机和大型机）扩充汉字功能，问题比微型机复杂得多，因为它们的系统软件非常庞大。每一个机型又可有多个操作系统，其上又可以有数据库系统和各种应用系统，任何一种系统都可以看成一个虚拟机器。从硬件上说，扩大汉字功能就是要把汉字设备加到虚拟机上去。那么在汉字系统的设计上，是否有什么准则呢？

我们可以从两个方面来考虑这个问题。一是汉字的代码问题，因为汉字作为一种数据在计算机内部表示，总要采用某一形式；另一个是如何对待原有的系统资源问题，因

为我们总是在某一个通用计算机上建立汉字系统的，因此，就有一个对原有的资源尽量加以利用呢，还是拚弃的问题。

关于汉字代码的问题，我们在10.2.2节已经阐明了。究竟以哪一种代码标识方法为好，已提出若干条参考意见，这里不妨重复一下：

（1）要从尽可能少的字节数来表示尽可能多的汉字个数；

（2）要跟GB2312码有简单的一一对应关系；

（3）在汉字数据（包括字母汉字混合数据）运算（查找、合并、截断、比较、替换、传输等）时不容易产生二义性和不确定性。

为了尽量利用原有资源，或尽量利用原有的系统软件，我们已经在10.3.1节中阐明了程序设计语言调用、处理汉字的标准。如果原有的程序设计语言都能符合10.3节提出的这四个标准，那么原操作系统下的各种系统软件一般不必作大改动就能处理汉字了，因为每一个软件，总是用某一种程序设计语言写成的。

此外，我们在10.3.3节中讲了建立汉字系统的三种方法。即：软件的方法；软、硬件结合的方法；硬件的方法。

自然，不论解决中、大型机还是小型机的汉字系统的问题，这三种方法原则上都是可以采用的。但是对于小型机以上的机型来说，采用硬件的方法是最有前途的。对此，下节再重述接插兼容的思想。

二、硬设备的设计方向

要扩大汉字功能就是要把汉字设备加到系统中去，汉字设备是汉字CRT或汉字TTY（包括键盘）和汉字印刷机。

（一）汉字CRT/汉字TTY

这种汉字终端应自带汉字字模库和不少于一种的汉字编码转换表，不需要有编辑功能。它发送给主机或从主机获得的数据采用10.2.2节所阐明的某一种表示方式。

这种终端要能够模拟字符数字终端（GB1988码终端或EBCDIC终端），使得原则上不修改操作系统的设备驱动模块而能输入输出汉字，象输入输出字母数字一样。

只要我们对10.3.1节或10.3.2节的汉字输入输出流程有清楚的了解，就不难设计这种接插兼容的汉字终端。前面提到过，要在汉字终端中妥善考虑汉字代码在汉字字符显示中所占空间以及字母数字代码在字母数字显示中所占空间的协调问题。协调得好，原有计算机系统中的字母数字编辑程序也可用作汉字的编辑程序。

接插兼容的汉字终端的意义是在一个计算机系统中，将一个字母数字终端替换出来，插入接插兼容的汉字终端，于是系统就能处理或输入输出汉字信息。

（二）汉字印刷机

一个起码的汉字系统也要做到从应用程序这一级能印刷出汉字。我们的目标是：汉字印刷机加到系统中去，使该印刷机成为作业排队输出的设备。

可以有两种途径来达到上述目标：一种方法是在操作系统中安装用户提供的外部设备的驱动模块；另一种方法是研制“接插兼容的汉字印刷机”。

对于接插兼容的汉字印刷机，汉字字模库一定要做在汉字印刷机中，而且一定要有一个功能模块（ROM）能区分汉字代码和字母数字码，或者在汉字印刷机中设置一个单片微处理机，以实现印刷机的控制功能和对汉字、字母数字的区分等功能。

第十一章　汉字情报检索系统

11.1　情报检索的一般概念

本章将以文档库（file library）㊀和数据库（data base）为中心，论述有关汉字情报检索系统（IRS；Information Retrieval System）的基本原理、方法、工程设计与软件编制等问题。

11.1.1　情报检索问题

我们举例说明这个问题。

用文献语言㊁中的词，即标引㊂（indexing）用的词（又称主题词㊃或关键词㊄（Key Word）或叙词㊅），例如自行车，或者按照自然语言中的词（例如脚踏车、钢丝车），或者使用方言（例如单车、洋驴等）来描述各种不相同的文献、情报㊆资料。如果按照计算机的数据存储结构将这些资料存放起来的话，在读者需要查找含有自行车一词作检索项的提问式，向机读㊇文档库（或数据库）索取含有自行车一词的文献、情报资料时，其结果必然是只能找到（如果有的话）仅含自行车一词的各种文献、情报资料，而无法找出含有与其同义的词脚踏车、钢丝车、单车或洋驴的各种文献、情报资料（当然，假定该检索系统没有词库㊈）。

本例说明为使文献情报资料能够按其主题内容从某个集合中被检索出来，必须用某种方式去描述这一主题内容。由此产生的这种描述的结果又必须采用某种适合机读文档中的存储结构形式，以便它能够被人或机器方便地检索出来。

由此引出下面几个主要问题：

（1）如何对文献情报资料进行描述或标引？

（2）如何将标引过的资料按某种存储形式存放起来？

（3）如何从已存储了的资料库中将所需要的东西取出来？

（4）如何组建词库，使存储和检索有统一的、规范化的描述，借以提高检索效率？

㊀ 根据国内情报界的惯用法，这里将file一词译成“文档”而不译成“文件”或“文卷”。

㊁ 文献语言是人工语言，它是术语的汇集，在必要时也包括一定的句法。为了分类和检索的目的，它用来描述文献的内容。

㊂ 标引是一种文献处理方法，它是使用一种受控语言（相对于自然语言）中的术语，以某种程度的压缩方式来表征一篇文献的有关全部内容。

㊃ 主题词是指文献语言中专门用来表达文献主题的词。

㊄ 关键词是指文献语言中能表达文献内容基本概念的哪种词。

㊅ 叙词是指（以词典形式表示文献语言的）文献词典中用的词、词组或表达式。

㊆ 这里，“情报”是指一种有条件的可用“信息”。

㊇ 机读是机器可读（machine readable）的缩写，它表示对象能为计算机所识别和读写。

㊈ 词库是一部文献词典，用于收藏自然语言的词、词组及规范化术语。它按语义上的类缘关系将术语分组，并给出相互之间的关系。这种文献词库中用的词称为“用词”，而为“用词”所排斥的词称为“非用词”。

这些主要问题就是本章将要讨论和解决的所谓情报检索问题。

产生情报检索的社会背景主要是：

（1）随着科学技术的发展，文献、情报品种的规模和数量增长速度很快。这种“资料爆炸”，导致文献、情报部门的工作陷于困境，由此而提出图书情报资料处理自动化的问题。

（2）一方面文献情报资料不断增加，另一方面任何人阅读文献情报资料的时间是不变的。有人估计，一个科学家在阅读科技情报上花费他一天工作时间的百分之十以跟上科学的发展，这个比例若干年来基本保持不变，那么他将发现，在自己的学科领域中，1976年发表的文献是1966年的两倍。除非他采用更有效的措施跟上最新的进展，否则他就必然越来越落后。这就是人们为什么需要在浩如烟海的文献情报资料中准确而又快速地查找出他们所需求的文献资料的理由。

总之，情报检索同其他学科一样，都是因社会需要而应时产生的。

今天，作为信息科学的一个分支，情报检索已经是一门有着相当广泛内容的新兴学科。它是专门研究关于情报信息的存储以及从存储的信息中提取有关情报信息的工作、方法或过程的一门科学。

无论是从宏观世界到微观世界，还是从自然系统到人工系统，到处都存在有信息的存储和检索问题。但本章重点是介绍汉字情报检索系统[一]。汉字情报检索系统的处理对象是指那些以汉语描述的文献情报资料（包括图表数据等）。

原则上，汉字、西文[二]情报检索系统两者的工作原理和处理过程没有根本的区别。只是汉字情报检索系统要有汉字输入输出和汉字字模库等设备，而在标引方式和词库组织等方面与西文有所不同（详见本章第四节的讨论内容）。

11.1.2 传统处理方式

情报检索（information retrieval），过去通常是指文献情报工作的总称。原来的情报检索也确实是以文献检索为主的。图11-1反映了文献检索的基本流程。

文献检索的基本步骤如下：

（1）情报的收集。采用不同收集方法，从地理分布、专业范围等不同性质的情报源高效率地收集情报。

（2）情报的评价选择。从收集到的情报中，选出那些在系统中有存储价值的情报。

（3）情报的分类编目。将经过评价选择后的情报整理分类、编目归档，使它便于存储和检索（如书目卡片）。这包括：用于编写二次情报（如索引、文摘）的整理加工；制作各种索引卡片；用于编写一次情报（如原本书刊报告或视频文档——象照片、缩微胶卷或市售磁带文档）的整理加工。

（4）情报的存储入库。将经过分析处理后的情报，按内部管理规则入库存储(索引卡片装入卡片柜、书刊等收藏入书库)。

（5）情报的检索流通。根据借阅规则和流通方法，对读者的询问要求，按已经找到

[一] 本章把在概念上既能反映文献、情报处理过程又能形成一个功能完整、效率得当和动作协调的有关情报信息的存储和检索的整体叫做情报检索系统。

[二] 这里，西文泛指欧美国家使用的英文、法文、德文等各种语言文字。

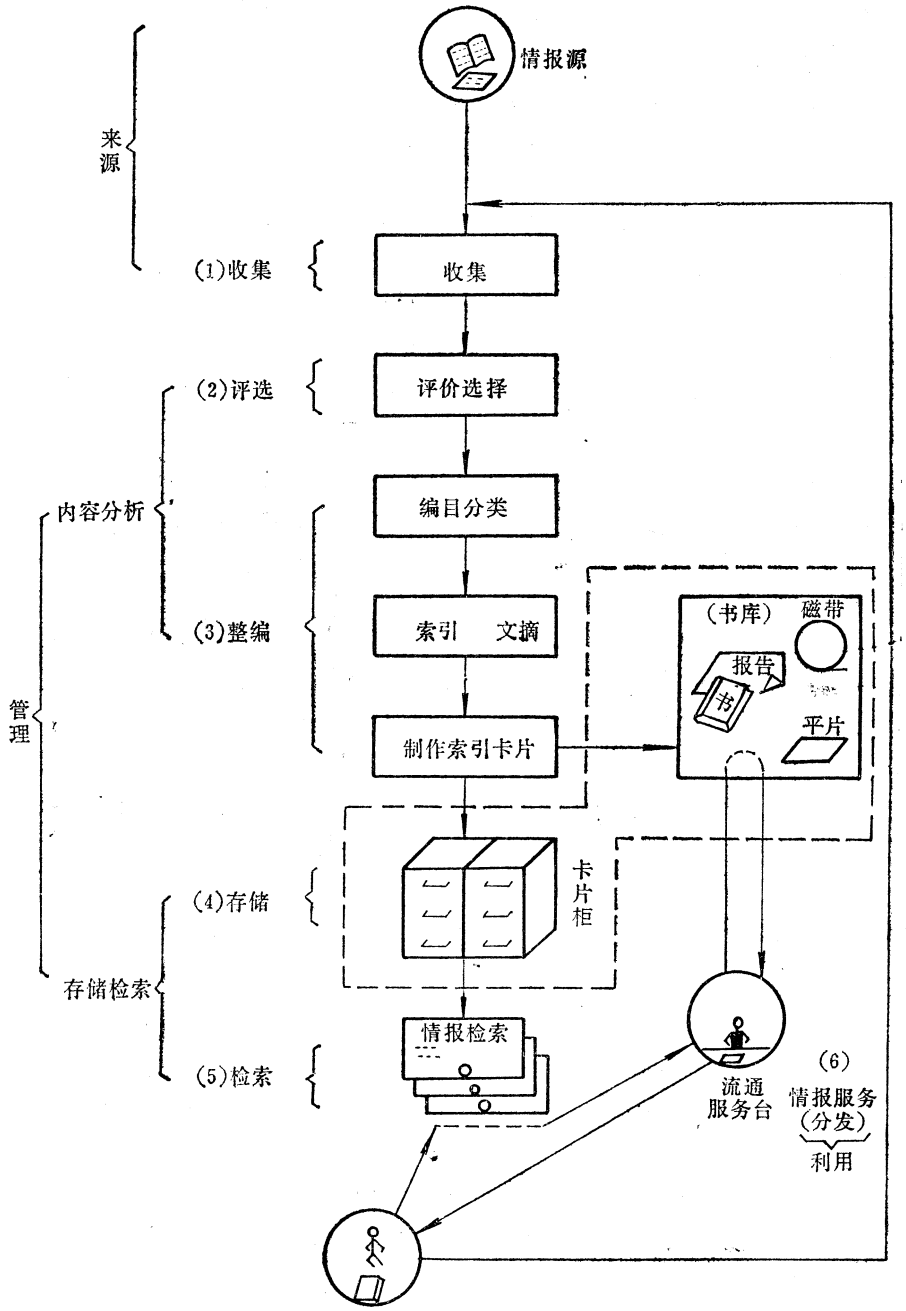


图11-1 文献检索加工流程图

的书目卡片，入库取书。对索取到的书，如果满意的话，就转到（6）去处理。如果不满意的话，读者可以对图书馆提出建议，要求收录新的或短缺的书刊资料。

（6）情报的复制分发。对于检索到的情报，将根据读者对一次情报的需要进行借阅和复制。

目前，利用计算机对文献内容进行自动分析还是一个比较薄弱的环节。而存储和检

索是当前自动化处理的重点工作。因此，从狭义来说，情报的存储和检索称之为情报检索。

如从设备性能和检索方法的自动化程度来看，情报检索大体分为手检、半自动和自动的三类。而从检索工具和方法史看，到目前为止，大致可分为五个发展阶段（四十年代以前；四十年代；五十年代；六十年代；七十年代）。

1．手检方式　1940年以前，检索工具是以书目卡片为主，检索方法是完全手工的，这是第一个发展阶段。

2．半自动检索系统　主要是指四十年代出现的手检穿孔卡和重叠比卡检索系统。这是第二个发展阶段。

3．自动化方法　自动化检索系统借助于三种设备：穿孔卡片机、胶片检索机和电子计算机。

五十年代的穿孔卡片系统、缩微胶卷系统，这是第三个发展阶段。

六十年代的批处理检索系统，是第四个发展阶段。

第一个以计算机为基础的情报检索系统创始于五十年代，但脱机● 批处理检索系统只是在六十年代才开始对情报服务产生重大的冲击。这期间仍使用经改进的缩微胶卷系统。

七十年代的联机检索系统，是第五个发展阶段。

虽然联机检索的试验至少可以追溯到1964年，但联机系统实际上得到承认是在七十年代。随着通信、数据库和计算机网络等技术的迅速发展，联机情报检索在走向信息化社会的时代里，将会越来越受到人们的重视。

11.1.3　计算机处理方式

图11-2说明了一个典型的计算机处理的（汉字）情报检索系统。

先从输入开始说起。

图中“文献”在它被输入和储存起来以前，已经是经过标引人员用文献语言标引过的文献。然后用某种输入方式把它输入到计算机里面去，同时按照逻辑要求加工成机读文档库或数据库，并存放到磁盘或磁带上，以备检索时用。

这样，用户需要查找文献资料前，应先根据系统提供的查询语言对他的检索要求进行描述和表示出来。如果有词库分系统的话（同文献一样要预先输入和存放起来），那么查询表达式将在用词管理下进行词的置换操作，即将非标准词置换成词库中的标准用词，或者按照词库中用词的相关关系进行词的扩充运算。此后查询是在拟定好的某种类型、例如布尔型的检索算法下进行处理。

当检索系统是联机时，用户在终端上进行人机对话查找期间，按照系统引导部分提供的信息，可以更改他的询问要求。

其次，看一看涉及到检索处理用的数据处理机（属系统组成部分）。这种检索处理可能还包括有用某种适当的途径将文献结构化、体系化，例如将文献进行自动分类等。但它主要是执行实际的检索功能，即实现对询问所制定的查找策略。

● 这里，脱机处理是指用户不直接对检索系统进行操作，而联机检索则是直接操作。

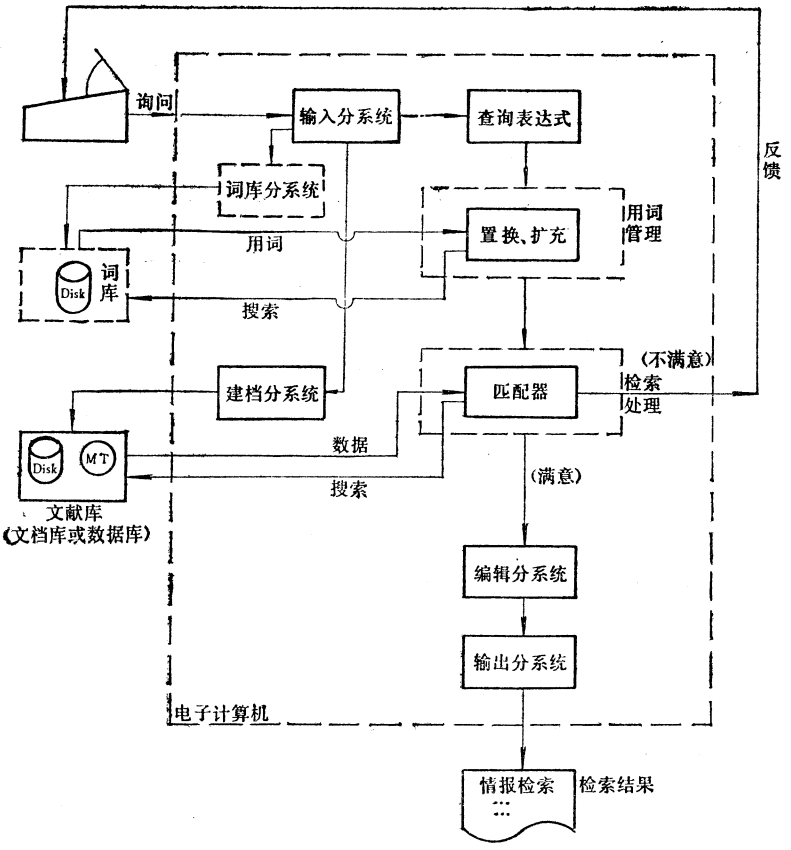


图11-2 一个典型的（汉字）情报检索系统

最后是输出部分。它通常是经过编辑排版加工以后，按照版面设计要求将检索结果显示或打印出来。其结果包括有一条或多条书目出处，或许还带有一些附加的信息，例如文摘等。

由此可见，情报检索自动化系统一般应有六个主要分系统组成，即

（1）建档分系统（组建文档库或数据库）；

（2）词库分系统（用词管理）；

（3）引导分系统（提供检索示范和介绍系统性能）；

（4）检索分系统（根据制定的检索算法对用户查询进行处理）；

（5）编辑分系统（包括索引刊物排印）；

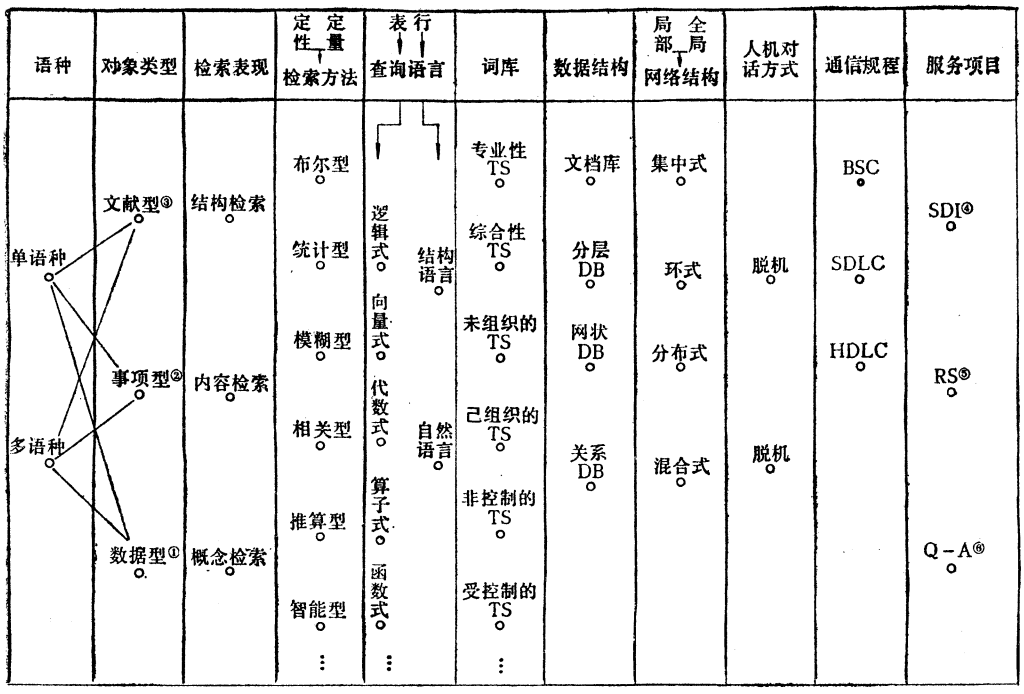
（6）管理分系统（检索系统的管理）。

此外，还有：

自动标引分系统；通信分系统，等等。

表11-1示明了情报检索系统的分类情况。

表11-1 情报检索系统的组合分类表



〔横向组合可以产生出不同类型的(汉字)情报检索系统〕

① 数据检索系指数值数据本身的检索。例如，人口调查数据，热物理性质数据或化学物性数据的存储和检索系统。

② 事项检索系指反映实际的事件或数据的存储检索。例如

（1）如果文档里存储有数据①素数是自然数，②5是素数，那么对于这一文档，一旦输入“5是自然数吗？”的查询时，系统将输出“是的”这个回答。

（2）如果黄山海拔1841米、泰山海拔1524米、珠穆朗玛峰海拔8848米等名山大川都已存储进计算机，那么当输入“哪座山最高？”的查询时，系统便输出“珠穆朗玛峰”这一答案。

③ 文献检索是正文检索的一种。它指的是文献的存储和检索。其目的是输出文献出处，以及在该文献上登录有所需要的信息（如文摘）等。

④ SDI服务（提供定题情报的服务）。它是由系统人员将预先登记好的提问编制成用户需求文档，每当新的文献进入系统时，定期地把符合查询要求的文献资料分发给用户的一种服务。也就是向特定用户定期提供特定的情报的一种服务方式。

⑤ RS服务，追溯性查找服务。是指用户对现在或过去已存储的文献情报进行（或脱机或联机）查找符合他所要求的情报资料的一种追溯性服务。

⑥ QA服务，人机对话服务。是指用户以联机方式直接输入提问，通过人机对话方式，随即进行检索并输出结果的一种问答式服务。

11.2 机读文档组织和检索策略

本节主要讨论情报检索中所使用的文档结构和检索策略，并介绍词库和系统设计与应用软件配置问题。我们将着重从逻辑要求和物理特性考虑文档的静态结构 而不着眼于实施中它的动态结构。这里，将按表11-2逐一加以讨论。

表11-2 文档结构类型表

静态结构	线性	线性结构：（一）顺序文档；（二）连续文档。
		链表结构：（一）串联文档；（二）（目录）单元文档。
	非线性	树形结构：（一）二元树；（二）平衡树；（三）霍夫曼树。
		网状结构：（一）索引文档；（二）倒排文档。
动态结构	散列函数和直接存取文档	

11.2.1 文档结构

一、文档设计

文档设计的任务在于把现实存在的、多种多样的、错综复杂的文献情报资料加以体系化、结构化，使之能为计算机所接受、加工、存储和检索。

但是在情报自动化管理中对文献的内容分析、标引作业、编制文摘和语言翻译等方面的自动化处理还存在着困难，至少目前对汉字文献情报资料是如此。这是因为选择文献、标引文献和这些文献的主题内容可能被分析或者能被描述到何种完备程度和分析及描述的确切程度有关，而且还和用来表达这些文献主题内容的词库是否选择恰当有关。所有这些都还有赖于人的智力工作，因此它们影响到文档的设计内容。

（一）设计文档时需要认真考虑的事项

1．经济性　在权衡文献情报资料使用价值和运用的经济性以后，决定收录对象(选材）并进行文档设计。

2．可靠性　存储的文献情报资料必须是可靠的。在选材时必须保证资料的可靠性。输入时，必须经过严格校对，确保进入机内的数据资料没有错误。

3．通用性　检索系统不是一成不变的。它会发生环境变化和改变目标要求。因此，必须使设计的文档具有适应系统变化的通用性。

4．交换性　为便于信息交换，设计的文档必须具有可交换性。

5．扩展性　系统建造一般是分期分批完成的。随着系统的不断扩充，文档也必须具有可扩展性。

6．管理性　删去低价值的文献情报资料，录入高价值的资料，赖以提高存储空间利用率。因此，管理性也很重要。

7．保密性　根据存取授权法则，一个文献情报服务中心有权允许或拒绝其它用户存取他的文档库或数据库数据。

8．安全性　根据系统用户协议和法律政策，确保文档库或数据库数据的存取控制、修改和传播得以正常进行。

关于文档设计的详细内容和设计方法，请读者查阅有关文献。

（二）文档设计中几个有关的概念和名词

通常资料（包括图书期刊、报纸杂志、公文档案和图表数据等）被收集起来以后，需要经过评价选择、整理加工、分类编目，进行主题标引和做文摘的工作，接着这些文献情报资料就进入到某种形式的文档库或数据库中。一篇标引过的文献成为一个“记录”(record)，是在计算机内部的加工单位。一个记录由许多文献分析时所确定的“特性标识项”(identifing item)，如由登录号、分类号、作者、语种或关键词（key word）等

所组成。这里特性标识项有时叫属性项（attribute item），它们在计算机科学术语中叫做“数据项”（data item）、“字段”（field）或“节段”（segment）。将各记录汇集起来形成一个计算机与外界作信息交换用的叫做“文档”（file）的逻辑单位，并用一文档名（file name）来引用这个记录集合。此后，根据需要可以从该文档中派生出“主题索引文档”、“分类号索引文档”和“作者索引文档”以及“关键词倒排文档”等等。将这些文档汇集起来又组成一个“机读文档库”，并用一个库名来引用该文档集合。注意，文档之间、记录与数据项之间都按照各自的某些关系进行组织和控制。对这些关系的研究是文档结构或数据结构的主要任务。在描述或定义文档结构时，除了逻辑描述外，有时为了反映物理特性还必须给出物理描述。

二、线性结构

文献情报资料按一定顺序排成一列的结构称为线性结构（linear structure）。

（一）顺序文档

一篇文献为一个记录。

以文献为单位来组织文档，例如以文献号码按递增顺序排列起来，而每个号码后面附有该篇文献的全部特性标识。这种按顺序排列记录的方法所组织的文档称为（线性结构的）顺序文档（sequential file）。

注意，记录内部的各个字段也是依次被存放起来的，并有一个确定的开头和结尾。

图11-3说明了一个顺序文档的组织情况。

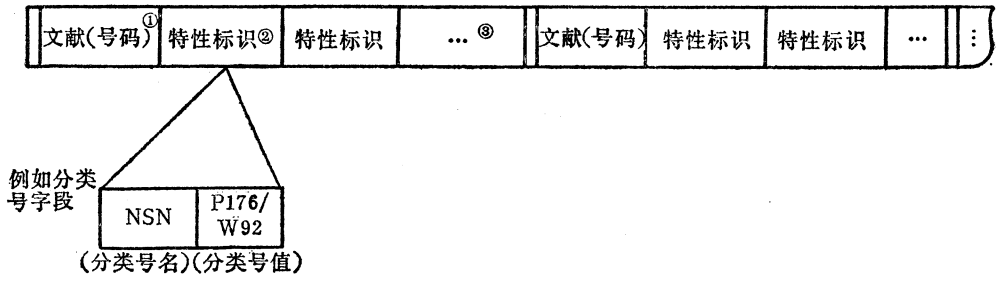


图11-3　一个顺序文档例子

① 文献（号码）可以是馆藏号、盘区地址或别的代表文献的相对存储地址。

② 特性标识可以是确定的文献标引事项，如分类号、作者、标题、关键词等。

③ 省略表示。

顺序文档是所有文档结构中最基本的结构。其主要优点是实现容易和能用字典编辑方式的次序给下一个记录提供快速存取。它的缺点是修改困难，当插入一个新记录时需要移动该文档的大部分，在需要随机存取时速度很慢。

（二）连续文档

连续文档（successive file）是顺序文档的一种变型。它是因使用磁盘存放而得名。在磁盘上连续文档是由若干个连续的盘区所组成的，见图11-4所示。

这种文档的缺点是文档长度一经固定就不能改变，因此它不宜用来处理可变长记录格式的文档。其优点是存取速度较高，并能直接访问任意盘区的内容。

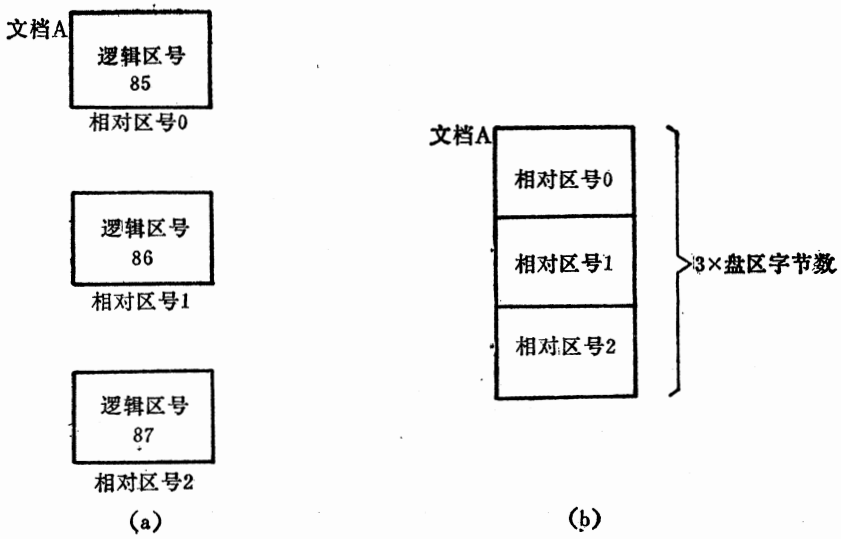


图11-4　一个连续文档例子

（a）连续文档的物理结构；（b）用户观点的连续文档。

三、链表结构

链表结构（linking list structure），将用到表处理中的某些基本知识。

一个表，是由若干个数不定的元素所组成的一个有序集合。如果元素之间具有邻接关系，则称为线性的；否则是非线性的。例如上述的顺序文档、连续文档都是一个线性表。其中一个记录为一个表元素。

但顺序文档、连续文档内各记录间并不一定是严格按照文献号码递增或递减进行排列的。即表元素之间的邻接并不意味着在存储媒体上也一定是邻接的。因此，对于这种不相邻接的元素，添加一个称之为指址器（pointer）的地址字段来指示表中下一个元素的地址，这就能形成一个媒体中不相邻接的元素所组成的链式线性表，也称链表文档（见图11-5）。图中∧表空指址器。

（一）串联文档

这类串联文档（serial file）是链表文档中的一种。在磁盘上它由若干个不一定连续的盘区组成（见图11-6）。每一盘区可取首字或尾字为指址器，指出文档中的下一个盘区，其余的用于存放数据。

实际上，每个盘区指址器中存放的是相连接的盘区的逻辑区号。

象这样的串联文档在表处理中叫做单链表（single linking list）（见图11-7）。

串联文档的优点是对任一盘区的可达性，即从当前盘区出发，通过链路可对文档进行正向或反向查找，从而能访问到所有其他盘区。它还便于增删改操作。其缺点是访问速度仍较低，它必须按相对区号的顺序，并通过缓冲区进行访问。

如果指址器中存放有连接的前后两个盘区的逻辑区号，则构成一种称之为双向循环链式线性表，简称双链表（见图11-8）。

这里，图11-8双链表用来表示处理具有可变长（存取）键项的索引文档。例如，设

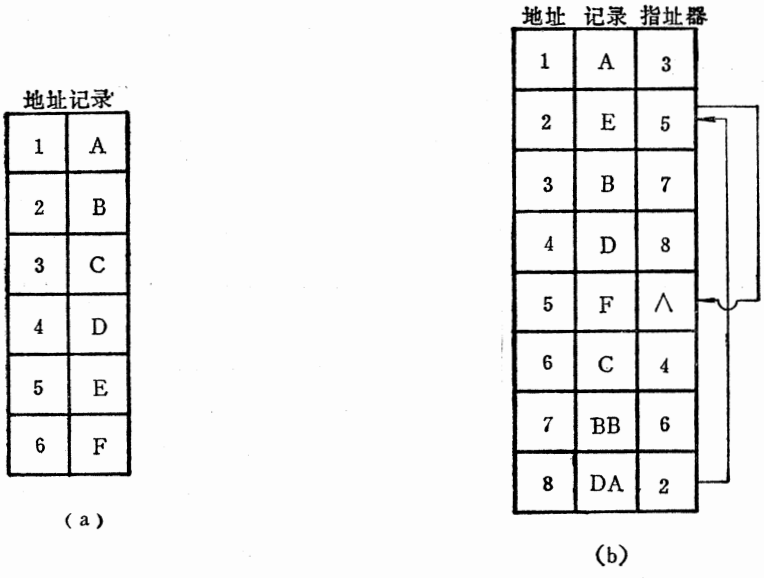


图11-5 链表文档

（a）线性表；（b）链式线性表。

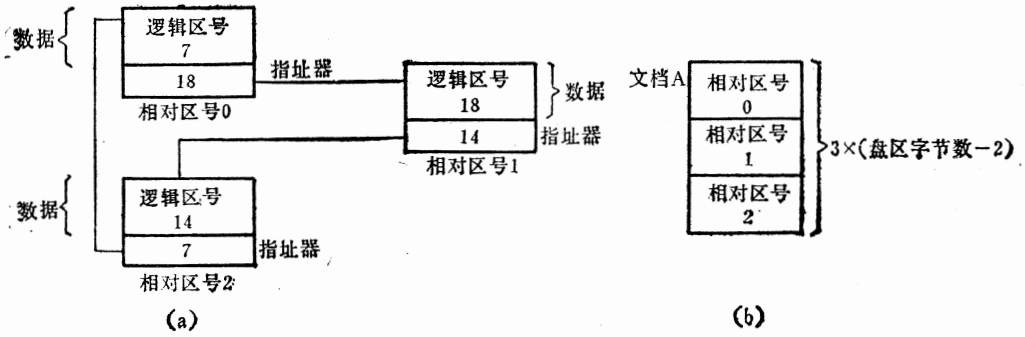


图11-6 串联文档

（a）串联文档的物理结构；（b）用户观点的串联文档。

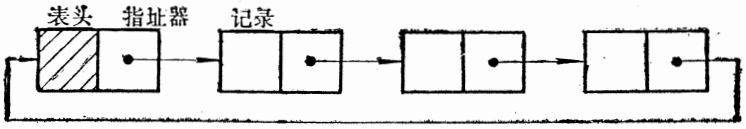


图11-7 （有表头的）单链表

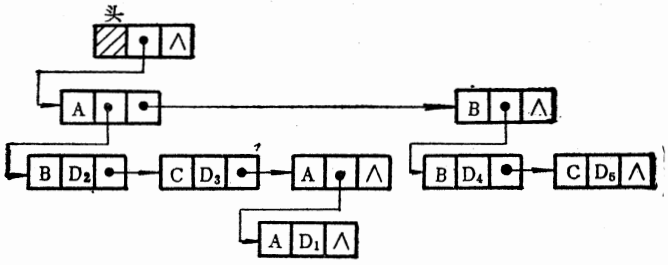


图11-8 一个双链表表示的索引文档例

{A，B，C}为索引键符号的集合，并设D_1、D_2、D_3、D_4、D_5为五个被储存起来的文献记录。给这些文献记录指定由这三个符号组成的存取键如下：AAA，D_1；AB，D_2；AC，D_3；BB，D_4；BC，D_5。

该双链表的检索按如下方式进行。

给定一个任意（查找）键（key），在模式识别中称它为子串（sub-string），它的出现与否通过对照该文档结构中的索引键（另一个子串）与之相匹配（matching）的方法来核对。子串匹配是逐层进行的，当在一个层里发现一个子串匹配（键）符号，则指址器P_i跟着就指到下一层里的二中择一的（键）符号集。然后，按以下情况结束该子串匹配操作。

（1）当该（查找）键扫描结束时，即没有更多的键符号需要进行匹配。

这里，又有以下两种情况：

① 如果在同一单元中，作为子串最后匹配（键）符号的指址器P_i现在指向一个文献记录D_j，则表明该索引键存在。也就是说该查找键在此文档中出现有与之相匹配的子串——索引键；

② 指址器p_i指向另外的（键）符号，即说明该查找键在索引键集合中并不存在。

（2）在当前层里没有找到子串匹配的（键）符号时，也表明在此文档中无此文献记录。

双链表虽然比单链表多占了一些存储单元，但这是一种用空间换取时间的办法。它给操作带来了很大方便，即查找可以自由地向左向右进行，而且还便于文档的增、删、改等操作。图11-9示明了在双链表中插入一个记录的例子。

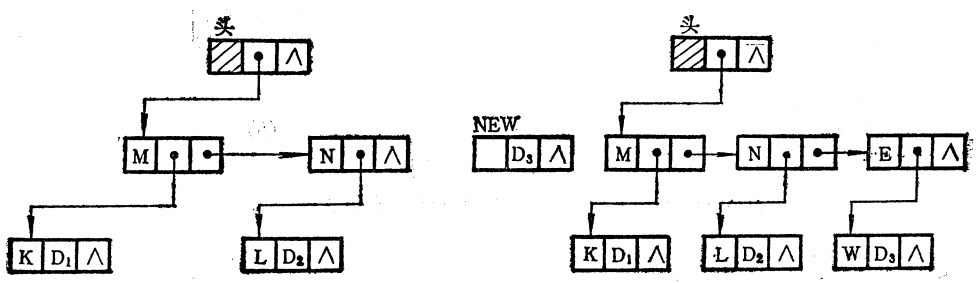


图11-9 双链表插入一个记录的例子

此外，如有必要可以根据要求设置多个指址器，以此将有关联的所有文献记录都链接起来，从而构成一种称之为多链表的文档。

（二）（目录）单元文档

（目录）单元文档是从手检方式中的单元词文档演变而来的。其方法是在一张卡片内，登录所有包含有这个特性标识的文献号码。一张单元词卡片上共有10列，编号为0到9（见图11-10）。

所有文献根据其文献号码的最后一位数列入相应的纵列里。凡是号码以0结尾的文献都记在0号纵列内；凡是号码以1结尾的文献都记在1号纵列内；依此类推，将各文

献都录入各自的纵列中。当进行复合概念（后组配）● 检索时，例如查找有关“计算机和情报检索”的文献资料，就只需取出这两张单元词卡片并找出其中录有相同的文献号码（70，79），然后根据文献号码（70，79）去找出所需要的原文文献资料来。

计算机

0	1	2	3	4	5	6	7	8	9
50	21	72	123	144	55	96	187	68	19
70	131			184	65		217	348	79
				234	185		247		
				274	235		307		

情报检索

0	1	2	3	4	5	6	7	8	9
70	141	82	133	84	45	106	87	88	29
110	151	102	163		195	116	237	188.	79
									109

图11-10　单元词卡片例

由于磁盘被划分成盘区页面，因此可以仿照单元词文档做法来组成（目录）单元文档。例如以柱面为单位，将文献号码按照个位数为0者置于0号柱面C_0中，为1者置入C_1中，依此类推，将所有文献都按文献号码的末位数而分别被录入C_0，C_1，…，C_9中（见图11-11）。

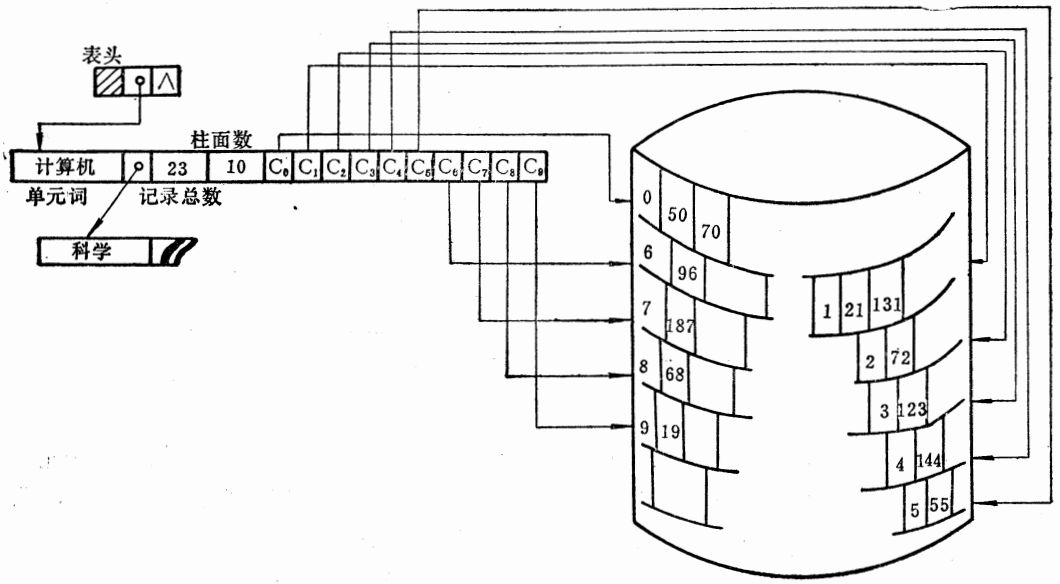


图11-11　（目录）单元文档

● 用单元词进行组合检索（后组配）时称为复合概念检索。

这种文档可以被看作是倒排文档（inverted file）的一种特例。

四、树形结构（tree structure）

树是图论中最重要的概念之一，它作为文档结构之一而被广泛地应用于计算机科学和系统工程之中。

有关树的定义很多。现在选择这样一种定义，即如果我们认为图是由这样一组节点和一组枝组成的，而这里每一个枝严格连接两个节点的话，则一棵树就定义为不带环的并至少占有两个节点的一有限连接图。为了说明一个环，先解释一下一条链的含义。用$u_R=〔x, y〕$表示连接两个节点x和y的边u_R。一条链是一个边的序列，其中每条边u_R有一个和前面的边u_{R-1}的共有节点，而与后继边u_{R+1}有另一共有节点。因此，$〔a, x_1〕$，$〔x_1, x_2〕$，$〔x_2, x_3〕$，$〔x_3, b〕$是一条链的例子。而一个环是从一个节点开始并终止于同一个节点的一条有限链（在本链例中$a=b$）。

注意，在实际应用中有一个节点是当作特殊的节点被挑选出来的，这个节点通常称为树的根。并且在这棵树中的任一其它节点只能由根开始来达到，即沿着边的一条链继续进行下去，直到被搜索到该节点为止。由根开始的每一路径（有向链）将最后终止于这样的一个特殊节点，由此再也没有分枝出现。这样的一些节点称为树叶。

树形结构在数据结构中是属于非线性的。

图11-12(a)是用树形结构表示的一个文献著录事项例子。

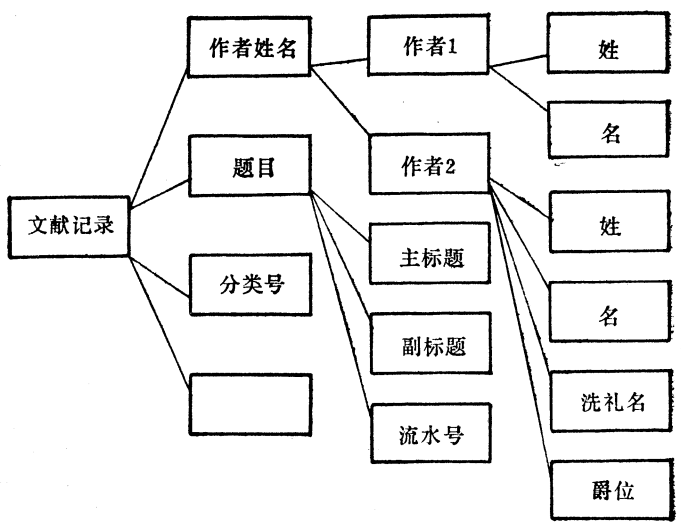


图11-12(a) 一个树形结构例子

（一）二元树

二元树（binary tree）是这样一种结构，在其中每个节点（除树叶那些节点外）恰好有由它出发的两个分枝。索引文档中的索引键可以用二元树表示出来。此时就可以利用对分搜索法进行检索。对分法用对该（索引）键集合逐次分半方式进行下去，在每次分半中丢掉一半作为不包含被查找的键的当前集合。当该集合包含有N个有分类的（索引）键时，检索时间大约是$\log_2 N$。

这种结构的优点是查找速度比较快。其缺点是增删比较麻烦并且不能进行多键

检索。

（二）平衡树

平衡树（balanced tree）是一种树结构，在它的任何节点处，左分枝上的子树和右分枝上的子树大体有同样多的个数（见图11-12 b）。在平衡树中检索路径比最优值稍长，但不会超过45%。期望的检索时间和插入时间大约还是$\log_2 N$。

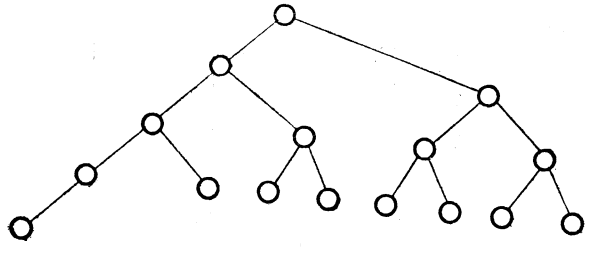


图11-12(b) 一个平衡树例（二元树）

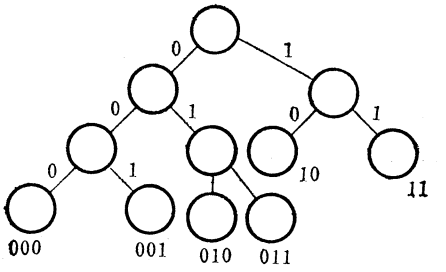


图11-12(c) 编码树例

（三）霍夫曼树

图11-12（c）给出了一棵二元树的树叶编码说明例。从一节点向左下行为0，向右下行为1。这样，树叶的编码为{000, 001, 010, 011, 10, 11}。用这些编码写的信息很容易区别开来，不至含混不清。编码的要求在于用比较短的代码来表示常用的符号，而少用的符号代码可以稍长一些，这样使得表达信息的代码位数较少。

如何设计最佳的编码？答案是归结为找一棵二元树，使得权 $m(T)=\sum_{(i)} p_i l_i$ 取最小值。这里，l_i是第 i 个树叶到树根的长度（每条边的长度设为1），p_i是第 i 个树叶的概率权，$\sum_{(i)}$ 是对所有的树叶求和。这样的二元树称为最优二元树。

霍夫曼（Huffman tree）树广泛应用于编码设计。但在信号检索中，到目前为止还未见到它有什么应用。

五、网状结构

上述线性结构、树形结构和链表结构都是网状结构（network structure）的特例。关于网状结构的数据组织可参见本章第11.3.2节的讨论和有关文献。这里，将着重介绍索引文档和倒排文档，尤其是关键词倒排文档。

（一）索引文档

人们在文献检索中通过文献查找有关的特性标识，这是常用的顺序查找方法。也能通过文献特性直接查找出全部具有该特性的有关文献资料，这是随机检索方法，并由此引出了索引文档、倒排文档等结构组织的检索用的文档。

如同双链表中所述，列出一个（索引）键和与之相应的文献记录的地址对照表，这张表就叫做索引（index），带有索引的文档简称为索引文档（indexing file）。

显然，索引项必须按照升序和降序进行顺序排列，而文档本身则可以按顺序或不按顺序组织，前者称为索引顺序文档，后者称为索引非顺序文档，统称为索引文档。

如果将索引方法和链接方法结合起来，则能组织一种称之为索引链接文档。

将索引本身当作文档再引出一个索引，即索引的索引，便能组成一种带有多级索引的文档，简称为多级索引文档（multiple indexing file）。

一般来说，在存储媒体上索引文档可分为两个区：索引区和数据区。

建立文档时，文献记录（数据）按物理次序存入，同时将文献号码作索引键和文献记录地址顺序记入索引项而自动建立起索引区。数据全部输入完后，将索引区进行一次排序。最后的索引区便是索引键（文献号码）与文献记录地址的一张对照表。该表是按索引键的升序或降序排列而成的。由于数据文档不一定顺序存放，所以必须每一个记录有一个索引项。

在多级索引里，各级索引键可以分层次先后按照升或降序进行排列，象COBOL语言中排序语句就具有这种功能。

图 11-13 是一个索引顺序文档例子。

实际应用中，这种文档分三个区：索引区，数据基本区和数据溢出区。数据溢出区是为了解决追加新的文献记录时进行插入而设置的。溢出区有两种安置方法，一是集中存放；一是分散存放，每一柱面（或盘区）有一溢出区。通常采用后一种方法。溢出区和基本区的比一般约为 1:6 左右、具体情况究竟为多少，这由系统设计决定。

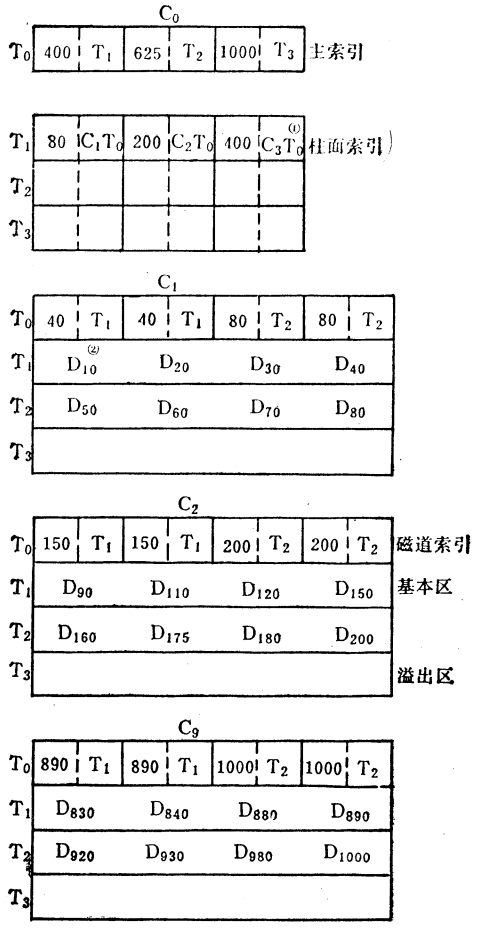


图11-13 一个索引顺序文档的实现例子

① C_3T_0表示 3 号柱面 0 号磁道；

② D_{10}等为文献记录地址。

本例中索引区是采用分级索引的，有主索引、柱面索引和磁道索引。

查找按下述情况进行。

从主索引查出柱面索引的分布，从柱面索引查出磁道索引的分布，从磁道索引可以查出所要检索的文献记录的地址。在使用时，主索引常驻于内存。

注意，磁道索引中，每一个索引项有两个子项。一为正常索引项，一为溢出索引项。在插入新记录之前，这两项相同；在插入后，要改变溢出索引项。例如在本例的 C_2 中插入一个文献记录 D_{190}，则将 D_{190} 置于溢出区中，而 C_2 的磁道索引改为：

150	T_1	150	T_1	200	T_2	190	T_3

当然，在情报检索中通常需要有主题索引、标题索引、题中关键词索引 KWIC 和题外关键词索引 KWOC 等。它们都可以通过自动索引产生出来。至少目前有四种轮

排法可以产生出主题索引或KWIC、KWOC索引。它们是置换主题法、循环轮排法、交替轮排法和换轨轮排法。

（二）倒排文档

倒排文档的重要性，在介绍布尔型检索时将显得更加清楚。

图11-14示明了一个简化的倒排文档例子。这里，特性标识可以是关键词或其它；文献记录可以是文献号码（相对地址）或盘区地址（绝对地址）或其它。

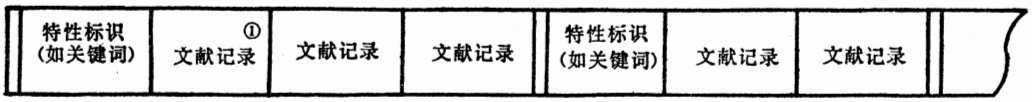


图11-14　一个简化的倒排文档例子

① 文献记录可以是文献原文，但多数为馆藏号或盘区地址等。

图11-15说明了倒排文档的一般结构形式。图11-16给出了一个典型的倒排文档例子。

特性标识	文献记录总数	同文献记录总数的单链表张数	首张单链表的地址	———————	末张单链表的地址 ④
计算机	23	23	19	～	106

图11-15　倒排结构示意图

注：① 取单元词例，但单链表是23张，且首地址有23个。

如果说顺序文档是“词从属于文献”的话，那么倒排文档就是“文献从属于词”了。从检索模式看，顺序文档需要连续读出所有文献记录，而倒排文档只要读出与所需求的特性标识有关的那部分文献记录。

这类文档的优点是文献特性可以按照任何次序排列，存取速度也比较快。它的缺点是增删改操作比较慢。

应该认为情报检索大多以倒排结构组成检索系统。例如：

（1）ADABAS（有自适应性的数据库），是一个大量使用倒排结构组织的数据库系统，也是一个运行比较好的系统。

（2）STAIRS（存储和信息检索系统），是一个成功运行着的类似倒排结构的检索系统。

六、散列函数和直接存取文档

与前面介绍过的有非常密切关系的，尚未讨论的一种文档结构就是常说的散列存储结构。按此实现文档的技术常称之为杂凑编址。假定可以通过一些（存取）键 k 来存取数

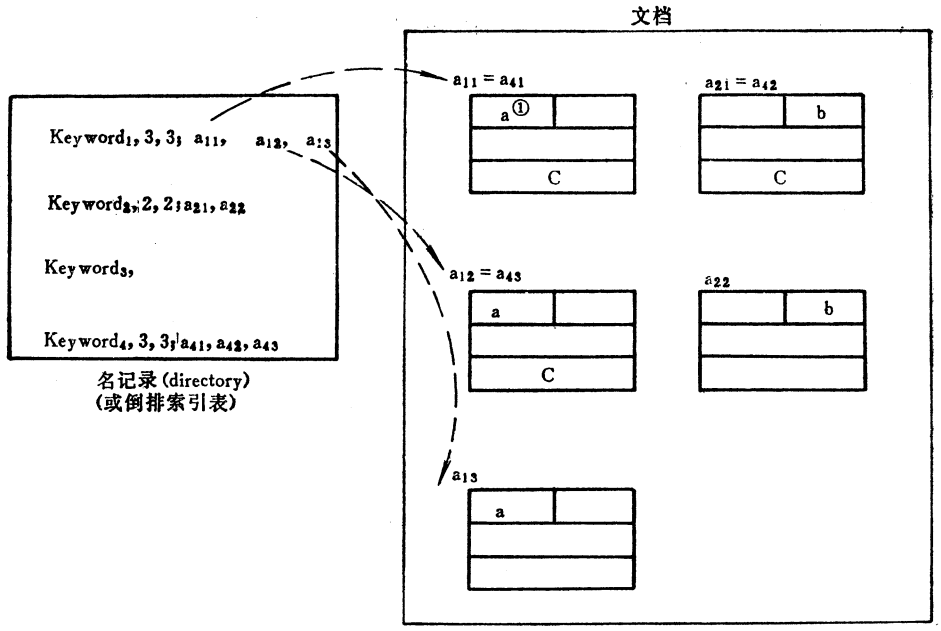


图11-16 一个典型的倒排文档例子

注① a，b，c 表示文献记录的特性字段值。

据 d（文献记录），那么在存储器里该数据 d 的地址就能通过一个键变换函数 H(k) 来定位。即寻找一个这样的杂凑函数（Hash function）H(k)，将键转换为地址 A(d_k)，这里 d_k 为对应于键 k 的文献记录数据，A(d_k) 为其杂凑地址。理想的 H(k) 应该是能使得键的集合连续均匀地散布在可能得到的存储单元上。当然，如果该函数是一对一的话，那么这一方法就很合乎理想了。但是由于可能的键值的分布区域比可能得到的存储单元地址为多。因此，给定任何杂凑函数●，会出现两个不同的键可能映照（mapping）到相同的地址 $A(d_{k_i})=A(d_{k_j})$ 上。这就引出了冲突问题。

有一系列处理冲突的方法。但从本质上说，可把它们分为两类：使用指址器将冲突的键链接起来；不使用指址器。有关处理冲突的细节和技术请读者参看这方面的文献。

杂凑编址和寻址方法的优点是存取快速，比较节约存储空间。因为这种方法只要调用一个算法过程，而无须占用存储空间作索引。缺点是不易找到一个良好的算法，而且有时冲突现象发生过多时，会延长存取时间。

在情报检索中，杂凑法的应用倾向于在表构造和查找过程方面。一个重要的应用是构造索引表和查找索引文档上。

所谓直接存取文档（directly access file)实际上是指利用杂凑法进行组织的文档。图 11-17 是一个随机文档例。它是一种直接存取的文档。

这里，随机文档是由若干个不一定连续的盘区组成，其中每个盘区都用于存放数据信息。为了表示各盘区之间的相对关系，因此随机文档还要包含一个按串联文档这类

● 这里散列（scatter）、哈希（Hash）、杂凑（cross）的函数、编址或方法，都是一个意思，即将存取键通过变换函数H(k) 映照到地址A(d_k) 上，d_k为对应于键 k 的文献记录数据。

结构组织起来的（多）键地址目录（本例中仅有一个主键）。

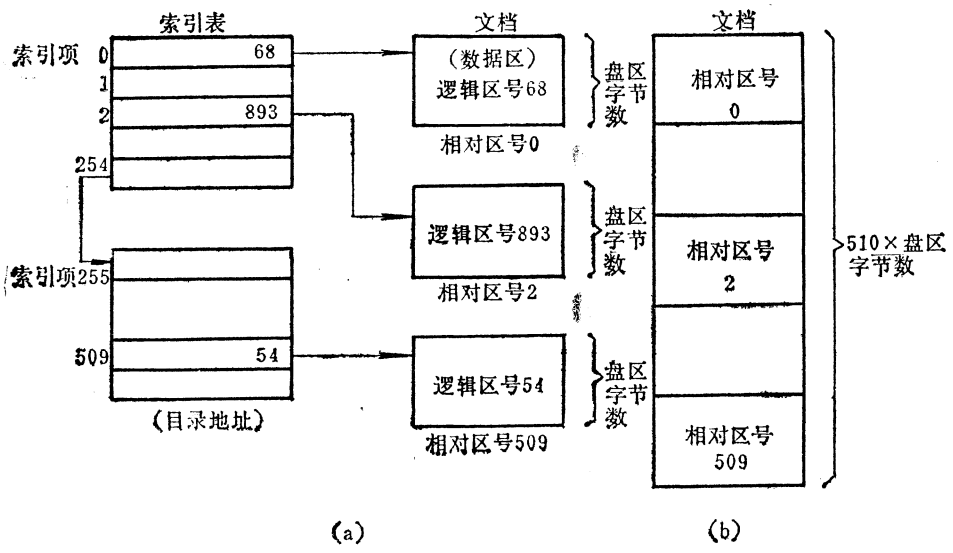


图11-17　一个随机文档例子

（a）随机文档的物理结构，（b）用户观点的随机文档。

查找随机文档时，先读取（多）键地址目录，随后根据索引地址存取相应盘区的文献记录内容。这种随机文档的优点是便于增删文档的数据区，而且可以直接读写任意盘区的内容。

以上叙述了关于存储情报所需的文档结构的基本内容。涉及文档的安全性和保密性等其他问题请参见有关文献。

11.2.2　词库组织

检索系统中的词库，在图 11-2 里已经用虚线画出来了。它对系统的性能起着重要的作用，对每个检索策略的构造和实际的检索过程也有很大作用。由于词库有表达文献中的概念的能力，因此它是影响标引的完备性和深度的主要因素。

其次，词库对用户向文献情报服务中心提出的询问也有影响，这是因为用户无法严格遵循索引语言的要求来表达和描述出他对情报的真实需要。

一、定义和解释

对一个词库（或称叙词表）可以按其功能及作用，也可以按其结构来下定义。

就功能而言，词库是一种术语控制手段，用以将文献情报、标引人员或用户的自然语言转换成更受约束的文献语言（或索引语言）。

就结构而言，词库是从语义学上和从类属意义上相关的受控的或动态的专业词语汇编(或综合词语汇编)。

这种作为用户询问与文献情报资料库的纽带的词库，是自然语言的一个有组织的结构式的子集，用以描述文献情报、研究对象或数据类集的主题内容。

一个词库所收入的词称作词条（词汇的项目），而只有哪些被词库允许用于标引的

词才称为叙词。叙词是以词库中作为词条而存在的有代表性的和已定形的词、词组、表达式和符号为特征的，用以明确地表达文献情报和询问的概念。

词库可以按下述四条准则分类：

（1）单语种的或多语种的；

（2）未组织的或已组织的；

（3）非控制的或受控制的；

（4）专业性的或综合性的。

一个单语种词库只限于收入一种人类语言的词汇。在一个多语种词库中则常常出现不止一种自然语言。可以有一种文献标引所必须使用的推荐语言（受控的多语种词库），或标引时各种允许使用的语言都是等效的（自由的多语种词库），都可以用作多种语言的词库。

一个未组织的词库，只包含叙词而不经任何组织。一个已组织的词库，则在词条之间保持着语义的和等级的相关性。词库中的相互参照使得各词汇在一个概念网络中的相互关系清晰。这就给人工建造词库时提供了解决词的预组配问题的条件（词的预组配的反面是词的后组配，它通常是出现在联机询问和词库对话上）。

一个非控制的词库不使用推荐词，也就是说词库中列为词条的所有的词在原则上都是叙词（即使用的是自由词）。

在受控词库中不许可用未经审定的词来标引文献情报。这些未经审定的词叫做非叙词。如果在文献情报的标引语言中带有非叙词，则文献情报将被当作出现差错而被放弃或不被认可。

一个专业性的词库是指在本专业范围内所收集到的并组织起来的词汇表，其中的各词条将不考虑它是否出现在其他专业上语义的相关性问题。反之，综合性的词库就是包罗万象的，囊括所有学科的相关性的总词汇表。

总之，不论哪类词库，它可能是基于语义关系建立起来的（例如人工词库），或者是基于语法和统计频度而建立起来的（例如自动词库）。

二、词库结构

在一个词库中词条的概念内容主要是通过与其他词的关系说明来表示的。这样，由于将词条放到语法学和语义学的位置上，一个词条对于其他许多词条关系的网络就能提供出词库的一类定义。

这里，应注意到三种类型的相互关系：语义学的类同；等级的关系；概念的联系或近似的关系。

因此，词库结构上应为词条间有关联的情况提供相互参照表，并可以考虑包含有以下几方面的内容：

（1）用参照，用代参照，用…组合参照；

（2）同义词、同形异义词、反义词、近义词、类义词和对义词；

（3）等级关系（上位概念、下位概念）、相关（类同）关系、联想（旁族）关系、双关义词、错用词、倒文同义词和正反共义词；

（4）潜义词、今古义词、词外义；

（5）翻译词、缩写词；

（6）范围注释等等。

总之，提供给情报检索用的词库、大体上可以归结为：

（1）词库对文献的术语或者用户的情报询问提供标准用词。

（2）词库必须有足够的专业深度，以便允许大多数检索能在一个可接受的查准率水平上进行。

（3）词库必须是充分预组配的，以避免假组配和不正确的词间关系的问题。

（4）词库必须通过控制同义词、近义词等来促进标引和检索中的一致性。

（5）词库必须通过区分同形异义词、正反义共存词等，以及通过对那些意义和范围不清楚的词（如潜义词等）加以定义，以此减少术语上的含混性。

（6）词库必须通过等级关系和相互参照关系，帮助标引人员和检索人员选择用来表述某一特定主题的最合适的词。

11.2.3 检索策略

以上讨论偶尔提到过检索效率和适合检索用的文档结构，但尚未涉及有关查找的实际过程。在文献检索中，查找结果是与询问有关的文献记录的子集合。实际上，情报检索系统是检索资料或者提供有关某一主题的文献、情报的出处（即存放地点）。因此，它起到了某种“通报”的作用。

所有的检索策略都是建立在询问与存储文献之间的比较基础之上的。当询问和文献代表（例如经过聚类分析所得到的分类口代表）进行比较时，这种匹配往往只能间接进行。

不同类型的检索策略之间的差异有时也理解为表达查询信息所用的语言不同。查询语言的性质往往决定了查找策略的性质。本节将介绍两种查询语言和布尔型检索方法。

一、查询语言（query language）和查询表达式（query expression）

（一）查询语言：表语言和行语言

存储情报不是情报检索的唯一目的。检索情报是为了让用户能索取到他所需要的文献情报或图表数据等资料。因此，组建机读文档库或数据库不是为了别的，而是为了能起到资料流通的作用。

为实现流通的目的，应该给用户设计一种能正确地表达情报要求和与系统之间进行对话的语言工具。查询语言和查询表达式是这种语言工具之一。

虽然，作为人类之间交流工具的汉语、日语、英语、法语、德语和俄语等自然语言灵活又练达，但由于历史悠久反而造成语法复杂、语义含混，因此不宜用来直接作为情报检索系统中的查询语言。

只有某种受限制的、没有语义含混的和有规则可循的经人工设计的语言才能作为查询语言或查询表达式。

一般地说，查询语言应具有：

（1）描述精炼，不含混；

（2）语法简单，便于掌握；

（3）能够具有描述各种情报要求的灵活性；

（4）语言便于扩充。

在表现形式和内容上，查询语言分为确定型的和不确定型的，专用型的和通用型的几种。

确定型的查询语言是指以固定方式频繁地提出情报要求，频繁地使用的一种检索语言。它必须简单可行和行之有效。

不确定型的查询语言是指用户每使用一次即改变他的要求，因此它必须具有能描述各种情报要求的灵活性。

专用型的查询语言是为专门用于一个系统而设计的，因此它对每个检索系统都各不相同。

通用型的查询语言适用于各个系统。它有两种表现形式和至少有六种以上功能。

1．通用型查询语言的表现形式　语言表现形式分表语言和行语言两种。

（1）表语言。采用固定格式的提问单，根据情报要求把必要的情报写入到固定的栏中。

表语言的优点是预先规定好什么样的情报在何处描述，所以按情报要求描述的量少，简单可行。其缺点是联机查询有困难。

本章第 11.2.5 节介绍批处理检索系统的例子中将采用这种表语言。

（2）行语言。它是一种采用自由格式描述的、接近口语化的、有限制的、结构式的语言。从终端的使用角度看，它是一种简易的语言工具。其缺点是要求对情报描述的量比表语言多，也比较容易产生描述错误。

表 11-3 说明用行语言描述的（也可以用其他语言描述，如英语）、以图 11-19 为例的查询要求。

表11-3　行语言描述例子

开始␣①1.␣谢谢您，␣请您回答口令!␣(系统)②
␣2.③␣南京大学数学系␣␣南京汉口路（用户)②
␣3.␣请输入您的提问!␣(系统)
␣4.␣表达式：␣Q1＝'操作系统'(φ1，φ1）或'PL/1'(φ1，φ1)␣(用户)
␣5.␣请再输入!␣(系统)
␣6.␣表达式：␣Q2＝'73.87221'(φ2，φ1）或'IBM'(φ2，φ1)␣(用户)
␣7.␣请再输入!␣(系统)
␣8.␣表达式：␣Q3＝5④×(Q1与⑤Q2)␣(用户)
␣9.␣请再输入！␣(系统)
␣10.␣N0！␣(用户)
␣11.␣␣输出条件：␣终端显示或打印？␣(系统)
␣12.␣␣显示！␣(用户)
␣13.␣␣请再次提问！␣（系统)
␣14.␣N0！␣␣谢谢!␣(用户)
␣15.␣␣承蒙关照！␣谢谢！␣下次见！␣(系统)
结束：␣

① ␣表示空格。　② 注释。　③ 数字或字母占半字。　④ 输出截值，这里 5 表示只要输出命中文献记录子集的前 5 篇。　⑤ 运算符（或，与）可以直接用'OR，AND'表示。

（3）处理方式：分解释执行和编译执行两种。

解释方式适用于要求响应快，能用会话形式进行询问的检索系统，大多用于行语言处理。编译方式适宜于处理复杂的有大量数据的查询系统，常用于表语言处理。

2．通用型查询语言的功能　它应具有如下六种功能：

（1）检索条件。象子串匹配中的一致条件、比较条件或对检索项加权处理等。

（2）运算处理。对带检索条件的项或记录能作（加权或不加权）任意组合的逻辑运算处理。

（3）编辑排版。对检索结果能按照用户编辑要求作格式输出。

（4）引导功能。根据需要能为用户进行检索处理时提供有关系统性能介绍的功能，包括关于查询语言的使用方法、菜单（menu）式检索示范和有关文档库或数据库的数据类型、数量等。

（5）建立用户文档的功能。能把涉及到用户本身的信息（例如姓名地址等）以及有保密授权的数据（例如用户职称、工资等）建成用户文档，等等。

（6）语言扩充功能。查询语言应有可扩充性，以便能适应系统环境变化和用户的新要求。

（二）查询表达式

查询表达式是指查询语言中描述查询的一种表达式，例如布尔表达式、统计加权的算术表达式，以及集合运算用的代数表达式等。

二、检索方法

不同的文档结构有不同的检索方法。按检索系统的形态和内容分至少有三种模式：结构检索、内容检索和概念检索。

（一）结构检索（按键查找）

根据存储要求，对各记录对象都要给定一个（存取）键（可以是相对符号地址或盘区绝对地址），以确定它们在文档库或数据库中的位置，以后只能按照这个（存取）键从库中取出它们。如果不知道这个键，就不能检索出一个特定的记录对象。这种“按键检索”模式适用于线性结构的（如顺序文档）或树形结构的（如分类号索引文档），以及链表结构的（如串联文档）的检索。

（二）内容检索（特性查找）

在库中寻找符合某种特定需要或者能表达若干特性要求的哪些记录对象。这种“按内容检索”模式通常用检索键（例如叙词）的逻辑组合来表达该查询要求，并且通过键地址变换法使检索键和索引键对应匹配方式来实现。如果匹配成功，则表示找到，而且找到的是一组满足特性要求的记录对象。反之，表示没有找到。按特性查找方法适用于非线性结构的文档检索，例如索引文档、倒排文档等。

（三）概念检索

如果使用词库，则可引入一种“概念检索”模式。这是考虑到一篇论文或文献总有一个它所论述的一定的涉及范围或者主题事项。我们将使用词库中的用词，它或多或少可以描述该主题事项。这些用词既应指明该文献所论及到的全部内容，又应指明该论文的主题概念是什么。

这种检索模式能给用户提供一种扩检（族性检索）和专检（特性检索）手段。而且是通过用词管理系统和用词的亲缘与等级关系来实现的。

但从方法论考虑，会有许多不同类型的检索方法。这里，只着重介绍一下布尔型检索方法。它是情报检索中最常用的一种检索方法，并将检索出询问为真的那些文献情报

资料子集来。

（四）一般数学模型

图 11-18 是对文档库进行检索的一般数学模型。文档库 FB 经过 T_1 变换建立起文献——词关联矩阵 D，D 可以表示为一系列文献——词关联向量 D_i（$i=1\cdots n$）。询问 Q 经过 T_2 变换成询问——词关联向量 q。μ 是对 D_i 和 q 进行模式匹配操作。R 是满足匹配条件要求的阶段检索结果向量。R 经过 T_4 变换便得到最终结果 A〔D：Q〕。A〔D：Q〕是满足询问要求的一组文献资料。

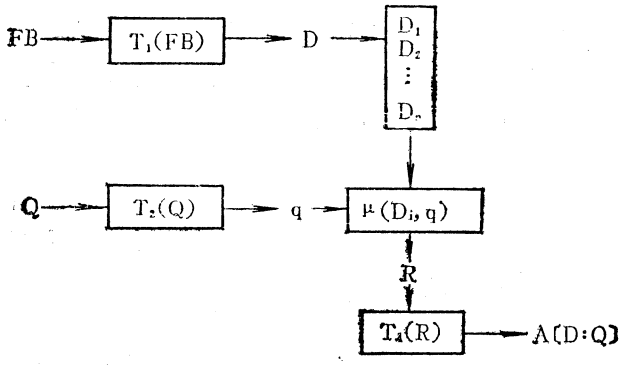


图11-18　检索的一般数学模型

（五）布尔型检索

以具体的文献——词关联矩阵和询问——词关联向量为例来说明布尔型检索处理过程。设

$$D=\begin{array}{c} \\ D_1\\ D_2\\ D_3\\ D_4\\ D_5\\ D_6 \end{array}\begin{array}{c} \begin{matrix} k_1 & k_2 & k_3 & k_4 \end{matrix}\\ \begin{pmatrix} 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 1 & 1\\ 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array},\quad T_2(Q)\to q=\begin{array}{c} \\ \\ \\ \\ \end{array}\begin{array}{c} \begin{matrix} \end{matrix}\\ \begin{bmatrix} 1\\ 1\\ 1\\ 1 \end{bmatrix} \end{array}\begin{array}{l} k_1\\ k_2\\ k_3\\ k_4 \end{array}$$

这里，矩阵 D 表示文档库 FB 经 T_1 变换所得到的文献——词关联矩阵，其行向量表示文献记录（D_1，$D_2\sim D_6$），其列向量表示文献特性（k_1，k_2，k_3，k_4）；询问——词关联向量 q 是原查询表达式 Q 经 T_2 变换后所得到的。注意凡出现有 1 的地方表示该文献包含有这种特性。

检索过程是，先取 q 中为 1 的元素 k_1，查找 D 中相应 k_1 列所对应的文献记录子集合 $R_1=\{D_1, D_2, D_3, D_4, D_5, D_6\}$。依此类推，取出 k_2、k_3 和 k_4 所对应的文献记录子集合 $R_2=\{D_1, D_2, D_5\}$，$R_3=\{D_1, D_2, D_3, D_4\}$ 和 $R_4=\{D_1, D_4\}$。这表示按向量分量进行模式匹配 μ 操作，并要求其匹配双方都为 1 才能满足检索要求。这样，便得到阶段检索结果向量 $R=\{R_1, R_2, R_3, R_4\}$。如果经 T_2 变换，上述 q 的原询问逻辑表达式成为 $Q=(k_1\wedge k_2)\vee(k_3\wedge \mathrm{NOT}\, k_4)$，则 R 经过 T_4 变换便得到最终检索结果 A〔D：Q〕=

$(R_1 \cap R_2) \cup (R_3 - R_4) = (\{D_1, D_2, D_3, D_4, D_5, D_6\} \cap \{D_1, D_2, D_5\}) \cup (\{D_1, D_2, D_3, D_4\} - \{D_1, D_4\}) = \{D_1, D_2, D_5\} \cup \{D_2, D_3\} = \{D_1, D_2, D_3, D_5\}$。这里的 $\cap$、$\cup$ 和 $-$，分别表示集合的交、并、差运算符；{ }表示集合符号；A〔D：Q〕表示询问 Q 对文档库 D 进行使之查找为真的检索结果表示。如果取 $Q = k_1 \wedge k_2 \wedge k_3 \wedge k_4$（以合取范式表示的查询表达式）并经 T_2 变换后仍然变成上述 q，则检索结果 A〔D：Q〕$= \{D_1\}$。

将它推广到一般文献——词关联矩阵和询问——词关联矩阵，便有如下表示法：

$$FB \xrightarrow{T_1} D = \begin{matrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{matrix} \begin{bmatrix} k_1 & k_2 & \cdots & k_t \\ & d_{ij} & & \end{bmatrix}, \quad Q \xrightarrow{T_2} q = \begin{bmatrix} q_1 & q_2 & \cdots & q_l \\ & q_{jk} & & \end{bmatrix} \begin{matrix} k_1 \\ k_2 \\ \vdots \\ k_t \end{matrix}$$

这里，n 为文献记录总数，t 为文献特性总数，l 为询问个数，q 为原查询表达式 Q（共有 l 个）经 T_2 变换后所得到的询问——词关联矩阵。

检索处理分为两步（如果以合取范式或析取范式表示查询表达式Q，则第二步只要做集合交或集合并运算）：

（1）对所有询问 q_i，取出全部特性 $k_j = 1$，并找出所对应的文献记录子集 $R_{ij} = \{D_p, d_{pj} = 1, p = 1 \cdots n\}$，$i = 1 \cdots e$，$j = 1 \cdots t$。

（2）对 l 个查询表达式Q，按照原逻辑式对所有 R_{ij} 进行集合运算，便得到检索结果 A〔D：Q〕$= \{2^D | \mu(D, Q) = 1\}$。这里，$d_{pj}$ 为特性值，μ 为匹配函数，2^D 表示集合 $\{D_1, D_2, \cdots, D_n\}$ 的幂集。

这种检索处理实际上等价于对D按列作成倒排文档，然后依照询问Q的逻辑表达式直接进行集合运算的结果。

11.2.4 系统设计和应用软件

一、系统设计

本节中将叙述有关情报检索系统设计的几个环节、主要内容和设计方法问题。

（一）设计环节

关于情报检索的系统设计包括有以下几个主要阶段：

1．形成概念阶段　从提出问题到调查研究和分析解剖系统的目标、环境和功能，研制出一个系统的“纸上模型”。

2．建立模型阶段　在这一阶段里，研制一个小规模的工作模型并上机试验。

3．模型的评价　通过数据模拟，对建立的模型进行局部性评价（如系统性能价格比、经济效益估计等）。

4．根据评价的结果修改或优化设计　这包括对计算机系统的配置和文档库或数据库的设计、标引和词库组织与检索方法的修改或优化工作。

5．全面执行系统　在这一阶段里，主要是全面实施阶段，进行程序设计和系统投入运行、维护、管理等几方面的工作。

6．提高运行效能　以用户反馈信息为依据，实行某种质量控制工作，以确保系统能适应服务对象的要求。

7．系统鉴定阶段　系统经过试用期考验之后，应正式通过技术鉴定，办理工程移交手续。

（二）设计内容

这里，主要的设计内容将从情报的存储、检索和词库等几方面来考虑。

在目前的技术条件下，已经有可能为用户提供一个良好的服务环境，即可以设计出一个有联机询问的实用性系统供用户使用。

在通过系统可行性报告之后，将从计算机系统的配置、输入对象的确定、机读文档库或数据库的组织、词库的管理和输出的设计，以及应用软件设计等几方面进行具体考虑。

1．可行性分析　设计单位对情报部门用户提出的自动化业务处理问题，结合现有系统进行深入调查研究，并就系统目标要求、实施的环境条件和系统功能三个方面作出全面分析，写出可行性报告。

2．一般设计　一旦用户单位审核批准可行性报告以后，决定投资建造系统时，可以委托设计单位负责拟定或提出计算机系统配置方案（包括硬件体系结构和系统软件要求），并就应用系统的功能要求和实施方案进行论证（属形成概念阶段）。

3．详细设计　在一般设计完成之后，接着就进入到详细设计阶段。这是在确定的分期分批的系统配置上，由设计单位以设计小组名义拟定或提出应用软件实施方案并着手进行系统模拟（属建立模型阶段）。

在这一阶段里，多数采用由顶向下的设计方法或者按照功能要求采用带反馈的分层输入输出和加工的设计技术进行设计。

在这一阶段必须解决：（1）数据准备工作（包括数据收集和操作方式）；（2）代码设计；（3）输入输出设计；（4）机读文档库或数据库的数据结构和描述语言的设计；（5）硬件接口和软件接口设计；（6）数据通信设计；（7）用户数据文档的保密设计；（8）系统安全性和保护措施的设计；（9）系统模拟工具的设计；（10）教育、培训等工作计划和进度表。

（三）设计方法

从总体设计到个别设计，一般是采用自顶向下的设计方法。不过由于了解计算机和文档库或数据库的人员并不很了解情报检索的业务知识，而且很可能他们还缺乏实际经验，再加上系统设计过程中有些用户要求，一时还没有办法得到解决，因而只好结合采用自底向上的设计方法。

对情报检索系统而言，将围绕着下述三方面进行系统设计方面的研究：（1）存储些什么，如何存储？（2）标引方式、词库组织和用词管理的决定；（3）用户和系统如何对话？

除常用的自顶向下方法和分层输入输出加工技术以外，还有很多其他的设计方法，例如分析法、探索法和计划评审技术等。

总之，从工程方法学看，系统设计应该达到这样的目的，即使得设计出的系统是一个功能齐全、实用可靠和经济效益较大的运行系统。

二、应用软件

情报检索系统最终是由所确认的应用软件来实现的。虽然，情报检索用的软件将随系统目标和功能的要求不同而有所不同，但情报的存储和检索是不会因系统要求不同而改变的。因此，它应该配有下列四个软件包：（1）建档分系统（存储）；（2）检索分

系统（检索）；（3）编辑分系统（输出）；（4）管理分系统。此外还视要求而定，配置下列软件包：（5）词库分系统；（6）引导分系统；待开发的软件包有：（7）自动标引分系统；（8）自动文摘分系统；（9）自动翻译分系统；（10）通信情报分系统；等等。

对应用软件设计时应该注意到下述几方面的要求：（1）方便用户，操作简单；（2）维护简便；（3）能降低应用软件开发和维护费用；（4）应用软件的可靠性高；（5）应用软件作业的自动化程度高；（6）应用软件应具有可扩展性、可移植性和通用性，并纳入系统软件之中；（7）标准化程度高等。

在情报检索系统中，常用于实现应用软件的程序语言有：COBOL 语言，PL/I 语言，Fortran 语言和汇编语言等。

11.2.5 实例

本节将通过实例示范性地说明一个情报检索系统的题材选择、标引作业、文档设计和存储检索的实现概况以及应用软件编制等有关问题。

一、脱机批处理检索系统

批处理检索系统，将介绍 1976 年研制的试验性情报检索系统，简称 ND76 系统为例。

（一）机读文档组织

当时考虑到图书馆的现状和程序设计的实现方便以及具体的检索要求，我们参照了美国国会图书馆机读目录 MARCⅡ，确定了 ND76 系统以“南京大学图书馆西文图书著录卡片”为收录对象（见图 11-19），采用段数固定、段长可变的半固定记录方式，形成机读文档记录（见图 11-20）。

P176　IBM 操作系统/360 程序语言 PL/Ⅰ：语言说明书
W92
INTERNATIONAL BUSINES MACHINES CORP.
IBM OPERATING SYSTEM/360;
PL/I: LANGUAGE SPECIFICATION.
NEW YORK, IBM, 1965, 168P.
(IBM SYSTEMS REFERENCE LIBRARY)
73-1347-08
7847;　　73.872221/I61

图11-19　南京大学图书卡片例

进行文档设计时，设定文献记录字段标识，如图 11-20 注中说明的情况。

这样，书目卡片内容按照图 11-20 注中标引要求作成原始数据。同时以计算机能接受的格式要求，在 80 列的穿孔卡片上穿孔，并制成输入文档的格式化数据卡片叠（见图 11.21 a、b）。

接着，通过读卡机将这些数据卡片输入到计算机，并加工成磁盘或磁带的顺序文档。

这种顺序文档的记录格式说明如同图 11-20 所示●。

记录格式	头标部									目录部					数据部			
	固定长									固定长					固定长		可变长	
字段说明	记录总长度	状态区分符	区分标识			指示符位数	分字段代码位数	数据基本地址	空白	字段标识	字段长度	字段起始地址	～(共15段)	字段分隔符	控制区		可变区	
			记录性质	书刊种类	空白										(1)	(2)	(3)～(15)	记录分隔符
长度	5	1	1	1	2	1	1	5	7	3	4	5		1				1
位置	1～5	6	7	8	9～10	11	12	13～17	18～24	25～204				205	206～265		266～4800	

图11-20 ND76系统机读文档记录说明

图11-22是对应于图11-19所示书目卡片的例子，经加工处理后建立在磁盘或磁带上的一个顺序文档记录样本。

ND76系统采用的是不分块记录方式，即一个物理存储记录块只存放一个逻辑记录。一般分有固定的或可变的（逻辑记录）、分块的或不分块的（物理记录）等组合存储方式和不作任何指定的存储方式。具体采用哪一种存储方式由文档设计时的要求确定。

（二）检索服务

ND76系统原先是以定题情报服务（SDI）为基础，即对特定的用户采用预先登录的方法，将有关用户本身的信息（例如姓名、地址等）和查询要求的信息，按照第11.2.3节介绍的表语言格式要求先作成用户提问文档（见图11-23）。然后将此提问文档加工成磁盘或磁带上的顺序结构的用户需求文档。这样，每当收进新的情报资料时，由系统定期地进行批处理，把符合查找要求的情报内容分发给各用户。其检索结果是以表11-7示明的输出格式、即以书目卡片形式提供给用户的，以便直接到图书馆索借已检索到的书刊资料。

其检索方法是采用表展开技术进行顺序查找的。做法是将用户需求文档展开成如同表11-4所示的一张提问表，并把一篇篇文献记录作成能被查找用的一张张检索标识表（见表11-5）。然后在两张表上（展开表和标识表）逐项按模式识别中的子串（项）进行匹配比较，直到查找成功或比较完毕为止。这种表展开查找过程有如同表11-6所示情况进行。

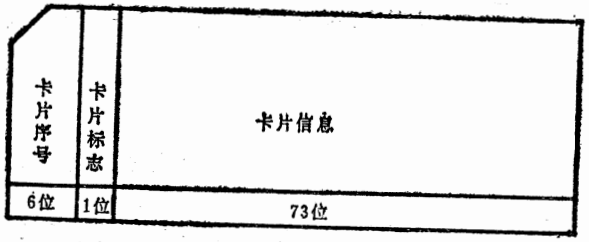
● 图中，数据部各字段说明如下：

（1）南京大学图书馆图书登录号长度15位置 1～15，标识 010。

（2）收藏日期（子段）长度 6 位置16～21，标识 020；发行年（子段）长度 6 位置22～27；出版国（子段）长度12位置28～39；文种（子段）长度12位置40～51；版本（子段）长度 4 位置 52～55；密级（子段）长度 4 位置 56～59；字段标记长度 1 位置60；（控制区字段）。

（3）南京大学图书馆图书分类号，标识 030。（4）北京图书馆图书分类号标识 050。

（5）中国科学院图书馆图书分类号标识 060。（6）北京航空学院图书馆图书分类号（或 UDC，国际十进分类法），标识 080。（7）作者项，标识 600。（8）篇名项，标识 245。（9）出版项（地点，出版机构和出版年代），标识 090。(10) 稽核项（页数、大小），标识 980。(11) 参考项，标识 970。(12) 备注项，标识 960。(13) 标题项，标识 690。(14) 文摘项，标识 520。(15) 关键词标识 955。



(a)

1 2 3 4 5	6	7	8 9 10	20	40	60	80
	1			73 -1 34 7- 08	, 76 07 01 $A 19 65 $B 0 02	01 I BM$C 0 01 00 E NG$D	AP $E 04 , $ AP 17 61 W9 2,
	2		0 $	AT 84 7, 1 $A	73 .8 72 21 / I 61 , 0$A ,	1$ AI BM, $ AI BM ϕ PE RA	TI NG S YS TE M1 36 0· $B PL
	3		/ 1:	L ANGUAG E	SP EC IF IC AT Iϕ N, 0 $AWE	W- YϕRK $B IB M$ C1 96 5, 1	$A 16 8 P. , 0$ A , 1$ A
	4		,	$AL IB M SY	ST EM S RE FE RE NC E LI BR	AR Y) ,0 6$ A ,1 6$ Al BM	$B ϕP ER AT IN G- SY ST EM /3
	5	*	60$	CP L/. 1$ DL AN	GUAGE*		

卡片结束标志　　(记录结束标志)　　(字母0)

(b)

图11-21　文献数据卡片组

（a）卡片格式；（b）数据卡片组。

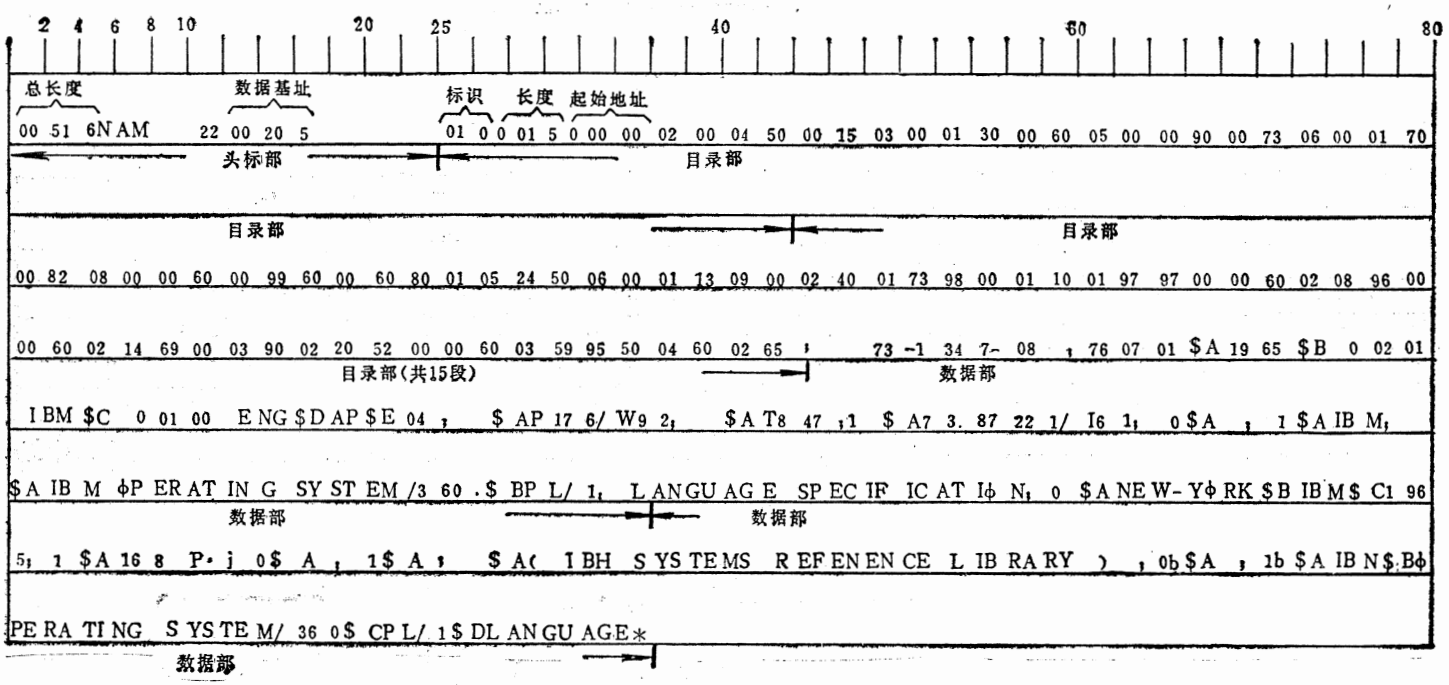


图11-22 顺序文档的一个记录样本

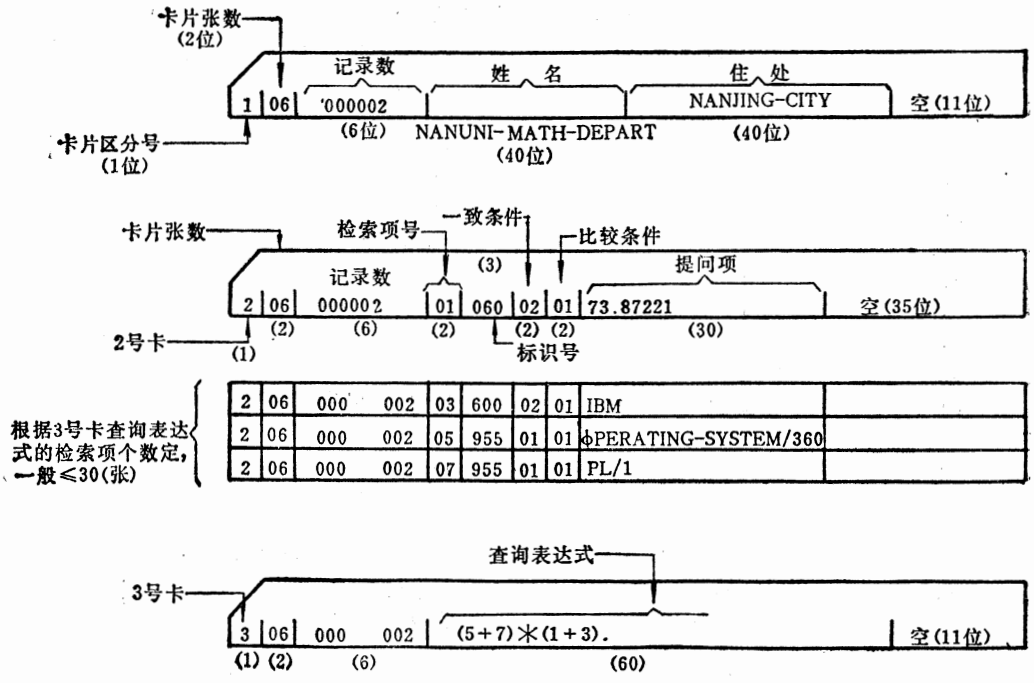


图11-23 用户提问文档数据卡片格式例

表11-4 提问展开表

项号	字段号	成功转向	失败转向	一致条件	比较条件	有效位	检索词
01	15	03	02	01③	01⑤	0020	ΦPERRTING-SYSTEM/360
02	15	03	90②	01	01	0004	PL/1
03	05	80①	04	02④	01	0008	73.87221
04	07	80	90	02	01	0003	IBM

① 90表示查找失败出口。 ② 80表示查找成功出口。 ③ 一致条件中01表示完全一致，即提问项与被检索项内容完全相同，如本例中OPERATING-SYSTEMS/360。 ④ 02表示前方一致，即提问项与被检索项内容前方部分相同，如本例中提问项 7387221，被检项 7387221/丨61；03表示任意一致，即提问项与被检索项内容任意一部分都相同，本例中没有（例如提问项“情报检索”，而被检查索项可以是“情报检索”，“情报检索系统”，“联机情报检索”和“联机情报检索系统”）；04表示后方一致，即提问项与被检索项内容后方部分相同，本例中没有（例如提问项“情报检索”，被检索项“联机情报检索”）。 ⑤ 比较条件中：01表示匹配双方字节（或字）长度相等；02表示长度不相等；03表示被检索项长度大于提问项长度；04表示被检索项长度小于提问项长度。

（三）应用软件

ND76系统的应用软件是采用模块结构的，并由四个独立的COBOL程序包组成，即汉字字模生成程序包、文档库建立程序包、提问文档建立程序包和检索程序包。

表11-7为ND76系统输出格式样本。ND76系统以后陆续有所改进，这里就不作详细介绍了。

二、联机汉字情报检索系统

联机检索系统将以1981年由南京大学研制的试验性质的ZD819系统为例，介绍如下。该系统的特点是：（1）联机操作；（2）系统具有汉字输入输出功能。

（一）系统配置

表11-5　检索标识表

字段号	有效位	被检索的词项
01	0010	73-1347-08
02	0004	1965
02	0003	IBM
02	0003	ENG
02	0002	AP
03	0008	P176/W92
04	0004	T847
05	0012	73.87221/I61
06	0001	
07	0003	IBM
09	0008	NEW-YORK
09	0003	IBM
09	0004	1965
14	0001	
15	0003	IBM
15	0020	ΦPERATING-SYSTEM/360
15	0004	PL/I
15	0008	LANGUAGE
……		

表11-6　表展开查找过程示意

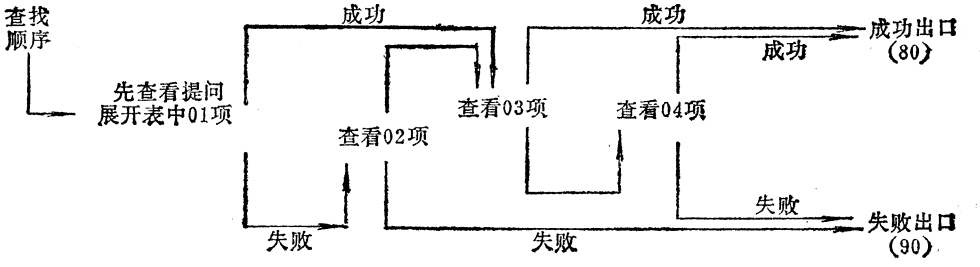


表11-7　ND76系统输出格式样本

NANUNI-MATH-DEPAT　NANJING-CITY
　　BIBIOGRAPHY CARD CONTENTS -ENGLISH-
73-1347-08
　　IBM
　　IBM OPERATING-SYSTEM/360 PL/I, LANGUAGE SPECIFICATION
　　NEW-YORK　　IBM　　1965
　　166P.
　　(IBM SYSTEMS REPENCE LIBRARY)
　　P176/W92
　　T847
　　73.87221/I61
＊＊＊＊＊　　ABSTRACT　　＊＊＊＊＊
KEYWORD:
　　IBM OPERATING-SYSTEM/360 PL/I LANGUAGE

1．硬件　系统硬件包括计算机系统和终端系统两部分。

（1）计算机系统。由带常规外部设备的DJS130机，配上两台中速磁带机、两台5MB磁盘机与一个32×32点阵的汉字字模库（共存4096个汉字，其中有256个汉字可以随机存储），以及一台每秒400个32×32点阵汉字、行宽为54个字的激光汉字印刷机组成。

（2）终端系统。以Z80微处理机为基础，配上一台2片200毫米（8英寸）1M字节软盘机、一台汉字针式打印机（24×24或18×16点阵汉字）、一台75个键的字母数字键盘和一台汉字显示装置（可显示1520个18×16点阵的汉字）等。

本系统配置如图11-24所示。

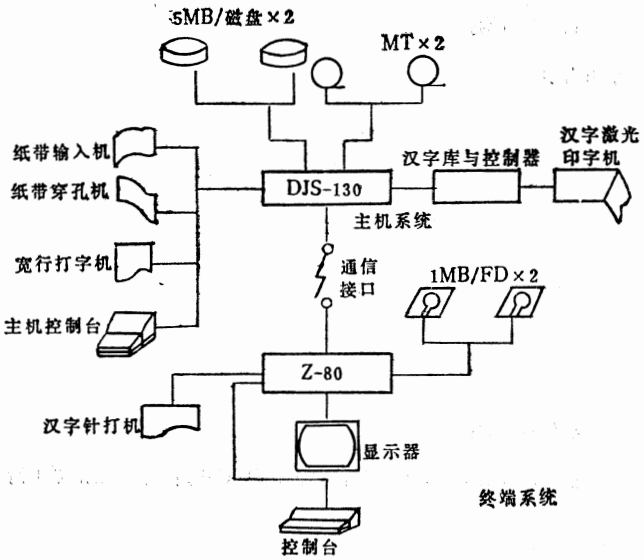


图11-24　ZD819系统配置示意图

2．应用软件　下面按计算机系统和终端系统两部分来介绍。

（1）计算机系统中的应用软件。在RDOS下用汇编语言写成的一个专用系统，包括有会话、管理、建档、分类、查找、编辑、印刷、复制、传递、通信和处理联机、脱机作业等功能的程序模块。

（2）终端系统中的应用软件。在CP/M下用BASIC语言写成的一个专用系统，包括有会话、管理、建档、查找、编辑、印刷、通信、代码转换，以及联机和脱机作业处理等功能的程序模块。

本系统工作方式分脱机与联机两种。脱机时，计算机与终端系统都可以各自处理本身的作业。联机时，终端系统上的联机作业是在主机系统的控制下进行处理的。其工作原理是：对于本系统要收集的文献情报，按传统标引方法加以标引，随后根据通用型文档系统的全可变的记录格式（即字段数和字段长度都不固定）要求，将原文档输入到计算机并存储在磁盘和磁带上。输入时，可用脱机或联机两种方式进行。脱机输入时是用一键九字的汉字整字键盘，双手操作，穿成一个汉字两排八单位孔的纸带，然后通过光电机输入到计算机系统。联机输入时又分计算机主控制台方式和终端方式两种，前者每一个汉字以四位字频码（按汉字综合使用频度高低排序的代码），送入计算机；后者每一汉字按汉字三角号码法编码，用普通字母数字键盘输入，再通过联机通信软接口，经译码（将汉字三角号码转换成字频码）后，按照显示屏幕上每行35个汉字的信息传输到计算机的磁盘和磁带上。当以此方式建立起一次文档后，就可以根据不同的要求，加工成各种二次文档（例如索引文档、分类号索引文档、倒排文档和编辑文档等），以供脱机或联机检索时使用。联机查询语言是直接用汉字在系统提供的标准显示格式表达出来

(见表11-8)。

表11-8 建立查询表达式样本(是通过终端针打机输出的)

```
═══════欢迎您联机检索！═══════

请问您的姓名：郭瑞枫

请告诉您的密码号：NDZD 02

请回答检索类型：书刊 (BJ)？公文 (DT)？名单 (NL)？NL

请开始查询：

中山大学

恕结束查询：

请给出查询结果处理意见：主机打印 (HP)？终端显示 (TD)？终端显示打印 (DP)？DP
```

(二) 查找算法

为了节省存储设备，因此不使用词库分系统，查找算法是用以下方法实现的：

(1) 在汉字终端系统上建立查询表达式 (见表11-8)，它是一种简单的行语言，用时指明查找的文档类型和给出结果处理要求。

(2) 查询表达式经过码转换后，通过通信软接口，将终端查询要求传送到计算机，并由计算机系统中的检索处理功能模块进行加工处理，即按照 (OP1, OP2●, OP3)● 三地址文本形式，通过快速扫描的随机处理方法，进行加工处理，直到遇上查询表达式的结束标志符为止。这里，扫描查询表达式是采用边扫描边处理的方式进行的，随机读取倒排文档时采用了对分搜索法。

(3) 如果结果不满意，可以重做步骤 (1)、(2)，直到满意为止。

小结

本节介绍了两个实例，一个是以脱机批处理为主的情报检索系统ND76，一个是以联机为主的汉字情报检索系统ZD819。它们都是实验性的、无词库的检索系统。这种使用自由词的检索，其效果是不很理想的。要想得到较高的查全率和查准率，除了考虑建造合适的能起信息的存储和检索桥梁作用的词库参照分系统外，检索系统还应该考虑尽量配有多种的检索算法，例如统计型检索法、模糊型检索法、线性相关型检索法等，以及提供有多种的查询语言。

● 表示集合运算符，它可以是集合交、并、差和中心差等；

● 表示操作数1，2，3。

总之，应该建立一个在词库基础上根据类似性理论●，对（图书、情报或数据）资料文档库或数据库进行单重或多重——可变相关询问处理的多语种的汉字联机情报检索系统来。

11.3 机读数据库组织和查询语言

本节将着重介绍常用的三种数据库模型，即分层模型、网状模型和关系模型的数据库。

11.3.1 必要性

随着计算机应用领域的不断扩大，文档系统越来越庞杂。例如，同一所学校里，有教职员工的工资、人事档案、教学管理、设备管理和技术档案、科技成果档案等。这样，势必会有大量的多余的重复的字段或数据项出现，从而导致整个系统在维护和管理上的困难，而且还浪费大量的存储空间。因此人们提出能否设计一种统一的文档系统来支持并产生出上述各类文档，以便实现统一维护、统一管理和节省存储空间、共享资源与数据的目的。其结果就导出了数据库这个概念。

在六十年代后期，数据系统语言委员会（CODASYL）在开发COBOL语言的过程中，产生了数据库管理系统DBMS。接着科德（Codd，E.F）发表了关系数据模型，为以后数据库发展和研究奠定了理论基础。

表11-9 数据库管理系统和文档管理系统的区别

数据库管理系统	文档管理系统	数据库管理系统	文档管理系统
1. 整体性。DBMS的文件当作整体被公用，数据的重复性不复存在。	1. 独立的文件。用户的文件仅供各自的应用程序使用。	6. 存取方式通过DBMS和操作系统（OS）发生联系。即按照“应用程序—DBMS—OS—数据文件”方式。DBMS存储组织是应用程序中数据语言操作的依据，它与OS存取方式对应接口。	6. 存取直接由OS提供，即“应用程序—OS—数据文件”方式。
2. 控制的冗余度。数据库在理论上应是数据项的无冗余集合。但在实际上，许多数据库为缩短访问时间和改善寻址方法还存在某种程度的冗余。然而，这种冗余度是受控制的或是最小的。	2. 自由的冗余度，文件间充满着冗余。	7. 数据独立性。数据独立于使用它们的应用程序。若数据存储结构或全局逻辑结构改变，其应用程序不需随之改变。因为应用程序通过DBMS作用。	7. 数据依赖性。数据存储结构改变，其应用程序也要随之改写。
3. 用户共享数据资料。	3. 独自的用户。	8. 恢复。若处理一大批数据，到中途时遇硬件故障或各类事故发生的突变DBMS将已处理的保存，待恢复运行时，从中断点再继续处理下去。	8.不能自动恢复。
4. 安全性由系统负责。在数据库中，安全性控制比起只存放简单的、无相互关系的文件来更需要且更困难。	4. 保护文件安全的责任在用户。		
5. 完整性是系统校验保证。所谓“完整性”指的是： ① 数据正确性； ② 数据一致性； ③ 各个终端同时删除、插入等并不影响他人对数据的正确使用	5. 完整性靠用户自己维持。		

数据库是在文档基础上发展起来的，表11-9说明了数据库与文档系统之间的一些主

● 指以集合中类似关系为基础的理论。这里，类似关系是指集合中的具有自反性和对称性两个性质。

要差别。

以下叙述三种数据库模型的基本原理。

11.3.2 分层模型和网状模型

一、分层模型

分层数据模型（hierarchial model）是指能用树形结构表示的一个相关联的数据集合。它的特点是记录类型之间只有简单的层次联系，并满足下列两个条件：

（1）在记录类型中只有一个没有首记录的类型；

（2）其它记录类型有且只有一个首记录的类型。

〔例一〕 大学课程数据库（见图11-25）。

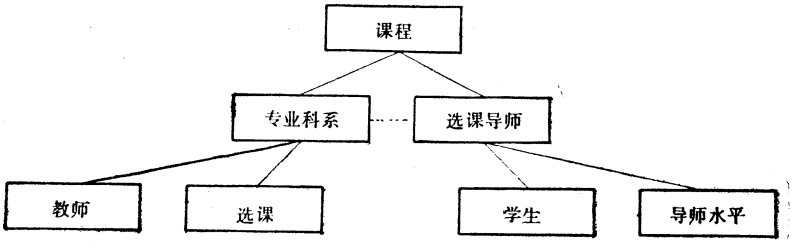


图11-25 大学课程数据库示意图

该数据库可以包括有：某一专业科系开那些选课的记录、学生的记录和教师的记录等。这样，选课记录可以按专业科系又集合成一个科系，学生记录也可以按选课导师集合成另一个科系。如图11-25中某个“专业科系”记录就是“教师”科系和“选课”科系的首记录。而“专业科系”首记录是“课程”记录。但“课程”记录却是一个唯一没有首记录的（这里科系指特性相同的记录集合）。

如果想访问有关某一课程（如数理逻辑学）在某一专业科系（如数学系）中的数据，就要查找属于“课程”记录的那个科系，直到找出“专业科系”记录。接着又要查找属于“专业科系”记录的“选课”记录，从中找到“数学系”这个记录，再接着查找属于这个“数学系”记录的那个科系，直至找到“数理逻辑学”记录为止。我们称这种从首记录找到属记录，再从属记录找出有关记录的查找过程为“导引”（navigation）。

二、网状模型

网状数据库比分层数据库更一般化。但它也是基本层次联系的集合，所不同的是：

（1）可以有一个以上的记录类型没有首记录；

（2）有的记录类型有多于一个的首记录。

图11-26是一个网状模型（network model）的例子。

在网状数据库中，一个节段，如“学生”的上面可以有一个以上的首记录。图中示明节段“学生”的首记录有节段“专业科系”和节段“选课导师”。这样，查找节段“学生”可以通过节段“专业科系”找到，也可以通过节段“选课导师”来找到。

现以网状模型的代表DBTG系统为例，简要介绍如下。

（一）DBTG系统简况

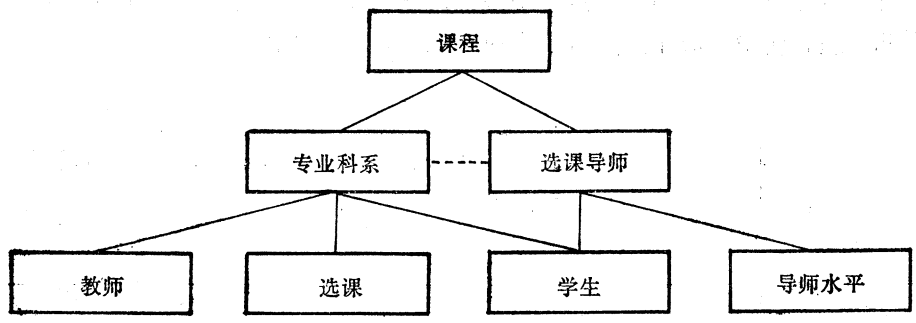


图11-26　一个网状模型数据库的例子

DBTG是CODASYL下属的程序语言委员会的一个数据库任务组。它于1969年提出了关于数据库系统的建议书，1971年四月作了修改，修改后的版本含有模式数据描述语言Schema DDL，子模式数据描述语言SubSchema DDL和数据操作语言 DML（嵌入COBOL语言中）。1971年CODASYL成立了新的数据描述语言委员会，下设数据库语言任务组，并于1973年，1975年又陆续提出了修改报告。

（二）DBTG系统的数据结构

在网状数据模型中，首先引进了“系”（set）的概念，并用它来描述网状的数据结构。这里，系是用来表示在网状结构中的联系类型的。图11-27示明了系概念的表示法。

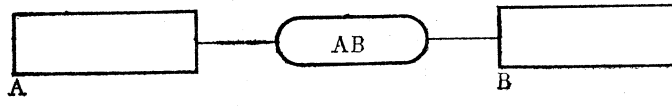


图11-27　系的表示法

这里，图中A为系AB的首记录，B为系AB的属记录。图11-28（a）、（b）、（c）分别示明了1-1联系条件，1－n联系条件，和首记录暂无属记录的表示法。

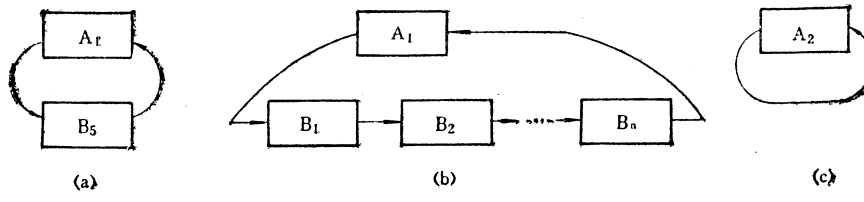


图11-28　1-1或1－n联系条件表示法

这样，可以用系描述实体间的关系。例如：

（1）一个记录类型能为多于一个属记录的首记录；

（2）一个属记录可为别的系类型的首记录。

再引进链记录类型来解决n－m联系条件，如图11-29（a）和（b）所示。

可见，CODASYL语言的基本结构是系，并由系即可以建立起复杂的结构来。正因为引进了系概念，使得网状结构与分层结构在查找方法上互不相同。分层结构查找数据是从（树）根节段上找下来的。而网状结构是通过每个记录类型都能找到所要的数

据，并且能找遍整个网；但每经过一个查找点就可能得出与许多系发生联系的记录类型，因此必须指明取道哪个系，否则会使查找路径发生紊乱。

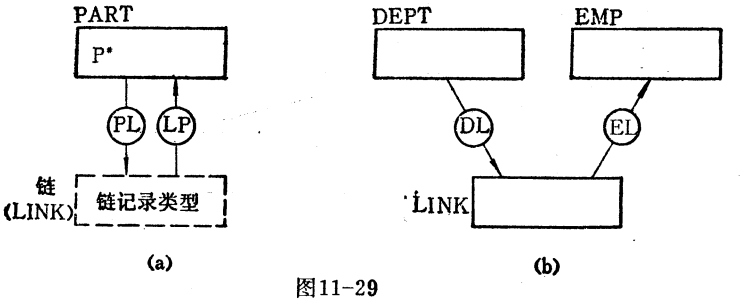


图11-29

（a）链记录类型；（b）n－m联系条件表示法。

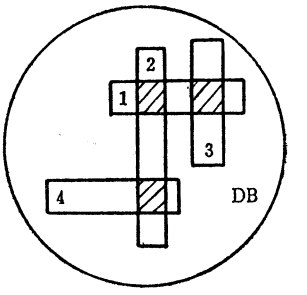


图11-30 共享数据库DB的数据例子

（三）数据库管理系统（DBMS）

设计数据库的目的在于为了能满足各个不同用户的要求，以此达到实现共享数据的目的，正如图11-30中所示，数据库DB中的矩形 i （1，2，3，4）为用户 i 的数据模式。阴影部分是用户共享的数据。

这说明数据库DB的实际结构对每个用户都能满足，但每个用户对数据库DB结构看法是不同的。

现在引进外部模型和概念模型两个术语。例如在一个为企业管理设计的数据库中：

（1）仓库管理员只看每一零件有多少张订货单；

（2）订货采购员只知道每张订货单需多少种零件。

他们各自的信息结构称为外部模型或叫用户视图。将两种结构统一在一个整体结构中，粗略地说，整体结构中既能看到每一种零件与多少张订货单发生联系，又能知道每张订货单订购多少种零件；这种整体的信息结构即概念模型，有时也叫做数据库管理员视图。

图11-31表示了DBMS与用户的关系示意图。

（四）数据描述语言（DDL）

CODASYL把对逻辑数据库描述称为模式，用它来表示存储在一个数据库里的所有数据项类型和记录类型的整体概念。子模式指应用程序员对他所用数据的概念。

用来对模式进行描述的语言称为数据库描述语言或模式语言，记为DDL。DBTG是用DDL来描述数据库的物理和逻辑结构的。用来对子模式进行描述的语言称为子模式语言，记为Sub DDL。

Schema DDL，一般包括有数据逻辑结构，安全性与完整性的规定，存储安排的指定和存取路径等四个方面。相应的由模式名、域（realm）、记录和系四部分组成DDL。

这里，域的概念是指将DBTG的数据库全部存储空间划分为许多部分，而其中每一部分称为域。注意，属于一个域的存储空间在物理上并不完全连接。至于哪个型号设备的哪一部分命名为什么域以及页面的大小、记录密度等，都由物理数据库描述语言

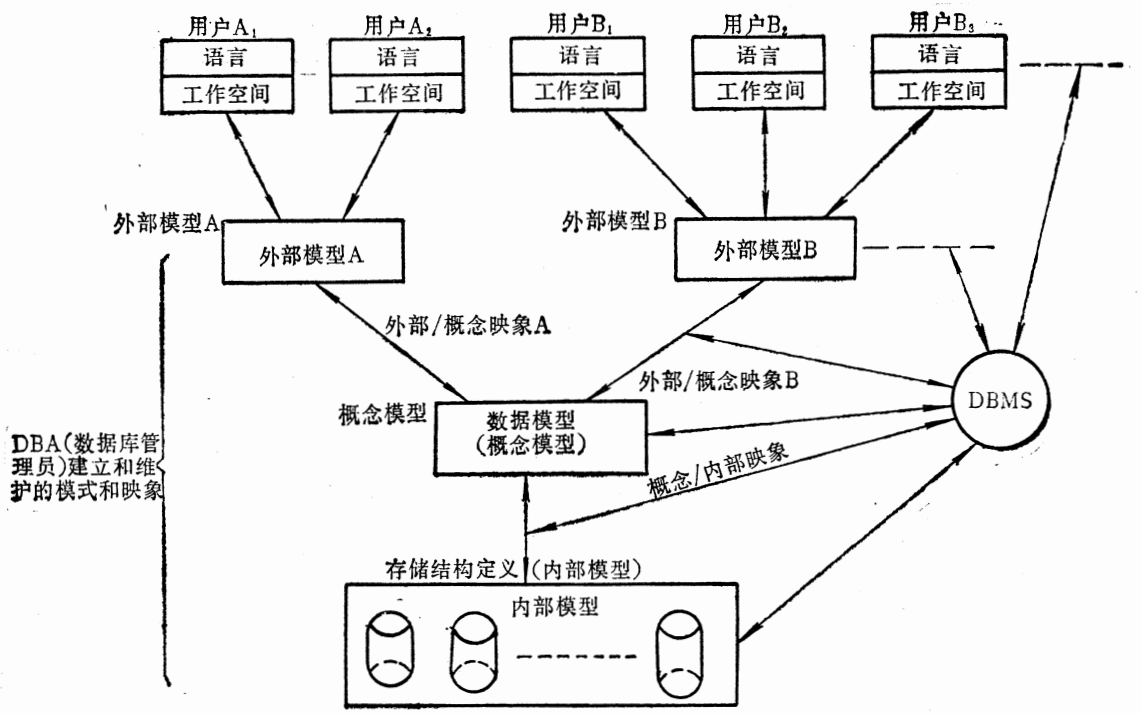


图11-31　DBMS与用户关系图和模式概念

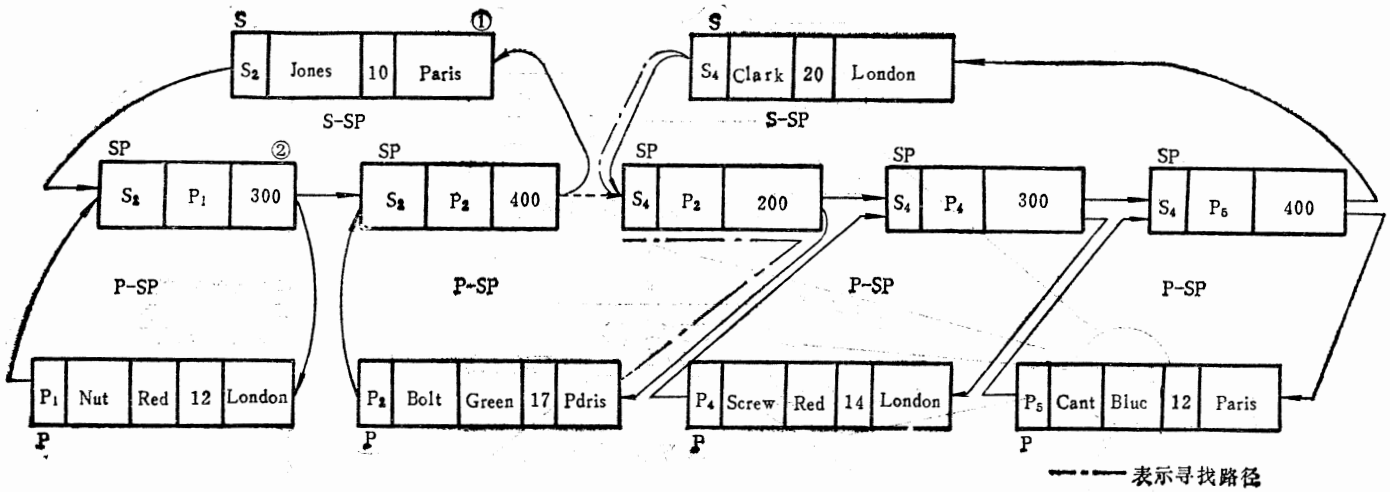


图11-32 供应者和零件的联系条件

① S_2为供应者号；Jones为供应者姓名；10表示状态；Paris为供应者所在城市。

② 记录SP中S_2为供应者号；P_1为零件号； 300为数量。

③ P_1为零件号；Nut为零件名；Red为零件颜色；12为零件重量；London为零件所在城市。

DMCL来定义。

DBTG的DDL书写的模式包括模式条目，一个或多个域条目、记录条目和系条目。其中

（1）模式条目规定了模式名，并对于某些操作规定了密码；

（2）域条目规定了域名，指定打开或关闭时要执行的数据库过程，以及使用域时规定的密码等；

（3）记录条目规定记录名、存放方式、位于什么域中，对记录或数据项执行某些操作时要调用的数据库过程和规定的密码，以及记录中各层数据的名字和字型的描述；

（4）系条目规定系名、首记录名、属记录名、系的顺序、隶属类型（这里隶属指属记录和系的关系，即按入系或出系的情况来划分）和系值选择等。

现以图11-32为例写出一个模式说明（见表11-10）。

表11-10 供应者和零件的模式说明的例子

```
SCHEMA  NAME IS SUPPLIERS-AND-PARTS.
AREA    NAME IS BASIC-DATA-AREA.
AREA    NAME IS LINK-DATA-AREA.
RECORD  NAMA IS S;
    LOCATION MODE IS CALC HASH-SNO USING SNO IN S;
    WITHIN BASIC-DATA-AREA;
    IDENTIFIER IS SNO IN S.
    02 SNO; TYPE IS CHARACTER 5.
    02 SNAME; TYPE IS CHARACTER 20.
    02 STATUS; TYPE IS FIXED DECIMAL 3.
    02 CITY; TYPE IS CHARACTER 15.
RECORD NAME IS P;
    LOCATION MODE IS CALC HASH-PNO USING PNO IN P;
    WITHIN BASTC-DATA-AREA;
    IDENTIFIER IS PNO IN P.
    02 PNO; TYPE IS CHARACTER 6.
    02 PNAME; TYPE IS CHARACTER 20.
    02 COLOR; TYPE IS CHARACTER 6.
    02 WEIGHT; TYPE IS FIXED DECIMAL 4.
    02 CITY; TYPE IS CHARACTER 15.
RECORD NAME IS SP;
    LOCATION MODE IS SYSTEM-DEFAULT;
    WITHIN LINK-DATA-AREA;
    IDENTIFIER IS SNO IN SP, PNO IN SP.
    02 SNO; TYPE IS CHARACTER 5.
    02 PNO; TYPE IS CHARACTER 6.
    02 QTY; TYPE IS FIXED DECIMAL 5.
SET     NAME IS S-SP;
    OWNER IS S;
    ORDER IS PERMANENT SORTED BY DEFINED KEYS.
    MEMBER IS SP;
        INSERTION IS AUTOMATIC.
        RETENTION IS MANDATORY;
        KEY IS ASCENDING PNO IN SP.
            DUPLICATES ARE NOT ALLOWED.
```

```
            NULL IS NOT ALLOWED;
        SET SELECTION IS THRU S-SP OWNER.
            IDENTIFIED BY IDENTIFER SNO IN S.
SET     NAME IS P-SP;
    OWNER IS P;
    ORDER IS PERMANENT SORTED BY DEFINER KEYS.
    MEMBER IS SP;
        INSERTION IS AUTOMATIC.
        RETENTION IS MANDATORY;
        KEY IS ASCENDING SNO IN SP.
            DUPLIGATES ARE NOT ALLOWED.
            NULL IS NOT ALLOWED;
        SET SELECTION IS THRU P-SP OWNER.
            IDENTIFIED BY IDENTIFIER PNO IN P.
```

表中，（1）模式条目指出其名为供应者和零件。

（2）域条目分别定义基本数据域和链数据域。前者存放S.P（S指Suppliers，P指Parts），后者存放SP。

（3）记录条目，以供应者S为例，语言说明对记录的存放是用供应者号码SNO、通过杂凑法计算确定；记录S的标识符是SNO；组成S的诸字段类型为S的子条目，所以前面冠以“0 2”；象SNO由5个字母组成，…，STATUS由3位固定长的十进数组成。

（4）系条目中指出系S-SP，其中，首记录是S，属记录是SP；SP插入是系统自动安排的，保留是人为的。SP的链按PNO由小到大排列，SP中重复和空缺是不允许的。系的选择通过S-SP的首记录，首记录由SNO标识。系P-SP的说明也相似。

（五）子模式数据描述语言（SUBDDL）

在DBTG报告中有Subschema DDL（COBOL），子模式以宿主语言（COBOL）书写。子模式是相应模式的子集。由一个模式可以推导出许多不同的子模式，子模式之间可以互相覆盖。子模式是面向于一个或多个应用程序需要的一部分数据的一种图式，也是一种程序员的文档组织。

子模式与模式的区别有如下几点。

（1）子模式为模式的子集，不论是域、记录、系和数据项都可以只取一部分；

（2）域、系、记录或数据项可以用别名；

（3）数据项的数据类型可以不同，而且可以选择一组数据项给以不同的组名；

（4）记录内数据项的次序可以更动；

（5）系内某些属记录类型可以不包括在子模式中；

（6）可以另行规定保密锁，这些保密锁优先于模式。

表11-11是表11.10例子中的一个子模式描述。该子模式名为SUPPLIES，并且指明出自于模式SUPPLIERS-AND-PARTS中。系S-SP的名字被改名为S-SS。

（六）数据操纵语言

DBTG为应用程序和DBMS之间的接口配置了一种语言，即数据操纵语言，记为DML。它被嵌入于一个COBOL之类的宿主语言之中。

当DML请求一个记录时，这个记录便被传送到应用程序的工作区中，并由该应用

程序对它进行操作。

表11-11　子模式说明的例子

```
1.  TITLE DIVISION.
2.  SS SUPPLIES WITHIN SUPPLIERS-AND-PARTS.
3.  MAPPING DIVISION.
4.  ALIAS SECTION.
5.  AD S-SP SET-NAME BECOMES S-SS.
6.  STRUCTURE DIVISION.
7.  REALM SECTION.
    RD BASIC-DATA-AREA.
    RD LINK-DATA-AREA.
8.  SET SECTION.
    SD S-SS.
9.  RECORD SECTION.
10. 01 S.
11. 02 SNO; PIC X(5).
12. 02 SNAME; PIC X(20).
13. 02 STATUS; PIC 9(4).
14. 02 CITY; PIC A(20).
15. 01 SP.
16. 02 SNO; PIC X(5).
17. 02 PNO; PIC X(6).
18. 02 QTY; PIC 9(5).
```

图 11-33 是使用 DBMS 过程说明一个包括 DML 命令的应用程序具体处理过程的例子。

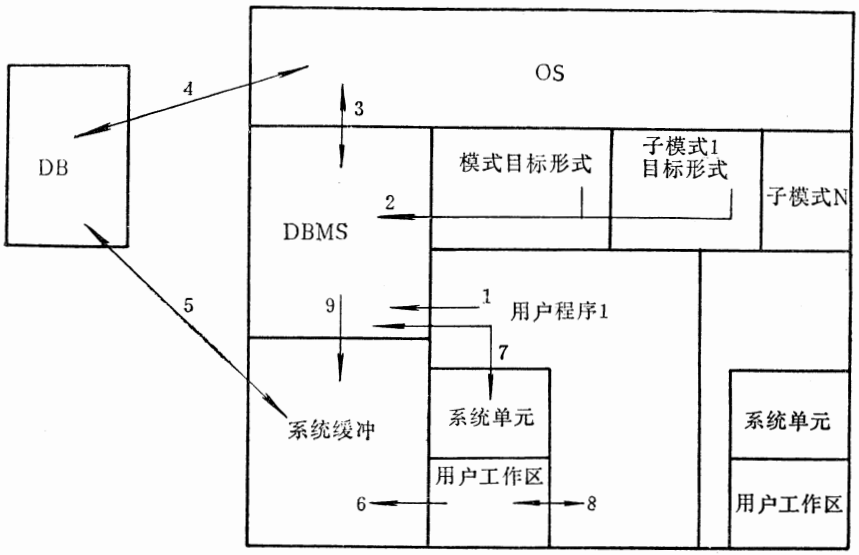


图11-33　DBMS过程说明的例子

图中，（1）用户程序 1 （包含 DML）向 DBMS 要求存取数据库中的数据；

（2）DBMS 根据模式，子模式 1 的描述决定存取路径；

（3）DBMS 的存取通过 OS 对应的存取方式进行；

（4）OS 向 DB 发命令；

（5）将 DB 中一组数据（其中有程序所需要的数据）取到主存内的缓冲区；

（6）数据从缓冲区进入用户工作区；

（7）系统单元里置状态，说明所要数据取到与否，并回复 DBMS；

（8）工作区中取到的数据与处理程序相互作用；

（9）若重新取一个数据，即步骤 1 执行后，做步骤 9：DBMS 先查缓冲区，上次取进一组数据内有否所要的新数据；如果有，则做步骤（6），（7），（8）；如果没有，则做步骤（3），（4），（5），（6），（7），（8）。

DBTG 设计的 DML(COBOL)中典型命令有：Find，Get，Modify，Insert/Connect，Remove/Disconnect，Delete/Erase 和 Store 等。

（七）物理数据库描述语言（DMCL）

DBTG 设计的 DMCL 是物理数据库描述语言，它是按照物理数据布局安排进行的。因有标准可循，这里就不介绍了。

11.3.3 关系模型、查询语言及其实例

一、关系模型

关系方法是应用数学理论处理数据库系统的方法。应用数学方法往往能使看来非常错综复杂的问题，变得非常简洁精炼。基于这种指导思想，1970年提出了一般化的关系数据模型。其后陆续发表了多篇关于引进规范化概念的论文。这一模型将数据的逻辑结构归结为满足一定条件的二维表的形式，同时还使用了关系代数和关系演算作为数据操纵语言，借以对用关系模型（relational model）建造的数据库进行数据存取。

例如，图 11-34 表示了一个用关系模型来表示的大学开设课程的数据库。其数据是按关系存放的，关系中的每个项目称为元组（tuple）或记录。为了找到数理逻辑学课程的教师姓名和教室，就按“选课——教师”这个关系查找到“数理逻辑学”这个记录，从中就能得知教师的名字叫“Wang Hao”。然后查找“选课——教室”关系，找到“数理逻辑学”这个记录，从中得知教室为“Mathematics 2”。这样，从数据库得到的答案是“Wang Hao，Mathematics 2”，这都是通过“数理逻辑学”这个记录值找到的。

选课-教师关系

选　课	教　师
微积分学	……
物理学	……
近世代数学	……
函数论	……
数理逻辑学	WANG HAO
…	…

其他关系

选课教室关系

选　课	教　室
函数论	……
近世代数学	……
微积分学	……
物理学	……
数理逻辑学	Mathematics 2
…	…

图11-34　关系数据库的例子

关系是个二维表，如何将这样的二维表以及对于表中数据的操作按严格的数据概念

定义，是研究关系数据库的基础。

（一）说明性定义

将一组数据排成一个 m 行、n 列的二维表，称为具有 m 个 n 元组的关系。每一行是一个 n 元组，相当于一个记录值，用以描述一个个体。每一列叫做域，相当于数据项类型，用以表示属性。域是命名的。

（二）数据描述语言（DDL）

在关系模型中，DDL 所描述的模型也叫做模式。DDL 是用来描述外部模型和概念模型中的关系的。表 11-12 是三种数据模型间的语义对照表。

表11-12　三种数据模型间的语义对照表

现实世界	分层模型	网状模型	关系模型
异质客体集合	逻辑数据库	数据库	表
同质客体集合	逻辑数据库记录（文档）	系（set）	子表（table）
客体	节段（Segment）	记录（值）（record）	行(tuple)
属性	字段（field）	数据项(data item)	列(row)

DDL 对于 1-1 联系条件，1－n 联系条件和 n－m 联系条件的描述是按以下法则规定的：

1．1-1 联系条件　例如选课与教师是 1-1 映照，则可用下面两张表来表示：

课程

C^*
C_1
C_2
⋮

教师

T^*	C^*
T_1	C_1
T_2	C_2
⋮	

2．1－n 联系条件　例如同一选课有多少教师开出，则可表示为：

课程

C^*
C_1
C_2
⋮

教师

T^*	C^*
T_1	C_1
T_2	C_1
T_3	C_3
T_4	C_2

3．n－m 联系　例如有多少选课有多少教师开出，则可表示为：

课程

C^*
C_1
C_3
⋮

教师

T^*
T_1
T_2
T_3
⋮

CT

C^*	T^*
C_1	T_1
C_2	T_1
C_3	T_2
	⋮

二、查询语言及例子

（一）数据操纵语言（DML）

关系模型中的 DML 有三种。现举例说明如下：

1．ALPHA 语言（简写 α 语言）　例如：

COURSE（课程）

（课程号） C^*	（课程名） CNAME
C_1	N_1
C_2	N_2
C_3	N_3
	⋮

TEACHER（教师）

（教师号） T^*	（教师名） TNAME⋯	（薪金） SALARY
T_1		61
T_2		77
T_4		77
T_{10}	⋮	61

CT

C^*	T^*
C_1	T_1
C_1	T_2
C_2	T_1
C_3	T_4
⋮	

ALPHA 语言是以谓词演算为基础的数据语言，系 IBM 公司的科德提出的。以

ALPHA 语言表示的数据操作的例子如下：

```
RANGE TEACHER TX.
RANGE COURSE CT CTX.
GET  W(COURSE, NAME): ∃TX∃CTX(CTX·C* = COURSE·C*∧TX·T*
                              = CTX·T*∧TX·SALARY = '61')。
```

这里，TX、CTX 为区域变量，前两句表示 TX，CTX 在 TEACHER，CT 关系各元组的区域内取值。第三句，GET 意味着从数据库找出数据置于用户工作区W中。W后括号内的关系名和属性名即为查询的目标，冒号“：”下面为限定条件。符号“∃”表示存在量词。因 TEACHER 中某些元组 TX 的薪金就是61元，CT 中某些元组 CTX 的教师编号就是相应于61元薪金的教师编号 TX，有61元薪金的教师的课程编号是 CTX 的课程编号，于是在 COURSE 中课程编号与 CTX 的课程编号相同的这些元组的COURSE 名——COURSE.NAME 就是查询所需的课程名。

由于α语言比较复杂，推广使用有一定困难，因此创造了较简单的结构式英语查询语言 SEQUEL2，它是按照科德所述的第一范式关系进行操作的。

2．SEQUEL2语言　这种语言接近口语化，是面向非程序员用户的，而查询能力较强。IBM 公司研制的 SYSTEMR 关系数据库就是用 SEQUEL2 进行操作的。SEQUEL 2 是 SEQUEL 1 上发展起来的。它属于非过程性语言。

例如上例，用 SEQUEL 2 写出为：

```
SELECT CNAME.
FROM COURSE.
WHERE C* IN
        SELECT C*.
        FROM CT.
        WHERE T* IN
              SELECT T*.
              FROM   TEACHER.
              WHERE SALARY = '61'.
```

这里的意思是从（FROM）课程 COURSE 中选出（SELECT）所需要的 CNAME(课程名)，它对应那里（WHERE）的 C#（课程编号)。其条件是该 C# 在下一级中。接着从 CTC 课程——教师关系）中选取所需的 C#。这里 C# 又对应于其中的 T#（教师编号)。而 T# 又在下一级中，从 TEACHER 选取所需的 T#。该 T# 对应的是满足条件薪金为61元的元组。

SEQUEL 2 语言从查询目标一级级往下对应映照。但它的操作步骤和检索数据的路径却正好相反。

3．关系代数语言　关系代数语言是对关系代数运算进行描述的语言。关系代数是对于整个关系进行运算的总体。关系是个以元组为元素的集合。对于关系可施行并、交、差等一般集合运算。此外，还可以施行关系所特有的投影、连接和除法等运算，其结果也是个关系。关系代数语言所需要的运算符有：SELECT(选取)、JOIN(连接)PROJECT（投影）等。例如上例可写作：

```
PROJECT (JOIN (JOIN((SELECT TEACHER WHERE SALARY = '61')
   AND CT OVER T*) AND COURSE OVER C*)) OVER CNAME
```

（二）规范化问题

所谓规范化是指用更单纯，结构更规则的关系取代原有关系的过程，其目的在于解决冗余度和二义性问题。

规范是逐步实现的。

1．冗余度问题　例如有关系表：

课程

C^*	课程名	…T^*	教师名	薪金
C_1	数学	T_1		
C_1	数学	T_2		
C_1	数学	T_3		

这里，由于将课程和教师的数据列为一张表，加上课程和教师间的联系是1－n，所以表中有教师T，课程名的大量重复问题。为此，分成两张关系表：

C^*	课程名…
C_1	数学
	⋮

T^*	教师名	薪金	C^*
T_1			C_1
T_2			C_2
T_3			C_3

就能达到减少数据冗余度的目的。

2．二义性问题　例如：

课程——教室

C^*	教室*	QTY
		7

这里，QTY的含义不清楚。它倒底是说明课程——教室间关系（某课程需要占用几个教室）还是仅仅说明现有教室有多少个？而这种表是决定不了的，就是存在二义性。

若QTY只说明教室数量，那单列下表也许更合适些：

教室

教室*	QTY
	7

为了避免二义性，关系什么时候分，什么时候合，要有一些原则可循，所以才提出来要进行关系规范化的问题。

在讨论规范化前，先引进两个函数相关概念。

（1）函数相关。设R是一个关系，X和Y是R中的两个属性。如果R中X的任何一个值，仅有一个Y的值与之对应，则称R的属性Y函数相关于属性X。这里，X可以是由几个属性组成的复合属性。

例如，COURSE(C#, CNAME, SPECIALITY, …)的关系中，属性CNAME, SPECIALITY, …，都函数相关于属性C*。箭头表示任何一个C*仅能找到一个CNAME（或SPECIALITY）与之对应。

（2）完全函数相关

如果属性Y函数相关于复合属性X，而且不与X的任何子集函数相关，则称属性Y完全函数相关于复合属性X。

例如，CT((C*, T*), GRADUATE-STUDENT, …)中，GRADUATE-STUDENT, …，都函数相关于复合属性（C*, T*），且它不与X的任何子集函数相关，

所以 GRADUATE-STUDENT 就是完全函数相关于（C#，T#）。

3．规范关系　规范关系是分级的。由于人们对属性联系的不同看法，形成了如图11-35所示的一级级规范关系。近来，规范关系越分越细，已经提出了第四范式(4NF)，第五范式（5NF）……

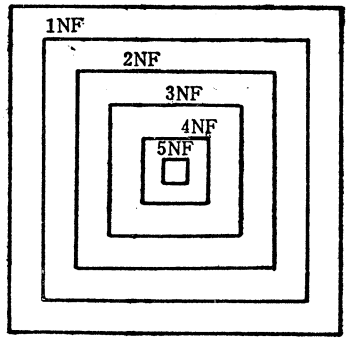


图11-35　范式分级图

一个关系是属于哪一种规范关系，有下述两个判断条件可循：

（1）非主属性● 是相互独立的，即任何非主属性间不存在函数相关。

（2）非主属性是完全函数相关于主键。

注意，必须满足条件（1）和（2）的为3NF，只满足（2）不满足（1）的为2NF，（1）和（2）都不满足的称为1NF。

例如，在2NF例中，TEACHER（T#，TNAME，STATUS，SAL）对数据操作处理妨碍很大。也就是说，例中

T^*	TNAME	STATUS	SAL
T_1	WANG	1	103
T_2	XU	3	77
T_3	SHAN	2	85
T_4	GUO	3	77
⋮	⋮		
T_9	GAO	4	68

（1）它不便于插入操作。例如在此关系中，要加入一个第5等级（STATUS）工资（SAL）是61元的，因无对应的教师编号，教师姓名，所以不能实现。

（2）它不便于删去操作。如果把下一行都删去，则STATUS SAL，有关第一等级的工资是103元的资料也一并丢失了。

（3）它不便于修改。如果提高工资，第2等级的加成96元，则凡是登记过的第2等级的85元的地方处处都要改为96元。

为此，如果我们将2NF化成3NF：

TEACHER(T^* TNAME STATUS)

T_1	WANG	1
T_2	XU	3
T_3	SHAN	2
T_4	GUO	3
⋮	⋮	⋮
T_9	GAO	4

SS(STATUS SAL)

1	103
2	85
3	77
4	68

则上述插入、删去、修改时的问题都解决了。

同理，对1NF，先规范成2NF，再规范成3NF，…。一般地说，这种逐步规范化用

● 这里，主键是指这样的键，即设Key为关系R中的一个属性组合。若R完全函数相关于Key，则称Key为关系R的键。一个关系R可能不止有一个键，我们把这些键称为候选键，可以选定其中之一作为关系R的主键，例如若例中教师编号和姓名都不同的话，则T^*，TNAME为候选键，并可选T^*为主键。主属性是指包括在候选键之中的属性，而把不包括候选键的属性，称为非主属性，如例中SAL。

两种方法进行。一种是将关系告诉系统，由系统自动分开，但这样做系统开销代价太高；另一种方法是用户自己将关系规范化后交给系统。这样做较妥当，用户经过仔细分析，关系清晰，有条不紊，便于实现。

限于篇幅，本节未能就数据独立性，数据保密性、数据安全性、数据通信和数据字典以及数据目录（*DD/DC*）等作一一介绍。注意，随着数据通信技术和计算机网络的发展，出现了分布式的数据处理系统，相应的需要有分布式数据库。因此又有（上述三种数据模型的）集中式数据库系统和（分段式或复制式的）分布式数据库系统之分。

本节所介绍的机读数据库组织和查询语言是情报检索系统随着计算机科学和通信技术的发展，对情报的存储结构和检索策略的不断提高所产生和演变的必然结果。

11.4 汉字和西文情报检索系统的异同点

汉字和西文情报检索系统同属于信息管理系统，原理上，两者并没有本质上的区别。仅在处理内容和实现的技术上，汉字情报检索有它的特点。内容上主要反映在题材选择、内容分析、标引方法和词库组织等几个方面，而技术上则表现在硬设备和软设备两个方面。

11.4.1 汉字情报的标引和表示

科学技术的发展，特别是计算机科学技术的发展，有助于发掘、整理、继承和发扬祖国文化遗产，实现象中国历代古籍文献资料和现代科技、文史资料等的自动化处理。因此汉字情报系统将有个选题取材（包括资料网罗范围和专业性）、分类编目、标引作业和词库确定以及在计算机内部如何表达文献内容等问题。这都将涉及到标引作业如何进行、索引语言如何确定和汉字关键词如何抽取等具体问题。

这里，标引作业又将涉及到内容分析和标引词的确定。通常作法是了解文献所讨论的主题内容，形成概念，并用标引词来代表这些内容。例如假设有一篇文献，它只论述六个课题（*A*，*B*，*C*，*D*，*E*，*F*），那么在标引作业中能被确认的有多少个，并用标引词来表示它能达到什么程度，这就是常说的标引网罗范围或综合性和专业性概括，或称为标引深度。本例中如果在内容分析阶段确认所有这六个课题，并用最合适的标引词来表示这六个课题，那么可以说，在标引这篇文献时达到了完全的高度综合性。显然，如果这六个课题都被标引了，那么无论怎样对这些课题的组合进行提问检索，这篇文献都能被查找出来。因此标引的高度综合性，有利于保证查全率，同时也倾向于低查准率。反之，标引的低度综合性导致低查全率和高查准率。

对于汉字文献情报资料的标引工作，从语言上看，中文白话文要比西文麻烦一些，古文的要比白话文的又困难一点。这主要表现在概念理解和标引词选定以及标引深度方面。

索引语言是在存储和检索原文文献时，为了确切掌握它们的意义，从参照系统中认识出来而抽取出最合适的元素（如关键词、主题词、叙词或特性标识）来表达的语言系统（也称文献语言、情报语言）。索引语言中的元素称为索引词。如果说标引词是标引人员标引文献时用的语言工具，那么索引词便是检索人员检索文献时用的语言工具。索引语言有时也叫做检索语言。标引语言（标引词的词汇表）、索引语言都是文献工作语言。文献语言按其内部组织方式分为两类：层次式（或分类式）的结构和组配式(或词汇式)

的结构。

关于汉字文献情报资料使用哪一种索引语言为好，是有待于解决和研究的一个课题。

对汉字文献进行自动抽词明显地要比西文的困难一些。从形式上看汉字没有词间隔标志，而西文一般以空格表示词间隔。从语法和语义关系看，汉字比西文要复杂。

总之，从题材选择、标引过程和索引编制到文献传播的整个加工流程中，说明汉字文献资料的表示方法与西文的不同，并在某些环节上汉字的要比西文的困难。

11.4.2 自动标引

自动标引原是西文情报检索中尚未解决的课题之一，对于汉字情报检索系统更是如此。至少在目前，我国还没有创建出自己的一套适合于用来进行自动标引的好方法。加上汉语，主要是指其语法和语义两方面所带来的复杂性和麻烦，使它的实现比西文要困难。同样，自动分类、自动文摘和自动翻译等的情况也是如此。

11.4.3 词库

从汉语词汇学、句法学和语义学等几方面看，汉字词库要比西文词库复杂和严谨些。这不是说西文词库就一定容易组建。只是要说明一下，汉语历史悠久，演变沿流不断和历代保存下来的大量的丰富的文献宝库给汉字词库自动组建增加了内容的复杂性和困难，以及我们在组建汉字词库方面还缺乏经验所产生的影响。尽管已出版的综合主题词表给我们提供了一个可参考的工具，但遗憾的是它并不很适合于用来进行机器自动检索。因此建立适合于机器检索用的汉字词库也是当前急待解决的一个课题。

11.4.4 汉字情报检索系统对汉字输入输出技术的要求

为了能处理汉字信息，在硬件设备上需要有汉字输入输出和汉字字模库，联机检索的情况要配用接插兼容的汉字终端设备。理想的情况要求主机的操作系统和程序设计语言能兼容汉字和西文数据处理能力，使它们能具有直接处理汉字信息的功能（详见本书第九、十章）。情报检索系统或数据库系统进行信息处理时的特点，同一般的数据处理系统没有本质上的区别。它们都要求在源程序中用汉字表示直接量和汉字注释，在记录或文档中也可以用汉字西文混用的字符串表示。软件能区别开汉字和西文代码，并在输出时作分别处理。因此，对于数据处理适用的汉字、西文兼容技术，也同样适用于这类系统。

第十二章　汉字联机通信网络

随着电子计算机的广泛应用和汉字信息处理技术的发展，计算机与通信技术的结合将势在必行。这就是说，建立汉字联机通信网络，已成为通向信息化社会的必经之路。

由于过去专业分工的限制，计算机专业工作者对通信技术并不熟悉。因此，这里有必要先对通信技术作简要介绍。

用特定的方法通过某种媒体或传输线将信息从一地传送到另一地的过程，这就是通信。通信可分为模拟通信和数据通信两大类。随着现代科学技术的发展，数据通信越来越成为通信技术的主流。

鉴于汉字数据通信技术与一般的数据通信技术并无本质上的区别，因此有必要首先介绍一下数据通信技术。

12.1　数据通信技术

所谓数据通信是，利用通信系统对二进制编码的字母、数字、符号以及数字化的声音、图象信息所进行的传输、交换和处理。通常是以计算机为中心，通过线路和远程终端直接连接起来形成联机系统，远程终端所产生的数据能及时地传送到中央处理机进行处理，而处理后的结果又能马上返送给远程终端，除数据源和交换装置外，整个过程不需人工干预。随着计算技术的发展，交换过程也可能完全自动化。

数据通信技术所涉及的内容很多，其中包括通信信道、调制与解调、传输方式、通信接口、数据链路的建立、信息代码、通信协议（或规程）等等。

实现数据通信过程一般至少需要包括以下五个环节：发送器或信息源；报文；二进制串行接口；通信信道（或链路）以及接收器。如图 12-1 所示。

为使二进制串行数据进入通信信道，必须满足一定的通信接口要求。

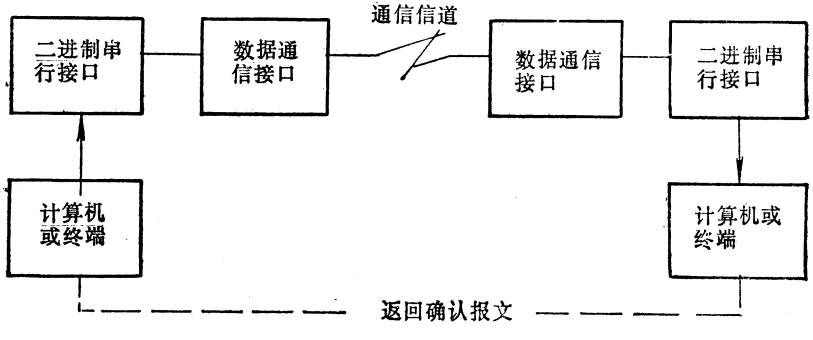


图12-1　数据通信过程

12.1.1　信道类型

信道（或通信链路）定义为双站或多站（终端）之间传输信息的路径。信道有三种基本类型：单工、半双工与双工信道。图 12-2 示出了用于A与B两点间信息传输的信

道类型。

单工信道是一种单向传输信道，例如它只能从A点向B点传输，而不能从B点向A点传输。半双工信道可实现不同时的双向传输，例如先从A点向B点传输，然后从B点向A点传输。如果采用两线制线路，必须将线路反转才能改变传输方向。如果采用四线制线路就不需要进行线路反转。双工信道能实现同时双向通信制，例如从A点向B点及同时从B点向A点进行传输。如果采用频率分隔技术，将信道分成接收信道与发送信道，则用双线线路也可进行双工通信。

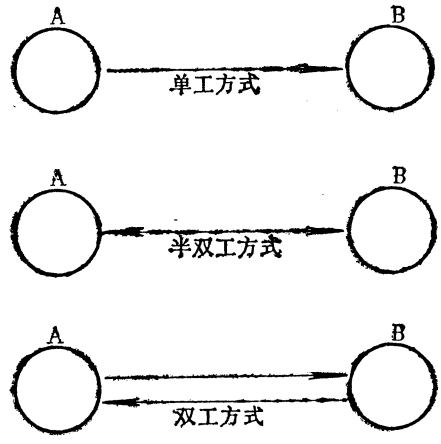


图12-2 信道类型

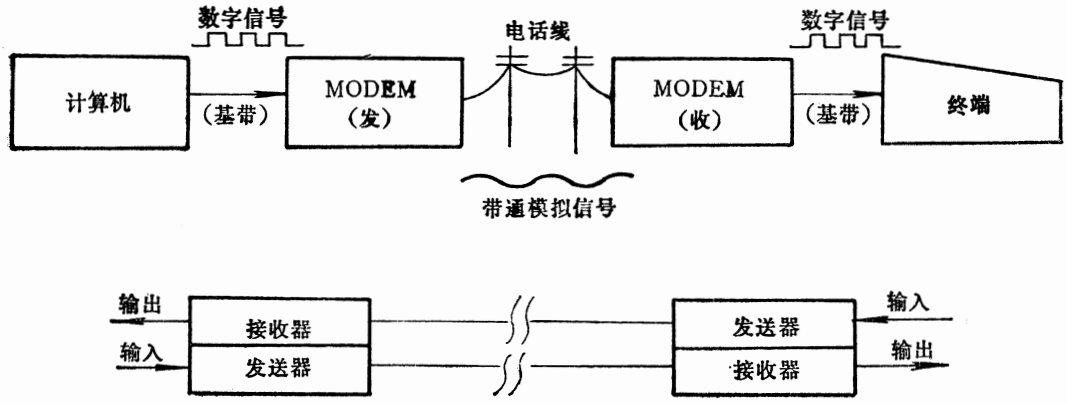
除了传输方向外，信道的另一特征是传输速度。传输速度通常是以每秒内信号码元数（或称波特速率）来测量的。如果信号码元是用二进制来表示，那么波特率就等于比特率。如果用两个以上的状态来表示，比特率就大于波特率。

12.1.2 调制解调器

最初，计算机之间的数据传递，是通过纸带或磁带等中间媒体来进行的。当计算机之间的距离较远时，便需要借助通信信道或电话线路来实现数据通信。这样，为使计算机的数据信息在载波线路上传输，就需要采用调制解调器（MODEM)。调制解调器是将来自计算机或终端的数字信号变换成通信信道所需要的调制载波信号而使用的一种设备。本质上，它是一个复杂的A/D和D/A变换器。如图 12-3(a) 所示。MODEM系统是由发送器、接收器和滤波器等组成（图 12-3 b)。发送器用来将来自计算机的数字（基带）信号变换成适合于电话线传输的模拟（带通）信号，而接收器用来检测出这些带通信号，并对它们加以解调，以恢复成原先的基带数字信号。

电话线的带宽为 300～3300 赫。工作在电话线上的 MODEM 称为话音级 MODEM。各种应用对 MODEM 提出了不同的要求。根据不同的数据率和数据量来选择传输调制技术（频率键控、幅度调制及相位调制等），并决定是使用公用电话线还是使用租用线路。制定 MODEM 的规格要求时，应在通信线路和硬件造价之间作出选择。例如，可使用造价低的 MODEM，以 300位/秒全天传送数据；也可使用造价高而性能好的 MODEM，以 9600位/秒的高速率短时间应用，以减少线路租用费。

在低速 MODEM 中，比特率与波特率相同。用户一般使用 50 波特或 110 波特的 MODEM。这些 MODEM 中的比特率也是 50 或 110位/秒。然而，对于 2400、4800 或 9600位/秒 的 MODEM，实际传输的波特数比上述数值要少。原因是为了获得较高的速率而使用了较复杂的调制技术。此技术是对每个波特使用较多的比特数。现代高速 MODEM 对每个波特可要求使用 2、3 或 4 个比特，以便实现比特率为 2400、4800 和 9600位/秒 的通信。



12-3(a) 利用MODEM传输模拟信号

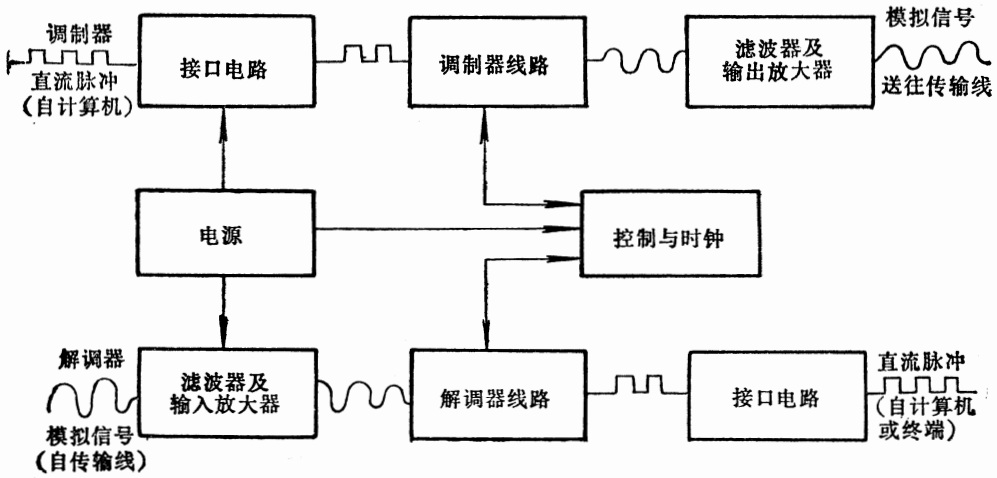


图12-3(b) 调制解调器的内部结构

相对于给定的信道带宽的信噪比，可向信道发送的信息量有个理论极限。例如，电话线的最大有效数据传输率为10000位/秒，这就是多数商用MODEM不超过9600位/秒的原因。

功能上MODEM分为两个逻辑部分（见图12-3b），即调制器和解调器。调制器接收来自计算机或远程终端的数字输入信号，并把直流方波脉冲转换成模拟音频信号，然后在通信链路上进行传送。在链路的另一端，第二个MODEM的解调器将模拟信号再转换成数字信号后输出。

短程MODEM是在较短的距离上用实心导线，在无带宽限制，无负荷的线路上工作（一般小于十公里）。实际上它们并不真正是MODEM。更确切地说，它们是一种发送与接收数字数据的线路驱动器和线路接收器。

话音级调制解调器有两大类：一类是租用线路上以2000工作的异步设备；另一类是在拨号线路上以4800位/秒最高数据率和在租用线路上以9600位/秒数据率工作的同步设备。表12-1列出了现有的MODEM分类。

表12-1　现有MODEM的分类

MODEM类型	通信信道	数据率（位/秒）	应　用
短　程	19.2K,1M	私人线路 租用线路	有限距离（6）
宽　带			
半　群	8803	19.2K	大容量电话线双工计算机-计算机链
群❶	8801	40.8K，50K	
超群❷	5700	230.4K	
话音级			
高速同步	租用线路	4.8K,7.2K,9.6K	计算机-终端，数据采集，过程控制
中速同步	租用线路 拨号	2K,2.4K,3.6K 4.8K	
低速异步	租用线路 拨号	1.2K 1.8K	

❶ 群：由12个话路合成一个基本群。

❷ 超群：由5个基本群合成一个超群。

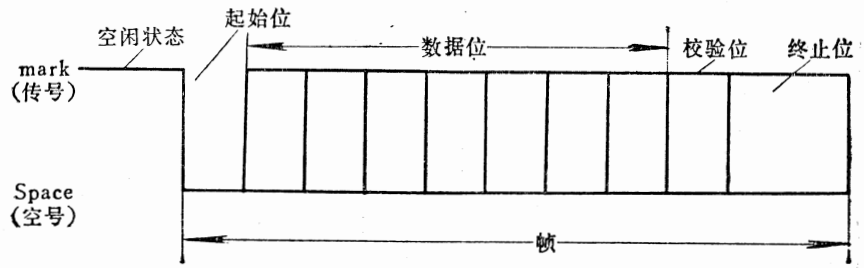


图12-4(a)　异步传输信号

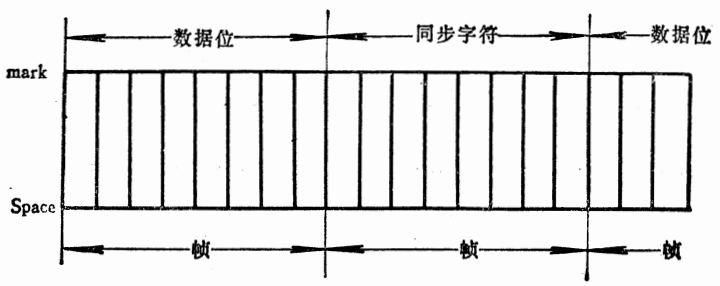


图12-4(b)　同步传输信号

12.1.3　同步与异步传输

数据率小于1200位/秒的低速终端使用异步传输。异步系统的传输格式示于

图12-4(a)，在其无效状态时处于传号状态(二进制1)。当发送每个字符时，最先是起始位或由传号转成空号(二进制0)，以便向接收终端指出字符已经发送。接收端检测出起始位以及数据位。在字符传输末尾，线路又返回到传号状态，并为下一个字符的起始作好准备。异步字符的长度随使用的信息代码而变化，有5位的Baudot代码、7位的GB 1988代码和8位的EBCDIC代码（汉字为双字节代码)。在传输过程中，从一个字符到另一个字符重复上述过程，直到整个报文发送完毕。使用起始位和终止位的目的是使接收终端与发送器同步。

同步系统的传输格式如图12-4(b)所示。为使接收器与发送器同步，在MODEM中使用了内部时钟。一旦接收终端收到同步字符（SYN)，数据传输便逐个字符地进行下去，无须插入起始位和停止位。输入的数据根据MODEM提供的接收时钟进行位采样。时钟同步通常是通过锁相环从接收数据中获得的。接收设备不断从MODEM接收数据，直到它检测出结束字符为止。

报文字组通常由1或2个同步字符、数据字符、控制字符、结束字符以及1个或2个出错控制字符组成。

报文之间，通信线路可能空闲地处于SYN字符或保持在传号状态（mark)。

异步传输主要用于人机接口，而同步传输可用于计算机间的高速数据通信。

同步MODEM可用于传送异步数据。反之，如果接收终端能从数据中获得时钟信号，异步MODEM也可用于传输同步数据。当数据传输无规律时（操作员从键盘打字输入)，异步传输便更能体现出它的优点。由于接口逻辑和线路简单，故异步传输成本较低。另一方面，同步传输由于消除了每个字符上的起始位和终止位，故可更有效地使用传输设备。同步MODEM提供较高的传输速率，但造价也较高，因为它需要精确的时钟。

当在通信链路的一端，有若干个输入输出设备时，可使用多路复用器（multiplexer）或MODEM共享设备，使许多设备共用一条通信线路，从而可降低通信费用。

多路复用器从多个终端获取低速输入数据，并将它们并入一个高速数据流，在四线全双工租用电话线上同时传送。在链路的另一端，多路复用器（实际上是一台信号分离器)，将高速数据流变换成一系列低速数据输入到计算机。信道分为时分多路(TDM)或频分多路（FDM）两种。一般可使多达12或24个设备共用一条线路。

MODEM共享设备（MSU）可使多台终端共用一台MODEM。它特别适用于远程城市集群终端的网络。

12.1.4 RS-232C接口

RS-232C是美国电子工业学会（EIA）制定的用于串行二进制数据交换用的数据终端设备和数据通信设备之间的接口标准。它规定了接口的电气及机械特性、数据、定时与控制电路的功能说明，它使用25芯插头连接，并与CCITT（国际电报电话咨询委员会）V24兼容。此标准使用的数据传输速率范围为0～20000位/秒。它适用于串行同步及异步数据通信系统。

目前国内外大部分汉字终端使用这种接口。但这一标准不能定义通过接口传输数据的方法，这种方法在不同系统中往往差别是很大的。

虽然EIA接口主要用于带有MODEM的电话线数据传输，但它还可用作终端与计

算机之间的串行接口，其电缆长度不超过 15 米。当计算机与终端的距离大于 15 米时，可使用带驱动器的双绞线。在终端一方，此双绞线带有线路驱动器，另一个驱动器放在计算机的前端。线路驱动器是一种信号变换器，它用来放大 EIA 接口驱动器送来的数字信号，以保证在大于 15 米距离上的可靠传输。EIA 接口一般限于用在几千米距离的大于几千米距离的数据通信，可用四线的双绞线对。当距离更远时，可通过 MODEM 并经长途载波线路把终端与主机连接起来。

一、RS-232C 的电气特性

图 12-5 示出了 RS-232 C 接口的等效电路。

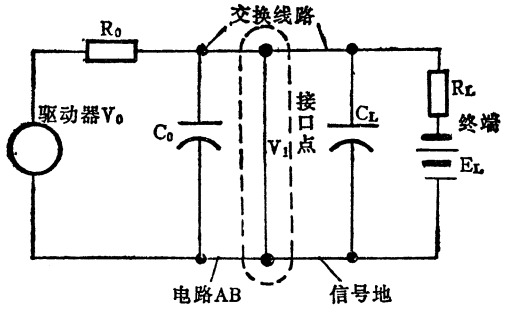


图12-5　RS-232C 等效电路

图 12-5 中：

V_0——驱动器开路电压；

R_0——驱动器内部直流电阻；

C_0——加到驱动器上的总有效电容。它在接口点上测得并包括连到接口点上的任何电缆；

V_1——接口点上的电压；

C_L——加到终端上的总有效电容。它在接口点上测得并包括连到接口点上的任何电缆；

R_L——终端负载直流电阻；

E_L——开路终端电压。

二、RS-232C 接口的信号规定

RS-232C 定义了与设备间传送的数据及控制信息有关的电平，以及有关通信的逻辑规定。交换数据时，在接口部位测得的电平低于－3 伏（相对于信号的地电平）的信号被认为是传号状态（marking），高于＋3 伏的信号被认为是空号状态（spacing）。传号状态用逻辑“1”表示，空号状态用逻辑“0”表示。此外，在定时和控制交换电路中，交换电路的电平高于＋3 伏（相对于信号的地电平）时，它的功能处于接通状态，而在低于－3 伏时处于断开状态。接通状态用逻辑“0”表示，断开状态用逻辑“1”表示。

RS-232C 是一种较常用的串行接口，它是一个 25 芯的连接插头。每一个芯脚和各种信号电平都已作了标准规定。因此便于互相连接使用。图 12-6 示出了 RS-232 C 连接图。

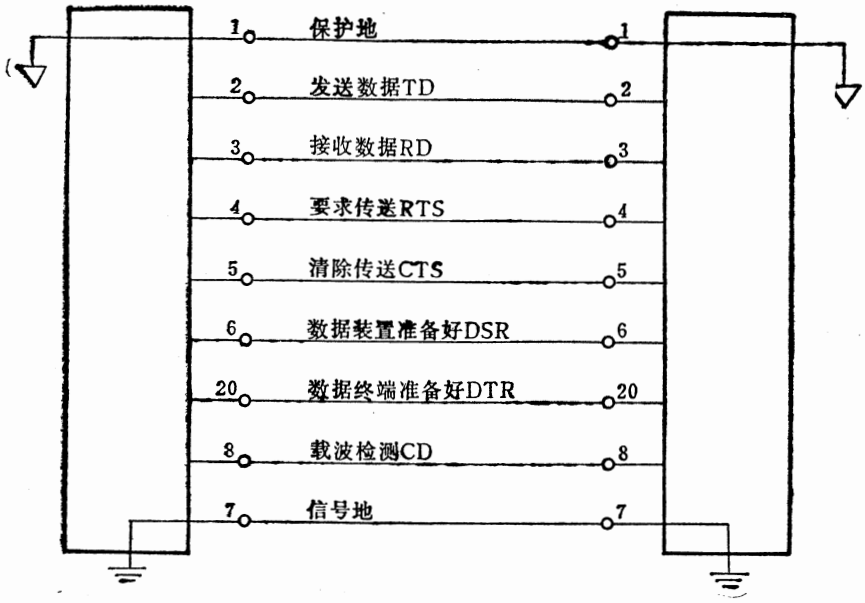


图12-6 RS-232C连接图

三、其它接口标准

虽然EIARS-232C是广泛采用的串行数据接口标准，但其连接长度局限在15米以内，且数据率低于20K位/秒。为此，EIA又颁布了新标准RS-449、RS-422A及RS-423A。RS-422A和RS-423A涉及到接口的电气特性，而RS-449规定了功能和机械特性。RS-232C的基本交换功能均包括在RS-449之内。其不同点于下：

（1）数据率扩充到2M位/秒；

（2）具有10种附加线路功能；

（3）去掉3种RS-232C线路功能；

（4）重新定义某些线路功能；

（5）接口的插头不同。

RS-232C中的DTE（数据终端设备）—DCE（数据通信设备）连接说明，均包括在RS-422A和RS-423A中。新标准完全与RS-232C兼容，并且，其最大数据率可从20K位/秒到2M位/秒。为此，RS-232C的不平衡信道线路将被RS-422A的平衡线路取代。如果不需要最大数据率时，RS-423A仍可采用不平衡线路。因此，RS-449功能标准之一是根据不同的数据率、电缆长度及噪音环境来使用RS-422A或RS-423A。

12.1.5 通信规程

在计算机网络中，为了使计算机或终端之间能够正确地传送信息，必须有一整套关于信息传输顺序、信息格式和信息内容等的约定，这一整套约定称为通信规程，或称通信协议（communication protocol）。

通信规程用来在通信系统中管理信息的流通。数据链路控制（DLC）有时也称为线路控制。它是计算机（终端）相互连接时，为保证有秩序地传输信息而必须遵循的一

组规则。

DLC 的基本功能为：

（1）在双站之间建立连接；

（2）通过误差检测保证信息的完整性；

（3）通过查询与选择判明发送者与接收者；

（4）处理专门控制功能和状态请求、站请求、起始回答和切断等。

目前国际标准化组织（ISO）制定的通信规程可分为如下两种：

通信规程
- 面向字符的规程
  - 基本型
  - 扩充基本型
    - 透明型
    - 会话型
    - 全双工型
- 面向比特的规程，如 SDLC、HDLC规程

面向字符的规程有国际标准化组织的基本型及 IBM 公司的二进制同步通信（BSC）规程。扩充基本型是在基本型的基础上，为适用于不同应用而进行了适当扩充的改进型。

面向比特的规程有 ISO 的 HDLC 和 IBM 公司的 SDLC。其它大计算机厂家也开发了自己相应的通信规程。

较高级的规程可用来实现报文缓存、代码变换、判别和报告终端或线路的故障状态、与主机的通信以及通信网络的管理等功能。这些规程由 IBM 公司的 SNA 和 CCITT 的 X_{25} 软件来实现。

IBM 系统网络结构（SNA）的组成部分为：

（1）IBM370 主机的虚拟远程通信存取方法（VTAM）；

（2）IBM370X 通信控制器的网络控制程序（NCP）；

（3）线路规则用的同步数据链路控制规程（SDLC）。

X_{25} 提供了在公用分组交换网内为多通道使用而设计的主机和数据终端使用的接口。

X_{25} 的突出特点有：

（1）线路规则采用高级数据链路控制（HDLC）；

（2）最小数据信息组为 128 个字节；

（3）具有分组终端与网络之间的多路接口。在终端与网络之间可同时有 4095 个逻辑通道工作，以启动虚拟线路服务。

12.1.6 数据链路构成

数据链路包括调制解调器（MODEM）、串行通信接口和通信信道。为了运转数据链路，需要数据链路控制设施，但不包括链路两端的计算机、终端或输入输出设备。以下介绍两种主要的数据通信链路。通信链路结构如图 12-7 所示。

一、点-点数据链路

它仅由两站之间的通信设备组成。数据链路上的所有传输必须在数据链路上的双站间进行。可在租用（非交换）通信线或交换网上建立点-点数据链路。在租用线上(永久

性连接），传输活动始终是在两个同样站之间进行的，并且可采用全双工或半双工方式。在环形网络结构中，只有一个方向的点-点通信（单工式）。

在交换网上双站完成传输任务后，要切断数据链路，用标准拨号方式（人工或自动），为下一个传输活动建立新数据链。尽管可与网络中的任一站建立新的数据链路，但在交换网上每次只能有一个信息流通方向（半双工）。如图 12-7(a)所示。如果在半双工方式的基础上增加另一个通信信道，则可构成双工方式。如图 12-7(b) 所示。

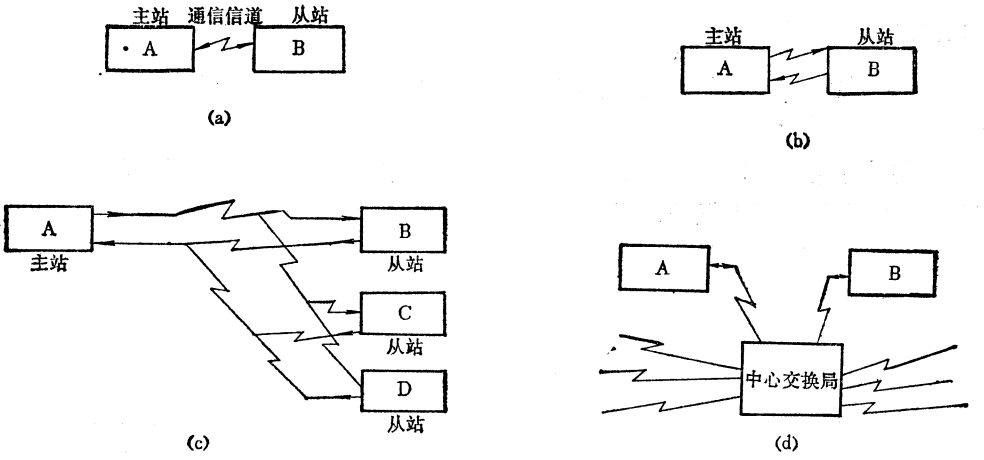


图12-7 通信链路结构

(a) 点-点、半双工（非交换式）；(b) 点-点、双工（非交换式）；
(c) 多点、双工（非交换式）；(d) 交换式（点-点、半双工）。

二、多点数据链路

对于多点工作，网络中始终有一个站被指定为主站（控制站），而其余的站则被指定为从站（或副站）。控制站用来控制多点数据链路中的全部传输。此数据链路通常建立在租用线路（非交换）上，此称为集中化多点工作方式。控制站通过选择或查询从副站来启动全部传输。数据链路的任何传输均是在指定的主站和一个副站间进行。如图 12-7(c) 所示。

网络中的其它站处于被动监视状态。多点信道可以采用全双工或半双工方式。通常，在多点信道上只有一个主站处于全双工方式，而其他从站都工作在半双工方式。

12.1.7 信息代码

为了实现组成数据链路的各种计算机和终端间的通信，需要有确定的信息交换方法。这就要求为解释字符和报文句法建立某种字符代码结构，并且采用某种实现报文字符和信息交换的数据通信控制方法。

在数据通信系统中，为了表示字符，使用了几种不同的编码方案。字符分为两类：用来表示符号的图形字符；用来控制终端和计算机功能或通信的控制字符。

目前通信系统使用的许多代码中，有 7 位加校验位的 ASCII 代码（美国标准信息交换代码），它与我国国家标准 GB1988 码兼容。

较广泛使用的另一类代码有：扩充的二进制编码的十进交换码（EBCDIC)；老式电传打字设备用的5单位代码；IBM穿孔卡片霍尔瑞斯代码；二进制编码的十进制代码（BCD)。

EBCDIC是一种与ASCII相似的8位代码。它使用第8位作为信息位。它可将字符扩充到256个。在ASCII中，用第8位作为校验位（同步传输为奇校验，异步传输为偶校验)。

在我国，汉字数据通信采用的是国家标准GB2312。每个汉字代码均用双字节表示。

12.1.8 汉字数据通信

现举一个主机与汉字终端之间、在专用线路上进行汉字数据通信的实例（见图12-8所示)。

汉字终端采用BSC通信规程同主机通信。汉字代码为双字节、ASCII代码使用单字节代码。报文组前附有报头，后面带有报尾。报文组最大长度为510字节（相当255个汉字代码)，报文中包括显示文字代码，位置确定代码及编辑代码等。

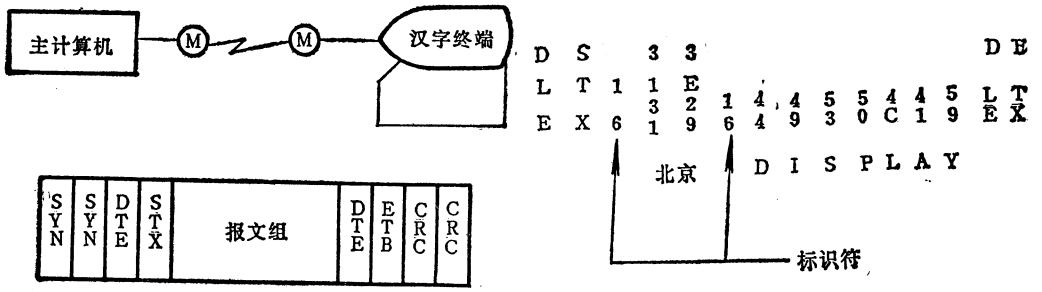


图12-8 汉字数据通信

12.2 汉字计算机网络

12.2.1 计算机网络系统

一、计算机网络组成

计算机网络已成为当前发展信息技术的重要环节，而汉字计算机网络将是我国走向信息化的一个重要标志。

由于计算机网络中的高级协议是面向比特数据的，故显然也可处理双字节的汉字数据。如果在同机种计算机网络中实现汉字与西文兼容，则可实现网络中的软件资源共享。因此，汉字计算机网络与一般计算机网络之间无本质上的差异。

计算机网络是利用通信线路把分布在不同地点的多台独立的计算机系统连接起来的一种网络。用户可共享网络中的所有软、硬件和数据资源。

计算机网络由下述三部分组成：

（1）计算机子系统；

（2）终端子系统；

（3）通信线路网子系统。

如图 12-9 所示，网络中主要有前置处理机、远程处理机等各种处理机和数据交换网。

图 12-9 中，

HOST——主计算机（CPU 及其外部设备等）；

FEP——前置处理机；

CCU——通信控制器；

BEP——后置处理机；

RP——远程处理机；

TC——终端控制器；

T——终端。

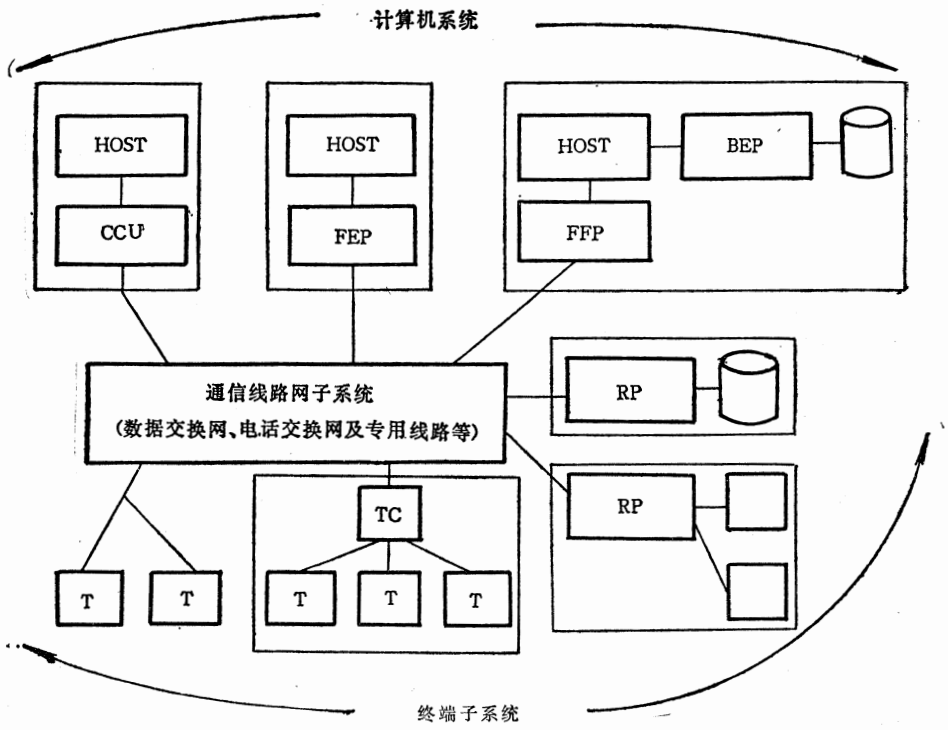


图12-9 计算机网络组成

具体来说有两种情况。一种是通过通信线路将一台中心计算机与多台终端连成的集中式数据通信系统。另一种是分布式数据通信系统。后者是经过通信线路网将原先独立设置在不同场所的多台计算机系统互连起来，以实现网络内的资源共享。

按通信线路网的构成，计算机网络大致可分为三类：一种是将中心与终端间的功能加以分散的层次分布式（或纵向分布式）；另一种是将多台中心计算机平等分布的水平分布式；第三种是将上述两种合并成的复合分布式（见图 12-10）。

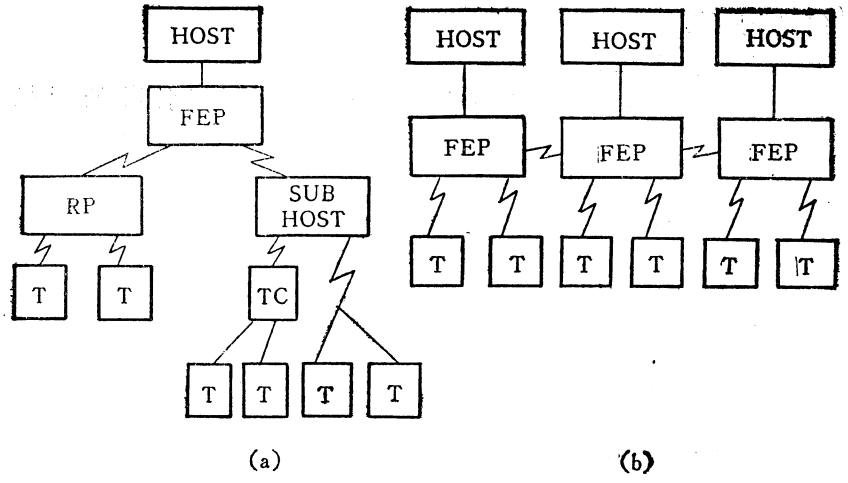


图12-10 分布式网络

（a）层次分布式；（b）水平分布式。

二、网络任务

计算机网络的主要任务是实现数据通信、信息处理与资源共享。它是用具有公共资源的主机为中心的计算机系统来保证的。终端子系统主要用来实现人机接口功能及信息输入输出功能。通信线路子系统用来实现上述系统的互连，提供数据传输手段。

三、网络

在计算机网络中，为了顺利地利用数据传输线路进行数据通信，需要在数据的发送端和接收端之间作出某些约定。我们称这种约定为通信规程。实际上通信规程是用来管理网络节点的物理链路，因此也称为数据通信链路控制规程，习惯上称为协议（protocol）。

网络中各同等层之间进行的通信是虚拟通信。这种通信，必须服从一个统一的网络协议。所以，计算机网络功能设计，大部分是制定节点同等层之间的通信协议。由于网络有多层，对应的网络协议也有多级。各级网络协议规定了该级虚拟网络的通信流程和完成相应功能的一组命令语言。实现网络各级协议的程序系统，就是网络软件(network software)。

用户级、服务级和传输级属于高级协议，负责解释和加工信息的内容。支持高级协议的程序系统一般都在主机上实现。分组级和链路级都属于低级协议，负责信息的传输。支持低级协议的软件称为通信软件，一般在前置机上实现。通信控制规程就是网络协议中的最低一级协议、即链路级协议，它实现链路上的通信控制。

四、网络控制

计算机网络其所以能够实现资源共享，主要在于计算机系统内的进程能够互相通信。因此，计算机网络的本质在于进程间的相互通信。所谓进程（process）是指一个程序的一次执行过程。它是一个动态概念。

在计算机网络内交换信息的主体（称为通信主体）是应用程序、文件和终端设备等。这些通信主体的功能构成及其交换信息的数量和格式，随着通信主体类型、应用程序的内容和实现系统的差异而有所不同。但是为了实现计算机网络的控制软件，尽可能

使这些通信主体能进行共同处理。实际上，数据传输功能的共同处理是可能的。这就是，把面向各种通信主体共同特性的逻辑模块也作为进程处理。

五、数据链路控制

它是在相邻节点间进行数据传输的一种控制。它并不约束收发报文的数据格式。为了顺利地实现各种控制功能，计算机网络的通信规程一般具有层次结构（图12-11）。按照这种层次结构，某层（如数据链路层）也把其上层（如传输层）的控制信息作为一部分数据送到对方。因此，在设计时必须考虑到：即使以后改变某层的规程，原则上不致影响其它层。

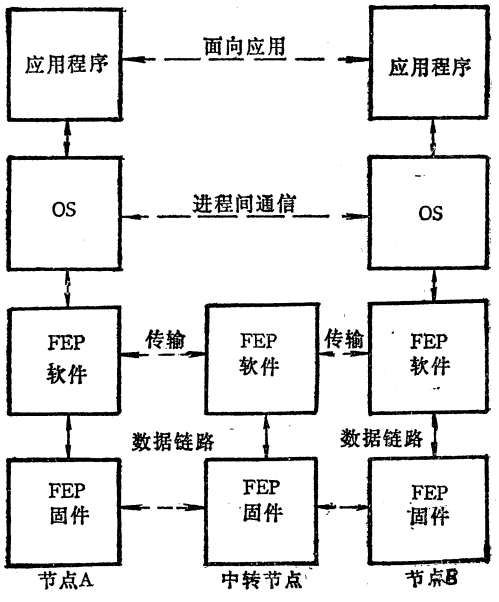


图12-11　规程与节点功能的层次结构

面向应用程序的规程，是根据通信主体种类及业务类型而设计的。对应通信主体种类的有：标准终端与主机应用程序之间交换信息用的虚拟终端规程；文件传输规程；作业传输规程等。

进程间的通信规程，用来实现下述功能：在两个进程节点间建立逻辑总线，以便使比特信息数据顺利地传输；实现信息流控制和顺序控制等。由于进程间通信规程同传输规程、信息流控制及顺序控制技术很类似，因此很多情况下是统一设计的。例如：在美国 ARPANET 网内，进程间规程中的 HOST/HOST 规程，是与通过子网络在节点间传输数据的 HOST/IMP 规程分开设计的；而在法国的 CYCLADES异种机网络中，进程间的传输同节点间的传输有着共同的传输规程。

在设计相邻节点间的传输差错检测与重新发送控制时，多半使用高级数据链路控制（HDLC）规程。下面将涉及到：作为报文分组交换网的接口 CCITT 建议——X.25，它是一种适用于 DCE-DTE 间的 HDLC 规程。

六、网络的互连

这是一种用于互连计算机网络的规程转换技术。对于具有不同体系结构的计算机网络之间，它的层次数和各层次的规程也有差异。

此外，为了将现有的各种终端与计算机网连接起来，需要通过应用程序将它们变换成具有标准规范的终端。所谓终端虚拟化，就是将终端的控制功能变换成按上述层次中指出的虚拟终端规程所需要的处理。网络中的终端虚拟化工作，大多是在各种计算机系统的前置机中进行的。

七、网络协议层次结构

为了扩大应用和使计算机网络多样化，设计了各种网络体系结构。网络体系结构用来表示计算机硬件系统的逻辑结构整体，此外还包括数据流动方式和控制方式等。它由网络系统的逻辑结构、规程及接口规范来决定。

逻辑结构规定的节点各层次称为层次规程（即同一层次节点间的通信约定）。网络体

系结构的接口是各节点共同层次间的接口。

国际标准化组织关于开放系统互连规定了如下的七层参考模式：

（1）物理层（physical layer）。它用来提供在一条物理媒体上传输信息的实际手段。它详细规定了设备之间的物理、机械及电气连接特性。

（2）数据链路层(data link layer)。它为物理链路的两端设备之间提供可靠的数据交换。

（3）网络层(network layer)。它通过连接系统的数据链路提供传送信息的手段，即为网络两端用户提供一条逻辑信道。

（4）传送层(transport layer)。它为两个开放系统之间的信息交换和端-端控制提供独立于网络的标准化协议。它包括流控制、差错控制及服务质量监督等，用于保证连接的可靠性。

（5）对话层(session layer)。当一个终点用户与另一个终点用户通信期间，称为一次对话。它具有进行通信的两个进程间的连接功能及数据传输同步功能。

（6）表示层(presentation layer)。它为两个进程之间传输数据提供数据格式、文件格式及终端显示格式等格式变换功能。

（7）应用层(application layer)。它为开放系统中的应用进程提供支持，包括各种用户协议及系统管理等。

八、网络协议层次结构举例

美国IBM公司的计算机网络体系SNA具有如图12-12所示的通信功能。

图中：

DLC——数据链路控制；

TC——传输控制；

DFC——数据流控制；

FM——功能管理。

SNA具有如下几个部分：

（1）应用业务层：执行与用户数据处理有关的应用功能。

（2）功能控制层：它属于信息格式变换及数据收发的流动控制。

（3）传输子系统：它包括传输控制、路径控制及数据链路控制的各层，并控制数据传送。

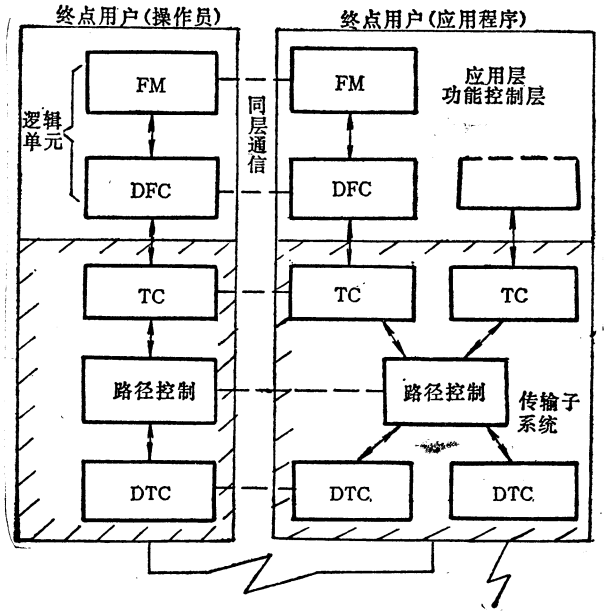


图12-12 SNA网络体系

12.2.2 数据交换方式

一、概述

数据通信系统的信道，一般使用专用线路和电话交换线路。但从功能、性能及经济性等方面看，满足多种数据通信需要还有困难，所以有必要实现最适用于数据通信的交

换网。

二、线路交换与包(packet)交换

(一) 线路交换

线路交换在电话系统中最常用，自动电话交换局就是一例。在这种交换技术中，通过电话的转接来提供需要的通信线路。线路接通之后，两个连接的终端之间可以直接通信。通话完毕之后，线路就可提供别的用户使用。

电话交换广泛用于传输模拟信息，也可传输数字信息。最初，线路交换采用机电或电子方式进行交换控制。近来又采用了全电子数字技术，将模拟信息变换成PCM（脉码调制）码传输，用时间分割复用技术进行交换连接。如图 12-13 所示。

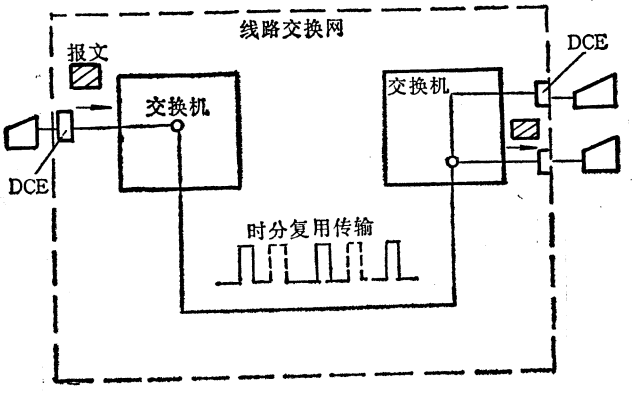


图12-13　线路交换原理

(二) 包交换

包交换是存储转发交换技术中的一种交换方式。这一交换方式是以称为包(packet)的报文组为单位进行的，如图 12-14 所示。从输入线路送来的数据，在交换中心内存储后，便选择输出线路。接着把存储的数据送往下个交换中心或数据终端设备。具体地说，发报方将报文分解成若干个包，选择空闲信道按照报头内容传输到目的地。收报方再将接收到的包信息组装成原先的报文格式，然后送往终端用户。这种交换方式的优点是，允许按不同传输控制规程工作的数据终端加入网络，使具有不同速度的数据终端也能借助网络进行通信，从而有利于实现资源共享。

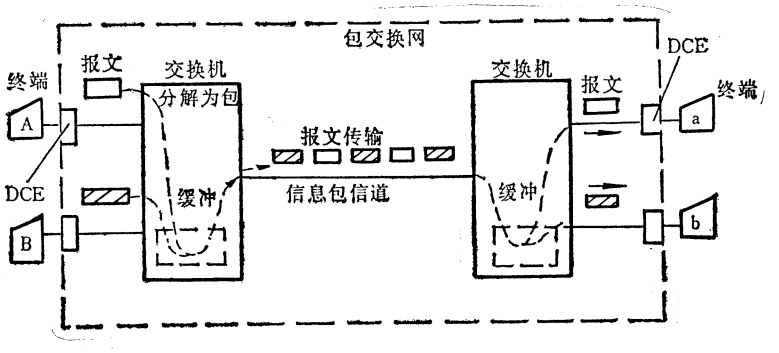


图12-14　包交换原理

三、交换网的构成

终端和计算机借助通信线路进行连接的方式有下列几种（如图 12-15 所示）。

(一) 交换连接

图 12-15(a)说明终端与交换中心之间的线路经常为终端所占用。在有数个交换中心的情况下（如图12-15 b），当全部终端都通信时，可占用交换中心与交换中心之间的

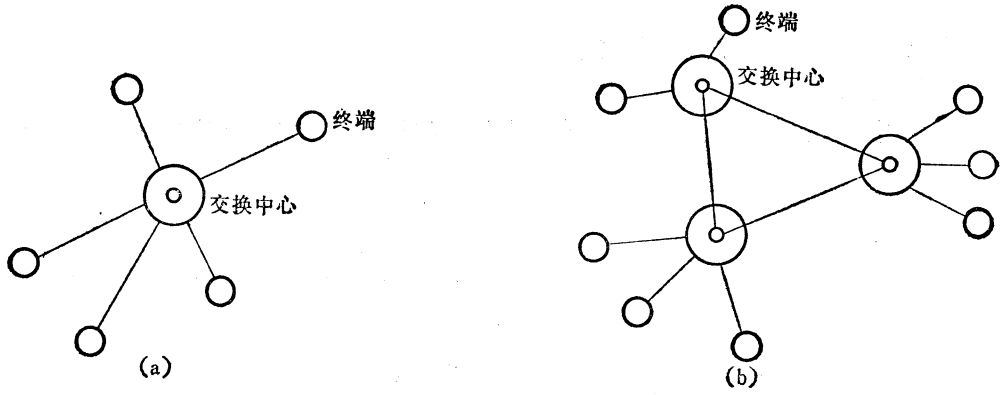


图12-15　交换连接

（a）只有一个交换中心；（b）有数个交换中心。

线路。因此，采用交换连接，可实现全部终端之间的通信。这种交换连接方式适合于大型网络。

（二）分支连接与集中连接

图 12-16 示出了几种连接方式。其中包括分支连接和集中连接。

环状连接是将计算机与终端用高速信道以环状形式连接在一起。它具有分支连接和集中连接两者的优点，适用于高速通信。

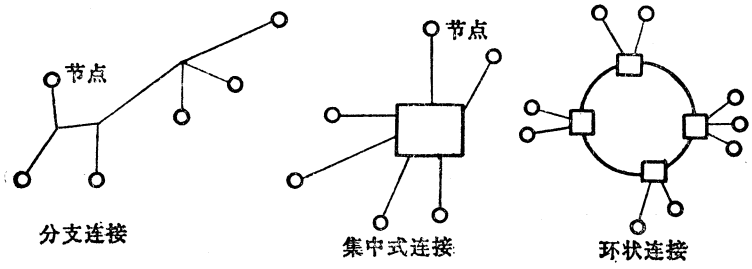


图12-16　几种连接方式

四、各种交换方式比较

交换网络的综合性能取决于网络的经济性、可靠性及网内传输延迟时间等因素。线路交换有赖于物理信道和载波延迟时间，而存储转发交换则取决于交换中心的存储时间。由于改变路径中的中继站数目可能有增减，所以传输延时会有变化。线路交换的连接时间长，而包交换的连接时间最小。由于存储转发交换方式容易实现地址确认、数据格式检验，以及交换中心的差错控制和数据记录等，所以，所传输的数据可靠性高。在存储转发交换时，可利用存储功能进行速度变换及字符格式变换，所以可与各种不同终端连接。但是，线路交换方式则不可能实现上述功能。存储转发交换方式可将终端出入的低速信息集中起来，使中继线路高速化，从而能扩大吞吐量。此外，由于线路的多路利用，可压缩判断时间及应答等待时间，故线路利用率高。而线路交换方式的线路利用率则很低。

12.2.3 传输控制

一、概述

数据传输方式分为串行与并行传输两种。按照ISO标准，字符用8位表示。串行传输是将8位信息以位为单位在时间上逐位传输。并行传输是同时传输8位信息。

为了通过通信线路正确地在数据终端设备(DTE)之间传输数据，首先要对MODEM等数据通信设备（DCE）进行必要的控制。DCE将DTE送来的数字信号转换成模拟信号，用编码形式在普通线路上传输，此过程称为调制。接收时，再进行解调，把模拟信号变换成原先的数据。

（一）数据信号速度

数据信号速度为单位时间内传输的信息量。其表达式如下：

$$S〔位/秒〕=\sum_{i=1}^{m}\frac{1}{T_i}\log_2 n_i$$

其中，m——并行信道数；

T_i——第 i 信道的最小间隔（脉冲宽度），秒；

n_i——第 i 信道的调制状态数。

串行传输时，脉冲只处于0与1两种状态，所以，$m=1$，$n_i=2$，从而有

$$S=\frac{1}{T}\log_2 2=\frac{1}{T}$$

（二）调制速度

它是单位时间内被传输信号的状态变化次数。单位用波特表示。

$$B〔波特〕=\frac{1}{T〔秒〕}$$

（三）数据传输速度

它是单位时间内被传输的数据量。如果M位数据在T（秒）时间内传输，则数据传输速度为M/T。数据量可用信息位表示，时间可用秒、分和小时表示。

二、通信规程

为了保证报文能在通信子网中的实际链路上正确有序地传输，需要制定数据通信传输控制顺序或通信规程。

从六十年代初期开始，为了适应批量通信的需要（即主机与远程终端之间的双向数据传输），发展了各种面向字符的通信规程。大量使用的规程有下列三种：

（1）美国国家标准协会ANSI的通信控制规程；

（2）国际标准化组织ISO的基本型传输控制规程；

（3）IBM公司的二进制同步通信规程BSC。

此类规程产生于终端网络的创立和发展时期，适用于主计算机与终端之间的通信。

六十年代末期，不仅由于通信量的增加，而且由于分时、查询响应、数据库及计算机网络等应用面的扩大，数据通信有了明显发展。

基本的面向字符的规程虽几经扩充，但仍不能适应新的应用。为了提供新的功能和操作模式，创立了面向位的通信规程。大量使用的通信规程有以下三种：

（1）1969年贝尔实验室提出、由IBM公司系统采用的同步数据链路控制规程SDLC。

（2）美国国家标准协会ANSI在SDLC的基础上提出了它的先进数据通信控制规程ADCCP。

（3）国际标准化组织ISO以SDLC为基础提出了它的高级数据链路控制规程HDLC。

这些规程产生于计算机网络的创立和发展时期，适用于计算机与计算机之间的通信。

三、面向字符的数据通信链路控制规程

它是国际标准化组织ISO的基本型传输控制规程。所谓面向字符是指所传输的信息是按一定的字符格式进行编写的。基本型传输控制规程采用国际电报电话咨询委员会CCITT的七位编码。所有的传输功能是使用特殊的传输控制字符来完成的。这些控制字符在这种编码表中被定义为TC_1～TC_{10}。

（一）适应环境

（1）传输可用任何一种速率来进行；

（2）可以串行和并行传输；

（3）同步方式可以是起止式或同步式；

（4）可以是点-点或多点线路；

（5）可以是专用线或交换线；

（6）物理接口：DTE与DCE之间的接口遵循CCITT的V 24建议。

（二）报文格式

一次传输的信息单位称为一个报文（message）。报文分二类。一类称为信息报文（I），用来向接收方传递用户数据；另一类称为监控报文，用来在收发双方之间传输监控信息。与信息报文传输方向一致的称为正向监控报文（F）。与信息报文传输方向相反的称为反向监控报文（B）。信息报文和监控报文都是字符序列，而且每个字符序列至少包含一个控制字符，用以区别报文类别和不同的监控功能。

基本型传输控制规程的10个控制字符如表12-2所列。

表12-2　基本型传输控制规程的10个控制字符

控制字符	名称	适用电文类别	功能
(TC_1)SOH	序始	I	表示信息报文的报头开始
(TC_2)STX	文始	I	在正文前，表示报头结束，正文开始
(TC_3)ETX	文终	I	在正文之后，表示结束
(TC_4)EOT	送毕	F、B、	通知对方，传输结束
(TC_5)ENQ	询问	F	询问对方，要求对方应答
(TC_6)ACK	承认	B	肯定应答
(TC_7)DLE	数据链转义		由后续字符赋给传输控制功能
(TC_8)NAK	否认	B	否定应答
(TC_9)SYN	同步		用于取得和保持同步
(TC_{10})ETB	组终	I	正文信息组结束

信息报文中含有一分用户数据，简称正文。有时正文之前还带有一个报文头，在报文头中可包含一些有关正文的辅助信息，例如发信站址、时间、序号、报文名称及优先

级等。报文头的长度和含义由用户自定。一个长信息报文也可以分段传输。信息报文的基本格式有五种，如表12-3所列。

表12-3　五种信息报文的基本格式

STX	SOH	SOH	SOH	STX
正 文 ETX (BCC)	报 头 STX 正 文 ETX (BCC)	报 头 ETB (BCC)	报 头 STX 正 文 ETB (BCC)	正 文 ETB (BCC)

一个完整的信息报文可以从这五种基本格式之一开始，相继发送相应形式的若干信息报文段。整个信息报文是一段一段发送的，最终出现 ETX 时，表示 报文结束。

BCC 为某个信息报文段的块码组校验字符。当采用方阵码校验方式时，它 的 长 度为一个字符；当使用循环冗余码校验方式时，它的长度为两个字符。当不采用校 验 时，就没有 BCC 字符。

为了对信息报文进行控制，需要在通信的双方之间传送监控报文。表12-4列出了几种格式。

表12-4　监控报文的几种格式

探　询	选　择	询　问	结　束	切　断	肯　定	否　定
EOT 探询地址 ENQ	EOT 选择地址 ENQ	(前缀) ENQ	(前缀) ENQ	(前缀) DLE EOT	(前缀) ACK	(前缀) NAK

（1）探询：在多点线路方式中用于命令对方发送。其探询地址为被探询方 的 站地址及发送设备地址。

（2）选择：在多点线路方式中用于命令对方接收。其选择地址为被选择方的 站 地址及接收设备地址。

在探询和选择中以 EOT 开头，表示链路上原来的收发关系必须结束，要 重新建立收发关系。

（3）询问：用于询问对方状态。在点-点线路方式中，用于建立数据链，向对方请求发送。

其前缀是由 EOT 开头的，最多包含 15 个字符所组成的序列，用以 表 示 监控报文的类别。也可以没有前缀。

（4）结束：用于结束数据链（即这次主/从关系结束）。也用于否定探询。

（5）切断：用于断开线路连接。

（6）肯定：用于肯定选择，也用于肯定信息报文正确接收。

（7）否定：用于否定选择，也用于表示接收的信息报文有错。

（三）数据通信的几个阶段

数据通信的完整操作过程可分为五个阶段，通信控制规程必须具体规定这五个阶段的一些手续。

面向字符的通信控制规程对通信过程的控制全部是用预先规定好的字符进行的，而且要对字符编码作统一的规定。通信控制规程还必须统一规定各种报文序列以统一报文含义，例如上述信息报文及监控报文等，否则双方无法对话。

数据通信的五个阶段与日常打电话过程完全相似。

第一阶段：接通线路

这相当于打电话时必须拨号，使对方电话机和自己电话机之间的线路接通。数据通信使用公用交换网时，必须使通信的两站在公用交换网上建立连接。这个连接是通过人工拨号或自动拨号由长途通信管理局来建立的。完成这个阶段后，主叫站负责组织通信，起主站或控制站的作用。在专用线路上，双方的线路总是接通的，所以本阶段可以省略。

第二阶段：建立数据链路

这就是确定通信的对象和收发状态。这相当于打电话时通知对方自己是谁，并确认对方是否为自己要找的对象，确定是自己先说，还是让对方先说。在数据通信中，此阶段是用正向监控报文和反向监控报文来进行的。当一方收到正向监控报文后，要根据情况作出相应的应答，发出反向监控报文，以这种方式给期望的对象确定收发状态。

（1）点-点方式。要发送用户数据的站先发出 ENQ 请求，若得到对方同意，并收到 ACK 回答后，就建立起数据链。然后进入第三阶段。若发送 ENQ 请求后，由于种种原因仍没有收到 ACK 回答，则重发 ENQ 请求。

在半双工方式中，两个站有可能都要求发送，都发出 ENQ 序列，从而出现所谓争夺线路现象。为防止争夺，通常事先指定一个站有优先权。

（2）多点方式。在多点方式通信中，第二阶段是采用探询序列来建立数据链路的。主站作好接收准备后发出探询序列：

EOT
SA
UA
ENQ

其中，EOT——使所有和线路相连的从站的数据链路都结束，然后由主站重新建立数据链路；

SA——被探询站的站址；

UA——被探询站的发送设备地址；

ENQ——询问。

尽管线路上的所有从站都可以收到这个探询序列，但是只有与SA、UA值一致的从站能辨别出是探询自己。若被探询的从站有用户数据要传送，则可立即进行数据传输，于是进入第三阶段。若被探询的从站没有用户数据要传送，则回答 EOT 序列，以结束数据链路。若经过一定时间，主站既未收到信息报文，也未收到 EOT 序列，则重发探询序列。若主站有用户数据要向某从站发送，则采用选择序列建立数据链路：

EOT

SA
UA
ENQ

此处，UA为被选择站的输入设备地址，其它内容同于探询序列。显然，可用UA值来判定是探询还是选择。

若被选择的从站已做好接收用户数据的准备，则回答ACK序列。主站收到ACK回答后，即进入第三阶段，发送用户数据。在整个第三阶段，只有被选择的从站接收用户数据。若被选择的从站不能接收用户数据，则回答NAK序列，致使主站重发选择序列。

第三阶段：信息传送

这相等于借助电话线路和对方谈话。通常谈话中要作适当的停顿，等待对方回答，判断对方是否已确实听清楚。在数据通信中，把要传输的用户数据划分为若干传输块，一块一块地传输。在传输完一块之后要等待从接收端发来的应答，然后决定是向对方重发这一块，还是发新的一块。

信息传输时，由于线路上有杂音干扰，可能使信息错误地到达对方。为此，在传输报文块时，常常要附加用于差错控制的冗余码。在接收端进行差错校验，用应答字符（即“肯定”或“否定”）把校验结果告知发送端。若收到肯定应答ACK，而信息报文尚未发送完，则继续发送下一个报文块，然后再转入等待应答状态。如果没有信息报文要发送，则进入第四阶段。若收到否定应答NAK，或经过一定时间仍未收到应答，则重新发送这一报文块。为了保证对方正确接收，采用了在不论由于什么原因而造成收不到ACK应答的情况下进行重发的措施。这种方法可能会造成信息报文块重复。为此，可以由用户约定，对报文块进行编号并附加在信息报文块中以作识别。此信息传输阶段一直继续到这次要传输的用户数据完毕。在此阶段中，主站发送信息报文，从站接收信息报文，不改变站点的状态及其职能，即所谓选择保持。

第四阶段：结束数据链路

这相当于打电话时，告知对方话已说完。在数据通信中，上述第三阶段结束就意味着信息报文发送完毕，并被从站正确接收。主站发送结束字符EOT，要求停止传送信息并放弃主站状态，交出通信控制权。如果本链路上的其它站要求发送，则可发ENQ序列，再次进入第二阶段，建立新的收发关系。如果没有其它站要求发送，便进入第五阶段。

第五阶段：切断线路

这相当于挂掉电话。在数据通信中，第五阶段是由控制站发出DLE·EOT序列，致使数据终端准备好下降线路上的电平，从而切断同公用交换网的连接。若采用专用线路，则因双方之间的线路总是接通的，故可以省略第五阶段。图12-17（a）、（b）、（c）示出了第二、三、四阶段的序列传输情况。

（四）基本型传输控制规程的扩充

基本型传输控制规程有两个突出缺点。一个缺点是代码相关，不允许十个控制字符出现在所传输的数据内容中，即用户不能传输任意的代码信息。另一个缺点是单向传输。线路接通后只能在一个方向上传输报文，为了改变方向必须重建数据链路，以确定新的

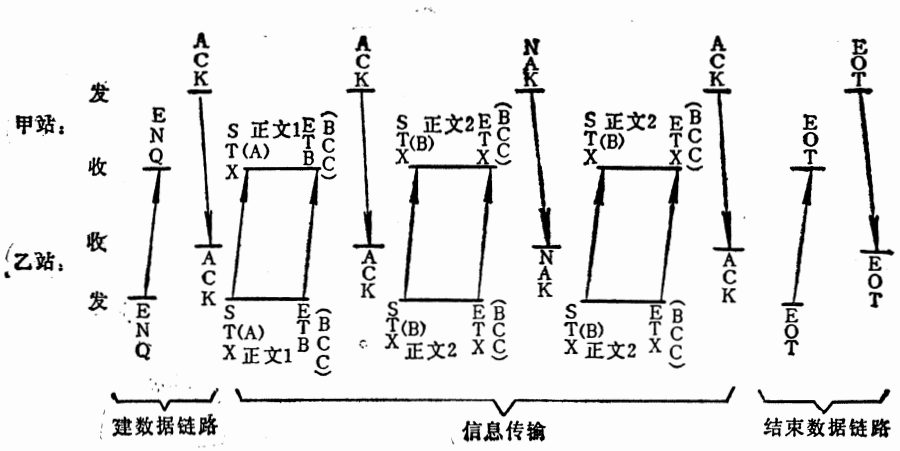


图12-17(a)　点-点方式

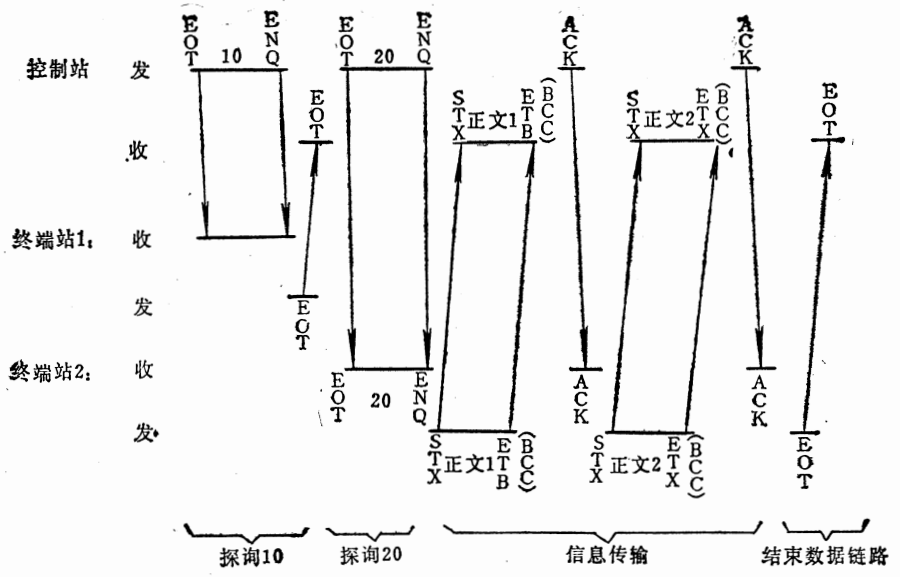


图12-17(b)　多点探询方式

收发关系（即采用选择保持方式）。这样就造成了如下局面：经常重复上述第二阶段到第四阶段的过程，从而增加了线路周转，降低了效率。为了克服这些缺点，以适应传输任意代码、便于会话操作，以及计算机之间的同时双向通信等功能的要求，改造了某些传输控制功能，产生了代码透明性传输控制规程、会话型及全双工传输控制规程。这些规程统称为扩充的基本型传输控制规程。

1．透明型（transparent）传输规程　这里，为了在数据内容中能传输任意的代码，仅对基本型传输控制规程的第三阶段加以改造和扩充。主要思想是：设法避免随机信息字符与控制字符之间的混淆。

用复合控制字符序列DLE·SOH，DLE·STX，DLE·ETB，DLE·ETX，DLE·SYN，代替控制字符SOH、STX、ETB、ETX和SYN。即在第三阶段中，作为传输

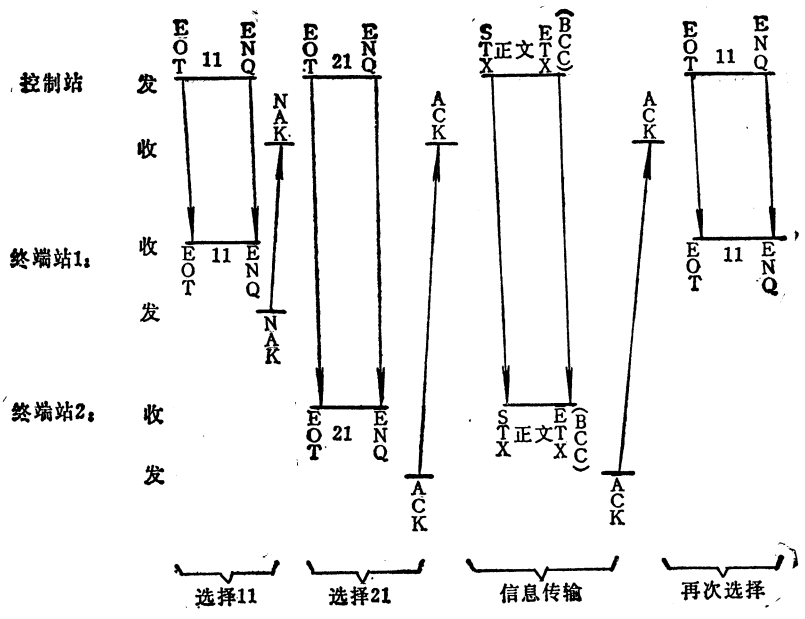


图12-17(c) 多点选择方式

控制用的真正控制字符，其前面必须冠以 DLE 转义字符。另外，为了区分出复合控制字符序列，规定在数据内部出现的 DLE 字符必须重复。也就是说，在发送端的数据内容中遇到 DLE 字符代码时，传输控制部分必须自动重复 DLE 字符代码；在接收端数据内容中遇到 DLE 字符代码时，必须是双个，而且要自动去掉一个。这就是透明性规则。

2．会话型传输控制规程　为了满足会话过程中反复改变传输信息报文方向的要求，而又要避免反复进行基本型传输控制规程中的第二阶段和第四阶段，就需要对第三阶段加以改造。这就是简单地用传送信息报文来代替 ACK。即当正确收到对方信息报文后，如果自己有报文要发送，就不发 ACK 序列，而直接发送报文。这是一种双方轮流发送一次报文的对话方式。中间省去了重建通信链，改变收发关系的过程。这个过程一直继续到双方都没有报文再需要发送时，才进入数据链路结束阶段。

3．改造全双工基本型传输控制规程　为了适应双向同时传输报文的要求，对全双工基本型传输控制规程作如下的改造。因为全双工没有竞争发送权的问题，故可省略第二阶段和第四阶段。其次，把给对方的监控报文插入对方的信息报文中。采用中断应答方式可以及时回答对方关于信息的正确与否，发出一个信息报文块后，应等待应答。根据应答情况，决定重发旧报文还是发新的报文。在等待时间内，以同步控制字符 SYN 来填充。开始时，需要建立链路，一旦建立，就保持收发关系。在无信息发送时，以同步字符填充，以保持同步关系，这样随时都可以发送信息。数据链结束和切断线路这两个阶段，与基本型相同。

4．差错检测规则　在传输速率为 1200位/秒 以下的起止式线路上，一般采用方阵码检测。在 1200位/秒 以上的同步式线路上，一般采用循环冗余码检测。在方阵码检测中，对于垂直奇偶校验，在同步系统中采用奇校验，在非同步系统中采用偶校验；对于水平

奇偶校验，都采用偶校验。在循环冗余码校验中，采用如下的优选生成多项式进行校验：

$$G(x)=x^{16}+x^{12}+x^5+1$$

四、面向位的数据链路控制规程

1969 年 IBM 公司首先提出同步数据链路控制（SDLC）规程。1972 年 ISO 提出了高级数据链路控制（HDLC）规程。两者都是面向位的传输规程。因 HDLC 已被推荐为国际标准，故在本段内容中主要介绍 HDLC。

早期终端网络使用的面向字符的数据链路控制规程，是以字符作为传输信息的基础。虽对其性能作了多方面的扩充，但仍不能满足在后来建立的计算机网络中，计算机与计算机之间传输任意长度二进制位流（*bit stream*）的要求。因此，又创立了面向位的数据链路控制规程。

这类规程提供了面向位的能力。也就是具有如下一些特点：既能传输任意位组合的用户数据，也能使数据的长度为任意的，而不必为某种字符编码的整数倍；可以按任意一种速率进行传输；可以按点-点方式工作，也可以按多点方式工作；既可以是半双工的，也可以是全双工的；既可用专用线路，也可用交换线路；它不能使用起止式，只能使用同步式；DTE 与 DCE 之间的物理接口遵照 CCITT 的 V 24 建议。

面向位的规程提出了几个新概念。其中之一是“主站”和“从站”概念。面向位的规程基于下述概念：在一个公用通信链路上，“主站”与一个或多个“从站”之间进行数据交换。主站负责初启、数据传输的组织及链路差错恢复等链路控制；从站只是执行主站指示的操作。主站和从站的分配是固定的，它们不能动态地改变。其优点是可将智能集中在主站。

（一）命令和响应

对链路操作的控制，取决于主站-从站间的命令和响应的交替情况。命令是由主站发向从站，要求从站执行指定的操作。只是在要求应答先前的命令时，才由从站向主站发出响应。某些命令和响应可以包含数据，其它内容仅用作链路操作的控制功能。

（二）帧的结构

帧是在主站与从站之间通过链路传输的一个完整的信息块，它是链路协议的数据单位。帧类似于面向字符规程中的一个报文段。用户信息或链路控制信息使用的帧，都具有标准的帧格式。

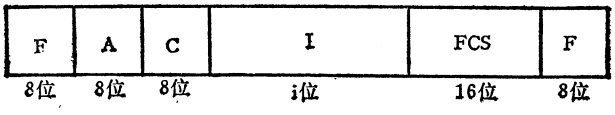


图12-18（a） 帧的格式

F—标志序列；A—地址段；C—控制字符；I—信息数据段；FCS—帧校验序列。

（三）帧的格式和内容

所有帧均采用统一的格式（见图12-18）。

下面说明一下帧格式中的各项内容。

1．标志序列F　采用一个唯一的位序列 01111110 来限定一帧的开始与结束，同时也用作帧同步。所有站不断监视信道，以发现F标志序列。实现这种功能的是所谓连续搜索系统。当发送时，工作站监视两个标志之间发送的位序列，如果发现有连续五个“1”位，则发送站必须在信息流中插入一个“0”位。由于这种“0”位插入的结果，在帧里将不会有多于连续五个“1”位的情况。此法适用于标志之间（但不包括标志）

的所有位。当接收时，工作站对连续五个“1”位后面的这一位进行检查。如果它为“0”，则接收站在向接收装置提供信息之前，把“0”位从信息流中去掉。如果它是“1”，则继续考察第7位。若该位为“0”，则接收站作为收到一个F结束标志来处理；若该位为“1”，则拒收该帧，因为若收到连续七个1，就意味着收到了取消序列。

2．地址段A　它用于指定收信地址。在向从站发送时，此地址为对方站址，用于指定由哪一个从站接收该命令帧。向主站发送时，此地址为本站站址，用以指出此响应帧是由哪一个从站发出的。

3．控制段C　它包含命令或应答代码及帧序号等，用以定义帧的类型和参数。采用帧序号制可以帮助检查帧的多余或缺少，从而可提高可靠性。主站使用控制字符向从站指示该帧为信息传输或监视命令，或者为无顺序命令。从站使用控制字符向主站表示该帧为信息传输或监视响应，或者为无顺序响应。控制字符格式见图12.18（b）。

4．信息数据段I　此段数据最长255字节。当帧的内容作监视用时，数据部分可为0。信息段从控制字符后面第1位起，直至FCS前最后一位止。信息段的长度除了受信息传输过程中有关站的缓冲容量限制外，一般可为任意长。信息段可含有任何形式的位序列，即具有完全透明性。帧中的各段由于是与位置相关，所以信息段除用来传输用户数据外，还可附带传输报头信息、控制信息和状态等，但不可包含链路控制信息。

5．帧校验序列FCS　每帧内部都含有此序列，用于差错检测。差错校验计算是对整个帧的内容（A、C、I）作循环冗余码校验（CRC），而标志序列和按透明规则插入的所有0，则不属校验范围。FCS是用CRC生成多项式按模2除帧的内容后所得的余数。

它是把要传输的原信息码多项式P，用生成多项式G除，所得余数R作为校验码加到原信息码上，编成线路传送码。在终端对接收到的传送码用同一生成多项式G除，若除尽，则说明传输无错，此时便去掉校验码字，而恢复成原来的信息码字。

其公式如下：

$$\frac{P}{G}=Q+R$$

式中　P——发送数据多项式；

G——固定生成多项式；

Q——商多项式；

R——余数多项式。

应当指出，A、C、I均由软件形成和处理。实现面向位的链路控制规程的软件是通信软件的一部分，其主要组成是帧格式的判别及帧的处理与形成。硬件仅完成二进位串缓冲、接收、发送以及F、FCS的产生和检测，还要负责透明性规则的实施。

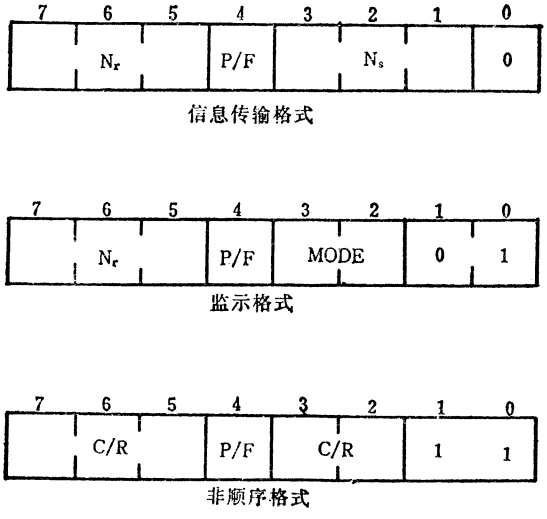


图12-18（b）　控制字符格式

（四）HDLC规程的特点

1．传输透明性好　HDLC中仅采用01111110作为标志。硬件方法中采用0位插入法，可较简单地保证标志的唯一性，而面向字符的规程则要搜索所有信息中是否出现十个控制字符。因此HDLC用来传输随机码是方便的，即传输透明性好。

2．报文格式统一　HDLC不论传输数据还是传输监视信息，均采用统一的帧格式，用以代替面向字符的控制规程中的控制字符功能的信息。

3．可靠性高　HDLC除了标志信息外，都要作CRC循环冗余校验。在面向字符的规程中，重要的控制字符仅有奇偶校验，但HDLC还用了帧序号校验来提高传输可靠性。

4．传输效率高　HDLC采用连续发送方式，可连发若干帧。由于有地址字段，一个站可同时与多个站通信，因此提高了传输率。用0位插入法作同步校验，这同面向字符的规程中插入同步码组的方法相比，控制简便并能提高传输效率。

5．扩充性好　HDLC要扩充功能时，只要改变控制字段内容和规定即可。而在面向字符的规程中，则需用转义符，这需要增加设备才能实现。

图12-18（b）示出了控制字符格式。

12.2.4　通信控制

一、概述

为了正确地进行通信，收发双方均需要按照事先约定的方法工作。按此方法进行报文通信的功能，称为通信控制的基本功能。它综合了下述功能：计算机经通信线路接收报文；接收发送报文控制；计算机或终端之间通信控制；传输控制等。

这一通信控制的基本功能，过去一般由计算机的操作系统和通信控制器承担。近年来，随着网络与分布技术的发展，各种专用的通信处理机、远程处理机和前置处理机得到了广泛应用。

典型数据通信系统示于图12-19。它由下列各项组成：处理信息的计算机系统；收发输入输出信息的数据终端；连接终端与计算机并传输信息的通信线路。早期，对通信线路的管理是由计算机中的CCU（通信控制器）完成的。

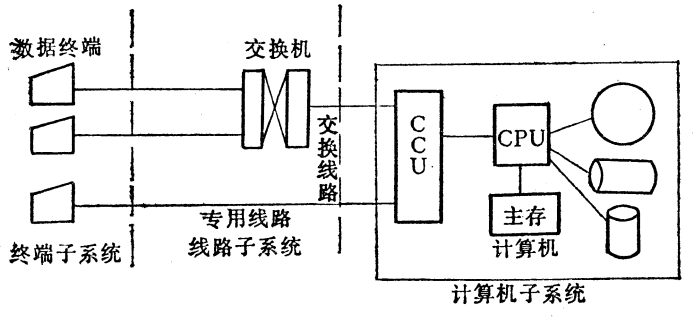


图12-19　数据通信系统的组成

通常，一个报文由若干字节组成。过长的数据要分成一定长度的数据块。计算机和终端内部按并行方式处理，而在通信线路上则按串行方式传输。这就需有并/串行变换及字符格式处理等功能。此功能均由CCU完成。图12-20示出了CCU的基本功能。

联机系统CCU的作用是实现一种I/O控制功能。它用来对接收的数据实现缓冲存储、进行速度变换与匹配，此外，它还要处理异步输入输出的数据，并且进行时分多路控制（此控制与多路通道功能相似）。六十年代，在CCU所采用的多路通道中，有一个子通道与一条通信线路连接，这个子通道不仅具有与线路之间的物理接口功能，还具有字符分解与组合等功能。在主计算机CPU内，具有传输控制等处理功能。

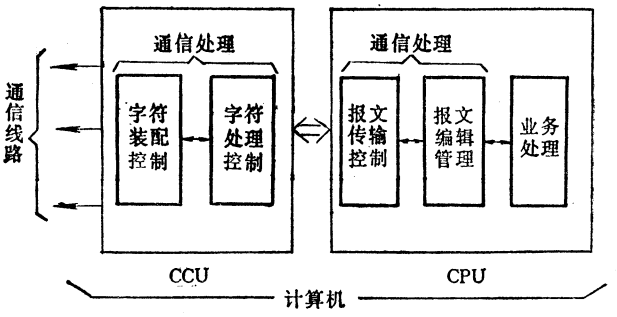


图12-20 通信控制器（CCU）功能

二、通信处理机

在早期的系统中，由于线路容量小和通信速率低等原因，计算机与通信网之间的接口功能是由软件完成的。其硬件仅完成基本的物理转接、数据缓冲等，而字符识别控制、传输控制、检错、传输块的装配与分解，以及超时检测等，都由主机软件完成。此种通信硬件只能称为简单的通信控制接口。随着大型联机系统、计算机网络及各种新型终端设备的出现，通信功能日益复杂化，因此，通常尽可能用硬件来实现通信控制功能（尤其是一些固定的、重复的适合硬件实现的功能）。这样，在CCU难于胜任的情况下就出现了各种功能完整的通信处理机。

通信处理机是通信子网中的主要部分。它是通过硬件与软件的结合来实现通信控制功能的。其硬件主要是一台适合于通信控制的专用处理机，它有通信控制所需要的专用命令。其软件附属于通信处理机内，一般可将软件固化。这种设备的功能较强，除了具有物理接口及字符缓冲功能外，还能识别控制字符，实现传输控制、块码缓冲及差错控制等，从而可大大减轻主机通信软件的负担。这种处理机的优点是：整体性强；线路容量大；处理速度快；逻辑表达能力强；有一定适应性和兼容性。其缺点是：通用性不如前置处理机，对主机有一定依赖性；其成本较高。

三、前置处理机（Front End Processor，FEP）

伴随计算机网络技术的发展，七十年代又出现了前置处理机和远程处理机（remote processor），用来作为计算机网络中的节点机。如果我们将通信处理机的功能再加以扩展，即实现有关链路级规程的功能（而主机只处理应用业务），则变成了前置处理机。前置处理机实际上是一台带有通信控制接口的能执行通信软件的计算机。它一般是在16位小型机的基础上配置通信接口而成的。图12-21示出了FEP的主要功能。

图12-22示出了有代表性的FEP的组成情况。它是以逻辑部件与存储器作为主体，加上通信线路接口、主机接口和外设接口而组成的。如果工作在远程集中器（remote concentrator）和远程处理机状态，则当FEP经过通信线路同主机连接时，不需要主机接口。根据FEP的用途与使用要求，也可以不使用外部设备。典型的FEP是根据需要由若干功能模块组装而成的。其线路适配器（line adaptor）数量随通信线路数目不同而有所不同。

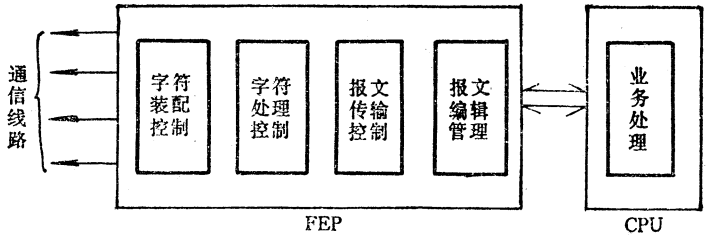


图12-21　前置处理机（FEP）的功能

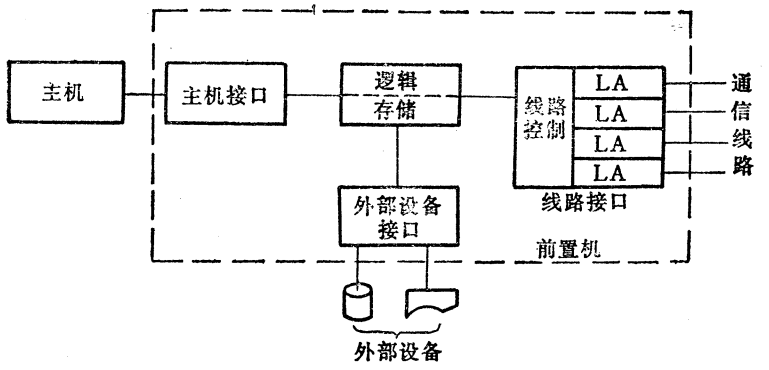


图12-22　前置处理机结构

（一）逻辑部件与存储部分

实用的 FEP 是在现有的小型计算机的基础上加上一些接口而设计成专用机。其小型机的存储容量并不大。近来也有在 FEP 中采用微处理机作为逻辑部件的，这可用多台微型机工作，以补充单台机能力的不足。对于普通的 CCU，仅有 30～50K字节的容量已足够了。为实现规程处理，报文块的分解与装配等高级通信控制功能，就需要扩充存储容量。而缓冲存储器容量也会随着线路数及功能的增加而增加。大型 FEP 一般有 100K字节以上的存储容量，高性能 FEP 大约需要 256～512K字节的存储容量。当然，还可根据功能，线路数及业务量，将 FEP 所需存储量按 32K 字节或 64K字节成倍扩充。

（二）通信线路接口

它由若干线路适配器（LA）和线路控制部分组成。每条线路对应一个 LA。逻辑部件与存储部分同 LA 之间的数据交换由线路控制部分承担。在通信线路和 FEP 之间放置 MODEM 或 DCE，而 LA 的线路接口就是 DCE 的接口。线路速度分为以下三档：低速（50位/秒～1.2K位/秒）；中速（1.2～9.6K位/秒）；高速（48K位/秒）。有的 LA 能同时适应高、低速工作。

（三）主机接口

FEP 与主机的连接方法有多种。如果现有主机不加以改造，那么一般就可象普通外部设备连接到数据通道那样来进行连接，即主机具有与外部设备控制器同样的面向数据通道的输入输出接口控制功能。与主机交换的数据，要在存储器内加以缓存处理。主

机接口对存储部分的数据具有读写功能。对存储部分的寻址方式有：经逻辑部分的方法；直接寻址法。后者适用于有大量数据交换的情况。

四、通信控制软件

图 12-23 示出了模块化的系统网络体系结构〔SNA（System Network Architecture)〕。其中，3705(或 3704）通信控制器用作 FEP 或 RP。主机提供的通信控制软件有 VTAM（虚拟远程通信存取法)；通信控制器提供的有 NCP（网络控制程序)。VTAM 与 NCP 之间的通信，应采用具有一定格式的数据。SNA 的基本思想在于：适应多种终端工作；采用的传输规程为 SDLC 型。

在 SNA 中，根据节点设备的功能可分为主机节点、通信控制器节点和终端节点。(1）主机节点是具有 VTAM、DOS/VS 的 370 系统，用以达到执行应用程序和数据库管理等多种目的。(2）通信控制器节点是具有 NCP/VS 的 IBM 3704 或 3705 通信控制器，用以负责通信线路的控制。此节点按存储转发方式工作。某通信控制器除了实现同主机和终端的连接外，还要实现同网络中其它通信控制器的连接，它是在 NCP/VS 管理下工作的。3705 可配置 240 K字节的存储器，最多可连接 352 条通信线路。3704的存储容量最多达 64K字节，可连接 32 条线路。(3)终端节点具有网络中的最小网络管理功能。例如 IBM 3767 通信终端，它可工作在 SDLC 或 BSC 方式。它可经非交换(租用）线路或交换（拨号）线路同 CCU 连接。因此，VTAM实现了一种层次式的树形网络。

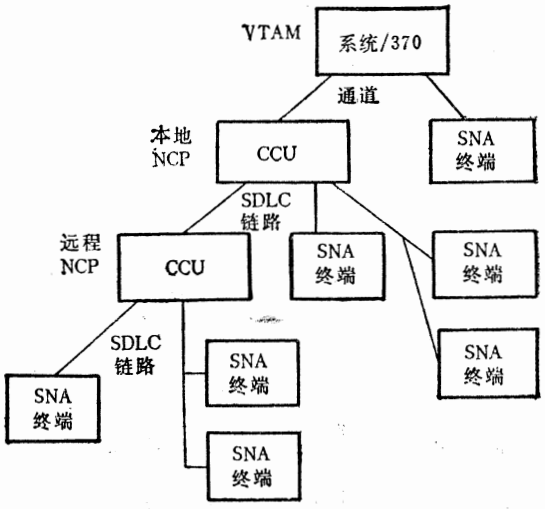


图12-23 SNA网络的通信控制

五、汉字联机网络

为了实现汉字网络通信系统，需在原 SNA 网上扩充汉字信息处理与通信功能。具体地说，必须解决汉字字符进入计算机网的问题，以实现汉字与西文兼容的汉字信息通信网络系统。

根据我国目前在几个系列机上实践的经验，一种简单而有效的方法是采用汉字联机仿真程序。这样就可以使一台经过通信线路同通用计算机（如 IBM4341）连接的汉字终端，能够有效地使汉字信息进入主机的联机编辑程序。于是，汉字终端就可以利用主机的各种软件资源处理汉字信息，并能将经处理后的内容在汉字打印机上打印输出；此外，汉字终端还可在通信处理机的控制下，进行网内计算机之间的汉字数据通信。图 12-24 是汉字联机网络的示意图。

具有汉字联机仿真程序的汉字终端远程适配器（或通信规程转换器)，可支持 BSC (或 SDLC) 通信规程。例如，汉字 3270 联机仿真器既能使一台汉字终端与主机兼容，并实现联机汉字通信，也可使网络内的用户充分共享汉字信息处理系统的资源。

在图 12-24 中，汉字终端经过汉字通信规程转换器（CPC）接入了 IBM 计算机网络。CPC 实际上是一台由微处理机构成的联机仿真器，内有汉字 3270 联机仿真程序。其功能主要是：进行规程转换与控制；预处理双字节的汉字国标码；进行传输代码转换；进行通信方式和速率转换；实现差错控制及数据缓存等。其主机采用探询/选择链路控制方式；其终端按报文中的设备地址响应主机探询并进行数据通信。CPC 功能详见表 12-5。

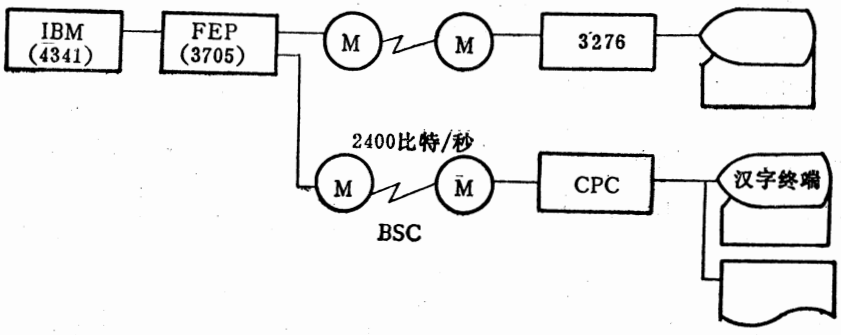


图12-24　汉字联机网络

表12-5　CPC 功能转换表

型　　号	IBM4341（主机）	ZD-2000（汉字终端）
通信速率	2400位/秒	9600位/秒
通信方式	双线半双工	全双工或半双工
同步方式	同　　步	异　　步
差错校验	CRC	字符奇偶校验
规　　程	BSC	无
控制字符	EBCDIC系列	专用控制符
代　　码	EBCDIC	GB-2312-80

12.3　汉字联机信息处理系统

联机系统系指输入数据不经过中间媒体（磁带、软盘、卡片、或纸带等）而直接经过通信线路进入中央处理机进行处理，处理结果又直接传送到用户的系统。

联机系统一般由以下三个部分组成：

（1）主机；

（2）通信前置机；

（3）终端网络。

联机系统主要有下列五种类型：

（1）实时控制系统。如军事指挥、航天测控、交通管制，以及一般工业控制所使用的系统。

（2）管理信息系统。如银行系统、订票系统、企业和政府机构的经营或事务管理系统。

（3）分时系统。它是多个联机用户同时使用一台计算机进行处理的系统。每个远离计算机的用户通过通信线路，用终端以问答的方式控制程序的运行，系统把处理机时间轮流地分配给各联机作业，每个作业只运行极短的一个时间片。如果在时间片结束之前处理还未完成，该程序就被时钟中断，等待下一轮再处理，此时处理机让给另一联机作业使用。这样，各用户的每次要求都能快速响应，给每个用户的印象是他自己独占该计算机系统。例如，大的科学计算中心或数据中心就有采用这种分时工作方式的系统。

（4）远程成批处理系统。它是一种把程序或数据从远程终端高速传送到中央处理机进行成批处理，并用联机方式将处理结果返送到远程终端的系统。

（5）数据采集系统。它用于各种观测、监视和实验测量数据的采集和记录，供事后分析处理。

由此可以看出，联机系统无论在军事或国民经济的各部门都有着广泛的应用。

目前在我国，由于对信息的采集、处理、存储、检索、交换的使用要求日益迫切，因此，在通用计算机上建立联机汉字信息处理系统的要求，也显得特别重要。实现联机汉字信息处理，对于提高工作效率和准确性，有着十分重要的意义。

联机汉字信息处理系统的组成与一般西文联机系统是类似的。但它主要是以使用汉字的用户为服务对象。因此，它的基本组成部分应该是汉字终端设备和具有汉字处理功能的主机系统。

下面具体讨论在通用计算机上实现汉字联机信息处理的主要途径。

12.3.1 汉字终端联机接口

在通用计算机上实现联机汉字信息处理的基本要求，是在应用程序一级上实现汉字的输入和输出。但是必须指出，对于通用计算机来说，仅仅作为这一点是不够的，这是因为在这类机型上有着丰富的软件资源，为了扩充汉字处理功能而妨碍或废弃这些资源的利用显然是不可取的。通用计算机所支持的汉字信息处理系统同微型机所支持的系统不一样，它具有庞大的汉字数据文件和大量的汉字应用软件。建立这些汉字数据文件和应用软件，对于程序人员是一个很大的负担，因为他们不得不花费相当多的精力去查汉字数据的代码。因此，对一个有效的汉字联机系统的进一步要求应该是：能够充分利用通用计算机系统的各种软件资源（如文本编辑程序，连接编辑程序等等）；能够在用高级程序语言编制的程序中直接写入汉字直接量和注释部分，而不致于破坏程序的可移植性。我们称这样的系统是汉字与西文兼容的系统。

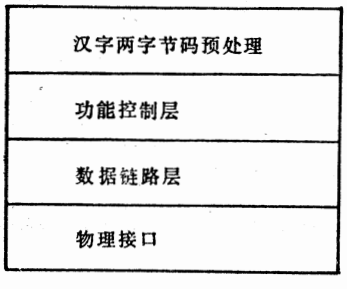


图12-25 联机终端与主机的界面

原则上，要实现这样一个汉字西文兼容的系统，就要求所引入的汉字终端在和主机的联机界面上同原有的西文终端兼容。这个界面可以用图 12-25 所示的层次关系来说明。

这个界面由下述四层组成：

一、物理层

在联机系统中，终端往往是经过调制解调器或主机的串行输入输出通道而进入主机。

的终端网络的。这种接口已经标准化，如 CCITT V24 或 EIA RS-232C 等。所要连接的汉字终端也应满足这些标准的要求。目前我国生产或研制的大多数汉字终端都具有这种接口。

二、数据链路层

数据链路是由终端设备和按特定方式工作以完成终端之间信息交换的互连网络所构成的整体。上述定义中的终端设备是广义的，在联机系统中，主机也看成是链路一侧的一种终端。其间的互连网络包括调制解调器，信道、通信前置处理机、主机的输入输出通道等部分，网络的工作方式则受主机的通信管理程序、通信前置机的链路协议等的支配。

联机系统与批量系统的主要差别在于增加了通信处理功能。因此，在它的系统软件中也增加了相应的处理模块(如图 12-26 所示)。

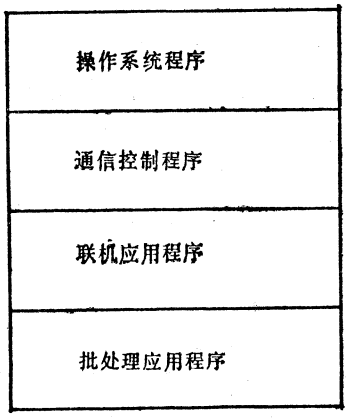


图12-26　联机系统的软件

主机的通信处理模块用来实现高级程序语言中的输入输出语句和终端的通信功能。这样，用户在编写应用程序时，不需了解通信的具体过程，而只要写简单的输入输出语句，就能完成远程信息的输入输出。这就是通用计算机系统为远程终端用户提供的各种远程存取方法，其中如 IBM 的 BTAM、TCAM、VTAM；UNIVAC 的 EXESⅡ；B-URROUGHS 的 MCP；DEC 的 COMTEX 等。

大型的联机系统有比较大的终端网络。在网络中，终端的处理要求是随机发生的。为了有效地管理多个终端的同时存取，必须制订一系列严格的通信协议（通常称为通信控制规程)。目前多数通用机采用基本型传输控制规程，这种规程是基于 10 个基本传输控制字符，即ISO 1745（见 12.2 节报文格式）来实现通信控制。为了减轻主机的负荷，通信规程的控制功能一般由通信前置机来实现。

不同的机型往往采取不同的通信控制规程，这种情况是由于两方面的原因造成的，（1）系统的复杂程度及应用环境的差别；（2）某些计算机厂家往往把这种规程作为保护其硬件资源的一种手段，用以防止别的外部设备生产厂的产品进入该机型的市场。

因此，要把汉字终端接入通用计算机系统的终端网络，首先要解决的问题是，在数据链路层能与主机兼容。

尽管不同的机型采用不同的通信控制规程，但作为一个完备的规程文本都应包含下列项目

（一）规程的适用范围

它必须指明：该规程是适用于点-点数据链路还是多点数据链路，采取何种传输控制方式，是争夺式还是探询/选择方式；通信速率范围；用全双工还是用半双工信道通信，以及具体的应用系统类型等。

（二）通信控制符

ISO 1745 所规定的 10 个基本传输控制字符已在表 12-2 中列出。这里的表 12-6 进一步列出 ASCII 及 EBCDIC 数据编码中的相应通信控制的代码。

表12-6 数据链路控制符

数据链路控制符	EBCDIC（十六进制）	ASCII（十六进制）
SOH	01	01
STX	02	02
ETX	03	03
EOT	37	04
ETB	26	17
ENQ	2D	05
SYN	32	16
ACKO/1	1070/1061	1030/1031
NAK	3D	15
DLE	10	10

对于这些基本传输控制符，在不同的系统中实施时，根据线路质量及系统的特殊要求，往往有一些差别。例如，有的系统采用双字符EOT EOT来表示传输结束，用DLE NAK来表示否认等。对于承认字符ACK的使用，有的规程要求用ACK_0和ACK_1交替应答等，这些都必须加以注意。

（三）同步方式

在数据通信系统中，使用异步和同步两种传输方式，但目前多数通用计算机采用同步传输。在同步传输方式中，实现串行传输的位及字符的同步，是用同步码组来实现的。在每次传输的开始，都必须传输同步码组使收发双方建立起位和字符的同步关系。

1．字符同步　一般用两个以上的连续的SYN字符来建立字符同步。

2．位同步　在接收字符同步码组以前，必须先建立位同步。对于采用内部时钟的数据传输设备（Modem），在字符同步码组前面必须发位同步码组，如两个连续的“5 5”字符（十六进制），用以提供16次“0”、“1”交替来完成位同步。

3．信息同步　在信息组的传输过程中，为了维持同步，往往要求在固定的时间间隔（如1秒钟）插入同步字符。这是必须注意的。

4．填塞字符　为了保证数据的正确传输而设置的。前填塞字符可与位同步码组合并，后填塞字符用以确保在数据传输设备的发送器关闭以前，使最后一个有效字符ETX BCC或ETB BCC等正确传输完毕。

（四）传输数据格式（或称电文格式）

大多数通用计算机所用的电文格式主要分成三种类型：

（1）信息电文；

（2）正向控制序列；

（3）反向控制序列。

在每一个具体的系统上，这种格式往往有很大的差别。在报文格式中，要特别注意透明及非透明传输格式问题，在透明传输方式下，对所传输的代码没有限制，全部数据链路控制符都可以作为透明数据传输，而不被误解为控制符。在透明传输方式下，数据链路控制符前面都要加DLE字符，才被看作是控制功能符。例如：

DLE　STX

DLE　ETB

DLE ETX
DLE SYN
DLE ENQ
DLE DLE
DLE ITB

透明数据块如图 12-27 所示。

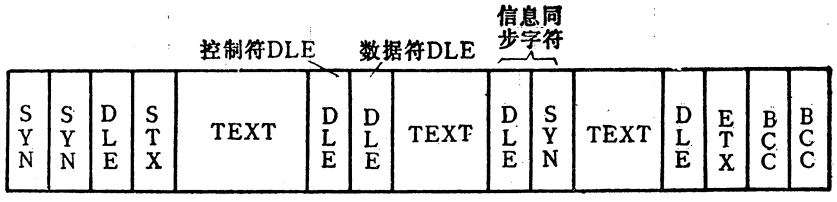


图12-27 透明数据块

（五）差错控制方式

在基本型传输控制过程中，除了字符的奇偶校验外，还采用纵向冗余校验（LRC）或循环冗余校验（CRC）来作为组校验。在了解一个具体的规程时，必须注意下列问题：

（1）校验的开始；
（2）校验的结束；
（3）哪些字符不参与校验计算；
（4）具体的校验计算方法，如 LRC、CRC-12、CRC-16 等。

一种有代表性的 BSC 规程的 CRC 校验计算范围如图 12-28 所示。

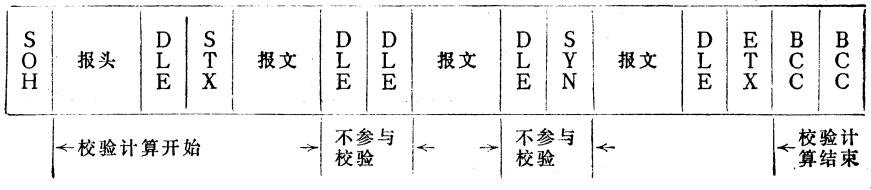


图12-28 CRC校验计算范围示意图

（六）传输控制顺序

它规定了在数据传输的不同阶段，收发双方信息来往的顺序。在规程文本中，通常用流程图或状态转移矩阵来表达。

IBM 3276 对探询的响应流程图如图 12-29 所示。

当缺乏有关传输控制顺序的完整资料时，需进行必要的测试。

（七）超时处理

为了发现及确立系统故障，以及确定这些故障的性质是偶然的或是固定性的，在很多规程中规定了时间越界值。当发生时间越界（超时）时，就进行询问或重执。用计数

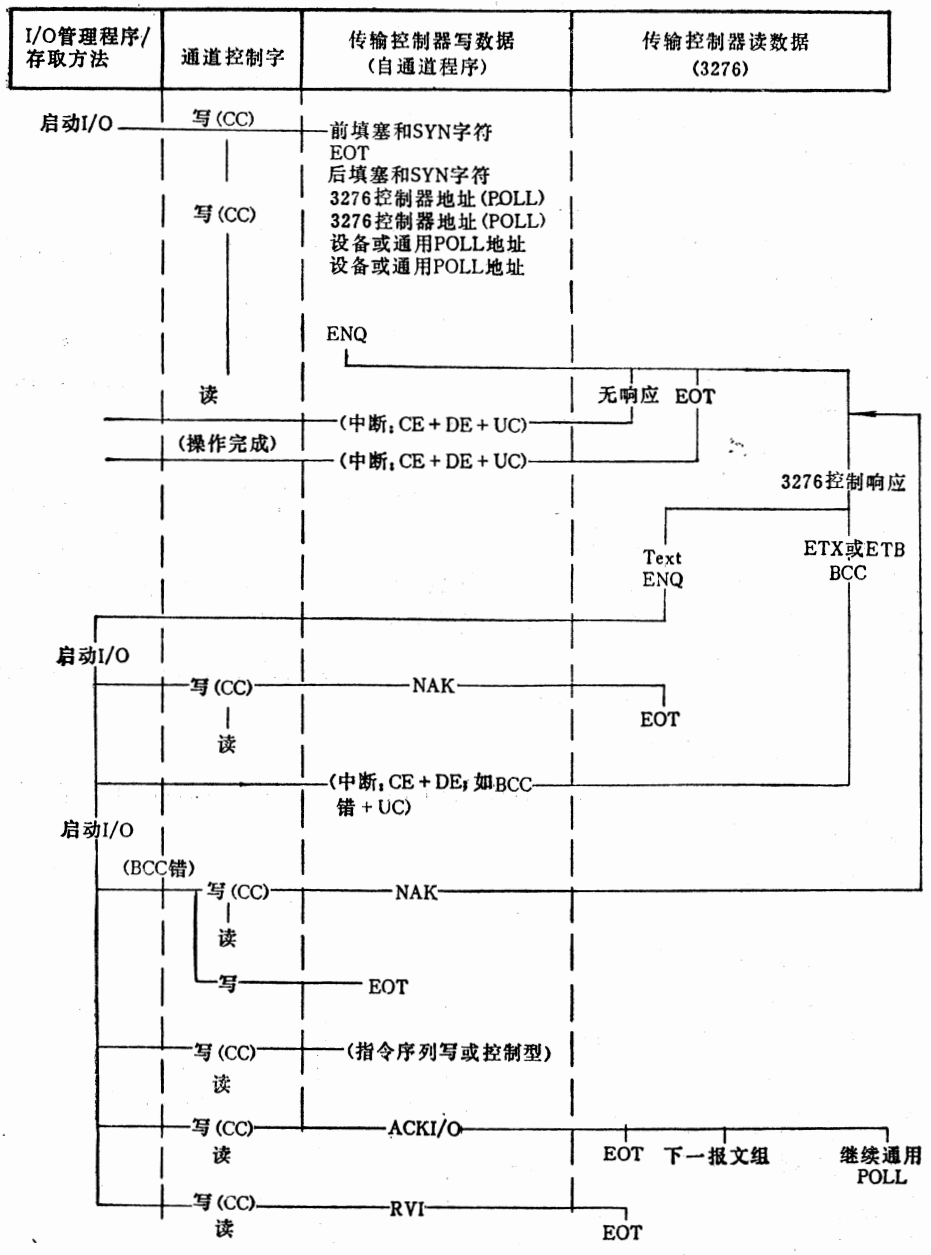


图12-29 IBM3276探询响应流程图

CC—(Chan Command) 链指令；CE—(Channel End) 通道结束；DE—(Device End) 设备结束；UC—(Unit Check) 单元检查。

器来记录询问或重执的次数，当重执超过规定次数后仍未成功，就作为固定性故障处理。

计时器的越界值和重执次数的规定是随不同的终端设备、不同的应用场合及线路质量而变化的，变动范围较大。

在一种 BSC 规程中规定了下列五种时间越界值：

等待应答计时	3 秒
等待报文计时	25秒
信息同步计时	3 秒
等待确认（WACK）发信计时	2 秒
报文暂时延时（TTD）发信计时	2 秒

由于目前大多数汉字终端只能提供异步通信接口，故实现汉字终端联机的关键是实现同步数据链路接口。其具体的作法是：

（1）在汉字终端上扩充同步通信接口，并开发针对某一规程的仿真程序。

（2）采用规程转换器，在终端与主机之间进行规程转换（包括同步方式、代码、速率、传输控制顺序、差错控制方式和数据格式等）。

三、功能控制层

数据链路层连通之后，实现数据的输入输出便不成问题。但是要使汉字终端能有效地联机工作，还必须要在功能控制层与主机兼容。这里所指的主要是屏幕编辑和设备控制符的转换问题。在一般情况下，主机和汉字终端所使用的编辑控制符是不一致的，要形成一定的屏幕格式，就需要在两者之间建立起一一对应的关系。设备控制主要是指终端的键盘功能的定义。对于某些机型，在汉字终端上难于实现全部对应的功能，这一点在联机汉字终端选型时要特别加以注意。

四、汉字双字节码预处理

从信息的传输和处理的角度来看，汉字与西文的主要差别在于汉字用了两字节或多字节编码。下面分别加以说明。

（一）汉字数据通信

汉字数据的传输和西文数据的传输本质上是没有差别的。作为信息交换用的汉字国标码（GB 2312-80 码）的两个字节，都安排在 ISO 646 码的图形字符区，共 94×94 个字符，它和 ISO 1745 的基本传输控制符是相容的。如图 12-30 所示。对于使用EBCDIC字符的机型，由于其控制符区和汉字码部分重叠，就会发生二义性问题。在这种情况下，要采取一定的解决办法。一是移动汉字交换码的码区，例如在 GB 2312-80 码上恒定加上十六进制数 8080（见图 12-31）。另一种解决办法是采取透明传输，即利用转义字符 DLE 来标识数据字节中出现的控制符。

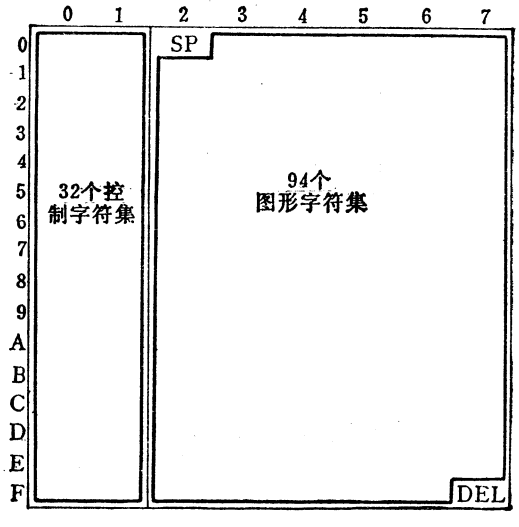


图12-30 ISO 646码区图

（二）汉字数据处理

汉字数据在计算机内部的表示一般有三种方式：

（1）两字节码：如GB2312-80 码或其它代码。

（2）三字节码。

（3）四字节码：如标准电报码。

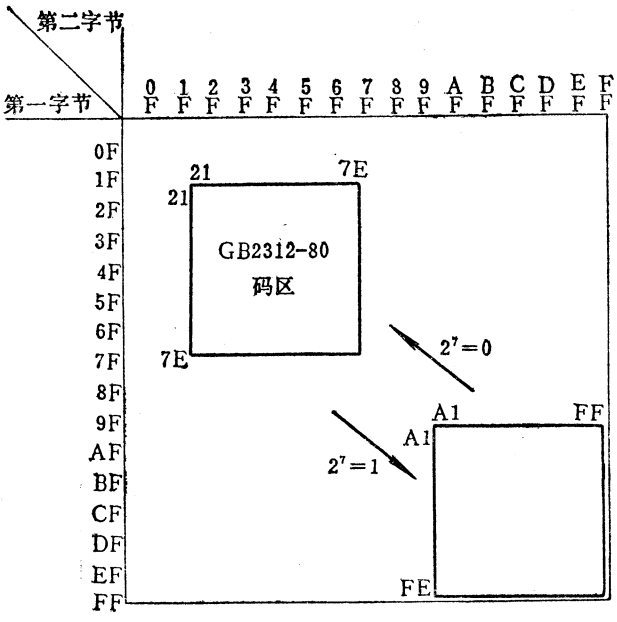


图12-31　码区平移示意图

在计算机内部，汉字数据代码的处理和西文无根本区别。但作为人机界面的汉字显示终端，在区分汉字和非汉字时，主要根据数据中的标识符来判定。例如国产 ZD-2000 汉字终端，采用了 ESC U 和 ESC V 来标识汉字和非汉字（见图 12-32）。由此不难看出，联机汉字终端只要能提供适当的标识符，便能在通用机上扩充汉字功能。对于三、四字节码，码组本身具有自名标识的作用。

在通用机上实现汉字处理，主要是指用高级程序语言如 COBOL、FORTRAN、PL/1 等所完成的汉字数据处理。具体地说，就是在高级程序语言编写的程序中写入汉字注释量和直接量，以及实现对汉字数据的传送、比较、置换、字符串处理、输入和输出、分类和合并等操作。

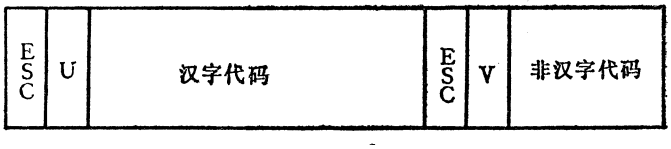


图12-32　ZD-2000汉字终端的汉字标识法

上面已经提到，要使汉字数据进入高级语言源程序中的注释量和直接量部分，就是要选择一种能为语言的编译系统所能接受的汉字标识符，同时这种标识符应具有唯一性，即它不应和汉字数据的任意组合发生重码。此外，还必须妥善解决汉字数据的某些部分同高级程序语言的字符常数的分界符撇号（' '）、或双撇号（" "）的重码问题。只要解决好这两个问题，就可以在高级程序语言中扩充汉字处理功能。

汉字数据的处理和西文也是类似的，但对于并置运算，由于采用了标识而造成二义性，这只能通过使用多字节代码来解决。此外，汉字的分类也具有很多特性，这要靠汉字处理实用程序来解决。

12.3.2　汉字终端远程适配器

根据上述的联机接口要求而设计的汉字终端远程适配器，它利用微处理机在汉字终

端和通用计算机之间进行多级联机仿真和转换处理。适配器提供双向的联机通信接口和必要的通信缓冲区。适配器的软件设计概念可以用图 12-33 来说明。主控制程序完成通道和工作单元的初始化，以及各个功能处理模块之间的调度。规程控制程序和功能码转换程序实现数据链路层与功能层的接口，汉字双字节码预处理程序完成汉字数据进入主机的高级程序语言所必要的预处理。

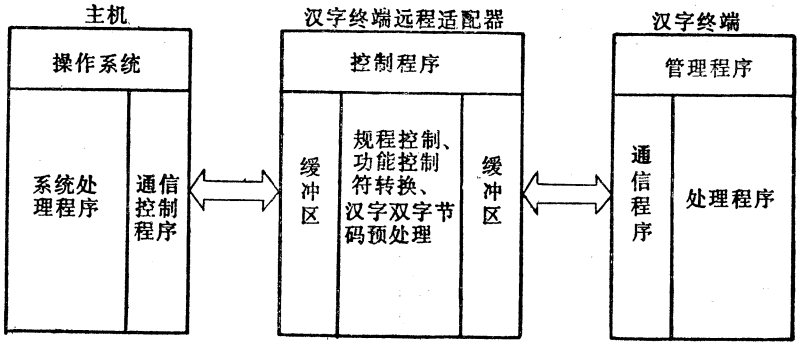


图12-33　汉字终端远程适配器概念图

12.4　汉字微型机局部网络

12.4.1　概　　述

近年来随着局部地区网络（Local Area Network）技术的不断发展，国外出现了许多商品化的微型机局部网络。我国以微型机为基础的汉字智能终端，在很多领域内开展了应用。目前国内用户也开始使用汉字微型机局部网络，以满足现代化管理的需要。

汉字微型机局部网络同普通微型机局部网络并无本质的区别。可以用同轴电缆或双绞线，将多台汉字终端加以连接，以实行相互间的通信和软、硬件资源共享。只要将双字节汉字作为字符串处理，在帧格式的数据段中用适当的标识码区分西文和汉字，将各汉字终端经过通信规程转换到 HDLC 高级链路规程，并实现网络软件的汉字化，便能实现汉字局部网络。因此，实际上仍然是使用一般微型机局部网络技术，来实现汉字微型机局部网络。下面将分别作简要介绍。

局部地区网络通常是某个单位所拥有的一种数据通信系统。它能使许多相同或不同的数字设备在共用的传输媒体上互相对话。它可在下列设备间实现通信：主机、小型计算机、微处理机、文字处理机、个人计算机、智能终端、工作站、打印机和磁盘驱动器等。

局部网络是在有限的地区范围内提供这种通信的。如：一层楼面、部分楼区、整个建筑、建筑群或工厂综合企业等，其通信距离可从几百米到几公里不等。数据通信速率可从 0.25K位/秒到 50M位/秒。它比在电话交换线路上的远程通信速率要高，而比许多近程计算机的输入和输出总线速率要低。总之，局部网络的通信距离大于输入和输出总线，而小于远程通信线路。

局部网络出现于七十年代末。由于微型机的成本不断下降，故局部网络能提供价格

低廉、使用方便的计算机或智能工作站，同时也能在用户中间共享昂贵的计算机软、硬件资源。

局部网络在数据处理及文字处理方面的有效应用，使它成为在八十年代的一种主要计算机通信技术。据统计，局部网络在1983年有13万个节点，预计到1990年将增加到近120万个节点。为满足这种急剧增加的需求，许多厂家开始生产各种性能价格比的局部网络系统。这些局部网络有各种网络拓朴结构。有些厂家只提供网络接口，如控制器（或网络接口装置）和收发器。使用这些设备和一定标准的通信规程，可将各种计算机连接到通信线路。目前还没有完全对局部网络实行标准化。实际上按性能价格比可将它们分成如下三类：

（1）使用双绞线作为传输媒体的低造价、低性能的系统（低档系统）。如美国的OMNINET、DESNET和CLUSTER/ONE MODEL A等。

（2）使用屏蔽基带同轴电缆作媒体的中等造价和中等性能的系统（中档系统）。如美国的ETHERNET、NET/ONE和Z-NET等。

（3）使用宽带或基带屏蔽同轴电缆作传输媒体的高造价、高性能的系统（高档系统）。如美国的CableNet、Domain、Ring Net和Wang Net等。

大量研究表明，办公室全部通信量的60%是在机关内进行的。局部网络可使办公室自动化大大推进一步，以促进当前业务处理中打印和人工传送等方面的自动化要求。例如，用于零售点业务、财务、企业管理、过程控制、CAD/CAM人机会话、科学计算，以及科研和实验室等等。

局部网络有下列特点：

（1）较高的数据率。可进行快速短时间传送和多站访问

（2）一定的地区覆盖。局部网络在机关各部门间提供快速通信并能为各种设施服务。

（3）传输误码率低。网络能可靠地适应大量的通信业务，能检测出现的差错并加以恢复。

（4）网络连接与安装造价低。在不需改变正常工作条件下，用户可自由地在网络内安装或拆除工作站。连接费用不超过工作站造价的(10～20)%。

（5）可为大量的用户服务。网络能支持上百个用户并有扩展余地。

（6）可靠性与可用性较好。个别故障不影响网络工作并能顺利排除。

（7）保密性较好。用户能限制对机密文件数据的网络访问。

（8）具有灵活的拓朴结构。用户能按机关扩充需要，修改网络结构。

（9）可利用多种媒体通信。网络能实现话音、视频和数据通信。

（10）具有多种产品的兼容性。网络内可包括各厂家的设备。

IEEE对局部网络提出的功能要求是：

（1）传输线长度不应小于1公里；每个电缆应能处理1M位/秒～20M位/秒之间的数据率；网络至少容纳150台设备。

（2）标准情况是应尽可能具有中等的独立性（独立于传输线等）。

（3）网络应当可靠工作，未检测出的数传差错每年不应多于1次。

（4）网络应提供同等通信，任何设备应能互相对话而不使用中间设备。

12.4.2 局部网络工作原理

一、局部网络的工作原理

下面我们将研究一下 IEEE802 委员会建议的具有冲突检测的载波监视多路 访问 技术（CSMA/CD）的总线/树状网络、标志探询网络以及标志环状网络的工 作原理。

图 12-34 示出局部网络的拓朴结构。星状结构主要用于中心控制的 PABX；总线/树状结构则多用于分布处理系统；环状网络既有中心控制式又有分布式。

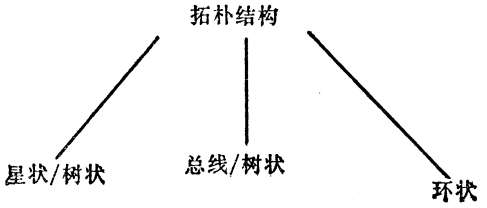


图12-34 网络拓朴结构

对传输媒体的访问方式有控制式访问（包括探询式，如标志通过）和随机式访问（如CSMA/CD)。在控制式访问中，由于采用了探询，故减少了竞争。在标志通过(token passing)探询中，控制受标志支配，即标志从一个用户传向另一个用户。凡具有标志的用户，可利用标志实现数据传输。最常用的随机式访问是所谓载波监视多路访问——冲突检测技术。使用此项技术的用户，只有当媒体空闲时才传输信息，一旦出现冲突，便停止传输（如图12-35所示)。

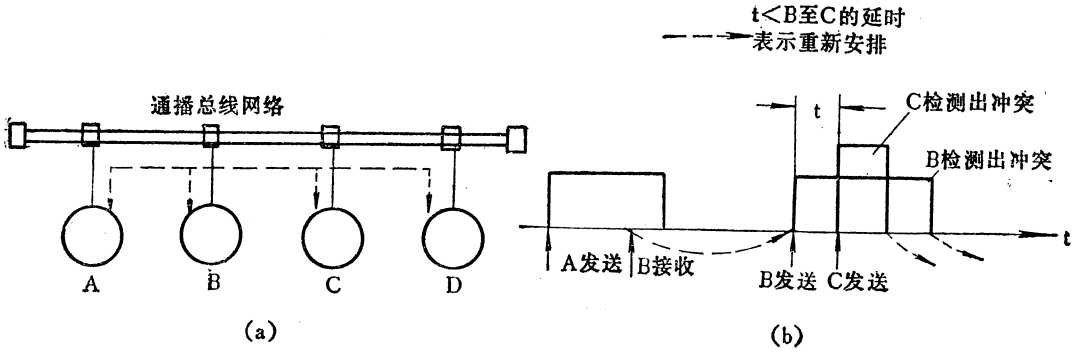


图12-35 CSMA/CD和标志探询的工作原理
（a）CSMA/CD；（b）标志探询。

环状网络（如图 12-36 所示）的全部链路构成了环状。对于信息交换网络(图12-36 a)，在给定的信息交换期发送报文：A向B发送，B向C发送，C向D发送，D向A发送。对于标志环状网络（图12-36 b)，如果A具有标志，就向环状网络发送报文，随后，A将标志送往B。如此循环进行。

网络中的关键设备是网络接口装置 NIU （或网络控制器)，它将用户计算机（或终端）与同轴电缆总线相连。如果终端打算访问其它节点，NIU就发送数据。如果目的地址的NIU收到此报文，它便用确认方式进行响应，而其它NIU就不响应。

图 12-37 示出几种选择NIU的方案。目前提供的大都是带RS-232-C接口的低速集群式终端，一台NIU连接 2 ～16个用户（见图12-37 c)。因此，NIU也作为终端集中器工作，以便有效地使用其带宽。

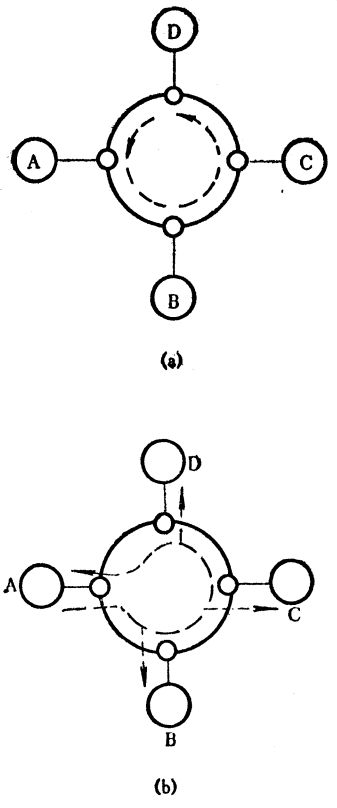


图12-36　环状网络

（a）信息交换环状网络；

（b）标志环状网络。

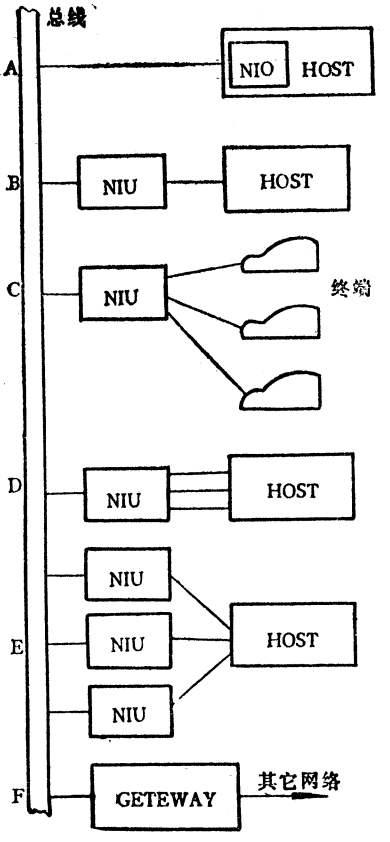


图12-37　NIU在局部网中的配置

图中A示出主机中I/O通信软件实现NIU功能。D型独立式NIU经过多端口连到主机并带有低级链路规程。B型是一种单路连接。在E型中大型主机的每个端口用一台NIU。总之，NIU的用途是完成与网络有关的功能〔例如物理层的访问、寻址和链路层的差错控制，非分组式终端的数据包装拆（PAD）功能、信息流控制和网络层上的内部连网等功能〕。这些功能对于低速终端的NIU是足够的，但不能提供主机-主机内部连接的较高级功能。

NIU的结构体系与规程功能有关。它可实现大部分低层功能。多数环状及树状NIU由终端接口装置(TIU)和收发器(TRU)两部分组成。

TIU是用户端的接口；TRU是供网络端使用的。如图12-38示出了一台NIU的组成。它将12台用户终端与带同轴电缆总线的网络相连。NIU包括3块TIU板和一块TRU板。每个TIU供4个终端接口使用。每个TIU中有自己的CPU。例如，一台8位NMOS微处理机，其最小指令周期为1微秒。TIU与TRU通过一条主总线相互连接，总线速率与同轴电缆的数据传输速率相同（约1～10M位/秒），TIU向TRU的传送按DMA方式进行。

NIU的构成随着设计不同而有差异。它大致可分为A、B两种类型。

A型NIU：TRU板上有CPU和存储器，它控制往来TIU的DMA传送。此外，还有

作为 CPU 支持的总线控制器的大规模集成电路芯片。

B 型NIU：TRU板上没有CPU和存储器。来往TIU和DMA传送是由TIU和 适当的仲裁逻辑控制。格式化功能如差错校验等，均由MSI硬件完成。

因此，TIU和TRU之间的规程功能是按以下方式分工的：在A 型NIU中 TRU 完成全部物理层、链路层和某些网络层功能，而TIU只完成数据包的装拆；在B 型 NIU 中，TRU多半完成物理层功能，而不是链路层功能。如果TIU和TRU工作在排队网络模式，随着TRU CPU链路层规程处理业务量的增加、A 型NIU会饱和。当多个TIU访问 TRU 流水线进行竞争而使业务量加大时，B 型NIU也会达到饱和。在A 型NIU中，链路层功能在串行工作的TIU和TRU之间分开实现，而B 型NIU并行处理，其中 TRU 作为收发器工作。

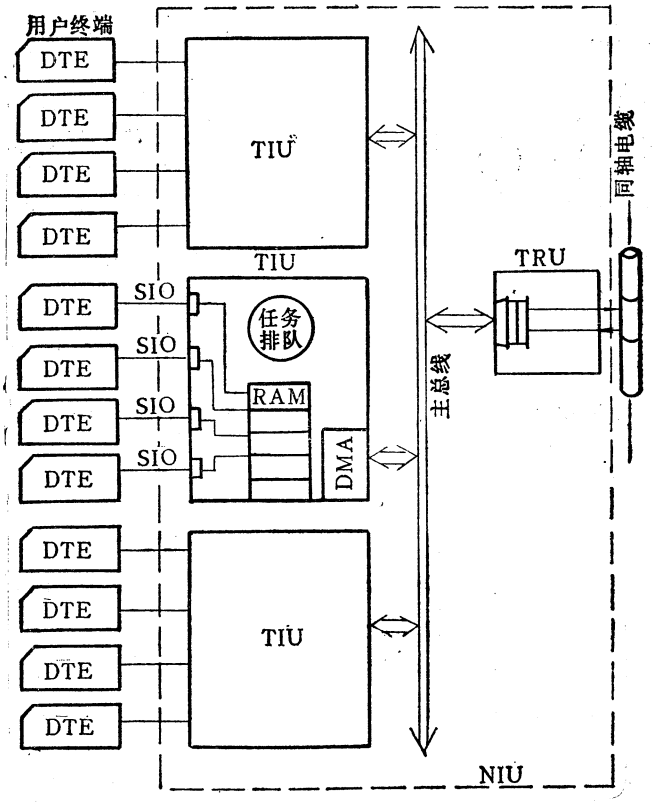


图12-38　网络接口装置NIU的基本框图

二、局部网络控制器的组成

控制器是将计算机（或用户终端）和网络连到一起的设备。它由一台微处理机（或专用芯片）、EPROM、RAM和一个高级链路控制器芯片组成（见图12-39）。每台控制器有两个与外部相连的接口：一个与计算机（或终端）连接；另一个用于连接网络的同轴电缆。串行接口 RS232-C 用于连接显示器和串行印刷机等，并行接口主要 用于连接用户计算机。

EPROM内存有控制程序，而 RAM 为网络中的收发数据提供缓冲区。ADLC 用于

使网络收发数据格式化。当传送数据时，ADLC从计算机数据总线接收并行数据，并转换成串行同步格式向网络传送。从网络接收数据时，ADLC把接收数据加以格式化，并送到数据总线，然后送给计算机。数据按高级数据链路控制规程(HDLC)分组传输。网络使用的HDLC数据帧格式如下：

标志	目的地址	源地址	数据	帧校验	标志
8位	8	8	$8n(n>0)$	16	8

起始标志、循环冗余码校验及结束标志都由ADLC产生。冲突检测电路是将ADLC送来的数据与网络读回的数据进行比较。如两个数据不同，就产生IRQ中断信号通知计算机发生了冲突。

控制器软件用来管理下列一些数据：接收网络的数据；发送到网络的数据；接收计算机的数据；送往计算机的数据。

当ADLC以中断方式通知处理机它已接收到网络数据包的起始标志时，就执行接收网络数据的程序。如果数据包的第1个字节地址等于控制器本身站址，就存储第一个字节以后的信息，然后再传送给接收设备。如果数据包地址与站地址不符，则去掉包的其余部分。当接收完一个包时，就检验ADLC的状态位，以确保没有任何CRC差错或其它错误。如果检测出错误，就放弃该数据块。

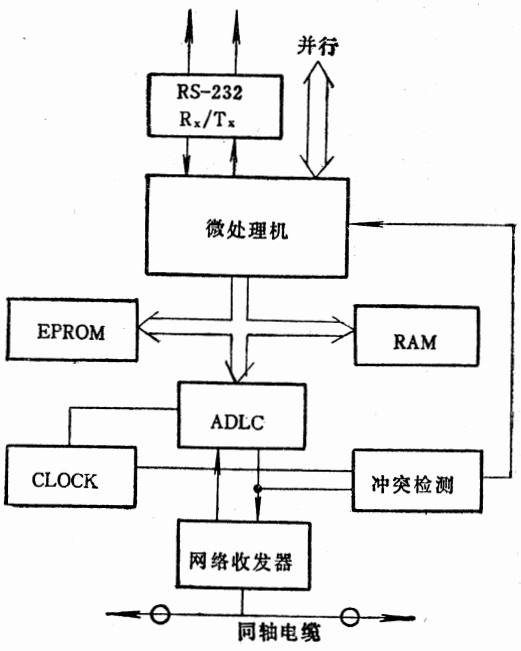


图12-39　控制器框图

当向网络传送数据块时，先检查网络是否处于忙态，如处于忙态就要等待，此时ADLC提供一状态位表示网络忙或闲。一旦网络处于空闲状态，发送程序便连续不断地将数据送给ADLC，并将这些数据转换成串行格式。

如果在发送过程中检测出数据冲突，便进入中断程序并对其进行处理。中断程序关闭网络一段时间，再等待某个随机时间才能重发数据包。

如果用户计算机要发送数据块，则当此数据块准备好后，便向控制器发出信号。此时，计算机必须等待，直到控制器发出回答信号才开始发送。同样，当控制器从网络接收数据块并把它发送给计算机时，也使用类似的信息交换协议。

如果用户只有一台显示器或行式打印机与网络连接，则使用控制器软件管理数据的收发。接口控制软件一般用汇编语言编写。

12.4.3　三种局部网络

一、CLUSTER/ONE MODEL A低档局部网络

它是八十年代初推出的廉价、低性能的个人计算机局部网络。它用于办公室自动化系

统、工程与软件开发、教育、旅行行业、地产业务和个人计算机的分布处理。

网络由连接65台AppleⅡ计算机用的扁形16线或15线圆形屏蔽电缆组成。这种连接是经过插入式网络接口板来实现的。每个板上有AppleⅡ接口电路、一个总线收发器、存有ISO通信软件的ROM(2K字节)、1K缓存存储器和8位地址和识别逻辑。为了降低成本、接口板没有使用微处理机或复杂的控制电路。其它一些接口板插到扩展槽上且与打印机、MODEM及软硬盘等外部设备相连。在300米内可组成任意网络，如星状、总线或树状网络。

每个工作站可作为服务器(server)或用户使用，但两种不能同时使用。

作为用户工作时它是独立工作的。此时使用8位6502微处理机、64K字节存储器、键盘、显示器和5英寸软盘。

作为软件控制下的服务器工作时，它能为用户提供网络共享。在网络工作期间，每个站都能共享与更新信息。为了数据保密，控制系统能保证其它站不能更新同一信息。一台多级口令保护机构使得只有指定的用户才能访问专用文件。每个站可按240K位/秒的最大速率将并行数据直接送给另一站，并可共享网络中心的磁盘资源。此存储设备是由数据录入用的1.2M字节软盘到大于40M字节的硬盘。数据以1～256字节变长度数据包的形式，按8位在单个基带通道上并行传送。这种格式安排要比串行电缆网络的吞吐量大8倍。

网络访问控制使用带回避冲突的CSMA技术。向网络内发送信息的工作站，首先在单独的控制线上检验载波信号的存在（而不象以太网络中检验数据的存在)。如果网络不忙，工作站就发送数据包报文。该报文由目的站地址、发送站地址、报文长度、变长度数据和16位校验和组成。

目的站计算发送报文的校验和，并将它与接收的校验和进行比较。根据结果，在载波信号未起作用以前，收站立即向发站发送ACK（肯定）或NACK（否定)。此法有立即确认的优点，减轻了冲突检测方法中有许多分开的短ACK/NACK报文的缺点。

所有站都能运行Apple DOS或Apple PASCAL系统，并透明地访问中央文件。网络文件服务台能运行操作系统以及维护与测试磁盘子系统的实用程序。

二、以太网络(ETHERNET)——一种中档局部地区网络

包交换的ETHERNET是美国Xerox公司研制的中等造价及中等性能的局部网络。它使用基带同轴电缆通信，并能定义ISO开放型互连网络基准中的6个层次。在美国已有二千多台工作站与此网连接。它是一种面向总线的通信系统。以太网络主要用于办公室自动化、分布数据处理及终端访问等，适用于高数据率的间断通信业务，但它不适于实时响应系统或通信量大的业务。

该网络在500米距离内能支持100个工作站，全部连接的工作站可在链路级上交换数据。任意双站间的最小距离为2.5米，点-点链路的最大距离为100米。网络由收发器、控制器、基带同轴电缆、接口电缆及工作站（计算机式终端）组成。图12-40示出了连到网络上的工作站，它使用10M位/秒的带宽，并用50欧姆同轴电缆作总线。计算机通过收发器和控制器同公用同轴电缆连接，以实现各计算机之间的通信。

传输子系统由公用基带同轴电缆、终端连接器、收发器和收发器电缆组成。收发器在控制器和共用同轴电缆间收发编码信号和提供一定的电气隔离，在收发器电缆和同轴

电缆之间提供电平转换。

控制器是工作站的接口。它有串-并和并-串转换、译码、缓冲、目的地址识别及CRC产生与校验等功能。

网络使用曼彻斯特相位编码技术传送数据。其基带方案使用带冲突检测的载波监视多路访问技术(CSMA/CD)，链路控制采用统计竞争方式。以太网络的物理层及链路层协议为高层协议提供一种有效的数据包传送系统。

为发送一个数据包，工作站要等待网络的空闲。当网络空闲时，才发送数据包。此时，工作站测定是否与其他发送器发生冲突。无冲突时，工作站便获得网络使用权。如果发现冲突，工作站发送4～6个附加数据字节、并放弃该数据包的发送。此时需重新安排发送计划。

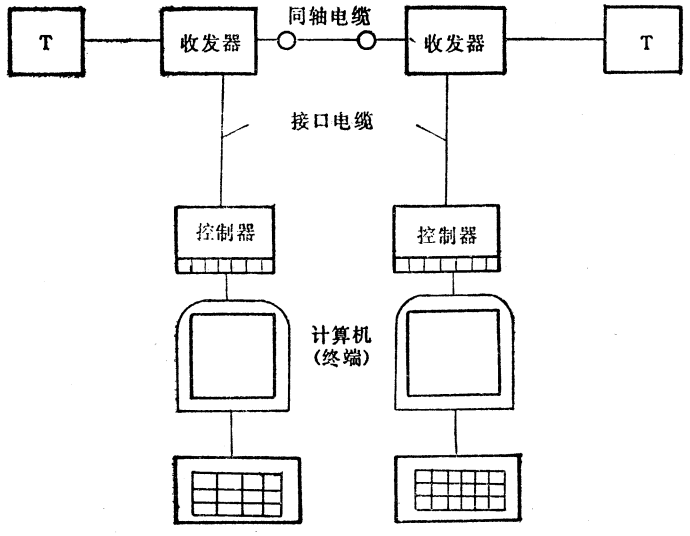


图12-40　以太网络结构

报文格式为可变长数据、它使用64位前同步信号(preamble)、48位目的地址、48位源地址、16位类型字段(type field)和1500字节的数据区，32位的CRC，用于整个区(但preamble除外)。因此，理论上48位源地址可提供281万亿个工作站工作。

以太网络的突出优点是：结构简单；造价低廉；易于安装及使用；通信速率高；延时较小；它有较大的访问灵活性，某个站出现故障并不影响整个系统工作；用户可随意在同轴电缆的任意点上增减工作站，而不会打乱网络工作。

以太网络可通过专门收发器和控制器接口提供不同终端的兼容性。它使用数据图表(datagram)概念在多站访问中间传送报文。由网络软件提供数据图表，以实现面向报文包的智能设备同时与任何数量的其它这类设备通信。所以，数据图表为智能设备与网络中任何地点的任意一批其它智能设备提供交换报文包的手段。以太网络的缺点是：传输速率固定；数据链路不能支持工作站的优先级；无数据加密措施等。

三、WangNet——一种高性能高造价的高档局部地区网络

前几种局部网络不能处理数字化的话音、视频及图象，也不能传送快速检索文件或

大数据块。但是，高性能、高造价的局部网络则可在宽带屏蔽电缆上完成数据，话音和视频信息处理。它的主要特点是：轻便；采用模块化；具有办公室及工厂通信应用的灵活性。

美国王安公司研制的Wang Net，使用了高价、高性能的宽带同轴电缆技术，能适应话音、视频、传真和大量数据处理。它使用频分多路(FDM)技术，以便将340兆赫的带宽分成只占有用频谱35%的三个独立的带宽：一个组成点-点数据通道的内部连接带宽；一个12M位/秒分时通道；用于仪表测量和监视的实用带宽。这就远远超过了双绞线技术（最大2M位/秒）及以太网络(10M位/秒)。

此网络是一种使用商用CATV电缆在600米距离工作的多点树状拓朴结构。网络按并行的双电缆系统工作；一路用于发送；另一路用于接收。

对于数据通路，内部连接通带提供一组3个通道。第一组为32个专线的通道工作在10～12兆赫，在数据率达9.6K位/秒的RS-232C设备之间提供全双工多点透明式或点-点通信。第二组是在12～22兆赫上为数据率达64K位/秒的RS-449设备之间提供同样的通信。以上两组都要求能提供每个用户设备的通信端口，以及一台晶体控制的固定频率调制解调器。用这种固定通道分配通带不需要主控制。第三组是交换式。它在数据率达9.6K位/秒的RS-232C设备（RS-366自动拨号设备）之间，能在48～82兆赫为全双工和点-点交换的透明通信提供256个通道。在第三组中，每个用户的设备的通信端口有一台频率可变的调制解调器，每个Wang Net系统有一台数据开关装置(dataswitch用于按用户要求管理调制解调器各通道的频率分配)。

Wang Net能在217～251兆赫上连接65535台计算机、数据处理及文字处理系统。它以最大2K字节的可变长HDLC数据包在虚拟线路上用12兆位/秒速率传送数据。

Wang Net网络访问使用改进的以太CSMA/CD技术，能通过一台电缆接口设备(CIU)提供全差错恢复和信息流控制。每个王安系统需要一台连接到同轴电缆上的CIU。CIU有一台带有128K字节存储器的Z80控制器，完成数据包的装/拆、差错检测/校正及数据缓存。

王安网络系统允许在所有相连的系统进行资源共享，并能通过WANG电子邮政系统进行文件传送、远程编辑和报文分配。

Wang Net还利用174～216兆赫作为实用通带，以支持7个6兆赫的视频通道。每个通道能适应用户视频设备的合成彩色视频及音频信号工作。

12.4.4 汉字微型机局部网络

一、概述

为了适应我国机关、工厂及商业各部门推广办公室日常事务处理的需要，建立一种汉字微型机局部网络已是当务之急。为了保障办公室汉字信息处理自动化，即在日常的大量事务处理中用微处理机实现汉字编辑、报表、统计、查询、存档、管理、计算、打印及检索。在楼内各办公室之间用局部网络将多台汉字终端设备连接起来，实现计算机资源共享和各个业务部门之间的业务联络和通信。

在第二节中阐述的局部网络是我国开发汉字微型机局部网络的基础。所不同之处在于汉字微型机局部网络应具有汉字系统管理软件、高级语言处理汉字能力及相应的应用

软件。此外，还要配有有效的汉字输入和输出手段及汉字字模库。

具体实现途径可通过长度为2公里的同轴电缆或双绞线将多台汉字终端加以连接。在建筑群或同一建筑内各办公室之间传输汉字信息并共享系统中的软硬件资源。其性能价格比要优于远距离计算机网络，其费用比每个办公室配置一台独立的汉字系统要低。

汉字局部网络的拓朴结构可以是分布式、集中式或混合式。它应具有中西文兼容性，即既能传输西文又能进行汉字通信。为此，在网络中应有多台控制器作为网络接口，它可以处理西文代码和双字节的汉字代码，并能满足我国汉字通信规程和信息交换用汉字控制字符集的要求。如果网络内进行不同类型的汉字终端通信，还需要进行预处理，如对控制码进行转换等。

汉字局部网络中可能有三种站。第一种是产生和接收数据的用户终端站。第二类是共享数据的终端站。它能为几个用户终端提供对磁盘和打印机的相互访问。第三类是处理不同型号主机之间的通信用的通用控制器。

每个用户处用一台收发器，用户可将上百台汉字终端连成几公里的电缆网络，并在基带数据通道上传输数据。网络采用HDLC面向位的数据格式。多站访问可使用以太网络的CSMA/CD方法。

在典型的网络节点中，有一个带微型机的网络接口电路（控制器）和一块存储器板。网络接口电路可使每种终端与网络连接，并完成速率和代码转换，以及为使非兼容性设备通信用的规程转换。它提供数据通信链路层规程（含有地址检测、数据包格式形成和CRC校验）。因此，用户可将一台显示终端、一台打印机加到上述任何一种终端用户，而不需要附加硬件接口。

在第一种用户终端站中，节点处理机也可作为主计算机工作，它具有操作系统、实用程序和COBOL高级语言。

第二种共享数据终端站支持多用户的数据共享访问。此节点处理其它节点的请求并带有联机程序库。其特点是数据与磁盘资源共享以及直接报文存取。

第三种终端是一台通用控制器。它提供若干RS-232C串行接口和8位并行打印机接口，以便把型号不同的设备连接到网络。

二、典型汉字局部网络

下面将介绍一种集中式与分布式相结合的汉字局部网络。在某机关大楼内配置的汉字局部网络中，有一台带有汉字数据库和高速汉字打印机的高档微型计算机作为主机。按局部网络总线拓朴结构，将上百台节点的汉字智能终端用2公里长的基带同轴电缆加以连接，构成如图12-41所示的网络。

设在各个办公室内的用户汉字智能终端，具有独立工作能力，可进行汉字输入、编辑、打印及各种报表制成等事务处理。除了各智能终端之间可相互进行汉字通信外，各用户站都可共享主机软、硬件资源。日常性文件经处理后存在终端用户软盘，较重要的全局性文档资料（如：全机关人事档案）放在主机数据库内，而用户站可对它们进行检索或更新。

主机中的汉字字模库具有上万个汉字，智能终端处的汉字字模库一般在6～7千字。主机还可通过调制解调器（电话线路）与上级机关的主机或高级计算机网络连接。

本网络既有各业务部门之间的分布处理与相互通信及数据共享能力，又有与主机之

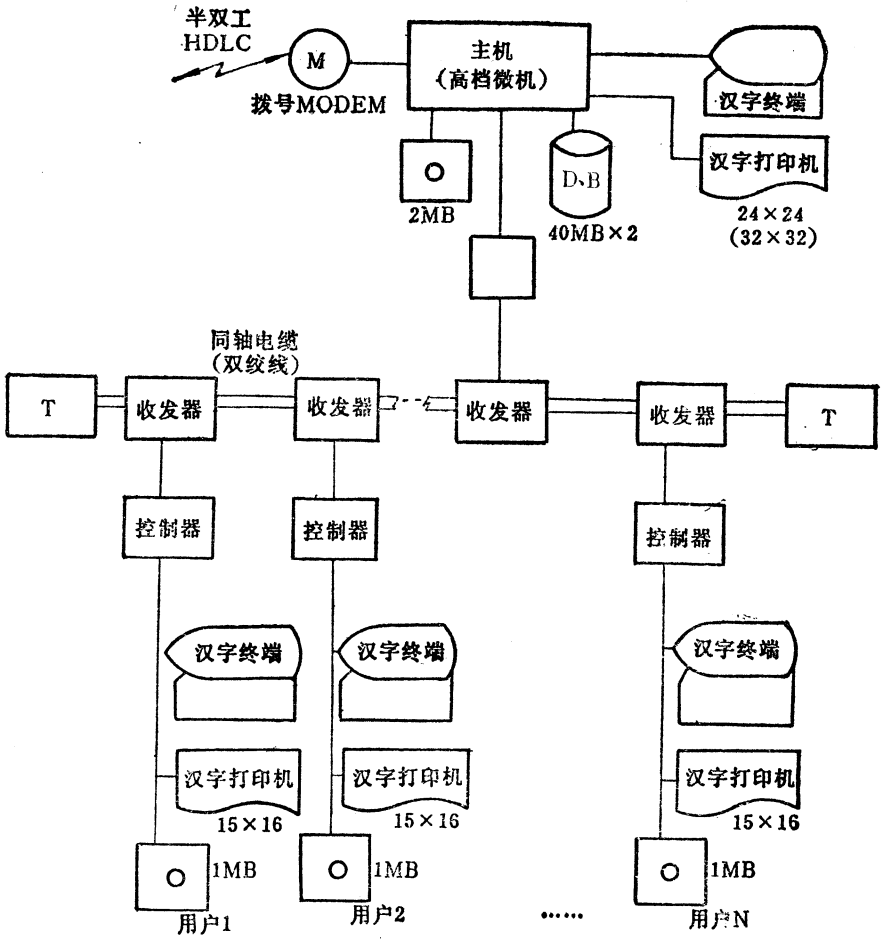


图12-41　汉字微型机局部网络

间的集中式层次结构。

三、对组成汉字信息处理局部网络的主机与汉字智能终端技术要求

（一）主机

1．硬件系统：

（1）采用16位微处理机；

（2）内存容量不小于 512K 字节；

（3）温式硬盘容量为 40M字节 × 2；

（4）软盘（8 英寸）容量为 2M字节 × 4；

（5）高速汉字打印机（汉字1000行/分）采用32×32或24×24点阵；

（6）通信接口采用RS-232-C　BSC通信插板等；

（7）汉字字模库可容纳 10000 字以上。

2．软件：

（1）采用可支持局部网络的多用户、多任务操作系统；

（2）采用数据库管理系统；

（3）采用各种高级语言；
（4）具有汉字处理实用程序（如：汉字输入编码变换表以及支持汉字高速打印软件等）；
（5）具有各类应用程序；
（6）具有大型计算机的联机仿真程序（如IBM3270仿真程序）；
（7）采用汉字信息交换用汉字控制字符集。

（二）用户汉字智能终端站

1．硬件系统：
（1）CPU采用8位或16位微处理机；
（2）内存容量为128K字节以上；
（3）MOS汉字字模库可容纳6000字以上（GB2312-80标准）；
（4）8英寸软盘（双面双密度）容量为1兆字节；
（5）汉字打印机采用三种字体15×16点阵针打；汉字打印速度为60字/秒以上；
（6）采用12英寸光栅扫描CRT，它带有64K字节RAM刷新存储器，可显示15×16点阵汉字
（7）输入键盘采用标准ASCII小键盘（或4千字笔触式大键盘）
（8）通信接口：网络传输器插板。

2．软件：
（1）采用CP/M操作系统（或MS-DOS），中英文混合编辑及文字处理软件，网络管理软件；
（2）采用BASIC、PASCAL、COBOL、FORTRAN等高级语言。利用BASIC和COBOL可编写汉字应用软件（有开发能力）；
（3）在汉字输入编码方面，配有4～5种国内能推广应用的输入编码方法及整字编码、电报码方式等；
（4）各种应用软件。

（三）业务范围

用户汉字智能终端的业务归纳如下：
（1）办公室自动化（文稿处理、文件传递等）；
（2）人事管理；
（3）银行业务；
（4）仓库管理；
（5）商业信贷；
（6）大、中型饭店管理；
（7）科学计算（含教学）；
（8）交通管理；
（9）海关业务；
（10）工资计算统计；
（11）图书管理（情报检索）；
（12）企业管理等。

第十三章　精密汉字编辑排版系统

活字印刷是我国的四大发明之一，但近几十年来我国的排版印刷技术大大落后了。在近二十多年来，美国、西德、英国相继发明了第二、三、四代照相排字机(phototypesetter，简称照排机)。这种照排机与计算机相连，组成编辑排版系统(editing-typesetting system)，可以取代铅字，实现自动排印书、报等正式出版物，从而显著提高了劳动生产率。

本章介绍精密照排机的几个发展阶段、激光照排系统的主要技术困难及其解决方法、系统的构成，以及支持编辑排版系统的大型软件的主要功能。

13.1　精密型照相排字机的几个发展阶段

13.1.1　手动（第一代）照排机

1946年美国Intertype公司研制成功第一台手动的西文照相排字机，称为Fotosetter。它与铅字活版的区别是，字模不是由铅字组成，而是制作在一块透明的模板上。在键盘的控制下，被选中的字符对准一个窗口，在强的灯光照射下使这个字符在底片上感光，然后底片移动相当于一个字符宽度的距离，准备照下一个字符。手动照排机的效率很低，一旦按错字，不得不修改底片或相纸，造成很多麻烦。近年来日本在手动照排机上配上微处理机，变成廉价的半自动照排机，提高了效率。

13.1.2　光机式（第二代）照排机

1951年美国研制成第一台光学机械式照排机，称为Photon 200，其控制和计算部件采用继电器。稍晚，又推出了Photon 260、513、540和560。在这些机型上，首先实现了在穿孔纸带控制下的连续照排。五十年代末，第二代照排机与计算机相连，构成计算机排字系统，从而进入了一个新阶段。

第二代光机式照排机可分成两类。

第一类：旋转圆筒型。上面提到的Photon系列属于这类，其基本的光学原理如图13-1所示。

西文字模制作在一个透明圆盘或圆筒上，在照排过程中圆盘作高速的匀速转动，圆盘的每个字模位置上都有一个标记，可用光学方法读出。对此标记计数，就可知道圆盘已经旋转到什么位置上。每圈的开始，将标记计数器清为0。

旋转透镜组一般包含十几个不同的透镜，以获得十几种不同大小的文字。当字号改变时，需要发命令，让透镜组旋转，使它对准窗口位置。

当所需要的字符的代码与标记计数器的值相同时，意味着当前处于窗口位置的字符正是需要照相的字符，于是启动闪光灯发出短暂而十分强烈的光，使窗口位置上的字符经透镜和反射镜在底片上成象。然后依靠反射镜旋转或移动底片准备使下一个字符

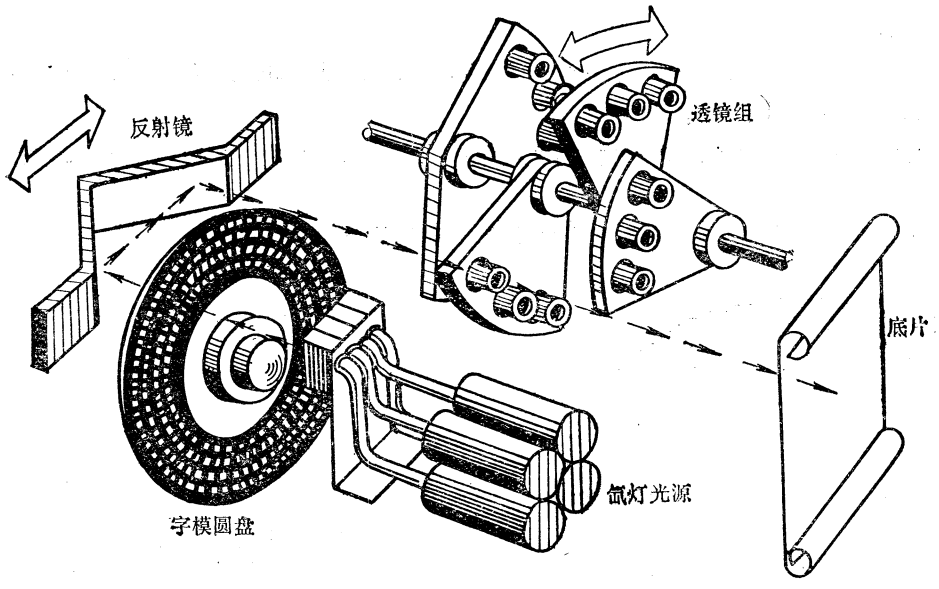


图13-1　光机式照排机原理图

曝光。

旋转的字模圆盘和用氙灯曝光是 Photon 公司在这项技术设备上的主要专利，这一发明在以后生产的二代机中得到了广泛的应用。

由于圆盘上的字模是在高速运动下被曝光的，所以曝光时间必须很短，一般为几个微秒，否则圆盘上下一位置的字符会被错误地曝光或者成象不清晰。这样就要求光很强，底片灵敏度比较高。随着第二代照排机的问世，美国很快研制出适合于这种曝光方式的比较灵敏的底片。

第二类：静止的字模板方式。

所有字模都制作在一块透明平板上，分成很多个区，每个区包含几十个或几百个字符，依靠下面两个机械动作选择所需的字符：

（1）先让平板移动，把包含所选字符的区移动到对准镜头的位置上；

（2）由于一个区内包含很多字符，在区内的选择则依靠一个网状的屏蔽装置。移动此装置，能把区内除一个字符以外的其余所有字符都屏蔽住，选中的字符通过网状屏蔽装置和透镜借助光源的照射在底片上成象。这里不必用闪光灯，只需要一个强光源就行了，因为在曝光过程中字符是静止的。为了保持光密度不变，当字号变化时光强也应有相应的变化，所以实际上需要几种不同强度的光源。

第二代照排机的缺点是机械动作太多，很难得到高可靠性，此外速度低，适应面窄，不能输出各种复杂的图案，更不可能输出黑白图片和有灰度层次的照片。

汉字字量比西文字母大得多，光机式汉字照排机的圆盘必然会做得更大，机械和光学系统将更复杂。七十年代初日本与美国 Photon 公司合作，研制成第二代汉字照排机，过去十年内在日本各种类型的第二代汉字照排机逐渐得到推广。

13.1.3　阴极射线管（第三代）照排机

一般有数字存储和模拟存储两种类型。

一、数字存储型

1965 年西德 Rudolf Hell 公司首先研制成第三代照排机的原理性样机，称为 Digiset，1968 年开始成为商品。

所有的字模以数字化形式存储在计算机内，图 13-2 给出一个西文字符 D 的数字化字形。

输出装置是一个高分辨率的阴极射线管，其原理如图 13-3 所示。

图13-2　西文字符的数字化字形

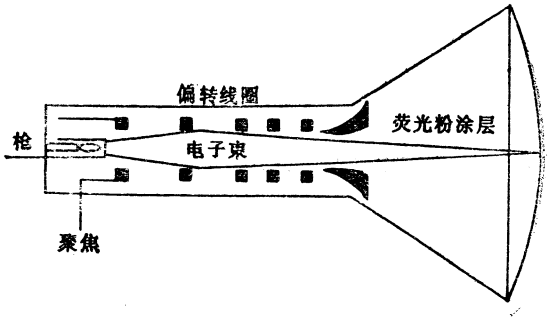


图13-3　阴极射线管照排机原理

与普通的电视显象管或一般终端设备中的 CRT 显示不同之处是：

1）分辨率要高得多。一般在 20 线/毫米 以上，甚至 50 线/毫米。因为以点阵表示的精密汉字字模，一个五号字需 100×100 点才能保证质量，要把一页版面都显示在荧光屏上，x 方向至少需显示 4000 个点以上，而电视显象管仅 600 多点。

2）不需要刷新存储器，不必反复扫描整个屏幕。每个字符只需扫描一次，在底片上感光就行。

3）不象电视机那样沿水平方向逐线扫描，而是逐字扫描，一个字符扫完后再扫描第二个字符。扫完一个字符后可以停下来，等待第二个字符的点阵信息到达计算机内存后再进行扫描。

4）有复杂的校正电路。

（1）动态聚焦：保证屏幕中心部分和四周有相同的聚焦效果。

（2）非线性失真校正：保证屏幕边缘部分与中间部分的扫描线性度一致。

（3）象散校正。

第三代照排机可以采用电子方法获得不同大小的字符。例如五号字用 100×100 点阵，当字号缩小至小五号时，点阵数仍不变，而用缩小扫描步长的方法，即缩小两个光点之间的距离，使实际显示的字缩小。当然，同时还需相应缩小光点的直径，相应地改变聚焦程度，这样才能保证光密度不变。

要求设计制造高分辨率显象管和复杂的校正电路，以及利用电子方法形成不同大小的字符，这是 CRT 西文照排机的三个主要困难。七十年代初，这些技术已趋成熟，CRT

西文照排机逐渐在市场上大量出现。日本于七十年代中期引进西德技术，研制成汉字的第三代照排机。

二、模拟存储型

它又分成下述两种方式

（一）飞点扫描方式

第一台这类设备是美国的Linotron 1010，后改进成 Linotron 505，其原理如图13-4所示。

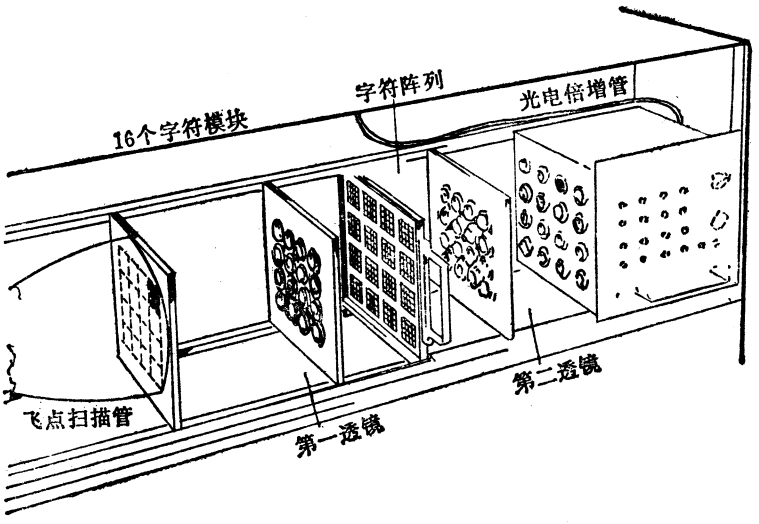


图13-4　飞点扫描方式照排机

所用的飞点扫描管（flying spot scanning tube），是一种特殊类型的阴极射线管，它能在屏幕的任一区域上作小范围的扫描，发出的一连串扫描光束，经透镜后射向字模板（阴图）上选中的那个字符。每束光由一个以二进制数表示的坐标位置来定位，然后光束在这个字符所占的范围内扫描，扫描过程中，若某点有笔划，则光透过，光电检测器输出为‘1’；若无笔划，则光透不过，输出为“0”。检测出的点阵信息立即送给输出用的 CRT，在屏幕上形成该字符，并在底片上感光。这里输出用的 CRT 与提取字符点阵的飞点扫描管是严格同步的。

（二）字模管方式

字模板并不是放在外面，而是直接放在 CRT 管内，插在电子枪与荧光屏之间。电子枪发出的电子束经选择偏转后，透过中间的字模板上所选中的那个字符，再在 CRT 屏幕的适当位置上显示出来。

七十年代美国研制成的字模管（charactron）照排机可存放一两千个西文字符，并能产生不同大小的字。再提高字符数将会遇到较大困难。

由于数字化存储器的价格大幅度下降，所以近年来数字化存储技术已成为 CRT 照排机中的主流。对于汉字 CRT 照排机，要用飞点扫描方式或字模管方式来实现，其困难都是很大的。日本曾研制过飞点扫描方式的 CRT 照排机，但未获良好的结果。

第三代照排机的优点是：机械动作极少；输出速度高；可同时输出黑白图片和照片；其缺点是利用荧光屏发光而在照相底片上感光，要求底片的灵敏度较高。

13.1.4 激光（第四代）照排机

最早研制激光照排机（Laser typesetter)的有美国 Dymo 公司和英国 Monotype公司，但 Dymo 公司于七十年代末中断了这一研究，英国 Monotype 公司则于 1976 年左右推出了平板转镜式激光照排机 Lasercomp，并投入了批量生产。1980 年又将该照排机初步具备汉字排版的某些功能。

第四代激光照排机分成滚筒型和平板转镜型两类。

一、滚筒型激光照排机

在滚筒型激光照排机运行时，底片贴在旋转的滚筒上，滚筒以每分钟 1500～3000 转的固定速度匀速转动。滚筒每转一圈在底片上扫描出一条线，有笔划处的信息为“1”，控制声光（或电光）调制器，使激光束通过，在底片上曝光一个点（1 位信息)；无笔划处的信息为“0”，控制调制器不让激光通过，因而底片上该点位置不曝光。这是主扫描系统，实现版面水平方向的扫描。副扫描系统则是带有激光源的部件在完成版面水平方向的扫描后，匀速移动一个光点的距离。当丝杠从起始位置移动到结束位置时完成整个版面的扫描。

为了提高滚筒型激光照排机的速度可采用两种途径。一种途径是提高滚筒转速，但滚筒转速过高会引起机械振动或造成各种误差而影响文字质量；另一种途径是用多路激光平行扫描，在滚筒转速不变的情况下，从一路变成四路激光平行扫描就能使输出速度提高四倍，当然相应的丝杠移动速度也要快四倍。多路平行扫描需解决有关的光学和电路问题，才能保证四束光的光强均匀一致和在底片上有正确的成象位置。

二、平板转镜型激光照排机

它的主扫描系统采用高速旋转的多面转镜，转镜每旋转一个面，在底片上扫描出一条线，同时底片往前移动一个光点的距离。通过调制器控制曝光的方法与滚筒型相同。

平板转镜型照排机可以达到更高的速度，也便于上、下片和底片的切割，但输出精度往往不如滚筒型。

第四代激光照排机的优点是：文字质量很高，输出速度高，适用面广，能同时输出黑白图片和照片，对底片灵敏度要求不高等；此外更吸引人的是能过渡到激光直接雕版。所以它是很有前途的新的一代照排机。激光扫描的缺点是扫描控制的灵活性不及 CRT 照排机。

13.2 高分辨率汉字字形的存储和几种信息压缩方案

13.2.1 对精密汉字照排的分辨率要求和数字化字模的存储量问题

第三代和第四代照排机都是用扫描打点的方式形成一个西文字符或汉字字模。其分辨率越高，字的质量越好。为了满足书、刊等正式出版物的质量要求，其分辨率至少需 25 线/毫米。北京大学等单位研制的系统中，分辨率定为 29.2 线/毫米。与 CRT 照排机不同，无论滚筒型还是平板转镜型的激光照排机，都不能瞬时改变激光光点的大小，因而不同大小的字符必须用不同的点阵表示。若按 29.2 线/毫米 计算，则印刷用各种字号所需点阵大小如表 13-1 所列。

表13-1 印刷用汉字字号与字身点阵的关系

字号	磅数	字身点阵大小	字号	磅数	字身点阵大小
七号	6	62×62	小二号	18	184×184
小六号	7	72×72	二号	21	216×216
六号	7.875	82×82	一号	28	288×288
小五号	9	92×92	小初号	31.5	324×324
五号	10.5	108×108	初号	35	370×370
小四号	12	124×124	小特号	42	432×432
四号	14	144×144	特号	49	494×494
三号	15.75	162×162	特大号	56	576×576

（1磅 = 0.35毫米）

印刷用的汉字字体也很多，有书版宋体、报版宋体、标题宋体、仿宋体、楷体、黑体、长宋体、扁宋体、长黑体、扁黑体、长方宋体、隶书体等。每种字体需7000以上汉字。考虑不同字体和不同字号在内，印刷用的汉字字模数量超过65万个，其对应的存储量超过2×10^9字节。存储量是发展第三代尤其是第四代汉字照排机的一个关键问题，需要研究有效的信息压缩技术以减少字形信息所需的存储量，下面三小节将介绍国外研究的三种信息压缩方法。

13.2.2 记录黑白段长度的压缩方法

假定点阵为256×256。不需死板地记录每个方格内的0或1信息，而是依次记录每行内各个白段和黑段的长度，如图13-5所示。这样行信息用如下压缩格式：

〔行标志〕 〔重复行数〕 〔白段〕 〔黑段〕 〔白段〕 〔黑段〕…， 〔白段〕 〔黑段〕

* n W_1 B_1 W_2 B_2 W_m B_m

当相邻n行的所有白、黑段数值完全相同，即全为空白行或者所有笔画均为竖直线时，只需一行信息便能表示，此时重复行数将指示n。

上述压缩格式中，行标志与重复行数占一个字节，每个白段占一个字节，每个黑段占一个字节。

若汉字每行的黑段数平均为3，则每行平均需7个字节，一个汉字共256行，即平均需256×7个字节表示，而原来需256×32个字节才能表示。显然，原来的点阵越大，压缩效果越明显；笔画越少(意味着黑段少)，尤其是非垂直笔画越少，压缩效果越明显。

这种压缩方法首先在1965年西德Hell公司生产的CRT西文照排机Digiset中使用(它与字稿的数字化扫描仪输出结果是一致的)，以后便在第三、第四代西文照排机中广泛应用。其优点是可以忠实地复原出原来的点阵而丝毫没有失真，复原硬件也比较简单。缺点是单字压缩倍数和总体压缩倍数都太低；而且不易从一种字号的信息不失真地得到其他字号的点

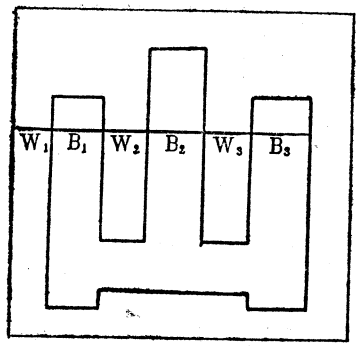


图13-5 记录黑、白段长度的压缩方法

阵，因而不同大小的字不得不占据不同的存储空间。由于有这些缺点，这一方法用于汉字字形信息的压缩有较大的局限性：一个 80 兆字节的磁盘只能提供几万个汉字字模，而速度低的磁盘将成为照排系统的输出瓶颈（bottle neck）。

13.2.3 霍夫曼压缩方法

把一个 128×128 点阵分成 32×32＝1024 个区，每个区含 4 × 4 个点。每个区内16个点可有 2^{16} 种排列，即 2^{16} 种模式，但这些模式出现的概率差别很大。例如，16 位全 1 的模式概率是非常高的，图 13-6 所列出的一些模式出现的概率也是较高的，这些模式对应于汉字的横、竖笔画。

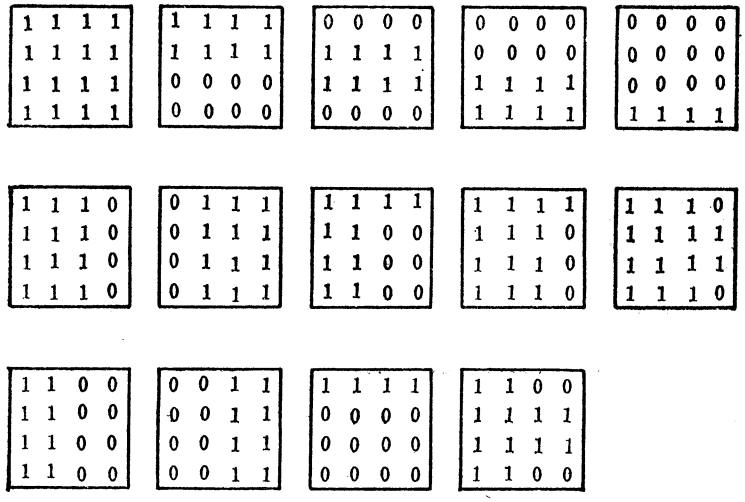


图13-6　概率较高的一些模式的例子

一般情况下，一个区需 16 位二进制位表示，但对于那些概率最高的模式可用 4 位二进位表示，概率次高的模式用 12 位二进位表示，概率最低的模式用 20 位二进位表示。

例如用 0000～1101 表示概率最高的 14 种模式，以此地址查表可找到他们所对应的 4 × 4 点阵；1110 则表示概率次高的 256 种模式，后面的 8 位表示该模式序号，以此序号作地址，查表可找到对应的 4 × 4 点阵；1111 则表示概率最低的一般的区，后面 16 位直接表示对应的 4 × 4 点阵。

这种压缩方法的基本思想是利用模式出现的概率有较大差别这一特点。需对大量汉字作统计，确定哪些模式概率最高，哪些模式概率次高。具体实现时还可以有很多变型和改进，这里不详细介绍了。

假定概率最高的14种模式占 50%，概率次高的 256 种模式占 25%，其余模式占25%。一个汉字平均需 1024×（4 ×0.5＋12×0.25＋20×0.25）位，即 1024×10 位，而原来需 1024×16 位。

对于高分辨率汉字字形而言，霍夫曼（Huffman）压缩方法比上面所说的记录黑白段的方法压缩倍数更低，而复原步骤却较复杂，因而优点不明显。七十年代初，大规模集成的 EPROM 和 ROM 还很昂贵，日本有些公司用霍夫曼方法对 32×32 点阵和 24×24 点阵进行信息压缩，当时有一定的价值。

13.2.4 字根组字的压缩方法

这一压缩方法的基本思想是把汉字看成由不同的字根组成，在计算机存储器内并不存放所有整字字模，而是存放几百个字根，所有的字都用这些字根的图形经适当放大、缩小、拉长、压扁后定位组合而成。例如“新”字由“立”、“木”、“斤”三个字根组成，把字根“立” x、y方向均缩小成1/2，放在左上角；把字根“木”在x、y方向均缩小成1/2，放在左下角；把字根“斤” x方向缩小成1/2，y方向不变放在右半部分，这样组成一个“新”字。由于字根数目比整字数目要少得多，所以可得到较高的压缩倍数。但这一方法有一个严重缺点：合成的字的质量会受到影响。质量下降主要由下述两个原因造成。

（1）印刷用汉字字模的结构十分严谨，由字根组合的字往往在结构上与标准字模有差别。我们曾对最简单的字根“艹”（草字头）作了统计，由于草字头下面部分的笔画不同，草字头的变化是多种多样的，用很少几个草字头来对付所有出现草字头的汉字，必然影响文字的质量。

（2）组合成的汉字往往笔画粗细比例不匀称。例如上面提到的“新”字中的横与整字“立”中的横的宽度应该是差不多的，其正确的宽度比约为3:4或4:5。但用字根组字法合成的“新”字中的横，与整字“立”中的横的宽度比却变成1:2，甚至2:5，这样就很难看。

由于以上原因，字根组字的压缩方法并未在精密照排系统中得到普遍应用，国外广泛使用的是记录黑白段长度的压缩方法。

既要得到高的压缩倍数，又要保证文字质量，这是一个困难的问题。北京大学等单位于1979年研制成的计算机-激光汉字编辑排版系统（原理性样机）中首次采用一种保证文字质量、同时压缩倍数很高的汉字字形信息压缩方法，显著改善了性能价格比，1983年又用大规模集成电路和双极型位片器件彻底更新了整个系统。下面几节将详细介绍这一技术的原理和实现方法。

13.3 一种保证文字质量的高倍数汉字字形信息压缩技术

13.3.1 汉字规则笔画和不规则笔画的压缩表示

一、规则笔画的信息压缩表示

把印刷体汉字的笔画分成横、竖、折等规则笔画和任意曲线形式的不规则笔画。规则笔画由直线段、起笔、收笔和转折等笔锋组成，其中直线段可有轻微的倾斜。随着笔锋附近笔画的疏密变化，一类笔锋可有很多不同形状，有大有小，有长有扁，但其总数是不大的；可以把它们的外形用轮廓折线表示，事先都存在内存中，然后编成号，用号码来表示笔锋。例如字体横笔画的形状如图13-7所示。

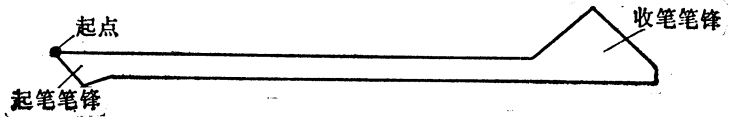


图13-7 宋体横笔画

这里收笔笔锋有七种变化就能概括全部宋体字的横，变化再多意味着字写得不规范，反而不美观。因此，在宋体横笔画的压缩表示中只需三位指示收笔笔锋的号码（0～6）；其余信息将指示横的起始 x、y 坐标，横的长度、宽度以及有没有起笔笔锋。这样，宋体的一笔横只需 3 或 4 个字节就能作精确描述而丝毫没有失真。

标题宋体的竖笔画的形状如图 13-8 所示。

注意直线段的右侧 B 点以下向外倾斜，这种倾斜在标题宋体的竖和折笔画中是常见的，显得气派大。

用 5 个字节指示起始 $x-y$ 坐标、长和宽、起笔和收笔笔锋号、A B 间距离和右侧倾斜度。当左右两侧均无向外倾斜时，只需 4 个字节。

据统计，规则笔画占汉字笔划总数的一半以上，上述十分紧凑的信息形式对提高压缩倍数起了重要作用；信息中细致描述宽度、倾斜度的作法，有利于保持文字变倍后规则笔画宽度一致性和防止缩小时笔划过密。这是确保变倍质量所必需的，因为横、竖、折的宽度和对比对文字质量影响很大。

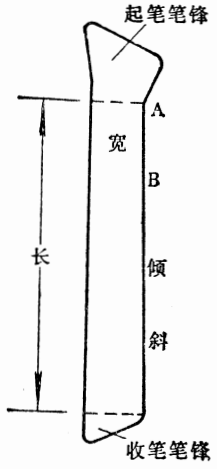


图13-8　标题宋体竖笔画

二、不规则笔画的信息压缩表示

对于不规则笔画，可用一连串折线逼近其轮廓曲线，只要折线与原曲线足够近似，这种逼近对文字质量没有影响。我们取五号字的点阵密度为 108×108，去掉边框（即字间距），字心点阵为 96×96。一般字体均以五号字为基准字号，计算机内只存五号字的压缩信息，其余字号的点阵由基准字号通过硬件变倍得到，因此每个汉字都看成画在 96×96 点阵的网格上。

图 13-9 示出了不规则笔画的例子。

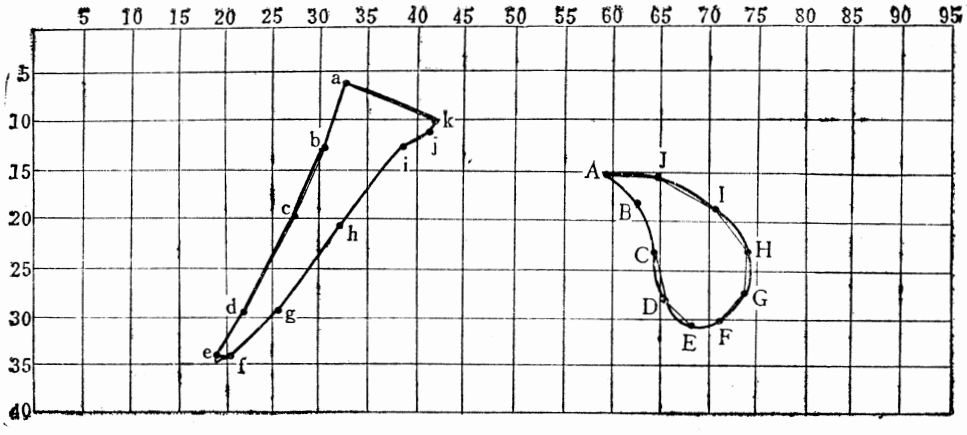


图13-9　不规则笔画例子

图 13-9 右面的不规则笔画“点”用一串向量 AB，BC, CD, DE, EF, FG, GH, HI, IJ，JA 表示。笔画的起点用 $x-y$ 坐标指示；每个向量用带符号的 Δx，Δy 表示。我们规定：第一象限 $\Delta x \geqslant 0$，$\Delta y < 0$；第二象限 $\Delta x < 0$，$\Delta y < 0$；第三象限 $\Delta x < 0$，$\Delta y \leqslant 0$；第四象限 $\Delta x \geqslant 0$，$\Delta y \geqslant 0$。上述一串向量的数字表示如下：

$+3$，$+3$；$+2$，$+5$；$+1$，$+5$；$+3$，$+2$；$+4$，-1；

$+3$，-3；$+0$，-3；-3，-3；-7，-4

最末一个向量 JA 是封口向量，不必在压缩信息中列出，可由点阵复原设备自动产生。可以看出，以上向量的 Δx 和 Δy 绝对值均小于 8，但象限经常变化。统计表明，这类情况在汉字不规则笔画中经常出现（例如“点”），其信息应更紧凑。为此我们用控制字节来区分向量长度的三种不同情况：小于 8，小于 16 和大于等于 16。

（1）若控制字节的开头两位为 00，则表示后面 N 个向量处于同一象限内，并且每个向量的 Δx、Δy 绝对值都小于 16；这里 N 由控制字节的最末四位指示，控制字节的中间两位指示四个象限中的哪一个。后面 N 个向量中的每个向量则用一个字节表示：其中四位指示 Δx 绝对值（0～15），四位指示 Δy 绝对值（0～15）。

（2）若控制字节的开头两位为 01，表示后面 N 个向量的 Δx、Δy 绝对值都小于 8；这里 N 由控制字节的最末五位指示。后面 N 个向量中的每个向量则用一个字节表示：其中两位指示该向量所处的象限，三位指示 Δx 绝对值（0～7），三位指示 Δy 绝对值（0～7）。

（3）若控制字节的开头两位为 10，表示当前向量的 Δx、Δy 绝对值中至少有一个大于等于 16；此时控制字节的末六位和下一字节合起来表示一个向量。

（4）若控制字节的开头两位为 11，表示不规则笔划的起始点，剩下六位和下一字节合起来，指示起始 $x-y$ 坐标。

这样，图 13-9 的不规则笔画‘点’由下列 12 个字节表示：

起点 A 的 $x-y$ 坐标（2 字节，第一字节为 11 控制字节）；

01 控制字节（$N=9$）；

9 个字节表示向量 AB，BC，CD，DE，EF，FG，GH，HI，IJ。

图 13-9 左面的不规则笔划‘撇’由下列 14 个字节表示：

起点 a 的 $x-y$ 坐标（2 字节）；

00 控制字节（$N=4$，第三象限）；

4 个字节表示向量 ab，bc，cd，de；

00 控制字节（$N=6$，第一象限）；

6 个字节表示向量 ef，fg，gh，hi，ij，jk。

对不规则笔画形状的大量统计表明：压缩表示中上述三种区分向量长度的情形有利于压缩信息。这种区分不会给字模制作带来附加麻烦，因为字模制作系统的软件将计算每个向量的长度，区分三种情形，自动形成对应的控制字节。

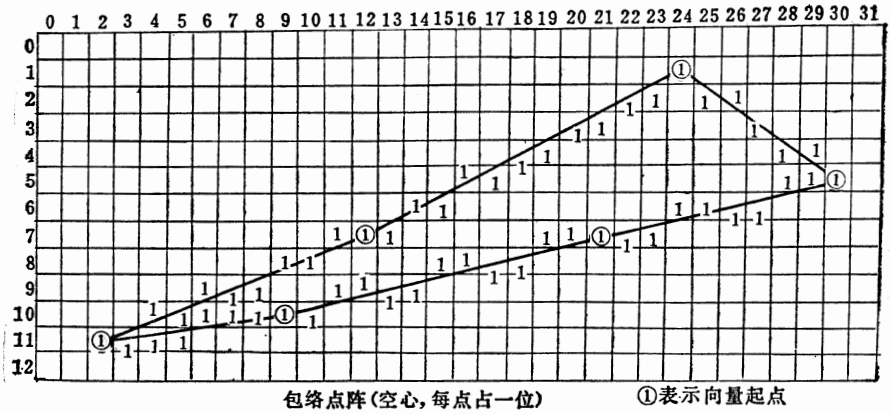
采用上述规则笔画和不规则笔画的压缩表示方法后，一个五号汉字所需存储量平均降为 100 个字节，而文字质量不受影响。

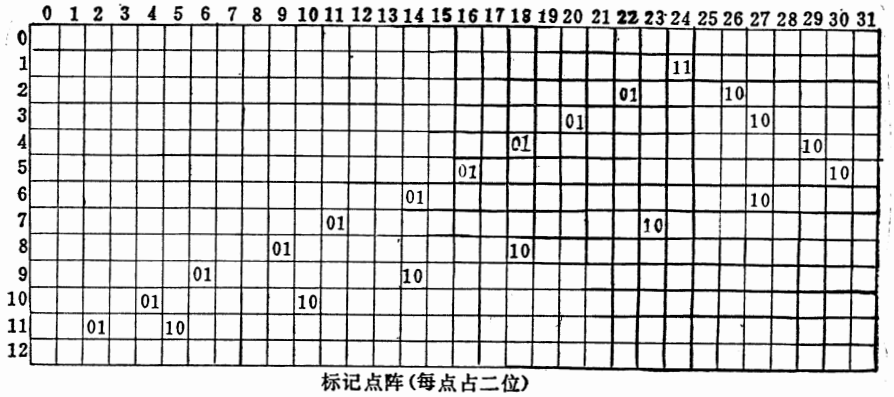
13.3.2 汉字字形复原技术

一、字形点阵的复原步骤和标记点阵

把汉字压缩信息转换成最终输出的字形点阵称为字形复原，其步骤如下：

（1）对规则笔画压缩信息进行分析，按照笔锋编号找到对应的向量串信息，并转换成标准形式；把不规则笔画压缩信息的三种不同形式的向量串转换成标准形式，并作文字变倍，变倍后的标准形式向量 Δx 和 Δy 各占 10 位（包括符号）。





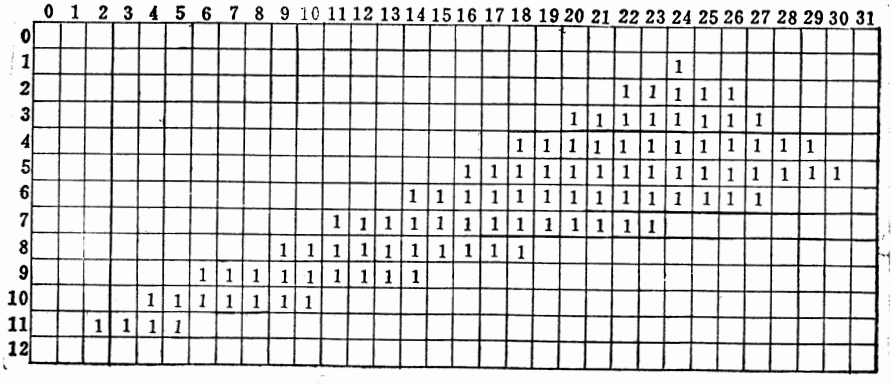


图13-10　包络点阵、标记点阵和最终输出点阵

（2）把标准形式的向量转换成与之最逼近的阶梯点，在经过的每个阶梯点上写入两位标记，形成标记点阵。标记点阵中每个点占两位二进位，表示该点的下面四种情况：

① 空点（00）；

② 黑段的始点（01）；

③ 黑段的终点（10）；

④ 孤立黑点（11）。

（3）把标记点阵转换成最终输出的点阵，并存入激光扫描缓冲存储器的对应单元中去。

复原步骤1、2和3分别由两台双极型位片微型计算机完成。复原设备应简单、高速和不引起文字失真。要做到这点，需在汉字信息压缩方案，复原步骤和算法，位片微型机的逻辑设计和微程序设计这个三方面下功夫。

复原设备输入的是汉字压缩信息，它以笔画为序，规则笔画在前，不规则笔画在后，一笔完了再开始下一笔；而复原设备输出的是点阵，它以水平方向扫描线为序，一线完了再开始下一线。输入数据与输出数据之间存在着“次序冲突”。解决这一冲突的办法是增加中间存储器存放中间数据，而中间数据结构的选择是至关重要的。这里我们采用标记点阵作为中间数据，使汉字压缩信息转换成标记点阵、以及把标记点阵转换成最终输出点阵，这两部分转换都简单快速。我们没有选择图13-10最上面的包络点阵作为中间数据，因为把包络点阵转换成最终输出点阵的算法很复杂，不易实现。

二、把一个向量转换成最近似的阶梯点的算法

研究折线方程 $y=\frac{\Delta y}{\Delta x}x$ （$x=1,2,\cdots,\Delta x$），假定 $\Delta x\geqslant\Delta y$。由于折线对应的阶梯点只能取整数值（即近似值），因此

$$y\text{近似值}=\begin{cases}\left[\frac{\Delta y}{\Delta x}x\right]+1 & \left(\text{当}y\text{准确值}\geqslant\left[\frac{\Delta y}{\Delta x}x\right]+\frac{1}{2}\right)\\ \left[\frac{\Delta y}{\Delta x}x\right] & \left(\text{当}y\text{准确值}<\left[\frac{\Delta y}{\Delta x}x\right]+\frac{1}{2}\right)\end{cases}$$

这一近似值是最优的。为了硬件实现时避免乘除运算，我们用逐次递推法获得 y 近似值。

起始时 $x=0$，$y=0$。

第一步 $x=1$，此时 $\left[\frac{\Delta y}{\Delta x}\times 1\right]=0$，所以门槛值(threshold)为 $0+\frac{1}{2}=\frac{1}{2}$。

若 $\frac{\Delta y}{\Delta x}\geqslant\frac{1}{2}$，则近似值 y 加1，即上升一格；若 $\frac{\Delta y}{\Delta x}<\frac{1}{2}$，则近似值 y 不变。每当近似值 y 加1后，门槛值也应加1。因此算法变成每一步时的 y 准确值与门槛值进行比较：

若大于等于门槛值，则 y 加1，门槛值加1；

若小于门槛值，则 y 不变，门槛值不变。

依次类推，第 i 步，即 $x=i$ 时，应判断 $\frac{\Delta y}{\Delta x}i$ 是否 $\geqslant y^*+\frac{1}{2}$（$y^*$ 是 $x=i-1$ 时的 y 近似值）。上式可改为判断：

$$\Delta y_i - y^*\Delta x \text{是否} \geqslant \frac{\Delta x}{2}$$

这里 Δy_i 表示：x 每前进一步，累加器应加 Δy；$-y^*\Delta x$ 表示：y 每前进一步，累加器应减 Δx。这样导出了由位片微型机实现的下述转换步骤。

向量起点时，累加器 ACC 的初值置为 0。

每一步中，ACC+Δy→ACC，然后ACC与$\frac{\Delta x}{2}$比较：

若 ACC$\geqslant\frac{\Delta x}{2}$，则 y 改变（第一、二象限加 1，第三、四象限减 1），且 ACC−Δx→ACC，继续做下一步；

若 ACC$<\frac{\Delta x}{2}$，则 y 不变，继续做下一步。

重复执行上述操作，直到做完 Δx 步为止。

对于 $\Delta x<\Delta y$ 的向量，所执行的操作与此类似。

图 13-10 '撇'的最高点（$x=24$，$y=1$）起始的那个向量 $\Delta x=-12$，$\Delta y=6$，转换成最近似的阶梯点时，x 每步都减 1；y 只有第一、三、五、七、九、十一步时才加 1，其余步 y 不变。x、y 的变化在包络点阵中可看出。

上述转换步骤保证字形失真为最小，而且只用加、减法和比较运算，便于用最简单的硬件实现高速转换。

三、在向量经过的阶梯点上写入二位标记的规则

除向量的起点和终点外，一个向量通过的阶梯点上，每一步写入什么样的二位标记由表 13-2 所列规则决定。

表13-2中“本步”指的是从该点开始的那步，“上步”指的是达到该点的那步。图13-10中以最高点（$x=24$，$y=1$）为起点的那个第三象限向量经过的阶梯点中，$x=23$，$y=2$ 的那点本步 y 不变，所以该点写入 00 标记（即不写）；$x=22$，$y=2$ 的那点本步 y 变，所以该点写入01标记。图中以 $x=30$，$y=5$ 为起点的第二象限向量经过的阶梯点中，$x=29$，$y=4$ 的那点上步 y 变，所以该点写入10标记；$x=28$，$y=4$ 的点上步 y 不变，所以在该点写入00标记（即不写）。

对于两个向量之间的结点（即向量起点或终点），写入什么样的二位标记由表 13-3 所列规则决定。表13-3的第一向量指以该点为终点的向量，第二向量指以该点为起点的向量。

表13-2可看作表13-3的特殊情况，此时两个向量的象限相同。

表13-2　二位标记的决定

象限	条　件	写入的标记	象限	条　件	写入的标记
I	本步 y 变 本步 y 不变	1 0 0 0	II	本步 y 变 本步 y 不变	0 1 0 0
III	上步 y 变 上步 y 不变	1 0 0 0	IV	上步 y 变 上步 y 不变	0 1 0 0

下面以图13-10中的结点为例说明表 13-3 的功能。

$x=24$，$y=1$ 的结点，第一向量属第二象限，第二向量属第三象限，上步和本步 y 都变，所以写入11标记。

$x=2$，$y=11$ 的结点，第一向量属第三象限，第二向量属第一象限，本步 y 不变，所以写入01标记。

表13-3　结点处二位标记的决定

第一向量象限	第二向量象限	条　件	写入的标记
Ⅰ	Ⅰ	本步 y 变 本步 y 不变	1 0 0 0
Ⅱ	Ⅱ	上步 y 变 上步 y 不变	1 0 0 0
Ⅲ	Ⅲ	本步 y 变 本步 y 不变	0 1 0 0
Ⅳ	Ⅳ	上步 y 变 上步 y 不变	0 1 0 0
Ⅰ	Ⅱ	任何情况	1 0
Ⅰ	Ⅲ	本步 y 变 本步 y 不变	1 1 1 0
Ⅰ	Ⅳ	任何情况	0 0
Ⅱ	Ⅰ	上步，本步 y 都变 其余情况	1 0 0 0
Ⅰ	Ⅱ	上步，本步 y 都变 上步，本步 y 都不变 上步 y 不变，本步 y 变 上步 y 变，本步 y 不变	1 1 0 0 0 1 1 0
Ⅰ	Ⅳ	上步 y 变 上步 y 不变	1 1 0 1
Ⅱ	Ⅰ	本步 y 变 本步 y 不变	1 1 0 1
Ⅱ	Ⅱ	任何情况	0 0
Ⅱ	Ⅳ	任何情况	0 1
Ⅳ	Ⅰ	上步，本步 y 都变 上步，本步 y 都不变 上步 y 不变，本步 y 变 上步 y 变，本步 y 不变	1 1 0 0 1 0 0 1
Ⅳ	Ⅱ	上步 y 变 上步 y 不变	1 1 1 0
Ⅳ	Ⅲ	上步、本步 y 都变 其余情况	0 1 0 0

由于汉字的笔画交错，对一个字执行复原步骤 1、2 时，在交错点的标记点阵可能被写入两次。图13-11是仿宋体的“十”字，当处理‘横’时，F 点写入 01 标记，G 点也写入01标记；当处理‘竖’时，在 F 点再写入01标记，但原来该点已有一个01标记了，此时应该在 F 点的右侧，即 H 点位置上再补写一个01标记，否则该行内01标记和10标记将不成对，造成复原步骤 3 的操作错误；当处理‘竖’时，G 点打算写入10标记，但发现该点已有一个01标记了，此时应在 G 点最终写入11标记。假定一个汉字复原开始时，存放标记点阵的中间存储器 e 清为 0，则最终写入的两位标记由准备写入的两位标记（由表13-3决定）和已经写入该点的两位标记联合决定，如表 13-4 所列。

综合表13-3和表13-4，两位最终标记由下列八位信息联合决定：

（1）第一向量的象限（2 位）；

（2）第二向量的象限（2 位）；

（3）上步 y 是否变（1 位）；

（4）本步 y 是否变（1 位）；

（5）该点的原来标记（2 位）。

对上述两位最终标记信息的存储，可用 256× 4 PROM实现，PROM 的地址即上述八位；PROM的内容即两位最终标记，以及附加两位指示是否需在右侧补写01或在左侧补写10。这种实现方法设备极省而速度很快。

表13-4　最终标记的决定

原标记	准备写入的标记	最终标记	补写信息
0 0	0 0	0 0	
0 0	0 1	0 1	
0 0	1 0	1 0	
0 0	1 1	1 1	
0 1	0 0	0 1	
0 1	0 1	0 1	该点右侧补写01
0 1	1 0	1 1	
0 1	1 1	0 1	
1 0	0 0	1 0	
1 0	0 1	1 1	该点左侧补写10
1 0	1 0	1 0	
1 0	1 1	1 0	
1 1	0 0	1 1	
1 1	0 1	0 1	
1 1	1 0	1 0	
1 1	1 1	1 1	

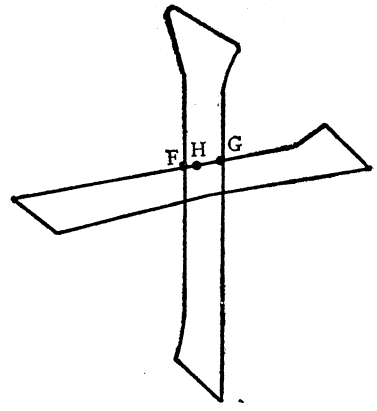


图13-11　某个点被写入两次的例

四、把标记点阵转换成最终输出点阵的方法

点阵的每一行从左至右扫描。当遇到01标记时，该点及其右面各点均转换成“1”，直到遇10 标记为止。11 标记永远转换成“1”，但对该行其余点无影响。标记点阵及其对应的最终输出点阵的例子见图 13-10。

有时，一对 01 标记和 10 标记可能嵌套在另一对 01 标记和 10 标记内，如图 13-12 所示。

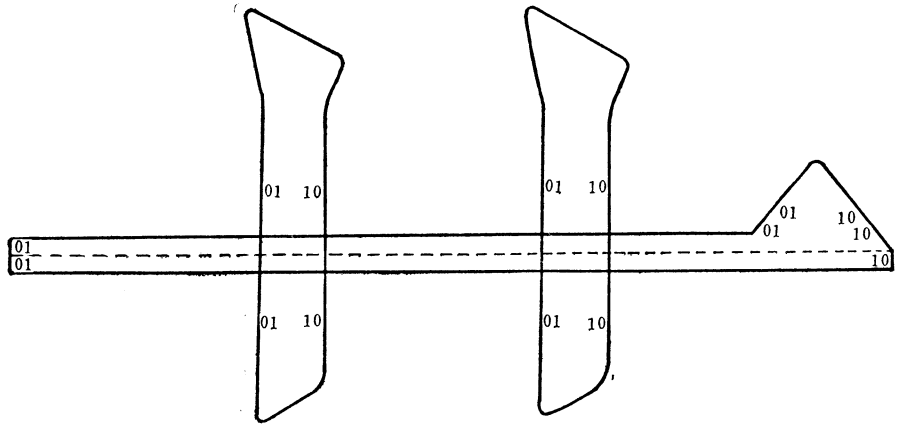


图13-12　01、10标记嵌套情况

用一个两位层数计数器记录01标记的深度，其操作如下：

（1）遇 01 标记时，层数计数器加 1；

（2）遇 10 标记时，层数计数器减 1；

（3）遇 00 或 11 标记时，层数计数器不变。

这样，标记点阵转换成最终输出点阵的规则如下：

（1）01、10 和 11 标记永远转换成“1”；

（2）当层数计数器≥1 时，00 标记转换成“1”，否则转换成“0”。

为提高转换速度，一拍内转换相邻两个点，用 64 × 4 PROM 实现：PROM 的六位

地址是被转换的标记点阵中相邻两个点（4位）和层数计数器的原来状态（2位）；PROM的内容是对应的最终输出点阵的两个点（2位）和层数计数器的新状态（2位）。作为例子，表13-5列出了若干单元的内容。

每线结束时，层数计数器必为0，否则是故障。

表13-5 PROM内容

PROM6位地址		PROM4位内容	
标记点阵两点	层数计数器原状态	最终点阵对应两点	层数计数器新状态
0 0 0 0	0 0	0 1	0 1
0 0 0 1	0 1	1 1	0 0
0 1 0 0	0 0	1 1	0 1
1 0 0 0	0 1	1 0	0 0
0 1 0 1	0 0	1 1	1 0
1 0 0 0	1 0	1 1	0 1
0 0 1 1	0 0	0 1	0 0
0 1 1 1	0 0	1 1	0 1
0 1 1 0	0 0	1 1	0 0
1 0 0 1	0 1	1 1	0 1

13.3.3 高分辨率汉字字形的放大和缩小技术

一、汉字的字号及其大小比例关系

北京大学等单位研制的系统中共容纳十六种字号，如表13-6所列。

表13-6 字号及其比例关系

字号	磅数	字心点阵	比例关系
七	6	56×56	小四号的一半
小六	7	64×64	四号的一半
六	7.875	72×72	三号的一半
小五	9	82×82	小二号的一半
五	10.5	96×96	六种字体的基本字号
小四	12	110×110	五号的 1.146倍
四	14	128×128	五号的 1.333倍，大字体的基本字号
三	15.75	144×144	五号的 1.5倍
小二	18	164×164	五号的 1.708倍
二	21	192×192	五号的2倍，或四号的 1.5倍
一	28	256×256	四号的2倍，或接近五号的 2.5倍
小初	31.5	288×288	五号的3倍
初	35	320×320	四号的 2.5倍
小特	42	384×384	四号的3倍
特	49	448×448	四号的 3.5倍
特大	56	512×512	四号的4倍

注：1磅 = 0.35毫米。

字心点阵不包括字间距，它比字身点阵略小，例如五号字的字心点阵为96×96，字身点阵为108×108。

书版宋体、报版宋体、标题宋体、仿宋体、黑体和楷体这六种字体均以五号字为基准字号，字模存储器中只存放五号字的压缩信息，其余字号都由五号字变倍而得。上述这六种字体原则上都允许放大到小特号，但有些字体不宜放得很大。例如报版宋体一般只宜放大到三号字，若再放大用作标题字，则由于报版宋体的笔画太细，显得气势不够，这种场合应该用书版宋体或标题宋体；正文黑体放大到一号字仍很合适，但再放大到特号，作报纸的通栏大标题字，则显得不够醒目。因此系统还需存储大标题宋体、大黑体、隶书、新魏体这几种字体，他们的管辖范围一般从一号到特大号。为了使特大号标题字笔画边缘仍光滑，我们以四号字作为这四种大字体的基本字号，其他字号由此变倍而得。上述十种基本字体又可通过拉长和压扁的变倍方法变化出各种长字和扁字体。

二、文字变倍的基本原理和防止失真的措施

规则笔画的起笔、收笔和转折等笔锋是被编成号的，它们的形状则以精细的轮廓折线形式存放在内存中。点阵复原设备的微程序将根据笔锋编号获得内存地址，访问这一笔锋所对应的一连串折线信息；不规则笔画本来就用一连串折线描述轮廓。用轮廓折线表示的图形（包括汉字）很易放大缩小。例如，要使图形放大 r 倍（这里 r 不一定是整数），只需把对应的每段轮廓折线放大 r 倍，也即对每条折线的 Δx，Δy 值都乘以 r。这种方法很易想象，也不难实现。但要得到高质量，还需对汉字字形特点作更多的分析，并采取一系列的相应措施，以防止变倍过程中引起的失真。

（一）防止文字变倍时的舍入误差积累

现在要从五号字出发，经变倍变成四号字，变倍率为 1.333 倍。假定采用最简单办法，对每段折线的 Δx、Δy 直接乘以变倍率 1.333，四舍五入获得变倍后的折线，其增量用 $\Delta x'$、$\Delta y'$ 表示。表13-7列出了当变倍率为 1.333 时，Δx 与 $\Delta x'$ 之间的关系。

表13-7　变倍时的舍入情况

Δx	未舍入的 $\Delta x'$	舍入后的 $\Delta x'$	舍还是入
1	1.333	1	舍
2	2.666	3	入
3	3.999	4	入
4	5.332	5	舍
5	6.665	7	入
6	7.998	8	入
7	9.331	9	舍
8	10.664	11	入
9	11.997	12	入
10	13.330	13	舍
11	14.663	15	入
12	15.996	16	入
13	17.329	17	舍
⋮	⋮	⋮	⋮

图13-13的曲线 AH 由 AB，CD，DE 等七段折线组成；曲线 HM 由 HI，IJ，JK 等五段折线组成。A 点和 M 点的 x 坐标相同。在变成四号字时，曲线 AH 所对应的七段

折线的 Δx 乘以 1.333 获得 $\Delta x'$ 时，都是舍；而曲线 HM 所对应的五段折线的 Δx 乘以 1.333获得$\Delta x'$时，都是入。结果变倍后的 A 点和 M 点（用 A' 和 M' 表示）的 x 坐标不仅不相等，甚至可能相差很多，造成严重失真。这就是误差积累的后果。

因此，对于规则笔画的笔锋部分和整个不规则笔画，不能用上述简单的变倍方法，而用结点变倍方法，即用下述公式得到变倍后的 $\Delta x'$，$\Delta y'$：

$A'B'$ 折线的$\Delta x' = B'$ 点 x 坐标 $- A'$ 点 x 坐标

$=$〔B 点 x 坐标 $\times$ 变倍率〕(舍入)$-$〔A 点 x 坐标 $\times$ 变倍率〕(舍入)

这里〔P〕(舍入)表示 P 经四舍五入后取整数值。

这样变倍过程中舍入误差不会有任何积累，永远不会超过坐标网格上的半个格范围。

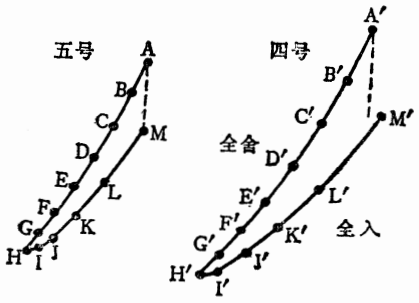


图13-13　误差积累的例子

（二）保证规则笔画宽度的一致性

用上节所述的结点变倍法对规则笔画‘横’作变倍时将产生严重问题。例如，一个字中的两笔横，本来宽度是一样的，在五号字时宽度均占两格。但由于这两笔横的 y 坐标不同，用结点变倍法变成小五号字（变倍率为 0.854）时，由于舍入，很可能一笔横的宽度仍占两格，另一笔横的宽度变成一格。本来宽度一样的两个规则笔画变倍后宽度会一样，这将对文字质量产生不良影响，而用结点变倍法是无法防止这种失真的。

因此，对于规则笔画横、竖、折的宽度部分，我们不采用结点变倍法，而按下述公式计算：

变倍后的宽度＝〔变倍前宽度×变倍率〕(舍入)。

（三）小号字横的宽度的控制和笔锋的细致描述

宋体字可能缩小到六号、小六号、甚至七号。在变倍过程中会引起某些部分变得过分密集，尤其当一个字内横的数量很多时。变倍后既要保持横的宽度一致，又要使小五号字笔画仍很清晰。我们在字体横的压缩信息中，用两位二进位细致刻画横宽：00表示1.7格宽；01表示 2 格宽；10表示2.6格宽；11表示 3 格宽。这里的格都指五号字96×96网格中的格。当然，00和01编号在五号字时只能都是两格宽，显不出差别，因为分数无法体现。但变成六号时，01编号仍是两格宽，而00编号成为一格宽。对于笔画很多，尤其是横很多的汉字，00编号的横将使这一汉字在各种小字号下都保持清晰美观。

对于黑体字，我们允许对横和竖的喇叭形笔锋作细致描述，允许半格、一格、一格半宽几种程度的喇叭形。这样当黑体字放大后，不同程度的喇叭形将能表现出笔锋的细致差别。

确保各种字号下的文字质量是高分辨率汉字信息压缩技术需要解决的最重要问题。上述这些防止失真的措施用中规模集成电路或常规方法实现，将是麻烦的，很难做到低价格和高速度。而用AM2900位片器件，因为 AM2900 微指令格式和内容可由系统设计者根据需要自己确定，可以做到低代价而高速、高质量的字号变倍效果。

三、用访问 PROM 的方法实现变倍计算

结点变倍时要把 x、y 坐标乘上变倍率，然后舍入取整。用 Am2900 位片实现乘法的拍节数与字长有关，字长 n 位需 n 拍，速度太低；用专门的乘法器组件可以加快速度，但仍不满足要求。我们用 PROM 代替乘法操作：对于每种变倍率，我们把 0～127 的坐标值乘此变倍率，然后舍入取整；把这128个数事先算好，写入PROM内。因此，对于每种变倍率，需128× 8 位PROM。这里，PROM 地址即变倍前坐标值，PROM 内容即变倍后坐标值，只需一拍就完成乘法和舍入取整操作。仔细分析表13-1的字号，可以看出，只需对小四号、四号、三号和小一号这四种字号建立PROM变倍表，正好占一片 512× 8 位的 PROM。而七号、小六号、六号和小五号这四种小字号可通过 访问上述四种字号的PROM，然后往低移一位得到变倍值；二号以上的字号都是基本字号的 1.5 倍、2 倍、 2.5 倍、 3 倍、 3.5 倍或 4 倍，很易实现而不必用乘法。例如

坐标×2.5＝坐标往高移一位后的结果＋坐标往低移一位后的结果。

13.4 逐段生成汉字技术和复杂版面形成技术

13.4.1 高分辨率汉字字模的两级存储和调度

滚筒式激光照排机的滚筒每转一圈，在底片上扫出一条线，同时带有激光扫描头的丝杠匀速前进。滚筒旋转电机和丝杠前进电机来自一个主振源，以保证严格的同步和得到所需的合作系数。平板转镜式激光照排机采用多面转镜扫描，每旋转一面，底片上扫出一条线，同时底片匀速前进。上述两种类型的照排机的副扫描系统都采用匀速运动的方式，这种方式机械简单，与“走走停停”的扫描方式相比，可得到更高的精度，但却给照排控制器的设计带来困难。由于在扫描一整页的过程中不许停顿，控制器必须保证任何时刻都及时提供版面点阵信息。这种控制方式比第三代 CRT 照排机要复杂得多。CRT照排机采用电子束定位，扫完一个字再扫另一个字，扫完一个字后若下一个字的点阵尚未准备好，则CRT可等待，暂停扫描。

为了适应第四代激光照排机的上述特点，在汉字字模的存储和调度方面要采取相应的措施，使得整个照排过程中（即输出一本书或一版报纸的整个过程中）不需要访问磁盘，因为从磁盘上取一个汉字字模需10毫秒到 100 多毫秒，一般情况下跟不上激光扫描速度。为此，本系统中汉字字模分两级存储，如图 13-14 所示。

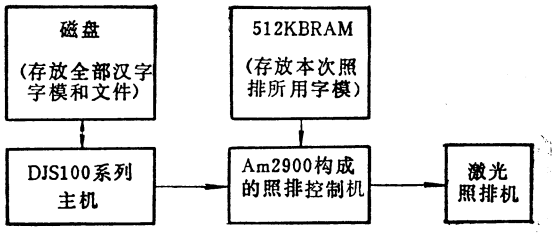


图13-14 汉字字模的两级存储

照排控制机有一个 512K 字节的局部存储器，存放本次照排所用的全部字模。一次照排可连续输出 n 页，n 可达几百。在输出开始前，由DJS100系列主机的软件把这次照排所用的全部字模从磁盘上挑出，经过数据通道，送到照排控制机局部存储器内；以后在连续输出 n 页的整个过程中，不再为读取字模而访问磁盘。需要注意的是，尽管印刷用汉字字模的数量极大，字体有十多种，每种字体需 6000～10000 字以上，但一本书中所用的字模数是不多的，一般不会超过 4000 个。据统计，四版报纸通常不超过 3000 字

模；一本几十万字的书往往以书版宋体作为正文字，用3000多个不同的字，以黑体字为标题字，用几百个不同的字。采用高倍数信息压缩技术后，一个汉字平均只需100字节便能精确表示，并能放大缩小。这样，512KB的局部存储器能够放下一次照排所需的全部字模。

主机软件对本次照排的文字进行分析时，对所用的每种字体建立一个布尔串，每个汉字的内部码对应布尔串中一位："1"表示该汉字出现过，"0"表示该汉字未出现过。然后根据此布尔串从磁盘上挑出本次照排所用到的全部字模，送到照排控制机局部存储器；同时建立汉字入口地址表，也送到照排控制机局部存储器的有关区中。每个汉字不管本次照排中是否用到，都对应入口地址表中的一个入口（占三个字节），若用到该汉字，则入口指示该汉字字模（压缩信息）在局部存储器中的起始地址。由于局部存储器中的字模位置随每次照排所用字模不同而变化，因此每次照排开始前必须由主机软件重建入口地址表；当然，对于未用到的汉字，因入口地址表的内容是不起作用或无关紧要的，故软件不必处理。

为了照排一本书，统计和建立布尔串，从磁盘上挑选字模，建立入口地址表这些工作全部加起来，一般只需20秒，而激光照排机输出八页32开书就需250秒，所以前者所花的时间是微不足道的。为了减少主机开销，我们还采取下述改进措施。

一本出版物往往有一种主体字，也即正文所用的字体。例如报纸通常以小五号或五号报版宋体作为主体字，科技书或文艺书一般以五号书版宋体作为主体字，小学课本常以楷体作为主体字，而各种文件常以仿宋体作为主体字。排版语言中有一个"版心注解"，除指示版心大小外还指示主体字及其字号。主机软件在照排开始前永远把内部码为0～3071的主体字从磁盘上取出，不加挑选地送往照排控制机局部存储器，而不管这些字是否在本次照排中都用到。据粗略统计，0～3071号主体字一般占出版物中所用字模的80%以上。对于这些字，统计和建立布尔串、挑选、建立入口地址表等工作统统省略了，因而显著减少了主机开销。

对于内部码为4096以上的汉字，不管是否用到，都占入口地址表的一个入口，也是浪费的，因为这种罕见字的出现概率为万分之一。软件扫视文字过程中，将把内部码为4096以上的汉字转换成一个新序号。例如本次照排用到五个罕见字，其内部码为4176、4290、5120、5122、5230，软件把这些内部码相应地转换成4096、4097、4098、4099、4100，这是本次照排中出现的罕见字的序号。这样罕见字入口地址表只需五个入口，节省了存储，但增加了处理时间。由于罕见字出现概率极低，这样做是值得的。

尽管本系统收容的汉字字模可多达100万个字头以上（计算了不同字体、不同字号后），依靠高倍数汉字字形信息压缩技术、硬件文字变倍技术和上述字模调度技术这三者的结合，我们首次实现了"照排过程中不需访盘"。这不仅提高了输出速度，克服了系统瓶颈，简化了照排输出设备；而且使得激光直接雕版设备、激光扫描的计算机输出缩微卡设备（即新的一代COM）等易于实现；这一方法还使实现高速大样输出机成为可能。大样输出机是用激光束在普通纸上印出逼真大样的设备，它与国外的激光印字机在功能上很不相同：通常的激光汉字印字机只能输出一、两种字体，三、四种字号，不太复杂的版面；而大样输出机则要求输出与最终底片形式一样的逼真大样，应允许十多种字体，十六种字号，能输出任意复杂的版面。八四年以前国外的照排系统都利用照排

机在照相纸上输出逼真大样。大样一般需多次修改才能满意，而照相纸成本高，使得照排系统的运转维持费用高于手工拣铅字排版费用，这不利于推广照排技术。采用硒鼓转印方式可在普通纸上输出文字，但带有文字潜影的旋转硒鼓一般是不能停顿的，很难设计成“转转停停”的工作方式，而缓冲一版报纸的点阵则需较大的存储量。“输出过程中不需访盘”的做法有助于实现廉价高速的大样控制器。

13.4.2 适合于激光扫描的版面描述方法

准确描述版面形式的信息称为编辑信息，它是由主机软件产生并提供给照排控制机的；照排控制机将根据这一信息，产生编辑人员所需的版面点阵。第三代CRT照排机采用逐字的扫描方式，因而编辑信息十分简单：对于每个版面上出现的字模，只需指示该字模在版面上的绝对 x 和 y 坐标；编辑信息内汉字出现的次序与文章内汉字次序是一致的。例如报纸上的一篇文章对应的编辑信息，通常是先指示标题字的位置（标题可能是竖排的），然后按文章内汉字次序逐个指示它们在版面上的位置（正文可能是分栏排的）。CRT 逐字扫描的次序与编辑信息内汉字次序是一致的。这样主机软件处理和照排控制硬件都很简单，没有任何困难。激光照排机只能逐线扫描，扫描一页报纸时只能从最左到最右、一线一线地扫描，而不可能象CRT那样，先扫描竖排标题大字，再扫描横排正文小字。但我们又要求激光照排机能够输出任意复杂的版面，这就需要在扫描控制上精心设计。

假如汉字字形点阵采用国外流行的低压缩倍数的黑、白段压缩方法，即记录一个字的每条水平方向扫描线内各个黑段的始点和长度，则激光扫描的次序与汉字压缩表示的信息次序基本一致；但我们采用的是高倍数压缩方案，压缩信息以笔画为序，而不是以点阵的扫描线为序。按照 Jackson 关于数据结构的观点，这种情形下在输入和输出数据之间既有边界冲突，又有次序冲突。

假如照排控制机内有一个 10^8 位的大容量内部存储器，则问题又好办了：我们可以事先产生一整页的全部点阵信息并存放在这一大内存中，然后再命令激光束开始扫描。但10^8位的RAM 代价太大，采用磁盘缓存速度又太慢。为了用低代价实现任意复杂版面的输出，采用了两项新技术：软硬件相结合的版面描述和处理技术，以及硬件逐段生成汉字技术。

为了节省照排控制机的扫描缓冲存储器，只缓冲相邻两个段，规定每段包含32条扫描线。照排输出版面的宽度为280毫米，分辨率为 29.2线/毫米，因此扫描缓冲存储器 SS 的容量为 $2\times32\times8192$位。主机软件把版面划分为行，这里的行与出版物中的自然行基本上是一致的，只是每行所含线数必须是32的倍数，一行可含 1～16 段，每段为 32 线。汉字点阵复原设备每次只生成一个汉字的一个段（32线）。例如一行内包含 N 个汉字，则生成汉字点阵的次序如下：先生成第一个汉字的第一段，再生成第二个汉字的第一段，直到 N 个汉字的第一段都生成完，并放在交替工作的扫描缓冲存储器之一中；当另一扫描缓冲存储的内容已全部由激光照排机输出完时，再生成第一个汉字的第二段，第二个汉字的第二段，直到第 N 个汉字的第二段，依此类推。这种方法称为逐段生成和缓存汉字点阵技术。

编辑信息以行为单位，由下列六种标记组成（每个标记占两个字节，其中最高三位

指示标记类型)：

（1）行开始标记，总是处在一行编辑信息之首，指示该行所含的段数；

（2）字体号标记，指示字体和x、y方向字号，当x字号与y字号相同时形成正方形字，当不同时形成长体字或扁体字；

（3）汉字内部码标记，微程序将根据内部码找到该汉字压缩信息在局部存储器中的始址；

（4）汉字点阵在版面上x位置标记，每个汉字都伴有此标记；

（5）汉字点阵在版面上相对y位置标记，如果后面的汉字y位置均相同，这个标记可为这串汉字所共享。

（6）行结束（或页结束、或底片结束、或照排结束）标记。

这里需要解释的是上述第五种标记，该标记指示本行始行与此标记后面的汉字的点阵始线之差：若此值为负，则表示该汉字在本行始线之下；若此值为正，则表示该汉字始线在本行始线之上，也即该汉字跨两行，上半部分点阵在上一行内，下半部分点阵在本行内。因为报纸上有竖排标题字，科技书中的数学公式、表格和框图十分复杂，故很难区分自然行。在这种版面情况下，软件无论怎样仔细地划分行，也不可能避免一个字跨两行的现象，因此，编辑信息的设计要适应这种情况。利用上面六种标记，只要版面上出现的文字和图形都是由字模组成的，就可以准确描述任意复杂的版面。这种形式的编辑信息适合于激光逐线扫描，有利于简化照排控制机硬件；同时又比较紧凑，主机软件也不难设计。

13.4.3 逐段生成汉字点阵

照排控制机由两台Am2900位片微处理机构成。微处理机G的微指令长56位，共2K条；微处理机F的微指令长48位，共512条。两台微处理机的运算器都是16位长，各用4片Am2901A；微程序控制器用一片Am2910。

微处理机G的功能是：

1）通过数据通道每次从主机内存中取出一行编辑信息，暂存在局部存储器的编辑信息区内；

2）逐个处理编辑信息的各种标记：

（1）遇行开始标记，微程序把标记中指示的段数往高移五位后送段数寄存器，并置段号寄存器为0（这里的段数、段号寄存器都是Am2901A的内部寄存器)；

（2）遇字体号标记，微程序把字体和y字号送内部寄存器R_7的高位，把x字号送R_6的高位；

（3）遇汉字内部码标记，微程序根据R_7内的字体和这里提供的内部码，找到此汉字的入口地址表，从表中取出此汉字压缩信息始址，然后转向汉字点阵复原微程序（见下面3)；

（4）遇x位置标记，微程序把此标记所指示的x坐标（13位）送给F微处理机；

（5）遇相对y位置标记，把段号寄存器的内容与此标记指示的相对y值相加后送R_{14}；

（6）遇行结束标记，令段号寄存器内容加32，并与段数寄存器内容比较，若相等，

则表示本行内的所有段都已生成完毕，再取一行编辑信息，转向1。

3）实现汉字点阵复原步骤中的开头两步是：

第一步把十分紧凑的规则笔画压缩信息和多种形式的不规则笔画压缩信息转换成标准形式的向量串；

第二步对向量串中的每个向量根据字号进行变倍，然后判断变倍后的向量是否在当前段内，若在本段内，则把该向量转换成最近似的阶梯点，在经过的每个阶梯点上写入两位标记，形成标记点阵。标记点阵中每个点占两位二进位，表示该点的下面四种情况：空点（00）；黑段始点（01）；黑段终点（10）；孤立黑点（11）。详见后面图12-10。

把标记点阵转换成最终输出点阵的操作很简单，由微处理机F实现（将在下节介绍）。标记点阵存储器位于微处理机G和微处理机F之间，由两个独立的存储体组成，每个容量为4K×8位，可存放一个特大号字（512×512点）的标记点阵的一段（32线）。

微处理机G实现点阵复原步骤的第二步中需判断向量是否在当前段内，其判断算法如下：

R_{14}中存放的是下限值，指示当前要处理的那段在汉字点阵中的y坐标范围的下限。例如图13-15的'人'字，相对y位置标记指示相对值为－20。当处理第1段时，段号寄存器内容为0，R_{14}内容为0－20＝－20，第一段在'人'字点阵中的y坐标范围为〔－20，11〕，只有向量AB、BC和向量TU、UV、VA在第一段内；当处理第2段时，段号寄存器内容已变成32，因而寄存器R_{14}的内容为32－20＝12，第2段在'人'字点阵中的y坐标范围为〔12，43〕，依此类推。

假定向量结点（始点或终点）变倍后的y坐标在R_{13}内，则判断该结点在段上、段内或段下的流程如图13-16所示。

按照表13-8来决定一个向量是否在段内，仅对段内向量才转换成最近似的阶梯点。

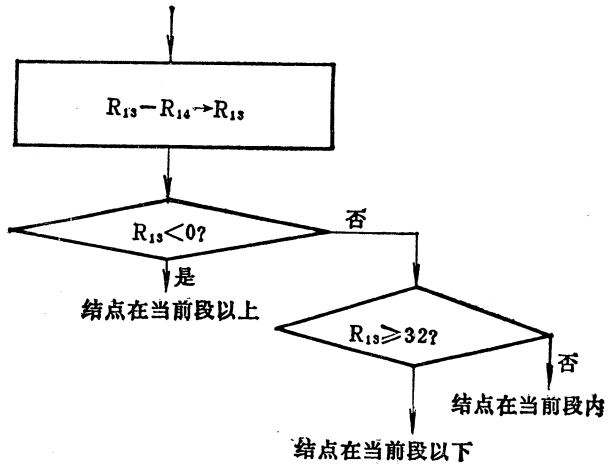


图13-15　逐段生成汉字的例子

判断向量是否在段内是一频繁出现的关键性操作，必须以高速实现，对应上述流程的微程序只需2～3拍就完成判断。由于一段只有32线，所以造成一个汉字的向量中的大多数不在当前段内，可以跳过，不必转换。一个汉字将被反复生成多次，每次只形成点阵的32线，这样无疑要求点阵复原设备有更高的处理速度，而双极型位片微处理机很适合于提高特种运算的速度。还需指出，即使扩大扫描缓冲器SS，使之缓存相邻两行（而不是两个32线），"判断向量是否在当前范围内"和"部分生成点阵"的做法仍是必需的。因为一个汉字跨两行的现象无法避免，必然要求点阵复原设备能够只生成在当前行内的那部分点阵。

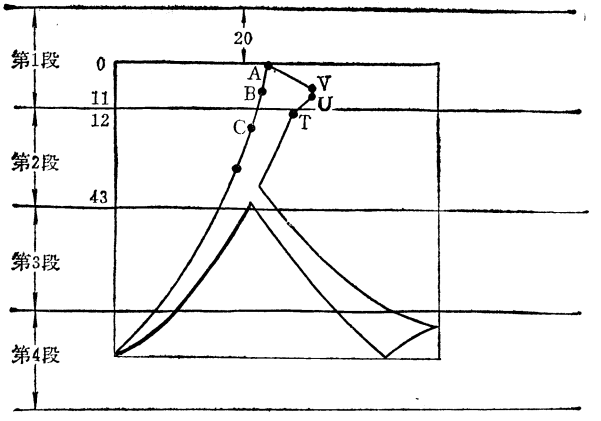


图13-16　判断结点在段上、段下或段内的流程

表13-8　向量是否在段内判别

向量始点	向量终点	相　应　的　操　作
在段内		转换成阶梯点
	在段内	转换成阶梯点
在段以上	在段以上	跳过
在段以上	在段以下	转换成阶梯点
在段以下	在段以上	转换成阶梯点
在段以下	在段以下	跳过

13.4.4　最终输出点阵的形成和激光扫描控制

微处理机F的框图如图13-17所示。其功能是

（1）实现汉字点阵复原步骤中的第三步（也是最后一步）：把标记点阵转换成最终输出点阵，并根据微处理机G提供的X位置标记中的x坐标，把形成的点阵送入扫描缓冲器SS的有关单元中；

（2）控制激光扫描。滚筒式四路激光平行扫描的激光照排机每扫完4×4个点（约8.4微秒）向微处理机F发出请求，F从SS中依次取出4×4位信息，提供给激光照排机。

标记点阵是一线一线地存放在4K字节存储器中的：第0～127单元存放点阵的第1线，…，第3968～4095单元存放点阵的第32线。而SS存储器则按四线方式存放，一个单元中16位分属四条扫描线。因此标记点阵不应该一线一线地转换，而应先转换第一线的头四个点，第二线的头四个点，第三线的头四个点，第四线的头四个点；接着再转换第一线下面的四个点，第二线下面的四个点…。为此需设四个层数计数器，分别记录平行四线的层数。

微处理机G和微处理机F使用同一时钟脉冲，独自平行地工作。当微处理机G对下一个汉字执行复原步骤的开头两步，把两位标记写入4K字节存储器的同时，微处理机F对上一个汉字执行复原步骤的第三步，从另一4K字节标记点阵存储器中逐个取出字

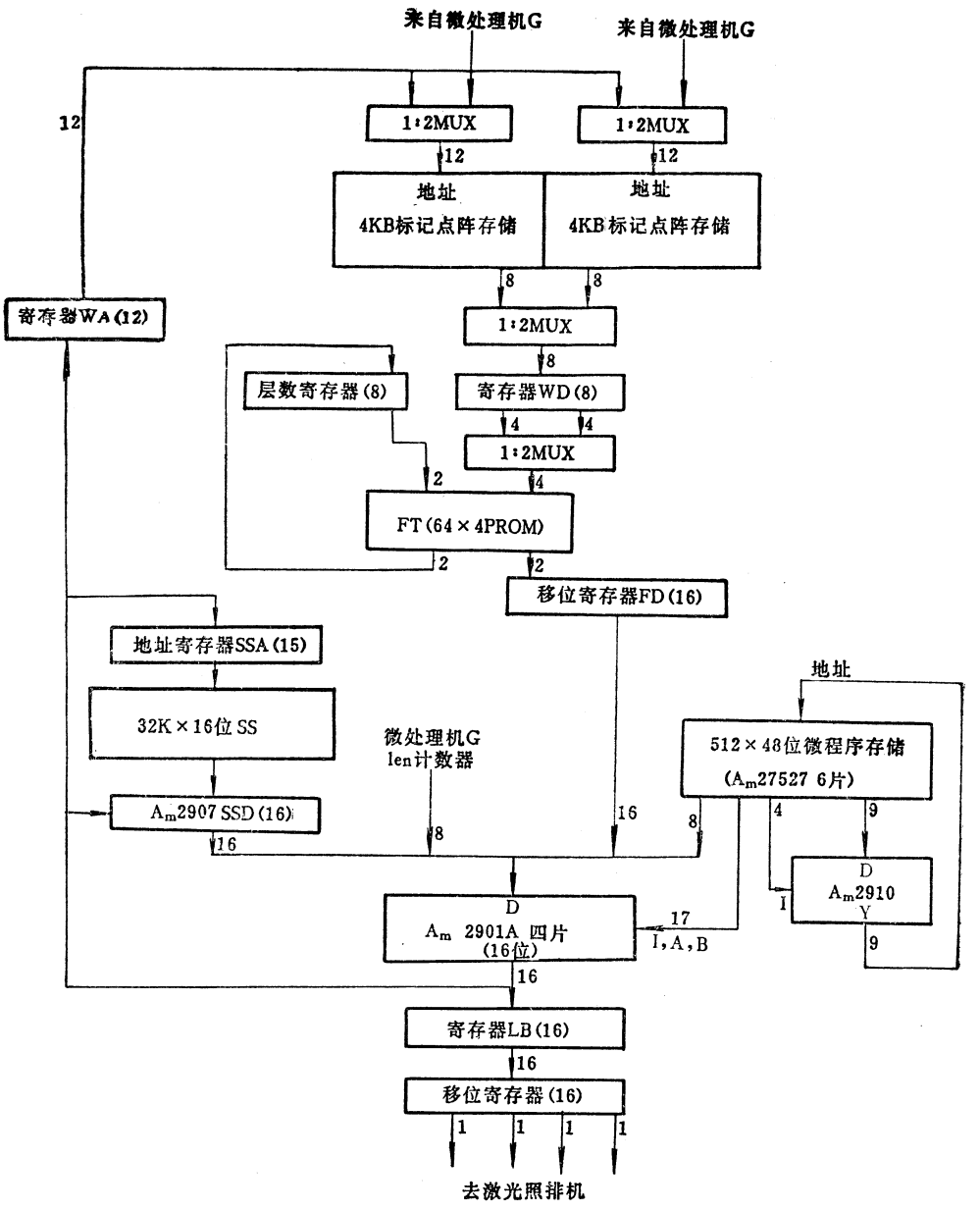


图13-17　微处理机 F 框图

节，转换成最终输出点阵送 SS。两台微处理机通过通信触发器 FBUSY 协调工作。

13.4.5　Am2900微处理机系统的设计方法

位片式微处理机系统的设计包括逻辑设计（画逻辑图）和微程序设计两部分，它不同于传统的硬件设计，也不同于软件设计。通常的软件设计是在固定的机型上进行的，指令系统是确定的；而位片微处理机系统中尽管 ALU（例如Am2901A）和微程序控制器（例如Am2901）的内部操作是组件内做死的，但微指令的格式、操作内容是由系统设计

者确定的，可以千变万化。因而逻辑图、微指令内容和微程序编制这三者互相影响，设计工作更为复杂。软件设计强调的是程序结构反应问题结构，可读性好，易于理解、修改；而把程序运行速度和长度放在次要地位或稍后再做这种优化。位片微处理机系统的设计则不然，必须把运行速度放在十分重要的地位；程序长度也应力求缩短，以节省容量有限的双极型PROM。因而往往采用“最关键部分先设计”的方法，而不是单纯的从顶到底的设计，实际工作就是这样做的，这是设计方法上的一大差别。

根据我们的经验，设计步骤大体如下：

（1）写出规范说明，包括输入输出数据结构和主要运算步骤；

（2）初步估算关键运算流程的拍节数，分析所选择的位片微处理机器件是否能满足速度要求；

（3）画出基本的逻辑框图，初步确定微指令格式和操作内容；

（4）仔细编出速度最紧张部分的微程序和最核心部分的微程序，并反复修改逻辑图和微指令格式和内容，在此基础上再修改和比较不同方案下的微程序，以求得最佳或接近最佳的方案；

（5）画出详细的逻辑图，确定微指令格式和内容；

（6）编出全部微程序，在编制过程中对逻辑图和微指令格式和内容作小的调整。

硬件故障检测和诊断措施，以及微程序的调试工具是十分重要的，这些考虑对逻辑图有影响，必须在第5步中加以研究。上述第4步是很关键的，在这一步中要确定哪些操作是频繁出现的，应增加哪些专门硬件，微指令中应设哪些专门位，控制执行哪些操作。这一步设计中最能体现位片微处理机的灵活性和高速度。

对于很多应用领域，位片微处理机能得到比单片MOS微处理机更高的处理速度，这不仅是因为双极型位片微处理机有更高的时钟频率，更重要的是位片微处理机具有下列灵活性：

（1）微指令长度和内容由设计者确定，因而很易增加附加硬件，在微指令的专门位控制下，附加硬件可与Am2901A和Am2910同时协调地工作，以提高关键操作的速度；

（2）Am2901A可选择任一外部来源作为算术和逻辑运算的操作数；

（3）Am2910可选择任一外部状态作为条件码，并检测此条件码，以决定下条微指令的地址。

这些特点使得Am2901A，Am2910与附加硬件能在每一拍上进行协调，这是很重要的。

13.5 编辑排版软件系统

13.5.1 编辑排版系统结构

一、编辑排版系统的硬件构成

计算机激光汉字编辑排版系统硬件构成，如图13-18所示。

本系统主机采用国产的DJS153机或进口NOVA3、NOVA4机，每台主机可带联机的汉字终端6～8台，每台终端由汉字键盘与汉字显示器组成，并配有一台针式汉字打印机，或几台终端合用一台针式汉字打印机。20兆字节以上的活动头磁盘，一部分作为

精密汉字的字模库，存放各种字体的汉字压缩信息，其余大部分用作存放系统程序、排版文稿及中间结果。照排控制器（简称TC）作为一个外部设备接在主机上，由它将汉字压缩信息还原成照排用的精密汉字点阵，并控制激光照排机将汉字按软件安排的位置在底片上扫描打点，形成一张可供印刷制版用的胶片，还可控制大样输出机在普通纸上印出一张逼真的排版结果——报纸版面或书版版面。

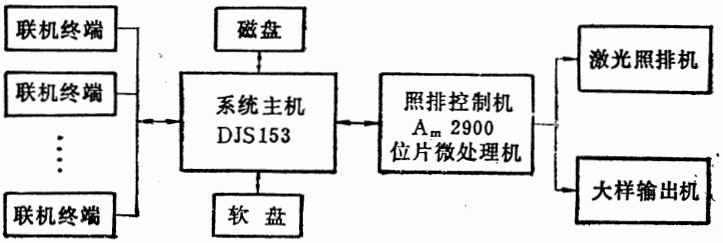


图13-18　系统硬件构成图

二、编辑排版系统的工作流程

用户首先在排版的文章中将排版要求以一定的格式（这些格式称为排版注解，将在13.5.3中叙述）写明，然后在脱机工作的汉字终端上，用汉字键盘将文章与排版要求一并输入到汉字终端的软盘存储器中。汉字终端的显示器可以在击键的同时将文稿与排版要求显示出来，以便随时纠正。在整篇文章输入到软盘后，可在显示器上重新显示，也可在汉字打印机上印出，以供校对。这种带有排版要求的文章称为小样文件。此时显示或印出的文字都是仅供辨认的简易汉字。脱机终端具有增、删、改等编辑功能，可以方便地对小样文件进行修改，当小样校改完毕后，即可在联机的汉字终端上向主系统发输入命令，将软盘上的小样文件输入到主系统的磁盘中。

一本书（或整版报纸）的所有小样输入完毕后，就可在联机终端上发排版命令，主系统按小样文件中的排版要求给出每个字的字体、字号信息，计算出它们的位置，形成供照排用的输出信息。当排版完成后，若是排书，在终端上给出完成回答，并显示出本书总页数；若是排报，给出每篇文稿是否正好适合划定的版面的回答（用针式打印机打印），并作版面情况的显示，使用户对文章在版面上的分布情况有所了解。当用户写错了版排要求，系统无法实现时，终端也会在针式打印机上将错误信息一一列出。若用户对版面不够满意，或有文章安排的位置不合适，或有错需要将小样修改时，可直接在联机终端上将文章显示出来，进行校对修改，改完后再发排版命令，由系统重新进行排版，直到得出一个满意的版面为止。整个过程只需几分钟即可完成。

为了看到排版后的逼真的报纸版面，系统配有大样输出机，在联机终端上发‘出大样’命令，就可将排版结果在大样输出机上输出，供用户仔细校阅，校阅满意后，再发照排命令，由激光照排机将版面输出在胶片上，系统的工作即告完成。当对大样甚至胶片上的结果有不满意处，仍可对小样进行修改，并重新排版，当然这种修改愈早愈节约。用户得到胶片后即可制版印刷。今后还可实现激光雕版，直接输出印刷用版，免去显影、定影及用胶片制版这些工序，从而使流程缩短。

三、面向用户的自动照排虚拟机

编辑排版系统配有一个大型软件，这一软件由专用操作系统、用户命令处理系统（包

括排版命令处理程序——排版语言编译程序）及汉字终端处理程序等组成，整个软件总长约140K字（故障检测、字模自动生成、字模调试等服务性程序尚不计在内）。

在这一软件系统的支持下，整个编辑排版系统面向用户的是若干台独立的自动编辑排版虚拟机，每个联机终端就是一个具有编辑校对、排版等功能的虚拟系统，排版结束后利用激光照排机（或大样机）输出。虚拟系统以人机对话方式完成上述功能，用户只须了解人机对话格式及排版要求的写法，就可使用本系统。

用户与系统的人机对话是以用户命令与机器回答组成的，也就是说虚拟系统由若干个用户命令构成，命令按其结构分共有两类：

1）无参数命令　对应键盘上设的功能键

2）一般命令　格式为：

〈命令动词〉〈参数表〉

其中，命令动词由两位助记的大写拼音字母组成，

〈参数表〉::＝〈参数〉〔{,〈参数〉}〕〔{㊀〈参数〉}〕

〈参数〉::＝〈文件名〉|〈字母数字符号串〉|〈文件名〉

〈字母数字符号串〉

〈文件名〉::＝〈字母〉〈字母数字串〉

（字母数字串长≤7）

㊀是键盘上表示空一格的符号，{ } 表示其中结构可能重复，〔 〕表示其中结构可能不出现（下同）。

〔例1〕　命令：写入

这是第一类命令，它对应键盘上一个‘写入’功能键。这一命令是要求将本终端的显示器上当前幅面内容写入文件中，刷新原来所存的文件内容，一般用于联机修改之后。此命令执行后，若为正常情况，在显示器上回答“完成”，异常情况，指明故障性质。

〔例2〕命令：PB SB〈文件名〉

这是第二类命令，PB为命令动词，是‘排版’的二位拼音，参数SB表示书版，文件名是组版文件的名字。在组版文件中列出排版的小样文件名，并规定书刊的大小规格，页码排法、有无书眉等。

此命令根据组版文件中所列小样文件的顺序，逐一按文件中的排版要求进行排版，确定全书每个字符所属页面及其在页面上的位置。执行此命令后，若正常结束，在终端的显示器上指明排版总页数，并在显示器上准备好下一步骤的命令，同时在针式打印机上印出每个小样文件的起始页码，以便用户编排目录，还印出供调整全书总页数等参考用的其它内容，包括留有较多空白处的文件名与页码、文件开始于一页较末尾处的文件名与页码。若需要调整总页数，可根据这些资料修改某些小样文件，也可修改组版文件，重发排版命令再排。当排版过程中发现排版要求写错或不合理，在终端的显示器上指出‘排版有错’，并在针式打印机上印出这次排版所发生的所有错误性质和发生错误的位置，以便用户修改后重新排版。

四、虚拟机的功能与用户命令一览表

自动照排虚拟机具有编辑校对、排版、照排等功能。可分以下两类：

（一）编辑校对功能

除了与脱机终端（见13.5.2）有相同的增、删、改等编辑功能外，还可以实现：

（1）保留某一小样文件的一段内容（任意字数），插入该文件或其它文件的任意指定的位置上；

（2）将若干个小样文件合并成一个文件；

（3）将一个小样文件分离成两个文件（任意指定分离点）；

（4）将一个小样文件复制成另一个文件，原文件保留不变；

（5）更改一个文件的文件名。

（二）排版功能

（1）排三十二开或十六开的书版（允许各种尺寸）；

（2）排整版的八开报纸，并在显示器上显示版面安排情况，包括每篇文章的位置、文章中标题的尺寸与位置、花边长短与位置（今后还将设计大报版的排版程序）；

（3）将排版结果在大样机上输出大样；

（4）将排版结果用激光照排机输出底片，供制版印刷；

（5）利用滚筒扫描装置输入图片、照片，在终端上对图片、照片进行剪裁、放大或缩小，并与排版结果一起输出大样或底片。

以上功能由下列各用户命令完成。

1．第一类命令（见表13-9）

表13-9　第一类命令

命令名	说明
前进一页	前进一个显示幅面
后退一页	后退一个显示幅面
前进一行	幅面前推一行，补一行新内容。
后退一行	幅面后退一行，前面补一行，
写入	当前幅面内容写入主系统中
保留始	当前光标指示处为保留始点
保留终	当前光标指示处为保留终点
保留且删终	同上，但删去文件中保留始点到终点内容
插入	将本终端保留内容插入光标指示处之前
改完	表示本文件修改完毕，给软件整理文件的信号，以便再次显示时段落整齐

2．第二类命令（见表13-10）

表13-10　第二类命令

命令名	参数	说明
SR(输入)	A/B〔空文件名$_1$，…，文件名$_n$〕($n\leqslant 11$)	将软盘A面或B面的全部或指定文件的信息输入
SC(输出)	A/B空文件名$_1$，…，文件名$_n$　($n\leqslant 11$)	将指定文件的信息输出到软盘的A面或B面
XS(显示)	文件名〔，n〕	从文件始点起或第n幅内容显示
QJ(前进)	n	前进n幅后显示
PB(排版)	SB|BB空组版文件名	排书版（用SB）或报版（用BB）

（续）

命 令 名	参　　数	说　　明
BM(版面)	组版文件名	作报版版面显示
ZB（准备）	组版文件名〔，页号〕〔·目录文件名〕	可从指定页号开始作准备，组成适合于激光照排的信息
ZP(照排)	组版文件名〔㊀页号〔·数〕〕	可从指定页号起补拍几页（页数用“·数”表示），可以补拍若干组，也可从某页起拍到末页。
DB (大样准备)	参数同ZB	
DY (大样输出)	参数同ZP	
PW(排完)	组版文件名	表示本次照排已完成，撤离所用的全部文件
GM(改名)	文件名$_1$，文件名$_2$	将以文件名$_1$命名的文件改名为文件名$_2$
HY(合一)	文件名$_1$，文件名$_2$…文件名$_n$〔；文件名〕	将n个文件按顺序合并为一个文件，有‘；文件名’时以此文件名命名，否则以文件名$_1$命名
FZ(复制)	文件名$_1$，文件名$_2$	第一个文件复制后以文件名$_2$命名
FK(分开)	文件名$_1$，〔文件名$_2$〕或〔文件名$_1$〕，文件名$_2$	从光标所指位置处将正在显示的文件分开，分别以文件名$_1$，文件名$_2$命名，缺一个名字时，所缺者以原文件命名
CL(撤离)	文件名$_1$，…文件名$_n$（$n\leqslant 11$）	撤离指定的文件

注：照片、图片的编辑命令尚未列入。

13.5.2　排版语言与排版编译程序

一、CL排版专用语言的结构特点

CL排版专用语言（以下简称CL语言）是用于中文书报的排版语言，目前主要用于排普通书刊与八开报纸，除了能排这类书刊的一般格式外，还有处理各类表格（如无线表、有线表以及带斜线的有线表）、划线、划方框图，以及实现用于期刊杂志的分栏或对照等功能。

CL语言由排版注解与字符组成。

（一）排版注解

排版注解相当于一般程序语言中的语句，它用以表示文章的各种排版要求。

排版注解由一般注解、注解符号及间隔符组成。一般注解的形式为：

〖〈注解名称〉〔〈参数表〉〕〗

其中，〈注解名称〉::＝〈大写字母〉〈大写字母〉

〈参数表〉::＝〈参数〉{〈参数〉}

〈参数〉::＝〈字母数字符号串〉

由两位大写字母组成的注解名称，是助记的拼音词，特殊括号〖 〗用以区分注解与字符。

注解符号是无参数注解，如换行、换段，它们用特殊符号表示，间隔符用于表格等结构中，以区分各项的内容。

〔例 1〕　页码注解　例如：〖YM5BZ·〗表示全书页码用五号白正体数字，数字两

边加圆点“·”。这里YM是注解名称，是页码的两位拼音；5BZ·为参数表，由参数5、BZ和·三者组成，BZ是白正体的二位拼音。页码在换页时自动插入，单页位于页右，双页位于页左，也可位于中间，只要另加参数说明；没有书眉时页码放于页末，有书眉时放在书眉线上。

〔**例2**〕 空格注解 例如：我〖KG3〗们表示“我”字与“们”字之间空开三个字的宽度。其中KG是注解名称，3是参数。

又如：社〖KG1·4〗会主义好，表示社字后四个字每字之间空一个字宽。应排成：

社　会　主　义　好

CL语言的每个注解有确定的定义域，它们既是独立的，又允许某些注解的作用域相互嵌套，某些注解允许递归引用。

（二）字符

字符相当于一般程序语言中的变量或数据，它包含汉字、标点、符号、字母、数字等类型，每种类型在编译时作相应的处理。可分为以下四类：

1）标点类 在换行时需判别是否属于行末禁排或行首禁排，并作对应的禁排处理；

2）字母类 区别于汉字，应作相应的字母宽度调整，并不得任意换行；

3）数字类 应区分单个数字与多个数字，后者一般应排对开数字，并在这一数字串前后各加四分空，且不得任意换行；

4）符号类 应判别某些符号的特殊要求，如%不能与其前面的数字分排在两行。

此外，系统接纳下列两类复合字符：

（1）盘外字符串。它用以表示键盘中未包含的字符，它用一个以上的键盘字符表示，应译成对应字符的内部码，并决定其类型。如((5))代表圈码⑤，((>>))代表≫(大于大于)等，特殊括号(( ))用以区分复合字符与普通字符。

（2）注解型字符串。它等价于包括注解在内的字符结合，例如((C↑2；ij))等价于下列内容：

C上角码区2〖KG-*3〗下角码区ij

〖KG-*3〗是退回到字母C之后，使i与上角码2对齐，这就排成$C^{2}_{i_{j}}$。

CL语言广泛地采用缺席属性，以便用户使用方便。CL语言的缺席属性共有两类：

1）一般注解中普遍允许无参数，对于无参时规定特定意义。如行距注解一般格式为〖HJ〈行距参数〉〗，当无参数时〖HJ〗表示恢复到版心行距（即由版心注解中规定的行距）；又如报纸的标题注解一般格式为〖BT〈行距比例〉〗，注明每行标题之间及标题上空与下空之间的比例，当无参数时〖BT〗表示采用当前所排报纸的常用比例，如对《参考消息》按以下规定排版：

（1）只有一行标题，上、下各空标题区空白的50%（以下百分比同此含义）；

（2）有两行标题，上空35%，两行标题之间空30%，下空35%；

（3）三行标题空白比例为30%，20%，20%，30%；

（4）二行标题下有提要，提要行距固定后，比例为26%，20%，24%，30%；

……

2）允许在特定情况下省略注解，注解省略后有确定的含义。例如对于没有注解说

明其字体号的汉字，认为它们的字体同版心注解规定的字体，字号同当前字号；若从未出现过任何字体号注解(包括数字、外文的字体号)，则除了字体同版心字体外，字号也按照版心注解的规定。又如遇竖排标题中的标点符号，尽管在文章输入时用的是横排符号，只要有标题竖排标记，自动将它们改为竖排符号而不需要另加注解说明。

CL语言设有自定义注解。可根据用户要求，对其常用排版格式设计成自定义注解。

例如新华社《参考资料》用自定义注解〖ZW〗(正文)，其含义相当于下列注解：

〖HT4F〗〖HJ5＊2〗〖KG2〗

表示字体号为四号仿宋，行间距为五号字的二分（1/2），并在排字符的第一行前空两个字宽。

还可以设计对不同排版对象灵活确定其排法的自定义注解。例如新华社《中国新闻》用自定义注解〖BT〗(标题)，表示以下意义：

标题字体号用四号黑体，居中排，若标题内容只有一行，则标题占正文四行；若标题有两行内容，则占正文六行，两行标题之间的行间距为五号字的一倍；若标题下有一行副题，共占正文六行，行间距也是五号字的一倍。

二、CL排版专用语言的编译程序

（一）主要功能

（1）按照排版注解的语法、语义规定，作语法、语义检查；

（2）按照排版注解的要求，确定每个字符的字体、字号和在版面上的位置；

（3）实现排版的基本规则：在换行时，遇标点、闭括弧等作行首禁排处理，遇开括弧等作行末禁排处理，并对外文字母、数字等作特殊处理；

（4）除实现自动换行外，对于报版可按分栏要求，实现自动换栏，并能作版面分布情况显示；对于书版可实现自动换页、自动形成页码，当有书眉时安上书眉，而不需要人工干预。

（二）程序结构

CL语言编译程序共分三次扫描完成。第一次扫描对输入的小样文件作仔细的语法分析，当发现错误时详尽地列出错误性质与出错地点；同时把小样文件转化为中间语言，把其中的复合字符进行译码和分解，把排版注解转化为中间形式；第二次扫描是对第一次扫描输出的中间语言作编译排版，实现上述（2）、（3）、（4）的功能；第三次扫描是将第二次扫描的输出结果，按照排控制机要求改造成适合激光输出的形式。

编译程序采用分层的模块结构，每个模块是相对独立的程序单位。采用模块结构不仅使编制和调试较为简便，而且便于对程序功能扩充和改进，还可将模块重新组合，组成新的排版程序。CL语言的书版编译程序中大部分模块可为报版编译程序公用，将它们与报版各专用模块组合在一起，组成了报版语言的编译程序。

三、排版注解与应用举例

排版注解以功能划分共有七类。

（1）用于规定排版尺寸（一页书的行数，一行字的字数）、行间距、字体、字号等基本注解。

其中有：版心（BX)；汉字、数字和外文字体号（HT、ST、WT)；行距（HJ)；全身（QS)；对开（DK)；行宽（HK)；改宽（GK）等注解。

（2）用于横向或纵向限定字符的位置。

其中有：空格（KG）；居中（JZ）；居右（JY）；前后（QH）；空行（KH）；行数（HS）；基线（JX）；行中（HZ）；对齐（DQ）；位标（WB）；对位（DW）等注解。

（3）用于划直线、括弧、花边与下加着重点（线）。

其中有：着重（ZZ）；长度（CD）；花边（HB）等注解。

（4）用于表格、框图的制作。

其中有：无线表（WX）；有线表（BG）；方框（FK）等注解。

（5）用于规定书版的书眉与页码的字体、字号、位置与内容。

其中有：书眉说明（MS）；页码说明(YM)；单眉(DM)；双眉(SM)；暗码(AM)；无码（WM）等注解。

（6）用于书版期刊的分栏排与几栏对照排，有分栏（FL）、对照（DZ）等注解。

（7）用于报纸的版面设计。

其中有：划定每篇文章区域的分区（FQ）；规定花边的排法的花边（HB）等注解。

现举例说明。

〔例1〕 排以下字样：

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

每行一种字号，第一行为初号黑体，第二行头号黑体，以下各行分别为二、三、四、小四、五、小五、六、七号黑体，每行之间的行间距为四号字高度的一倍，每行的字间距为该号字的四分，各行均居中排。其排版注解写法如下：〖HJ 4:1〗〖HT 0 DH〗〖JZ（*4〗汉字编辑排版〖HT 1 DH〗↙汉字编辑排版〖HT2DH〗↙汉字编辑排版〖HT3H〗↙汉字编辑排版〖HT 4 H〗↙汉字编辑排版〖HT4"H〗↙汉字编辑排版〖HT 5 H〗↙汉字编辑排版〖HT 5 " H〗↙汉字编辑排版〖HT 6 H〗↙汉字编辑排版〖HT 7 H〗↙汉字编辑排版↙〖JZ)〗

这里〖JZ（*4〗与〖JZ)〗是居中括弧时，表示其间内容居中排，参数*4表示字间距为该行字号的四分。↙表示换行。初号、头号、二号用大黑体，其余用黑体。

系统中还允许特大号、特号、小特号、小初号、小二号、小六号。

〔例 2〕 排下段文字

对于齐次线性方程组，综合以上两点即得，解的线性组合还是方程的解。这个性质说明了，如果方程组有几个解，那么这些解的所有可能的线性组合就给出了很多的解。基于这个事实，我们要问：齐次方程组的全部解是否能够通过它的有限的几个解的线性组合给出来？回答是肯定的。

在此段文字中有两句加着重点，可分别在文字中插入着重注解如下：

对于…，…即得，〖ZZ(〗解的线性组合还是方程的解〖ZZ)〗。这个…，…我们要问：〖ZZ(〗齐次…，…给出来〖ZZ)〗？回答是肯定的。

当加着重点的文字少（例如 5 个），可不用括弧时，直接指出字数（例如〖ZZ5〗）。

〔例 3〕 两行左右对齐与在两行中间排字。例如要排：

中国工艺美术公司
中国美术家协会 联合举办

这里，上面一行八个字与下面一行七个字排同样宽度，都是八个字宽，而“联合举办”四个字排在其后空一字宽度处的两行当中。其注解可这样书写：

〖DQ*(〗中国工艺美术公司↙中国美术家协会〖DQ)〗㊀联合举办

㊀表示空一格。

〔例 4〕 排表格。例如表 13-10 所列的内容

表13-10 排一个表格的例子

品名 / 年份	小麦	黑麦	大麦	燕麦	玉米	稻谷	总计
1960	737.4	16.8	186.8	334.8	1984.8	49.6	3310.2
1970	750	20	178	264	2088	75	3375

此表除顶线外，横向有三条行线，此三条横线间距依次为五号字的三个字高、两个字高、两个字高；纵向连左右墙线共九条栏线，除左墙线与第一条栏线之间距离为五号字的五个字宽外，其余八条间七个距离均为五号字的四个字宽；还有一条从左顶点到第一条行线与第一条栏线交点的斜线（顶线、左墙线称为第 0 条行线、栏线，它们的交点为（0·0）就是左顶点）。斜线项中内容称为表首（BS）内容，用起点XY位置与XY方向分布长度及字数给出。表中其余每项用间隔符〖〗或〖$m\cdot n$〗相隔，〖$m\cdot n$〗表示右下角是第 m 条行线与第 n 条栏线交点的那一项。表格中每项居中排（用！表示），数字的小数点不对位（用＊表示）。注解写法如下：

〖BG!*(3，2·2；5，4·7；(1·1)〗〖BS〗$X2*2Y*2X2Y1*2$，2，品名〖BS〗$X*2Y1X2*2Y1*2/3$，2，年份〖0.1〗小㊀麦〖〗黑㊀麦〖〗大㊀麦〖〗燕㊀麦〖〗玉㊀米〖〗稻㊀谷〖〗总㊀计〖〗1960〖〗737.4〖〗16.8〖〗186.8〖〗334.8〖〗1984.8〖〗49.6〖〗3310.2〖〗1970〖〗750〖〗20〖〗178〖〗264〖〗2088〖〗75〖〗3375〖BG)〗

5。排公式

例如排下列方程式：

$$f(x_1, x_2, x_3)=2(y_1+y_2)(y_1-y_2)-6(y_1-y_2)y_3+2(y_1+y_2)y_3$$

注解写法如下：

f(((x↑; 1)), ((x↑; 2)), ((x↑; 3)))= 2 (((y↑; 1))+((y↑; 2)))
(((y↑; 1))−((y↑; 2)))−6(((y↑; 1))−((y↑; 2)))((y↑; 3))+ 2
(((y↑; 1))+((y↑; 2)))((y↑; 3))

用复合符号((y↑ i ; j))表示y_j^i。

〔**例 6**〕 排公式

例如排以下算式：

$$a+i\times b:=\frac{p+i+q}{r+i\times s}$$

其注解写法为

$a+i\times b:=$〖HZ(〗〖ZZ5−〗$p+i\times q$↙$r+i\times s$〖HZ)〗

将$a+i\times b:=$排在两行中，用行中注解控制，分数线用着重注解来划，行中注解还能使分母与分子左边对齐。

〔**例 7**〕 排以下一段程序。

下列程序是检查一个字母并确定它是一个十进制数。字母是在累加器中，测试时不被破坏。累加器AC_x和AC_y被破坏：

```
LDA     ACx,    C60             ; ACx= 0
LDA     ACy,    C71             ; ACy= 9
ADCZ#   ACy,    ACS,    SNC     ; 空一格，如 (ACS)> 9
ADCZ#   ACS,    ACx,    SZC     ; 空一格，如 (ACS)≥ 0
JMP     ……                      ; 不是数字
……      ……                      ; 是数字
C60:    60                      ; 0
C71:    71                      ; 9
```

这是一段NOVA机上的汇编程序，它的排版注解可用排无线表写法：

〖WX(4, 3·3, 11〗 LDA〖 〗ACx, 〖 〗C60〖 〗〖 〗; ACx= 0〖 〗LDA〖 〗ACy, 〖 〗C71〖 〗〖 〗; ACy= 9〖 〗ADCZ#〖 〗ACy, 〖 〗ACS, 〖 〗SNC〖 〗; 空一格，如 (ACS)> 9〖 〗ADCZ#〖 〗ACS, 〖 〗ACX, 〖 〗SZC〖 〗; 空一格，如 (ACS)((→>)) 0〖 〗JMP〖 〗……〖 〗〖 〗〖 〗; 不是数字〖 〗……〖 〗……〖 〗〖 〗〖 〗是数字〖 〗C60:〖 〗60〖 〗〖 〗〖 〗; 0〖 〗C71:〖 〗71〖 〗〖 〗〖 〗〖 〗; 9〖WX)〗

此无线表第一栏为四个字宽，二、三、四栏均为三个字宽，末栏为十一个字宽。无线表每项以间隔符〖 〗相隔，按行逐项排版，排满规定的栏数后，自动换行。空项则在两个间隔〖 〗之间不写内容。复合符号((→>))代表≥。

〔**例 8**〕 排以下一段内容：

```
M1:DELTA 03:= 2;
  FOR i:=1,2DO
  BEGIN 地址 (A〔i
    +1〕):=A0+i+1;
    存入 (D105+
    DELTA  03)中;
    DELTA03:=
    DELTA  03-1;
  END;
  STRUTURE order
```

预先计算A_0及ΔA，循环时跳过

排版注解写法为：

((M↑; 1)) 〖ZK(〗: DELTA〳〳03:=2; 〖ZH(〗〖ZZ3-〗FOR〳〳i:=1,2 〖ZZ2-〗DO↙ 〖ZZ5-〗BEGIN地址 (A〔i + 〖ZK (1〗↙1〕):=((A↑; 0))+ i + 1; ↙存入(((D↑; 105))+↙DELTA〳〳03) 中; ↙DELTA〳〳03:=↙DELTA〳〳 03-1; 〖ZK)〗〖ZZ3-〗END; ↙STRUTURE〳〳order 〖ZH)〗〖ZK)〗〖CD〉-10〗 预先计算 ((A↑; 0)) 及 ((DL))A，循环时 跳过。

用自空（ZK）注解控制前空位置，使换行时左边对齐。自空注解可放在要求 对 齐的位置，如第一行用“((M↑(1)) 〖ZK(〗:”就是要求以下各行左边与“:”对齐；自空注解也可用参数直接写出前空的距离，如第四行“〗):=…”要求比上行缩排一个字宽，可用 〖ZK(1〗控制。用自换（ZH）注解控制右边换行的位置，使以下各行 不 得超过此位置。它们都以对应的闭括弧取消这种控制。大括弧〉用长度（CD）注解 实 现，〖CD〉-10〗 表 示 长 10 行，“-”表示从下往上画，画完后在中线出口，〳〳是外文间隔控制符，它属于排版注解类，用于表示一个外文字的 结 束，与下一外文字（或数字）之间应空半个字宽。

除以上所述注解外，系统还接受用户自定义注解，这里不再详述。

13.5.3 报纸的版面设计

报纸版面设计是对一版文章（包括图片、花边）的排版格式的整体性考虑，既要使每篇文章有一个恰当的位置，又要使每篇文章的排版格式组成一个满意的版面。

CL排版语言用专门注解刻画一版八开报纸文章区域的分布及文章标题的格 式，CL语言编译程序的执行结果能一次形成整个版面，并能显示版面分布情况，以便用户用人机对话方式了解和调整版面，最后得到满意的结果。

一、用 CL 语言描述报纸版面

为了用户使用方便，CL语言把报纸的版面设计（每篇文章在版面上的 分 布）及标题的格式（包括标题的字号、字体，每个字在标题区里位置的安排，提要的字号、字体等）独立于文章之外，用专门的排版注解“分区（FQ）”与“排题（PT）”等描述，这些注解组成“组版文件”，当用户需要改变版面的布局或标题的格式时，只要修改 组 版文件就可达到目的。

以下说明版面设计的两个主要注解——分区与排题。

（一）分区注解

其一般格式如下：

〖FQ〈版数〉B：｛〈文件名〉（〔X〈上接版数〉〕〈行号〉〈列号〉Q〈文件面积与位置〉〔〈标题区面积与位置〉〕）｝〗

其中，花括号｛ ｝表示其中内容可重复；方括号〔 〕表示其中内容可能不出现。这些和前述情况一样。

〈版数〉*B* 说明排第几版报纸。

〈上接版数〉是当本文为上接别版的文章时，说明上接第几版的文章。其前面的*X*表示“续”。

〈行号〉〈列号〉*Q* 说明本文在版面上位置的起始点，*Q*表示“起”。

〈文件面积与位置〉说明本文在版面上所占的面积与位置。

〈标题区面积与位置〉说明本文的标题所在的位置与所占的面积，当本文是上接别版的文章时，允许没有标题。

这一注解刻画一版报纸所有文章的位置。

例如：〖FQ2B：C1304（0101Q41×10＋25×10·3〔Z 4×3 L〕）C1305（X42712Q16×10·3）C1306（0145Q41×21〔Y23×4〕）C1307（4401Q33×21·2－ZX10×10＋33×21〔Z4×2L；YX 3 ×1L〕）〗

这一注解给出如图 13-19 所示的版面情况。

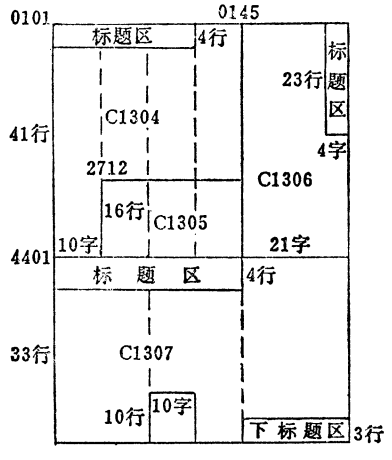


图13-19　版面分布图

这是报纸的第二版，共有四篇文章，它们的文件名分别是C1304、C1305、C1306、C1307。其中C1304从第 1 行第 1 字排起，共分四栏，第一栏宽10个字、高 41 行，第二、三、四栏都是宽10个字、高25行，标题区在文章的左上方，共占 4 行 3 栏。C1305 是上接第四版的文章，从第27行第12字排起，共三栏，每栏都是宽10个字高16行，没有标题。C1306从第 1 行第45字排起，不分栏，高41行宽 21字，标题区在文章的右上角，宽 4 字高23行。C1307从第44行第 1 字排起，共三栏，第一、二栏都是宽 21字高33行，但第二栏的左下角去掉宽10个字高10行，第三栏紧接在第二栏之右，都是从第44行起，但面积不同，为宽21字高33行，标题区有两个：第一个在左上方，高 4 行宽 3 栏；第二个在右下方，高三行宽一栏。

还有许多分区形式，不能一一举例。

（二）排题注解

其一般格式为：

〖PT{〈文件名〉（〔！〕{〔〈一行标题排法〉〕}〔〈提要排法〉〕〔；{〔〈一行标题排法〉〕}〕〔〈提要排法〉{〔〈一行标题排法〉〕}〔；{〔〈一行标题排法〉〕}〕）}〗

其中，

文件名后，若有“！”，则表示标题竖排，若无“！”，则为横排，本语言允许出现竖排标题与提要。

〈一行标题排法〉说明这一行标题分几段，每段标题的字体号与格式。允许有若干行标题。

〈提要排法〉说明提要文字的字体号及位置。提要可以在标题的上面或下面。

当有两个标题区时，第二个标题区中的标题用“；”与第一个标题区内容相间隔。

这一注解可用以说明一篇文章全部标题的排法，也可以用于说明同一版上所有文章标题的排法。

二、报纸的版面显示

报版排版以后，可用显示器显示排版结果，用针式汉字打印机印出说明文字。版面显示器用三种点阵（或三种颜色的光点）分别表示整个版面正文文字的分布、标题位置、花边位置及图片位置(留出图片空白)。打印机印出每篇文章及某标题的排版情况，说明那篇文章排不下，那篇文章排不满，或那篇文章的那行标题排不下；当文章排不下时，在何处形成了断点，第几段，共多出多少字等。显示器与针式打印机输出的内容形成一幅图文并茂的版面。用户就可以对本次的排版结果一目了然，做到心中有数，便于对文章或版面作必要的修改。

上述分区注解的版面情况，如图 13-20 所示。

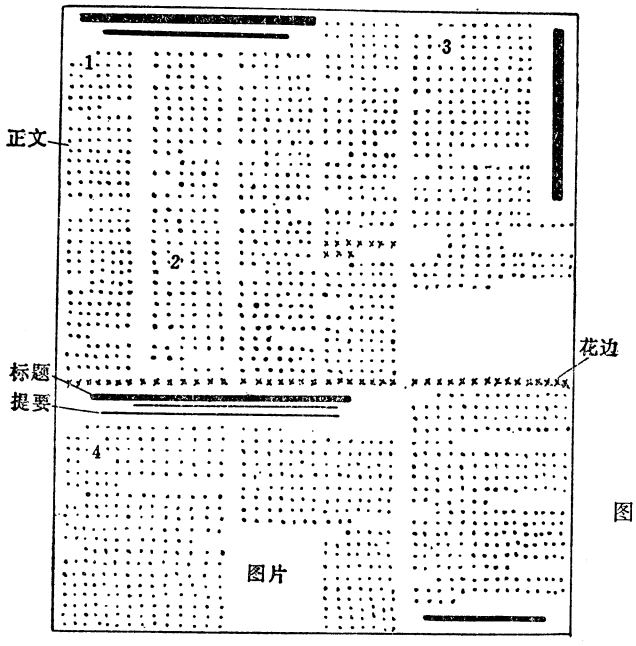


图13-20　版面显示图

每篇文章在针式打印机上印出两行说明，第一行为文件名与文章排版情况，第二行为标题排版情况。如第一行文件名后为空白，表示文章正好排下，如第二行空，表示这篇文章所有标题都能排。

三、人机对话调整版面过程

在排版过程中，可能因为版面的区域划分不合理，造成有的文章排不满，有的文章过长而排不下，需要压缩或者需加设断点（在排版前断点的位置是未知的），或需要更改区域划分；也可能因为标题的字号、体号选择不当，或行距、字距等参数选择不合适，造成标题排法不够恰当；也可能由于用户疏忽造成花边与文章重叠等。当出现这些问题，在用手工铅字排版时，由编辑人员重新修改版面布局，或修改某些文章中标题排法，或修改文章内容，此后再进行改版。改版时拆拼原来版面或局部版面重排，甚至重拣部分铅字，才可重新组成新的版面。当这一版面还不合适时，又要重复上述过程。

用本系统排报版，当需要修改版面时，可利用联机终端修改组版文件中分区或排题等注解，或修改花边位置，不涉及文章内容就可变更版面的分布情况。当要修改文章内容时，也可用联机终端的编辑功能完成（见13.6.3节），修改后再发排版命令，由CL语言编译程序重新排版。即使是全部重排，也可在一分钟内重新排出一版八开报纸，排完后又可通过显示器和针式打印机了解排版结果。如果需要看到排版后的逼真的报纸版面，可用大样输出机在普通纸上印出排版后的版面，供用户仔细校阅。如此反复直到得到满意的版面后才进行照排，由激光照排机输出制版用的胶片，再制版印刷。

这一人机对话的修改版面过程，只需编辑人员考虑版面的修改方案，修改相应的内容就可用很快的速度调整完毕，既提高效率又可免除繁重的手工劳动。

13.6 编辑排版系统中的汉字终端子系统

13.6.1 终端子系统的类型

对于计算机-激光汉字编辑排版系统的汉字终端子系统，按其功能可分为两类：一为脱机使用的汉字终端；一为联机使用的汉字终端。前者用于脱机校对修改文章，并将它们存于软盘存储器上，经过校改的文章从软盘输入到主系统，供排版使用；后者与主系统相联，用户在其上以人机对话方式，完成书、报的照排，它同样也具有编辑校对功能，以实现对已进入主系统的文稿的校改。

目前采用的终端是CMPT-Ⅱ型终端，它由M6800微处理机控制。采用笔触式汉字字盘，盘面共收汉字及字符2880个，另有功能键64个，词组键128个。汉字显示器每幅显示16行，每行32字，每幅共显示512字，汉字显示器可兼用于显示八开报纸的版面设计情况。终端上显示与打印的简易汉字均为15×18点阵。

终端所收汉字共7000字左右。其中属于盘面已收入的盘内字采用一字一键方式输入；盘外字按部首偏旁规则用盘内字拼写，即采用一字多键方式输入，如“挛”用“亦”、“手”两个盘内字拼成，显示器与打印机输出的是拼好的“挛”字，进入系统的是该字的内部编码。

设立词组键以提高输入效率，它以一键多字方式输入汉字或字符。其中一部分为固定词组，存放常用词组；另一部分为自定义词组，供用户根据需要随时定义使用。

13.6.2 脱机终端的使用及其编辑功能

脱机终端使用的目的是将文章组成文件输入到软盘上。在终端的键盘上击键输入文章过程中，可用“写盘”功能键随时写入软盘中。记入软盘的文章可通过命令在打印机上印出全部或部分内容，供用户校阅；也可用显示命令在显示器上显示，供用户校对和修改，修改后再重写入软盘中。将存有若干篇校对无误的文章的软盘内容卸出后妥善保存，在用作排版时送到主机上输入，供主系统照排。

脱机终端上设以下两类功能键，实现对显示于屏幕上文章的校对修改。

一、光标功能键

光标位于字符之下与字符同宽的一条光点组成，它用于指示字符位置。它的移动由以下十个功能键指挥：

（1）向左　移至左边一个字符。若光标当前位于行首，则左移一字到本行行末。

（2）向右　移至右边一个字符。若光标当前位于行末，则右移一字到本行行首。

（3）向上　移至上行同列字符。若光标当前位于第一行，则移到本列末行。

（4）向下　移至下行同列字符。若光标当前位于最末行，则移到本列第一行。

（5）还原　移至第一行第一列。

（6）行中　移至本行中间。

（7）列中　移至本列中间。

（8）行首　移至本行第一列。

（9）列首　移至本列第一行。

（10）清除　清除整个屏幕内容，并将光标还原到第一行第一列。

二、编辑功能键

第一类　删。共六个功能键：

（1）删一字　将光标所指的字符删去，其后的字符前移。若删去的是分段符↙，则两段合并。

（2）删一行　将光标所在行全部字符删去，将其后行前移。

（3）删一句　将光标所在句全部字符删去，其后字符前移。句以标点符号，、；。•（外文句号）：？！之一为结束。

（4）删一段　将光标所在段字符全部删去，其后字符前移。

（5）删光标之左上　删去本行位于光标之左及以上所有行的字符，其后字符前移。

（6）删光标之右下　删去本行位于光标之右及以下所有行的字符。

在字符前移时，仍保持段落，并可能在屏幕之末造成空白。

第二类　增。一个功能键。

按此键后，再按字符或词组，就将字符或词组插入光标指示位置之前，其后字符后移。字符后移时，仍保持段落。若增加的是分段符↙，则将一段内容分成两段。为便于实现，规定不允许将增加（或修改）的内容因字符后移而移出幅面。

第三类　改。一个功能键。

按此键后，再按字符或词组键，就将字符或词组替换光标所指的字符，且将光标后

移一字（或同词组字数）。如果换入的是分段符↙，则将一段文字分成两段。

脱机终端有独立的文件系统。可以显示或打印指定的文件，对文件实现上述编辑功能。还设有实现以下功能的命令：

（1）将正在显示的文件前进（或后退）一行，前进（或后退）若干页（幅面）、连续前进或后退（即滚动）；

（2）打印出文件中的某页或若干行；

（3）撤离指定的文件；

（4）命名新文件。

13.6.3 联机终端的功能及其使用方式

联机终端有以下功能：

1．具有比脱机终端更强的编辑功能　除了可以使用13.6.2节中所述两类功能键，实现脱机终端具有的全部编辑功能外，还可使用联机命令实现文件的拆开、合并以及将一段文字插入同一文件的某处或其它文件中等功能。只是编辑校改的对象，不是终端软盘上的文件，而是存于主系统中的小样文件。

2．接受用户输入的命令　除命令功能键外，对一段命令先进行语法检查，检查正确后向主系统发送。

3．接受主系统发送来的信息，并按信息类型进行加工后显示或打印。主系统发来的信息共有以下四类：

（1）小样显示类。主系统发来小样文件内容，终端按当前命令分别作不同处理。

若当前为‘显示’命令，则将主系统发来信息直接送幅面显示；

若当前为‘前进一页’命令，则将原幅面末行内容保留在显示区第一行，主系统发来信息补在后面显示；

若当前为‘后退一页’命令，则将原幅面第一行内容保留在末行，主系统发来信息补在前面显示；

若当前为‘前进一行’命令，则将原幅面内容前推一行，主系统发来信息补在末行；

若当前为‘后退一行’命令，则将原幅面内容后退一行，主系统发来信息补在前面。

（2）版面显示类。主系统发来一版八开报纸的版面设计信息，终端按编码分别译为表示正文、标题、花边的点阵后显示。

（3）印字信息类。将主系统发来信息送针式汉字打印机印出（简易汉字）。

（4）命令回答类。主系统发来命令回答编号，终端译成对应的汉字显示出来。例如主系统发来编号为01，终端译成‘完成’两字在屏幕的第二行显示出来。命令回答有时还带信息，如‘缺文件’其后带文件名，终端也需将它们显示出来。

主系统在发来信息的同时，指示当前执行的命令是否结束。若已结束，则终端终止命令执行状态，准备接收新命令；若未结束，则终端处于等待命令执行状态，不能接收新命令。

计算机-激光汉字编辑排版系统同时允许八个用户联机使用。每个联机终端设有下列控制类功能键，实现与主系统连接、从主系统撤出，以及接收命令等功能。

（1）终端启用键。此键向主系统发申请启用终端的信号，主系统响应后，发送来用

户名表，终端接收后显示出“准备好，请打用户名”，通知用户输入自己的用户名。

（2）终端结束键。用户使用终端完毕，通知主系统将本终端从系统中撤出，切断与主系统的通路，此后才可关闭终端。终端结束键应在一个或一批命令结束后使用。

（3）命令开始键。表示以下击键输入的是命令，终端处于接收命令状态。

（4）命令结束键。表示一个命令或一个用户名的输入结束。对于命令，终端按命令动词、参数逐步进行分析，发现错误时，发“命令动词错”或“命令格式错”警告，用户可利用终端的编辑功能对命令进行修改，修改后的命令必须重新按命令结束键。只有对于正确的命令，终端才向主系统发送。对于用户名，终端查用户名表，查不到时显示“用户名不正确”，查到时向主系统登录并显示“可以开始工作”。

联机终端使用过程为：

开启汉字键盘、汉字显示器等各电源开关后，在汉字键盘上按“终端启用”键，主系统响应后，在显示器上显出：

准备好，请打用户名

当用户打入用户名，并在显示器上亮出“可以开始工作”的通知后，用户即可打入命令，待命令执行时，向用户报告执行情况，如此以人机对话方式完成编辑排版等功能。

除了一般用户命令外，系统还设有特权用户命令，以了解系统运行情况和对系统的控制，一般可利用控打作特权用户终端。这些只供特权用户使用的命令有：

（1）增删用户名。其命令格式为：

ZS 〔Z|S〕 用户名

（2）询问某终端上命令执行情况。其命令格式为：

XW 终端号

（3）撤消某终端已输入但主系统尚未执行的命令。其命令格式为：

CX 终端号

等等。

还要指出的是，主机的控制打字机除了作为上述特权用户终端外，它同样可以作为一般用户终端使用，其差别是没有汉字显示器、汉字打印机和汉字键盘，不能输入输出汉字，只能用宽行输出汉字的内部编码。这在调试主系统阶段和出现某些意外情况，终端不能接入时，使用比较灵活和单纯。此外，还可利用它了解主系统内部情况，如打出用户名表、文件目录等。

本章所述精密汉字编辑排版系统的主要技术困难是：精密字模存储量大，且要求能快速存取；输出精度要求很高；为了排各种复杂的版面（例如数学公式、化学结构式、框图和乐谱等）需要开发功能较强的软件。一旦这些困难得到解决，计算机排版将显示出巨大的优越性，这是一个具有广阔前景的应用领域。

随着大规模集成电路技术的发展，采用1兆位的存储器片、双极型Am29116微处理器芯片和门阵列技术，可以把精密汉字字模，点阵复原设备和照排控制设备缩小到一块插件板的规模。精密照排系统不再成为印刷厂的专用设备，而将成为办公室系统和中、小型计算机上都可以增设的廉价设备。激光直接雕版被称为第五代照排机，免除了底片感光和底片显影、定影等工序，将进一步提高效率和降低成本。照排系统与激光输出缩微机相连，有利于出版物的长期存档；照排系统与情报检索系统相连，有利于对出版物的检索。这些都是很有意义的研究领域。

第十四章　汉字企业管理系统

14.1　现代企业管理系统简介

14.1.1　现代企业管理系统的主要功能

现代化企业的根本任务是根据国内外市场的需要，为社会提供物质产品和劳务，为社会创造更多的财富。这一切活动都是以生产为中心的。在生产过程中，一方面要加速资金周转，努力降低成本和物资消耗，提高劳动生产率，逐步改善职工的劳动条件、提高职工的福利；另一方面，要加强职工的技术教育，采用新技术、新工艺、新材料，不断提高生产技术能力。为了更好地完成上述任务，必须努力学习和应用现代化的科学管理方法，不断提高管理水平。

简言之，现代化企业要按经济规律办事，充分发挥企业的职能和作用，科学地组织生产，用最少的消耗，取得更大的经济效果。

企业管理的职能是通过计划、生产、质量、技术、设备、劳动、物资和财务等诸方面的科学管理来实现的。如表14-1所示。

表14-1　现代企业经营管理系统一览表

现代企业经营管理系统
- 计划管理
  - 市场预测
  - 管理计划
  - 投资计划
- 生产管理
  - 作业计划
  - 生产准备
  - 生产调度
  - 生产统计
- 质量管理
  - 质量控制
  - 计量标准和量具管理
- 技术管理
  - 技术资料管理
  - 工艺资料管理
  - 产品结构文件管理
  - 新产品试制管理
- 劳动工资管理
  - 劳动组织管理
  - 工时定额管理
  - 工资、奖金管理
- 设备管理
  - 设备维修计划及实施
  - 设备更新计划及实施
  - 设备运行记录管理
- 物资管理
  - 物资供应计划
  - 库存管理
  - 销售管理
  - 供销运输管理
- 财务管理
  - 财务资金管理
  - 成本管理
  - 经济核算

下面就现代企业经营管理系统的主要职责作简单介绍：

一、计划管理

根据上级下达的任务、用户的定货合同和市场预测资料，在充分利用本企业资源的基础上，确定企业的发展计划和投产计划。按时间又可分长远计划和年度、季度、月度计划等。前者又可称为企业的经营战略，后者称为管理计划。计算机在计划管理方面的应用属于经营决策，它是高级的判断性方面的工作。经营战略和管理计划的好坏，关系到企业的根本利益，甚至和企业的兴衰休戚相关。因此计划管理是整个管理的出发点。

二、生产管理

它的职责是要合理使用企业的各种资源（如人力、物力和财力等），生产优质、高产低耗的产品和劳务。生产管理要确保生产计划的贯彻，必须制订周密的生产准备计划和合理的作业计划。生产准备工作做好了，生产计划的执行才有保证。作业计划是生产计划更具体的表现形式，在作业计划中要明确车间、工种内部每一个工人、每一台设备，在每一个工作时间内做什么，因此它是生产管理中最主要的任务。在实际的生产过程中，往往还有很多预想不到的情况发生，例如紧急任务的下达，生产任务完成的好坏、设备的故障以及其他生产事故的发生等，都有可能改变作业计划的正常执行。此时，生产调度必须根据新出现的情况，对作业计划作适当的调整，使生产活动在新的情况下，继续保持有条不紊地进行下去。生产统计既是企业及车间领导人进行调度决策的重要依据，又是各项考核指标和统计报表的基础信息。

三、质量管理

按技术标准和订货合同对产品质量进行严格的检验，负责各主要工序以及原材料、外购件、半成品等的检验工作。

四、技术管理

一方面对企业当前生产提供技术保证，另一方面负责技术开发工作。作为信息管理系统对各种技术文件、工艺文件的管理显得特别重要，因为他们是整个企业各个生产环节的依据。

五、劳动工资管理

根据生产工人日报表进行工时统计，作为制定工时定额的参考。根据工时统计编制劳动统计报表。进一步根据劳动统计和人事档案进行工资管理和奖金分配，供财务管理进行工资核算。

六、设备管理

要保证各种设备经常处于良好状态，为此必须做好设备运转的台时统计，制订大修计划。同时要采用新的先进的设备，为企业提供更好的劳动手段和工具，为此必须制订切实可行的更新计划。

七、物资管理

物资管理包括仓库管理和销售管理两大部分。一般企业有三类仓库：供应仓库、销售（成品）仓库以及废品仓库。主要的是前两类仓库。供应仓库又可分为原材料仓库，原器材仓库，零件仓库以及各种半成品仓库等。供应仓库的目的是要适时、适地、适量地供给各种生产活动所需要的材料物资。在保证生产正常进行的前提下，尽可能将各种物资的库存压缩到最低限度，以减少存储费用。销售仓库应该在保证供应（用户需求）的

情况下，也应该尽可能地压缩库存量。为此，必须做好市场预测和严格执行用户合同的前提下，制订一个对于产品的数量、品种、质量在时间和需求上相配合的生产计划大纲，以保证销售的需要。另一方面，根据生产大纲的要求，制定一张合理的物料采购计划。

八、财务管理

财务管理主要包括三方面的职责。其一是根据生产计划、劳动工资计划、工时定额、材料定额和物资供应计划等，产生成本计划、利润计划和流动资金计划等；其二，是会计核算，它汇总企业各个部门的各种原始凭证，经过货币资金核算，得到各种记帐凭证，然后编制各种分类明细帐、序事帐以及各种统计报表；其三、根据各种记帐凭证进行成本核算、材料核算、固定资产核算、工资核算和奖金发放等工作。

14.1.2 计算机在企业管理中的应用概况

一、采用计算机进行企业管理可收到明显的经济效果

电子计算机是企业管理现代化的重要工具。现代化生产的特点是：自动化程度高；生产部门的分工越来越细；各种经济问题的决定因素越来越复杂；要求及时得到对情况的反映并作出决定，这些都对管理工作提出了更高的要求。对这样错综复杂问题的解决只有用电子计算机才能做到。事实上，用不用电子计算机进行企业管理，效果是大不一样的。例如，美国霍尼威尔公司的一个工厂，六十年代实现用计算机管理，年资金周转率从3.78提高到5.50，库存减少360万美元，按期交货率从59％提高到96％。又如新日本制铁公司的君津工厂有职工七千人，年产钢一千万吨，生产过程和管理工作广泛采用电子计算机，节约了职工两千多人。该厂共用计算机51台。其中，使用4台大型机中的2台IBM370/158独立操作，主要负责订货处理，材料计算，年、月、旬、日生产计划编制；另外的2台IBM370/158用于联机操作，负责工作指令、操作要求及数据收集；47台控制用计算机，进行过程控制、处理工作指令，记录数据等。表14-2列出了该厂在管理和过程控制中采用计算机以后得出的指标，可以看出，其效果是十分明显的。

表14-2 新日本制铁公司的君津工厂在管理和过程控制中采用计算机后所取得的经济效果

	采用计算机管理	计算机用于过程控制
节省人力	42%	3.5%
增产	36%	84.7%
提高成品率	13%	4.6%
降低作业费	3%	—
压缩库存	2%	—
其他	4%	7.2%

据苏联统计，用电子计算机管理企业，劳动生产率普遍增长5％。据罗马尼亚的实践结果，若一个工厂用电子计算机管理，则在不增加人员和设备的基础上，生产一般可提高3～5％。在西方国家中，已有80％的电子计算机是用于数据处理的，而其中管理信息的处理业务又占极大部分。

在我国，电子计算机在企业管理中的应用也有不少成功的例子。

某汽车制造厂用计算机对全厂金属材料进行管理、物资供应计划管理、合同管理与库存管理。达到数据准确、处理及时，提高了工作效率。该厂采用微型机管理汽车零配件的自动立体仓库，使货格的利用率从60%提高到80%，加快了零配件的出库速度，同时减少了库存，加速了资金周转，并能按品种及货格进行盘库，显示、打印明细帐。使仓库管理更加现代化。在全国这方面的例子还很多，这里就不一一介绍了。

采用电子计算机进行辅助企业管理，尽管在开始阶段要有较多的投资，但由于它有明显的经济效果，全部投资一般在几年内即可回收。

我国是采用表意文字的国家，汉字拼音还不普及，这对计算机处理文字信息带来了困难。然而近几年来，由于对汉字信息处理技术的大量研究和试用，解决了很多技术难题，积累了不少经验，目前已趋于实用阶段，这样，汉字企业管理系统也就有了普及的可能。

二、计算机用于企业管理的几个阶段

计算机企业管理系统是计算机技术设备、人和管理对象组合而成的人机系统。利用计算机进行数据采集、记录、显示、传输，运用经济数学的方法，实现对企业的最优管理。对于大量信息的快速处理和重复性的劳动都让计算机来完成，而对于处理结果的分析、判断、决策等创造性的劳动，则由人来完成。这种人尽其才，物尽其用的人机系统是计算机用于企业管理的格局。

从计算机用于企业管理的信息结构来看，有组成文件系统的，也有基于数据库之上的。

在利用计算机进行企业管理的初级阶段，往往是针对某一方面的管理问题而设计的，设计人员大多采用文件系统方式来组织信息。信息的组织可以优化地与应用系统结合起来。目前国内外用小型、微型计算机上进行企业管理时，广泛采用这种系统。然而文件系统有一定的缺点，例如数据的一致性、安全性、保密性等缺乏有效的、统一的控制办法。文件系统的不足，导致数据库的研究和应用。

建立在数据库上的企业管理系统可以对企业中的信息进行统一管理，因此，数据可以做到没有冗余性，从而可保证数据的一致性。在数据库技术中，提供了对数据的定义、存储、检索、修改等操作，以及对数据安全性、完整性、保密性进行统一控制的数据库管理系统（DBMS），使得对数据的应用更为有效。因此这是计算机用于企业管理的高级阶段。数据库上的企业管理系统大致又可分为两类。一类是采用中、大型计算机及若干台中、小型计算机构成的计算机系统，它一般适用于中、大型规模的企业管理；另一类是采用若干台高档微型机在局部网络支持下构成的微型机系统，它一般可用于中、小型规模的企业管理。

14.1.3 实现计算机企业管理系统的条件和步骤

要管理好一个大型企业是一件很不容易的事情。其原因如下：①在企业的内部有很多职能部门，各职能部门之间有密切的联系，它们相互依赖、相互制约，所以，企业管理系统是一个整体性系统；②一个企业又不能是一个封闭系统，它需要和上级部门、供货单位、各类用户等交往，因此，企业与外部环境有密切的联系，这种联系也是相互依赖、相互制约、相互作用的，此外还有很多交往存在着不确定的因素（例如能不能按期得到

所需的原材料，市场到底需要本企业提供多少产品，什么时间生产最好等等），所以说，企业管理系统又是一个开放性系统。③企业管理系统还是一个反馈系统，也就是说，受控对象的行为要返回来影响控制决策。

综上所述，企业管理系统是一个整体性、开放性和反馈性的系统。因此，要建立一个计算机的辅助企业管理系统必须从系统工程的角度来进行系统分析，系统设计。站在系统的立场上，而不再片面地追求某一单项或局部的最优，而是追求系统目标的最优。

每一个企业都要建立自己的计算机企业管理系统，这是大势所趋，但对一个具体的企业，它是否具备了实现计算机辅助企业管理的条件，这还必须在进行了认真的分析以后才能确定。

一、实现计算机辅助企业管理系统，从企业的角度来看应具备的条件

1．要有明确的系统目标

要建立一个比较完善的计算机辅助企业管理系统，就必须有明确的系统目标。一般都是为了解决企业现有的资源（如人力、设备、资金等）和要完成的任务之间的矛盾。即使是建立单项或局部的计算机辅助企业管理系统，也要提出明确的目标。

2．要有上下一致的积极性

企业的领导，有关的各职能部门以及计算机工作者都要有实现系统的积极性。其中领导的积极性尤为重要，没有领导的积极性，其他的积极性既不可能得到发挥，也不可能持久。

3．要有资金和技术环境条件的保证

企业要具有足够的资金、人员和实现系统所必需的工作环境的保证。企业为实现系统所要的投资（资金、人员和工作环境）同系统的目标密切相关。这里指的人员包括系统分析员、数据库管理员、系统维护员（软件、硬件）以及操作员等。

企业的系统分析员应该由精通整个企业管理业务的人员来承担。对于一个中、大型企业这项工作往往很难由一个人来承担，在这情形下可成立系统分析员小组，其组成人员包括厂部（或总厂）系统分析员和各职能部门（或各分厂）的系统分析员。企业的系统分析员的职责是编写系统分析报告，协助系统设计人员进行系统调查以及参加系统设计。系统设计人员一般由计算机专家来承担，由于他们不可能对整个企业管理了解得十分透彻，因此要使被设计的系统切实可行必须要有企业的系统分析员参加系统设计。

数据库管理员是对数据库全局控制负有责任的人员（对于大型系统，也可组成一个小组）。数据库管理员的主要职责是：

（1）决定数据库的信息内容；

（2）决定信息在数据库中的存储结构和存取策略；

（3）和使用者建立联系；

（4）定义授权；

（5）实现恢复策略，以防止危害数据库（的任一部分）事故的发生，一旦发生了事故，就得修补数据；

（6）监视系统性能的变化，使系统经常处于最佳工作状态。

4．企业领导直接承担计算机企业管理系统的筹建工作

要由熟悉全企业生产业务的副厂长或副总工程师这一级或更高一级的领导直接参加

这项工作。要成立计算机管理委员会。计算机管理委员会由厂长、总工程师、直接参加具体领导的副厂长或副总工程师，系统分析员、数据库管理员、计算中心、各有关职能部门负责人和参谋人员组成。

5．为实现系统目标所必需的数据信息和资料要齐全。

6．要在企业的全部中、高级管理人员中进行对计算机辅助企业管理知识的教育。

7．要提出系统分析报告（或用户需求书） 下面对系统分析报告的有关问题分别加以阐述：

1）编写系统分析报告的目的：

编写系统分析报告的目的是从企业的角度论证建立新系统的必要性，并提出初步设想。对新系统在技术上、经济上、资源上和行政上的可行性进行必要分析。

2）系统分析报告的用途：

（1）给上级部门审阅、备案，以争取上级部门的支持；

（2）作为企业实现计算机辅助企业管理后用来衡量效益的依据之一；

（3）作为同设计单位洽谈合作的依据。

3）系统分析报告的主要内容：

（1）企业的概况

其中包括企业的历史、现状、组织、生产管理状况、今后发展规划等。其细目可以列举如下：

① 企业成长的统计资料（人事、销售和利润等）；

② 管理形态（组织结构、部门职责）；

③ 信息流程（绘制各种信息流程图、收集所有的票证、单据、凭证和帐册等样品）；

④ 企业外部环境描述（国家和上级部门对企业的政策要求和变化、企业和外部的各种交往活动情况等）。

（2）对要设计的计算机辅助企业管理系统在功能上、时间上、效益上以及实现效益的可能性进行分析。

（3）投资预算。包括财力、物力和人力等。

（4）其他。如对领导部门、对设计单位的希望和要求等。

二、实现计算机辅助企业管理系统的步骤

（一）系统调查

要成立由系统设计员、委托的系统分析员和企业的系统分析员（作为企业代表）以及数据库管理员等组成的联合调查组，对企业管理业务作全面的调查。其调查内容大致包括：

（1）系统分析报告中涉及的内容；

（2）对企业组织结构的调查；

（3）对企业职能部门的调查；

（4）对企业的信息源，即原有管理系统的调查；

（5）对物资流的调查；

（6）对企业的计划和统计工作的方法的调查；

（7）对企业的外部信息的调查；

（8）收集有关企业的目标信息；

（9）分析企业实现计算机管理的条件是否成熟。

系统调查强调从系统整体的角度来描述和分析企业内部和外部的工作情况。

（二）整体设想

（1）分析系统规模及配置；

（2）分析系统实施后可能达到的经济效果；

（3）系统预算、进度安排；

（4）提出系统可行性报告；

（5）签订合同。

（三）逻辑系统设计

（1）系统的整体结构及其各子系统的设计；

（2）描述输出说明；

（3）描述处理说明；

（4）制定决策准则；

（5）描述输入说明；

（6）制订各子系统和职能部门承担的各项任务一览表；

（7）编制各子系统和职能部门之间文件周转路线图。

在上述逻辑系统设计的基础上进一步提出

（8）系统所需应用软件概述；

（9）系统所需的支持软件；

（10）系统的硬件配置；

（11）制订系统设计和实施综合进度表。

逻辑系统设计先不依赖计算机硬件系统来规划和制定准则。

逻辑系统设计的结果得出的是实体系统设计的说明书。

（四）实体系统设计

（1）文件格式的确定、周转路线的制定和转换算法的描述（包括经济数学模型的描述）；

（2）信息文件的组织；

（3）编制全部框图；

（4）编写全部程序功能的说明；

（5）建立编码手册；

（6）确定标准查询格式的填写规则和实例；

（7）编写系统使用说明书和工作流程；

（8）实现全部硬件描述（包括性能、技术规范）；

（9）实现系统支持软件的描述；

（10）制定系统工作人员的职责条例；

（11）系统出现局部故障和系统全局性故障时的应急措施条例；

实体系统设计是把逻辑系统设计转变为一种能为计算机有效和经济地处理的形式。

（五）系统试运行

（1）运行数据的采集；

（2）人员培训（使用人员）；

（3）运行报告和技术参数测定。

（六）系统鉴定

一般在试运行半年以后进行。

（七）系统维护

本节讨论的计算机辅助企业管理系统设计，是针对要建立的系统对企业的各项业务比较全面的管理而说的。目前对大多数企业要做到上述要求还存在一定的困难。因此，先建立在文件系统之上对企业的单项或局部进行计算机管理，再向更高级的阶段发展，建立一个比较全面的计算机企业管理系统。

本书不是计算机辅助企业管理的专著，不可能全面加以论述，下面对建立在文件系统上的计划管理和生产管理作概要介绍，其他方面的管理就不一一介绍了，可参考有关的专著。

14.2 计 划 管 理

14.2.1 管理计划的基本概念

每一个企业从决定经营开始，就离不开计划。对现代化企业更是如此。计划管理包括经营决策和管理计划两大方面。对于经营决策的制订，因其根据主要不在于企业内部因素，故这里不作讨论，而只讨论管理计划。

一个企业要生产一定种类的产品，还要有质量、数量方面的要求，就需要有一定种类、一定质量和一定数量的原料、材料、外购件和能源（电力、燃料）等，还要有各种具有生产技能的职工以及各种生产工具（机器、仪器、设备）等，这些方面客观上都存在着一定的比例关系。此外，还要保证进行正常生产的动力供应、维护设备等辅助生产单位的能力，它们和直接制造产品的基本生产单位的能力之间也存在着一定的比例关系。

对于一个医院、旅馆和运输公司等，虽然它们并不生产任何产品，没有现代化工厂那样复杂的工艺过程，但同样也存在计划管理的问题。

不难想象，一个企业如果对上述许多比例关系，事前不进行精确的计算和保持一定的平衡，不编制出一个周密的计划表，在生产过程中又不按统一的计划去指挥、调度、控制和管理，那就肯定会乱套，从而造成严重的经济损失。所以说，凡是现代化的企业，都必须实行计划管理。

有计划按比例地发展是我国整个国民经济的根本出发点，而企业的计划是整个国民经济计划的细胞，因此，企业计划的制定和执行的好坏，不仅是一个企业的问题，同时直接关系到整个国民经济计划的执行结果。

企业计划管理的任务有以下两个方面：

（1）必须保证全面完成和超额完成国家下达的计划任务，只有每个企业都做到这一条，才能保证整个国民经济计划的全面完成和超额完成。

（2）在国家指导性计划的基础上，根据本企业的特点和市场预测资料，合理地利用

人力、物力和财力，充分发挥它们的作用，取得最好的经济效果。

作为企业计划，按时间来划分，有长远计划、年（季、月）度计划和作业计划等。

长远计划是企业一年以上的计划，其中包括企业在较长时间内生产、技术和经济发展的重大问题。例如：产品发展方向；生产发展规模；技术发展水平；技术改造措施；设备现代化措施；主要技术经济指标要达到的水平；生产组织；劳动组织；安全设施；环境保护；人才投资；福利设施等等。

年（季）度计划是企业全部生产经济活动的纲领，它主要包括下列一些内容：

（1）生产计划；

（2）辅助生产计划；

（3）技术组织措施计划；

（4）设备维修计划；

（5）劳动工资计划；

（6）物资供应计划；

（7）产品销售计划；

（8）运输计划；

（9）成本计划；

（10）财务计划。

以上列举的是十项基本计划。各企业可根据本企业的具体情况，制订上述十项中若干个计划，也可以根据特殊需要，增订若干个计划，如新产品试制计划、质量计划和基建计划等等。

企业的年（季）度计划的各个部分之间相互依存、相互制约和相互促进的密切关系，其中以生产计划为核心。而生产计划的制订是以企业的计划管理的任务为考虑问题的出发点。作业计划是企业长期计划和年度计划的具体化，是企业用以指导和组织日常生产经营活动的一种计划。企业各种生产经营活动都要编制作业计划。它包括生产、物资供应、生产力平衡、设备台时平衡、动力平衡和运行平衡等方面的月度计划和进度计划。其中，以生产作业计划为主体。它要规定每个生产单位（车间、工段和小组）及每台主要设备在每一个工作单元时间内完成哪些任务。

14.2.2 计划管理的实例

一、计划管理的出发点

对于一个企业来说，制定一个最佳的年度计划是至关重要的，然而这是一件十分复杂的事。它必须考虑到：

（1）国家的经济政策和上级主管机关下达的计划任务和各项技术经济指标；

（2）国内外市场的需求情况；

（3）已订的经济合同；

（4）本企业能够承受的生产能力。

根据上述四方面的考虑因素，可以提出各种各样的投产方案。相应地要研究以下两个问题：

（一）综合测算

对任何一个投产方案而言，要确切掌握本企业在计划期内能否完成。为此，必须对企业的生产能力进行综合测算。在综合测算中主要考虑的因素如下：

（1）对任何一个投产方案，所需要的劳动力、设备生产能力和原材料等同计划期内能提供的劳动力、设备生产能力和原材料等之间是否平衡；

（2）对任何一个投产方案，所需要的各个部门、各个生产环节同企业在计划期内能提供的各个部门、各个生产环节之间是否适应。这里所指的各个部门、各个生产环节是指基本生产部门和辅助生产部门；生产部门内部的各车间；辅助生产部门内部各个环节以及生产准备等等。

（二）对投产方案要作出评价

对投产方案的评价指标主要有：

（1）产品的品种指标；

（2）产品的产量指标；

（3）产品的质量指标；

（4）产品的产值指标；

（5）全员劳动生产率指标；

（6）利润指标等。

通过分析，可了解对一个投产方案是否能够达到上述六项指标。对于前两项指标，在提出备选投产方案时就可以给予评价。至于质量指标，那是在实施计划的过程中加以控制的内容。因此，在确定选择最优方案时，后三项是最根本的，可以作为考核方案优劣的指标。图14-1表示在若干个备选投产方案中选出N个符合考核指标的方案，然后从中选出最优投产方案的算法：下面对示意图作简单说明：

框1中设置的初始考核指标I_0反映了企业领导的意图，I_0一般是由产品的产值指标、全员劳动生产率指标以及利润指标所组成的三元组。

框2及框5都表示输入一个备选的投产方案。

框3，对备选投产方案进行全面综合测算是困难的，一般只能对其中最主要的因素进行平衡测算，例如劳动工时和设备台时要求平衡，其他项目就不作限制，即使作此简化，在实践中也可收到较明显的经济效果。本框在下面将再作介绍，这里就从简了。

对于框4中所指的“可行”，可以用模型来判断，也可以由人来干预作出判断。实践证明，用后一种判断更为实用。

对于框7中的“比较”，在这里是对目标的比较，类同框4，可在两种模式中任选一种。考核指标的比较结果，给出了框8中当前所考察的方案Pn是否达到标准。

框11中的N是事先规定的参数。

在整个示意图中工作量最大的是框3。为了对框3有一个直观了解，下面就具体进行介绍。

下面先介绍有关的几个名词：

定额工人数是指有工时考核指标的生产工人的人数。

设备数是指现有的可用于生产的机器设备台数。对某些有设备的工种而言，有设备就必有操作该设备的定额工人。反之，有些工种是手工操作，虽有定额工人，但没有机器设备。

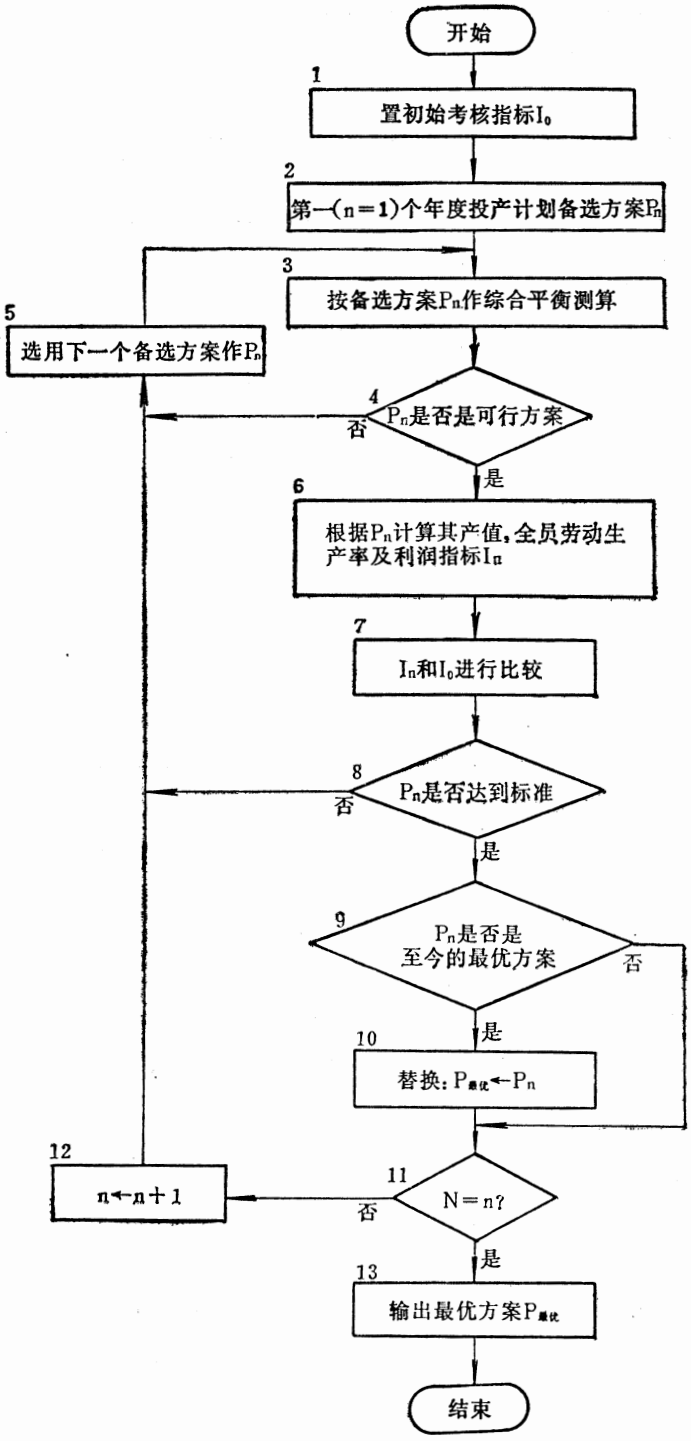


图14-1　选择最优投产方案示意框图

工时利用率是指工人的制度工作时间被利用于生产的程度。可用公式表示为：

工时利用率＝(制度工作时间内实际工作时数/制度工作时数)×100%。

台时利用率是指制定设备开动台时被利用于生产的程度。可用公式表示为：

台时利用率＝(制度工作时间内使用设备的台时数/制定设备台时数)＊100%

工时（或台时）压缩系数是指对原来工时（或台时）定额的压缩程度。例如，若取某工种压缩系数为90.9%，则原工时定额为100，而其压缩工时定额为90.9；原工时定额为120，而其压缩工时定额为120＊90.9%＝109.08。

（三）对备选的投产方案进行测算的主要步骤

下面以电子工厂为例来说明这一测算的主要步骤。

备选方案的输入可采用两种方式，一种是按月份输入投产产品的套数；另一种是把投产的产品分成零件、组件和总调三档（对应工厂中三种类型的车间）输入（以后也称测算表），如有必要对零件、组件再细分成前、后两道工序，对总调也可分成总装、调试两道工序（如图14-2所示）。前者称为粗测，后者称为细测。

序号	任务代号	年度计划	工种类别		月份 1	2	3	4	5	6	7	8	9	10	11	12
1	AT-4	200	零件	(前)			20	20	25	25	(90)					
				(后)					30	30	30	(90)				
			组件	(前)						30	30	30	(90)			
				(后)	30	30	30	(90)			30	30	30	(90)		
			总装		20	30	30	30	(110)			30	30	30	(90)	
			调试			20	30	30	30	(110)			30	30	30	(90)
2	B791	500	零件	(前)	200	200	(400)			200	200	200	(600)			
				(后)		200	200	(400)			200	200	200	(600)		
			组件	(前)			150	150	200	(500)						
				(后)				150	150	200	(500)					
			总装						110	130	130	130	(500)			
			调试								200	200	100	(500)		
⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	H60	8万5千	零件	(前)	1万1	1万1	1万1	1万1	(4万4千)				1万1	1万1	1万1	1万1
				(后)											(4万4千)	
			组件	(前)	1万	1万	1万	1万	(4万)					1万	1万	1万
				(后)											(3万)	
			总装			1万	1万2	1万2	1万2	(4万6千)				1万2	1万3	1万4
			调试												(3万9千)	

图14-2　备选投产方案实例

测算所要的输出有两种表格，一类是对一个产品而言，看每月每季及全年各车间各工种对给定的投产方案需要多少工时、多少设备台时；第二类每一个投产产品或全年投产的所有产品（设有T个），要按车间、工种输出每月、每季及全年需要多少工时、压缩工时、需要多少工人数，根据目前的劳动力布局对每个工种每个月份将余多少人或缺多少人。对设备而言，也要回答上述各个问题。

〔例子〕

测算过程可用图14-3所示的例子来说明。本例的实现是采用FORTRAN语言写成，因此绘制框图按FORTRAN程序要求进行。

（一）数据结构

工作日历文件（CAL文件）：一般是指一年内每个月中扣除厂休日及国定假日的天数以后的每月劳动天数所构成的文件，每年更新一次。

车间文件（WP_i（$i=1,2,\cdots,N$）文件）：设N是一个工厂中车间的最大编号，而实际存在的车间数往往小于N。也就是说，存在着某些不需要参加测算或不存在的车间编号。

在每一个参加测算的车间文件WP_i中，其信息有两部分组成，第一部分是车间资源，其中包括：

1）车间工种数M_i（i为车间编号）；

2）工种的机内编码$WK_i(j)$（$j=1,2,\cdots,M_i$）。

其中，$WK_i(j)$=（工种性质标识符，工种编号）。

工种性质包括零件工种、组件工种及总调工种三大类（即三种标识符），如前所述，每类还可分为两种（即共有6种标识符）。

3）j工种的参数$WKP_i(l,j)$（$l=1,2,\cdots,8$；$j=1,2,\cdots,M_i$）

其中：

$WKP_i(1,j)$——i车间j工种工人数；

$WKP_i(2,j)$——i车间j工种定额工人数；

$WKP_i(3,j)$——i车间j工种工时利用率；

$WKP_i(4,j)$——i车间j工种工时压缩系数；

$WKP_i(5,j)$——i车间j工种设备台数；

$WKP_i(6,j)$——i车间j工种设备班次数/天；

$WKP_i(7,j)$——i车间j工种设备利用率；

$WKP_i(8,j)$——i车间j工种设备台时压缩系数。

车间文件的第二部分是经过若干个产品测算以后，要本车间提供的劳动工时数和设备台时数：

（1）劳动工时数 $WESW_i(k,j)$（$k=1,2,\cdots,17$；$j=1,2,\cdots,M_i$）。

其中，当$k=1,2,\cdots,12$，表示i车间j工种从一月份到12月份测算的工时；当$k=13,14,15,16$，表示i车间j工种一至四季度测算的工时；$k=17$表示i车间j工种全年测算的工时数。

（2）设备台时数 $WESM_i(k,j)$（$k=1,2,\cdots,17$；$j=1,2,\cdots,M_i$）。

其中，当$k=1,2,\cdots,12$表示i车间j工种从一月份到12月份测算的台时数。当$k=13,14,15,16$表示i车间j工种一至四季度测算的台时数；当$k=17$表示i车间j工

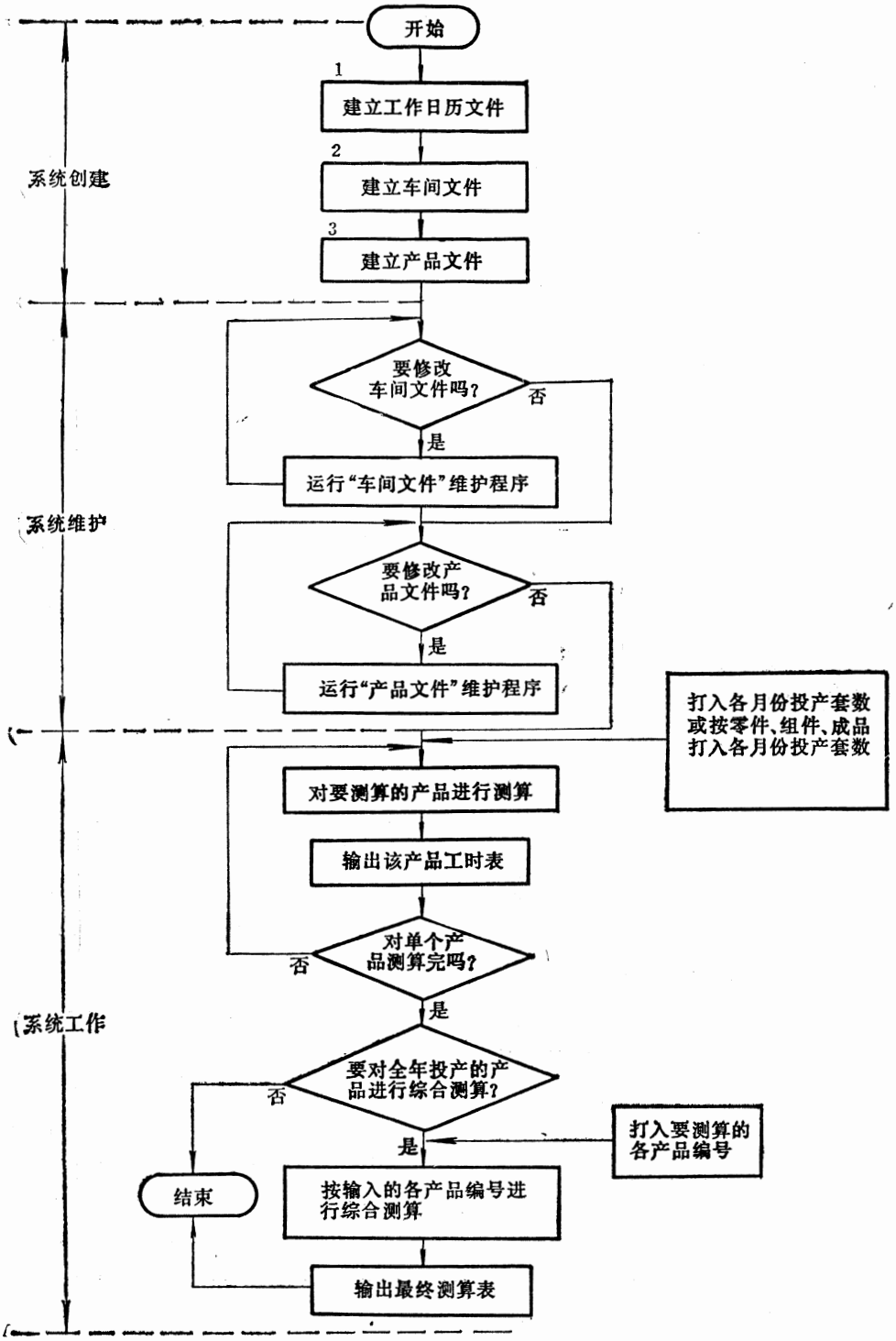


图14-3　备选方案测算过程

种全年测算的台时数。

（3）产品文件 PR_s （$S=1, 2, \cdots, T$）。

设T是具有投产条件的产品数。但本年度或本季度是否投产，要看需要而定。对于每一个具备投产条件的产品，有了工艺流程和数据，可以先组织成产品文件，存储在磁盘中，当需要投产进行测算时，随时可以调出使用。每一个产品文件PR_s有三部分组成。第一部分是关于加工产品的结构信息，其中包括：

PM_s（S号产品所涉及的加工车间数）

$PA_s(u, v)$（$u=1, 2$；$v=1, 2, \cdots, PM_s$）

其中，$PA_s(1, v)$表示S号产品涉及第v个加工车间的编号；$PA_s(2, v)$表示S号产品在$PA_s(1, v)$车间中参与加工的工种数；

令$JS=\sum_{i=1}^{PM_s} PA_s(2, i)$表示$S$号产品所涉及的总的加工工种数，

$PB_s(y)$（$y=1, 2, \cdots, JS$）表示参与加工s号产品的工种编号。这里对于工种编号，在涉及的每一车间的内部以递增次序排列，在车间之间以$PA_s(1, r)$（$v=1, 2, \cdots, PM_s$）顺序排列。

第二部分是产品的基本数据区及输入测算表的存储区：

$PE_s(x, y)$（$x=1, 2, \cdots, 16$；$y=1, 2, \cdots, JS$）

其中，

$PE_s(1, y)$表示$PA_s(1, y)$　　车间$PB_s(y)$工种工时定额；

$PE_s(2, y)$表示$PA_s(1, y)$　　车间$PB_s(y)$工种准备工时；

$PE_s(3, y)$表示$PA_s(1, y)$　　车间$PB_s(y)$工种台时定额；

$PE_s(4, y)$表示$PA_s(1, y)$　　车间$PB_s(y)$工种准备台时；

$PE_s(x, y)$，当$x=5, 6, \cdots, 16$时，依次表示$PA_s(1, y)$车间$PB_s(y)$工种从一月份到十二月份投产套数（测算表）。

第三部分是测算后的工时数、台时数：

$PESW_s(x, y)$（$x=1, 2, \cdots, 12$；$y=1, 2, \cdots, JS$）依次表示$PA_s(1, y)$车间$PB_s(y)$工种x月份测算后的工时数；

$PESM_s(x, y)$（$x=1, 2, \cdots, 12$；$y=1, 2, \cdots, JS$）依次表示$PA_s(1, y)$车间$PB_s(y)$工种x月份测算后的台时数。

（二）系统结构

系统结构大致可分成三大部分：系统创建、系统维护和系统工作。系统工作中又可分对一个产品的测算，对多个产品的综合测算，以及表格输出等（见图 14-3 所示）。

1．系统创建　建立工作日历文件、车间文件和产品文件。在系统创建时建立的车间文件中实际上只包含车间资源部分，对其第二部分的信息只有进行综合测算时才能形成；对产品文件也类似在系统创建时只包含关于加工产品的结构信息和产品的基本数据，其它部分都是在系统工作时形成。

2．系统维护　对于车间文件、产品文件（包含输入的测算表）中某些信息有变动，例如人员的变动，设备的增减，工时利用率和台时利用率的更改等等都可通过人机对话方式进行逐项修改。平时均作为磁盘文件而被长期保存。

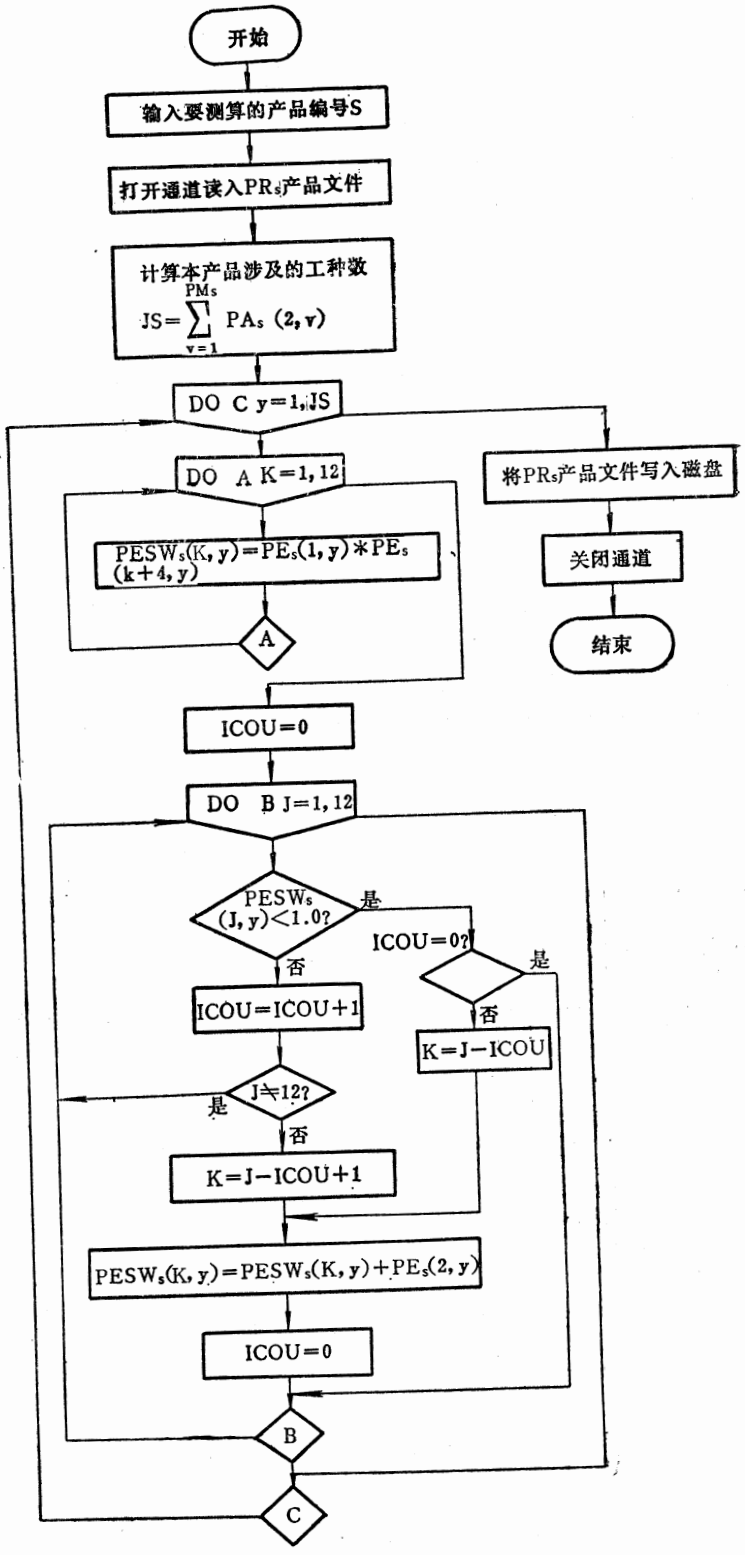


图14-4　一个产品工时的测算框图

（三）系统工作——对一个产品的测算（图12-4）

首先把要投产产品的测算表，即备选投产方案中一个产品 S 月份和工种类别输入计算机，程序将它们送入 S 产品文件的 $PE_s(x, y)$（$x=5, 6, \cdots, 16$）中相应的存储单元内。当测算表输入结束后，自动打印出刚输入的测算表，以便校对和存档。然后，对每一个投产产品，在每一月份中所要花费的工时、台时进行测算。测算模型如下：

如果 k 月不是投产的头一个月，则

$$PESW_s(k, y)=PE_s(1, y)*PE_s(k+4, y)$$
$$PESM_s(k, y)=PE_s(3, y)*PE_s(k+4, y)$$

如果 k 月是投产的头一个月，则

$$PESW_s(k, y)=PE_s(1, y)*PE_s(k+4, y)+PE_s(2, y)$$
$$PESM_s(k, y)=PE_s(3, y)*PE_s(k+4, y)+PE_s(4, y)$$

程序实现一个产品的工时测算如图 14-4 所示。

要输出一个产品的测算工时，现在已可进行。但是对一个产品进行测算时对计划成本的核算是必要的，对选择“最佳的”备选投产方案也是必要的，但这样还没有达到最后目的，要对本年度（或本季度）所有投产产品进行测算，才能看出输入备选方案的好坏。为此，要进行综合测算。

（四）系统工作——对全年投产产品的综合测算（图14-5）

综合测算前，要求操作员打入所有要测算的产品编号是完全合格的（这些产品都需单独测算过）。

为了叙述的方便，继前面符号规定，且不失一般性，假设 PR_s 产品文件中的 y 指向 i 车间 j 工种，那末 PR_s 需要 i 车间 j 工种付出的工时和台时数分别为

$$\left.\begin{aligned} WESW_i(k, j)|_s &= PESW_s(k, y) \\ WESM_i(k, j)|_s &= PESM_s(k, y) \end{aligned}\right\} (k=1, 2, \cdots, 12)$$

对于添加一个测算产品 S 以后，i 车间 j 工种要付出的工时、台时数，可用下列递归表达式表示：

$$WESW_i(k, j)_{\text{添加后}}=WESW_i(k, j)_{\text{添加前}}+WESW_i(k, y)|_s$$
$$WESM_i(k, j)_{\text{添加后}}=WESM_i(k, j)_{\text{添加前}}+WESM_i(k, y)|_s$$

递归过程执行到本年度投产的产品全部测算结束为止。

设 i 车间 j 工种在 k 月份压缩工时为 $CWT_i(k, j)$，压缩台时为 $CMT_i(k, j)$，余缺工人数为 $RW_i(k, j)$，余缺设备台数为 $RM_i(k, j)$，则它们之间的关系为

$$CWT_i(k, j)=WESW_i(k, j)*WKP_i(4, j)$$
$$CMT_i(k, j)=WESM_i(k, j)*WKP_i(8, j)$$

而

$$RW_i(k, j)=WKP_i(2, j)-CWT_i(k, j)/(8*WKP_i(3, j)*CAL(k))$$
$$RM_i(k, j)=WKP_i(5, j)-CMT_i(k, j)/(PWKP_i(6, j)*WKP_i(7, j)*CAL(k))$$

其中

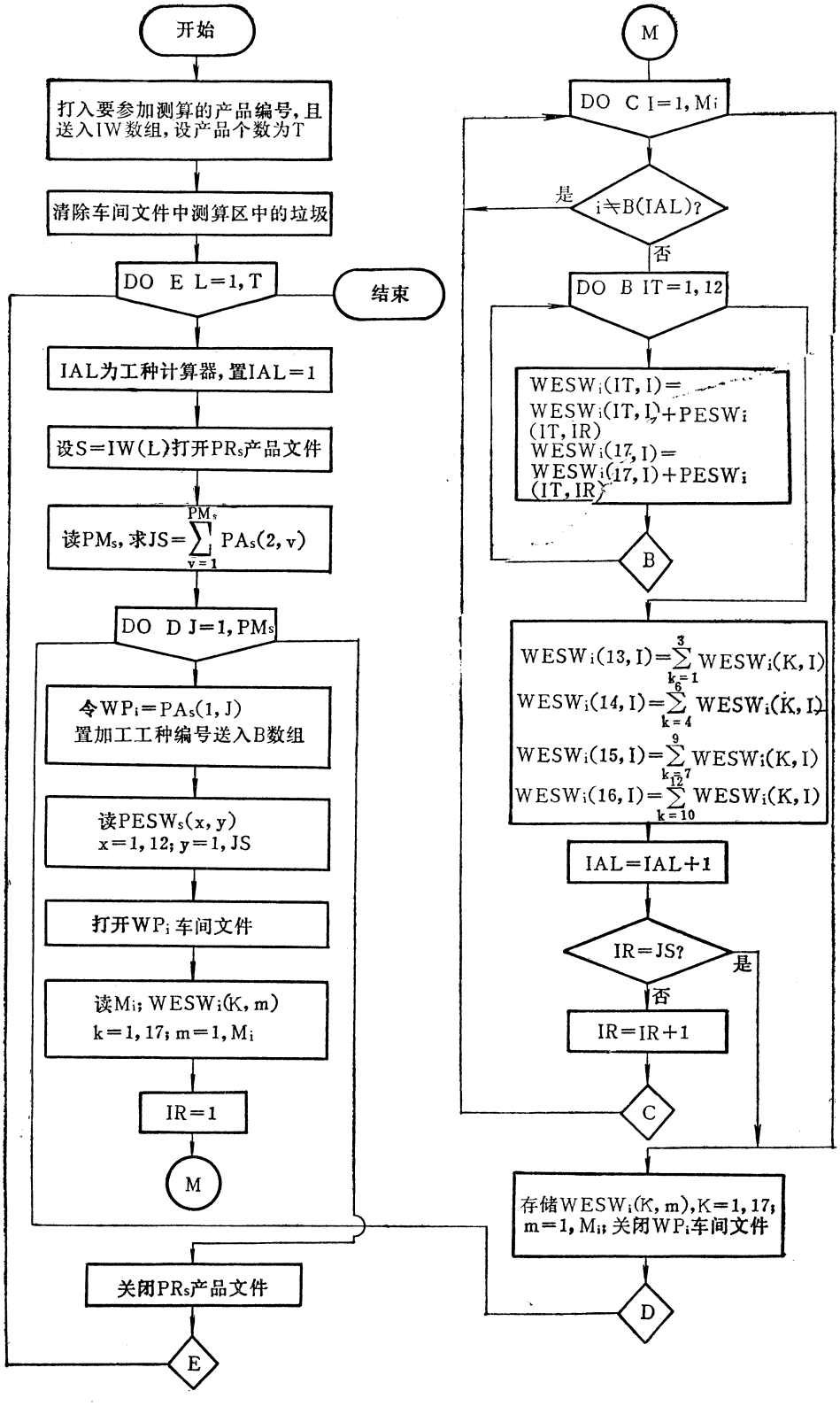


图14-5　T个产品关于工时测算程序框图的核心部分

$$PWKP_i(6,\ j)=\begin{cases}8 & \text{当}WKP_i(6,\ j)=1\\ 15.5 & \text{当}WKP_i(6,\ j)=2\\ 22.5 & \text{当}WKP_i(6,\ j)=3\end{cases}$$

对于 i 车间 j 工种第 l 季度的压缩工时数、台时数、余缺工人数以及余缺设备台数分别为：

$$\left.\begin{aligned}CWT_i(12+l,\ j)&=\sum_{m=1}^{3}CWT_i(3*(l-1)+m,\ j)\\ CMT_i(12+l,\ j)&=\sum_{m=1}^{3}CMT_i(3*(l-1)+m,\ j)\\ RW_i(12+l,\ j)&=\sum_{m=1}^{3}RW_i(3*(l-1)+m,\ j)\\ RM_i(12+l,\ j)&=\sum_{m=1}^{3}RM_i(3*(l-1)+m,\ j)\end{aligned}\right\}(l=1,2,3,4)$$

i 车间 j 工种全年的压缩工时、台时、余缺工人数以及设备台数分别为

$$CWT_i(17,\ j)=\sum_{l=1}^{4}CWT_i(12+l,\ j)$$

$$CMT_i(17,\ j)=\sum_{l=1}^{4}CMT_i(12+l,\ j)$$

$$RW_i(17,\ j)=\sum_{l=1}^{4}RW_i(12+l,\ j)$$

$$RM_i(17,\ j)=\sum_{l=1}^{4}RM_i(12+l,\ j)$$

图 14-5 表示 T 个产品关于工时测算程序框图的核心部分。

（五）表格输出

根据上面各个算式，对于投入的测算表，可以正确地得出在有定额的车间、工种中工时、台时的需求以及造成劳动力和设备的余缺情况。按年度、季度以及月份完全根据用户的要求格式输出各种类型的表格，对这些技术上容易实现的工作，这里就不一一赘述了。

14.3 生 产 管 理

14.3.1 生产管理的基本概念

生产管理是指如何组织实现企业生产计划的一系列管理工作，即对企业的基本生产过程的日常作业活动进行计划、组织、控制和调整，以保证全面、均衡地完成企业的生产任务。

所谓生产过程是从准备生产起，直到把成品生产出来的全过程，或者到完成劳务的全过程。构成生产过程主要有生产准备、基本生产、辅助生产和生产服务四个过程，它们之间既有区别又有联系，而其中基本生产过程占主导地位，其他过程是围绕基本生产过程进行的，是为基本生产过程的执行创造良好的条件。产品生产的基本生产过程不仅被划分为各个工艺阶段或局部过程，而且每一个工艺阶段或局部过程又可分为不同的工种和一系列上下联系的工序。工序是生产过程中的基本环节，因此，工序也是生产过程中的最小单元，又是配备工人、计算劳动量、确定生产组织形式、编制作业计划和进行质量管理的基本单位。所以工序划分是否合理，对企业的技术经济效果有直接影响。

前面介绍的年度生产计划，规定了企业在一年内的生产总任务和各项指标。但它不可能预见到生产过程中的各个细节和可能产生的变化，要作出符合实际的安排，还必须有更可靠的行动依据。要使生产过程的不同工艺阶段之间，不同的工序之间，在时间上衔接并在空间布局上合理，还要求使生产过程满足连续性、比例性、平衡性和节奏性。因此在年（季）度生产计划的基础上制订出把计划落实到每个车间、工段、班组的作业计划是十分重要的。进一步，在按作业计划生产的过程中，又往往会出现很多意想不到的问题，如某些产品（或零、部件）在某一段时间内未完成计划、而另一些产品(或零、部件）却超过了计划的要求，这时生产管理又要调整作业计划。因此，及时地收集生产过程中的各种信息在生产管理中显得十分必要，它们是调整作业计划的重要依据。

14.3.2 生产管理实例

车间作业计划是车间生产管理的一项重要内容，作业计划安排得好坏，在执行过程中调度情况的好坏，对完成车间各项技术经济指标有很大影响，弄得不好要影响其他车间和全厂生产任务的完成。这里以机械加工车间为例，说明车间作业计划的编制原理。

车间作业计划要解决的问题是在给定车间劳动力和设备数量的情况下，要求加工的零件的工艺流程和工时定额是已知的，在一个计划期内要求加工完成一定数量的零件，问如何安排作业计划，才能如期或提前完成任务？

当然，实际情况比较复杂。首先，它是一个动态的问题，即在制订作业计划时，各车间、各班组的生产活动并没有停止。此外，加工零件的定额工时可能有改变；参加加工的工人数以及设备数也可能有改变，这些问题使得安排作业计划增加了困难。但是，有计划比没有计划好，定出合理的计划，然后归结成通用程序，让计算机去执行，就可以收到良好的效果。这里我们简述它的设计原理。为了简化问题，假设每一设备若加工某一零件就必须要加工完，不可中途撤下来。此外，还假设机器不会中途出故障，劳动力、原材料和零件是足够的等等。

安排作业计划的原则规定为：

（1）加工时间紧迫的、重要的、难加工的零件优先安排；

（2）大部件、工序较多的先安排；

（3）小件，灵活机动，见缝插针。

我们根据这些原则来安排作业计划。例如车间内有车床2台、刨床2台及铣床一台。零件加工顺序及其相应的加工时间如表14-3所列。表中“对应的加工时间”内包括加工的准备时间，时间的单位为天，月初作为计划期开始。

表14-3 零件加工顺序及相应的加工时间表

零件编号	加 工 顺 序	对应加工时间	总 计
1	车→铣→车→刨→铣	1，3，4，2，6	16
2	铣→刨→车→铣→车	5，4，1，2，3	15
3	车→刨→铣→车→刨	4，3，2，1，1	11
4	车→铣→车→刨	1，2，3，4	10
5	车→刨→铣→车→刨→铣	3，4，1，2，5，6	21
6	车→铣→刨→车→铣→刨	2，3，4，1，2，5	17

安排这一作业计划中第一个碰到的问题是：一开始有5个零件要求在车床上加工，而车床只有2台，先加工哪一个呢？根据安排作业计划的原则，时间紧迫的先加工。那么，时间紧迫程度如何用数量来描述呢？如果一个零件在车间加工工序多，时间长，整个的加工时间离要求完成的期限就近，这样我们说这个零件加工时间紧迫。例如一个零件要求用两道工序，先在刨床上加工13天，而后在磨床上要花16天，但任务要求30天完成，那末这个零件只有一天可以松动的时间，这个零件的加工当然是比较紧迫的。一般可描述为：

设T_i是零件i要求完成的时刻，t是某一时刻。$\sum_{(k)>t} a_k^i$是零件i从时刻t以后所有工序所需要的加工时间的总和。定义

$$S_t^i = T_i - t - \sum_{(k)>t} a_k^i$$

其中，a_k^i表示零件i的第k次加工时所需的时间；$(k)>t$表示时刻t以后的工序，称S_t^i为零件i在时刻t的松紧指标，也称为上机优先指标。此值越小，说明时间限得越紧。实际上，S_t^i是零件i在时刻t以后的空闲时间。本例中一开始有5个零件要求在车床上加工，若所有零件都要求第28天完成（各个零件可以不同），而车床只有2台，那么先加工哪个零件呢？为此，要先计算松紧指标，这时$t=0$，$T_i=28(i, 1,3, 4, 5,6)$。所以

$$S_0^1 = 28 - 0 - 16 = 12$$
$$S_0^3 = 28 - 0 - 11 = 17$$
$$S_0^4 = 28 - 0 - 10 = 18$$
$$S_0^5 = 28 - 0 - 21 = 7$$
$$S_0^6 = 28 - 0 - 17 = 11$$

下面我们要用的指标是每道工序平均松紧指标。为简单计，以后所说的指标或松紧指标都是指工序的平均松紧指标

$$\overline{S}_t^i = S_t^i / N_t^i$$

其中，N_t^i为时间t以后零件i所要加工的工序数。于是有：

$$\overline{S}_0^1 = 12/5 = 2.4$$
$$\overline{S}_0^3 = 17/5 = 3.4$$
$$\overline{S}_0^4 = 18/4 = 4.5$$
$$\overline{S}_0^5 = 7/6 \approx 1.17$$
$$\overline{S}_0^6 = 11/6 \approx 1.83$$

松紧指标是零件加工时间限得是否紧迫的一种量度。从上面可以看出，零件5是限得最紧的，其次是零件6。所以，首先要在车床上先加工这两件零件。例如让车床1加工零件5，车床2加工零件6。

假设有零件在机器旁“排队”等待加工，我们就可以根据一定的法则，知道哪一个零件先进行加工，哪一个零件排在后面。作业计划就是根据这种思想来制定的。当几个零件的松紧指标都同样为最小时，原则上可以随意挑选，但一般先上零件编号小的。图14-6所示实际上是车间加工作业的一个过程。

根据前面的计算，先安排零件5和6在车床上加工，零件1，3，4排队等待上车床加工，零件2在铣床上加工。进行到时刻2，零件6在车床上加工完毕，它要求在铣床上加工，可是铣床不空，所以零件只好等待，此时车床2空，可以在排队等待它加工的零件中挑一个进行加工，它们的松紧指标分别是：

$$\overline{S}_2^1=(28-2-16)/5=2$$

$$\overline{S}_2^3=(28-2-11)/5=3$$

$$\overline{S}_2^4=(28-2-10)/4=4$$

所以，其中零件1上车床2加工。继续进行到时刻3，车床1，2上加工的两个零件都完成了，当然此时一台上零件3，另一台上零件4。这时车床上完工的两个零件，一个是零件5，另一个是零件1。零件3要求在刨床上加工，此时刨床空闲，零件3可立即上去加工。另一个零件1，要求在铣床上加工，而铣床此时不空，它和零件6一起排队等待上铣床加工。到时刻4，在车床上加工完毕，它又要求在铣床上加工，也参加等待，此时等待铣床加工的零件有三个，即零件1、4、6。到时刻5，在铣床上加工的零件2完工，要在等待的零件中选一个，其松紧指标是：

$$\overline{S}_5^6=(28-5-15)/5=1.6$$

$$\overline{S}_5^1=(28-5-15)/4=2$$

$$\overline{S}_5^4=(28-5-9)/3=14/3\approx4.7$$

选中零件6上去加工。进行到时刻7，零件3在车床上完工，它又要求在刨床上加工，就立刻可以安排。零件5在刨床上加工完毕，又要求在铣床上加工，所以它和零件1、4一起排队等待。进行到时刻8，零件6在铣床上加工完毕，又要求在刨床上加工，此时刨床不空，就等待。这时铣床又要挑选加工零件。在等待它加工的三个零件的松紧指标分别为：

$$\overline{S}_8^1=(28-8-15)/4=1.25$$

$$\overline{S}_8^5=(28-8-14)/4=1.5$$

$$\overline{S}_8^4=(28-8-9)/3\approx3.67$$

选中零件1上铣床加工。

照上面办法做下去，其结果如图14-6所示。

表14-4是各个零件在此作业计划中完成的时间，这样安排作业计划无疑是比较好的，而且有很强的适应性。

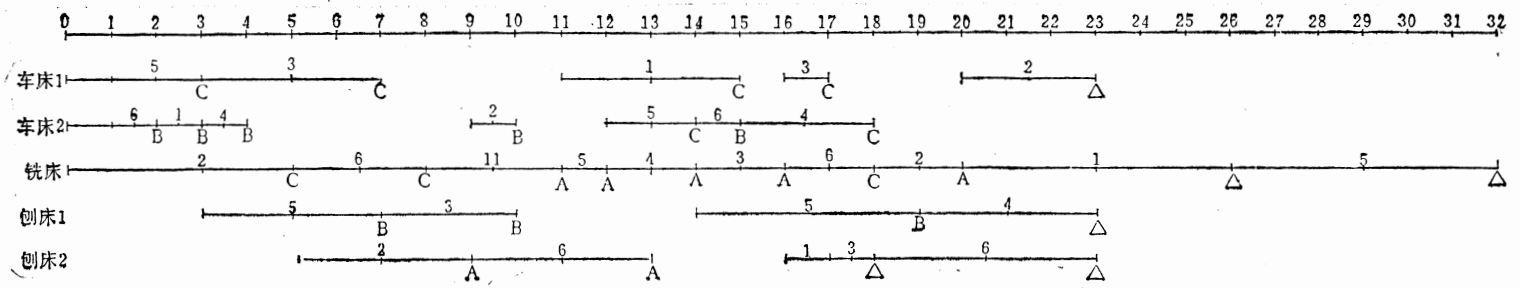


图14-6　车间作业计划安排实例

表14-4　实例中各零件完成时间表

零　件	完成时间(天)	零　件	完成时间(天)	零　件	完成时间(天)
1	26	3	18	5	32
2	23	4	23	6	23

平均完成时间是24.17天

14.4 现代企业管理中的汉字信息处理实例

14.4.1 概述

在我国，企业管理的各个方面，不论是计划、生产、物资或财务等票证、表格的处理，都不可避免要用到汉字。不少单位曾在计算机企业管理的输出中采用汉字拼音表示，但因为汉字拼音目前使用还不普及，所以都未收到良好的效果。那末，在现代化企业管理系统中是如何处理汉字信息的呢？我们不是针对企业管理中每一功能去解决它的汉字信息的处理方法，而是综合起来进行分类，然后根据不同的类型采用不同的处理技术。这样做的结果肯定要比针对每一功能去解决要有效得多。

（一）我国现代企业管理中所要用到的汉字信息形式

主要有以下三个方面：

1．汉字文件档案　这里所说的文件档案是指诸如报告、公文、记录、通报、函件之类的东西。

2．汉字表格　系指统计报表、发票、合同、账册等。

3．汉字统计图表　系指附有汉字说明的点图、折线图、条状图和圆形图等。

（二）企业管理中需要使用的汉字信息处理设备

1．汉字输入设备　诸如：汉字键盘（整字键盘、字根键盘或标准字母数字键盘）；手写体字、表格、图形的识别设备；汉语声音输入设备等。

2．汉字输出设备　诸如：汉字显示器；汉字打印机；汉字制表机；汉字绘图仪；汉语声音输出设备等。

3．汉字显示终端设备　它同时具有汉字输入、输出功能。

4．汉字字模库　在有些汉字企业管理系统中，汉字字模是放在磁盘上的，故又称之为汉字软字模库。采用这种方式可以减少硬件费用，但汉字字模的存取速度要慢得多。现在逐步被大规模集成电路字形发生器所替代。此外还可以采用压缩信息存储的汉字字模库。

（三）企业管理中所需要的汉字软件

在企业管理中所用的汉字信息处理系统所需的汉字软件一般包括：

（1）汉字文件档案处理程序；

（2）汉字表格生成程序；

（3）汉字图表生成程序。

上述三种处理程序一般采用非过程性语言或命令来实现。

对汉字信息的处理程序进行优化的目的是为了尽可能少地占用存储空间和提高处理速度。但因这两方面的要求往往是矛盾的，因此应针对具体情况有所侧重。此外，若汉字字模的存储采用软字库的形式，则必须附有建库程序和字模库的维护程序等。

下面就汉字文件档案和汉字表格的处理为例作简单介绍，因为这两种形式在企业管理中的使用面最广。

14.4.2 汉字文件档案处理实例

这里介绍的面向企业管理的汉字文件档案处理实例被配置在国产DJS100 系列机上，适用于一些对汉字输出字形要求不高，但又是必需的应用例子。

鉴于上述考虑，以及我国目前硬件费用比较昂贵这一情况，确定以下一些原则：

一、设计的原则

（1）要处理的信息符合国家标准“信息交换用的汉字编码字符集——基本集”，并要考虑到国家标准的汉字编码的扩充；

（2）汉字点阵采用15×16，点阵每列点数与机器字长一致，这样既提高存储效率，又便于程序处理；

（3）为了降低成本，除了可用价格低廉的硬件来实现的功能外，其余功能尽可能采用软件方式来实现；

（4）程序设计中采用数据库设计思想。在程序编制中尽可能注意少占内存空间。

二、汉字文件档案处理中软、硬件的配置

（一）软件配置

一般应配置如下一些软件：

（1）实时磁盘操作系统（RDOS）；

（2）汉字输入处理程序（CHIP）；

（3）汉字编码程序（CEDIT）；

（4）汉字输出处理程序（CHOP）；

（5）汉字软字模库（$CWB_i(i=1, 2, \cdots, N)$）及维护程序。

（二）硬件配置

一般应配置如下一些硬件：

（1）DJS100 系列机（内存 32K字）；

（2）磁盘（DISC），容量为 5 兆字节；

（3）汉字针式打印机（CHO）；

（4）汉字键盘输入装置（CHI）；

（5）DJS100 系列机常规外部设备。其中包括：电传机（TTI，TTO）；光电机（PTR）；纸带凿孔机（PTP）等。

三、对汉字针式打印机（CHO）的基本要求及主机调用格式规定

1．基本要求　至少要有 256 字的静态随机存储器；受控点针不少于 16 根；有对 GB1988 代码的处理能力；可代替行式打印机；其输出点阵为 8 × 8，比一般打印的汉字字形小一半。此外，还可以借助软件实现正反打印和绘图功能。

2．主机调用格式　用 DOA 指令传送点阵码；用 DOBS 传送操作控制，其规定如图 14-7 所示。

在主机对操作的控制调用中，按照回车、移头、打印及走纸之间的关系规定其优先级为

回车 | 优先级高
移头
走纸 ↓ 优先级低

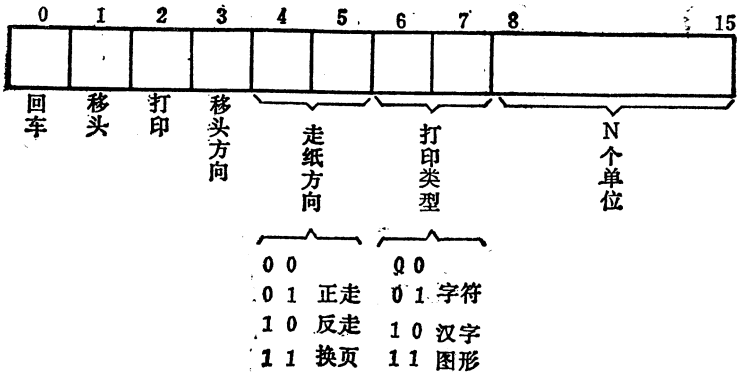


图14-7 控制码格式

当主机对操作的控制调用中有移头及走纸时，指令中的N表示移头的单位长度数，这时硬性规定正走纸一个行距。

四、汉字输出处理程序（CHOP）

（一）CHOP的结构层次

CHOP的结构大致分为三层。内层是指对汉字字模库的物理存储形式，由于本系统是在实时磁盘操作系统控制下工作的，因此对具体的物理存储由操作系统来承担。至于操作系统按什么方式存储这是由设计者决定的。这里，汉字字模库采用的是连续文件，直接读写方式。外层也就是用户界面。概念层处于用户界面和操作系统界面之间，它通过USF/CON、CON/DOS两个映照进行联结，实现信息的转换及传递。图14-8示出了汉字文档处理结构实例。其中的软接口（DRP）实际上就是驱动程序。

（二）用户调用CHOP的方式

我们在叙述各层次内部结构前，先介绍一下用户调用方式。

当用户调用CHOP时，首先要做两件事：（1）汉字针式打印机要借助于IDEF命令介绍给（RDOS）操作系统，以作为用户定义设备；（2）打开汉字字模库文件通道。然后进行调用。调用方式的规定如下●：

（1）AC_0：置要CHO打印的汉字、西文和数字字符编码区字首地址。

（2）AC_1：置0，表示汉字、西文和数字字符编码采用格式1；置1，表示汉字、英文和数字字符编码采用格式2。（格式说明见下文）

（3）AC_2：第0位为0，表示本次输出从CHO左首开始打印；第0位为1，表示本次输出从CHO当前位置继续打印；第1位到第15位全为0，表示未装配FSC命令（只适用于行写）；其他表示FSC(功能区）字首地址。

```
JSR CWRL          (调用子程序)
错返              (出口)
正返              (出口)
```

● AC_0，AC_1，AC_2均为DJS100系列机的累加器。

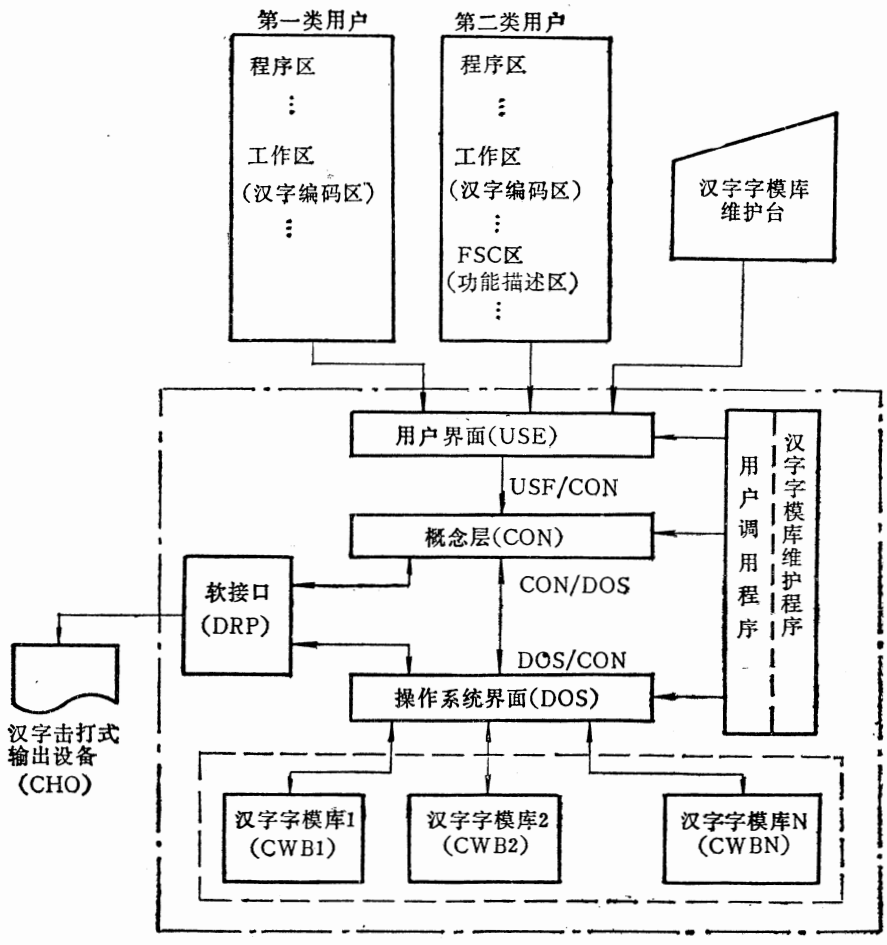


图14-8　汉字文档处理结构实例

这种调用方式类似于 RDOS 下的系统调用命令。RDOS 控制下的输出命令都是用行写。在相当于此调用中 $(AC_2)_{1\sim15}=0$ 的情况下，可扩充行写命令为篇写命令。也就是说，当用户要汉字输出设备打印数行，以至整篇文章时，只要一次调用就可以了。而对于文章中的空格、换行、换页等操作，则可以通过 FSC 功能描述来实现。在汉字编码区中，只要依次填入有效汉字编码即可。这样，既提高了编码区的存储效率，又加快了处理速度。为此，只需在调用时设置输出功能描述命令，并在 AC_2 累加器中存储其字首地址。

FSC 区中的命令采用非过程描述形式，每一命令占两个字节，左字节存放实现控制有关参数的二进制码 N，右字节存放功能码 F(用 GB1988-80代码表示)：

左字节	右字节
N	F

其中，F 表示 W、M、L、C、A 中的任意一个；

NW 表示连续打印 N 个汉字；

NM 表示使打印头移动 N 个汉字位置；

NL 表示走纸 N 行；

NC 表示回车并走纸 N 行；

NA 表示连续打印 N 个字符。

FSC 区中的命令的存储方式同汉字编码区存储方式一样，根据先后次序顺序存放，最后加空白表示结束。

五、从用户程序传送到用户界面（USF）的编码信息的格式

其格式为

FA：$B_1P_1\llcorner B_2P_2\llcorner\cdots\llcorner B_NP_NE$

这里，FA 表示存放编码信息区的首地址；B_iP_i 表示国标码；$\llcorner$ 表示空格符；N 表示本次要输出的汉字个数，E 是空白字符，表示结束。

对任一国标码 $Y=BP$，这里 B 称为国标码 Y 的高位，P 称为国标码的低位，均以十六进制表示。进一步，根据国标码的结构有：

$\overline{B}=(B-20_{16})_{10}$ 称为国标码 Y 的区号；

$\overline{P}=(P-20_{16})_{10}$ 称为国标码 Y 的位号。

这里，$\overline{B}$，$\overline{P}$ 为十进制表示。

我们定义

$$\overline{Y}=(\overline{B}-1)*95+\overline{P}$$

$$\widetilde{Y}=(\overline{Y})_8$$

$\widetilde{Y}$ 称为国标码 Y 的机内码。

因此，对于来自 RDOS 控制下的保存文件，或来自通信网络的代码，或来自用户终端的国标码 Y，在 USF/CON 映照下将其变换成 $\widetilde{Y}$，

$$\widetilde{Y}=f(Y)=\{[(B-20_{16})_{10}-1]*95_{10}+(P-20_{16})_{10}\}_8$$

六、概念层的四项基本任务

（1）接收 USF/CON 映照 f 所获得的机内码。通过散列函数，求出该汉字机内码对应的汉字字模子库名；

（2）知道了该汉字编码的子库名，通过 CON/DOS 映照，通知 RDOS，对汉字字模子库的物理存储进行检索、读出或写入等操作；

（3）由 RDOS 读出一组汉字点阵信息，从中检索出需要的汉字点阵信息，把它们装配到点阵码输出缓冲区或用新的汉字点阵进行替换；

（4）通过软接口（驱动程序）使汉字点阵码的输出缓冲区和汉字针式打印机（CHO）的静态随机存储器进行通信联络，实现汉字的输出打印。

下面围绕上述任务作扼要说明：

设 $\widetilde{Y}$ 是某汉字编码 Y 的机内码。又设 M 是每一汉字字模子库中存有汉字个数的上限，$M\equiv 0\ (\mathrm{mod}16)$，则

$I=\lfloor\widetilde{Y}/M+1\rfloor$● 表示汉字字模子库的编号，它说明汉字编码 Y 的点阵属于汉字字模子库 CWBI(这是个组合字作为汉字字模子库名；其中 I 是由上述散列函数求得的整数)。

● $\lfloor x\rfloor$ 表示 x 的底限，取最大整数值。

令 $Y'=\widetilde{Y}-(I-1)*M$，表示汉字编码Y在汉字字模子库 CWBI 中的分库码，现在已把检索汉字编码Y在汉字字模库中点阵码的问题化为检索分库码 Y'在汉字字模子库 CWBI 中点阵码的问题。

由于 RDOS 所提供的直接读写数据的命令中，最小单位是一个盘区，因此，在概念层中有必要安排中间暂存区，将 Y'所在的盘区信息读出后暂存在里面，程序从中挑选 Y'所对应的一组点阵码，当需要更新时，就进行更新。当需要装配到输出缓冲区时，就进行装配。为了提高速度，减少访盘次数，不是说对每一个汉字编码都要进行访库，而首先询问在当前中间暂存区中是否已有该汉字编码的点阵码。若有，则可直接装入输出缓冲区；若没有，则再进行访盘。

CHOP 和 CHO 之间的软接口由驱动程序 DRP 来实现。如前所述 CHOP 建立在实时磁盘操作系统（RDOS）上，当要输出汉字信息时，首先面临怎样调用汉字针式打印机的问题。解决这个问题的方法是借助于 RDOS 系统命令 · IDEF 将 CHO 介绍给 RDOS，使它作为 RDOS 的一个有定义的用户设备，而 CHOP 把检索到要 CHO 输出的点阵码通过 DRP 接口程序送到 CHO 的缓冲存储器中。因此，DRP 的设计就是起 CHOP、CHO 及 RDOS 之间的协调和联系作用。

上面提到的输出缓冲区是用来暂存检索到的点阵信息。为了提高输出缓冲区的使用效率，可以把它设计成动态循环的存取工作方式。即从概念层送来的点阵信息可以不断地被装入缓冲区，而 DRP 也可以不断地从中取出信息交给 CHO 输出打印。为了实现缓冲区的动态循环存取，设置了信息装填指针 LOAD 和信息取出指针 UNLOAD，同时还设置操作码存储区域 OPRA，用来存放概念层送来的操作码（操作码是具有实现 CHO 作移头、走纸、换行、回车和打印等操作的功能代码）和该操作码前输出信息的长度。若 LOAD=UNLOAD 且 OPRA≒0，则表示缓冲区装满，这时只允许取而不允许装；而 OPRA=0 则表示缓冲区空，这时可以装入信息。一旦 LOAD≒UNLOAD，则表示可装可取。软接口（DRP）示意图如图14-9所示。

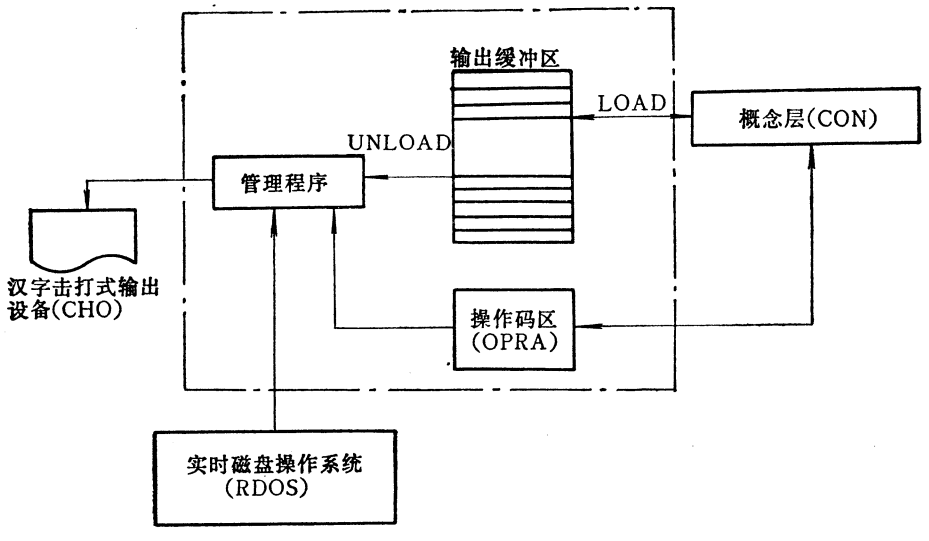


图14-9　软接口示意图

七、CON/DOS 映照

现在我们来介绍 CON/DOS 映照，设 Y' 是 Y 在 CWBI 中的分库码，则 $D=\lfloor Y'/16_{10} \rfloor$ 是编码 Y 的点阵码在 CWBI 中的相对盘区号，而 $A=(Y'-D*16_{10})*16_{10}$ 就是 Y 的点阵码在相对盘区号 D 中的相对首地址（参阅图14-10的实例）。

这里，汉字字模库 CWB 是由 N 个子库 CWBI($I=1, 2, \cdots, N$) 组成的。根据 RDOS 使用方式，要建立任何一个子库只需建立对应的一个连续的磁盘文件即可。

上面我们花费了较长的篇幅讨论了汉字文件档案的输出问题。采用这种方法的特点是以“软”为主，即汉字字模库存放在磁盘介质上，当需要使用汉字字模时，再通过程序调用。这是一种比较经济的实现方法。另一种方法是以“硬”为主，即把常用的汉字字模信息存入只读存储器中作为固件，再适当配置一些控制电路做成插件板放在主机或终端内，把不常用的汉字字模存放在磁盘介质上。采用这种方法的优点是汉字输出速度快，输出占用的机器时间少，尽管造价要贵一些，但随着硬件价格的逐年下降，将会得到广泛的应用。

对现代化企业管理系统要解决的另一个问题，报表的功能也是很需要的。

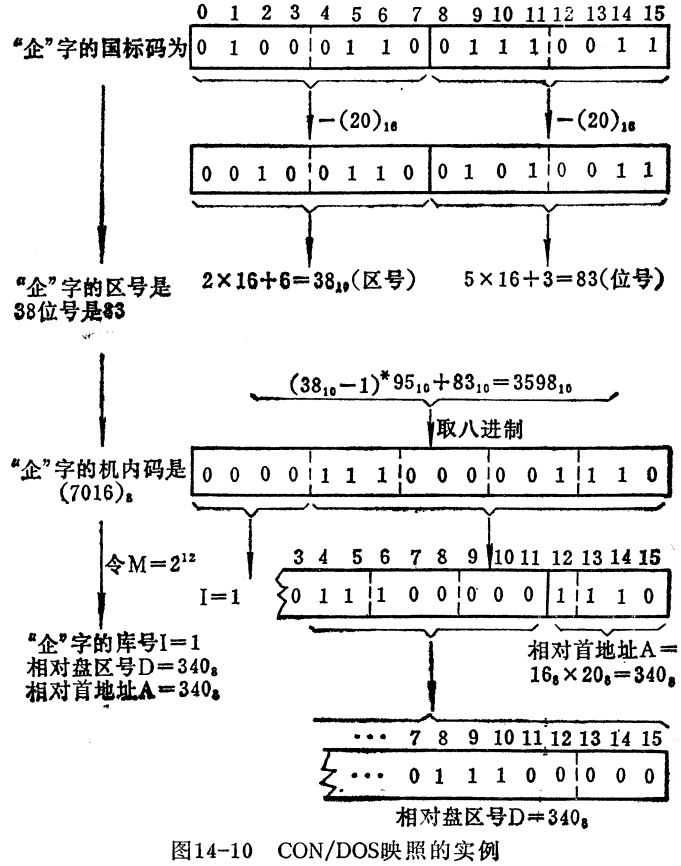


图14-10　CON/DOS映照的实例

14.4.3　汉字制表语言

现代化企业管理中，作为人与人之间信息交往大多采用单据、票证和统计报表。在计算机辅助企业管理系统中，尽管它已减少了许多内部信息和单据的往来，但是作为输

出结果还是有许多单据、票证和统计报表需要制作，因此制表语言对于企业管理是不可少的。目前很多计算机系统上大多配置了制表语言或报表生成程序，但都只能处理西文，不能处理汉字。在我国，计算机辅助企业管理系统若不配置汉字制表语言，则无疑是一个很大的缺陷，甚至不能成为一个实用的系统。

下面介绍一个简单的汉字制表语言，而后举些实例。

一、汉字制表语言概述

这里介绍建立在关系型数据库基础上的汉字制表语言。

（一）标识符和变量的构造

标识符包括诸如数据库名、文件名和属性域名等。

变量名由长度不超过50个字符或25个汉字所组成。变量名中可包含前缀或后缀，中间用特殊符号“—”来连接。

（二）定义语句

```
DEFINE（参数|变量）标识符
  TYPE(INTEGER|CHAR n|CHCHAR n|REAL)
END
```

其中，

INTEGER——表示整型量；

CHAR n——表示长度为 n 的字符串；

CHCHAR n——表示长度为 n 的汉字串；

REAL——表示实型量。

定义语句是程序中的可选语句，它可以用来定义变量或参数。例如

```
DEFINE
  variable counter TYPE INTEGER
  variable selection value TYPE CHCHAR 10
END
```

（三）输入语句

输入语句仅用于已被定义的变量在运行时置初值，典型的输入语句如下：

```
INPUT
  prompt for selection value using
  “请选择项目”
END
```

当程序运行到该语句时，印出提醒的字符串，在终端上用户置 selection value 的初值后，系统得到响应，运行进行下去。

INPUT 语句是可选语句。

（四）输出语句

制表语句隐式规定了页输出顶部、底部、左边等该留的空白，当选用输出语句时，可以改变隐式规定。例如

```
OUTPUT
  page length 24          {典型的字符显示器尺寸}
  right margin 80
  leff margin  0
  top margin  2
  bottom margin  1
  report to “screenfile”
END
```

这一输出语句写一个报告送到一个文件（可以把设备看作为一个文件），通过显示终端输出报告，一次处理一个文件的24个字符行（或12个汉字）。

如果一个输出语句要直接发送一行去印出，则可使用下面的语句：

```
OUTPUT
  report to printer
END
```

输出语句也是可选语句。

（五）读语句

```
READ〔INTO x〕域名表
    {whereclause}
    {joinclause}
ENDSYMBOL
```

其中，两个大括号的次序是任意的，可以不选或选多个。

域名表是一个或多个域名或文件名，用“，”、空格式“AND”来分隔。

Joinclause 是域名＝域名的联结。

Whereclause 是由布尔表达式组成的条件子句，这里的布尔表达式是由关系运算符（REOP）

〈　〉　≤　≥　〈〉　(≒)　＝

和逻辑运算符 AND、OR 组成。

ENDSYMBOL 是由 END 或分号“；”组成。

x 是一个中间文件。若需要，可建立此中间文件，也可不建立。此时删去〔INTO x〕。

（六）分类语句

SORT 按域名表 END

域名表即域名〔上升/下降〕；

域名是域名或域名〔n，m〕，这里$m>n>0$，域名〔n，m〕是按域名中第n到第m个字符组成的子串，根据上升或下降进行分类排序的。域名表之间的分隔用空格、逗号、AND 或重新列一行。

（七）格式语句

格式语句由下面七个子句组合而成：

```
first page header
page header
page trailer
on every record
before group of
after group of
on last record
```

现逐一进行介绍

1. FIRST PAGE HEADER 子句　它是为打印整个表格的第一页而设置的子句。使用此子句可使表格的第一页的首部比以后页的首部打印更多的信息。

2. PAGE HEADER子句　它是在使用FIRST PAGE HEADER子句的情况下使用的。PAGE HEADER 是为打印后面每一页首部而设置的。如果有 PAGE HEADER 子句而未用 FIRST PAGE HEADER 子句，则所有的页有相同的首部。

3. ON EVERY RECORD 子句　该子句用来处理表格中的“细目”行。

4. BEFORE GROUP OF 和 AFTER GROUP OF 子句　分类语句可以包含八个等级进行分类排序，每一个等级按记录组划分。BEFORE GROUP OF 和 AFTER GROUP OF 分别表示在执行处理一组记录以前或以后的控制子句。分类语句可以对域进行嵌套，而 BEFORE 和 AFTER GROUP OF 子句也可以进行嵌套。例如

```
    ⋮
SORT BY STATE, CITY END
    ⋮
FORMAT
  BEFORE GROUP OF STATE
    ⋮
  AFTER GROUP OF STATE
    ⋮
  BEFORE GROUP OF CITY
    ⋮
  AFTER GROUP OF CITY
    ⋮
END
```

5. ON LAST RECORD 子句　ON LAST RECORD 子句和 AFIER GROUP OF 子句一样，但它的控制级别高于任一其他子句，这一子句对于打印总量值非常有用。

6. PAGE TRAILER 子句　这一子句的打印条件是在每一页的底部，底空白的上面。

（八）打印语句

1. 简单的打印语句

PRINT{字符（包括汉字）串|域}

此语句表示用来打印括号中的字符（包括汉字）串和域中所指定的内容。例如，若 item 是域名，占 6 个汉字位置，“教科书”是其域中的内容，则执行

PRINT item

的结果是打印：

教科书

实际上是打印了 6 个汉字，在“教科书”的右边还有 3 个汉字位置是空白。

2. 打印数字域和表达式　打印数字域要根据格式定义，格式定义子句中的格式符有 *、&、#、，、。、－、＋、（、）、￥等，这里

*　填空格符，尤其对金额前空余位数填 *

&　填数字 0

#　填空格

，　数字分隔符

．　小数点

－　负（减）号

＋　正（加）号

（　左括号

) 右括号

¥ 货币标记

例如：

格式串	数值	格式输出结果
"##, ###.##"	1234.56	61,234.56
"##, ###.##"	0.01	bbbbb.b1
"&&, &&&.&&"	1234.56	01,234.56
"&&, &&&.&&"	0.01	00000.01
"¥***, ***, &&"	1234.56	¥**1,234.56
"(¥¥¥, ¥¥¥.&&)"	123.45	¥123.45

表达式中可出现 COUNT(计数)，AVERAGE(平均值)，PERCENT (百分比)，MAX(最大值)，MIN(最小值)，TOTAL(总计)等几个函数及其他算术运算，操作对象为域、常数或参数等。

3. LET 语句

LET 标识符＝表达式

赋值语句是一个简单的语句，通常左边是算术变量，右边是任一算术表达式。

4. IF-THEN-ELSE 语句　这个语句在很多语言中使用。这里只举一例略作说明：

```
IF (售价>售价的平均值)
THEN
   BEGIN
     IF (售价>1,000,000.00)
     THEN
       BEGIN
         PRINT"这是永远超过平均售价"
       END
         ELSE
           BEGIN
         PRINT"超过平均售价"
           END
   END
 ELSE
   BEGIN
     PRINT"低于平均售价"
   END
```

5. SKIP子句　此子句主要有两种形式：

SKIP　n　LINES

表示跳过 n 行。

SKIP　TO　TOPE　OF　NEXT　PAGE

表示跳到下一页的首部。

以上是制表语言或称报表生成程序中的主要语句，不是语句的全部。

二、汉字制表语言的一个输出实例

下面是输出提升工资通知单的实例：

某工厂的数据库中有：

```
FILE EMPLOYEES
FIELD emp-name TYPE CHCHAR    10
```

```
FIELD emp-number TYPE INTEGER
FIELD emp-address 1 TYPE CHCHAR    10
FIELD emp-address 2 TYPE CHCHAR    10
FIELD emp-address 3 TYPE CHCHAR    10
FIELD emp-phone TYPE CHAR    6
FIELD emp-birth TYPE CHAR    6 {yymmdd}
FIELD emp-title TYPE CHCHAR    10
FIELD emp-dept-num TYPE INTEGER
FIELD emp-hire-date TYPE CHAR    6 {yymmdd}
FIELD emp-term-date TYPE CHAR    6 {yymmdd}
FIELD emp-last-review date TYPE CHAR    6 {yymmdd}
FIELD emp-salary-effective-date TYPE CHAR    6 {yymmdd}
FIELD emp-salary TYPE REAL
FIELD emp-status TYPE CHAR    1 {"e" = 在职}
                              {"t" = 退职}
                              {"s" = 事假}
                              {"b" = 病假}
FIELD emp-rating TYPE CHAR    8 {"poor"}
                              {"fair"}
                              {"good"}
                              {"superior"}
FILE departments
FIELD dpt-dept-num TYPE INTEGER
FIELD dpt-dept-desc TYPE CHCHAR    20
FIELD dpt-address TYPE CHCHAR    10
END
```

输出程序

```
OUTPUT
  report to "salary·rpt"
  top margin    6
END
{READ from the employee file}
READ employee dpt-dept-desc
  {select only those who are currently employee.}
  where emp-status = "e"
  {Join the employee information}
  {with each employee's departmant name.}
    from join on emp-dept-num = dpt-dept-num
END
{sort the result by the employee's name.}
sort by dpt-dept-desc emp-name END
FORMAT
OR EVERY RECORD
PRINT 20 SPACES, "提升工资通知单"
SKIP 3 LINES
PRINT "姓名：", emp-name, 10 SPACES,
      "参加工作日期："
       emp-hire-date[3, 4], "/",
       emp-hire-date[5, 6], "/",
       emp-hire-date[1, 2],
SKIP 1 LINE
```

```
PRINT "上一次工调日期: "
        emp-last-review-date[3, 4], "/",
        emp-last-review-date[5, 6], "/",
        emp-last-review-date[1, 2], 10 spaces,
        "工号: ", emp-number using "####"
SKIP 1 LINE
PRINT "现在工资: ", emp-salary using "¥¥¥、##"
SKIP 1 LINE
PRINT "提升以后的工资: ",
IF (emp-rating = "poor")
  THEN PRINT emp-salary * 1.00 using "¥¥¥.##"
IF (emp-rating = "fair")
  THEN PRINT emp-salary * 1.05 using "¥¥¥.##"
IF (emp-rating = "good")
  THEN PRINT emp-salary * 1.10 using "¥¥¥.##"
IF (emp-rating = "superior")
  THEN PRINT emp-salary * 1.15 using "¥¥¥.##"
SKIP 1 LINE
PRINT "生效日期", emp-salary-effective-date[3, 4], "/",
                  emp-salary-effective-date[5, 6], "/",
                  emp-salary-effective-date[1, 2]
SKIP 1 LINE
PRINT "工作部门: ", dpt-dept-desc
SKIP 1 LINE
PRINT "职务: " emp-title
SKIP 4 LINES
PRINT "新的职务和职责建议"
SKIP 4 LINES
PRINT "备注: "
SKIP 4 LINES
PRINT "厂长签名_____"
SKIP 1 LINE
PRINT "劳动工资科长签名_____"
SKIP 1 LINE
PRINT "部门负责人签名_____"
SKIP TO TOP OF NEXT PAGE
END
```

本程序是对在职职工按其工作情况输出提升工资通知单。

运行该程序的结果如下:

提升工资通知单

姓名: 魏师化　　参加工作日期: 04/01/75

上一次工调日期: 06/01/81　　工号: 5443

现在工资: ¥70.00

提升以后的工资: ¥77.00

生效日期: 01/01/83

工作部门: 供销科

职务: 产品推销员

新的职务和职责建议: 产品推销员兼收用户对产品的意见

备注：
厂长签名：刘毅
劳动工资科长签名：李一明
部门负责人签名：黄建民

14.5 汉字企业管理系统的硬件配置实例

汉字企业管理系统的硬件配置不能一概而论，而应该根据以下几个因素来进行综合考虑，作权衡以后来决定：

（一）系统要实现的目标

是只做单项管理，还是打算进行全面管理，即使是全面管理，还存在一个要求管理的深度问题，是只做厂部一级的管理，还是要实现厂部、车间以及班组三级管理，或者是总厂、分厂和车间三级管理。

（二）企业的规模

如果只考虑单项管理，不考虑企业的全面管理，而且只考虑到眼前，而不考虑到今后系统的可扩展性，企业的规模是不用考虑的。要现实企业的全面管理或在现实单项、局部管理时，同时考虑到以后系统的可扩展性，则必须将企业的规模考虑在内。

作业企业的规模，也是由多种因素决定的，企业的职工人数是一个因素，除此以外还有企业所占有的地理区域，企业的产品种类、产品的复杂程度，投产的方式以及设备的情况等等因素综合决定的。

（三）今后和上级部门的计算机联网的可能性。

除了上面三点技术因素外，还要考虑企业的经济实力(即能够用于企管项目的投资)和人力配备等。

图 12-11 所示是汉字企业管理系统的硬件配置的一个实例。在一个系统中，应该配多少台处理机，用哪些类型的处理机，处理机之间采用什么联结方式，以及每台处理机应该带多少外围设备等，这些问题都要经过十分认真的考虑。

从处理机之间的联结方式来说，除了像图 14-11 中所示的树状结构外，还有环形结构、网状结构和总线结构等。

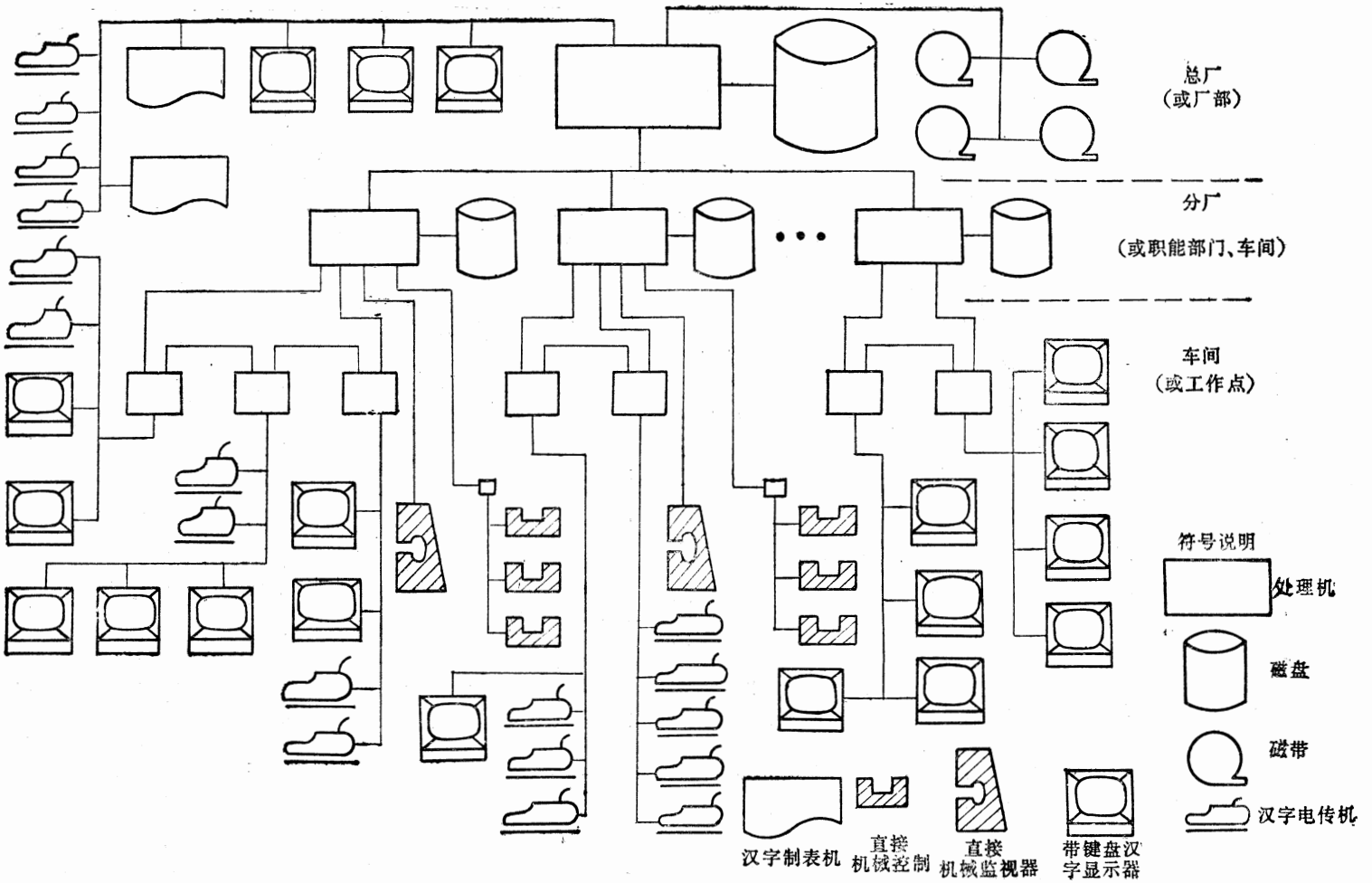


图14-11　汉字企业管理系统硬件配置实例

第十五章　中医诊疗系统

15.1　诊疗系统发展概况

随着计算机科学的不断进步和人工智能理论的飞速发展，作为人工智能领域的一个重要分支——诊疗系统也在日臻完善，并已开始在实际诊疗中使用。

我国的诊疗系统的研制工作是在七十年代后期开始的，目前已取得了很大进展，但主要集中于中医的诊断、医疗方面。已研制成功了几十个中医诊疗系统，其中有关幼波的肝病诊疗系统，路志正的眩晕病诊疗系统，林如高骨伤诊疗系统，林沛湘外感咳嗽诊疗系统，刘风谦温病部分诊疗程序，梁乃津胃脘痛诊疗系统，陈可望冠心病诊疗系统，李敬之治疗冠心病程序，陆南山眼科角膜病诊疗系统，赵思俭糖尿病诊疗系统等等。研制这些系统的共同特点是模拟行医几十年的著名中医师的逻辑推理和诊断过程。

就中医诊疗系统而言，目前大多尚处于单项的实验性阶段。基本理论的研究方面，已有了几种模型，但尚未形成严密的、完整的、体系性的理论，因此还有待于进一步发展。这里就目前的研制情况向读者作简单介绍。

15.2　建立计算机中医诊疗系统的意义

中国医药学是一个伟大的宝库，有悠久的历史和光辉的成就，它是我国人民几千年来同疾病斗争的极其丰富的经验总结。运用近代科学知识和方法来整理、研究中医中药，创造中国统一的新医学新药学，这是历史赋于我们的重任。中医诊疗系统的研究，为完成这一历史重任迈出了可喜的一步，其意义是巨大的，主要表现在下述几个方面：

（1）每一位著名的老中医对某种疾病一般都积累了几十年的诊疗经验，有独特见解，疗效显著，深受广大患者欢迎。但是，能受到老中医亲自诊治的患者毕竟是少数，中医诊疗系统能如实地再现老中医一整套辨证施治的医学思想。此外电子计算机有丰富的逻辑功能，精确而高速的处理能力，因此，中医诊疗系统可使著名中医专家的经验为更多的患者服务，从而大大提高诊疗服务效能。即使是偏僻的山区或遥远的边疆，只需要有相应的软件包和硬件设备，即可得到相当于名中医亲临现场辨证治疗。另一方面，也减轻了名中医大量重复性门诊工作，使他们有更多的时间和精力致力于研究、开发工作。

（2）计算机中医诊疗系统，在解决中医后继乏人这一急迫问题上有独特价值，目前已找到了描述中医辨证施治过程中的数学模型，它就象再现名演员的艺术集锦片，以“活”的形式再现著名中医的诊断过程。当我们把这种方法用于我国著名中医诸家之后，即可使他们的宝贵经验长存于世，为广大的患者服务。目前，采取的“传、帮、带”及组织力量著书立说的办法诚然重要，但是在数以万计的后继者尚未达到名师水平之前，将不可避免地出现一段漫长的空白，而中医诊疗系统却能弥补这一欠缺。

（3）中医诊疗系统的研究把祖国医学引入计算机时代，对实现祖国医学现代化是

一个有力的推动。使中华民族的古老的医学宝库得到充分地开发。

15.3 中医诊疗的理论基础

中医诊疗的基础是整体观念和辨证施治。

15.3.1 整体观念

中医的整体观念含有下面两个方面的意思：

（1）中医认为人体是一个复杂的矛盾的统一体，人体的内部脏腑之间，脏腑和体表的感觉器官之间是相互联系的，一旦某一部分有病变，就会影响到其他部分，以至影响到整个身体。因此，治病决不能头痛医头，脚痛医脚，应该从整体出发去观察局部的病变现象。否则不可能取得较好的疗效。

（2）中医认为人的生存决不是孤立的，它和外界环境有密切的联系，如饮食中可能有不洁之物乘机而入。体表和其他东西接触，可能受到感染。环境对人体有很大影响，当人们不适应环境的变化（如季节变化、淋雨，受寒或烈日爆晒等）时，就可能出现疾病。因此，医治疾病必须注意周围环境对人体的影响。

15.3.2 辨证施治

病人患病都是通过一定的症状得到反应，例如头晕、目赤、耳鸣、咽干，胸闷、胁胀和脉弦等等。中医通过“四诊”收集症状。所谓四诊是望诊、闻诊、问诊、切诊。下面分别予以介绍。

（一）望诊

医生用肉眼直接观察患者全身和局部情况，藉以了解病人的精神、动态、体质和营养情况等。局部情况主要观察病人的肤色、舌（包括舌体和舌苔）、眼、鼻、唇、牙齿等，通过它们了解病人体内脏腑的病变，判断疾病的性质和发展等。

（二）闻诊

医生用耳听鼻嗅等感觉来检查病人声音（指语言、呼吸、咳嗽、呃逆）和嗅气（指口臭、痰、大小便、白带）等，藉以了解病人的病情。

（三）问诊

医生询问病人的主要症状、伴随症状，以及它们之间的相互关系，此外还要了解症状的演变过程。同时还要询问与上述四个方面有关的问题，如寒热、出汗、二便、饮食、头身疼痛和胸腹等情况，因为他们是分辨寒热、表里、虚实和阴阳等的重要根据。

（四）切诊

中医的切脉是区别于西医的一大特点，切诊主要是切脉。中医通常辨认的是寸口脉，它位于手腕部桡动脉搏动处，分为分寸、关和尺三个部位。如图 15-1 所示。中医认为，左右两手的寸、关、尺六部脉分别和一定的脏腑相对应，见表15-1。切脉的脉象种类繁多，一般可归为浮、沉、迟、数、虚和实六类作为脉纲。

辨证就是藉助四诊对病人的病理和临床现象作正确的判断，主要应分辨出以下内容：

病变的部位；

人体对致病因子的反应能力；

致病因子和疾病的性质；

疾病已进入的深度。

然后，用八纲（阴阳、表里、寒热、虚实）加以综合辨证。

施治是在辨证的基础上进行的，这就是通常说的要对症下药。如何下药？中医治疗的基本原则是“扶正”与“祛邪”。“正”是指正气，即指人体所具有的抵抗力，自身复原力，对内、外环境的适应力；“扶正”是采用各种调理（如滋阴、补阳、益气、补血等）配合饮食去增加人体对疾病的防御能力。“邪”是指病邪——致病因子，“祛邪”是采取各种治疗方法来消除致病因子，限制和制止病情的发展。在具体治疗时，到底采取“扶正祛邪”（先扶正后祛邪），还是“祛邪扶正”（先祛邪后扶正），或是扶正、祛邪同时进行，这必须根据病人的具体情况灵活掌握。进一步，如何祛邪，如何扶正，这要根据疾病的缓急程度以及疾病所在部位等具体情况，分别遵循若干准则。例如“急则治其标，缓则治其本”，“调理脾肝肾，中焦要当先”等等。

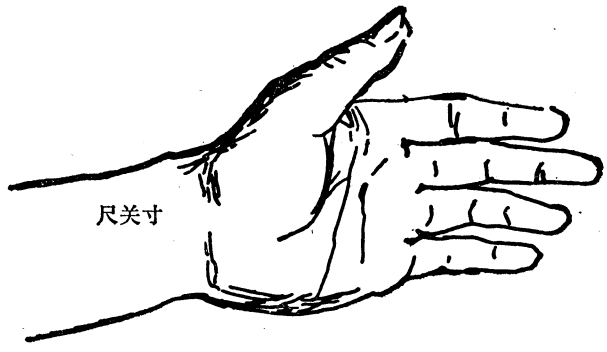


图15-1　寸、关、尺的部位

表15-1　切脉部位和对应的内脏

部位		脏腑
左手	寸	心、小肠
	关	肝、胆
	尺	肾、膀胱
右手	寸	肺、大肠
	关	脾、胃
	尺	肾、命门

15.4　中医诊疗的数学模型 I

中医诊疗系统的第一种设计思想是根据模糊数学模型，利用电子计算机来模拟疾病诊断时最直观的方法，对各种症型确定一个阈值 O_j $(O<O_j\leqslant 1)$，并计算某一病人属于第 j 型病的隶属（程度）函数 $\mathscr{C}_{A_j}(x_i)$，若

$$\mathscr{C}_{A_j}(x_i)\geqslant O_j$$

则说该病人患了第 j 型病。

下面我们阐述中医诊疗的模糊数学模型。

15.4.1　症候群空间

在 n 维欧氏空间中，取 n 个正交基作为该空间的基底。由这个基底建立坐标系。四诊及某些化验结果所获得的症状对应一个坐标轴。某种症状出现时，对应该轴上值是 1；

该症状不出现时，对应轴上的值是0。这样，每一个“症候群”就对应 n 维空间中的一个点。这 n 维空间就称为症候群空间。

显然，对于 n 个症状的情况，对应的症候群空间由 2^n 个点组成的集合。以X来表示这些点所组成的集合，就有

$$X=\{x_i \mid i=1,2,\cdots,2^n\}$$

当把不同类型的“病”看成是 x 上不同的模糊子集时，医学诊断的问题就归结为确定X上的某一元素，x_i（$i\in\{1,2,\cdots,2^n\}$）以多大的程度属于哪一个模糊子集的问题。

假设在X上找出 m 个模糊子集，$A_1, A_2, \cdots, A_m$，显然，

$$A_j \subseteq X^{\bullet},\quad j\in\{1,2,\cdots,m\}$$

现在，我们对每一模糊子集可定义一个隶属函数$\mathscr{K}_{A_j}(x_i)$，$x_i\in X$，即表示X的元素 x_i 隶属于 A_j 程度的大小。$\mathscr{K}_{A_j}(x_i)$ 的值越接近于1，x_i 隶属于 A_j 的程度越高。我们可以知道，A_j 型具有的最标准症候群所对应的点 x_0^j，就是模糊子集 A_j 的核，即

$$\mathscr{K}_{A_j}(x_0^j)=\underset{x_i\in X}{\mathrm{SUP}}\,\mathscr{K}_{A_j}(x_i)\qquad j\in\{1,2,\cdots,m\}$$

（$\underset{x_i\in X}{\mathrm{SUP}}\mathscr{K}_{A_j}(x_i)$表示对每一个 $x_i\in X$，属于 j 型病的隶属函数$\mathscr{K}_{A_j}(x_i)$中取上确界）。

15.4.2 隶属函数$\mathscr{K}_{A_j}(x_i)$的计算

一、确定权系数

我们研究的病种中均有不同的型，每一型又有一系列症状。也就是说，每一型有它相应的标准症候群，其中有主症、次症、兼症、舌苔脉象等，每一症状对诊断为该病的重要性各不相同，症状与症状之间又有内在联系。主症的值较高，次症的值较低，这种所起的作用程度大小的量称为权系数。

实践中还发现：有些症状在有另一症状出现时，显得很重要，而一旦另一症状不出现时，就显得不重要了。权系数的取得没有一般的方法，只能根据老中医多年的诊疗经验和通过分析成百上千的病案，经过综合评定来得到。每一种症状在各型病中的权系数（即经验条件概率），要在试验中反复调整，直至被合理地确认。

二、隶属函数$\mathscr{K}_{A_j}(x_i)$的计算

先求和：$p^0_{A_j}=\sum_{i=1}^{P}\alpha_i a_i$

式中，$a_1, a_2, \cdots, a_p$ 是 x_0^j 含有的所有症状，即第 j 型病所有可能出现的标准症候群，在上式中取1，而 $\alpha_1, \alpha_2, \cdots, \alpha_p$ 是 x_0^j 的各症状在第 j 型病中相应的权系数。

对前来就诊的某一具体病人而言，设其对应的所有症状为 $a'_1, a'_2, \cdots, a'_q$（$q\leqslant p$）。当其中的症状在第 j 型病的标准症候群中有相同含义的症状时，取值为1，否则取值0；$\alpha'_1, \alpha'_2, \cdots, \alpha'_q$ 是症状 a_i（$i\in\{1, 2, \cdots, q\}$）在第 j 型病中相应的权系数。作和式

$$p_{A_j}=\sum_{i=1}^{q}\alpha'_i a'_i$$

● $A\subseteq X$ 表示集合 A 是集合 X 的子集。

最后，计算隶属函数如下：

$$\mathscr{K}_{A_j}(x_i)=p_{A_j}/p^0_{A_j}$$

15.4.3 阈值的确定

医学诊断问题，实质上是一个模式识别问题。我们对那些需要通过计算隶属函数来加以辨证分型的，就使用确定阈值的办法来完成这种模式识别。所采用的直观做法是对各种病型确定一个值 $\theta_j(0\leqslant\theta_j\leqslant1,\ j\in\{1,\ 2,\ \cdots,\ m\})$。

当

$$\mathscr{K}_{A_j}(x_i)>\theta_j$$

时，就可将病人 x_i 诊断为患有第 j 型病。

与确定权系数一样，对于确定 θ_j 值也没有一般的方法，只能按老中医的经验和实验来确定，力求最大限度地反映老中医诊断的观点，达到比较理想的模式识别的目的。

15.4.4 浮动阈值技术

在临床病例的调试中发现，当阈值 θ_j 选择较大时，会出现对某一病人采集的 x_i 不属于任何一个病型 A_j，反之，当 θ_j 取值较小时，会出现某一病人 x_i，其隶属函数 $\mathscr{K}_{A_j}(x_i)\geqslant\theta_j$ 对几种不同的病型 A_j 都成立。这样会使我们得不到正确的诊断，为了达到只归入一个病型 A_j，以便能抓住病人的主要矛盾，给出正确诊断和有效施治，就可以使用浮动阈值技术。

所谓浮动阈值技术就是让 θ_j 从选择的某一个上限 $\bar{\theta}_j$ 到某一个下限 $\underline{\theta}_j$ 之间浮动。对不同的 j，阈值的上下限可以不同，浮动的速度也可以不同，在阈值从上限向下限方向的浮动过程中，可以根据实际情况的需要，重新安排 A_j 的顺序，依次检查哪一个 j 能使 $\mathscr{K}_{A_j}(x_i)>\theta_j$ 先得到满足。

使用了浮动阈值技术，可以使判断更准确，更易于处理各种病人千变万化的情况。

15.5 中医诊疗的数学模型 I 的程序流程及输出

对应于模型 I 的程序流程如图15-2所示。

程序由以下六个部分组成：

（一）信息输入

用人机对话形式输入日期（每天开始门诊时）、姓名、性别、年龄、病历号等。接着是症状输入，将望、闻、问、切这四诊中所能得到的一切症状以及其他化验结果输入计算机。

（二）机器自检输入信息

如果输入的症状、体证、检查结果的信息不够或发现一些症状内容重复、矛盾以及有其他疑问，计算机就在这部分程序中自动进行自检，并询问一些有关症状的情况，待再次对病人复查并输入症状后，计算机才进行辨证分型处理。

（三）辨证分型

根据已输入的某个病人的各种症状来模拟老中医的诊断方法并进行辨证分型。确定病人是患哪一病型的病或基本康复状况等，为对症下药做准备。

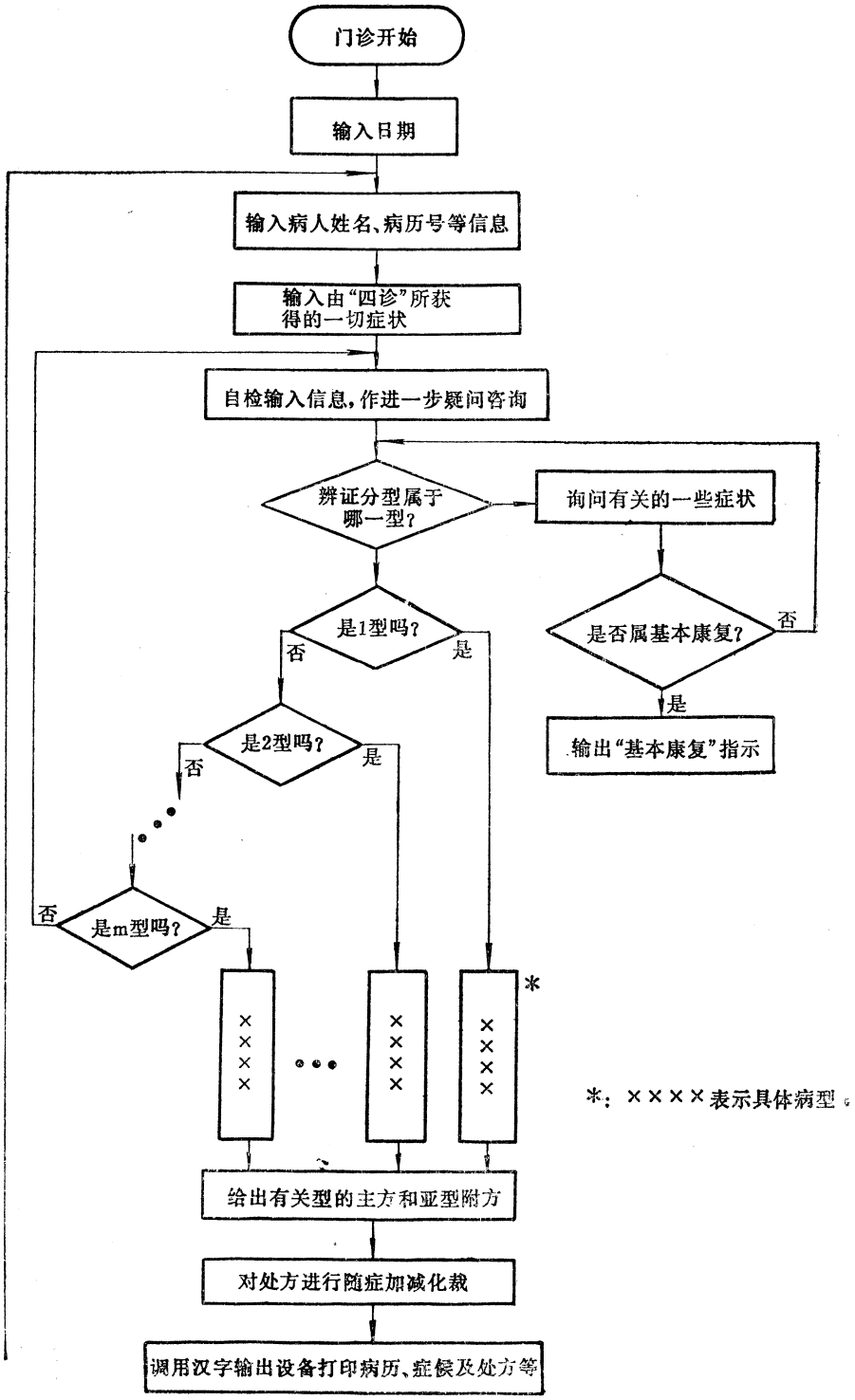


图15-2　中医诊疗模型Ⅰ的程序流程

本部分程序中还应设计这样的询问功能：当计算机有疑问，还不能准确地确定病人是属于哪一种病型时，计算机能自动地询问一些有关的未输入的主要症状，要求作再次复查之后，进行辨证分型。在这里同时设置了浮动阈值技术。

（四）给出施治主方

根据阈值分型，得到正确辨证后可相应地给出施治的主方。

（五）随症加减化裁

病人的症候群和标准症候群之间若有差异，则说明对主方有修正的必要。此外，必须考虑到病人的性别、年龄、季节以至患病的部位，也要对主方进行修正，其原则做法是根据上述情况对主方中的药味作适当调整或在用药的分量上作适当加减，使处方更符合病人的具体情况。

（六）输出处理

采用针式汉字打印机输出下述一些内容：

（1）病历档案　其内容有病历号、姓名、年龄、性别、工作单位、症状、诊断、处方、药费计价、医嘱、假条、复诊日期等。

（2）处方笺　其内容有病历号、姓名、年龄、性别、处方、药费计价、服药贴数、复诊日期等。

（3）疾病证明书　其内容中除去症状、处方、药费计价外、保留项（1）中其他项目。

实用程序中还有一些附属项目，如修改药价，缺药代换等等，这里不一一赘述。

15.6　中医诊疗的数学模型Ⅱ

中医诊疗系统的第二种设计思想是使用数据库、知识表示、生成规则、推理网络、控制策略和自然语言理解等近年来迅速发展起来的人工智能的理论和技术。系统不仅在诊断的结论和用药的处方上和名中医吻合，而且致力于系统在整个推理过程中也要和名中医的思路相一致。为此，采用多层推理网络的数学模型是合适的。

多层推理网络是藉助于生成规则，用不同种类的链来沟通和激发。各个层次是根据中医辨证施治的思维过程和由表及里的深化过程来划分的。各个层次之间的链相当于思维过程中的联想、判断和推理。多层推理网络是从初始症状集开始的。初始症状集包括医生的“四诊”、患者主诉及有关信息，初诊时系统首先借助加工链的加工机制，把初始症状加工成综合症状。通过提升链的功能，把第一、二两层的相对信息提升为子型，接着运用分类链，把取自第二、三层里的诸信息分类成合适的症属。最后依靠合成链和加减链得到诊断结果，治疗方案与方剂等。除因果链之外，所有的链都是为了把层与层沟通起来，使系统能从第一层信息出发依次使用这些链，顺次得到第二、三、四层及第五层信息，以至最终完成初诊。因果链和上述诸链不同，它仅局限于子型、脏腑机层，描述了该层中诸子型及初始症状之间的因果关系。

图 15-3 所示的是具体病例的多层推理网络。病例是一位 29 岁的初诊女患者，主诉脸部、前胸、四肢屈侧等部位有红斑、红色疱疹、稍有渗水、皮肤灼热、瘙痒、心烦口渴、大便秘塞、小便赤黄、舌质红、苔黄、心绪不宁。皮损具体分布、面积、程度从略。

系统对这个具体患者的初诊推理过程为：

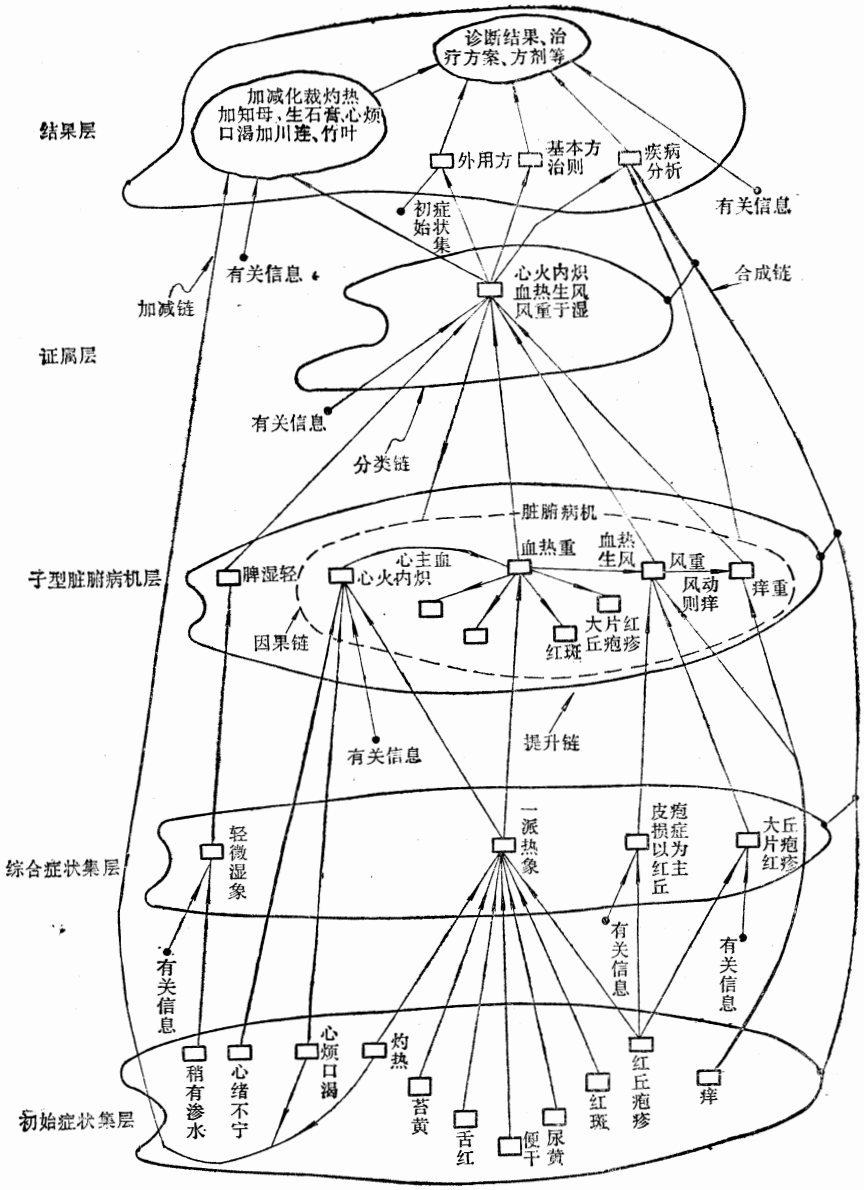


图15-3　初诊多层推理网络实例

第一步是加工操作。从稍有渗水、苔不腻，渗水来源于红丘疱疹（不是水疱），从皮损类型、分布等信息加工出综合症状——轻微湿象。从红丘疱疹的分布面积、密集程度等信息加工出大片红丘疱疹。还加工出以红丘疱疹为主的皮损及一派热象等综合症状。

第二步是提升操作。提升主要依据脏腑病变，风湿热燥四邪和初始症状、综合症状及相关信息间的对应关系进行操作。具体地说，例如把心绪不宁、心烦口渴、一派热象及有关信息提升为心火内炽这一子型。由提升操作得出的子型还有血热重、风重、痒重等。这里脏腑病机理的因果推理从略。

第三步是分类操作。分类操作主要依照中医鉴别诊断学，运用直译、挑选等比较方法确定疾病的证属。本例中系统首先通过比较法确定诸子型的主次，定出心火热象为主，风重次之，湿象居末。从而排除了湿热俱之证属，又从疾病发生、发展的过程及无过敏源，排除了毒热人营之证属。

最后对合成操作作简单说明。这一部分用于对诊断作深入解释和证明。它综合了第一、二、三及第四层的一些信息得出对本疾病的分析如下：根据《内经》“诸痛痒疮，皆属于心”。因为心主火，心主血脉，所以心绪不宁、心火内炽，逐生血热。就会身起红丘疱疹，并且灼热、舌红、苔黄、大便干秘、小便赤黄等呈现一派血热现象，复因血热生风，风动则痒。从而用生地、丹皮、赤芍凉血清热；知母、生石膏清肌热；川连、竹叶以清心火；佐以豨签草、海桐皮、苦参、苍耳子、六一散，祛风除湿止痒。

推理网络是系统的核心。它设计的好坏，直接关系到系统诊断的正确性和对老中医思想的逼近程度。

上面所述的多层推理网络中涉及的诊断仅仅是针对一次独立的诊断或者是初次诊断而言。而实际临床处理中的复诊比初诊要复杂得多。它不仅囊括了初诊推理网络中的所有内容，而且要求能把当前和前一次的病情比较，以确定病情变化。它把以往诸次的诊断、疗效等信息进行分析和总结，并结合当前的病情作出诊断。为实现上述内容，在系统复诊的模块中应具备病历存储、检索，病情比较，诊断分析和综合辨证等功能。

复诊时，系统首先借助病情比较操作判断出病情是趋于好转、无变化还是恶化。接着再按具体情况，分别进行处理。

当和上次诊断比较后发现病情好转时，系统运用自身掌握的中医理论及临床经验，对上次诊断时确定的证属、治则、方剂等，从确定为该证属后就医治的累计次数、治疗效果，病情好转程度以及患者的各种反映等信息进行分析、综合，再决定方剂的更换与加减化裁。

经上次诊疗后至今无明显的变化时，系统要按照上次选取的证属的性质、符合性(症状到证属的对应与名中医诊断的一致程度)、连续选择该证属的次数、疾病的特点及患者情况等信息，判断出是上次诊断不当，还是疾病顽固疗程较长，需坚持治疗下去。在这种情况下，可以发现系统与名中医之间的差异，从而提请专家对系统进行修改。

如果经上次治疗后，系统发现病情恶化，则系统要判断出是某种外因引起了病情反复，还是疾病本身向高峰发展，而并非诊断错误，或者是上次诊断不当。根根三种不同情况给予妥善处理。

由于复诊功能的存在，使系统能够发现并记录那些诊断治疗不当和治疗无效的病例。然后请专家来分析和确定哪些病例反映了系统的不完善或错误，哪些病例是需要医学上进一步发展才能解决的问题。系统的这种自动发现、记录、存储本身同名中医诊断不符合之处及其他问题的本领，说明它具有一定程度的学习功能。

15.7 多层推理网络中的生成规则和元规则

系统在推理过程中大量使用以“IF…Then…”形式生成规则。生成规则在推理网络中所处的位置、使用的成分和执行的先后大致可分为三种：

(1) 初始症状生成综合症状规则。它对应系统网络中的加工链。

（2）初始、综合症状、中间结果生成中间结果规则。它对应提升链、因果链。

（3）所有信息生成结论规则。它对应系统网络中的分类链、合成链和加减链。

所有的生成规则都由控制策略来控制执行。

元规则是为了建立、维护和扩展多层推理网络而对输入的自然语言进行分析和处理的规则。系统接受自然语言的输入后，先进行如下检查：

1．检查输入语言的用词是否是系统设计时由专家确定的限用词，若是，便进一步分析。

（1）检查输入语句是否符合系统原先规定的各项句法规则，然后根据系统设计的安排，执行那些符合规则的操作（如删除等）；

（2）如果输入语句符合系统原先规定的各项规则，则根据语言中的用词，将其转换成系统的内部表示，于是在系统的知识库中就可以搜索出有关的中医诊断知识或术语的内部表示，从而进一步分辨语句中的因果关系，将其转换成“IF…THEN”语句来表示。例如有A、B症状，同时无C症状，或者有A、D、C症状，则属X型病；否则有A、B、D症状，属X_1型病。这里假定现有库中对A、B、C、D四种症状，可能出现的病型是X或X_1两种。将上述自然语言用“IF…THEN”语句来表达即为

```
IF NOT C AND A AND B OR A AND D AND C
THEN X
ELSE IF A AND B AND D
    THEN X1
    ELSE “系统无法诊断”
```

对于系统输出信息“系统无法诊断”，是针对某一病人说出不符合上述症状的组合形式之一。那末，这时有两种可能性：一种可能性是确诊为X_2型病，但遗漏送入系统；另一种可能性是专家从未碰到过。不管是哪一种可能性，在已给的库中还没有确切的断言，此时系统只能输出信息：“系统无法诊断”。要进一步完善系统的话，就要在知识库中添加新的知识。

若是添加，则要找到添加语句的适当位置，但这是比较困难的。不过可以根据语义，先确定添加语句所在的层或区段，然后系统分析要添加的是什么语句，加在什么位置，对原有的安排该作如何调整。

若是修改，则根据语义，找到将被修改的语句，然后在这个语句的位置上输入新的规则，当然对原有的安排也要作适当的调整。

在上述动作完成之后，系统还要检查这样的处理与已有的其他规则是否有矛盾。具体地说，就是要寻找和新的生成规则中“IF”部分相同，而“THEN”部分不同的规则。若有，再作进一步处理；

2．输入语言的用词中具有不是专家确定的限用词，即有系统不能理解或未加说明的术语和知识。对此只能求助于专家，增加新的术语和知识。

元规则的引用，使医学专家不必具备计算机程序设计和语言方面的知识，而只需了解对输入自然语言的某些限制，就可以对系统进行修改和扩充，而通常不必求助于软件人员来作为同系统对话的桥梁。这就使系统易于完善、推广，从而可提高实用性，给医学专家带来很多方便。

15.8 中医诊疗系统的硬件配置和输出实例

目前最简易的中医诊疗系统的硬件配置可用一台带软盘和显示器的微处理机来完成，人机对话通过显示器的字母数字键盘和显示器屏幕，如果配用一台汉字印刷机，则诊断结果和处方等均可方便地由印刷机打印出来。

一个比较实用的最小中医诊疗系统的硬件配置如图15-4所示。

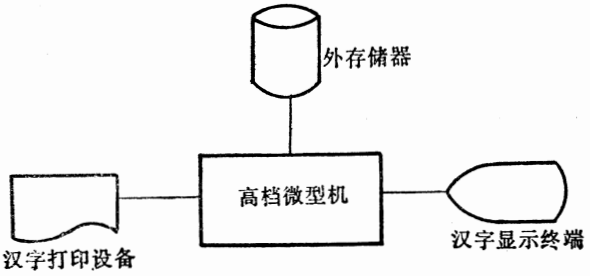


图15-4 实用的最小中医诊疗系统的硬件配置

下面我们举一个中医诊疗输出的实例：

病员杨××，男，24岁，病历号为887429。

9个月来左足疼痛，行走后明显加重，并出现跛行，当跛行疼痛时需休息一刻钟左右才能缓解，两个月后左趾青紫、溃破、流脓水，经治疗现溃破表浅，结痂未固，疼痛较少。

检查结果表明：

双足趾甲变形，左足跗阳脉（足背动脉）（一），太溪脉（胫后动脉）（一）；右足跗阳脉（一），太溪脉（±）。舌质红苔白，脉细数。

把上述症状和体症送入计算机后，立刻可以生成病历档案，同时在汉字印刷机上输出：

NO887429　杨××　　19××年×月×日

诊断

血栓闭塞性脉管炎Ⅱ期脱疽毒热伤阴型

处方

双花	30g	公英	30g	连翘	20g	元参	30g	石斛	30g
苏木	30g	归尾	15g	赤芍	15g	红花	15g	猪苓	10g
丹参	30g	牛夕	10g	元胡	10g	川楝子	30g	生芪	30g

7剂

外用药

止痛生肌散

医师　（签名）

15.9 结束语

国际上诊疗系统的研究开始于七十年代初，目前已取得了很大进展，比较著名的西医诊疗系统有：已广泛应用于临床的诊断肺病的PUFF系统，诊断100多种传染性疾病的MYCIN系统，这两个系统都是斯坦福大学研制的。麻省理工学院研制出关于心脏病用药的DIGITAL IS ADVISOR系统和诊断胃病的PIP系统。拉特哥斯大学研制出用因

果联想语义网络方法的诊断青光眼的CASNET系统和所有生成规则的诊断甲状腺和风湿性关节炎的EXPERT系统。日本KIKUCHI$_1$等人研制的自适应的血糖调节系统等等。除此以外，还有肠胃病、先天性残废、胰腺癌等等的诊断咨询系统。

我国的中医诊疗系统的研究工作虽然比国际上的西医诊疗系统迟几年，但同样取得了很大的进展。为了向更高级、更完善的方向发展，还有大量的诸如下述一些工作要做：

（一）设计综合性的中医诊疗系统

目前的中医诊疗系统绝大多数都是立足于总结某一位名中医的辨证施治。因此对某一位就诊患者来说，实质上已有了一个倾向性的辨证估计。患有肝炎的病人要到关幼波的肝病诊疗系统上就诊，跌伤的病人要到林如高骨伤诊疗系统上就诊。如果实质上患有肝病的病人在心脏病诊疗系统上就诊，在诊疗上就可能出现偏差。因此，应该设计出综合性中医诊疗系统。通过对病人的“四诊”及其他化验体检，经过综合分析，然后进行辨证施治。

（二）开展症状自动采集的研究

目前的中医诊疗系统确实能重现名中医的辨证施治，但对症状的采集还需靠人工获得，症状采集是否准确，对于辨证施治关系极大。就以切脉而言，就可能有细、弱、弦伏、濡、缓、滑、沉、迟、滥、数、微、软、虚、浮、结、代、有力、无力、尺弱、两尺弱、两寸弱、左脉兼数、右脉兼数等二十多种之多，而相互之间的差别则十分微小。对此，若没有丰富的经验，是很难分辨正确的。要通过问诊来得到正确的症状就更困难，这里，除了医生本身的经验外，还包含患者的自我感觉和心理因素。有些属于患者分辨不清，有些因病人原来体质的差异在症状的程度上描述不精确，有些患者因思想顾虑不愿意如实向医生作实事求是的回答。这些都对正确地采集症状带来了困难。总之，研究各种症状的采集装置是非常重要的。

（三）把历代名医的病例、处方纳入到中医诊疗系统中来

充分地开发祖国的医学宝库，古为今用，具有重大的意义，这样做不但当代名中医的医术，而且使历代名中医的医术都能造福于人类。这方面的收集、开发工作已有人在做，但与要做的工作相比，差距甚大。

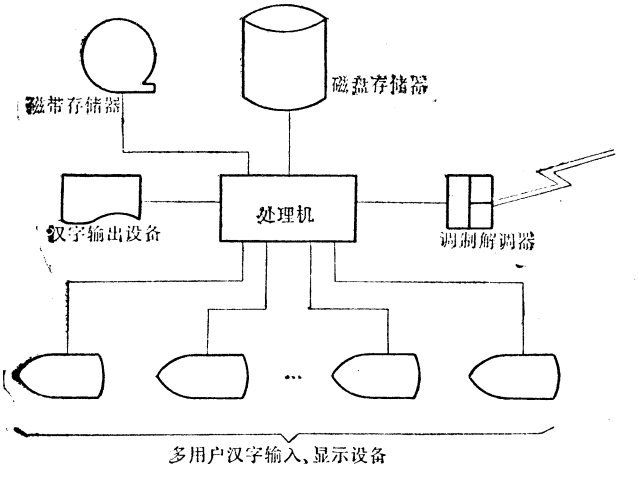


图15-5　中医诊疗系统中一个节点的硬件配置实例

（四）建立全国性或地区性的中医联机网

图15-5是联机网中一个节点的硬件系统配置实例。建立这样的联机网络，最大的优点是可以实现资源共享，使每一个节点都具有很强的诊疗功能。

第十六章　其他应用

随着社会的进步和科学技术的发展，计算机的应用范围越来越广阔。可以毫不夸张地说，计算机正在渗透到人类社会的一切领域。

计算机用于科学计算，解决了许多无法用人工实现的计算课题，从而大大促进了科研生产和科学技术的发展。

计算机用于辅助设计和制造，大大缩短了产品设计和制造的周期，提高了设计效果和产品质量。

计算机用于辅助管理，使管理人员可以摆脱繁琐的事务，以便从事更为高级的决策活动。

计算机用于信息检索，使人类更有效地管理科技情报和图书资料，从而可以更好地为科研和生产服务。

人工智能和自动控制技术的发展，出现了更高级的机器人，从而可以替代人类在繁重或恶劣环境下的劳动。

计算机已成为探索宇宙秘密，打开新知识大门的钥匙。

总之，电子计算机，特别是微型计算机是当前新的技术革命的先导。它的广泛应用，将大大加速我国四化建设的进程。

汉字信息处理技术的发展和日趋成熟，在更为广泛的范围内为我国推广计算机应用创造了条件。计算机的应用实例举不胜举。除了前面已介绍的若干应用外，这里再简要介绍几种应用系统，作为本书的结尾。

16.1　旅馆服务系统

近年来，国家投资或与外商合资的现代化大型旅馆、大型饭店正在全国各地建造。为了提高房间和床位的利用率，减少旅客的等待时间，旅馆的总服务台必须及时地掌握各房间和床位的使用情况，实现对旅馆业务的全面管理。此外，为了方便对旅客有关旅游、交通等问题的查询，也需要靠计算机系统作出迅速和正确的回答。计算机用于旅馆服务已成为不可缺少的工具。

16.1.1　旅馆服务业务简介

一、旅客登记表

一位旅客要住进一个旅馆都需填写一张登记表，登记表中主要包括以下几项内容：

姓名，性别，出生年月或年龄，国籍，护照编号，工作单位，职务，来自何地，计划逗留天数，旅行下一站地点，起住时间，实际进馆时间，房间（床位）号码，费用结算，登记者编号等十五项，其中前十一项应由旅客自己填写。

一个旅馆的房间号码（包括床位号），一般都要进行统一编码，其中包括楼号、层

号、房间号、床位号等。

就目前国内的宾馆、旅馆和大饭店的房间类型有：

单人房，单人套房，单人（夫妻）房，双人房，双人套房，三人房，四人房 等 等，当然对不同的旅馆、不同的房间其收费是不相同的。

二、房间的状态

可以使用的房间（床位）大致可以分为三种状态：

（1）已出租给旅客：即旅馆正在使用的房间；

（2）待整理：即旅客已经退房，但尚未打扫整理，目前暂时还不能出租；

（3）备用：即房间已打扫整洁，可以租给旅客。

三、旅馆服务系统

一个完整的旅馆服务系统，就其服务对象来说有两类：一类是面向旅馆内部的，例如旅馆发展规则、旅馆内部的财务管理、设备管理和人事档案管理等等；另一类是面向旅客的，它主要包括旅客服务管理、自动控制系统（当然电梯控制系统也是一种自动控制系统，但一般它自成体系）。这里所指的自动控制系统是指取暖和空调等控制。但这类自动控制系统对汉字的要求关系不大，超出本书讨论的范围。而旅客服务管理和汉字的关系就十分密切。

16.1.2 旅馆服务系统的硬件组成

旅馆服务系统的硬件组成一般包括以下几部分：

（1）主机（根据要处理的业务量的大小可以选不同的机种，如高档微型 机、小 型机或中、大型机）；

（2）外存储器（可配硬盘机、软盘机和磁带机等）；

（3）带有汉字输入键盘的显示器；

（4）汉字打印输出设备；

（5）各层数据终端设备；

各层数据终端设备是各层服务员和总服务台进行通信联系的显示终端设备，用以汇报各个楼层、各个房间（床位）的状态。

（6）自动控制系统等。

图 16-1 是旅馆服务系统硬件配置示意图。

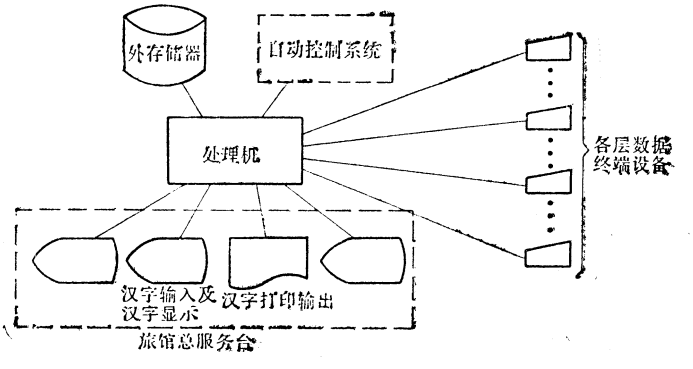


图16-1 旅馆服务系统硬件配置

16.1.3 建立在关系型数据库上的旅馆服务系统

为了叙述上的简便起见，只涉及到房间号码而不涉及床位。关于关系型数据库的知识请参阅第十一章有关内容。

一、关系的定义和撤销

作为关系型数据库上的旅馆服务系统的例子，可以设计如下四个关系模式：

（一）“房间出租”关系

这个关系中包括如下各个属性项：房间号码；房间类型（单价）；住客姓名；性别；出身年月（年龄）；国籍；护照编号；工作单位；职务；来自何地；计划逗留天数；下一站旅行地点；起住时间；实际离馆时间；房租结算；登记表编号。

（二）“房间待整理”关系

这个关系中只包括房间号码和房间类型（单价）两个属性项即可。

（三）“备用房间”关系

它包括的属性项和“房间待整理”中的属性项相同。

（四）“服务记录”关系

在这个关系中包括下列属性项：服务项目；房间号码（服务对象）；单价；数量；服务日期。

定义关系时要指出关系名、属性名及其类型说明，对不允许未定义值的属性（主关键词）用 NONULL 显示指出。否则用 VAR，以示区别。属性名的类型一般可分为：INTEGER（表示整型量）；REAL（表示实型量）；BOOLEAN（表示布尔量；只可取真、假值）；CHAR（n）(表示 n 个字符的字符串)；CHCHAR（n）(表示由 n 个汉字组成的汉字串或由 2 n 个字符组成的字符串，统称汉字字符串)。

例如，建立名为房间出租的关系可以如下方式进行：

```
CREATE  TABLE  房间出租
房间号码        (INTEGER, NONULL),
房间类型        (INTEGER, VAR),
住客姓名        (CHCHAR (20), VAR),
性别            (CHCHAR ( 3 ), VAR),
年龄            (CHAR (10), VAR),
国籍            (CHCHAR ( 8 ), VAR),
护照编号        (CHAR (10), VAR),
工作单位        (CHCHAR (20), VAR),
职务            (CHCHAR (10), VAR),
来自何地        (CHCHAR (10), VAR),
计划逗留天数    (INTEGER, VAR),
下一站地点      (CHCHAR (10), VAR),
起住时间        (CHAR (12), VAR),
实际离馆时间    (CHAR (20), VAR),
房租结算        (REAL, VAR),
```

登记者编号　(INTEGER, VAR))

要撤销一个关系可用 DROP 命令。如果我们在关系型数据库中曾建立过某个英国旅客的关系，而这个关系现在用不着了而要撤销这个关系，我们就可以用这个命令。

二、关系中扩充新属性

由于事先考虑不周，往往在建立一个关系时会遗忘某些属性，而在以后的使用中又发现了这一遗漏。这时，可使用扩充语句，对已定义的关系中扩充新的属性。

例如，在房间出租关系中，费用结算除了房租外，还要增添其他费用一项，则可用命令

EXPAND TABLE 服务记录

ADD COLUMN 总计 (REAL, VAR)

三、插入操作

如果 138 房间的旅客于 83.5.20 交旅馆洗两件衬衣。这一事实应该在服务记录中进行登记。这时可用命令 INSERT。

例如：INSERT 服务记录 (编号，服务项目，单价，数量，日期)(138, 洗衬衣, 0.50, 2, 83.5.20)

INSERT 命令中属姓名的次序是不重要的，重要的是属姓名的次序要和下面具体插入的元组（记录）中的属性值的次序相一致。

四、删除操作

删除操作就是删去某关系中某个元组（记录），被删去的元组应满足的条件由 WHERE 子句说明。若无 WHERE 子句，则认为由它说明的条件恒真，将删去所有的元组，该关系就成了空集合，即没有一个实在的元组（记录）存在，但此时作为关系模式仍被登记在库目录中，因此它是有别于撤销操作的。

例如：编号为 20840 的旅客已结帐离开旅馆，此时可用下列命令：

DELETE 旅客名

WHERE 编号＝20840

也可以把该记录拷贝到磁盘上作为历史记录存档。

INSERT 房间待整理 (房间号码，单价)(20840, 18.00)

UPDATE 床位状态

SET　　状态＝待整理

WHERE 编号＝20840

五、修改操作

执行修改操作时，首先按 WHERE 子句选出需要修改的元组或元组集，然后按 SET 子句规定的修改表达式进行修改。

例如：1983 年 5 月 20 日进旅馆的两位英国旅客，其护照号码为 830172, 830179, 因故需延长逗留期 2 天，可用如下命令：

UPDATE 房间出租

SET　　计划逗留天数＝计划逗留天数＋2

WHERE 起住时间＝1983.5.20

　　AND 国籍＝ENGLISH

　AND (护照号码＝830172 OR

护照号码＝830179）

六、检索操作

检索是数据库上最基本的操作，可通过SELECT语句——FROM子句——WHERE子句来描述。

例如：有人到本旅馆询问名为李明的旅客，可用命令：

SELECT ALL
FROM 房间出租
WHERE 姓名＝李明

如果旅馆中有两个李明（同名同姓），则上述操作能得到两份旅客姓名都是李明的登记表，因为姓名不是主关键字。

例如：询问1983年5月20日以后来旅馆至今未离馆的日本东京旅客的姓名、职业、性别和工作单位。这时检索语句为：

SELECT 姓名，职务，性别，工作单位
FROM 房间出租
WHERE 起住时间≥83.5.20.
AND来自何地＝日本东京

又例如：查询本旅馆目前所有的空房间号码。则可用以下语句：

SELECT 房间号码
FROM 房间待整理，备用房间

WHERE子句未列出限制条件恒真，通过此查询可以列出本旅馆已整理或未整理的所有空房间的号码。

16.2 订票系统

飞机、火车或轮船的订票系统仅仅是航空、铁路或水运服务系统中一个面向旅客的（子）系统。订票问题不只是运输部门存在，在公共娱乐场所等部门也同样存在。

有代表性的订票系统是美国航空公司于1952年采用的Magnetronic Reservisor系统，随后于1959年开始研究塞泊（SABRE）系统，该系统于1963年正式投入营业，它是美国航空公司的一个综合性系统。该系统不仅用于订票，而且还从事预约信息、航空信息等各种信息的处理，能造出经营管理方面的报表。该系统由一台大型计算机联结1300个终端设备。日本研制的MARS-1国营铁路订票系统，该系统在1960年投入使用。紧接着投入对MARS-102，103的研制，于1963年投入使用，它可实现在一个星期内对全国507次列车车票的预订工作。它是一个多重并行处理的多机系统，共有852台终端设备。作为一个实用的订票系统，它必定是一个联机实时系统。

下面就以民航订票系统为例介绍系统设计思想。图16-2是局部民用航线示意图。为了叙述的方便，本节中所说的本地、某城市均指航空班机实际停靠的城市（名）。

作为一个民航订票系统一般要提供以下四个功能：

（1）向本地民航售票处订购到某城市的飞机票；

（2）对飞机航班的查询；

（3）退票处理；

（4）从本地出发，要求不重复地周游指定的几个城市，选择合适航线，使其费用（指购买飞机票的总的钱数）最少，然后根据每一城市要逗留的时间预订机票。

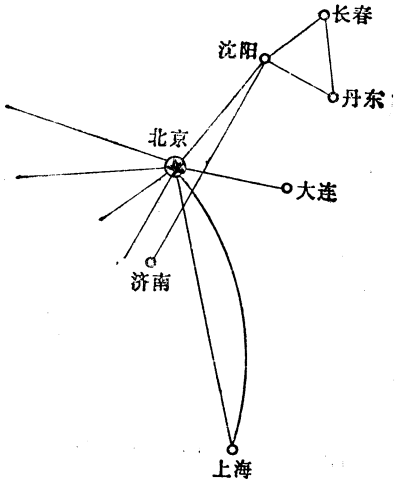


图16-2　局部民用航线示意图

16.2.1　数据结构

首先要建立若干个数据文件

1）对每一个城市建立一个航班文件，不妨用该城市名作为该城市航班文件名。

该文件中主要包括：下一站名；开航日期；开航时间；到站时间；航程；头等舱基本票价；二等舱基本票价等。

头等（或二等）舱基本票价是指对国内全票旅客的票价，半票或其他票价均可依此进行折算。

2）对每一航班要构造一张航班情况登记表，该登记表中主要包含下列两类信息：

（1）航班信息。主要有航班号、机型、机长姓名、驾驶员姓名、航行情况记录等。

（2）旅客信息。主要有座位号、旅客姓名、国籍、工作单位、职务、客票种类、客票号码、票价、代购单位、购票人、联系电话等。

16.2.2　算法设计

航空线路同铁路、公路相比，其差别是：通过第三城市到达目的城市，在路程和票价上均大于或等于直达目的城市。为了减少算法的复杂性，可以作以下假设：

一、采用深度为主的搜索法

为了便于理解，下面用实例来说明。

用户问题：要订购上海到丹东的飞机票。

用户提示：北京是必须经过的中转站（简称必经点）。

（1）取出上海文件，首先找上海文件是否有直达丹东的航班（见图16-3）。若发现没有直达航班，则作下一步。

（2）查是否有上海到必经点北京的航班。若发现有此航班，则用户终端显示上海

到北京的所有航班供旅客选择订票，并在航班文件中登录。

（3）取出北京文件，查北京文件是否有直达丹东的航班。若发现没有直达航班，则作下一步。

（4）查是否有北京到达其他必经点的航班。若发现没有其他必经点，也就没有这种航班。

（5）在北京文件中依次查下一站名。北京文件如表16-1所列。

根据深度为主搜索，沿大连进行搜索。但是，由于大连文件中除了有到北京的航班外无其他航班，故返回北京文件，取下一站为上海（1）、（2）。上海是整个航程的出发点，无疑不合要求，故舍去。再取北京文件下一站为沈阳。

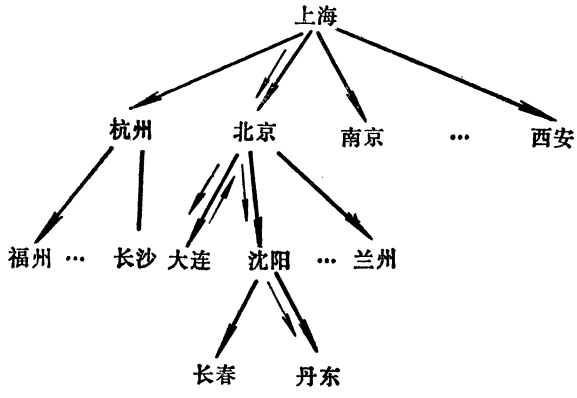


图16-3　上海→丹东的航线搜索过程

表16-1　北京航班文件

下一站名	开航日期	开航时间	到站时间	……
大连	每　天	18.20	19.50	……
上海(1)	1、2、3、4、7	10.25	12.05	……
上海(2)	1、3、4、5、6	15.55	17.35	……
沈阳	2、4	15.00	16.50	……
……❶	……	……	……	……

❶ 为了简化问题，不考虑其他航班。

（6）取沈阳文件。在沈阳文件中有三个下一站，它们是长春、济南和丹东，其中有目的站，故其他站可不考虑。

（7）在用户终端上显示上海到北京，北京到沈阳，沈阳到丹东的各次航班供用户选择。然后，一方面在各选中的航班中进行登录，另一方面通过汉字印刷设备，输出预订的飞机票。

由于在本例中有用户提示，使问题的算法大为简化，否则，根据深度为主搜索法需较长的检索时间。

实现功能（2）、（3）的算法比较容易，下面讨论实现上节功能（4）的算法。

二、用户问题

（1）从本地（规定编号为n）出发，不重复地经过$n-1$个城市，使购买飞机票的费用最省；

（2）根据每一城市要逗留的时间进行预订机票。

为了解决问题（1），对选定的n个城市，分别用$1\sim n$进行编号，本地城市用整数n表示。根据城市文件，构造$n\times n$矩阵，矩阵中的元素C_{ij}表示从城市i到城市j的飞机票的票价。若城市i到城市j之间有p条航线可通，则选取其中费用最省的一条。

又因为整个航线图是一张连通图，因此这样的 C_{ij} 一定存在，当 $i=j$ 的 $C_{ij}=\infty$。● $C=(C_{ij})\ n\times n$ 称票价矩阵。现在我们看到，每一条航线实际上和整数编号 1，2，…，$n-1$ 的一个排列 p 之间存在一一对应的关系。因此，对一个给定的排列，我们可以很容易地追踪出与这个网络模型相对应的航线，进一步可解决用户问题（2）的要求。为此，对所有的排列表一个一个地计算，并把到目前为止找出的费用最省的路线存储起来。如果我们找到一条费用更省的路线，就把这条路线作为下一次比较的依据。

设飞行路线为 ROUTE（1：n）；最少飞机票价为 MIN。票价矩阵中的行列次序和 $n-1$ 个城市编号相一致。一个排列 p 和一条航线 T（p）相对应，同时可以计算总的飞机票价为 COST(T(p))，若当前航线的总的飞机票价小于当前最小飞机票价 MIN，则 MIN←COST(T(p))，否则继续选下一条航线，最后得到最小飞机票价 MIN 以及对应的航线 ROUTE。其算法（为了描述算法的方便起见，这里采用拟 ALGOL 语言）如下：

```
Begin
  I←1; ROUTE←0; MIN←∞      /初始化/
  While I≤(N-1)!do
    Begin
      P←PERMUTE(N-1, I)     /PERMUTE(N-1, I) 是生成1到N-1的第I个排列的子算法/
      COST(T(P))←EFP(C, T)    /EFP(C, T)是由票价矩阵C及飞行路线T(P)所算得的总的飞机票价子
                               算法/
      if COST(T(P))<MIN then   /比较/
        Begin
        ROUTE←T(P)
        MIN←COST(T(P))
      End
      I←I+1
    End
    PRINT MIN, ROUTE     /输出打印/
  End
```

根据已选得的总的飞机票价最短的航线 ROUTE，按问题（2）的要求，不难得到旅客满意的订票。

16.2.3 订票系统的硬件配置

一般可取下述两个方案：

（1）采用一台大型机，再配置几十组到几百组终端设备，每一个城市（根据我国一个城市有一个民航售票处的特点）配备一组终端设备。每一组终端设备由一台汉字显示器及一台汉字印刷机组成。

（2）采用分布式远程网络。每一个城市为一个网络结点，每一结点有一台小型机或一台高档微型机（带有汉字显示器及汉字印刷机）。

16.3 电话查号系统

凡是装有电话机的城市和地区都不可避免地存在要求查询某某单位的电话号码的问

● 对于∞，具体实现时，可用计算机能表示的充分大的数字来替代。

题。出现这种问题的原因有两个，其一不可能要求每一个要打电话的用户手头都有一本电话簿；其二电话经常出现装拆、迁移，这样电话号码就成为一种动态记录，即使定期出版的“电话号码簿”也无法正确无误地反映实际情况。所以无论是从为用户服务，或是为使电话线路畅通着想，实行电话查号服务是十分必要的。目前世界各地的电话局为此都投入相当多的人力、物力，它已成为邮电事业中一个重要的业务部门。查号效率的高低已成为有关电话局服务水平的标志之一。以往沿用的人工查号方法效率低，话务员的劳动强度大，因此采用自动电话查询具有重要意义。迄今为止，自动电话查询有两种途径，一是利用缩微胶卷（MF），另一种是电子计算机电话查号。下面介绍计算机电话查号系统。

16.3.1 电话查号的基本环节

采用电子计算机进行电话查号有三个基本环节，它们是接受用户查询；电子计算机进行自动查号；回答用户的询问。

不管用户如何询问，其中至少包含一个他希望知道某个单位的电话号码。我们把这种单位的名称称为用户名信息。计算机接受这一户名信息最理想的方式是直接接受用户的声音，但目前的技术水平还不可能达到适应于各种用户（各种方言、不同年龄、不同性别)声音都能让计算机直接接受，随着人工智能技术的不断进步，今后的发展终将得到实现。目前还只能请话务员去听取用户的询问，随后把户名信息通过汉字输入装置（键盘）送入计算机。在计算机内部，不论采用何种转换方式，都要找到户名信息所对应的单位的电话号码。根据电话号码，自动声音报号装置便实现自动报号。这里，采用了数字信息合成的办法，以 8192 位二进制信息来表示一个声音，将所有这些信息都存入固定只读存储器，作为声源。其中，只要有十二个音符就可以实现报号，这些音符包括数字 0 到 9，间隔音和“局”音。需要时，就可取出来组成电话号码的语音，并控制其连续向查询者报发多次。

16.3.2 户名信息模式

在实际实现电话查号时，为了提高查询速度，作为查询依据的户名信息可以分为以下两类，并采用不同的方法输入。

一、常用户名

对那些被用户经常查询的单位称为常用户名，如火车站、轮船码头和民航局的问询处，水电、煤气急修站等。对一个地区或一个城市而言，称得上常用户的不会很多。我们将上海的这些用户名称加以编号，如火车站问询处编为 1 号，轮船码头问询处编为 2 号，青年宫编为 54 号，总工会编为 51 号，中共一大会址编为 71 号等等。话务员必须将这些编号和对应户名关系记牢，当要查询时，只需在键盘上按下

N↵

这里N是要求查询的常用户名的编号（十进制数），“↵”表示操作结束符。常用户名及编号可根据地区或城市不同而由话务员自行规定。

二、普通户名

常用户名的数量不宜太多，多了话务员记不住。除此以外的用户称为普通户名，普

通户名面向话务员的规范形式（如图 16-4 所示）。

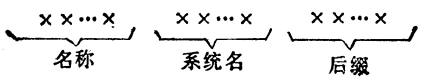


图16-4　普通户名规范形式

其中，系统名是用来表示该单位所在的实际分类，如“大学”、“中学”、“研究所”、“饭店”、“百货商店”等等，在这一部分中出现的每一汉字简单地对应汉语拼音的首字母列出即可，如“DX”(大学)、“ZX”(中学)、“YJS”(研究所）及“FD”(饭店）等等。系统名最多占四个字符，对应于每一个系统名，可以组成一个电话号码文件，系统名可以和文件名相一致。

面向话务员的普通用户名的规范形式是符合人们呼名习惯的。在对于每一个普通户名的规范形式中，可以没有后缀，但肯定有名称和系统名。例如：

（名称）	（系统名）	（后缀）
北京	火车站	问询处 ↵
市	政府	接待室 ↵
北京	饭店	服务台 ↵
朱乃进	私人（电话）	↵
魏师华	私人	↵

为了查找的需要，在计算机内要存储“电话号码簿”。倘若按照出版的电话号码簿顺序存储，则作为一个文件，检索时无疑要花费很长的时间。如果把“电话号码簿”构成一个文件系统，而把常用电话作为一种特例，则由于查找它特别频繁，故可在内存中开辟一个区域，把常用电话号码依次集中登录，如表 16-2 所列。

表16-2　常用电话号码格式

编　码	电话号码

对于普通户名，要采用多级查找方式。

16.3.3　查号流程图

整个系统的查号过程如图 16-5 所示。

若键盘输入的户名信息是某一整数N（说明它是常用户名），则可以直接查常用电话号码表，从中检出电话号码。

若键盘输入的户名信息是普通户名，则先看是否有后缀（把电话号码文件分成两大类，第一类电话文件对应的户名信息中无后缀）。若无后缀，则查找时先根据系统名作一级索引，然后根据户名信息中的名称，即可找到对应的电话号码。如果户名信息中有后缀，则根据系统名作一级索引后，再根据名称作二级索引，然后根据后缀找到对应的电话号码（见图 16-6）。

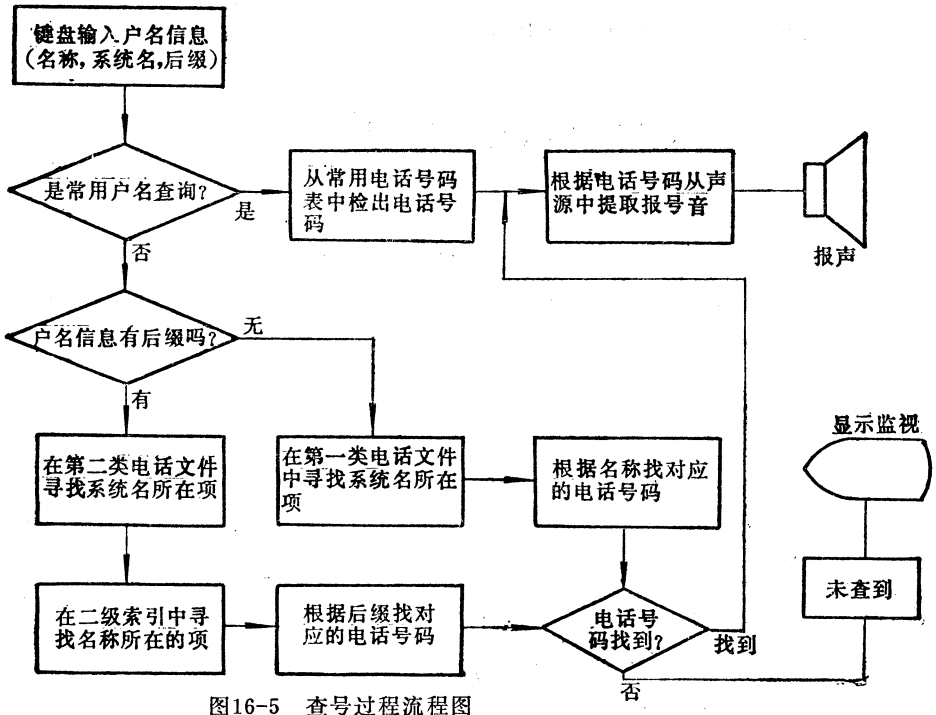


图16-5　查号过程流程图

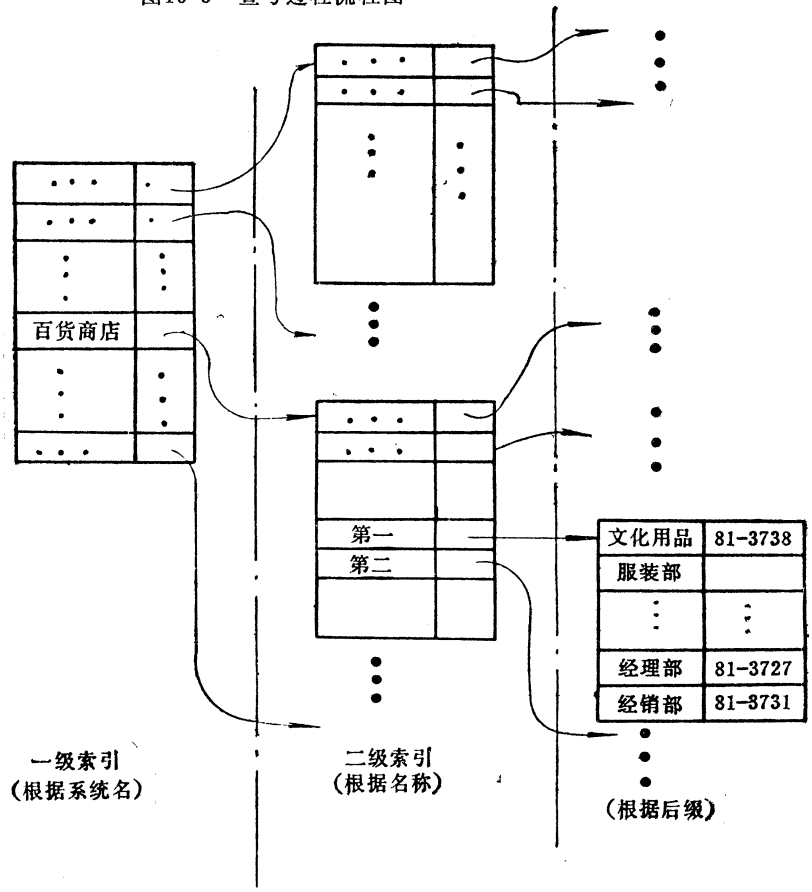


图16-6　带后缀的普通户名的内部查询装置报出要查的电话号码

16.3.4 查号系统硬件配置

电话查号采用一个多任务实时系统。在图 16-7 中，有四个并行接口和处理机联接。每一个接口可联接八个操作台，八个操作台之间是串行的。每一个操作台又管辖六条中继线。这样，一个电话查号系统可以对付 192 路询查。

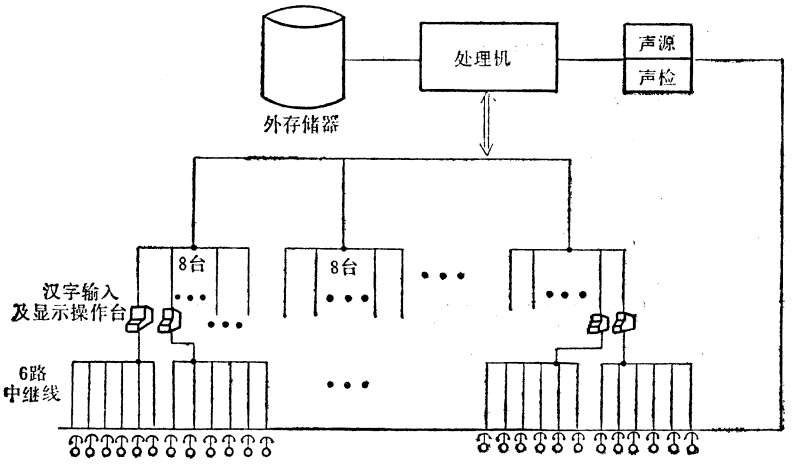


图16-7 查号系统硬件配置

16.4 汉字语言自动处理系统

16.4.1 什么是语言自动处理系统

在信息社会中，表达和记录各种信息的最基本的工具是语言。语言是一个民族、一个国家或一个社会团体中通用的、用来表达和交流思想的特殊系统。语言又可以分为言语和文字两大类。言语是由表达语言的声音组成，文字是由记录语言的书面符号组成。语言本身也是信息。言语是声音信息，文字是（符号）图象信息。汉字语言自动处理就是采用电子计算机，对语言内部的各个构件（语音、文字、词汇、标点符号和非文字符号等），对一段特定的语言片断（又称基础原文）以及各种规则（词法、语法）等进行分类、分析、统计、归纳、综合、比较，用以帮助人们更深刻、更科学地认识和研究语言内部规律的一门新型技术。对于处理文字输入基础原文的是书面语言自动处理系统；对于处理言语输入基础原文的是言语自动处理系统。

汉字语言自动处理系统的研究，为建立汉语计算语言学以及汉语语言理论提供了自动化的研究手段，它开拓了计算机应用的新领域。它也将为我国少数民族文字（如藏文、蒙文、维吾尔文等）的自动处理、机器翻释、自然语言理解等研究提供了先例。

16.4.2 语言自动处理系统的构成

语言自动处理系统的构成如图16-8所示。

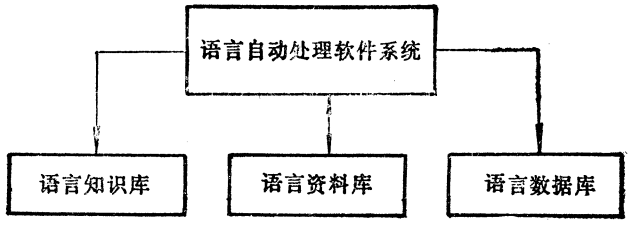


图16-8　语言自动处理系统

由图可见，语言自动处理系统必须由下述四部分组成：

（1）语言自动处理软件系统；

（2）语言资料库；

（3）语言知识库；

（4）语言数据库。

语言自动处理软件系统具备各种分析、统计、综合、归纳、编目、检索等功能。语言知识库是人赋予计算机的各种语言知识，包括语言的各种规则和模式，计算机就是根据这些知识对基础原文进行各种自动处理的。语言资料库是存入计算机的各种著作（也就是人们交给计算机的待处理的语言资料，是计算机进行自动处理的对象），它是人们建立的机器“图书馆”。语言数据库用以储存语言资料经过自动处理后的信息，用户可借助显示器、打印机或其他输出设备输出索引、图表、卡片等各种语言数据。

必须特别指出的是，知识库的知识和资料库的资料都要不断地充实积累，随着资料、知识的增加，数据库中的各种数据也相应地在自动增加或更新。

资料库的语言资料和数据库的语言数据将构成一个现代化的语言资料中心。它可随时向语言研究者提供所需要的资料和数据。

下面分别介绍语言自动处理系统的各个组成部分。

16.4.3　语言自动处理软件系统

语言自动处理软件系统主要包括原文输入编辑程序、系统管理程序、自动处理程序库、咨询服务系统，以及信息转储和通信系统等（见图16-9）。

自动处理程序库是自动处理软件系统的核心。它所处理的内容十分广泛。例如：

（1）书面语言处理和口头语言的处理；

（2）书面语言的自动处理又分为汉语、英语、日语、法语、俄语等等语种的处理；

（3）语种之间的自动转换（机器翻译）；

（4）汉语的自动处理又可按年代进行分类，例如现代汉语处理、近代汉语处理和古代汉语处理等；

（5）当代、现代、近代、古代汉语处理都是断代自动处理，而在它们之间建立的综合比较处理是历史比较法的自动化，它属于汉语史的自动处理，将为语音史、文字史、词汇史和语法史的研究提供各个时期的综合或比较的数据；

（6）以现代汉语的处理而言，又包括语音处理、文字处理、词汇处理、语法处理、非语言符号处理等。在这些项目的自动处理中，以语法处理为最困难；

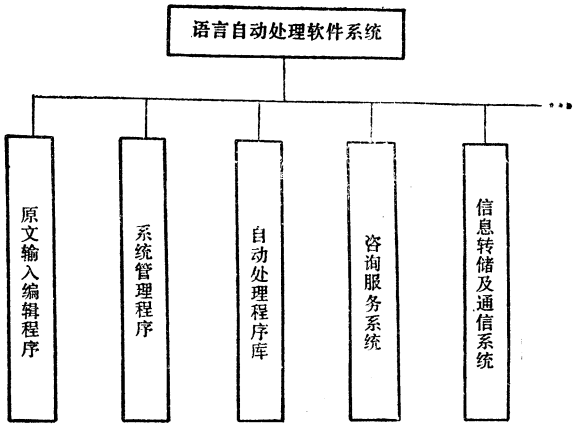


图16-9　自动处理软件系统的组成

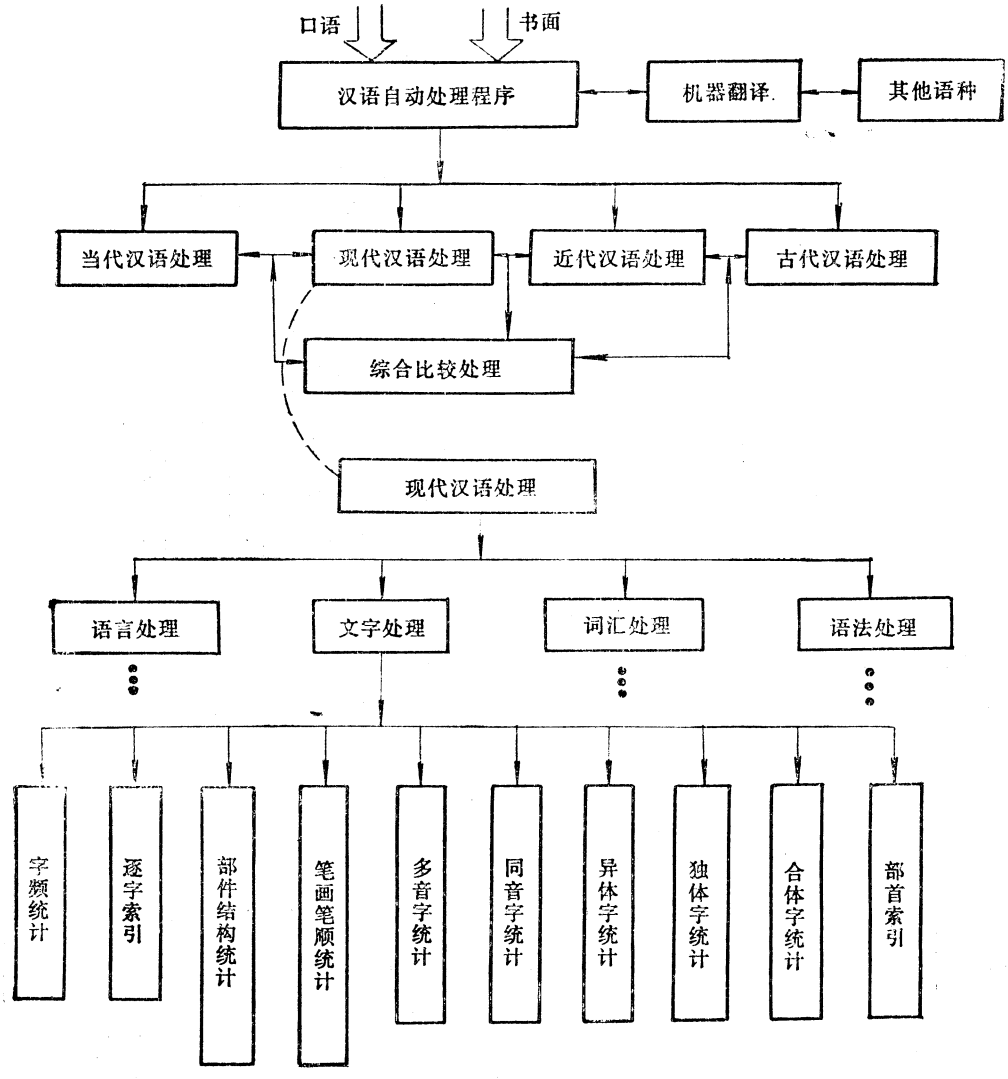


图16-10　汉字自动处理程序示意图

（7）各个语种、同一语言的不同时期、不同的应用领域又可包括许多细目的自动处理，以汉语文字系统处理而言，就包括字频统计、部件和结构统计、笔画和笔顺统计、多音字统计、同音字统计、异体字统计、独体字统计、合体字统计、自动编制逐字索引、单字音序索引、部首索引、笔画索引等。图 16-10 是汉字自动处理程序的构成。

16.4.4 语言资料库的构成

语言资料库是自动化的机器“图书馆”。其中存储的各种语言资料，是自动处理系统的信息源，加工的原材料。语言资料的存储应允许各种分类的实现和检索要求。语言资料库的资料既可以按时期分类提取，也可以按学科、作者、文体或其他分类方法来提取。每一大类下面还可以再分小类，例如按学科分类可以分社会科学和自然科学两类；社会科学又可分为文、史、哲、经等各小类；自然科学又可分为数学、力学、物理学、化学、天文学、生物科学、电气和电子技术、机械工程等等。各小类还可以进一步细分，如图16-11所示。

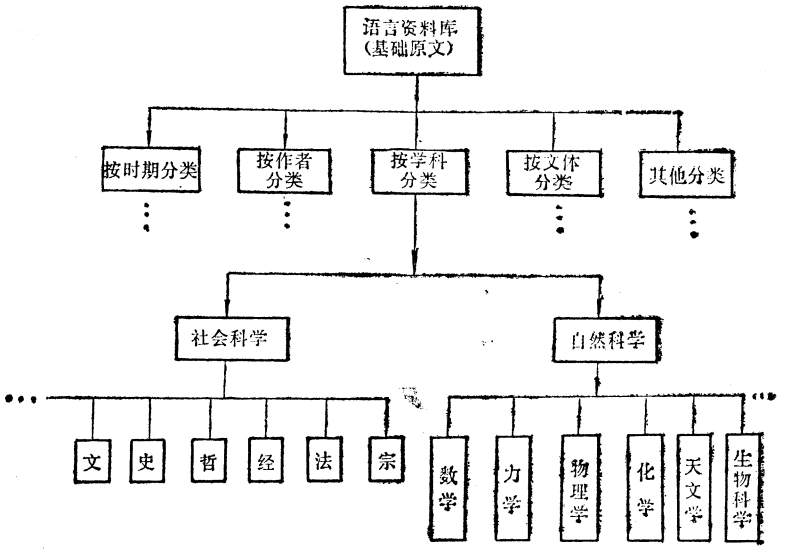


图16-11 语言资料库示意图

用语言自动处理系统的综合比较程序对各种分类资料分别进行综合或比较的处理，就是对语言的各种比较研究。按时期进行综合比较就是对语言的历史作比较研究；按作者进行综合比较的是对语言的风格作比较研究；按学科的综合比较是对语言的学科比较研究；按文体的综合比较是对语言的体裁作比较研究等等。

16.4.5 语言知识库的构成

语言知识库中存储着人们赋予计算机的各种语言文字“知识”，计算机运用已掌握的这些知识，对资料库中的语言资料从语音、文字、词汇、语法等各个角度进行加工处理。各种处理的结果数据反过来又不断充实和更新知识库中的“知识”。

例如要处理某基础原文中“圆周率”一词，在知识库中你必须有22/7，3.141592654…等数据信息与“圆周率”一词相联系。又例如，语法知识库中有一结构规则是：（主）+

(状)+(谓)+(补)+(定)+(宾)。现分析基础原文中一个句子："我们又筑成一条铁路"。显然该句子符合上述语法结构。但若有一句子是："我们又筑成铁路一条"。该句子就是一个病句，因其语序不符合规则。

对知识库中的知识可分为语音知识、文字知识、词汇知识、语法知识和其他知识等。各类"知识"又可包括许多小类。以文字知识为例，它应包括各种字典、词典、各种字义、繁简字对照表、新旧字形对照表、姓氏字表、地名字表等等（见图 16-12 所示）。

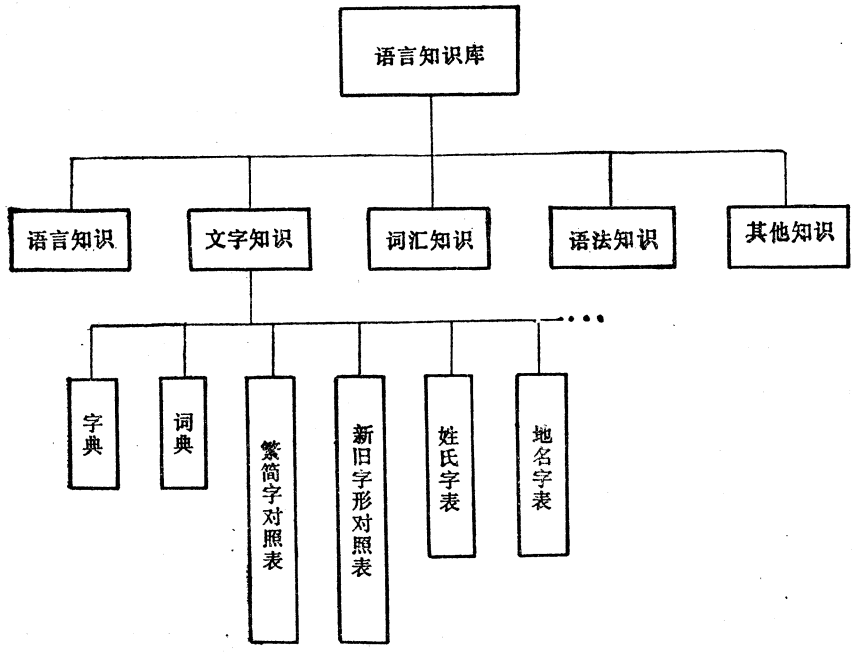


图16-12　语言知识库示意

16.4.6　语言数据库的构成

语言数据库应按较好的数据结构科学地存储各种经自动处理的语言信息，以供人们从各种角度灵活地检索，达到咨询服务的目的。究竟建立怎样的语言数据库，这同用户要求密切相关，决不能闭门造车。

这里所指的用户需要，可以是中、小学语文教材的编纂、情报自动检索、文献自动检索、名人、公安、法律、病例等档案自动检索、文字自动识别、语音自动识别、自然语言理解、机器翻译、词典编纂、文字改革等等。

以词典编纂为例，目前国内花费大量人力物力制作了大量的资料卡片，据估计国内各地共有二千万张卡片，不少是重复的，不重复的那部分有用卡片又分散在全国各地，无法统一管理，统一调用。如果把这些卡片送入计算机的语言资料库，要进行各种词典的编纂就显得十分容易，从而可大大提高编纂速度。

目前汉字语言自动处理系统的基础研究刚刚开始，有大量的研究课题需要数以万计的各类专家去探索。

16.5 结 束 语

随着社会的进步和科学的发展，计算机的应用正在渗透到人类社会的一切领域。

除了前几章介绍的与汉字信息处理技术关系甚为密切的汉字情报检索系统、汉字企业管理系统、精密汉字编辑照排系统、汉字联机通信网络系统、中医诊疗系统等之外，类似的系统还可以举出很多。下面列举的几例也是应用面很广的：

（一）计算机辅助军事指挥系统

该系统的功能包括：存储大量的军事地图，敌我双方各部队的番号、编制、驻地及战斗特长等信息；收集已有的军事情报进行汇总、分析和分类；对敌方军事行动的预测，对我方军事行动的决策选择和决策评价等等。指挥部向前沿阵地发布的口头的或手写的命令通过密码传送到前沿阵地进行解密后作语音转换或作文字显示。

（二）公用信息服务系统

公用信息服务系统为每个用户（个人、家庭、单位）提供新闻报导、影剧预告、各种广告、导游、购货指南以及市场商品价格等咨询业务。

（三）计算机教育系统

用户在取得计算机教育系统对某些教学环节的使用权后，用户可在家里或办公室内面对计算机终端进行某门课程的学习，用户可通过口述、键盘或光笔手写向计算机提出问题，请求答疑。计算机通过语音、显示屏幕或印刷机输出给予回答。进一步，计算机将根据用户提出的问题回答情况，自动编制思考题供用户参考。

（四）交通控制系统

（五）各种自动检测系统

（六）邮件分检系统

本书所列举的计算机应用系统，对整个计算机的应用领域来说只是海边拾贝。据悉，目前世界上计算机应用的类型已达五千多种，大量的应用项目需要我们去研究和开发。在我国，汉字信息处理技术的发展和应用，将为我国进入信息化社会显示出强大的生命力。