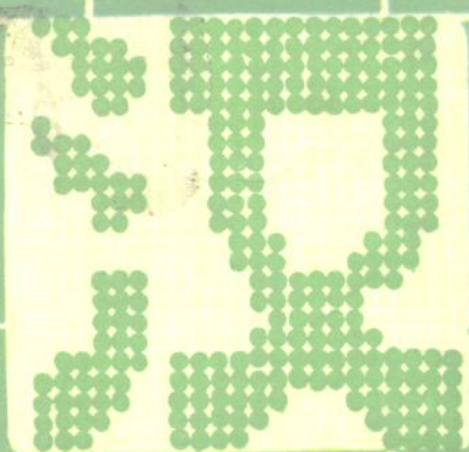


郭平欣 张淞芝 主编

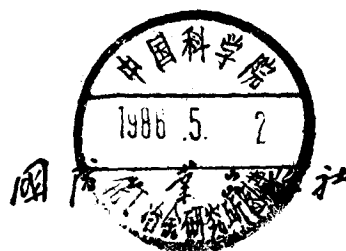
汉字信息处理技术



国防工业出版社

汉字信息处理技术

郭平欣 张淞芝 主编



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 张淦芝 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34³/₈ 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

JS/00/23

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 蒋攢平 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 陈堃铎
- 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 and 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

1.1 汉字信息处理的意義和任务	1
1.1.1 什么叫汉字信息处理	1
1.1.2 汉字信息处理技术涉及的范围	2
1.1.3 汉字信息处理技术要解决的问题	5
1.2 汉字信息处理系统的构成和分类	13
1.2.1 汉字信息处理系统的构成	13
1.2.2 汉字信息处理系统的分类	14
1.2.3 汉字信息处理技术标准化问题	17
1.3 汉字信息处理技术的现状和展望	18
1.3.1 国内汉字信息处理技术的现状	18
1.3.2 国外汉字信息处理技术状况	21
1.3.3 汉字信息处理技术的发展前景	24

第二章 汉字属性

2.1 汉字演变概况	26
2.2 汉字字量	26
2.2.1 汉字的累积字量	26
2.2.2 汉字实用处理 (一)	27
2.2.3 汉字实用处理 (二)	27
2.2.4 汉字实用处理 (三)	28
2.3 汉字字形	28
2.3.1 汉字的形体结构及其分解	28
2.3.2 汉字图象的细胞——位点	29
2.3.3 笔画	29
2.3.4 字根	31
2.3.5 部首和字首	33
2.3.6 单字	36
2.4 汉字字音	39
2.4.1 汉语和汉字	39
2.4.2 汉字反切法和拉丁字母式双拼法	39
2.5 汉字字义	40
2.6 汉字排序	41
2.6.1 汉字排序的意义	41
2.6.2 汉字排序法	41
2.7 汉字信息处理与汉字属性	44

第三章 汉字输入编码方法

3.1 汉字输入编码概述	45
3.2 汉字集及其划分	46
3.2.1 按汉字天然属性划分的子集	46
3.2.2 字母集上的有序组	48
3.2.3 汉字的笛卡尔积集	49
3.2.4 汉字代码集	50
3.2.5 汉字输入编码的简单模型	50
3.3 汉字输入编码的设计	52
3.3.1 字种的确定	52
3.3.2 汉字的熵值	53
3.3.3 选择键盘类型	54
3.3.4 选择汉字属性类型	54
3.3.5 汉字属性元素在键位上的配置	55
3.3.6 重码数量预测	56
3.3.7 代码表的编制	56
3.4 汉字输入代码的类型	57
3.4.1 概述	57
3.4.2 字根代码类	57
3.4.3 角形代码类	59
3.4.4 笔形代码类	60
3.4.5 字音代码类	61
3.4.6 音、形等相结合的代码类	62
3.5 汉字输入编码方法的计算机辅助设计	63
3.6 汉字输入编码方法的评测	64

第四章 信息处理交换用的汉字代码

4.1 概述	68
4.2 汉字代码种类	68
4.2.1 汉字输入码	69
4.2.2 汉字内部码	69
4.2.3 汉字地址码	71
4.2.4 汉字交换码	71
4.2.5 汉字控制功能码	72
4.2.6 汉字扩充码	73
4.2.7 汉字字形码	73
4.3 汉字代码的标准化	74
4.3.1 GB1988《信息处理交换用的七位编码字符集》	75
4.3.2 GB2311《信息处理交换用七位编码字符集的扩充方法》	77
4.3.3 GB2312《信息交换用汉字编码字符集——基本集》	80
4.3.4 汉字点阵字模的设计与标准化	81

4.3.5 汉字交换码辅助集的标准化	83
4.3.6 文字图形设备增补控制功能的标准化	86

第五章 汉字输入方法和设备

5.1 汉字键盘输入方法	95
5.1.1 汉字整字键盘	96
5.1.2 笔触式汉字字盘	97
5.1.3 中文打字机式汉字键盘	105
5.1.4 汉字字根键盘	106
5.1.5 标准字母数字键盘	110
5.1.6 联想式人机对话汉字输入方法	114
5.1.7 键盘控制器	115
5.1.8 键盘式汉字输入设备的操作性能	121
5.2 汉字语音输入方法	123
5.2.1 声音产生的基本物理原理	124
5.2.2 语音识别的特征参量	125
5.2.3 语音识别方法	129
5.2.4 语音识别系统实例	133
5.3 汉字字形输入方法	137
5.3.1 概述	137
5.3.2 印刷体汉字的字形输入	138
5.3.3 手写体汉字的字形输入	149
5.3.4 联机手写体汉字的识别	155
5.3.5 小结	157

第六章 汉字字形发生器

6.1 汉字字形的数字化表示	158
6.1.1 汉字字形的特点	158
6.1.2 汉字字形的数字化	159
6.1.3 汉字点阵字模的制作	161
6.2 汉字字形发生器	162
6.2.1 存储器概述	163
6.2.2 汉字字形发生器的存储容量	165
6.2.3 汉字字形发生器的缓冲存储器	166
6.3 汉字字形的压缩存储方法	171
6.3.1 部件组字法	171
6.3.2 向量存储法	172

第七章 汉字输出设备

7.1 汉字印刷技术概述	178
7.1.1 汉字印刷输出设备的功能	178

7.1.2 汉字印刷输出设备的类型	179
7.1.3 汉字印刷机的发展概况	180
7.2 针式汉字打印机	180
7.2.1 针式汉字打印机的性能特点	180
7.2.2 针式汉字打印机的工作原理	181
7.2.3 针式汉字打印机的组成与动作	183
7.2.4 针式汉字打印机的接口及其控制信息	186
7.2.5 针式汉字打印机的选用	191
7.3 激光汉字印刷机	193
7.3.1 激光汉字印刷机的工作原理	193
7.3.2 激光汉字印刷机的组成	194
7.3.3 激光汉字印刷机的动作与性能	197
7.3.4 激光汉字印刷机的选用	202
7.4 喷墨、热感、静电、光纤管等式样的汉字印刷机	203
7.4.1 喷墨式汉字印刷机	204
7.4.2 热感式汉字印刷机	206
7.4.3 静电式汉字印刷机	208
7.4.4 光纤管转印汉字印刷机	210
7.5 汉字语音输出技术和设备	211
7.5.1 汉语语音合成装置的组成	211
7.5.2 语音的合成	213

第八章 汉字显示终端

8.1 汉字显示器	218
8.1.1 CRT 显示器的扫描方式	218
8.1.2 显示器的刷新方法	223
8.2 汉字显示终端	231
8.2.1 概述	231
8.2.2 汉字显示终端的硬件结构	233

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	237
9.1.1 系统配置的基本考虑	237
9.1.2 微型机汉字信息处理系统的配置	237
9.1.3 小型机汉字信息处理系统的配置	242
9.1.4 中、大型机汉字信息处理系统配置	244
9.2 汉字输入处理	246
9.2.1 汉字输入方式	246
9.2.2 汉字输入程序	249
9.2.3 汉字输入编码转换程序	251
9.3 汉字输出处理	256

9.3.1 汉字字形输出	256
9.3.2 汉字输出程序	258
9.3.3 访问汉字字模库程序	262
9.4 扩充的汉字信息处理程序	268
9.4.1 编制汉字信息典程序	268
9.4.2 汉字字形旋转程序	271
9.4.3 汉字尺寸变倍程序	273
9.4.4 汉字文本编辑程序	281
9.4.5 汉字文件加密、解密程序	297

第十章 汉字数据处理系统软件

10.1 什么是系统软件	304
10.2 汉字数据处理问题的提出	305
10.2.1 汉字与西文兼容问题	306
10.2.2 汉字和字母数字在计算机内部的表示	306
10.3 汉字数据处理系统软件建立的方法	311
10.3.1 汉字数据处理系统软件建立的发展阶段	311
10.3.2 几种汉字数据处理系统软件的建立方法	320
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	325
10.4 大、中、小和微型机系统扩充汉字功能的考虑	327
10.4.1 在微型机上扩充汉字功能	327
10.4.2 在大、中、小型机上扩充汉字功能	328

第十一章 汉字情报检索系统

11.1 情报检索的一般概念	330
11.1.1 情报检索问题	330
11.1.2 传统处理方式	331
11.1.3 计算机处理方式	333
11.2 机读文档组织和检索策略	335
11.2.1 文档结构	336
11.2.2 词库组织	347
11.2.3 检索策略	349
11.2.4 系统设计和应用软件	353
11.2.5 实例	355
11.3 机读数据库组织和查询语言	363
11.3.1 必要性	363
11.3.2 分层模型和网状模型	364
11.3.3 关系模型、查询语言及其实例	372
11.4 汉字和西文情报检索系统的异同点	377
11.4.1 汉字情报的标引和表示	377
11.4.2 自动标引	378

11.4.3	词库	378
11.4.4	汉字情报检索系统对汉字输入输出技术的要求	378

第十二章 汉字联机通信网络

12.1	数据通信技术	379
12.1.1	信道类型	379
12.1.2	调制解调器	380
12.1.3	同步与异步传输	382
12.1.4	RS-232C接口	383
12.1.5	通信规程	385
12.1.6	数据链路构成	386
12.1.7	信息代码	387
12.1.8	汉字数据通信	388
12.2	汉字计算机网络	388
12.2.1	计算机网络系统	388
12.2.2	数据交换方式	392
12.2.3	传输控制	395
12.2.4	通信控制	404
12.3	汉字联机信息处理系统	408
12.3.1	汉字终端联机接口	409
12.3.2	汉字终端远程适配器	415
12.4	汉字微型机局部网络	416
12.4.1	概述	416
12.4.2	局部网络工作原理	418
12.4.3	三种局部网络	421
12.4.4	汉字微型机局部网络	424

第十三章 精密汉字编辑排版系统

13.1	精密型照相排字机的几个发展阶段	428
13.1.1	手动(第一代)照排机	428
13.1.2	光机式(第二代)照排机	428
13.1.3	阴极射线管(第三代)照排机	429
13.1.4	激光(第四代)照排机	432
13.2	高分辨率汉字字形的存储和几种信息压缩方案	432
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	432
13.2.2	记录黑白段长度的压缩方法	433
13.2.3	霍夫曼压缩方法	434
13.2.4	字根组字的压缩方法	435
13.3	一种保证文字质量的高倍数汉字字形信息压缩技术	435
13.3.1	汉字规则笔画和不规则笔画的压缩表示	435
13.3.2	汉字字形复原技术	438
13.3.3	高分辨率汉字字形的放大和缩小技术	443

13.4 逐段生成汉字技术和复杂版面形成技术	446
13.4.1 高分辨率汉字字模的两级存储和调度	446
13.4.2 适合于激光扫描的版面描述方法	448
13.4.3 逐段生成汉字点阵	449
13.4.4 最终输出点阵的形成和激光扫描控制	451
13.4.5 Am2900 微处理机系统的设计方法	452
13.5 编辑排版软件系统	453
13.5.1 编辑排版系统结构	453
13.5.2 排版语言与排版编译程序	457
13.5.3 报纸的版面设计	463
13.6 编辑排版系统中的汉字终端子系统	466
13.6.1 终端子系统的类型	466
13.6.2 脱机终端的使用及其编辑功能	467
13.6.3 联机终端的功能及其使用方式	468

第十四章 汉字企业管理系统

14.1 现代企业管理系统简介	470
14.1.1 现代企业管理系统的主要功能	470
14.1.2 计算机在企业管理中的应用概况	472
14.1.3 实现计算机企业管理系统的条件和步骤	473
14.2 计划管理	477
14.2.1 计划管理的基本概念	477
14.2.2 计划管理的实例	478
14.3 生产管理	488
14.3.1 生产管理的基本概念	488
14.3.2 生产管理实例	489
14.4 现代企业管理中的汉字信息处理实例	493
14.4.1 概述	493
14.4.2 汉字文件档案处理实例	494
14.4.3 汉字制表语言	499
14.5 汉字企业管理系统的硬件配置实例	506

第十五章 中医诊疗系统

15.1 诊疗系统发展概况	508
15.2 建立计算机中医诊疗系统的意义	508
15.3 中医诊疗的理论基础	509
15.3.1 整体观念	509
15.3.2 辨证施治	509
15.4 中医诊疗的数学模型 I	510
15.4.1 症候群空间	510
15.4.2 隶属函数 $\mathcal{N}_{A_j}(x_i)$ 的计算	511

15.4.3 阈值的确定	512
15.4.4 浮动阈值技术	512
15.5 中医诊疗的数学模型 I 的程序流程及输出	512
15.6 中医诊疗的数学模型 II	514
15.7 多层推理网络中的生成规则和元规则	516
15.8 中医诊疗系统的硬件配置和输出实例	518
15.9 结束语	518

第十六章 其他应用

16.1 旅馆服务系统	520
16.1.1 旅馆服务业务简介	520
16.1.2 旅馆服务系统的硬件组成	521
16.1.3 建立在关系型数据库上的旅馆服务系统	522
16.2 订票系统	524
16.2.1 数据结构	525
16.2.2 算法设计	525
16.2.3 订票系统的硬件配置	527
16.3 电话查号系统	527
16.3.1 电话查号的基本环节	528
16.3.2 户名信息模式	528
16.3.3 查号流程图	529
16.3.4 查号系统硬件配置	531
16.4 汉字语言自动处理系统	531
16.4.1 什么是语言自动处理系统	531
16.4.2 语言自动处理系统的构成	531
16.4.3 语言自动处理软件系统	532
16.4.4 语言资料库的构成	534
16.4.5 语言知识库的构成	534
16.4.6 语言数据库的构成	535
16.5 结束语	536

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