汉字信息处理系统

东南大学出版社　曾庆辉 主编

汉字信息处理系统

曾庆辉　雍殿书
李文忠　迟鲁艺
编

东南大学出版社

内 容 提 要

本书分十章系统地介绍计算机汉字信息处理系统的基本知识、技术和应用。其内容深入浅出，通俗易懂，适合各类汉字信息处理人员阅读，亦可供计算机专业和非计算机专业选用教材。

汉字信息处理系统

曾庆耀 主编

东南大学出版社出版
南京四牌楼2号
江苏省新华书店发行　　大丰县印刷二厂印刷
开本787×1092毫米1/32　印张18×5/16　字数41[illegible]千字
1989年3月第1版　　1989年3月第1次印刷
印数：1—5000册

ISBN 7-81023-104-9

TP·5　　定价3.70元
责任编辑：张 克
责任校对：吕 岚

前　言

“文字”是信息的主要表达方式。在以计算机为支柱的信息化社会里，利用计算机对各种文字信息进行处理，如存贮、分类、统计、检索、转换、传输和控制等，已对人类社会的发展产生了深刻而巨大的影响。

我国汉字的历史源远流长，当今世界上，每三个人就有一个人使用汉字。因此，计算机系统是否具有汉字处理能力，将直接影响到计算机的推广和使用。汉字有其自身的构造特点，随着社会的发展，汉字本身也在演变过程之中。如何用计算机处理汉字信息，这一课题涉及到许多学科的交叉配合，它吸引了国内外众多学者在进行研究和探索。新的研究成果不断涌现，呈现出一派生机勃勃的繁荣景象。

为了教学工作的需要以及便于对此方面感兴趣的科技工作者能够比较系统地了解计算机汉字信息处理的基本知识，在编写本书时，力求做到深入浅出，通俗易懂，并在书后附有思考题以帮助读者加深理解。

在本书的内容编排上，首先介绍了汉字信息处理的概貌，以期让读者有一个完整的系统印象，然后介绍了汉字编码，汉字输入方式及设备，汉字库的建立及存贮，汉字显示终端，汉字印字技术，汉字信息处理系统的软件以及微机汉字信息处理系统的实现和计算机情报检索系统，最后为了便于读者能够联系实际，深入理解汉字信息处理系统，在本书中还列举了一些应用实例。

本书由曾庆辉主编和修订并编写了第五章，雍殿书编写了第一、二、三、七、九章和第十章的部分章节，李文忠编写了第一章和第十章的部分章节，迟鲁艺编写第四、六、八章和第十章的部分章节，全书由北京计算机系统工程研究所高级工程师吴克忠主审。

由于编者水平所限，错误与不足之处在所难免，敬希读者批评指正。

编　者

1988年5月

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第四章　汉字库的建立及存贮

第五章　汉字显示终端

第六章　汉字印字技术

第七章　汉字信息处理系统软件

第八章 微机汉字信息处理系统的实现

第九章 计算机情报检索系统

第十章 汉字信息处理技术的应用

第一章　汉字信息处理概论

语言是人类思维表达和思想交流的工具，文字是语言的书面记录，语言文字都是符号系统，是信息的主要表达方式，语言文字信息处理就是用现代化技术——计算机，对语言文字信息进行各种处理，如存贮、分类、统计、检索、转换、传输、控制和模拟等，使语言文字得到最佳利用，使凝聚在语言文字中的知识信息发挥最大的效能。

我国是汉字的发源地，使用汉字已有几千年的历史，目前世界上使用汉字的人口约占全世界总人口的36%，因此，改善汉字信息处理手段就显得更加迫切了。电子计算机的出现为汉字信息的自动化处理提供了极好的条件，目前国内已普遍开始用计算机来处理汉字信息，但如何研制出适合于我国国情的有实用价值的汉字信息处理系统是人们普遍关注的一个重要课题。

语言文字信息处理是由语言学、文字学、心理学、控制论、信息论、情报学、声学、通信技术、激光技术、微电子技术、自动化技术……多种学科相结合而产生的多边缘交叉性学科，它是信息科学的一个重要分支。

由于方块汉字的字体特殊，构形复杂，字种繁多，处理困难等原因，我国的语言文字处理基本上还是以手抄笔写或机械处理为主，远远落后于拼音文字的文字处理。

传统的文字信息处理采用手抄笔写，消耗了人们大量的精力和时间。要改变现状，及时沟通情报，提高办事效率，

把我们的管理工作搞好，有效的途径就是广泛地使用电子计算机。尤其在那些信息量大，要求时间短和信息复杂的事务处理方面，人工是代替不了的，更需要用计算机进行中文信息处理。因此尽快解决好中文信息处理技术，让电子计算机具有高效率处理汉字业务的能力，是实现四个现代化的需要。

我们必须根据我国的国情，充分认识汉字结构的特点，加强汉字信息的计算机处理研究，争取在不太长的时间内，用电子计算机技术武装我国古老的语言和文字。

1.1 汉字信息处理系统的研究内容和发展方向

汉字信息处理是一个涉及面较广的综合性学科。它包括文字结构分析、文字改革、汉字编码、词语统计、机器检索、汉语理解、汉字自动识别、语音识别、语音合成、中文信息交换与传输等一系列研究课题。

汉字信息处理系统可按如下办法分类。

1）按实用范围分

（1）专用系统　这类系统一般是大型汉字信息处理系统，它具有精美多变的输出字形和灵活方便的编辑功能（如激光照排系统）。

（2）通用系统　这种系统一般是指具有汉字输入输出和一定的文字编辑功能，能用于数据处理和文字处理，并具有中西文兼容的功能。

2）按系统功能分

（1）文字处理系统　这种系统侧重于文稿编辑功能。

（2）数据处理系统　这种系统要求具有很强的检索功能和严格的格式按排。

3）按实现方法分

（1）图形系统　这种系统按点元素方式处理图形，进而可处理汉字。它常常是专用或通用系统支持的应用软件包。

（2）字符系统　通用计算机系统是一个处理英文数字的字符系统，各种类型的数据都必须以字母或数字的形式输入，经过计算机处理后再以字母或数字的形式输出。汉字作为现有字符集的一个扩充集也要用字母或数字表示，以便计算机处理。

汉字信息处理的内容主要包括字形库的建立和管理、代码的转换和校验、汉字字符串的操作处理、汉字输出操作处理、汉字编辑处理、汉字标志符的识别处理等。目前国内的汉字信息处理系统，除专用的汉字信息处理系统外，一般是在原有系统基础上改造和扩充其软、硬件来实现的。汉字信息处理的实现方法可以归纳为三种类别。

（1）在语言编译系统上进行修改和扩充，增加汉字代码识别、变换和汉字处理程序，在ROM或磁盘上建立字模库。

（2）在操作系统层进行修改和扩充，把汉字的处理纳入原系统I/O管理和驱动模块内。这种方法能充分利用系统的其它软件资源，使用时也比较方便。

（3）插接兼容技术。这种方法是将汉字处理模块附加到计算机系统的内核——裸机部分，即附加汉字处理模块的字符设备。这种设备包括硬字库、编码转换、汉字存贮代码、标志的设定和识别，以及汉字发生器地址合成等功能。此法基本上不受机型和操作系统类型的限制，能够适用于多种高级语言和数据库的汉字处理。实际上汉字处理功能是在终端及打印机中实现的，它不额外占用主机内存和CPU的工作

此为试读，需要完整PDF请访问：www.ertongbook.com

时间。

具有汉字处理功能的计算机，从系统结构来看，就是具有汉字数据类型，它表现为带标志符的数据表示。针对这种数据表示，要设置相应的指令或语句来实现汉字处理操作，如汉字字符串处理、汉字代码的传送、比较、替换、分类、合并、排序、各种逻辑运算操作以及汉字代码的输入输出操作等。

近年来，我国对汉字信息处理技术所做的大量工作，可以概括为以下三个方面。

1．汉字信息处理基础理论

• 汉字图形模型和识别（包括印刷体字、限制性手写体字及一般手写体字）；

• 汉字语音模型和识别（包括语音波形编码和解码、语音的分解和合成）；

• 自然语言的理解和处理；

• 自然语言的机器翻译；

• 自动标引和自动作文摘；

• 汉字用字频度的研究；

• 汉字字根、词组的研究；

• 汉字编码理论及评测标准；

• 人机对话系统理论；

• 汉字信息处理的程序设计语言和软件的理论；

• 数据库和中文语言库的研究等。

2．汉字信息处理系统工程研究

• 将所用汉字以简便快速的方式输入到计算机。有整字键盘输入、半自动编码输入、自动口授语音输入、自动印刷体字输入及手写体字输入等方式；

• 给所用的汉字设计最佳编码（包括输入码、机内码和交换码的优化）以及实现编码间的转换；

• 设计优良的汉字字模信息库（包括载体结构设计、字库生成程序、信息压缩和还原方法等）；

• 实现操作系统和程序语言的中文化（包括解决它跟原有计算机软件的兼容性）；

• 建立汉字数据库管理系统，中文文件管理系统；

• 合理地改造已有的英文数字处理系统，并使之具备汉字信息处理功能；

• 实现汉字信息的本地通信和远程通信；

• 制定各种汉字信息的国家标准，包括信息交换用汉字编码字符集标准、控制文字符号标准、汉字字模点阵标准、汉字编码评测标准、汉字属性信息字典、汉字印刷体字标准等；

• 制定汉字智能终端和汉字信息处理系统的系列型谱等等。

3．汉字信息处理应用研究

• 在企业管理自动化系统和办公室自动化系统等信息管理系统中都要有较全的汉字数据库，力求汉字化；

• 图书情报管理系统要选择简捷的检索方式和海量的存贮器；

• 照像排版系统要具有规格齐全的、字形合乎规范的汉字库；

• 信息服务网要有较快的系统响应；

• 汉字、词典及属性字典的编纂系统要有合适的输入手段及庞大的汉字数据库等。

目前对汉字信息处理技术的研究主要致力于四个方面。

1．汉字编码输入方案的优选和优化

目前我国的汉字输入方法主要是汉字编码输入和整字输入两种。

汉字编码输入方案已有400多种，当前用户采用较多的有十几种。汉字编码输入法是目前采用的主要输入手段。

笔触式整字键盘输入汉字在国内已引起用户的注意，其中压感式和静电耦合式很受欢迎。整字键盘将二三千常用字组成字表矩阵，使汉字编码在用户面前被比较直观的位置信息所取代，进入键盘后由微处理机转换为内码或交换码。这种方法的优点是规则简单，人们乐于接受。但整字键盘不易实现盲打（字数多、键位面积小），平均键入速度不高。

无论是汉字编码输入还是整字键盘输入，速度上都与计算机运行速度不相匹配，成为汉字输入的“瓶颈”。因此，汉字自动识别、汉语语音识别等已成为开辟汉字输入的新技术，现已取得一定的成果。

2．汉字信息处理系统与英文数字处理系统的兼容

据不完全统计，我国已研制了120余种汉字信息处理系统和设备，形成了一定规模的生产能力。但从系统角度上看，构成系统形成系列的较少，大部分是在国产或引进原装微型机上进行改造，使之具有汉字处理功能。而现有计算机系统拥有多种通用的程序设计语言及其编译程序或解释程序，拥有各种成熟的操作系统或监控程序，并且开发了丰富的应用程序，这些都应该充分加以利用、继承。发展汉字信息处理系统，要最大程度地与现有的英文数字处理系统兼容，使它既具有汉字信息处理功能，又符合通用化、标准化原则。

根据计算机系统分层结构的概念，可以在以下四个层次上扩充改造，以获得汉字处理功能。

(1) 应用程序层。在应用程序中增加用汇编语言和系统调用组成的汉字输入输出模块实现。

(2) 高级语言层。在原有高级语言中增加处理汉字的语句，增加专门的处理模块。

(3) 操作系统层。在操作系统的内层增加汉字处理的功能模块，使系统管理下的各种高级语言均可实现汉字处理。

以上三个层次都是以软件方式获得汉字信息处理功能的，它们的硬件未作较大改动。处理功能的强弱决定于附加汉字功能的层次，愈往里层便愈强。目前原有英文数字系统的改造大量采用的是（1）、（2）两种方式。它的问题主要是系统开销大，处理效率不高，需要加以解决。

(4) 硬件层。它是在原系统软硬件基本不动的基础上获得汉字处理功能的一种方法。有代表性的就是插接兼容式汉字信息处理系统（汉字终端、打印机、脱机工作站）。该系统的特点是将汉字处理模块（包括硬字库和编码转换、汉字存贮代码）附加到计算机系统的内核上，从系统的高层上获得很强的汉字处理功能；对不同机型适应性也强，处理效率大为提高，因而日益受到重视。

3．汉字专用设备性能的提高

我国的汉字终端设备已初步分为四个型谱：简易型、基本型、智能型、工作站型。这将为我国汉字终端设备的研制与生产走向系列化、标准化、通用化奠定了基础。

为了适应汉字信息处理的新要求，还需要大力研制各种适合于汉字特点的新型专用设备。例如，文件和图形输入机，手写字读入机，汉语语音识别装置，汉语发音装置，标准字库，汉字输出缩微设备等。

4．汉字字模及图形识别和汉语语音识别

汉字信息处理技术主要用于快速处理大量信息。浩瀚的日益激增的文件（印刷体文字）、文稿（手写体文字）、图形、实时的语音输入（会议讲话、同声翻译）等的输入是不能依赖于人工编码的输入方式，必须采用自动输入中文信息的新手段。

回顾近年来汉字信息处理技术发展的历程和所取得的成果是令人鼓舞的。瞻望汉字信息处理技术的未来，有许多新的课题，有待于探讨和开拓。我们相信通过国内外广大科技人员的努力工作，汉字处理系统必将日臻完善。

1.2 汉字信息处理的特点

世界上的文字有三种，即表意文字、音节文字和音素文字。表意文字单独成为一个体系，音节文字和音素文字合称为拼音文字体系。全世界所有文字符号都可以包括在这两大体系之中。在使用拼音文字的国家里，计算机已扩展到社会生活的各个领域。

拼音文字体系之所以利于计算机处理，主要有两个原因：

(1) 拼音文字的字形信息和语言信息的关系密切。

例如：

英文（26个字母），英语（48个音素），字母与音素基本对应；

俄文（32个字母），俄语（30个音素），字母和音素几乎完全对应；日文（51个假名——汉字除外），日语（70个音素），字母与音素对应。

拼音文字体系的绝大多数单字，可以通过拼音读出对应的字音，根据字音也容易写出对应的字形。

(2) 拼音文字由少数有限的字母图形，自左而右，一个方向，在一维空间里依线性关系顺序排列成单词。无论拼音文字有多少万个词，都可以由几十个字母线性组合而成。

因此，在计算机上比较容易实现拼音文字的输入、输出和信息存贮。

汉语汉字能自立于世界语言之林，经几千年历史和数以亿计的人使用的考验，是有其特点和演变规律的。

汉字属表意文字，是文字发展史中最早出现的一种。此外，古苏末尔人的象形文字，巴比伦钉头字，古埃及文字，古玛雅文字都属于表意文字。由于表意文字和语言的语音之间没有直接的关系，所以汉字可以用来表达不同的民族语言。换句话说，表意文字是适合于民族语言众多的国家。

方块汉字是把形、音、意三个方面有机地结合起来的一种优美文字。每个汉字都有特定的形体，一定的读音并能表示一定的意义，这三者是不可分割的统一整体，常称为文字的三要素，可用图 1-1 表示。

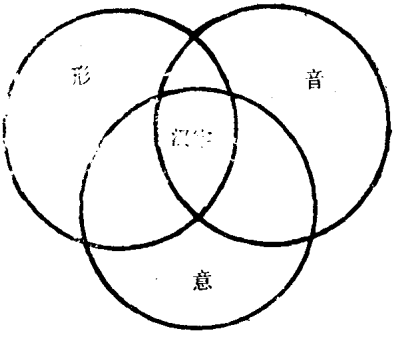

图 1-1 汉字的三要素

实际上，汉字位于形、音、意三集合的公共部份，用逻辑式可表示为：

汉字::={形}AND{音}AND{意}

据统计，形音字约占80%。（逻辑式中::=表示定义，{ }表示集合，AND表示逻辑“与”。）

汉字信息言简义深。要传达同一思想，表达一个概念，

此为试读，需要完整PDF请访问：www.ertongbook.com