汉字信息处理概说

王绪龙 编著

南京大学出版社

汉字信息处理概说

王绪龙　编著

南京大学出版社

1988·南京

内 容 简 介

本书介绍了汉字信息处理技术的基本知识，基本理论和技术实现中的基本方法，包括汉字的有关属性、汉字输入输出、汉字信息处理系统及其典型应用，并对今后发展中的一些具体问题作了探讨。

全书内容深入浅出，通俗易懂，稍具计算机知识的读者即可阅读。可作为高等学校非计算机专业学生学习汉字信息处理课程的教学参考书，也可供语文科学工作者和有关科技人员参考。

汉字信息处理概说

王绪龙 编著

责任编辑 顾其兵

南京大学出版社出版

（南京大学校内）

江苏省新华书店发行 江苏省丹阳第二彩印厂印刷

1988年3月第1版 1988年3月第1次印刷

开本：787×1092 1/32 字数：138000

印张：6.125 印数：1—3000

ISBN 7-305-00142-2

TP·12 定价：1.05元

前　言

现代社会，人们无时无刻不需要信息。信息的生产、管理、传播和分配已经直接影响着经济的发展。信息资源的有效利用是一个国家社会生产力发展水平的明显标志。

语言文字是当今人类社会信息交换的最主要、最高级的媒介。对语言文字信息用计算机进行高效处理，是我国“四化”建设的需要，是全球性信息交往的需要，是全世界信息界极为关注的课题。

我国是汉字(汉语)的发源地，全世界使用汉字的人口占世界总人口的1/4以上。汉语还是联合国法定的六种正式工作语言之一。汉字信息资源极其丰富。计算机能否高效地对汉字信息进行处理，是计算机这一时代的宠儿能否在祖国大地普及、开花结果的至关重要的问题，也是各行各业立志改革创新、面向现代化、面向世界、面向未来的志士仁人普遍关心的问题。

经过多年的努力，我国的汉字信息处理事业蓬勃发展，取得了众多举世瞩目的成绩，部分基础理论和技术方法已在国际上处于领先地位。但是，从语言文字信息处理角度出发，这仅仅是开始，还有更多、更新的课题需要更多的关心这一事业的人们去探索、去开拓、去发展。

汉字信息处理技术是一门综合多个领域现代科学成就的新兴学科，它涉及到统计学、管理学、工程学、心理学、社会学、数学、语言文字学和计算机科学，等等。

笔者结合多年从事汉字信息处理技术的实际工作经验和教学工作经验，参阅了大量国内外有关文献资料和技术成果，立足于基本知识、基本理论和基本方法的介绍，避免冗长的概念模式推演，力图使读者通过阅读本书对汉字信息处理技术有个整体的了解，并通过这一了解，从多个侧面为推动我国汉字信息处理业的全面发展作出贡献。

鲍明炜先生对本书编写提出过许多指导性意见，王俊华同志、雪梅同志、王琦同志等为本书的文稿、图稿作了大量具体工作，顾其兵同志在本书编辑过程中一丝不苟、认真负责，并作了必要的修改，在此一并表示感谢。

因水平有限，错误和不当之处在所难免，恳请读者诸君不吝斧正。

编 著 者

1987 年 10 月于南京大学

目　录

第一章 绪 论

1.1 研究汉字信息处理技术的意义

当今的人类社会，除了物质、能源以外，信息也成了重要的社会资源。信息的生产、管理、传播和分配直接影响着经济的发展，是社会生产力发展水平的明显标志。

语言文字是当今人类社会信息交换的最主要、最高级的媒介(当然也包括图象、造型等)。随着科学的进步，用文字记载的信息量日益增多。我们的前人曾经用"汗牛充栋"这句话来形容过书本、文献之多，可是到了今天，这句话已远不能表达出人们的惊叹心情。今天的社会信息量已远不是几条牛能拉得走、几间房能容得下了，事实上已经形成了所谓"爆炸性"局面。仅以科技资料为例，目前的年产量就达一千万件以上，而且每年正以5%—10%的速度递增。一个化学家要是想浏览一下一年内的化学论文和著作，那么将把其毕生的精力耗尽。信息是客观存在的，信息量的急速膨胀也是客观存在的，我们正是处在这样的人类社会中。

"汪洋大海"般的信息量，是人类社会的无限宝藏。但是，这并不是说，所有社会信息对任何人、在任何时间、任何场合都能发挥效用的。要把它作为资源加以利用，还必须对它进行加工和处理，才能在特定情况下充分发挥其资源效益。一般说，这些加工和处理主要包括对信息进行识别、统计、

分类、存储、比较、推理、检索、转换、传输、控制和模拟等等。过去，这些加工和处理，主要是人们靠手工进行的，效率低、速度慢。只是当电子计算机问世以后，信息处理才有了崭新的有力工具，使信息处理技术取得了突飞猛进的发展。

今天，以电子计算机为中心的信息处理业已在世界范围内形成了巨大的产业部门。在美国，1984 年信息处理业的产值为 1063 亿美元，超过了汽车业，而成为美国仅次于石油业的第二大产业。在日本、欧洲以及其他一些国家和地区，也都已经或正在上升为国民经济中的显赫部门。据统计，到 1984 年，全世界各类计算机已有 4000 万台，这些计算机中的90%是服务于非典型科学计算的信息处理业的。它的应用已遍及人类社会生产和生活的各个方面。例如：经济规划、军事指挥、企业管理、过程控制、统计分析、情报检索、编辑排版、自动翻译、语言研究、智能模拟、辅助教学、金融保险、自动销售、旅游服务、文艺创作、医疗卫生、体育竞技、办公事务、家庭事务、个人事务等等，不胜枚举。在所有这些领域都取得了巨大成就，获得了极大的经济效益和社会效益。单就信息处理作业量而言，1981 年美国拥有各种计算机 200 万台(到 1984 年美国已有各种计算机 1400 万台)，而它一年所完成的总作业量，相当于 400 亿人一年的总工作量，或者说，相当于美国二亿人口 200 年的总工作量。这样的效益是何等可观！

以计算机为中心的信息处理技术，可以大量地节省时间、精力和费用，可以有效地提高工作效率和管理效率。而一切信息处理的一个最明显、最主要的特征，则是用计算机这一强有力的工具，对语言文字(也包括图形)信息进行加工和处

理，使语言文字信息得到最佳利用，使凝聚在语言文字媒介中的信息发挥出最大效能。这正是我国急需大力研究开发汉字信息处理技术的原因所在。

用计算机对汉字信息进行处理的一般过程是：把要处理的事项(如，处理什么、怎样处理等)，用文字、数字和一些规定的符号等表示出来，通过适当的方式送进计算机(即所谓输入)，计算机被启动后，便根据已经输入的人的指示，高速地进行大量分析、比较、计算、判断，直至得到合乎人们需要的结果(即所谓处理)，这些处理结果再用文字、数字和符号表示，并自动印刷在纸上，或显示在荧光屏上，或记录在磁盘、磁带上(即所谓输出)，供人们直接选用或再次送进计算机进行处理。事实上，输入—处理—输出这样的流程，不仅对方块汉字，对所有语言文字信息处理都是一样的。而且在计算机内部进行处理时，直接处理的对象都已变换成二进制代码的形式，即用一系列的"0"和"1"的组合形式来表示千变万化、千差万别的各种事项，并对它们进行处理。因此，从本质上讲，不论哪种语言文字信息，从计算机处理角度来看，应该是一致的。这就从根本上对"汉字信息能否用计算机处理"作出了肯定的回答。

但是，在实际的以计算机为中心的文字信息处理技术中，拼音文字(姑且以英语为例)和汉字确实有着明显的差别。概括地说，这种差别主要表现在两个方面：

其一，从计算机方面看，目前计算机的终端输入键盘，基本上沿用了在英语社会早已普及的英文打字机结构形式，用这种键盘打键输入是目前文字信息处理的最主要输入手段，它是人和计算机联系的最直接的设备。因而，对英语圈的人来说，学习操作和使用计算机要方便得多。中文(汉字)

打字机的基本结构形式也可作为计算机的终端设备使用，但除非经过专门训练，一般人是不容易熟练掌握的。另外，目前计算机程序设计语言及其命令系统是以英语为基础发展起来的，对于熟悉英语的人比较直观、易于掌握。计算机能直接接受英语信息，处理也较方便，对汉字来讲则要困难得多。

其二，从语言文字来看，英语是拼音文字，独立使用的字种是字母，总数为26个，加上大小写字母则为52个。汉语中独立使用的字种是汉字，总数高达6万，常用的也有8000个上下，同时汉字的字形也要比英文字母复杂得多。另外，适合于计算机处理的英语语法系统研究得已较为深入，它是否适合汉语？怎样才能使汉语系统与计算机处理协调起来？都还有待于多种学科的志士仁人深入进行探讨和研究。

英语和汉语的这种差别，至少在目前使得汉字的输入、存储、转换、加工和输出等等都要比英语复杂得多、困难得多。也许正是这些差别，成了汉字信息处理技术发展缓慢的主要原因。

但是，我国是汉字的故乡。汉字是当今世界各国使用文字中最古老的文字，已有3000多年的历史。它在我国以及世界其他一些国家的发展史上建树了不可磨灭的功勋，记载了社会、政治、文艺、经济、技术等等大量的光辉文献。这些文献是先人留给我们的无价之宝，它的发掘和处理必然对我国文化发展产生积极影响。

汉字也是世界上使用人数最多的文字，除我国（包括台湾、港澳地区）外，还有亚洲的一些国家（如新加坡等）及海外华侨也使用汉字，日本和南朝鲜也是使用汉字的（读音和字意与我国汉字有所不同）。就使用汉字人口数而言，超过了世界总人口的四分之一。汉语还是联合国法定的六种正式语言和

工作语言之一。近年来，世界各国学习汉语的人数日益增多，在美国和欧洲的一些学校里，有的已把汉语作为第二外语列入了教学计划。无疑地，加速汉字信息处理技术的研究和开发，对加强我国国际地位、发展国际交往，将有着深远的意义。

目前我国正进行着现代化建设的伟大事业，汉字信息处理技术有着极为繁重的任务。试看：我国现有30个省、直辖市、自治区，353个市，2000多个县，它们的纵向指导和汇报，横向交流和促进，便构成了一个庞大的汉字信息网。我国有40余万个大中小企业和蓬勃发展着的乡镇企业，它们的管理、生产、销售、发展和竞争，都和高效地处理汉字信息有着直接的联系。我国有近500种报纸、4000种杂志、400家出版社，高效地出版印刷以及高效地对这些汉字信息进行加工处理并转化为资源加以充分利用，是汉字信息处理技术的当务之急。为加强国际间的科技合作和文化交流，汉语和外语间的自动翻译十分急迫。汉字通讯的落后状况急需改变，能适应新技术革命需要的数以亿计的人材要培养，要让开始普及的计算机适合我国社会特点，在我国现代化建设中被尽可能多的人掌握，彻底改变目前我国计算机还很少的情况下，竟有70%以上的计算机不能发挥作用的不正常状况，加速发展汉字信息处理技术就更显得十分重要了。

1.2 我国汉字信息处理技术发展的简单回顾

回溯起来，我国汉字信息处理技术的研究和我国的计算机历史差不多一样长。1958年，在我国诞生了第一台电子管

计算机——103型电子计算机，它的运算速度为每秒钟1000次。次年研制成功了每秒钟运算1万次的104型电子计算机。这一成就成了我国计算机科学技术发展的奠基石。差不多与此同时，在我国就已开展了用计算机进行由俄语到汉语自动翻译的研究工作。这是我国最早的用计算机对语言文字信息进行处理的研究。当时因为还没有能与计算机相联的汉字印字设备，译出的汉语是用拼音字母形式输出的。50年代开始的用计算机进行自动翻译的研究，不仅在中国，在世界许多国家，如美国、法国、加拿大、苏联、英国、联邦德国、捷克斯洛伐克、意大利、荷兰、日本等十多个国家，都曾掀起过一股热潮，持续了10年左右。这些翻译研究的对象语言有英语、德语、俄语、法语、日语等，也有我国的汉语。但是，受限于当时的计算机技术水平，主要是计算机的存储容量有限，运算速度较慢，同时对语言本身的分析研究和适合于计算机处理的语言模式研究不够，以致一场世界范围的轰轰烈烈的机器翻译研究，在这一时期内，未能结出理想之果。但是，在我国的汉字信息处理研究史上写下了光辉的一页，汉字、汉语已开始和计算机结下了不解之缘。

1969年，汉字电报译码机在我国电报业务中开始使用。这种译码机从通讯电路接收到由发报局发来的汉字电报代码，自动转换成该代码所对应的汉字字形点阵代码，驱动印字机构在纸上打印出以点阵方式组成的汉字，取代了收报局将收到的电报代码用人工进行译码和抄写的作业，有效地提高了电报通讯的效率和可靠性。电报译码机的这种使用方法，也是目前大量使用的点阵式汉字打印机的基本方法，事实上，70年代我国研制的某些汉字信息处理系统的印字输出设备，就曾采用过电报译码机的原型。

70年代伊始，我国已日益明显地面临着国际信息化社会的严峻考验，我国汉字信息处理业再也不能缓缓而行了。在一批立志开发我国汉字信息处理事业有识之士的倡导下，在国家6个有关部、委的领导与资助下，经过一段时间的准备，于1974年制订和组织了我国第一个大型的汉字信息处理工程项目——748工程。分别在汉字情报检索、汉字照相排版印刷、汉字远程通讯等诸方面开始了艰苦努力。与此密切相关，还打破了闭关自守、故步自封的不正常状况，从发展我国汉字信息处理技术的需要出发，首次由国外引进了一套汉字信息处理系统设备。该设备是以小型计算机为前处理机的高速汉字印刷系统。机内常存汉字4096个，在软件支撑下，这些汉字的一部分或者全部都可任意选择，给系统能处理的汉字字种数增加了灵活性。脱机的输入设备采用整字式汉字输入键盘，记录在纸带或磁带上。印刷汉字采用光导纤维管的电子照相方式，每秒钟可印2000个汉字。748工程历经数年，涌现了一批人才，取得了一批成果，部分成果取得了国际公认的先进水平，也探索到了该领域的一些问题和难点。这一大型工程项目吹响了我国汉字信息处理开始向深度和广度发展的进军号。

70年代末期以来，我国汉字信息处理业的蓬勃发展，除了我国社会环境发生了历史性变化等诸多决定性因素以外，各种专业学术团体的纷纷成立及其学术活动，无疑是一股巨大的推动力量。1978年全国汉字编码研究会成立，1980年中国仪器仪表学会汉字信息处理系统研究会成立，这一年中国电子学会计算机委员会成立了汉字信息处理专业组，1981年中国中文信息研究会成立，其后全国各地不断成立了许多地区性的专业学术团体。这些学术团体在全国范围内组织了大

量的学术研究活动和学术交流活动，组织和参加了许多国际学术交流活动。仅以中国中文信息研究会为例，自成立起至1985年初，该会就组织召开了19次全国性学术交流会议，交流论文达1000多篇，1983年还在北京和联合国教科文组织联合召开了“中文信息处理国际研讨会”。这些学术团体以及它们成立前的筹备组织还参加1980年在香港召开的“中文资料与文稿处理国际电算机学术会议”，1982年在华盛顿召开的“中文计算机学会国际会议”，1983年在东京召开的“1983年大字符集文本处理国际会议”，1983年参与组织在上海召开的“多国仪器仪表展览会”，以及更多的双边交流和考察的国际的学术界和产业界活动。这些活动为使我国汉字信息处理事业面向现代化、面向世界、面向未来作出了重大贡献。

伴随着80年代以来微型计算机在我国的拥有量急剧增长，微型机汉字信息系统和汉字终端系统吸引了广大研究开发人员、广大生产厂家和广大用户。据不完全统计，至1985年初，我国已有近百个单位研制生产了120余种汉字信息处理系统和设备，已形成相当规模的生产能力。几十种微型计算机和某些大型计算机实现了中英文的兼容处理，改造已有的微型计算机使之能处理汉字信息的成绩十分显著，在多种汉字卡的设计、汉字操作系统、汉字终端、汉字通讯以及研制汉字打印机技术等方面，达到了较先进的水平。远程汉字信息处理联机网络系统也于1984年在我国投入运行，数百字的电文可以在不到一分钟的时间内准确无误地显示在几千公里以外的接收终端上。

80年代以来，在多种学科的渗透下，汉字信息处理领域的各项研究都取得了许多可喜的进展。汉字整字输入键盘技

术不断更新。汉字编码输入方案与设想层出不穷，这些方案与设想已超过400种，40余种在计算机上作过验证，10余种已为众多用户所采用。汉字字形识别的研究已进入了实验研究的中、后期，已公布的实验识别率不论在印刷体汉字还是在手写体汉字方面，均已超过90%，部分成果的识别率已高达98%以上，多种字体识别系统的识别率也已在90%以上，正面临着向商品化过渡的进程。汉字语音识别的研究，在单呼语言的识别方面，识别率已达99.5%，连呼语言的识别也开始了实验研究的进程。在汉字情报检索方面，有数十个单位和部门开展着踏实的研究，有的已开始了定题情报服务的应用阶段。汉语自然语言理解的研究和语音合成技术取得了很大进展。汉字照排印刷系统中的某些核心环节，目前已达到国际公认的先进水平。机器自动翻译方面，据不完全统计，我国目前已有俄汉、英汉、日汉、德汉等10余个实验型翻译系统。20多所大学还研制了30多种计算机辅助教学系统，计算机辅助汉语教学系统，对来我国留学的各国留学生学习汉语将愈益发挥更大的作用。利用汉字信息处理系统对我国的古典文学著作和现代文学著作的研究、对诗词创作的研究、对语言文字学的计量研究、对汉语语言模式的研究等等都取得了大量可喜的成果。

1983年底我国已安装大、中、小型计算机约4千台，微型计算机2万多台。1984年底，各类计算机猛增到近20万台(其中绝大多数是微型机)。这些计算机除了分布在第一产业(即农、林、牧、渔等业)和第二产业(即制造业、采掘业、建筑业等)外，也为行业很多、范围很广的第三产业部门(即流通部门、为生产和生活服务的部门、为提高科学文化水平和居民素质服务的部门、为社会公共需要服务的部门等，根

据国家统计局规定，军队和警察也是第三产业部门)所大量拥有。在第一和第二产业部门中的管理决策机构也有大量计算机。据有关资料分析，在这些部门的信息处理作业大体可分为两个方面，即数字信息计算和文字信息处理，而70%以上属于后者。它成了近年来微型机汉字信息系统研究开发的一股巨大推动力量。办公自动化系统的开发已纳入国家计划。汉字数据库、汉字数据库管理系统、汉字信息联机网络等已开始进入有效的应用阶段。随着"办公自动化"计划的开展，一批重点技术(如计算机技术、接口技术、数据库系统技术、通讯技术、网络技术、系统工程、人机工程、光电子技术、文字处理技术等)和重点设备(如激光印字机、光盘、光电转换、扫描设备、程控电话交换机、轻印刷系统等)将获得更快的发展，一些重要的技术基础，如统一编码标准，制定文件格式、数据和报表格式的标准、办公业务流程的规范化等，也将取得更快的进展。一种适合我国国情的，具有中国特点的，多层次办公自动化系统将扎根于中国大地。

汉字信息处理技术是一门新兴学科，经过多年的努力，我们已取得了举世瞩目的许多成绩，同时也有着许多更新的课题，有待于更多的人去探讨和开拓。我们相信，汉字信息处理技术广泛地进入我国社会生产和社会生活的时代已为期不远了。