中文信息的计算机处理

张普董振东 张强生著

Fundamentals of Chinese Language Computing

中文信息的计算机处理

张寿萱　徐建毅　张建生　著

1984年

内 容 简 介

本书是第一本完全由电子计算机进行信息处理，而正式出版的中文书籍。本书介绍了如何由电子计算机作为工具，进行中文信息自动处理的全过程。本书全面地、系统地介绍了中文信息处理技术的概况、汉字结构、汉字数据结构、代码映照方法、输入方法和输入设备、输出方法和输出设备、系统设计和实现、中文数据库和情报检索、语言理解和计算机文化、汉字的计算机识别等内容。本书收入许多中文信息处理系统的具体方法，易于读者移植到具体计算机上。本书内容理论联系实际，既适合于读者泛读，以了解中文信息处理技术的概念和原理；也适合于读者精读，以系统掌握中文信息处理技术的细节和技巧。

本书既可供从事中文信息研究、计算机科学、工程及应用、企业管理现代化的科技人员阅读，也可作高等院校有关专业的教学参考书。

中 文 信 息 的 计 算 机 处 理

张寿萱　徐建毅　张建生　著

*

学报出版社出版

上海印刷技术研究所激光照排实验室排版

上海第七印刷厂印刷

新华书店北京发行所发行　各地新华书店经销

*

开本：787×1092　1/16　　印张：19　　字数：460 千字

1984 年 10 月第一版第一次印刷　印数：布面精装 1—10000 册
平　　装 1—15000 册

统一书号：15244—0010

定价 精装 6.40 元
平装 4.70 元

这是第一本完全由电子计算机进行信息处理，而正式出版的中文书籍。本书介绍了如何由电子计算机作为工具，进行中文信息自动处理的全过程。

在本书编写、审校、制版印刷等过程中，使用了以下三个系统：

1. 中国船舶工业总公司应用软件开发中心的 System C 中文信息综合处理系统。(在 WANG VS-100 电子计算机上运行)

2. 上海仪器仪表研究所与上海电表厂的 Olympia 1011 汉字自动打字机系统。

3. 上海印刷技术研究所的中文计算机激光照排系统。

《序》

人类历史上最古老的文字开始于象形体，汉字仅是其中之一，还有苏木尔楔形字，古埃及文字，但是这些文字在世界上早已绝迹，唯独汉字几千年来在我国盛行不衰，不因社会体制和地区语言的不同而变化，仍然一直是我国统一的文字，始终是我国人民最通用的记载信息的工具，它能适应我国不同时代文风和词句的变革，记录了我国丰富多采的文学创作和诸子百家的学术思想。它对保留中华文化遗产做出了宝贵贡献。今天使用汉字的人口越来越多，地区越来越广。据统计目前世界上约有四分之一的人口使用汉字，汉语也是联合国五种通用文字之一。因此可以说，汉字是一种生命力最强，使用率最高，适用性最广的文字。

在人类社会中，信息是以文字、语言、声音等形式出现的，现在国外拼音文字国家早已实现了应用计算机进行文字信息处理。古老的汉字是否能适应当前新形势的挑战，对汉字是一个严峻考验，也是举世拭目以待的一个谜。自然引起了中外科学工作者极大兴趣，而争相研究与探讨。我国对中文信息处理的研究起步较迟，1975年对汉字信息化才开始有了研究成果的报告，但此后发展很快，研究的项目也很广泛，如汉字编码、输入终端、汉字库的压缩信息、输出装置、处理系统等等，并已取得突出的成果。但是从单独的，个别的成果中很难一窥全豹难以断言汉字已能和拼音文字比美，完全能适应信息时代的需要。

但是当这本书的作者坐在计算机屏幕前，边思考，边键入，现编，现改。由计算机进行自动排版、自动分页、自动产生表格，最后制版印刷，装订成册，正如他在美国时经常用计算机编写英文稿件一样，但是不一样的，也是值得骄傲的是他键入的是古老的汉字，他已成功地应用计算机，实现了中文处理系统，能做到和拼音文字一样的处理过程，得到同样或者更好的效果。他不仅进行了大量实践，而且还把整个实践过程，设计思想，基本原理，汉字数据结构，采用的硬件和软件的成果，编写成书，详细介绍，公诸于世。这一既有实践，又有理论基础的成果，有足够的根据揭开了这一世界之谜，证实了古老汉字的生命力是旺盛的，完全能适应当前新时代的挑战，承担我国未来的信息社会的新使命。

当然这一可贵成果的取得，是与我国海峡两岸汉字信息专家们的努力所获得的杰出成果分不开的，也与海外华裔汉字信息专家们的贡献分不开。本书的作者结合现有的诸家成就集其大成，成功地组成已能实际使用的系统。这说明作者具有丰富的专业理论基础和熟练的操作经验，这将是读者们阅读这本书后所能体会的。但更重要的是作者对科学事业的正确处理方法和高度的学术修养，他能真诚地承认人家的学术成果，诚恳地容纳人家的成果，把人家的成果推到实用阶段，同时也加强自己的成果在实践中的生命力。好比一个有成就的文艺家，只有在精通文学经典，具有艺术的修养，才能写出美好诗篇，这就是我对这一本书的粗浅的评价。

古老汉字能置身于信息社会不仅是依靠人为的加工，主要是汉字本身的素质和结构的直接关系。汉字结构是以笔划为基础的字元有规则的组合，很类似拼音字的字母，字母易于转换成信息，同样字元也就可以转换成信息。如果人们对汉字结构较有认识，就不难理解汉

字本身是具有规律性的自然信息，如果我们能简易识别汉字固有信息，那就不需要象电报码死记硬背，而做到见字识码。

汉字是象形字，但与图象信息有别，图象信息不是汉字信息，用处理图象信息的方法来处理汉字信息是杀鸡用牛刀，造成信息通道的大比率浪费，难度大，用途也有局限性。对文字信息的处理而言，键入的码不仅是字码，而且有功能码。普遍用于计算机输入终端，一律用代码键入。中外一致，使用方便，这是我的肤浅之见。

汉字码与拼音文字码比较还有下列几个优点：

1．码短，汉字码目前一般平均不大于四个字母，而拼音字则较长，英文平均字长为 6.7 个字母。码短，打字输入的速度就较快。

2．中文的造句比西文精炼紧凑含义深广，更没有画蛇添足的文法语标，一页的英文稿件，翻译成中文往往只有半页。数据库文件信息的存贮量就可以压缩一半。

3．今天中文的发展趋势是常用汉字减少了，而词汇和专业名词术语大大地丰富了，表达的宏观和微观能力更强了，用汉字固有的信息组成的词汇的信息可以更加压缩，现在一篇文件平均字的码长缩短到 2.5 个字母，打字最高速度每分钟可达到 100—120 个汉字。无怪作者习惯用英文键入，而惊呼汉字键入特快。确实是有道理的。

在此值得一提的是我国文字改革工作对汉字信息化也有很大的贡献，简化汉字和汉语拼音，对我们今天能易于识别汉字固有信息，提供了可能与条件。因此我想如果今后在简化汉字工作上能考虑到汉字信息化的要求，则汉字固有信息可以大大减少重码。

为迎接信息时代的到来，为实现我国四个现代化的大业，信息交换、文件的存贮与检索、情报的提供、企业管理、数据的建立等等汉字信息处理系统，必将广泛地开展。本书适时完成出版，提供专业的需要数据，应能起一个开山劈路，抛砖引玉的作用。其中值得提出的是全国有不少专家也有丰硕成果，但贵在迅速实践也要取长补短，相互配合，在高水平上接力作战，而不应在同一水平上重复劳动。取得成果也要及时写成专著，相互交流，我相信我国信息处理系统必将百花齐放，百家争鸣，广泛地应用在各行各业。如果通过网络建立国家数据库把信息送到祖国各地，改善我们工作方法，管理体制，社会生活习惯，缩短距离，节约时间，扩大视野，发展智能，促进我们早日进入新的时代。

支秉彝

1984.6.18

作者的话

一九八二年五月至八月，我们在 WANG VS-100 超级小型计算机上完成了 System C 中文信息综合处理系统以后，曾用 System C 系统，非正式地出版了几本介绍该系统设计思想的论文集。这些论文都是我们自己坐在计算机终端屏幕前，边思考，边键入，现编现改，由计算机进行自动排版，自动分页，自动产生表格，最后打印输出，装订成册的。

一九八三年五月，宇航出版社的廖寿琪同志建议我们将这几本小册子综合编辑成一本专著，以向广大读者介绍中文信息如何由电子计算机进行自动处理，这样一个很有意义的过程。

一九八三年十月，在“中文信息北京国际研讨会”期间，我们看到有那么多国内外的专家和学者正在从事中文信息处理技术的研究，并取得了那么多的丰硕成果，很受鼓舞。正如“中国中文信息研究会理事长”钱伟长教授在开幕词中所指出的：“中文信息处理技术及汉字计算机，作为一个新兴学科在世界各国迅速发展起来了”。在此会议期间，廖寿琪同志再次提起本书创作计划。几经磋商以后，一九八三年十二月，我们在屏幕终端上按下了第一个键，开始了这本书的编写工作。

一九八三年，上海印刷技术研究所激光照排实验室成功地完成了中文计算机激光照排系统，通过了鉴定。在这个系统的支持下，也已经试排出版了几本小册子和二十万字以上的会议论文集。一九八四年一月，张佑陵同志建议我们搞一次科研合作，直接将在 System C 系统上完成的稿件，通过软盘交换方式，送入激光照排系统进行激光照排制版。这是一项很有意义的尝试，也使得我们这本书有可能成为第一本完全由电子计算机进行中文信息处理，而正式出版的中文书籍。在双方领导的热情支持下，签定了科研合作协议书，保证了这项工作的顺利完成。如果没有张佑陵同志和她的同事们支持和帮助，这本书是不可能在这样短的时间内就与读者见面的。

一九八三年，我们在王安机组的张靖新、尹明康等同志和上海电表厂赵传勋等同志的合作协助下，完成了 WANG VS-100, WANG MVP, OLYMPIA 1011 这三个中文系统之间的联机通讯试验，构成了一个小小的分布式中文信息处理系统。这样，WANG VS-100 上的 System C 系统就可以将全部信息分批送往 OLYMPIA 1011 系统排版处理和喷墨打印输出。相反地，在 OLYMPIA 1011 系统上输入的中文资料也可以方便地送入 System C 系统进行再处理。由于采用了支秉彝博士的“见字识码”输入方法，大大地提高了输入效率和速度，也保证了本书的创作和审校工作的顺利完成。

这样，我们也就对从写作、改稿、审校、发稿、编辑、排版、插图、制版直到印刷，这样一本书诞生的全过程都完全采用了中文计算机系统来进行中文信息的自动处理。从键入第一个字开始起，直到装订成册，上市发行的整个周期为十一个月。显然，如果没有中文计算机系统这样强有力的工具，这种高效率是无法实现的。我们向所有为本书的出版和为开发这几个系统作出贡献的专家和同行们致以诚致的敬意和衷心感谢！

当前，国际上正在酝酿着一次新的产业革命，开始了一场从工业社会向信息社会的革命

性转移。我们炎黄子孙能不能有效地用我们自己国家和民族的文字进行信息处理，进入"机器信息世界"之门，来迎接这一场现代文明浪潮的挑战，是关系到我们国家和民族兴亡的大问题。时代赋予新中国的计算机科学家们光荣而艰巨的使命，振兴中华需要我们加倍地努力，我们将本书敬奉给读者，希望用实际行动来打破至今仍然在一些人头脑中存在的，对"中文能不能适应计算机处理"的怀疑；我们希望用实际行动来建立中文信息处理系统的信誉，并以此来鼓励各行各业的同胞们增长使用中文信息处理的信心和兴趣。坚冰已经打碎，航船已经开通，可以相信中文信息处理技术研究和应用的春天已经到来。

我们非常荣幸和感激的是，中国仪器仪表学会汉字信息处理系统研究会付理事长支秉彝博士欣然答应为本书写了序言。在这二年工作中，支秉彝博士曾给予我们许多指导和帮助，使我们很受鼓舞。

我们也要感谢我们学生时代的老师、上海自动化学会名誉理事长胡汝鼎教授为本书题写了书名。早在十年前，我们就曾聆听过当时作为我们母校副教务长的胡老先生的学术教导。

我们还要感谢几位作为我们前辈的老师：美国 Purdue 大学付京孙教授，美国 Princeton 大学刘必治教授、罗无念教授，中国科学院计算技术研究所唐稚松研究员，他们在理论上曾经给我们许多方向性的指导，在具体实现方法上给予我们许多启发性的建议。我们还要感谢中国科学院、电子工业部的一些老一辈科学家、学会的几位理事长、秘书长和许多同行们对我们工作的鼓励和帮助。

还要感谢下列同行们的帮助和建议。在本书的写作过程中，在与王景寅副研究员、姚天顺副教授和朱邦复先生的多次学术讨论中，我们得到了许多有启发性的观点和见解，也充实了本书的内容。在我们的研究工作中，中国仪器仪表学会汉字信息处理系统研究会刘建国同志曾给予很多帮助，我们的同事彭仲林和周钦铭同志认真阅读和校对了全书的打印稿，提出了不少改进意见，在此谨表诚挚谢意。本书关于激光照排系统和激光打印机部份的内容是根据张佑陵同志提供的资料和观点撰写的。关于显示输出系统部份的内容是根据徐加明和明璐琳同志所提供的资料和草图撰写的。关于输入设备方面的内容曾经与张清才同志作过多次有意义的学术讨论。在全文种情报检索系统的建立过程中，孔陶华、秦美华、杨振铎 、陈鸿嘉等同志做了大量的工作，其中一部份程序是戎引、周钦铭同志的工作。感谢支爱悌教授、李进祺、张明等同志协助完成了本书精装本序言的翻译工作。

全书的图例是在田在华同志帮助下完成的。

最后，但显然不是最微薄的，要感谢宇航出版社同志在审稿、编辑、出版方面作的努力。没有他们的帮助支持，本书也是不可能在如此短时间内完成的。

目　录

第一章　概　　述

1.1 什么是“中文计算机”？

目前，电子计算机的应用已超出了科学计算的范畴，“计算机”已经不再是一种单纯计算的“机器”，而是一种加工“信息”的“自动机”；是一种“电脑”；是人们大脑思维活动的模拟、延伸与再创造。以计算机为标志的“信息革命”时代已经开始。仅有不到四十年历史的计算机，正在刺激着一门已有四千多年文明历史的古老语言，目前仍为世界上四分之一以上人口交流思维信息所用的语言 —— “中文”，为适应信息革命的时代要求而提出了新问题，有了新发展。现在，已经有越来越多的人们重视这个问题，从各个有关学科，各个领域，各个行业作为出发点在讨论“计算机中文化与中文计算机化”这个大问题。毫无疑问，我们国家的现代化，离不开计算机的中文化与中文的计算机化，离不开“中文计算机”。

什么是我们所说的“中文计算机”呢？为什么要研究“中文计算机”呢？怎样开发“中文计算机”及其应用系统呢？我们企图以如下一件小事引起读者的直观感觉和思考兴趣，并且一起来讨论如何回答这些问题。

作者曾有过这样一次经历，这是 1980 年的某一天，在一个美国教授的办公室里，作者与他正在讨论一个学术问题，大家都很有启发，互有心得。忽然，这位教授坦然地说：“请等一等”。于是，他转过椅子，坐到放在他办公室一角的计算机终端屏幕前，熟练而迅速地从键盘上打入了一批信息内容。经由计算机处理排版后，去机房取来了一式二份关于刚才讨论内容的备忘录。约两天后，以这份备忘录为基础的一篇学术论文，已经通过计算机的处理而完

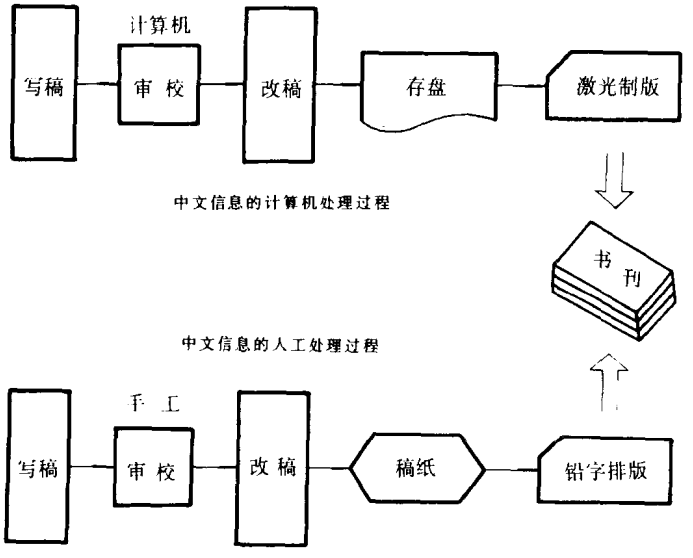

图 1.1　中文信息的计算机处理过程与人工处理过程的比较

成了。其效率之高，使有些教授达到了有可能一星期出三篇论文的水平！

不久，我们也习惯了使用电子计算机这样的工具写英文论文了。坐在绿色的终端屏幕前，通过全屏幕编辑系统，边思考、边按键、边写边改，几易其稿之后，一篇篇数万字的学术论文的原稿就完成了。经过运行自动页面排版系统，将写好的论文原稿由电子计算机进行自动排版后，交给印字机分页打印输出，就得到了一本完全符合装订要求的全文了。当然，也可以为每篇论文取一个合适的文件名字，作为计算机的文件存入系统的磁盘、磁带中。对于需要出版发行的论文，则可以将其文件名字交出版部门，由出版工作者自行到电子计算机系统中，按名找到记录该论文信息的计算机文件，送往激光照排系统加工。几天以后，就可以得到一本本装订美观的书籍、论文和技术报告了。

当时，我们远离重洋，身有重任，肩负着振兴中华的光荣使命，非常羡慕和欣赏在计算机上写论文的先进技术。决心汲取人家先进科学和技术，并渴望着有一天，能在祖国故乡，用中国文字在计算机上写论文。现在，敬奉给读者的这本书，正是通过许多中国计算机科学工作者的努力，实践了这一过程的硕果。在写作、审校、排版印刷过程中，我们使用了如图 1.2所示的三个建立在“中文计算机”上的“中文信息处理系统”。

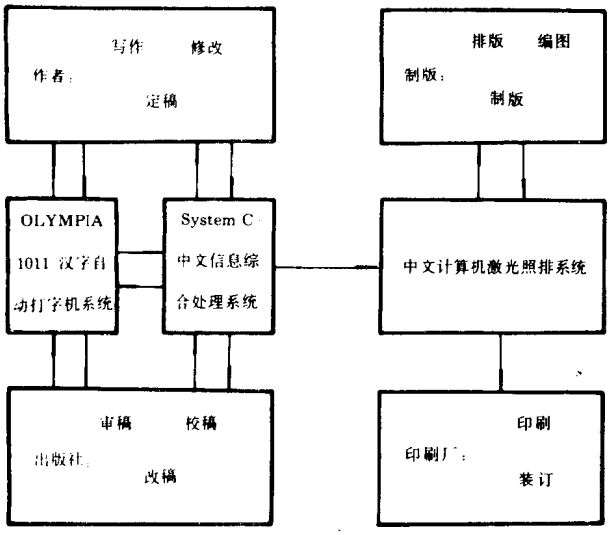

图 1.2 本书出版使用的三个中文计算机系统

当然，这只不过是计算机“字处理”功能的一个小小的应用。从这样一件普通的小例子中，可以看出，要真正达到普及、推广和应用，至少可以归结出如下四条要求：

1. 计算机要多，要普及，要价廉物美，要随手可得。

由于现在的中小型计算机一般都可带几台，几十台，以至上百台以上的屏幕终端。特别是随着微型计算机日新月异的发展，计算机系统的普及情况必定会越来越好。在系统的概念下，用户总是通过工作站来使用计算机的。所以更确切的说法是：工作站要多，要普及，价格要相对便宜，才能随手可得，并进入每个企业、每个办公室甚至每个家庭。

2. 要人人会用，要使用简单、方便。

配备专业操作人员的工作方式在有些场合是必要的，但是，要使任何人都能熟练应用，才是普及的真正目标。

3．满足了上述二点要求，就有可能使人们随时随地，即席地将他们的思维活动信息化。也就是说，人们有可能不再需要过分地依赖笔和纸，也不按传统的方式，用笔将思维活动记录在纸上，而是直接坐在工作站旁，直接将思维活动输入计算机，想到什么就可以做什么。

4．由于计算机的诞生，一开始是与西方语言，特别是英文联系在一起的。而且，计算机市场上的通用产品，主要都是"西文计算机"，其输入键盘是与西文打字机键盘完全兼容的。因此在西文社会中，其应用普及就有十分广泛的基础。如果将这些通用计算机称为"西文计算机"的话，那末，在我们中国经济建设中，就不得不开发一种新型的计算机，则称为"中文计算机"。

定义：具备处理中文信息能力的一类计算机，称为中文计算机。

这个定义很简练，一目了然，但是也有"毛病"。因为还没有定义什么是"中文信息"。由于我们的祖国是一个多种民族的国家，中国人民除了使用汉字以外，还有蒙古文、藏文、朝鲜文等等多民族语言文字。特别是我们国家，多年来还推广了汉语拼音文字系统。如果说用汉语拼音文字描述的信息也是"中文信息"，那末传统的西文计算机，毫无疑问也就具备了处理"中文信息"的能力。语言是思维的外壳，文字是表达语言的符号系统。随着各种文字系统依据的基本符号和系统构成原则的差异，世界各国所采用的文字可划分为二大系统。通常所说的"西文"是指拼音文字系统，因此"西文计算机"就是适合处理西文信息的电子计算机。而"中文"是一种拼形文字系统，汉字是一种有代表性的拼形文字。一般地，我们说"中文信息"，就是指以拼形文字为表现形式的系统传输和处理对象，"中文信息"就是"汉字信息"。由于这一类计算机，不可能如传统的"西文计算机"那样建立在一个几十个字母的小字符集上，因此也有人称这一类计算机为"大字符集计算机"(computer with a large character set)。

"中文计算机"(Chinese Language Computer)亦称为"中文电脑"，或者称为"面向中文的计算机"(Chinese—Oriented Computer)。由于这类计算机具备处理中文信息的能力，因此，在这类计算机上开发的系统称为中文计算机系统 (Chinese Language Computer System)，或"中文信息处理系统" (Chinese Information Processing System)，或"汉字信息处理系统" (Chinese Character Information Processing System)。

目前，国外有一种将信息处理技术从"计算机科学"(Computer Science) 中分离出来，建立"计算科学"，或"信息处理科学"的趋势。其英文译名为 (Computing Science)，故在本书的英译名中已将"中文信息的计算机处理"译为"Chinese language Computing"，或 "Computing of Chinese language"，或更简写为"Chinese Computing"。

请注意，这并不是说，应该发展一种与"西文计算机"完全不同的"中文计算机"。我们的真正目标是要使传统的西文计算机都能具有处理中文的能力，发展新型的兼容于"西文计算机"的"中文计算机"。这才是"计算机中文化"的真正含意。这种新型的"中文计算机"应该保留和实现所有"西文计算机"的全部功能，应该可以运行全部成熟的商品软件。因此，有必要在一开始就提出从体系结构、数据结构、系统结构上强调中西文的"兼容"性。事实证明，这种"兼容"性是任何一种开发"中文计算机"系统方案能否卓有成效的关键，也是鉴别一个"中文计算机"系统优劣的主要标准。

在我国已经有专业学会从事这方面的专业研究工作。一个是"中国中文信息研究会"(The Chinese Information Processing Society of China)；一个是"中国仪器仪表学会汉字信息处理系统研究会"(The Chinese Character Information Processing Society of China Instrument Society). 该两大学会近年来在国内均召开了多次大型学术会议。并于1983年4月和10月分别在上海和北京组织了二次大型国际活动，即4月在上海举行的多国仪器仪表学术会议(Miconex,83)专题讨论和10月在北京举行的大型"中文信息处理国际研讨会"(ICCIP, 1983)。并有两本专业杂志，"中文信息学报"和"中文信息"公开广泛发行。同时台湾省的电脑学会也十分重视汉字信息处理课题，于1973年以来，每隔二年一次的"国际计算机会议"(Interntional Computer Symposium,ICS)均有不少论文发表。旅居国外的华人计算机科学家亦组织了"中文计算机学会"(Chinese Language Computer Society CLCS),每隔二年召开一次国际会议，并有一本专门的杂志。

本书的组织结构如下：

第一章是概论。第一节给出我们所说的"中文计算机"的概述。第二节以计算机中文化和中文计算机化为题，提出三个关于中文信息处理系统开发的一般问题。第三节以三个目前比较广泛使用的中文计算机系统为例子，使读者从一开始就对一个实际的系统有比较全面的了解。它们分别是：在WANG VS系列计算机上运行的多功能中文信息综合处理系统System C; Olympia 1011汉字自动打字机系统；APPLE II汉字系统。选取这三个系统作为例子，完全是由于我们比较熟悉的缘故，而且在这三个系统上我们也已经做了一些应用开发工作，对它们的优缺点有不少自己的体会。第四节简要介绍结构化系统设计语言SPARKS。

第二章讨论汉字结构与计算机表示法。从形音意公式出发，着重分析象形文字系统的层次组织结构，阐述了关于层次结构的树表示法。然后在第三节中，给出了树结构的计算机表示法。

第三章的题目是汉字的数据结构。由于汉字的数据结构问题，是直接影响着整个中文信息处理系统效率的根本性问题，因此，该章较深入地探讨了汉字字形表示、内部码和输入码的各种实现方法和设计准则，阐述了保证整个系统的中西文信息兼容的意义。最后，在第六节中，专门提出了作者关于实现所谓"中文化程序设计"问题的观点。

第四章阐述了二类代码映照的各种方法，其一是从输入码到内部码的输入映照；其二是从内部码到字形表示的输出映照。从一些实例出发，逐步引出了实现公式映照法、二叉树映照法、索引树映照法和trie-索引树映照法的许多算法，并对这些算法的复杂性作了分析和比较。

第五章讨论汉字的输入方法和输入设备。第一节强调指出，必须建立和推广独立于字典的输入子系统的理由。第二至四节分别讨论大键盘或笔触整字输入、中键盘纯形符字根部首拼字输入、标准小键盘编码输入、图象识别和声音识别输入等，各种输入方法和具体的输入设备。第六节讨论一个颇有意思的动态键盘的设计和实现问题。

第六章讨论中文计算机的输出方法和输出设备。由于汉字是一种象形文字，因此，汉字与图形有着密切的关系。一直存在着两种实现汉字和图形的方法，一种是图形汉字系统，以图做字；另一种是汉字图形系统，以字做图。该章的第一节即从这二种不同的实现方法出发，

讨论了屏幕显示输出过程的一般原理。第二、三两节介绍各种汉字输出设备的工作原理。第四节阐述联机通讯与分布式中文信息处理系统的构造。

第七章从软件工程的角度，阐述了关于信息综合处理系统构造和实现方法的一些主要技术问题。比较详细地介绍了关于实现全屏幕编辑、页面自动排版、表格自动生成等软件工具的程序设计技巧和具体算法。

第八章讨论关于中文数据库系统研究的一些技术问题。第一节从一些数据库的基本概念出发，引出一些中文数据库特有的技术问题，指出为什么要专门研究中文数据库系统的理由。并在以后的各节中分别讨论屏幕查询法、模糊匹配与智能查询、中文数据库设计和用户界面等问题。最后，以一个全文种情报检索系统作为中文数据库的实际应用，较全面地介绍一个中文数据库系统的构造和实现方法。

第九章的问题是语言理解和计算机文化。主要讨论一些涉及中文信息的人工智能研究，其中包括现代语言学知识、句法分析知识、中文信息的知识结构、全信息词典、自然语言理介和机器翻译、计算机诗词创作等有现实意义的研究课题。其中专门列了一节介绍有广泛实际应用的面向文法系统的设计和实现方法。

第十章讨论汉字的计算机自动识别。主要介绍一些句法模式识别的基本原理，各种关于汉字的分割法和基元选取与识别的方法。分别讨论了面向点阵字形，边写描述，骨架向里的各种颇有成效的汉字识别方法的实现技术，也介绍了结构汉字识别与基于图象数据库技术的汉字识别方法。

1.2 计算机中文化与中文计算机化

人类之所以能成为万物之灵，是因为有思维活动，即有思想、知识、智慧。人类不但能自觉地认识自然世界，能根据外界情况进行有意识地适应，而且可以通过劳动创造财富，创造世界，也创造了人类自己。语言是思维的物质外壳。思维如果不能用语言表达出来，就成了人类头脑中的空想，也就失去了存在的意义。人们用语言作为相互交流思维活动的工具和媒介时，一定有两方面的人员参加，一方是讲的人，另一方是听的人。讲的一方将其智慧输出，听的一方则进行智慧输入。但是语言这种传播媒介，却将随着时间与空间的变化而消失，于是人类创造了记录语言的符号，即文字。文字是表达语言的符号系统。这个符号系统，作为另一种记录思维活动和智慧的媒介，将人类的知识和智慧，作为一种“信息”，不受时间和空间的限制，逐年累月的积累，形成了一个庞大的信息库。近百年来，尤其是近三十年来，自然科学得到迅速的发展。人类通过自己的劳动，开始对自然界的奥秘有了进一步的认识与了解。在这个庞大的“信息库”中，人们存入了关于宇宙、星系、星云、恒星、行星、中子星、黑洞、太阳黑子等知识。人们逐步掌握了关于地球本身的资源开发、环境保护、天气预测方面的大量知识信息。人们有了原子能反应堆、火箭、卫星、宇航船、电子计算机等发明创造。在这个庞大的“信息库”中，人们也有了对微观世界中的分子、原子、质子、中子、电子、离子、光子、夸克、光电子等量子力学方面的知识。人类也对生物自身的规律加以研究，从微生物学、细菌学、抗疫学、到生命起源、生命时钟、人造胰岛素、基因分离和移植，以及最近赫有成果的遗传工程学、生命延长、生命创造等等，从而使人们在这个庞大的“信息库”中，不但存贮了对客观世界认识，

也对人类自身有了不断的认识。在向自然界挑战的同时，人类自身又相互加以组织，形成了关于政治、经济、教育、法律、军事、交通、生活、分配、管理、人事、福利等等社会知识信息。人类自身要长期地生存下去，自然界也将长期地生存下去，人类形成的社会也将长期地生存下去。无穷的新知识信息不断地涌现，人类已经陷入了知识信息的海洋，不时发出"信息爆炸"，"信息危机"的呼声。人类为了拯救自己，发明创造了电子计算机，开创了"信息革命"的时代。计算机成了时代的宠儿。而人类又是分成国家和民族共同生活在世界上的，每个国家和民族都有自己的语言和表达自己语言的文字。计算机的科学技术水平，生产规模和应用程度，已经成为衡量一个国家现代化水平的显著标志。能否有效地用自己国家和民族的文字对信息进行处理，进入"机器信息世界"之门，关系到一个国家和民族的兴亡。因此，计算机的中文化和中文的计算机化，成了我们炎黄子孙投以极大兴趣，并有重大意义的研究课题。事实已经证明，我们中国科学工作者有志气，也有能力，在信息处理方面赶上和超过世界先进水平。

既然电子计算机是一种由人类所创造，被人类所应用，为人类服务的工具。因此，计算机的中文化工作，正是讨论如何使计算机具有处理中文信息的能力，即发展一种有别于所谓传统"西文计算机"的新型计算机。然而，计算机的中文化只是一种手段。我们发展中文计算机这一工具的真正目的，在于应用这种手段使中国语言文字的信息处理机器化。因此，研究中文的计算机化，就是研究中国语言文字及其应用的现代化。从六十年代初期就开始研究的机器翻译，到现在已初具规模的机器字典、机器检索、机器教学等等，都属于中文计算机化的广泛范围之中。以前，国内外的一些学者曾狭义地认为："中文电脑化"就是指使汉字适合计算机要求的文字改革，就是推广中文的拼音化，因此，褒"计算机中文化"，贬"中文计算机化"是没有道理的。

"工欲善其事，必先利其器"。中国的语言文字经历漫长的历史年代演变至今，都是伴随着书写工具的改革而发展的。从三千多年前殷商时代刻在龟骨、兽骨上的卜辞和记事，到战国时期青铜食器、乐器、礼器上镌的铭文。从周秦时代鼓形石上的"刻石"到汉唐的"碑文"。从竹梃点漆著于锦帛到黑石磨汁、松烟制墨、蔡伦造纸，一直发展到今天的自来水笔和园珠笔。工具在改革，文字也在改革，由繁至简，越来越利于书写。但汉字的结构没有改变，手工写字

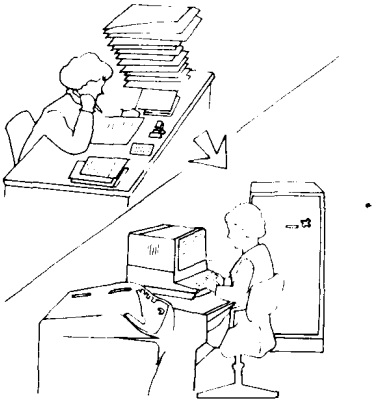

图 1.3 "机器写字"是对"手工写字"传统的伟大挑战

的过程也没有改变。也就是说，所有这些书写的工具的改革，在这几千年来并没有影响人们手工写字的传统习惯。但是，中文计算机这一工具的出现，却对人们的手工书写传统提出了伟大的挑战。它要求人们逐步适应这种新的情况，熟悉用按键的方式进行“机器写字”。现在我们还无法预料，中文计算机将对中国语言文字的改革带来多大的影响。但我们却有可能清楚地预见到，在中文计算机全面普及的未来世界里，人们将和现在的人不熟悉计算机一样，不熟悉使用笔和纸来手工写字了。当然，对于未来的他们，书法将仅仅作为一种专门的艺术享受而继承下来。未来的智能计算机和机器人也许还会进行机器书法的创作，可以和人进行书法比赛了。

值得引起我们注意的是，我们进入“机器信息世界”之门，从传统习惯上要比西方人困难。因为，西方人自从发明了西文打字机以后，从本质上已经开始了由手工书写到机器书写的转化过程。在西方计算机普及之前，随着西文打字机的普及，许多西方人本来已经丢掉了笔杆子，习惯了按键式的机器书写方式。因此，西文计算机的普及推广，要比中文计算机的普及推广有更好的基础。因此我们中国的计算机科学工作者，应该清醒地认识到，在进行计算机中文化工作时，我们有责任帮助人们从心理上克服手工书写的习惯势力，跨越中文打字机不成功的阶段，使人们能直接适应按键式机器书写方式。

西文打字机的发明与成功应用，虽然开始了机器写字阶段，却没有引起“革命”。电传打字机的发明与成功应用，虽然开始了机器通讯的阶段，但是从本质上说，也没有形成完整的信息交换系统。只有电子计算机才真正扮演了“信息革命”标志的角色。其本质区别在于：计算机有记忆存贮信息的能力以及自动管理信息的能力。因此，只有在计算机系统上才有可能模拟人的大脑，建立知识信息数据库，并发展一套完整的检索策略。当许多计算机通过通讯网络建立起分布式网络系统时，才有可能实现全部信息的电子化，由计算机来管理整个系统中的信息交换。电子帐册和电子档案等将使企业的管理实现现代化，电子电话和电子公文等将使办公室工作自动化，电子报纸和电子订票等将使人们文化生活现代化。整个世界的信息结构，人们的思维观念都将与“传统观念”彻底决裂，一场伟大的“信息革命”已经开始。记得

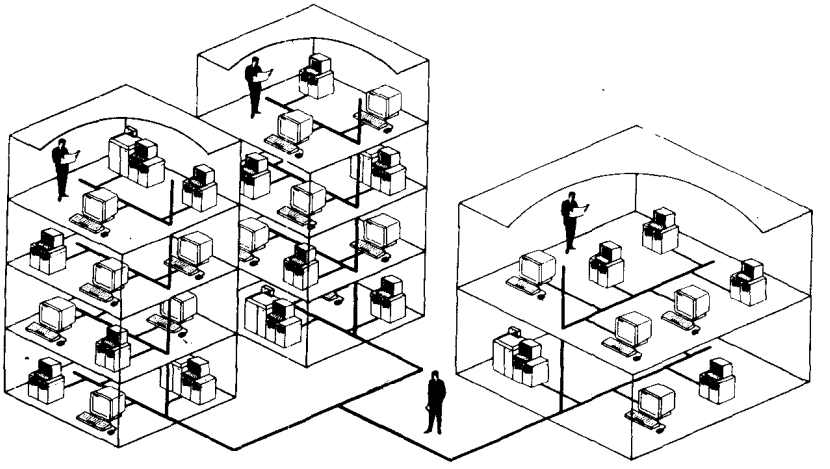

图 1.4　信息社会的电子化信息管理与交换