院士科普书系

教电脑识字

——浅谈汉字识别

吴佑寿　著

清华大学出版社
暨南大学出版社

(京)新登字 158 号

出 版 者: 清华大学出版社(北京清华大学学研大厦,邮编 100084)
http://www.tup.tsinghua.edu.cn
暨南大学出版社(广州天河,邮编 510630)
http://www.jnu.edu.cn
责任编辑: 宋成斌
印 刷 者: 北京鑫丰华彩印有限公司
发 行 者: 新华书店总店北京发行所
开 本: 850×1168 1/32 印张: 5 字数: 97 千字
版 次: 2000 年 12 月第 1 版 2002 年 7 月第 2 次印刷
书 号: ISBN 7-302-04216-0/G·183
印 数: 5001～7000
定 价: 12.00 元

《院士科普书系》编委会(第二届)

提高全民族的科学素质

——序《院士科普书系》

人类走到了又一个千年之交。

人类的文明进程至少已有 6000 余年。地球上各个民族共同创造了人类文明的灿烂之花。中华文明同古埃及文明、古巴比伦文明、古印度文明、古希腊文明等一起，是人类文明的发源地。

15 世纪之前，以中华文明为代表的东方文明曾遥遥领先于当时的西方文明。从汉代到明代初期，中国的科学技术在世界上一直领先长达 14 个世纪以上。在那个时期，影响世界文明进程的重要发明中，相当部分是中华民族的贡献。

后来，中国逐渐落后了。中国为什么落后？近代从林则徐以来许多志士仁人就不断提出和思索这个历史课题。但都没有找到正确的答案。以毛泽东同志、邓小平同志为代表的中国共产党人作出了唯一正确的回答：中国落后，是由于生产力的落后和社会政治的腐朽。西方列强对中国的欺凌，更加剧了中国经济的落后和国家的衰败。而落后就要挨打。所以要进行革命，通过革命从根本上改变旧的生产关系和政

治上层建筑,为解放和发展生产力开辟道路。于是,就有了80多年前孙中山先生领导的辛亥革命,就有了50年前我们党领导的新民主主义革命的胜利,以及随后进行的社会主义革命的成功。无论是革命还是我们正在进行的社会主义改革,都是为了解放和发展生产力。

邓小平同志提出的"科学技术是第一生产力"的著名论断,使我们对科学技术在经济和社会发展中的地位与作用的认识,有了新的飞跃。我们应该运用这一真理性的认识,深刻总结以往科学技术发展的历史经验,把我国科技事业更好地推向前进。中国古代科技有过辉煌的成果,但也有不足,主要是没有形成实验科学传统和完整的学科体系,科学技术没有取得应有的社会地位,更缺乏通过科技促进社会生产力发展的动力和机制。为什么近代科学技术首先在文艺复兴后的欧洲出现,而未能在中国出现,这可能是原因之一吧。而且,我国历史上虽然有着伟大而丰富的文明成果和优良的文化传统,但相对说来,全社会的科学精神不足也是一个缺陷。鉴往开来,继承以往的优秀文化,弥补历史的不足,是当代中国人的社会责任。

在新的世纪中,中华民族将实现伟大的复兴。在一个占世界人口五分之一的发展中大国里,再用50年的时间基本实现现代化,这又是一项惊天动地的伟业。为实现这个光辉

的目标,我们应该充分发挥社会主义制度的优越性,坚持不懈地实施科教兴国战略。

科教兴国,全社会都要参与,科学家和教育家更应奋勇当先,在全社会带头弘扬科学精神,传播科学思想,倡导科学方法,普及科学知识。科教兴国也要抓好基本建设。编辑出版高质量的科普图书,就是一项基本建设,对于提高全民族的科学素质,是很有意义的。在《院士科普书系》出版之际,写了上面这些话,是为序。

一九九九年十二月二十三日

人民交给的课题

——写在《院士科普书系》出版之际

世界正在发生深刻的变化。这一变化是20世纪以来科学技术革命不断深入的必然结果。从马克思主义的观点看来,生产力的发展是人类社会发展与文明进步的根本动力;而"科学技术是第一生产力",因此,科学技术是推动社会发展与文明进步的革命性力量。从生产力发展的阶段看,人类走过了农业经济时代、工业经济时代,正在进入知识经济时代。

知识经济时代,知识取代土地或资本成为生产力构成的第一要素。知识不同于土地或资本,不仅仅是一种物质的形态,知识同时还是一种精神的形态。知识,首先是科学技术知识,将不仅渗透到生产过程、流通过程等经济领域,同时还将渗透到政治、法律、外交、军事、教育、文化和社会生活等一切领域。可以说,在新的历史时期,一个国家、一个民族能否掌握当代最先进的科技知识以及这些科技知识在国民中普及的程度将决定其国力的强弱与社会文明程度的高低。科技创新与科普工作是关系到一个国家、一个民族兴衰的

大事。

对于我们科技工作者来说,我们的工作应当包含两个方面:发展科技与普及科技;或者说应当贯穿于知识的生产、传播及应用的全过程。我们所说的科普工作,不仅是普及科学知识,更应包括普及科学精神和科学方法。

我们的党和政府历来都十分重视科普工作。党的十五大更是把树立科学精神、掌握科学方法、普及科技知识作为实施科教兴国战略和社会主义文化建设的一项重要任务提到了全党、全国人民和全体科学工作者的面前。

正是在这样的背景下,1998 年春由科学时报社(当时叫“中国科学报社”)提出创意,暨南大学出版社和清华大学出版社积极筹划,会同中国科学院学部联合办公室和中国工程院学部工作部,共同发起《院士科普书系》这一重大科普工程。

1998 年 6 月,中国科学院与中国工程院“两院”院士大会改选各学部领导班子,《院士科普书系》编委会正式成立,各学部主任均为编委会委员。编委会办公室在广泛征求意见的基础上拟出 150 个“提议书目”,在“两院”院士大会上向 1000 多名院士发出题为《请科学家为 21 世纪写科普书》的“约稿信”,得到了院士们的热烈响应。在此后的半年多时间里,有 176 名院士同编委会办公室和出版社签订了 175 本书的写作出版协议,开始了《院士科普书系》艰辛的创作过程。

《院士科普书系》的定位是结合当代学科前沿和我国经济建设与社会发展的热点问题，普及科技知识、科学方法。科学性、知识性、实用性和趣味性是编写的总要求。

编写科普书对我国大多数院士来说是一个新课题。他们惯于撰写学术论文。如何把专业的知识和方法写成生动、有趣、有文采的科普读物，于科技知识中融入人文教育，不是一件容易的事。不少院士反映：写科普书比写学术专著还难。但院士们还是以感人的精神完成自己的书稿。在此过程中，科学时报社和中国科学院学部联合办公室、中国工程院学部工作部以及清华大学出版社、暨南大学出版社也付出了辛勤的劳动。

《院士科普书系》首辑终于出版了。这是人民交给科学家课题，科学家向人民交出答卷。江泽民总书记专门为《院士科普书系》撰写了序言，指出科普是科教兴国的基础工程，勉励科学家、教育家“在全社会带头弘扬科学精神，传播科学思想，倡导科学方法，普及科学知识”，充分表达了党的第三代领导集体对科普的重视，对提高全民族科技素质的殷殷期望。

《院士科普书系》将采取滚动出版的模式。一方面随着院士们的创作进程，成熟一批出版一批；另一方面随着科学技术的进步和创新，不断有新的题材由新的院士作者撰写。因此，《院士科普书系》将是一个长期的、系统的科普工程。

这一庞大的工程，不但需要院士们积极投入，还需要各界人士和广大读者的支持——对我们的选题和内容提出修订、完善的建议，帮助我们不断提高《院士科普书系》的水平与质量，使之成为国民科技素质教育的系统而经典的读本。在科学家群体撰写科普书方面，我们也要以此为起点为开端，参与国际竞争与合作，勇攀世界科普创作的高峰。

中国科学院院长
《院士科普书系》编委会主任

路甬祥

2000 年 1 月 8 日

代　　序

——方块字会被拼音文字取代吗？

“方块汉字是否将被拼音文字所取代”这个题目曾经在国内外不太大的知识分子圈内流传过。为了说明它的来龙去脉，必须先从古老的方块汉字跟先进的计算机的关系（包括用计算机识别方块汉字的关系）说起。

汉字识别，通俗地讲就是教计算机“识字”，其目的是把汉字自动转换成计算机内部的编码，以便于用计算机对汉字所携带的信息做进一步处理，如查询检索、提取摘要、编辑、出版、翻译、建立数据库等等。

文字是信息的载体，是人们表达和交流思想，传播知识和情报，保存资料和典籍的媒介。方块汉字已有数千年的历史，也是世界上使用人数最多的文字之一。汉字对中华民族灿烂文化的形成和发展，以及对世界文化和科学技术的影响，都有着不可磨灭的功绩。

但是方块汉字也有着突出的弱点。它一字一形，结构复杂，而且字量多，字体不一，书写印刷都十分不便。在相当长的历史时期内，汉字书写工具主要是被称之为“文房四宝”的笔墨纸砚，相当原始。在人类逐步进入信息社会，通信非常发达，计算机得到广泛应用的时代，方块汉字的弱点就显得更为突出。摆在我们面前的问题是：大量的资料、文献、典籍需要整理、传送、利用或保存，许多书刊、文章和法规，要求能

自动检索查阅或翻译为其他文字,办公及管理自动化要求能及时编制、处理和传送各种文件,检索并获取有关情报资料,迅速做出预测与决策,等等。此外,电子信函、电子出版、电子商务之类的应用也日益普及。凡此种种依靠落后工具和人工方法是难以办到的。如何利用当今科学技术的成果,特别是计算机等先进工具,使汉字这一人类的瑰宝继续并更好地发挥作用,就成为我们必须解决的问题。

电子计算机是西方国家发明并发展起来的,它的基础自然是西方的语言文字,其键盘也是从西文打字机衍变而来的。对于习惯于用打字机打字的西方人来说,用键盘往计算机输入西文是很自然和简单的事情。在我国,情况却迥然不同。虽然我国是活字排版印刷的发源地,但是用键盘"打"方块汉字,在相当长一段时间内却被人们视为"畏途",有的人甚至认为方块汉字将要消亡而被拼音文字所代替!解决这个问题的关键就是研究计算机如何适应汉字信息处理的需要,使它不但能取代"笔墨纸砚"等传统工具,还能实现诸如信息交换、处理、存储、翻译等功能。在信息技术迅猛发展的今天,这个问题的重要性已是不言而喻了。

用计算机来处理汉字信息的系统大体分为汉字输入、汉字信息加工与处理、汉字输出三个部分。汉字输入是把方块汉字转换成为计算机便于处理的代码,这是汉字信息处理的基础,也是汉字信息处理系统的"瓶颈"。这是人和机器(计算机)连接的关键部位,是人把信息或命令送入计算机的"咽喉",技术上叫做"人机接口"。这是两个不同系统或不同设备之间的一个公共"边界"或交接部分,其功能是使两个连接着的系统或设备兼容,使它们能够互相衔接,互换信息。

从技术上说,人和计算机是两种不同的系统。人向计算

机传送信息或下达命令,可以用文字,也可以用声音,甚至可以用“气味”、“手势”和某种“动作”,等等。《阿里巴巴和四十大盗》中的强盗向山洞的大门高喊:“芝麻,开门!”就是用语音向藏着珠宝的山洞下达命令。由于某种魔法的作用,山神能够听懂强盗们的口令,迅速把洞门打开。这是一种典型的“话音接口”。“芝麻,开门!”是美丽的神话,而在信息技术高度发展的今天,用计算机来识别话音、文字,以及其他形式的信号,则已经成为事实。

近 20 年来,我国科技工作者在包括汉字识别在内的中文信息处理的各方面做了很多富有创造性的工作,使方块汉字无法跟计算机相结合的思想障碍不攻自破。汉字键盘输入已可以跟英文字母的键盘输入相媲美,汉字计算机识别研究也已取得突破性进展,并在实际中得到应用。目前市场上出售的微机已普遍具有汉字输入和汉字信息处理功能,各种汉字数据库已不断建立,汉字照排系统已成功地用于各种印刷系统,并已远销海外。在基础研究方面,利用计算机对汉字字频加以统计并分析其分布规律,也取得丰硕成果。关于词切分、书面汉语分析与理解、机器翻译等方面的研究工作也不断深入,特别是对汉字信息处理起着重要的规范作用的各种标准,如国家标准 GB 2312-80《信息交换汉字编码字符集——基本集》、《信息处理用 GB 13000.1 字符集——汉字部件规范》等,也已经或正在逐步建立之中,“方块汉字”不仅能适应现代信息社会的需求,而且汉字识别技术独辟蹊径。

撰写这本小册子的目的主要是介绍计算机识字的原理和国内外有关的研究工作。除此之外,作者还希望通过介绍我国汉字识别研究工作的发展历程和成就,来具体表明:我国科技工作者不仅有责任、有能力解决计算机中文信息处理

的问题,而且会使经过五千年磨炼并对人类做出巨大贡献的方块汉字,能在不断发展的人类社会中继续发挥作用,为人类文明和进步谱写更瑰丽的诗篇。

本书在主要介绍计算机汉字自动识别的原理、方法,以及国内外有关的研究工作的过程中,特别着重介绍我国几种联机手写汉字识别系统(笔输入装置)和印刷汉字识别系统(光符阅读器)的构成及应用。全书分为 6 章:第 1 章是一般性介绍,主要讨论汉字识别的用途、分类和工作原理;第 2、3、4 章是本书的主要内容,分别介绍几种类型的汉字识别系统的特点、系统构成和问题,着重介绍它们的识别方法及目前的技术水平;第 5 章介绍一种实用系统,简单讨论组成识别系统的一些外围设备,以及进一步提高识别系统性能的方法(如后处理法等);第 6 章讨论利用人工神经网络识别汉字的方法。为了使读者能根据自己的需求和兴趣有选择地阅读,本书各部分尽量做到相互独立,读者不一定对本书从头到尾阅读。对于只希望对汉字识别有粗浅了解的读者,也许只看看第 1 章的内容就可以了;要求更深入了解有关问题的读者,除第 1 章外,也可按自己的需求再选读相应章节。这样,各章的内容有少许重复是难以避免的。考虑到这是一本近 10 万字的科普书籍,读者很难一口气花较长时间去阅读,因此,上述安排也许对读者较为方便。

吴佑寿[1]

2000 年 1 月

目　　录

1　计算机是怎样识字的

1.1　从国际象棋人机大战说起 …… 1
1.2　机器是怎样识字的 …… 4
1.3　汉字识别系统的分类 …… 6
1.4　对汉字识别系统的要求 …… 7
1.5　汉字的基本知识 …… 10
1.6　关于模式识别的讨论 …… 15
1.7　汉字识别的关键问题 …… 19

2　联机手写汉字识别——笔输入

2.1　什么是笔输入——从键盘输入说起 …… 22
2.2　计算机是怎样识别手写汉字的 …… 25
2.3　联机手写汉字识别的困难 …… 27
2.4　国内外联机手写汉字识别研究简况 …… 32
2.5　笔画编码法 …… 34
2.6　汉王笔 …… 39
2.7　文通笔 …… 41
2.8　书写板及其他 …… 43
2.9　个人数字助理和智能通信手机 …… 46

2.10 智能笔 …… 47

3 汉字光符阅读器——印刷汉字识别

3.1 从超市收款机说起 …… 50
3.2 脱机汉字识别的困难 …… 52
3.3 汉字 OCR 系统的构成 …… 56
3.4 第一个实验性印刷汉字识别系统 …… 61
3.5 能识别 6763 个印刷汉字的系统 …… 65
3.6 让汉字光符阅读器变得更聪明——关于多体汉字识别 …… 70
3.7 “研”以致用——把“智能”更高的汉字 OCR 投入市场 …… 75

4 “最后的堡垒”——脱机手写汉字识别

4.1 攻克堡垒待创新 …… 81
4.2 手写汉字脱机识别的困难 …… 82
4.3 脱机手写汉字识别系统的“课本”——手写汉字样本库 …… 86
4.4 脱机手写汉字识别的主要问题及其解决办法 …… 89

5 实用印刷汉字识别系统

5.1 沿“实现产业化”的方向做不懈努力 …… 94
5.2 印刷汉字识别系统的构成 …… 96

5.3 实用印刷汉字识别系统的总体技术指标 …… 99
5.4 汉字识别系统的输入装置 …… 104
5.5 文本的版面分析和理解 …… 108
5.6 汉字的行切割和字切割 …… 111
5.7 汉字图像的归一化 …… 115
5.8 怎样建立识别系统中的字典 …… 117
5.9 单字识别中的匹配判决 …… 119
5.10 后处理纠错 …… 122

6 人工神经网络汉字识别系统
6.1 电脑能“变成”人脑吗 …… 125
6.2 生物神经元结构 …… 127
6.3 神经元模型 …… 129
6.4 能识别两种模式的人工神经网络 …… 130
6.5 能识别3755个汉字的人工神经网络识别器 …… 133
6.6 “类脑计算机”离实用尚远 …… 135

参考文献 …… 138