高等学校教材·计算机科学与技术

可下载课件
http://www.tup.tsinghua.edu.cn

中文信息处理技术
——原理与应用

李宝安　李 燕　孟庆昌 编著

清华大学出版社

高等学校教材·计算机科学与技术

中文信息处理技术
——原理与应用

李宝安　李　燕　孟庆昌　编著

清华大学出版社

北京

内 容 简 介

本书以简单、实用、易于理解为原则，内容力求全面、新颖，涵盖了中文信息处理的主要相关技术和研究成果。读者阅读本书之后，能够系统地了解汉字的编码、字形压缩与还原、光学汉字识别、中西文兼容处理、汉语自然语言处理等技术，以及中文信息处理技术的典型应用系统的原理与使用，如电子排版印刷系统、办公自动化系统、Internet 网络搜索引擎、智能检索系统等，最终达到对中文信息处理技术的系统性了解。本书附录中还提供了该领域常用的各项国家标准。

本书可作为大专院校计算机、信息管理、系统工程等专业的本科教材，也可以供从事中文信息系统研发工作的科研人员参考。

图书在版编目(CIP)数据

中文信息处理技术：原理与应用/李宝安，孟庆昌编著. —北京：清华大学出版社，2005.7
(高等学校教材·计算机科学与技术)
ISBN 7-302-11200-2

I. 中… II. ①李… ②孟… III. 汉定信息处理—高等学校—教材 IV. TP391.12

中国版本图书馆 CIP 数据核字(2005)第 061892 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社总机：010-62770175　　**客户服务**：010-62776969
责任编辑：付弘宇
印 刷 者：北京市清华园胶印厂
装 订 者：北京市密云县京文制本装订厂
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：23.75　**字数**：590 千字
版　　次：2005 年 7 月第 1 版　2005 年 7 月第 1 次印刷
书　　号：ISBN 7-302-11200-2/TP·7394
印　　数：1～3000
定　　价：35.00 元

高等学校教材·计算机科学与技术

编审委员会成员

（按地区排序）

清华大学	周立柱　教授
	覃　征　教授
	王建民　教授
	刘　强　副教授
	冯建华　副教授
北京大学	杨冬青　教授
	陈　钟　教授
	陈立军　副教授
北京航空航天大学	马殿富　教授
	吴超英　副教授
	姚淑珍　教授
中国人民大学	王　珊　教授
	孟小峰　教授
	陈　红　教授
北京交通大学	阮秋琦　教授
北京信息工程学院	孟庆昌　教授
北京科技大学	杨炳儒　教授
石油大学	陈　明　教授
天津大学	艾德才　教授
复旦大学	吴立德　教授
	吴百锋　教授
	杨卫东　副教授
华东理工大学	邵志清　教授
华东师范大学	杨宗源　教授
	应吉康　教授
东华大学	乐嘉锦　教授
上海第二工业大学	蒋川群　教授
浙江大学	吴朝晖　教授
	李善平　教授
南京大学	骆　斌　教授
南京航空航天大学	秦小麟　教授
南京理工大学	张功萱　教授

南京邮电学院	朱秀昌　教授
苏州大学	龚声蓉　教授
江苏大学	宋余庆　教授
武汉大学	何炎祥　教授
华中科技大学	刘乐善　教授
中南财经政法大学	刘腾红　教授
华中师范大学	王林平　副教授
	魏开平　教授
武汉理工大学	李中年　教授
国防科技大学	赵克佳　教授
	肖　侬　副教授
中南大学	陈松乔　教授
湖南大学	林亚平　教授
	邹北骥　教授
西安交通大学	沈钧毅　教授
	齐　勇　教授
西北大学	周明全　教授
长安大学	巨永峰　教授
西安石油学院	方　明　教授
西安邮电学院	陈莉君　副教授
哈尔滨工业大学	郭茂祖　教授
吉林大学	徐一平　教授
	毕　强　教授
长春工程学院	沙胜贤　教授
山东大学	孟祥旭　教授
	郝兴伟　教授
山东科技大学	郑永果　教授
中山大学	潘小轰　教授
厦门大学	冯少荣　教授
福州大学	林世平　副教授
云南大学	刘惟一　教授
重庆邮电学院	王国胤　教授
西南交通大学	杨　燕　副教授

版说明

改革开放以来，特别是党的十五大以来，我国教育事业取得了举世瞩目的辉煌成就，高等教育实现了历史性的跨越，已由精英教育阶段进入国际公认的大众化教育阶段。在质量不断提高的基础上，高等教育规模取得如此快速的发展，创造了世界教育发展史上的奇迹。当前，教育工作既面临着千载难逢的良好机遇，同时也面临着前所未有的严峻挑战。社会不断增长的高等教育需求同教育供给特别是优质教育供给不足的矛盾，是现阶段教育发展面临的基本矛盾。

教育部一直十分重视高等教育质量工作。2001 年 8 月，教育部下发了《关于加强高等学校本科教学工作，提高教学质量的若干意见》，提出了十二条加强本科教学工作提高教学质量的措施和意见。2003 年 6 月和 2004 年 2 月，教育部分别下发了《关于启动高等学校教学质量与教学改革工程精品课程建设工作的通知》和《教育部实施精品课程建设提高高校教学质量和人才培养质量》文件，指出“高等学校教学质量和教学改革工程”是教育部正在制订的《2003—2007 年教育振兴行动计划》的重要组成部分，精品课程建设是“质量工程”的重要内容之一。教育部计划用五年时间（2003—2007 年）建设 1500 门国家级精品课程，利用现代化的教育信息技术手段将精品课程的相关内容上网并免费开放，以实现优质教学资源共享，提高高等学校教学质量和人才培养质量。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上；精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学

科体系有实质性的改革和发展、顺应并符合新世纪教学发展的规律、代表并引领课程发展的趋势和方向)、示范性(教材所体现的课程体系具有较广泛的辐射性和示范性)和一定的前瞻性。教材由个人申报或各校推荐(通过所在高校的“编委会”成员推荐),经“编委会”认真评审,最后由清华大学出版社审定出版。

目前,针对计算机类和电子信息类相关专业成立了两个“编委会”,即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。首批推出的特色精品教材包括:

(1) 高等学校教材·计算机应用——高等学校各类专业,特别是非计算机专业的计算机应用类教材。

(2) 高等学校教材·计算机科学与技术——高等学校计算机相关专业的教材。

(3) 高等学校教材·电子信息——高等学校电子信息相关专业的教材。

(4) 高等学校教材·软件工程——高等学校软件工程相关专业的教材。

(5) 高等学校教材·信息管理与信息系统

清华大学出版社经过近 20 年的努力,在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌,为我国的高等教育事业做出了重要贡献。清华版教材经过 20 多年的精雕细刻,形成了技术准确、内容严谨的独特风格,这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

E-mail: dingl@tup.tsinghua.edu.cn

前　言

计算机中文信息处理技术，是我国特有的、利用计算机系统来处理中文信息的技术。

从最早发现的甲骨文到现在，中文已有三千多年的历史，可以说中文的历史就是中华民族灿烂文化的历史。目前，人类已经跨入21世纪，当前的社会是一个信息社会，是计算机科学、通信技术、Internet等众多先进技术和学科快速、普及发展的时代。作为当今世界上人口最多、国民经济快速发展的国家，怎样抓住机遇，大力发展信息技术，进一步提升我国的综合竞争力，缩小与发达国家的技术差距，已成为我国政府亟待解决的问题。面对如此众多的中文人口，要想在我国普及计算机信息系统，解决好中文信息的计算机处理问题是必要的基础和先导。

中文信息处理技术不仅涉及到计算机体系结构、操作系统、程序设计语言、数据库和网络通信技术，还涉及到语言文字学、语音学、词汇学、人工智能机器学习、文字识别技术、语音识别技术、排版印刷技术等，是一门多学科交叉的科学。

目前，中文信息的处理已经在很多方面取得了不小的进展，而且还在不断完善与发展，如汉字编码输入处理技术、汉字字形压缩与还原技术、光学汉字识别技术、中文文字与词语处理系统、汉字设备、中文通信系统、机器翻译系统、中西文兼容处理技术、电子排版印刷系统、办公自动化系统、Internet搜索引擎、智能检索系统等。

目前，不少高校已经开设了“中文信息处理技术”课程，但由于各种原因，有关“中文信息处理技术”课程的教材非常缺乏，特别是由于中文信息处理技术是伴随着计算机技术的发展而不断变化的，能够紧跟计算机技术的发展，反映最新研究和应用成果的专著或教材就更少了。笔者根据多年来从事“汉字信息处理”课程的教学实践、技术研究和实际应用开发的经验，编写了《中文信息处理技术——原理与应用》一书，一方面可作为大专院校的教材，另一方面可以满足从事计算机系统研究与开发的广大科研、工程技术人员的需要，供大家参考。

全书共七章：第1章中文信息处理技术概论，介绍信息处理的实质、汉字编码的种类与变换、汉字内码体系、Unicode与Unicode汉字、中西文兼容处理问题、中文信息处理系统五层结构模型等；第2章汉字编码输入原理，对汉字及汉字的属性进行深入地分析与刻画，介绍汉字编码输入方法的分类，给出汉字键盘码的笛卡儿积集与汉字信息的熵值的概念，并对汉字编码方案给出简易及专业的评测方法，对中文输入技术的一些问题进行探讨；第3章汉字字形存储与压缩技术，介绍汉字字形码的整字存储和压缩存储问题，常见压缩与还原技术及其重要指标，如矢量存储、部件组字、子信息块哈夫曼树等；第4章汉字识别技术，对OCR技术的发展概况、汉字识别的种类与原理、印刷体和联机手写汉字识别的方法给出详细说明，并给出一些汉字识别产品的应用实例；第5章中西文兼容处理技术，介绍中西文兼容处理的概念、中文信息处理系统结构、常见汉字编码辨析与编码转换、系统级兼容处理方法、应用级兼容处理方法、终端级兼容处理方法、UNIX操作系统的汉化与国际化；第6章汉语自然语言理解，对汉语自然语言理解概述、语言的分类与理解语言

的过程、中文理解的单位等问题进行探讨，介绍自然语言理解的国内外研究现状、汉语自然语言理解与生成的进展难点与问题、基于语法语义的汉语自然理解系统等；第 7 章中文信息处理技术的应用，介绍中文应用系统发展概况、中文电子印刷排版系统原理与相关产品、中文信息检索系统的原理与组成及中文文献的自动分类、检索方法与技术、搜索引擎概述与分类、使用方法与总体评价等。

本书的最后提供了五个附录，分别介绍几个常用国家标准，即中华人民共和国国家标准 GB 2312—1980《信息技术　信息交换用汉字编码字符集　基本集》、GB/T 7589－1987《信息交换用汉字编码字符集　第二辅助集》、GB /T7590－1987《信息交换用汉字编码字符集　第四辅助集》、GB 13000.1—1993《信息技术　通用多八位编码字符集(UCS)　第一部分：体系结构与基本多文种平面》、GB18030—2000《信息技术　信息交换用汉字编码字符集　基本集的扩充》，供读者参考。

本书第 1~3、5~7 章由李宝安老师编写，第 4 章和附录由李燕副研究员编写，全书由孟庆昌教授审定，陈晓玲老师和王友兰老师参与了书稿的校阅和部分录入工作。

本书的读者对象为普通高校大学本科或专科的学生，适合作为现阶段高校计算机、自动化、信息管理、系统工程等专业的中文信息处理技术相关课程的教材或教学参考书，也可供从事中文信息系统研究和开发的广大工程技术人员自学或参考。

笔者在编写本书的过程中，努力跟踪中文信息处理学科的新发展、新技术，把它们纳入到本书中来，以保持本书的先进性和实用性。笔者要特别感谢国内外中文信息处理领域的专著、教材和许多高水平论文、报告的作者们，以及为我国中文信息处理技术的发展做出过杰出贡献的单位和个人，正是由于他们的不懈努力和辛勤工作，我国的中文信息处理领域呈现出了前所未有的大好局面，研制出了许多具有中国特色的计算机产品，并且向着更高的目标发展。同时也使本书能够取各家之长，较全面地反映中文信息处理各个应用领域的最新进展。但由于笔者学识浅陋，加之编写时间紧迫，难免有许多不足之处，在此诚恳地希望各位专家和读者不吝指教。

李宝安　孟庆昌

2004 年 8 月

目　录

第 1 章　中文信息处理技术概论

1.1　信息处理的实质

1.1.1　信息和信息技术

1. 信息

在信息技术领域，信息是指对事物之间相互联系、相互作用的状态的描述。

信息的性质有普遍性、无限性、相对性、转移性、共享性、变换性、动态性、转换性。

信息的传播及利用可以追溯到古代的烽火台、飞鸽传书，近代的邮政、电报、电话以及现代的计算机、计算机网络、互联网、无线通信等等。

人类认识世界的过程，实际上就是获得外部世界信息并对这些信息进行加工的过程；而改造世界的过程，则是由认识主体把加工所形成的信息(目标和策略)反作用于外部世界、并不断按照策略信息来引导外部事物达到目标的过程。因此，人类认识世界和改造世界的过程本质上就是一个信息处理过程。一个完备的控制系统必然也是一个完备的信息处理过程。

信息的基本功能是作为生存的要素、社会的资源、认识的向导、实践的指南、决策的依据、控制的基础、智慧的源流、系统的灵魂。

2. 信息技术

信息技术就是用以扩展人的信息器官功能的技术。

人的信息器官及其功能分别是：感觉器官完成获取信息功能；传导神经网络完成传递信息功能；思维器官完成加工和再生信息功能；效应器官完成使用信息功能。

信息技术的基本内容就是所谓的信息技术四基元，即感测技术、通信技术、智能技术及控制技术。

信息系统的工作流程如图 1-1 所示。

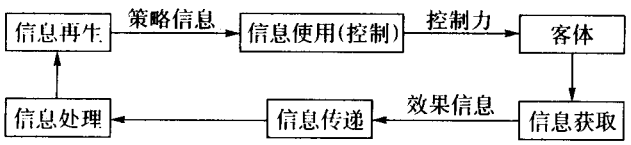

图 1-1　完备的信息系统的工作流程

信息是自然环境和人类的一切活动所产生的各种状态和消息的总称。人们很早就已知道信息这一概念。从定性的意义上说，人们在得知某个消息后，他在事前认为消息中所包含的事件发生的可能性愈小，则认为这个消息给他带来的信息量愈大。可见信息的量值与

事件的随机性有关。

信息在人类社会活动的各方面都很重要。但是，在科技不甚发达的时代，信息的作用及其利用价值被限制在较低的程度上。例如，信息技术的一种手段为传递，在电信技术发明以前，人们只能用人工通信，或者其他简单的表示方式或各种约定来传递信息。而电气通信技术的发展，从电话电报到传真、电视，从有线通信发展到无线通信，直到微波、光纤通信、卫星通信，信息的传输速率大大提高，性能也在改善，但只限于传输信息。信息技术的另一方面为信息处理技术。20 世纪 40 年代发明了电子计算机，开始只是利用它处理数值运算，但是很快就意识到可以利用数据代表广义的信息，从而发展了数据信息处理这一意义深远的应用技术。利用计算机处理数据信息，不只是作单纯的信息传输，而主要是对信息按某种规律做某种意义的加工，使它适应某种特定目的的需要。例如，气象预报中的信息处理，结合信息传感技术，对采集到的原始信息按预先设计的数学模型进行处理，得出的结果可以作为气象预报的资料。对信息进行加工处理离不开计算机技术，所以信息处理这一术语就和计算机技术联系在了一起。用计算机处理或加工信息，扩大了信息的利用范围，使信息的利用价值也大为提高。这一意义深远的科技成果的应用，使信息日益成为现代社会科技进步、经济发展、人类文明进程所不可缺少的社会财富。它和物质、能源被列于同等重要的地位，被看作现代人类社会生存和发展的三大要素。科技进步的国家已经建立起强大的信息产业，并仍在高速发展，在整个国民经济中的份额日益增大。信息处理技术在人类文明和科学技术现代化的进程中正在发挥重要的作用。

广义的信息涉及多种范畴。例如，一些自然现象所包含的各种信息；人类社会活动，如政治、经济、军事、文化、商业等活动所产生的各种信息；科学技术和生产活动，如揭示自然和物质结构的奥秘，从事地质研究、探矿等产生的各种信息。它们涉及人们生存的环境和从事科研、生产、生活等活动的一切方面。在这些含义丰富的信息中，信息的表示形式又是多样性的。例如，信息可以有数据、文字、声音、图形等多种形式，这称为信息的多元化表示。

用计算机处理多元化信息，是信息处理技术的范畴。根据信息处理技术的发展情况，可以分为传统的信息处理和通信技术，以及现代的信息处理技术。传统的信息处理指狭义的信息处理，如信息的存储和检索；传统的通信技术只是完成信息的传输或转移；而现代化的通信技术(即广义的信息处理技术)则兼有信息处理和信息传输的功能。

传统的信息处理技术在近十多年来有了很大的发展。这要归功于微电子技术和计算机技术的飞速进步。微电子技术的进步体现在超大规模集成电路的技术水平日益提高，各种大容量存储器芯片和具有复杂逻辑运算功能的集成电路芯片日益增多，并迅速推广使用。计算机技术的进步体现在计算机硬件性能价格比的大幅度提高，微型机和以微型机技术为基础的各种终端设备的日益普及。这些因素大大推进了信息处理技术的实用化进程。另一方面，计算机软件技术也有很大进步，例如，软件工程、第四代程序设计语言和各种先进的软件工具的实用化，数据库管理系统等各种公共支持软件技术的进步和普及。人工智能软件技术的发展以及各种应用软件的开发和利用，不仅使数据和文字信息处理技术更加完善，应用更为广泛，而且开拓了信息处理技术的新的应用领域，如图像信息处理、模式识别、语音识别和语音合成、自然语言处理、语言的翻译等高技术领域。

传统的通信技术以传输模拟信号为主，自从数据通信技术出现之后，经计算机存储和

处理的信息可以在两台或多台计算机或数据处理设备之间互相传输，从而增强了信息处理和传输的能力，特别是互联网时代的到来，更扩展了信息处理技术的范畴。

1.1.2　文字信息处理

信息的表示形式是多样的。那么，当前人们最关心什么形式的信息？根据 IBM 公司的调查，当前人们最关心的仍是文本信息。在多元化的信息中，文字信息是一种最通用、最普遍的表示形式。

各种信息的特点如何？各种信息的表现形式如何？各有什么特点？

视频、音频的特点是表现形式直观，表达的信息易于被不同层次的对象接受。超视声频和其他形式的信息现在还鲜有用于计算机处理的例子。其中，文本信息的特点是易于传播、所需存储空间小。但是由于世界各国语言文字存在较大差异，交流的群体受到限制。现在的公文、文件、信函、报表、各种印刷出版物等绝大多数都使用文字的形式来记录。文字也是一个国家或民族文化的象征，在社会和历史的发展中有着特殊的地位。计算机从处理数据发展到处理文字信息，代表了应用技术上的一个重大进展，否则计算机的应用将局限在一个较狭小的范围内。文字信息处理的应用范围非常广泛，从编辑文稿、建立文件档案资料、排版印刷到行政管理、办公室自动化，凡是需要用文字表达信息的应用场合，都可以利用文字信息处理技术。随着个人计算机应用的普及，以这类计算机为基础构成的文字处理机目前已有了很大的发展。文字处理机依据其应用的不同要求，可以设计成不同的档次。使用最为普遍的一种是便携式的文字处理机，或称为电子打字机，其使用范围正在日益扩大。和传统的机械式打字机相比，电子打字机具有编辑功能丰富、灵活的独特优点，并且可以提供一定数量的文件存档，价格也在逐渐降低，今后有望能逐步取代机械式打字机。高档次的文字处理机更具有传统的机械式打字机无法比拟的优点。随着微型机性能和软件技术水平的不断提高，文字处理机的功能也会不断扩展。如高级的文字处理机可以利用计算机人工智能，在字、词处理的基础上增添语法和句法处理、书面和自然语言处理等新功能。随着高技术的开发和工业生产的发展，文字处理技术的推广应用前景是乐观的。

文字信息处理的实质，是先把文字信息数字化，即用一个固定的数码代表一个字母或文字。例如，在英文信息中，以 26 个字母作为文字信息处理的单位，因此要对 26 个字母逐个地确定代替它的数码。在汉字的情况下，一般是以一个整字作为文字信息处理的单位，因此要对每一个整字惟一地确定代表它的数码。这一数码统称为代码(code)。在计算机内部处理文字信息时，就像处理数据一样对待。处理完毕后，再把替代的数码还原成相应的字母或文字。利用计算机能够调整处理数据的性能，使文字信息处理也能够分享计算机技术的这一独特优点，从而实现文字信息处理的高效化。

计算机之所以能有较高的运算和处理能力，是由于它利用了电子处理技术以及二进制数运算这一法则。计算机中的运算器，利用半导体器件的两个状态(通和断)的变化，代表二进制数字串中的一个二进制数位上的“1”或“0”的变化，从而能够高速地执行二进制数的数值或逻辑运算。实际上，计算机无论做数值的或任何种类信息的运算或处理，最基本的运算操作就是这种二进制数的演算。