普通高等教育“十五”国家级规划教材

语言文字信息处理

盛玉麒 编著

山东大学出版社

普通高等教育“十五”国家级规划教材

语言文字信息处理

盛玉麒　编著

山东大学出版社

图书在版编目(CIP)数据

语言文字信息处理/盛玉麒主编．—济南：山东大学出版社，2006.5
ISBN 7-5607-3173-2

Ⅰ．语...
Ⅱ．盛...
Ⅲ．汉字信息处理—高等学校—教材
Ⅳ．TP391.12

中国版本图书馆 CIP 数据核字(2006)第 046374 号

山东大学出版社出版发行
(山东省济南市山大南路 27 号　邮政编码：250100)
山 东 省 新 华 书 店 经 销
莱芜市圣龙印务有限责任公司印刷
787×980 毫米　1/16　17.25 印张　313 千字
2006 年 5 月第 1 版　2006 年 5 月第 1 次印刷
定价：32.00 元

内容介绍

本书针对高校文科学生的知识背景和认知习惯，系统介绍了语言文字信息处理主要分支领域的基本原理、基础知识和基本技能，包括绪论、中文电脑基础、汉字编码原理、文稿编辑技巧、中文模式识别与语音处理、中文信息库、中文计算机辅助教学以及中文电子出版等八章，结合作者多年的研究成果，既有知识的介绍，又有现状的分析和研究课题的展望，可作为文科高年级学生的选修课教材或相关专业研究生的辅助教材或参考书，也可作为其他非计算机专业学生选修课教材或辅助教材。

信息是人类最伟大的发现

（代前言）

信息是构成世界的三大要素之一

在信息被发现之前，人们对世界的认识是二元的，即世界万物都是由物质和能量组成的。物质不灭，能量转换，周而复始，生生不息。根据对两者关系的认识，人们的世界观分为唯物主义和唯心主义两种。不管是唯物主义还是唯心主义，都是在物质与能量的二元范围之内。

信息被发现之后，人们逐渐认识到，信息是与物质和能量一起构成世界的三大基本要素之一。在所有的自然现象和社会现象中，信息都起着非常关键而特殊的作用。

物质的结构、大小、状态以及与其他物质之间的关系等信息，决定了该物质的性质和它所具有的能量，于是，决定了它的运动和存在的方式。在运动和变化中，都表现为信息数量和信息属性的变化。如果没有信息的传递和转换，就不会有所谓“物质不灭，能量转换”。

信息变化的结果导致物质的性质、状态的变化和能量的转化。金刚石和石墨是同素异构体，它们之间截然不同的物理属性，就是因为结构方式不同，也就是结构信息不同，这是“信息观”从理论上给出的科学合理的解释。

今天，人们仍然无法忘记20世纪末对电脑“千年虫”的恐惧。仅仅因为电脑系统中的一个“信息位”的问题，竟然会给金融、航空、航天、交通、商

业甚至家用电器系统安全造成巨大的隐患。问题还远不止于此，凡是使用电脑技术的地方，包括医院、矿井、冶金、物理化学反应工程、制造业等等，都面临千年虫问题的挑战。这是“信息”在向人类社会开的一个不大不小的玩笑。如果亿万财富竟会在瞬间化为乌有，如果庞大的设备、复杂的系统转眼之间一片混乱，人们是不是该认真思考一下：在物质和能量的世界中，是否还有一个更为重要、更能表现本质特征的存在呢？

信息改变了我们的观念系统

进化论所揭示的物种进化和演变的原因由于基因的发现而得到根本的解释，而遗传基因，就是DNA的组织和结构。现在，通过辐射等方式改变DNA结构，已经可以创造出许多新的“物种”。这项技术被广泛地运用在生物工程中，许多新品种的水果、蔬菜在普通人的餐桌上也不难见到。与其说这是生物科学技术的新发展，不如说是信息理论及其处理技术在生物科学中的应用。

爱因斯坦相对论的世界观是对牛顿力学世界观的否定和超越。问题在于：相对论究竟是客观的存在还是主观的认识？多维思维方式的客观基础表现在哪里？静止是相对的，物质运动也是相对的，在三维空间中如此，在多维空间中更是如此。一切都是相对的，只有相对是绝对的。这些令人费解的问题，都可以描述为不同信息场及其关系的变换问题，因此，都可以从信息系统的角度得到科学的解释。

现在人们常说“世界在变小”，完全是因为信息处理方式所引起的观念的改变。不同的交通工具曾经使每一个人都感受到了速度的作用。速度在改变时间的同时，也改变了空间。速度就是物体运动的方式和状态，这正是控制论创始人维纳给“信息”所下的权威定义的核心概念。不难发现，不同的工具和不同的处理方式，所有的发明和创造，都可以归结为通过对“速度”的改变而改变我们在时空范围中的生存方式和生存质量。

速度是时间和距离的函数，距离是空间方位的一维函数。世界万事万物都存在于时间和空间之中，因此，我们又可以说，速度就是世界万事万物的函数。不同的事物存在于不同的信息场中，各个信息场中的物体以不同的速度和状态表现各自的存在方式。速度和状态就是万事万物的信息参数，改变这些信息参数，就改变了一切。著名的摩尔定律充分证明

了这一点。

信息是控制物质和能量的重要符号

人类社会的经济活动是一个复杂性系统。众多政治经济学理论提出的商品、货币、价值等基本概念，实际上都是经济信息的不同载体形式。在财富表现为货币形式的时候，是价值信息的载体由实物向“通货”的一种转换，是发生在有形载体的不同形式之间的转换。马克思在商品价值中发现的“一般社会劳动”，实际上是经济信息的真正内涵——高度符号化的价值载体。今天，我们会发现，有哪一个百万富翁天天在数自己的钱袋呢？全世界的金融投资者和企业家们，几乎都在密切关注股市大盘上变化不定的数据信息，密切关注着政治经济领域中的各种新闻消息。瞬息万变的数据所代表的实际上就是投资者们的金钱和财富的交换和转移，政治社会生活甚至自然界动态的信息也许就直接或间接地影响着每个人的实际生活质量和生活方式。

某条高速公路的建设方案刚刚发布，就会立刻引起周边地区房地产的升值，所有居民的生活将因此有不同程度、不同方式的改变；同时，还会引来无数外部资金的关注，引起相关企业的效益和股价的改变，等等。

北京申奥成功的消息，不但改变了北京市的建设规划和建设速度，而且改变了周边许多地区的发展规划，也改变了无数企业的发展方向。这个信息究竟隐含了多少商机，实在无法预料。

因此，我们说，在信息时代，人们不再是通过屯积聚奇来积累财富，而是通过对信息的占有和控制来实现对物质和能量的占有与控制。

语言文字是人类社会最重要的信息载体

语言文字具有以下八个方面的特点：

- 逻辑思维的符号
- 认知交流的媒介
- 智力开发的工具
- 科教发展的基础
- 文化传承的载体
- 信息处理的重点

- 民族心理的寄托
- 国家主权的标志

其中,“信息处理的重点”正是信息时代语言文字功能的最大化。在多媒体信息网络时代,在图、文、声、像等所有的信息媒体中,我们社会和生活中绝大多数信息都是以语言文字的形式出现的。因此可以说,进入信息时代的根本标志就是语言文字信息的数字化。对于汉语来说,就是“电脑中文化”和“中文电脑化”。

语言文字信息处理是一门新兴学科

在社会信息化的进程中,语言文字信息处理成为应运而生的一门重要的多边缘新兴学科。

我国语言文字信息处理始于“748 工程”。三十年来经历了字处理、词语处理的探索发展,当前正进入语句和篇章处理的新阶段,这是汉语言文字信息处理的基础与应用研究综合发展的阶段。新的阶段面临新的挑战,这就是计算机处理汉语信息时所遇到的知识短缺问题。

由于汉语自身没有形态标志,加上汉语的研究历来是面向人的,而且是会说汉语的人,而不是面向机器的,所以,计算机处理汉语信息时许多已有的知识需要加工和重组,更多的还需要探索和挖掘。这就需要大量精通语言学与计算机信息处理的人才。过去的三十年中,计算机信息技术领域的专家学者们,在语言文字信息处理的研究上,表现出极大热情和执著的精神。他们对汉语汉字的研究给传统语言学注入了生机和活力。可惜的是,汉语言文字学对信息处理缺乏足够的关注和投入,成为影响学科发展的一个重要原因。另一方面,面向理科学生的计算机信息处理类教材,在很大程度上偏离文科学生的知识背景和接受能力,也影响了文科学生的理解掌握信息处理方面人才的培养。.

信息处理需要文理兼通的知识控制型人才

多边缘交叉学科需要具有多边缘学科知识的复合型人才,对文科学生来说,首先是文理兼通。因此,针对文科学生的知识背景和思维特点,结合当前学生普遍掌握一定电脑知识的实际情况,学习本课程应特别注重以下三点:

1. “两个系统”相结合

文科在知识表达、逻辑推导、认知习惯等方面都存在许多与理科不同的特点，因此，帮助文科学生学习和了解理科的思维习惯，从而优化自己的知识结构就显得十分重要；语言文字信息处理涉及到“人际系统”和“人机系统”的转换，要帮助学生了解和掌握两个系统之间信息处理方式方法的共性和个性。

2. 知识和技能相结合

复合型人才要具有理论知识和实践能力。因此，要把计算机信息处理知识的学习和技能训练结合起来，保证一定的上机实习时间。实习的内容与文科学生的专业学习、语言文字应用研究和将来从事文字工作等密切相关，例如文稿编辑的方法和技巧、模式识别与语音处理、中文信息库和语料库以及计算机辅助教学等。

通过学习让文科学生了解计算机能够做什么，从而知道如何根据需要充分利用计算机这一现代化工具，进一步还应知道自己能为计算机做什么。

3. 学习和探索相结合

素质教育和能力培养的一个重要方面是扩大学生的知识面。考虑到学科知识体系的系统性，本教材内容涉及到语言文字信息处理的许多方面。因课时有限，内容丰富，只有变知识积累型人才培养模式为知识控制型人才培养模式，努力让学生掌握控制知识的方法。实际教学过程中，不可能做到面面俱到，要针对学生的实际，突出重点，详略得当。

控制知识的目的在于更好地应用。应用的过程就是发现问题、分析问题和解决问题的过程，就是探索和创新的过程。要把学习和探索结合起来，把学习和研究结合起来，把学习和创新结合起来。

突破智能化中文信息处理“瓶颈”的重任，历史地落在成长中的一代人的肩上。希望这本书能够为文科学生步入语言文字信息处理的领域指明方向和路径，希望有更多的有文科背景的人能够参加到汉语言文字信息处理科研攻关的队伍中来，因为我们坚信——走的人越多，这条路就越宽广。

目 录

第一章 绪 论

教学大纲与教学建议

[目的] 概述语言文字信息处理的性质，特别是汉语言文字信息处理的特点，使学生了解学习本课程的目的、任务和学习方法，开阔视野，提高学习的积极性。

[重点] 信息和信息处理的基本概念，信息革命，信息处理方式与人类文明进步的关系，中文信息处理的系统结构。

[难点] 信息处理方式，信息处理系统的结构，汉语言文字信息处理的特点。

[教学建议]

1. 突出汉语言文字信息处理的特点。

2. 针对学生的知识背景，做好内容剪裁和教学设计，注意详略得当，把学生的表面知识引向深入。

3. 注重理论与实践相结合。

[课时建议] 讲解 2 课时。

语言文字信息处理是以语言文字学为基础，以计算机和远程通信为核心技术的一门多边缘交叉的新兴应用型学科。计算机技术从数值运算到非数值运算的发展是催生语言文字信息处理学科诞生的关键，远程通信技术则是保证语言文字信息处理实现社会化效应的核心。这一学科还广泛涉及到语言学、计算机科学、通信技术、信息科学以及信息论、系统论、控制论、逻辑学、哲学、心理学、数

学、电子学、仿生学、脑神经科学、声学、自动化技术和人工智能等许多边缘学科。

以计算机和远程通信为核心的现代科学技术的发展，从根本上改变了人们的“时空”环境。20世纪末，西方未来学家用“第三次浪潮”、“计算机革命”、“信息化社会”、“信息革命”、“信息爆炸”等词语来表述当代社会的特点。我们不去推敲各个词语的内涵及其差异，只需强调一下被普遍发现并承认的事实，这就是“信息”和“信息处理”的确已经成为当代社会众多领域谈论的热门话题。

“信息”不仅由一个新生词进入到常用词的行列，而且已经成为一个具有很强构词能力的常用词。“信息化”、“信息场”、“信息源”、“信息量”、“信息网”、“信息员”、“信息社会”、“信息科学”、“信息技术”、“信息处理”、“信息资源”、“信息交换”、“信息传输”、“信息反馈”、“信息载体”及“经济信息”、“商品信息”、“管理信息”、“股市信息”、“市场信息”、“遗传信息”等等，魔术般地出现在当代社会生活的各个领域。

由于“中文”(汉字)本身所具有的特殊性，在“时”、“空”、“量”(单位数量和使用者人数)等方面堪称“世界之最”。因此，自“748工程”提出中文信息处理科技攻关项目以来，三十年的时间里，汉语言文字信息处理已经发展为一个成熟的新兴学科，在基础研究和应用研究方面都取得了一系列举世瞩目的成就，表现出强大的生命力和多彩多姿的无穷魅力。本书将与大家共同探索和领略这门新兴多边缘交叉学科的神奇与风采。

第一节　信息与信息处理系统

教学大纲与教学建议

[目的]　了解信息与信息处理的基本概念、性质。

[重点]　信息的性质，信息的类型，信息处理系统。

[难点]　信息的性质，信息处理系统。

[教学建议]　结合现代社会政治、经济、文化生活的例子，突出语言文字的重要信息属性。

一、信息处理的基本概念

(一)信息

信息是客观物质世界存在的形式、状态及各种关系，是与物质、能量共同构