

智能技术与 信息处理

凌云 王勋 费玉莲 著

- ▶▶ 智能信息处理领域发展的**最新技术**
- ▶▶ 具体应用领域中的**数据仓库**和数据挖掘
- ▶▶ 智能搜索和智能决策支持系统等**最新发展方向**



科学出版社
www.sciencep.com

智能技术与信息处理

凌 云 王 勋 费玉莲 著

科 学 出 版 社

北 京

内 容 简 介

本书是信息处理技术领域的一部力作。全书内容紧凑、连贯,内容新颖,理论性较强。旨在让读者通过阅读本书,了解信息处理与智能技术的基本知识和前沿应用。本书首先讲述了信息处理和智能技术的基本理论,然后重点讲解了智能化推理技术、数据仓库和数据挖掘技术,并对智能决策支持系统和智能计算机等前沿性应用进行了探讨。为了便于读者对自己感兴趣的部分作更深入的研究,每章后面都列出了相关的参考文献,读者可以自己查阅。

本书可作为计算机和信息系统专业研究生的教材,也是这些领域的专业工作者知识更新的参考资料。

图书在版编目(CIP)数据

智能技术与信息处理/凌云,王勋,费玉莲著.—北京:科学出版社,2003
ISBN 7-03-011255-5

I.智...II.①凌...②王...③费...III.信息处理—信息技术 IV.G202

中国版本图书馆(CIP)数据核字(2003)第015101号

责任编辑:李佩乾 / 责任校对:都岚

责任印制:吕春珉 / 封面设计:立日新设计公司

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003年4月第 一 版 开本: B5(720×1000)

2003年4月第一次印刷 印张: 16

印数: 1—2 000 字数: 303 000

定价: 25.00 元

(如有印装质量问题,我社负责调换〈路通〉)

序

在世界进入信息化社会的高速发展过程中，我们回顾一些值得思考的历史：

首先，信息化的基础技术不是线性常速，而是指数式加速前进：决定了存储芯片容量和处理器芯片处理能力的集成电路芯片上晶体管的集成度，按照 1965 年提出的 Moore 定律，每 18 个月翻一番，即一年增加 60%。其他(如计算机硬盘容量)也以每年 50% 的增量增长。因特网兴起后通信设备的带宽发展一直也能够跟上网民的增加。软件在和硬件的相互推拉中，也是同步增长，20 世纪 80 年代的字处理软件只占几万字节，现在已要数千万字节。

经济学家观察到了—个良性循环：技术进步导致更好的产品和更低的价格，由此可能开发出新的应用，而新应用开辟的新市场吸引许多新的公司参与竞争，而竞争又导致对技术创新的渴求。当然良性的新市场必须要能带来大量的新用户群体。

1946 年出现的第一台电子数字计算机 ENIAC，具有自动高速计算能力，但它是一个昂贵的庞然大物，只能让它承担重要的国防和工程计算任务。此后数年在少数大学和研究所内相继建造出来的计算机，也都是如此。直到 20 世纪 50 年代中期，当企业中的数据处理或信息处理的市场被打开后，计算机才由专业工厂大批量地生产。信息处理变成计算机的主导，多数的应用与信息产业和研发在良性循环中互相促进，使信息产业的发展在发达国家超越了传统工业，也开发出丰富的信息处理应用技术，使信息技术和广大群众的生活息息相关，它极大地提高了社会生产力和个人生活品质，开始了社会信息化的进程。当然还要提到 80 年代的个人计算机 PC 和 90 年代的因特网兴起所带来两场群众潮对加速信息化的决定性的推动：PC 使计算机从庭院深处的计算中心中解放出来，使每个人都可拥有，而因特网则使个人可以访问和使用分布在全世界的无数计算机资源。

人工智能的研究也经历了曲折的历程，它研究用计算机的智能程序模拟人的智能，这是一个宏伟目标，在实验室小范围内也获得一些可喜的成功，但是面对复杂的实际世界，此后十多年始终没有突破性的进展。直到 70 年代，一些学者转而针对有限范围内的实际应用，运用人工智能问题求解和知识推理等技术，开发出模拟专家在特定领域内决策能力的“专家系统”，取得了有实用价值的成功，才使人工智能再度引起人们的普遍兴趣，而七八十年代专家系统也在各行各业中遍地开花。专家系统中的核心技术——知识处理也成为人工智能其他研究领域(如语音识别、图像识别、自然语言理解、机器人规划等方面)一个有力的工具。90 年代知识处理更进一步与已经十分繁荣的信息处理相结合成为智能信息处理技

术,更高效地提高了信息处理的自动化水平,这将使人工智能应用更上一层楼。

本书的作者杭州商学院计算机与信息工程学院的凌云等老师,长期从事信息处理和人工智能的教学和科研,对于这两个学科有深刻的了解和体会,因此他们能融会这两个方面写出本书。本书在扼要地介绍了智能技术的基础以后,重点论述了近年来出现的一些新的重要的信息系统,如数据仓库、数据挖掘、决策支持和信息搜索等除了阐述它们的具体实现原理以外,还分析了智能技术的运用以提高信息处理的能力和效率,最后展望了智能系统的结构和智能计算机的发展前景。本书内容涉及面广,可作为计算机和信息系统等专业研究生的教材,也是这些领域的专业工作者知识更新的有价值的参考书。

何志均

2002年11月20日

前 言

近年来, 计算机网络特别是 Internet 和 WWW 的迅猛发展和广泛应用, 为人工智能提供了广阔发展的新天地。信息化需要智能化的支持, 一个智能化信息处理的新时代正向我们走来。

本书从人工智能及信息处理两大主题出发, 在系统全面地介绍了人工智能的基本理论、方法及其实现技术的基础上, 着重于智能信息处理领域发展的最新技术, 结合具体的应用领域, 介绍了数据仓库和数据挖掘的基本理论和最新的进展, 智能搜索和智能决策支持系统的基本方法及发展方向, Agent 理论及计算机技术的智能化。

全书共分为 8 章, 涵盖了智能信息处理技术的各个主要方面。

第 1 章简要地介绍了信息处理技术、智能理论和智能信息技术, 包括信息处理技术的一般理论、中文信息处理技术、经典人工智能理论、计算智能、智能网技术、智能计算机技术及以 Agent 为代表的新一代智能技术的应用。通过本章的讨论希望读者能对智能理论有一个概括的了解, 并建立起一些基本概念。

第 2 章是关于经典人工智能中知识的表示问题, 对各种主要的表示方法进行系统地介绍, 包括谓词逻辑表示法、产生式系统、框架理论、语义网络及面向对象的知识表示方法。

第 3 章阐述计算机要具有某种程度的智能, 重要的是要具有一定程度的思维能力, 即运用已掌握的知识来推理出新知识的能力, 以使问题得到解决。本章在介绍了智能化推理技术中推理的一般概念的基础上, 讨论了经典逻辑推理的各种技术、搜索的基本策略及不确定性和非单调推理理论。

第 4 章介绍随着计算机应用及 Internet 的日益普及, “丰富的数据与贫乏的知识”问题也日见突出, 人们期待着从堆积如山的数据中得到自己想要的答案。本章在介绍了数据仓库和数据挖掘的一般概念的基础上, 给出它们的基本结构和方法, 详细地讨论了数据仓库中挖掘关联规则的方法、数据的分类和预测方法、聚类分析方法、基于 Web 的挖掘、可视化挖掘及基于神经网络的挖掘方法等。

第 5 章首先阐述了信息搜索的基本原理和技术, 接着对网络信息的搜索技术作了全面系统的介绍, 然后对智能搜索引擎所涉及的自然语言理解、对称搜索技术、基于 XML 的技术等进行了展望, 最后对当前的热点图像和动态网页的搜索进行了讨论。

第 6 章系统和科学地阐述了决策支持系统的一些基本概念, 提出了结合数据仓库、数据挖掘和知识库系统而形成的综合集成决策支持系统, 进而向更高形式

的智能化方向发展。

第 7 章针对 AI 在方法和技术上还只能处理那些较“成型”的问题，而对现实环境的变化无能为力时，引入 Agent 和多 Agent，以试图走出一条 AI 的研究路线，从而在智能 Agent 的水平上实现 AI 学科的核心内容与最终目标。本章将对 Agent 理论、多 Agent 求解机制和有关技术作较系统的讨论。

随着计算机技术和人工智能研究的不断深入，人们对于计算机的要求也不断提高，建造能模拟人类智能行为的智能计算机系统成为人类的梦想。为此，包括我国在内的许多国家都开展了新一代计算机的研究工作，取得了一定的进展。第 8 章在讨论了新一代智能计算机有关概念的基础上，介绍几种智能计算机的研究进展，提出几个未来智能计算机系统可能的发展方向。

全书以信息处理技术的智能化为中心，对具体的技术按其地位、作用及与其他技术间的联系安排讨论的篇幅，力求能简明地阐述清楚智能技术新理论和新方法，跟踪各研究领域发展的新趋势。

目 录

第 1 章 信息处理与智能技术.....	1
1.1 信息与信息处理.....	1
1.1.1 信息的定义	1
1.1.2 信息处理	1
1.1.3 信息技术	2
1.1.4 中文信息处理	4
1.1.5 信息处理向知识处理过渡.....	6
1.2 智能理论	7
1.2.1 智能理论概述	7
1.2.2 经典人工智能	8
1.2.3 计算智能	9
1.2.4 其他智能理论	10
1.3 智能信息技术	12
1.3.1 智能网技术	12
1.3.2 网络管理和控制技术智能化.....	13
1.3.3 网络信息搜索智能技术.....	14
1.3.4 智能计算机技术.....	16
1.3.5 智能控制技术	19
1.3.6 智能 Agent 技术及其在网络搜索中的应用	20
第 2 章 信息的智能化表示	22
2.1 综 述.....	22
2.2 一阶谓词逻辑表示法.....	22
2.2.1 一阶谓词逻辑的理论基础.....	23
2.2.2 基于一阶谓词逻辑的知识表达.....	27
2.3 产生式表示法	28
2.3.1 产生式规则的基本形式.....	28
2.3.2 产生式系统	28
2.3.3 产生式系统的推理过程.....	30
2.4 框架表示法.....	31

2.4.1 知识的框架表示.....	31
2.4.2 框架推理	32
2.4.3 框架与规则的结合.....	32
2.5 语义网络表示法.....	33
2.5.1 语义网络的知识表示.....	33
2.5.2 常用的语义联系.....	33
2.5.3 语义网络的推理.....	34
2.5.4 语义网络知识表示的特点.....	35
2.6 面向对象表示法.....	35
2.6.1 面向对象的基本概念和特征.....	36
2.6.2 面向对象的知识表示.....	37
第3章 智能化推理技术.....	39
3.1 基本概念.....	39
3.1.1 什么是推理	39
3.1.2 推理的发展概述.....	39
3.1.3 推理的控制策略.....	40
3.2 经典逻辑推理技术.....	43
3.2.1 自然演绎推理	43
3.2.2 归结演绎推理	45
3.2.3 基于规则的演绎推理.....	52
3.3 搜索的基本策略.....	58
3.3.1 状态空间表示的搜索策略.....	58
3.3.2 与/树表示的搜索策略.....	70
3.3.3 博弈树搜索策略.....	76
3.4 不确定与非单调推理.....	81
3.4.1 什么是不确定性推理.....	81
3.4.2 不确定性推理中的基本问题.....	82
3.4.3 不确定性推理方法.....	85
3.4.4 什么是非单调推理.....	87
3.4.5 非单调推理方法.....	88
第4章 数据仓库和数据挖掘.....	91
4.1 数据仓库概述	91
4.1.1 数据仓库的起源.....	91
4.1.2 数据仓库的定义.....	92

4.1.3 数据仓库的特征.....	92
4.2 数据仓库构建和优化.....	94
4.2.1 数据仓库开发生命周期.....	94
4.2.2 数据仓库的数据结构.....	96
4.2.3 高性能数据仓库平台的建立.....	99
4.3 数据仓库管理和维护.....	105
4.3.1 数据的更新和复制.....	106
4.3.2 数据源的同步化.....	107
4.3.3 故障恢复.....	107
4.3.4 访问控制与安全性.....	107
4.3.5 数据增长的管理.....	108
4.3.6 数据库性能的管理.....	108
4.3.7 数据仓库的增强与扩充.....	109
4.4 数据挖掘概述.....	110
4.4.1 数据挖掘概念.....	110
4.4.2 数据挖掘的必要性.....	111
4.4.3 数据挖掘的发展前景.....	112
4.5 数据挖掘技术简介.....	113
4.5.1 分类和预测.....	113
4.5.2 基于关联规则的挖掘.....	116
4.5.3 聚类分析技术.....	121
4.5.4 基于 Web 的挖掘.....	125
4.5.5 其他挖掘技术.....	128
4.6 数据仓库与数据挖掘.....	130
4.6.1 数据挖掘和数据仓库的关系.....	130
4.6.2 基于数据仓库的数据挖掘.....	131
第 5 章 信息搜索及其智能化.....	135
5.1 信息搜索概述.....	135
5.1.1 基本概念.....	135
5.1.2 存在的问题.....	137
5.2 搜索引擎概述.....	138
5.2.1 搜索引擎发展历程.....	138
5.2.2 搜索引擎的工作原理.....	140
5.2.3 搜索模型.....	144

5.2.4 搜索引擎的分类.....	147
5.2.5 搜索引擎的现状及面临的挑战.....	149
5.3 智能搜索引擎.....	151
5.3.1 智能搜索引擎特征.....	151
5.3.2 智能搜索引擎技术.....	154
5.4 图像和动态网页的搜索.....	159
5.4.1 图像信息的搜索.....	159
5.4.2 动态网页的搜索.....	162
第 6 章 智能决策支持系统	165
6.1 决策支持系统概述.....	165
6.1.1 决策支持系统的基本概念.....	165
6.1.2 决策支持系统的形成和发展.....	168
6.1.3 智能决策支持系统.....	170
6.2 智能决策支持系统的体系结构.....	171
6.2.1 智能决策支持系统的基本构件.....	172
6.2.2 智能决策支持系统的系统结构.....	177
6.3 智能决策支持系统集成技术.....	180
6.3.1 模型生成技术和模型管理.....	180
6.3.2 模型管理和数据管理的接口技术.....	183
6.3.3 知识库与数据库的结合技术.....	184
6.4 智能决策支持系统的建造和实现.....	185
6.4.1 人机界面的设计与实现.....	185
6.4.2 数据库的设计与实现.....	186
6.4.3 模型库的设计与实现.....	188
6.4.4 知识库的设计与实现.....	190
第 7 章 Agent 与多 Agent	193
7.1 智能体概述.....	193
7.1.1 基本概念.....	193
7.1.2 Agent 理论.....	194
7.1.3 Agent 系统结构.....	197
7.2 Agent 的分类与应用.....	199
7.2.1 接口型 Agent.....	200
7.2.2 流动型 Agent.....	201
7.2.3 信息型 Agent.....	203

7.3 多 Agent 系统.....	205
7.3.1 多 Agent 系统结构.....	205
7.3.2 多 Agent 特点.....	207
7.4 多 Agent 的通信机制.....	208
7.4.1 本体论.....	208
7.4.2 KIF 格式.....	209
7.4.3 KQML.....	210
7.4.4 SHADE.....	211
7.5 多 Agent 间的协商与协调.....	213
7.5.1 多 Agent 的协商.....	213
7.5.2 多 Agent 的协调.....	214
第 8 章 智能计算机.....	218
8.1 概述.....	218
8.1.1 智能计算机概念.....	218
8.1.2 智能计算机体系结构研究.....	219
8.2 人工神经网络计算机.....	221
8.2.1 神经计算机系统背景及研究现状.....	221
8.2.2 神经网络计算机的体系结构.....	223
8.2.3 电子神经网络计算机.....	224
8.2.4 光神经计算机.....	228
8.2.5 总结.....	231
8.3 网格计算机系统.....	232
8.3.1 网格的起源与发展.....	232
8.3.2 网格的关键技术问题.....	233
8.3.3 研究现状.....	236
8.4 生物计算机.....	236
8.5 纳米计算机.....	238
8.5.1 纳米计算机的分类.....	238
8.5.2 纳米级的二态器件.....	240
8.5.3 最新研究进展.....	241

第 1 章 信息处理与智能技术

1.1 信息与信息处理

1.1.1 信息的定义

什么是信息？像其他司空见惯的重要术语一样，迄今为止还没有一个公认的定义。人们常常从“自身”关注的角度给予各种限定和概括，从不同的角度产生了 3 种不同的信息概念：

- “技术信息”概念，认为信息是物质属性的反映。
- “语义信息”概念，认为信息是人们为适应外部世界，并同外部世界进行交换的内容的标记。
- “价值信息”概念，认为信息是具有价值性、有效性、经济性及其他特性的知识。

人们在实践中发现，信息的含义比从各个角度定义包含的内容广泛得多，一切含有内容的信号都是信息。即用语言、文字、符号、声音和图像等形式所表示出来的事物的特征，如人的名字、报纸上的新闻和电视上的画面等都是信息。并且信息应具有可识别性、可转换性、可传递性、可处理性和可存储性。

由于信息的含义非常广泛，人们每天都离不开信息，可以说，人类是生活在信息的海洋中。因此，对信息的处理也就自然成为人们日常生活中最常见的工作之一。

1.1.2 信息处理

信息处理（Information Processing）也称数据处理（Data Processing），是计算机应用的一个重要方面。最初是指在计算机上对商业、企业的信息和数据进行加工。随着计算机应用的广度和深度不断发展，现在常用来泛指计算机在非科技工程方面的所有计算、管理和各种应用，如企业管理、库存管理、账目计算和信息情报检索等，几乎无所不包的应用都认为是信息处理的范畴。下面从信息本身的角度而不是应用范畴的角度对信息处理含义作一概括。

信息处理是指用某种手段获取原始信息之后，采用一定的方法和设备，按一定的目的和步骤对信息进行加工，使之转变成能被利用的有效信息过程的总称。可见，对信息的处理应包括信息的采集、储存、加工及传输。

- 信息的采集

信息采集是整个信息处理的第 1 个环节，也是以后各种处理的基础。信息采集过程中要解决的第 1 个问题是信息的识别，也叫确定信息需求。这是因为想得到关于客观情况的全部信息往往是不可能的，有时也是不必要的，因此必须对信息进行识别和筛选；在信息识别后，下一步就是信息的采集，根据不同的目标需求可选择与之相适应的信息采集方法；信息采集的最后一个问题是信息的表达。信息的表达目前一般有 3 种形式：一种是文字或符号表达，另一种是数字或编码表达，再一种就是声像方式表达。

- 信息的存储

信息存储是将信息暂时或长期保存起来，以备需要时应用。信息存储的主要问题是确定需要存储哪些信息，存多长时间，以什么方式存储，以及存储在什么介质上等。

- 信息的加工

信息加工是信息处理的中心环节，它的任务是依据某项任务的要求，以计算机为工具对信息进行鉴别、选择、比较、分类、计算和编写等工作，使之成为有用的信息。

- 信息的传输

信息的传输是利用一定的装置和设备，实现信息的有目的流动，以满足对信息的需要。

一般来说，人本身也具备了一定的信息处理能力，例如，学生在完成作业时，首先将老师布置的作业抄到作业本上，然后进行分析与解答，最后把解题过程写在作业本上，其处理过程可归结为 3 个主要步骤，如图 1.1 所示。

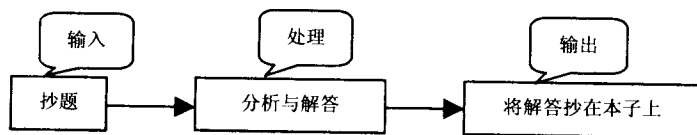


图 1.1 信息处理过程

1.1.3 信息技术

信息技术是一种综合技术，是一个多层次、多专业的复杂矩阵技术体系，它的发展是由许多单元技术所支撑的。具体地说，信息技术是指感测技术、通信技术、智能技术（包括计算机硬件、软件、人工智能）和控制技术的整体。其中，感测技术包括传感技术和测量技术，如遥感和遥测技术等，它们是使人们可以更好地从外部世界获取各种有用的信息的技术。通信技术就是传递信息的技术。智

能技术包括计算机硬件技术、软件技术、人工智能技术和人工神经网络等,能更好地加工和再生信息。而控制技术则能根据输入的指令信息(决策信息)对外部事物的运动状态和方式实施干预,即利用信息的技术。因此这个定义不但给出了信息技术的内容,也明确了信息技术的获取、传递、处理和利用的体系,还明了感测、通信、计算机和控制等概念比较明确,领域比较清晰,大众比较有感性认识的技术在信息系统中的作用和相互关系。

信息系统中的感测、通信、智能和控制这4大技术虽然各司其职,但也是一个有机的整体,它们和谐有机地合作,共同完成扩展人的智力功能的任务。它们之间的关系如图1.2所示。

由图1.2可见,信息系统的4大技术是相互包含、相互交叉和相互融合的,但从技术要素层次上看,它们的作用并不在相同层次上,通信技术和智能技术处在整个信息技术的核心地位,而感测技术和控制技术则是核心与外部世界之间的接口。没有通信和智能技术,信息技术当然就失去了基本的意义,而没有感测技术和控制技术,信息技术同样也是没有意义的:一方面没有信息的来源,另一方面,信息又没有归宿。可见,信息系统的4大技术是一个完整的体系。

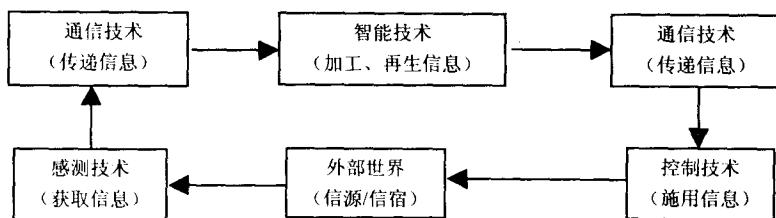


图 1.2 信息系统

处于核心地位的智能技术的高速发展带动了整个信息技术的高速发展。事实上,在计算机技术产生之前,遥感技术、通信技术和控制技术就已经产生了。但那时这些技术的水平还是比较低的,很多操作还需要人工进行。计算机技术产生以后,遥感技术、通信技术和控制技术的水平得到了极大的提高,不仅自动化水平不能与过去同日而语,而且通过程序控制实现了越来越强大、越来越复杂、越来越便利、越来越高效的功能和服务。可以说,当前信息技术的基本特征就是计算机程序控制化。

程序控制化的优点是:第一,能够可靠地长期运转。因为一个程序一旦正确地设计完成,就可以一劳永逸地反复执行,完成指定的功能或服务,可靠性远远超过人工。第二,技术进步快。任何复杂的功能强大的程序都是由多个简单的、功能单一的程序组成的,程序的不断开发就自然地积累下来,而程序的积累就会直接带来技术的进步;并且技术更新往往只是软件的更新,花费的成本和代价较小。第三,便于构造大系统。大系统可以被分解为众多的子系统来构造,子系统

内部靠内部程序控制, 子系统之间靠数据或协议来联系和协调。依靠这种方式, 一个覆盖全球的大信息系统也不难建立。

正是由于这些优点, 自计算机技术产生以来, 信息技术便有了突飞猛进的进步。它们之间相互促进、渗透、覆盖和影响, 必将促使信息高技术急速发展, 进而对整个社会、经济和军事等一切方面的发展产生巨大而深远的影响。人们越来越清楚地看到, 一些发达国家正是依靠先进信息科学技术的推动, 把自己国家从后工业化时期逐步推向先进的信息时代。

虽然当今的信息技术已经给人们带来了不曾奢望的利益, 但技术前进的脚步是不会停止的。在面向 21 世纪的技术变革中, 信息技术的发展方向将是智能化。

当今的信息技术在某些方面已经超过了人脑在信息处理方面的能力, 如记忆能力、计算能力等, 但在许多方面, 却仍然逊色于人脑, 如文字识别、语音识别、模糊判断和模糊推理等等。尤其重要的是, 人脑可以通过自学习、自组织和自适应来不断提高信息处理的能力; 而存储程序式计算机的所有能力都是人们通过编程赋予给它的, 与人脑相比是机械的、死板的和无法自我提高的。

针对以上问题, 人们从多年以前就开始研究智能理论与技术, 探索人脑信息处理的机制, 以使用机器更好地模拟人脑的功能。通过几十年的努力, 智能理论与技术已经取得了很大进展。正是这些进展为信息技术的智能化提供了基础。

1.1.4 中文信息处理

1. 中文信息处理现状

计算机中文信息处理技术从 20 世纪 70 年代蓬勃发展至今, 仅仅经历了短短 20 多年的时间, 便完成了由初级阶段向比较成熟阶段的过渡, 这是微电子技术和 IT 技术高速发展以及迫切的应用需求所促成的。

中文信息处理技术处于初级阶段的主要特征是以研究为主。在 20 世纪 70~80 年代, 国内曾出现汉字输入方法研究千军万“码”的局面, 上千种输入方法应运而生。在汉字字型方面, 从 15×16 点阵到 256×256 点阵, 仿宋、宋、楷、黑等各种字体不断涌现, 以 CCDOS 为代表的 20 余种汉化 DOS 不断出台, 各具特色, 联想汉卡、巨人汉卡、四通汉卡等曾风靡一时。

从 20 世纪 90 年代初开始, 中文信息处理技术开始进入比较成熟的阶段。其主要标志是初步实现了标准化、一体化、工程化和产品化。

- 标准化

指国家相应出台了一系列有关中文信息处理方面的标准。

- 一体化

指中文信息处理多项技术实现了有机、合理的结合。

- 工程化和产品化

指中文信息处理解决了在大规模应用、大规模生产以及市场营销中出现的问题。

经过 20 多年的努力,我国在中文信息处理方面已取得了十分可喜的成绩,在某些方面的研究已处于世界领先地位。如北大方正的激光照排技术,其市场份额独占鳌头。汉王公司的手写汉字识别技术及产品,在经历了 1996 年和 1997 年的“春秋争霸”之后,超过了实力强大的摩托罗拉的“慧笔”和台湾的“蒙恬”,占据了最大的市场份额。另外,由国家语言工作委员会和二炮二所研制的 7 000 万字现代汉语语料库也已基本完成,首批将颁布 2 000 万字语料库投入社会应用。国家 863 和国家“九五”预研工作在智能中文人机接口、跨平台中文处理、中文虚拟平台、中文浏览器和中文文本处理等方面的研究也都取得了令人鼓舞的进展。

在看到这些成绩的同时,我们还必须十分清醒地看到存在的差距和问题。如在语音输入方面,IBM 公司开发的 Viavoice 产品占据了主要地位;国外公司利用我国“海外兵团”开发的中文产品不断涌入国内;微软和 IBM 公司分别在中国成立了研究院和研究中心,广揽人才,其主要研究方向就是中文信息处理。

另一方面,各类具有广阔市场的信息产品的研发面临着中文信息处理的新问题,包括中文显示、中文输入和中文语音识别、手写识别等。我国互联网用户和应用的高速发展,也将促使中文信息处理方面的高效中文搜索引擎、实时机器翻译、大规模中文文本处理、跨平台中西文自动识别转换、泛中文语义理解、中文电子商务等技术实现重大突破。

2. 中文信息处理新技术

新的电子设备和新的市场需求,促进了中文信息处理诸多新技术的发展。

- 汉字字库

汉字字库主要包括点阵、矢量和曲线轮廓 3 种。经过多年的努力,在国家信息标准委员会推出相应点阵字型标准之后,造字工作趋于平稳。但随着移动电话、POS 等电子产品的出现,市场急需小于 15×16 点阵的小点阵显示字库。由于在小点阵下的某些汉字难免缺笔画,因此,如何确保用户既可识别又不失汉字的规范,有许多问题需要解决。此外,随着多媒体、电子印刷技术的发展,市场还需要各种字型的高质量曲线字。

- 数字输入方法

目前,数字码输入技术大体可分为 3 类,即基于笔画的、有基于拼音的、音码与形码结合的。与大键盘相同,数字小键盘的编码主要源于“音、形、义”三方面,由于“义”缺乏直观性,多用于智能处理;“音”码