CENTURY 21

高等学校教材

智能数据挖掘技术

薛惠锋 张文宇 寇晓东 编著

西北工业大学出版社

高等学校教材

智能数据挖掘技术

[illegible]

智能数据挖掘技术

薛惠锋　张文宇　寇晓东　编著

西北工业大学出版社

【内容简介】 数据挖掘是人工智能、机器学习、数据库技术等多学科相结合的产物，是由计算机自动从已有数据中发现以前未知的、具有潜在应用价值的信息或模式的技术。数据挖掘是知识发现过程的重要内容。本书系统地介绍了数据挖掘技术的原理、方法和计算技术，可作为系统工程、控制工程及计算机类专业研究生的教材，也可供相关专业技术人员参考使用。

图书在版编目(CIP)数据

智能数据挖掘技术/薛惠锋，张文宇，寇晓东编著．西安：西北工业大学出版社，2005.2

ISBN 7-5612-1891-5

Ⅰ.智… Ⅱ.①薛… ②张… ③寇… Ⅲ.数据采集—研究生—教材 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2005)第 009766 号

出版发行：西北工业大学出版社

通信地址：西安市友谊西路 127 号　　**邮编**：710072

电　　话：029-88493844　88471757

网　　址：www.nwpup.com

印 刷 者　陕西丰源印务有限公司

开　　本：850 mm×1 168 mm　1/32

印　　张：8.625

字　　数：208 千字

版　　次：2005 年 2 月第 1 版　2005 年 2 月第 1 次印刷

定　　价：16.00 元

前 言

当今世界正处在信息爆炸的时代，人类在如日中天的信息化浪潮中开启了21世纪的现代化之旅，信息技术以其在各个领域应用中的强大渗透性和高附加值成为时代发展的强劲动力。作为人类智慧的象征，智能技术是信息技术发展中最具诱惑力的研究前沿，知识发现(KDD)是其重要的研究课题之一。

如何在浩瀚的信息海洋中获取有用的信息是困扰人们有效利用信息资源的“瓶颈”问题。数据挖掘(Date Mining)就是要从已有的海量数据中发现以前未知的、具有潜在应用价值的信息和模式，解决高噪声数据与低音量知识的矛盾。它是人工智能、数据库技术、机器学习等多种技术的结合，是知识发现过程的关键步骤。从数据挖掘的基本概念和原理，到挖掘方法、算法及软件工具，许多学者、研究部门和IT公司都进行了广泛深入的研究，形成了较为完整的数据挖掘理论与方法体系，并出现了一些实用的数据挖掘工具。在石油勘探、金融、商业销售、产品制造、医疗、保险、化工等领域，这些研究成果得到广泛应用，取得了巨大的经济效益。

作为一个非常具有活力的研究领域，数据挖掘技术的知识在不断地翻新。作者虽涉足此领域研究多年，但也只是对其有了一个初步的把握。本书将作者了解的数据挖掘的有关技术及自己的研究成果融合成一个易于理解的体系呈现给大家。

书中第1章先介绍了人工智能的基本概念及研究领域，然后是知识发现和数据挖掘的概念与结构，以及相关的问题。第2章介绍了数据仓库的基本概念、结构和设计，以及工程规划方面的内

容。从第 3～7 章开始，详细介绍了数据挖掘的相关技术，即是：基于概率统计与神经网络的数据挖掘技术、基于信息论的数据挖掘技术、基于关联规则的数据挖掘技术、基于分类规则的数据挖掘技术、基于聚类规则的数据挖掘技术。第 8 章和第 9 章是数据挖掘应用方面的内容，分别介绍了基于 Web 的数据挖掘技术和基于数据挖掘的智能决策研究框架。

本书在写作过程中，参考了众多同行专家的论著，西北工业大学资源与环境信息化工程研究所的荣群山博士对书稿的完成也做了大量工作，在此一并向他们表示感谢。

由于我们时间仓促，才疏学浅，加之数据挖掘是一项新技术，书中难免有错误和不当之处，恳请广大读者批评指正。

作者

2004 年 10 月

目 录

第1章 绪 论

半个多世纪以来，人工智能(AI)获得了很大发展并引起众多学科的日益重视，成为一门具有广泛应用的交叉学科和前沿学科。

1.1 人工智能综述

1.1.1 人工智能的发展历史

人工智能是计算机科学、控制论、信息论、神经生理学、语言学等多种学科互相渗透而发展起来的一门学科。人工智能的发展虽然已走过了40多年的历程，但是，人工智能至今尚无统一的定义。尽管学术界有各种各样的说法和定义，但就其本质而言，人工智能是研究、设计和应用智能机器或智能系统，来模拟人类智能活动的能力，以延伸人类智能的科学。人类智能活动的能力是指人类在认识世界和改造世界的活动中，经过脑力劳动表现出来的能力。一般地说，可概括如下：

(1)通过视觉、听觉、触觉等感官活动，接受并理解文字、图像、声音、语言等各种外界信息，这就是认识和理解外界环境的能力。

(2)通过人脑的生理与心理活动以及有关的信息处理过程，将感性知识抽象为理性知识，并能对事物运行的规律进行分析、判断和推理，这就是提出概念、建立方法、进行演绎和归纳推理、做出决策的能力。

(3)通过教育、训练和学习，日益丰富自身的知识和技能，这就

是学习的能力。

(4)对不断变化的外界环境(如干扰、刺激等外界作用)能灵活地做出正确的反应,这就是自适应能力。

不论从什么角度来研究人工智能,都是通过计算机等现代工具来实现的。计算机科学与技术的飞速发展和计算机应用的日益普及,为人工智能的研究和应用奠定了良好的物质基础。人工智能的发展使计算机更聪明、更有效,与人更接近。

1. 人工智能的起源

自古以来,人类对人工智能就有持久、狂热的追求,并凭借当时的认识水平和技术条件,设法用机器来代替人的部分脑力劳动,用机器来延伸和扩展人类的某种智能行为。例如,公元前 900 多年,我国就有歌舞机器人传说的记载。12 世纪末至 13 世纪初,西班牙的一位神学家和逻辑学家曾试图制造能解决各种问题的通用逻辑机。17 世纪,法国物理学家和数学家巴斯卡(B. Pascal)制成了世界第一台会演算的机械加法器并获得实际应用。随后,德国数学家和哲学家莱布尼兹(G. W. Leibniz)在加法器的基础上发展并制成了进行全部四则运算的计算机,他还提出了逻辑机的设计思想,即通过形式逻辑符号化,对思维进行推理计算。这种"万能符号"和"基于符号的推理计算"的思想是"智能机器"的萌芽,因而他被誉为数理逻辑的奠基人。进入 20 世纪后,人工智能相继出现了若干开创性的工作。1936 年,年仅 24 岁的英国数学家图灵(A. M. Turing)在他的一篇"理想计算机"的论文中,提出了著名的图灵机模型。1945 年,他进一步论述了电子数字计算机的设计思想。1950 年,他又在"计算机能思维吗?"一文中提出了机器能够思维的论述。1938 年,德国工程师苏斯(Zuse)研制成第一台累计数字计算机 Z-1。1946 年,在美国诞生了世界上第一台电子数字计算机 ENIAC。在同一时代,控制论和信息论的创立,生物学家设计的脑模型等,都为人工智能学科的诞生做出了理论和实验

工具的巨大贡献。

1956年的一次历史性聚会被认为是人工智能学科诞生的标志。1956年夏季，在美国达特莫斯(Dartmouth)大学，由当时的年轻数学助教、现任斯坦福大学教授的麦卡锡(J. McCarthy)联合他的三位朋友——哈佛大学年轻数学和神经学家、现任麻省理工学院教授的明斯基(M. L. Minsky)，IBM公司信息研究中心负责人洛切斯特(N. Lochester)和贝尔实验室信息部数学研究员香农(C. E. Shannoon)——共同发起，邀请IBM公司的莫尔(T. Moore)和塞缪尔(A. L. Samuel)、麻省理工学院的塞尔夫利奇(O. Selfridge)和索罗莫夫(R. Solomonff)、兰德(RAND)公司和卡内基工科大学的纽厄尔(A. Newell)和西蒙(H. A. Simon)等10名年轻学者，举办了为期两个月的学术讨论会，讨论机器智能问题。经麦卡锡提议，在会上正式决定使用“人工智能”(Artificial Intelligence)这一术语，从而开创了人工智能作为一门独立学科的研究方向。麦卡锡因而被称为人工智能之父。从此在美国开始形成了以人工智能为研究目标的几个研究组，如纽厄尔和西蒙的Carnegie-RAND协作组，明斯基和麦卡锡的MIT研究组，塞缪尔的IBM工程研究组等。

1956年，人工智能的研究取得了两项重大突破。第一项是纽厄尔、肖(J. Shaw)和西蒙的研究组编制了一个逻辑理论程序LT(The Logic Theory Machine)，模拟人们用数理逻辑证明定理的思想，采用分解、代入、替换等规则，证明了怀特海(A. N. Whitehead)和罗素(B. A. W. Russell)合著的《数学原理》第二章中的38条定理。1963年，修订的程序在大机器上终于完成了该章中全部52条定理的证明。一般认为，这是用计算机模拟人的高级思维活动的一个重大成果，是人工智能研究的真正开端。第二项是IBM工程研究组的塞缪尔研制的西洋跳棋程序。这个程序可以像一个优秀棋手那样，向前看几步来下棋。尤其是它具有自学习、自组

织、自适应的能力，能在下棋过程中积累经验，不断提高棋艺。它能学习棋谱，在学习了 175 000 多个棋局后，可以根据棋局猜测棋谱所有推荐的走步，准确度达 48%，这是机器模拟人类学习过程的一次极有意义的探索。1959 年，这个程序战胜了设计者本人；1962 年，它又击败了美国某个州的跳棋冠军。

1957 年，纽厄尔、肖和西蒙通过心理学实验，发现了人在问题求解时思维过程的一般规律，大致可分为三个阶段：

(1)先思考出大致的解题计划；

(2)根据记忆中的公理、定理和推理规则组织解题过程；

(3)进行方法和目的分析，不断修正解题计划。

基于这一规律，他们于 1960 年合作编制成功一种不依赖于具体领域的通用问题求解程序 GPS(General Problem Solver)，能求解 11 种不同类型的问题。

1959 年，麻省理工学院研究组的麦卡锡发表了表处理语言 LISP。由于 LISP 语言可以方便地处理符号，很快成为人工智能程序设计的主要语言。LISP 语言武装了一代人工智能科学家，时至今日，仍然是研究人工智能的重要工具。

一连串的研究成果使醉心于人工智能远景的学者们做出了过于乐观的预言。1958 年，纽厄尔和西蒙曾充满自信地认为：在 10 年内，计算机将成为世界的象棋冠军；计算机将要发现和证明重要的数学定理；计算机将能谱写具有优秀作曲家水平的乐曲；大多数心理学理论将在计算机上形成。有人甚至断言：20 世纪 80 年代将是全面实现人工智能的年代，到了 2000 年机器的智能可以超过人的智能。

但是，事情的发展远非如此理想。塞缪尔的下棋程序在获得州的冠军之后，再也没有当上全国冠军。自然语言的机器翻译是人工智能研究最早并取得实验性成果的研究方向之一。人们以为只要用一部双向互译字典和某些语法知识即可很快地解决自然语

言之间的互译问题，实际上，由机器翻译出来的文字有时会出现十分荒谬的错误。

自从人工智能形成一个学科之后，许多学者遵循的指导思想是：研究和总结人类思维的普遍规律，并用计算机来模拟人类的思维活动。他们认为，实现这种计算机智能模拟的关键是建立一种通用的符号逻辑运算体系。但是，由于人类的认知和思维过程是一种非常复杂的行为，至今仍未能被完全解释，也由于现实世界的复杂性和问题的多样性，老一辈人工智能科学家为之奋斗的通用逻辑推理体系至今也没有创造出来。他们早期的代表作——通用问题求解程序 GPS——的通用性受到严格的限制，只能对具有相当小的状态集和良定义的形式规则的问题有效。人工智能的早期研究只能停留在实验室里，作为研究的实验系统或演示系统，不能解决实际问题。科学家们开始对人工智能探索人类思维普遍规律的研究战略思想进行反思。

2. 人工智能的发展

20 世纪 60 年代中期以后，人工智能由追求万能、通用的一般研究转入特定的具体研究，通用的解题策略与特定领域的专业知识及实际经验结合，产生了以专家系统(ES, Expert System)为代表的基于知识的各种人工智能系统，使人工智能走向社会，走向实际应用研究。斯坦福大学当时的年轻教授费根鲍姆(E. A. Feigenbaum)重新举起了英国 16 世纪哲学家和自然科学家培根(Francis. Bacon)的旗帜："知识就是力量"，于 1965 年开创了基于知识的专家系统这一人工智能研究的新领域。与通用问题求解程序 GPS 的系统不同，专家系统并不试图发现很强有力的和很通用的问题求解方法，而是把研究范围缩小在一个特定的相对狭小的专业领域中。人类专家之所以成为专家，是因为他拥有解决自己专业领域问题的大量专门知识，包括各种有用的经验。

在费根鲍姆的主持下，第一个专家系统课题 DENDRAL 化学

分子结构分析系统于1965年在斯坦福大学开始研究，1968年研制成功。该系统能根据质谱仪数据推断未知有机化合物的分子结构。它是一个启发式系统，把化学专家关于分子结构质谱测定法的知识结合到控制搜索的规则中，从而能迅速消去不可能为真的分子结构，避免了搜索空间以指数级膨胀。通过产生全部可能为真的分子结构，它甚至可以找出那些人类专家可能漏掉的结构。DENDRAL及附属的CONGEN系统商品化后，每天为上百个国际用户提供化学结构的解释。这一研究成果使人们看到，在某个专门领域里，以知识为基础的计算机系统完全可能相当于这个领域里的人类专家的作用。

MACSYMA系统是麻省理工学院于1968年开始研制的大型符号数学专家系统。该系统从应用数学家那里获得了几百条关于一个表达式与另一等价表达式之间转换的规则，擅长于易引起组合爆炸的符号表达式的化简，能执行微分、积分、解方程、泰勒级数展开、矩阵运算、向量代数等600多种不同的数学符号运算。1971年研制成功后，由于它具有很强的与应用分析相结合的符号运算能力，很多数学和物理学的研究人员及各类工程师争相使用MACSYMA系统，遍及美国各地的很多用户每天都通过ARPA网与它联机工作达数小时。

在DENDRAL和MACSYMA的影响下，在化学、数学、医学、生物工程、地质探矿、石油勘探、气象预报、地震分析、过程控制、计算机配置、集成电路测试、电子线路分析、情报处理、法律咨询和军事决策等各方面出现了一大批专家系统。著名的MYCIN系统就是斯坦福大学人工智能研究所于1973年开始研制的一个诊断和治疗细菌感染性血液病的专家咨询系统。该系统可以看成是DENDRAL系统的直接后继者，并且具有更广泛的影响。MYCIN拥有的知识包括约450条“前提-结论”型的关于细菌性血液感染的诊疗规则，系统根据提供的数据和主动向医生询问获得