中国计算机学会
学术著作丛书

知识发现

史忠植 著

清华大学出版社

中国计算机学会学术著作丛书

知识发现

史忠植 著

清华大学出版社

(京)新登字 158 号

内 容 简 介

知识发现是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。知识发现将信息变为知识,从数据矿山中找到蕴藏的知识金块,将为知识创新和知识经济的发展作出贡献。

本书全面而又系统地介绍了知识发现的方法和技术,反映了当前知识发现研究的最新成果。

全书共分 14 章。第 1 章是绪论,介绍知识发现的重要概念和任务。第 2 章讨论决策树,它是归纳学习方法中最实用的一种技术。关联规则挖掘是近几年应用最为广泛的方法,第 3 章将对重要的关联规则挖掘算法进行讨论。第 4 章讨论范例推理,它是一种有效的实用技术。第 5 章探讨模糊聚类法。第 6 章讨论粗糙集。第 7 章是贝叶斯网络,贝叶斯网络可以处理不完整和带有噪声的数据集,它用概率测度的权重来描述数据间的相关性。第 8 章探讨支持向量机,它在近几年知识发现研究中是极其活跃的研究课题。第 9 章讨论隐马尔科夫模型。第 10 章是神经网络,书中着重介绍几种实用的算法。第 11 章讨论进化和遗传算法。第 12 章介绍知识发现平台 MSMiner。接着,以 Web 知识发现、生物信息处理为例,介绍知识发现的应用。第 13 章关于 Web 知识发现。第 14 章介绍生物信息处理中基因组模式的发现。

本书内容新颖,认真总结了作者的科研成果,取材国内外最新资料,反映了当前该领域的研究水平。论述力求概念清晰,表达准确,突出理论联系实际,通过实例说明原理,富有启发性。本书对从事知识发现、数据挖掘、机器学习、人工智能研究和知识管理的科技人员具有重要参考价值,可以用作计算机、信息技术等专业博士生、硕士生的教材。

书　　名:知识发现
作　　者:史忠植　著
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
责任编辑:薛慧
印 刷 者:清华大学印刷厂
发 行 者:新华书店总店北京发行所
开　　本:787×1092 1/16　**印张**:26　**字数**:601 千字
版　　次:2002 年 1 月第 1 版　2002 年 1 月第 1 次印刷
书　　号:ISBN 7-302-05061-9/TP・2961
印　　数:0001～6000
定　　价:38.00 元

前　言

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，世界上的数据正以惊人的速度增长，堆积如山。不同领域的人们都期待着从这些数据中得到自己想要的答案，将信息变为知识，从数据矿山中找到蕴藏的知识金块。

知识发现正是这样一种从数据中挖掘知识的工具，它集数据收集、数据清洁、降维、规则归纳、模式识别、数据/结果分析及评估、可视化输出等多种过程于一身，是统计学、计算机科学、模式识别、人工智能、机器学习及其他学科相结合的产物。它不仅被许多研究人员看作是数据库系统和机器学习方面一个重要的研究课题，而且被许多工商界人士看作是一个能带来巨大回报的重要领域。从数据库中发现出来的知识可以用在信息管理、查询响应、决策支持、过程控制等许多方面。从 20 世纪 80 年代中期的小范围研究到如今的蓬勃兴起，知识发现已经在企业界与科学界占据了一席之地。事实上，世界 500 强企业中的 80%都涉足知识发现的前瞻性研究或拥有一个或多个知识发现产品系统。它们帮助企业进行客户关系管理，减少不必要的投资，提高资金周转和回报；帮助人们迅速获取所需的知识和信息，提高工作效率，改进服务质量。

知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程，它与数据仓库有着密切的联系。数据仓库是源于可操作数据的一个数据存储中心。数据仓库中的信息是面向主题的、稳定的且具有历史数据属性的，因此数据仓库用于存储大规模的数据集。知识发现与数据仓库、决策支持的结合预示着信息和知识管理的一个全新的变革。

本书全面而又系统地介绍了知识发现的方法和技术。全书共分 14 章。第 1 章是绪论，介绍知识发现的重要概念和任务。第 2 章讨论决策树，它是归纳学习方法中最实用的一种技术。关联规则挖掘是近几年应用最为广泛的方法，第 3 章将对重要的关联规则挖掘算法进行讨论。第 4 章讨论范例推理，它是一种有效的实用技术。第 5 章探讨模糊聚类法。第 6 章讨论粗糙集。第 7 章是贝叶斯网络，贝叶斯网络可以处理不完整和带有噪声的数据集，它用概率测度的权重来描述数据间的相关性。第 8 章探讨支持向量机，它在近几年知识发现研究中是极其活跃的研究课题。第 9 章讨论隐马尔科夫模型。第 10 章是神经网络，书中着重介绍几种实用的算法。第 11 章讨论进化和遗传算法。第 12 章介绍知识发现平台 MSMiner。接着，以 Web 知识发现、生物信息处理为例，介绍知识发现的应用。第 13 章关于 Web 知识发现。第 14 章介绍生物信息处理中基因组模式的发现。

本书是中国科学院计算技术研究所智能信息处理重点实验室有关机器学习和知识发现研究工作的总结。涉及的研究项目得到国家自然科学基金、国家 863 高技术计划、北京市自然科学基金、国家重点科技攻关项目的资助。参加该项研究工作的人员

有叶世伟副教授、何清副教授、李晓黎博士、叶施仁博士、宫秀军博士生、刘少辉博士生、郑毅、郑金华教授、张建博士、王军博士、张颖博士、李云峰博士、潘谦红博士、吴斌博士生、贾自艳博士生、游湘涛、任力安、李宝东等。本书得到清华大学出版社计算机专著出版基金的资助。

史忠植
2001 年 9 月

目　录

前言 …… I

第 1 章　绪论 …… 1

1.1　知识 …… 1

1.2　知识发现 …… 2

1.3　知识发现的任务 …… 4

1.3.1　数据总结 …… 4

1.3.2　概念描述 …… 5

1.3.3　分类 …… 5

1.3.4　聚类 …… 6

1.3.5　相关性分析 …… 6

1.3.6　偏差分析 …… 6

1.3.7　建模 …… 7

1.4　知识发现的方法 …… 7

1.4.1　统计方法 …… 7

1.4.2　机器学习 …… 9

1.4.3　神经计算 …… 11

1.4.4　可视化 …… 12

1.5　知识发现的对象 …… 13

1.5.1　数据库 …… 13

1.5.2　文本 …… 14

1.5.3　Web 信息 …… 15

1.5.4　空间数据 …… 15

1.5.5　图像和视频数据 …… 16

1.6　知识发现与创新 …… 17

第 2 章　决策树 …… 21

2.1　归纳学习 …… 21

2.2　决策树学习 …… 21

2.3　CLS 学习算法 …… 23

2.4　ID3 学习算法 …… 24

2.4.1　信息论简介 …… 24

2.4.2　信息论在决策树学习中的意义及应用 …… 25

2.4.3　ID3 算法 …… 26

2.4.4　ID3 算法应用举例 …… 26

2.5 决策树的改进算法 …… 28
2.5.1 二叉树判定算法 …… 28
2.5.2 按信息比值进行估计的方法 …… 29
2.5.3 按分类信息估值 …… 29
2.5.4 按划分距离估值的方法 …… 30
2.6 决策树的评价 …… 31
2.7 简化决策树 …… 32
2.7.1 简化决策树的动机 …… 33
2.7.2 决策树过大的原因 …… 33
2.7.3 控制树的大小 …… 34
2.7.4 修改测试属性空间 …… 36
2.7.5 改进测试属性选择方法 …… 38
2.7.6 对数据进行限制 …… 40
2.7.7 改变数据结构 …… 41
2.8 连续型属性离散化 …… 44
2.9 基于偏置变换的决策树学习算法 BSDT …… 45
2.9.1 偏置的形式化 …… 46
2.9.2 表示偏置变换 …… 47
2.9.3 算法描述 …… 48
2.9.4 过程偏置变换 …… 49
2.9.5 基于偏置变换的决策树学习算法 BSDT …… 51
2.9.6 经典范例库维护算法 TCBM …… 51
2.9.7 偏置特征抽取算法 …… 52
2.9.8 改进的决策树生成算法 GSD …… 53
2.9.9 实验结果 …… 55
2.10 归纳学习中的问题 …… 56
第 3 章 关联规则 …… 57
3.1 关联规则挖掘概述 …… 57
3.1.1 关联规则的意义和度量 …… 57
3.1.2 经典的挖掘算法 …… 59
3.2 广义模糊关联规则的挖掘 …… 61
3.3 挖掘关联规则的数组方法 …… 64
3.4 任意多表间关联规则的并行挖掘 …… 64
3.4.1 问题的形式描述 …… 65
3.4.2 单表内大项集的并行计算 …… 65
3.4.3 任意多表间大项集的生成 …… 67
3.4.4 跨表间关联规则的提取 …… 68
3.5 基于分布式系统的关联规则挖掘算法 …… 68

3.5.1 候选集的生成 …… 69
3.5.2 候选数据集的局部剪枝 …… 71
3.5.3 候选数据集的全局剪枝 …… 73
3.5.4 合计数轮流检测 …… 75
3.5.5 分布式挖掘关联规则的算法 …… 76
3.6 词性标注规则的挖掘算法与应用 …… 78
3.6.1 汉语词性标注 …… 78
3.6.2 问题的描述 …… 79
3.6.3 挖掘算法 …… 80
3.6.4 试验结果 …… 83
第4章 基于范例的推理 …… 85
4.1 概述 …… 85
4.2 过程模型 …… 86
4.3 范例的表示 …… 88
4.3.1 语义记忆单元 …… 89
4.3.2 记忆网 …… 89
4.4 范例的索引 …… 91
4.5 范例的检索 …… 92
4.6 相似性关系 …… 93
4.6.1 语义相似性 …… 94
4.6.2 结构相似性 …… 94
4.6.3 目标特征 …… 94
4.6.4 个体相似性 …… 95
4.6.5 相似性计算 …… 95
4.7 范例的复用 …… 96
4.8 范例的保存 …… 98
4.9 基于例示的学习 …… 99
4.9.1 基于例示学习的任务 …… 99
4.9.2 IB1 算法 …… 100
4.9.3 降低存储要求 …… 102
4.10 范例工程 …… 104
4.11 范例约简算法 …… 106
4.12 中心渔场预报专家系统 …… 109
4.12.1 问题分析与范例表示 …… 109
4.12.2 相似性度量 …… 111
4.12.3 索引与检索 …… 111
4.12.4 基于框架的修正 …… 112
4.12.5 实验结果 …… 114

第 5 章　模糊聚类 …… 116
5.1　概述 …… 116
5.1.1　聚类结果的表示 …… 116
5.1.2　模糊聚类的一般模型 …… 116
5.2　传递闭包法 …… 117
5.2.1　模糊相似系数的标定 …… 117
5.2.2　传递闭包法 …… 120
5.2.3　动态直接聚类法 …… 120
5.2.4　最大树法 …… 121
5.3　FCMBP 聚类法 …… 122
5.3.1　问题背景 …… 122
5.3.2　Fuzzy 等价标准型 …… 124
5.3.3　置换等价类与平移等价类的记数公式 …… 128
5.3.4　X_n 的结构 …… 129
5.3.5　模糊最优等价阵的存在性 …… 130
5.3.6　最优模糊等价阵的算法步骤 …… 131
5.3.7　基于 FCMBP 模糊聚类的语音识别 …… 135
5.4　系统聚类法 …… 136
5.5　C-均值聚类法 …… 137
5.6　聚类有效性 …… 140
5.7　聚类方法的比较 …… 141
第 6 章　粗糙集 …… 143
6.1　概述 …… 143
6.1.1　知识的分类观点 …… 145
6.1.2　新型的隶属关系 …… 145
6.1.3　概念的边界观点 …… 146
6.2　知识的约简 …… 147
6.2.1　一般约简 …… 147
6.2.2　相对约简 …… 147
6.2.3　知识的依赖性 …… 148
6.3　决策逻辑 …… 149
6.3.1　决策表的公式化定义 …… 149
6.3.2　决策逻辑语言 …… 150
6.3.3　决策逻辑语言的语义 …… 151
6.3.4　决策逻辑的推演 …… 152
6.3.5　规范表达形式 …… 153
6.3.6　决策规则和决策算法 …… 153
6.3.7　决策规则中的一致性和不分明性 …… 154

6.4 决策表的约简 …… 154
6.4.1 属性的依赖性 …… 155
6.4.2 一致决策表的约简 …… 155
6.4.3 非一致决策表的约简 …… 160
6.5 粗糙集的扩展模型 …… 163
6.5.1 可变精度粗糙集模型 …… 164
6.5.2 相似模型 …… 165
6.5.3 基于粗糙集的非单调逻辑 …… 165
6.5.4 与其他数学工具的结合 …… 166
6.6 粗糙集的实验系统 …… 166
6.7 粗糙集的展望 …… 168
第 7 章 贝叶斯网络 …… 169
7.1 概述 …… 169
7.1.1 贝叶斯网络的发展历史 …… 169
7.1.2 贝叶斯方法的基本观点 …… 170
7.1.3 贝叶斯网络在数据挖掘中的应用 …… 170
7.2 贝叶斯概率基础 …… 172
7.2.1 概率论基础 …… 172
7.2.2 贝叶斯概率 …… 174
7.3 贝叶斯学习理论 …… 176
7.3.1 几种常用的先验分布选取方法 …… 177
7.3.2 计算学习机制 …… 179
7.3.3 贝叶斯问题求解 …… 181
7.4 简单贝叶斯学习模型 …… 183
7.4.1 简单贝叶斯学习模型 …… 183
7.4.2 简单贝叶斯模型的提升 …… 185
7.4.3 提升简单贝叶斯分类的计算复杂性 …… 187
7.5 贝叶斯网络的建造 …… 187
7.5.1 贝叶斯网络的结构及建立方法 …… 187
7.5.2 学习贝叶斯网络的概率分布 …… 188
7.5.3 学习贝叶斯网络的网络结构 …… 190
7.6 贝叶斯潜在语义模型 …… 193
7.7 半监督文本挖掘算法 …… 196
7.7.1 网页聚类 …… 196
7.7.2 对含有潜在类别主题词的文档的类别标注 …… 197
7.7.3 基于简单贝叶斯模型学习标注和未标注样本 …… 198
第 8 章 支持向量机 …… 203
8.1 统计学习问题 …… 203

8.1.1 经验风险 …… 203
8.1.2 VC 维 …… 203
8.2 学习过程的一致性 …… 204
8.2.1 学习一致性的经典定义 …… 204
8.2.2 学习理论的重要定理 …… 204
8.2.3 VC 熵 …… 205
8.3 结构风险最小归纳原理 …… 206
8.4 支持向量机 …… 208
8.4.1 线性可分 …… 208
8.4.2 线性不可分 …… 209
8.5 核函数 …… 211
8.5.1 多项式核函数 …… 211
8.5.2 径向基函数 …… 211
8.5.3 多层感知机 …… 211
8.5.4 动态核函数 …… 212
8.6 基于分类超曲面的海量数据分类方法 …… 213
8.6.1 Jordan 曲线定理 …… 213
8.6.2 SVM 直接方法基本思想 …… 214
8.6.3 实现算法 …… 215
8.6.4 实验结果分析 …… 215
第 9 章 隐马尔科夫模型 …… 219
9.1 马尔科夫过程 …… 219
9.2 隐马尔科夫模型 …… 220
9.3 似然概率和前反向算法 …… 221
9.3.1 前向算法 …… 222
9.3.2 反向算法 …… 222
9.3.3 Viterbi 算法 …… 223
9.3.4 计算期望 …… 223
9.4 学习算法 …… 224
9.4.1 EM 算法 …… 224
9.4.2 梯度下降 …… 225
9.4.3 Viterbi 学习 …… 226
9.5 基于状态驻留时间的分段概率模型 …… 226
9.5.1 SDSPM 模型的构成 …… 227
第 10 章 神经网络 …… 230
10.1 概述 …… 230
10.1.1 基本的神经网络模型 …… 230
10.1.2 神经网络的学习方法 …… 230

10.2 人工神经元及感知机模型 …… 232
10.2.1 基本神经元 …… 232
10.2.2 感知机模型 …… 233
10.3 前向神经网络 …… 234
10.3.1 前向神经网络模型 …… 234
10.3.2 多层前向神经网络的误差反向传播(BP)算法 …… 235
10.3.3 BP 算法的若干改进 …… 237
10.4 径向基函数神经网络 …… 241
10.4.1 插值问题 …… 242
10.4.2 正规化问题 …… 242
10.4.3 RBF 网络学习方法 …… 244
10.5 反馈神经网络 …… 247
10.5.1 离散 Hopfield 网络 …… 247
10.5.2 连续 Hopfield 网络 …… 253
10.5.3 Hopfield 网络应用 …… 255
10.5.4 双向联想记忆模型 …… 256
10.6 随机神经网络 …… 257
10.6.1 模拟退火算法 …… 257
10.6.2 玻尔兹曼机 …… 260
10.7 自组织特征映射神经网络 …… 263
10.7.1 网络的拓扑结构 …… 263
10.7.2 网络自组织算法 …… 263
10.7.3 有教师学习 …… 264
第 11 章 进化和遗传算法 …… 265
11.1 概述 …… 265
11.2 基本遗传算法 …… 267
11.2.1 基本遗传算法的构成要素 …… 267
11.2.2 基本遗传算法的一般框架 …… 268
11.3 遗传算法的数学理论 …… 270
11.3.1 模式定理 …… 271
11.3.2 积木块假设 …… 273
11.3.3 遗传算法欺骗问题 …… 274
11.3.4 隐并行性 …… 274
11.4 遗传算法的基本实现技术 …… 275
11.4.1 编码方法 …… 275
11.4.2 适应度函数 …… 278
11.4.3 选择算子 …… 280
11.4.4 交叉算子 …… 282

11.4.5 变异算子 …… 284
11.4.6 约束条件的处理方法 …… 285
11.5 遗传算法的高级实现技术 …… 285
11.5.1 反转操作 …… 285
11.5.2 变长度染色体遗传算法 …… 286
11.5.3 小生境遗传算法 …… 286
11.5.4 混合遗传算法 …… 287
11.5.5 改进遗传算法 …… 290
11.6 并行遗传算法 …… 291
11.7 遗传算法应用 …… 292
11.7.1 优化神经网络连接权值 …… 292
11.7.2 用遗传算法优化神经网络连接结构 …… 293
第 12 章 知识发现平台 MSMiner …… 295
12.1 概述 …… 295
12.2 数据仓库 …… 297
12.2.1 数据仓库含义 …… 297
12.2.2 元数据 …… 298
12.2.3 OLAP …… 299
12.2.4 数据仓库和数据挖掘技术的结合 …… 299
12.3 MSMiner 的体系结构 …… 300
12.3.1 数据挖掘模型 …… 300
12.3.2 系统功能 …… 301
12.3.3 体系结构 …… 302
12.4 元数据管理 …… 303
12.4.1 MSMiner 元数据的内容 …… 303
12.4.2 MSMiner 元数据库 …… 304
12.4.3 MSMiner 元数据对象模型 …… 304
12.5 数据仓库管理器 …… 307
12.5.1 MSMiner 数据仓库的基本结构 …… 308
12.5.2 主题 …… 309
12.5.3 数据抽取和集成 …… 310
12.5.4 数据抽取和集成的元数据 …… 313
12.5.5 数据仓库建模及 OLAP 的实现 …… 314
12.6 算法库管理 …… 318
12.6.1 数据挖掘算法的元数据 …… 318
12.6.2 可扩展性的实现 …… 319
12.6.3 挖掘算法的接口规范 …… 320
12.7 数据挖掘任务规划 …… 322

12.7.1 面向对象的数据挖掘任务模型 …… 322
12.7.2 数据挖掘任务模型的处理 …… 326
12.8 关系数据库知识发现查询语言 KDSQL …… 328
12.8.1 知识对象 …… 328
12.8.2 知识发现查询语言定义 …… 329
12.8.3 扩充的 CREATE 命令语句 …… 330
12.8.4 扩充的 SELECT 命令语句 …… 332
第 13 章 Web 知识发现 …… 334
13.1 概述 …… 334
13.2 Web 知识发现的任务 …… 336
13.2.1 Web 知识发现任务的分类 …… 336
13.2.2 Web 内容发现 …… 337
13.2.3 Web 结构挖掘 …… 338
13.3 Web 知识发现方法 …… 338
13.3.1 文本的特征表示 …… 339
13.3.2 TFIDF 向量表示法 …… 340
13.3.3 特征子集的选取 …… 342
13.4 模型质量评价 …… 343
13.5 文本分析功能 …… 344
13.5.1 名字提取 …… 345
13.5.2 术语提取 …… 346
13.5.3 缩写词识别器 …… 346
13.5.4 其他提取器 …… 347
13.6 文本特征的提取 …… 347
13.6.1 一般特征项的提取 …… 347
13.6.2 专有特征项的提取 …… 348
13.7 基于文本挖掘的汉语词性自动标注研究 …… 351
13.8 文本分类 …… 352
13.9 文本聚类 …… 356
13.9.1 层次凝聚法 …… 356
13.9.2 平面划分法 …… 357
13.9.3 简单贝叶斯聚类算法 …… 358
13.9.4 *k*-最近邻参照聚类算法 …… 359
13.9.5 分级聚类 …… 359
13.9.6 基于概念的文本聚类 …… 359
13.10 文本摘要 …… 361
13.11 用户兴趣挖掘 …… 362

第 14 章　生物信息知识发现 …… 364
14.1　概述 …… 364
14.2　基因的基本结构 …… 366
14.3　生物信息数据库与查询 …… 367
14.3.1　基因和基因组数据库 …… 367
14.3.2　蛋白质数据库 …… 369
14.3.3　功能数据库 …… 370
14.4　序列比对 …… 371
14.4.1　序列两两比对 …… 371
14.4.2　多序列比对 …… 373
14.5　核酸与蛋白质结构和功能的预测分析 …… 374
14.5.1　核酸序列的预测方法 …… 374
14.5.2　针对蛋白质的预测方法 …… 375
14.6　基因组序列信息分析 …… 377
14.7　功能基因组相关信息分析 …… 380
14.7.1　大规模基因表达谱分析 …… 380
14.7.2　基因组水平蛋白质功能综合预测 …… 381
14.8　Internet 资源和公共数据库 …… 382
参考文献 …… 387
索引 …… 398

第1章 绪 论

1.1 知 识

人类从工业社会向知识社会演进时，政治经济中心正从“生产”转向“发现、发明和创新”。知识正在成为创新的核心，知识创新成为知识经济发展的最主要的动力源泉。知识经济对物质文明发展能够发挥巨大的推动作用，依靠无形资产的投入来实现可持续发展的，推动经济全球化发展。

在信息科学中，信息是根据表示数据所用的约定，赋予数据的意义。数据是事物、概念或指令的一种形式化的表示形式，以适合于用人工或自然方式进行通信、解释或处理。信息是数据所表达的客观事实。数据是信息的载体，与具体的介质和编码方法有关。20世纪40年代，Shannon对信息的数学本质进行过研究，提出了著名的Shannon信息论。他用熵的概念来研究信息的容量，采用比特作为度量信息的单位。

信息经过加工和改造形成知识。知识是人类在实践的基础上产生又经过实践检验的对客观实际的可靠的反映。知识是人脑创新的成果，是人类智慧的结晶。智慧是人类文明的源泉，是推动历史发展的永衡动力，是生产力诸要素中的核心。知识一般可分为陈述性知识、过程性知识和控制性知识。陈述性知识提供概念和事实，描述系统状态、环境和条件，使人们知道是什么。例如，在一个知识检索系统中，陈述性知识包括陈述具体事实的数据库内容。过程性知识提供有关状态的变化、问题求解过程的操作、演算和动作的知识。智能信息检索系统利用过程性知识处理陈述性知识。用控制策略表示问题的知识常称为控制性知识。控制性知识，即元知识，包含有关各种处理过程、策略和结构的知识，常用来协调整个问题求解的过程。

知识具有下列特性：

(1) 知识的客观性。虽然知识是人脑对信息加工的成果，但这些成果是客观的，人类对自然、社会、思维规律的认识是客观的，这些规律的运行是不以人的意志为转移的。

(2) 知识的相对性。人类对自然、社会、思维规律的认识必须有一个过程。在一段时间内认为正确的东西，经过变革，可能发生变化。1991年第12届国际人工智能大会上，IJCAI的计算机和思维奖授予MIT的Brooks。他提出基于行为的人工智能，认为智能不要知识表示，智能不要推理[20]。2001年第17届国际人工智能大会上，IJCAI的计算机和思维奖授予斯坦福大学的Koller，以表彰她在概率推理、机器学习方面的贡献[85]。

(3) 知识的进化性。人类在认识客观世界和主观世界的过程中，不断对真理的长河加入新的内容，知识不断更新，例如对物质结构的认识，对基因的认识等。

(4) 知识的依附性。知识有载体，载体分层次。离开载体的知识是没有的。随着载体的消失，知识也跟着消失。

(5) 知识的可重用性。在使用过程中知识可以反复重用。当然，要根据具体情况作

具体分析，灵活应用知识。

(6) 知识的共享性。基础研究一般由政府进行投资，所得到的科学知识具有共享性；但最新的技术知识受到知识产权法保护，使用者只有支付一定的费用，才能获得这种知识的使用权。知识产权的保护对发展技术和知识经济是非常重要的国策。

1.2 知识发现

知识发现是从数据集中抽取和精化新的模式。知识发现的范围非常广泛，可以是经济、工业、农业、军事、社会、商业、科学的数据或卫星观测得到的数据。数据的形态有数字、符号、图形、图像、声音等。数据组织方式也各不相同，可以是有结构、半结构或非结构的。知识发现的结果可以表示成各种形式，包括规则、法则、科学规律、方程或概念网等。

目前，关系型数据库应用广泛，并且具有统一的组织结构，一体化的查询语言，关系之间及属性之间具有平等性等优点。因此，数据库知识发现(knowledge discovery in databases，KDD)的研究非常活跃。该术语于1989年出现，Fayyad定义为“KDD是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程”[39]。在上面的定义中，涉及几个需要进一步解释的概念：“数据集”、“模式”、“过程”、“有效性”、“新颖性”、“潜在有用性”和“最终可理解性”。数据集是一组事实 F(如关系数据库中的记录)。模式是一个用语言 L 来表示的一个表达式 E，它可用来描述数据集 F 的某个子集 F_E，E 作为一个模式要求它比对数据子集 F_E 的枚举要简单(所用的描述信息量要少)。过程在KDD中通常指多阶段的处理，涉及数据准备、模式搜索、知识评价以及反复的修改求精；该过程要求是非平凡的，意思是要有一定程度的智能性、自动性(仅仅给出所有数据的总和不能算作是一个发现过程)。有效性是指发现的模式对于新的数据仍保持有一定的可信度。新颖性要求发现的模式应该是新的。潜在有用性是指发现的知识将来有实际效用，如用于决策支持系统里可提高经济效益。最终可理解性要求发现的模式能被用户理解，目前它主要是体现在简洁性上。有效性、新颖性、潜在有用性和最终可理解性综合在一起称为兴趣性。

由于知识发现是一门受到来自各种不同领域的研究者关注的交叉性学科，因此导致了很多不同的术语名称。除了KDD外，主要还有如下若干种称法：“数据挖掘”(data mining)，“知识抽取”(information extraction)、“信息发现”(information discovery)、“智能数据分析”(intelligent data analysis)、“探索式数据分析”(exploratory data analysis)、“信息收获”(information harvesting)和“数据考古”(data archeology)等等。其中，最常用的术语是“知识发现”和“数据挖掘”。相对来讲，数据挖掘主要流行于统计界(最早出现于统计文献中)、数据分析、数据库和管理信息系统界；而知识发现则主要流行于人工智能和机器学习界。

知识发现过程可粗略地理解为三部曲：数据准备(data preparation)、数据开采以及结果的解释评估(interpretation and evaluation)(见图1.1)。