高水平大学
重点学科建设教材

电子信息类

数据挖掘

朱　明　编著

中国科学技术大学出版社

高水平大学重点学科建设教材·电子信息类

数 据 挖 掘

朱 明 编著

中国科学技术大学出版社
合 肥

内 容 简 介

数据挖掘技术，又称数据库知识发现，是20世纪90年代在信息技术领域开始迅速兴起的计算机技术。作者结合自己10余年来所从事的专家系统、机器学习、数据挖掘，以及互联网信息智能处理等方面科研与教学经验，编著完成了本书。

本书系统地介绍了数据挖掘中的主要挖掘方法和对复杂数据进行挖掘的方法，以及在互联网信息智能处理领域中，数据挖掘方法与技术的主要应用。

本书的主要内容包括：数据挖掘基本知识、数据挖掘处理流程、数据预处理方法、定性概念归纳、决策树分类方法、回归统计预测方法、关联规则发现方法、各种聚类算法，以及复杂数据，尤其是多媒体数据挖掘方法的最新研究成果；此外还详细介绍了利用数据挖掘方法获取互联网信息，挖掘互联网使用知识，以及网络安全中数据挖掘方法应用等。

本书作为学习、掌握和应用数据挖掘方法和技术的综合指导书，是从事数据挖掘研究与设计人员、开发人员，以及需要了解数据挖掘有关方法与技术的IT技术人员的良师益友。同时也是一本较好的大学高年级或研究生相关课程的教材和参考书。

图书在版编目(CIP)数据

数据挖掘/朱明编著. —合肥：中国科学技术大学出版社，2002.5
高水平大学重点学科建设教材·电子信息类
ISBN 7-312-01364-3

Ⅰ.数… Ⅱ.朱… Ⅲ.数据采集-高等学校-教材 Ⅳ.TP274

中国版本图书馆CIP数据核字(2001)第089052号

中国科学技术大学出版社出版发行
(安徽省合肥市金寨路96号，邮编：230026)
中国科学技术大学印刷厂印刷
全国新华书店经销

开本：787mm×1092mm 1/16 印张：20.75 字数：531千
2002年5月第1版 2002年5月第1次印刷
印数：1—4000册
ISBN 7-312-01364-3/TP·284 定价：23.00元

前　言

随着数据库应用的普及，人们正逐步陷入“数据丰富，知识贫乏”的尴尬境地。而近年来互联网的发展与快速普及，使得人类第一次真正体会到了数据海洋，无边无际。面对如此巨大的数据资源，人们迫切需要一种新技术和自动工具，以便能够利用智能技术帮助我们将这巨大数据资源转换为有用的知识与信息资源，从而可以帮助我们科学地进行各种决策。

数据挖掘(Data Mining，简称 DM)作为 20 世纪末刚刚兴起的数据智能分析技术，由于其所具有的广阔应用前景而备受关注，作为数据库与数据仓库研究与应用中的一个新兴的富有前途领域，数据挖掘，常常也被称为数据库知识发现(Knowledge Discovery from Database，简称 KDD)，它可以从数据库，或数据仓库，以及其它各种数据库的大量各种类型数据中，自动抽取或发现出有用的模式知识。

数据挖掘是一个多领域交叉的研究与应用领域，所涉及的领域包括：数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识系统、知识获取、信息检索、高性能计算以及可视化计算等领域。本书将主要介绍能够对大量数据进行挖掘处理的有关方法与技术，其中包括数据预处理、定性概念归纳、决策树分类方法、回归统计预测方法、关联规则发现方法、各种聚类算法，以及复杂数据，尤其是多媒体数据挖掘方法的最新研究成果；此外，还详细介绍了利用数据挖掘方法获取互联网信息，挖掘互联网使用知识，以及网络安全中数据挖掘方法应用等情况。

数据挖掘作为一门新兴的学科，它的发展与完善需要较长的过程；但是在它的形成与发展过程中却表现出强大的生命力，广大从事数据库应用与决策支持，以及模式识别、机器学习、专家系统、自动化等学科的科研工作者和工程技术人员迫切需要了解和掌握它。为此我们结合自己 10 多年来所从事的专家系统、机器学习、数据挖掘，以及互联网信息智能处理等方面科研与教学经验，以本人的博士论文及 20 余篇相关学术论文为基础，并以加拿大 Simon Fraser 大学韩家威教授的《Data Mining: Concepts and Techniques》专著为主要参考资料，编著完成了这本书，以飨读者。希望本书能对高等院校信息技术、模式识别、计算机技术、自动化等专业的教师、研究生和高年级本科生，以及从事数据库应用与决策支持系统设计与应用的广大科技人员有所帮助。

全书共分十章，主要论述数据挖掘的基本概念、方法、技术与应用等方面的内容。前七章主要介绍数据挖掘的主要方法；后三章侧重介绍数据挖掘在互联网信息智能信息处理方法的应用情况。

第一章主要介绍数据挖掘的基本概念，主要数据挖掘处理流程，目前数据挖掘研究现状与存在的问题，以及数据挖掘系统的分类与实际应用等内容。

第二章主要介绍数据挖掘中的一个主要处理：数据预处理中的主要方法。其中包括：数据清洗方法、数据集成方法、数据转换方法和数据消减方法。

第三章主要介绍定性归纳中两种主要基于属性的归纳方法，即定性概念描述和定性对比概念描述方法。

第四章主要介绍分类学习方法(决策树方法)、简单贝叶斯分类器、神经网络分类与规则抽取，以及其它如 k-最近邻方法、基于关联的分类方法等一些常见分类方法，此外，还介绍了基于回归模型的预测方法。最后还讨论了分类学习的结果评估问题。

第五章主要介绍 Apriori 关联规则挖掘算法。并在此基础上介绍了多层次、多属性等多种类型的关联规则挖掘方法。

第六章主要介绍常用的 k-means，k-medoids 和 CLARAN 聚类算法，在此基础上，介绍了基于层次、基于密度、基于网格、基于模型等其它一些典型的聚类方法。

第七章主要介绍对复杂类型的数据，如：对多媒体数据、空间地理数据、序列数据、时序数据，以及文本与互联网信息等，进行挖掘有关的方法。

第八章主要介绍互联网信息挖掘中有关网页智能搜索算法，以及网页信息智能抽取方法，并在此基础上，介绍了实现互联网信息个性化服务的一种解决方案。

第九章主要介绍互联网使用信息的数据挖掘方法，主要包括使用信息的预处理(用户的识别、会话的识别等)，以及挖掘结果的自动评估(主观评价标准的量化)方法。

第十章主要介绍网络安全中数据挖掘方法的应用，主要包括入侵检测中数据挖掘方法的应用，以及病毒检测中的数据挖掘方法应用。

本书在编写过程中得到了中国科学技术大学信息科学学院有关同志的鼎力相助，其中，周津同志帮助进行了本书第二、六、七章的文字校对工作；明鸣同志帮助进行了本书第三、四、九章的文字校对工作；严捷丰同志帮助进行了本书第一、五、八章的文字校对工作，在此对他们所给予的帮助表示感谢。

加拿大 Simon Fraser 大学韩家威教授，非常热心地向我赠送了他刚完成的 *Data Mining: Concepts and Techniques* 数据挖掘专著，使我受益匪浅，并得以将这几年数据挖掘课程的教学经验，充实到本书中，在此特向韩家威教授致以崇高的敬意和感谢。

本书在编写过程中，得到了中国科学技术大学研究生院和自动化系有关同志的大力支持，在此一并表示感谢。此外，本书在编写过程中，我们还参阅和摘引了许多先行者的学术论文，在此谨致谢意。

尽管我们十分努力，以求得本书在内容上尽善尽美，但是由于作者的水平有限，加之为了教学急需，时间较仓促，书中存在一些错误和不足肯定在所难免，祈请广大读者和同行专家不吝赐教。

朱 明

2002 年 2 月于中国科技大学

目　次

前　言 …… Ⅰ
第一章　数据挖掘导论 …… 1
1.1　数据挖掘发展简述 …… 1
1.1.1　数据丰富与知识匮乏 …… 1
1.1.2　从数据到知识 …… 2
1.1.3　数据挖掘产生 …… 3
1.2　数据挖掘基本知识 …… 5
1.2.1　数据挖掘定义 …… 5
1.2.2　数据挖掘深入 …… 7
1.3　数据挖掘功能 …… 9
1.3.1　概念描述：定性与对比 …… 9
1.3.2　关联分析 …… 11
1.3.3　分类与预测 …… 11
1.3.4　聚类分析 …… 12
1.3.5　异类分析 …… 13
1.3.6　演化分析 …… 13
1.4　数据挖掘结果的评估 …… 14
1.5　数据挖掘系统 …… 15
1.5.1　数据挖掘系统分类 …… 15
1.5.2　数据挖掘系统应用 …… 15
1.6　数据挖掘研究重点 …… 18
1.7　本章小结 …… 20
参考文献 …… 20
第二章　数据预处理 …… 22
2.1　数据预处理的重要性 …… 22
2.2　数据清洗 …… 23
2.2.1　遗漏数据处理 …… 24
2.2.2　噪声数据处理 …… 24
2.2.3　不一致数据处理 …… 26
2.3　数据集成与转换 …… 26
2.3.1　数据集成处理 …… 26
2.3.2　数据转换处理 …… 27

2.4 数据消减…… 28
2.4.1 数据立方合计…… 29
2.4.2 维数消减…… 30
2.4.3 数据压缩…… 31
2.4.4 数据块消减…… 32
2.5 离散化和概念层次树生成…… 35
2.5.1 数值概念层次树生成…… 36
2.5.2 类别概念层次树生成…… 38
2.6 本章小结…… 39
参考文献 …… 40
第三章 定性归纳 …… 42
3.1 概念描述基本知识…… 42
3.2 数据泛化与概要描述…… 43
3.2.1 数据泛化中的数据立方方法…… 43
3.2.2 基于属性归纳方法…… 44
3.2.3 基于属性归纳算法…… 47
3.2.4 基于属性归纳结果的表示…… 48
3.3 属性相关分析…… 50
3.3.1 属性相关分析意义…… 50
3.3.2 属性相关分析方法…… 51
3.3.3 分析定性描述示例…… 52
3.4 挖掘概念对比描述…… 54
3.4.1 概念对比方法与实现…… 54
3.4.2 概念对比描述的表示…… 56
3.4.3 概念的定性与对比描述的表示…… 57
3.5 挖掘大数据库的描述型统计信息…… 58
3.5.1 计算中心趋势…… 59
3.5.2 计算数据分布…… 59
3.6 方法讨论…… 60
3.6.1 概念描述:经典机器学习比较 …… 61
3.6.2 概念描述的递增和并行挖掘…… 61
3.7 本章小结…… 62
参考文献 …… 62
第四章 分类与预测 …… 64
4.1 分类与预测基本知识…… 64
4.2 有关分类和预测的若干问题…… 66
4.3 基于决策树的分类…… 67
4.3.1 决策树生成算法…… 67
4.3.2 属性选择方法…… 68

4.3.3 树枝修剪 …… 71
4.3.4 决策树中分类规则获取 …… 72
4.3.5 基本决策树方法的改进 …… 72
4.3.6 决策树归纳的可扩展性 …… 73
4.3.7 数据仓库技术与决策树归纳的结合 …… 74
4.4 贝叶斯分类方法 …… 76
4.4.1 贝叶斯定理 …… 76
4.4.2 基本贝叶斯分类方法 …… 77
4.4.3 贝叶斯信念网络 …… 78
4.4.4 贝叶斯信念网络的学习 …… 80
4.5 神经网络分类方法 …… 81
4.5.1 多层前馈神经网络 …… 81
4.5.2 神经网络结构 …… 82
4.5.3 后传方法 …… 82
4.5.4 后传方法和可理解性 …… 85
4.6 基于关联的分类方法 …… 87
4.7 其它分类方法 …… 88
4.7.1 k-最近邻方法 …… 88
4.7.2 基于示例推理 …… 89
4.7.3 遗传算法 …… 89
4.7.4 粗糙集方法 …… 90
4.7.5 模糊集合方法 …… 90
4.8 预测方法 …… 91
4.8.1 线性与多变量回归 …… 91
4.8.2 非线性回归 …… 93
4.8.3 其它回归模型 …… 93
4.9 分类器准确性 …… 93
4.9.1 分类器准确性估计 …… 94
4.9.2 提高分类器准确性 …… 95
4.9.3 有关分类器准确性的若干问题 …… 95
4.10 本章小结 …… 96
参考文献 …… 97
第五章 关联挖掘 …… 100
5.1 关联规则挖掘 …… 100
5.1.1 购物分析:关联挖掘 …… 100
5.1.2 基本概念 …… 101
5.1.3 关联规则挖掘分类 …… 102
5.2 单维布尔关联规则挖掘 …… 103
5.2.1 Apriori 算法 …… 103

5.2.2 关联规则的生成 …… 107
5.2.3 Apriori 算法的改进 …… 108
5.3 挖掘多层次关联规则 …… 110
5.3.1 多层次关联规则 …… 110
5.3.2 挖掘多层次关联规则方法 …… 111
5.3.3 多层次关联规则的冗余 …… 114
5.4 多维关联规则的挖掘 …… 115
5.4.1 多维关联规则 …… 115
5.4.2 利用静态离散挖掘多维关联规则 …… 116
5.4.3 挖掘定量关联规则 …… 117
5.4.4 挖掘基于距离的关联规则 …… 118
5.5 关联挖掘中的相关分析 …… 120
5.5.1 无意义强关联规则示例 …… 120
5.5.2 从关联分析到相关分析 …… 122
5.6 基于约束的关联挖掘 …… 122
5.6.1 基于元规则的关联挖掘 …… 122
5.6.2 基于规则约束的关联挖掘 …… 123
5.7 本章小结 …… 125
参考文献 …… 126
第六章 聚类分析 …… 129
6.1 聚类分析概念 …… 129
6.2 聚类分析中的数据类型 …… 131
6.2.1 间隔数值属性 …… 132
6.2.2 二值属性 …… 133
6.2.3 符号、顺序和比例数值属性 …… 135
6.2.4 混合类型属性 …… 136
6.3 主要聚类方法 …… 137
6.4 划分方法 …… 138
6.4.1 传统划分方法 …… 139
6.4.2 大数据库的划分方法 …… 142
6.5 层次方法 …… 143
6.5.1 两种基本层次聚类方法 …… 143
6.5.2 两种层次聚类方法 …… 145
6.5.3 层次聚类方法:CURE …… 146
6.5.4 层次聚类方法:CHAMALEON …… 147
6.6 基于密度方法 …… 149
6.6.1 基于密度方法:DBSCAN …… 149
6.6.2 基于密度方法:OPTICS …… 150
6.7 基于网格方法 …… 151

6.7.1 基于网格方法:STING …… 151
6.7.2 基于网格方法:CLIQUE …… 152
6.8 基于模型聚类方法 …… 154
6.8.1 统计方法 …… 154
6.8.2 神经网络方法 …… 155
6.9 异常数据分析 …… 157
6.9.1 基于统计的异常检测方法 …… 157
6.9.2 基于距离的异常检测方法 …… 158
6.9.3 基于偏差的异常检测方法 …… 159
6.10 本章小结 …… 161
参考文献 …… 162
第七章 复杂数据的挖掘 …… 164
7.1 多维分析与描述性知识挖掘 …… 164
7.1.1 结构数据的泛化 …… 164
7.1.2 空间和多媒体数据的泛化 …… 165
7.1.3 对象类/子类层次的泛化 …… 166
7.1.4 继承和产生性质的泛化 …… 166
7.1.5 类组成结构的泛化 …… 167
7.1.6 对象立方的构造与挖掘 …… 167
7.1.7 基于泛化的挖掘 …… 167
7.2 空间数据库挖掘 …… 170
7.2.1 空间数据立方与 OLAP …… 170
7.2.2 空间关联分析 …… 173
7.2.3 空间聚类分析 …… 174
7.2.4 空间分类与趋势分析 …… 174
7.2.5 光栅数据库挖掘 …… 174
7.3 多媒体数据库挖掘 …… 175
7.3.1 多媒体数据的相似搜索 …… 175
7.3.2 多媒体的多维分析 …… 176
7.3.3 多媒体数据分类与预测分析 …… 177
7.3.4 多媒体数据的关联分析 …… 178
7.4 时序数据和序列数据挖掘 …… 178
7.4.1 趋势分析 …… 179
7.4.2 时序数据中的相似搜索 …… 180
7.4.3 序列模式挖掘 …… 182
7.4.4 周期性分析 …… 183
7.5 文本数据库挖掘 …… 183
7.5.1 文本数据分析和信息检索 …… 184
7.5.2 文本挖掘 …… 187

7.6 互联网挖掘 …… 188
7.6.1 Web 链接挖掘 …… 189
7.6.2 Web 文档自动分类 …… 190
7.6.3 构造多层次 Web 信息库 …… 191
7.6.4 Web 使用的挖掘 …… 192
7.7 本章小结 …… 192
参考文献 …… 193
第八章 互联网信息挖掘 …… 196
8.1 Web 信息挖掘简介 …… 196
8.1.1 Web 信息挖掘意义 …… 196
8.1.2 Web 网页基本搜索方法 …… 197
8.2 Web 网页智能搜索 …… 201
8.2.1 Web 信息搜索工具包:WebSuite …… 201
8.2.2 基于主题的 Web 信息搜索 …… 202
8.2.3 基于强化学习的 Web 搜索 …… 204
8.3 Web 网页信息抽取 …… 206
8.3.1 基于层次结构的信息抽取:STALKER …… 206
8.3.2 可视化网页信息抽取:W4F …… 210
8.3.3 基于概念模型的多记录信息抽取 …… 213
8.4 Web 信息的自主搜索 …… 218
8.4.1 自主搜索的重要性 …… 218
8.4.2 自主搜索问题描述 …… 219
8.4.3 自主搜索知识表示 …… 219
8.4.4 自主搜索算法 …… 223
8.4.5 搜索知识的获取方法 …… 226
8.5 Web 信息的自主抽取 …… 229
8.5.1 信息抽取的应用 …… 229
8.5.2 信息抽取问题描述 …… 230
8.5.3 抽取知识表示方法 …… 231
8.5.4 Web 抽取知识表示方法 …… 235
8.5.5 信息抽取算法 …… 238
8.5.6 抽取知识的获取 …… 243
8.6 Web 个性化信息服务 …… 245
8.6.1 个性化信息服务意义 …… 245
8.6.2 个性化信息服务问题描述 …… 246
8.6.3 个性化信息需求描述 …… 247
8.6.4 信息的搜索与获取 …… 249
8.6.5 信息集成 …… 253
8.7 本章小结 …… 254

参考文献 …… 255
第九章 互联网使用挖掘 …… 259
9.1 Web使用挖掘的应用 …… 259
9.2 数据源与数据模型 …… 262
9.2.1 Web数据源 …… 262
9.2.2 数据建模 …… 264
9.3 网站结构与内容的预处理 …… 265
9.3.1 结构预处理 …… 265
9.3.2 内容预处理 …… 267
9.4 网站使用数据的预处理 …… 268
9.4.1 数据清洗 …… 269
9.4.2 用户与会话识别 …… 269
9.4.3 网页浏览识别 …… 273
9.4.4 补全路径 …… 275
9.5 使用模式挖掘方法 …… 278
9.5.1 模式发现方法概述 …… 278
9.5.2 模式挖掘算法 …… 279
9.6 使用模式评估 …… 282
9.6.1 有趣性评估标准 …… 282
9.6.2 信息过滤器 …… 283
9.6.3 证据量化 …… 285
9.6.4 结构证据量化 …… 287
9.6.5 内容证据量化 …… 289
9.7 本章小结 …… 290
参考文献 …… 290
第十章 网络安全数据挖掘 …… 293
10.1 入侵检测中的数据挖掘 …… 293
10.1.1 网络安全概述 …… 293
10.1.2 存在的问题 …… 294
10.1.3 审计数据的挖掘 …… 295
10.1.4 基于属性趣味的挖掘 …… 295
10.1.5 挖掘模式的使用 …… 299
10.1.6 错误使用的检测应用 …… 303
10.2 邮件病毒检测中的数据挖掘 …… 304
10.2.1 恶意邮件过滤器概述 …… 304
10.2.2 与Procmail的结合 …… 306
10.2.3 监视邮件附件的传播 …… 307
10.2.4 基于数据挖掘的检测模型 …… 307
10.2.5 恶意邮件附件的检测方法 …… 308

10.2.6 恶意邮件附件的检测实验结果…… 308
10.3 病毒程序检测中的数据挖掘…… 309
10.3.1 恶意程序检测概述…… 309
10.3.2 恶意程序的检测方法…… 310
10.3.3 恶意程序的特征抽取…… 311
10.3.4 恶意程序的检测算法…… 313
10.3.5 恶意程序的检测模型…… 314
10.3.6 恶意程序的检测结果…… 315
10.4 本章小结…… 316
参考文献…… 317

第一章 数据挖掘导论

数据挖掘作为一个新兴的多学科交叉应用领域，正在各行各业的决策支持活动中扮演着越来越重要的角色。本书将介绍数据挖掘(Data Mining)与数据库知识发现(Knowledge Discovery from Databases)的基本知识，以及从大量有噪声，不完整，甚至是不一致数据集合中，挖掘出有意义的模式知识所涉及的概念与技术方法。

本章将从数据管理技术演化角度，介绍数据挖掘的由来。以及数据挖掘的作用和意义。同时还将介绍数据挖掘系统的结构、数据挖掘所获得的知识种类，以及数据挖掘系统的分类。最后还简要介绍了当前数据挖掘领域存在的一些热点问题。

1.1 数据挖掘发展简述

1.1.1 数据丰富与知识匮乏

计算机与信息技术经历了半个世纪的发展，给人类社会带来了巨大的变化与影响。在支配人类社会三大要素(能源、材料和信息)中，信息愈来愈显示出其重要性和支配力，它将人类社会由工业化时代推向信息化时代。随着人类活动范围的扩展，生活节奏的加快，以及技术的进步，人们能以更快速、更容易、更廉价的方式获取和存储数据，这就使得数据及其信息量以指数方式增长。据粗略估算，早在 20 世纪 80 年代，全球信息量每隔 20 个月就要增加一倍。而进入 90 年代，全世界所拥有的数据库及其所存储的数据规模增长更快。一个中等规模企业每天要产生 100MB 以上来自各生产经营等多方面的商业数据。美国政府部门的一个典型大数据库每天要接收约 5TB 数据量，在 15 秒到 1 分钟时间里，要维持的数据量达到 300TB，存档数据达 15PB ～ 100PB。在科研方面，以美国宇航局的数据库为例，每天从卫星下载的数据量就达 3TB ～ 4TB 之多；而为了研究的需要，这些数据要保存七年之久。90 年代互联网(Internet)的出现与发展，以及随之而来的企业内部网(Intranet)和企业外部网(Extranet)以及虚拟私有网(VPN：Virtual Private network)的产生和应用，使整个世界互联形成一个小小的地球村，人们可以跨越时空，在网上交换信息和协同工作。这样，展现在人们面前的已不是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋。据估计，1993 年全球数据存储容量约为 2000TB，到 2000 年增加到 300 万 TB，面对这极度膨胀的数据信息量，人们受到“信息爆炸”、“混沌信息空间”(Information Chaotic Space)和“数据过剩”(Data glut)的巨大压力。

然而，人类的各项活动都是基于人类的智慧和知识，即对外部世界的观察和了解，做出正确的判断和决策以及采取正确的行动，而数据仅仅是人们用各种工具和手段观察外部世界所得到的原始材料，它本身没有太多意义。从数据到知识再到智慧，需要经过分析加工处理精炼

的过程。如图 1.1 所示，数据是原材料，它只是描述发生了什么事情，并不能构成决策或行动的可靠基础。通过对数据进行分析找出其中关系，赋予数据以某种意义和关联，这就形成所谓信息。信息虽然给出了数据中一些有一定意义的东西，但是它往往和人们需要完成的任务没有直接的联系，也还不能作为判断、决策和行动的依据。对信息进行再加工，即进行更深入的归纳分析，方能获得更有用的信息，即知识。而所谓知识，可定义为“信息块中的一组逻辑联系，其关系是通过上下文或过程的贴近度发现的”。从信息中理解其模式，即形成知识。在大量知识积累基础上，总结出原理和法则，就形成所谓智慧(Wisdom)。事实上，一部人类文明发展史，就是在各种活动中，知识的创造、交流，再创造不断积累的螺旋式上升的历史。

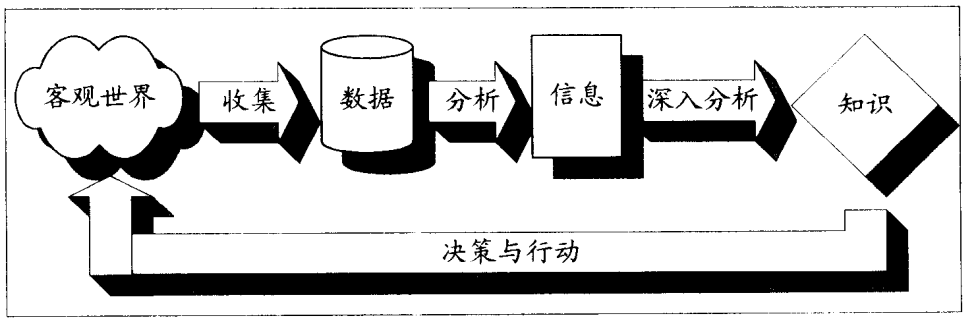

图 1.1 人类活动所涉及数据与知识之间的关系描述

计算机与信息技术的发展，加速了人类知识创造与交流的这种进程，据德国《世界报》的资料分析，如果说 19 世纪时科学定律(包括新的化学分子式，新的物理关系和新的医学认识)的认识数量 100 年增长一倍，到 20 世纪 60 年代中期以后，每 5 年就增加一倍。这其中知识起着关键的作用。当数据量极度增长时，如果没有有效的方法，利用计算机及信息技术来帮助人们从中提取有用的信息和知识，那么人类显然就会感到像大海捞针一样束手无策。据估计，目前一个大型企业数据库中的数据，约只有 7%得到很好地应用。因此目前人类陷入了一个尴尬的境地，即“丰富的数据”(data rich)和“贫乏的知识”(knowledge poor)并存。

1.1.2 从数据到知识

早在 20 世纪 80 年代，人们在“物竞天择，适者生存”的大原则下，就认识到“谁最先从外部世界获得有用信息并加以利用，谁就可能成为赢家”。而今置身市场经济且面向全球性剧烈竞争的环境下，任何商家的优势都不是单纯地取决于如产品、服务、地区等方面因素，而在于创新。用知识作为创新的原动力，就能使商家长期持续地保持竞争优势。因此要能及时迅速地从日积月累庞大的数据库中，以及互联网上获取与经营决策相关的知识，自然而然就成为满足易变的客户需求以及因市场快速变化而引起激烈竞争局面的唯一武器。因此，如何对数据与信息快速有效地进行分析、加工、提炼以获取所需知识，就成为计算机及信息技术领域的重要研究课题。

事实上计算机及信息技术发展的历史，也是数据和信息加工手段不断更新和改善的历史。早年受技术条件限制，一般用人工方法进行统计分析和用批处理程序进行汇总和提出报告。在当时市场情况下，月度和季度报告已能满足决策所需信息要求。随着数据量的增长，多数据源带来的各种数据格式不相容性，为了便于获得决策所需信息，就有必要将整个机构内的数据以统一形式集成存储在一起，这就是形成了数据仓库(data warehousing)。数据仓库不同于管

理日常工作数据的数据库，它是为了便于分析针对特定主题(subject-oriented)的集成化的、时变的(time-variant)即提供存储 5 ～ 10 年或更长时间的数据，这些数据一旦存入就不再发生变化。

数据仓库的出现，为更深入地对数据进行分析提供了条件，针对市场变化的加速，人们提出了能进行实时分析和产生相应报表的在线分析工具 OLAP(On Line Analytical Processing)。OLAP 能允许用户以交互方式浏览数据仓库内容，并对其中数据进行多维分析，且能及时地从变化和不太完整的数据中提取出与企业经营活动密切相关的信息。例如，OLAP 能对不同时期、不同地域的商业数据中变化趋势进行对比分析。

OLAP 是数据分析手段的一大进步，以往的分析工具所得到的报告结果只能回答“什么”(What)，而 OLAP 的分析结果能回答“为什么”(Why)。但 OLAP 分析过程是建立在用户对深藏在数据中的某种知识有预感和假设的前提下，由用户指导的信息分析与知识发现过程。但由于数据仓库(通常数据贮藏量以 TB 计)内容来源于多个数据源，因此，其中埋藏着丰富的不为用户所知的有用信息和知识，而要使企业能及时准确地做出科学的经营决策，以适应变化迅速的市场环境，就需要有基于计算机与信息技术的智能化自动工具，来帮助挖掘隐藏在数据中的各类知识。这类工具不应再基于用户假设，而应能自身生成多种假设；再用数据仓库(或大型数据库)中的数据进行检验或验证；然后返回用户最有价值的检验结果。此外，这类工具还应能适应现实世界中数据的多种特性(即量大、含噪声、不完整、动态、稀疏性、异质、非线性等)。要达到上述要求，只借助于一般数学分析方法是不能达到的。多年来，数理统计技术方法以及人工智能和知识工程等领域的研究成果，诸如推理、机器学习、知识获取、模糊理论、神经网络、进化计算、模式识别、粗糙集理论等等诸多研究分支，给开发满足这类要求的数据深度分析工具提供了坚实而丰富的理论和技术基础。

20 世纪 90 年代中期以来，许多软件开发商，基于数理统计、人工智能、机器学习、神经网络、进化计算和模式识别等多种技术和市场需求，开发了许多数据挖掘与知识发现软件工具，从而形成了近年来软件开发市场的热点。目前数据挖掘工具已开始向智能化整体数据分析解决方案发展，这是从数据到知识演化过程中的一个重要里程碑。如图 1.2 所示。

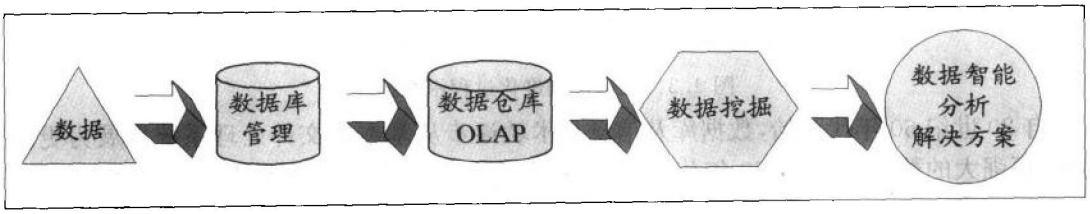

图 1.2 数据到知识的演化过程示意描述

1.1.3 数据挖掘产生

随着计算机硬件和软件的飞速发展，尤其是数据库技术与应用的日益普及，人们面临着快速扩张的数据海洋，如何有效利用这一丰富数据海洋的宝藏为人类服务，业已成为广大信息技术工作者所关注的焦点之一。与日趋成熟的数据管理技术与软件工具相比，人们所依赖的数据分析工具功能，却无法有效地为决策者提供其决策支持所需要的相关知识，从而形成了一种“丰富的数据，贫乏的知识”之独特的现象。为有效解决这一问题，自 20 世纪 80 年代开始，数

据挖掘技术逐步发展起来，数据挖掘技术的迅速发展，得益于目前全世界所拥有的巨大数据资源，以及对将这些数据资源转换为信息和知识资源的巨大需求，对信息和知识的需求来自各行各业，从商业管理、生产控制、市场分析到工程设计、科学探索等等。数据挖掘可以视为是数据管理与分析技术的自然进化产物，如图 1.3 所示。

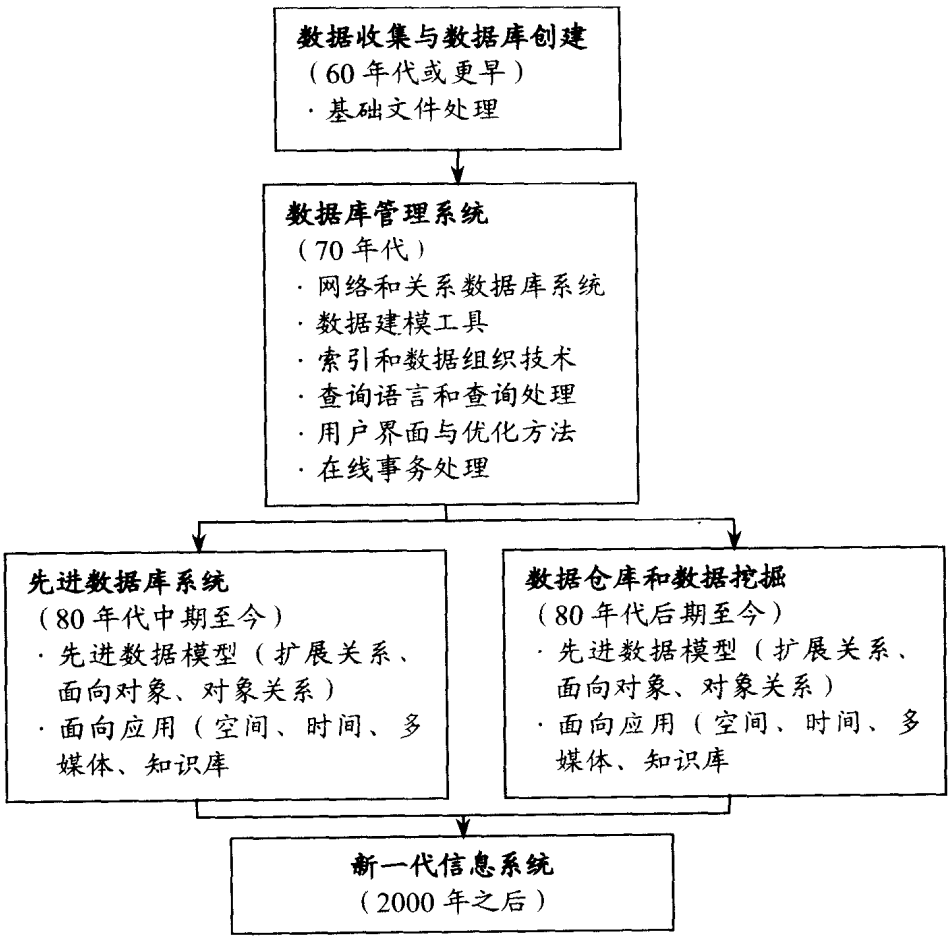

图 1.3 数据挖掘进化过程示意描述

自 20 世纪 60 年代开始，数据库及信息技术就逐步从基本的文件处理系统发展为更复杂功能、更强大的数据库系统；70 年代的数据库系统的研究与发展，最终导致了关系数据库系统、数据建模工具、索引与数据组织技术的迅速发展，这时用户获得了更方便灵活的数据存取语言和界面；此外在线事务处理(OLTP：on-line transaction processing)手段的出现也极大地推动了关系数据库技术的应用普及，尤其是在大数据量存储、检索和管理的实际应用领域。

自 20 世纪 80 年代中期开始，关系数据库技术被普遍采用，新一轮研究与开发新型与强大的数据库系统悄然兴起，并提出了许多先进的数据模型：扩展关系模型、面向对象模型、演绎模型等，以及应用数据库系统：空间数据库、时序数据库、多媒体数据库等。目前异构数据库系统和基于互联网的全球信息系统也已开始出现并在信息工业中开始扮演重要角色。

被收集并存储在众多数据库中且正在快速增长的庞大数据，已远远超过人类的处理和分析理解能力(在不借助功能强大的工具情况下)，这样存储在数据库中的数据就成为“数据坟