

基于内容的 视频结构挖掘

付畅俭◎著

湘潭大学出版社

基于内容的 视频结构挖掘

付畅俭◎著

湘潭大学出版社

图书在版编目(CIP)数据

基于内容的视频结构挖掘 / 付畅俭著. —湘潭:
湘潭大学出版社, 2016.8

ISBN 978-7-81128-992-3

I. ①基… II. ①付… III. ①视频编辑软件—数据处
理—研究 IV. ①TP317.53

中国版本图书馆 CIP 数据核字(2016)第 200802 号

基于内容的视频结构挖掘

JIYU NEIRONG DE SHIPIN JIEGOU WAJUE

付畅俭 著

责任编辑: 姚海琼

封面设计: 曾天怡

出版发行: 湘潭大学出版社

社 址: 湖南省湘潭市 湘潭大学出版大楼

电话(传真): 0731-58298966 邮编: 411105

网 址: <http://press.xtu.edu.cn/>

印 刷: 长沙理工大印刷厂

经 销: 湖南省新华书店

开 本: 787×1092 1/16

印 张: 18.25

字 数: 386 千字

版 次: 2016 年 8 月第 1 版

印 次: 2016 年 8 月第 1 次印刷

书 号: ISBN 978-7-81128-992-3

定 价: 48.00 元

(版权所有 严禁翻印)

目 录

第一篇 数据挖掘基础	1
第1章 绪论	1
1.1 数据挖掘的发展	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	6
1.2 什么是数据挖掘	7
1.2.1 数据挖掘的定义	7
1.2.2 与数据挖掘有关的举例	8
1.3 数据挖掘的主要功能	11
1.3.1 挖掘频繁模式、关联和相关性	11
1.3.2 用于预测分析的分类与回归	13
1.3.3 聚类分析	14
1.3.4 孤立点分析	15
1.3.5 趋势和演变分析	15
1.4 数据挖掘步骤	16
1.4.1 数据挖掘的体系结构	16
1.4.2 知识发现过程	17
1.4.3 数据挖掘标准流程——CRISP - DM	18
1.4.4 SEMMA	20
1.5 数据挖掘软件	22
1.5.1 第一类:由统计软件发展而来	22
1.5.2 第二类:由数据库软件发展而来	23

1.5.3	第三类:第三方软件	23
1.5.4	第四类:开源数据挖掘软件	24
第2章	数据认识与数据预处理	30
2.1	数据属性	30
2.2	数据统计描述	31
2.2.1	均值	32
2.2.2	中位数	32
2.2.3	众数	33
2.2.4	根差	34
2.2.5	四分位距	34
2.2.6	方差和标准差	35
2.3	数据预处理	35
2.3.1	数据集成	36
2.3.2	数据变换	37
2.3.3	数据归约	38
2.4	概念描述	40
2.4.1	面向属性的归纳(Attribute – Oriented Induction, AOI)	40
2.4.2	面向属性归纳的基本算法	42
2.4.3	挖掘类比较:区分不同的类	46
第3章	SQL SERVER 建立数据仓库	50
3.1	数据仓库:基本概念	50
3.1.1	什么是数据仓库	50
3.1.2	什么是 OLAP	51
3.1.3	数据仓库的类型	51
3.1.4	数据仓库与传统数据库的比较	52
3.1.5	数据仓库的设计	53
3.2	SQL SERVER 建立数据仓库	54
3.2.1	基本概念	54
3.2.2	建立数据仓库	56
3.3	应用举例	74
3.3.1	使用 SQL SERVER 实现贝叶斯分类方法	74

3.3.2 使用 Excel 实现关联规则分析	75
3.3.3 使用 Excel 实现其他数据挖掘分析	77
第4章 关联规则挖掘	94
4.1 关联规则的定义和属性	94
4.1.1 关联规则的定义	94
4.1.2 关联规则的属性	95
4.1.3 关联规则挖掘过程	96
4.1.4 关联规则挖掘的经典算法	96
4.1.5 关联规则应用	97
4.2 Apriori 算法	98
4.2.1 Apriori 算法伪代码	99
4.2.2 Apriori 算法举例	100
4.2.3 由频繁项集产生关联规则	102
4.2.4 Apriori 算法的 python 实现	103
4.2.5 Apriori 算法改进	106
4.3 Fp - Growth 算法	107
4.3.1 构造 FP 树	107
4.3.2 FP - growth 挖掘过程	109
4.4 关联规则评价	111
4.5 关联规则挖掘进展	112
4.5.1 增量式更新挖掘算法	112
4.5.2 并行/分布式关联规则挖掘算法	112
4.5.3 多层关联规则挖掘算法	113
4.5.4 在数据挖掘类型方面的拓展	113
4.5.5 基于约束的频繁模式挖掘	113
第5章 分类与预测	115
5.1 分类与预测基本知识	115
5.2 ID3 算法	117
5.2.1 ID3 算法伪代码	117
5.2.2 ID3 算法的 python 实现	121
5.2.3 ID3 的优缺点	127

5.3	C4.5 算法	128
5.3.1	C4.5 的改进	128
5.3.2	C4.5 算法的 python 实现	130
5.4	Python 中 Scikit - Learn 机器学习简介	138
5.4.1	数据预处理	138
5.4.2	主要分类方法	140
5.4.3	算法参数优化	142
5.4.4	利用 graphviz 和 pydot 数据可视化	144
第 6 章	聚类分析	145
6.1	聚类分析概述	145
6.1.1	聚类分析的定义	145
6.1.2	聚类方法概述	146
6.1.3	聚类分析的一般步骤	147
6.1.4	聚类方法的要求	148
6.2	聚类的主要算法	149
6.2.1	划分方法	149
6.2.2	层次法	153
6.2.3	基于密度的方法	158
6.2.4	模糊聚类	163
6.3	聚类进展	170
第 7 章	时间序列变动趋势挖掘	172
7.1	时间序列影响因素及分析模型	172
7.1.1	时间序列及其作用	173
7.1.2	时间序列的影响因素	173
7.1.3	时间序列分解模型	174
7.1.4	时间序列趋势挖掘步骤	175
7.2	时间序列季节调整	176
7.2.1	季节调整处理方法	176
7.2.2	季节因子	177
7.2.3	季节调整	179
7.3	时间序列趋势挖掘	182

7.3.1	时距扩大法	183
7.3.2	移动平均法	184
7.3.3	指数平滑法	185
7.3.4	最小平方法	187
7.4	时间序列循环因子测定	193
7.4.1	循环变动的测定目的	194
7.4.2	循环变动的测定方法	194
7.4.3	乘法模型	195
7.4.4	加法模型	197
7.4.5	乘加模型	197
 第二篇 基于内容的视频结构挖掘		199
 第1章 视频结构挖掘概念框架		199
1.1	视频结构挖掘的概念框架	199
1.2	视频结构挖掘的系统结构	203
1.2.1	传统数据挖掘过程	203
1.2.2	视频结构挖掘系统结构	205
1.3	视频结构挖掘的功能结构	207
1.4	视频结构挖掘方法	208
1.4.1	基本结构单元的发现和表现	208
1.4.2	视频单元分类	209
1.4.3	视频单元聚类	209
1.4.4	视频单元关联规则挖掘	210
1.5	本章小结	211
 第2章 视频基本结构挖掘		214
2.1	视频基本结构	214
2.2	镜头分割及关键帧提取	215
2.2.1	镜头分割常用方法	215
2.2.2	视频镜头关键帧抽取	219
2.2.3	镜头分割颜色模型选择及量化	220
2.2.4	自适应双直方图两次判别镜头分割算法	223

2.3	场景构造	227
2.3.1	镜头特征提取及相似性度量	228
2.3.2	基于镜头多特征聚类视频场景分割	234
2.3.3	基于竞争力的场景分割	235
2.4	本章小结	237
第3章	视频结构语法挖掘	240
3.1	视频结构语法	240
3.2	视频数据预处理	242
3.2.1	镜头分割	242
3.2.2	特征提取	243
3.2.3	镜头符号序列化	244
3.3	视频结构模式挖掘	246
3.3.1	传统关联规则挖掘	246
3.3.2	视频结构关联规则挖掘	249
3.3.3	视频结构语法模式分析	251
3.4	本章小结	254
第4章	视频结构语义挖掘	257
4.1	视频结构语义	257
4.2	镜头语义概念挖掘	259
4.2.1	镜头语义概念挖掘主要图模型	261
4.2.2	多概念判别随机场模型	263
4.2.3	广义多概念判别随机场模型	268
4.3	视频高层结构语义事件挖掘	269
4.3.1	视频高层结构语义事件	269
4.3.2	HMM 模型	271
4.3.3	HHMM 模型	273
4.3.4	篮球视频高层事件 HHMM 模型	275
4.4	本章小结	279

第一篇 数据挖掘基础

第1章 绪 论

1.1 数据挖掘的发展

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 个月就增加一倍,而进入九十年代,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 分钟的时间里,要维持的数据量达到 300TB,存档数据达 15 - 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 3 - 4TB 之多;而为了研究的需要,这些数据要保存七年之久。九十年代互联网(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。今天,全球数据存储量呈现爆炸式增长,企业及互联网数据以每年 50% 的速率在增长,据 Gartner 预测,到 2020 年,全球数据量将达到 35ZB,等于 80 亿块 4TB 硬盘。数据结构变化给存储系统带来新的挑战。非结

结构化数据在存储系统中所占据比例已接近 80%。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,它本身没有任何意义。从数据到知识到智慧,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。信息虽给出了数据中一些有一定意义的东西,但它往往和人们需要完成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进度发现的”。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。

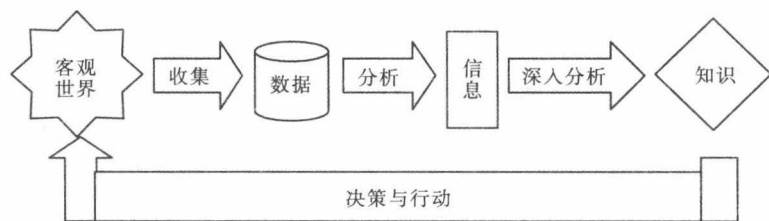


图 1-1 人类活动所涉及数据与知识之间的关系描述

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)的认识数量一百年增长一倍,到 20 世纪 60 年代中期以后,每五年就增加一倍。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,由计算机及信息技术来帮助从中提取有用的信息和知识,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数据,只有约百分之七得到很好地应用。因此,目前人类陷入了一个尴尬的境地,即“丰富的数据”(data rich)而“贫乏的知识”(knowledge poor)。

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,任何商家的优势不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的要求。随着数据量的增长,多数据源所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),即提供存贮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(OnLine 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的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,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无法达到的。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。

九十年代中期以来,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图

图 1-3 左图是包含单个数据“21”的 Excel 表,我们假定那是你的年龄,它仅是一个数据,一种描述你已经出生的时间。图 1-3 中图是一个商业活动的顾客年龄列表,这就是商业活动可能用到的有价值的信息。通过这组年龄列表,你可以得到这组顾客年龄的平均值、最小值、最大值以及各年龄段的频率分布图。

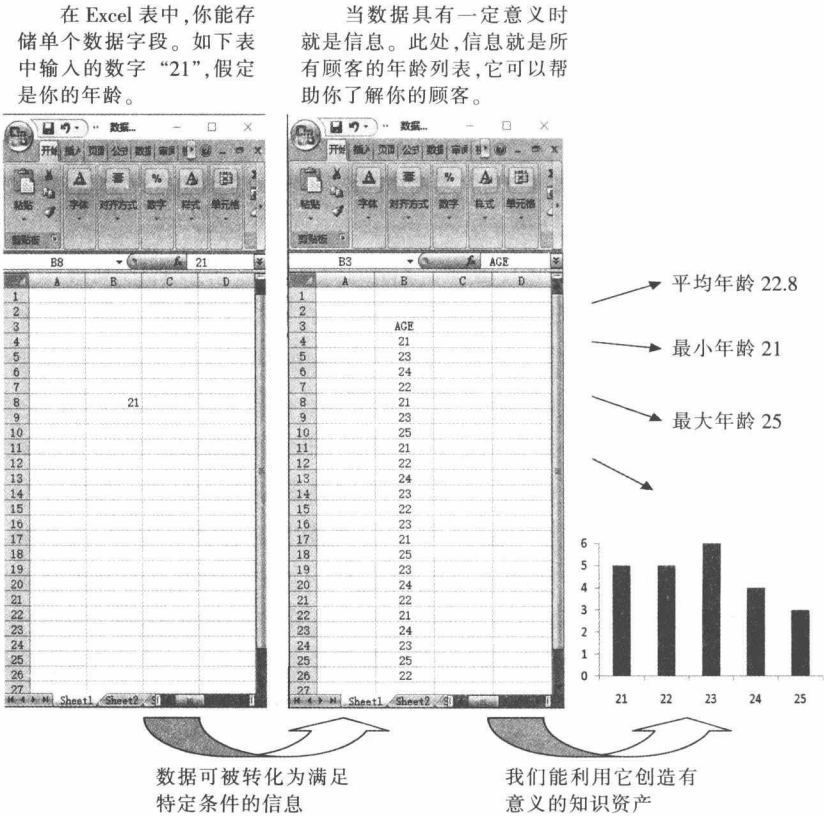
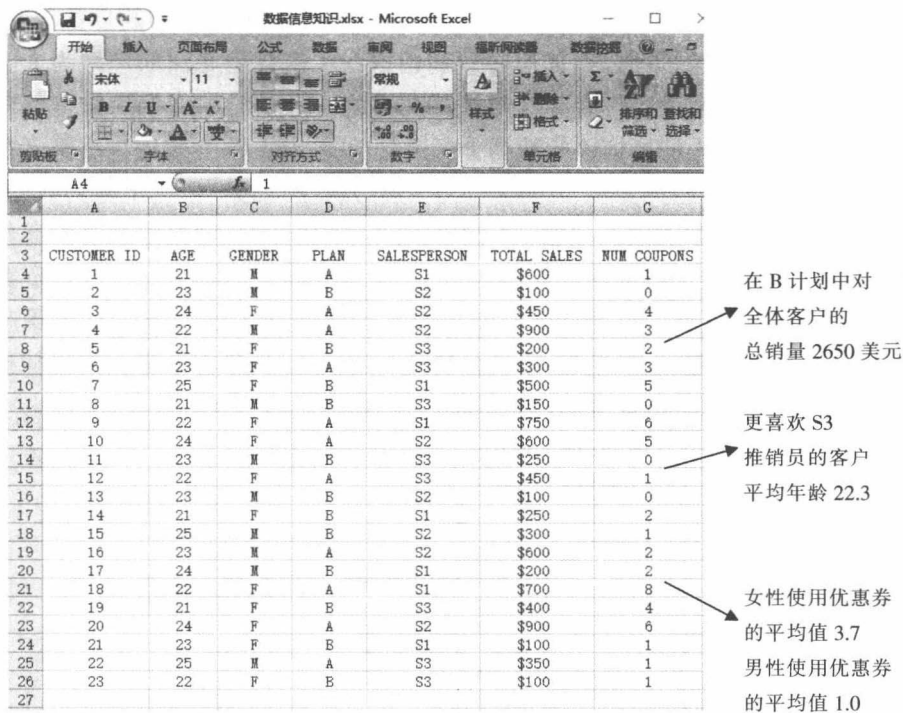


图 1-3 数据与信息

而在图 1-4 中,Excel 表单包含每个顾客的许多类信息,这就是商务智能。细心观察信息的条目,我们可以发现每位顾客偏好哪位收银员和使用的赠券数目,并且我们还可以获得一些更有意义的信息——商务智能,比如对比男性顾客和女性顾客使用赠券的差别、偏好某一收银员的顾客的平均年龄。



当你把若干组详细组合起来时,你就可获得商务智能信息,帮助你制定更有效的战略决策。

图 1-4 商务智能

注意数据、信息和商务智能彼此之间的关系。信息是多个数据点的合理组合,如上例中,一个年龄是单独的一段数据,而信息是所有顾客的年龄集;商务智能将顾客年龄信息扩展包括性别信息、赠券使用量、偏好的收银员以及总购买额;知识基于以上所有,你可以通过实践利用信息和智慧在某一业务或行业中获取知识。接下来对知识的解释可能会更有利于你的理解。

知识是一个比较宽泛的概念:(1)它包括对商务智能的内容解释;(2)也指为获得商务智能而采取的行动;(3)包括知识资产,如专利和商标;(4)也包括在组织内什么是最佳的实践行为。

以图 1-3 和图 1-4 为例,能解释女性消费者比男性消费者使用更多的赠券的原因是大部分赠券都附赠予在女性杂志上,这就是知识。也可以通过图 1-4 展示的商务智能信息制定营销战略的过程来创造知识。再以图 1-4 商务智能信息

为例,组织中具有相关知识的人能促进组织利用这些信息,如市场战略应该制定让 B 计划中的更多顾客增加总购买量。

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,已成为广大信息技术工作者重点关注的焦点之一。与日趋成熟的数据管理技术和软件工具相比,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,从而形成了一种独特的现象——“丰富的数据,贫乏的知识”。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。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 1-5 所示。

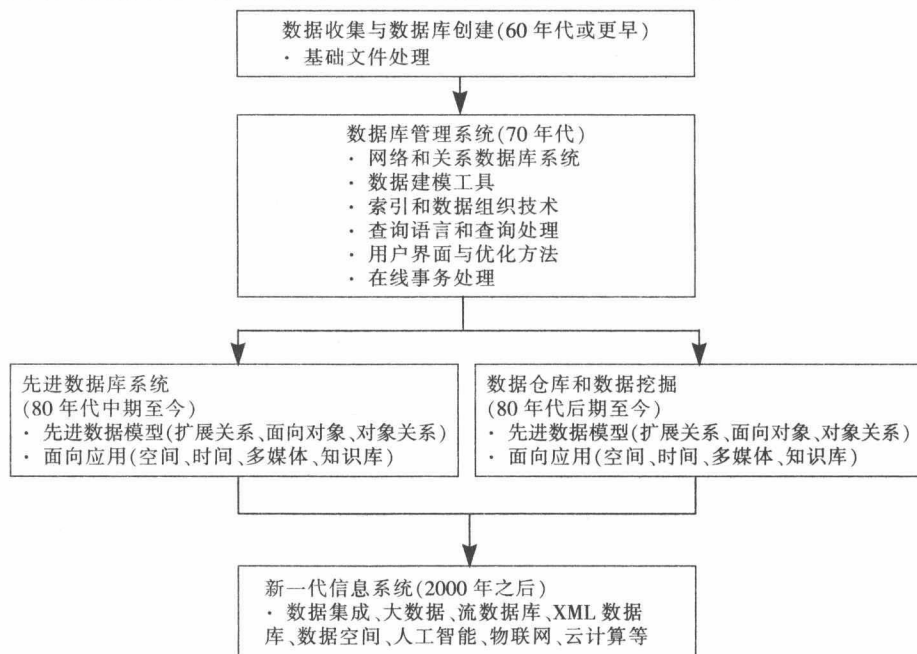


图 1-5 数据挖掘进化过程示意描述

自 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: OnLine 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80年代中期开始,关系数据库技术被普遍采用,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、多媒体数据库系统等),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“数据坟墓”,即这些数据极少被访问,结果许多重要的决策不是基于这些基础数据而是依赖决策者的直觉而制定的,其中的原因很简单,这些决策的制定者没有合适的工具帮助其从数据中抽取出所需的信息知识。而数据挖掘工具可以帮助从大量数据中发现所存在的特定模式规律,从而可以为商业活动、科学探索和医学研究等诸多领域提供所必需的信息知识。数据与信息知识之间的巨大差距迫切需要系统地开发数据挖掘工具,来帮助实现将“数据坟墓”中的数据转化为知识财富。

1.2 什么是数据挖掘

1.2.1 数据挖掘的定义

1. 数据挖掘的技术定义

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。

数据挖掘和信息检索:信息检索和数据挖掘的相同点是从档案文件或数据库中抽取感兴趣的数据和信息。区别在于数据检索对信息的抽取规则是事先定义好的,抽取的是外在信息。数据挖掘在于挖掘寻找现象之间事先未知的关系和关联。

2. 数据挖掘的商业定义

按企业既定业务目标,对大量的企业数据进行探索和分析,揭示隐藏的、未知的或验证已知的规律性,并进一步将其模型化的先进有效的方法。数据挖掘是从海量数据中提取隐含在其中的有用信息和知识的过程。它可以帮助企业对数据进行微观、中观乃至宏观的统计、分析、综合和推理,从而利用已有数据预测未来,帮助企业赢得竞争优势。

应用实例:某经营公司对多年来的客户资料进行挖掘后发现,大多数购买电脑的客户具有下面的特点:①年轻(20—45岁之间);②收入高;③居住地:城市;④学历高。基于此,此经营公司可以根据这些客户的特点有目的的做一些广告或者促销。

1.2.2 与数据挖掘有关的举例

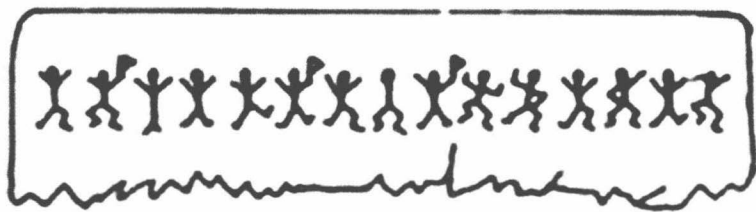
例1 Google 成功预测冬季流感

数据挖掘把大型数据集转换成知识。像 Google 这样的搜索引擎每天接受数亿次查询。每个查询都被看作一个事务,用户通过事务描述他们的信息需求。随着时间的推移,搜索引擎可以从这些大量的搜索查询中学到什么样的新颖的、有用的知识?有趣的是,从众多用户查询中发现的某些模式能够揭示无价的知识,这些知识无法通过仅读取个体数据项得到。例如,2009年,Google 通过分析 5000 万条美国人最频繁检索的词汇,将之和美国疾病中心在 2003 年到 2008 年间季节性流感传播时期的数据进行比较,并建立一个特定的数学模型。最终 Google 成功预测了 2009 冬季流感的传播甚至可以具体到特定的地区和州。Google 的 Flu Trends(流感趋势)使用特殊的搜索项作为流感活动的指示器。它发现了搜索流感相关信息的人数与实际具有流感症状的人数之间的紧密联系。当与流感相关的所有搜索都聚集在一起时,一个模式就出现了。使用聚集的搜索数据,Google 的 Flu Trends 可以比传统的系统早两周对流感活动做出评估。这个例子表明,数据挖掘如何把大型数据集转化成知识,帮助我们应对当代的全球性挑战。

例2 跳舞的人

亚瑟·柯南道尔所写的广为流传的神奇故事——《福尔摩斯探案集之跳舞的人》,大名鼎鼎的福尔摩斯通过破解加密体制,展示了他非凡的聪明才智。下面是故事中关于密码部分的概要。

希尔顿最近结婚了,他给福尔摩斯发去了一封信,信中有一张纸是他在花园中发现的,这张纸是用跳舞的棒形小人所写的。



两个星期后,又发现有人用粉笔在他的工具间的门上写下了另外一些小人的信息: