

信息管理与信息系统专业
核心课程精品教材系列

DATA WAREHOUSE AND DATA MINING

数据仓库 与数据挖掘

IMIS

周根贵 主编
浙江大学出版社

图书在版编目 (CIP) 数据

数据仓库与数据挖掘 / 周根贵主编. — 杭州: 浙江大学出版社, 2004.8

(信息管理 with 信息系统专业核心课程精品教材系列)

ISBN 7-308-03831-9

I. 数... II. 周... III. ①数据库系统 - 高等学校 - 教材②数据采集 - 高等学校 - 教材

IV. ①TP311.13②TP274

中国版本图书馆 CIP 数据核字 (2004) 第 081182 号

丛书策划 樊晓燕

封面设计 俞亚彤

责任编辑 杜玲玲

出版发行 浙江大学出版社

(杭州天目山路 148 号 邮政编码 310028)

(E-mail: zupress@mail. hz. zj. cn)

(网址: <http://www.zjupress.com>)

排 版 杭州兴邦印务有限公司

印 刷 富阳市育才印刷有限公司

开 本 787mm×960mm 1/16

印 张 19.25

字 数 335 千

版 次 2004 年 8 月第 1 版 2006 年 4 月第 2 次印刷

印 数 3001—4500

书 号 ISBN 7-308-03831-9/TP·254

定 价 28.00 元

信息管理与信息系统专业核心课程精品教材系列

编 委 会

主 任 吴晓波

副主任 陈畴镛 周根贵 琚春华

编 委 王晓耘 卢向南 叶 枫

李小东 邵 雷 凌 云

序

关于信息化与后工业社会的话题已经被世人热烈地讨论了二十余年,正式的“信息管理与信息系统”专业在我国大学中的设立亦已是最近十来年的事。而近三年来,该专业却是我国大学本科乃至专科专业中发展和普及最快的专业之一。它表明了一个重要的事实:信息化在我国新型工业化进程中正在扮演一个极其重要的角色。我国社会、经济的迅速发展对于既掌握信息技术,同时又拥有管理知识,能够运用信息技术于管理实践的复合型人才的需求正日益高涨。

信息管理与信息系统是一门具有交叉性的复合型学科,它融合了计算机科学、信息技术、管理学、经济学、系统科学、运筹学、组织行为学等学科的知识,它强调运用定性与定量相结合的方法及相关学科的研究手段,深入研究并解决各类社会系统中的信息管理问题。该专业直接以满足信息化建设人才与新型复合型管理人才的需求为目标,培养具有现代管理科学理论知识,具备较强的计算机及网络技术运用能力的,适合在经济管理部门、各类企事业单位从事信息系统建设和管理以及从事相应科学研究等工作的综合型高级专门人才。1998年教育部调整学科专业目录,把原有的管理信息系统、经济信息管理、科技信息、图书情报检索、信息学及林业信息管理等专业改为“信息管理与信息系统”专业,作为管理学门类中“管理科学与工程”一级学科之下的

一个二级学科,使之在培养目标、内容和方向上均得到了进一步的凝练和提升。此后,由于社会、经济信息化进程的加速,该专业得到了快速的催生和发展。

众所周知,信息技术是当代发展最快的技术之一,相应的管理应用及经济、法律体系亦正面临着深刻的变化,因此,相关的教材亦面临技术基础多变和快速更新的挑战。为促进和支持该专业的发展,浙江大学出版社及时组织了有关专家在充分讨论和酝酿的基础上,精心组织出版了这套“信息管理与信息系统专业核心课程精品教材”。本人作为教育部高校管理科学与工程类教学指导委员会成员,参与了全国高校“信息管理与信息系统专业核心课程及其教学大纲”的修订工作。在组织本系列教材的编写时,亦强调了与教育部教学指导委员会规范化要求的一致与契合。本系列教材的编著者均有长期从事该专业教学和科研的经验,他们的前期工作经验使本系列教材的质量有了可靠的保证。相信这套教材的出版能为高校信息管理与信息系统专业教学水平的提高和规范化发展起到积极的推进作用。

吴晓波

2004年8月于求是园

前 言

计算机网络与数据库技术的迅速发展和广泛应用,使得各行各业的管理工作进入了一个崭新的时代。广大基层管理人员摆脱了繁重的制表业务和数据处理工作,管理工作进一步规范化。各种在线事务处理信息系统的建立,对各种日常业务处理提供了有效的支持。然而,面对当今竞争日趋激烈与瞬息万变的市场,各级管理人员迫切需要根据组织的现状和历史数据做出判断和决策。他们希望能够从组织的信息系统中获取有效的、一致的决策支持信息,及时准确地把握市场变化的脉搏,做出正确有效的判断和抉择。概括地说,数据处理的重点应该从传统的业务处理扩展到在线分析处理,并从中得到面向各种主题的统一统计信息和决策支持信息。

数据仓库和数据挖掘技术就是针对上述问题而产生的一种技术方案,它是基于大规模数据库的决策支持系统环境的核心。数据仓库是面向主题的、集成的、不可更新的、随时间不断变化的数据集合,用以支持经营管理中的决策制定过程;而数据挖掘是从大量的数据中提取出隐含的、以前不为人所知的、可信而有效的知识。它能够对数据进行再分析,以期获得更加深入的了解。它具有预测功能,可通过已有数据预测未来。数据仓库与数据挖掘技术相结合,与现代的管理决策方法相结合,就能使数据仓库在组织的经营管理决策中发挥巨大的作用。

本书的编写旨在提供数据仓库和数据挖掘领域的一个广博的且也是深入的概览。全书共分10章。第1章主要介绍数据仓库和数据挖掘的产生背景、应用和发展;第2章介绍数据仓库的组成和结构等技术与开发的基本概念以及第3章介绍数据仓库的技术管理;第4章详细介绍数据仓库的查询工具OLAP;第5章介绍数据仓库的应用与开发工具SQL Server;第6章介绍数据挖掘和知识发现的基本概念,以及数据挖掘的方法与技术;第7章详细介绍统计类的数据挖掘技术;第8章介绍知识类的数据挖掘技术;第9章进一步介绍21世纪各类新的数据挖掘技术;第10章代表性地介绍一些数据仓库与数据挖掘的综合应用场合和实例。

本书第1章由浙江工业大学周根贵编写;第2,4,5章由浙江工业大学黄洪编写;第3,6,7,8章由杭州电子科技大学陈朵玲编写;第9章由浙江科技学院李崇岩编写;第10章由浙江科技学院顾忠伟编写。全书最后由周根

贵教授修改、统稿。

在本书的编写过程中,得到浙江工业大学朱艺华教授和杭州电子科技大学陈畴镛教授的大力支持与帮助,在此表示衷心的感谢。

我们在编写本书的过程中,尽可能做到深入浅出,力求概念正确,理论联系实际。由于数据库与数据挖掘是一个新的领域,发展非常迅速,加之我们水平有限,书中一定存在许多不足之处,恳切希望各位读者批评指正。

作者

2004年5月

目 录

第 1 章 概论	1
1.1 决策支持技术与数据库的发展	2
1.1.1 决策支持技术的发展.....	2
1.1.2 数据库技术的发展.....	4
1.2 数据仓库概述	7
1.2.1 数据仓库概念的提出.....	8
1.2.2 数据仓库的定义	10
1.2.3 数据仓库的特征	11
1.2.4 数据仓库的应用和发展	13
1.3. 数据挖掘概述	18
1.3.1 数据挖掘的定义.....	19
1.3.2 数据挖掘与数据仓库的关系.....	21
1.3.3 数据挖掘的应用和发展.....	22
本章小结	27
习题	27
第 2 章 数据仓库的技术与开发	29
2.1 数据仓库的体系结构.....	30
2.1.1 用户眼中的数据仓库结构	30
2.1.2 数据仓库系统的体系结构	31
2.1.3 数据集市	35
2.2 元数据.....	36
2.2.1 元数据的定义	36
2.2.2 元数据的主要作用	37
2.2.3 元数据分类	37
2.3 数据仓库的数据模型.....	39
2.3.1 概念模型	39
2.3.2 逻辑模型	43
2.3.3 物理模型	45
2.4 粒度和分割.....	46

2.4.1	粒度的确定	46
2.4.2	粒度划分实例	47
2.4.3	数据分割	49
2.5	数据仓库和开发流程	50
2.6	总线型结构的数据仓库	52
2.6.1	统一的维	52
2.6.2	统一的事实	52
2.6.3	数据仓库总线	53
本章小结	54
习题	54
第3章	数据仓库管理技术	55
3.1	数据仓库管理的基本内容	56
3.2	休眠数据管理	57
3.2.1	休眠数据的定义与理解	58
3.2.2	休眠数据的处理	60
3.3	元数据的管理	62
3.3.1	传统的元数据管理方法	63
3.3.2	企业级中心知识库的管理方法	63
3.4	数据清理	65
3.4.1	脏数据的来源和清理	65
3.4.2	过期数据的清理	67
本章小结	67
习题	68
第4章	联机分析处理	69
4.1	概述	70
4.1.1	OLAP 的定义	70
4.1.2	OLAP 的基本概念	72
4.1.3	OLAP 的基本分析操作	75
4.1.4	OLAP 和 OLTP 的比较	78
4.2	多维 OLAP 和关系 OLAP	79
4.2.1	数据存储	79
4.2.2	MOLAP 和 ROLAP 的特征	83
4.2.3	星型模式	84
4.3	OLAP 的新发展——OLAM	86



4.3.1	OLAM 应该具有的功能特征	87
4.3.2	OLAM 的主要发展方向	87
4.3.3	基于 Web 的 OLAM 须解决的问题	88
	本章小结	88
	习题	88
第 5 章	SQL Server 数据仓库的应用与开发	89
5.1	概述	90
5.2	连接数据源	92
5.3	建多维数据集	94
5.3.1	建立数据库	94
5.3.2	建立数据源	94
5.3.3	建立多维数据集	96
5.3.4	编辑多维数据集	107
5.3.5	设计存储和处理多维数据集	108
5.4	浏览多维数据集数据	109
5.5	创建、使用数据挖掘模型	112
	本章小结	126
	习题	126
第 6 章	数据挖掘与知识发现	127
6.1	知识发现与数据挖掘的概念	128
6.1.1	数据挖掘的任务	131
6.1.2	数据挖掘的分类	133
6.1.3	数据挖掘的对象	137
6.1.4	数据挖掘与专家系统的区别	139
6.2	数据挖掘方法与技术	140
6.2.1	归纳学习方法	141
6.2.2	仿生物技术	142
6.2.3	公式发现	144
6.2.4	统计分析方法	145
6.2.5	模糊数学方法	147
6.2.6	可视化技术	148
6.3	数据挖掘的知识表示	148
6.3.1	规则	149
6.3.2	决策树	149

6.3.3	知识基(浓缩数据).....	150
6.3.4	网络权值.....	151
6.3.5	公式.....	151
	本章小结.....	152
	习题.....	152
第7章	统计类数据挖掘技术	153
7.1	基本概念.....	154
7.1.1	统计学.....	154
7.1.2	统计类数据挖掘技术.....	155
7.2	最简单的统计类挖掘技术.....	155
7.2.1	聚集函数与度量.....	156
7.2.2	柱状图.....	156
7.3	回归分析数据挖掘技术.....	156
7.3.1	线性回归数据挖掘技术.....	157
7.3.2	非线性回归数据挖掘技术.....	159
7.4	聚类分析与最近邻挖掘技术.....	163
7.4.1	聚类的概念.....	163
7.4.2	最近邻技术.....	164
7.4.3	聚类分析与最近邻技术的运用.....	165
7.4.4	聚类分析应用示例.....	172
7.5	统计分析工具及其使用——SPSS.....	174
7.5.1	统计分析工具.....	174
7.5.2	统计分析工具应用.....	178
7.5.3	SPSS 及其应用.....	182
	本章小结.....	185
	习题.....	186
第8章	知识类数据挖掘技术	187
8.1	知识发现系统的结构.....	188
8.2	关联规则的数据挖掘技术.....	190
8.2.1	关联规则描述.....	191
8.2.2	关联规则的定义.....	192
8.2.3	关联规则的种类.....	192
8.2.4	关联规则挖掘算法——频繁集方法.....	193
8.2.5	关联规则应用举例.....	195

8.3	神经网络的数据挖掘技术	197
8.3.1	人工神经元及其互连结构	198
8.3.2	神经网络模型	202
8.3.3	神经网络的应用	208
8.4	遗传算法的数据挖掘技术	209
8.4.1	遗传算法概述	210
8.4.2	遗传算子	212
8.4.3	遗传算法的应用	217
8.5	粗糙集的数据挖掘技术	219
8.5.1	粗糙集概念	220
8.5.2	粗糙集分类规则发现模式	223
8.5.3	粗糙集的应用	224
8.6	知识发现工具简介	225
8.6.1	知识发现工具的系统结构	225
8.6.2	知识发现工具运用中的问题	227
8.6.3	知识发现的作用	230
8.6.4	知识类数据挖掘工具简介	230
	本章小结	232
	习题	233
第9章	21世纪的数据挖掘技术	235
9.1	文本挖掘技术	236
9.1.1	文本挖掘的概述	236
9.1.2	信息检索系统	237
9.1.3	文本挖掘	239
9.2	Web数据挖掘技术	241
9.2.1	Web的特点	241
9.2.2	Web结构挖掘	242
9.2.3	Web内容挖掘	243
9.2.4	Web日志挖掘	244
9.3	可视化数据挖掘技术	245
9.3.1	数据可视化技术	245
9.3.2	可视化数据挖掘技术的应用	246
9.4	基于GIS的空间数据挖掘技术	247
9.4.1	地理信息系统	247

9.4.2	空间数据挖掘	249
9.5	分布式数据挖掘技术	251
9.5.1	概述	251
9.5.2	适合水平式数据划分的分布式挖掘方法	252
9.5.3	适合垂直式数据划分的分布式挖掘方法	253
9.6	数据挖掘的其他主题	254
9.6.1	视频和音频数据挖掘	254
9.6.2	科学和统计数据挖掘	255
9.6.3	数据挖掘的理论基础	256
9.6.4	数据挖掘和智能查询应答	257
	本章小结	259
	习题	259
第 10 章	数据仓库与数据挖掘的综合应用	261
10.1	数据仓库在信息管理中的实际应用	262
10.1.1	应用数据仓库弥补 ERP 的不足	264
10.1.2	建立良好的客户关系——数据仓库实现分析型 CRM	266
10.1.3	数据仓库提高供应链管理的效率	267
10.1.4	用数据仓库支持企业决策	269
10.1.5	数据仓库促使企业重构业务过程	271
10.2	金融业中的数据挖掘	272
10.2.1	数据挖掘在银行领域的应用	273
10.2.2	数据挖掘在证券领域的应用	275
10.2.3	数据挖掘在保险领域的应用	279
10.3	零售业中的数据挖掘	280
10.4	电信业中的数据挖掘	283
	本章小结	288
	习题	288
	参考文献	289



第1章

概 论

学习目标

- 决策支持技术的发展
- 数据库技术的发展
- 数据仓库的应用和发展
- 数据挖掘的应用和发展



信息技术的快速发展将人类社会带入了以信息的存储、交换、处理、使用为特征的信息时代。以事务处理系统(TPS)、管理信息系统(MIS)为代表的_{事务型}信息处理主要是对管理信息进行日常的收集、传递、存储、加工、维护和使用等操作处理;而以决策支持系统(DSS)、高级经理支持系统(ESS)为代表的_{信息型}处理主要是通过访问大量的历史数据,做出相应的逻辑分析,为管理者的决策服务。正是由于这种信息型分析处理的快速发展,使得管理信息的处理从原来的以单一数据库为中心的数据环境发展为需要一种新的数据环境。数据仓库与数据挖掘正是为了构建这种新的分析处理环境而出现的一种数据存储和组织技术。


1.1 决策支持技术与数据库的发展

在各行各业,每日、每时、每刻都有大量的管理信息需要去处理、去使用。这些管理信息的处理主要有操作型(事务型)处理和分析型(信息型)处理两种。人们为了在复杂的经营环境中,做出快速响应的有效决策,以应对各种由于市场变化所带来的挑战,一方面不断地加强事务型处理的功能,开发出更加完善的事务处理的信息系统;另一方面不断探讨和开发具有分析处理功能的信息系统。这种新型的信息系统不仅是信息系统对决策支持技术发展的结果,也是信息处理所基于的数据库技术发展的结果。

1.1.1 决策支持技术的发展

数据处理(EDP)是电子计算机应用中最广泛的领域,约占70%。一个国家的现代化水平越高,数据处理的面越宽、量越大,数据处理所占的比例就越高。随着20世纪50年代到60年代数据处理领域应用的成功,20世纪60年代到70年代西方国家兴起了管理信息系统(MIS)的热潮。我国是在20世纪70年代末到80年代初才兴起了管理信息系统的应用。管理信息系统是在管理科学利用计算机后发展起来的,它使计算机的应用由数值计算领域拓宽到数据处理(非数值计算)领域,使计算机走向社会和家庭。具体地说,管理信息系统是一个由人、计算机等组成的能进行数据的收集、传递、存储、加工、维护和使用_{的系统},它主要支持_{事务型}的数据处理和信息管理。

20世纪70年代初,运筹学和系统工程工作者利用计算机形成了模型



辅助决策系统。由于采用的模型主要是数学模型,所以其辅助决策的能力主要表现在定量分析上,从而发展起把管理信息系统和模型辅助决策系统结合起来的决策支持系统(DSS)。DSS主要是进行分析处理,使得数值计算和数据处理融为一体,提高了辅助决策的能力。它将数据、复杂的分析模型和用户友好的软件集成在一起,形成能够很好地支持各种复杂决策问题的信息系统,其目的是辅助管理决策。

另一方面,随着计算机技术的迅猛发展,20世纪60年代末兴起了一个新研究领域——专家系统(Expert System, ES),它是对20世纪50年代人工智能的进一步发展。专家系统是利用专家的知识在计算机上进行推理,达到专家解决问题的能力。1968年E. A. Feigenhanm等人研制了DEN-DRAL专家系统,用来帮助化学家推断分子结构。1974年E. H. Shortliffe等人研制了MYCIN专家系统,用来诊断和治疗感染性疾病。同一时期,人们还研制出不少其他专家系统。专家系统的出现使人工智能走上了实用化阶段。

专家的知识表现为产生式规则和语义网络等形式。知识的推理是采用符号逻辑中的假言推理。在搜索知识的时候,采用了深度优先或启发式搜索方法。专家系统也是一种很有效的辅助决策系统,它是利用专家的知识,特别是经验知识,经过推理得出辅助决策信息。对于专家知识,不规定它是数值的,更多的是不精确的定性知识。因此,专家系统辅助决策的方式属于定性分析。

专家系统和决策支持系统几乎是同时兴起,并沿着各自的道路发展起来的,它们都能起到辅助决策的作用,但辅助决策的方式完全不同。专家系统辅助决策的方式属于定性分析,决策支持系统辅助决策的方式属于定量分析。如果把这两者结合起来,辅助决策的效果将会大大改善,即达到定性辅助决策和定量辅助决策相结合。这种专家系统与决策支持系统相结合而形成的系统称为智能决策支持系统(IDSS),它是决策支持系统的发展方向。

决策支持系统和专家系统的结合,并不是那样容易实现的,因为它们自成体系,要结合它们将有一些技术难题需要解决。专家系统结构中核心的部分由推理机、知识库和动态数据库三部分组成。知识库存放大量的专家知识;推理机完成对知识的搜索和推理;动态数据库存放已知的事实和推出的结果。专家系统中的动态数据库不同于决策支持系统中的数据库,相对来说,决策支持系统中的数据库是静态数据库。两系统中各部件之间的接口以及两系统的集成是形成智能决策支持系统的关键。

因此,不仅决策支持技术本身的发展对数据环境提出了更高的要求,而且决策支持技术与专家系统、人工智能的结合更要求能提供一种新的数据环境,为决策分析提供必要的数据源。只有当模型技术、专家系统以及这种新的数据环境的全方位的有机集成,才使得决策支持技术无论是在体系结构还是在信息处理能力上都产生了较大的变化,形成了人们熟悉而期望的智能决策支持系统。

1.1.2 数据库技术的发展

数据库(Database)一词起源于20世纪50年代。当时美国因战争需要,把各种情报集中在一起,存放在计算机中,称为Information Base或Database。数据库技术是研究数据库结构、存储、设计和使用的—门软件科学,于20世纪60年代中期产生,经过短短30年它已从第一代的网状、层状数据库,第二代的关系数据库系统,发展到第三代以面向对象模型为主要特征的数据库(尽管其在学术上和技术上都尚不够成熟,但1990年DBMS功能委员会发表的“第三代数据库宣言”已标志了第三代数据库的出现),再到目前的数据仓库和数据集市等几个阶段。

数据模型是数据库系统的核心和基础。因此,数据库发展阶段划分应以数据模型的进展为主要依据和标志。数据模型根据其应用不同,可分为两大类或两个层次:①概念数据模型;②结构数据模型。其中概念数据模型只强调信息特征和语义,是现实世界到信息世界的第一层抽象,而用于划分数据库发展阶段的是用于机器世界的第二层抽象,即结构数据模型。层次数据库系统和网状数据库系统的结构数据模型(以下均简称为数据模型或模型)都是在20世纪60年代后期研究和开发的。它们从体系结构、数据库语言到数据存储管理均具有共同特征,可称为第一代数据库系统(见图1.1,图1.2)。

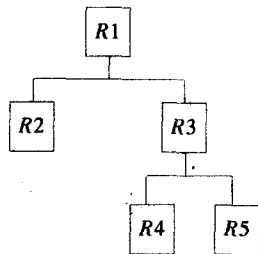


图 1.1 层次模型

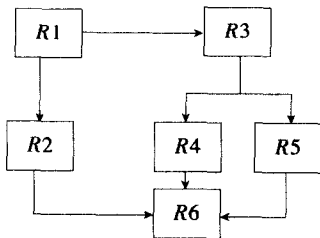


图 1.2 网状模型