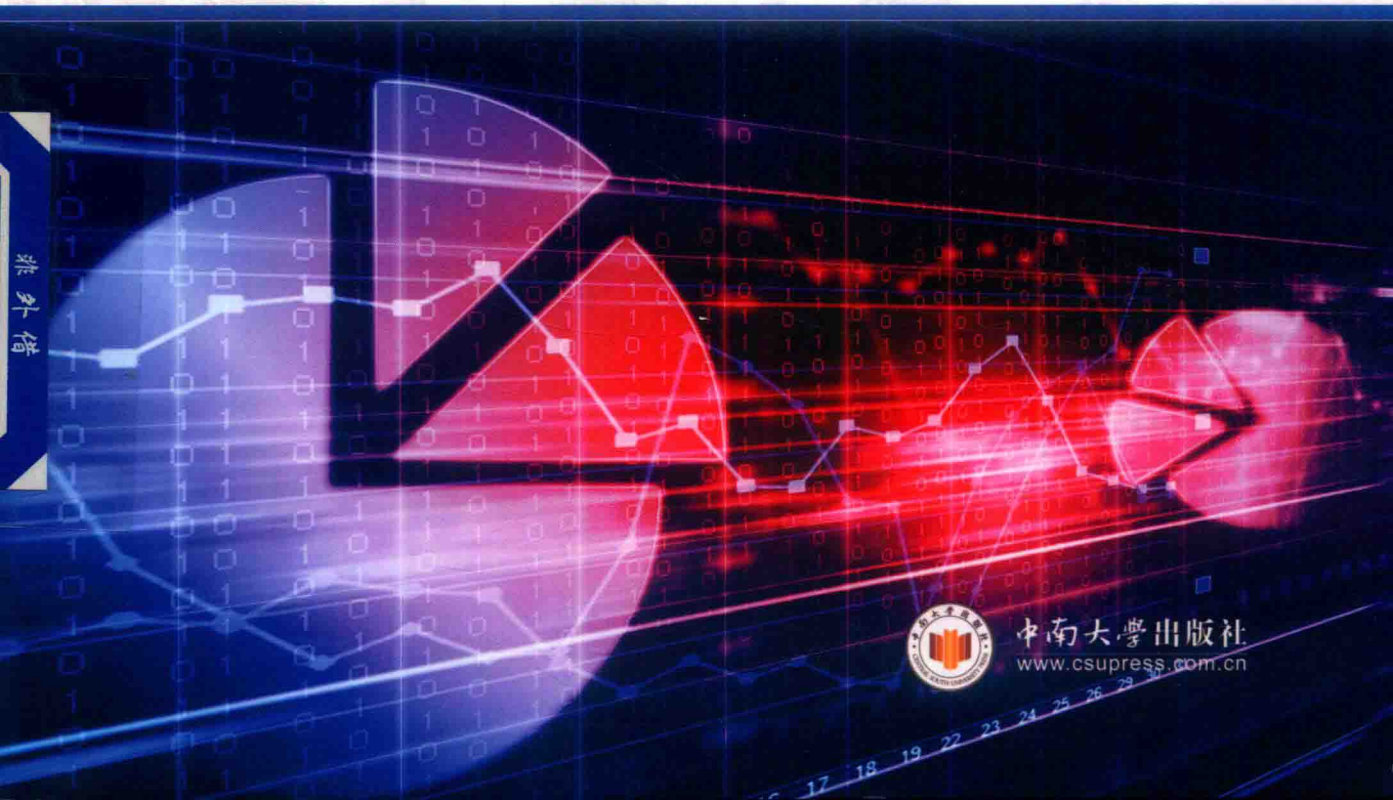
普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据仓库与数据挖掘

SHUJU CANGKU 与 SHUJU WAJUE

龙　军　章成源 ⊙ 编著



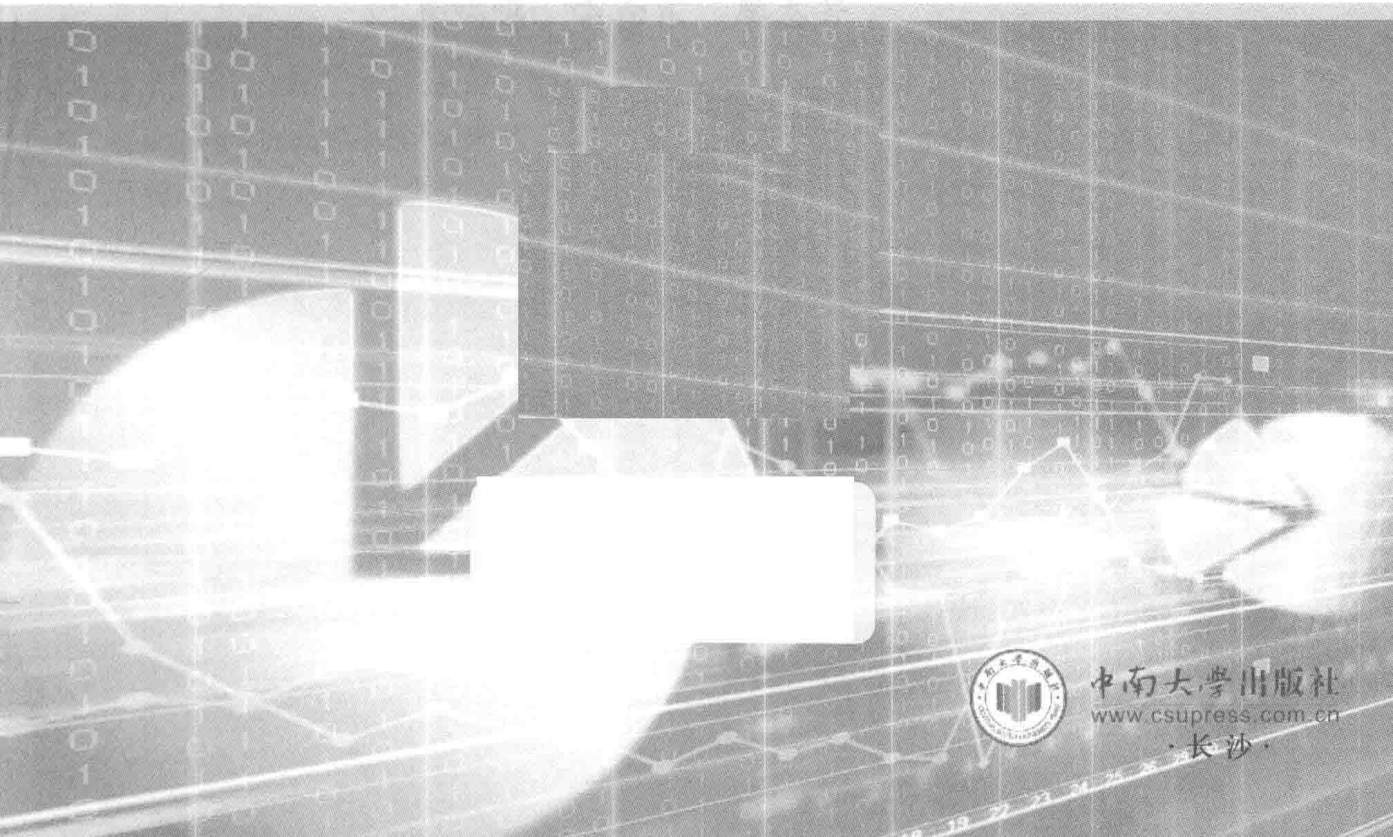
中南大学出版社
www.csupress.com.cn

普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据仓库与数据挖掘

SHUJU CANGKU 与 SHUJU WAJUE

龙　军　章成源 ⊙ 编著



中南大学出版社
www.csupress.com.cn
·长沙·

图书在版编目（CIP）数据

数据仓库与数据挖掘 / 龙军，章成源编著. --长沙：中南大学出版社，2018.12

ISBN 978-7-5487-3171-9

Ⅰ.①数… Ⅱ.①龙… ②章… Ⅲ.①数据库系统 ②数据采集 Ⅳ.①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字(2018)第 071236 号

数据仓库与数据挖掘

龙 军 章成源 编著

□责任编辑 韩 雪
□责任印制 易红卫
□出版发行 中南大学出版社
社址：长沙市麓山南路 邮编：410083
发行科电话：0731-88876770 传真：0731-88710482
□印 装 长沙德三印刷有限公司

□开 本 787×1092 1/16 □印张 17.25 □字数 435 千字
□版 次 2018 年 12 月第 1 版 □印次 2018 年 12 月第 1 次印刷
□书 号 ISBN 978-7-5487-3171-9
□定 价 45.00 元

普通高等院校数据科学与大数据技术专业“十三五”规划教材

编委会

总序

Preface

随着移动互联网的兴起，全球数据呈爆炸性增长，目前90%以上的数据是近年产生的，数据规模大约每两年翻一番；而随着人工智能下物联网生态圈的形成，数据的采集、存储及分析处理、融合共享等技术需求都能得到响应，各行各业都在体验大数据带来的革命，“大数据时代”真正来临。这是一个产生大数据的时代，更是需要大数据力量的时代。

大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据从产生、分析到利用提出了前所未有的新要求。高等教育只有转变观念，更新方法与手段，寻求变革与突破，才能在大数据与人工智能的信息大潮面前立于不败之地。据预测，中国近年来大数据相关人才缺口达200万人，全世界相关人才缺口更超过1000万人之多。我国教育部门为了响应社会发展需要，率先于2016年开始正式开设“数据科学与大数据技术”本科专业及“大数据技术与应用”专科专业，近几年，全国形成了申报与建设大数据相关专业的热潮。随着专业建设的深入，大家发现一个共同的难题：没有成系列的大数据相关教材。

中南大学作为首批申报大数据专业的学校，2015年在我校计算机科学与技术专业设立大数据方向时，信息科学与工程学院院领导便意识到系列教材缺失的严重问题，因此院领导规划由课程团队在教学的同时积累素材，形成面向大数据专业知识体系与能力体系、老师自己愿意用、同学觉得买得值、关联性强的系列教材。经过两年的准备，针对2017年《教育部办公厅关于推荐新工科研究与实践项目的通知》的精神，中南大学出版社组织对系列教材文稿进行相应的打磨，最终于2018年底出版“高等院校数据科学与大数据技术专业‘十三五’规划教材”。

该套系列教材具有如下特点：

1. 本套教材主要参照“数据科学与大数据技术”本科专业的培养方案，综合考虑专业的来源，如从计算机类专业、数学统计类专业以及经济类专业发展而来；同时适当兼顾了专科类偏向实际应用的特点。

2. 注重理论联系实际，注重能力培养。该系列教材中既有理论教材也有配套的实践教程。力图通过理论或原理教学、案例教学、课堂讨论、课程实验与实训实习等多个环节，训练学生掌握知识、运用知识分析并解决实际问题的能力，以满足学生今后就业或科研的需求；同时兼顾“全国工程教育专业认证”对学生基本能力的培养要求与复杂问题求解能力的

要求。

3. 在规范教材编写体例的同时，注重写作风格的灵活性。本套系列教材中每本书的内容都由教学目的、本章小结、思考题或练习题、实验要求等组成。每本教材都配有 PPT 电子教案及相关的电子资源，如实验要求及 DEMO、配套的实验资源管理与服务平台等。本套系列教材的文本层次分明、逻辑性强、概念清晰、图文并茂、表达准确、可读性强，同时相关配套电子资源与教材的相关性强，形成了新媒体式的立体型系列教材。

4. 响应了教育部“新工科”研究与实践项目的要求。本套教材从专业导论课开始设立相关的实验环节，作为知识主线与技术主线把相关课程串接起来，力争让学生尽早具有培养自己动手能力的意识、综合利用各种技术与平台的能力。同时为了避免新技术发展太快、教材纸质文字内容容易过时的问题，在相关技术及平台的叙述与实践中，融合了网络电子资源容易更新的特点，使新技术保持时效性。

5. 本套丛书配有丰富的多媒体教学资源，将扩展知识、习题解析思路等内容做成二维码放在书中，丰富了教材内容，增强了教学互动，增加了学生的学习积极性与主动性。

本套丛书吸纳了数据科学与大数据技术教育工作者多年的教学与科研成果，凝聚了作者们的辛勤劳动，同时也得到了中南大学等院校领导和专家的大力支持。我相信本套教材的出版，对我国数据科学与大数据技术专业本科、专科教学质量的提高将有很好的促进作用。

桂卫华

2018 年 11 月

前言

Foreword

随着计算机、网络和通信等信息技术的发展，数据采集的方法越来越丰富，人类收集、存储和访问数据的能力大大增强。存储设备的容量不断提升而成本逐年下降，特别是数据库技术在各行各业的普及应用，使人类积累了海量的数据。快速增长的海量数据集已经远远超出了人类的理解能力，人类步入了大数据时代却陷入了“数据丰富、知识贫乏”的困境。人们迫切希望从所拥有的数据中获取有用的知识，以帮助其更好地进行决策。针对这一问题，数据仓库和数据挖掘技术应运而生，并且显示出强大的生命力。要将海量数据转换为有用的信息和知识，首先要有效地收集和组织数据。数据仓库是良好的数据收集和组织工具，它的任务是搜集来自各个业务系统的有用数据，存放在一个集成的储存区内。在数据仓库丰富完整的数据基础上，数据挖掘技术可以从中挖掘出有价值的知识，从而帮助决策者做出正确决策。“数据仓库与数据挖掘”已成为普通高等院校计算机、经贸管理和信息类相关专业研究生和高年级本科生的学位课程或选修科目。

按照教育部关于高等学校本科教育以培养更多应用型人才为目标的教学改革方向，以及全日制研究生以学术型和专业型两大类进行有差别培养的要求，我们迫切需要一本在教学课时限制严格的条件下，理论叙述深入浅出、实际应用具体完整、算法描述自然易懂、计算实例详略得当的数据仓库与数据挖掘方面的教材。

本书正是在这种社会需求背景和实际教学需要的情况下，在笔者总结多年教学改革与实践经验所编写的讲义基础上修改而成的。本书兼顾了应用型人才与学术型人才培养的需求，介绍了数据仓库原理、数据仓库设计和实现方法，为读者真正架起了理论与实践的桥梁。本书还以大量的计算实例来增加读者对数据挖掘原理及各种挖掘算法的理解深度。

本书主要介绍数据仓库和数据挖掘技术的基本原理和应用方法。全书共分为 13 章，主要内容包括数据仓库的概念与体系结构、数据、数据存储、OLAP 与数据立方体、数据挖掘基础、关联挖掘、聚类分析、分类、神经网络、统计分析、非结构化数据挖掘、知识图谱、大数据挖掘算法。其中，第 1 章为数据仓库的概念与体系结构，内容包括数据挖掘的兴起、数据仓库的基本概念及数据仓库的特点与组成等。第 2 ~ 4 章主要介绍数据仓库的基本原理和数据仓库系统的组建方法。第 5 ~ 12 章介绍当前流行的数据挖掘算法的主要思想和理论基础，并且给出丰富的应用实例。第 13 章为数据挖掘创新篇，内容主要为基于 MapReduce 的数据

挖掘算法及 Hadoop 的介绍等。

本书紧跟数据仓库和数据挖掘技术的发展和人才培养的目标，有以下几个特点：

(1)可读性强，文字叙述深入浅出，易读易用。

(2)概念清晰，条理清楚，内容取舍合理。

(3)本书强调基础，重视实例，各章节都以经典算法为主，介绍其主要思想和基本原理，并且给出恰当和丰富的实例。

(4)书中实例和课后习题实用、丰富，通过练习，读者可以对各个知识点从不同角度得到训练，掌握和巩固所学知识。

(5)本书教学资源丰富，提供了多媒体教学课件及实验平台，方便教学。

本书各章节之间衔接自然，同时各章节又有一定的独立性，读者可按本书的自然顺序学习，也可以根据实际情况挑选需要的章节学习。

限于作者水平，加之数据仓库与数据挖掘理论技术的内容十分丰富，且发展非常迅速，疏漏和不当之处在所难免，殷切希望广大师生和读者批评指正。

编 者

2018 年 10 月

目 录

Contents

第1章　数据仓库的概念与体系结构

任何企业都希望在市场竞争中通过利用全面的数据分析能力来获取更大更持久的竞争优势。例如，银行希望知道如何有效规避信贷风险，发现欺诈和洗钱等不合法行为；电信公司希望知道如何对市场业务发展和竞争环境进行精准分析，从而为市场决策提供深入有力的分析支撑，提升营销活动的精准性；保险公司希望知道哪些理赔客户骗保的可能性更高，以及哪些客户是高价值低风险的客户群；等等。以上这些问题的解决都离不开数据的支撑和对现有数据的分析利用。

传统的信息资源管理主要依靠数据库技术，利用数据库技术进行数据的组织与存储，并使用基于数据库的信息系统进行信息资源的有效利用。但是随着计算机技术的飞速发展，一些新的需求不断提出，这些新的需求是传统数据库技术难以满足的。

传统的数据库技术以数据库为中心进行事务处理、批处理到决策分析等各种类型的数据处理工作。数据库系统作为数据管理手段，从它诞生开始，就主要用于事务处理。近年来，随着计算机应用的拓展，人们对计算机数据处理的能力提出了更高的要求，希望计算机能够更多地参与到数据分析与决策支持中。但是事务处理和分析处理有着不同的性质，直接使用事务处理环境来支持分析决策有着一定的局限性。因此，需要一种新的数据管理技术——数据仓库，来实现分析决策的目的。

本章主要介绍数据仓库的兴起(1.1 节)、数据仓库的基本概念(1.2 节)、数据仓库的特点与组成(1.3 节)以及数据仓库的体系结构(1.4 节)。

1.1　数据仓库的兴起

任何一项新技术的出现都有其深厚的背景，数据仓库技术作为一种新型的数据管理技术，它的出现也并非偶然，而是随着人们对数据处理技术新需求的不断提出，最终衍生出了数据仓库。本小节主要介绍数据仓库的兴起，并从数据管理技术的发展(1.1.1 节)、数据仓库的萌芽(1.1.2 节)来展开讲述。

1.1.1　数据管理技术的发展

如图 1 -1 所示，从 1946 年计算机的诞生到 20 世纪 80 年代，数据管理技术经历了人工管理、文件系统以及数据库系统三个阶段，三个阶段之间彼此联系。由于人工管理的不足导致了文件系统的出现，同样由于文件系统的不足导致了数据库系统的出现。

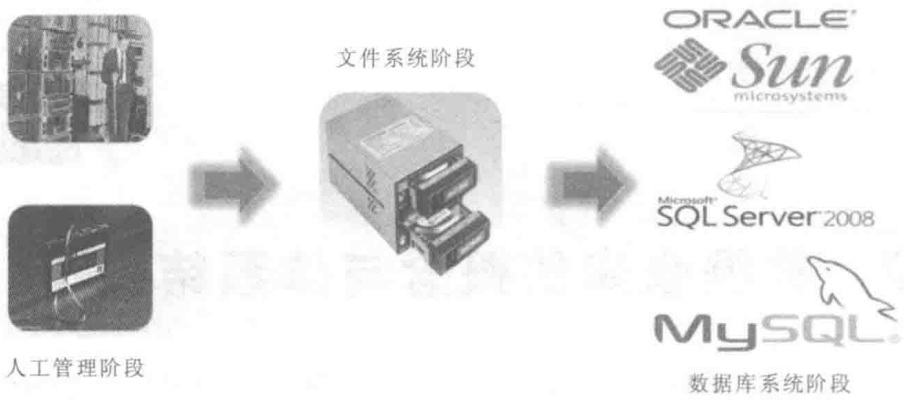


图 1-1　数据管理技术发展的三个阶段

1. 人工管理阶段

在 20 世纪 60 年代初期，创建运行于主文件上的单个应用是计算领域的主要工作。这些应用的特点表现在报表和程序上，常用的语言是 COBOL。穿孔卡是当时常用的介质，主文件存放在磁带文件上。磁带由于其廉价的特性适合于存放大容量数据，但其缺点是需要顺序地访问。访问整条磁带的文件可能要花去 20～30 分钟时间，这取决于我们所需数据在磁带上的存放位置。事实上，在最糟糕的情况下，一次磁带文件的操作需要访问 100% 的记录，然而只有 1% 的记录是真正需要的。

大约在 20 世纪 60 年代中期，主文件和磁带的使用数量迅速增长，很快，处处都是主文件。随着主文件数量的增长，数据出现了大量的冗余，并由此引出了一些严重的问题：

(1)数据更新需要保持数据的一致性。

(2)程序维护的复杂性。

(3)开发新程序的复杂性。

(4)支持所有主文件所需要的硬件数量。

简言之，由于介质本身存在固有缺陷，使得主文件成为了发展的障碍。如果仍然只用磁带作为数据存储的唯一介质，很难想象现在的数据管理会是什么样子。如果除了磁带文件以外没有别的介质可以存储大量数据，那么现实生活中将永远不会有大型、快速的系统。

2. 文件系统阶段

20 世纪 60 年代中期，计算机硬件有了磁盘直接存储设备，软件也有了操作系统，于是人们将数据文件长期存储到磁盘直接存储设备上，并利用操作系统所提供的文件系统对数据进行快速的访问。磁盘存储从根本上不同于磁带存储，磁盘存储不需要经过第 1 条记录，第 2 条记录，……，第 n 条记录，才能得到第 $n+1$ 条记录。只需要知道第 $n+1$ 条记录的地址，就可以轻而易举地直接访问它。磁盘直接存储设备以及操作系统的出现，解决了磁带存储设备数据访问效率低下的问题。但是它并没有解决主文件技术所带来的问题，只是更改存储介质提升了数据访问的速度，亟需一种新的数据存储与访问技术来解决主文件技术所面临的问题。