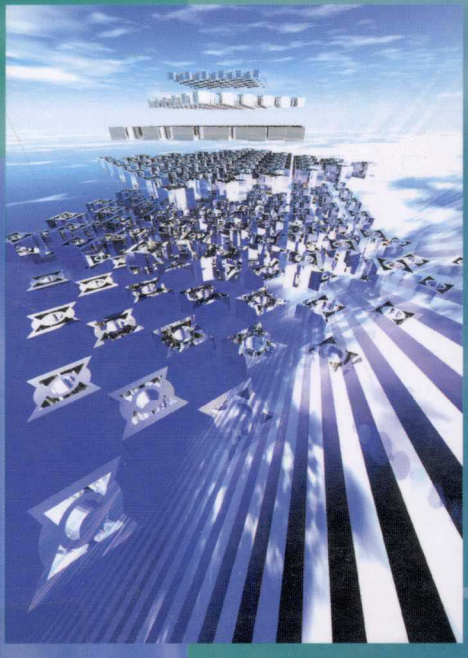
普通高等教育“十一五”国家级规划教材
信息管理与信息系统专业规划教材

数据仓库与数据挖掘技术

（第二版）



夏火松　编著

科学出版社
www.sciencep.com

普通高等教育“十一五”国家级规划教材

信息管理与信息系统专业规划教材

数据仓库与数据挖掘技术

（第二版）

夏火松　编著

科学出版社

北京

内 容 简 介

本书详细阐述了数据仓库与数据挖掘的基本原理，系统而全面地介绍了数据仓库与数据挖掘的概念、作用、算法和应用举例，并且给出了信息分析所涉及到的若干问题及框架。本书介绍了最新的信息分析技术研究成果，如小波分析、Rough 分析、蚁群分析、分形技术、Agent、数据挖掘的进化算法、聚类分析、非结构数据的挖掘、离群数据挖掘，但并未详细描述，而将介绍重点放在其应用上，起到抛砖引玉的作用。

本书既可以作为信息管理与信息系统、计算机应用、经济管理等专业的高年级本科生和研究生的教材，又可以作为有关在经济管理领域中应用信息分析技术提高决策人员的参考。

图书在版编目（CIP）数据

数据仓库与数据挖掘技术/夏火松编著. —2 版. —北京：科学出版社，2009

（普通高等教育“十一五”国家级规划教材·信息管理与信息系统专业规划教材）

ISBN 978-7-03-012934-5

Ⅰ. 数… Ⅱ. 夏… 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. TP. 311. 13

中国版本图书馆 CIP 数据核字（2004）第 011010 号

责任编辑：陈晓萍/责任校对：耿 耘

责任印制：吕春珉/封面设计：耕者设计工作室

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新 蕾 印 刷 厂 印刷

科学出版社发行 各地新华书店经销

*

2004 年 3 月第 一 版 开本：B5（720×1000）

2009 年 2 月第 二 版 印张：19 1/2

2009 年 2 月第四次印刷 字数：393 000

印数：8 001—11 000

定价：30.00 元

（如有印装质量问题，我社负责调换<新蕾>）

销售部电话 010-62134988 编辑部电话 010-62135763-8003

序

国家教育部于1998年7月6日公布了新的《普通高等学校本科专业目录》，将原来的经济信息管理、信息学、科技信息管理、林业信息管理和管理信息系统等专业合并为管理学科门类中的信息管理与信息系统专业。可以认为，这次合并既是学科相融的必然，也是国家信息化发展的需要。据有关资料介绍，到目前为止，全国已有超过200所高校开设了信息管理与信息系统专业。

自20世纪40年代以来，信息技术经过60余年的高速发展，它对人类社会各个领域的影响越来越广泛和深入，其影响最大、受益最多的当属管理和经济领域。信息作为最主要的经济资源，已经被人们所接受，并且愈来愈受到重视。信息技术的普及和推广，信息资源的组织、开发和利用，促进了企业的发展和产业结构的调整。当前所实施的电子商务、电子政务和数字图书馆等工程直接加速了生产力的发展和促进了社会的进步。我国政府提出的“以信息化带动工业化”的战略举措，必将有力提升我国的综合国力，同时也为信息管理与信息系统专业带来极大的发展机遇和发展空间。

信息管理与信息系统是一门交叉学科，它不是信息技术和管理科学的简单组合，而需要融合管理学、经济学、系统科学、运筹学和计算机科学于一体，因此，必须要有一套具有本专业特点的知识结构体系和适合本专业需要的教材体系。

信息管理与信息系统专业从1998年设立至今的10年来，许多专家学者在专业建设和教材建设方面倾注了大量的心血，有力地促进了专业和学科的发展。但是，由于该专业具有跨度大、内容新和变化快等特点，如何培养适应现代信息技术高速发展需要的、具有创新能力的、既懂信息技术又懂管理的复合型人才，对广大教育工作者而言是一个巨大的挑战。

在科学出版社的直接推动下，在我国信息管理领域的知名学者薛华成教授、侯炳辉教授和马费成教授的指导下，在湖北省信息产业厅和经济贸易委员会及相关企业的支持下，武汉地区包括华中科技大学、武汉大学、华中师范大学、中南财经政法大学和武汉理工大学等20余所高校联合编写了这套针对本科生的信息管理与信息系统专业规划教材。

这套教材共22本，除了数学基础类的《运筹学》外，大致可以归为以下3类。

计算机技术类（8本）：《数据库技术》、《计算机网络技术》、《数据结构与算

法（C语言实现）》、《面向对象的开发方法》、《数据仓库与数据挖掘技术》、《操作系统》、《多媒体信息管理技术基础》和《实用软件工具》。

信息系统类（6本）：《信息系统分析与设计》、《信息系统案例分析》、《项目管理》、《管理信息系统》、《信息系统原理》和《决策支持系统》。

信息管理类（7本）：《信息管理学基础》、《信息资源管理》、《信息经济学》、《信息政策与法规》、《信息组织学》、《信息检索》和《信息安全》。

这套教材具有以下特点：

（1）内容新。正如前面所指出的一样，这套教材并不是简单地分门别类讲解信息技术和管理科学知识，而是站在信息管理与信息系统专业这个全新的角度上，力求全面、及时地反映国内外信息管理与信息系统领域的最新发展和研究成果。

（2）体系全。为保证本系列教材体系的完整性和内容的系统性，编委会曾多次开会讨论并广泛征求国内信息管理与信息系统领域的有关专家的意见，该套教材主要集中于专业基础课和专业课方面，并考虑了这些课程之间的相互衔接和整体上的协调。

（3）注重基础。本系列教材从选题到编写均充分考虑到当前我国本科生的知识结构和知识背景及其后续发展的需要，着重于讲解信息管理与信息系统专业的基础知识，注意培养学生的能力。

（4）结合实际，多采用案例教学。本系列教材的作者都是从事一线教学工作的教师，了解本科生的特点和需求，大多数作者又有从事信息系统开发和信息资源管理的经验，了解实际工作对本专业的需求。因此，在编写过程中作者们能注意理论与实践相结合，通过引入适当的案例和实验，加深学生对理论知识的理解和掌握。

我们希望，这套教材的成功出版，能为推动我国信息管理与信息系统专业教育工作的发展、促进信息化人才的培养起到积极的作用。

这套教材是我们不同类型的学校，不同专业背景、但同属信息管理与信息系统专业教师合作的一种尝试。我们欢迎信息管理和信息系统及相关专业的教师、学生和科研工作者以及有关人士提出宝贵的意见和建议，以便进一步提高我们的教材质量。

本套规划教材编委会主任
华中科技大学管理学院院长
管理信息研究所所长
张金隆　教授

第二版前言

本书自2004年3月第一版出版以来，得到了广大读者的支持和认可，不仅被很多学校选为本科生教材或研究生教材，而且被不少读者在撰写学术论文时大量引用，并于2008年申报到普通高等教育“十一五”国家级规划教材，在此表示衷心的感谢。

值此本书重修出版之际，我们对部分章节进行了修改。全书主要修改和增加部分如下：第2章“数据仓库的分析”增加了2.1节；第3章“数据仓库的设计与实施”增加了3.1节和3.3.3小节；第7章“非结构化数据挖掘”增加了7.1节。

本书第二版由夏火松博士任主编并负责全书总纂与定稿工作。在本书的编写过程中，我们参考了国内外不少的文献资料，在此向这些文献的作者表示衷心的感谢。同时，我还要感谢美国亚利桑那大学提供的学习和研究的经历。

由于编者的水平有限，书中难免存在缺点和错误，敬请各位专家学者和读者批评指正。

夏火松

2008年12月

第一版前言

随着信息技术的深入应用，社会迫切需要更多的人能够掌握从数据中获取有用知识的理论与方法，从而更好地进行有效的决策。本书是系统阐述数据仓库和数据挖掘的理论、方法与实践的专业书籍，其内容融合了先进的数据库技术、Web 技术、数理统计技术、人工智能技术、现代的管理思想和系统的科学方法。本书的写作目的并非是要深入到每一种算法是如何编程的具体细节当中，而是以企业中正在从事或将要从事营销管理、经营决策和管理信息系统的深入开发等方面的工作者和 IT 人员作为对象，为其提供较为详细的信息分析技术、方法与总体思路。

本书在组织材料上，力求做到系统性、准确性、完整性、先进性、实用性，把培养读者对信息进行管理和利用的能力作为出发点。本书所涉及的知识点既可促进管理创新，又可使信息技术在管理中得到更广泛深入的应用。要求读者在阅读本书前，应具备数理统计、数据结构、数据库技术和至少一门程序设计语言等方面的知识，还应具有一定的经营管理方面的知识。书中有部分章节难度较大，读者根据实际情况可跳过。

全书共 10 章。主要从功能方面介绍数据仓库和数据挖掘的定义，分析使用数据仓库和数据挖掘的原因，从数据仓库的生命周期入手，分析如何建立数据仓库的逻辑结构和进行数据仓库的设计；对信息分析的基本技术进行了描述，分析了数据挖掘的方法和成功使用数据挖掘的过程；较为详细地介绍了数据挖掘的基本算法、非结构数据挖掘和离群数据挖掘；对数据挖掘的语言、数据挖掘的工具选择和研究的前景进行了介绍。最后对知识、知识管理和知识管理系统也做了分析。本书将应用重点放在经济管理领域中的信息分析和知识的获取上，对算法的叙述力求简洁明了，既重点叙述几个操作性很强的算法，又全面给出了其他算法的核心思想与原理，为了更清楚地深入有关算法，书中提供了详细的资源。这就体现出本书的三大特色：第一解决了数据仓库与数据挖掘的基本原理与在经济管理领域中的应用脱节，使读者可以先抛开繁琐的算法而树立一种用信息分析的方法进行思考问题的方式，为科学的决策提供知识获取的方案；第二解决了广泛而复杂的算法与提供较大的扩展空间的矛盾，在提供详细的经典算法中，如决策树的 ID3 算法和关联规则的 Apriori 算法的同时，对个别算法只提供基本的思想，而更深入的放在提供详细的资源和思路上，起到“指路”的作用，读者根据自己的基础可以跳过有关章节阅读，但并不影响对本书基本知识的掌握；第三解决了

最新研究成果与基本知识的掌握间的矛盾。

本书可作为高等院校信息管理与信息系统专业、计算机应用专业的教材，也可作为从事信息系统建设和计算机应用工作的技术人员、管理人员的参考书，还可作为研究生的教学参考资料。

夏火松博士主编本书并负责全书总纂与定稿工作，陈冈、胡新明、蔡辉、罗建军、陈智洁老师和何茂丹同学分别对本书进行了校对并提出了宝贵的建议，最后由华中科技大学博士生导师蔡淑琴教授对本书进行了审定。

在本书的编写过程中，我们参考了国内外不少的文献资料，得到了华中科技大学管理学院院长、博士生导师张金隆教授的帮助，得到了科学出版社的大力支持和帮助，在此表示衷心的感谢。对于一些并未一一提到的朋友给予的支持，我们在此表示深深的谢意。

由于本书的写作时间较短，加之编者的水平有限，书中难免存在缺点和错误，敬请各位专家学者和读者批评指正。

夏火松

2004 年 1 月

目　录

第 1 章　数据仓库与数据挖掘概述

1.1　数据仓库引论

1.1.1　为什么要建立数据仓库

随着信息处理技术的不断发展，信息的存储、管理、使用和维护显得越来越重要，而传统的数据库管理系统（database management system，DBMS）很难满足其要求，表现为：数据量成几何级数增长；不同部分的数据难以集成；访问这些数据的响应性能不断降低。而决策支持系统（decision support system，DSS）所需数据必须预先经过提取、转换、过滤并与其他数据源整合，按主题存放在中央数据库中。客户查询时只访问中央数据库（database，DB），而不访问其他数据库。要想使数据能够发挥其最佳效用，更好地为用户服务，数据也必须经过严格的准备、组织和显示等几个步骤。完成这些工作的场所通常被称为数据仓库（data warehouse，DW）。数据仓库早在 20 世纪 90 年代起就开始流行。由于它为最终用户处理所需要的决策信息提供了一种有效方法，因此数据仓库被广泛应用，并且得到很好的发展。

1. 数据仓库的作用

首先我们看看哪些人需要数据仓库：

1）从大量的数据中得出结论并以大量的数据为依据来做出决策的人。

2）以定制方法实现有用的信息与知识获取的人，而这类人不必（或者不能够）为了这个目的而进行数据的寻找与组织操作。

3）希望以简单的信息技术就能访问数据库的人。

4）科学的决策对企业是非常有价值的，基于数据仓库能为企业做出更好科学决策的人。

在商界中有许多具备这些特征的人。这造就了数以百万计的潜在的数据仓库用户，也对与之相应的各种类型的商务数据仓库造成了巨大的影响。

数据仓库具有两个主要作用：一是从各信息源提取决策需要的数据，加工处理后，存储到数据仓库中；二是提供用户的查询和决策分析的依据。

决策所需要的信息来自不同地点的数据库或其他信息源，而这些信息源可能具有分布式和异构式的特点。数据仓库是存储数据的一种组织形式，但数据仓库中的存储并不是简单的存储，而是对来源于不同系统异构的数据进行加工和集成处理后的再存储。例如数据仓库管理系统预先把实体化视图对应的数据从内部、

外部数据库中提取、加工、综合，然后物理地存储到数据仓库中，使这些视图成为物理存储的数据实体。

数据仓库作为信息分析的基础，为数据驱动型的决策支持提供了数据基础。例如对客户进行分析，通过客户数据仓库，可以分析客户的购买习惯，预报客户的流失的概率，分析客户的购买趋势、季节性购买模式、广告的成功率和其他战略性的信息。沃尔玛(Wal-Mart)的数据仓库始建于20世纪80年代，1988年数据仓库容量为12千兆字节(GB)，1989年为24GB，1996年已达7.5兆兆字节(TB)，1997年达到24TB，所以沃尔玛的成功较大程度上取决于利用数据仓库对商品购物篮进行分析(marketing basket analysis)，找到了不同商品购买的相关性。由此可见数据仓库的出现能使企业能够准确分析和把握消费心理。

数据仓库对数据库中的数据进行再加工，形成一个综合的、面向分析的环境，以更好地支持决策。

2. 建立数据仓库的好处

建立数据仓库不仅在商业上产生巨大的影响，而且在技术上能够得以实现。由于用户大多是业务领域的专家，而不是计算机专业人员，利用所有可能的数据快速而正确地做出决策需要数据仓库。同时企业数据每18个月翻一番，需要有一种有效的访问这些数据的方法，这也需要数据仓库。随着竞争的加剧，数据仓库可以进行商务智能(business intelligence，BI)分析。

在技术上，由于存储介质价格的下跌，计算机的计算能力越来越强，网络带宽的增长，网络的传输能力越来越好，整个企业的计算机环境越来越复杂；各个时代各个不同厂家的应用系统同时存在，新的应用系统需要访问其他应用系统的数据。这些问题使数据仓库可以应用于数据定位、数据呈现(报表和图表)、检验假设、知识发现和共享分析。

总之，建立数据仓库可以产生如下的好处：

有形的好处是改善产品库存控制；降低产品推广费；更加高效地制定决策；能提供一个关于整个企业的整体构架。

无形的好处是通过把所有的数据放在一个地方，方便存取，提高生产效率；减少重复数据处理和分析；数据仓库提高用户对数据的应用程度；为商务流程再造提供支持。

1.1.2 什么是数据仓库

1. 数据仓库的概念

我们知道几百年前仓库的概念一般指的是存储货物和商品的地方或设施，当人们检查货物确定它们的用途以后，就将货物运进常规的仓库中去。货物只要放

进仓库，就被整齐地排放在一起，以备在需要时取用。人们很可能为了方便而有效地存取，将可能的要求放在一起的货物分到一组。仓库系统具有找出货物在某处的功能，也具有核查在该处有哪些具有潜在价值货物的功能。另外，仓库系统还具有必要的关联功能，以便告知用户它所包含的内容以及运用管理方式所控制的有价值的资产。数据仓库中的特征对应于真实的仓库的每一个特征。但什么是数据仓库，目前有以下不同的看法：

定义 1　W. H. Inmon 在《Building the Data Warehouse》中定义数据仓库为："数据仓库是面向主题的、集成的、随时间变化的、历史的、稳定的、支持决策制定过程的数据集合。"即数据仓库是在管理人员决策中的面向主题的、集成的、非易失的并且随时间变化而变化的数据集合。下面是这个定义中一些术语的基本含义：

"面向主题"是指数据是由业务主题组织的，例如，事务数据组织的就不是按主题组织的。

"集成"是指数据是作为一个整体进行存储的，而不是以可能有不同结构或组织方式的文件集合存储的。

"非易失"是指数据保持不变，即按计划添加新数据，而原数据不会丢弃。

"随时间而变化"是指时间量度明确地包含在数据中，使得数据随时间的趋向和变化可以用于分析研究。这并不意味着数据元素随时间改变，正如用户核对账目余额数据在用户银行操作数据库中不会随意改变那样，许多数据仓库也让数据含有地理空间这一维。

所以从逻辑上讲，数据仓库又是一个多维数据库，它为信息分析提供了良好的基础。

定义 2　数据仓库是作为 DSS 基础的分析型 DB，用来存放大容量的只读数据，为制定决策提供所需的信息。

定义 3　数据仓库是与操作型系统相分离的、基于标准企业模型集成的、带有时间属性的。即与企业定义的时间区段相关，面向主题且不可更新的数据集合。

定义 4　数据仓库是一种来源于各种渠道的单一的、完整的、稳定的数据存储。这种数据存储是一种能够提供允许最终用户在其业务范畴中理解并使用的方式。

定义 5　数据仓库是大量有关公司数据的数据存储。

定义 6　数据仓库提供公司数据以及组织数据的访问功能，其中的数据是一致的，并且可以按每种可能的商业度量方式分解和组合；数据仓库也是一套查询、分析和呈现信息的工具；数据仓库是我们发布所用数据的场所，其中数据的质量是业务再工程的驱动器(driver of business reengineering)。

这些定义的共同特征：首先，数据仓库包含大量数据，其中一些数据来源于组织中的操作数据，也有一些数据可能来自于组织外部；其次，组织数据仓库是为了