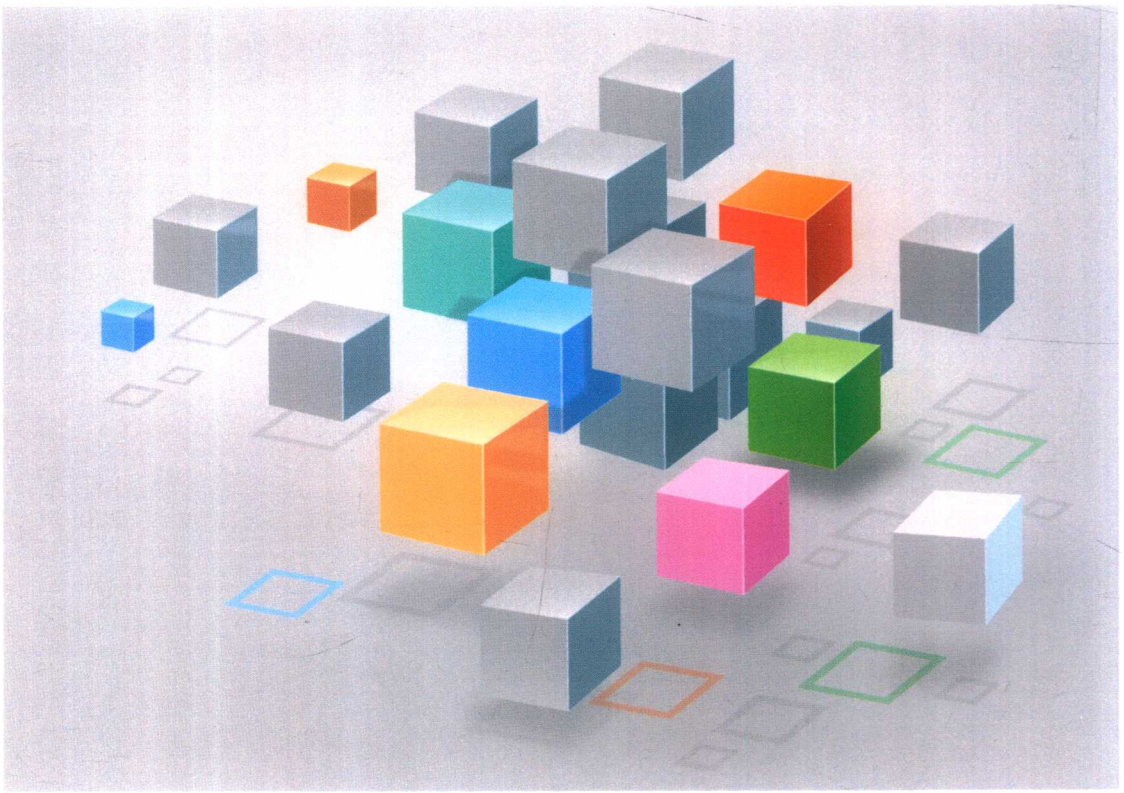

数据仓库与数据挖掘

概念、方法及图书馆应用

SHUJU CANGKU YU SHUJU WAJUE
GAINIAN FANGFA JI TUSHUGUAN YINGYONG

朱东妹◎著

安徽师范大学出版社

安徽师范大学出版基金项目资助出版

数据仓库与数据挖掘
概念、方法及图书馆应用

朱东妹◎著

安徽师范大学出版社
·芜湖·

图书在版编目(CIP)数据

数据仓库与数据挖掘概念、方法及图书馆应用/朱东妹著.—芜湖:安徽师范大学出版社,2017.8

ISBN 978-7-5676-3022-2

Ⅰ.①数… Ⅱ.①朱… Ⅲ.①数据库系统－应用－图书馆工作－研究Ⅳ.①G25-39

中国版本图书馆CIP数据核字（2017）第161571号

数据仓库与数据挖掘概念、方法及图书馆应用

朱东妹 著

责任编辑:胡志立

封面设计:周 敏

出版发行:安徽师范大学出版社

芜湖市九华南路189号安徽师范大学花津校区 邮政编码:241000

网 址:http://www.ahnupress.com/

发 行 部:0553-3883578 5910327 5910310(传真) E-mail:asdcbsfxb@126.com

印 刷:虎彩印艺股份有限公司

版 次:2017年8月第1版

印 次:2017年8月第1次印刷

规 格:700 mm × 1000 mm 1/16

印 张:15.75

字 数:275千字

书 号:ISBN 978-7-5676-3022-2

定 价:42.50元

凡安徽师范大学出版社版图书有缺漏页、残破等质量问题,本社负责调换。

前　言

随着信息技术不断的发展及在图书馆的广泛应用，图书馆的工作内容、服务模式、技术服务手段和管理机制等都发生了巨大的变化，同时图书馆在新技术环境下开始积累了大量的数据。如何高效运用这些数据并从中挖掘出有意义的信息和知识，进而为图书馆管理人员提供决策依据是本书的目标。

基于此，本书注重理论与实践相结合，力求突出以下特征：

（1）采用浅显易懂的语言表达相关的概念与方法。

（2）理论与实际相结合，使概念和方法具体化、实用化。除了第1章，其余各章最后一节都是案例，意在学以致用、学用结合。

（3）可操作性强。本书在介绍相关技术时，以ILASIII图书馆自动化集成管理系统中的数据，在Microsoft SQL Server 2012数据仓库开发及数据挖掘操作环境下，作了丰富的操作讲解和图示，读者可以把这些方法应用到自己需要处理的问题中。

本书主要分为三部分内容。

第一部分为概述，是第1章的内容，简要介绍了数据仓库和数据挖掘的基本概念和发展等相关知识。

第二部分为数据仓库、ETL数据抽取转换加载和OLAP联机分析，包含第2、3、4章的内容，主要是对数据仓库的建立、数据抽取转换加载以及联机分析处理技术的基本方法和实例的具体实现。

第三部分为数据挖掘，包含第5、6、7、8、9章的内容，主要是对数据挖掘中的关联规则、分类、聚类、线性回归、时序等方法的相关知识和实例的具体实现。

本书的亮点在于，除了第1章，其余各章的最后一节都是本章理论方法

在图书馆应用中的一个具体实现，便于读者深入掌握。

尽管作者为本书付出了努力，但是由于水平有限，加上时间仓促，书中不妥之处，期待您的批评和建议。如有任何意见或建议，请发邮件到zdm5180@163.com，谢谢！

作　者

2017年5月于芜湖

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第1章　概述

随着数据库技术的发展，信息系统的用户除了需要计算机为其处理日常事务外，更需要从大量实际存在的数据中归纳出业务的规律性及其发展趋势去帮助管理决策，传统的数据库的处理方式不能满足决策分析的需求，数据仓库在这样的背景下应运而生。针对数据的复杂化与海量化，如何将这些海量的数据从数据仓库中提取出来，并转为有用的信息，需要更灵活、效率更高及理论更完善的方法和工具。多年来，数理统计方法、人工智能及知识工程等领域的丰硕成果，为开发对数据进行深度分析的工具提供了坚实的理论与技术基础，数据挖掘理论与技术应运而生。

1.1　初识数据仓库

1.1.1　数据仓库的产生过程

数据仓库（Data Warehouse）是一种信息管理技术，是一种新的数据处理体系结构，它为企业决策支持系统提供所需信息。数据量越大，数据仓库的作用就越大。数据仓库的产生主要经历了以下几个阶段[①②]：

数据仓库概念最早可追溯到20世纪70年代。麻省理工学院的研究员致力于研究一种优化的技术架构，该架构试图将业务处理系统和分析系统分开，即将业务处理和分析处理分为不同层次，针对各自的特点采取不同的架构设计原则。麻省理工学院的研究员认为这两种信息处理的方式具有显著差别，

① 潘华，项同德.数据仓库与数据挖掘原理、工具及应用[M].北京：中国电力出版社，2007：9-12.
② 钱星常.远程教与学策略和案例[M].北京：科学出版社，2008：179.

此为试读，需要完整PDF请访问：www.ertongbook.com

以至于必须采取完全不同的架构和设计方法，但受限于当时的信息处理能力，这个研究仅仅停留在理论层面。

20世纪80年代中后期，美国数字设备公司已经开始采用分布式网络架构来支持其业务应用，并且首先将业务系统移植到其自身的RDBMS产品RdBA上。同时，结合麻省理工学院的研究结论，建立了TA2（Technical Architecture2）规范。该规范定义了分析系统的四个组成部分：数据获取、数据访问、目录服务和用户服务。这是系统架构的一次重大转变，第一次明确提出了分析系统架构并将其运用于实践。1988年，为解决全企业集成问题，IBM公司第一次提出了信息仓库（Information Warehouse）的概念，并称之为VITAL规范（Virtually Integrated Technical Architecture Lifecycle）。VITAL定义了85种信息仓库组件，包括PC、图形化界面、面向对象的组件以及局域网等。至此，数据仓库的基本原理、技术架构以及分析系统的主要原则都已确定，数据仓库初具雏形。

1991年Willian Inmon出版了关于数据仓库的第一本书*Building the Data Warehouse*，标志着数据仓库概念的确立。这本书不仅说明为什么要建数据仓库、数据仓库能带来什么，还详细定义了数据仓库的具体原则：数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化的（Time Variant）数据集合，用于支持管理决策（Decision-Making Support）。这些原则至今仍然是指导数据仓库建设的最基本原则，因此Willian Inmon被称为数据仓库之父。

1994—1996年，由于企业级数据仓库的设计、实施及其坚持第3范式设计要求，从而无法支持决策支持系统对性能和数据易访问性的要求，因此数据仓库的建设者和分析师开始考虑只建设企业级数据仓库的一部分，提出了“数据集市”的概念，主要采用多维数据模型，在传统的关系型数据模型和多维联机分析之间建立了很好的桥梁。此后，建立企业级数据仓库还是部门级数据集市？关系型还是多维？这个问题在很长时间一直被争论着，相应地出现了“自底向上”和“自顶向下”两种实施方法的争议。“自顶向下”是从企业的整体来考虑数据库的主题和实施，是一种系统的解决方法，并能够最大限度地减少集成问题。然而，它费用高，需要长时间开发，而且缺乏灵活性，因为整个组织的共同数据模型达到一致是较困难的；而“自底向上”是从建造某个部门的数据集市开始，逐步扩充数据仓库所包含的主题和范围，

最后形成一个能够完全反映企业全貌的企业级数据库，花费低，并能够得到快速的投资回报，但存在的问题是将分散的数据集市集成，形成一个一致的企业数据仓库并不容易。事实证明，比较切合实际的做法是将“自底向上”和“自顶向下”两种实施方法结合起来构建数据仓库。

2006年，Willian Inmon总结了20年来数据仓库实践经验和存在的问题，提出了DW2.0的概念。DW2.0提出了数据仓库生命周期概念，把整个系统分为四个区：交互区、整合区、近线区、归档区；提出了集成非结构化数据，要求将非结构化文本数据存放在数据仓库中，并与结构化数据整合在一起；提出了要对数据仓库数据进行监视；更加重视元数据的作用，认为元数据是数据仓库基本结构中一个主要且极为重要的部分。

从目前形势看，数据仓库已成为继因特网之后，信息社会中获得企业竞争优势的关键。国外许多厂家和公司相继推出了各自的数据仓库解决方案，例如，IBM所推崇的商业智能，其核心就是数据仓库；微软的SQL server7.0就开始绑定了OLAP服务器，将数据仓库功能集成到数据库中，并建立了数据仓库联盟。

1.1.2 数据仓库的体系结构

数据仓库的体系结构由数据获取层、数据处理层、数据存储层以及数据分析层组成。数据仓库的体系结构如图1-1所示。

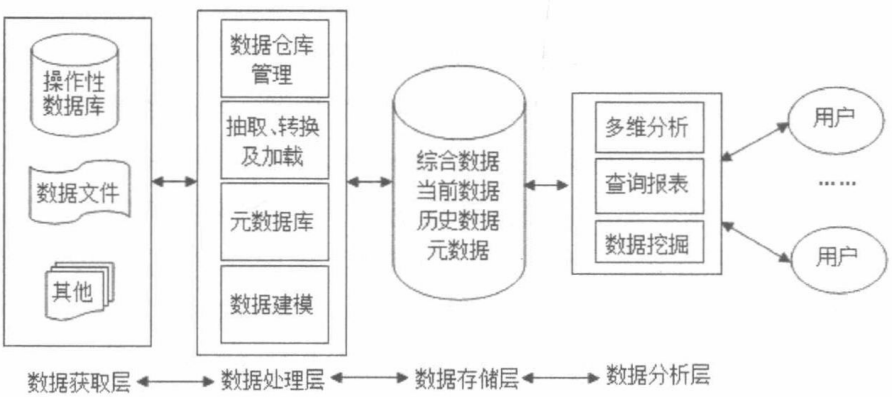

图1-1 数据仓库体系结构

（1）数据获取层：该层是数据仓库的数据获取来源。同一数据仓库可以有多种不同的数据源，包括业务操作数据和其他外部数据源。数据源是数据

仓库的基础。

（2）数据处理层：负责将数据源中对决策分析有用的数据进行清洗、转换和加载到数据仓库中，同时还负责监视数据源的数据变化，随时对新的或变化的源数据进行分析、转换并将其更新到数据仓库中。

（3）数据存储层：负责对数据仓库中的数据及元数据进行归档、备份及安全管理。数据仓库中保存的数据量相对传统的数据库来说要大得多，需要有效地进行组织。数据仓库的数据组织把从各数据源获取的数据以不同的粒度级别进行存放，分为历史数据、当前数据及综合数据。粒度越大，数据的综合程度越高；粒度越小，数据的细节程度越高。

（4）数据分析层：数据分析层面向终端用户，为其提供数据查询、协助其分析和评估决策等服务。

1.1.3 数据仓库的关键技术

1. 数据抽取、转换及加载

数据抽取、转换及加载（Extract-Transform-Load，ETL）是数据仓库体系结构中数据处理层的一项关键技术。用户从数据源抽取出所需的数据，经过数据清洗、转换，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去，成为联机分析处理、数据挖掘的基础。

例如，不同的图书馆在构建数据仓库过程中，涉及的数据可能有读者基本信息数据库、流通数据库、采购数据库、查询数据库、门禁系统数据库等本地数据，以及网络服务平台中形成的Web服务器数据、用户登记信息、代理服务器数据、读者访问电子资源信息等网络数据。这些数据源格式多种多样，有文本文件、电子表格文件、SQL Server及Oracle等数据库文件，必须通过抽取、清洗、转换，然后装载到数据仓库中。ETL策略的制定必须考虑到源系统、目标系统及业务规则等多方面因素。实现异构数据库抽取数据后，还需建立统一标准进行数据存储。比如我们在处理读者信息时，有的读者身份证数据位数是15位，有的是18位；日期型的数据在部分图书管理系统中是以字符型存储的，要把它转换成日期型；有的读者所在院系名称写的是全称，而有的写的是简称；对于读者属性的一些分类，在图书管理系统中一般用不同的代码表示而不是用文字，例如，用“001”代表读者流通类型为“教师”，“n”代表读者证状态为“有效”，“F”代表性别为“女”；有的读者部分

信息不全等这些问题都要在整个数据进入数据仓库前根据实际需要制定统一规则进行清洗、转换。另外，数据进入数据仓库之后，需制定数据定期更新及维护策略。

2. 联机分析处理

联机分析处理（On-Line Analysis Processing，OLAP）是数据仓库体系结构中数据分析层的一项关键技术。OLAP是在多维数据结构上进行数据分析的，支持决策人员从不同的角度，迅速、灵活地对数据仓库中的数据进行复杂查询和多维分析，并且以直观、容易理解的形式将查询和分析结果提供给各种决策人员。

例如，图书馆构建好图书流通主题的OLAP模型后，可以非常快速地从不同维度查询日常流通业务活动信息，如某周、某月或某年某个馆藏地点图书入库数，某周、某月、某年各类图书的馆藏量及其借还量，各类型读者借还图书情况等，并且可以通过OLAP追踪查询某类图书、某类读者借还量变化存在的原因。

1.2 初识数据挖掘

1.2.1 数据挖掘对象

数据挖掘就是从大量数据中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。简言之，数据挖掘就是一深层次的数据分析方法，是要在数据中发现知识。从应用领域的角度看，数据挖掘对象主要包括以下几大类型[①②]。

1. 关系数据库

关系数据库，是建立在关系数据库模型基础上的数据库，借助于集合代数等概念和方法来处理数据库中的数据，同时也是一个被组织成一组拥有正式描述性的表格。每个表格，也称为关系，包含用列表示的一个或更多的数据种类，每行包含一个唯一的数据实体，这些数据是被列定义的种类。关系

① 蒋盛益，李霞，郑琪.数据挖掘原理与实践[M].北京：电子工业出版社，2011：6-20.
② 苏新宁，杨建林.数据挖掘理论与技术[M].北京：科学技术文献出版社，2003：9-11.

数据库分为两类：一类是桌面数据库，例如Access、FoxPro和dBase等；另一类是客户/服务器数据库，例如SQL Server、Oracle和Sybase等。关系数据库是数据挖掘最常见的数据源。

2. 数据仓库

数据仓库是面向决策支持的，其目的是根据不同的主题集成多种异构数据源，建立一种高度一体化的数据存贮处理环境，包括详细和汇总性的数据、历史数据、整合性数据及解释数据的数据。针对不同主题联机分析处理技术提供了对数据仓库中的数据进行复杂显示和分析的方法。数据仓库为数据挖掘准备了良好的数据源。

3. 文本数据库

随着信息技术的不断进步，以电子文本为载体保存下来的信息越来越多，于是形成了文本数据库。文本数据库存储的内容均为文字，是长句、段落甚至全文。文本数据类型多数为非结构化的（如文章摘要和内容），也有些半结构化的（如XML数据、Email邮件、学术期刊数据库等）。部分文本数据如果结构良好，也可用关系型数据库来实现（如文档的标题、作者、出版单位及分类号等）。挖掘内容包括文本分类、文本聚类、文本特征提取等。

4. 空间数据库

空间数据库以描述空间位置和点、线、面、体特征的拓扑结构的位置数据为对象的数据库系统。对空间数据库的挖掘可以为城市规划、环境和资源管理、商业网络、森林保护、人口调查、交通及税收等领域的管理提供决策支持。

5. 时序数据库

时序数据库主要用于存放与时间相关的数据，它可用来反映随时间变化的即时数据或不同时间发生的不同事件。例如，图书馆中连续多年存放的即时的图书借阅信息、门禁记录的读者入馆信息以及电子资源下载等信息。对时序数据的挖掘，可以发现事物的演变过程、隐藏特征及发展趋势。

6. Web数据库

Web数据库指在互联网中以Web查询接口方式访问的数据库资源。Web可以描述为在互联网上运行的、全球的、交互的、动态的、跨平台的、分布式的、图形化的超文本信息系统。Web数据库中的数据类别有网页内的结

构、网页间的结构、网页的内容、用户的注册信息及用户访问网页规律等数据，挖掘内容包括Web内容挖掘、Web结构挖掘及Web使用挖掘等。由于Web本身具备超大量性、高度复杂性、动态性和用户群体的多样性等特点，Web挖掘具有挑战性。

1.2.2 数据挖掘过程

数据挖掘项目的成功实施有很多决定性因素，如问题如何界定，数据如何选取，生成的模型能否嵌入到现有业务流程中等。为了使数据挖掘过程标准化，数据挖掘软件提供商们提出了各自的数据挖掘过程的方法论。CRISP-DM（Cross Industry Standard Process for Data Mining，跨行业数据挖掘标准流程）是其中的优秀代表，现成为行业通用的模型标准。这个方法论可以在数据挖掘项目的整个生命周期为用户提供指导。CRISP-DM将数据挖掘过程分为以下6个阶段[①②]。

1. 理解问题

理解问题是数据挖掘的第一步。这个阶段是了解用户业务问题，明确用户真正需要达到的目的，然后将这些理解转化为数据挖掘的问题定义，并制定项目计划。项目计划应该细化、明确，便于监督。

2. 理解数据

理解数据阶段首先要理解所有与业务对象有关的内、外部数据，在此基础上收集、描述原始数据，产生数据收集、描述报告。对数据质量进行鉴定，产生数据质量报告。

3. 数据准备

数据准备阶段是数据挖掘工作的最关键阶段。现实业务数据往往被存储在不同的数据库或不同的部门中。这一步骤需将这些数据进行抽取、清理、重构、整合及格式化，生成可以建立数据挖掘模型的数据集。

4. 建立模型

建立模型是数据挖掘工作的核心阶段。在模型的建立过程中要选择适当的建模技术，一般一个类型的数据挖掘问题需要选择和应用几种不同的建模技术，需要将不同模型的参数调整到最佳值，在此过程中结合具体业务实践

① 崔雷.医学数据挖掘[M].北京:高等教育出版社,2006:24-25.
② 张文彤,钟云飞.IBM SPSS数据分析与挖掘实战案例精粹[M].北京:清华大学出版社,2013:6-10.

此为试读,需要完整PDF请访问: www.ertongbook.com