智能信息系统

——以关联知识优化数据建模的方法和实践

任　明◆著

Intelligent Information System

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

智能信息系统

——以关联知识优化数据建模的方法和实践

任　明　著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

智能信息系统：以关联知识优化数据建模的方法和实践 / 任明著. —杭州：浙江大学出版社，2012.3
ISBN 978-7-308-09977-6

Ⅰ.①智… Ⅱ.①任… Ⅲ.①智能系统—管理信息系统—应用—企业管理 Ⅳ.①F270.7

中国版本图书馆 CIP 数据核字（2012）第 097623 号

智能信息系统

——以关联知识优化数据建模的方法和实践

任 明 著

责任编辑 许佳颖
文字编辑 陈静毅
封面设计 黄晓意
出版发行 浙江大学出版社
（杭州市天目山路 148 号 邮政编码 310007）
（网址：http://www.zjupress.com）
排 版 杭州中大图文设计有限公司
印 刷 浙江省邮电印刷股份有限公司
开 本 880mm×1230mm 1/32
印 张 6.75
字 数 175 千
版 印 次 2012 年 3 月第 1 版 2012 年 3 月第 1 次印刷
书 号 ISBN 978-7-308-09977-6
定 价 29.00 元

前　言

1993年，三金工程开启了我国国民经济的信息化进程。时至今日，信息系统应用的数量和重要性的巨大增长是有目共睹的，包括商业、医院、教育、政府和图书馆在内的几乎每一类组织都开始使用信息系统存储、处理和检索数据。如今，一大批组织和企业建立了信息交流的基础平台，构建了比较完善的数据库系统，比如财务资产数据库、人力资源数据库、物资管理数据库，以及一些面向专业技术应用的数据库系统。随着信息技术和网络通信技术的进一步发展，以及大批信息化建设成果的示范性展示，人们对信息化建设的深远意义和现实作用的认识也逐渐深刻，信息化能够创造效益、增强竞争力的观念已得到广泛认同。国家“十二五”规划明确阐述了建设下一代国家信息基础设施，推动信息化和工业化深度融合，推进经济社会各领域信息化的目标。信息化已经成为国家经济社会发展的总体发展战略中的重要组成部分。

飞速发展的网络技术和智能技术不断孕育着信息系统的创新和突破。在信息系统集成和信息资源整合的基础上融合数据挖掘技术，构建智能化的信息系统，正在成为新的突破点。传统的信息系统及数据模型是人为构造的，然而现实世界的复杂性和不确定性使得对一个领域的理解并非易事，有关的数据和领域知识是通过逐渐的学习、探索得来的，信息系统数据模式的不断演化是非常必要的。数据挖掘能够发现数据中隐藏的、新颖的、有趣的、有用的知识，是对传统的数据建模所依赖的用户和系统分析人员的知识的重要补充，可以反映到系统中优化现有的数据模

型，从而不断演化出蕴含了独特领域知识的信息系统，使组织和企业具有核心的竞争优势。

本书构建了信息系统建模和数据挖掘的桥梁，这是随着信息技术的飞速发展而兴起的两个研究领域。研究聚焦到数据挖掘领域一种重要且常见的知识——关联规则，挖掘到的关联规则对于数据模型来讲是新颖的、有意义的补充。研究关注概念层次上的数据建模，并选取广泛用于关系数据库建模的实体关系（ER）模型作为基础，它有助于捕捉现实世界的结构和语义，通常是技术人员与用户之间的沟通桥梁。本书建立了关联规则和ER模型之间的术语和语义的对照，探讨了使用关联规则丰富ER模型的方法和技术，对特殊化建模机制进行扩展，引入了关系特殊化，用于表达关联规则的语义，提高模型的语义表达力。

在概念模型结构扩展的基础上，本书具体研究了使用关联规则知识丰富ER模型的相关技术，重点解决在交叉领域下新的概念模型的新问题、新性质，以及由此产生的对数据库建模的影响和相关扩展。第一个问题与新概念模型中的语义约束有关，具体研究了秩约束的建模、推理、定义和一致性检查，用于保证在新概念模型中正确的、合理的定义秩约束，不影响模型中的一致性。第二个问题与关系数据库建模有关，具体地设计了从概念模型到关系数据模式的转换方法，该方法是保持信息含量不变，并得到满足范式要求的关系数据模式。此外，在动态的业务环境中，信息系统的模型需要持续优化，因此我们重点关注数据模式演化，并具体研究了层次结构中的推理、模式的改动、模式演化中的可理解性变化，相关工作有助于数据模式的演化和精化。本书搭建的数据建模优化的框架能够优化数据管理和数据存储，支持对领域理解的不断深化，使信息系统针对用户需求提供高效的信息服务成为可能。

本书还关注零售业信息系统应用，重点研究了以关联知识优

化零售信息系统、提供个性化推荐的信息服务的问题。近年来个性化推荐系统备受关注，不仅在社会经济中具有重要的应用价值，也是一个非常值得研究的科学问题。本书以网上商城为例，将得到的关联知识嵌入零售信息系统，展示了数据模式的演进内容，使信息系统更加智能地支持个性化推荐以及客户关系管理。

本书对信息系统建模和数据挖掘两个领域进行了全新角度的探索，研究内容涉及信息系统、概念建模、数据挖掘、客户关系管理等多学科的理论和方法。本书关注理论和实践的结合，构建了使用关联知识进行信息系统的数据模型演进的理论框架，并研究了相关的若干技术问题，进而展示了在零售业背景下进行数据模式演化、提供个性化推荐服务的应用，相关内容具有理论以及应用的参考价值。

书中大部分内容来自作者近年来在清华大学和中国人民大学进行的大量专题研究的成果，在此作者要感谢导师清华大学经济管理学院的陈国青教授曾给予的学术指导，使该研究课题不断向前推动和深化。感谢清华大学经济管理学院的李飞教授从营销科学的角度给予的宝贵建议，不断夯实相关研究的应用根基。

感谢中国人民大学信息资源管理学院的大力支持，感谢中国人民大学 985 工程的资助，使本书得以顺利出版。感谢浙江大学出版社的各位老师为本书的编辑出版付出的艰辛的劳动。最后，还要感谢爱人的理解支持和父母无微不至的关怀，不断给我信心和动力完成本书的撰写。

面对这样一项探索性的课题研究，作者深感水平有限，书中疏漏之处在所难免，恳请读者批评指正。

2011 年 12 月于中国人民大学

目　录

第 1 章　引　　言 …… 1

1.1　信息管理与商务智能 …… 1
1.2　信息系统的智能化趋势 …… 4
1.3　本书的研究内容 …… 8

第 2 章　信息系统的数据建模 …… 11

2.1　信息系统生命周期 …… 12
　2.1.1　瀑布模型 …… 14
　2.1.2　快速原型法 …… 16
　2.1.3　增量模型 …… 19
　2.1.4　螺旋模型 …… 21
2.2　实体关系模型 …… 23
　2.2.1　ER 模型 …… 24
　2.2.2　EER 模型 …… 32
2.3　业务规则 …… 38
　2.3.1　完整性约束与数据依赖 …… 39
　2.3.2　秩约束 …… 42
　2.3.3　覆盖约束 …… 46

第 3 章　数据模式 …… 54

3.1　数据模型与数据模式 …… 55

3.2 模式转换 …… 57
3.2.1 关系模型 …… 58
3.2.2 从 ER 模型到关系模型 …… 60
3.2.3 模式转换方法对比 …… 64
3.3 模式演化 …… 66
3.3.1 模式修改、演化与版本化 …… 68
3.3.2 模式的更改 …… 70
3.3.3 模式演化的框架 …… 75

第 4 章 数据挖掘的若干方法 …… 81

4.1 数据挖掘概述 …… 82
4.1.1 数据挖掘、知识发现、商务智能 …… 82
4.1.2 数据挖掘的过程和体系 …… 83
4.1.3 数据挖掘的任务 …… 86
4.2 关联规则 …… 89
4.2.1 关联规则概述 …… 89
4.2.2 关联规则的挖掘 …… 90
4.2.3 关联规则的形式扩展 …… 94
4.3 聚类 …… 95
4.3.1 聚类分析概述 …… 96
4.3.2 聚类方法 …… 97

第 5 章 基于关联知识的 AR-EER 建模 …… 108

5.1 概念建模方法 …… 108
5.1.1 关联规则的语义 …… 108
5.1.2 实体特殊化与关系特殊化 …… 112
5.1.3 扩展的实体关系模型:AR-EER …… 115
5.1.4 AR-EER 模型评价 …… 118

5.2 语义建模方法 …… 122
5.2.1 AR-EER 模型中的秩约束 …… 122
5.2.2 秩约束的蕴涵与推理 …… 123
5.2.3 秩约束的一致性 …… 133
5.3 基于 AR-EER 的数据库建模 …… 145
5.3.1 基于 AR-EER 的数据库设计 …… 145
5.3.2 模式转换算法及性质 …… 149
5.3.3 关系模式的规范化 …… 152
5.4 基于 AR-EER 的模式更改 …… 154
5.4.1 模式的更改 …… 155
5.4.2 层次结构中的推理 …… 157

第 6 章 以关联知识增强零售信息系统 …… 161

6.1 零售业与客户关系管理 …… 162
6.2 个性化推荐系统 …… 165
6.2.1 零售业中的商务智能应用 …… 166
6.2.2 个性化推荐方法概述 …… 170
6.2.3 基于关联规则的推荐方法 …… 173
6.3 以关联知识支持 CRM …… 183
6.3.1 以顾客为中心 …… 184
6.3.2 网上商城的数据模式演化 …… 186

第 7 章 结 语 …… 195

7.1 信息系统演进与竞争优势 …… 195
7.2 实践面临的挑战 …… 197

术语表 …… 200

第1章 引 言

信息化是当今世界经济和社会发展的大趋势，随着计算机技术、网络技术、通信技术等现代化信息技术的发展和应用，人类社会步入了信息时代。如何更好地利用信息技术促进经济和社会发展是全球政府和企业的重要关注点。信息化战略已经成为国家发展的重要战略，“十二五”规划已明确提出要推动信息化与工业化融合，推进经济社会各领域信息化，全面提高信息化水平。在我国，大型企业普遍已经建立起信息系统支持企业运作，中小型企业的信息化建设也正如火如荼地展开。[1]但信息系统应用的深度和业务支撑能力有待进一步提高。

1.1 信息管理与商务智能

信息管理的思想、内容、方法等随着技术的进步与应用的深化而不断发展和革新。追溯信息管理的发展沿革，按照企业利用数字化信息支持决策的程度，信息管理基本上可以分为三个层次（见图 1-1）：事务处理（Transaction Processing，TP），分析处理（Analytical Processing，AP）和商务智能（Business Intelligence，BI）。事务处理指的是围绕企业的基本业务和生产过程的自动化，对数据和信息进行加工和处理，在管理中，它能够回答“发生了什么”的问题；分析处理指的是围绕企业的分析和控制功能，对数据和信息进行回溯、分维、切片和 What-if 分析，从而回答“为何会发生”的问题；商务智能则是围绕企业的经营策略和竞争优势，

对数据和信息进行挖掘和整理以便获得支持决策的知识，从而回答“将会发生什么”的问题。信息管理利用的三个应用层次的逐级提升，是围绕对信息资源的开发、管理和利用展开的，在很大程度上代表着现代企业竞争力水平的三个台阶。

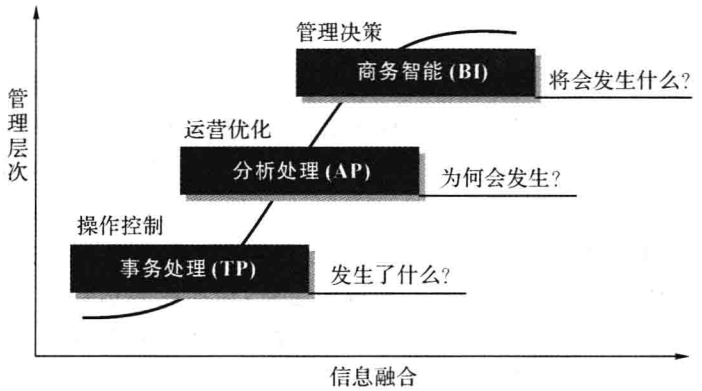

图 1-1 信息融合与管理层次[2]

企业信息管理初期是事务处理的层次，在这一层次上的信息系统是事务处理系统(Transaction Processing System，TPS)，用于处理发生在组织内部的事务，实现数据的电子化采集、交换和处理，其主要职能是捕获数据、创建数据和存储数据，并将数据传递给用户。在各种组织中都可以发现 TPS 在许多职能中发挥着作用，比如有负责处理薪酬工资、库存、订单的事务处理系统，它们用于提高事务处理效率与保证其正确性。

在分析处理层次上的信息系统是管理信息系统(Management Information System，MIS)，它能够通过多维度的、分析性的处理提供数据汇总报告的信息，从而向用户提醒存在的问题或者机会。一家应用了 MIS 系统的主管人员这样说：“我们将整个组织的信息都存储在大型主机上，我可以在任何时间登录，进行

各种类型的查询，以生成我需要的精确的 MIS 报告。这比在一个确定时间输出预定报告的传统方法好得多，我甚至可以建立自己的报告系列，在需要的时候从菜单上选中它就可以。”[3]广为熟知的企业资源计划（Enterprise Resource Planning，ERP）系统将采购、生产、销售、财务、人力资源等功能结合为一体，以系统化的管理思想对企业资源进行综合平衡和优化管理。它对于改善企业业务流程、提高企业核心竞争力具有显著作用。

企业在现实中面临的很多决策任务要复杂得多，为了支持最佳决策的制定，要求对数据进行深入的、智能化的挖掘分析，寻找潜在的未知知识进行预测，这就发展到了商务智能的层次。商务智能是信息资源利用从数据和信息层次上升到知识和智能层次的关键性纽带。在这个层次上，信息系统应具有很强的利用模型和智能技术的能力，并为用户提供友好的接口，智能化地提供辅助决策信息。此时，用户也已经认识到信息系统以及数据的价值所在。例如，企业可以通过挖掘详细的销售数据库以发现顾客的购物模式，并将其作为广告和市场促销的基础。

随着电子商务的发展，网上大宗货物交易和配送体系为零售企业提供了大量的信息和商业机会，同时也对传统的经营方式构成了很大的挑战，商业透明度加大，竞争更加激烈，市场更加复杂。同时，大型企业普遍已经建立起信息系统支持经营运作，业务管理系统、销售 POS 终端、互联网、供应商、客户、政府等方方面面都在不断地向企业提供信息。数据爆炸（data explosion）已成为信息时代一个突出的特点。如何有效地利用海量信息提高工作效率、增强核心竞争力，已成为企业当前关注的重要问题。企业尤其是大型企业目前已经开始接受并采纳先进的管理和技术理念，希望通过有效的挖掘和使用海量数据中的知识，以支持企业决策。数据挖掘技术因能够从大量数据中提取隐藏的、未知的、新颖的、有潜在应用价值的知识或模式，为决策者提供业务规

律、未来趋势、行为特征等策略支持，从而得到了业界和学术界的广泛关注，被认为是很有发展潜力的应用与研究领域。数据挖掘的应用领域非常广泛，主要有市场营销管理（如目标市场选定、市场购物篮分析、交叉销售等）、风险分析和管理、质量控制、信用管理、科研中的知识发现、智能查询、文本挖掘和 Web 挖掘，等等。

当信息技术与信息系统应用推动着企业的信息资源管理与利用水平发展到智能化的层次，全面的信息管理与核心竞争力的形成才成为可能。在现实世界中，各种复杂性和不确定性使得识别机会和风险并非易事；有关客户的数据和信息也常常是不完全的、不断获取的，有关的领域知识都是通过逐渐学习、探索得来的。探索得来的知识不仅仅应用于决策支持，还要与信息系统有效地集成起来，固化在系统中，使信息系统自动地完成查询、推理等任务，帮助企业加速精细化管理和提高企业的核心竞争力。比如，零售企业在积累了大量的顾客数据和交易数据后，使用数据挖掘技术寻找诸如购买模式和特征等知识，并把目标群体聚焦到某客户群体，在此探索过程中得到的对该群体的理解和知识逐渐被反映到系统中，又作为客户群跟踪和下一步研究的基础，从而演化出蕴含了独特的客户群体知识的信息系统，使企业具有核心的竞争优势成为可能。

本书关注商务智能层次上的信息管理，对信息系统和商务智能进行全新角度的探索，提出将数据挖掘得到的知识嵌入到信息系统中，推进信息系统向智能化系统的演化，使信息系统更动态、更灵活、更具针对性地满足用户需求，从而具有更高的应用价值。

1.2 信息系统的智能化趋势

随着信息技术应用的不断发展，信息系统在向着智能化方向发展。智能化的信息系统能够综合应用信息管理、数据挖掘等多

领域技术和方法，实现数据和知识的获取、存储和利用，提供高效率的信息服务。其优势在于它能够根据不完全的信息资料，预测所研究的物质、过程、现象的属性。[3]在业界，信息系统市场与商务智能市场出现了大规模的整合趋势。在2007年一年中，全球三大商务智能厂商Business Objects、Cognos和Hyperion分别被SAP、IBM和Oracle所收购。这也从一个侧面上反映出，在企业集成化信息系统的基础上融合具有绩效管理和决策支持能力的商务智能应用，正在成为提升企业竞争力水平的一个新的突破点。

在学术界，信息技术的飞速进步和广泛应用促进了信息系统建模和数据挖掘两个领域的发展。信息系统建模旨在构造数据模型，以反映现实世界中的对象及其属性和关系，即“有什么、是什么”的问题；数据挖掘能够发现隐藏的、有趣的关系。如图1-2所示，设计人员根据用户提出的业务需求而设计开发出信息系统及其核心数据库，把数据不断累积到数据库，作为对现实世界及业务活动的结构化记录。数据加工之后得到信息，被进一步组织起来用于支持决策并有利用率的信息则称为知识。在海量数据的背景下，数据挖掘技术可以帮助我们从数据中发现隐藏的知识。但是，数据、信息和知识三者不是一个简单的台阶式的概念，而是一个相互转换、螺旋发展的事实。[4]

本书从新的角度对信息系统、数据挖掘的交叉领域进行研究。数据挖掘得到的知识反映了数据中隐藏的关系，是对传统的数据建模所依赖的用户和系统分析人员的领域知识的重要补充，可以用于丰富现有的数据模型。具体来说，数据库是人为构造的，用于存储用户需要和感兴趣的数据；在此基础上挖掘得到新颖的知识，这些知识在系统设计阶段由于分析人员提取数据语义能力所限未能发现，或者由于用户未提出相应需求未被关注，但随着业务环境的不断变化有可能是用户感兴趣的。在新知识引

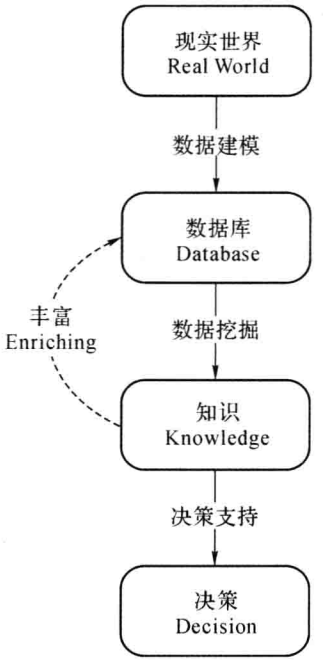

图 1-2 信息系统建模与数据挖掘

发了新的关注点时，用户也有可能提出新的数据需求以做进一步研究，但是现有数据库未必能够满足这种需求。那么对新发现的知识和新的变化，信息系统需要及时作出优化和补充，存储新的数据和知识。从数据模式的角度来讲，最初的数据模式可能是相对粗糙的，因为人们无法对数据中体现的规律作出预见，或者对某维度的分布不感兴趣；只有在数据积累到一定程度时，某些规律才会显现，并被挖掘出来展示给用户。如图 1-3 所示，提取出的数据中的关联和规律，可以用于对数据模式进行重组和梳理，并在必要的时候添加新的结构（如属性）来满足数据需求，以此得到扩展的数据模型将具有更丰富、更新颖、更准确的语义表达力，进

而提升数据模型的有用性，最终使信息系统更动态、更智能、更具针对性地提供信息服务。

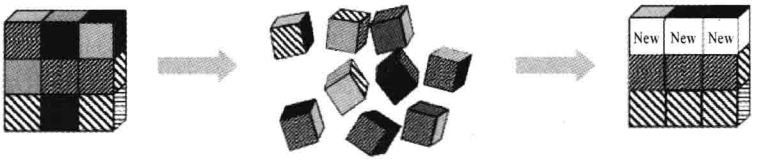

图 1-3　数据模式

将数据挖掘发现的新知识作为对领域知识的重要补充固化到系统中，增强的信息系统将具有以下优点：①优化数据存储，并且是建立在数据本身体现的规律上的优化；②面向用户，针对需求提供高效的信息服务；③支持对领域的理解不断深化，支持多层次的数据管理和信息服务。

数据库领域著名学者、ER 模型创始人 Peter Chen 在 2002 年的一次国际大会报告中总结 ER 建模的成果、发展与未来时也提到这个研究契合点："(数据挖掘得到数据中潜在的)关系已经存在，我们需要发现它们并利用起来……"近年来，也出现了一些将数据挖掘得到的知识用于数据建模的工作，主要是针对新语义扩展领域本体(domain ontology)模型[4,5]，以及针对新数据依赖修正数据库[6]，本书的工作是用数据挖掘得到的知识(如新模式、新结构)丰富数据模型，这是在相关领域探索性的一步。

有必要说明的是，数据挖掘得到的知识并不是都适合、或者有必要用于数据建模。在动态的商业环境中，模式不是一成不变的，一般来说，将选择那些具有一定的稳定性、兴趣性和重要性的模式来丰富数据建模。而且这个过程需要领域专家的参与。

1.3 本书的研究内容

本书提出了根据数据挖掘知识进行信息系统演进的框架，并聚焦到一种重要、常见的知识——关联规则，具体探讨了使用关联规则知识丰富概念模型的方法和技术。

首先，本书构建了数据挖掘和信息系统建模两个领域的桥梁。数据挖掘和信息系统建模原本是随着信息技术的飞速发展而兴起的两个研究领域，本书选取了两个研究领域中的重要概念——关联规则和实体关系(ER)概念模型，建立了两者的术语和语义的对照，阐述了根据关联规则知识丰富 ER 模型、进一步提高概念模型的语义表达力的思路。数据挖掘和数据建模的交叉领域的课题，提出根据新发现的关联规则丰富 ER/EER 概念模型。

其次，本书具体研究了使用关联规则知识丰富 ER 模型的基本框架和相关技术。在考察关联规则的语义的基础上，首次明确了 ER 模型中不同形式的特殊化——实体特殊化与关系特殊化，其中引入关系特殊化的建模原则和概念是为了表达关联规则的语义，提高 ER 模型的语义表达力。在对 ER 模型进行结构扩展的基础上，具体围绕信息系统演进中的相关问题进行了探讨，以解决在交叉领域下新的概念模型的新问题、新性质，包括由此产生的对数据库建模的影响和相关扩展。第一个问题是新的概念模型中的秩约束建模，这是在概念建模中常见的约束。在特殊化结构中存在秩约束蕴涵，即现有的秩约束对新结构上的秩约束构成了一定的约束，对此我们给出了相关推理规则，对新结构上的秩约束的定义给出建议。秩约束的一致性问题也是我们所关心的，通过一套不等式系统证明了新定义的秩约束不影响原有的秩约束一致性。第二个问题是新的概念模型的新结构对数据库建