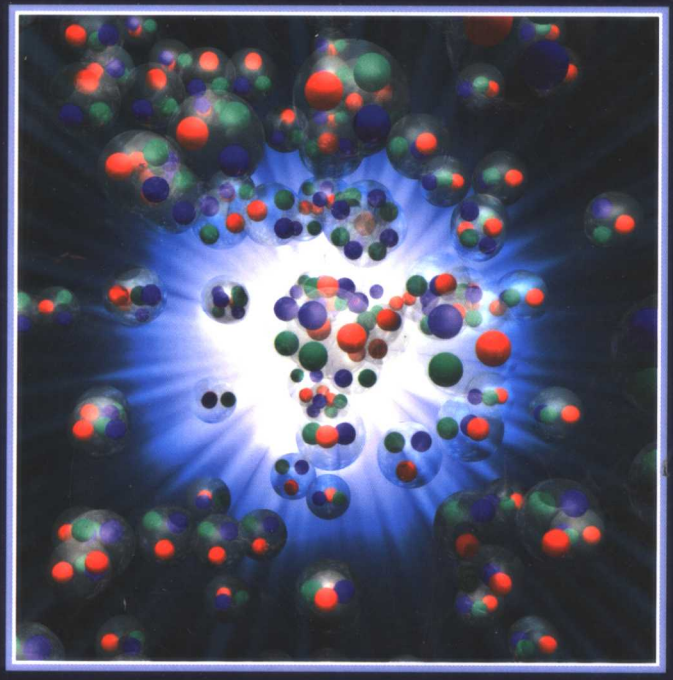
数据库技术丛书

业务建模与数据挖掘

Business Modeling and Data Mining



(美) Dorian Pyle 著

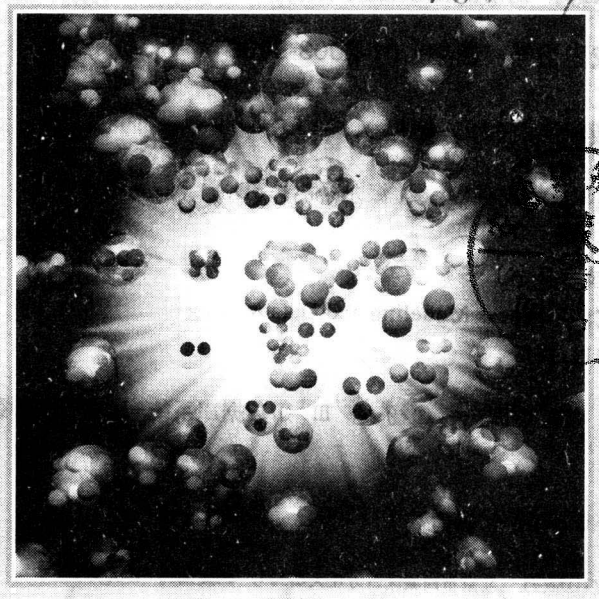
杨冬青 马秀莉 唐世渭 等译

机械工业出版社
China Machine Press

数据库技术丛书

业务建模与数据挖掘

Business Modeling and Data Mining



(美) Dorian Pyle 著
杨冬青 马秀莉 唐世渭 等译

机械工业出版社
China Machine Press

本书系统介绍业务建模与数据挖掘技术。

内容涵盖了如何发现、构建和提炼在业务情景中有用的模型；如何设计、发现和开发挖掘所需的数据；如何提供为各种业务情景挖掘数据的实用的方法等。

本书适合从事业务建模和数据挖掘以及相关领域的专业技术人员参考。

本书版权登记号：图字：01-2003-5009

图书在版编目（CIP）数据

业务建模与数据挖掘/（美）派尔（Pyle, D.）著；杨冬青等译. -北京：机械工业出版社，2005.4

（数据库技术丛书）

书名原文：Business Modeling and Data Mining

ISBN 7-111-16194-7

Ⅰ.业…　Ⅱ.①派…　②杨…　Ⅲ.①数据库系统-建立模型②数据库系统-数据采集　Ⅳ.TP311.13

中国版本图书馆CIP数据核字（2005）第016501号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：王镇元　梁开莉

北京昌平奔腾印刷厂印刷·新华书店北京发行所发行

2005年4月第1版第1次印刷

787mm × 1092mm 1/16 · 28印张

印数：0 001-4000册

定价：55.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

译 者 序

随着数据库技术的发展和数据库应用的普及，全世界范围内，在业务管理、政府管理、科学与工程数据管理和其他应用领域所使用的数据库的数量和规模都在迅速增大。于是，如何从数据库所积累的大量数据中及时有效地提取对管理决策有用的信息和知识，成为了几乎所有经营管理者所关心的一个共同问题。数据挖掘的目标就是从数据库的数据中发现这样的有用的规则和模式。数据挖掘研究作为数据库和相关领域近年来非常重要和非常活跃的研究领域之一，吸引了来自数据库系统、人工智能、机器学习、统计学、数据可视化……等许多领域的研究人员进行跨学科、跨领域的综合研究。这些研究的重点主要集中在分类、聚类、关联、概要等挖掘操作的高效算法上，产生了各种期刊和会议上发表的高水平学术论文，所研究出的方法和技术也已经成了DBMS的主要厂商的数据挖掘产品的一个部分，并且已被各行各业大大小小的用户所使用。

然而，各行各业各个层次的经营管理者——数据挖掘工具的真正用户所面临的许多非常切实的问题，在传统的数据挖掘研究和教学中却没有得到足够的关注，例如，数据挖掘技术能够在哪些领域中最有效地应用？如何建立挖掘模型，从而把业务问题转化为数据挖掘能够解决的形式？如何为数据挖掘工具准备数据等等。数据库领域29位高级研究人员于2003年5月在美国马萨诸塞州的Lowell举行了评估当前数据库研究状态、推荐值得特别关注的问题领域的第6次不定期讨论会，基于他们的讨论所形成的“Lowell数据库研究自评估报告”中也谈到“我们希望数据挖掘超出基本运算算法的范畴，同时我们还有一个感觉，那就是，计算机科学和IT课程应该包括更多的关于数据挖掘工具的正确使用的内容。”

本书正是针对上述问题而著成的。作者Dorian Pyle具有超过25年从事数据挖掘工作的经验，担任着若干个数据挖掘工具公司、信用卡业务公司、制造业公司的顾问。他对各个行业，各种层次和角色的经营管理人员所面临的业务情景和挖掘问题具有深入的了解，本书给出了解决问题的方法学和实际步骤。我们深切感到国内从事数据挖掘工具开发和使用数据挖掘工具解决实际业务问题的经营管理人员都需要这样的一本参考书。坦率地说，我们近年来在数据挖掘领域中的研究也主要是前面所说的在不同类型的数据上进行挖掘的高效算法的研究。对于实际业务环境中为了挖掘而建模和数据挖掘工具在实际业务中的应用，我们也缺乏实际的经验和体会，本书的翻译过程也是我们的一个学习过程。我们希望本书对于数据挖掘在我国各行各业的实际应用起到推动作用。

杨冬青、马秀莉、唐世渭组织并参加了本书的翻译和审校工作；参加翻译的还有遇辉、姜力争、张德辉、李双峰、梅源、李希婷、田枫、袁征、赵翔宇、武纬。

限于译者水平，译文中疏漏和错误难免，欢迎批评指正。

译 者

2004年12月于北京大学

前　言

一次在研讨会上发言后，有位听众走上来跟我探讨一个问题。他说自己在一家银行工作，并且还在教一门研究生数据挖掘课程。他问我对于如何把一个问题的业务化描述转变为能用数据挖掘和数据回答的形式有何建议。确实，同一组的好些其他人也问过我同样的问题。从那以后，我多次在不同场合被问到类似的问题，这些场合包括从email讨论、学术会议上的听众提问、业务上的指导演讲，到专门研究如何挖掘数据的个人。*如何表达现实世界中的业务问题以便通过数据挖掘来回答它们？*显然这已经变成一个紧迫的问题。然而在这之上还有一个更基本的问题：*数据挖掘能够有用地解决哪些问题——以及如何解决？*

发现数据挖掘最有效的应用领域，然后把业务问题转化为数据挖掘和建模能够解决的形式，这非常类似于每一个初学代数学的人所面临的问题——那就是令人烦恼的文字问题。如何把一个用文字描述的句子转换成实际上的一个数学公式？在某种意义上，本书是对这些问题的一个详尽解答——如何使用，哪里适用，如何在战略的和战术的商业应用中最有效地使用数据挖掘和建模。

本书的内容

本书的核心目标在于，帮助挖掘者或建模者构造和优化那些在商业背景下有用的可挖掘模型。当然，除了数据挖掘以外还有许多不同的构造模型的方法，而且存在大量不同的模型——远比采用数据挖掘方法可以构造出来的要多。然而，当利用数据挖掘来解决商业背景下的问题时，人们要么构造一个新的模型，要么理清一个已经存在的模型。在本书中，你将找到一个实用、可行的方法来从挖掘数据中取得最好的结果，以及足够的必要的理论来得到实际可用的答案。

模型是人类生活中不可或缺的一部分。数据挖掘中用到的模型就是人类知识的结构。无论我们是否意识到，所有我们关于这个世界的知识——个人的、社会的、商业的、科学的、政治的、情感的、无论何种形式——都可以被构造成模型。我们使用了太多的工具来改变、调整或阐明我们的知识——我们对这个世界建立的模型。数据挖掘只是其中一个工具，尽管是非同寻常以及有些陌生的一个——因为它是一个自动的过程。这些模型跟我们所说的“思考”这一过程很相近，而对于思考过程的任一部分能够自动完成这一概念，我们会非常不习惯。确实，现代的技术对我们关于思考如何组成的固有观念是如此的具有挑战性，以致我们开始重新定义思想的本质。

不管这些工具看起来在根本上对我们多么具有威胁性，它们在这个越来越复杂的世界中仍然非常有用。本质上，人们试图做到：改善我们对具体事物的有导向性和有目的的控制，从而达到大家都认为有益的结果。我们的基本假设是，如果我们对因与果的本质以及它们之间的关系有更好的认识，而且可以在某种程度上对因实施一定的控制，那么我们就可以影响得到的结果。

由此，这本书交织了两个主题——挖掘和建模。前面我曾经做过一个关于代数学的类比，代数学本身就是一种模型。它是一种能够用来以给出有效答案的方式对符号化表示进行操作的工具。对于不熟练的人来说，把一个领域中表示的问题（文字叙述的问题）转换为另一个领域中

的问题（代数学中的形式化符号）是很困难的。对业务模型的数据挖掘来说，幸运的是，我们所有人开始时就对本书中讨论到的模型有直观的了解，因此它们比代数模型更容易懂。不过，为了把挖掘和建模两个主题交织起来，我们必须勾勒出这些知识模型是什么，以及如何在实践中使用它们的一个轮廓。

我们并不经常把模型当作对象，或者关注如何去构建和使用它们。然而理解在不同场合中使用的适当模型的范围，这构成了一个数据挖掘者用来表达挖掘结果的基本“语言”。但是模型并不只是挖掘的结果，它更像一种“输出”。在许多情况下，挖掘者需要在挖掘开始之前对一种业务情形建模。这个挖掘前建立的模型用于确定：一个业务情形中哪里存在不确定性，决定挖掘在哪里能够贡献出最大价值，还有发现哪些数据需要被恰当地挖掘以找到答案。其他挖掘前建立的模型可能定义：数据需要如何进行扩展或丰富，或者它们可能对决定什么特征能被提取出来有用。实际上，挖掘发生在一个完全由各种不同模型组成的范畴。

由于模型对挖掘者，以及对寻求业务情形答案的挖掘过程都是如此的重要，本书既概略地说明了什么是模型，以及我们如何使用模型来描绘世界，同时又详细地介绍了那些在挖掘中起重要作用的特定模型。

数据挖掘和建模都是通用的技术，可以应用于广泛的问题。这里描述的技术能够，并确实已经被应用于非常广泛的领域；然而，如今数据挖掘主要被应用于发现商机以及解决商业问题。鉴于目前对发现商机和解决商业问题方面的关注，本书使用的例子主要引自商业世界。选择聚焦商业的实例是经过深思熟虑的，尽管这些方法和技术本身对于生物信息学或工业自动化等领域都同样适用。尽管那些领域的实例不那么丰富，你仍旧可以找到许多有用、有价值和有趣的东西。

数据挖掘目前在商业中被用作战术性的工具，显然在战术层面上它非常有价值；不过，核心的商业过程都发生在战略的层面上，并且正是对数据挖掘的战略性运用有望给一个公司带来最大回报。因此，在设置了挖掘和使用模型的框架后，本书着重于在一个企业的任何层次中既出于模型本身的目的，同时也在模型中使用挖掘，从战术性层面直到核心的战略性层面。

到目前为止，我讨论了这本书是关于什么内容的。为提供另一个不同的视角，与本书不包括什么内容做一个比较是很有帮助的。许多关于数据挖掘的书集中讨论算法的细节和如何尽可能有效地利用这些算法和数据。在本书中你很少会找到关于算法的内容。这个主题被简单地带过，因为不同类型的模型决定了挖掘者偏向于使用不同的工具。那些主要讨论算法的书都假设你已经理解了问题存在的框架。它们也假设你了解数据与手边问题的关系，数据应当如何被操作以得到期待的结果，为何认为这些数据反映了相关的问题，挖掘模型如何与商业环境相符合等。而这些假设正是本书所没有做的。实际上，本书阐述的恰恰就是上述问题。它直接阐述了如何动手处理、如何思考，以及如何组织各种问题来从你为挖掘付出的努力中得到最好的结果。

这本书覆盖了挖掘者和建模者关心的所有问题。它包括了对业务情形建模的方法论，还有挖掘数据的方法论，并且把两者结合在一起。但是这样还不足以让建模者和挖掘者在现实世界情形中装备自己，因此本书也涉及了以下实际需求：如何识别问题，如何获取基层和管理层对项目的支持，如何鉴别和量化一个项目的商业利益，如何选择适当的工具以获益，以及如何在建模和挖掘过程的每一步中构造和细化有用的模型——但不包括特定情形下的具体商业知识。对于这些，读者需要应用这里讲到的技术来发现和揭示自己在特定情形下的需要。

本书的读者对象

如果你懂得计算机，而且需要通过处理数据来做出明智的决策，那么，这本书是为你而写的。本书提供了一个框架，读者可以根据这个框架来分析自己的数据，并理解如何以及在何处应用它们。本书解释了怎样用模型来理性地形成一个决策，以及挖掘在何时何处才是一个提供决策所需信息的合适工具。尽管本书中有许多内容会有助于做出任何基于数据的决策（甚至可能做出任何类型的决策！），然而确实，这本书的主要读者对象有着严格的设定。下面我给出几个简短的角色描绘来说明谁会从本书受益：

- “安娜”每天都与数据打交道，为其他部门提供所需的数据和数据的汇总。她的老板读了一些关于数据挖掘的文章，叫她对数据挖掘做更多的了解，最终构建一些模型，从而帮助他们每天为之提供数据的那些部门。她对数据挖掘一无所知，对建模也是一样。*这正是适合她的书。*
- “巴里”是一个商业分析员。他的工作对象是新近创建的数据仓库——与一项已完成的任务相比，这更接近于一个不断发展的项目——并且他已经做了不少报告。他了解商业形势的建模，对如何从数据中产生商业报告也了解一些。他希望在现有简单数据统计的基础上更进一步，同时，他听说过数据挖掘；然而，他目前仅仅知道 1）数据挖掘具有商业方面的应用；2）数据挖掘使用了一些非常复杂的算法。巴里希望能了解处理数据的算法如何能被实际地用来解决商业上的问题。*这本书也适合他。*
- “凯瑟琳”是一家公司的财务分析员。她的MBA课程包括了几门统计学方面的课，而她也习惯于对她的数据做统计学的解释。作为一个高级分析员，她的任务是支持管理团队决定公司的战略方向。她很清楚仅仅提供详细的财务报表是远远不够的，即使这些报告确实包括了一些预想问题的讨论。由于存在大量可获得的数据，她作为管理团队的顾问组中最熟悉统计学的人，大家期望她能够使用这些数据来获取信息以及使局势明朗化。尽管她对财务数据如鱼得水，但是她没有在提出和解决商业问题时和非财务数据打过交道。*这本书适合她。*
- “戴维”是一位资深中层经理，他一直在公司里提倡客户关系管理（CRM）的潜在商业优势。高层管理人员让戴维给他们演示采用这样的模型后，公司的业绩将如何得到改善，潜在的易犯错误和困难是什么，以及他们如何把自己已经知道的东西结合到这个新的手段中。*这本书是适合他的。*
- “伊丽莎白”一直负责实现公司的数据仓库。付出了大量的精力，进行了艰苦奋斗，经历了无数的挫折后，基本完成了一个有用的至少是初级阶段的数据仓库项目。除了数据仓库本身，还提供了一个联机分析处理（OLAP）工具作为主要的访问仓库的途径。好几个经理都对伊丽莎白表示了他们的诧异，因为在投资了如此大量的时间、精力和金钱后，这个项目传递出的商务智能似乎达不到他们预期的水准。她被要求从公司的数据仓库中调查如何改善业务智能的创建与产生。*这本书也是适合她的。*
- “弗雷德”是一个数据挖掘者和分析员，他上过几门数据挖掘的课，全都是教他不同的算法以及如何把这些算法应用到数据上。这些都充实了他原有的统计分析知识，上完课后，他觉得已经有足够的知识来应付现实问题。几个经理热切地想看新的工具能够产生什么，于是

把他们的数据交给弗雷德，看他能提出什么精辟见解。就所得到的结果来看，尽管弗雷德在技术上对这些结果已经满意了，然而却没能给他的顾客——那几个经理——留下深刻印象。弗雷德被要求确认是否能够提高所产生的业务智能的价值。这本书正是为他而写的。

- “吉莉安”在加入她的公司之前以社会学学位毕业。尽管没在这一领域工作，她学到了许多关于创建和解释不同情形的模型，如，监视干扰，还有评估结果等。作为一个经理，她觉得这些工具和技术一定对提高她的业务运作的效率、组织和满意度是可行的。然而，如何定义最适当的问题以及使用手边的数据令她感到困难。不知何故，她尝试去牢固地掌握她所看到的问题，并产生可信的结果来提供给别的经理，可是产生的实质结果不如预期。吉莉安在寻找一些定义她看到的业务问题的方法以及获得改善的渠道。这本书适合她。
- “哈利”了解定性的以及定量的建模。他确定了大量需要解决的业务问题——实际上是太多了。呈现在他面前的选择的数目如此之多以至于无法一一应付。而那些他处理过的问题又有出乎预料的结果。他需要某种组织和评估这些问题的方法，然后用一种令人满意的方法解决它们。这本书也是适合他的。
- “英格”做数据挖掘已经有一段时间了，主要是为营销部门服务。目前，她的工作包括无穷无尽的顾客履历描绘，和为公司的顾客吸引、员工激励和长期保持进行市场细分。她知道她正在使用的这些工具能够对公司更有益处，同时她说服了经理给她机会显示这些工具能做什么。然而，寻找一个最好的地方入手原比她预想的要困难。这本书同样适合她。

上述所有的人物和情景都是基于我所遇到的在建模和挖掘相关活动中的真人真事。这些人都面临许多不同的问题，但是他们都有某些共同的东西：他们都被要求建立（或发现）商机或者问题，并给出受数据支持的有效解决方案。其中一些人比其他人更熟悉部分过程，然而他们所有人都被要求扩展视野，以涵盖自己之前从未接触过的领域。

他们都面临一个基本问题：如何发现最佳的行动路线。他们面临的问题是，探索当前的形势和环境，从而使那些工具和技术能够最好地应用到数据上，最终产生能够用于改善形势的结果。这个问题的成分包括：

1）他们目前的洞察力、理解力和知识。
2）他们面临的形势与环境。
3）评估形势和环境的工具。
4）可能相关的数据。
5）评估这些数据的工具。
6）应用这些工具的技术。
7）如何发现正确的问题。

当然，直到正确的问题被发现为止，前6个部分才能被正确应用。这本书提供了相关的框架和指引，把所有这些部分紧密捆绑成一个密不可分的整体。本书中，这整个过程就被称为建模。

总而言之，技术目前的发展状况把工作任务在信息发现者和决策制定者之间分开了。越来越多在商业机构里的人，或者在内部或者在外部，都有把数据挖掘应用到数据的工作任务，工作的结果将被那些自己既不挖掘也不建模的决策制定者参考。正是部分地由于这种分割把数据挖掘放到了战术的而不是战略的领域。举例来说，一个销售经理很容易可以把发现一个新的市

场细分的任务分派给“数据挖掘组”或者“量化分析组”。然而对同样的销售经理来说，要看到数据挖掘在发现新的产品、或公司应追求的全新方向中扮演什么角色的话就比较困难了。还有，公司的会计很可能认为数据挖掘可以检测出欺诈性的事务，却发现挖掘在公司资源配置或者人事安排组织方面几乎不起任何作用。确实，除非挖掘者直接参与管理链，否则很难通过能使挖掘增加见解的方式来对战略性问题进行形式化。不过，挖掘和建模两者都在所有层次上扮演一个完整而重要的角色，包括在核心的公司战略层次。

为了完成从战术到战略的飞跃，挖掘者和建模者，无论是雇员还是承包商，都要扮演顾问的角色。这本书就是主要针对这群越来越多的内部或外部“顾问”的。它能为你提供一个框架，该框架把问题从提出者表述的形式转化为挖掘或建模能够解析并反馈所需结果的结构。自然，这里也有很多适合计算机和数据处理经理的内容，这里描述的框架最终都会被公司的管理层使用和实现。

本书的结构

描述如何开始建模过程需要几条线的交织理解。欣赏一块布的图案就必须把这块布所有的线看成一个整体。然而，对于这块布本身的织就来说，每次只能织一条线；对于模型挖掘的解释也是同样的。不是一条线的解释，而是多条线的综合才能创造出最终的图案。然而将要被交织起来的这些线——主题——是什么呢？

令人意外的是，其中一根线是：*理解我们已经知道的事情*。我们本身具有的知识、洞察力、偏见、喜好、渴求、希望、恐惧、志向等都直接影响了我们的所见。在很大程度上，我们已经了解的事物决定了我们所能够了解的事物。实际上，这根线是如此重要——尽管不完全是直觉的思路——以致本书开始的几段下了很大工夫，至少较为详细地介绍这个领域。

另一根线是：*解决问题的过程*。这里使用的术语问题仅仅是为了表达找到潜在行动的一个合适过程或一些过程的困难。换句话说，问题在于为了达到特定的期望的目标，我们要发现我们能做什么。在这个意义上，术语问题也涵盖了商机的发现，正如“问题在于寻找最好的发展机会”。我们将看到，永远不会只有一个可能的行动过程，其他的被交织起来的解释线尝试去发现应该选择哪一个可用的行动过程。尽管看起来发现可用的行动过程很简单，实际上有时候会非常困难。

另一根线涉及到：*按照我们对于什么是重要的和对于事物如何整体协作的感觉来刻画一个情形的特点*。一个对此更加正式的描述是定性的建模，不过其实质比它那令人望而生畏的名字所蕴含的内容要简单。我们所直觉知道的、思考的、感受到的和产生的情绪在处理商业情形时是非常重要的。有多少时候一些项目完成了，由于它是总裁偏好的项目？有多少时候一个项目被忽略，由于它感觉上就不对，或者由于“不是我们这里办事的方式”？又或者，“头儿”经常没能看出那些直观上就很明显的东西。简单地把潜在方案排队，然后编号寻找解答并不总会成功——可能永远不会成功。建模这块织布中需要找一个能合理地把这根线编织进去的地方。只有沿着这根线，模型才会整合与评估这种“软”信息。

当然，*数据挖掘*必须有它自己的线。然而数据挖掘只是近来对定量的建模工具的一个扩展，主要是数学和统计学的工具。尽管不是全部，一般许多沿着这根线引入的内容都可以应用到定量的方法上。然而，数据挖掘有自己不同于其他定量方法的特殊技术。不过无论是用了什么样

的定量方法，这都只是整幅织布中的一根线。它必须与其他的线在平等的基础上组织起来。在许多商业建模的情形中，定量的数据可能不能得到或不必用到；然而这本书关注于那些能够获得定量的数据并将其有效使用的地方。由此，定量地组织数据是本书关注焦点的结果，实际上它并非所有情况下的一个要求。

在整本书的各个部分中，这些线都会或多或少地被阐述。为了便于解释，一些线不时地比别的线移至更加突出的地位。然而，这种关注点的改变仅仅是为了解释的简便。正如一块布是由它所有的丝线构成，挖掘和建模的解释也是如此。没有哪一个比另外一个更重要，要构成整体需要所有的线。

这些线就是书中的主题，贯穿于整本书——一种不同主题的叙述性解释的综合。然而，对事物是如何起作用的解释很难进行实践。最后一根线是*实际应用*，它解释了如何实际地建立适当的模型，以及挖掘特定的数据来为模型添加信息。你需要的指示越具体越好。基础工作完成后——也就是说，这块布的基本形状已经形成——这根线跃至显要位置，并揭示你应当思考些什么，如何确定你需要和谁交谈，你应当问什么问题，以及如何问这些问题。它还告诉你如何阐明概念和印象，确定创造情形的概略图的工具，解释说明那些概略图，以及交互式地完善这些概略图。当然，这条线也讨论到如何发现数据，如何处理它们，如何对它们建模等。因此这条线解决该做什么，但是它的细节要等到“为什么”的基础工作完成才会涉及。

这本书展示了挖掘者和建模者之间的明显区别。本书中，*建模者*总是与商业问题相关——与构造商业问题或机遇、与商业过程、与数据的商业主题、与对合适的商业过程的模型的应用、与利益相关群体的联系、与获取商业价值，以及与投资的回报相关。在发现了业务的框架后，*挖掘者*关注于挖掘数据——数据质量、工具的选择、合适的技术、关系的发现、可信度级别，还有模型的清晰度。可见，这是两个不同的角色，尽管他们之间有一定的交叉，出于解释和讨论的目的他们被最大限度地分开了。在许多项目中，同一个人仍然很有可能担任这两个角色。然而，每个角色的主题、关注点和活动都是不同的，即使它们体现在同一个人身上。分清这两个角色有助于完成项目。

第一部分关注于整体的、全面的东西。它提供了一些概括性的介绍材料。阅读这一部分后，挖掘者和建模者并不能立即行动，或者甚至学到任何具体的挖掘或建模技术。那么为什么把它放在这儿？除了介绍挖掘和建模的概貌，它还提供了一个建模者必须牢记于心的环境或方法，否则会有被骗或被困的风险。一个建模者就像一个寻找线索的侦探，他需要避免错误的假设，忽略不重要的现象，最后找到解释线索。对于一个侦探，不管是虚构的还是真实的，运用逻辑推理很少会得出诸如“Mustard上校在图书馆里拿着一把匕首”这样简单的结论。同样对于一个建模者，通往发现的道路很少做了清晰的标记，也不会是明显的。

第一部分显示了为了构造成功的模型，一个建模者应该考虑什么。它为后来更加详细地探索铺好了舞台。这里包括关于构造一个模型意味着什么的介绍。这里也是对模型表现现实世界的方式，以及模型及其环境之间交互方式的查看。在这部分中也有对数据挖掘这个全新活动的高层次的介绍。结果显示数据挖掘是人类精神活动的一个扩展，而这种精神活动自人类有史以来便已存在。在某种意义上，数据挖掘可能看来是涉及到以前被认为本质是人类的活动，但像我们将看到的，尽管它扩展了我们触及的范围，它仍然是完全由我们自己掌握的。第一部分以

通过对建模与挖掘技术有用的方式对设计问题进行了一个简要介绍，并以此作为结束。从全局来看，这是一个有意的简短介绍，它仅仅是点到为止。然而，它为理解后面的内容提供了一个必要的介绍和一个框架。

第二部分综合地介绍了如何对一个业务情形建模。业务建模的本质是创造一个能用最低成本和最低风险得到最优回报，并激起所有利益相关者的热情支持的结构。实现能够完全满足这一点的设计就是第二部分的本质。

第三部分详细地讲述了数据挖掘。在一个业务环境下，数据挖掘是一种艺术，这种艺术是发现、集成以及预处理数据，然后在这些数据上应用工具和技术，从而揭示其内在的规律和关系，这些规律和关系对于在数据和业务条件允许的情况下完成业务模型的目标是必要的。

尽管第一部分至第三部分的解释、阐述和例子都非常丰富和有用，但是第四部分才是本书的关键部分。第四部分提供了两个互补的和综合的方法论，这些方法论能够带领任何建模者和挖掘者一步一步走完创建业务模型和挖掘数据的所有阶段。

本书提出的方法论是全面的，因此你永远都会知道下一步该做什么。从项目开始到结束，该方法论为所有建模者和挖掘者提供了一系列详细的步骤。

本书特色

整本书形成的综合的催化（Catalyst）建模以及挖掘方法论可以看做一个超文本文档。你作为一个建模者或挖掘者在任何特定的阶段所需要做的取决于你到那时为止发现了什么。没有两个项目是完全相同的，而你发现的问题、见解和情形将决定你接下来需要做什么。在所有行动、选择、测试和技术中你的确切步骤都会因项目而不同。这些方法论表现为交互的文档，你会把它们用于获得无论你处于什么建模或挖掘情形下都适合的见解和建议——为了得到更多的细节，这些文档还引用了本书相应的章节。这个方法论的交互版本可以从作者的网站（www.modelandmine.com）下载。另外，这个站点还有许多其他的资源材料以支持本书，包括数据集、可用的例子、图示说明以及工具，许多在本书中都有讨论。

致谢

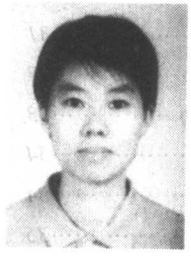
我要感谢我的客户们，尽管在本书中没有提到他们的名字，他们不仅给予我教学的机会，在与他们一起做业务建模和数据挖掘项目的时候尤其给了我学习的机会。完全是由于我做过许多项目——在许多不同的业务领域和行业中——的经验，本书中讲到的这些知识才能被结合到一起。

献给我亲爱的妻子Pat，她为这本书做了许多工作，并且负责理清了原稿中许多相当晦涩的部分，她还是支持这本书的网站的管理员，我实在是无法充分表达我的感激之情。这个项目同时也是属于她的。

我的朋友和同事Tom Breur、Marcelo Ferreyra、Elena Irina、Neaga和Ralphe Wiggins阅读了我的手稿并给出许多宝贵的建议，这些建议进一步完善了这本书。我还要特别感谢我在Morgan Kaufmann的编辑们：Dianne Cerra，她没能作为本书的编辑看到这本书的出版，还有Lothórien Homet，他中途接手了这个项目。我感谢他们的耐心和帮助，最终使这本书得以问世。

译者简介

杨冬青　1969年毕业于北京大学数学力学系数学专业，现任北京大学信息科学技术学院教授，博士生导师，网络与信息系统研究所副所长，数据库与信息系统研究室主任，中国计算机学会数据库专委会委员。多年来承担并完成“973”、“863”、国家科技攻关、国家自然科学基金等多项国家重点科研项目，曾获国家科技进步二等奖、三等奖和多项省部级奖，在国内外杂志及会议上发表论文百余篇，著译作十余部。目前主要研究方向为数据库系统实现技术、Web环境下的信息集成与共享、数据仓库和数据挖掘、典型应用领域的数据库技术等。



马秀莉　博士，1972年出生，2003年毕业于北京大学信息科学技术学院，获理学博士学位。先后参加过多项国家自然科学基金项目、国家重点基础研究发展规划（973）课题等科研项目及典型应用领域的应用研究项目，在国内外杂志及国际会议发表论文近20篇。主要研究领域为数据库系统实现技术、数据仓库、联机分析处理、数据挖掘等。

唐世渭　1964年毕业于北京大学数学力学系计算数学专业，毕业后留校任教至今，现为北京大学信息科学技术学院教授，博士生导师，中国计算机学会数据库专委会副主任。多年来承担并完成“973”、“863”、国家科技攻关、国家自然科学基金等多项国家重点科研项目，曾获国家科技进步二等奖、三等奖各1项，省部级科技进步奖多项，在国内外杂志及会议上发表论文百余篇，著译作多部。至今已培养硕士、博士、博士后60余名。目前主要研究方向为数据库系统、数据仓库和数据挖掘、Web环境下的信息集成与共享、典型应用领域的信息系统等。

目　录

第二部分 业务建模

第三部分 数据挖掘

第四部分 方法论

第一部分　本领域的概要

从某种程度上讲，本书是关于在通常情况下如何构建业务情形模型的一个详细的路线图，尽管本书主要针对的是那些能够使用计算量密集的分析技术——统称为数据挖掘技术——来进行有用的探索、知悉、阐明以及应用的特定业务模型。发现机遇或者解决难题是对业务形势进行建模的基本原理；然而，业务形势中的任何一个组成部分都不是在真空中存在的——也即它们都带有相当的负荷。

开门见山地对形成、探索、勾勒、阐明以及构建业务情形模型所使用的工具和技术进行详细说明而不通过任何导言，这是完全可以做到的。同样，直接深入到数据挖掘中的实质问题而不理会相关的论题或告诫也是可以做到的。然而，这样做则忽略了大量潜在有用的知识。直接深入到数据挖掘的工具和技术就完全遗漏了它们在面临现实生活时的一般意义上的基本认知：什么类型的思想和方法通常都是有价值的；什么样的前提可能一直潜伏着未被发现直到为时已晚；什么样的思路是建模者或挖掘者有可能需要遵循、探求、或者必须小心运用的。换句话说，直接深入介绍工具和技术，缺少的是对实际业务建模过程的更广概貌的了解——这个过程包括业务情形建模，挖掘数据以阐明模型，然后把从明晰的模型得到的推断应用回到对业务问题的实际可用的回答，首先产生效果。这本书的第一部分从广阔的范围来看待什么叫做对一个业务情形建模，以及什么叫做挖掘数据。

如果本部分成功地达到了它的本意，读者最后了解到的不是如何去对业务情形建模和挖掘，而是知道该去思考什么，应该注意什么，以及如何进行业务情形建模和挖掘数据的过程。其余的三个部分将讲述细节知识——在业务背景下进行建模与挖掘所需要的一系列方法。第一部分提供了建模和挖掘发生的上下文环境，形象地说这好比是对一个森林的总览，而后续章节关注的是这个森林下的树木个体。