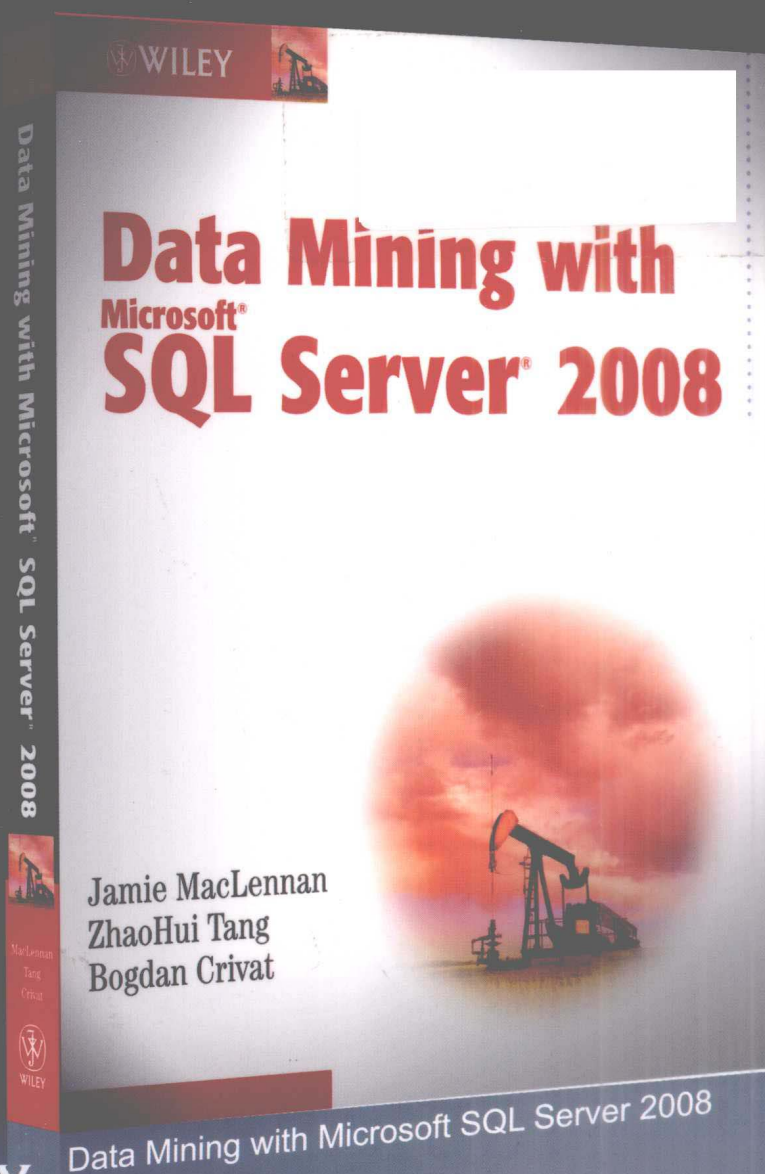


# 数据挖掘 原理与应用 (第2版)

## ——SQL Server 2008数据库

Jamie MacLennan  
(美) ZhaoHui Tang 著 董艳 程文俊 译 杨大川 审校  
Bogdan Crivat



国外计算机科学经典教材

# 数据挖掘原理与应用(第2版) —— SQL Server 2008 数据库

清华大学出版社

北 京

Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat

Data Mining with Microsoft SQL Server 2008

EISBN: 978-0-470-27774-4

Copyright © 2009 by Wiley Publishing, Inc.

All Rights Reserved. This translation published under license.

本书中文简体字版由 Wiley Publishing, Inc. 授权清华大学出版社出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

北京市版权局著作权合同登记号 图字: 01-2009-1938

本书封面贴有 Wiley 公司防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

数据挖掘原理与应用(第 2 版)——SQL Server 2008 数据库/(美)迈克伦南(MacLennan, J.), (美)唐朝晖(Tang, Z. H.), (美)克里沃茨(Crivat, B.)著; 董艳, 程文俊 译. —北京: 清华大学出版社, 2010.7

书名原文: Data Mining with Microsoft SQL Server 2008

(国外计算机科学经典教材)

ISBN 978-7-302-22842-4

I. 数… II. ①迈…②唐…③克…④董…⑤程… III. 关系数据库—数据库管理系统, SQL Server 2008—教材 IV. TP311.138

中国版本图书馆 CIP 数据核字(2010)第 096610 号

责任编辑: 王 军 崔 伟

装帧设计: 孔祥丰

责任校对: 胡雁翎

责任印制: 王秀菊

出版发行: 清华大学出版社

地 址: 北京清华大学学研大厦 A 座

<http://www.tup.com.cn>

邮 编: 100084

社 总 机: 010-62770175

邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者: 清华大学印刷厂

装 订 者: 三河市金元印装有限公司

经 销: 全国新华书店

开 本: 185×260 印 张: 32.25 字 数: 785 千字

版 次: 2010 年 7 月第 1 版 印 次: 2010 年 7 月第 1 次印刷

印 数: 1~4000

定 价: 58.00 元

---

产品编号: 030086-01

# 出版说明

近年来，我国的高等教育特别是计算机学科教育，进行了一系列大的调整 and 改革，亟需一批门类齐全、具有国际先进水平的计算机经典教材，以适应我国当前计算机科学的教學需要。通过使用国外优秀的计算机科学经典教材，可以了解并吸收国际先进的教學思想和教學方法，使我国的计算机科学教育能够跟上国际计算机教育发展的步伐，从而培养出更多具有国际水准的计算机专业人才，增强我国计算机产业的核心竞争力。为此，我们从国外多家知名的出版机构 Pearson、McGraw-Hill、John Wiley & Sons、Springer、Cengage Learning 等精选、引进了这套“国外计算机科学经典教材”。

作为世界级的图书出版机构，Pearson、McGraw-Hill、John Wiley & Sons、Springer、Cengage Learning 通过与世界级的计算机教育大师携手，每年都為全球的计算机高等教育奉獻大量的优秀教材。清华大学出版社和这些世界知名的出版机构长期保持着紧密友好的合作关系，这次引进的“国外计算机科学经典教材”便全是出自上述这些出版机构。同时，为了组织该套教材的出版，我们在国内聘请了一批知名的专家和教授，成立了专门的教材编审委员会。

教材编审委员会的运作从教材的选题阶段即开始启动，各位委员根据国内外高等院校计算机科学及相关专业的现有课程体系，并结合各个专业的培养方向，从上述这些出版机构出版的计算机系列教材中精心挑选针对性强的题材，以保证该套教材的优秀性和领先性，避免出现“低质重复引进”或“高质消化不良”的现象。

为了保证出版质量，我们为这套教材配备了一批经验丰富的编辑、排版、校对人员，制定了更加严格的出版流程。本套教材的译者，全部由对应专业的高校教师或拥有相关经验的 IT 专家担任。每本教材的责编在翻译伊始，就定期不间断地与该书的译者进行交流与反馈。为了尽可能地保留与发扬教材原著的精华，在经过翻译、排版和传统的三审三校之后，我们还请编审委员或相关的专家教授对文稿进行审读，以最大程度地弥补和修正在前面一系列加工过程中对教材造成的误差和瑕疵。

由于时间紧迫和受全体制作人员自身能力所限，该套教材在出版过程中很可能还存在一些遗憾，欢迎广大师生来电来信批评指正。同时，也欢迎读者朋友积极向我们推荐各类优秀的国外计算机教材，共同为我国高等院校计算机教育事业贡献力量。

清华大学出版社

# 审校者序

数据挖掘，作为商业智能(Business Intelligence)实现的最深层次，在商业智能解决方案的体系中占据着重要的位置。

数据库中存在着的是数据，对于业务人员来说，只是一些无法看懂的天书，他们需要的是可以更容易理解的信息。那么，我们以前如何解决这个矛盾的呢？一般的答案是报表系统。简单地说，业务人员看到的是美观的界面，便捷的操作，鼠标点击后，报表系统生成 SQL 语句，数据库服务器收到以后，返回所需要的信息。不错，报表系统已经可以称作是 BI 了，它是 BI 的低端实现。

现在国外的企业，大部分已经进入了更深层次商业智能，叫做数据分析，即基于多维数据库的在线分析系统(OLAP)。还有一些企业已经开始进入更深层次商业智能，叫做数据挖掘。从广义上说，任何从数据库中挖掘信息的过程都叫做数据挖掘。从这点看来，报表、多维分析和深度的挖掘都是挖掘数据的手段。但是，从技术术语上说，数据挖掘(Data Mining)特指的是：源数据经过清洗和转换等成为适合于挖掘的数据集，然后建立特定的挖掘模型，利用这些数据集训练模型，最后利用模型找出的知识模式进行预测，从而辅助决策工作。

过去，谈起数据挖掘，大家想到的往往是那些专业数学家、统计学家，一般的技术人员和业务人员望而却步。现在，随着 IT 技术的发展和工业化，SQL Server 2008 提炼了上述的各种复杂知识，加工成友好的视窗工具，嵌入到分析服务(Analysis Services)中，使得数据挖掘的用户扩展到了大量的开发者人群，甚至是经过培训的业务人员。它使我们的员工和程序变得更聪明了。

我在国外学习、工作了多年之后，深感商业智能即将成为未来几年 IT 领域的核心价值，因此从 2003 开始创建了北京迈思奇科技有限公司，致力于将国外的先进 BI 技术和工具引进国内，帮助国内的企业提高数据分析效率、增强竞争实力。公司成立六年来，与微软密切合作，通过数百次讲座和培训，为企业培养 BI 专业人员；同时，在承担 BI 项目开发的过程中，公司也积累了优秀的团队和丰富的项目案例，创立了国内一流的数据挖掘品牌。

本书作者唐朝晖(ZhaoHui Tang)博士是微软公司 SQL Server 数据挖掘产品的产品经

理，唐博士具有数据挖掘技术的深厚理论根基，同时对该产品有丰富的第一手资料。几年前，唐博士曾著《数据挖掘原理与应用——SQL Server 2005 数据库》一书，获得了很高的评价。本书虽然是新版本的升级版，我们在审校的过程中，依然感到有很多新的收获。

感谢清华大学出版社，及时为我们引进了这部优秀的教材；也感谢本书的译者，准确而清晰地传达了原著的精华。

杨大川  
北京迈思奇科技有限公司

# 序

这个世界到处都是数据。金融事务、在线广告分析、顾客喜好信息和科学发现的结果都意味着目前有海量的数据存在于结构化和非结构化的库中，还增长得比以往任何时候都快，这是因为得到了技术的支持，而且现代人在生活的各个方面都采用和集成了技术。

商业智能从业者力争使他们掌管的数据有益于他们的工作，更好地理解影响结果的因素。这种趋势一直在演化，而且变化得比以往任何时候都快，仅查看历史数据不再能推断出过去发生过什么。通过数据挖掘，可以更方便地理解事情发生的原因，找出(好的或坏的)历史是否会重演的区别，因为现在趋势的变化非常快，所以识别出趋势、找出异常的地方、预测未来的自动分析和专业算法也在获胜和仅参加竞争之间有了区别。数据挖掘可以对常见的列进行自动分类和聚类，找出数据中的趋势和异常，根据这些因素来预测，使海量的数据有了实际的意义。

很高兴可以与 Jamie MacLennan 和 Bogdan Crivat 一起工作和学习。他们热衷于研究技术为我们的生活所带来的变化，并提供必要的工具，使不断增加的数据对每个人都有一定的意义。本书清楚地解释了数据挖掘概念，论述了如何在常见的情形下使用 Microsoft SQL Server 中的算法和工具来应用它们。本书还提供了直接学习源代码的机会，我确信读者会认为本书是一本极具价值的参考资料。

Tom Casey

微软公司 SQL Server 商业智能部总经理

# 前言

Microsoft SQL Server 2008 是集成了数据挖掘技术的第 3 版 SQL Server。自从在 SQL Server 2000 中引入数据挖掘以来，它就成为这个大型产品的重要部分。数据挖掘从带有两个算法的 SQL Server Analysis Services 的一个独立部分中诞生，并成长为 SQL Server Business Intelligence(BI)平台的一个固有部分，与 OLAP、Integration Services 和 Reporting Services 完全集成起来。其他 Microsoft 应用程序，例如 Microsoft Dynamics CRM 和 Microsoft Performance Point Server，也与 SQL Server Data Mining 无缝地集成起来，重点使用其预测功能。

SQL Server Data Mining 是业界部署最广泛的数据挖掘服务器，许多第三方软件和咨询公司也在建立、专业化和扩展该平台。企事业单位、中小企业甚至学术人员和科学家也都采用或转而使用 SQL Server Data Mining，因为它的可伸缩性很大，容易获得，功能可以扩展，使用起来也很简单。

本书是 SQL Server Data Mining 的一个指南，解释了其工作原理，提供了 SQL Server Data Mining 的详细技术和实践论述，论证了为什么应部署和使用 SQL Server Data Mining。

---

## 本书组织结构

本书介绍了 SQL Server Data Mining 的所有方面，包括使用 SQL Server 实现成功的数据挖掘解决方案的所有必要知识，带读者熟悉其中的工具，给出了 Microsoft 数据挖掘算法的深度和广度，然后详细探讨了实现数据挖掘解决方案的各种方式。

本书的前半部分概述了利用 SQL Server Data Mining 所需的工具、技术和理念，详细描述了每个 SQL Server 数据挖掘算法，然后讨论如何把 SQL Server Data Mining 集成到 SQL Server BI 套件的其他部分中。本书的后半部分着重处理体系结构和编程问题，并举例说明了一些数据挖掘实现方案。

下面对各个章节进行简要介绍。

- 第 1 章不仅概述了本书的内容，还对这项技术进行了简要介绍。其中包含“数据挖掘”的详细定义，讨论了这项技术可以解决的问题种类。
- 第 2 章概述了 Table Analysis Tools for Office 2007 插件，这是任何信息工作者都可以使用的、用于 Excel 的一组工具。本章解释了使用这些工具的原因和方法，提供了获得最佳结果的指南。

- 第3章对于理解 SQL Server Data Mining 平台非常重要,它解释了考虑数据挖掘问题的底层概念,为 Data Mining Extensions (DMX) to SQL 提供了一个通过示例来学习的框架。
- 第4章介绍了如何使用 Business Intelligence Development Studio (BI Dev Studio)建立数据挖掘解决方案。除了基本概述之外,还提供了许多提示和技巧,以区别成功和失败的项目。本章还讨论了如何使用 SQL Server Management Studio 访问和保护数据挖掘对象。另外,还说明了如何通过 SQL Server Reporting Services 展示数据挖掘模型。
- 第5章探讨了 Data Mining Add-ins for Office 2007 中的其他工具。如该章所述,这些工具提供的功能比 BI Dev Studio 和 SQL Server Management Studio 多,但没有提供 SQL Server Data Mining 的全部功能。尽管如此,本章仍允许充分利用 Microsoft Office 工具进行数据挖掘。
- 第6~12章都介绍了 SQL Server Data Mining 中的一个或多个算法。在每一章中,都包含算法的基本描述,之后的用法有助于理解应用每个算法的方式、时间和场合。每一章都描述了如何使用特定的算法创建、训练、解释和应用模型,还包括这些算法的工作原理的深入技术讨论。
- 第13章简要介绍了 Online Analytical Processing (OLAP)和 SQL Server Analysis Services 的 OLAP 功能,解释了在 OLAP 立方体上进行数据挖掘的方式和时间,还详细说明了如何实现常见的 OLAP 挖掘场景。
- 第14章介绍了 SQL Server Integration Services (SSIS),描述了其各种组件,接着详细论述了在数据集成软件包中实现数据挖掘解决方案所使用的任务和转换。本章还描述了如何使用文本挖掘组件,为数据挖掘场景准备未结构化的数据。
- 第15章是从工具和概念转向 SQL Server Data Mining 编程和管理方面的第一章。本章讨论了基于服务器的数据挖掘系统的体系结构,包括作为所有客户-服务器通信基础的 XML for Analysis (XMLA)协议。本章还描述了数据挖掘服务器的管理,包括对 SQL Server Data Mining 和数据挖掘安全角色非常重要的服务器属性。
- 第16章详细介绍了 SQL Server Data Mining 的编程接口,包括编程创建、训练和应用数据挖掘对象的几个例子。
- 第17章描述了如何用自己的功能扩展 SQL Server Data Mining,说明了如何创建存储过程,给 DMX 添加操作,还论述了如何实现自己的数据挖掘算法,插入到 SQL Server Data Mining 中,并利用其功能。另外,本章还探讨了如何编写自己的数据挖掘可视化界面,来显示内置算法或自己的算法中的模式,把它们嵌入到 BI Dev Studio 和 SQL Server Management Studio 中。
- 第18章介绍了一个常见的数据挖掘场景:实现一个推荐信息引擎,把它集成到一个零售网站上。其中包含开始这个应用程序的示例查询和代码。
- 第19章除了总结本书之外,本章还提供了一个有价值的链接列表,其中包含数据挖掘项目的其他信息和帮助。本章包括其他一些参考资料,如果需要学习数据挖掘的更多知识,就可以阅读这些参考资料。

本书还有两个附录。

- 附录 A 简要描述了本书使用的各种数据集。
- 附录 B 列出了所有支持的 DMX 函数,还包含可以从 DMX 调用的所有 Visual Basic for Applications (VBA)和 Excel 函数,描述了一些附加的存储过程,以帮助理解本书中的示例查询。

---

## 本书读者对象

本书主要用于对数据挖掘有兴趣的 SQL Server 用户。有 SQL 工作经验非常有助于理解本书中的 DMX 和 DMX 查询。但是,非 SQL 用户仍可以从 Office 2007 和算法章节中获益。对 SQL Server Data Mining 编程感兴趣的读者应理解.NET 和 C#语言在相关章节中的应用。

对于阅读过本书前一版本《数据挖掘原理与应用——SQL Server 2005 数据库》(清华大学出版社引进并出版)的读者,欢迎回来!本书包含 Microsoft SQL Server 2008 Data Mining 新功能的完整介绍,还添加了描述大多数算法和场景的新示例。

---

## 需要的工具

为了更好地掌握本书的内容,需要具有 SQL Server 2008 Analysis Services 软件的访问权限。SQL Server 2008 Analysis Services 包含在 Microsoft SQL Server 2008 的标准版、企业版 and 开发版中。有时间限制的体验版可以从 <http://www.microsoft.com/sql> 上下载。要学习第 2 和第 5 章,还需要 Microsoft Office 2007 和 SQL Server 2008 Data Mining Add-Ins for Office 2007。Microsoft Office 2007 体验版可以从 [www.microsoft.com/office](http://www.microsoft.com/office) 上获得, Data Mining Add-Ins 的免费版本可以从 [www.microsoft.com/sql/dm](http://www.microsoft.com/sql/dm) 上获得。

此外,还需要安装 AdventureWorksDW2008 数据库。访问这个数据库的指导方法在网站 [http://www.wiley.com/go/data\\_mining\\_SQL\\_2008](http://www.wiley.com/go/data_mining_SQL_2008) 的 readme 文件中。

---

## 网站上的内容

读者可以从 [www.wiley.com/go/data\\_mining\\_SQL\\_2008](http://www.wiley.com/go/data_mining_SQL_2008) 和 <http://www.tupwk.com.cn> 网站下载本书的相关资料。网站中包含 SQL Server 数据库备份、SQL Server Analysis Services 数据库备份、项目文件、DMX 查询文件和/或源代码。每个章节的目录下都包含一个 readme 文件,它描述了如何使用该章的下载文件。

本书开启了 SQL Server Data Mining 世界的大门,消化吸收了本书的所有信息之后,就可以给日常生活添加预测和描述性的分析了。Microsoft SQL Server Data Mining 的强大开发环境和 API 可以改变每个用户查看数据和与数据交互的方式。赶快从积攒多年的数据中找出隐藏的、有价值的信息——这一定是一种很新鲜的体验!

# 目 录

## 第 1 章 SQL Server 2008 数据

### 挖掘介绍..... 1

#### 1.1 数据挖掘解决的商业问题..... 3

#### 1.2 数据挖掘的任务..... 4

##### 1.2.1 分类..... 4

##### 1.2.2 聚类..... 5

##### 1.2.3 关联..... 5

##### 1.2.4 回归..... 6

##### 1.2.5 预测..... 6

##### 1.2.6 序列分析..... 7

##### 1.2.7 偏差分析..... 7

#### 1.3 数据挖掘项目的生命周期..... 8

##### 1.3.1 商业问题的形成..... 8

##### 1.3.2 数据收集..... 8

##### 1.3.3 数据清理和转换..... 8

##### 1.3.4 模型构建..... 9

##### 1.3.5 模型评估..... 10

##### 1.3.6 报告和预测..... 10

##### 1.3.7 应用集成..... 10

##### 1.3.8 模型管理..... 10

#### 1.4 本章小结..... 11

## 第 2 章 用 Microsoft Excel 2007

### 进行数据挖掘..... 13

#### 2.1 表分析工具的安装..... 14

##### 2.1.1 配置具有管理权限的 Analysis Services..... 14

##### 2.1.2 配置没有管理权限的 Analysis Services..... 15

##### 2.1.3 使用该插件的要求..... 15

##### 2.1.4 寻求帮助..... 18

#### 2.2 分析关键影响因素工具..... 18

##### 2.2.1 主影响因素报表..... 20

##### 2.2.2 区别报表..... 21

##### 2.2.3 分析关键影响因素 任务小结..... 23

#### 2.3 检测类别工具..... 23

##### 2.3.1 启动工具..... 23

##### 2.3.2 类别报表..... 24

##### 2.3.3 检测类别工具小结..... 27

#### 2.4 从示例填充工具..... 28

##### 2.4.1 运行工具, 解释结果..... 28

##### 2.4.2 细调结果..... 30

##### 2.4.3 从示例填充工具小结..... 31

#### 2.5 预测工具..... 31

##### 2.5.1 启动工具, 指定选项..... 32

##### 2.5.2 解释结果..... 33

##### 2.5.3 预测工具小结..... 35

#### 2.6 突出显示异常值工具..... 35

##### 2.6.1 使用工具..... 35

##### 2.6.2 更复杂的交互操作..... 37

##### 2.6.3 限制和故障排除..... 39

##### 2.6.4 突出显示异常值工具小结..... 39

#### 2.7 应用场景分析工具..... 40

##### 2.7.1 目标查找工具..... 41

##### 2.7.2 对数字目标使用目标 查找工具..... 43

##### 2.7.3 对整个表使用目标查找工具..... 43

##### 2.7.4 假设工具..... 44

##### 2.7.5 对整个表使用假设工具..... 46

##### 2.7.6 应用场景分析工具小结..... 47

|                         |                                                 |
|-------------------------|-------------------------------------------------|
| 2.8 预测计算器工具..... 48     | 第4章 使用 SQL Server 进行<br>数据挖掘.....99             |
| 2.8.1 运行工具..... 49      | 4.1 BI Dev Studio 介绍..... 99                    |
| 2.8.2 细调结果..... 52      | 4.1.1 用户界面.....100                              |
| 2.8.3 使用结果..... 56      | 4.1.2 脱机模式和即时模式.....102                         |
| 2.8.4 预测计算器工具小结..... 56 | 4.1.3 创建数据挖掘对象.....105                          |
| 2.9 购物篮分析工具..... 56     | 4.2 设置数据源..... 105                              |
| 2.9.1 使用工具..... 57      | 4.2.1 数据源.....106                               |
| 2.9.2 捆绑销售商品报表..... 58  | 4.2.2 使用数据源视图.....107                           |
| 2.9.3 购物篮推荐报表..... 59   | 4.3 创建和编辑模型..... 116                            |
| 2.9.4 调整工具..... 60      | 4.3.1 结构和模型.....116                             |
| 2.9.5 购物篮分析工具小结..... 61 | 4.3.2 使用数据挖掘向导.....116                          |
| 2.10 表分析工具的技术概述..... 61 | 4.3.3 创建 MovieClick 挖掘结构<br>和挖掘模型.....122       |
| 2.11 本章小结..... 62       | 4.3.4 使用数据挖掘设计器.....123                         |
| 第3章 数据挖掘概念与 DMX..... 63 | 4.4 处理..... 129                                 |
| 3.1 DMX 的发展历史..... 63   | 4.5 使用模型..... 130                               |
| 3.2 为什么使用 DMX..... 64   | 4.5.1 了解模型查看器.....130                           |
| 3.3 数据挖掘过程..... 64      | 4.5.2 使用挖掘准确性图表.....132                         |
| 3.4 关键概念..... 65        | 4.5.3 为 MovieClick 模型创建<br>一个提升图.....136        |
| 3.4.1 属性..... 65        | 4.5.4 使用交叉验证.....137                            |
| 3.4.2 状态..... 66        | 4.5.5 使用挖掘模型预测.....139                          |
| 3.4.3 事例..... 66        | 4.5.6 针对 MovieClick 模型<br>执行查询.....140          |
| 3.4.4 键..... 68         | 4.5.7 创建数据挖掘报表.....140                          |
| 3.4.5 输入和输出..... 69     | 4.6 使用 SQL Server Management<br>Studio..... 142 |
| 3.5 DMX 对象..... 70      | 4.6.1 了解 Management Studio<br>用户界面.....143      |
| 3.5.1 挖掘结构..... 71      | 4.6.2 使用服务器资源管理器.....143                        |
| 3.5.2 挖掘模型..... 71      | 4.6.3 使用对象资源管理器.....144                         |
| 3.6 DMX 查询语法..... 72    | 4.6.4 使用查询编辑器.....144                           |
| 3.6.1 创建挖掘结构..... 72    | 4.7 本章小结..... 145                               |
| 3.6.2 创建挖掘模型..... 75    | 第5章 使用 Office 2007 执行数据<br>挖掘过程.....147         |
| 3.6.3 填充挖掘结构..... 83    | 5.1 数据挖掘客户端简介..... 147                          |
| 3.7 预测..... 88          | 5.2 用数据挖掘客户端导入数据..... 149                       |
| 3.7.1 预测连接..... 88      |                                                 |
| 3.7.2 预测查询语法..... 89    |                                                 |
| 3.7.3 预测函数..... 92      |                                                 |
| 3.7.4 嵌套表上的预测..... 94   |                                                 |
| 3.7.5 预测嵌套值列..... 95    |                                                 |
| 3.8 本章小结..... 96        |                                                 |

|                                          |            |                                                |            |
|------------------------------------------|------------|------------------------------------------------|------------|
| 5.3 数据浏览和准备 .....                        | 150        | 6.4.2 MAXIMUM_OUTPUT<br>_ATTRIBUTES .....      | 184        |
| 5.3.1 用浏览数据工具离散数据 .....                  | 150        | 6.4.3 MAXIMUN_STATES .....                     | 184        |
| 5.3.2 砍掉长长的尾巴 .....                      | 150        | 6.4.4 MINIMUM_DEPENDENCY<br>_PROBABILITY ..... | 184        |
| 5.3.3 合并多个含义 .....                       | 151        | 6.5 本章小结 .....                                 | 185        |
| 5.3.4 去除不合逻辑的值 .....                     | 153        |                                                |            |
| 5.3.5 配平数据 .....                         | 153        |                                                |            |
| 5.4 建模 .....                             | 154        | <b>第 7 章 Microsoft 决策树算法 .....</b>             | <b>187</b> |
| 5.4.1 基于任务的建模 .....                      | 155        | 7.1 决策树算法介绍 .....                              | 187        |
| 5.4.2 数据挖掘客户端中的<br>高级建模 .....            | 157        | 7.2 使用决策树算法 .....                              | 188        |
| 5.5 准确性和验证 .....                         | 159        | 7.2.1 创建决策树模型 .....                            | 188        |
| 5.6 模型用法 .....                           | 160        | 7.2.2 DMX 查询 .....                             | 188        |
| 5.6.1 浏览模型 .....                         | 160        | 7.2.3 模型内容 .....                               | 194        |
| 5.6.2 用 Visio 查看模型 .....                 | 161        | 7.2.4 解释模型 .....                               | 195        |
| 5.6.3 查询模型 .....                         | 163        | 7.3 决策树算法的基本原理 .....                           | 197        |
| 5.6.4 查询向导 .....                         | 163        | 7.3.1 决策树生成的基本思想 .....                         | 197        |
| 5.7 数据挖掘单元格函数 .....                      | 165        | 7.3.2 处理变量中的多个状态 .....                         | 200        |
| 5.7.1 DMPREDICT 函数 .....                 | 165        | 7.3.3 避免过度训练 .....                             | 201        |
| 5.7.2 DMPREDICTTABLEROW<br>函数 .....      | 166        | 7.3.4 结合先验知识 .....                             | 201        |
| 5.7.3 DMCONTENTQUERY<br>函数 .....         | 166        | 7.3.5 特征选择 .....                               | 201        |
| 5.8 管理模型 .....                           | 167        | 7.3.6 使用连续的输入属性 .....                          | 202        |
| 5.9 跟踪 .....                             | 167        | 7.3.7 回归 .....                                 | 202        |
| 5.10 本章小结 .....                          | 167        | 7.3.8 使用 Microsoft 决策树<br>算法进行关联分析 .....       | 203        |
| <b>第 6 章 Microsoft 贝叶斯算法 .....</b>       | <b>169</b> | 7.4 算法参数 .....                                 | 204        |
| 6.1 贝叶斯算法介绍 .....                        | 169        | 7.4.1 Complexity_Penalty .....                 | 204        |
| 6.2 使用贝叶斯算法 .....                        | 170        | 7.4.2 Minimum_Support .....                    | 204        |
| 6.2.1 创建预测模型 .....                       | 171        | 7.4.3 Score_Method .....                       | 205        |
| 6.2.2 数据浏览 .....                         | 172        | 7.4.4 Split_Method .....                       | 205        |
| 6.2.3 关键影响因子的分析 .....                    | 173        | 7.4.5 Maximum_Input_Attribute .....            | 205        |
| 6.2.4 文档分类 .....                         | 173        | 7.4.6 Maximum_Output_Attribute .....           | 205        |
| 6.2.5 DMX .....                          | 175        | 7.4.7 Force_Regressor .....                    | 206        |
| 6.2.6 理解贝叶斯模型的内容 .....                   | 175        | 7.5 存储过程 .....                                 | 206        |
| 6.2.7 浏览贝叶斯模型 .....                      | 177        | 7.6 本章小结 .....                                 | 207        |
| 6.3 理解贝叶斯算法的基本原理 .....                   | 180        |                                                |            |
| 6.4 贝叶斯算法的参数 .....                       | 184        | <b>第 8 章 Microsoft 时序算法 .....</b>              | <b>209</b> |
| 6.4.1 MAXIMUM_INPUT_<br>ATTRIBUTES ..... | 184        | 8.1 Microsoft 时序算法介绍 .....                     | 209        |
|                                          |            | 8.2 用法 .....                                   | 210        |
|                                          |            | 8.3 DMX .....                                  | 214        |
|                                          |            | 8.3.1 模型的创建 .....                              | 214        |

|       |                            |     |        |                          |     |
|-------|----------------------------|-----|--------|--------------------------|-----|
| 8.3.2 | 模型的处理                      | 216 | 9.3.4  | 聚类预测                     | 251 |
| 8.3.3 | 预测                         | 218 | 9.4    | 聚类算法的参数                  | 251 |
| 8.3.4 | 钻取功能                       | 222 | 9.4.1  | Clustering_Method        | 251 |
| 8.4   | Microsoft 时序算法的            |     | 9.4.2  | Cluster_Count            | 252 |
|       | 基本原理                       | 222 | 9.4.3  | Minimum_Cluster_Cases    | 252 |
| 8.4.1 | 自动回归                       | 223 | 9.4.4  | Modelling_Cardinality    | 253 |
| 8.4.2 | 周期                         | 224 | 9.4.5  | Stopping_Tolerance       | 253 |
| 8.4.3 | 自动回归树                      | 224 | 9.4.6  | Sample_Size              | 254 |
| 8.4.4 | 预测                         | 225 | 9.4.7  | Cluster_Seed             | 254 |
| 8.5   | 参数                         | 226 | 9.4.8  | Maximum_Input_Attributes | 254 |
| 8.5.1 | Missing_Value_Substitution | 226 | 9.4.9  | Maximum_States           | 254 |
| 8.5.2 | Periodicity_Hint           | 227 | 9.5    | 本章小结                     | 254 |
| 8.5.3 | Auto_Detect_Periodicity    | 227 |        |                          |     |
| 8.5.4 | Minimum_Series_Value 和     |     | 第 10 章 | Microsoft 序列聚类算法         | 257 |
|       | Maximum_Series_Value       | 227 | 10.1   | Microsoft 序列聚类           |     |
| 8.5.5 | Forecast_Method            | 227 |        | 算法介绍                     | 257 |
| 8.5.6 | Prediction_Smoothing       | 227 | 10.2   | 使用 Microsoft 序列聚类        |     |
| 8.5.7 | Instability_Sensitivity    | 228 |        | 算法                       | 258 |
| 8.5.8 | Historic_Model_Count       |     | 10.2.1 | 创建序列聚类模型                 | 258 |
|       | 和 Historic_Model_Gap       | 228 | 10.2.2 | DMX 查询                   | 259 |
| 8.5.9 | Complexity_Penalty         |     | 10.2.3 | 解释模型                     | 266 |
|       | 和 Minimum_Support          | 229 | 10.3   | Microsoft 序列聚类算法的        |     |
| 8.6   | 模型的内容                      | 229 |        | 基本原理                     | 270 |
| 8.7   | 本章小结                       | 230 | 10.3.1 | 什么是马尔可夫链                 | 270 |
| 第 9 章 | Microsoft 聚类算法             | 233 | 10.3.2 | 马尔可夫链的阶                  | 271 |
| 9.1   | Microsoft 聚类算法介绍           | 234 | 10.3.3 | 状态转移矩阵                   | 272 |
| 9.2   | 使用聚类模型                     | 235 | 10.3.4 | 使用马尔可夫链来                 |     |
| 9.2.1 | 进行聚类                       | 236 |        | 进行聚类                     | 273 |
| 9.2.2 | 将聚类作为一个分析步骤                | 237 | 10.3.5 | 聚类分解                     | 274 |
| 9.2.3 | 利用聚类检查异常                   | 238 | 10.4   | 模型内容                     | 274 |
| 9.2.4 | DMX                        | 240 | 10.5   | 序列聚类算法的参数                | 275 |
| 9.2.5 | 模型内容                       | 243 | 10.5.1 | Cluster_Count            | 275 |
| 9.2.6 | 理解聚类模型                     | 244 | 10.5.2 | Minimum_Support          | 275 |
| 9.3   | 聚类算法的基本原理                  | 248 | 10.5.3 | Maximum_States           | 275 |
| 9.3.1 | 硬聚类算法与软聚类算法                | 248 | 10.5.4 | Maximum_Sequence         |     |
| 9.3.2 | 离散聚类                       | 249 |        | _States                  | 275 |
| 9.3.3 | 可伸缩聚类                      | 250 | 10.6   | 本章小结                     | 276 |

|                                      |     |                                 |     |
|--------------------------------------|-----|---------------------------------|-----|
| <b>第 11 章 Microsoft 关联规则算法</b> ..... | 277 | 12.2.3 DMX 查询                   | 306 |
| 11.1 Microsoft 关联规则                  |     | 12.3 模型内容                       | 309 |
| 算法介绍                                 | 278 | 12.4 解释模型                       | 310 |
| 11.2 使用关联规则算法                        | 278 | 12.5 Microsoft 神经网络算法的          |     |
| 11.2.1 数据研究模型                        | 279 | 基本原理                            | 312 |
| 11.2.2 一个简单的推荐引擎                     | 280 | 12.5.1 什么是神经网络                  | 312 |
| 11.2.3 高级交叉销售的分析                     | 282 | 12.5.2 组合和激活                    | 314 |
| 11.2.4 DMX                           | 283 | 12.5.3 反向传播、误差函数和               |     |
| 11.2.5 模型内容                          | 287 | 共轭梯度                            | 315 |
| 11.2.6 解释模型                          | 288 | 12.5.4 处理神经网络的                  |     |
| 11.3 关联规则算法的基本原理                     | 290 | 简单示例                            | 316 |
| 11.3.1 理解关联规则算法的                     |     | 12.5.5 规范化和映射                   | 317 |
| 基本概念                                 | 290 | 12.5.6 网络拓扑                     | 319 |
| 11.3.2 挖掘频繁项集                        | 293 | 12.5.7 训练终止条件                   | 319 |
| 11.3.3 生成关联规则                        | 295 | 12.6 非线性可分类                     | 320 |
| 11.3.4 预测                            | 296 | 12.7 神经网络算法的参数                  | 321 |
| 11.4 关联算法的参数                         | 297 | 12.7.1 Maximum_Input_           |     |
| 11.4.1 Minimum_Support               | 297 | Attributes                      | 321 |
| 11.4.2 Maximum_Support               | 297 | 12.7.2 Maximum_Output_          |     |
| 11.4.3 Minimum_Probability           | 297 | Attributes                      | 321 |
| 11.4.4 Minimum_Importance            | 297 | 12.7.3 Maximum_States           | 321 |
| 11.4.5 Maximum_Itemset_Size          | 298 | 12.7.4 Holdout_Percentage       | 321 |
| 11.4.6 Minimum_Itemset_Size          | 298 | 12.7.5 Holdout_Seed             | 321 |
| 11.4.7 Maximum_Itemset               |     | 12.7.6 Hidden_Node_Ratio        | 321 |
| _Count                               | 298 | 12.7.7 Sample_Size              | 322 |
| 11.4.8 Optimized_Prediction_         |     | 12.8 本章小结                       | 322 |
| Count                                | 298 |                                 |     |
| 11.4.9 Autodetect_Minimum            |     | <b>第 13 章 挖掘 OLAP 立方体</b> ..... | 323 |
| _Support                             | 298 | 13.1 OLAP 介绍                    | 324 |
| 11.5 本章小结                            | 298 | 13.1.1 理解星型模式和                  |     |
|                                      |     | 雪花模式                            | 325 |
| <b>第 12 章 Microsoft 神经网络算法和</b>      |     | 13.1.2 理解维和层次                   | 326 |
| <b>逻辑回归算法</b> .....                  | 301 | 13.1.3 理解度量和度量组                 | 326 |
| 12.1 相同的基本原理，                        |     | 13.1.4 理解立方体的处理                 |     |
| 两个算法                                 | 301 | 和存储                             | 327 |
| 12.2 使用 Microsoft 神经网络               |     | 13.1.5 使用前摄缓存                   | 328 |
| 算法                                   | 302 | 13.1.6 查询立方体                    | 328 |
| 12.2.1 文本分类模型                        | 302 | 13.2 执行计算                       | 329 |
| 12.2.2 实用模型                          | 306 | 13.3 浏览立方体                      | 330 |

|        |                              |     |        |                                            |     |
|--------|------------------------------|-----|--------|--------------------------------------------|-----|
| 13.4   | 理解统一维度模型                     | 331 | 15.5   | 数据挖掘管理                                     | 395 |
| 13.5   | 理解 OLAP 和数据挖掘<br>之间的关系       | 334 | 15.5.1 | 服务器配置                                      | 395 |
| 13.5.1 | 挖掘聚集的数据                      | 335 | 15.5.2 | 数据挖掘安全                                     | 397 |
| 13.5.2 | OLAP 模式发现的需求                 | 335 | 15.5.3 | 创建和训练挖掘对象<br>的安全需求                         | 398 |
| 13.5.3 | OLAP 挖掘与关系挖掘                 | 336 | 15.5.4 | 各种部署场景的安全性                                 | 399 |
| 13.6   | 使用向导和编辑器来<br>构建 OLAP 挖掘模型    | 336 | 15.6   | 本章小结                                       | 402 |
| 13.6.1 | 使用数据挖掘向导                     | 337 | 第 16 章 | SQL Server 数据挖掘编程                          | 403 |
| 13.6.2 | 使用数据挖掘设计器                    | 345 | 16.1   | 数据挖掘 API                                   | 404 |
| 13.7   | 理解数据挖掘维                      | 346 | 16.1.1 | ADO                                        | 405 |
| 13.8   | 在 DMX 查询内部使用<br>MDX          | 348 | 16.1.2 | ADO.NET                                    | 405 |
| 13.9   | 将 AMO 用于 OLAP<br>挖掘模型        | 350 | 16.1.3 | ADOMD.NET                                  | 405 |
| 13.10  | 本章小结                         | 354 | 16.1.4 | Server ADOMD                               | 406 |
| 第 14 章 | SQL Server 集成服务<br>数据挖掘      | 355 | 16.1.5 | AMO                                        | 406 |
| 14.1   | SSIS 介绍                      | 356 | 16.2   | 使用 Analysis Services 的<br>API              | 406 |
| 14.1.1 | 理解 SSIS 包                    | 357 | 16.3   | 使用 Microsoft.AnalysisServices<br>创建和管理挖掘模型 | 407 |
| 14.1.2 | 任务流                          | 357 | 16.3.1 | AMO 的基本原理                                  | 408 |
| 14.1.3 | 数据流                          | 359 | 16.3.2 | AMO 应用程序和安全                                | 409 |
| 14.2   | 在 SSIS 环境中进行<br>数据挖掘         | 361 | 16.3.3 | 对象的创建                                      | 410 |
| 14.2.1 | 数据挖掘任务                       | 362 | 16.4   | 浏览和查询挖掘模型                                  | 420 |
| 14.2.2 | 数据挖掘转换                       | 368 | 16.4.1 | 使用 ADOMD.NET<br>来预测                        | 420 |
| 14.2.3 | 文本挖掘转换                       | 375 | 16.4.2 | ADOMD.NET 中的<br>表值参数                       | 425 |
| 14.3   | 本章小结                         | 383 | 16.4.3 | 浏览模型                                       | 427 |
| 第 15 章 | SQL Server 数据挖掘的<br>体系结构     | 385 | 16.5   | 存储过程                                       | 429 |
| 15.1   | Analysis Services 体系<br>结构介绍 | 385 | 16.5.1 | 编写存储过程                                     | 430 |
| 15.2   | XML for Analysis             | 387 | 16.5.2 | 存储过程示例                                     | 432 |
| 15.2.1 | XMLA 的 API                   | 387 | 16.5.3 | 在存储过程内部<br>执行查询                            | 434 |
| 15.2.2 | XMLA 和 Analysis<br>Services  | 390 | 16.5.4 | 从存储过程中返回<br>数据集                            | 435 |
| 15.3   | 处理体系结构                       | 391 | 16.5.5 | 部署和调试存储过程<br>程序集                           | 438 |
| 15.4   | 预测                           | 393 | 16.6   | 本章小结                                       | 439 |

|                                              |     |
|----------------------------------------------|-----|
| <b>第 17 章 扩展 SQL Server</b>                  |     |
| <b>数据挖掘</b> .....                            | 441 |
| 17.1 理解插件算法 .....                            | 441 |
| 17.1.1 插件算法的架构 .....                         | 442 |
| 17.1.2 插件算法实例的<br>生命周期 .....                 | 443 |
| 17.1.3 插件算法的概念 .....                         | 444 |
| 17.1.4 模型的创建和处理 .....                        | 447 |
| 17.1.5 预测 .....                              | 450 |
| 17.1.6 内容导航 .....                            | 451 |
| 17.1.7 自定义函数 .....                           | 451 |
| 17.1.8 PMML .....                            | 453 |
| 17.1.9 受托管的插件和<br>内部插件 .....                 | 453 |
| 17.1.10 安装插件算法 .....                         | 453 |
| 17.1.11 插件算法的更多信息 .....                      | 454 |
| 17.2 使用数据挖掘查看器 .....                         | 454 |
| 17.2.1 要实现的接口 .....                          | 454 |
| 17.2.2 显示信息 .....                            | 455 |
| 17.2.3 从 Analysis Services<br>中检索信息 .....    | 456 |
| 17.2.4 注册查看器 .....                           | 456 |
| 17.2.5 插件查看器的<br>更多信息 .....                  | 457 |
| 17.3 本章小结 .....                              | 457 |
| <b>第 18 章 实现 Web 交叉销售<br/>    应用程序</b> ..... | 459 |
| 18.1 源数据描述 .....                             | 459 |
| 18.2 构建模型 .....                              | 460 |
| 18.2.1 确定数据挖掘任务 .....                        | 460 |
| 18.2.2 将决策树算法应用于<br>关联任务 .....               | 460 |
| 18.2.3 使用关联规则算法 .....                        | 462 |
| 18.2.4 两个模型的比较 .....                         | 463 |
| 18.3 执行预测 .....                              | 465 |
| 18.3.1 批量预测查询 .....                          | 465 |
| 18.3.2 使用单例预测查询 .....                        | 466 |
| 18.4 在 Web 应用程序中集成<br>预测功能 .....             | 467 |
| 18.4.1 理解 Web 应用程序的<br>体系结构 .....            | 467 |
| 18.4.2 设置权限 .....                            | 467 |
| 18.4.3 分析 Web 推荐应用<br>程序的样例代码 .....          | 469 |
| 18.5 本章小结 .....                              | 472 |
| <b>第 19 章 总结与其他资源</b> .....                  | 473 |
| 19.1 回顾 SQL Server 2008 数据<br>挖掘的亮点 .....    | 473 |
| 19.1.1 一流的算法 .....                           | 474 |
| 19.1.2 易于使用的工具 .....                         | 474 |
| 19.1.3 简单而强大的 API .....                      | 475 |
| 19.1.4 与同类 BI 技术的集成 .....                    | 475 |
| 19.2 探讨数据挖掘的新领域<br>及应用 .....                 | 475 |
| 19.3 延伸阅读 .....                              | 476 |
| 19.3.1 Microsoft 数据挖掘<br>的资源 .....           | 476 |
| 19.3.2 数据挖掘的其他资源 .....                       | 477 |
| <b>附录 A 数据集</b> .....                        | 479 |
| <b>附录 B 支持的函数</b> .....                      | 483 |
| <b>附录 C 学习资源</b> .....                       | 491 |