国外计算机科学经典教材

Data Mining with SQL Server 2005

数据挖掘原理与应用

——SQL Server 2005 数据库

(美) ZhaoHui Tang 著
Jamie MacLennan
邝祝芳 焦贤龙 高升 译
杨大川 审校

WILEY
清华大学出版社

国外计算机科学经典教材

数据挖掘原理与应用

—— SQL Server 2005 数据库

(美) ZhaoHui Tang
Jamie MacLennan 著

邝祝芳 焦贤龙 高升 译

杨大川 审校

清华大学出版社

北 京

ZhaoHui Tang and Jamie MacLennan
Data Mining with SQL Server 2005
EISBN：0-471-46261-6

图书在版编目(CIP)数据

数据挖掘原理与应用——SQL Server 2005 数据库/(美)唐(Tang, Z. H.)，(美)麦克雷南(MacLennan, J.)著；邝祝芳，焦贤龙，高升译. —北京：清华大学出版社，2007.1
书名原文：Data Mining with SQL Server 2005
ISBN 978-7-302-14000-9

I. 数… II. ①唐… ②麦… ③邝… ④焦… III. 关系数据库—数据库管理系统，SQL Server 2005
IV. TP311.138

中国版本图书馆 CIP 数据核字(2006)第 121563 号

责任编辑：王 军 郑雪梅
装帧设计：孔祥丰
责任校对：成凤进
责任印制：杜 波
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
c-service@tup.tsinghua.edu.cn
社 总 机：010-62770175 **邮购热线**：010-62786544
投稿咨询：010-62772015 **客户服务**：010-62776969
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市新茂装订有限公司
经 销：全国新华书店
开 本：185×260 **印 张**：24.5 **字 数**：536 千字
版 次：2007 年 1 月第 1 版 **印 次**：2007 年 1 月第1 次印刷
印 数：1～4000
定 价：46.00 元

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和改革，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学需要。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学思想和教学方法，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、Thomson 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、Thomson 通过与世界级的计算机教育大师携手，每年都为全球的计算机高等教育奉献大量的优秀教材。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，我们为该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

国外计算机科学经典教材

编审委员会

审 校 者 序

接受这本书的审校任务时，我正在美国进行商务访问。期间，遇到了我的好友，本书的作者之一：ZhaoHui Tang。谈起这本书已被译为中文，并很快在国内出版，大家都感到非常的高兴和欣慰。借此机会，我不妨谈谈自己的感想。

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。

数据库中存在的是数据，对于业务人员来说，只是一些无法看懂的天书，没有人会去拿放大镜分析数据库服务器硬盘上的磁轨。他们需要的是信息。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一些层次的商业智能，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一个层次的商业智能，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作数据挖掘。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 2005 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。公司成立三年来，与微软密切合作，通过近百次讲座和培训，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。

唐朝晖(ZhaoHui Tang)博士是微软公司 SQL Server 2005 数据挖掘产品的产品经理，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。这本书著成以后，在国外获得了很高的评价。虽然已经多次拜读过英文原著，这次在审校的过程中，

依然感到有新的收获。需要提出的一点是：原书定稿的时候，SQL Server 2005 还是 Beta 版，实际上，我们审校的时候，已经是正式发布版本(加上 SP1 补丁)了。这个细节的变化包括了算法程序的改进，反映在书中，大家会发现不少实际案例的截图与原书是不一样的。这并非翻译和审校的错误。实际上，译书中的所有案例都是我和我公司的同事们基于 SQL Server 2005 SP1 版本全部重新搭建并且重新截图的，付出了相当的精力。希望读者朋友们也能够亲手尝试搭建这些案例，这对于掌握这门技术是很有帮助的。

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

译 者 序

存储技术的迅速发展，特别是硬件价格的下降，使得数据的积累速度不断提高，面对日益庞大的数据资源，我们迫切需要强有力的工具来挖掘其中有用的信息。Microsoft最新的数据库平台 SQL Server 2005 中的数据挖掘组件是数据挖掘工具的典型代表。

SQL Server 2000 中包括的数据挖掘算法只有决策树算法和聚类算法，与之相比，SQL Server 2005 中引入了多个新的数据挖掘算法，包括贝叶斯算法、时间序列算法、序列聚类算法、关联规则算法和神经网络算法。与传统数据挖掘工具相比，SQL Server 2005 数据挖掘功能具备许多的优势，SQL Server 2005 数据挖掘功能与所有 SQL Server 产品实现了集成，包括 SQL Server、SQL Server Integration Services 和 Analysis Services。SQL Server 2005 数据挖掘功能具有易用性、可伸缩性和可扩展性等特点，同时它包含简单而丰富的 API。

从 SQL Server 2000 到 SQL Server2005，经历 5 年 SQL Server 数据挖掘的功能实现了一个跨跃式的发展，可谓“五年磨一剑”，也正是本书的 ZhaoHui Tang 和 Jamie MacLennan 两位作者带领他们的团队紧密合作的顶峰。对于本书讲述算法的每一章(包括 4、5、6、7、8、9、10 章)，作者不仅对每一个算法进行了详细的讲述，还在每一章的引言部分给您描述本章算法能应用的一个实际场景，很容易吸引读者往下阅读。

我们作为数据挖掘的研究者，在翻译 SQL Server 2005 数据挖掘时，发觉在 SQL Server 2005 中包含的数据挖掘算法是如此之多，以至于超过了 SAS、SPSS 和 IBM 的 Intelligent Miner 等数据挖掘工具。同时我们作为数据挖掘的工作者，在使用 SQL Server 2005 进行数据挖掘的同时，切身体会到了 SQL Server 2005 数据挖掘功能与所有 SQL Server 产品实现集成给我们的数据挖掘工作带来的便利，以及它的易用性、可伸缩性和可扩展性。

正如本书前言所述，本书对于使用 SQL Serve 2005 进行数据挖掘的用户很有用，可谓及时性、实用性、可靠性集于一体。其中，第 1 章对数据挖掘进行了一个基本的介绍，第 2 章讲述了 OLE DB for DM 规范，第 3 章讲述了如何使用 SQL Server 2005 的数据挖掘工具，第 4 章到第 10 章每一章讲述了一个数据挖掘算法，第 11 章讲述了如何对 OLAP 立方体进行数据挖掘，第 12 章讲述了如何使用 SSIS，第 13 章和第 14 章对数据挖掘的体系结构和 API 进行了讲述，第 15 章实现了一个 Web 交叉销售应用程序，第 16 章讲述了如何使用 Microsoft Excel 进行高级的预测，第 17 章讲述了有关扩展 SQL Server 数据挖掘的知识，第 18 章对本书以及 SQL Server 2005 中数据挖掘的功能进行了总结，以及给出了一些附加资源。附录 A 描述了本书使用的 4 个数据集以及如何导入这些数据集；附录 B 描述了 SQL Server 2005 支持的 VBA 函数和 Excel 函数。对于打算采用本书作为

教材的老师，则建议您讲述第 1 章到第 10 章，以及第 15 章，如果时间允许，您可以讲述第 11、12、17 章，对其他章感兴趣的学生可以自学。

译者

2006 年 5 月于国防科技大学

前　言

数据库系统在过去的20年当中取得了巨大的成功。越来越多的数据被收集并且存储在数据库中。一个数据库拥有海量的数据是很平常的事。在这些数据库中找到有用的信息已经成为许多企业面临的重点问题；数据挖掘作为一个挖掘这些信息的关键组件越来越受到人们的重视。数据挖掘算法和可视化工具适用于挖掘数据中的重要模式，并且提供有价值的预测。这种技术实质上适用于各行各业，包括银行、电信、制造业、营销和电子商务。

在SQL Server 2000中引入了数据挖掘算法和可视化工具。从此以后，大多数关系数据库系统包含了数据挖掘的功能。在将数据挖掘技术与数据库技术进行集成方面，SQL Server 2005中的数据挖掘功能实现了一个跨越式的发展，也正是SQL Server产品团队和Microsoft Research 5年来紧密合作的一个巅峰。来自这两个组织的项目人员和研究人员为SQL Server共同开发了经典的、最新的和前沿的数据挖掘工具。本书作者，ZhaoHui Tang和Jamie MacLennan，是这两个组织合作的重要驱动者。

对于使用SQL Server 2005进行数据挖掘的用户而言，本书将成为他们非常宝贵的参考手册。作者阐述了每个数据挖掘算法的基本原理和各种可视化工具，并且提供了实用的示例。我确信大多数数据库开发人员、数据库管理人员、IT专业人员和数据挖掘方面的学生都会从本书中获益。

David Heckerman
Research Manager
Microsoft Research, Redmond

目　录

CHAPTER 1

数据挖掘导论

在当今的商业组织中数据挖掘变得越来越受关注。您可能会常听说，“我们应该利用数据挖掘工具对我们的客户进行细分”，“数据挖掘将会增加客户对我们的满意程度”，甚至还有，“我们的竞争对手正在使用数据挖掘工具获得更多的市场份额——我们必须加油！”

那么，什么是数据挖掘？使用数据挖掘将带来什么好处？如何利用这种技术来解决商业问题？数据挖掘使用了什么技术？一个经典的数据挖掘项目的生命周期是怎样的？在本章中，将回答所有这些问题，并且加以延伸，带领您进入数据挖掘世界。

在本章中，将会学习：

- 数据挖掘的定义
- 确定哪些商业问题可以通过数据挖掘来解决
- 数据挖掘的任务
- 使用各种数据挖掘技术
- 数据挖掘流
- 数据挖掘项目的生命周期
- 当前的数据挖掘标准
- 数据挖掘领域的几种新趋势

1.1 什么是数据挖掘

数据挖掘是商务智能(Business Intelligence，BI)产品系列中的关键成员，BI 中的关键成员还包括联机分析处理(Online Analytical Processing，OLAP)、企业报表和 ETL。

数据挖掘指的是分析数据，使用自动化或半自动化的工具来挖掘隐含的模式。在过