<<数据挖掘>>

图书基本信息

书名：<<数据挖掘>>

13位ISBN编号：9787111391401

10位ISBN编号：7111391403

出版时间：2012-8

出版时间：机械工业出版社

作者：（美）Jiawei Han,（加）Micheline Kamber,（加）Jian Pei

页数：468

译者：范明,孟小峰

前言

前言： 社会的计算机化显著地增强了我们产生和收集数据的能力。
大量数据从我们生活的每个角落涌出。
存储的或瞬态的数据的爆炸性增长已激起对新技术和自动工具的需求，以帮助我们智能地将海量数据转换成有用的信息和知识。
这导致称做数据挖掘的一个计算机科学前沿学科的产生，这是一个充满希望和欣欣向荣并具有广泛应用的学科。
数据挖掘通常又称为数据中的知识发现（KDD），是自动地或方便地提取代表知识的模式；这些模式隐藏在大型数据库、数据仓库、Web、其他大量信息库或数据流中。
本书考察知识发现和数据挖掘的基本概念和技术。
作为一个多学科领域，数据挖掘从多个学科汲取营养。
这些学科包括统计学、机器学习、模式识别、数据库技术、信息检索、网络科学、知识库系统、人工智能、高性能计算和数据可视化。
我们提供发现隐藏在大型数据集中的模式的技术，关注可行性、有用性、有效性和可伸缩性问题。
因此，本书不打算作为数据库系统、机器学习、统计学或其他某领域的导论，尽管我们确实提供了这些领域的必要背景材料，以便读者理解它们各自在数据挖掘中的作用。
本书是对数据挖掘的全面介绍。
对于计算科学的学生、应用开发人员、行业专业人员以及涉及以上列举的学科的研究人员，本书应当是有用的。
数据挖掘出现于20世纪80年代后期，20世纪90年代有了突飞猛进的发展，并可望在新千年继续繁荣。
本书全面展示该领域，介绍有趣的数据挖掘技术和系统，并讨论数据挖掘的应用和研究方向。
写本书的重要动机是需要建立一个学习数据挖掘的有组织的框架——由于这个快速发展领域的多学科特点，这是一项具有挑战性的任务。
我们希望本书有助于具有不同背景和经验的人交换关于数据挖掘的见解，为进一步促进这个令人激动的、不断发展的领域的成长做出贡献。
本书的组织 自本书第1版、第2版出版以来，数据挖掘领域已经取得了重大进展，开发出了许多新的数据挖掘方法、系统和应用，特别是对于处理包括信息网络、图、复杂结构和数据流，以及文本、Web、多媒体、时间序列、时间空间数据在内的新的数据类型。
这种快速发展、新技术不断涌现使得在一本书中涵盖整个领域的广泛内容非常困难。
因此，我们决定与其继续扩大本书的涵盖面，还不如让本书以足够的广度和深度涵盖该领域的核心内容，而把复杂数据类型的处理留给另一本即将面世的书。
第3版对本书的前两版做了全面修订，加强和重新组织了全书的技术内容，显著地扩充和加强处理一般数据类型挖掘的核心技术。
第2版中讨论特定主题的章节（例如，数据预处理、频繁模式挖掘、分类和聚类）在这一版都被扩充，每章都分成两章。
对于这些主题，一章囊括基本概念和技术，而另一章提供高级概念和方法。
第2版关于复杂数据类型的章节（例如，流数据、序列数据、图结构数据、社会网络数据和多重关系数据，以及文本、Web、多媒体和时间空间数据）现在保留给专门介绍数据挖掘的高级课题的新书。
为了支持读者学习这些高级课题，我们把第2版的相关章节的电子版放在本书的网站上，作为第3版的配套材料。
第3版各章的简要内容如下（重点介绍新的内容）： 第1章提供关于数据挖掘的多学科领域的导论。
该章讨论导致需要数据挖掘的数据库技术的发展历程和数据挖掘应用的重要性。
该章考察挖掘的数据类型，包括关系的、事务的和数据仓库数据，以及复杂的数据类型，如时间序列、序列、数据流、时间空间数据、多媒体数据、文本数据、图、社会网络和Web数据。

该章根据所挖掘的知识类型、所使用的技术以及目标应用的类型，对数据挖掘任务进行了一般分类。最后讨论该领域的主要挑战。

第2章介绍一般数据特征。

该章首先讨论数据对象和属性类型，然后介绍基本统计数据描述的典型度量。

该章概述各种类型数据的数据可视化技术。

除了数值数据的可视化方法外，还介绍文本、标签、图和多维数据的可视化方法。

第2章还介绍度量各种类型数据的相似性和相异性的方法。

第3章介绍数据预处理技术。

该章首先介绍数据质量的概念，然后讨论数据清理、数据集成、数据归约、数据变换和数据离散化的方法。

第4章和第5章是数据仓库、OLAP（联机分析处理）和数据立方体技术的引论。

第4章介绍数据仓库和OLAP的基本概念、建模、结构、一般实现，以及数据仓库和其他数据泛化的关系。

第5章更深入地考察数据立方体技术，详细地研究数据立方体的计算方法，包括Star-Cubing和高维OLAP方法。

该章还讨论数据立方体和OLAP技术的进一步研究，如抽样立方体、排序立方体、预测立方体、用于复杂数据挖掘查询的多特征立方体和发现驱动的数据立方体的探查。

第6章和第7章介绍挖掘大型数据集中的频繁模式、关联和相关性的方法。

第6章介绍基本概念，如购物篮分析，还有条理地提供了许多频繁项集挖掘技术。

这些涵盖从基本Apriori算法和它的变形，到改进性能的更高级的方法，包括频繁模式增长方法，使用数据的垂直形式的频繁模式挖掘，挖掘闭频繁项集和极大频繁项集。

该章还讨论模式评估方法并介绍挖掘相关模式的度量。

第7章介绍高级模式挖掘方法。

该章讨论多层和多维空间中的模式挖掘，挖掘稀有和负模式，挖掘巨型模式和高维空间数据，基于约束的模式挖掘和挖掘压缩或近似模式。

该章还介绍模式探查和应用的方法，包括频繁模式的语义注解。

第8章和第9章介绍数据分类方法。

由于分类方法的重要性和多样性，内容被划分成两章。

第8章介绍分类的基本概念和方法，包括决策树归纳、贝叶斯分类和基于规则的分类。

该章还讨论模型评估和选择方法，以及提高分类准确率的方法，包括组合方法和处理不平衡数据。

第9章讨论分类的高级方法，包括贝叶斯信念网络、后向传播的神经网络技术、支持向量机、使用频繁模式的分类、k-最邻近分类、基于案例的推理、遗传算法、粗糙集理论和模糊集方法。

附加的主题包括多类分类、半监督分类、主动学习和迁移学习。

聚类分析是第10章和第11章的主题。

第10章介绍数据聚类的基本概念和方法，包括基本聚类分析方法的概述、划分方法、层次方法、基于密度的方法和基于网格的方法。

该章还介绍聚类评估方法。

第11章讨论聚类的高级方法，包括基于概率模型的聚类、聚类高维数据、聚类图和网络数据，以及基于约束的聚类。

第12章专门讨论离群点检测。

本章介绍离群点的基本概念和离群点分析，并从各种监督力度（监督的、半监督的和无监督的）以及方法角度（统计学方法、基于邻近性的方法、基于聚类的方法和基于分类的方法）讨论离群点检测方法。

该章还讨论挖掘情境离群点和集体离群点，以及高维数据中的离群点检测。

最后，在第13章我们讨论数据挖掘的趋势、应用和研究前沿。

我们简略地介绍挖掘复杂数据类型，包括挖掘序列数据（例如，时间序列、符号序列和生物学序列），挖掘图和网络，以及挖掘空间、多媒体、文本和Web数据。

这些数据挖掘方法的深入讨论留给正在撰写的数据挖掘高级课题一书。
然后，该章转向讨论其他数据挖掘方法学，包括统计学数据挖掘、数据挖掘基础、可视和听觉数据挖掘，以及数据挖掘的应用。
讨论数据挖掘在金融数据分析、零售和电信产业、科学与工程，以及入侵检测和预防方面的应用。
该章还讨论数据挖掘与推荐系统的联系。
由于数据挖掘出现在我们日常生活的方方面面，所以我们讨论数据挖掘与社会，包括无处不在和无形的数据挖掘，以及隐私、安全和数据挖掘对社会的影响。
我们用考察数据挖掘的发展趋势结束本书。
书中楷体字用于强调定义的术语，而黑体字用于突出主要思想。
本书与其他数据挖掘教材相比具有一些显著特点：它广泛、深入地讨论了数据挖掘原理。
各章尽可能是自包含的，使得读者可以按自己感兴趣的次序阅读。
高级章节提供了更大的视野，感兴趣的读者可以选读。
本书提供了数据挖掘的所有主要方法，还提供了关于多维OLAP分析等数据挖掘的重要主题，这些主题在其他书中常常被忽略或很少提及。
本书还维护了一个网站，其中包含大量在线资源，为教师、学生和该领域的专业人员提供支持。
这些将在下面介绍。
致教师　本书旨在提供数据挖掘领域的一个广泛而深入的概览，可以作为高年级本科生或一年级研究生的数据挖掘导论。
除了讲稿、教师指南和阅读材料列表等教学资源之外，本书网站（www.cs.uiuc.edu/~hanj/bk3或www.booksite.mkp.com/datamining3e）还提供了一个样本课程安排。
根据授课学时、学生的背景和你的兴趣，你可以选取章节的子集，以不同的顺序进行讲授。
例如，如果你只打算给学生讲授数据挖掘入门导论，可以按照图P.1的建议。
注意，根据需要，必要时可以省略其中某些节或某些小节。
图P.1　入门导论课程的建议章节序列　根据学时和讲授范围，你可以有选择地把更多的章节增加到这个基本序列中。
例如，对高级分类方法更感兴趣的教师可以首先增加“第9章　分类：高级方法”；对模式挖掘更感兴趣的教师可以选择包括“第7章　高级模式挖掘”；而对OLAP和数据立方体技术感兴趣的教师可以增加“第4章　数据仓库与联机分析处理”和“第5章　数据立方体技术”。
或者，你可以选择在两个学期的系列课程中讲授整本书，包括本书的所有章节，时间允许的话，加上图和网络挖掘这样的高级课题。
这些高级课题可以从本书网站提供的配套材料选择，辅以挑选的研究论文。
本书的每一章都可以用做自学材料，或者用做数据库系统、机器学习、模式识别和数据智能分析等相关课程的专题。
每章后面都有一些习题，适合作为家庭作业。
这些习题或者是用于测验对内容的掌握情况的小问题，或者是需要分析思考的大问题，或者是实现设计。
有些习题也可以用做研究讨论课题。
每章后面的文献注释可以用来查找包含正文中提供的概念和方法的来源、相关课题的深入讨论和可能的扩展的研究文献。
致学生　我们希望本书将激发你对年青，但正在快速发展的数据挖掘领域的兴趣。
我们试图以清晰的方式提供材料，仔细地解释所涵盖的主题。
每一章后面都附有一个小结，总结要点。
全书包含了许多图和解释，以便使本书更加有趣和便于阅读。
尽管本书是作为教材编写的，但是我们也试图把它组织成一本有用的参考书或手册，以有助于你今后在数据挖掘方面进行深入研究和求职。
为阅读本书，你需要知道什么？
你应当具有关于统计学、数据库系统和机器学习的概念和术语方面的知识。

然而，我们尽力提供这些基础知识的足够背景，以便在读者对这些领域不太熟悉或者记忆有些淡忘时，也能够理解本书的讨论。
　你应当具有一些程序设计经验。
特别是你应当能够阅读伪代码，能够理解像多维数组这样的简单数据结构。
致专业人员　本书旨在涵盖数据挖掘领域的广泛主题。
因此，本书是关于该主题的一本优秀手册。
由于每一章的编写都尽可能独立，所以读者可以关注自己最感兴趣的课题。
希望学习数据挖掘关键思想的应用程序员和信息服务管理人员可以使用本书。
对于有兴趣使用数据挖掘技术解决其业务问题的银行、保险、医药和零售业的数据分析人员，本书也是有用的。
此外，本书也可以作为数据挖掘领域的全面综述，有助于研究人员提升数据挖掘技巧，扩展数据挖掘的应用范围。
　本书所提供的技术和算法是实用的，介绍的算法适合于发现隐藏在大型、现实数据集中的模式和知识，而不是挑选在小型“玩具”数据库上运行良好的算法。
本书提供的每个算法都用伪代码解释。
伪代码类似于程序设计语言C，但也精心加以策划，使得不熟悉C或C++的程序员易于理解。
如果你想实现算法，你会发现将我们的伪代码转换成选定的程序设计语言程序是一项非常简单的任务。
本书资源网站　本书网站的地址是www.cs.uiuc.edu/~hanj/bk3，另一个是Morgan Kaufmann出版社的网站www.booksite.mkp.com/datamining3e。
这些网站为本书的读者和对数据挖掘感兴趣的人提供了一些附加材料，资源包括：　每章的幻灯片。
提供了用微软的PowerPoint制作的每章教案。
　高级数据挖掘的配套章节。
本书第2版的第8～10章涵盖了挖掘复杂的数据类型，这超出了本书的主题，对这些高级主题感兴趣的读者可从网站上获取。
　教师手册。
本书习题的完整答案通过出版社的网站只向教师提供。
　课程提纲和教学计划。
使用本书和幻灯片用于数据挖掘导论课程和高级教程的本科生和研究生，可以获取这些资源。
　带超链接的辅助阅读文献列表。
补充读物的原创性文章按章组织。
　到数据挖掘数据集和软件的链接。
我们将提供到数据挖掘数据集和某些包含有趣的数据挖掘软件包的站点的链接，如到伊利诺伊大学厄巴纳-尚佩恩分校IlliMine的链接(http://illimine.cs.uiuc.edu)。
　作业、考试和课程设计样本。
一组作业、考试和课程设计样本将在出版社的网站上向教师提供。
　本书的插图。
这可能有助于你制作自己的课堂教学幻灯片。
　本书目录。
PDF格式。
　本书不同印次的勘误表。
欢迎读者指出本书中的错误。
一旦错误被证实，我们将更新勘误表，并对你的贡献致谢。
　评论或建议请发往hanj@cs.uiuc.edu。
我们很高兴听到你的建议。

<<数据挖掘>>

内容概要

本书完整全面地讲述数据挖掘的概念、方法、技术和最新研究进展。
本书对前两版做了全面修订，加强和重新组织了全书的技术内容，重点论述了数据预处理、频繁模式挖掘、分类和聚类等的内容，还全面讲述了OLAP和离群点检测，并研讨了挖掘网络、复杂数据类型以及重要应用领域。

本书是数据挖掘和知识发现领域内的所有教师、研究人员、开发人员和用户都必读的参考书，是一本适用于数据分析、数据挖掘和知识发现课程的优秀教材，可以用做高年级本科生或者一年级研究生的数据挖掘导论教材。

作者简介

Jiawei Han（韩家炜），是伊利诺伊大学厄巴纳-尚佩恩分校计算机科学系的Bliss教授。
他因知识发现和数据挖掘研究方面的贡献而获得许多奖励，包括ACM SIGKDD创新奖（2004）、IEEE计算机学会技术成就奖（2005）和IEEE W.Wallace McDowell奖（2009）。
他是ACM和IEEE会士。
他还担任《ACM Transactions on Knowledge Discovery from Data》的执行主编（2006—2011）和许多杂志的编委，包括《IEEE Transactions on Knowledge and Data Engineering》和《Data Mining Knowledge Discovery》。

Micheline Kamber，由加拿大魁北克蒙特利尔Concordia大学获计算机科学（人工智能专业）硕士学位。
她曾是NSERC学者，作为研究者在McGill大学、西蒙-弗雷泽大学和瑞士工作。
她的数据挖掘背景和以易于理解的形式写作的热情使得本书更受专业人员、教师和学生的欢迎。

Jian Pei（裴健），现在是西蒙-弗雷泽大学计算机科学学院教授。
他在Jiawei Han的指导下，于2002年获西蒙-弗雷泽大学计算科学博士学位。
他在数据挖掘、数据库、Web搜索和信息检索的主要学术论坛发表了大量文章，并积极服务于学术团体。
他的文章被引用数千次，并获多次荣誉奖。
他是多种数据挖掘和数据分析杂志的助理编辑。

<<数据挖掘>>

书籍目录

章节摘录

<<数据挖掘>>

编辑推荐

·数据挖掘领域最具里程碑意义的经典著作 ·完整全面阐述该领域的重要知识和技术创新海报：

版权说明