普通高等教育“十一五”国家级规划教材

21世纪高等学校计算机专业核心课程规划教材

数据挖掘原理与算法（第3版）

毛国君 段立娟 编著

清华大学出版社

普通高等教育“十一五”国家级规划教材

21世纪高等学校计算机专业核心课程规划教材

数据挖掘原理与算法

（第 3 版）

毛国君　段立娟　编著

清华大学出版社
北 京

内 容 简 介

本书是一本全面介绍数据挖掘和知识发现技术的专业书籍，系统地阐述了数据挖掘和知识发现技术的产生、发展、应用以及相关概念、原理和算法，对数据挖掘中的主要技术分支，包括关联规则、分类、聚类、序列、空间以及 Web 挖掘等进行了理论剖析和算法描述。本书的许多内容是作者们在攻读博士学位期间的工作总结，一方面，对于相关概念和技术的阐述尽量先从理论分析入手，在此基础上进行技术归纳；另一方面，为了保证技术的系统性，所有的挖掘模型和算法描述都在统一的技术归纳框架下进行。同时，为了避免抽象算法描述给读者带来的理解困难，本书的所有典型算法都通过具体跟踪执行实例来进一步说明。

全书共分 8 章，各章相对独立成篇，以利于读者选择性学习。在每章后面都设置专门一节来对本章内容和文献引用情况进行归纳，它不仅可以帮助读者对相关内容进行整理，而且也起到对本章内容相关文献的注释性索引功能。

本书可作为计算机专业研究生或高年级本科生教材，也可以作为从事计算机研究和开发人员的参考资料。作为教材，教师可以根据课时安排进行选择性教学。为了更好地让教师进行选择性教学，本书配有专门的教师用书，对内容的重点、难点和课时分配给出了对应的建议，对重要的和难度较大的习题进行了分析和解答。对于研究人员，本书是一本高参考价值的专业书籍。对于软件技术人员，可以把它当作提高用书或参考资料，一些算法可以通过改造用于实际的应用系统中。

图书在版编目(CIP)数据

数据挖掘原理与算法/毛国君，段立娟编著. --3 版. --北京：清华大学出版社，2016
21 世纪高等学校计算机专业核心课程规划教材
ISBN 978-7-302-41581-7

Ⅰ. ①数…　Ⅱ. ①毛… ②段…　Ⅲ. ①数据采集－高等学校－教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 220315 号

责任编辑：刘　星
封面设计：杨　兮
责任校对：焦丽丽
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**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
课 件 下 载：http://www.tup.com.cn，010-62795954
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：21.75　　**字　数**：504 千字
版　　次：2005 年 6 月第 1 版　2016 年 1 月第 3 版　　**印　次**：2016 年 1 月第1 次印刷
印　　数：17801～19800
定　　价：39.50 元

产品编号：063843-01

FOREWORD

前言

《数据挖掘原理与算法》经过第 1 版和第 2 版，历经十几年的历程，得到了研究者、教师、学生及计算机从业者的肯定和鼓励，在此表示衷心的感谢。据不完全统计，前两版已经被国内二十多所高校作为研究生或者本科生教材使用。在使用过程中，许多人也对第 2 版中的文字错误、内容编排等提出一些很好的建议。加之数据挖掘技术本身的发展对再次改版提出了强烈需求。第 3 版除了对必要的文字等错误进行修正外，重点增加了大数据挖掘等新的数据挖掘的需求和技术分析，对 Web 挖掘的内容进行了重新编排，并增加了必要的新方法。这样，第 3 版的内容及其编排更趋合理，近年来出现的公认的典型算法和技术也得到加强，使之很好地适应读者在教学或者学习中的新需求。

数据库技术从 20 世纪 80 年代开始，已经得到广泛的普及和应用。随着数据库容量的膨胀，特别是数据仓库以及 Web 等新型数据源的日益普及，人们面临的主要问题不再是缺乏足够的信息可以使用，而是面对浩瀚的数据海洋如何有效地利用这些数据。面对这一挑战，数据挖掘和知识发现技术应运而生，并显示出强大的生命力。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，而且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地解决决策、预测等问题。历经十几年的发展，数据挖掘技术本身已经积累了一批有价值的理论和技术成果。同时，包括统计学、人工智能等在内的相关学科的发展，从某种程度上对数据挖掘技术的发展起到了极大的推动作用。根据麻省理工学院的《科技评论》评估，“数据挖掘”技术是对未来人类产生重大影响的十大新兴技术之一。毫不夸张地说，如今的数据挖掘已经成为计算机、信息科学以及相关领域的一个时髦名词，而且在诸如银行、电信、保险、交通、零售（如超级市场）以及天文学、分子生物学等领域得到应用。可以预见，随着大数据概念的提出和应用，数据挖掘也必将是支撑大数据分析的最重要和最核心的技术之一。

诚然，要真正理解数据挖掘技术并不是一件容易的事。一方面，数据

挖掘技术覆盖范围很广泛,需要从理论到应用、从概念到算法的完整过程;另一方面,作为比较新的交叉研究领域,不同背景的研究人员(数据库、人工智能、数学等)可能提供不同的视点,而且本身仍在发展中。本书第一作者长期从事相关方面的教学工作,其中面临的问题之一就是教材的选择。由于目前相关书籍较少,而且侧重点不同,内容的完整性和科学性有待商榷。由于没有合适的教材可用,在教学的初期不得不通过指定大量参考书或文献来解决,之后也采用补充讲义的形式来扩充。同时,对于一些软件工程师或工程硕士、在职硕士进修班等要求提高实践能力的人员来说,也需要在科学的理论(原理)框架下理解和掌握数据挖掘技术。基于这样的要求,第一作者在多年各类教学和软件工程的实践基础上,对积累的素材进行了整理和加工,并且邀请段立娟博士、王实博士和石云博士参与本书的编写。本书的许多内容是作者们在攻读博士学位期间的工作总结。这些保证了本书的系统性、先进性和实用性。

本书可作为计算机专业研究生教材、高年级本科生的选修教材,也可以作为从事计算机研究和开发人员的参考资料。为了保证内容的先进性和深度,对重点内容进行了重点阐述。本书内容相对全面,各章之间耦合度小。作为教材,教师可以根据学生类型、学时安排等进行选择性教学。作为参考书,读者可以根据自己的基础进行选择性学习或查阅。在每章后面都设置专门一节来对本章内容和文献引用情况进行归纳,它不仅可以帮助读者对相关内容进行整理,而且对读者,特别是研究人员,也起到文献的注释性索引功能。本书的所有典型算法都通过具体跟踪执行实例来进一步说明,这对于读者正确理解和应用算法是有益的。对于工程技术人员来说,这些算法完全可以在理解的基础上进行改进或改造应用到实际工作中。

全书共8章。第1章是绪论,系统地介绍了数据挖掘的概念、产生背景以及应用价值;第2章给出了知识发现的过程分析和应用体系结构设计,并对数据挖掘应用系统的主要功能部件和关键步骤进行了较为详尽的剖析;第3章全面阐述了关联规则挖掘的原理和算法,并对一些新的焦点问题(如多维、数量、约束关联规则挖掘)的最新成果尽可能地加以介绍;第4章给出分类的主要理论和算法描述;第5章讨论聚类的常用技术和算法;第6章对时间序列分析技术和序列挖掘算法进行论述;第7章系统地介绍了Web挖掘的主要研究领域和相关技术及算法;第8章是对空间数据挖掘技术和算法的分析和讲解。

特别感谢北京工业大学刘椿年教授和中国科学院高文和孙玉方研究员,作为作者的导师,他们在作者攻读博士学位期间对本书素材的积累提供了极大的帮助。本书也凝聚了北京工业大学和中央财经大学一些研究生的心血,他们在本书算法实例整理和验证等方面做了很多工作,在此就不一一列举了。此外,也感谢使用第2版图书的教师和学生,他们的使用给予我们进一步编好该书的动力,同时提出的许多意见也提升了第3版内容编排的质量。

作　者

2015年9月于北京

CONTENTS

目录

CHAPTER 1

绪　论　第1章

数据挖掘(Data Mining)是一个多学科交叉研究领域,它融合了数据库(Database)技术、人工智能(Artificial Intelligence)、机器学习(Machine Learning)、统计学(Statistics)、知识工程(Knowledge Engineering)、面向对象方法(Object-Oriented Method)、信息检索(Information Retrieval)、高性能计算(High-Performance Computing)以及数据可视化(Data Visualization)等最新技术的研究成果。经过十几年的研究,产生了许多新概念和新方法。特别是最近几年,一些基本概念和方法趋于清晰,它的研究正向着更深入的方向发展。

数据挖掘之所以被称为未来信息处理的骨干技术之一,主要在于它以一种全新的概念改变着人类利用数据的方式。20世纪,数据库技术取得了决定性的成果并且已经得到广泛的应用。但是,数据库技术作为一种基本的信息存储和管理方式,仍然以联机事务处理(On-Line Transaction Processing,OLTP)为核心应用,缺少对决策、分析、预测等高级功能的支持机制。众所周知,随着数据库容量的膨胀,特别是数据仓库(Data Warehouse)以及Web等新型数据源的日益普及,联机分析处理(On-Line Analytic Processing, OLAP)、决策支持(Decision Support)以及分类(Classification)、聚类(Clustering)等复杂应用成为必然。面对这一挑战,数据挖掘和知识发现(Knowledge Discovery)技术应运而生,并显示出强大的生命力。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询,并且能够找出过去数据之间的潜在联系,进行更高层次的分析,以便更好地做出理想的决策、预测未来的发展趋势等。通过数据挖掘,有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来,从而使大型数据库作为一个丰富、可靠的资源为知识的提取服务。

特别需要指出的是,数据挖掘技术从一开始就是面向应用的。它不仅仅是面向特定数据库的简单检索查询应用,而是要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理,进而发现潜在的知识。这里所说

的知识发现,不是要求发现放之四海而皆准的真理,也不是要去发现崭新的自然科学定理和纯数学公式。所有发现的知识都是相对的,是面向特定领域的,同时还要能够易于被用户理解。

1.1 数据挖掘技术的产生与发展

1.1.1 数据挖掘技术的商业需求分析

数据挖掘之所以吸引专家学者的研究兴趣和引起商业厂家的广泛关注,主要在于大型数据系统的广泛使用和把数据转换成有用知识的迫切需要。20世纪60年代,为了适应信息的电子化要求,信息技术一直从简单的文件处理系统向有效的数据库系统变革。20世纪70年代,数据库系统的三个主要模式:层次、网络和关系型数据库的研究和开发取得了重要进展。20世纪80年代,关系型数据库及其相关的数据模型工具、数据索引及数据组织技术被广泛采用,并且成为了整个数据库市场的主导。从20世纪80年代中期开始,关系型数据库技术和新型技术的结合成为数据库研究和开发的重要标志。从数据模型上看,诸如扩展关系、面向对象、对象-关系(Object-Relation)以及演绎模型等被应用到数据库系统中。从应用的数据类型上看,包括空间、时态、多媒体以及Web等新型数据成为数据库应用的重要数据源。同时,事务数据库(Transaction Database)、主动数据库(Active Database)、知识库(Knowledge Base)、办公信息库(Information Base)等技术也得到蓬勃发展。从数据的分布角度看,分布式数据库(Distributed Database)及其透明性、并发控制、并行处理等成为必须面对的课题。进入90年代,分布式数据库理论上趋于成熟,分布式数据库技术得到了广泛应用。目前,由于各种新型技术与数据库技术的有机结合,使数据库领域中的新内容、新应用、新技术层出不穷,形成了庞大的数据库家族。但是,这些数据库的应用都是以实时查询处理技术为基础的。从本质上说,查询是对数据库的被动使用。由于简单查询只是数据库内容的选择性输出,因此它和人们期望的分析预测、决策支持等高级应用仍有很大距离。

新的需求推动新的技术的诞生。随着信息技术的高速发展,数据库应用的规模、范围和深度不断扩大,已经从单台机器发展到网络环境。近年来由于数据采集技术的更新,如商业条码的推广、企业和政府利用计算机管理事务的能力增强,产生了大规模的数据。数以百万计的数据库系统在运行,而且每天都在增加。决策所面对的数据量在不断增长,即使像使用IC卡和打电话这样简单的事务也能产生大量的数据。随着数据的急剧增长,现有信息管理系统中的数据分析工具已无法适应新的需求。因为无论是查询、统计还是报表,其处理方式都是对指定的数据进行简单的数字处理,而不能对这些数据所包含的内在信息进行提取。人们希望能够提供更高层次的数据分析功能,自动和智能地将待处理的数据转化为有用的信息和知识。

数据挖掘的基础是数据分析方法。数据分析是科学研究的基础,许多科学研究都是建立在数据收集和分析基础上的。同时在目前的商业活动中,数据分析总是和一些特殊的人群的高智商行为联系起来,因为并不是每个人都能从过去的销售情况预测将来的发

展趋势或做出正确决策的。但是，随着一个企业或行业业务数据的不断积累，特别是由于数据库的普及，人工去整理和理解如此大的数据源已经存在效率、准确性等问题。因此，探讨自动化的数据分析技术，为企业提供能带来商业利润的决策信息就成为了必然。

事实上，可以将数据(Data)、信息(Information)和知识(Knowledge)看作是广义数据表现的不同形式。毫不夸张地说，人们对于数据的拥有欲是贪婪的，特别是计算机存储技术和网络技术的发展加速了人们收集数据的范围和容量。这种贪婪的结果导致了“数据丰富而信息贫乏(Data Rich & Information Poor)”现象的产生。数据库是目前组织和存储数据的最有效方法之一，但是面对日益膨胀的数据，数据库查询技术已表现出它的局限性。直观上说，信息或称有效信息是指对人们有帮助的数据。例如，在现实社会中，如果人均日阅读时间在30分钟的话，一个人一天最快只能浏览一份20版左右的报纸。如果你订阅了100份报纸，其实你每天也不过只阅读了一份而已。面对计算机中的海量数据，人们也处于同样的尴尬境地，缺乏获取有效信息的手段。知识是一种概念、规则、模式和规律等，它不会像数据或信息那么具体，但是它却是人们一直不懈追求的目标。事实上，在我们的生活中，人们只是把数据看作是形成知识的源泉。我们是通过正面的或反面的数据或信息来形成和验证知识的，同时又不断地利用知识来获得新的信息。因此，随着数据的膨胀和技术环境的进步，人们对联机决策和分析等高级信息处理的要求越来越迫切。在强大的商业需求的驱动下，商家们开始注意到有效地解决大容量数据的利用问题具有巨大的商机。学者们开始思考如何从大容量数据集中获取有用信息和知识的方法。因此，在20世纪80年代后期，产生了数据仓库和数据挖掘等信息处理思想。

1.1.2 数据挖掘产生的技术背景分析

任何技术的产生总是有它的技术背景的。数据挖掘技术的提出和普遍接受是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。

归纳数据挖掘产生的技术背景，是下面一些相关技术的发展起到了决定性的作用：

- 数据库、数据仓库和Internet等信息技术的发展。
- 计算机性能的提高和先进的体系结构的发展。
- 统计学和人工智能等方法在数据分析中的研究和应用。

数据库技术从20世纪80年代开始，已经得到广泛的普及和应用。在关系型数据库的研究和产品提升过程中，人们一直在探索组织大型数据和快速访问的相关技术。高性能关系型数据库引擎以及相关的分布式查询、并发控制等技术的使用，已经提升了数据库的应用能力。在数据的快速访问、集成与抽取等问题的解决上积累了经验。数据仓库作为一种新型的数据存储和处理手段，被数据库厂商普遍接受并且相关辅助建模和管理工具快速推向市场，成为多数据源集成的一种有效的技术支撑环境。另外，Internet的普及也为人们提供了丰富的数据源。据说，在美国，电视普及达到5000万户大约用了15年，而Internet上网普及达到5000万户仅用了4年。而且Internet技术本身的发展，已经不光是简单的信息浏览，以Web计算为核心的信息处理技术可以处理Internet环境下的多种信息源。因此，人们已经具备利用多种方式存储海量数据的能力。只有这样，数据挖掘技术才能有它的用武之地。这些丰富多彩的数据存储、管理以及访问技术的发展，为数据

挖掘技术的研究和应用提供了丰富的土壤。

计算机芯片技术的发展,使计算机的处理和存储能力日益提高。大家熟知的摩尔定律告诉我们,计算机硬件的关键指标大约以每18个月翻一番的速度在增长,而且现在看来仍有日益加速增长的趋势。随之而来的是硬盘、CPU等关键部件的价格大幅度下降,使得人们收集、存储和处理数据的能力和欲望不断提高。经过几十年的发展,计算机的体系结构,特别是并行处理技术已经逐渐成熟并获得普遍应用,而且成为支持大型数据处理应用的基础。计算机性能的提高和先进的体系结构的发展使数据挖掘技术的研究和应用成为可能。

历经了十几年的发展,包括基于统计学、人工智能等在内的理论与技术成果已经被成功地应用到商业处理和分析中。这些应用从某种程度上为数据挖掘技术的提出和发展起到了极大的推动作用。数据挖掘系统的核心模块技术和算法都离不开这些理论和技术的支持。从某种意义上讲,这些理论本身的发展和应用为数据挖掘提供了有价值的理论和应用积累。数理统计是一个有几百年发展历史的应用数学学科,至今仍然是应用数学中最重要、最活跃的学科之一。如今相当强大有效的数理统计方法和工具,已成为信息咨询业的基础。然而它和数据库技术的结合性研究应该说是近十几年才被重视。以前的基于数理统计方法的应用大多都是通过专用程序来实现的。我们知道,大多数的统计分析技术是基于严格的数学理论和高超的应用技巧的,这使得一般的用户很难从容地驾驭它。一旦人们有了从数据查询到知识发现、从数据演绎到数据归纳的要求,概率论和数理统计就获得了新的生命力。从这个意义上说,数据挖掘技术是数理统计分析应用的延伸和发展。假如人们利用数据库的方式从被动地查询变成了主动地发现知识的话,那么概率论和数理统计这一古老的学科可以为我们从数据归纳到知识发现提供理论基础。

人工智能是计算机科学研究中争议最多而又始终保持强大生命力的研究领域。专家系统曾经是人工智能研究工作者的骄傲。专家系统实质上是一个问题求解系统。领域专家长期以来面向一个特定领域的经验世界,通过人脑的思维活动积累了大量有用信息。在研制一个专家系统时,首先,知识工程师要从领域专家那里获取知识,这一过程是非常复杂的个人到个人之间的交流过程,有很强的个性和随机性。因此,知识获取成为专家系统研究中公认的瓶颈问题。其次,知识工程师在整理表达从领域专家那里获得的知识时,一般用if-then等规则表达,这种表达局限性太大,勉强抽象出来的规则有很强的工艺色彩,知识表示又成为一大难题。此外,即使某个领域的知识通过一定手段获取并表达了,但这样做成的专家系统对常识和百科知识出奇地贫乏,而人类专家的知识是以大量常识知识为基础的。人工智能学家Feigenbaum估计,一般人拥有的常识存入计算机大约有100万条事实和抽象经验法则,离开常识的专家系统有时会比傻子还傻。另外,由于专家系统是主观整理知识,因此这种机制不可避免地带有偏见和错误。以上诸多难题大大限制了专家系统的应用。数据挖掘继承了专家系统的高度实用性的特点,并且以数据为基本出发点,客观地挖掘知识。机器学习应该说是得到了充分的研究和发展,从事机器学习的科学家们,不再满足自己构造的小样本学习模式的象牙塔,开始正视现实生活中大量的、不完全的、有噪声的、模糊的、随机的大数据样本,进而也走上了数据挖掘的道路。因

此，可以说，数据挖掘研究在继承已有的人工智能相关领域的研究成果的基础上，摆脱了以前象牙塔式的研究模式，真正客观地开始从数据集中发现蕴藏的知识。

1.1.3 大数据时代的数据挖掘技术需求分析

大数据(Big Data)概念虽然最早是在20世纪80年代提出的，来自于*Nature* 2008年推出的Big Data专刊，但是真正受到广泛研究与应用探索应该算是在2011年，其中重要的标志是麦肯锡咨询公司发布的《大数据：下一个创新、竞争和生产率的前沿》报告。然而，大数据如此迅猛发展，形成的是目前“边研究边应用”的局面，当然也带来新的问题。

毋庸置疑，数据挖掘技术将是大数据分析的核心和骨干技术之一。当然，大数据时代也对数据挖掘技术的发展提出新的挑战性的问题。我们可以从大数据的发展历史、对应的概念演变入手，分析大数据时代对数据挖掘的技术需求。

大数据研究的发展可以粗略划分成三个阶段。

① 2000年及以前，称为“大数据概念萌芽阶段”。就科技文献而言，从主流的学术引擎上(Google学术)，以Big Data为关键词，检索出的2000年及以前的学术论文不超过50篇。这足以说明该时期虽然有了大数据这个名词，但是并没有受到学术界和商界的广泛重视。当然该时期有了“大数据的萌芽”，其中一个重要标志是：开始将互联网作为大数据的一个重要来源。由于互联网数据的特别关注，使得计算机界早已提出的海量数据处理问题得到扩展。众所周知，海量数据处理起源于大容量科学数据计算需要，所以其研究和应用主要还是面向单一的数据结构(如数据库表)。

② 2001—2010年，大数据概念得到广泛讨论，应用价值获得共识，我们把它称为“大数据概念探索阶段”。例如，这一阶段的年均科技文献已经超过100篇。除美国以外，包括中国在内的其他国家的论文数量也显著增长。说明大数据概念已经得到普遍认可。当然，该阶段大数据的概念还是在探索中，和我们今天认识的大数据还是有差距的。特别地，由于当时数据处理中的“数据挖掘”研究已经进入高峰时期，因此，提出和讨论的大数据概念总是有数据挖掘的影子。然而，从科学发展的历史的角度来说，任何概念总要经过初始提出和不断探讨才能越来越清楚，而这中间数据挖掘概念及其丰富的研究成果为大数据概念的逐步演化起到了关键的作用：提供了理论和方法上的储备。

③ 2011年及以后，大数据的概念进一步深化，已经成为学术研究的焦点，成为许多应用的支撑概念。特别地，检索2011—2013年的论文，其年均论文数量已经接近1000篇。可谓百花齐放、齐头并进，可以看出大数据已经受到学界和商界的高度重视。

下面摘录了一些引用比较多的关于大数据概念的解释。

① 2011年，麦肯锡公司发布的《大数据：下一个创新、竞争和生产率的前沿》报告认为：“大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集”。

② 2012年，国际数据公司(IDC)则认为大数据有四个特征，即4V属性：数据规模大(Volume)、数据高速聚集(Velocity)、数据类型多样(Variety)、数据价值巨大(Value)。关于第四个属性，IBM的研究报告也有真实性(Veracity)之说，所以很多学者认为，前三个属性是最关键的。

③ 2012年，顾能(Gartner)公司技术报告则认为：“大数据是需要新处理模式才能具