教育部“产学合作、协同育人”项目成果教材
“十三五”江苏省高等学校重点教材

大数据分析与挖掘

BIG DATA ANALYSIS AND MINING

www.cmpedu.com
配PPT课件

书内扫码 观看视频
扫码观看视频

朱晓峰 主编
北京络捷斯特科技发展股份有限公司 组编

机械工业出版社
CHINA MACHINE PRESS

教育部“产学合作、协同育人”项目成果教材
“十三五”江苏省高等学校重点教材（教材编号：2018-2-024）

大数据分析与挖掘

组编　北京络捷斯特科技发展股份有限公司

主　编　朱晓峰
副主编　王晓艳　李宇航
参　编　张　琳　郑　乐　冷凯峰　潘海兰
　　殷延海　陈向阳　黎浩东　王志峰

机械工业出版社

本书分为理论篇、工具篇和实训篇。理论篇主要介绍数据挖掘的基础知识、基本任务和常用方法，侧重培养学生对于数据挖掘基本概念等理论知识的正确理解；工具篇主要介绍PMT这一优秀的数据挖掘工具，通过功能简介、分类预测认知实验等内容，侧重培养学生对于数据挖掘基本操作的准确认知；实训篇主要介绍了7个来自企业实际需求的大数据挖掘案例，侧重培养学生对于使用数据挖掘方法解决实际问题的应用能力。

本书结构严密、内容较新、叙述清晰、强调实践，可作为各类院校大数据及相关专业教材，也可作为企事业单位大数据分析培训教材，以及企业管理、电子商务、市场营销、国际贸易等相关从业人员的参考用书。

本书配有电子课件，选用本书作为教材的教师可以从机械工业出版社教育服务网（www.cmpedu.com）免费下载或联系编辑（010-88379194）咨询。本书还配有二维码视频，读者可扫描二维码在线观看。

图书在版编目（CIP）数据

大数据分析与挖掘/北京络捷斯特科技发展股份有限公司组编；朱晓峰主编. —北京：机械工业出版社，2019.3

教育部“产学合作、协同育人”项目成果教材

ISBN 978-7-111-62102-7

Ⅰ.①大… Ⅱ.①北… ②朱… Ⅲ.①数据处理－高等职业教育－教材 Ⅳ.①TP274

中国版本图书馆CIP数据核字(2019)第035930号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）

策划编辑：梁 伟 责任编辑：郑 华 李绍坤

责任校对：杨清清 封面设计：鞠 杨

责任印制：李 昂

河北鹏盛贤印刷有限公司印刷

2019年3月第1版第1次印刷

184mm×260mm · 15印张 · 356千字

0001—3000册

标准书号：ISBN 978-7-111-62102-7

定价：42.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务 网络服务

服务咨询热线：010-88379833 机 工 官 网：www.cmpbook.com

读者购书热线：010-88379649 机 工 官 博：weibo.com/cmp1952

教育服务网：www.cmpedu.com

金 书 网：www.golden-book.com

前　言

大数据分析、数据挖掘是当今科技行业非常受欢迎的流行语，也是各领域人士极为关注的话题。飞速发展的中国，同样将大数据作为行业重点，企业实践成果不断涌现。

本书是数据科学领域为数不多的理论与实践相结合的教材，它通过详细剖析数据挖掘的基础理论、数据挖掘工具基本功能和企业的实训实例，全面展现了大数据分析与挖掘的基础知识、基本任务、常见方法、实用场景和主要流程等。

本书分为三篇。理论篇，大数据分析与挖掘的理论部分，包括数据挖掘概述、数据挖掘任务和方法；工具篇，大数据分析与挖掘的工具部分，包括数据挖掘平台 PMT、数据挖掘认知实验；实训篇，大数据分析与挖掘的实训部分，包括基于时间序列的分仓商品需求预测、基于聚类分析（K-means）的快递企业客户群识别、基于关联规则的超市顾客购物行为分析、基于决策树的电信流失客户预警与分析、基于神经网络算法的共享单车需求预测、基于逻辑回归算法的信用风险预测、深度学习在图像识别及图像分类领域中的应用 7 个不同实际场景的实训。每个实训都包括实训背景、实训分析、核心知识点、实训步骤、拓展与思考 5 个部分。

初识大数据

本书由北京络捷斯特科技发展股份有限公司组编。朱晓峰担任主编，王晓艳和李宇航担任副主编，参加编写的还有张琳、郑乐、冷凯峰、潘海兰、殷延海、陈向阳、黎浩东和王志峰。

由于编者水平有限，本书难免有疏漏和不妥之处，恳请广大读者提出宝贵意见，以期不断改进。

编　者

目　录

理论篇

第 1 章

数据挖掘概述

随着计算机技术、网络技术、通信技术和 Internet 技术的发展，以及各行各业业务操作流程的自动化，企业内积累了大量业务数据，这些数据动辄以 TB 计算。这些数据和由此产生的信息是企业的财富，如实地记录着企业运作的状况。面对大量的数据，人们不断寻找新的工具，来对企业的运营规律进行探索，为商业决策提供有价值的信息，使企业获得利润。能满足企业这一迫切需求的有力工具就是数据挖掘。对于企业而言，数据挖掘有助于发现业务的趋势，揭示已知的事实，预测未知的结果。从这个意义上讲，知识是力量，数据挖掘是财富。

二维码 1-1-1　何为数据挖掘

1.1 数据挖掘的基本概念

1.1.1 数据挖掘的界定

1. 数据挖掘的定义

关于什么是数据挖掘（Data Mining，DM），很多学者和专家给出了不同的定义，包括：

1）Gartner Group 提出："数据挖掘是通过仔细分析大量数据来揭示有意义的新的关系、模式和趋势的过程。它使用模式认知技术、统计技术和数学技术。"

2）The META Group 的 Aaron Zornes 表示："数据挖掘是一个从大型数据库中提取以前不知道的可操作性信息的知识挖掘过程。"

3）J. Han and M. Kamber 认为：数据挖掘是从大量数据中提取或"挖掘"知识。该术语实际上有点用词不当，数据挖掘应当更正确地命名为"从数据中挖掘知识"，不幸的是它有点长。许多人把数据挖掘视为另一个常用的术语"数据库中的知识发现"即 KDD（Knowledge Discoveryin Database）的同义词。而另一些人只是把数据挖掘视为数据库中知识发现过程的一个基本步骤。

4）David Hand 认为：数据挖掘就是对观测到的数据集（经常是很庞大的）进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据。

5）Mehmed Kantardzic 认为：运用基于计算机的方法，包括新技术，从而在数据中获得有用知识的整个过程，就叫作数据挖掘。

综上所述，数据挖掘又译为资料探勘、数据采矿，就是从大量数据（包括文本）中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程；是利用各种分析工具在海量数据中发现模型和数据之间关系的过程。这些模型和关系可以被企业用来分析风险、进行预测。

2. 数据挖掘的本质

什么是数据挖掘，不同的人会给出不同的答案。从本质上而言，往往会给出相似的答案。

（1）数据挖掘是个交叉复合领域

数据挖掘不是有限的几种工具或算法，例如聚类、分类和预测等，它是一个目的性导向的学科，目的是从数据中获取知识、规则或其他可直接、间接用以产生效益的信息。广义上的数据挖掘是和概率统计、高等数学、数学分析、离散数学等数学分支无法清楚分割的，也是和数据库、网络、大数据等技术无法分割的，更是和各行各业的专业知识和业务需求无法分割的。

（2）数据挖掘不追求处理方法，只是为了获取知识

数据挖掘的目的是为了获得知识，至于用了什么手段获得，那只是从愿望到目的的桥梁，重要的是结果。在数据挖掘应用中，不是处理方法越复杂就越好，有时即使是非常简单的方法也可以睿智地理解数据。例如，当统计学家沃德在被咨询飞机上什么部位的钢板需要加强时，他画出飞机的轮廓，标出返航战斗机上受敌军创伤的弹孔位置。统计积累一段时间后，机身各部位几乎都被标满了。最后，沃德建议，把剩下少数几个没有弹孔的位

置加强，因为被击中这些位置的飞机都没有返航。最后实践验证了沃德对飞机改进的良好效果。

（3）数据挖掘是一种探索性的活动

由数据所表达的大量事物中通常可能蕴含了一些规律或知识，但谁也不敢保证一定有。另外，挖掘大量数据中所隐含的知识本身，无论从技术上还是从专业上都是一项极富挑战性的工作。因此，数据挖掘是一种探索性质的活动。探索性质的活动意味着过程可能会很艰辛，结果可能不可预料。所以，如果数据挖掘的结果达不到人们的预期，一种可能是技术、方法不行，一种可能是数据没有能够真实描绘、反映事物，还有一种可能是事物中没有蕴含想要的东西。但是，由于隐含知识通常比表象知识具有更大的价值，而需求引导不断地去追求，因此，数据挖掘会不停地探索。

（4）数据挖掘是有目的的活动

数据挖掘的方向是由业务需求所引领的，知识发现是一项目的性很强的工作。不同的数据挖掘目的所涉及的技术、方法，甚至投入的人力、物力都大有不同，要选择恰当的目的，使得数据挖掘工作可控、成本可控。因此，数据挖掘通常分为评估性初探、计划、评估、实施、再评估、部署、维护等过程。如果数据挖掘目的不明确、缺乏效果评估和风险评估，则项目的失败就会在所难免。

1.1.2　数据挖掘的特征

1. 应用性

数据挖掘是理论算法和应用实践的完美结合。数据挖掘源于实际生产生活中应用的需求，挖掘的数据来自于具体应用，同时通过数据挖掘发现的知识又要运用到实践中去，辅助实际决策。所以，数据挖掘来自于应用实践，同时也服务于应用实践。

2. 工程性

数据挖掘是一个由多个步骤组成的工程化过程。数据挖掘的应用特性决定了数据挖掘不仅是算法分析和应用，而且是一个包含数据准备和管理、数据预处理和转换、挖掘算法开发和应用、结果展示和验证以及知识积累和使用的完整过程。而且在实际应用中，典型的数据挖掘过程还是一个交互和循环的过程。

3. 集合性

数据挖掘是多种功能的集合。常用的数据挖掘功能包括数据探索分析、关联规则挖掘、时间序列模式挖掘、分类预测、聚类分析、异常检测、数据可视化和链接分析等。一个具体的应用案例往往涉及多个不同的功能。不同的功能通常有不同的理论和技术基础，而且每一个功能都有不同的算法支撑。

4. 交叉性

数据挖掘是一个交叉学科，它利用了来自统计分析、模式识别、机器学习、人工智能、信息检索、数据库等诸多不同领域的研究成果和学术思想。同时，一些其他领域如随机算法、信息论、可视化、分布式计算和最优化也对数据挖掘的发展起到重要的作用。数据挖掘与这些相关领域的区别可以由前面提到的数据挖掘的 3 个特性来总结，最重要的是它更侧重于应用。

1.1.3 数据挖掘的基本对象

从字面而言，数据挖掘包含数据和挖掘，二者同样重要，缺一不可。因此，数据挖掘的基本对象就是数据本身。数据作为数据挖掘的基础素材，可以被分为大数据、小数据、宽数据、深数据。

1. 大数据

经典意义上的数据挖掘，通常是指对海量数据进行分析。怎么样才算是海量数据？目前还没有明确的标准。而近几年，类似于海量数据，又产生了大数据的提法，其概念无论从内涵和外延上都有了扩展。但从本质上而言，大数据和海量数据是相似的。在实践中，不单单是记录数多的就称为大数据，通常大数据是指数据量和数据维度均很大，数据形式很广泛，如数字、文本、图像、声音等。而大数据往往可能蕴含着丰富的规律和知识，所以在大数据之上应用数据挖掘就成了理所当然的活动。

2. 小数据

相对于大数据，在实践中还存在不少特殊情况。例如，在医学上有些疾病极为少见，只出现几百例，甚至几十例就几乎是该病的总体了，它们被称为小数据。业务中需要对这些小数据进行深入分析和探索，以便挖掘出罕见疾病的特征，并为相应的临床应对提供依据。对于这样规模的数据进行分析，如果按照记录数，依照传统数据挖掘的观念、方法和技术，则根本无法开展探索性的分析工作。需求引领观念和技术，数据挖掘的一个发展分支应该是从规模较小的、有限的数据中探索其中的规律和知识，尽管目前的技术发展还很有限。

3. 宽数据

还有一种情况是小数据高维度，小样本大信息，称之为宽数据。如某些基因组信息，数据量很少，通常只有几十例到几百例，但维度很高，通常有几百个到几千个。同样，个人大信息，也是单个记录下的高维信息，如从宽带、移动支付、物联网、手机等媒介收集的个人信息。在不远的将来会出现单独个体的高维数据，并需要解决此类数据挖掘的新理论和新算法。

4. 深数据

如果数据涉及维度不是很宽，但是在某几个维度上跨度非常大，历史数据非常多或者数据量的增长速度非常快，可称之为深数据。如医学检查中 24h 心电图监测、较长时段（如 1h 以上）的脑电图监测，每小时会产生几十万至几百万条数据；再如，互联网服务商的 DNS 服务器对互联网访问事件的日志记录，也是每小时会产生几十万至几百万条数据。这类数据，有时也称为流数据。对这些深数据进行挖掘也是非常具有挑战性的，一方面由于它的数据量非常大，另一方面也由于对这类数据进行挖掘的实时性要求较高。

这些随着数据收集手段的进步而形成的各有特色的数据，正在逐步进入数据挖掘研究的视野。所以说，数据挖掘应包括大数据挖掘、小数据挖掘、宽数据挖掘和深数据挖掘。人们需要做的是处理好各类数据来获取知识，研究解决各类型数据的挖掘新理论和新算法，这些数据的分析算法不完全与经典大数据挖掘相同。例如，医学上的个性化精确治疗，就离不开涉及个人的宽数据和深数据。

1.2　数据挖掘的起源与发展

1.2.1　数据挖掘的起源

1. 数据挖掘起源的时代背景

（1）数据爆炸但知识贫乏

《纽约时报》由 20 世纪 60 年代的 10～20 版扩张至现在的 100～200 版，最高曾达 1572 版;《北京青年报》也已是 16～40 版；市场营销报已达 100 版。然而在现实社会中，人均日阅读时间通常为 30～45min，只能浏览一份 24 版的报纸。大量信息在给人们带来方便的同时也带来了新的问题：第一是信息过量，难以消化；第二是信息真假难以辨识；第三是信息安全难以保证；第四是信息形式不一致，难以统一处理。人们开始提出一个新的口号："要学会抛弃信息。"人们开始考虑，如何才能不被信息淹没，而是从中及时发现有用的知识、提高信息利用率?

（2）传统技术不能满足用户需求

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，缺乏挖掘数据背后隐藏知识的手段。

2. 数据挖掘起源的学科背景

由于数据挖掘理论涉及的面很广，它实际上起源于多个学科，如建模部分主要起源于统计学和机器学习。统计学方法以模型为驱动，常常建立一个能够产生数据的模型；而机器学习则以算法为驱动，让计算机通过执行算法来发现知识。而且，数据挖掘除了建模外，还涉及不少其他知识，如图 1-1-1 所示。

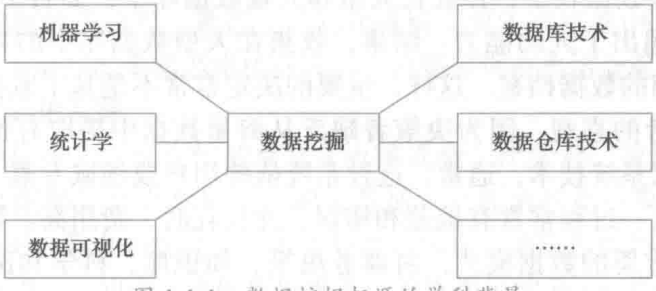

图 1-1-1　数据挖掘起源的学科背景

"数据挖掘"这个术语是在什么时候被大家普遍接受的已经难以考证，它大约在 20 世纪 90 年代开始兴起。最初一直沿用"数据库中的知识发现"。在第一届 KDD 国际会议中，委员会曾经展开讨论，是继续沿用 KDD，还是改名为 Data Mining（数据挖掘）？最后大家决定投票表决，采纳票数多的一方的选择。投票结果颇有戏剧性，一共 14 名委员，其中 7 位投票赞成 KDD，另 7 位赞成 Data Mining。最后一位元老提出"数据挖掘这个术语过于含糊，做科研应该要有知识"，于是在业界便继续沿用 KDD 这个术语。而在商用领域，因为

“数据库中的知识发现”显得过于冗长，就普遍采用了更加通俗简单的术语——“数据挖掘”。严格地说，数据挖掘并不是一个全新的领域，它颇有点“新瓶装旧酒”的意味。组成数据挖掘的三大支柱是统计学、机器学习和数据库，数据挖掘纳入了统计学中的回归分析、判别分析、聚类分析以及置信区间等技术，机器学习中的决策树、神经网络等技术，数据库中的关联分析、序列分析等技术。另外，它还包含了可视化、信息科学等内容。

1.2.2 数据挖掘的发展

1. 数据挖掘的发展历程

（1）数据挖掘的发展，是信息技术自然进化的结果

20 世纪 60 年代以来，数据库和信息技术已经系统地从原始的文件处理进化到复杂的、功能强大的数据库系统。自 20 世纪 70 年代以来，数据库系统的研究和开发已经从层次和网状数据库发展到开发关系数据库系统、数据建模工具、索引和数据组织技术。此外，用户通过查询语言、用户界面、优化的查询处理和事务管理，可以方便、灵活地访问数据。联机事务处理（OLTP）将查询看作只读事务，对于关系技术的发展和广泛地将关系技术作为大量数据的有效存储、提取和管理的主要工具作出了重要贡献。

自 20 世纪 80 年代中期以来，数据库技术的特点是广泛接受关系技术，研究和开发新的、功能强大的数据库系统。这些使用了先进的数据模型，如扩充关系、面向对象、对象—关系和演绎模型。包括空间的、时间的、多媒体的、主动的和科学的数据库、知识库、办公信息库在内的面向应用的数据库系统百花齐放。分布性、多样性和数据共享问题被广泛研究。异种数据库和基于 Internet 的全球信息系统，如 WWW 也已出现，并成为信息工业的生力军。

在过去几十年中，计算机硬件稳定的、令人吃惊的进步导致了功能强大的计算机、数据收集设备和存储介质的大量供应。这些技术大大推动了数据库和信息产业的发展，使得大量数据库和信息存储用于事务管理、信息提取和数据分析。

快速增长的海量数据收集、存放在大型和大量数据库中，如若不依靠强有力的工具，理解它们已经远远超出了人的能力。结果，收集在大型数据库中的数据变成了“数据坟墓”——难得再访问的数据档案。这样，重要的决定常常不是基于数据库中信息丰富的数据，而是基于决策者的直观，因为决策者缺乏从海量数据中提取有价值知识的工具。此外，考虑当前的专家系统技术，通常，这种系统依赖用户或领域专家人工地将知识输入知识库，不幸的是，这一过程常常有偏差和错误，并且耗时、费用高。数据挖掘工具进行数据分析，可以发现重要的数据模式，对商务决策、知识库、科学和医学研究作出巨大贡献。数据和信息之间的鸿沟要求系统地开发数据挖掘工具，将数据坟墓转换成知识“金块”。数据挖掘的进化过程见表 1-1-1。

表 1-1-1 数据挖掘的进化过程

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集（20 世纪 60 年代）	过去五年中我的总收入是多少?	计算机、磁带和磁盘	IBM，CDC	提供历史性的、静态的数据信息

（续）

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据访问（20 世纪 80 年代）	在新英格兰的分部去年三月的销售额是多少？	关系数据库（RDBMS），结构化查询语言（SQL），ODBC 开放数据库连接	Oracle、Sybase、Informix、IBM、Microsoft	在记录层级提供历史性的、动态的数据信息
数据仓库；决策支持（20 世纪 90 年代）	在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论？	联机分析处理（OLAP）、多维数据库、数据仓库	Pilot、Comshare、Arbor、Cognos、Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘（当前）	下个月波士顿的销售会怎么样？为什么？	高级算法、多处理器计算机、海量数据库	Pilot、Lockheed、IBM、SGI、其他初创公司	提供预测性的信息

（2）数据挖掘的发展历程，是一个逐渐演变的过程

电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习成为人们关心的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些范例总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类的问题。随后，随着神经网络技术的形成和发展，人们的注意力转向知识工程，知识工程不同于机器学习那样给计算机输入范例，让它生成出规则，而是直接给计算机输入已被代码化的规则，计算机通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它有投资大、效果不甚理想等不足。20 世纪 80 年代人们又在新的神经网络理论的指导下重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。80 年代末出现了一个新的术语，它就是“数据库中的知识发现”，简称 KDD。它泛指所有从源数据中发掘模式或联系的方法，人们接受了这个术语，并用 KDD 来描述整个数据发掘的过程，包括最开始的制订业务目标到最终的结果分析，而用数据挖掘来描述使用挖掘算法进行数据挖掘的子过程。最近，人们却逐渐发现数据挖掘中有许多工作可以由统计方法来完成，并认为最好的策略是将统计方法与数据挖掘有机结合起来。

2. 数据挖掘的主要里程碑

数据挖掘现在随处可见，而它的故事在《点球成金》出版和“棱镜门”事件发生之前就已经开始了。数据挖掘是在大数据集（即大数据）上探索和揭示模式规律的计算过程。它是计算机科学的分支，融合了统计学、数据科学、数据库理论和机器学习等众多技术。

1763 年，Thomas Bayes 的论文在他死后发表，他所提出的 Bayes 理论将当前概率与先验概率联系起来。因为 Bayes 理论能够帮助理解基于概率估计的复杂现况，所以它成了数据挖掘和概率论的基础。

1805 年，Adrien-Marie Legendre 和 Carl Friedrich Gauss 使用回归分析确定了天体（彗星和行星）绕行太阳的轨道。回归分析的目标是估计变量之间的关系，在这个例子中采用的方法是最小二乘法。自此，回归分析成为数据挖掘的重要工具之一。

1936 年，计算机时代到来，它让海量数据的收集和处理成为可能。在 1936 年发表的论文《论可计算数（On Computable Numbers）》中，Alan Turing 介绍了通用机（通用图灵机）

的构想，通用机具有像今天的计算机一般的计算能力。现代计算机就是在图灵这一开创性概念上建立起来的。

1943 年，Warren McCullon 和 Walter Pitts 首先构建出神经网络的概念模型。在名为《A logical calculus of the ideas immanent in nervous activity》的论文中，他们阐述了网络中神经元的概念。每一个神经元可以做三件事情：接受输入、处理输入和生成输出。

1965 年，Lawrence J. Fogel 成立了一个新的公司，名为 Decision Science, Inc.，目的是对进化规划进行应用。这是第一家专门将进化计算应用于解决现实世界问题的公司。

20 世纪 70 年代，随着数据库管理系统趋于成熟，存储和查询百万兆字节甚至千万亿字节成为可能。而且，数据仓库允许用户从面向事物处理的思维方式向更注重数据分析的方式进行转变。然而，从这些多维模型的数据仓库中提取复杂深度信息的能力是非常有限的。

1975 年，John Henry Holland 所著的《自然与人工系统中的适应》问世，成为遗传算法领域具有开创意义的著作。这本书讲解了遗传算法领域的基本知识，阐述理论基础，探索其应用。

20 世纪 80 年代，HNC 将“数据挖掘”这个短语注册了商标。注册这个商标的目的是为了保护名为“数据挖掘工作站”的产品的知识产权。该工作站是一种构建神经网络模型的通用工具，不过现在早已销声匿迹。也正是在这个时期，出现了一些成熟的算法，能够“学习”数据间关系，相关领域的专家能够从中推测出各种数据关系的实际意义。

1989 年，术语“数据库中的知识发现”（KDD）被 Gregory Piatetsky-Shapiro 提出。这个时期，他合作建立起第一个名为 KDD 的研讨会。

20 世纪 90 年代，“数据挖掘”这个术语出现在数据库社区。零售公司和金融团体使用数据挖掘分析数据和观察趋势以扩大客源，预测利率的波动、股票价格以及顾客需求。

1992 年，Berhard E. Boser、Isabelle M. Guyon 和 Vladimir N. Vanik 对原始的支持向量机提出了一种改进办法，新的支持向量机充分考虑到非线性分类器的构建。支持向量机是一种监督学习方法，用分类和回归分析的方法进行数据分析和模式识别。

1993 年，Gregory Piatetsky-Shapiro 创立“Knowledge Discovery Nuggets（KDnuggets）”通讯。其本意是联系参加 KDD 研讨会的研究者，然而 KDnuggets.com 的读者群现在似乎很多。

2001 年，尽管“数据科学”这个术语在 20 世纪 60 年代就已存在，但直至 2001 年，William S. Cleveland 才以一个独立的概念介绍它。根据《Building Data Science Teams》所述，DJ Patil 和 Jeff Hammerbacher 随后使用这个术语介绍他们在 LinkedIn 和 Facebook 中承担的角色 。

2003 年，Micheal Lewis 写的《点球成金》出版，同时它也改变了许多主流联赛决策层的工作方式。奥克兰运动家队（美国职业棒球大联盟球队）使用一种统计的、数据驱动的方式针对球员的素质进行筛选，这些球员被低估或者身价更低。以这种方式，他们成功组建了一支打进 2002 和 2003 年季后赛的队伍，而他们的薪金总额只有对手的 1/3。

2015 年 2 月，DJ Patil 成为白宫第一位数据科学家。

如今，数据挖掘的应用已经遍布商业、科学、工程和医药领域，这还只是一小部分。信用卡交易、股票市场流动、国家安全、基因组测序以及临床试验方面的挖掘，都只是数据挖掘应用的冰山一角。

1.3　数据挖掘的应用产业与行业

数据挖掘所要处理的问题就是在庞大的数据中找出有价值的隐藏事件并加以分析，获取有意义的信息和模式，为决策提供依据。数据挖掘应用的产业和行业非常广泛，只要有分析价值与需求的数据，都可以利用挖掘工具进行发掘分析。目前，数据挖掘应用最集中的产业包括物流、电商、零售、金融；应用行业包括医疗和电商、电信和交通等。而且每个产业和行业都有特定的应用背景，也都有自己的成功案例（见表 1-1-2）。

表 1-1-2　数据挖掘的应用产业与行业

应用产业	应用方式与成功案例
信用卡公司	信用卡公司可使用数据挖掘来增加信用卡的应用、作购买授权决定、分析持卡人的购买行为并侦测诈骗行为，成功的案例有 Amercian Express 及 Citibank
零售商	了解客户购买行为及偏好对零售商来说是必需的，数据挖掘可以为其提供所需要的信息。像菜篮分析（MBA）或采购篮分析（SBA），或是利用电子销售点（EPOS）数据，并根据其结果来投入有效的促销及广告，有些商店也会应用数据挖掘技术来侦测收银员的诈骗行为，成功的案例有 Wal-Mart 及 Victoria's Secret
金融服务机构	证券分析师广泛使用数据挖掘来分析大量的财务数据以建立交易及风险模型来发展投资策略。许多公司的财务部门已经试着去使用数据挖掘的产品，而且都有不错的效果
银行	虽然数据挖掘在银行业有非常大的应用潜力，但仍处于起步阶段，大约只有 11% 的银行懂得使用数据仓库来促进数据挖掘的活动。银行应该以它们自有的能力来搜集并分析详细的客户信息，然后将结果整合成为营销策略。银行也可以使用数据挖掘以识别客户的贷款活动、调整金融商品以符合客户需求、寻找新的客户及加强客户服务。成功的案例如美国银行，较小的银行因其资源及技术有限，可以通过外包来进行数据挖掘及数据仓库活动
电话销售及直销	电话销售及直销公司因使用数据挖掘已节省许多金钱并且能够精确地取得目标客户，电话销售公司现在不但能够减少通话数，也可以增加成功通话的概率。直销公司正依客户过去的购买数据及地理数据来设置及邮寄它们的产品目录，而直销营销也可利用数据挖掘分析客户群的消费行为与交易记录，结合基本数据，并依其对品牌价值等级的高低来细分客户，进而达到差异化营销的目的
航空业	当前航空公司不断增多，竞争也越来越激烈了，了解客户需求已经变得极为重要，航空公司要取得客户数据以制定因应策略
制造业	数据挖掘已广泛地应用于制造工业的控制和流程，全美第三大的钢铁公司 LTV Steel Corp. 使用数据挖掘来侦测潜在的质量问题，使得他们的不良产品减少了 99%
电信公司	电信公司过去最有名的就是降价策略，但新的策略是了解他们的客户将会比过去来得好。使用数据挖掘，电信公司可以为客户提供各种他们想购买的新服务，电信巨人如 AT&T 和 GTE 正在应用这些快速侦测不寻常行为的技术来防止盗打
保险公司	数据对于保险公司来说是极为重要的，数据挖掘可以使保险公司从大型数据库中取得有价值的信息用以进行决策，这些信息能够让保险公司了解他们的客户并有效地侦测保险欺诈
医疗业	预测手术、用药、诊断或是流程控制的效率