商业分析

Business Analytics

# 商业数据挖掘

许　鑫　万家华◎编著

华东师范大学出版社

商业分析

Business Analytics

# 商业数据挖掘

许　鑫　万家华◎编著

华东师范大学出版社

**图书在版编目（CIP）数据**

商业数据挖掘/许鑫，万家华编著. —上海：华东师范大学出版社，2015.8
（商业分析丛书）
ISBN 978-7-5675-4020-0

Ⅰ.①商… Ⅱ.①许… ②万… Ⅲ.①商业信息—数据采集 Ⅳ.①F713.51

中国版本图书馆 CIP 数据核字（2015）第 192084 号

**商业数据挖掘**

编　　著　许　鑫　万家华
策划组稿　孙小帆
项目编辑　汪　芳
审读编辑　陈　震
装帧设计　卢晓红

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
网　　址　www.ecnupress.com.cn
电　　话　021-60821666　行政传真 021-62572105
客服电话　021-62865537　门市（邮购）电话 021-62869887
地　　址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　店　http://hdsdcbs.tmall.com/

印 刷 者　苏州工业园区美柯乐制版印务有限责任公司
开　　本　787×1092　16 开
印　　张　15.5
字　　数　301 千字
版　　次　2015 年 10 月第 1 版
印　　次　2015 年 10 月第 1 次
书　　号　ISBN 978-7-5675-4020-0/F·341
定　　价　39.00 元

出 版 人　王　焰

# 本书简介

## 目标：

通过结合商业环境中的实例，对数据挖掘概念、流程、方法、工具、应用和最佳实践进行全面介绍，帮助读者了解数据挖掘基本概念、掌握数据挖掘流程和各阶段具体任务及其实现方法、了解数据挖掘在解决实际业务问题中的价值，最终提升商业数据挖掘的能力。

## 内容组织：

本书分为基础编、流程编和应用编三部分。

基础编包括数据挖掘基础（第 1 章）、了解和管理好你的数据（第 2 章）、数据挖掘常用算法（第 3 章），介绍了数据挖掘基本概念、数据相关知识、常用算法和工具等，帮助读者整体上把握商业数据挖掘的应用背景和整体框架。

流程编包括数据挖掘过程（第 4 章）、商业理解（第 5 章）、数据准备（第 6 章）、数据理解（第 7 章）、模型构建（第 8 章）、模型评估（第 9 章）和模型应用（第 10 章），结合电信行业的实例把数据挖掘的整个流程串起来，介绍分类预测模型是如何一步一步实施的，其中还加入了一些电信行业数据挖掘的最佳实践和经验总结。

应用编从数据挖掘的商业应用（第 11 章）和如何做专题分析（第 12 章）两方面展开，通过具体案例展示数据挖掘是如何解决现实商业环境中业务问题的，同时帮助读者学会怎样进一步完成企业专题分析。

## 体例特点：

本书在论述介绍的过程中结合了具体案例及其实例数据，实操性比较强，同时为了帮助读者总结和思考，每章结尾部分均有本章小结和启发思考题。

# 目录

# 基础编

# 1 数据挖掘基础

当前，市场竞争异常激烈，企业以及相关的一些组织机构为了能在竞争中占据优势而费尽心思，数据挖掘技术就是能给企业带来新的生机和活力的利器之一，其也被越来越广泛地应用于各个领域。例如，NBA 的教练利用 IBM 公司提供的数据挖掘工具临场决定替换队员。数据挖掘还能够深入股市去预测股票走势，比如找出一只股票的走势与另一只股票走势相关联的潜在规律，数据挖掘就曾经得到过这么一个结论：如果微软的股票下跌 4%，那么 IBM 的股票将在两周内下跌 5%。再来看一个发生在我们周边的实例。

**小例子**

## 推荐系统——了解客户的伙计

你是否曾有过这样的购物体验？当你在一个图书网站上浏览一本你感兴趣的书籍时，网站总会给你推荐一些你可能感兴趣的其他书籍，不知不觉从你口袋中拿走了更多的钱。这就是当今互联网无处不在的推荐系统，她就像了解你的伙计一样，总能推荐一些你感兴趣的东西。

她是如何做到的呢？

数据挖掘的最近邻算法：

(1) 基于客户的最近邻思想，通过客户的历史图书评价数据，计算客户之间的相似性，找到一个客户的最近邻居，把客户最近邻居喜欢的图书推荐给该客户；

(2) 基于图书的最近邻思想，通过图书的历史评价信息，计算图书之间的相似性，当客户购买某本图书时，同时把与该书最相似的图书推荐给客户。

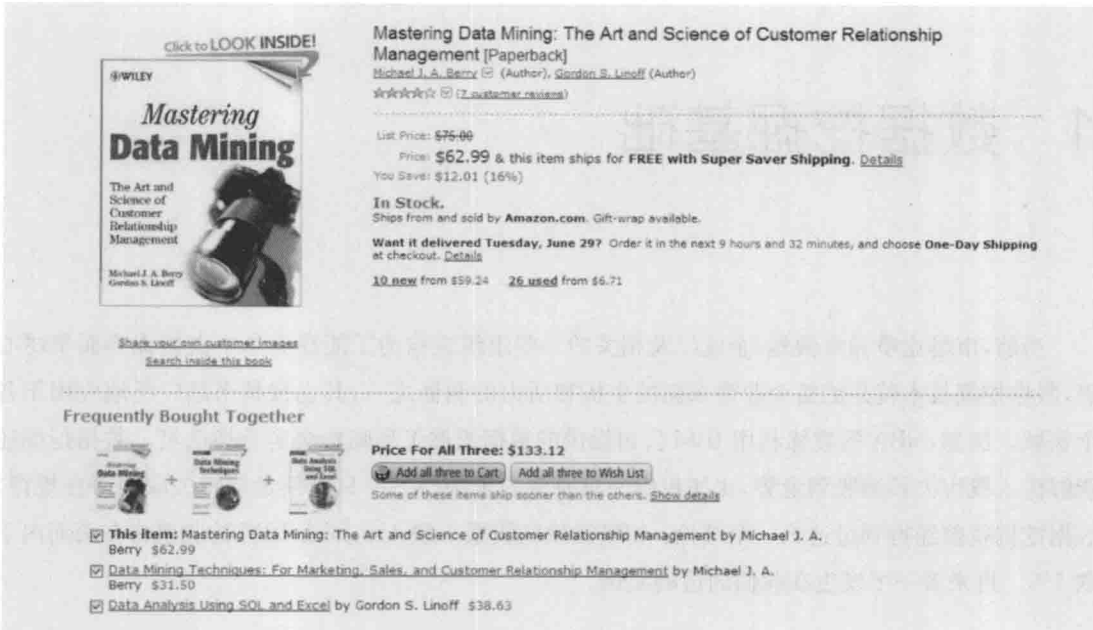


图 1-1　你正在亚马逊浏览一本你感兴趣的数据挖掘图书

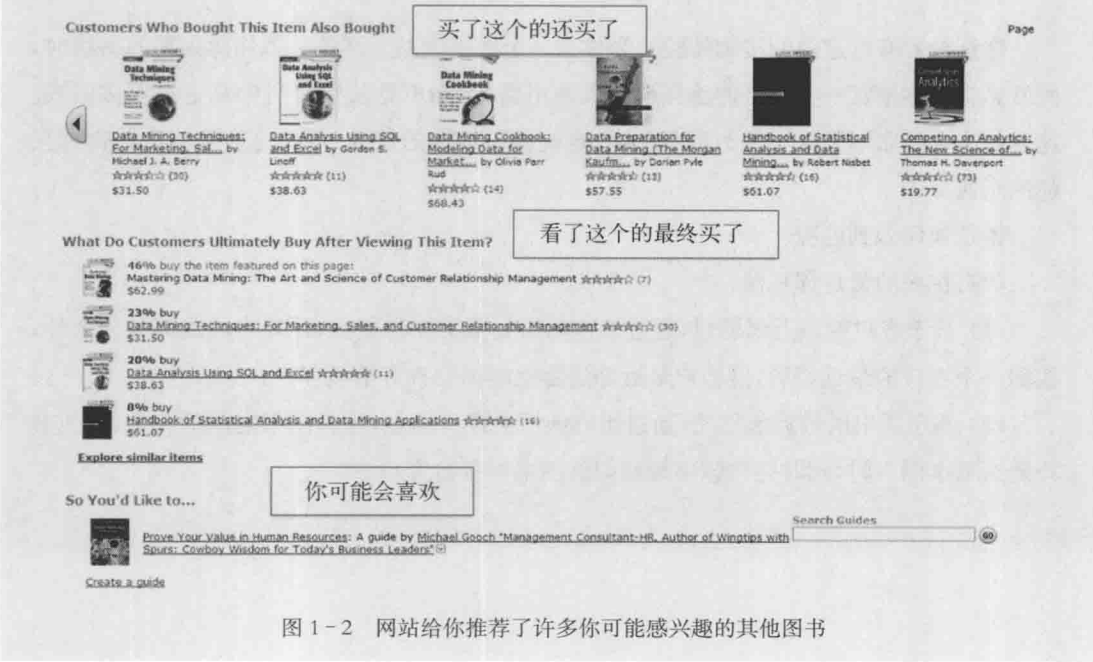


图 1-2　网站给你推荐了许多你可能感兴趣的其他图书

数据挖掘对于提升竞争力的新关联、新模式、新知识的发现是各类机构关注它的主要原因；另外一方面，现实的数据环境也迫使人们要利用诸如数据挖掘这样的技术。信息爆炸的时代，人们被淹没在数据的海洋中，面对浩如烟海的数据，人们往往手足无措，如何从各种各样海量的数据中获取所需信息和知识正是数据挖掘所擅长的。

既然现实的商业环境和数据环境客观上都对数据挖掘有所期待，那么数据挖掘有哪些方法？这些方法的应用场景有哪些？如何实际地运用这些方法开展商业分析？上述问题都是企业中与数据打交道的同仁们亟须了解的。

## 1.1 数据挖掘的产生背景

数据挖掘的兴起有着它的应用背景，当全球向信息化社会迈进之时，人类利用信息技术收集、加工、组织、生产信息的能力也大大提高，数以万计的各种类型的数据库诞生，它们在科学研究、技术开发、生产管理、市场扩张、商业运营、政府办公等方面发挥着巨大作用。然而，随着信息量的不断增多，特别是网络信息资源的迅猛扩张，人类面临着新的挑战。如何不被堆积如山的信息所淹没？如何能够迅速地从海量信息中获取有用数据？如何能够充分提高信息的利用率？由此，数据挖掘技术应运而生。从目前的发展趋势来看，数据挖掘技术的研究与应用越来越显示出强大的生命力。

随着世界信息技术的迅猛发展，信息量也呈几何指数增长。特别是随着云时代的来临，海量数据发展到大数据(big data)已日益明显，现在许多单位与组织在日常运营中生成、累积的各种数据，规模是如此庞大，以至于不能用 G 或 T 来衡量。例如，一天之中，互联网上产生的全部内容可以刻满 1.6 亿多张 DVD；发出的邮件有 2 940 亿封之多(相当于美国两年的纸质信件数量)；发出的社区帖子达 200 万个(相当于《时代》杂志 770 年的文字量)；卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万……

截止到 2012 年，数据量已经从 TB(1 TB=1 024 GB)级别跃升到 PB(1 PB=1 024 TB)、EB(1 EB=1 024 PB)乃至 ZB(1 ZB=1 024 EB)级别。国际数据公司(IDC)的研究结果表明，2008 年全球产生的数据量为 0.49 ZB，2009 年为 0.8 ZB，2010 年增长为 1.2 ZB，2011 年的数量更是高达 1.82 ZB，相当于全球每人产生 200 GB 以上的数据。而到 2012 年为止，人类生产的所有印刷材料的数据量是 200 PB，全人类历史上说过的所有话的数据量大约是 5 EB。IBM 的研究称，整个人类文明所获得的全部数据中，有 90%是过去两年内产生的。而到 2020 年，全世界所产生的数据规模将达到今天的 44 倍。如何从巨量、复杂的数据中获取有用的信息，成为信息技术研究领域的热

门课题。在这样的背景下，数据挖掘技术诞生并成为近年来的研究热点。机器学习、数据库技术和数理统计是数据挖掘的三个技术支柱。今天，这些技术已经相当成熟，加上高性能关系数据库引擎、数据仓库、文档数据库和广泛的数据集成，让数据挖掘技术得到了广泛的实际应用。

目前数据挖掘相关研究文献越来越多、可用技术也层出不穷，数据挖掘的理论体系正在形成，相信很快就会成为一种主流信息技术。当然，数据挖掘面向应用领域要做的事还很多，比如需要开发更多数据挖掘系统和产品，需要建立行业内的数据标准和通用挖掘平台，需要建立可交换信息和共享知识的通用数据仓库等。应该说，数据挖掘包含的内容很多，值得研究的方面也很多。但是，我们也注意到，就目前而言，注重多种策略和技术的集成，以及各个学科之间的相互渗透是目前的研究热点。传统机器学习方法一般使用研究者按照条件和结论事先组织好的数据，但是数据挖掘却需要面对现实的数据，通常具有不完整、带有噪声、数量大，甚至还不断增加等特点，因此传统机器学习方法需要改进后才能用于数据挖掘。所以，目前数据挖掘的研究重点应该是针对应用实践，综合借鉴交叉学科中的技术和方法，互相渗透，发现新的方法或进行多种策略和技术的集成。

数据挖掘研究不仅来自对“堆积如山”信息量的处理需求，更是由于社会发展各方面的迫切需要而发展起来的。如，企业为了提高自己的竞争力开展良好的商业运作、信息提供商对网络信息资源的组织等都需要研究数据挖掘技术。

(1) “信息爆炸”引发对数据挖掘的需求

在现代社会中，信息的激增主要表现在两个方面。一方面，信息的获取渠道大大增加，包括网络、广播(电台数量增多)、电视(频道大大增加)、广告(网络广告、街头广告、报纸广告、电视电台广告)、报纸杂志、各类会议等。另一方面，信息量成倍增长，例如：报纸杂志无论是种类还是版面都较过去大大增加；数据库的种类、形式及其各自规模的扩充与发展非常迅猛；网络信息更是难以估计；等等。面对如此繁杂巨量的信息源，人们往往无所适从，大量的“信息垃圾”扰得人们烦躁不安。在这样一个“信息爆炸”的社会中，人们深感过去传统的信息获取手段是如此“笨拙”。因此，迫切希望有一种技术能够从各类信息源中准确、全面、有效地得到有用的数据和信息，这就是计算机应用领域、信息检索领域目前正在寻求和研究的技术——数据挖掘。

(2) 解决数据爆炸和知识贫乏这对矛盾的需要

数据积累得越多，隐藏在数据背后的知识和信息就越多，同样，数据间的各种关联关系也就越多。过去的数据组织和信息检索技术难以将数据间的关联关系表现出来，虽然数据库系统具备高效的数据查询和统计等功能，但无法发现数据间存在的联系和规则，更难以发现隐藏在大量数据背后的知识，也无法根据现有数据预测它的未来趋势。有人把这一现象称作“数据激增带来

了知识贫乏”。要解决上述问题，必须研究新的数据组织工具和数据分析、处理与检索技术。由于技术上的需要，数据仓库技术由此而诞生，它能够在数据间建立更加有机的联系，所具备的联机分析处理技术(OLAP, online analytical processing)能够获得数据间的关联与规则，为数据挖掘的研究提供了分析处理工具。

(3) 企业竞争中数据挖掘的需求

知识经济时代，企业竞争更加激烈，如何实现企业内部的知识管理、挖掘和发现企业内部的隐性知识、抓取与企业竞争相关的外部数据和信息，是企业在信息化时代保证竞争处于有利地位所面临的紧迫问题。解决这一问题，除了管理上的改革(实现知识管理)以外，在信息技术方面也需要有一种技术来支持，这就是数据挖掘与知识发现。目前，企业所面临的共同问题是：与企业相关的信息量非常大，但真正具有直接利用价值的数据却很少，需要对大量的信息进行深层分析，发现其内在关联，挖掘其中有利于企业竞争的信息，好比在沙石中淘金。这就是企业竞争对数据挖掘和知识发现的需要。

(4) 商业运作的数据挖掘需求

商业管理和运作的自动化和计算机化，将会在这一领域产生大量的数据，包括商品本身的信息和商业运作过程中产生的大量业务数据。这些数据并不是专门为了分析而收集的，而是纯粹在商业运作过程中记录下来的。分析这些数据、从这些数据中挖掘知识将会对企业的商业决策提供非常有价值的信息。例如，在网上书店，可以通过购买记录分析某一本书的购买群体，这些读者买了此书后又买了哪些书，等等，通过挖掘，当再有读者购买该书时，就显示出他可能还会买哪些书。另一个典型的例子就是，一家连锁店通过对销售记录的数据挖掘，发现了婴儿尿布和啤酒之间有着惊人的联系。可以认为，数据挖掘是为了商业决策、商业运作、商业竞争而自然产生的一种新型数据处理技术，企业在市场竞争中需要利用数据挖掘技术分析研究市场行为、提高对市场运作过程的控制、加强对客户关系的管理等。随着数据挖掘技术带来的商业利润，人们对它在商业领域中的运用将更加深入。

小贴士

### 相关技术的发展和演变

数据挖掘实际上是相关技术逐渐发展和演变的一个过程。早在 20 世纪 50 年代末，IBM 公司的 H. P. Luhn 先生就对文献信息的自动标引、自动摘要做过研究与实践；1960 年，M. E. Maron 先生发表了第一篇有关文献信息的自动分类的文章。这些研究都可以看

作是当今数据挖掘技术的最早雏形。随后，人工智能领域开展的机器学习研究，也就是将某一应用中已被成功解决的许多范例和有关规则输入计算机，然后通过它们解决问题，这就是早期的专家系统。20 世纪 80 年代神经网络理论形成，人们在该理论的指导下，将前期的相关研究成果应用于对大型数据库的数据处理，并对源数据进行关联分析和统计，这项研究产生了一个新的术语——数据库知识发现（KDD，Knowledge Discovery in Database），而目前人们往往把数据挖掘和知识发现相提并论。

数据挖掘出现于 20 世纪 80 年代末，最早是从数据库知识发现研究起步，KDD 一词首先出现在 1989 年人工智能国际会议上，以后这一研究逐渐成为热点，由于这项研究对象的扩展，人们更多地称之为数据挖掘。1995 年，召开了第一届知识发现与数据挖掘国际会议，以后每年召开一届。我国从事数据挖掘的研究起步较晚，大约在 20 世纪 90 年代中期，许多高校、科研院所在这一领域内开展研究，并取得了许多成绩。

20 世纪 90 年代末，随着数据挖掘研究的深入，数据仓库技术随之出现，并促使着数据挖掘研究更加深入和广泛。近年来，数据挖掘研究已拓展到数据库以外的非结构化数据源的挖掘研究，如网络信息的 Web 挖掘、全文信息的文本挖掘、多媒体信息的数据挖掘等。

如上所述，数据挖掘是由于多方面的应用需求和技术发展演变而产生的。实际上，我们也可以将其看作是多项技术发展的必然会合。例如，基于人工智能的知识发现技术已基本成熟，但由于缺乏应用的土壤，研究路子越来越窄。同样，积累了无数资源的数据库，其技术虽已相当成熟，但由于缺乏发现数据内在规律和相互关联以及挖掘隐藏在数据之中知识的工具，发展受到阻碍。两者结合实现了互补（双赢），知识发现技术有了应用的领域，数据库技术也得到了充分的发展。

## 1.2 数据挖掘的概念定义

数据挖掘（data mining）是指从大量的数据（结构化和非结构化）中提取有用的信息和知识的过程。在这个定义中，数据是大量的、真实的、不完全的、有噪声的、模糊的、随机的实际应用数据；所发现的信息和知识是潜在的并隐藏在大量数据背后的，是用户感兴趣的、可理解、可运用的知识。所以，数据挖掘有时也被人们称为知识挖掘、知识提取、知识发现等，可以说数据挖掘的本质就是知识发现，事实上，人们往往不严格区分数据挖掘和数据库知识发现这两个概念，常常将两者混同使用。一般在科学研究领域中称为 KDD，而在工程应用领域则称为数据挖掘。不过不

要认为这里所指的知识发现是发现放之四海而皆准的真理，也不是去发现新的物质或新的自然科学定理，更不是利用计算机证明某个定理是否正确。实际上，它所有发现的知识都是隐藏在大量数据之中的关联信息，所有的知识都是有特定前提和约束条件的，是面向特定领域的，而且，这些知识还要能够易于被用户理解，能用自然语言表达所发现的结果。数据挖掘是一门交叉学科，涉及机器学习、统计学、人工智能、模式识别、数据库、信息检索、信息可视化和专家系统等多个领域。

数据挖掘也可视为是一类深层次的新型数据分析方法，它与传统的数据分析（查询、报表、联机应用分析）的本质区别在于：数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识，所得到的信息通常是预先未知又难以预料到的，甚至是与人的直觉相违背的，但又是非常有用的；而传统的数据分析得到的信息则是浮于表面并与人的直觉较为接近的。

数据挖掘也可以当作一个在海量数据中探索数据间的关系、利用各种分析工具构建数据分析模型、发现隐藏于数据之中知识的过程。这一过程由若干步骤组成：数据清理（消除冗余数据，排除噪声，保证数据的干净和一致）；数据集成（多种数据源组合，并建立数据间的关系）；数据选择（根据问题选择欲挖掘的数据库）；数据转换（把数据变成适合于挖掘的形式，如关联、集聚、重组等）；数据挖掘（具体的挖掘步骤：构造数据提取模式，检索数据知识）；挖掘结果评估（利用兴趣度测量法，确定表示知识最有趣的模式）；知识表示（将挖掘出的知识向用户展示，可采取可视化的知识表达技术）。

被挖掘的原始数据其形式是复杂多样化的，有结构化数据、异构化数据、半结构化数据，甚至是非结构化数据（包括多媒体数据）。使用的数据挖掘方法也是丰富多彩的：有数学方法，也有非数学的方法；有演绎的方法，也有归纳的方法；有聚类的方法，也有分类的方法；有关联的方法，也有孤立点分析方法；有对文本数据的挖掘方法，也有对复杂数据的挖掘方法；等等。被挖掘出来的数据和知识的用途是广泛的，可以用于知识管理、信息服务的知识推送、企业竞争、客户关系管理以及决策支持和过程控制等领域。

在进行数据挖掘和知识发现的学习与研究过程中，数据、信息、知识是我们直接接触的三个概念，这三者之间既有区别又有联系，在受到其他因素的作用之后，它们之间将会进行转化。图1-3显示了数据到知识的转化过程。在实际的数据挖掘中，从数据到知识也要经过这样的转化过程，但它是采用各种算法和模式实现的。


图1-3 数据、信息、知识的转化