大数据掘金

挖掘商业世界中的数据价值

［美］杜尔森·德伦（Dursun Delen） 著

丁晓松 宋冰玉 译

REAL-WORLD DATA MINING

Applied Business Analytics and Decision Making

在滚滚而来的数据洪流中沙里淘金

挖掘大数据背后的价值洼地，为企业带来下一个增长红利

中国人民大学出版社

大数据掘金

挖掘商业世界中的数据价值

[美] 杜尔森·德伦（Dursun Delen） 著

丁晓松 宋冰玉 译

REAL-WORLD DATA MINING

Applied Business Analytics and Decision Making

中国人民大学出版社

• 北京 •

图书在版编目（CIP）数据

大数据掘金：挖掘商业世界中的数据价值 /（美）德伦著；丁晓松，宋冰玉译. —北京：中国人民大学出版社，2016.1

ISBN 978-7-300-22031-4

Ⅰ.①大… Ⅱ.①德… ②丁… ③宋… Ⅲ.①商业信息—数据处理 Ⅳ.①F715.51

中国版本图书馆 CIP 数据核字（2015）第 252672 号

大数据掘金：挖掘商业世界中的数据价值

[美] 杜尔森·德伦 著

丁晓松 宋冰玉 译

Dashuju Juejin: Wajue Shangye Shijie Zhong de Shuju Jiazhi

出版发行	中国人民大学出版社		
社　　址	北京市中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）		010-62511770（质管部）
	010-82501766（邮购部）		010-62514148（门市部）
	010-62515195（发行公司）		010-62515275（盗版举报）
网　　址	http:// www. crup. com. cn		
	http:// www. ttrnet. com（人大教研网）		
经　　销	新华书店		
印　　刷	北京中印联印务有限公司		
规　　格	170 mm × 230 mm 16开本	**版　　次**	2016 年 1 月第 1 版
印　　张	14.25 插页 1	**印　　次**	2016 年 1 月第 1 次印刷
字　　数	180 000	**定　　价**	49.00 元

版权所有　侵权必究　印装差错　负责调换

推荐序

杜尔森·德伦博士的著作简明清晰、内容丰富，为渴望了解数据分析、数据挖掘和“大数据”的读者提供了实用的学习工具。在商业活动越来越复杂、越来越趋向全球化的今天，决策者必须依靠现有的信息采取快速准确的行动，而这必须依靠现代数据挖掘和分析。本书明确了该领域当前的最佳做法，向读者——主要是学生和从业者——展示了如何应用数据的挖掘与分析发现数据隐含的规律与联系，如何利用这些信息改进并提升整个决策过程。

作者选取了适量的概念、技术和案例帮助读者真正理解数据挖掘技术的运行原理。这些技术包括：数据挖掘过程、方法与技术，数据的作用与管理，工具与量表，文本与网页挖掘，情感分析，以及接下来与最新大数据分析方法的整合。

在第 1 章中，作者巧妙地将数据分析的源头追溯到了第二次世界大战时期（见图 1—2），使用下列期刊的读者信息作为数据：20 世纪 70 年代的《决策支持系统》（*Decision Support Systems*）、20 世纪 80 年代的《企业 / 高管 IS 系统》（*Enterprise/Executive IS Systems*）以及我们都听说过的 20 世纪 90 年代和 21 世纪初期的《商务智能》（*Business Intelligence*），最后还有当前的《分析》

（*Analytics*）和《大数据时代》（*Big Data*）。第 1 章的内容为后续即将论述的数据挖掘打下基础。

在第 2 章中，作者对数据挖掘进行了简明易懂的描述，并进行了准确的分类，将数据挖掘与其他几个相关的术语区分开来，明确表示了数据挖掘的实际意义是发现知识。认识到数据挖掘实质上是在坚持许多原则的基础上解决问题与制定决策，无疑是思维上的一次洗礼，许多人都认为数据挖掘本身是一种新概念。这一章运用现实生活中的真实案例、具有启发性的图表以及平实的语言，向广大读者揭开了数据挖掘的神秘面纱。这种方法十分巧妙，将数据挖掘这样看似复杂而又富有技术含量的话题介绍给了普罗大众。

在第 3 章中，德伦博士以浅显易懂的形式向读者展示了规范数据挖掘过程的不同方法。该章介绍的第一种方法是数据库知识获取（Knowledge Discovery in Database，KDD），这种方法由业界先驱尤萨马·法雅德（Usama Fayyad）首创。德伦博士在讨论中展示了 KDD 技术，并用图表加以说明（见图 3—1），清楚地显示了运用 KDD 技术进行数据挖掘的过程。与此同时，这一章还介绍了众多团体或个人提出的其他数据挖掘方案，这些方案共同构成了数据挖掘这一领域基本思想的沿革发展。为了显示这些方案的实用性，德伦博士还在最后提供了一个案例研究——“挖掘癌症数据，获取最新知识”。

第 4 章主要研究数据挖掘中使用到的数据，包括目前越来越频繁使用的文本数据（即非结构化、非数字性的数据，占当今世界可用数据的近 90%）。数据准备是数据挖掘最重要的一步，要建立实际可用的模型，所用的数据必须经过处理统计，否则就像俗语中说的“无用输入，无用输出”。因此，在数据挖掘过程中近乎 90% 以上的时间都花在了数据准备这一环节。德伦博士竭尽所能采取种种方法统计整理数据，为进一步的数据分析做好准备，这些准备包括打造数据链，测试数据组，为学习者提供最人性化的 k 倍交叉核实界面（见图 4—6）。

此为试读，需要完整PDF请访问：www.ertongbook.com

在第 5 章中，德伦博士介绍了最常见的数据挖掘运算，其讲解简明易懂，外行人也能看出门道。此外，他还全面介绍了神经网络与支持向量机（Support Vector Machines，SVM），使这些原本晦涩难懂的数学工具变得生动易学。其中，德伦博士亲自设计的演算示例也让本书物超所值。

第 6 章详细讲述了文本挖掘（即文本分析）。一开始，德伦博士引用了我们在 2012 年出版的《实用数据挖掘》（*Practical Data Mining*，我本人是这本书的主编）首次使用的图表。博士成功地将我们 1 100 页的著作浓缩成短短一章——事实上，这样的浓缩版本对初学者而言更有意义。干得漂亮，德伦！

最后，在第 7 章中，德伦博士介绍了当前分析领域一个炙手可热的名词——大数据分析。我们几乎每天都能在新闻中听到“大数据”这个词，它到底是什么意思呢？对不同的人而言，这个词有着不同的含义。但作为一个在数据挖掘领域活跃了 15 年以上的人，我可以说每时每刻都与大数据打交道。数据存储空间的成本越来越低，云存储逐渐进入人们的生活，一台小小的笔记本电脑都能够进行数据分析中的分配步骤和多线程运算。轻薄的平板电脑甚至能够胜过几十年前存放在开着冷气的库房中的主服务器。现在人们甚至可以用智能手机管理几个服务器和云存储。数据正日渐变“大”，而处理数据所需的物理实体却越来越“小”。

但是大多数人对大数据都存在着误解，至少在我看来是这样的。许多人认为数据挖掘必须用到大数据。我与住院医师有过 10 年的合作，他们希望在为期一年的项目中研究尽可能多的案例，但在有限的时间内只能找到一部分所需的材料。以传统统计学标准来看，这些小型数据组的研究是没有任何意义的，但是我发现，使用工具学习这种现代数据挖掘方法，往往能够从小数据组中得到有用的假设，获得从前使用传统费雪学派 p 值统计法不可能得到的信息。在 20 世纪，传统统计学还被认为是非主流的统计方法，而在 20 世纪以前，贝叶斯统计法（Bayesian statistics）曾统领了数据分析领域长达几百年之久。随着 21 世

纪的到来，贝叶斯统计的现代形式，包括SVM、NN及其他工具学习模型卷土重来，我们又回到了贝叶斯的时代。虽然对于“传统统计训练”而言，还需要一定时间来理解和跟上时代的潮流，但是统计领域的前沿阵地无疑是属于贝叶斯统计法、数据挖掘和大数据的。

所有想要了解数据挖掘并在这一方面掌握一技之长的读者都应该选择这本书，当阅读到本书的最后一页就会发现，你已经完全了解这一领域，如蛹化蝶飞。

加里·麦尼博士（Dr. Gary D. Miner）
戴尔信息管理集团软件事业部
高级分析师、医疗保健应用专家
（其两部著作曾经获得PROSE奖）

目 录

REAL-WORLD DATA MINING

APPLIED BUSINESS ANALYTICS AND DECISION MAKING

第1章　分析学入门

尽管商务分析学是个新名词，但最近却在商业世界中以前所未有的势头迅速升温。一般而言，分析学是指发现信息的方式和技术，即利用复杂的数学模型、数据和专业知识进行有效而及时的决策制定。从某种程度来讲，分析学的意义就是制定决策和解决问题。近年来，分析学也可以被简单定义为“发掘数据中有意义的规律”。在当今互联网时代，分析学所用的数据也逐渐向数量大、种类多的方向发展。尽管分析学更多地关注数据，然而许多分析学的应用对数据需求却很少甚至不需要数据。恰恰相反，这些应用使用的是依赖过程描述和专业知识发挥作用的数学模型（比如优化与仿真模型）。

商务分析学利用分析工具、技术以及原理来解决复杂的商业问题。企业往往通过分析数据来描述、预测和改善企业绩效。数据分析在企业中有众多的应用，具体如下：

- 改善企业与客户（此效应贯穿采购、退货、添货等顾客关系管理的所有过程）、

此为试读，需要完整PDF请访问：www.ertongbook.com

员工及其他利益相关者的关系；

- 明确欺诈交易及不正当行为，以节省开支；
- 改善产品和服务质量与定价，提高顾客满意度，提升效益；
- 优化市场营销与宣传策略，在成本最小化的前提下通过准确的信息和宣传精准定位顾客；
- 优化库存管理和资源分配，利用优化驱动模型将资源在需要的时候运送到需要的地点，同时将成本降到最低；
- 在处理顾客关系或顾客相关问题时，为员工提供所需信息以便进行更好更快的决策。

“分析学”一词在短期内迅速成为一个备受关注的热词，在很多情况下替代了原来使用的术语，例如，情报、挖掘和发现。“商业智能”现在变成了“商务分析学”；“顾客信息”变成了“顾客分析学”；“网页挖掘”变成了“网页分析学”；“知识发现”变成了“数据分析学”，等等。由于现代数据（我们也称之为大数据）有着数量大、种类多、流动速度快的特点，因而数据分析学需要大量的计算。而分析项目所用的工具、技术、运算必须采用各行业技术水平最先进的方法，涉及到管理科学、计算机科学、统计学、数据科学以及数学等领域的知识。图 1—1 展示了与分析学和大数据相关的“词汇云”。

图 1—1　分析与大数据词汇云

分析学与分析有区别吗

尽管分析学（analytics）与分析（analysis）二者之间常常可以互用，但二者并不完全相同。

从根本上讲，分析指的是将一个问题分解为若干个小问题，再对各个小问题采取各个击破的方法解决问题。这种方法往往适用于对整个系统的调查不方便或是不切实际，需要将其分解成更基本的部分的情况。一旦完成了部分的优化和检验，就可以通过合成方式将部分合成为整体。

而分析学指的是利用一系列研究方法、技术和相关工具发现新知识，解决复杂问题，进行更好更快的决策。从本质上来讲，分析学是一种多方面、多学科交融的解决复杂问题的方法。分析学利用数据以及数学建模来理解我们所生活的世界。虽然分析学在研究活动的不同阶段需要进行不同种类的分析，但它并不仅仅是分析，还包括合成等其他许多任务和步骤。

数据挖掘该归何处

数据挖掘指在大型数据组蕴涵的规律和联系中发现新知识的过程。分析学的目的是将数据或事实转化为具体可行的信息或情报，数据挖掘正是协助其达成该目标的关键。数据挖掘比分析学存在的时间要长得多，至少比现代意义上的分析学历史还要悠久。当分析学成为决策支持和问题解决技巧中首当其冲的术语时，数据挖掘则在更为广阔的领域里发挥着作用，包括判别变量（例如，市场篮子分析）之间关系的描述性研究以及建立模型估计相关变量未来值。本章稍后会介绍，

在分析学的相关术语中，数据挖掘在从简单到复杂的各个层次上都扮演着至关重要的角色。

分析学何以突然受到追捧

如今，分析学是一个炙手可热的新词，无论你看哪本商业周刊或杂志，都能发现关于分析学或是关于分析学如何改变管理决策的文章。它是循证管理（evidence-based management，指基于事实或数据进行的决策活动）的一个新标签。但是，分析学何以变得如此受欢迎？时机为何偏偏是现在？这一名气的来源有三：需求、可用性与可负担性、文化变化。

来自商业活动的需求

众所周知，当今的商业再不会有“一成不变”的说法。过去的商业竞争往往是本地级、区域级、国家级，而如今的商业竞争已扩展为全球级别。无论是大型、中型还是小型商业，每个企业都承担着全球竞争的重担。过去曾在其地理范围内保护企业的关税与交通成本壁垒现在已经逐渐失去效力。除了全球竞争，消费者的需求也越来越高，甚至前者很可能导致了后者问题的激化。消费者想要以最低的价格出售最高品质的商品与服务，并且尽可能地在最短时间内送达。企业的成功乃至存活取决于其灵活机智的行动，及其管理者顺应市场驱动力（例如，快速发现并处理问题，快速发现并抓住机会）及时采取解决问题的最佳方案。因此，基于事实、质量更高、速度更快的决策显现出了前所未有的重要性。面对无法改变的市场环境，数据分析学将帮助管理者获取信息，更好更快地做出决策，提高企业的市场地位。目前，分析学已经被广泛看做是在国际商业活动中帮助管理者

的救命稻草。

数据的可用性与可负担性

随着科技的进步，软件硬件的成本不断下降，企业能够大规模地收集数据。基于一系列感应器和 RFID 系统的自动化数据收集系统，大大增加了企业数据的数量和质量，再加上社交媒体等互联网技术提供了内容更为丰富的数据信息，如今企业收集的数据已经远远超过了他们能够处理的数据规模。正如俗语所说：“他们沉浸在数据的海洋却仍然渴望知识。”

随着数据收集技术的进步，数据处理技术也得到了长足的发展。目前的处理工具有数不胜数的处理器以及大规模的存储能力，因此能够在合理的时间范围内（通常是即时）迅速处理大量复杂数据。软硬件技术的进步同时也反映在定价上面，此类处理系统的价格一降再降。除了购买处理系统，企业还可以使用服务型软（硬）件商业模式，允许企业（尤其是中小型财力有限的企业）租借分析技术，并根据其使用的部分付费。

文化改变

企业从很早开始就致力于摒弃传统的由灵感决定的决策方式，转而使用基于事实的新时代决策方法。业内大多数领军企业都曾有意识地进行基于数据或事实的商业活动。随着时代的进步，企业掌握的数据越来越多，高新技术设施越来越发达，使得这种观念上的转变正以人们意想不到的速度在发生。随着新一代有着量化思维的管理者取代“婴儿潮”一代管理者，这样基于事实的管理观念转变将会越来越多。

分析学的应用领域

商务分析学的浪潮虽然方兴未艾，却在很多方面得到了大量应用，使用范围几乎涵盖了商业活动的全部领域。举例来说，在顾客关系管理方面，我们有许许多多成功的案例，讲述企业如何通过制定精妙的模型来定位新客户、寻找追加销售（up-sell）或交叉销售（cross-sell）的机会、辨识消耗量大的顾客。企业利用社交媒体分析学以及情感分析，试图控制公众对其商品服务与品牌的舆论导向。产品检测、缓和风险、产品定价、优化营销策略、融资计划、员工留任、新人招聘甚至保险估计都属于分析学在商业方面的应用范围。从商业报告到数据存储，从数据挖掘到优化分析，在任何一个商业活动中都可能找到分析学应用的身影。

分析学面临的主要挑战

尽管分析学的优势是显而易见的，但应用分析学的主要弊端也导致了许多企业仍然踯躅不前，其弊端包括以下几点。

- **分析学人才**。数据分析师，即能将数据转化为实际信息或情报的数据天才，在市场上十分罕见，找到真正适合的优秀人才十分困难。分析学本身是一门新兴学科，其人才资源也正在发展，许多大学推出了本科与硕士项目以弥补这一人才空缺。随着分析学的不断升温，企业需要将大数据转变为信息和知识以应对实际问题，对这方面的人才需求也会越来越大。
- **文化**。俗话说：“江山易改，本性难移”。企业要从以灵感为基准进行决策的传统管理方式转变为基于数据和科学模型进行管理决策、收集企业知

识的现代管理方式是十分困难的。人们往往不喜欢改变。改变往往意味着放弃我们过去已有或已经掌握的知识，重新学习如何进行工作，意味着我们经年积累的知识（也可以说是能力）有朝一日会全部或部分丧失。文化的转变也许是采用新型管理模式中最困难的一部分。

- **投资回报**。应用分析学的另一困难是很难确定其投资回报。分析学项目十分复杂，成本也较高，其投资回报并不能够马上见效，许多企业管理者在进行分析学投资时都会遇到重重阻碍，特别是大规模的投资。分析学的投资回报能够超过成本吗？如果可以，那何时能够开始盈利？要将分析学带来的好处转化为可测量的数据是十分困难的，它带来的大部分利益都是无形且作用于企业整体的。若使用得当，分析学可以使整个企业得以转型，将其提升到一个新的高度。要使投资回报量化，将企业活动向分析相关的管理活动转变，需要一系列有形与无形因素的共同作用。
- **数据**。现代媒体对数据有着极大的信心，将其视为改善企业行为的无价之宝。从很大程度上来说这是正确的，尤其当企业知道如何合理使用这些数据，其价值就更加珍贵。然而，对于那些不知该如何使用数据的企业而言，大数据反而成为了挑战。大数据不仅仅是数量“大”，而且还是非结构化的，其发展速度之快使传统收集处理的方式都望尘莫及，而且往往未经处理、杂乱无章。企业要在分析学上取得优势，就要具备经过深思熟虑的大数据处理方案，将数据及时转化为有价值的信息或情报。
- **科学技术**。尽管科学技术正逐渐变得可行、可用以及相对而言可负担，对于一些技术能力稍弱的企业，科学技术仍然是其使用分析学的一个壁垒。虽然进行数据分析的基础设施价格不再高不可攀，但其成本仍然是很大的

此为试读，需要完整PDF请访问：www.ertongbook.com