大数据丛书@国家十二五重点图书出版规划项目

Broadview
www.broadview.com.cn

发现数据之美

数据分析原理与实践

彭鸿涛 聂磊 著

DISCOVER
THE BEAUTY OF DATA

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据丛书@国家十二五重点图书出版规划项目

发现数据之美

数据分析原理与实践

彭鸿涛 聂磊 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

大数据时代已经来临，这将引起深刻的行业变革。但是，大数据的真意在于数据分析，即从繁多的数据中找出洞见，并将其应用于实际决策中，以产生更明智的决策。这是一个看起来简单、做起来较难的事情。

本书从一个自底向上的角度，全面地阐述了数据分析所涉及的知识和技术，对于经典算法和工具的介绍也不止于泛泛而谈，而是加入了作者的经验和理解。所谓自底向上的角度，即从数据分析实践开始时所需要的数据准备、数据探查、数据再处理等，到经典的统计分析和数据挖掘算法及应用，还讲述了模型的部署，优化技术的引入，最终到决策自动化。

本书对企业管理者、数据分析从业者及高校的学生都有参考意义。管理者能看到一个较全面的数据分析的阐述，明确自身的需求；从业者能看到经验的总结及经典工具的使用；高校学生能看到数据分析所涉及的知识，对数据分析有一个全面的认识。

图书在版编目（CIP）数据

发现数据之美：数据分析管理与实践 / 彭鸿涛，聂春霞 著 —北京：电子工业出版社，2014.8
（大数据丛书）
ISBN 978-7-121-23558-0

Ⅰ. ①发… Ⅱ. ①彭… ②聂… Ⅲ. ①数据处理Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2014)第 132179 号

策划编辑：刘　皎
责任编辑：徐津平
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：20.75　字数：416 千字
版　　次：2014 年 8 月第 1 版
印　　次：2014 年 8 月第 1 次印刷
印　　数：3000 册　　定价：75.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

推荐序一

“数据正成为一项重要的自然资源”，正如 IBM CEO 罗睿兰所言，“越来越多的决策将基于预测分析，而不是直觉或者经验。”大数据及分析将深刻改变企业的运营方式，企业也会因为大数据及分析的应用而获得竞争优势。

首先，大数据及分析将改变企业做决策的方式。从数据中挖掘洞察并用于指导实际的决策，是数据分析的价值所在。面对海量的、多样的、快速增长变化的及准确的数据，分析技术的成功应用将极大升华数据的价值，企业的各种决策将更智慧。

其次，大数据及分析将改变企业创造价值的方式，即把分析应用于每一件事情上，创造出更好且独特的价值。在不同行业的各种活动中，如设计、生产、仓储、营销、金融等，分析技术的运用会带来非常积极的结果。如在设计之初就依赖分析技术研究用户的喜好，又如在营销活动中预测用户流失的可能性，这些例子已举不胜举。对企业而言，分析技术正为企业创造前所未有的价值。

最后，大数据及分析将改变企业向每一个客户提供价值的方式。在准确的时间，将准确的产品推荐给准确的客户，是一个典型的营销问题。分析技术则会帮助企业寻找出合适的潜在客户，针对每一个客户提供定制化的产品和服务，这将极大提高用户体验和营销成功率。

IBM 非常重视在大数据及分析领域的发展，每年在该领域研发的投入高达 30~40 亿美元。同时通过不断地精准收购，吸收了许多非常优秀的员工和技术，已经成为大数据及分析领域的领导者，是目前业界唯一一家集咨询、服务、软硬件的综合实力于一身，能够提供全面的解决方案的公司。在电信、交通、交通、医疗、零售等行业领域，IBM 帮助客户成功实施了诸多的大数据及分析案例，这不但成就了客户的商业目标，还积累了丰富的行业洞察及经验。

在数据分析领域，IBM SPSS 四十多年以来一直是重要的领跑者，具有一系列经典的工具和广泛的用户群。IBM SPSS 数据分析的能力也被应用到众多成功的业务分析解决方

案中，给客户带来了巨大的价值。在大数据时代，分析技术的突飞猛进注定不会缺少 IBM SPSS 的贡献。

本书由 IBM SPSS 资深软件工程师和数据分析师撰写，里面分享了他们多年来对数据分析的经验和理解，并阐述了数据分析所涉及的主要内容和过程，包括“数据收集—数据预处理—经典的统计分析—经典的数据挖掘方法—优化技术的引入—决策自动化”等。希望这些内容能对决策者或者初学者有所帮助，同时也欢迎读者能对数据分析的相关内容与 IBM 的工程师和数据分析师们进行深入探讨。

王 阳 博士

IBM 全球副总裁兼中国开发中心总经理

推荐序二

自从 2004 年创建了 IBM SPSS 西安实验室之后，我一直关注着那里的发展，每年都会回国内看看，特别是关心和鼓励那里年轻工程师的成长。两年前，当鸿涛告诉我，他计划写本关于业务分析的书时，我惊喜于他的勇气，同时又担心他在繁忙的工作之余，很难坚持完成这样一本系统介绍业务分析技术和方法的著作。没想到今天就收到了他和聂磊的书稿，并邀我作序，欣喜之余，感叹他们的艰苦付出，希望他们在业务分析领域能有更大的发展。

三十多年前，当我在美国初次进入统计分析和数据挖掘领域时，它仅仅是由少数统计专家和数学精英掌握的神奇的高端工具。随着计算机技术、人工智能和机器学习等领域的发展，业务分析已经逐渐渗透到人们的工作、学习和生活中，并将在未来很长一段时间里，作为企业、组织和个人不可或缺的决策帮手。在过去的几年中，业务分析相关的各种名词，如大数据、预测分析、认知分析、决策自动化等一直闪烁着耀眼的光芒，越来越成为人们关注的焦点。我深信大放异彩的业务分析绝不会是昙花一现的时髦潮流，而是历史发展的必然，并且将深深地影响人们的生活。

业务分析这件事已经被研究、应用了很多年，给人类社会带来了巨大的价值，但这还仅仅只是开始。数据的积累、算法的不断精进、各种运算平台的巨大改进，都给业务分析的大发展提供了肥沃的土壤。业务分析方面各种时髦名词的涌现，就如同一棵茁壮成长的小树上不断结出果实，一片欣欣向荣。在我多年的业务分析从业经历中，从没出现过今天这样的景象——人们到处谈论各种业务分析的事情。

基于业务分析，各种新奇的应用将不断涌现，也会对人们的生活产生更深刻的影响。智慧城市和精准营销就是两个典型的例子，前者从市政建设、公共服务等方面，应用业务分析提供更智慧的方案和决策；后者则根据用户的特质和行为定制产品或服务，显著提高用户体验。还有太多例子，相信读者会逐步感受到这些应用带来的便利。

数据与算法，构成了业务分析的主干，并且相辅相成，缺一不可。有时人们看重数据的价值，那是因为合适的数据往往很难获取；有时人们又看重算法，那是因为在没有真正

得到由业务分析而来的洞见之前，人们无法判断算法的优劣。无论如何，二者都是不可或缺的。对于业务分析者来说，数据与算法就是其工作的主要内容。

人们渴望成功部署业务分析的应用，但业务分析依然具有较高的门槛，它从数据、人员技能和应用环境等方面提出了较高的要求，特别是对业务分析从业者的要求更高。我很欣喜地看到，鸿涛和聂磊能写出这样一本书来，采用自底向上（从数据探查到决策自动化）、理论结合工具的视角，全面介绍业务分析的主要方面。这是一本很难得的书籍，对管理者、业务分析从业者及学生都有很大的参考价值。感谢作者的艰苦付出，希望看到更多的年轻人积极地投身到业务分析的实践中，为这一领域的发展继续添砖加瓦。

石静云

IBM SPSS 首席统计师，IBM DE，IBM SPSS 西安实验室创始人

前　言

这个世界每天都在发生各种奇妙的事情，特别是当很多人每天坐在屏幕前，不断敲击键盘的时候，各种新奇的事物以前所未有的速度不断涌现。

多年以前人们可能不会想到，手机会以非常智能的方式出现，但当乔布斯说“Today, we are-invent the phone”时，手机智能化便成为现实。人与计算机之间的语言交流，在无数科幻电影中被反复演绎，当 IBM 的 Watson 再次战胜人类的时候，这似乎就在眼前。

仔细想想，这真是一件有意思的事情。人们将各种电子元器件集成起来，并将各种计算包含其中，然后定义出各种用于人类与计算机进行交流的计算机编程语言，这便成了一个放大、成就人们各种奇思妙想的利器。

计算机的广泛应用自不必细说，人们已经得到了其诸多的便利。在商业应用领域，计算机能帮人们做很多事情，比如将各种数据存放起来，自动化地处理各种业务，生成各种报表以供人们参考，等等。可以说，人们已经离不开计算机技术的帮助。然而，这一切的重要基础是数据。

我们已经进入了大数据时代

早在二十年前，尼葛洛庞蒂就在《数字化生存》中描述和预言了当今的生活——人们已经离不开数字，人们的生活已经与数字息息相关。数字代表了一定的数据信息，是各种定性指标的表达，人们与数字已经紧密地捆绑在了一起。

数字化是计算机用来表述事物的方式，或至微至细，或宏观概括。就像人类的语言，可以表达很多复杂事物。如今，数字化已经不是一个技术问题，而是一个意愿问题。大量的事物本来就存在，且都能用数字的方式来表达，问题是人们是否愿意来数字化它们。如今一个显而易见的趋势是人们对数据的渴望似乎是无止境的，即使数据是巨量的，人们似乎也愿意存储和处理。

海量数据的产生一方面是积累而来，另一方面是人们开始愿意并且能够收集、存储和处理它们。在过去的几十年间，不论是企业、机构还是国家，都在努力地收集和存储数据。从企业层面来讲，数据的收集和积累大多来自于信息化系统的应用，如各种业务系统等。维克托·迈尔·舍恩伯格和肯尼思·库克耶合著的《大数据时代》中宣称：世界的本质就是数据；并且基于了解世界的渴望，人们不断地扩大数据的收集规模。数据已经成为了一种商业资本，一项重要的经济投入，可以创造新的经济利益。事实上，一旦思维转变过来，数据就能被巧妙地用来激发新产品和新型服务。数据的奥妙只为谦逊、愿意聆听且掌握了聆听手段的人所知。

很多看起来很酷的应用都必须有一定数量的数据基础。这也非常暗合于计算机世界的形态：计算机的绝大多数组件都在处理各种数据而不是在产生各种数据，字节进入计算单元，然后流出另外一组经过计算的字节。计算单元就是消费数据的组件，而数据则是需要收集和积累的。如果没有足够的数据，有再多的计算单元也只是个摆设。

很久以前就有这样的观点：数据将成为比自然资源更重要的资源。这个观点强调了数据中包含着具有巨大价值的信息、知识，这些信息和知识的应用会带来非常可观的价值。一个非常简单的例子就能说明数据的重要性。例如，有两家不同的公司，一家从一开始就不断收集和积累各种数据，并且愿意不断扩大数据收集的规模，那么它就有可能从数据中找到一些知识：什么样的用户会喜欢什么样的产品，他们可能对哪些营销活动感兴趣，等等；另外一家公司则不注重数据的积累和收集，显然它不大可能从残缺的、低质量的数据中找到有用的洞见。这洞见具有巨大的使用价值，比其拥有的其他资源更重要。

数据分析的意义所在

数据分析是拥有数据之后要做的最有意义的事情。数据分析是个比较广泛的概念，数据挖掘、统计分析、商业智能（Business Intelligence）、业务分析等都属于数据分析的范畴。数据分析的最终目的是从数据中找出有用的信息和知识，以支持、帮助决策。其基本的步骤有数据探查、数据清洗、数据转化和建模等。

数据挖掘是个使用频度非常高的名字，并且经常和很多名词混用，如人工智能、机器学习和商业智能等。其实数据挖掘最为显著的特征是发现，即从冗繁的数据中找到有用的模式（pattern）。这个寻找的过程可能是人工智能和机器学习的实践过程。

统计分析是关于数据收集、组织、分析、解释和描述的科学。统计分析的方法可以分

为三个：描述性的统计分析、探查性的数据分析和证实性的数据分析。描述性的统计分析用来给出给定数据集合的主要特征，如样本大小等；探查性的数据分析主要用来发现数据的一些特征，如数据的分布等；证实性的数据分析用来验证一些假设是否成立，如假设检验等。

相对来说，商业智能是比数据挖掘和统计分析大很多的概念。商业智能包含了一系列的理论、方法论、过程、架构和技术，将数据转化为有实际意义的信息，这些信息能够帮助决策者确定和开发各种市场机会，企业能够利用这些机会巩固和发展市场地位。商业智能在具体实施过程中也需要引入一些统计分析和数据挖掘的应用。

业务分析这个名字在最近的使用频度很高，其含义在利用数据的层次上较商业智能更进一步。业务分析代表了从数据中持续探查、挖掘，从而得到洞察以帮助人们进行决策的一系列技巧、技术、应用和实施，其着重强调了利用数据和数据分析去发现新的洞察，以提升人们的决策质量。

通常，商业智能利用的工具是查询、报告、OLAP（On-Line Analytical Processing，联机分析处理）和预警，回答一些诸如“过去发生了什么”、“发生了多少”、“发生的频率”、“问题出在哪里”、“下一步应采取哪些措施”的问题。业务分析则着重利用数据分析工具来回答“为什么会发生这样的问题”、“接下来还可能发生什么”、“能够采取的最优措施是什么”等问题。如之前提到的，从利用数据的难度这个层次来说，业务分析较商业智能高。

从传统意义上讲，预测分析是利用统计分析、数据挖掘等技术的一个子方法，其对历史数据进行分析，从而对未来可能发生的事情进行预测。然而，近年来人们对其含义进行了大量扩充，其包含了分析很多相关内容，如描述性的建模、预测建模、决策建模、优化，等等。这种扩大对于偏信传统概念的人来说，有点困难。我也曾对预测分析含义的扩充感到不适应，但事实确实发生了。这可能也是为了强调业务分析中最重要的特色，与商业智能有所区别吧。

近几年，还有一个比较新的概念——决策管理，它是业务分析这个大的范畴下的一个分支。如果说商业智能主要完成决策支持的话，属于业务分析的决策管理则强调了决策自动化，即根据数据所代表的情况自动做出决策，而不是人为的。决策自动化是一个很复杂的过程，涉及分析、建模等技术，还有一个很重要的就是优化技术的引入。优化技术能够回答类似“什么样的决策才是最优的决策”这样的问题。让机器做决策，听起来是个很神奇的事情，但仔细想想，像苹果的 Siri、IBM 的 Waston 等能够和人进行交流的应用出现后，看起来很神奇的事情如今也可成为现实。虽然决策管理和 Siri、Waston 没有可比性，但是通过一系列的数据分析，让机器在特定的领域自动做出决策，已经有很多实现案例了。

以上这些分类，只是非常粗略地概述了一些数据分析的分类，从这些分类中我们能看到数据分析的益处。

这是一本关于 SPSS 的书籍

SPSS 在计算机世界是一个有很长历史的公司，早在 1968 年，几个创始人发布了 Statistical Package for the Social Sciences（简称 SPSS）的第一个版本。这个产品就是后来大家耳熟能详的统计分析的 SPSS 软件。在 1975 年，以 SPSS 这个名称注册了公司。

在 2000 年前后，SPSS 软件有了新的含义——Statistical Product and Service Solutions。在 2008 年，SPSS 公司对已有产品进行重新命名，将原来的 SPSS 软件命名为 SPSS Statistics，这样一来，意思更明确，不至于让粗心的用户分不清 SPSS 公司和 SPSS 软件。2009 年 IBM 收购了 SPSS，此时 SPSS Statistics 的名字又变成了 IBM SPSS Statistics。

除了著名的 IBM SPSS Statistics，SPSS 公司还有一些其他知名软件，如 IBM SPSS Modeler、IBM SPSS Data Collection、IBM Analytical Decision Management，等等。每一个产品都有其特长及专注解决的方面，特别是最近几年，SPSS 在企业级业务分析的应用上，投入了很多。除此之外，对于大数据的分析，SPSS 的动作也非常之大（为便于读者阅读，我们在后续描述中，会用 Statistics 指代 IBM SPSS Statistics，用 Modeler 指代 IBM SPSS Modeler，用 ADM 指代 IBM Analytical Decision Management）。

目前国内大多数读者对 SPSS 的概念还停留在 SPSS 的经典工具上，对 SPSS 的其他能力并不了解，特别是 SPSS 针对决策管理、优化技术的引入等方面的能力。就目前来说，市面上还没有一本全面介绍 SPSS 的书（从数据分析到决策管理），而这个过程涉及数个软件。

我试图从“数据分析”到“决策管理”给出一个概要描述，并且突出 SPSS 工具的特点。让国内的用户能够较为全面地了解这个过程所涉及的要点，对设计、实施业务分析相关的应用有所帮助。

你将从这本书中得到什么

市面上已经有很多与统计分析、数据挖掘等相关的书籍，那么本书的特色是什么呢？人们对于数据分析的印象大多是“从数据中找到真知灼见并将其应用于实际的问题解决

中”。“从数据中找到真知灼见”其实包含了很多内容，比如针对问题的不同而采用统计分析的相关技术（如假设检验），或者采用数据挖掘中的典型相关技术（如聚类）。除了技术的不同，我们还需要考虑方法论的问题，例如，如何挑选数据、如何探查数据的质量、该选择哪种模型哪种算法、模型的部署和更新，等等。研究这些问题需要知识的准备和时间的积累。本书就试图给出一个全景式的描述，按照我的经验和理解对典型问题逐一探讨。

“将真知灼见应用于实际的问题解决中”也是一个值得深入探讨的问题。最为浅显的想法就是将模型部署，让模型返回一些预测值等类似的值，作为进一步决策的新依据。模型的部署也需要考虑一个方法论的问题，如本书中讨论的 CRISP-DM 参考模型。

但是，如果模型仅能返回一些预测值，只能说我们达到了决策支持的阶段，能不能让模型直接返回决策建议呢？或者直接实现决策自动化呢？这就属于决策管理的范畴。决策管理绝不仅仅是一个模型复杂化的问题，也有方法论的因素，这是本书讨论的重点之一。

总之，我试图给出一个全景式的描述，对上述方面做一些介绍。数据分析的应用级别是分层次的，最简单的是数据探查，只看看数据的分布、特征等；其次是统计分析和数据挖掘，这些都属于决策支持的范畴。除此之外，要采用优化技术做出最优决策，实现决策自动化的决策管理，又是比较高的应用层次了。本书以自底向上的叙述方式，对上述方面都进行了描述。初学者、管理者，或者数据分析从业人员，都可以通过本书对数据分析的重要方面和阶段有一个清晰的了解。初学者可以了解数据分析有哪些主要的技术需要学习，管理者可以根据企业自身的情况了解其真实的需求是什么——简单的统计分析还是决策自动化，数据分析从业人员可以将本书作为一本参考书，了解相关的产品。

欢迎指正

我在 IBM ADM 项目组成立之初就加入了这个团队，至今已有 6 年。由于项目的需要，我对统计分析和数据挖掘都有所了解，也经历了数个企业级决策管理应用的开发和部署。但是，几年来，我发现人们对数据分析、特别是决策管理的理解和重视远远不足，所以萌生了写书来介绍的想法。

在写书的过程中，我查阅了很多相关材料，由于我在 SPSS 的产品线上工作，可以查阅到各种 SPSS 的文档，所以作者试图结合自己工作的便利，全面地介绍数据分析的相关方面，并且深入浅出地介绍这些晦涩的内容。

即便如此，我深信一些描述错误是不可避免的，读者若发现任何值得商榷的地方，真心期望读者能够指出，我将在今后的写作中改进。

感谢

在吃晚饭时，我说我要写书。父母和妻子先是惊喜，然后是鼓励。在接下来的一年多里，每逢周末，他们都帮我腾出大量时间，我深信他们付出了很多，非常感谢他们。

还要感谢我的合作写书人，聂磊，他是一个非常聪明的帅小伙，当我邀请他一起写作时，他欣然同意，并积极完成了本书第 3 章、第 7 章和第 8 章的写作。在本书的写作过程中，我俩经常积极讨论、相互学习，我们非常享受这样的过程！

感谢 IBM 全球副总裁兼中国开发中心总经理王阳博士，能在百忙之中为本书作序。感谢 SPSS 的首席统计师、IBM DE、SPSS 西安公司的创始人石静云女士，当我告诉她我要写一本关于 SPSS 的书时，她非常高兴并答应给本书写序，这大大增加了我写作的信心。

感谢 IBM CDL BA 主管、资深经理吉燕勇的鼓励和肯定。当我告诉他写书这件事情的时候，他非常肯定这件事情，并积极帮忙安排各项事宜。他的帮助和鼓励，至关重要。还要感谢 IBM CDL 的资深经理王俊波、蒋俭，他们的帮助也很大。另外，非常感谢我的经理李慨的支持。

感谢来自 IBM 大学合作部，美丽、聪慧的杨敏同事关于如何出书给予的帮助，以及 IBM Academic Initiative 社团的同事史俊辉的给力支持，他在不断地帮助和协调关于出书的各项事宜。IBM Academic Initiative 社团的其他同事也给予了很大帮助，没有他们的帮助，本书可能不会这么快与读者见面。

非常感谢电子工业出版社的编辑刘皎，她提出了非常有用的意见，在她的帮助下，本书得以进入“十二五国家重点图书出版规划项目”。这对没有出书经验的作者来说，是莫大的鼓舞和支持！

彭鸿涛

2014 年 5 月于西安

目　录