纵观大数据

建模、分析及应用

牛 琨◎著

北京邮电大学出版社
www.buptpress.com

数据，是比文字更早出现的工具，它帮助人类不断拓展对客观世界的认知，是社会生活中不可缺少的关键要素。置身大数据时代的我们，享受着大数据分析带来的种种便利的同时，也被可能的隐私泄漏所困扰。为了更好的描述、理解和管理数据，我们需要掌握一定的相关知识，本书将成为打开数据分析之门的这把钥匙。

牛琨，女，博士，副教授，硕士研究生导师，2007 年 7 月毕业于北京邮电大学网络与交换技术国家重点实验室，获通信与信息系统专业博士学位。第一批北京市高校青年英才，北京邮电大学“周炯槃优秀青年教师励志奖”获得者。致力于大数据分析挖掘、智能信息处理及行业应用等相关研究。发表中英文学术论文 30 余篇，授权发明专利 2 项。主研包括国家自然科学基金、国家科技支撑计划在内的项目 20 余项。

授课经验包括：《数据挖掘》《大数据分析与挖掘》《智能信息处理》《数据分析工具使用》《数据仓库与知识发现》《大数据技术实践》《离散数学》《数据分析那些事》《形式语言与自动机》《Data Warehouse and Knowledge Discovery》

纵观大数据：
建模、分析及应用

THE BIG DATA

牛　琨◎著

北京邮电大学出版社
www. buptpress. com

内容简介

大数据分析是个入门容易但精专颇难的领域。本书以大数据分析为主线，以电信行业应用为背景，以一线操作者为对象，系统阐述了大数据分析的理论、方法和实践。从思维创新开始，依次介绍了大数据挖掘、经营分析和营销策划三个主题。本书聚合了作者多年实战操作的经验，加上执教和内训的总结，辅以真实的案例，对于电信行业的分析师而言是一部较为实用的工具类书籍。

图书在版编目（CIP）数据

纵观大数据：建模、分析及应用 / 牛琨著. -- 北京：北京邮电大学出版社，2017.9

ISBN 978-7-5635-5130-9

Ⅰ. ①纵… Ⅱ. ①牛… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2017）第148154号

书　　名：纵观大数据：建模、分析及应用

著作责任者：牛　琨　著

责 任 编 辑：满志文　穆晓寒

出 版 发 行：北京邮电大学出版社

社　　址：北京市海淀区西土城路 10 号（邮编：100876）

发 行 部：电话：010-62282185　传真：010-62283578

E-mail：publish@bupt.edu.cn

经　　销：各地新华书店

印　　刷：

开　　本：720 mm × 1 000 mm　1/16

印　　张：18.75

字　　数：316千字

版　　次：2017 年9月第 1 版　2017 年9月第 1 次印刷

ISBN 978-7-5635-5130-9　　定价：48.00元

前言

Scan the big data

数据，是比文字出现更早的工具，它帮助人类不断拓展对客观世界的认知，是社会生活中不可缺少的关键要素。身处大数据时代的我们，更加受到数据及其分析模型带来的影响，既有各种生活的便利，情景化的舒爽，也有隐私泄露的不快。为了更好地掌握数据，正确地分析数据，精准地描述规律，我们必须掌握一定的数据分析知识，而本书将是打开这扇门的一把钥匙。

执教十年，经历了从数据挖掘到大数据的云卷云舒，一代代的技术更迭，不变的是对数据知识探索的执着初心。但是，咨询者众，待解惑者也不少，一一解答既无效率又没效果，因此在去年萌生了写本书的想法。与理论型书籍不同，本书的方法论是来自传统理论但充分考虑了实战环境而进行了适配性的修订。希望读者在阅读时要注意，因地制宜，随机应变，重神不重形，切不可机械照搬。

第一章讲创新思维。这是因为数据分析的起点就是头脑，是思考，想做好数据分析，打开头脑是最重要的，没有之一。

第二章至第八章讲数据分析。从数据本身开始，评述了统计分析、数据挖掘和大数据等分析技术，还介绍了由浅入深的三种主要工具软件的使用技巧，非常适合有一定操作经验但亟须进阶的操作者。

第九章至第十六章则聚焦经营分析。经营分析是企业经营最重要的分析工具组合，可能融合了非常复杂的分析技术。本书抽

丝剥茧，先从理论高度系统介绍了定性和定量方法，再从主题和专题两个角度来演示经营分析过程，最后通过案例来说明具体步骤。

第十七章至第十九章介绍营销策划。一切公司的目标都是赚钱，不以盈利为目标的组织是公益和公共服务部门。数据分析最核心的价值就体现在对营销策划的强力支撑上。另外，本书重心在于介绍数据分析如何用在营销策划中，而不在于制定营销策略和最终决策。

限于行业服务背景和本人的水平与经验，书中不足之处，恳请读者和专家不吝赐教。

作　者

第一章

思维能力特训

头脑是人类最大的财富。万物之灵能主宰蓝色星球，靠的是地表最强的思维能力。虽然我们每天都运用头脑，但却很少思考其运作机制。我们经常评价某人聪明或是愚钝，而背后的一切都源于对个体头脑的开发和利用程度。大数据分析是头脑使用的高级阶段，为了更好地掌控世界，我们先从思维能力开始讲起。

第一节　大脑如何转弯

在作者的少年时代，“脑筋急转弯”风靡一时，早熟的少男少女借用这一技巧卖弄着自己的智力同时完成搭讪这一社交领域的高难度动作。为了更好地达成不可言传之目的，有好事者干脆去书店买一本脑筋急转弯大全之类的书来背。典型的问题如“冬瓜、黄瓜、西瓜、南瓜都能吃，什么瓜不能吃？”答案是“傻瓜”。

这样的问题可称为弄巧之技，因为这类问题根本无法准确、可复现地衡量一个人的智商水平。公认的智商测试应包括对观察、记忆、想象、创造、分析判断、思维、应变、推理等能力的测量，而与这类脑筋急转弯问题的相关系数

趋近于 0。这是因为，几乎所有人都可以通过机械地背诵来应对此类问题，而记忆力只是智商中的一部分；另外，此类问题无法通过严密的逻辑推导来找到答案，越是逻辑能力强的人往往越容易被愚弄。

其实，人类的思维能力分为两大部分，这两大部分缺一不可。一部分是线性思维，又称硬性思维，在这种情况下 1+1 必须等于 2，二进制下的 01+01 必须等于 10，如果不遵循这个原则，则目前计算机系统的基础架构将全部坍塌。线性思维能力强的人非常适合学习理工科，尤其是当程序员。另一部分是非线性思维，又称软性思维，此时 1+1 等于多少呢？我们想等于几都可以，非线性思维能力强的人适合进行艺术类创作，擅长文史艺术类学科，就业方向可以是作家编剧、广告设计师等。

值得庆幸的是，还有这样一些人，同时具备很强的线性思维能力和非线性思维能力，兼具理性思辨和超凡想象力，留下一些震古烁今的历史影响力。亚里士多德是哲学家、科学家和教育家，其著作构建了西方哲学的第一个广泛系统，包含道德、美学、逻辑和科学、政治和玄学，绝对的百科全书式的超级牛人。张衡以发明浑天仪闻名于世，头衔有天文学家、数学家、发明家、地理学家、文学家，与司马相如、扬雄、班固并称“汉赋四大家”，这个跨界范围不小，让后人无法模仿和超越。还有机械妖孽达·芬奇，作为画家、天文学家、发明家、建筑师，擅长雕刻、音乐、发明、建筑，通晓数学、生理、物理、天文、地质，具有超越当时科技 30～50 年的技术实力。

看到这里，如果有人提出，每个普通人都能通过一定的训练加强自己并不擅长的另一方面的思维能力，从而获得更大的成就，肯定是一碗非常给力的心灵鸡汤。趁着鸡汤还没凉，接下来，我们需要一把巨大的智慧之匙。

第二节　智慧之匙

这把智慧的钥匙，其实就掌握在我们自己手里，它的名字叫作创新思维。这个所谓的创新思维，是真的可以训练出来的吗？

先看三个小例子。

① Take one from nine，you can get ten，Why?（九中去一得十，为何？）

② What is One−half of thirteen?（十三的一半是多少？）

③下式不是一个有效的数学表达式：2+7−118=129。请在上式中加一条直线，使之成为一个有效的数学表达式。

第一个问题，在作者十几年的内训生涯几千个学员中，只有一位立刻、准确地答出了正确答案，作者的感受是此人确实天赋异禀。那我们这些貌似没什么天赋的普通人如何是好？

解：

第一步，已知 9−1=8，9+1=10，即按照某种逻辑关系，正负号反转。最常用的阿拉伯数字、汉字、英语，一一否决……（严密的逻辑推导）

第二步，有什么语言的数字编码方式具有反转正负号的逻辑呢？（这一步至关重要）

第三步，这题应该考的是常用的，而非小语种或罕见的数字编码模式。（限定条件）

第四步，除了阿拉伯数字，就属罗马数字在西方最常用（这可是一道英语题），早期的钟表和当代的很多高档手表仍然沿用。而罗马数字中的“左减右加”恰好就实现了反转正负号！（没有文科知识行吗？）

第五步，9 在罗马数字中表示为 IX，本身就是 10−1 的意思，那么把 X（表示 10）左边的 I（表示 1）去掉，恰好得到了 X（10），而这里面的 take from（去）词组不是指代 minus（减），实际说的是 remove（移）。

上述解答过程告诉我们，严密的逻辑推导，在解决大部分问题的时候是有效的。

第二个问题答上来的学员较多，“6.5”“1”“3”“thir”“teen”等都是正确答案。这个问题表明，发散性思维也是解决某些问题的利器。

第三个问题难住了很多人，尤其是笃信逻辑推导的那些人，他们试图通过小时候擅长的速算 24 经验来套用，这样会很自然地陷入思维的陷阱。答案很简单，一根小小的斜线，把“=”变为“≠”即可，思维的关键点在于“有效的数学表达式”可不一定都是“等式”。

通过三个小例子，我们知道严密的逻辑推导、发散性思维和突破性思维都

能有效地解决问题，这正是人类创新思维的起点。而且幸运的是，它可以被训练出来。

第三节　人人皆可创新

提出人类需求层次理论的著名心理学家马斯洛认为，创造性分为“特殊才能的创造性（Special Talent Creativity）”和“自我实现的创造性（Self-actualizing Creativity）”。

爱迪生名垂科学史，靠的是“特殊才能的创造性”，也就是天才的创新能力。我们从小就知道，他发明的“电灯”是经过上千次材料实验测试出来的。问题是，爱迪生是否具备把这么多种材料加工成灯丝的动手能力。

显然，背后另有高人，那就是爱迪生的团队，他们就是不那么有名的美国技工约翰·沃特、英国车工巴契拉、瑞士钟表匠巴格曼等人，这些人是爱迪生两千多项发明、一千多项专利的坚实基础。这些伟大的工匠，面对人类历史上前所未有的新挑战，发挥了强大的创新能力，这就是“自我实现的创造性”，用实践证明了“人人皆可创新”。

第四节　阻碍创新的因素

既然人人都可以创新，那么为什么创新大师们如此卓越而大多数人却显得缺乏创新能力？这是因为，我们生活中存在很多阻碍创新的因素。

创新的第一个枷锁是“思维标准化”。

思维标准化的第一种表现是“功能固着”，即严格遵守对象与功能函数的对应关系。在第二次世界大战的北非战场上，由于技术兵器的严重缺乏，面对