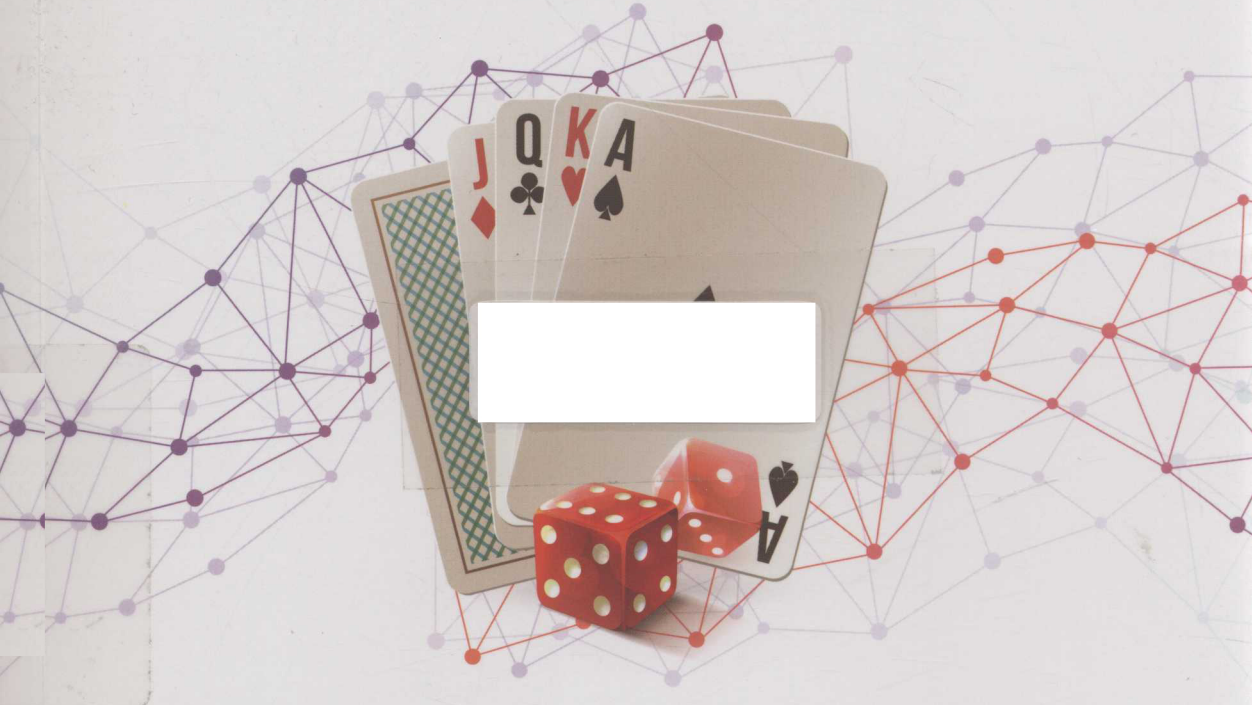
Broadview®
www.broadview.com.cn

CDA数据分析师系列丛书

大数据思维

从掷骰子到纸牌屋

马继华 著



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

CDA数据分析师 系列丛书

大数据思维

从掷骰子到纸牌屋

马继华 著

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

内 容 简 介

数据分析不在于你掌握了多少先进的软件工具，也不在于你拥有多么高智商的头脑，而是要靠更大视野、更宽角度和更具有逻辑性的思维。本书不是一本介绍大数据概念的流行读物，也不是开讲编程工具高深理论的专业教材，而是立足于大数据之上的思维模式的普及。读者不需要任何统计学知识，也没必要掌握复杂的公式与算法，在最通俗易懂的案例介绍和娓娓道来中就可以轻松理解大数据分析的基本模式与方法。

作为读者，你可以是大中专院校的数据分析专业学生，也可以是企事业单位的经营分析人员，或者是任何行业任何职业中喜欢“头头是道”的分析爱好者。开卷有益，即便你从来不需要大数据，也可以从本书中领悟到思维魔力，因此让工作与生活更充满智慧与乐趣。

图书在版编目（CIP）数据

大数据思维：从掷骰子到纸牌屋 / 马继华著. —北京：电子工业出版社，2016.7
（CDA 数据分析师系列丛书）
ISBN 978-7-121-29407-5

Ⅰ. ①大… Ⅱ. ①马… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 163950 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
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：720×1000　1/16　印张：17.5　字数：281 千字
版　　次：2016 年 7 月第 1 版
印　　次：2016 年 7 月第 1 次印刷
定　　价：55.00 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）88254888，88258888。

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。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

前言

早就想写一本关于数据分析的书，最主要的原因就是，自己是统计专业毕业，又从事过多年数据分析的工作。工作几经变迁，现在已经很少用软件重操旧业，但却越来越感觉到数据分析的重要性。

经常看网络、电视和报纸上的很多分析，在信誓旦旦的说教与言之凿凿的数字之外，很多却是惨不忍睹的分析过程，甚至说是误人子弟也不为过。因为自媒体的流行，很多人根本没有基本的分析方法和技巧，在违背常理的情况下做出了很多奇异的解释，将大家引导到错误的方向。

最为可笑的，曾经有一次看到某知名报纸上的文章，分析的是中国信息分类领域的两家互联网巨头：58 同城与赶集网（这两家公司在 2015 年宣布合并）。当时，58 同城刚刚上市，这家报纸的专栏作者发表了一篇针对性的分析文章，文中称，他查阅了 ALEX 网站，58 同城的流量排名在世界网站的第 300 名，而赶集网排名是第 900 名。于是，这位作者就果断地下结论说，以上数据足以证明 58 同城的网络流量是赶集网的 3 倍。呜呼，如此分析竟然逃过了多少编辑的眼睛，甚至还

被众多读者接受，是多么可悲！

在实际工作中，一些人虽然科班毕业，通晓各种分析工具，甚至对各种各样的软件如数家珍，编程造模轻车熟路，但却对具体的分析套路与方法形同陌路，只能机械刻板地对数字结论进行解读。实际上，这样的数据分析还不如不做，错误的分析和错误的解读同样都是害人不浅。

当然，由于分析能力不到位，让自己吃亏上当丢人的案例更是不胜枚举。中国足协就是典型案例。2013 年，人所共知的原因，中国足球终于迎来了出人头地的机会，中国足协更是喜出望外。为了配合隆重的节日气氛，也是要彰显一下中国足球有雄起的能力，中国足协费尽心思地组织了一场国际足球友谊赛。

中国足协应该在邀请友谊赛的对手方面煞费苦心。邀请德国队？肯定不行，严谨的德国人不明就里的职业精神会破坏比赛气氛。邀请西班牙队？鼎盛时期的西班牙与中国队比赛也必须让自己有一个可以接受的成绩，否则被人笑掉大牙。于是，中国足球邀请了我们的近邻，泰国队，可怕的比赛开始了。估计包括中国足协官员在内的中国球迷都没有想到，一场友谊赛进了 6 个球，更重要的是，我们只进了一个，泰国队进了 5 个。

如果中国足协进行了充分的数据分析，也许就会避免这场悲剧的发生。历史数据证明，中国队此前已经多年没有胜过泰国队。如今的中国队不再是以前的那支“中国头球队”，依靠身高与体重就可以战胜东南亚球队，几年来学西班牙控制脚下球的中国队既没有学到技术，也忘记了本分，对付泰国这样的小老虎已经心有余而力不足。或者，这场比赛还不如邀请韩国，场面也不会失控。

如果我们非要挖苦一下数学水平奇差的中国足协，那也是可以的。

因为，某年某月某日的世界杯外围赛亚洲区预选赛，中国与黎巴嫩同组，在最后一轮比净胜球决定出线的关键时刻，中国足协竟然鬼使神差地算错了账。当全场球迷因为中国队 7：0 战胜中国香港而成功惊险获得出线权而欢呼的时候，足协才明白过来，8：0 才出线，我们已经被淘汰出局。这样的数据分析能力怎有能力让中国足球拿下大力神杯？

从历史上看，中国一直不是一个靠数据化进行管理的国家，我们太多的中庸之道和模糊分辨，“好好好”、“是是是”、“差不多”，贯穿着经济和社会管理的始终，这个模式也对中国的国家统计局产生着潜移默化的影响，也直接造成了人们对国家统计机构数字的不信任。

数据分析是每个人生活与工作的基本功，小时候对父母的察言观色也是在分析，长大以后的相亲娶妻也要分析，工作中的汇报决策更需要分析，炒股理财也离不开分析。数据分析无处不在，数据分析无时不在，数据分析伴随我们生命的始终。

我们生活的世界变化是如此之快。电力引入美国 46 年后，才覆盖 1/4 国民；电话花了 35 年；电视机 26 年；宽带呢？只用了 6 年。2007 年，数码世界容纳了 2810 亿 GB 的数据，全球平均每人 45GB，数码资料首次超越保存空间总量，目前，互联网每小时处理的数据量已经超过 1EB。

要给美国国会图书馆填满逾 5700 万份手稿、2900 万册书籍和期刊、1200 万张照片及其他，需时 2 个世纪，现在全球每日生成的数码资料几乎是这些的 100 倍。人类 5000 年的文字记载总共是 5EB，今后每年将产生的数字内容超过 1000EB。

我们所拥有的数据量在海量暴增，我们认识世界的水平也在不断提高。大数据时代来了，我们的思维是不是也应该有所改变？

目录

第1章

大数据与人脑的较量

BAT 为何如此了解我们

开篇，我们来讲一个简单的问题，你知道腾讯的 QQ 与微信的重要区别是什么吗？

现在的中国人，如果有人问你，你用 QQ 或者微信吗？估计很少有人会回答“否”。因为，QQ 或者微信已经深入到我们生活的各个方面，成为工作与生活的必需品。

可是，如果问你，QQ 与微信有什么区别？估计很多人答不上来。或者有人会说，QQ 有空间，微信有朋友圈；还有人会说，QQ 能穿衣服，微信没有。这些也是差别，但却没看到本质。

通过大数据的分析，我们也许能得到更为靠谱的答案。我们试着再提示一下，你在使用 QQ 的时候，使用频率最高的词是什么？这个问题如果问腾讯，腾讯可以通过系统地查询很容易地得到答案。我们普通用户实际上也能说得出来。一些人说，QQ 上使用频率最高的词是“呵呵”或者“哈哈”，还有“哦”，但更多人会联想到一个词，那就是“在吗？”

是的，我们需要的答案就是“在吗”。因为，我们可以对比一下，你在使用微信的时候，还会经常使用“在吗”吗？答案是，不会。

以上的分析，我们就是使用了最简单的词频分析，以最简单的数数的方式获得了最佳的分析路径，因为一句“在吗”就能充分地展示 QQ 与微信的本质差别。

我们通过进一步分析可知，因为 QQ 是互联网时代的产物，后来与移动互联网相结合，因此，QQ 有电脑客户端，也有手机客户端。大家使用 QQ 的时候之所以经常第一句说“在吗”，是因为我们无法判断

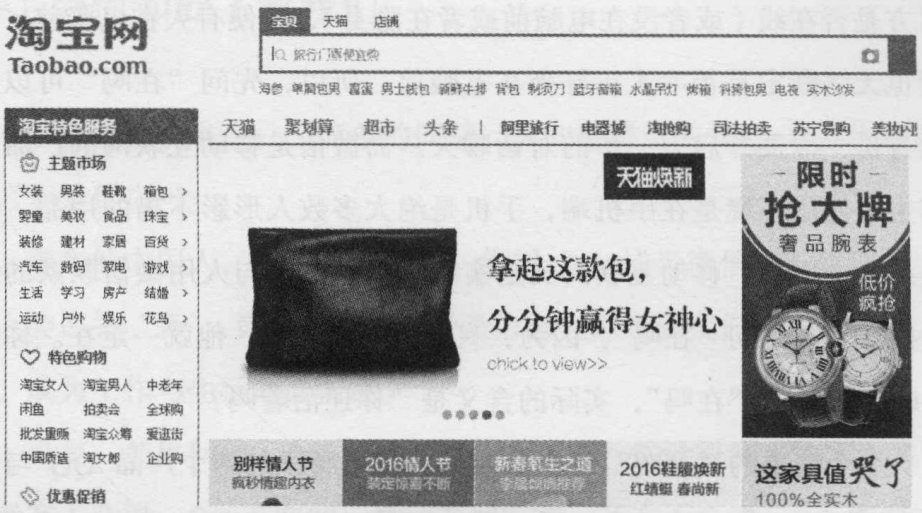
对方是否在线（或者没在电脑前或者在隐身），即便有人在电脑前，我们也无法断定是否本人正好坐在电脑前，所以，先问“在吗”可以确认身份，以便开启下一步的对话聊天。而微信是移动互联网的产品，其主要使用环境是在手机端，手机是绝大多数人形影不离的用品，而且是个人用品，移动互联网又是实时在线，我们与人用微信联系的时候根本无需先问“在吗”，因为，只要这个人还在，他就一定在。你这个时候问对方“在吗”，实际的含义是“你还活着吗？”

一个简单的“在吗”就形象地刻画出了腾讯的两个产品 QQ 与微信的代差，也找到了互联网与移动互联网产品分析的钥匙，这是多么神奇？

接下来，如果你是中国移动的员工，或者是通信行业的分析师，如果要分析中国移动的飞信产品，那与之进行对比分析的产品应该是 QQ 还是微信？很简单，应该是 QQ，而不是同样有一个“信”字的微信，因为，飞信与 QQ 同样都是互联网时代的产品，都拥有电脑客户端和手机客户端，而且都可以同时在线。

分析就是如此，只要你找到了窍门，四两拨千斤，简单的方法可以解释大道理，何必非要扎在数据堆里当无头苍蝇呢？

对用户的使用行为研究最充分的，无疑是阿里巴巴。很多人都发现，只要你打开淘宝，首页上的推荐就让你欲罢不能，特别是网页中间那张跳动的大图，怎么看都是自己想要的商品。是的，淘宝说要实现千人千面，每个人看到的网页都是不一样的，因为那个页面就是根据你最近的搜索、下单等历史行为结合你的各种资料进行“定制”的。



有这样一个小故事：一个连锁商店，专门有一个铺子卖婴幼儿产品。因为客户信息很多，就发现当人怀孕之后，行为会出现改变。比如会更多选择没有香味的洗发水，买营养品的时候口味也和怀孕前有不同。商店便可以根据客人购买行为的变化，预测是否可能怀孕了，然后给可能怀孕的客人寄婴幼儿产品广告，说买我的尿布吧，买我的奶粉吧。一天，一个父亲很愤怒地过来说，“我女儿还在高中，你们现在天天给她寄婴儿尿布、奶粉的广告，什么意思？你鼓励未婚怀孕啊？”然后商场说，“对不起，我们搞错了！”过了一个星期，这个爸爸又回来，说：“对不起，我搞错了，我女儿已经向我坦白了，她真的怀孕了。”

在现代企业经营中，电子商务都非常重视针对性的产品推荐，比如淘宝，更具有大数据应用意义的就是信用评价，比如芝麻信用分。芝麻信用公布了基本的计算模型，综合考虑了个人用户的信用历史、行为偏好、履约能力、身份特质、人脉关系五个维度的信息，没有任何一个单项信息能够直接或完全决定个人的芝麻分，其五个维度包含

的内容举例如下：

（1）信用历史：过往信用账户还款记录及信用账户历史；

（2）行为偏好：在购物、缴费、转账、理财等活动中的偏好及稳定性；

（3）履约能力：享用各类信用服务并确保及时履约；

（4）身份特质：在使用相关服务过程中留下的足够丰富和可靠的个人基本信息；

（5）人脉关系：好友的身份特征，以及跟好友互动的程度。

根据这个计算模型，我们大概可以总结出一些规律，能够帮助个人提高自己的信用得分。

（1）你要至少办一张信用卡，并经常在网上进行消费，特别重要的是要记得按时还款，如果你是使用支付宝进行按时还款，那么肯定会增加信用分。

（2）即便你有钱，也要使用下“花呗”、小额信用贷款等，并设置自动还款，保证你的账户里有这笔钱到时候准时还上，如果你不设置自动还款却能按时手动还款，那信用的分数肯定会提高。

（3）使用支付宝进行慈善捐款，如果是每年每月都坚持下来，即便数额不大，也会对信用分数帮助不小，因为理论认为做慈善的人信用比较好。

（4）发发红包，不管是定向发还是抢红包，都表明你乐善好施并且不差钱，信用不会差。

（5）多交几个有钱的朋友，并经常在网络上互动，如果发现谁经常信用卡不还，赶快绝交，至少也要在网络上不要来往。

（6）在网上买东西，要记得收到货物之后尽早地主动支付而不是

等系统默认付款，最好要给买家进行评价，如果能不厌其烦地多写几句话，就更好了。

（7）网购时的收货地址要力争保持稳定，如果你是租房或经常变换居住地，或者是房子太多经常换地方住，那也要选最稳定的地址来收货，比如办公室的地址，或者直接是一个居住稳定的朋友代收。经常换地方收网购商品对信用影响很大。

（8）如果可能，就把自己的网购账户的信息多填点，那些多人或家人公用一个账号的自然在个人信用评分上会受到影响。

（9）如果你有钱，在各互联网公司的理财产品里放些闲钱，既能保障收益，也可以让自己看起来是个有钱人。

怎么样？数据分析很有用吧，不仅可以帮助企业了解客户需求，还可以帮助客户找到针对性地提升自己社会信用的方法。掌握简单的科学的数据分析方法，对所有人都是必要的。

战争是各种矛盾最为激烈的表达，而数据分析更是战场指挥员不可缺少的工具。最为著名的案例就是，林彪靠战利品分析意外地快速结束了辽沈战役。

据资料记载，在中国革命战争年代的十大元帅中，林彪非常有特点，从白山黑水到天涯海角，战功卓著。据说，林彪从红军带兵时起，身上就有个小本子，上面记载着每次战斗的缴获、歼敌数量，其实这就是在积累大数据。1948 年的辽沈战役，是决定国共命运的大决战开端。每天深夜，林彪都在东北野战军前线指挥所里听取军情汇报，由值班参谋读出下属各个纵队、师、团用电台报告的当日战况和缴获情况，而林彪则认真细致地记录着他的大数据：每支部队歼敌多少、俘虏多少；缴获的火炮多少、车辆多少、枪支多少、物资多少……作为

司令员，林彪的要求很细，俘虏要分清军官和士兵，缴获的枪支，要统计出机枪、长枪、短枪，击毁和缴获尚能使用的汽车，也要分出大小和类别。

一天深夜，值班参谋正在读着下面某师上报下属部队的战报，说他们的部队碰到了一个难度不大的胡家窝棚遭遇战，歼敌部分，其余逃走。与其他之前所读的战报看上去并无明显异样，值班参谋就这样读着读着，林彪突然叫了一声“停!”。林彪接连问了三句：“为什么那里缴获的短枪与长枪的比例比其他战斗略高？”“为什么那里缴获和击毁的小车与大车的比例比其他战斗略高？”“为什么在那里俘虏和击毙的军官与士兵的比例比其他战斗略高？”林彪不等别人回答，指着地图上的那个点说：“我猜想，不，我断定！敌人的指挥所就在这里!”结果，部队集中兵力攻击，很快抓获了廖耀湘。从大批杂乱无序的数据中将信息集中、提炼，分析出研究对象的内在规律，找到蛛丝马迹的异常变动，从而为决策提供最强支撑。

神奇的不仅仅是林彪，还有柳传志，更是擅长根据蛛丝马迹的数字做出自己的判断。据说，柳传志的创业起因非常具有传奇色彩，只是因为看了一张再普通不过的报纸。有时候，借助敏锐的数据分析能力就可以发现别人不易察觉的变化，从而让自己的人生大不相同。

1978 年 11 月 27 日，中国科学院计算所 34 岁的工程技术员柳传志按时上班，走进办公室前他先到传达室拎了一个热水瓶，跟老保安开了几句玩笑，然后从写着自己名字的信格里取出了当日的《人民日报》，一般来说他整个上午都将在读报中度过。20 多年后，他回忆说：

“记得 1978 年，我第一次在《人民日报》上看到一篇关于如何养牛的文章，让我激动不已。自打‘文化大革命’以来，报纸一登就全

是革命，全是斗争，全是社论。在当时养鸡、种菜全被看成是资本主义尾巴，是要被割掉的，而《人民日报》竟然登载养牛的文章，气候真是要变了！”

从现在查阅的资料看，日后创办了赫赫有名的联想集团的柳传志可能有点记忆上的差失。因为在已经泛黄的1978年的《人民日报》中，并没有如何养牛的文章，而有一篇科学养猪的新闻。在这天报纸的第三版上，有一篇长篇报道是“群众创造了加快养猪事业的经验”，上面细致地介绍了广西和北京通县如何提高养猪效益的新办法，如“交售一头可自宰一头”、“实行公有分养的新办法”，等等。柳传志看到的应该是这一篇新闻稿。

不过，是养牛还是养猪似乎并不重要，重要的是，举国之内，确有一批像柳传志这样的人，“春江水暖鸭先知”，他们在这个寒意料峭的早冬，感觉到了季节和时代的变迁（节选自《激荡三十年》）。

还有更神奇的大数据应用，即便是很多美女最喜欢玩的自拍，也有可能成为大数据应用的先驱，因为网络上忽悠你做明星脸对比的，往往都是一些人脸识别的程序在收集素材训练“机器人”。

媒体报道，史上最昂贵的自拍照应该是诞生于2007年。两名美国大兵在伊拉克的军营中玩自拍传到了社交网络上，结果几天之后，这个秘密的驻扎地就遭到了恐怖分子火箭弹的袭击。四架“阿帕奇”直升机惨遭击落，两亿美金灰飞烟灭。美军情报部门百思不得其解，最后才发现：原来是大兵的自拍照中附带了经纬度信息，让“好友”轻易掌握了他们的位置。但是就在2015年，某ISIS成员在其“总部大楼”自拍，并且在社交网络上大肆吹嘘这里的指挥能力有多么“炸裂”。结果一语成谶，22个小时之后，这幢大楼就被美军三枚导弹“强拆”了，