O'REILLY®

社交媒体大数据分析

理解并影响消费者行为

Ask Measure Learn

[美] Lutz Finger
Soumitra Dutta 著
杨旸 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

社交媒体大数据分析
——理解并影响消费者行为

[美] Lutz Finger 著
Soumitra Dutta
杨 旸 译

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

社交媒体大数据分析 ：理解并影响消费者行为 /（美）芬格尔（Lutz Finger），（美）杜塔（Soumitra Dutta）著 ；杨旸译. -- 北京 ：人民邮电出版社，2016.9
ISBN 978-7-115-42084-8

Ⅰ. ①社… Ⅱ. ①芬… ②杜… ③杨… Ⅲ. ①传播媒介－数据处理 Ⅳ. ①G206.2②TP274

中国版本图书馆CIP数据核字(2016)第157318号

版权声明

◆ 著　[美] Lutz Finger　Soumitra Dutta
译　杨　旸
责任编辑　陈冀康
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：20.25
字数：238 千字　2016 年 9 月第 1 版
印数：1－2 500 册　2016 年 9 月河北第 1 次印刷

著作权合同登记号　图字：01-2013-9304 号

定价：59.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

在如今这个大数据时代，个人和企业的社交网站活动越来越活跃，社交网站在数据挖掘方面的挑战和需求变得越来越迫切。

本书从业务角度出发，而不是从技术角度出发，介绍了如何挖掘社交网络数据并且对业务规划进行有用的分析。全书分两个部分共 10 章，第 1 部分介绍了市场营销、销售、公共关系、客户服务、社交化的客户关系管理、与系统博弈、预测、提出恰当问题；第 2 部分介绍了使用正确数据以及定义正确的度量标准。

在社交媒体数据挖掘需求旺盛的今天，本书非常具有实用价值。本书适合数据挖掘技术人员、数据分析师以及市场营销领域的人士参考阅读。

对本书的赞誉

大数据很快就将成为众多公司最宝贵的资源，大数据的运用将成为很多业务模型的必备之物。但此时此刻大多数的公司仅仅只是坐拥成堆的数据，却没有使用它们的明确策略，也不知道如何从这些数据之中提取出它们所蕴含的信息。在这本书之中，两位作者给出了若干个明确的方向，帮助我们理解大数据带来的挑战以及如何提出恰当的问题来处理它们。

——蒂姆·韦伯（Tim Weber），英国广播公司新闻互动部前商务与技术编辑

对个人以及企业来说，大数据意味着一个可以提升自己的绝佳机会。本书将带领读者踏上从理解大数据的概念到从大数据中导出价值的奇妙之旅。

——N. R. 娜瑞娅娜·穆斯（N. R. Narayana Murthy），Infosys公司联合创始人

大数据是一个谜团——对于同一个问题有太多的答案。而这本书从根本上改变了这个现象，并且阅读它也是愉快至极。

——本·维瑞恩（Ben Verwaayen），阿尔卡特–朗讯公司前CEO

本书是那些想将大数据引入自己公司的人的必读之选。它就大数据之中的问题给出了一个比较公允而全面的看法。本书的作者不仅对大数据本身有深入理解，他们更能看到世界范围内未来在这一领域可能的变化。

——艾利克斯·皮特兰（Alex Pentland），麻省理工学院教授

大数据可能是一个大家都在谈论的热门话题，但是真正理解它含义的人却屈指可数。本书对人们如何理解和运用大数据提供了莫大的帮助。

——弗兰克·布朗（J.Frank Brown），General Atlantic Partners 公司总经理兼 COO

无论是在社交媒体还是大数据中，洞见都是一切分析必不可少的组成部分。商界领袖需要洞见，而本书正是引导他们获得洞见。本书是一个非常有用的框架和案例集，恰当地使用它就可以避免那些因分析而分析的情况。

——罗伊科·雷·米尔（Loic Le Meur），Leweb 会议 CEO

本书把围绕着大数据的所有迷雾都吹得烟消云散。本书也就如何使用大数据来获取真正的商业价值做出了阐述。对于那些正在面对着将大数据应用于自己公司商业策略相关挑战的公司来讲，本书可以说是一本必读书籍。

——阿勒特·艾瑞斯（Annet Aris），欧洲工商管理学院战略学副教授

我认为这是一本应时应景的书。数据革命才刚刚开始，每一个与大数据打交道的或者试图成为数据驱动型业务的公司，都在飞速膨胀的数据海洋之中，搜寻数据的第四重维度—— 价值。我发现本书阐述了很多对我来说受益匪浅的观点，因此强烈地推荐给那些想要定义数据驱动策略的人阅读。

——乌维·维斯（Uwe Weiss），Blue Yonder 公司 CEO

对于每一个组织来说，本书都是一个绝佳的框架以及必读的书籍。很多时候，我们看到商务智能得出的结论都是从数据以及图表出发的。我们

常常忘了，最本质的东西是，我们找寻的不是图表而是洞见。本书帮助我们把关注重点放到商业价值上，其中包括了很多实际的案列分析。通过这种方式，作者向我们展示了数据科学的趣味性和简单易用性。

——斯蒂芬·波佩尔（Stephan Poppel），Tchibo 公司电子商务主管

本书以实例的形式帮助我们将纯粹的数据变为可执行的洞见。对于那些和社交媒体以及大数据分析打交道的人来说，本书是必读书籍。两位作者展示了他们的核心思想，即询问恰当问题的重要性。

——伯恩·奥格里贝里（Björn Ognibeni），Buzzrank 公司 CEO

最近一段时间，科技的突破性发展以及社交媒体的广泛应用让我们可以获取前所未有的海量数据。但与此同时也带来问题：在这海量数据之中，我们无法抓取出有效的数据，并做出恰当的选择。本书通过着手处理这些问题来帮助商界领袖们提升自己业务的价值。

——伯恩·赫尔曼（Björn Hermann），Compass 公司 CEO

作者简介

卢茨·芬格尔（Lutz Finger）是一位在社交媒体以及大数据分析领域的权威人士，担任社交媒体 LinkedIn 的数据分析主管。他之前是位于新加坡的数据分析公司 Fisheye Analytics 的 CEO 和联合创始人，该公司每月会为众多的政府以及非政府机构处理 70TB 的公共社交媒体数据。卢茨更为戴尔欧洲建立了一个多达 700 人的销售中心，该中心后来成为爱立信的移动软件的孵化中心。卢茨也是很多在欧洲以及美国的数据中心机构的顾问以及董事会成员。

作为一位顶级的数据分析专家，芬格尔经常在加利福尼亚大学伯克利分校等各大顶级名校做关于数据分析的演讲。他拥有欧洲工商管理学院的 MBA 学位，以及柏林工业大学的量子物理学的理学硕士学位。

苏米特拉·杜德（Soumitra Dutta）是康奈尔大学克里特斯·杰克森管理学院的系主任。他曾以教授的身份就职于欧洲工商管理学院的 eLab 实验室，这是一个世界顶级的研究生商学院，其校区分布于枫丹白露、新加坡以及阿布扎比。杜德也是 Fisheye Analytics 公司的联合创始人以及前主席。他在新技术对商业的影响方面是具有相当权威性的，社交媒体和社交网络以及数字化经济的战略规划都是他的特长。他也是两本在技术与发明方面有影响力的报告的联合编辑以及作者，这两本报告分别是《全球信息技术报告》（与世界经济论坛联合出版）以及《全球发明索引》（与世界知识产权组织联合出版），这些报告在全球范围内帮助了多个政府就技术和创新策略进行了评估以及规划。苏米特拉获得了位于新德里的印度科技学院电气工程与计算机科学学士学位。他还获得了工商管理硕士学位以及计算机科学的硕士学位，还是加利福尼亚大学伯克利分校的计算机科学博士。

译者序

大数据、云、社交网络，还有很多其他时髦但又让你不知所云的新名词在最近几年之中不断地出现在我们的视野之中，一遍又一遍地冲击着我们的想法、观念，甚至生活方式。

而我不止一次地看到来自社会各个领域的人们热情高涨地振臂高呼："我们要社交化，我们要向云迁移，我们要使用大数据。"但奇怪的是，这些如革命者般充满激情的"变革者"中，鲜有能够定义或者描述清楚什么是"大数据"、"云"或者"社交化"的，就更不用谈能理清其中利弊了。而那些能够理解自身业务特点，并能够最大限度地利用这些技术的从业者就更是凤毛麟角了。

这种还没有真正理解即将涉足的领域，就争先恐后、不计后果地向前冲的现象，在我们不长的现代科技史之中却十分常见。20 世纪 90 年代的互联网泡沫是这样，随后的平板化热潮也是这样，更不必提本书中将着重介绍的大数据和社交媒体了。

为什么会产生这种现象，并且还会一而再地重复呢？为什么总会有后来者不厌其烦地重蹈前人的覆辙呢？这是一个值得深思的问题。

每一个社会现象都不会是孤立的，它们的发生都与其所在时代以及时代的特质是密不可分的。我们生活在一个飞速发展的时代，而其发展速度在可以预见的未来之中都将会越来越快。快已经成为了这个时代的特质。生活在这个时代之中，作为渺小的个体，我们只有去适应这种速率以便能够生活下去，并且活得更好。这也就是"赶上最后一班车"的心理能够产生的最重要因素之一。

但是，我们要理解的是，快虽然是这个时代的特质，但这并不代表这个时代是为了快而快。快只是一种手段，而目的是为了又快又好地达到目

标，这才是这个时代所要求我们具备的。往往“又好”这一部分很容易被人们忽略，这也导致了很多人并没有明确的目标就冲入时代的浪潮之中，在经过了许久的冲刷之后才发现自己都不知道为什么身处其中。

如果我们能稍微退一步，先对前方的情况有一个宏观的认识，观察一下那些前人犯过的错误，避开那些危险的陷阱，然后再带着明确的目标前进，这难道不是一种更好的“快”吗？

本书就秉承这种“退一步、看得清”的思路，为我们详细讲述大数据产生的来龙去脉，为我们展示大数据背景下市场营销的诸多误区。让我们在进入这一领域之前，能够对它的全貌有一个整体认识。

本书也提供了针对具体问题的一些可能的解决方案，并详细描述这些解决方案背后所蕴藏的思想。通过这些案例以及思想的学习，相信在读完本书之后，你要建立起一套不错的数据度量体系应该不是什么难事了。

杨　旸

2016 年 2 月

序　言

2011 年 4 月，美国特种部队击毙了基地组织领导人——奥萨马·本·拉登。这次猎捕行动是在本·拉登东躲西藏 10 年之后才成功的。那么，到底是谁知道他藏身何处的呢？

答案让人非常意外：我们都知道。

根据伊利诺伊大学厄本那香槟分校研究员凯文·莉塔瑞（Kalev Leetaru）的一个分析公共新闻的报告的结论，本·拉登的藏身之处可以定位在精度为 200 千米的范围内。这也意味着，这个世界上最隐秘的藏身地竟然被其自身独特的数据特性所出卖了。每一个记者对于本·拉登的藏身地都有自己的看法，而这些意见的集合却可以构建出真正正确的答案。而在这个过程中，并没有开展任何问卷调查，也没有去采访任何记者。这些记者通过自己的文章来揭露本·拉登的藏身地，这就是公共数据以及非结构性数据的威力。其分析结果如图 i-1 所示。

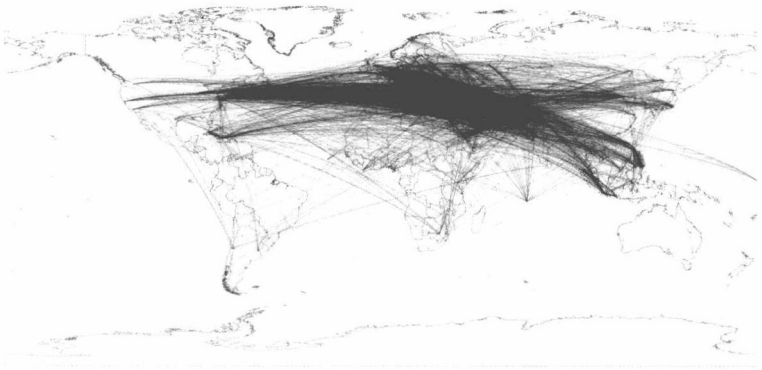

图 i-1　社交媒体之中揭露本·拉登藏身地文章的地理编码（由凯文·莉塔瑞提供）

美军很有可能并不是依赖这些众包形式的智慧得出本·拉登的藏身地的。我们知道美国政府拥有如国家安全局这样的机构，对各种信息源都会

进行监听，包括最高级别政府要员的通话以及普通百姓通过邮件进行的正常交流。但是，其背后的原则都是一样的：行动情报都是从整合到一起的大量的独立个体中得出的，在这个案例中，就是那些看起来随机的数据点。

而这就是我们所知道的大数据的愿景。大数据已经变成了如今数字世界最热门的一个流行词了。大数据涵盖了商务智能之前由于数据量太大而无法进行处理或者以数据库形式进行维护的海量数据，这些数据常常是以百万兆字节甚至千万兆字节为单位的。今天，社交媒体数据就是这种海量数据最极端的体现形式。社交媒体就是一面镜子，它可以揭示我们所想、所需、所爱。我们在互联网上的冲浪、对手机的使用以及我们的地址位置信息都在不断地充实着它的内容，这些数据会产生出关于我们以及我们未来的深刻见解。

作为本书作者的我们，为政府以及非政府组织分析大数据，特别是社交媒体数据，并从中得出有价值的结论。我们已经在这个领域工作有半个世纪之久了。我们也成立了一家叫作 Fisheye Analytics 的公司来为分析社交媒体数据提供配套的软件以及服务，我们每月会为客户分析高达 70TB 的文本数据。我们从中学到的是，大数据的根本并不是数量，而更多的是关于如何从正确的数据之中，分析恰当的问题。

在本书之中，我们会与你一起通过一个个案例来揭示数据的本来面目，以及其中的价值。其重点，无关数据，也无关其大小，重要的是数据能够带来的价值。

数据的第四重维度

数据必然有其战略价值，但今天可用的海量数据以及我们处理这些数据的能力，都变成了一种新的资源形式。从某种意义上来说，数据现今就如同石油和黄金一样。今天的数据热正如得克萨斯州在 20 世纪爆发的石油热或者旧金山在 19 世纪初爆发的淘金热一样。这股热潮孕育出了一个崭新的行业，而且这个行业抓住了所有商业人士的眼球。

我们所说的这种新的资源——大数据，通常被描述为具有“三重维度”。大数据，就是一种拥有海量数据且变化周期短、涵盖多种复杂信息的数据集。而在这 3 种传统定义的“维度”之后，我们在这里加入数据的“第四重维度”——价值。这才是所有人都在苦苦寻觅的东西，而且这也是大数据在今天获得这么大关注度的原因。我们所谈到的大数据，可以是以结构性数据的形式存在的（如金融交易数据）；它也可以以诸如图片或者博客文章这样的非结构性数据的形式存在。如我们在抓捕本·拉登的例子中看到的，它可以是众包式，也可以是独立收集的（如保险公司长期以来所做的事情）。而让人感到矛盾的是，大数据的价值往往是以小数据的形式出现的。例如，“是与非”这样的问题，“我是否应该收购这家公司”或者诸如定位本·拉登藏身地的地理坐标信息。对于挖掘价值来讲，其首要任务就是收缩大数据，只有这样它才能成为“有价值”的数据。

大数据在 21 世纪最火爆的社交媒体的推波助澜之下变得异常流行。我们作为一个集体所作出的讨论、评价、点赞以及社交媒体之中与他人的联系，这一切都变成了数据，而且是数量巨大的数据。如果所有 Facebook 的用户都来自一个国家，那么这个拥有 10 亿以上活跃用户的国家，就将是世界上最大的国家，而 Twitter 的用户在 2013 年早期发送的消息总量每月就达到了数以百万计。目前，这是有史以来我们第一次可以深入地研究人类之间的讨论和交互。每一个 Twitter 或者新浪微博的用户都会留下一个公开的数据轨迹。而我们在 Facebook 或者 Qzone 上进行的私人对话也会揭露很多关于我们自身的内幕信息：我们在搜寻什么？我们读什么书？我们去过什么地方？我们与谁一起？我们吃什么？我们买什么？简而言之，任何你可以想象的人类交互行为，都可以在社交媒体之中被找到，并加以研究。如果我们可以对所有的数据进行挖掘，那么其结果将是无穷无尽的。这就使得从公共数据之中揪出本·拉登的藏身地变成了可能，而社交媒体也就成为了所有秘密的终结者。

与此同时，和我们身边的其他流行科技一样，在大数据以及社交媒体

领域也存在很多炒作。在社交媒体分析兴起的初期，人们坚信只要采用恰当的分析手段，我们可以通过社交媒体说服任何人做任何事。这显然是无稽之谈，即便最好的预测分析也无法拯救那些糟糕的产品。社交媒体有的时候被市场营销人员当作一个魔法武器，他们期望可以为产品构建一个像今天社交媒体这样的热潮。我们会在本书之中解释为什么市场营销人员的这种愿望并没有实现。今天的预测性分析以及社交媒体度量与互联网泡沫时期的网站运营非常类似，那个时候（1996 年）人们以为只要有一个网站就稳操胜券了。而大数据的基本技术和社交媒体分析将成为我们很多人的常用技术，就如电话和互联网一样。

大数据的时代已经到来，它正在改变我们的生活，改变我们做生意的方式。但是，要成功，我们需要的就不仅仅是数据了。正如美军需要决定是使用来自社交媒体的数据，还是使用内部数据，也要在数据的使用上做出取舍。不同企业的数据是具有其自身特性，而互不相同的，这些数据包括从日志文件到 GPS 数据再到客户与机器或者机器与机器之间的通信数据，每个企业都需要决定到底要使用哪种数据。这就更需要我们拥有一个恰当的方式来解构这些数据，然后才能进一步对这些数据进行分析。这就要求我们知道如何把重要信息从炒作信息之中分离出来。这也是本书的目的：指导并告诉你，各种研究得出的真正可以工作的方式，及其背后的原则；并且帮助你将大数据应用于业务中并获得成功。

数据的世界仿若汪洋，各行各业都需要专注于自己的数据集。贯穿本书的，有大量的各种社交媒体度量的例子，这并不是因为我们认定社交媒体就是最有潜力的预测性分析的数据源。事实上，这句话反过来说是成立的。正如我们将在后面的章节之中看到的，社交媒体为我们提供了最难以处理的数据。但是，由于社交媒体对每个人来说几乎都是公开的，而在本书中学到的原理、数据结构和其他东西都可以轻易地移植到你自己对数据的需求和可用性上。那么首先，让我们来看一下大数据的愿景是如果影响我们的业务的。

愿　景

今天的哲学就是数据主义。

——大卫·布鲁克斯（David Brooks）

数据分析的倡导者为我们的生活规划了一个美好的明天。他们承诺我们可以对从价格指数到军情的任何事情进行预测，而他们是正确的。例如，在加利福尼亚的圣克鲁兹，一款软件就可以对一天中最可能发生犯罪的时间、地点进行预测。近来，警察逮捕了正在窥视汽车的两位女性，发现她们是在逃的罪犯，并且携带了大量的毒品。她们没有意识到自己是败给了大数据应用，而这种软件已经成功地阻止了几起犯罪的发生。其实预测性的数据已经在警察部门使用了多年，这里只是列举一个真实的案例而已。

当然，预测性的治安管理只是一个例子，还有很多其他领域都有类似的例子，它们都在向我们展示大数据的威力。如今，我们看到数据应用越来越多。

- Google 使用数据来预测下一波流感的爆发。
- IBM 使用数据来优化斯德哥尔摩的交通，从而让这个城市获得更好的空气质量。
- Zafu、2Style4You 以及其他一些使用自助采集身体数据的公司可以向你推荐适合的衣服。
- 来自新泽西的内科医生杰弗里·勃伦纳（Jeffery Brenner）通过使用医疗付款数据绘制出其所在城市发生最复杂和最昂贵医疗案例的热点地区，这是作为降低医疗费用项目的一部分。
- 美国国家学术改革中心运用了数据挖掘技术来帮助大学生明白他或她最可能取得成功的领域。
- 保险公司为那些在车上安装了 GPS 的客户降低了保费。他们使用数据来预测你是否将会遇到车祸，并以此对你的保险条款进行调整。

- 很多零售商都使用数据来进行推荐，并会进行针对诸如孕妇等人群的定向广告。

如今，我们生活在一个什么都可以被度量的世界。“数据”已经成为了一种新的意识形态。对于这个漫长的旅程，我们才刚刚起步。我们将会度量越来越多的事物，并分析越来越多的信息，这样让我们可以更好地驱动我们的业务以及做出更明智的决定。

这个世界也成为了一个关注焦点。这些关于隐私及社会其他方面的数据库给我们带来什么，目前还不得而知，但著名的批评家杰伦·拉尼尔（Jaron Lanier）就告诫人们不要相信那些所谓的“群众智慧”。而且，对于在警务以及军事情报之中使用的数据也让大众对数据隐私的关注更高了。曾几何时，美国特工甚至通过在电话之中安装窃听装置，来监听他们最亲密的盟友，很多人都认为政府和大公司已经越过雷池了。我们会在本书中对这几个方面的事情进行一些讨论，而目前，数据透明和公开化是能够对抗这些问题的唯一办法。

即便有这么多的警告和担忧，“数据驱动”对很多人来说已经成为了一种新的管理哲学。经济学智库就发布了关于大数据对雇主和雇员的帮助程度，人们的感受是什么的报告（见图 i-2）。大约 2/3 的人觉得大数据会帮助我们找到新的市场机遇，并且可以协助我们做出更好的决定。而有近一半的人觉得大数据可以帮助我们在竞争中获得优势，更有多达 1/3 的人认为大数据可以让我们的经济腾飞并且创造更多的机会。

但是，愿景往往是言过其实的。和很多新出现的技术一样，围绕着大数据也有很多的炒作行为。如果你信以为真，那么这个世界以及你的业务的问题，就都可以通过增加数据量或者调查最新的推文得到完美解决了。时任《连线》杂志主编的克里斯·安德森（Chris Anderson）就发布过一个大胆的声明，称如果我们只要拥有足够的数据，我们就会走向一个“理论的终结”的时代：“Google 的哲学就是，我们并不知道这个页面为什么比其他页面好，如果连入这个页面的统计数据说是，那么就足够了。”未来

是美好的，但是不会那么美好。在第 9 章中，我们会讨论相关性和因果性之间的差异，以及为什么度量因果关系总是那么困难。

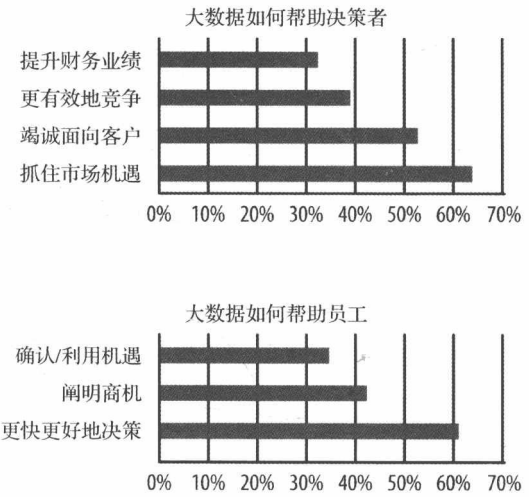

图 i-2　大数据是如何影响业务的

本书就是为打破围绕着大数据的迷雾，看清那些炒作背后的真相而设计的。本书会帮助我们认识到大数据的第四重维度，也就是价值。这个价值并不是指所谓的“群众智慧”，也不是“更多的数据”。为了找出大数据之中的价值，我们就需要拥有正确、精心构建的问题，还要有恰当的方法，以及合适的数据。只有拥有这些之后，我们才能够获得竞争优势。

专注数据

数据实实在在地为我们所做的每一件事提供支持。

——杰夫·魏娜（Jeff Weiner）

你可能会说你从来都是结果导向的。要是结果是可度量的，那么你就必须是数据驱动的。对吧？预测性分析其实并不是什么新东西，随便哪个保险公司都用这种技术很长时间了。那么，为什么我们突然就把关注点转移到了数据以及预测之上了呢？这其中有两个原因：

1．现在有更多公共的可用数据了。