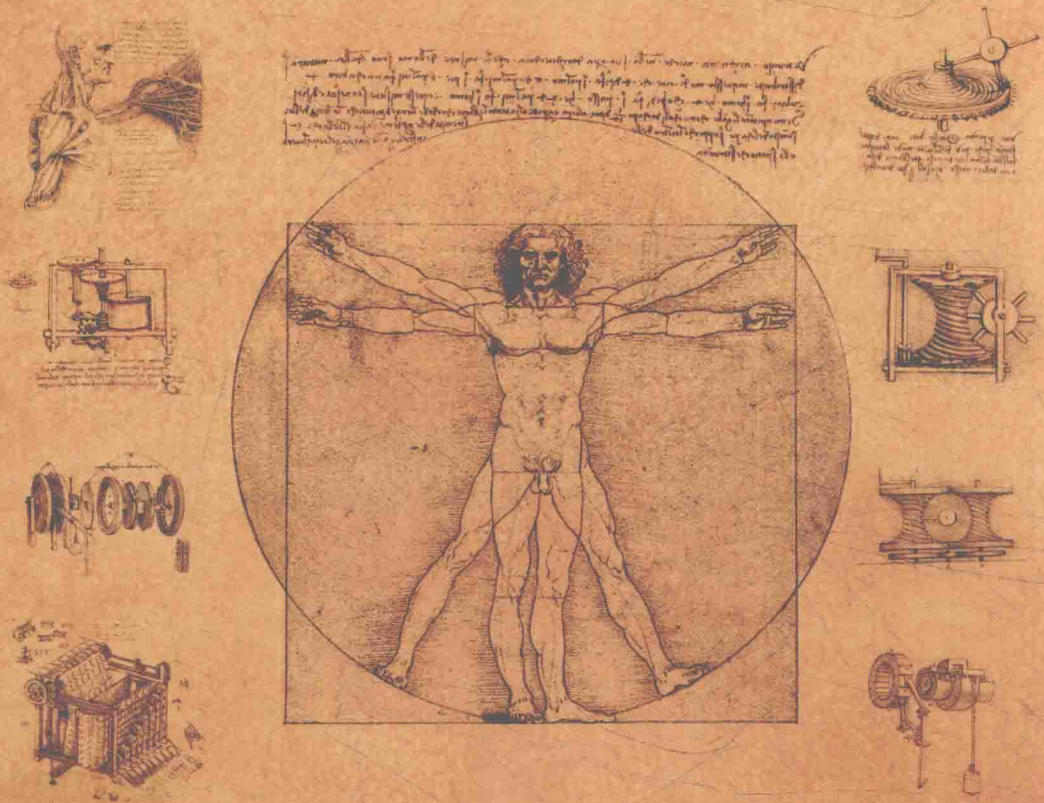
TURING 图灵原创

百度外卖技术委员会主席力作
微观分析大数据智能驱动增长方法论

智能增长

蒋凡 著

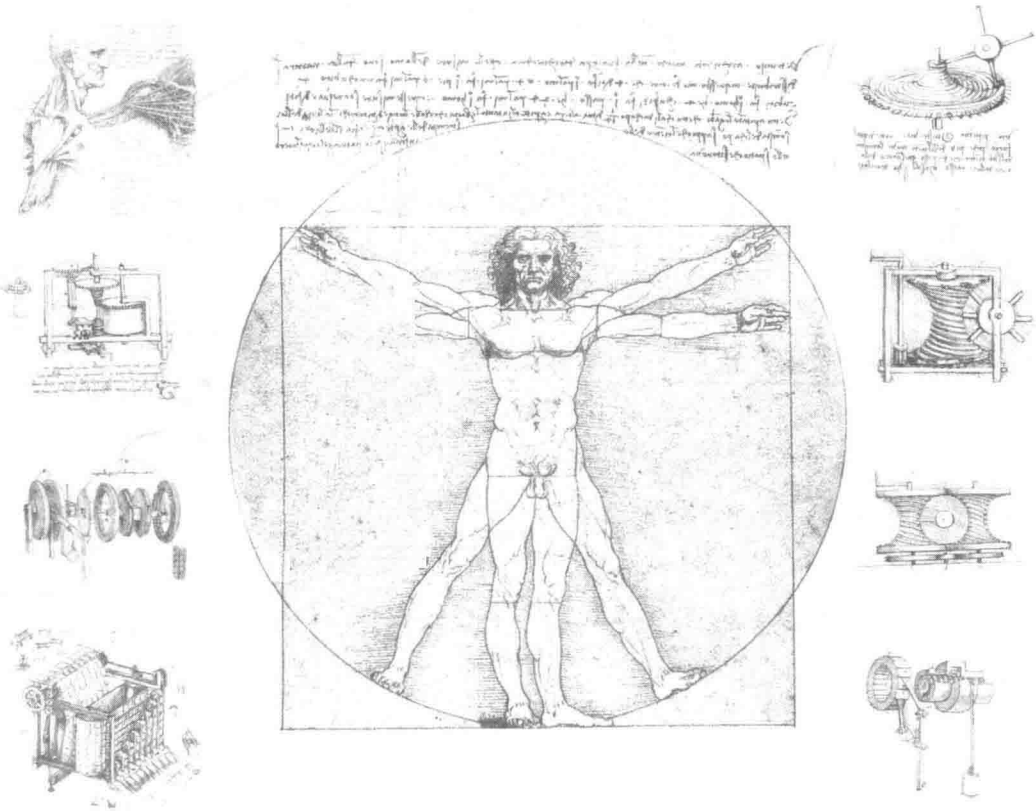


中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原创

智能增长

蒋凡 著



人民邮电出版社
北 京

图书在版编目（CIP）数据

智能增长 / 蒋凡著. -- 北京 : 人民邮电出版社, 2017.12
（图灵原创）
ISBN 978-7-115-47142-0

Ⅰ. ①智… Ⅱ. ①蒋… Ⅲ. ①网络公司－企业管理 Ⅳ. ①F276.6

中国版本图书馆CIP数据核字(2017)第264908号

内 容 提 要

本书共分 5 部分：数据篇从智能驱动增长方案的基础出发，沿着实际工作链条，介绍数据收集、加工、存储和访问；模型篇衔接基础数据到上层应用，全面介绍数据建模，包括生命周期、RFM、AARRR 和地理信息模型；场景篇真正介绍智能增长怎么做，这里以完整的商业运营链条为例拆解 9 大运营场景，从业务、财务和技术的角度详述增长实践；团队篇解释了数据驱动增长在人物层面需要具备的必要因素；结语篇探讨了智能增长对经济的促进作用和作用方式。

本书适合所有从事移动互联网行业研发、产品和运营的人员阅读，对智能增长、互联网 + 经济分析预测的观察者也有借鉴意义。

◆ 著　　蒋 凡
责任编辑　王军花
执行编辑　陈兴璐
责任印制　彭志环

◆ 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
大厂聚鑫印刷有限责任公司印刷

◆ 开本：720×960　1/16
印张：16
字数：343千字　　2017年 12 月第 1 版
印数：1－3 500册　　2017年 12 月河北第 1 次印刷

定价：69.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前　言

智能革命越来越成为时下热门的话题，每个人都迫切想知道人工智能技术正在以及接下来会对人类社会的生产、运输、销售和消费活动产生什么样的影响。原因是不言自明的。最初，人们觉得自己的生活变得方便了一些，可以便捷地查询到详尽的信息和知识，可以轻松建立信任，购买到各种各样的商品，可以几近免费地即时和亲朋好友沟通联络。然后，人们逐渐发现也可以将生活中繁重细琐的事务交给更加智能友好的应用，像差使佣人一般，让手机App为自己预定餐厅、叫出租车、点外卖、呈阅新闻，似乎有一种魔力唤出了蕴藏在地下的无尽宝藏，每个人都在享受这个时代的美好生活。可是这些佣人并没有停下脚步，而是坚定地按照某个方向改善或者修正着我们本以为只能是那样的生活。佣人们开始学着掌控驾驶汽车、查阅病例做出诊断、辅导孩子学习功课、决定银行审核放贷，甚至坐下来和我们对话聊天。

在社会生活中，每一个人既是消费者，同时也是生产者。当我们作为消费者体验到更便捷高效的服务时，一般不会想到那些被服务取代了的生产者已经退出历史舞台。直到有一天，当自己作为一个生产者时，才惶恐地发现，似乎有更先进的替代者已经隐隐出现。于是，一些人学习并实践了几十年的知识和经验不再重要，一些机械重复性的工作逐渐被取代，甚至今天看起来体面稳定的工作也许会在未来的某一天成为历史。还在享受生产效率提升带来丰厚收益的人们不禁开始为明天担忧起来。

有经验的人会联想到历史上曾经发生过的几次技术革命，每次似乎也都经历了人类社会生活水平提升、原有的工作职位消失、新的工作机会出现的过程。然而这一次的变化似乎更加剧烈、更加不可预测。

有识之士会指出大数据和人工智能技术是这场变革的推动力量。的确如此。然而，虚拟存储的数据和程序运行的算法究竟如何改变真实世界面貌，如何驱动物质世界增长？这对很多人来说仍然是模糊不清的。

通常，会有两类图书试图回答这个问题。

- 一类是由行业领袖、趋势分析专家和咨询顾问，从宏观层面分析智能革命的意义、价值和前景，提供很多源于各行各业的统计数据、发展图景和判断预测。读者可以由此加深对大数据和人工智能概念的理解，形成属于自己的理念认识和未来判断。然而，为求甚解的读者还是想知道微观层面发生的事情。
- 一类是由创业公司合伙人、传媒记者和观察家，从经验层面归纳智能革命在某个领域出现过的景况和规律，提供一些通过采访、摘录和体验而来的具体案例和分析结论。读者可以由此将这些经验转化为自身体验，丰富自己对某个行业发生变革的感性认识。不过，欲知其所以然的读者还是希望能掌握到方法论的工具。

本书则从微观层面分析了互联网企业面临的增长业务体量、提高财务收入等问题，探讨了大数据和人工智能技术所能起到的作用和需要采取的做法，从方法论的角度提供了一些普遍适用的衡量体系和思考框架来处理各种纷繁复杂的业务难题和盈亏困境。通过本书，读者能够自行搭建起一套大数据智能驱动增长的行动框架，了解其中各个组成部件的机能和运作规律，并可以结合自身需求进行调整。

因此，本书的基本内容和组成逻辑如下。

- 数据篇从智能驱动增长方案的基础出发，沿着增长型团队内部发掘数据价值的实际工作链条，分成收集、加工、存储和访问这4章，分别回答了数据来源和性质的差异对收集方式的影响问题、数据加工中要注意的忽略噪声和发现隐藏信息的问题、数据存储中考虑到时间与空间因素的不同解决方案的问题，以及衡量数据访问效率的工具、指标和优化设计问题。

- 模型篇则沿着从基础数据到上层应用发生转变的需求，分成生命周期模型、RFM模型、AARRR模型和地理信息模型这4章内容，带领我们进入数据驱动增长的承接阶段。

  每个数据模型自成体系，又相互关联：生命周期模型摆脱传统客户关系概念的束缚，将消费者和经营者看成自有其成长规律的循环体；RFM模型则迅速化繁为简，用三维指标提供了一套普适可用人群分类方案，并能足够灵活地衍生出各种变体；AARRR模型与前两者相似，但又有所不同，结合互联网产品用户行为特点，用形象的漏洞概念强调了留存和转化这两大关键指标，更加便于设计更多行之有效的营销策略；地理信息模型相对来说更强调移动通信和人工智能技术的进步，提供了一整套将实体世界数字化的技术方案，使得移动互联网下的各种应用方案成为可能。

- 场景篇用9章内容重点描述智能增长的关键场景，介绍智能增长怎么做，该如何实践。为了便于理解，再从公司内部的三个不同视角分开理解。

  其中，业务视角主要考虑扩大体量，因此对应了新用户获取、老用户促活、商户运营、商户赋能等各个促进业绩数量增长的领域；财务视角主要考虑增加盈利，因此对应了帮助运营和产品人员设计更加智能的定价、补贴力度、反作弊、调度策略等场景；技术视角主要考虑效率，因此需要逐一探究这些场景背后依赖的搜索引擎、推荐系统、自然语言处理、计算广告、调度规划等技术方案的秘密。

- 团队篇中，为了真正将以上理念落实到每个具体的公司和团队，并在实践中发挥作用，我们专门用两章内容来解释数据驱动增长在团队层面需要具备的必要因素。

  其重要性在于，再成熟的技术方案、再丰富的业务经验也需要整个增长团队的合作与执行才能落地。越来越多的公司开始重视组建负责增长指标的专业团队，采用更有效的方法促进各部门协同完成业务增长任务。其必要性在于组建增长团队的过程中还存在很多未知领域，不同公司也都面临各自不同的问题。来自于原先的数据分析团队、营销团队的成员或多或少都会按照之前的理解行事，因此未免会出现很多不能同心同力的问题。本章则给出了相应的答案。

- 结语篇则回到了出发的原点，带领读者继续探讨智能增长力量对整体经济发展的促进作用，包括在不同历史发展阶段逐渐展现自身力量规模和作用方式的区分。最后，讨论完智能增长的本质之后，再回过头来关注这一历程中人应该体现的价值和可能的归宿。

本书面向的读者群体包括如下三部分。

- 对所处时代正在发生的变化拥有好奇心的消费者和从业者。你在经历了科技改变传统商业模式过程、体验了便利满意的服务之后，愿意花些时间了解这些变革背后所依赖的机制。
- 对参与这场技术变革充满信心和热情的参与者，以及从事移动互联网行业研发、产品和运营的同学。你想知道如何将最前沿的数据挖掘、数学建模和机器学习等人工智能技术应用到传统行业，以催生更多互联网新经济增长点。
- 对智能增长、互联网+经济分析预测的观察者。本书从微观和方法论层面提出了一套崭新的理论和实践框架，通过这个框架，你可以得知诸多增长团队所面临的实际问题及其解决方案。

那么，现在就让我们逐章解读吧。

目　录

第一部分　数据篇

第二部分 模型篇

第三部分 场景篇

第五部分 结语篇

第一部分

数据篇

数据驱动增长的能量来源于丰沛充足的海量数据，数据对于智能时代的意义就像煤炭和石油对于工业时代。然而若没有高效的开采技术，一片蕴含极其丰富矿藏的土地，对于渴望借助它力量的人们来说都只是无意义的荒漠。大数据就是如此。我们必须掌握必备的数据收集、加工、存储和访问技术，才能将潜埋在大数据地下的洪荒之力唤醒，从中梳理出逐渐清晰可见的纹理脉络，让大数据成为即时可用的数据资源。

数据篇从智能驱动增长方案的基础出发，沿着增长型团队内部发掘数据价值的实际工作链条，分 4 章阐述原始数据从平台和用户的交互过程中收集而来、在系统内部为了满足各种约束条件而精细加工、要适应各种环境压力而存储自身，以及最终为了满足上层应用需求而提供访问能力的整个过程。

数据篇主要解释大数据从哪里来的问题，能够帮助运营和产品人员更好地了解，究竟是什么力量支撑着我们脚下的坚实大地。

第 1 章

数据收集

俗话说，万事开头难。大数据增长的第一步就是收集数据。有什么样的数据，就会有什么样的结论。数据收集阶段工作的质量会直接影响随后的一系列工作的方法、难度和效果。因此，从一开始就强调数据收集技术的重要性，是怎么都不为过的。

相比信息时代，数据收集技术在大数据时代发生了显著变化。

- **大数据的收集变得更容易**：以射频识别、红外感应器、全球定位系统、激光扫描器、气体感应器等信息传感设备为代表的物联网技术的广泛运用，极大提高了计算机系统获取外部数据的能力；移动智能手机设备的广泛普及，极大地拓展了用户随时随地接入互联网服务的时空界限；第 4 代移动通信技术的商用化和 WiFi 热点数量的快速增长，极大便捷了实时数据的传送，使得更为复杂的商业行为交互逻辑成为可能。
- **大数据的处理变得更便宜**：以 MapReduce、Hadoop 为代表的并行计算解决方案的出现，使得具备一定分布式编程能力的工程师，就可以在业界领先的工业级水平上进行操作；GPU 计算、云计算服务平台等硬件在计算能力和规模上数量级的提升，使得个人或中小企业可以用极为低廉的价格和降低的门槛，获得以前只有大公司才能组建和维护得起的数据处理能力。

硬件物理层面和系统架构层面的技术革新，在短短几年里迅速改变了信息时代数据收集技术面临的问题，也提出了新的挑战。在这个背景下，企业级应用的数据收集工作会遇到新问题：如何在实际工作中遵循一种普遍适用的准则去消繁就简，支持业务以较少的代价进入可用状态；如何尽可能多地利用成熟技术工具的现有解决方案，获取大数据环境下的自然红利；如何打开上帝视角，以全新的角度重新审视既有的流程和模式，提出对多源数据颠覆传统的理解。

本着实用和便于理解的原则，我们不以教科书知识点的形式，去阐述数据收集每个技术点的定义和相互关系，而是从商业应用业务最关心的核心问题切入。不管你是工程师、产品经理，还是业务运营人员，当搭建一个能提供商业价值，且具备一定技术含量的大数据收集子系统时，都要了解在这一过程中需要判断的因素、关注的要点和留意的风险，从而减少不必要的损失，提高系统部署成功的效率。

具体来说，本章会从用户行为、商业交易和标签属性三个方面，分别介绍收集这些数据过程中要注意的问题。这三大类数据已经基本覆盖了商业应用中的主要来源数据，比如在滴滴预约出租车出行、在百度外卖点餐送到家、在今日头条阅读新闻资讯、在携程预定旅游地酒店，这些主流的移动互联网应用都是围绕着如何获取这些数据展开的；常见的算法模型，比如协同过滤推荐、CTR 点击预估、用户画像，所依赖的基础数据也都来源于这些数据指标。因此，如果相关人员能以这样的分类视角，弄清楚上层数据形态所依赖的底层数据来源的收集技术，会更有利于从源头熟悉大数据增长技术的来龙去脉，获得更加全面整体的概念。

1.1 行为数据

行为数据，特指用户登录应用后在操作过程中产生的数据，而不包含这些操作行为所达成消费交易和社交关系建立等与平台相关的衍生数据。因此，在讨论行为数据时，会更侧重用户在与应用交互过程中特有的行为习惯对收集数据产生的影响。同时，由于用户在广义上的操作行为非常琐细繁复，在收集行为数据时我们也必须考虑到采集效率，尽量避免给采集过程带来较高的成本。

1.1.1　传统获取方式

目前，互联网应用收集用户行为数据主要包括后端服务器和前端客户端两种方式。

1. 基于后端服务器日志收集

当用户在浏览器里输入 URL 向网站服务器发出 HTTP 请求，或者登录手机 App 向应用服务器上传数据时，后端日志收集行为就开始了。服务器在接收到请求数据后，会自动在后端的日志文件里追加记录，记录下请求主机 IP 地址、登录账号、日期、时间、请求类型及详细指标等信息。服务器也会记录下返回给用户操作的每一次记录，包括响应主机 IP 地址、服务耗时、服务类型及详细指标等信息。

比如，谁在访问网站、用户访问网站的路径、用户在每页停留的时间、用户离开站点的位置、用户是否成功完成自己想要做的事情，等等。具体到某个应用领域，后端日志收集的范围可以包括用户浏览过的物品 ID、物品展现在哪个频道页、物品所在位序、用户加入购物车的物品 ID、是否最终下单、下单前取消过哪些物品、有哪些附言和特殊要求（比如是否要发票、配送时间）等。

2. 基于客户端用户行为收集

与后端服务器日志收集不同，前端用户行为收集需要采用通常所说的“埋点”技术，也就是在网页编码上嵌入 JavaScript 代码，或者是在 iOS 和 Android 应用代码中基于事件机制开发相应操作的记录代码。当用户访问网页，或者在手机 App 客户端上操作时，触发嵌入代码向单独的日志收集服务器发送请求，从而记录用户访问的数据。这段代码会从访问者硬件设备的 cookie 中取得详细信息（比如，访问时间、浏览器信息、工具厂商赋予当前访问者的 userID 等），并发送到数据收集服务器。数据收集服务器将收集到的数据处理后存入数据库中。网站经营人员通过访问分析报表系统查看这些数据。埋点技术以其快捷性和精确性得到很多数据收集方的青睐，已经发展成为当前最流行的数据收集方式之一。

1.1.2 获取方式对比

后端日志服务器的方法，比较适合以较低的成本获取数据，同时保持对外部数据获取方更加谨慎的态度。

- 它对于数据源的形态要求不高，开发成本较低。
- 它比较方便对历史数据进行离线的后续处理，也可以用于收集搜索引擎爬虫的访问记录。

从客户端收集到的数据更加精确、更加丰富。

- 它不受动态分配 IP 地址或代理服务器的影响。可以采用更加灵活的客户端跟踪技术，比如，由 Web 服务器对每个访问站点的客户机自动分配 ID，并将其记录在客户端的 cookie 中。这样，每当用户浏览网站时，Web 服务器可通过访问客户端的 cookie，就知道此客户端是否访问过本网站。
- 它能够记录正确的用户浏览路径。由于是在客户端记录用户行为，所以客户端代码可自动跟踪用户的浏览路径，不管是否通过本机缓存或通过代理服务器。这种收集数据的方法更加灵活、可定制性强，可以记录缓存、代理服务器访问，而且对访问者的行动追踪更为准确。

但是，第二种方式需要付出一定代价，即需要提前在网页或应用程序中加入大量定制化的脚本或代码，增加了前端服务的性能开销，频繁地记录下载和重定向数据会比较困难。当业务需求发生变化时，系统需要上线新的前端服务才能满足新的埋点数据收集要求，迭代周期长。

1.1.3 无需埋点的数据收集

“埋点”技术使得大规模高效灵活的数据采集成为可能，但也会遇到各种各样的问题。

- 它需要将用户行为轨迹的追踪变量、信号指示，或者 JavaScript 脚本、应用代码嵌入在网页和程序内部，这增加了原始代码的复杂度。由于是有别