

谭磊 著



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

013066542

F713. 36

860

数据掘金

电子商务运营突围

谭磊 著

藏书
图书馆



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

F 713.36

860

内 容 简 介

电商坐拥互联网行业最丰富的用户数据金矿，却很少有人从中挖掘出真金白银。本书旨在打破这一困境，一步一步引导从业者以数据为核心来运营网站或网店。本书用浅显的文字与独特的视角，不仅成功解读电商数据运营之惑，更呈现大量数据分析和挖掘的必要基础知识及实用相关工具。在通过阅读轻松掌握电商数据运营须关注的要点与方法之后，读者还可有针对性地从中选择学习如何利用数据来完成——流量获取优化、广告投放、客户分析，以及客户价值提升等一系列电商运营要务。

《数据掘金：电子商务运营突围》一书主要写给电商从业人员，无论是中小电子商务的运营人员、数据分析人员，还是大公司负责电子商务的策略官、市场官和运营官，都能从本书中找到自己所需且急需的有价值内容。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

数据掘金：电子商务运营突围 / 谭磊著. —北京：电子工业出版社，2013.6
ISBN 978-7-121-13897-3

I. ①数… II. ①谭… III. ①电子商务—数据处理—研究 IV. ①F713.36②TP274

中国版本图书馆 CIP 数据核字（2013）第 106920 号

策划编辑：张春雨

责任编辑：李云静

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：720×1000 1/16 印张：29 字数：404 千字

印 次：2013 年 8 月第 2 次印刷

印 数：4001~6000 册 定价：65.00 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 1

谭磊写了一本叫《数据掘金》的书，叫我做序。我知道他是客气，给我个待遇，但我这个电子商务的门外汉，却老实不客气，把这活儿应承下来了。这一应承不要紧，双方就都犯了难：我这边不知怎么写，他那边是热锅上的蚂蚁，书等着开印呢。这可是写电商的书，互联网电子商务的发展，一日千里，说不定被我一耽搁，书还在印刷厂里，书中的不少东西已经作古了。这也是我佩服谭磊的地方，这玩意儿也敢写！今年春节期间几次见到他，都是在咖啡馆里，每次我都不能确定准确的时间，我很内疚地以为他等了我一两个小时，谁知他一杯咖啡已经喝了半天了，号称是在等我，其实是在写他这本美其名曰“掘金”的书。我是读中文出身，对方块字始终怀着一种敬畏，若打算将它们一个一个码起来时，就有流芳的想头，最起码也会意淫到 20 年以后去。这会儿我就像一个穿了长衫马褂的学究，走在横陈了无数比基尼美女的沙滩上，脑子是过时了，内心却“鸡冻”着。前几天在陆兆禧的婚宴上听马云谈网络，与 10 年前听他讲黄页，感觉隔了好几个世纪。互联网是一种新的世界观，它没有停留，更没有永久，大量的东西还未出生，便已过时。这天咱估摸着，瞬息要变。于是我就央着谭磊老弟，扶我起来，咱把长衫这就脱了，到水里去练练。是为序。

通策集团董事长 吕建明

推荐序 2

电子商务和数据挖掘，这两个在最近一年很火的词，都覆盖在 Raymond 这本书里了。我认同 Raymond 书中的观点，在未来的若干年，电子商务运营一定会转向以数据为基础的运用，而社会化电商和移动电商会是电子商务未来发展的两个大方向。

社交网络与移动互联时代给媒体带来了极大的演变，2013 年纸媒和中小电视媒体将面临生死挑战。定位在稀缺资讯、智慧与思想供应商角色的凤凰网，以主流网民为首选用户，做品质内容提供商。随着 ifeng.com 的移动收入占比迅速提升，我们正在努力打造中国移动互联网广告的新格局。凤凰网如何能够更好地为电子商务服务，也是我们一直在思索的问题。

认识 Raymond 快 20 年了，脑中的印象还是他当年在杜克大学意气风发的文艺青年样子。作为理工科出身的 Raymond，书中的文字能够有这样的水准，应当说是相当难能可贵的。

凤凰新媒体 CEO 刘爽

推荐序 3

“内容为王”，这条蛇曾经盘踞在我大脑中很久。2006 以前，我一直小心翼翼地用虔诚和敬畏喂养它。2006 年 19 楼网络公司正式创立以后，我开始每天都与这条蛇搏斗。去年，在坚持了 18 年之后，我辞去了《都市快报》的日常采编工作，全身心投入 19 楼。那条蛇终于彻底从我的脑海中消失。

谭磊说，数据是最底层的东西，它是 0 和 1，是对事物最原始的记录，数据之上才有信息。我恍然：数据之上才有信息，信息之上才有内容。如果内容是王，那么数据就是王的王。如果内容是传统媒体的王，那么数据就是互联网的王。

相比传统媒体，互联网是一个更加纷繁而深邃的世界，蕴含着无限的可能。任何一位互联网用户，你可以不生产内容，但你不可能不生产数据。因此，你可以是互联网的奴隶，但同时又可以这个世界的主人。

19 楼 CEO 林煜

推荐序 4

在互联网、金融领域工作了将近二十年，我的体会是：数据挖掘、数据分析、数据仓库、商务智能的概念一直很热，但一直是一个美好的乌托邦，可望而不可即。两个最大的问题是无法做到精准和快速。最赚钱的金融行业和最潮的互联网行业，如何能够有效结合，并利用好数据，一直是我思考的问题。

我一直坚信精准的营销是服务，错误的营销是垃圾短信。之前，给美国一家大银行做咨询的时候，我们曾经设计过一个精准营销的规则，先从各家实体超市及网上商城搜集数据，如果一个20~40岁之间的人，在一个月内买了两次或两次以上的婴儿奶粉或尿布，这个人很有可能是家里有小孩（如果一个月只买一次，可能是买了送给生小孩的朋友），就向这个人营销可以免税的儿童教育基金产品。结果，有个朋友后来过来聊天，很高兴地告诉我，这家银行真贴心，他太太刚生了小孩，以前并不知道这种免税方法，得到推荐信息后，马上去买了基金，一年省了不少税。当时，我心里非常开心，甚至有些小得意。不幸的是，这种正面的例子并不多，而错误的营销，被当成垃圾短信的反面例子，在3.15消费者权益日的时候，我们听得太多了，这里就不再赘述了。

时效性方面，传统的数据分析技术，往往分析起来，要进行批处理，要半天、一天，甚至好几天的时间才能出结果。上面这个例子，时效性要求并不高。但是，在互联网上，这样的速度就是灾难了。现在的互联网用户打开一个网站，往往只浏览几分钟时间，如果网站不能精准地推荐用户感兴趣的产品，用户就跑掉了。用户下了订单，在最终的订单显示页，如果没有推荐很好的

相关产品，用户几秒钟就关掉网页了。

所幸，现在的大数据技术，终于可以比较好地解决精准和快速的问题了。具体怎么做，Raymond 在书中做了非常好的解答。

其实，IT、互联网方面的书也需要做到准确和快速。很不幸的是，出书的人，往往是学院派，没有实际的经验，信息和内容也比较滞后，所以，准确度和时效性比较差。而实战派的人，往往因为沉不下心来，或者没有时间，或者没有文字的功底，写不出书来。难得 Raymond 写出了这样有实战经验、实际、实在的一本书。认识 Raymond 24 年了，每一次，都以为已经很了解他了，可是每一次，他都让我对他新的潜能和能力惊讶不已。

高伟达集团副总经理兼首席技术官 程励箭

业内推荐

《数据掘金》是一本很有趣的书。作者谭磊在中国和美国的顶尖大学接受了计算机科班教育，之后在微软总部工作多年，回国后在技术和管理岗位上都业绩斐然。在繁忙的工作之余，谭磊乐于向读者分享他在业界的丰富经验与洞察力，本书已经是他的第二部著作了。《数据掘金》针对电商从业人员，对数据分析和数据挖掘的理论和算法做了通俗易懂的介绍，让电商运营和分析人员了解数据的意义，并能够使用主流工具来对数据进行分析 and 挖掘。本书还提供了大量电商数据分析的实例，对于有志于进入电商行业的技术人员和学生也有很好的参考价值。

——复旦大学计算机学院教授、博导 黄萱菁

未来的互联网世界一定是属于数据的。我们要将社交平台做到极致，数据分析和数据挖掘的作用也是必不可少的。Raymond 的这本书是电子商务数据分析领域有独特观点的好书，而且给从业者提供了切实可行的运营方案。希望能多看到一些像这样的好书。

——微软全球社交平台首席开发总监 李津

我们对数据分析和数据挖掘的关注度很高，也很高兴看到有这样一部专注于电子商务数据分析领域的好书面世。Raymond 的这本书深浅适中，既符合技术人员的需求，对于非技术的电商从业人员帮助也非常大。

——阿里巴巴集团资深总监 陈宜

认识 Raymond 十多年了，从西雅图经北京到杭州，从十几年前的微软总部软件研发工程师到现在的通策集团首席运营官，

Raymond 好像是朋友中变化最明显的，但聪明和对技术的敏锐把握一如既往。

通读了一遍 Raymond 的最新书稿，两个评语：有货，靠谱。数据分析是个大题目，但非常容易变成分享管理理论和统计公式的剪报夹。如何在建立高度的同时紧扣实践，是个难题。Raymond 的这本书做得不错：有货同时靠谱；靠谱不忘有货。

目前数据是最热的题目，Raymond 的这本书出来的正是时候。

——ITIL 专家、ITpreneurs 大中华区首席代表 蒋胜

这本书全面讲解了在电子商务中运用数据分析和数据挖掘。我反复看了若干遍，每次都能学到新的概念。如果您从事的行业是电子商务，又对数据的作用感兴趣，那么这本书是您一定要读的。

——中国工业设计协会秘书长、

前海尔数字产品集团全球营销总监 刘宁

《数据掘金》一书所介绍的是在互联网数据爆炸的背景下，如何运用新技术打造全新的盈利模式，摆脱日益同质化的电商竞争的困扰。数据运营目前是一个全球化的难题，本书对数据运营的介绍是与世界同步的，从这个意义上说，所有的电商企业与全球竞争对手重新站在了同一条起跑线上，未来 10 年的竞争图景将会更加令人期待。

——高德 CTO 崔海

谭磊的这本书给出了很多关于电子商务数据运营方面的独特观点，令人耳目一新。在他上一本书《New Internet：大数据挖掘》之上给出了更多实际的运营方案，是电子商务和运营领域从业者不得不读的一本好书。

——英特尔数据中心软件部大数据软件服务部总监 范磊

数据分析在国外电商的经营中受到无与伦比的重视，但在国内的电商企业中能理解数据重要性的依然很少。很荣幸能够比大多数

的读者提前看到这本书。本书不但全面讲解了数据分析、数据挖掘的各种概念、方法，而且还详细讲解了吸引客户和提升客户访问时间的各种有效办法，推荐给众多的电商从业者及负责人员，非常值得一看。

——赢时代总裁 曾荣群

我们一直在研究数据在电子商务运营中的应用，而我们所有的大客户也都进入了电子商务领域。读了 Raymond 这本关于电子商务数据挖掘的书，受益匪浅，因为只有在电子商务运营中充分利用数据，我们才能无往不利。

——麦尚广告公司总裁 邵腾

在全球宏观经济不景气的大环境下，电子商务无疑是中国经济的“兴奋剂”，几乎所有的行业都在探索和实践电子商务的运营。数据分析和数据挖掘将贯穿电子商务运营的价值链，尤其是在社会化电商和移动化电商的趋势下，如何更智慧地运营将是未来虚拟经济与实体经济实现无缝整合以及全面竞争力提升的必备武功。令我们惊喜的是，从 Raymond 这本书里，我们找到了答案！书中有大量鲜活的案例，以及前瞻性的剖析和策略性的思考。认真研读，你一定会茅塞顿开！Raymond 在本书中的一些重要观点对于我们互联网金融的业务发展有非常大的启发！

——资深广告人、中金在线副总裁 陈刚

数据在电子商务中的作用是毋庸置疑的，而它也是我们在做电子商务运营中最关注的。市场中关于电子商务和数据挖掘的书很多，有的只是讲述了概念，而有的只是操作手册，而 Raymond 的这本书不但把相关的理论知识用通俗易懂的语言讲解得很清晰，而且有大量的实际案例供我们参考，是我们从业人员很有价值的参考书。

——聚流电商总裁 周为

数据分析究竟是一种艺术还是一门科学？我不知道，但我知道数据是死的，而分析数据的人是活的。随着千人千面的理解和对于数据的不断追溯、剖析，它在某个时间的小结也是因人而异的。我很欣喜地能够在这个日新月异的数字时代读到这部由谭磊编写的《数据掘金》。它针对我的工作很有启发、指导意义；我向电商卖家、广告人及尊重数据营销的朋友们诚恳地推荐它！也许你正为疲软的市场、萎靡的金融环境和未知的销售势态而烦恼，那么《数据掘金》一定是你驶向蓝海的风向标！

——UTC 行家市场总监、淘宝知名大卖家 于振伟

之前和 Raymond 在闲聊的时候就听过他关于数据运营的概念，而《数据掘金》这本书更是系统介绍了数据在电子商务运营各个环节的作用。这本书内容丰富，而最关键的一点是所有的信息和数据都来自于实际操作。

——知名广告人、互联网营销专家 钱岛

电商在今天非常火，越来越多的人在进入这个行业，无论你已经经验丰富、身居要职，还是仍在苦苦摸索，仔细研读这本书，会让你思路更加清晰。电商运营就是从数据分析和挖掘出发，这是一本理论和方法并重的书，推荐给想在这个行业持续发展的同仁。

——杭州风牛马电子商务公司总经理 姚中伟

在未来，电子商务已是必然，而数据更是决定企业胜负最重要的因素。一家企业如果要成功，必须在数据运营上有突破。Raymond 的这本书牢牢把握住了“电子商务”和“数据运营”两个要点，是行业人士必读的好书。

——资深互联网营销专家 熊炜

作者的话

还记得 1998 年我第一次在网上购买 DELL 电脑，从我在网上做电脑的配置选择开始，一直到组装成的电脑从德克萨斯州寄送到西雅图的家里，大概有两个星期的时间。那个时候我颇为惊叹 DELL 的效率，而 DELL 公司确实也因为高效配送而在第一次互联网泡沫（Internet Bubble）中引领所有的硬件制造商。可到了今天，这个速度显然会被人给差评的。早上在电子商务网站下的订单，最快当天下午甚至中午可能就已经收到货了。

周围有不少做电商的朋友，有一些做得还是颇为成功的。不过其中成功运用数据提升运营效果的却是少之又少。他们对“数据挖掘”都很感兴趣，不过缺的是“know-how”，不知道怎么做。

我在给一些企业做咨询的时候也发现企业家朋友们对于是否需要进入电商领域这个疑问基本上已经没有了，但是对于怎么做和是否能够做好还是颇有不解的，特别对于电子商务数据运营的概念是很模糊的。

本书主要针对的人群是对数据运营感兴趣的电商朋友们，目的就是为了解读和电子商务有关的各类数据，让他们能够深入了解数据的意义，并能够使用一些工具来做数据分析和数据挖掘。对于还没有做过电商运营的读者，本书可能有一定的难度。亚马逊的总裁 Jeff Bezos 曾说过，“如果我在网上有 300 万个客户，我将要建立 300 万个商店”，这个想法在不久的将来会实现，靠的就是数据。

电子商务既充满挑战，也蕴涵机遇。在数据分析和数据挖掘的帮助下，我们希望能够帮助朋友们以最少的投入在最短的时间

内引入最大的流量、提高新增客户数量、找出高价值用户、提高客单价、提高重复购买率等。

这个世界唯一不变的就是变化，互联网的变化本身就很大，而本书讲述的电子商务又是在互联网领域中变化较大的一块。在笔者写作的时候足够准确和新颖的运营方法，很有可能在图书出版之际就会发生变化。如果你对书中的说法和提出的数据运营方法有不同的观点，欢迎和我交流。

这本书不好写，侠少和我在写作初期讨论了很久，数据分析和数据挖掘的理论和算法如果讲述得太深，主要工作是电子商务运营的读者可能会知难而退；而另一方面，数据分析和数据挖掘概念如果说得太浅，有技术背景的读者可能又会觉得无聊。另外从内容角度来说，我们并不想把这本书做成泛泛的知识罗列和知识点的堆砌，而是以实际案例和运作方式为主。我们的整体提纲经过了十数次的修改，最后勉强达到我们想要的效果，既相对浅显，又能给各种背景的读者提供符合他们要求的知识。如果读者能够从本书中获得一些有用的信息，余愿足矣。

在本书的写作过程中，得到了很多人的帮助。首先要感谢韩冬、曹晓波、江峰、荷铁勇、汪华、王海、楼建强、李嘉骅、牛文军和王磊等同学帮我查找数据分析和数据挖掘相关资料，还要感谢微软搜索中心的李津同学和通策集团吕笋同学提出的宝贵意见，感谢戴霖和李悌等同学帮我校验一些章节，以及姚中伟和彭毅同学指出书中的一些修改环节。另外，特别要鸣谢鲍佳和王剑同学为本书配的部分插图。

聚流电商的周为同学、思美传媒的江山同学、淘宝开放平台的冯光同学、UTC 的曹轶宁和于振伟同学、MyGift.com 的 Stephen Lai 及首正信息的罗俊峰同学为本书提供了大量精彩的案例和数据，在此一并表示特别的谢意。

谭磊 (Raymond)
@CarelessWhisper

目录

第 1 章 引言：电子商务运营和数据	1
1.1 2012 年最大的赌局	2
1.2 为 300 万人建 300 万个网站	6
1.2.1 电子商务的 RUPI 概念	7
1.2.2 在互联网上卖米	8
1.2.3 电子商务怎么能离开数据	10
1.2.4 淘宝店的四个核心数据	13
1.3 让电商运营不再那么辛苦	16
1.3.1 电商人的蓝精灵之歌	17
1.3.2 电子商务运营入学考试	17
1.3.3 店铺诊断——我的网店能挣更多钱吗	19
1.3.4 让你的网店脱颖而出	23
1.3.5 为你的网店装上“业务雷达”	27
1.4 电子商务数据运营的五大应用	30
1.4.1 让网站更吸引人	32
1.4.2 把潜在客户转化成真正的客户	33
1.4.3 挖掘老客户价值	35
1.4.4 推荐系统的设计和应用	36
1.4.5 针对不同客户提供个性化的产品	36
1.5 关于电商数据的六个“W”和一个“H”	37
1.6 本书的内容	39
1.7 本章相关资源	42
第 2 章 我们需要知道的数据分析	43
2.1 从数据分析专家林彪说起	44
2.2 数据分析基本概念	45
2.2.1 就这么简单：三种基础数据	46

2.2.2	我们这样来理解数据	47
2.2.3	概率并不可怕	50
2.3	让我们开始加工数据	52
2.3.1	数据集成——把所有数据都拿过来	52
2.3.2	数据清洗——给数据玩“洗刷刷”	54
2.3.3	数据转换——给数据换个“马甲”	59
2.3.4	数据规约——有时候也要丢掉数据	62
2.4	用向量表示数据	63
2.5	网站日志的收集和处理	66
2.5.1	网站日志信息分类	66
2.5.2	网站日志实例	70
2.5.3	网站日志预处理	76
2.6	最好的分析方法——看图说话	82
2.6.1	起起伏伏用折线图	83
2.6.2	简单比较用柱状图	83
2.6.3	转化率用漏斗图表示最直观	83
2.6.4	雷达图显示用户偏好	85
2.6.5	表示比例最好的饼图和环形图	86
2.7	本章相关资源	89
第3章	我们需要知道的数据挖掘	90
3.1	什么是数据挖掘	90
3.1.1	尿不湿和啤酒	92
3.1.2	Target 和怀孕预测指数	94
3.1.3	从数据分析到数据挖掘	95
3.1.4	数据挖掘的一般过程	97
3.2	人人都能做数据挖掘	100
3.3	我们需要知道的四类数据挖掘算法	101
3.3.1	分类——人以群分	101
3.3.2	聚类——物以类聚	108
3.3.3	关联——马原告诉我们事物是普遍联系的	111
3.3.4	序列——排队的规律，中国人最明白	119
3.4	Web 挖掘和信息检索	121
3.4.1	Web 挖掘和信息检索概览	122

3.4.2 协同过滤——推测同类客户的行为	124
3.4.3 个性化推荐和推荐系统——我们要更懂客户	126
3.5 本章相关资源	130
第4章 数据分析和数据挖掘工具的选择	132
4.1 数据分析工具	132
4.1.1 用 Excel 做数据分析	132
4.1.2 MATLAB	136
4.2 网站分析工具	139
4.2.1 用 GA 做分析	139
4.2.2 GA 的限制	142
4.2.3 各种站长工具	143
4.3 用 R 语言制作的工具	144
4.3.1 用 R 做数据分析的优势	145
4.3.2 用 R 绘制热力图	148
4.3.3 用 Rattle 分析广告投放数据	150
4.4 其他的开源数据挖掘工具	154
4.4.1 Weka 数据挖掘工具	154
4.4.2 Google 提供的数据挖掘工具	158
4.5 电商平台上的各种工具	159
4.5.1 用量子恒道分析淘宝网店	159
4.5.2 淘宝上的数据魔方	161
4.5.3 开放平台上的工具	165
4.6 数据展示工具	165
4.7 本章相关资源	168
第5章 电子商务数据运营入门	170
5.1 在讨论数据运营之前	170
5.1.1 数据运营的四大障碍	170
5.1.2 数据不是万能的	171
5.2 电子商务运营中重要的数据点	173
5.2.1 访客数	175
5.2.2 转化率	176
5.2.3 客单价	180