Broadview®
www.broadview.com.cn

无数据不运营，跨越数据天堑，
洞悉竞争对手运营机密

数据运营

深度揭秘SEO电商数据抓取技术

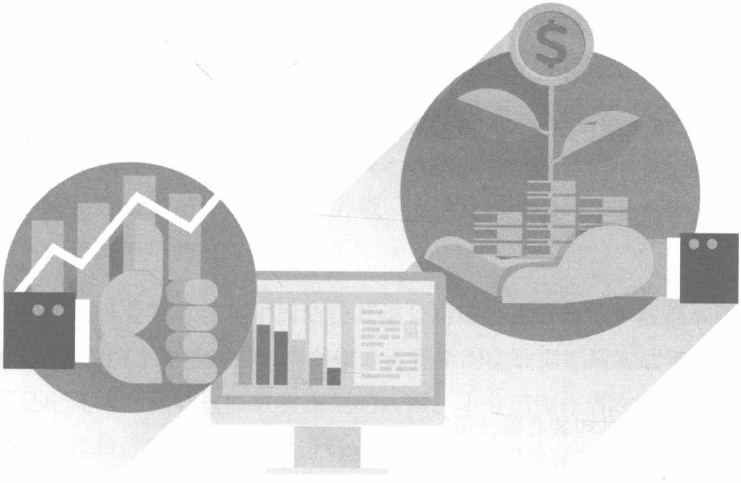
邢波涛　郭娟　著

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

数据运营

深度揭秘SEO电商数据抓取技术

邢波涛　郭娟　著



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

内 容 简 介

本书的写作目的就是针对电商运营相关人员，告诉他们如何获取淘宝 SEO 优化所必需的运营数据，使得运营能够更好地跟自己的 SEO 优化绝技相结合，从而使自己从手动获取运营数据的海量工作中解放出来，做一些更有意义的事情。本书内容主要包括：如何实时抓取淘宝搜索排名数据、淘宝无线排名数据、直通车关键词排位、宝贝的订单数据、宝贝的评价数据，淘宝组合关键词的拆分，查询整店动销率，淘宝关键词获取技术，以及抓取生 e 经数据和数据魔方数据辅助 SEO 优化等。

本书适合已经熟悉淘宝 SEO 优化技巧，又想深度了解如何获得 SEO 优化技巧背后数据秘密的运营人员阅读，希望能给他们带来帮助。

图书在版编目（CIP）数据

数据运营：深度揭秘 SEO 电商数据抓取技术 / 邢波涛，郭娟著. —北京：电子工业出版社，2015.10
ISBN 978-7-121-27158-8

Ⅰ. ①数… Ⅱ. ①邢… ②郭… Ⅲ. ①电子商务－网络营销－数据处理－研究 Ⅳ. ①F713.36

中国版本图书馆 CIP 数据核字（2015）第 224232 号

策划编辑：张春雨
责任编辑：葛　娜
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
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：12.75　　字数：243 千字
版　　次：2015 年 10 月第 1 版
印　　次：2015 年 10 月第 1 次印刷
印　　数：3000 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。

序

认识本书作者老邢很多年了，他是个无论在工作还是生活中都特别认真的人。在电商这个十分年轻的行业里，老邢是较早进入淘宝生态体系的技术人员，对淘宝的发展变化有着自己的理解和认识。经营过淘宝店铺的人都知道，数据对于电商来说极其重要，整个店铺运营策略都是围绕数据展开的。流量、转化、销量、DSR 这些数据的变化都会影响下一步的推广策略，可以说“无数据不运营”。

马云说世界经济正迅速从 IT 走向 DT（数据处理技术）时代，未来属于那些掌握数据的人。从人类技术文明的发展过程来看，技术总是从复杂向简单、从高端向普及演进。数据抓取及分析最初只是 IBM、谷歌、百度这些大公司才有机会使用的技术，现在已经进入到普遍使用的阶段。

通常电商网站也会采取技术手段防止其他公司或个人以技术手段抓取数据（淘宝早在 2008 年就屏蔽了百度爬虫），本书作者有五年淘宝数据抓取实战经验，对网站数据抓取尤其是电商网站的数据抓取有深刻的认识。十分感谢作者在书中毫无保留地把淘宝数据抓取的经验进行了分享，促进行业交流。

作为一个资深技术人员，作者有深厚的技术背景，能够跳出技术人员固有的思维方式，从店铺运营的视角，通过图文并茂的形式，深入浅出地介绍淘宝页面上主要运营数据的抓取方式，十分难得。对于有心学习如何抓取淘宝数据的运营人员，本书可以让你零基础掌握如何抓取淘宝页面上丰富的运营数据；对于初入电商领域的技术人员，本书也可以帮助你快速掌握淘宝核心数据抓取方式，不必重复发明轮子。

淘宝页面上的数据信息极其丰富，本书涉及内容基本覆盖了其中最重要的搜索结果、宝贝信息、销量信息、DSR、无线端、直通车、评价等，详细讲解了每一个数据的抓取方式，读来有疱丁解牛之感。

淘宝网站本身在不断发展变化，本书在讲解案例的同时，意在让读者通过参与分析一个个具体数据抓取过程，了解其本质，淘宝页面万般变化，变化的仅是数据表现方式，离不开本质。对于一个小白读者来说，可以轻松地掌握淘宝数据抓取的关键点，稍加分析，即可根据自己的需要对书中没有涉及的数据内容进行抓取。

随着技术的进步与普及，会有越来越多的数据被保存下来，人们的行为习惯会越来越多地通过量化数据的方式被分析出来，数据产生的价值也会越来越大。

掩卷而思，从社交网络到移动互联网，再到线上线下结合（O2O），再到李克强总理提出的“互联网+”，整个人类社会近一百年创造的文明成果已远远超过过去历史的总和。数据时代的来临不可避免，希望本书可以帮助更多的人加深对数据的理解，感谢老邢为数据抓取技术普及所做的努力。

华北区淘宝最大代运营商之一洪海龙腾公司首席运营官（COO）

陈权国

前 言

我 1996 年大学毕业到现在（2015 年），工作 19 年了，一直从事的都是软件开发工作。在 2011 年，我出版了自己的第一本面向程序员的图书《B2B2C 网上商城开发指南——基于 SaaS 和淘宝 API 开放平台》，今天我为什么想写一本面向电商运营的图书呢？

这其实与我最近 5 年的工作经历有关。从 2011 年开始的前三年（2011 年至 2013 年），我一直都是在围绕淘宝的 App 后台做工具软件的开发，例如进销存软件和微博营销软件。不过，由于种种原因，这款进销存软件和微博营销软件虽然试用客户上万，却并没有带来能够支撑团队运作的资金。而到了 2013 年年末，我在师弟天津商业大学宝德学院高学争老师的影响下，开始研究基于淘宝、天猫的 SEO 优化工作一直至今。在这个过程当中，我发现了一个有趣的问题，那就是从事淘宝、天猫 SEO 优化的运营人员和淘宝、天猫店长们，他们对业务能力的把握是很到位的。也就是说，运营人员对如何针对淘宝、天猫 SEO 优化的奇门绝技是很熟悉的。但是运营人员也有个致命的缺陷，就是 SEO 优化技巧所依赖的数据支撑，他们是没有能力获得的，他们只知道优化理论和技巧，却无法获得相应的数据支撑和验证。例如，针对大类目下的淘宝 C 店 SEO 优化最常见的下架时间问题，大家都知道下架时间很关键，却对如何获得关键词综合排名前 1 页甚至前 5 页、前 10 页宝贝的下架时间的数据，无能为力，那么宝贝根据下架时间卡位也就无从谈起了。当然，目前市面上有的公司也出了一些可以批量采集这些数据的工具软件，运营人员可以利用这些工具软件来完成相应的工作，但是他们对这些软件背后获取数据的原理是一无所知的。这也是我写这本书的目的，深度揭秘淘宝、天猫 SEO 优化背后数据获取的秘密。所以，本书就是针对电商运营相关人员，告诉他们如何获取淘宝 SEO 优化所必需的运营数据，使得运营能够更好地跟自己的

SEO 优化绝技相结合，从而使自己从手动获取运营数据的海量工作中解放出来，做一些更有意义的事情。

我写这本书的第二个目的，是因为关于淘宝 SEO 优化技巧的图书，市面上也出了很多本，但是如何批量获取 SEO 优化技巧背后所需要的运营数据，却一本也没有。这是因为绝大部分运营人员都没有软件程序员的工作经历。而我有 19 年的一线软件开发经验，又深度投入到了淘宝 SEO 优化的工作当中去，所以对运营人员（即使是全国最顶级的运营）来讲，我是一个牛 X 的资深程序员，而针对我的程序员同事和其他程序员，我又是一个已经入门的淘宝 SEO 优化和运营“专家”。从这方面来讲，我也算是“魔”“道”双修的，用朋友的一句话来说就是：流氓会武术，谁也挡不住。其实，从软件研发的角度来讲，我的职责其实是一个业务架构师+系统架构师的角色。

话又说回来，针对电商 SEO 优化运营技巧，从技术上讲，我也算入门和“专家”（针对我的程序员同行来讲）了，但是对真正的一线运营高手来讲，我又是只懂皮毛的外行。所以，这本书并不适用于想获得淘宝、天猫运营技巧的运营人员，本书适用的是已经熟悉淘宝 SEO 优化技巧，又想深度了解如何获得 SEO 优化技巧背后数据秘密的运营人员。

在我写这本书的过程当中，淘宝本身也在“丧心病狂”地采取各类反爬虫技术手段，妄图防止别人从淘宝公开的网页上很容易地获取到这些公开数据。所以，很多运营人员也可以看到，以前很好用的一些工具软件，现在也不能用了。不过，有矛就有盾，敌人再任性，狐狸再狡猾，我们也是可以找到一些特定的解决方案的。所以，本书写作的过程，也是跟淘宝技术做斗争的过程，但愿这本书能给已经熟悉淘宝 SEO 优化技巧，又想深度了解如何获得 SEO 优化技巧背后数据秘密的运营人员带来帮助。

目 录

第 1 章

淘宝/天猫做 SEO 优化对数据的需求

- 天猫、淘宝数据抓取背景
- 天猫、淘宝运营数据抓取技术概述

1.1 天猫、淘宝数据抓取背景

在针对淘宝/天猫 SEO 优化的问题上，有两类截然不同的观点：一类观点认为淘宝 SEO 优化至关重要，商品好坏无所谓，再差的宝贝（淘宝对商品的称呼，以下如果无特殊说明，宝贝和商品表达的是同一个意思，根据习惯不同就混合使用了）也能卖出去，所以在淘宝发展历史上也出现了一些专门针对淘宝 SEO 深度优化，甚至利用淘宝搜索引擎的缺陷而进行 SEO 作弊，从而达到快速出货的神店；另一类观点认为电商的本质是商，商品（宝贝）的好坏才是本质，一味追求电商 SEO 优化技巧是不足取的。个人认为，这两类观点都太偏执了。

针对第一类观点，在淘宝发展某个阶段也许是适用的，但由于产品不行，这些店铺也很难持续发展，所以淘宝历史上出现的那些神店，也只能昙花一现；而针对第二类观点，除非你真的拥有自己的知名产品品牌，这些品牌拥有大量的忠实粉丝，例如化妆品中的雅诗兰黛、箱包中的路易威登（LV），那么只考虑产品是没有问题的。而现实中淘宝 800 万卖家（有的说 700 万，有的说 900 万，估计 2015 年已经过千万了）中有多少卖家真正拥有自己的知名品牌产品呢？没有自己的核心产品，就不要谈电商的本质是商，商品（宝贝）的好坏才是本质。这就跟“好好学习，天天向上”这八个字是巨正确无比一样，落地是非常难的，对绝大部分卖家来讲，是不可能完成的任务。

所以，在货源和产品质量相对稳定的基础上，研究淘宝/天猫 SEO 优化技巧还是有意义的。电商大号@吴蚊米也在自己的微博上说——

“还没有开工，有卖家来问 2015 年怎么做，一般这种探讨宇宙起源的大问题我是拒绝回答的。可他说到了一点，活动效果越来越差。确实！活动真没几个流量，看 2014 年几个大促就知道流量基本靠搜索和老客，搜索能爆发 5~10 倍，看你怎么节骨眼卡位，算 7 天权重、30 天权重，老客全部在自主访问里增长……所以你懂了吧”。

针对绝大部分卖家来讲，“搜索”和“老顾客”就是 2015 年最核心的两个板块。搜索说的其实就是 SEO 优化，例如@吴蚊米在微博中提到的卡位（下架时间、价格等）、7 天权重、30 天权重（包括产品销量、DSR 评分、关键词质量得分等）。而“老顾客”

我这里也多说两句：非品牌店或者妖店以及重复购买率极高的类目，就不要再幻想老顾客了。品牌店就是拥有自己核心知名品牌的店铺（一般是指线下传统知名品牌在淘宝/天猫开店）。妖店其实也是以产品为核心，但不像线下传统知名品牌那样面向全网的顾客，而是只针对喜欢自家产品风格的特定的顾客，只满足少数人的需求，这类店做大的有例如裂帛，还有不少没有裂帛那么大，但是每年销售额也过亿的小品牌店。而对于重复购买率极高的类目，例如化妆品类目、干果类目，这类商品的重复购买率极高，所以无论如何都需要关注老顾客。而绝大部分类目和卖家，即使是女装这个占淘宝 40%销量的类目，针对老顾客，传统的营销措施（例如发短信、老客户分级分类、建立所谓的 RFM 模型）都是没有任何意义的。对绝大部分店铺来讲，顾客只忠实于价格和淘宝，所谓老顾客是淘宝全网的老顾客，而绝不是自己店铺的老顾客。

当然，我也不是只指出问题，而不给解决思路的专家，抱怨和找问题谁都会，问题是找出这些问题后如何落地？针对老顾客，我有两个建议。

建议一是熟悉阿里妈妈的达摩盘[1]。阿里妈妈内部的直通车和钻展一直在黑盒化地使用数据，即通过算法和数据来帮助广告主去做一些精准的定向。这虽然解决了一大批中小商家的营销需求，但是对于一些定制化的营销需求却仍然很难满足，一些商家希望根据他们特定的营销需求去定制自己的投放人群。因此，由商家自定义组合标签、选择目标人群进行投放，成为了达摩盘这个工具型基础设施平台的基本功能。例如，我们可以通过达摩盘针对全网做具有某类标签的老客户（即针对淘宝的老客户，而不是针对自己的店铺）的广告投放。

建议二是针对自己店铺的老顾客，要获取到这些老顾客的全网购买数据，从而给自己的老顾客打标签，打标签的目的并不是按照 RFM 模型给出所谓的购买频率、购买金额和找出所谓的 VIP 顾客，而是按照特定的标签（例如，这些老顾客全网平均购买价格、购买最多的类目、全网购买频次等）给自己的店铺重新定位。例如，如果自己店铺的老顾客全网平均购买价格高于自己店铺的主推宝贝价格，那么自己店铺的主推

1. DMP，所谓 DMP 就是数据管理平台（Data Management Platform）的英文缩写，阿里妈妈为 DMP 取了一个颇具中国特色的名字：达摩盘。

款就可以提价（当然是在产品质量提高的基础上），或者也可以根据这些老顾客的喜好，并结合淘宝指数（shu.taobao.com），对自己的店铺重新定位，找到一小部分忠实粉丝，针对他们的特点，更好地服务好他们，就像前文中说的妖店一样，找到自己店铺真正的老顾客，并可以导入到微店端。做微店的一个好处是自己本身就是一个闭环，而不像买家在淘宝购物那样，绝大部分买家只忠实于价格，只需简单搜索，就可以找到价格永远比自己店铺宝贝价格低的商品。

综上所述，对于绝大部分卖家来讲，都不可以放弃 SEO 优化，如果说产品质量和供应链需要放在战略地位上的话，那么针对淘宝/天猫的搜索引擎优化，则必须放在战术地位上。如果不是茅台，酒香也怕巷子深，我们必须要研究淘宝/天猫搜索引擎规则，从而使得我们的宝贝在买家搜索的时候，宝贝对应的主关键词能够达到前 3 页，甚至前 2 页。而研究天猫、淘宝搜索引擎规则，我们就必须要抓取天猫、淘宝的数据，从大数据和无序的数据当中找到规律，从而使无序的数据变得有序。

1.2　天猫、淘宝运营数据抓取技术概述

我们明白了天猫、淘宝数据抓取的必要性，但是天猫、淘宝的数据抓取也不是那么容易实现的。在 2014 年 10 月份之前，天猫、淘宝的数据抓取相对来说还是非常容易的，这也造就了不少以天猫、淘宝数据抓取为卖点的卖家工具。但是到了 2014 年年底，淘宝“丧心病狂”地加强了网页反爬虫技术手段，使得抓取天猫、淘宝数据变得异常艰难。如果说爬取一个宝贝的信息还不是很困难的话，那么连续爬取数百个，甚至即使只爬取数十个宝贝信息，就会触发天猫、淘宝反爬虫作弊引擎，常见的有：弹出验证码、返回淘宝登录界面、返回“爬虫无意义”的警告信息，或者当我们通过浏览器查看网页源代码时，汉字转化为 Unicode 编码等，让我们看得一头雾水。

尽管爬取天猫、淘宝的数据越来越难了，但是我这里还是给出一些通用的爬取数据的方法，供卖家老板们参考。由于本书定位的读者对象是卖家老板或者运营人员，而不是具有编程能力的程序员，所以我这里尽可能通俗地讲一些原理性的东西，供具有编程能力的卖家老板参考。另外，我会附一个我和高老师一起做的免费工具，供大家参考使用（下载地址：http://42.120.17.31/xuntu.zip 或者 http://pan.baidu.com/s/1c00NPde）。

因为淘宝会持续改进自己的网页加密算法，所以工具箱也会经常升级，感兴趣的同学，请加 QQ 群：15509455，大家一起讨论。

上面讲了一些天猫、淘宝数据抓取痛点，下面讲讲天猫、淘宝数据抓取的基本步骤，以及针对天猫、淘宝对数据抓取疯狂的屏蔽，我们会采取哪些措施，避开天猫、淘宝的反爬虫。

对通用网站的数据抓取，比如谷歌和百度，都有自己的爬虫，当然，爬虫也都是由程序写出来的。根据百度百科的定义：网络爬虫（又被称为网页蜘蛛、网络机器人），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫。不过，淘宝为了屏蔽网络爬虫对自身数据（例如商品价格、月销量、收藏量、评价、月成交记录等）的抓取，往往是采取 Ajax 技术，在网页加载完成后，再次加载这些数据，所以通用的网络爬虫抓取技术对抓取淘宝的敏感数据是无效的。针对淘宝本身的特点，天猫、淘宝数据抓取技术无外乎以下 4 种。

（1）通用的网页解析技术，适合解析一些常见的数据，例如关键词排名数据、宝贝标题、宝贝下架时间等。

（2）通过浏览器插件技术。无论是 IE、火狐（Firefox）还是谷歌浏览器（Chrome），都有自己的插件技术。淘宝无论如何增强反爬虫技术，数据最终总是要在浏览器里按照正常的数据格式显示出来。所有这些数据（例如商品价格、月销量、收藏量、评价、月成交记录等）在浏览器里正常显示后，通过浏览器插件接口就可以抓取到这些数据了。有的公司就是这么做的。

（3）做一个客户端，在客户端里模拟一个浏览器，然后模拟用户搜索。还是那句话，淘宝无论如何增强反爬虫技术，最终总是要在浏览器里按照正常的数据格式显示数据，现在很多的刷流量工具就是这么做的。

（4）通过一些网页分析工具，分析淘宝网页显示过程，找到呈现商品价格、月销量、收藏量、评价、月成交记录等的 Ajax 链接。同样模拟一个针对这些 Ajax 链接的浏览器请求，从而无须解析网页，直接解析这些 Ajax 返回来的数据就可以达到目的。

由于淘宝对数据的抓取采取的措施越来越严，只用某一种方法有时是不能达到目的的。例如，最简便的无疑是第四种方法，通过网页分析工具，直接找到这些 Ajax 调用，但是淘宝对通过 Ajax 链接调用的次数是有限制的，调用次数一多，触发了淘宝反爬虫作弊引擎，就会出现弹出验证码，或者返回“你已经被反爬虫作弊引擎发现”等声明，就抓取不到想要的这些数据了。所以，最好的数据抓取方法就是 4 种方法相结合。在后续章节中，我会重点介绍客户端数据抓取技术，以及通过一些网页分析工具，分析淘宝网页显示过程，找到呈现商品价格、月销量、收藏量、评价、月成交记录等的 Ajax 链接技术。