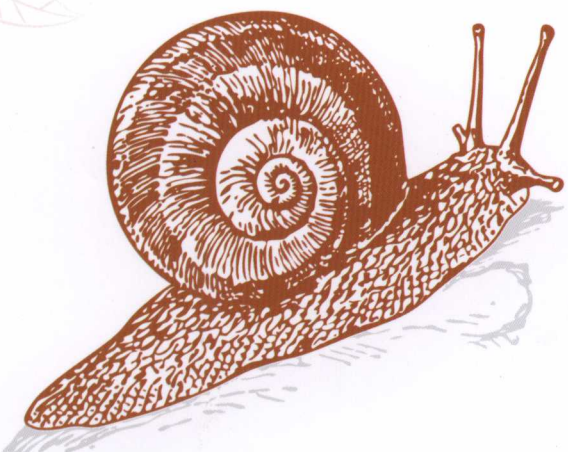
轻松学大数据挖掘

算法、场景与数据产品

汪榕 著



不依赖工具包，结合场景个性化构建业务模型

有数据情怀，更有深刻认知

是数据圈的一股清流，是初学者的入门指南，也是传统挖掘者的进阶之路

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

汪榕（@乐平汪二），一个充满大数据情怀的程序员，致力于分享自己的所感所悟，为数据生态圈的健康发展贡献自己的一份力量。拥有6年的业务建模经验，曾率队夺得全国大学生数据建模比赛一等奖，并代表重庆高校队伍与全国优秀名校的队伍一起参与深圳夏令营建模比赛。

目前从事互联网金融行业，专注于大数据挖掘与数据产品。同时也是大数据挖掘杂谈社区的创建人，该社区会集了全球各地的数据爱好者，共同探索数据的价值。

轻松学大数据挖掘

算法、场景与数据产品

汪榕 著

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

内 容 简 介

伴随着大数据时代的发展，数据价值的挖掘以及产品化逐渐被重视起来。本书作为该领域的入门教程，打破以往的数据工具与技术的介绍模式，凭借作者在大数据价值探索过程中的所感所悟，以故事的形式和读者分享一个又一个的数据经历，引人深思、耐人寻味。全书共 9 章，第 1～2 章介绍数据情怀与数据入门；第 3～6 章讨论大数据挖掘相关的一系列学习体系；第 7～9 章为实践应用与数据产品的介绍。让所有学习大数据挖掘的朋友清楚如何落地，以及在整个数据生态圈所需要扮演的角色，全面了解数据的上下游。

本书可作为相关工作经验在 3 年以内的数据挖掘工程师、转型入门做大数据挖掘的人士或者对数据感兴趣的追逐者的轻松学习教程，引导大家有一个正确的学习方向，也可供对数据产品感兴趣的产品经理和数据挖掘工程师阅读参考。

图书在版编目（CIP）数据

轻松学大数据挖掘：算法、场景与数据产品 / 汪榕著. —北京：电子工业出版社，2018.1
ISBN 978-7-121-32926-5

Ⅰ. ①轻… Ⅱ. ①汪… Ⅲ. ①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2017）第 258260 号

策划编辑：黄爱萍
责任编辑：牛 勇
印 刷：三河市华成印务有限公司
装 订：三河市华成印务有限公司
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：13 字数：230 千字
版 次：2018 年 1 月第 1 版
印 次：2018 年 1 月第 1 次印刷
定 价：59.00 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

这是一本关于大数据挖掘与数据产品的参考读物，为了使尽可能多的读者通过本书对大数据应用有所了解，笔者以个人所感所悟引导初学者正确学习大数据挖掘。但是基础知识归纳、开发环境部署、算法原理的介绍都是不可避免的。因此，本书更适合于工作经验在 3 年以内的数据挖掘工程师，以及转型入门做数据挖掘的人士，或者是对数据产品感兴趣的追逐者阅读。

全书共 9 章，第 1～2 章介绍数据情怀与数据入门；第 3～6 章讨论大数据挖掘相关的一系列学习体系；第 7～9 章为实践应用与数据产品的介绍。

本书在内容上尽可能以故事的形式，轻松愉快地介绍大数据、数据挖掘与数据产品实践应用的各方面内容。但作为学习方向性的引导读物且考虑到本书主题，很多常见的算法、技术知识点未能覆盖，毕竟相关的内容在网上已经有很多了，但大多数内容只是“术”，而缺乏“神”。所以本书才另寻思路，以笔者的真实经历告诉读者在学习过程中可能会遇到的“坑”，以及该如何正确学习。因此，建议有兴趣的读者进一步钻研探索，结合更多的学习资料实践应用。

笔者认为，大数据时代的发展，已经逐渐从基础性的建设、数据的积累，慢慢转变成对于数据价值的探索以及业务痛点的落地解决。因此，建议更多的数据挖掘学习者要结合业务场景思考，多了解数据生态圈的上下游，认清数据产品价值的重要性，以及知晓自身在整个数据流程中所扮演的角色的重要性。阅读这些内容的意义远远超过对数据分析工具、算法模型的熟练度的意义。

大数据、人工智能发展极为迅速，但是数据价值的输出仍然存在瓶颈，极大的原因是由于广大追逐者在对数据探索时走向了误区，把更多心思放在了“玩转数据”，而不是真正地解决业务痛点。所以，希望阅读本书的每一位读者都能够从笔者的过往经历和所感所悟中感受到数据之禅。参与本书编写的人员还有王勇老师，在此表示感谢。

笔者自认自己还有许多需要学习的地方，同时时间和精力有限，书中不足之处在所难免，望广大读者批评指正，不胜感激。

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/32926

目　录

第1章 数据情怀篇

1.1 数据之禅

大数据不是新概念，它一直存在，且不以人的意识为转移。
大数据的价值并不在于积累，而在于用更全面的角度去解读事物本身。
业务场景对于数据而言极其重要，它决定了你的分析思路。
当你沉迷于令人眼花缭乱的技术时，要记得数据才是最本质的一切。
浮躁时，找个时间去观察数据，你会得到意想不到的惊喜。
对待数据，要有敬畏之心。因为假的真不了，真的篡改不了。
不要试图去猜测数据，在你没读懂时，肯定还有一层层迷雾遮挡着你。
世间的万物皆有规律，有因有果，数据的表现也是这个道理。
要做好一个数据人，就要懂得沉淀，这样才能透过现象看到本质。

1.2 数据情怀

谈起大数据，知晓它的人都会说：势头猛、高科技、待遇好。“圈外”的人，迫不及待想一头扎进来。殊不知，“圈里”的大部分人却在坐以待毙，茫然无方向。

这些年，笔者接触过很多工作，如数据开发、数据分析、数据挖掘和产品经理，但都与数据产品相关，从来没改变过。近些年，随着“数据”概念的火热，越来越多的人涌向数据这个领域。

1.2.1 数据情怀这股劲

自始至终，国内真正领悟到大数据产品精髓核心的人并不多，有价值的数据产品更是屈指可数。难道大数据的价值在一款跨时代的数据产品身上这么难体现吗？

归根结底，关键性因素是“数据情怀”惹的祸。为什么这样说？很多身处大数据领域的人，不管是做培训，还是做产品，缺乏真正意义上的那一股劲——“数据情怀”，而这股劲，直接影响着你在为这个领域的蓬勃发展贡献多大的力量。

1.2.2 对数据情怀的理解

数据情怀都体现在哪些方面？概括起来，有以下几个词：

- 初心
- 使命感
- 快感
- 共鸣与傲娇

这是笔者对待大数据的一种态度。下面分别讲几个故事。

初心：不忘初心，方得始终。

有位朋友向我提过这样的问题：你是如何赶上机遇，选择这个领域的？是热爱，还是偶然？我很理解这个问题被提出的出发点，因为我知道现在大数据圈子里有这样一个现象：

- 很大一群“准大数据人”，正在培训班里接受培训或者自己学习。
- 一部分转型做数据开发的大数据人，工作年限在 5 年以上，很多人是从 Java 开发转行过来做大数据框架的，真正接触大数据的时间不会超过两年。
- 一部分转型做数据仓库或数据分析的大数据人，是从传统 BI 数据转过来的。

这样转型，除职业发展中的规划外，也有薪酬水平的原因，很幸运自己就算是其中一个。

故事一：笔者与数学的藕断丝连

笔者是学通信专业的，从小到大数学都很厉害，一路以来，转变过很多方向，都是在寻找一个答案——学数学的意义。

笔者在上大学以前，数学一直不错。上了大学后，还曾经熬夜钻研过哥德巴赫猜想，十分兴奋。但后来想明白了，数学公式的计算、求证和推导，并不是我

感兴趣的。

在大学有机会接触数学建模，顷刻间觉得它是应用数学在实践中的真正应用，是一种知识的融合和思考问题的突破。笔者参加了 11 次比赛，除在深圳参加夏令营遗憾地获得了三等奖，最后一次参加比赛获得美国建模二等奖外，剩余都是一等奖（其中也包括全国大学生数学建模一等奖）。

这时大数据时代来临，笔者觉得从大数据中或许能够找到数学乃至数据真正的意义，这的确是笔者喜欢瞎折腾的一个初心，太想在自己身上找到数学存在的意义了。所以，当时第一个想法是玩转数学。刚开始总是围绕数据源打转，做一些类似阿里指数那样的大数据报表，总想把各种大数据生态圈底层的开发技术都了解到，但这么做费力不讨好，也没有体现出大数据真正的价值在何处。

后来，在从事大数据领域工作的过程中，又转变了一些方向，有幸多次参与对一家美妆公司，甚至是一些高层的调研。花了一个多月的时间，慢慢领悟到业务真正需要数据为它做什么和业务方需要什么样的数据产品。数据真正的价值潜力很大，只是还很少有人去探索成功罢了。

这是自己目前折腾的事，至少这一路的初心，都是在寻找数学乃至数据的价值。并不是每个从事大数据工作的人，都必须要像笔者这样折腾，但至少你需要思考一下，当初选择进入这个圈子是自己的初心，还是执着，或者只是追潮流？

使命感：人这一辈子，能折腾的事不多，用心做好每一件事。

故事二：笔者的朋友圈，一些活跃的、典型的数据人

在笔者的朋友圈有位特别专注于智能金融的“捷哥”，一个从国外回来创业，想在互联网金融这个行业探索数据价值的人；有天天吟诗作乐，深深陶醉在大数据情怀的高总，同时他也有着大数据人才思维培养的重任；有从事自由职业，却天天飞这飞那做培训的黄老师，一直重视着业务与数据紧密结合，推广着自己写的书；有想在培训行业做出一番贡献，一直默默筹备着机会的老李，充满了情怀，立志于打破目前大数据培训的混乱局面。

这些人充满了使命感，即使迷途惆怅，也坚信光明就在远方。我喜欢这样的一群人，只是这样的人在大数据的圈子里面太少太少了。

故事三：特立独行的数据人

有些特立独行的数据人踏入大数据圈子仅仅是为了转型，为了薪酬，为了养老，并不想真正做出点什么。他们拥有一定的专业技能，但总在小圈子里钻，认为不断学习技术才是存在感，却不知技术本身真正的意义和价值，难应用于业务。

快感：一种想到就会小抽搐，跌宕起伏的兴奋。

故事四：最近上线的数据产品，让笔者充满了快感

几年前，领导私下问每个新人，对工作有什么规划，如下类似的答案从别人口中说出：想做资深 Hadoop 运维工程师、架构师、数据仓库大牛等。笔者的回答是：想做一款数据产品。结果被笑不切实际（却没人知道，笔者当初为了面试数据产品经理，整整准备了两大页自己的构思和知识点的整合）。

前些日子，由于个人发展方面的原因，笔者跳槽了，在面试过程中，还是有人问职业规划的问题。笔者认为，会有人相信了，所以说了自己这几年做了很多准备，就是想以后成为数据产品经理，做一款有自己特色的大数据产品。结果出乎意料，都被一一质疑，以及婉拒了。后面我变聪明了，改口说要成为资深数据挖掘师，沉醉于技术海洋里。听者兴奋，说者无心。

很幸运，来目前这家公司的这段时间里，花了半年多的时间，真切地拥有属于自己特色的数据产品了。从无到有，从需求的调研和分析、系统功能的规划和确定，到前后端功能的开发、推动和联调。

共鸣与傲娇：我们天生傲娇，却在渴望寻找着共鸣的声音。

老罗在一次发布会上提到了傲娇这个词，那种由心而然的底气很强烈，每次看发布会直播，笔者都能深深感受到，因为在大数据圈子里也有这样的一面。就像锤子手机，从创办至今，虽然不被一些人看好，但却在办每一次发布会时引起全国、全世界的关注。

能感受到老罗内心里的渴望，渴望共鸣的声音。即使声音很弱、很小，但却急切期待懂他的人能够共鸣，老罗找到了这样一些共鸣。每次听他发布会的“锤粉”们，因为懂他，也都会替他紧紧捏着一把汗。

回到大数据圈子里，每一个圈子里面的人，都在做着改变未来世界的事，都有可能引领大数据科技与生活的完美融合，不管是互联网+、生物医疗、基因工程、智能家居还是人工智能等，太多新领域充满了未知，充满了使命感。所以，我们真正天生傲娇，每个人都是自己的英雄。

1.3 大数据时代的我们

有人说人之所以痛苦，原因在于追求错误的东西。可是笔者认为，很多时候痛苦来源于迷茫和无奈。对笔者而言，不管是生活还是工作，更看重先做一件正

确的事，并且不顾一切地做下去。

2016 年 11 月 16 日到 18 日，世界互联网大会在浙江乌镇举行，全世界都在关注此次大会，笔者也一样很期待。

在李彦宏演讲结束以后，笔者深思了很久，伴随着互联网的日新月异，从移动互联网，到人工智能，在大数据思维全面灌输的时代，我们何时能够追上科技发展的步伐？

不管是战胜李世石的 AlphaGo，还是因为锤子 M1 一炮走红的科大讯飞，这些都是推动大数据实践落地的优秀先驱者。整个大数据环境，从萌芽到逐渐尝试突破层层泥土，让人们看到它有价值的一面，这是好事。

笔者也在做类似的尝试，相信数据产品能够服务于业务，应用于生活，彰显大数据更有价值的一面。就拿反欺诈产品来说，其能够整合全渠道，甚至是第三方的数据源，通过分析用户在平台上的一举一动，以及多个用户之间的强关联性，实时精准地监控用户在生命周期内的异常行为，甚至是识别恶意诈骗团伙。对于公司的运营成本，这样的大数据应用意义非凡。

还有很多这样有价值的应用，它们致力于服务消费者。就像淘宝推出的“聚星台”，随着移动互联网的发展，越来越多的用户群体被从 PC 端引流到手机端，购物，看新闻，寻找饮食。

聪明淘宝人，提出了用户画像和商品画像，精准推荐另一个应用场景——千人千面，异于传统模式下的协同过滤（基于人、物，甚至是商品之间的推荐），更人性化地展示给用户不同的商品宝贝，精准地推荐商品，缩短了用户的购物路径。这样对于具有“选择困难症”的朋友来说是一个福音。

虽然在和淘宝对接的过程中，这样的大数据应用落地效果并不完美，还有待优化，但是，这表示大数据时代带来的价值，已经轰轰烈烈地来了。

不少人会感觉到恐慌、陌生，甚至是无助。因为他们对大数据思维没有任何概念，但是他们都有一个信念，期待能够获得大数据时代“豪华游轮”的一张船票。而笔者想说的是，只要你足够走心，有大数据情怀，并找到正确的方向，你就能登上这艘游轮。

因为现在的大数据环境，还需要更多先驱者来推动这个领域的发展，打破外界对它的偏见，大数据并不是“大忽悠”，而是一种必然。

笔者想到了几年前的那个冬天。

市场：那时候大数据的整体氛围还没有这么强烈，很多公司都是在当时的业务方向上做小数据量的数据分析和挖掘工作，更多的时候是借助一些分析软件来