



热销图书《大数据架构商业之路：从业务需求到技术方案》的续作，深入解析大数据架构和算法在电商环境中的技术实现。作者荣获美国政府颁发的“美国杰出人才”称号，本书集其十多年科研经验之精华。

源码资本合伙人、前金山软件CEO、前微软亚太研发集团首席技术官张宏江先生作序力荐！



技术丛书



Big Data Architecture and Algorithm in Action
The Implementation in e-Commerce Systems

大数据架构和算法实现之路

电商系统的技术实战

黄申◎著



机械工业出版社
China Machine Press



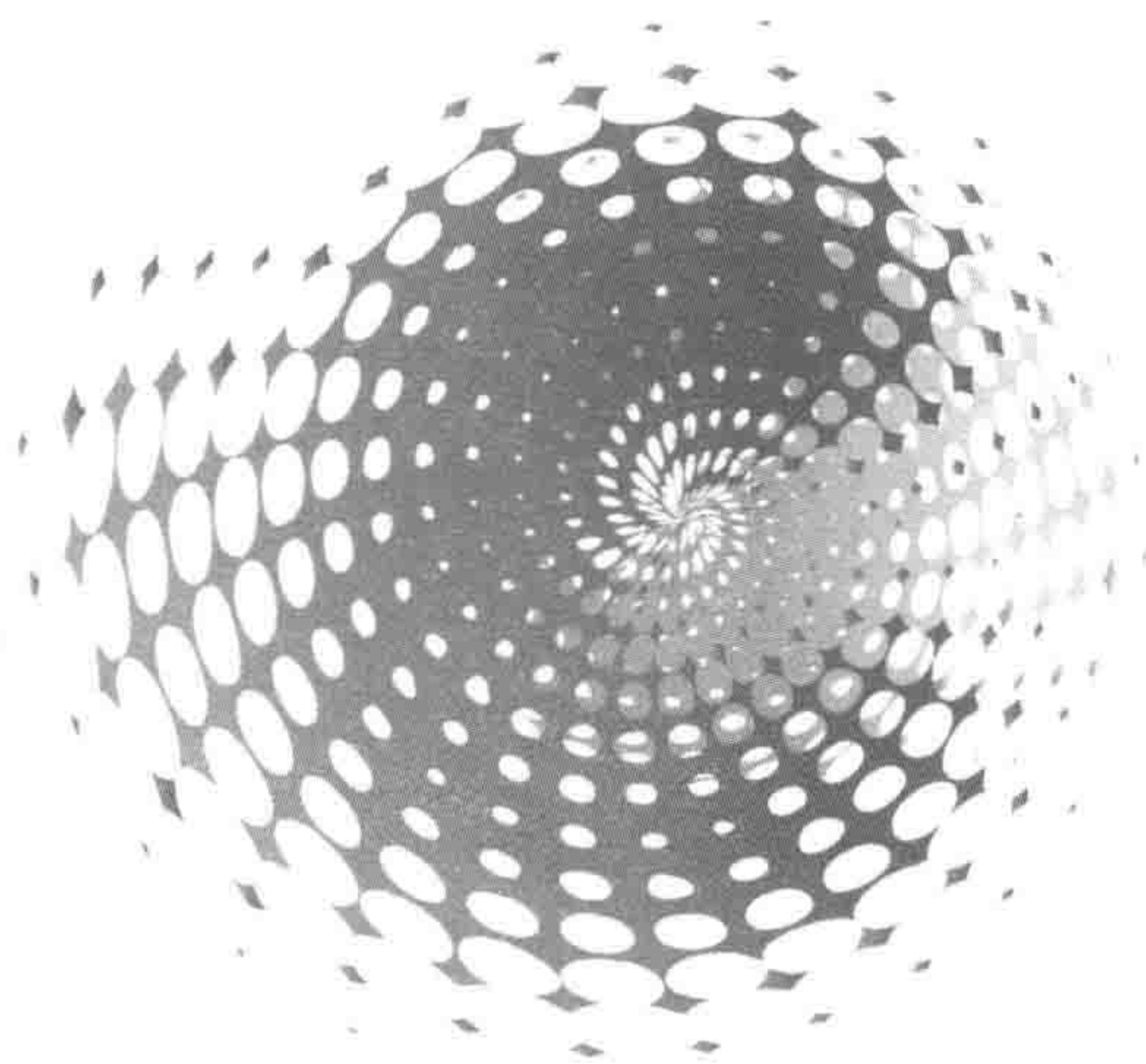
技术丛书

Big Data Architecture and Algorithm in Action
The Implementation in e-Commerce Systems

大数据架构和算法实现之路

电商系统的技术实战

黄申◎著



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

大数据架构和算法实现之路：电商系统的技术实战 / 黄申著. —北京：机械工业出版社，2017.6

(大数据技术丛书)

ISBN 978-7-111-56969-5

I. 大… II. 黄… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字 (2017) 第 101843 号

大数据架构和算法实现之路：电商系统的技术实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：杨绣国 陈佳媛

责任校对：殷虹

印刷：中国电影出版社印刷厂

版次：2017 年 6 月第 1 版第 1 次印刷

开本：186mm × 240mm 1/16

印张：27.5

书号：ISBN 978-7-111-56969-5

定价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010) 88379426 88361066

投稿热线：(010) 88379604

购书热线：(010) 68326294 88379649 68995259

读者信箱：hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

Forward 推荐序

最近的这几年，我们见证了大数据和人工智能如何推动企业的转型和升级。大数据的获取、处理和运营逐渐融入不同规模企业的日常业务中，并成为它们的创新引擎。之前我们就已经看到 Google 的广告业务，它背后存在许多大数据的技术作为支撑，因此，它能够比较精确地预测在什么时候给你推荐什么内容的广告。时至今日，这样的大数据技术越来越多地应用到生活中的各个领域，包括电商、金融、旅游、健康，甚至是游戏和娱乐产业。

不过，在利用大数据技术创新的时候，人们往往面临这样的困惑：对于某类技术，如何找到合适的应用场景？反之亦然。所以，无论是在微软还是金山时，我们都非常强调将科研成果转变为实际的产品过程。在创新的同时，需要找到合理的产品解决方案和定位。本书的作者黄申曾经在微软亚洲研究院工作，从事机器学习相关的研究。之后他加入了 eBay 中国等多家电子商务公司，对于大数据技术在电商领域的应用有着自己独到的见解。相信本书能够从电商业务的需求出发，解析技术实战的难点，探讨大数据和商业的结合之道，帮助大家打造更多实用型的创新产品。

张宏江先生，源码资本合伙人，前金山软件 CEO、前微软亚太研发集团 CTO

2017 年 4 月

前 言 Preface

为什么要写这本书

首先要感谢机械工业出版社华章公司的编辑们，在他们的大力支持下，我于2016年出版了《大数据架构商业之路：从业务需求到技术方案》一书，并获得了良好的销售额和口碑。不少读者主动和我联系，表示从书中学习到了如何使用大数据的知识，来制定合理的技术方案。能够让读者从书中获益，我也感到非常欣慰。与此同时，也有部分读者表示对于技术的细节很感兴趣，对此书未能包含实现部分深感遗憾。对此，我一直在犹豫是否需要重新写一版，包含更多的实战内容。因为《大数据架构商业之路：从业务需求到技术方案》一书的定位是最大程度地弥补业务需求和技术方案之间的空白，针对的读者主要是互联网公司的技术管理人员、产品经理、初级的架构师等。如果直接加入过多的技术细节，可能会导致该书的定位不清，让读者难以获得最佳的阅读体验。

与本书的策划编辑杨老师再三讨论之后，我决定不在原书中加入更多的实现部分，而是重新撰写一本兄弟篇。这本全新的书，仍然会沿用前作的故事背景和应用场景，不过读者对象改为资深的程序员、算法工程师、数据科学家和系统架构师。因此，新作将大幅缩减基础知识的详细介绍以及业务需求的逐步分析，而是直接进入实战的主题，包括系统架构、算法设计，甚至是重要的代码部分。当然，我也不希望该书全由代码堆砌而成，因此主要针对核心代码进行了讲解。全部的实例代码会以其他形式来提供。

虽然定位有所不同，但是我仍然希望保持前作深入浅出的特点。

- 易读易懂。黄小明和杨大宝的创业故事在稍作修改的基础之上得以保留，继续使用生动的案例和形象的比喻来解读难点，降低理解的门槛。
- 可实践性强。本书选取了电子商务的平台，通过分享大量实践才能积累的宝贵经验和重点代码，最大程度地弥补业务需求和技术方案之间的空白。与此同时，针对频繁升

级的开源软件，我也采用了 2016 年年底到 2017 年年初最新的版本。因此，部分代码甚至可作为中小公司创业起步的参考模板。这有利于技术人员针对不同的业务需求，规划更为合理的技术方案。

最后，我们衷心希望本书成为相关领域技术专家的良好益友，大家在阅读之后，对电商大数据的实践能有更加深入的理解，并对自己所从事的项目有所裨益。

读者对象

根据本书撰写的起心动念，我们觉得其内容适合如下的读者。

- 大数据相关领域的程序开发者和技术骨干。从本书中，他们可以看到常见的互联网公司从创业初期到中期，应该怎样设计数据平台、如何解决技术上的难题，才能最终满足业务需求。
- 中小互联网创业公司的数据科学家或者算法工程师。算法是数据平台的一个关键因素。最近几年，人工智能、机器学习乃至深度学习都是学术界和工业界的一大热点，而数据科学家也成为受人追捧的职业。合理地运用智能算法将从很大程度上节约重复劳动的成本，提高效率和转化率，最终增加商业的价值。
- 架构工程师。架构是数据平台的另一个关键因素，很多刚刚从院校毕业、工作没多久的朋友，学了一身的本领，对新技术也很有热情，可惜没有太多实践的机会。本书中的案例，浓缩了不少业界实践的经验心得，如能融会贯通，对他们的工作将有很大帮助。同时，覆盖面较广的技术课题概述，也为他们继续深入研究提供了方向和可能。

总之，本书适合钻研实现细节的程序员、工程师和算法专家。和前作的侧重点有所不同，本书并不适合作为入门教程使用。因此建议没有相关基础知识的读者，读完前作之后再次阅读此书。

如何阅读本书

本书介绍了一些主流技术在商业项目中的应用，包括机器学习中的分类、聚类和线性回归，搜索引擎，推荐系统，用户行为跟踪，架构设计的基本理念及常用的消息和缓存机制。在这个过程中，我们有机会实践 R、Mahout、Solr、Elasticsearch、Hadoop、HBase、Hive、Flume、Kafka、Storm 等系统。如前所述，本书最大的特色就是，从商业需求出发演变到合理的技术方案和实现，因此根据不同的应用场景、不同的数据集合、不同的进阶难度，我们为读者提供了反复温习和加深印象的机会。

勘误和支持

众所周知，大数据的发展实在是太快了。可能就在你阅读这段文字的同时，又有一项新的技术诞生了，N项技术升级了，M项技术被淘汰了。再加之笔者的水平有限，书中难免会出现一些不够准确或遗漏的地方，恳请读者通过如下的渠道积极建议和斧正，我们很期待能够收到你们的真挚反馈。

QQ: 36638279

微信: 18616692855

邮箱: s_huang790228@hotmail.com

LinkedIn: <https://cn.linkedin.com/in/shuang790228>

致谢

首先要感谢上海交通大学和俞勇教授，你们给予我不断学习的机会，带领我进入了大数据的世界。同时，感谢阿里云的高级总监薛贵荣，你的指导让我树立了良好的科研态度。

还要感谢微软亚洲研究院、eBay 中国研发中心、沃尔玛 1 号店、大润发飞牛网和 IBM 中国研发中心，在这些公司十多年的实战经验让我收获颇丰，也为本书的铸就打下了坚实的基础。

感谢曾经的微软战友陈正、孙建涛、Ling Bao、曾华军、张本宇、沈抖、刘宁、严峻、曹云波、王琼华、康亚滨、胡健、季蕾等，eBay 的战友逢伟、王强、王骁、沈丹、Yongzheng Zhang、Catherine Baudin、Alvaro Bolivar、Xiaodi Zhang、吴晓元、周洋、胡文彦、宋荣、刘文、Lily Yu 等，沃尔玛 1 号店的战友韩军、王欣磊、胡茂华、付艳超、张旭强、黄哲铿、沙燕霖、郭占星、聂巍、邵汉成、张珺、胡毅、邱仔松、孙灵飞、凌昱、王善良、廖川、杨平、余迁、周航、吴敏、李峰，熊健等，大润发飞牛网的战友王俊杰、陈俞安、蔡伯璟、陈慧文、夏吉吉、文燕军、杨立生、张飞、代伟、陈静、赵瑜、李航等，IBM 的战友李伟、谢欣、周健、马坚、刘钧、唐显莉等。要感谢的同仁太多，如有遗漏敬请谅解，很怀念和你们并肩作战的日子，那段时间让我学习到了很多。

感谢机械工业出版社华章公司的编辑杨绣国（Lisa）老师，感谢你的魄力和远见，在最近的 3 个月中始终支持我的写作，你的鼓励和帮助引导我顺利完成了全部书稿。也要感谢凌云为我引荐了如此优秀的出版社和编辑。

衷心感谢源码资本合伙人、前金山软件 CEO、前微软亚太研发集团 CTO 张宏江先生，非常荣幸他能在百忙之中抽空为本书作序。也衷心感谢 Apache Kylin 联合创建者及 CEO 韩卿先生，饿了么 CTO 张雪峰先生、CloudBrain 的创始人张本宇先生为本书撰写推荐语。

还要感谢我和太太双方的父母，感谢你们对我写书的理解和支持。

最后我一定要谢谢我的太太 Stephanie 和宝贝儿子 Polaris，为了此书我周末陪伴你们的时间更少了。你们不但没有怨言，而且时时刻刻为我灌输着信心和力量，感谢你们！

谨以此书，献给我最亲爱的家人，以及众多热爱大数据的朋友们。

黄 申

美国，硅谷，2017年3月

目 录 *Contents*

推荐序	
前言	
引子	1
第一篇 支持高效的运营	
第 1 章 方案设计和选型：分类	5
1.1 分类的基本概念	6
1.2 分类任务的流程	7
1.3 算法：朴素贝叶斯和 K 最近邻	8
1.3.1 朴素贝叶斯	8
1.3.2 K 最近邻	9
1.4 分类效果评估	10
1.5 相关软件：R 和 Mahout	12
1.5.1 R 简介	12
1.5.2 Mahout 简介	13
1.5.3 Hadoop 简介	14
1.6 案例实践	17
1.6.1 实验环境设置	17
1.6.2 中文分词	18
1.6.3 使用 R 进行朴素贝叶斯分类	22
1.6.4 使用 R 进行 K 最近邻分类	37
1.6.5 单机环境使用 Mahout 运行朴素贝叶斯分类	39
1.6.6 多机环境使用 Mahout 运行朴素贝叶斯分类	47
1.7 更多的思考	58
第 2 章 方案设计和选型：聚类	60
2.1 聚类的基本概念	60
2.2 算法：K 均值和层次型聚类	61
2.2.1 K 均值聚类	61
2.2.2 层次型聚类	62
2.3 聚类的效果评估	64
2.4 案例实践	66
2.4.1 使用 R 进行 K 均值聚类	66
2.4.2 使用 Mahout 进行 K 均值聚类	69
第 3 章 方案设计和选型：因变量连续的回归分析	74
3.1 线性回归的基本概念	74
3.2 案例实践	76
3.2.1 实验环境设置	76

- 3.2.2 R 中数据的标准化 78
- 3.2.3 使用 R 的线性回归分析 81

第二篇 为顾客发现喜欢的商品： 基础篇

第 4 章 方案设计和选型：搜索 94

- 4.1 搜索引擎的基本概念 94
 - 4.1.1 相关性 95
 - 4.1.2 及时性 97
- 4.2 搜索引擎的评估 100
- 4.3 为什么不是数据库 103
- 4.4 系统框架 104
 - 4.4.1 离线预处理 104
 - 4.4.2 在线查询 107
- 4.5 常见的搜索引擎实现 108
 - 4.5.1 Lucene 简介 108
 - 4.5.2 Solr 简介 113
 - 4.5.3 Elasticsearch 简介 120
- 4.6 案例实践 123
 - 4.6.1 实验环境设置 123
 - 4.6.2 基于 Solr 的实现 123
 - 4.6.3 基于 Elasticsearch 的实现 154
 - 4.6.4 统一的搜索 API 175

第三篇 为顾客发现喜欢的商品： 高级篇

第 5 章 方案设计和选型： NoSQL 和搜索的整合 195

- 5.1 问题分析 195

- 5.2 HBase 简介 196
- 5.3 结合 HBase 和搜索引擎 203
- 5.4 案例实践 204
 - 5.4.1 实验环境设置 204
 - 5.4.2 HBase 的部署 205
 - 5.4.3 HBase 和搜索引擎的集成 211

第 6 章 方案设计和选型： 查询分类和搜索的整合 219

- 6.1 问题分析 219
- 6.2 结合分类器和搜索引擎 219
- 6.3 案例实践 225
 - 6.3.1 实验环境设置 225
 - 6.3.2 构建查询分类器 226
 - 6.3.3 定制化的搜索排序 229
 - 6.3.4 整合查询分类和定制化
排序 236

第 7 章 方案设计和选型： 个性化搜索 245

- 7.1 问题分析 245
- 7.2 结合用户画像和搜索引擎 245
- 7.3 案例实践 249
 - 7.3.1 用户画像的读取 250
 - 7.3.2 个性化搜索引擎 253
 - 7.3.3 结果对比 260

第 8 章 方案设计和选型： 搜索分片 267

- 8.1 问题分析 267
- 8.2 利用搜索的分片机制 269

8.3 案例实践	271
8.3.1 Solr 路由的实现	271
8.3.2 Elasticsearch 路由的实现	278

第9章 方案设计和技术选型：

搜索提示

9.1 问题分析	283
9.2 案例实践：基础方案	284
9.2.1 Solr 搜索建议和拼写纠错的实现	284
9.2.2 Elasticsearch 搜索建议和拼写纠错的实现	286
9.3 改进方案	291
9.4 案例实践：改进方案	294

第10章 方案设计和技术选型：

推荐

10.1 推荐系统的基本概念	305
10.2 推荐的核心要素	306
10.2.1 系统角色	306
10.2.2 相似度	307
10.2.3 相似度传播框架	307
10.3 推荐系统的分类	307
10.4 混合模型	311
10.5 系统架构	312
10.6 Mahout 中的推荐算法	313
10.7 电商常见的推荐系统方案	314
10.7.1 电商常见的推荐系统方案	314
10.7.2 相似度的计算	317

10.7.3 协同过滤	319
10.7.4 结果的查询	320

10.8 案例实践	321
10.8.1 基于内容特征的推荐	321
10.8.2 基于行为特征的推荐	341

第四篇 获取数据，跟踪效果

第11章 方案设计和技术选型：

行为跟踪

11.1 基本概念	370
11.1.1 网站的核心框架	370
11.1.2 行为数据的类型	371
11.1.3 行为数据的模式	372
11.1.4 设计理念	374
11.2 使用谷歌分析	375
11.3 自行设计之 Flume、HDFS 和 Hive 的整合	378
11.3.1 数据的收集——Flume 简介	378
11.3.2 数据的存储——Hadoop HDFS 回顾	382
11.3.3 批量数据分析——Hive 简介	383
11.3.4 Flume、HDFS 和 Hive 的整合方案	386
11.4 自行设计之 Flume、Kafka 和 Storm 的整合	386
11.4.1 实时性数据分析之 Kafka 简介	386

11.4.2	实时性数据分析之 Storm 简介.....	388	11.5.4	自主设计实战之 Flume、 HDFS 和 Hive 的整合.....	401
11.4.3	Flume、Kafka 和 Storm 的 整合方案.....	390	11.5.5	自主设计实战之 Flume、 Kafka 和 Storm 的整合.....	410
11.5	案例实践.....	391	11.6	更多的思考.....	424
11.5.1	数据模式的设计.....	392	后记.....		425
11.5.2	实验环境设置.....	392			
11.5.3	谷歌分析实战.....	394			

引子

上海，又是一个春天，阳光透过薄薄的窗帘，懒懒散散地洒入屋内。当一缕光线偷偷地爬到杨大宝的眼角时，他睁开了朦胧的双眼。

等等！杨大宝是何许人也？

杨大宝，姓杨名大宝，土生土长的上海人，从小就喜欢玩电子产品，大学的专业是计算机科学，酷爱信息技术和互联网。自从大学毕业后，就一直就职于一家大型IT公司。最近，他面临着人生的一项重大选择。原来，有几位志同道合的朋友，想拉他一起开创公司。大宝很清楚，这几年中国迎来了创业的黄金时代。李克强总理提出的“大众创业，万众创新”，明确了政策对创业的大力支持。同时，老百姓的生活水平正在不断提高，各方面的需求也在不断增加，各种风险投资非常充裕。在这样的大背景下，大家的创业热情空前高涨，尤其是互联网，简直可以用“疯狂”来形容。大宝觉得这正是实现自己梦想的一个好契机。不过，放弃目前优厚的薪资待遇和受人尊敬的公司职位，和几个小伙伴去闯荡江湖，也是要冒不少风险的，最终是否能够成功也充满了变数，这样做到底值得吗？大宝这几天夜不能寐，就连晚上做梦也要纠结一番。若不是淘气的阳光溜进来，可能他还要继续在梦里思考。

洗漱完毕，大宝一边吃着早餐，一边接着整理思路。首先，创业的点子是不错的，主要思想是做线上线下O2O（Offline to Online）的社区商业模式：将大型社区周边的各种服务行业进行线上化，让用户足不出户，就可以叫外卖、订座，享受美甲、按摩等服务，还可以购买商品。用户的生活需求能够得到更大程度的满足，商家也可以吸引到更多的线上客流，而公司的平台也能从双方的交易中获得收益，形成多方互赢的局势，市场前景光明。其次，因为大宝是团队里唯一懂得IT技术的骨干，那么公司里整个庞大的网络系统架构肯定会由他来负责了。这几年的工作经历，让他也积攒了不少设计和开发的实战经验。对于后端的例如数据库、ERP（Enterprise Resource Planning）系统、图片服务器，前端的例如会员注册、购物流程、页面展示等，大宝都有很深入的了解。不过他还是隐约觉得缺了些什么。

吃完早餐后，大宝熟练地打开电脑，开始飞快地在网上查阅资料，钻研成功的互联网站点是如何设计和架构的。就这样，时钟滴滴答答地走着，不知不觉一天又过去了。随着夜幕的降临，望着窗外柔和的街灯，大宝深深地吐了口气。“还缺一个关键词：大数据”，这是他一天研究下来得出来的结论。

等等？大数据又是什么？

好问题，其实此刻大宝心里也没谱，但是他看到好多资料都反复提到这个词。他隐约觉得，如果没有摸清这一点，那么对于这个初创公司而言就会存在很大的不确定性。可是，目

前创业的团队也很多，竞争相当激烈，从来都不缺乏好的创意，就看谁能先做得出、做得好、做得快。没有太多的时间留给大宝了。该如何是好呢？突然，大宝想到一个人，也许能为他解决心中的这个疑惑。

此人就是黄小明，是大宝的表哥。他是知青子女，从小随父母到武汉生活和读书，到16岁的时候回到上海，考入了知名的高校，并且获得了计算机科学的博士学位，可谓知识渊博。毕业后他在几家世界知名的互联网和电子商务公司任职，有十多年的科研和开发经验，目前正在带领团队攻关几个核心项目。去年还出版了《大数据架构商业之路》一书，口碑很赞。

终于，在一个美好的周末下午茶时间，大宝约到了小明。大宝开门见山，针对自己目前的状况和思考的问题进行了说明。

“嗯……大宝，大数据的确是个非常重要的领域，而且想要上手也有一定的难度。”

“哦，为什么呢？”

“大数据入门的门槛比较高，原因有几点：知识面非常广，技术含量也比较高，此外发展和更新的速度也快得惊人。更为关键的是，这些技术一般都是开源的，很多都需要自己去摸索和积累。除非你们考虑直接使用一些大公司比较成熟的付费方案。”

“嗯，因为是创业起步阶段，我们肯定是不会考虑昂贵的商业解决方案的。”

“那问题就更加复杂了……不过……”

“不过什么？”

“如果你肯花些功夫学习，或许我能给你一些建议和启发。”

“哈哈，小明哥，搞了半天是你要自卖自夸啊！”

“这都被你看出来了。其实我在去年就出版过一本关于大数据的书，其中介绍了不少有关的基础知识和理论，并融入了这些年的心得体会，你有兴趣的话可以先看看这本书。”

“哈哈，你说的是《大数据架构商业之路》那本书吧，我已经开始拜读啦！不过，那本书偏重于理论知识，对于实际开发介绍得太少了。”

“那这样吧，结合你的实际工作需要，以及项目中的难点和挑战，我们一起来实践下如何？”

“那当然求之不得！”

第一篇 *Part 1*

支持高效的运营

- 第1章 方案设计和选型：分类
- 第2章 方案设计和选型：聚类
- 第3章 方案设计和选型：因变量
连续的回归分析

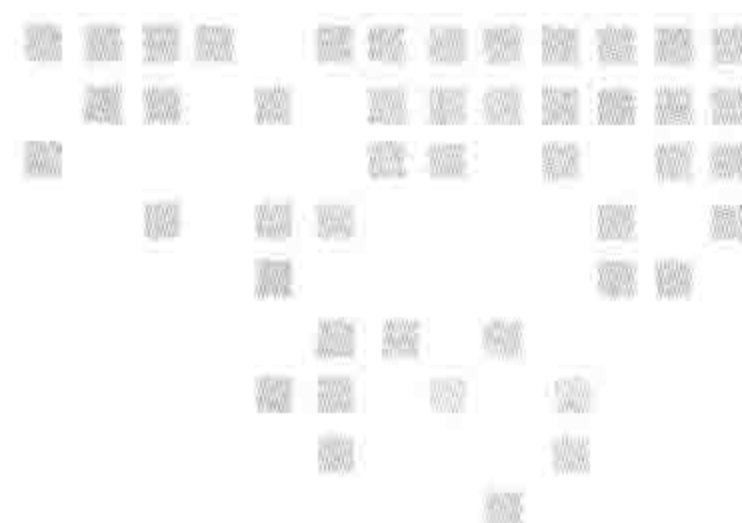
大宝和伙伴们的创业很快就开始了，由于其提供了线上线下无缝结合的社区商业模式，公司业务发展得相当顺利，陆续接入了几个大型社区和商圈周边的各种服务行业。整个线上系统的商品丰富度也相当不错，涵盖了衣食住行多个方面。然而，随着在线商品的不断增长，商家们发现已有的运营工具存在非常大的局限性，工作效率很难得到提高。随着商家抱怨程度的加剧，团队开始急于找到问题的根源所在，于是指派负责运营的小丽对商家进行了深入的访谈和调研。在收集完众多商家的反馈之后，小丽找到了大宝。

“大宝，早上好。有时间吗？最近我们部门对商家进行了一些访谈，深入讨论了他们提出的运营效率问题。其中有一些可能需要你们技术部的大力支持，所以我今天特地来和你沟通一下。”

“嗨，小丽，你好。没问题，你将问题说来听听”。

“我稍微整理了一下，大体上可以分为三个主要的痛点。第一点，缺乏帮助商家找到准确分类的工具。你也知道，目前我们的业务飞速发展，上架的商品琳琅满目，商品最细粒度的分类都超过了 5000 项。这对公司的成长来说无疑是好消息，不过对于运营人员而言可谓是噩梦。海量的类目信息让他们难以为自己的新商品找到合适的分类，偶尔还会发生错放类目的现象。第二点，缺乏帮助商家进行合理关键词 SEO (Search Engine Optimization) 的工具。我们最近也引进了不少新的商家，他们在传统的线下行业中有很强的竞争力，但是对线上的电子商务运营却知之甚少，甚至都不知道怎样合理地在文描中阐述自己的商品。第三点，缺乏可以预测商品转化率的工具。对于零售等消费领域而言，销量和转化率无疑是衡量业绩最为关键的因素。传统行业的商家大多数还是依靠经验和有限的销量报告来预估未来的销售情况。他们想知道，在电子商务的大环境下，是否能够利用大量的历史数据，来实现同样的目标。”

小明听完后感觉有些迷茫，他觉得普通的 IT 技术好像无法解决这些问题。于是，他找到了小明，希望小明能从大数据技术的角度，为他提供一些指导。



方案设计和技术选型：分类

听完大宝关于第一点的描述，小明很肯定地说：“你们的商家应该是需要这样的一个功能：在他们发布商品的时候，系统会自动地为其推荐合适的商品分类，其界面示意图如图 1-1 所示。如果商家希望出售一台苹果的 Mac Pro 笔记本电脑，输入 ‘MacBook Pro’ 后，系统能够自动为其提示最为相关的三个分类 ‘笔记本电脑’、‘笔记本配件’ 和 ‘其他数码’。这是由后台的分类算法来实现的，如果该算法足够聪明，那么它推荐的第一个分类就应该是正确的，商家只需要点击选择即可。这样，既方便了商家的商品发布，又避免了粗心大意而导致的错误分类。而且，对于少数企图违规操作的商家，如果他们选择了和系统默认推荐相差甚远的分类选项，其行为也会被系统记录在案，然后定期生成报表，提交给运营部门进行核查。如此一来，人们就不用再在纷繁复杂的类目中痛苦摸索，工作的效率也会大幅提升。”

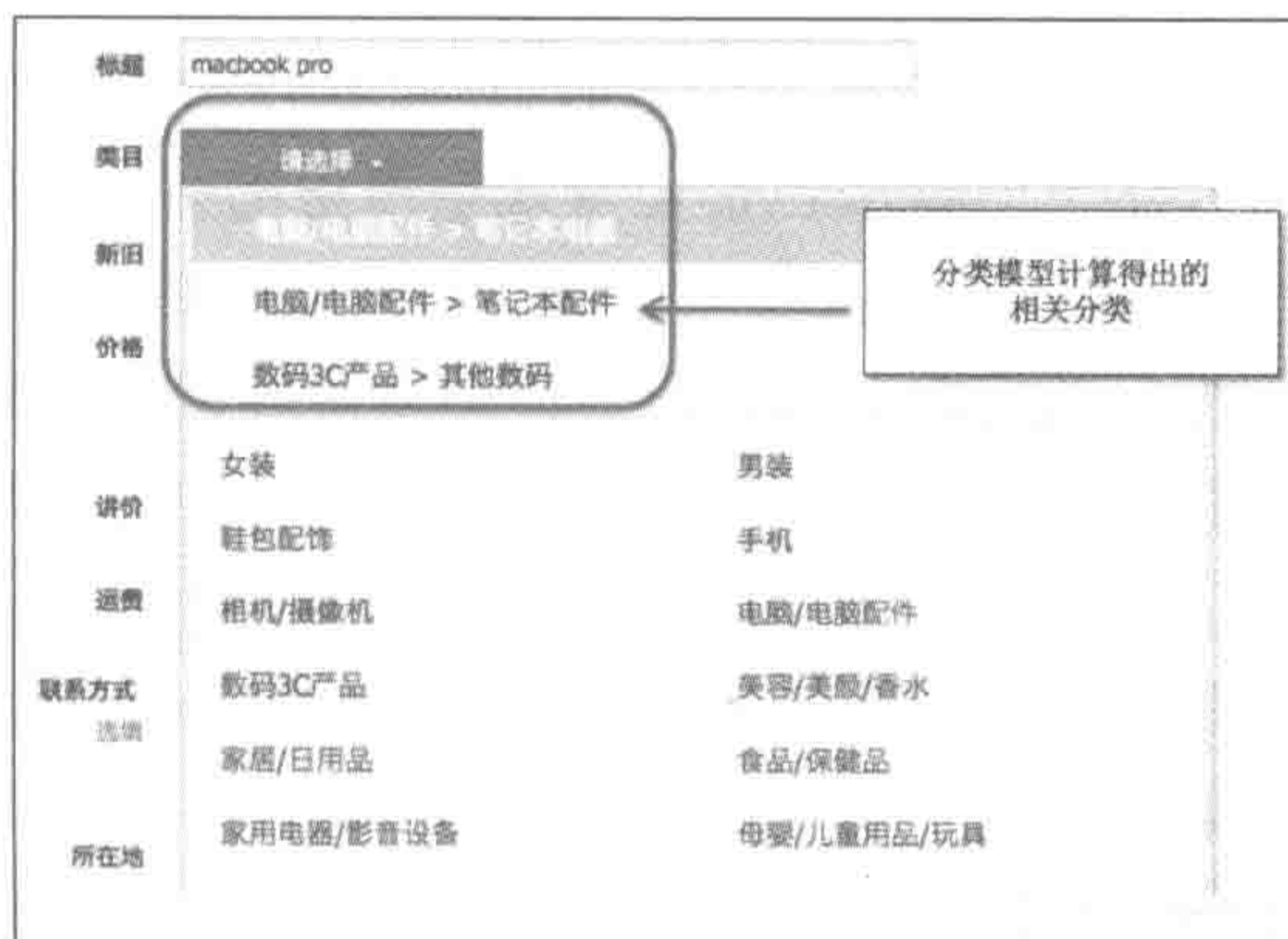


图 1-1 类目自动化分类的应用

“没错，这应该是商家愿意使用的工具，如果真能实现那就太棒了。不过，你刚刚提到的分类算法是什么？”

“分类，是一个典型的监督式机器学习方法”。

“哦，什么是机器学习？什么是监督式的学习？”