阿里巴巴数据分析专家卢辉撰写，多年数据挖掘实践经验总结

阿里巴巴数据委员会会长，中国商业智能领域最具知名度的领军人物车品觉热情推荐

实战性强，从数据分析师的角度对商业实战进行了总结和归纳，以大量事实和案例展现了“以业务为核心，以思路为重点，以挖掘技术为辅佐”的数据挖掘商业实践思想

华章科技 HZ BOOKS

大数据 技术丛书

Data Mining, Leading with Data Driven Practice

数据挖掘与数据化运营实战

思路、方法、技巧与应用

卢辉◎著

机械工业出版社
China Machine Press

大数据技术丛书

数据挖掘与数据化运营实战：思路、方法、技巧与应用

卢辉 著

HZ BOOKS
华章图书

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘与数据化运营实战：思路、方法、技巧与应用 / 卢辉著. —北京：机械工业出版社，2013.6
（大数据技术丛书）

ISBN 978-7-111-42650-9

I. 数… II. 卢… III. 数据采集 IV. TP274

中国版本图书馆CIP数据核字（2013）第111479号

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问 北京市展达律师事务所

本书是目前有关数据挖掘在数据化运营实践领域比较全面和系统的著作，也是诸多数据挖掘书籍中为数不多的穿插大量真实的实践应用案例和场景的著作，更是创造性地针对数据化运营中不同分析挖掘课题类型，推出一一对应的分析思路集锦和相应的分析技巧集成，为读者提供“菜单化”实战锦囊的著作。作者结合自己数据化运营实践中大量的项目经验，用通俗易懂的“非技术”语言和大量活泼生动的案例，围绕数据分析挖掘中的思路、方法、技巧与应用，全方位整理、总结、分享，帮助读者深刻领会和掌握“以业务为核心，以思路为重点，以分析技术为辅佐”的数据挖掘实践应用宝典。

全书共19章，分为三个部分：基础篇（第1～4章）系统介绍了数据分析挖掘和数据化运营的相关背景、数据化运营中“协调配合”的核心，以及实践中常见分析项目类型；实战篇（第6～13章）主要介绍实践中常见的分析挖掘技术的实用技巧，并对大量的实践案例进行了全程分享展示；思想意识篇（第5章，第14～19章）主要是有关数据分析师的责任、意识、思维的培养和提升的总结和探索，以及一些有效的项目质控制度和经典的方法论介绍。

HZ BOOKS
华章图书

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：朱秀英
印刷
2013年6月第1版第1次印刷
186mm × 240 mm · 17.25印张
标准书号：ISBN 978-7-111-42650-9
定 价：59.00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）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

推 荐 序

所谓，自知者明。

一个数据分析师，在面对海量数据时，偶尔把自己也当做对象去分析、思考、总结，才能成为一位有那么点儿味道的数据分析师，才能不断地审视、提升分析水平，才能在数据分析的道路上走得更远。

本书就是作者卢辉对过去 10 年数据挖掘职业生涯的自省、总结、提炼。

以前看的数据挖掘书籍，很难看到国内企业的完整实例。而本书分享的数据化运营实战案例都是来自阿里巴巴 B2B 近 3 年来的商业实践，有立竿见影的案例，也有充满了波折和反复的案例。面对这些实战中的挫折和曲折，作者分享了如何调整思路、调整方法，如何与业务方一起寻找新方案，最终如何达成满意的商业应用效果。这些分享都非常真实、非常可贵，相信这些完整的实战案例将给你全新的阅读体验，还你一个真实清楚的有关数据挖掘商业应用的原貌，也会对读者今后的数据挖掘商业实践起到很好的启迪和参考作用。

从这个角度看，本书就是作者摸索出的一系列有关数据挖掘和数据化运营的规律，是作者对数据分析师有效工作方法的框架和总结。

如果你是新入行（或者有兴趣进入数据分析行业）的读者，这本书对你是非常有参考和指导意义的：帮助你尽快入门，尽快成长。如果你是已具有一定工作经验的数据分析专业人士，本书亦可作为一面“镜子”，去引发你对于“自己的思考”、“自己的总结”。

通过阅读本书，读者朋友们可以问问自己：

- ❑ 数据分析挖掘的技巧，掌握了多少？
- ❑ 书中的实战案例，有实操过吗？
- ❑ 数据分析师对分析 / 数据的态度，你是否具备？

❑ 如何有效管理团队？

如果上述某些方面你没有想过，这本书会给你有意义的启迪。

最后，请允许我再唠叨些数据的未来吧：

关于分析师

不久的将来，或就是现在，数据分析师将直面新的挑战（也是一次转型机会）：在原有分析师职业定位上，为了与业务应用更加贴合，开始逐步融入产品经理“角色”：善于总结、善于提炼、善于推而广之、善于把自己的分析“产品化”。要做到这些，就要求数据分析师必须对数据的理解更透彻，对商业的理解更深入。

在成熟阶段，数据分析师们将是一群具备了商业理解、数据分析、商业应用思考这三大核心能力的综合体。

关于数据质量

在数据化运营道路上，有不少难题亟待解决。其中最棘手、最突出的就是数据质量。

企业的数据化商业实践中，“数据给自己用”与“数据给别人去用”是两个完全不同难度的课题，好比在家做几道家常菜和开餐厅，后者对于口味质量更为严格苛刻，食客们的眼睛都盯着呢。

这本书写了“自己使用数据、分析数据”的心得；在未来，当数据可以很容易地被大家使用的时候，我们会发现数据的力量已经渗透到每个人的决策环节里了。

车品觉

阿里巴巴数据委员会会长

华章图书

前　言

为什么要写这本书

自从 2002 年第一次接触“数据挖掘”（Data Mining）这个新名词以来，转眼之间我已经在数据挖掘商业应用相关领域度过了 11 年。这 11 年里我既见识了国外数据挖掘商业应用如火如荼地开展；又经历了从 21 世纪开始，国内企业在数据挖掘商业应用中的摸索起步，到如今方兴未艾的局面；更有幸在经历了传统行业的数据挖掘商业应用之后，投身到互联网行业（当今数据分析商业应用热火朝天、发展最快，并且对数据和数据挖掘的商业应用依赖性最强的行业）的数据挖掘商业实践中。这 11 年是我职业生涯中最为重要的一段时光，从个人生存的角度来说，我找到了谋生和养家糊口的饭碗——数据挖掘工作；从个人归属的角度来说，我很幸运地碰到了职业与兴趣的重合点。

在国内，“数据挖掘”作为一门复合型应用学科，其在商业领域的实践应用及推广只有十几年的时间，在此期间，国内虽然陆续出版了一些相关的书籍，但是绝大多数都是基于理论或者国外经验来阐述的，少有针对国内企业相关商业实战的详细介绍和分享，更缺乏从数据分析师的角度对商业实战所进行的总结和归纳。因此，从商业应用出发，基于大量的商业实战案例而不是基于理论探讨的数据挖掘应用书籍成为当今图书市场和广大“数据挖掘”学习者的共同需求。

同时，在有幸与数据挖掘商业实践相伴 11 年之后，我也想稍微放慢些脚步，正如一段长途跋涉之后需要停下脚步，整理一路经历的收获和感悟一样，我希望将自己一路走来的心得与体会、经验与教训、挫折与成绩整理出来。

基于以上原因，我决定从数据挖掘的商业需求和商业实战出发，结合我 10 多年来在不同行业（尤其是最近 4 年在互联网行业）的大量数据挖掘商业实战项目，将自己这些年来积累的经验和总结分享出来，希望能够起到抛砖引玉的作用，为对数据挖掘商业实践感兴趣的朋友、

爱好者、数据分析师提供点滴的参考和借鉴。同时，鉴于“数据化运营”在当今大数据时代已经成为众多（以后必将越来越多）现代企业的普遍经营战略，相信本书所分享的大量有关数据化运营的商业实践项目也可以为企业的管理层、决策层提供一定程度的参考和借鉴。

我相信，本书总结的心得与体会，可以推动自己今后的工作，会成为我的财富；同时，这些心得与体会对于部分数据分析师来说也可以起到不同程度的参考和借鉴作用；对于广大对数据挖掘商业应用感兴趣的初学者来说也未尝不是一种宝贵经验。

我是从机械制造工艺与设备这个与“数据挖掘”八竿子打不着的专业转行到数据挖掘商业应用行业的，这与目前国内绝大多数的数据分析挖掘专业人士的背景有较大差别（国内绝大多数数据分析挖掘专业人士主要来自统计专业、数学专业或者计算机专业）。我的职业道路很曲折，之所以放弃了自己没兴趣的机械制造工艺与设备专业，是因为自己喜欢市场营销。有幸在国外学习市场营销专业时了解并亲近了国外市场营销中的核心和基石——市场营销信息学（Marketing Informatics）。当然，这是国外 10 多年前的说法，换成行业内与时俱进的新说法，就是时下耳熟能详的“数据分析挖掘在市场营销领域的商业实践应用”）。说这么多，其实只是想告诉有缘的对数据挖掘商业实践感兴趣的朋友，“以业务为核心，以思路为重点，以挖掘技术为辅佐”就是该领域的有效成长之路。

很多初学者总以为掌握了某些分析软件，就可以成为数据分析师。其实，一个成功的数据挖掘商业实践，核心的因素不是技术，而是业务理解和分析思路。本书自始至终都在力图用大量的事实和案例来证明“以业务为核心，以思路为重点，以挖掘技术为辅佐”才是数据挖掘商业实践成功的宝典。

另外，现代企业面对大数据时代的数据化运营绝不仅仅是数据分析部门和数据分析师的事情，它需要企业各部门的共同参与，更需要企业决策层的支持和推动。

读者对象

- ❑ 对数据分析和数据挖掘的商业实践感兴趣的大专院校师生、对其感兴趣的初学者。
- ❑ 互联网行业对数据分析挖掘商业实践感兴趣的运营人员以及其他专业的人士。
- ❑ 实施数据化运营的现代企业的运营人员以及其他专业的人士，尤其是企业的管理者、决策者（数据化运营战略的制定者和推动者）。
- ❑ 各行各业的数据分析师、数据挖掘师。

勘误和支持

由于作者水平和能力有限，编写时间仓促，不妥之处在所难免，在此恳请读者批评指正。作者有关数据挖掘商业实践应用的专业博客“数据挖掘　人在旅途”地址为 http://shzxqdj.blog.163.com，欢迎读者和数据挖掘商业实践的爱好者不吝赐教。另外，如果您有关于数据挖掘商业实践的任何话题，也可以发送邮件到邮箱 chinadmer@163.com ，期待你们的反馈意见。

如何阅读本书

本书分为 19 章。

第 1 ~ 4 章为基础和背景部分，主要介绍数据分析挖掘和数据化运营的相关背景、数据化运营中“协调配合”的本质，以及实践中常见的分析项目类型。

第 6 ~ 13 章是数据分析挖掘中的具体技巧和案例分享部分，主要介绍实践中常见的分析挖掘技术的实用技巧，并对大量的实践案例进行了全程分享展示。

第 5 章，第 14 ~ 19 章是有关数据分析师的责任、意识、思维的培养和提升的总结与探索，以及一些有效的项目质控制度和经典的方法论。

本书几乎每章都会用至少一个完整翔实的实战案例来进行说明、反复强化“以业务为核心，以思路为重点，以挖掘技术为辅佐”，希望能给读者留下深刻印象，因为这是数据挖掘商业实践成功的宝典。

HZ BOOKS
华章图书

致谢

首先要感谢机械工业出版社华章公司的杨绣国（Lisa）编辑，没有您的首倡和持续的鼓励，我不会想到要写这样一本来自实践的书，也不会顺利地完成这本书。写作过程中，您的帮助让我对“编辑”这个职业有了新的认识，编辑就是作者背后的无名英雄。在本书出版之际，我向 Lisa 表达我深深的感谢和祝福。同时感谢朱秀英编辑在本书后期编辑过程中付出的辛劳，您的专业、敬业和细心使得书稿中诸多不完善之处得以修正和提高。

作为一名 30 多岁才从机械工程师转行，进入数据挖掘及其商业实践的迟到者，我在数据挖掘的道路上一路走来，得到了无数贵人的帮助和提携。

感谢我的启蒙导师，加拿大 Dalhousie University 的数据挖掘课程教授 Tony Schellinck。他风趣幽默的授课风格，严谨扎实的专业功底，随手拈来的大量亲身经历的商业实战案例，以及对待学生的耐心和热情，让我作为一名外国学生能有效克服语言和生活环境的挑战，比

较顺利地进入数据挖掘的职业发展道路。

感谢回国后给我第一份专业工作机会的前 CCG 集团（Communication Central Group）商业智能应用事业部总经理 Justin Jencks。中国通 Justin 在我们一起共事的那段日子里，果敢放手让我尝试多个跨行业的探索性商业应用项目，给了我许多宝贵的机会，使我迅速熟悉本土市场，积累了不同行业的实战案例，这些对我的专业成长非常重要。

感谢 4 年前给我机会，让我得以从传统行业进入互联网行业的阿里巴巴集团 ITBU 事业部的前商业智能部门总监李红伟（菠萝）。进入互联网行业之后，我才深深懂得作为一名数据分析师，相比传统行业来说，互联网行业有太多的机会可以去尝试不同的项目，去亲历数不清的“一竿子插到底”的落地应用，去学习面对日新月异的需求和挑战。

在本书的编写过程中，得到了淘宝网的商品推荐高级算法工程师陈凡（微博地址为 hppt://weibo.com/bicloud）和阿里巴巴 B2B 的数据仓库专家蒿亮（微博地址为 http://weibo.com/airjam；E-mail：airjam.hao@gmail.com）热情而专业的帮助和支持。陈凡友情编写了本书的 3.11 节，蒿亮友情编写了本书的 1.4.1 节和 13.1 节。

感谢一路走来，在项目合作和交流中给我帮助和支持的各位前辈、领导、朋友和伙伴，包括：上海第一医药连锁经营有限公司总经理顾咏晟先生、新华信国际信息咨询北京有限公司副总裁欧万德先生（Alvin）、上海联都集团的创始人冯铁军先生、上海通方管理咨询有限公司总经理李步峰女士和总监张国安先生、鼎和保险公司的张霖霏先生、盛大文学的数据分析经理张仙鹤先生、途牛网高级运营专家焦延伍先生，以及来自阿里巴巴的数据分析团队的领导和伙伴（资深总监车品觉先生、高级专家范国栋先生、资深经理张高峰先生、数据分析专家樊宁先生、资深数据分析师曹俊杰先生、数据分析师宫尚宝先生，等等，尤其要感谢阿里巴巴数据委员会会长车品觉老师在百忙中热情地为本书作推荐序，并在序言里为广大读者分享了数据分析师当前面临的最新机遇和挑战），以及这个仓促列出的名单之外的更多前辈、领导、朋友和伙伴。

感谢我的父母、姐姐、姐夫和外甥，他们给予了我一贯的支持和鼓励。

我将把深深的感谢给予我的妻子王艳和女儿露璐。露璐虽然只是初中一年级的学生，但是在本书的写作过程中，她多次主动放弃外出玩耍，帮我改稿，给我提建议，给我鼓励，甚至还为本书设计了一款封面，在此向露璐同学表达我衷心的感谢！而我的妻子，则将家里的一切事情打理得井井有条，使我可以将充分的时间和精力投入本书的写作中。谨以此书献给她们！

卢辉

中国 杭州

目　录

华章图书

HZ BOOKS 华章图书

第1章
什么是数据化运营

21 世纪核心的竞争就是数据的竞争，谁拥有数据，谁就拥有未来。

——马云

HZ BOOKS 华章图书

数据化运营是当前企业管理和企业战略里非常热门的一个词汇。其实施的前提条件包括企业级海量数据存储的实现、精细化运营的需求（与传统的粗放型运营相对比）、数据分析和数据挖掘技术的有效应用等，并且还要得到企业决策层和管理层的支持及推动。

数据化运营是现代企业从粗放经营向精细化管理发展的必然要求，是大数据时代企业保持市场核心竞争力的必要手段，要进行数据化运营，必须要企业全员的参与和配合。本书讨论的数据化运营主要是指互联网行业的数据化运营，所以，**除非特别申明，本书所有的“数据化运营”专指互联网数据化运营，尽管本书涉及的分析挖掘技术同样也适用于互联网行业之外的其他行业。**

数据化运营来源于现代营销管理，但是在“营销”之外有着更广的含义。

1.1 现代营销理论的发展历程

1.1.1 从 4P 到 4C

以 4P 为代表的现代营销理论可以追溯到 1960 年出版的（《基础营销》英文书名为 *Basic Marketing*）一书，该理论是由作者杰罗姆·麦卡锡（E.Jerome McCarthy）在该书中提出的。到了 1967 年，“现代营销学之父”菲利普·科特勒（Philip Kotler）在其代表作《营销管理》（*Marketing Management: Application, Planning, Implementation and Control*）第 1 版里进一步确认了以 4P 为核心的营销组合方法论。随后，该理论风靡世界，成为近半个世纪的现代营销核心思想，影响并左右了当时无数的企业营销战略。

4P 指的是 Product（产品）、Price（价格）、Place（渠道）和 Promotion（促销），如图 1-1 所示。4P 的内容简要概括如下。

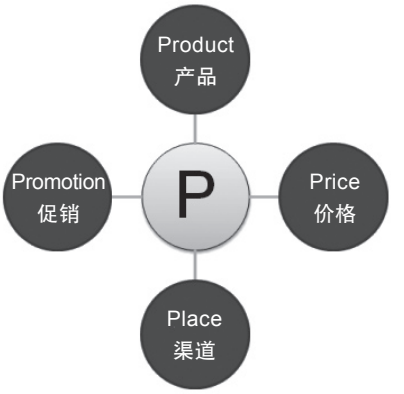


图 1-1　4P 理论结构图

- ❑ Product：表示注重产品功能，强调独特卖点。
- ❑ Price：指根据不同的市场定位，制定不同的价格策略。
- ❑ Place：指要注重分销商的培养和销售网络的建设。
- ❑ Promotion：指企业可以通过改变销售行为来刺激消费者，以短期的行为（如让利、买一送一、调动营销现场气氛等）促成消费的增长，吸引其他品牌的消费者前来消费，或者促使老主顾提前来消费，从而达到销售增长的目的。

4P 理论的核心是 Product（产品）。因此，以 4P 理论为核心营销思想的企业营销战略又可以简称为“**以产品为中心**”的营销战略。

随着时代的发展，商品逐渐丰富起来，市场竞争也日益激烈，尤其进入 21 世纪后，消费者已成为商业世界的核心。在当今这个充满个性化的商业时代，传统的 4P 营销组合已经无法适应时代发展的需求，营销界开始研究新的营销理论和营销要素。其中，最具代表性的理论就是 4 C 理论，这里的 4C 包括 Consumer（消费者）、Cost（成本）、Convenience（方便性）和 Communication（沟通交流），如图 1-2 所示，4C 的内容简要概括如下：

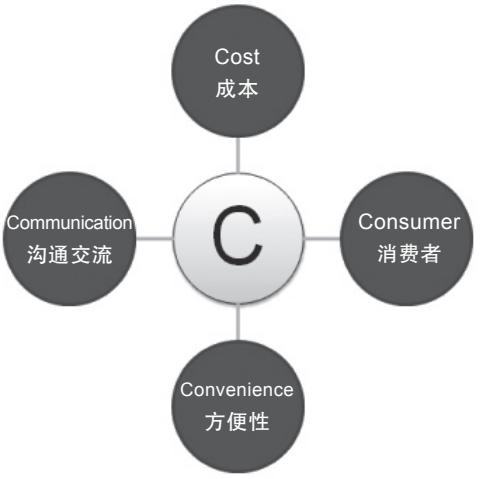


图 1-2　4C 理论结构图

- ❑ 消费者的需求与愿望（Customer’s Needs and Wants）。
- ❑ 消费者得到满足的成本（Cost and Value to Satisfy Consumer’s Needs and Wants）。
- ❑ 用户购买的方便性（Convenience to Buy）。
- ❑ 与用户的沟通交流（Communication with Consumer）。

4 C 理论的核心是 Consumer 消费者。因此，以 4C 理论为核心营销思想的企业营销战略又可以简称为“**以消费者为中心**”的营销战略。

1.1.2　从 4C 到 3P3C

4 C 理论虽然成功找到了从“以产品为中心”转化为“以消费者为中心”的思路和要素，但是随着社会的进步，科技的发展，大数据时代的来临，4 C 理论再次落后于时代发展的需要。大数据时代，日益白热化的市场竞争、越来越严苛的营销预算、海量的数据堆积和存储等，迫使现代企业不得不寻找更合适、更可控、更可量化、更可预测的营销思路和方法论。于是在基本思路上融合了 4P 理论和 4C 理论的 nPnC 形式的理论出现了。

具体到典型的互联网行业，虽然学术界对于到底是几个 P 和几个 C 仍存在着争议，没有定论，但是这并不妨碍企业积极探索并付诸实践应用，本书姑且以 3P3C 为例，如图 1-3 所示，概述互联网行业运营的典型理论探索。

图 1-3　3P3C 理论结构图

在 3P3C 理论中，数据化运营 6 要素的内容如下。

❑ Probability（概率）：营销、运营活动以概率为核心，追求精细化和精准率。

❑ Product（产品）：注重产品功能，强调产品卖点。

❑ Prospects（消费者，目标用户）。

❑ Creative（创意，包括文案、活动等）。

❑ Channel（渠道）。

❑ Cost/Price（成本 / 价格）。

而在这其中，以数据分析挖掘所支撑的目标响应概率（Probability）是核心，在此基础上将会围绕产品功能优化、目标用户细分、活动（文案）创意、渠道优化、成本的调整等重要环节和要素，共同促使数据化运营持续完善，直至成功。

需要指出的是，这里的目标响应概率（Probability）不应狭义理解为仅仅是预测响应模型之类的响应概率，它有更宽泛的含义，既可以从宏观上来理解，又可以从微观上来诠释。从宏观上来理解，概率可以是特定消费群体整体上的概率或可能性。比如，我们常见的通过卡方检验发现某个特定类别群体在某个消费行为指标上具有的显著性特征，这种显著性特征可以帮助我们进行目标市场的选择、寻找具有相似特征的潜在目标用户，制定相应的细分营销措施和运营方案等，这种方法可以有效提升运营的效率和效果；从微观上来理解，概率可以是具体到某个特定消费者的“预期响应概率”，比如我们常见的通过逻辑回归算法搭建一个预测响应模型，得到每个用户的预计响应概率，然后，根据运营计划和预算，抽取响应概率分数的消费者，进

行有针对性的运营活动等，这种方法也可以有效提升运营的效率和效果。

宏观的概率更加有效，还是微观的概率更加有效，这需要结合项目的资源计划、业务背景、项目目的等多种因素来权衡，不可一概而论。虽然微观的概率常常更为精细、更加准确，但是在实践应用中，宏观的群体性概率也可以有效提升运营效果，也是属于数据化运营的思路。所以在实践过程中如何选择，要根据具体的业务场景和具体的数据分析解决方案来决定。更多延伸性的分析探讨，将在后面章节的具体项目类型分析、技术分享中详细介绍。

上述 3P3C 理论有效锁定了影响运营效果的主要因素、来源，可以帮助运营人员、管理人员、数据分析人员快速区分实践中的思考维度和着力点，提高思考效率和分析效率。

1.2　数据化运营的主要内容

虽然目前企业界和学术界对于“数据化运营”的定义没有达成共识，但这并不妨碍“数据化运营”思想和实践在当今企业界尤其是互联网行业如火如荼地展开。阿里巴巴集团早在 2010 年就已经在全集团范围内正式提出了“数据化运营”的战略方针并逐步实施数据化运营，腾讯公司也在“2012 年腾讯智慧上海主题日”高调宣布“大数据化运营的黄金时期已经到来，如何整合这些数据成为未来的关键任务”。

综合业界尤其是互联网行业的数据化运营实践来看，尽管各行业对“数据化运营”的定义有所区别，但其基本要素和核心是一致的，那就是“**以企业级海量数据的存储和分析挖掘应用为核心支持的，企业全员参与的，以精准、细分和精细化为特点的企业运营制度和战略**”。换种思路，可以将其浅层次地理解为，在企业常规运营的基础上革命性地增添数据分析和数据挖掘的精准支持。这是从宏观意义上对数据化运营的理解，其中会涉及企业各部门，以及数据在企业中所有部门的应用。但是必须指出，本书所要分享的实战项目涉及的数据化运营，主要落实在微观意义的数据化运营上，即主要针对运营、销售、客服等部门的互联网运营的数据分析、挖掘和支持上。

注意：这种宏观和微观上的区别在本质上对于数据化运营的核心没有影响，只是在本书的技术和案例分享中更多聚焦于运营部门、销售部门、客服部门而已，特此说明。

针对互联网运营部门的数据化运营，具体包括**“网站流量监控分析、目标用户行为研究、网站日常更新内容编辑、网络营销策划推广”等，并且，这些内容是在以企业级海量数据的存储、分析、挖掘和应用为核心技术支持的基础上，通过可量化、可细分、可预测等一系列精细化的方式来进行的**。

数据化运营，首先是要有企业全员参与意识，要达成这种全员的数据参与意识比单纯地

执行数据挖掘技术显然是要困难得多，也重要得多的。只有在达成企业全员的自觉参与意识后，才可能将其转化为企业全体员工的自觉行动，才可能真正落实到运营的具体工作中。举例来说，阿里巴巴集团正在实施的数据化运营，就要求所有部门所有岗位的员工都要贯彻此战略：从产品开发人员到用户体验部门，到产品运营团队，到客户服务部门，到销售团队和支持团队，每个人每个岗位都能真正从数据应用、数据管理和数据发现的高度经营各自的本职工作，也就类似于各个岗位的员工，都在各自的工作中自觉利用或简单或复杂的数据分析工具，进行大大小小的数据分析挖掘，这才是真正的数据化运营的场面，才是真正的从数据中发现信息财富并直接助力于企业的全方位提升。也只有这样，产品开发人员所提出的新概念才不是拍脑袋拍出来的，而是来自于用户反馈数据的提炼；产品运营人员也不再仅仅是每天被动地抄报运营的 KPI 指标，通过数据意识的培养，他们将在运营前的准备，运营中的把握，运营后的反馈、修正、提升上有充分的预见性和掌控力；客户服务部门不仅仅满足于为客户提供满意的服务，他们学会了从服务中有意识地发现有代表性的、有新概念价值的客户新需求；销售部门则不再只是具有吃苦耐劳的精神，他们可通过数据分析挖掘模型的实施来实现有的放矢、精准营销的销售效益最大化。而企业的数据挖掘团队也不再仅仅局限于单纯的数据挖掘技术工作及项目工作，而是肩负在企业全员中推广普及数据意识、数据运用技巧的责任，这种责任对于企业而言比单纯的一两个数据挖掘项目更有价值，更能体现一个数据挖掘团队或者一个数据挖掘职业人的水准、眼界以及胸怀，俗话说“只有能发动人民战争，才是真正的英雄”，所以只有让企业全员都参与并支持你的数据挖掘分析工作，才能够真正有效地挖掘企业的数据资源。现代企业的领导者，应该有这种远见和智慧，明白全员的数据挖掘才是企业最有价值的数据挖掘，全员的数据化运营才是现代企业的竞争新核心。

数据化运营，其次是一种常态化的制度和流程，包括企业各个岗位和工种的数据收集和数据分析应用的框架和制度等。从员工日常工作中所使用的数据结构和层次，就基本上可以判断出企业的数据应用水准和效率。在传统行业的大多数企业里，绝大多数员工在其工作中很少（甚至基本不）分析使用业务数据支持自己的工作效率，但是在互联网行业，对数据的重视和深度应用使得该行业数据化运营的能力和水平远远超过传统行业的应用水平。

数据化运营更是来自企业决策者、高层管理者的直接倡导和实质性的持续推动。由于数据化运营一方面涉及企业全员的参与，另一方面涉及企业海量数据的战略性开发和应用，同时又是真正跨多部门、多技术、多专业的整合性流程，所有这些挑战都是企业内部任何单个部门所无法独立承担的。只有来自企业决策层的直接倡导和实质性的持续推动，才可以在企业建立、推广、实施、完善真正的全员参与、跨部门跨专业、具有战略竞争意义的数据化运营。所以，我们不难发现，阿里巴巴集团也好，腾讯也罢，这些互联网行业的巨人，之所以能在大数据时代如火如荼地进行企业数据化运营，自始至终都离不开企业决策层的直接倡导与持续推动，其在各种场合中对数据的重要性、对数据化运营的核心竞争力价值的强调和分

享，都证明了决策层是推动数据化运营的关键所在。2012 年 7 月 10 日，阿里巴巴集团宣布设立“首席数据官”岗位（Chief Data Officer），阿里巴巴 B2B 公司的 CEO 陆兆禧出任此职位，并会向集团 CEO 马云直接汇报。陆兆禧将主要负责全面推进阿里巴巴集团成为“数据分享平台”的战略，其主要职责是规划和实施未来数据战略，推进支持集团各事业群的数据业务发展。“将阿里巴巴集团变成一家真正意义上的数据公司”目前已经是阿里巴巴集团的战略共识，阿里巴巴集团旗下的支付宝、淘宝、阿里金融、B2B 的数据都会成为这个巨大的数据分享平台的一部分。而**这个战略的核心就是如何挖掘、分析和运用这些数据，并和全社会分享**。

1.3　为什么要数据化运营

数据化运营首先是现代企业竞争白热化、商业环境变成以消费者为主的“买方市场”等一系列竞争因素所呼唤的管理革命和技术革命。中国有句古语“穷则思变”，当传统的营销手段、运营方法已经被同行普遍采用，当常规的营销技术、运营方法已经很难明显提升企业的运营效率时，竞争必然呼唤革命性的改变去设法提升企业的运营效率，从而提升企业的市场竞争力。时势造英雄，生逢其时的“数据化运营”恰如及时雨，登上了大数据时代企业运营的大舞台，在互联网运营的舞台上尤其光彩夺目。

其次，数据化运营是飞速发展的数据挖掘技术、数据存储技术等诸多先进数据技术直接推动的结果。数据技术的飞速发展，使得大数据的存储、分析挖掘变得成熟、可靠，成熟的挖掘算法和技术给了现代企业足够的底气去尝试海量数据的分析、挖掘、提炼、应用。有了数据分析、数据挖掘的强有力支持，企业的运营不再盲目，可以真正做到运营流程自始至终都心中有数、有的放矢。比如，在传统行业的市场营销活动中，有一个无解又无奈的问题：“我知道广告费浪费了一半，但是我不知道到底是哪一半”。这里的无奈其实反映的恰好就是传统行业粗放型营销的缺点：无法真正细分受众，无法科学监控营销各环节，无法准确预测营销效果；但是，在大数据时代的互联网行业，这种无奈已经可以有效地降低，乃至避免，原因在于通过数据挖掘分析，广告主可以精细划分出正确的目标受众，可以及时（甚至实时）监控广告投放环节的流失量，可以针对相应的环节采取优化、提升措施，可以建立预测模型准确预测广告效果。

数据化运营更是互联网企业得天独厚的“神器”。互联网行业与生俱来的特点就是大数据，而信息时代最大的财富也正是海量的大数据。阿里巴巴集团董事局主席兼首席行政官马云曾经多次宣称，阿里巴巴集团最大的财富和今后核心竞争力的源泉，正是阿里巴巴集团（包括淘宝、支付宝、阿里巴巴等所属企业）已经产生的和今后继续积累的海量的买卖双方的交易数据、支付数据、互动数据、行为数据等。2010 年 3 月 31 日，淘宝网在上海正式宣

布向全球开放数据，未来电子商务的核心竞争优势来源于对数据的解读能力，以及配合数据变化的快速反应能力，而开放淘宝数据正是有效帮助企业建立数据的应用能力。2010 年 5 月 14 日阿里巴巴集团在深圳举行的 2010 年全球股东大会上，马云进一步指出“21 世纪核心的竞争就是数据的竞争”，“谁拥有数据，谁就拥有未来”。企业决策者对数据价值的高度认同，必然会首先落实在自身的企业运营实践中，这也是“因地制宜”战略思想在互联网时代的最新体现，我们也可以理解成“近水楼台先得月”在互联网时代的最新诠释。

1.4 数据化运营的必要条件

虽然从上面的分析可以看出，数据化运营有如此多的优越性，但并不是每个企业都可以采取这种新战略和新管理制度，也不是每个企业都可以从中受益。个中原因在于成功的数据化运营必须依赖几个重要的前提条件。

1.4.1 企业级海量数据存储的实现[⊖]

21 世纪核心的竞争就是数据的竞争，2012 年 3 月 29 日，美国奥巴马政府正式宣布了“大数据的研究和发展计划”(Big Data Research and Development Initiative)，该计划旨在**通过提高我们从大型复杂数据集中提取知识和观点的能力，承诺帮助加快在科学和工程中探索发现的步伐，加强国家安全**。从国家到企业，数据就是生产力。但是，具体到某一个企业，海量数据的存储是必须要面对的第一个挑战。数据存储技术的飞速发展，需要企业与时俱进。根据预测到 2020 年，全球以电子形式存储的数据量将达到 35ZB，是 2009 年全球存储量的 40 倍。而在 2010 年年底，根据 IDC 的统计，全球数据量已经达到了 1 200 000PB 或 1.2ZB。如果将这些数据都刻录在 DVD 上，那么光把这些 DVD 盘片堆叠起来就可以从地球到月球打一个来回（单程约 24 万英里，即 386 242.56 千米）。海量的数据推动了数据存储技术的不断发展与飞跃。

我们一起来回顾一下数据存储技术的发展历程：

1951 年：Univac 系统使用磁带和穿孔卡片作为数据存储。

1956 年：IBM 公司在其 Model 305 RAMAC 中第一次引入了磁盘驱动器。

1961 年：美国通用电气公司（General Electric）的 Charles Bachman 开发了第一个数据库管理系统——IDS。

⊖ 本节内容由阿里巴巴B2B的数据仓库专家蒿亮编写，蒿亮的微博地址为http://weibo.com/airjam，电子邮件为airjam.hao@gmail.com。

1969 年：E.F. Codd 发明了关系数据库。

1973 年：由 John J.Cullinane 领导的 Cullinane 公司开发了 IDMS——一个针对 IBM 主机的基于网络模型的数据库。

1976 年：Honeywell 公司推出了 Multics Relational Data Store——第一个商用关系数据库产品。

1979 年：Oracle 公司引入了第一个商用 SQL 关系数据库管理系统。

1983 年：IBM 推出了 DB2 数据库产品。

1985 年：为 Procter & Gamble 系统设计的第一个商务智能系统产生。

1991 年：W.H. BillInmon 发表了文章《构建数据仓库》。

2012 年：最新的存储技术为分布式数据仓库、海量数据存储技术和流计算的实时数据仓库技术。

回首中国企业的数据存储之路，国内的数据存储技术的发展经历了将近 30 年，而真正的飞速发展则是最近 10 年。

国内的数据存储的先驱是国有银行，在 21 世纪初，四大国有银行的全国数据中心项目（将分布在全国各个省行和直属一级分行的数据集中到数据中心）拉开了数据技术飞速发展的帷幕。

以发展最具代表性的中国工商银行为例，中国工商银行从 2001 年开始启动数据集中项目，刚开始考虑集中中国北部的数据到北京，中国南部的数据到上海，最终在 2004 年将全部数据集中到了上海，而北京则作为灾备中心，海外数据中心则安置在深圳。中国工商银行的数据量在当时是全中国最大的，大约每天的数据量都在 TB 级别。由于银行业存在一定的特殊性（性能要求低于安全和稳定要求），又因为当时业内可选的技术不多，因此中国工商银行选择了大型机 +DB2 的技术方案，实际上就是以关系型数据库作为数据存储的核心。

在 3 年的数据集中和后续 5 年基于主题模型（NCR 金融模型）的数据仓库建设期间，中国工商银行无论在硬件网络和软件人力上都投入了巨大的资源，其数据仓库也终于成为中国第一个真正意义上的企业级数据中心和数据仓库。

其他银行和证券保险，甚至电信行业以及房地产行业的数据仓库建设，基本上也都是采用与工商银行相似的思路和做法在进行。

不过，随着时间的推移，数据量变得越来越大，硬件的更新换代也越来越快，于是，这类数据仓库逐渐显现出了问题，主要表现如下：

- ❑ 少数几台大型机已经无法满足日益增加的日终计算任务的执行需求，导致很多数据结果为 T-2（当天数据要延后 2 天才完成），甚至是 T-3（当天数据要延后 3 天才完成）。
- ❑ 硬件升级和存储升级的成本非常昂贵，维护、系统开发以及数据开发的人力资源开支也逐年加大。
- ❑ 由于全国金融发展的进程差异很大，数据需求各不一样，加上成本等原因，不得不将一些数据计算任务下放到各个一级分行或者省分行进行，数据中心不堪重负。

随着互联网行业的逐渐蓬勃兴盛，占领数据存储技术领域巅峰的行业也从原有的国有银行企业转移到了阿里巴巴、腾讯、盛大、百度这样的新兴互联网企业。以阿里巴巴为例，阿里巴巴数据仓库也是经历了坎坷的发展历程，在多次重建后才最终站在了中国甚至世界的顶峰。

最开始的阿里巴巴互联网数据仓库建设，几乎就是中国工商银行的缩小版，互联网的数据从业人员几乎全部来自国内各大银行或电信行业，或者来自国外类似微软、yahoo 这样的传统 IT 企业。

随着分布式技术的逐渐成熟和工业化，互联网数据仓库迎来了飞速发展的春天。现在，抛弃大型机 + 关系型数据库的模型，采用分布式的服务器集群 + 分布式存储的海量存储器，无论是从硬件成本、软件成本还是从硬件升级、日常维护上来讲，都是一次飞跃。更重要的是，解决了困扰数据仓库发展的一个非常重要的问题，即计算能力不足的问题，当 100~200 台网络服务器一起工作的时候，无论是什么样的大型机，都已经无法与之比拟了。

拿现在阿里云（阿里巴巴集团数据中心服务提供者）来讲，近 1000 台网络服务器分布式并行，支持着每日淘宝、支付宝、阿里巴巴三大子公司超过 PB 级别的数据量，随着技术的日益成熟和硬件成本的逐渐降低，未来的数据仓库将是以流计算为主的实时数据仓库和分布式计算为主流的准实时数据仓库。

1.4.2 精细化运营的需求

大数据时代的互联网行业所面临的竞争压力甚至已超过了传统行业。主要原因在于互联网行业的技术真正体现了日新月异、飞速发展的特点。以中国互联网行业的发展为例，作为第一代互联网企业的代表，新浪、搜狐、雅虎等门户网站的 Web 1.0 模式（传统媒体的电子化）从产生到被以 Google、百度等搜索引擎企业的 Web 2.0 模式（制造者与使用者的合一）所超越，前后不过 10 年左右的时间，而目前 Web 2.0 模式已经逐渐有被以微博为代表的 Web 3.0 模式（SNS 模式）超越的趋势。

互联网行业近乎颠覆性模式的进化演绎、技术的更新换代，既为互联网企业提供了机

遇，又带给其沉重的竞争压力与生存的挑战。面对这种日新月异的竞争格局，互联网企业必须寻找比传统的粗放型运营更加有效的精细化运营制度和思路，以提升企业的效益和效率，而数据化运营就是精细化运营，它强调的是更细分、更准确、更个性化。没有精细化运营的需求，就不需要数据化运营；只有数据化运营，才可以满足精细化的效益提升。

1.4.3　数据分析和数据挖掘技术的有效应用

数据分析和数据挖掘技术的有效应用是数据化运营的基础和技术保障，没有这个基础保障，数据化运营就是空话，就是无本之水，无缘之木。

这里的有效应用包括以下两层含义。

一是企业必须拥有一支能够胜任数据分析和数据挖掘工作的团队和一群出色的数据分析师。一名出色的数据分析师必须是多面手，他不仅要具备统计技能（能熟练使用统计技术和统计工具进行分析挖掘）、数据仓库知识（比如熟悉主流数据库基本技术，可以自助取数，可以有效与数据仓库团队沟通）、数据挖掘技能（熟练掌握主流数据挖掘技术和工具），更重要的是他还要具有针对具体业务的理解能力和快速学习能力，并且要善于与业务方沟通、交流。数据分析挖掘绝不是数据分析师或团队的闭门造车，要想让项目成功应用，必须要自始至终与业务团队并肩作战，从这点来看，业务理解力和沟通交流能力的重要性甚至要远远超过技术层面的能力（诸如统计技能、挖掘技能、数据仓库的技能）。从之前的分析可以看出，一名出色的数据分析师是需要时间、项目经验去磨砺去锻炼成长的，而作为企业来说，如何选择、培养、配备这样一支合格的分析师队伍，才是数据化运营的基础保障。

二是企业的数据化运营只有在分析团队与业务团队协同配合下才可能做出成绩，取得效果。分析团队做出的分析方案、数据模型，必须要在业务应用中得到检验，这不仅要求业务方主观的参与和支持，也要求业务方的团队和员工同样要具有相应的数据化运营能力和水平，运营团队的人员需要具备哪些与数据化运营相关的技能呢？这个问题我们将在第 4 章阐述。

无论是数据分析团队的专业能力，运营团队的专业能力，还是其他业务团队的专业能力，所体现的都是互联网企业的人才价值，这个人才价值与数据的价值一样，都是属于互联网行业的核心竞争力，正如阿里巴巴集团董事会主席兼 CEO 马云在多个场合强调的那样，“人才和数据是阿里巴巴集团最大的财富和最强大的核心竞争力”。

1.4.4　企业决策层的倡导与持续支持

在关乎企业数据化运营的诸多必要条件里，最核心且最具决定性的条件就是来自企业决策层的倡导和持续支持。

在传统行业的现代企业里，也有很多采用了先进的数据分析技术来支持企业运营的，支持企业的营销、客服、产品开发等工作。但是总的来说，这些数据挖掘应用效果参差不齐，或者说应该体现的业务贡献价值在很多情况下并没有真正体现出来，总体的应用还是停留在项目管理的层面，缺乏全员的参与与真正跨部门的战略协调配合。这种项目层面的管理，存在的不足如下：

首先，由于参与分析挖掘的团队与提出分析需求的业务团队分属不同的职能部门，缺乏高层实质性的协调与管理，常会出现分析建模工作与真正的业务需求配合不紧密，各打各的锣，各唱各的歌。由于各部门和员工 KPI 考核的内容不同，数据分析团队完成的分析方案、模型、建议、报告很多时候只是纸上谈兵，无法转化成业务应用的实际操作。举个简单的例子，销售部门的年度 KPI 考核是销售额和付费人数，那么为了这个年度 KPI 考核，销售部门必然把工作的重心放在扩大销售额，扩大付费人数，维护续费人数，降低流失率等关键指标上，他们自然希望数据分析部门围绕年度（短期的）KPI 目标提供分析和模型支持，提高销售部门的业绩和效率。但是数据分析部门的年度 KPI 考核可能跟年度销售额和付费人数没有关系，而跟通过数据分析、建模，完善产品开发与优化，完善销售部门的业务流程与资源配置等相关。很显然，这里数据分析团队的 KPI 考核是着眼于企业长期发展的，这跟销售部门短期的以销售额为重点的考核在很大程度上是有冲突的。在这种情况下，怎么指望两者的数据化运营能落地开花呢？

其次，因为处于项目层面的管理，所以数据分析挖掘的规划也就只能局限在特定业务部门的范围内，缺乏真正符合企业发展方向的数据分析挖掘规划。俗话说得好站得高，方能看得远，起点低，视野浅，自然约束了数据分析的有效发挥。

无论是组织架构的缺陷，还是战略规划的缺失，其本质都能表现出缺乏来自企业决策层的倡导和持续支持。只有得到企业决策层的倡导和支持，上述组织管理方面的缺陷和战略规划的缺失才可以有效避免。如前所述，2012 年 7 月 10 日阿里巴巴集团宣布设置首席数据官的岗位，并将其作为企业的核心管理岗位之一，其目的就是进一步夯实企业的数据战略，规划和实施企业整体的数据化运营能力和水平，使之真正成为阿里巴巴集团未来的核心竞争力。

1.5 数据化运营的新现象与新发展

时代在发展，技术在进步，企业的数据化运营也在不断增添新的内容、不断响应新的需求。目前，从世界范围来看，数据化运营至少在下列几个方面已经出现了实质性的新发展，这些新发展扩大了数据化运营的应用场景、扩充了数据化运营的发展思路、也给当前（以及未来）数据化运营的参与者提供了更多的发展方向的选择。这些新发展包括的内容如下：

- 数据产品作为商业智能的一个单独的发展方向和专业领域，在国内外的商业智能和数据分析行业里已经成为共识，并且正在企业的数据化运营实践中发挥着越来越大的作用。数据产品是指通过数据分析和数据模型的应用而开发出来的，提供给用户使用的一系列的帮助用户更好理解和使用数据的工具产品，这些工具产品的使用让用户在某些特定场景或面对某些特定的数据时，可以独立进行分析和展示结果，而不需要依赖数据分析师的帮助。虽然在多年以前，类似的数据产品已被开发并投入了应用，但是在数据分析行业世界范围内达成共识，并作为商业智能的一个独立发展方向和专业领域，还只是近一两年的事情。淘宝网上的卖家所使用的“量子恒道”就是一个非常不错的数据产品，通过使用量子恒道，淘宝卖家可以自己随时监控店铺的流量来源、买家逗留的时间、买家区域、浏览时间、各页面的流量大小、各产品的成交转化率等一系列跟店铺的实时基础数据相关的数据分析和报告，从而有效帮助卖家制定和完善相应的经营方向和经验策略。数据产品作为数据分析和商业智能里一个专门的领域得以确立和发展，其实是跟数据化运营的全民参与的特征相辅相成的。数据产品帮助企业全员更好、更有效地利用数据，而数据化运营的全民参与也呼唤更多更好的数据产品，企业成功的数据化运营建设一定会同时产生一大批深受用户欢迎和信赖的数据产品。

- 数据 PD 作为数据分析和商业智能的一个细分的职业岗位，已经在越来越多的大规模数据化运营的企业得以专门设立并日益强化。与上述的数据产品相配套的，就是数据 PD 作为一个专门的细分的职业岗位和专业方向，正逐渐为广大的数据化运营的企业所熟悉并采用。PD（Product Designer）是产品设计师的英文缩写，而数据 PD，顾名思义就是数据产品的产品设计师。数据 PD 作为数据分析和商业智能中一个新的职业方向和职业岗位，需要从业者兼具数据分析师和产品设计师双重的专业知识、专业背景、技能和素质，有志从事数据 PD 工作的新人，可以抓住这个崭新的职业，几乎还是一张白纸的无限空间，快速成长，迅速成才。

- 泛 BI 的概念在大规模数据化运营的企业里正在越来越深入人心。泛 BI 其实就是逐渐淡化数据分析师团队作为企业数据分析应用的唯一专业队伍的印象，让更多的业务部门也逐渐参与数据分析和数据探索，让更多业务部门的员工也逐渐掌握数据分析的技能和意识。泛 BI 其实也是数据化运营的全民参与的特征所要求的，是更高一级的数据化运营的全民参与。在这个阶段，业务部门的员工不仅要积极参与数据分析和模型的具体应用实践，更要求他们能自主自发地进行一些力所能及的数据分析和数据探索。泛 BI 概念的逐渐深入普及，向数据分析师和数据分析团队提出了新的要求，数据分析师和数据分析团队承担了向业务部门及其员工指导、传授有关数据分析和数据探索的能力培养的工作，这是一种授人以渔的崇高行为，值得数据分析师为之奉献。

1.6 关于互联网和电子商务的最新数据

2012 年 12 月 3 日，阿里巴巴集团在杭州宣布，截至 2012 年 11 月 30 日 21:50，其旗下淘宝和天猫的交易额本年度突破 10 000 亿元。为支撑这巨大规模业务量的直接与间接的就业人员已经超过 1000 万人。

根据国家统计局的数据显示，2011 年全国各省社会消费品零售总额为 18.39 万亿元，10 000 亿元相当于其总量的 5.4%，而根据国家统计局公布的 2011 年全国各省社会消费品零售总额排行，可以排列第 5 位，仅次于广东、山东、江苏和浙江。电子商务已经成为一个庞大的新经济主体，并在未来相当长的时间里依然会高速发展，这意味着过去的不可能已经成为现实，而这才是刚刚开始。

阿里巴巴集团董事局主席马云表示：“我们很幸运，能够适逢互联网这个时代，一起见证并参与互联网及电子商务给我们社会带来的一次次惊喜和改变。10 000 亿只是刚刚开始，我们正在步入 10 万亿的时代，未来电子商务在中国，必将产生 1000 万个企业，具备服务全球 10 亿消费者的能力。”

HZ BOOKS 华章图书

第2章
数据挖掘概述

数据挖掘是指从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程，这些信息的表现形式为规则、概念、规律及模式等。

在第 1 章中介绍了什么是数据化运营，为什么要实现数据化运营，以及数据化运营的主要内容和必要条件。我们知道数据分析和数据挖掘技术是支撑企业数据化运营的基础和技术保障，没有有效的数据挖掘支持，企业的数据化运营就是无源之水，无本之木。

本章将为读者简单回顾一下数据挖掘作为一门学科的发展历史，并具体探讨统计分析与数据挖掘的主要区别，同时，将力求用简单、通俗、明了的文字把目前主流的、成熟的、在数据化运营中常用的统计分析和数据挖掘的算法、原理以及主要的应用场景做出总结和分类。

最后，针对互联网数据化运营中数据挖掘应用的特点进行梳理和总结。

2.1 数据挖掘的发展历史

数据挖掘起始于 20 世纪下半叶，是在当时多个学科发展的基础上发展起来的。随着数据库技术的发展应用，数据的积累不断膨胀，导致简单的查询和统计已经无法满足企业的商业需求，急需一些革命性的技术去挖掘数据背后的信息。同时，这期间计算机领域的人工智能（Artificial Intelligence）也取得了巨大进展，进入了机器学习的阶段。因此，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，并且尝试挖掘数据背后的信息。这两者的结合促生了一门新的学科，即数据库中的知识发现（Knowledge Discovery in Databases，KDD）。1989 年 8 月召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现了知识发现（KDD）这个术语，到目前为止，KDD 的重点已经从发现方法转向了实践应用。

而数据挖掘（Data Mining）则是知识发现（KDD）的核心部分，它指的是从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程，这些信息的表现形式为：规则、概念、规律及模式等。进入 21 世纪，数据挖掘已经成为一门比较成熟的交叉学科，并且数据挖掘技术也伴随着信息技术的发展日益成熟起来。

总体来说，数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的理论和技术，是 21 世纪初期对人类产生重大影响的十大新兴技术之一。

2.2 统计分析与数据挖掘的主要区别

统计分析与数据挖掘有什么区别呢？从实践应用和商业实战的角度来看，这个问题并没有很大的意义，正如“不管白猫还是黑猫，抓住老鼠才是好猫”一样，在企业的商业实战中，数据分析师分析问题、解决问题时，首先考虑的是思路，其次才会对与思路匹配的分析挖掘技术进行筛选，而不是先考虑到底是用统计技术还是用数据挖掘技术来解决这个问题。

从两者的理论来源来看，它们在很多情况下都是同根同源的。比如，在属于典型的数据挖掘技术的决策树里，CART、CHAID 等理论和方法都是基于统计理论所发展和延伸的；并且数据挖掘中的技术有相当比例是用统计学中的多变量分析来支撑的。

相对于传统的统计分析技术，数据挖掘有如下一些特点：

- ❑ 数据挖掘特别擅长于处理大数据，尤其是几十万行、几百万行，甚至更多更大的数据。
- ❑ 数据挖掘在实践应用中一般都会借助数据挖掘工具，而这些挖掘工具的使用，很多时候并不需要特别专业的统计背景作为必要条件。不过，需要强调的是基本的统计知识和技能是必需的。
- ❑ 在信息化时代，数据分析应用的趋势是从大型数据库中抓取数据，并通过专业软件进行分析，所以数据挖掘工具的应用更加符合企业实践和实战的需要。
- ❑ 从操作者来看，数据挖掘技术更多是企业的数据分析师、业务分析师在使用，而不是统计学家用于检测。

更主流的观点普遍认为，数据挖掘是统计分析技术的延伸和发展，如果一定要加以区分，它们又有哪些区别呢？数据挖掘在如下几个方面与统计分析形成了比较明显的差异：

- ❑ 统计分析的基础之一就是概率论，在对数据进行统计分析时，分析人员常常需要对数据分布和变量间的关系做假设，确定用什么概率函数来描述变量间的关系，以及如何检验参数的统计显著性；但是，在数据挖掘的应用中，分析人员不需要对数据分布做任何假设，数据挖掘中的算法会自动寻找变量间的关系。因此，相对于海量、杂乱的数据，数据挖掘技术有明显的应用优势。
- ❑ 统计分析在预测中的应用常表现为一个或一组函数关系式，而数据挖掘在预测应用中的重点在于预测的结果，很多时候并不会从结果中产生明确的函数关系式，有时候甚至不知道到底是哪些变量在起作用，又是如何起作用的。最典型的例子就是“神经网络”挖掘技术，它里面的隐蔽层就是一个“黑箱”，没有人能在所有的情况下读懂里面的非线性函数是如何对自变量进行组合的。在实践应用中，这种情况常会让习惯统计分析公式的分析师或者业务人员感到困惑，这也确实影响了模型在实践应用中的可理解性和可接受度。不过，如果能换种思维方式，从实战的角度考虑，只要模型能正确预测客户行为，能为精细化运营提供准确的细分人群和目标客户，业务部门、运营部门不了解模型的技术细节，又有何不可呢？
- ❑ 在实践应用中，统计分析常需要分析人员先做假设或判断，然后利用数据分析技术来验证该假设是否成立。但是，在数据挖掘中，分析人员并不需要对数据的内在关系做

任何假设或判断，而是会让挖掘工具中的算法自动去寻找数据中隐藏的关系或规律。两者的思维方式并不相同，这给数据挖掘带来了更灵活、更宽广的思路和舞台。

虽然上面详细阐述了统计分析与数据挖掘的区别，但是在企业的实践应用中，我们不应该硬性地把两者割裂开来，也无法割裂，在实践应用中，没有哪个分析师会说，“我只用数据挖掘技术来分析”，或者“我只用统计分析技术来分析”。正确的思路和方法应该是：针对具体的业务分析需求，先确定分析思路，然后根据这个分析思路去挑选和匹配合适的分析算法、分析技术，而且一个具体的分析需求一般都会有两种以上不同的思路和算法可以去探索，最后可根据验证的效果和资源匹配等一系列因素进行综合权衡，从而决定最终的思路、算法和解决方案。

鉴于实践应用中，统计分析与数据挖掘技术并不能完全被割裂开来，并且本书侧重于数据化运营的实践分享。所以在后续各章节的讨论中，将不再人为地给一个算法、技术贴上“统计分析”或“数据挖掘”的标签，后续各章节的技术分享和实战应用举例，都会本着针对不同的分析目的、项目类型来介绍主流的、有效的分析挖掘技术以及相应的特点和技巧。统计分析也罢，数据挖掘也好，只要有价值，只要在实战中有效，都会是我们所关注的，都会是我们所要分析分享的。

2.3 数据挖掘的主要成熟技术以及在数据化运营中的主要应用

2.3.1 决策树

决策树（Decision Tree）是一种非常成熟的、普遍采用的数据挖掘技术。之所以称为树，是因为其建模过程类似一棵树的成长过程，即从根部开始，到树干，到分枝，再到细枝末节的分叉，最终生长出一片片的树叶。在决策树里，所分析的数据样本先是集成为一个树根，然后经过层层分枝，最终形成若干个结点，每个结点代表一个结论。

决策树算法之所以在数据分析挖掘应用中如此流行，主要原因在于决策树的构造不需要任何领域的知识，很适合探索式的知识发掘，并且可以处理高维度的数据。在众多的数据挖掘、统计分析算法中，决策树最大的优点在于它所产生的一系列从树根到树枝（或树叶）的规则，可以很容易地被分析师和业务人员理解，而且这些典型的规则甚至不用整理（或稍加整理），就是现成的可以应用的业务优化策略和业务优化路径。另外，决策树技术对数据的分布甚至缺失非常宽容，不容易受到极值的影响。

目前，最常用的 3 种决策树算法分别是 CHAID、CART 和 ID3（包括后来的 C4.5，乃至 C5.0）。

CHAID(Chi-square Automatic Interaction Detector) 算法的历史较长，中文简称为卡方自动相互关系检测。CHAID 依据局部最优原则，利用卡方检验来选择对因变量最有影响的自变量，CHAID 应用的前提是因变量为类别型变量（Category）。

CART(Classification and Regression Tree) 算法产生于 20 世纪 80 年代中期，中文简称为分类与回归树，CART 的分割逻辑与 CHAID 相同，每一层的划分都是基于对所有自变量的检验和选择上的。但是，CART 采用的检验标准不是卡方检验，而是基尼系数（Gini）等不纯度的指标。两者最大的区别在于 CHAID 采用的是局部最优原则，即结点之间互不相干，一个结点确定了之后，下面的生长过程完全在结点内进行。而 CART 则着眼于总体优化，即先让树尽可能地生长，然后再回过头来对树进行修剪（Prune），这一点非常类似统计分析中回归算法里的反向选择（Backward Selection）。CART 所生产的决策树是二分的，每个结点只能分出两枝，并且在树的生长过程中，同一个自变量可以反复使用多次（分割），这些都是不同于 CHAID 的特点。另外，如果是自变量存在数据缺失（Missing）的情况，CART 的处理方式将会是寻找一个替代数据来代替（填充）缺失值，而 CHAID 则是把缺失数值作为单独的一类数值。

ID3（Iterative Dichotomiser）算法与 CART 是同一时期产生的，中文简称为迭代的二分器，其最大的特点在于自变量的挑选标准是：基于信息增益的度量选择具有最高信息增益的属性作为结点的分裂（分割）属性，其结果就是对分割后的结点进行分类所需的信息量最小，这也是一种划分纯度的思想。至于之后发展起来的 C4.5 可以理解为 ID3 的发展版（后继版），两者的主要区别在于 C4.5 采用信息增益率（Gain Ratio）代替了 ID3 中的信息增益度量，如此替换的主要原因是信息增益度量有个缺点，就是倾向于选择具有大量值的属性。这里给个极端的例子，对于 Member_Id 的划分，每个 Id 都是一个最纯的组，但是这样的划分没有任何实际意义。而 C4.5 所采用的信息增益率就可以较好地克服这个缺点，它在信息增益的基础上，增加了一个分裂信息（SplitInformation）对其进行规范化约束。

决策树技术在数据化运营中的主要用途体现在：作为分类、预测问题的典型支持技术，它在用户划分、行为预测、规则梳理等方面具有广泛的应用前景，决策树甚至可以作为其他建模技术前期进行变量筛选的一种方法，即通过决策树的分割来筛选有效地输入自变量。

关于决策树的详细介绍和实践中的注意事项，可参考本书 10.2 节。

2.3.2　神经网络

神经网络（Neural Network）是通过数学算法来模仿人脑思维的，它是数据挖掘中机器学习的典型代表。神经网络是人脑的抽象计算模型，我们知道人脑中有数以百亿个神经元（人脑处理信息的微单元），这些神经元之间相互连接，使得人的大脑产生精密的逻辑思维。

而数据挖掘中的“神经网络”也是由大量并行分布的人工神经元（微处理单元）组成的，它有通过调整连接强度从经验知识中进行学习的能力，并可以将这些知识进行应用。

简单来讲，“神经网络”就是通过输入多个非线性模型以及不同模型之间的加权互联（加权的过程在隐蔽层完成），最终得到一个输出模型。其中，隐蔽层所包含的就是非线性函数。

目前最主流的“神经网络”算法是反馈传播（Backpropagation），该算法在多层前向型（Multilayer Feed-Forward）神经网络上进行学习，而多层前向型神经网络又是由一个输入层、一个或多个隐蔽层以及一个输出层组成的，“神经网络”的典型结构如图 2-1 所示。

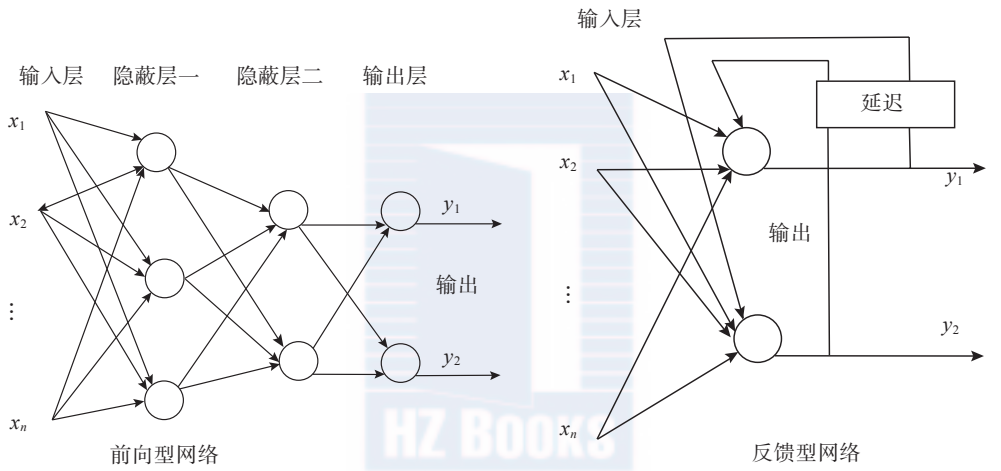


图 2-1 “神经网络”的典型结构图

由于“神经网络”拥有特有的大规模并行结构和信息的并行处理等特点，因此它具有良好的自适应性、自组织性和高容错性，并且具有较强的学习、记忆和识别功能。目前神经网络已经在信号处理、模式识别、专家系统、预测系统等众多领域中得到广泛的应用。

“神经网络”的主要缺点就是其知识和结果的不可解释性，没有人知道隐蔽层里的非线性函数到底是如何处理自变量的，“神经网络”应用中的产出物在很多时候让人看不清其中的逻辑关系。但是，它的这个缺点并没有影响该技术在数据化运营中的广泛应用，甚至可以这样认为，正是因为其结果具有不可解释性，反而更有可能促使我们发现新的没有认识到的规律和关系。

在利用“神经网络”技术建模的过程中，有以下 5 个因素对模型结果有着重大影响：

- 层数。
- 每层中输入变量的数量。

❑ 联系的种类。

❑ 联系的程度。

❑ 转换函数，又称激活函数或挤压函数。

关于这 5 个因素的详细说明，请参考本书 10.1.1 节。

“神经网络”技术在数据化运营中的主要用途体现在：作为分类、预测问题的重要技术支持，在用户划分、行为预测、营销响应等诸多方面具有广泛的应用前景。

关于神经网络的详细介绍和实践中的注意事项，可参考本书 10.1 节。

2.3.3　回归

回归（Regression）分析包括线性回归（Linear Regression），这里主要是指多元线性回归和逻辑斯蒂回归（Logistic Regression）。其中，在数据化运营中更多使用的是逻辑斯蒂回归，它又包括响应预测、分类划分等内容。

多元线性回归主要描述一个因变量如何随着一批自变量的变化而变化，其回归公式（回归方程）就是因变量与自变量关系的数据反映。因变量的变化包括两部分：系统性变化与随机变化，其中，系统性变化是由自变量引起的（自变量可以解释的），随机变化是不能由自变量解释的，通常也称作残值。

在用来估算多元线性回归方程中自变量系数的方法中，最常用的是最小二乘法，即找出一组对应自变量的相应参数，以使因变量的实际观测值与回归方程的预测值之间的总方差减到最小。

对多元线性回归方程的参数估计，是基于下列假设的：

❑ 输入变量是确定的变量，不是随机变量，而且输入的变量间无线性相关，即无共线性。

❑ 随机误差的期望值总和为零，即随机误差与自变量不相关。

❑ 随机误差呈现正态分布[⊖]。

如果不满足上述假设，就不能用最小二乘法进行回归系数的估算了。

逻辑斯蒂回归（Logistic Regression）相比于线性回归来说，在数据化运营中有更主流更频繁的应用，主要是因为该分析技术可以很好地回答诸如预测、分类等数据化运营常见的

⊖ 正态分布也称常态分布，是具有两个参数μ和σ^2的连续型随机变量分布，第一个参数μ是服从正态分布的随机变量的均值，第二个参数σ^2是此随机变量的方差，服从正态分布的随机变量的概率规律为取与μ邻近的值的概率大，而取离μ越远的值的概率越小；σ越小，分布越集中在μ附近，σ越大，分布越分散。

分析项目主题。简单来讲，凡是预测“两选一”事件的可能性（比如，“响应”还是“不响应”；“买”还是“不买”；“流失”还是“不流失”），都可以采用逻辑斯蒂回归方程。

逻辑斯蒂回归预测的因变量是介于 0 和 1 之间的概率，如果对这个概率进行换算，就可以用线性公式描述因变量与自变量的关系了，具体公式如下：

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$$

与多元线性回归所采用的最小二乘法的参数估计方法相对应，最大似然法是逻辑斯蒂回归所采用的参数估计方法，其原理是找到这样一个参数，可以让样本数据所包含的观察值被观察到的可能性最大。这种寻找最大可能性的方法需要反复计算，对计算能力有很高的要求。最大似然法的优点是在大样本数据中参数的估值稳定、偏差小，估值方差小。

关于线性回归和逻辑回归的详细介绍和在实践应用中的注意事项，可参考本书 10.3 节和 10.4 节。

2.3.4 关联规则

关联规则（Association Rule）是在数据库和数据挖掘领域中被发明并被广泛研究的一种重要模型，关联规则数据挖掘的主要目的是找出数据集中的频繁模式（Frequent Pattern），即多次重复出现的模式和并发关系（Cooccurrence Relationships），即同时出现的关系，频繁和并发关系也称作关联（Association）。

应用关联规则最经典的案例就是购物篮分析（Basket Analysis），通过分析顾客购物篮中商品之间的关联，可以挖掘顾客的购物习惯，从而帮助零售商更好地制定有针对性的营销策略。

以下列举一个简单的关联规则的例子：

婴儿尿不湿→啤酒 [支持度 =10%，置信度 =70%]

这个规则表明，在所有顾客中，有 10% 的顾客同时购买了婴儿尿不湿和啤酒，而在所有购买了婴儿尿不湿的顾客中，占 70% 的人同时还购买了啤酒。发现这个关联规则后，超市零售商决定把婴儿尿不湿和啤酒摆放在一起进行促销，结果明显提升了销售额，这就是发生在沃尔玛超市中“啤酒和尿不湿”的经典营销案例。

上面的案例是否让你对支持度和置信度有了一定的了解？事实上，支持度（Support）和置信度（Confidence）是衡量关联规则强度的两个重要指标，它们分别反映着所发现规则的有用性和确定性。其中支持度：规则 $X \rightarrow Y$ 的支持度是指事物全集中包含 $X \cup Y$ 的事物百分比。支持度主要衡量规则的有用性，如果支持度太小，则说明相应规则只是偶发事件。在

商业实战中，偶发事件很可能没有商业价值；置信度：规则 $X \rightarrow Y$ 的置信度是指既包含了 X 又包含了 Y 的事物数量占所有包含了 X 的事物数量的百分比。置信度主要衡量规则的确定性（可预测性），如果置信度太低，那么从 X 就很难可靠地推断出 Y 来，置信度太低的规则在实践应用中也没有太大用处。

在众多的关联规则数据挖掘算法中，最著名的就是 Apriori 算法，该算法具体分为以下两步进行：

（1）生成所有的频繁项目集。一个频繁项目集（Frequent Itemset）是一个支持度高于最小支持度阀值（min-sup）的项目集。

（2）从频繁项目集中生成所有的可信关联规则。这里可信关联规则是指置信度大于最小置信度阀值（min-conf）的规则。

关联规则算法不但在数值型数据集的分析中有很大用途，而且在纯文本文档和网页文件中，也有着重要用途。比如发现单词间的并发关系以及 Web 的使用模式等，这些都是 Web 数据挖掘、搜索及推荐的基础。

2.3.5　聚类

聚类（Clustering）分析有一个通俗的解释和比喻，那就是“物以类聚，人以群分”。针对几个特定的业务指标，可以将观察对象的群体按照相似性和相异性进行不同群组的划分。经过划分后，每个群组内部各对象间的相似度会很高，而在不同群组之间的对象彼此间将具有很高的相异度。

聚类分析的算法可以分为划分的方法（Partitioning Method）、层次的方法（Hierarchical Method）、基于密度的方法（Density-based Method）、基于网格的方法（Grid-based Method）、基于模型的方法（Model-based Method）等，其中，前面两种方法最为常用。

对于划分的方法（Partitioning Method），当给定 m 个对象的数据集，以及希望生成的细分群体数量 K 后，即可采用这种方法将这些对象分成 K 组（$K \leqslant m$），使得每个组内对象是相似的，而组间的对象是相异的。最常用的划分方法是 K-Means 方法，其具体原理是：首先，随机选择 K 个对象，并且所选择的每个对象都代表一个组的初始均值或初始的组中心值；对剩余的每个对象，根据其与各个组初始均值的距离，将它们分配给最近的（最相似）小组；然后，重新计算每个小组新的均值；这个过程不断重复，直到所有的对象在 K 组分布中都找到离自己最近的组。

层次的方法（Hierarchical Method）则是指依次让最相似的数据对象两两合并，这样不断地合并，最后就形成了一棵聚类树。

聚类技术在数据分析和数据化运营中的主要用途表现在：既可以直接作为模型对观察对象进行群体划分，为业务方的精细化运营提供具体的细分依据和相应的运营方案建议，又可在数据处理阶段用作数据探索的工具，包括发现离群点、孤立点，数据降维的手段和方法，通过聚类发现数据间的深层次的关系等。

关于聚类技术的详细介绍和应用实践中的注意事项，可参考本书第 9 章。

2.3.6 贝叶斯分类方法

贝叶斯分类方法（Bayesian Classifier）是非常成熟的统计学分类方法，它主要用来预测类成员间关系的可能性。比如通过一个给定观察值的相关属性来判断其属于一个特定类别的概率。贝叶斯分类方法是基于贝叶斯定理的，已经有研究表明，朴素贝叶斯分类方法作为一种简单贝叶斯分类算法甚至可以跟决策树和神经网络算法相媲美。

贝叶斯定理的公式如下：

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

其中，X 表示 n 个属性的测量描述；H 为某种假设，比如假设某观察值 X 属于某个特定的类别 C；对于分类问题，希望确定 $P(H|X)$，即能通过给定的 X 的测量描述，来得到 H 成立的概率，也就是给出 X 的属性值，计算出该观察值属于类别 C 的概率。因为 $P(H|X)$ 是后验概率（Posterior Probability），所以又称其为在条件 X 下，H 的后验概率。

举例来说，假设数据属性仅限于用教育背景和收入来描述顾客，而 X 是一位硕士学历，收入 10 万元的顾客。假定 H 表示假设我们的顾客将购买苹果手机，则 $P(H|X)$ 表示当我们知道顾客的教育背景和收入情况后，该顾客将购买苹果手机的概率；相反，$P(X|H)$ 则表示如果已知顾客购买苹果手机，则该顾客是硕士学历并且收入 10 万元的概率；而 $P(X)$ 则是 X 的先验概率，表示顾客中的某个人属于硕士学历且收入 10 万元的概率；$P(H)$ 也是先验概率，只不过是任意给定顾客将购买苹果手机的概率，而不会去管他们的教育背景和收入情况。

从上面的介绍可见，相比于先验概率 $P(H)$，后验概率 $P(H|X)$ 基于了更多的信息（比如顾客的信息属性），而 $P(H)$ 是独立于 X 的。

贝叶斯定理是朴素贝叶斯分类法（Naive Bayesian Classifier）的基础，如果给定数据集里有 M 个分类类别，通过朴素贝叶斯分类法，可以预测给定观察值是否属于具有最高后验概率的特定类别，也就是说，朴素贝叶斯分类方法预测 X 属于类别 C_i 时，表示当且仅当

$$P(C_i \mid X) > P(C_j \mid X) \quad 1 \leqslant j \leqslant m,\ j \neq i$$

此时如果最大化 $P(C_i|X)$，其 $P(C_i|X)$ 最大的类 C_i 被称为最大后验假设，根据贝叶斯定理

$$P(C_i \mid X)=\frac{P(X \mid C_i)P(C_i)}{P(X)}$$

可知，由于 $P(X)$ 对于所有的类别是均等的，因此只需要 $P(X|C_i)P(C_i)$ 取最大即可。

为了预测一个未知样本 X 的类别，可对每个类别 C_i 估算相应的 $P(X|C_i)P(C_i)$。样本 X 归属于类别 C_i，当且仅当

$$P(C_i \mid X)>P(C_j \mid X)\ \ 1 \leqslant j \leqslant m,\ j \neq i$$

贝叶斯分类方法在数据化运营实践中主要用于分类问题的归类等应用场景。

2.3.7　支持向量机

支持向量机（Support Vector Machine）是 Vapnik 等人于 1995 年率先提出的，是近年来机器学习研究的一个重大成果。与传统的神经网络技术相比，支持向量机不仅结构简单，而且各项技术的性能也明显提升，因此它成为当今机器学习领域的热点之一。

作为一种新的分类方法，支持向量机以结构风险最小为原则。在线性的情况下，就在原空间寻找两类样本的最优分类超平面。在非线性的情况下，它使用一种非线性的映射，将原训练集数据映射到较高的维上。在新的维上，它搜索线性最佳分离超平面。使用一个适当的对足够高维的非线性映射，两类数据总可以被超平面分开。

支持向量机的基本概念如下：

设给定的训练样本集为 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，其中 $x_i \in R^n$, $y \in \{-1,1\}$。

再假设该训练集可被一个超平面线性划分，设该超平面记为 $(w, x)+b=0$。

支持向量机的基本思想可用图 2-2 的两维情况举例说明。

图中圆形和方形代表两类样本，H 为分类线，H_1、H_2，分别为过各类样本中离分类线最近的样本并且平行于分类线的直线，它们之间的距离叫做分类间隔（Margin）。所谓的最优分类线就是要求分类线不但能将两类正确分开（训练错误为 0），而且能使分类间隔最大。推广到高维空间，最优分类线就成了最优分类面。

其中，距离超平面最近的一类向量被称为支持向量（Support Vector），一组支持向量可以唯一地确定一个超平面。通过学习算法，SVM 可以自动寻找出那些对分类有较好区分能力的支持向量，由此构造出的分类器则可以最大化类与类的间隔，因而有较好的适应能力和较高的分类准确率。

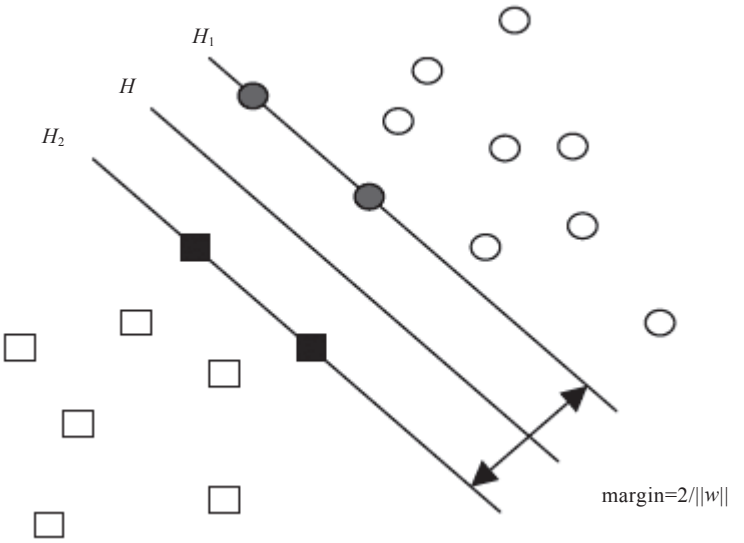


图 2-2　线性可分情况下的最优分类线

支持向量机的缺点是训练数据较大，但是，它的优点也是很明显的——对于复杂的非线性的决策边界的建模能力高度准确，并且也不太容易过拟合[⊖]。

支持向量机主要用在预测、分类这样的实际分析需求场景中。

2.3.8　主成分分析

严格意义上讲，主成分分析（Principal Components Analysis）属于传统的统计分析技术范畴，但是正如本章前面所阐述的，统计分析与数据挖掘并没有严格的分割，因此在数据挖掘实战应用中也常常会用到这种方式，从这个角度讲，主成分分析也是数据挖掘商业实战中常用的一种分析技术和数据处理技术。

主成分分析会通过线性组合将多个原始变量合并成若干个主成分，这样每个主成分都变成了原始变量的线性组合。这种转变的目的，一方面是可以大幅降低原始数据的维度，同时也在此过程中发现原始数据属性之间的关系。

主成分分析的主要步骤如下：

1）通常要先进行各变量的标准化工作，标准化的目的是将数据按照比例进行缩放，使之落入一个小的区间范围之内，从而让不同的变量经过标准化处理后可以有平等的分析和比

⊖ 过拟合，是指模型在训练的时候对样本“模拟”过好，不能反映真实的输入输出函数关系，所以一旦模型面对新的应用数据的时候，就表现为不准确的程度较大。

较基础。关于数据标准化的详细介绍，可参考本书 8.5.4 节和 9.3.2 节。

2）选择协方差阵或者相关阵计算特征根及对应的特征向量。

3）计算方差贡献率，并根据方差贡献率的阀值选取合适的主成分个数。

4）根据主成分载荷的大小对选择的主成分进行命名。

5）根据主成分载荷计算各个主成分的得分。

将主成分进行推广和延伸即成为因子分析（Factor Analysis），因子分析在综合原始变量信息的基础上将会力图构筑若干个意义较为明确的公因子；也就是说，采用少数几个因子描述多个指标之间的联系，将比较密切的变量归为同一类中，每类变量即是一个因子。之所以称其为因子，是因为它们实际上是不可测量的，只能解释。

主成分分析是因子分析的一个特例，两者的区别和联系主要表现在以下方面：

- ❑ 主成分分析会把主成分表示成各个原始变量的线性组合，而因子分析则把原始变量表示成各个因子的线性组合。这个区别最直观也最容易记住。
- ❑ 主成分分析的重点在于解释原始变量的总方差，而因子分析的重点在于解释原始变量的协方差。
- ❑ 在主成分分析中，有几个原始变量就有几个主成分，而在因子分析中，因子个数可以根据业务场景的需要人为指定，并且指定的因子数量不同，则分析结果也会有差异。
- ❑ 在主成分分析中，给定的协方差矩阵或者相关矩阵的特征值是唯一时，主成分也是唯一的，但是在因子分析中，因子不是唯一的，并且通过旋转可以得到不同的因子。

主成分分析和因子分析在数据化运营实践中主要用于数据处理、降维、变量间关系的探索等方面，同时作为统计学里的基本而重要的分析工具和分析方法，它们在一些专题分析中也有着广泛的应用。

2.3.9　假设检验

假设检验（Hypothesis Test）是现代统计学的基础和核心之一，其主要研究在一定的条件下，总体是否具备某些特定特征。

假设检验的基本原理就是小概率事件原理，即观测小概率事件在假设成立的情况下是否发生。如果在一次试验中，小概率事件发生了，那么说明假设在一定的显著性水平下不可靠或者不成立；如果在一次试验中，小概率事件没有发生，那么也只能说明没有足够理由相信假设是错误的，但是也并不能说明假设是正确的，因为无法收集到所有的证据来证明假设是

正确的。

假设检验的结论是在一定的显著性水平下得出的。因此，当采用此方法观测事件并下结论时，有可能会犯错，这些错误主要有两大类：

- 第Ⅰ类错误：当原假设为真时，却否定它而犯的错误，即拒绝正确假设的错误，也叫弃真错误。犯第Ⅰ类错误的概率记为α，通常也叫α错误，α=1− 置信度。
- 第Ⅱ类错误：当原假设为假时，却肯定它而犯的错误，即接受错误假设的错误，也叫纳伪错误。犯第Ⅱ类错误的概率记为β，通常也叫β错误。

上述这两类错误在其他条件不变的情况下是相反的，即α增大时，β就减小；α减小时，β就增大。α错误容易受数据分析人员的控制，因此在假设检验中，通常会先控制第Ⅰ类错误发生的概率α，具体表现为：在做假设检验之前先指定一个α的具体数值，通常取0.05，也可以取0.1或0.001。

在数据化运营的商业实践中，假设检验最常用的场景就是用于“运营效果的评估”上，本书第12章将针对最常见、最基本的假设检验形式和技术做出比较详细的梳理和举例。

2.4 互联网行业数据挖掘应用的特点

相对于传统行业而言，互联网行业的数据挖掘和数据化运营有如下的一些主要特点：

- 数据的海量性。互联网行业相比传统行业第一个区别就是收集、存储的数据是海量的，这一方面是因为互联网的使用已经成为普通人日常生活和工作中不可或缺的一部分，另一方面更是因为用户网络行为的每一步都会被作为网络日志记录下来。海量的数据、海量的字段、海量的信息，尤其是海量的字段，使得分析之前对于分析字段的挑选和排查工作显得无比重要，无以复加。如何大浪淘沙挑选变量则为重中之重，对此很难一言以蔽之的进行总结，还是用三分技术，七分业务来理解吧。本书从第7～12章，几乎每章都用大量的篇幅讨论如何在具体的分析课题和项目中选择变量、评估变量、转换变量，乃至如何通过清洗后的核心变量完成最终的分析结论（挖掘模型）。
- 数据分析（挖掘）的周期短。鉴于互联网行业白热化的市场竞争格局，以及该行业相对成熟的高级数据化运营实践，该行业的数据分析（挖掘）通常允许的分析周期（项目周期）要明显短于传统行业。行业技术应用飞速发展，产品和竞争一日千里，都使该行业的数据挖掘项目的时间进度比传统行业的项目模式快得多。一方面要保证挖掘结果的起码质量，另一方面要满足这个行业超快的行业节奏，这也使得传统的挖掘分析思路和步调必须改革和升华，从而具有鲜明的Internet色彩。

- 数据分析（挖掘）成果的时效性明显变短。由于互联网行业的用户行为相对于传统行业而言变化非常快，导致相应的数据分析挖掘成果的时效性也比传统行业明显缩短。举例来说，互联网行业的产品更新换代很多是以月为单位的，新产品层出不穷，老产品要及时下线，因此，针对具体产品的数据分析（挖掘）成果的时效性也明显变短；或者说，用户行为变化快，网络环境变化快，导致模型的维护和优化的时间周期也明显变短，传统行业里的“用户流失预测模型”可能只需要每年更新优化一次，但是在互联网行业里类似的模型可能 3 个月左右就有必要更新优化了。

- 互联网行业新技术、新应用、新模式的更新换代相比于传统行业而言更加迅速、周期更短、更加具有颠覆性，相应地对数据分析挖掘的应用需求也更为苛刻，且要多样化。以中国互联网行业的发展为例，作为第一代互联网企业的代表，新浪、搜狐、雅虎等门户网站的 Web 1.0 模式（传统媒体的电子化）从产生到被以 Google、百度等搜索引擎企业的 Web 2.0 模式（制造者与使用者的合一）所超越，前后不过 10 年左右的时间，而目前这个 Web 2.0 模式已经逐渐有被以微博为代表 Web 3.0 模式（SNS 模式）超越的趋势。具体到数据分析所服务的互联网业务和应用来说，从最初的常规、主流的分析挖掘支持，到以微博应用为代表的新的分析需求，再到目前风头正健的移动互联网的数据分析和应用，互联网行业的数据分析大显身手的天地在不断扩大，新的应用源源不断，新的挑战让人们应接不暇，这一切都要求数据分析师自觉、主动去学习、去充实、去提升自己、去跟上互联网发展的脚步。

HZ BOOKS
华章图书

第3章
数据化运营中常见的数据分析项目类型

千举万变，其道一也。

——《荀子·儒效》

HZ BOOKS 华章图书

数据化运营中的数据分析项目类型比较多，涉及不同的业务场景、业务目的和分析技术。在本章中，按照业务用途的不同将其做了一个大概的分类，并针对每一类项目的特点和具体采用的分析挖掘技术进行了详细的说明和举例示范。

一个成功的数据分析挖掘项目，首先要有准确的业务需求描述，之后则要求项目相关人员自始至终对业务有正确的理解和判断，所以对于本章所分享的所有分析项目类型以及对应的分析挖掘技术，读者只有在深刻理解和掌握相应业务背景的基础上才可以真正理解项目类型的特点、目的，以及相应的分析挖掘技术合适与否。

对业务的理解和思考，永远高于项目的分类和分析技术的选择。

3.1　目标客户的特征分析

目标客户的特征分析几乎是数据化运营企业实践中最普遍、频率最高的业务分析需求之一，原因在于数据化运营的第一步（最基础的步骤）就是要找准你的目标客户、目标受众，然后才是相应的运营方案、个性化的产品与服务等。是不加区别的普遍运营还是有目标有重点的精细化运营，这是传统的粗放模式与精细的数据化运营最直接、最显性的区别。

在目标客户的典型特征分析中，业务场景可以是试运营之前的虚拟特征探索，也可以是试运营之后来自真实运营数据基础上的分析、挖掘与提炼，两者目标一致，只是思路不同、数据来源不同而已。另外，分析技术也有一定的差异。

对于试运营之前的虚拟特征探索，是指目标客户在真实的业务环境里还没有产生，并没有一个与真实业务环境一致的数据来源可以用于分析目标客户的特点，因此只能通过简化、类比、假设等手段，来寻找一个与真实业务环境近似的数据来源，从而进行模拟、探索，并从中发现一些似乎可以借鉴和参考的目标用户特征，然后把这些特征放到真实的业务环境中去试运营。之后根据真实的效果反馈数据，修正我们的目标用户特征。一个典型的业务场景举例就是 A 公司推出了一个在线转账产品，用户通过该产品在线转账时产生的交易费用相比于普通的网银要便宜些。在正式上线该转账产品之前，产品运营团队需要一个初步的目标客户特征报告。很明显，在这个时刻，产品还没有上线，是无法拥有真实使用该产品的用户的，自然也没有相应数据的积累，那这个时候所做的目标客户特征分析只能是按照产品设计的初衷、产品定位，以及运营团队心中理想化的猜测，从企业历史数据中模拟、近似地整理出前期期望中的目标客户典型特征，很明显这里的数据并非来自该产品正式上线后的实际用户数据（还没有这些真实的数据产生），所以这类场景的分析只能是虚拟的特征分析。具体来说，本项目先要从企业历史数据中寻找有在线交易历史的买卖双方，在线行为活跃的用户，以及相应的一些网站行为、捆绑了某知名的第三方支付工具的用户等，然后根据这些行

为字段和模拟的人群，去分析我们期望的目标客户特征，在通过历史数据仓库的对比后，准确掌握该目标群体的规模和层次，从而提交运营业务团队正式运营。

对于试运营之后的来自真实运营数据基础上的用户特征分析，相对而言，就比上述的模拟数据分析来得更真实更可行，也更贴近业务实际。在该业务场景下，数据的提取完全符合业务需求，且收集到的用户也是真实使用了该产品的用户，基于这些真实用户的分析就不是虚拟的猜测和模拟了，而是有根有据的铁的事实。在企业的数据化运营实践中，这后一种场景更加普遍，也更加可靠。

对于上面提到的案例，在经过一段时间的试运营之后，企业积累了一定数量使用该产品的用户数据。现在产品运营团队需要基于该批实际的用户数据，整理分析出该产品的核心目标用户特征分析报告，以供后期运营团队、产品开发团队、服务团队更有针对性、更有效地进行运营和服务。在这种基于真实的业务场景数据基础上的客户特征分析，有很多分析技术可以采用（本书第 11 章将针对“用户特征分析”进行专题介绍，分享其中最主要的一些分析技术），但是其中采用预测模型的思路是该场景与上述“虚拟场景”数据分析的一个不同，上述“虚拟场景”数据分析一般来说是无法进行预测模型思路的探索的。

关于目标客户特征分析的具体技术、思路、实例分享，可参考本书第 11 章。

3.2 目标客户的预测（响应、分类）模型

这里的预测（响应、分类）模型包括流失预警模型、付费预测模型、续费预测模型、运营活动响应模型等。

预测（响应、分类）模型是数据挖掘中最常用的一种模型类型，几乎成了数据挖掘技术应用的一个主要代名词。很多书籍介绍到数据挖掘的技术和应用，首先都会列举预测（响应、分类）模型，主要的原因可能是响应模型的核心就是响应概率，而响应概率其实就是我们在第 1 章中介绍的数据化运营六要素里的核心要素——概率（Probability），数据化运营 6 要素的核心是以数据分析挖掘支撑的目标响应概率（Probability），在此基础上围绕产品功能优化、目标用户细分、活动（文案）创意、渠道优化、成本的调整等重要环节、要素，共同达成数据化运营的持续完善、成功。

预测（响应、分类）模型基于真实业务场景产生的数据而进行的预测（响应、分类）模型搭建，其中涉及的主要数据挖掘技术包括逻辑回归、决策树、神经网络、支持向量机等。有没有一个算法总是优先于其他算法呢？答案是否定的，没有哪个算法在任何场景下都总能最优胜任响应模型的搭建，所以在通常的建模过程中，数据分析师都会尝试多种不同的算法，然后根据随后的验证效果以及具体业务项目的资源和价值进行权衡，并做出最终的选择。

根据建模数据中实际响应比例的大小进行分类，响应模型还可以细分为普通响应模型和稀有事件响应模型，一般来讲，如果响应比例低于 1%，则应当作为稀有事件响应模型来进行处理，其中的核心就是抽样，通过抽样技术人为放大分析数据样本里响应事件的比例，增加响应事件的浓度，从而在建模过程中更好地捕捉、拟合其中自变量与因变量的关系。

预测（响应、分类）模型除了可以有效预测个体响应的概率之外，模型本身显示出的重要输入变量与目标变量的关系也有重要的业务价值，比如说可以转化成伴随（甚至导致）发生响应（生成事件）的关联因素、重要因素的提炼。而很多时候，这种重要因素的提炼，是可以作为数据化运营中的新规则、新启发，甚至是运营的“新抓手”的。诚然，从严格的统计学角度来看，预测响应模型中的输入变量与目标变量之间的重要关系并不一定是因果关系，严格意义上的因果关系还需要后期进行深入的分析和实验；即便如此，这种输入变量与目标变量之间的重要关系也常常会对数据化运营具有重要的参考和启发价值。

比如说，我们通过对在线交易的卖家进行深入分析挖掘，建立了预测响应模型，从而根据一系列特定行为和属性的组合，来判断在特定时间段内发生在线交易的可能性。这个响应模型除了生成每个 Member_Id 在特定时间段发生在线交易的可能性之外，从模型中提炼出来的一些重要输入变量与目标变量（是否发生在线交易），以及它们之间的关系（包括正向或负向关系，重要性的强弱等）对数据化运营也有着很重要的参考和启发。在本案例中，我们发现输入变量近 30 天店铺曝光量、店铺装修打分超过 25 分等与是否在线交易有着最大的正相关。根据这些发现和规则整理，尽管不能肯定这些输入变量与是否在线交易有因果关系，但这些正向的强烈的关联性也足以为提升在线交易的数据化运营提供重要的启发和抓手。我们有一定的理由相信，如果卖家提升店铺的曝光量，如果卖家把自己的店铺装修得更好，促进卖家在线成交的可能性会加大。

3.3　运营群体的活跃度定义

运营群体（目标群体）的活跃度定义，这也是数据化运营基本的普遍的要求。数据化运营与传统的粗放型运营最主要的区别（核心）就是前者是可以准确地用数据衡量，而且这种衡量是自始至终地贯穿于数据化运营的全过程；而在运营全过程的衡量监控中，活跃度作为一个综合的判断指标，又在数据化运营实践中有着广泛的应用和曝光。活跃度的定义没有统一的描述，一般都是根据特定的业务场景和运营需求来量身订做的。但是，纵观无数场景中的活跃度定义，可以发现其中是有一些固定的骨架作为基础和核心的。其中最重要、最常见的两个基本点如下。

1）活跃度的组成指标应该是该业务场景中最核心的行为因素。

2）衡量活跃度的定义合适与否的重要判断依据是其能否有效回答业务需求的终极目标。

下面我们用具体的案例来解释上述两个基本点。

案例：PM 产品是一款在线的 SAAS 产品，其用途在于协助卖家实时捕捉买家访问店铺的情况，并且通过该 PM 产品可以实现跟买家对话、交换联系方式等功能。作为 PM 产品的运营方，其运营策略是向所有平台的卖家免费提供 PM 产品的基本功能（每天只能联系一位到访的买家，也即限制了联系多位到访买家的功能）、向部分优质卖家提供一定期限内免费的 PM 产品全功能（这部分优质卖家免费获赠 PM 产品，可以享受跟付费一样的全功能）、向目标卖家在线售卖 PM 产品。

经过一段时间的运营，现在管理层需要数据分析团队**定义一个合理的“PM 产品用户活跃度”，使得满足一定活跃度分值的用户能比较容易转化成为 PM 产品的付费用户，同时这个合适的定义还可以帮助有效监控每天 PM 产品的运营效果和效率**。

根据上面的案例背景描述，以及之前的活跃度定义的两个基本点来看，在本案例中，该业务场景中最核心的行为因素就是卖家使用该 PM 产品与到访买家的洽谈动作（表现形式为洽谈的次数）、在线登录该 PM 产品的登录次数等。而该分析需求的终极目的就是促成付费用户的转化，所以项目最终活跃度的定义是否合适，是否满足业务需求，一个最重要的评估依据就是按照该活跃度定义出来的活跃用户群体里，可以覆盖多少实际的 PM 产品付费用户。从理论上来说，覆盖率越高越好，如果覆盖率不高，比如，实际付费用户群体里只有 50% 包含在活跃度定义的活跃群体里，那么这个活跃度的定义是不能满足当初的业务需求的，也就是说这是一个不成功的定义。

活跃度的定义所涉及的统计技术主要有两个，一个是主成分分析，另一个是数据的标准化。其中，主成分分析的目的，就是把多个核心行为指标转化为一个或少数几个主成分，并最终转化成一个综合的分数，来作为活跃度的定义，到底是取第一个主成分，还是前两个或前三个，这要取决于主成分分析的特征根和累计方差贡献率，一般来说，如果前面几个特征根的累计方差贡献率达到 80% 以上，就可以基本认为前面几个主成分就可以相应地代表原始数据的大部分信息了；至于数据标准化技术得到了普遍采用，主要是因为不同的指标有不同的度量尺度，只有在标准化之后，才可以将数据按照比例进行缩放，使之落入一个小的区间范围之内，这样，不同变量经过标准化处理后就可以有平等的分析和比较基础了。**关于数据标准化的详细介绍，可参看本书 8.5.4 节和 9.3.2 节。**

3.4 用户路径分析

用户路径分析是互联网行业特有的分析专题，主要是分析用户在网页上流转的规律和特

点，发现频繁访问的路径模式，这些路径的发现可以有很多业务用途，包括提炼特定用户群体的主流路径、网页设计的优化和改版、用户可能浏览的下一个页面的预测、特定群体的浏览特征等。从这些典型的用途示例中可以看到，数据化运营中的很多业务部门都需要应用用户路径分析，包括运营部门、产品设计部门（PD）、用户体验设计部门（User Experience Design，UED）等。

路径分析所用的数据主要是 Web 服务器中的日志数据，不过，互联网的特性使得日志数据的规模通常都是海量的。据预测，到 2020 年，全球以电子形式存储的数据量将达到 35ZB（相当于 10 亿块 1TB 的硬盘的容量），是 2009 年全球存储量的 40 倍。而在 2010 年年底，根据 IDC 的统计，全球的数据量已经达到了 120 万 PB，或 1.2ZB。如果将这些数据都刻录在 DVD 上，那么光把这些 DVD 盘片堆叠起来就可以从地球往月球一个来回（单程约 24 万英里）。

路径分析常用的分析技术有两类，一类是有算法支持的，另一类是严格按照步骤顺序遍历主要路径的。**关于路径分析中具体的算法和示例将在第 13 章做详细的说明。**

在互联网数据化运营的实践中，如果能把单纯的路径分析技术、算法与其他相关的数据分析技术、挖掘技术相融合，那么将会产生更大的应用价值和更为广阔的前景。这种融合的思路包括通过聚类技术划分出不同的群体，然后分析不同群体的路径特征，针对特定人群进行的路径分析，比如，对比付费人群的主要路径与非付费人群的主要路径，优化页面布局等、根据下单付费路径中频繁出现的异常模式可能来对付费环境的页面设计进行优化，提升付费转化率，减少下单后的流失风险等。

在运营团队看来，路径分析的主要用途之一，即为监控运营活动（或者目标客户）的典型路径，看是否与当初的运营设想一致。如果不一致，就继续深入分析原因，调整运营思路或页面布局，最终目的就是提升用户点击页面的效率；其二就是通过路径分析，提炼新的有价值的频繁路径模式，并且在以后的运营中对这些模式加以应用，提升运营的效率和特定效果。比如，通过某次运营活动的路径分析，我们发现从 A 入口进来的用户有 30% 会进入 C 页面，然后再进入 B 页面，而 A 入口是系列运营活动的主要入口之一，基于这个 C 页面的重要性发现，运营人员在该页面设置了新的提醒动作，取得了较好的深度转化率。

在产品设计部门（PD）看来，路径分析是实现产品优化的一个重要依据和工具，被路径分析证明是冷僻的功能点和路径的，或许可以被进一步考虑是否有必要取消或优化。对于 UED 来说，路径分析也是这样帮助他们优化页面设计的。

3.5　交叉销售模型

交叉销售这个概念在传统行业里其实已经非常成熟了，也已被普遍应用，其背后的理论

依据是一旦客户购买了商品（或者成为付费用户），企业就会想方设法保留和延长这些客户在企业的生命周期和客户的利润贡献，一般会有两个运营选择方向，一是延缓客户流失，让客户尽可能长久地留存，在该场景下，通常就是客户流失预警模型发挥作用，利用流失预警模型，提前锁定最可能流失的有价值的用户，然后客户服务团队采用各种客户关怀措施，尽量挽留客户，从而最终降低客户流失率；二是让客户消费更多的商品和服务，从而更大地提升客户的商业价值，挖掘客户利润，这种尽量挖掘客户利润的说法在以客户为中心的激烈竞争的 2.0 时代显得有些赤裸裸，所以，更加温和的说法就是通过数据分析挖掘，找出客户进一步的消费需求（潜在需求），从而更好及更主动地引导、满足、迎合客户需求，创造企业和客户的双赢。在这第二类场景中，涉及的主要应用模型就是交叉销售模型。

交叉销售模型通过对用户历史消费数据的分析挖掘，找出有明显关联性质的商品组合，然后用不同的建模方法，去构建消费者购买这些关联商品组合的可能性模型，再用其中优秀的模型去预测新客户中购买特定商品组合的可能性。这里的商品组合可以是同时购买，也可以有先后顺序，不可一概而论，关键要看具体的业务场景和业务背景。

不同的交叉销售模型有不同的思路和不同的建模技术，但是前提一般都是通过数据分析找出有明显意义和商业价值的商品组合，可以同时购买，也可以有先后顺序，然后根据找出的这些特性去建模。

综合数据挖掘的中外企业实践来看，最少有 4 种完全不同的思路，可以分别在不同的项目背景中圆满完成建立交叉销售模型的这个任务。一是按照关联技术（Association Analysis），也即通常所说的购物篮分析，发现那些有较大可能被一起采购的商品，将它们进行有针对性的促销和捆绑，这就是交叉销售；二是借鉴响应模型的思路，为某几种重要商品分别建立预测模型，对潜在消费者通过这些特定预测模型进行过滤，然后针对最有可能的前 5% 的消费者进行精确的营销推广；三是仍然借鉴预测响应模型的思路，让重要商品两两组合，找出那些最有可能消费的潜在客户；四是通过决策树清晰的树状规则，发现基于具体数据资源的具体规则（有的多，有的少），国外很多营销方案的制订和执行实际上都是通过这种方式找到灵感和思路的。

相应的建模技术主要包括关联分析（Association Analysis）、序列分析（Sequence Analysis），即在关联分析的基础上，增加了先后顺序的考虑，以及预测（响应、分类）模型技术，诸如逻辑回归、决策树等。

上面总结的是基于传统行业的实践，这些经验事实上也成功地应用到了互联网行业的数据化运营中。无论是多种在线产品的交叉销售，还是电子商务中的交叉销售，抑或各种服务的推广、运营中的商品捆绑策略，都可以从中看到交叉销售的影子，这种理念正在深入地影响着数据化运营的效果和进程。

下面针对典型的交叉销售模型的应用场景来举个例子：A 产品与 B 产品都是公司 SAAS 系列产品线上的重点产品，经过分析发现两者付费用户的重合度高达 40%，现在运营方需要一个数据分析解决方案，可以有效识别出最可能在消费 A 产品的基础上也消费 B 产品的潜在优质用户。本案例的分析需求，实际上就是一个典型的交叉销售模型的搭建需求，数据分析师在与业务团队充分沟通后，通过现有数据进行分析，找出了同时消费 A 产品和 B 产品（**注意，是同时消费，还是有先后次序，这个具体的定义取决于业务需求的判断，两者取数逻辑不同，应用场景也不同，不过分析建模技术还是可以相同的**）用户的相关的网站行为、商业行为、客户属性等，之后再进行数据分析和挖掘建模，最后得到了一个有效的预测模型，通过该模型可以对新的用户数据进行预测，找出最可能消费 A 产品同时也消费 B 产品的潜在付费用户人群（或名单）。这样，运营方就可以进行精准的目标运营，从而有效提升运营效果，有效提升付费用户数量和付费转化率了。

3.6　信息质量模型

信息质量模型在互联网行业和互联网数据化运营中也是有着广泛基础性应用的。具体来说，电商行业和电商平台连接买卖双方最直接、最关键的纽带就是海量的商品目录、商品 Offer、商品展示等，无论是 B2C（如当当网、凡客网），还是 C2C（如淘宝网），或者是 B2B（如阿里巴巴），只要是以商业为目的，以交易为目的的，都需要采用有效手段去提升海量商业信息（商品目录、商品 Offer、商品展示等）的质量和结构，从而促进交易。在同等条件下，一个要素齐备、布局合理、界面友好的网上店铺或商品展示一定比不具备核心要素、布局不合理、界面不友好的更加容易达成交易，更加容易获得买家的好感，这里揭示的其实就是信息质量的重要价值。

为让读者更加直观了解信息质量的含义，下面通过某网站的截图来举例说明什么是信息质量好的 Offer 效果，如图 3-1 和图 3-2 所示。

不难发现，相对于图 3-2 来说，图 3-1 中有更多的商品要素展示，包括付款方式、产品品牌、产品型号等，另外在详细信息栏目里，所包含的信息也更多更全。也就是说，图 3-1 中商品 Offer 的信息质量要明显好于图 3-2。

互联网行业的信息质量模型所应用的场合主要包括商品 Offer 质量优化、网上店铺质量优化、网上论坛的发帖质量优化、违禁信息的过滤优化等，凡是涉及信息质量监控和优化的场景都是适用（或借鉴）信息质量模型的解决方案的。

图 3-1　信息质量较好的 Offer 界面图

图 3-2　信息质量较差的 Offer 界面图

构建信息质量模型所涉及的主要还是常规的数据挖掘技术，比如回归算法、决策树等。但是对于信息质量模型的需求，由于其目标变量具有一定的特殊性，因此它与目标客户预测（响应）模型在思路和方法上会有一些不同之处，具体内容如下。

任何模型的搭建都是用于响应特定的业务场景和业务需求的，有时候搭建信息质量模型

的目标变量是该信息（如商品 Offer）是否在特定的时间段产生了交易，此时，目标变量就是二元的，即是与否；更多的时候，信息质量模型的目标变量与是否交易没有直接关系（这其实很容易理解，因为影响成交的因素太多），甚至有些时候信息质量本身是主观的判断，在这种情况下，**没有明确的来自实际数据的目标变量。那如何定义目标变量呢？专家打分，模型拟合是一个比较合适的变通策略。**

对于专家打分，模型拟合的具体操作，下面以“商品 Offer 的星级划分”项目为例来进行具体的解释和示范。商品 Offer 其实就是网上交易中，卖家针对每种出售的商品展示具体的商品细节、交易条款、图片细节等，使其构成的一个完整的页面，一般来说买家浏览了某种具体的商品 Offer 以后，只要点击“加入购物车”就可以进行后续的购买付费流程了。在某次“商品 Offer 的星级划分”项目中，目标变量就是专家打分，由业务专家、行业专家基于行业的专业背景知识，针对商品 Offer 构成要素的权重进行人为打分，这些构成要素包括标题长度、图片数量、属性选填的比例、是否有分层价格区间、是否填写供货总量信息、是否有混批说明、是否有运营说明、是否支持在线第三方支付等。首先抽取一定数量的样本，请行业专家对这些样本逐个打分赋值，在取得每种商品 Offer 的具体分数后，把这些分数作为目标变量，利用数据挖掘的各种模型去拟合这些要素与总分数的关系，最终形成一个合适的模型，该模型比较有效地综合了专家打分的意见并且有效拟合 Offer 构成要素与总分数的关系。为了更加准确，在专家打分的基础上，还可以辅之以客户调研，从而对专家的打分和各要素的权重进行修正，最后在修正的基础上进行模型的搭建和拟合，这属于项目的技术细节，不是项目核心，故不做深入的讲解。

信息质量模型是电子商务和网上交易的基本保障，其主要目的是确保商品基本信息的优质和高效，让买家更容易全面、清楚、高效地了解商品的主要细节，让卖家更容易、更高效地展示自己的商品。无论是 C2C（如淘宝），还是 B2B（如阿里巴巴），抑或是 B2C（如当当网、凡客网），都可以用类似的方法去优化、提升自己的商品展示质量和效果，有效提升和保障交易的转化率。

3.7　服务保障模型

服务保障模型主要是站在为客户服务的角度来说的，出发点是为了让客户（平台的卖家）更好地做生意，达成更多的交易，我们（平台）应该为他们提供哪些有价值的服务去支持、保障卖家生意的发展，这里的服务方向就可以有很多的空间去想象了。比如，让卖家购买合适的增值产品，让卖家续费合适的增值产品、卖家商业信息的违禁过滤、卖家社区发帖的冷热判断等，凡是可以更好地武装卖家的，可以让卖家更好地服务买家的措施，无论是产品武装，还是宣传帮助，都属于服务保障的范畴，都是服务保障模型可以并且应该出力的方向。

针对服务保障模型的示例将会在随后的预测（响应、分类）模型里专门进行介绍，所以这里不展开讨论，但是对于服务保障环节，我们还是应该有一定的认识，无论从数据化运营的管理、客户关系管理，还是数据分析挖掘应用上，服务保障环节都是不能忽视的一个方面。

3.8 用户（买家、卖家）分层模型

用户（买家、卖家）分层模型也是数据化运营中常见的解决方案之一，它与数据化运营的本质是密切相关的。精细化运营必然会要求区别对待，而分层（分群）则是区别对待的基本形式。

分层模型是介于粗放运营与基于个体概率预测模型之间的一种折中和过渡模型，其既兼顾了（相对粗放经营而言比较）精细化的需要，又不需要（太多资源）投入到预测模型的搭建和维护中，因而在数据化运营的初期以及在战略层面的分析中，分层模型有着比较广泛的应用和较大的价值。

正如预测模型有特定的目标变量和模型应用场景一样，分层模型也有具体的分层目的和特定用途，这些具体的目的和用途就决定了分层模型的构建思路和评价依据。其常用的场景为：客户服务团队需要根据分层模型来针对不同的群体提供不同的说辞和相应的服务套餐；企业管理层需要基于在线交易卖家数量来形成以其为核心的卖家分层进化视图；运营团队需要通过客户分层模型来指导相应的运营方案的制订和执行，从而提高运营效率和付费转化率等。这些分层模型既可以为管理层、决策层提供基于特定目的的统一进化视图，又可以给业务部门做具体的数据化运营提供分群（分层）依据和参考。

分层模型常用的技术既包括统计分析技术（比如相关性分析、主成分分析等），又可以含有预测（响应、分类）模型的技术（比如通过搭建预测模型发现最重要的输入变量及其排序情况，然后根据这些变量对分层进行大致的划分，并通过实际数据进行验证），这要视具体的分析目的、业务背景和数据结构而定，同时要强调的是，一个好的分层模型的搭建一定是需要业务方的参与和贡献的，而且其中的业务逻辑和业务思考远远胜过分析技术本身。

下面我们分别用两个典型的案例来说明分层模型是如何搭建和应用的。

案例一：以交易卖家数量为核心的卖家分层进化视图

背景：某互联网公司作为买卖双方的交易平台，其最终的价值体现在买卖双方在该平台上达成交易（从而真正让买卖双方双赢，满意）。现在，管理层希望针对在线成交的卖家（群体）形成一个分层进化的视图。其基本目标就是，从免费注册的卖家开始，通过该视图可以粗略地、有代表性地勾画出卖家一步一步成长、进步乃至最终达成交易的全过程。这里

的每一层都是一个或几个有代表性的重要指标门槛，顺着不同的门槛逐步进化，越往上走，人群越少，越有可能成为有交易的卖家，而最后最高一层将是近 30 天来有交易的卖家。从这个背景和目标描述里，我们可以大致想象出这个分层模型是一个类似金字塔的形状（底部人数多，越往上越小，表示人群在减少）。

这个分层模型的主要价值体现在：可以让管理层、决策层对交易卖家的成长、进化、过滤的过程有个清晰、直观的把握，并且可以从中直观地了解影响卖家交易的一系列核心因素，以及相应的大致门槛阀值，也可以让具体的业务部门直观地了解“培养成交卖家，让卖家能在线成交”的主要因素，以及相应的运营抓手。

在本案例中，有必要了解一些关键的业务背景和业务因素，比如要想在线交易，卖家的 Offer 必须是“可在线交易 Offer”。这个条件很关键，所谓“可在线交易 Offer”是指该商品的 Offer 支持支付宝等第三方在线支付手段，如果卖家的 Offer 不支持这些手段，那就无法在线交易，也就无法满足本课题的目标了。所以，这里的“卖家 Offer 必须是可在线交易 Offer”是一个前期的重要门槛和阀值，从此也可以看出，对业务背景的了解非常重要，它决定了课题是否成功。

下面来谈谈具体的分析思路，先是从最基本的免费注册的卖家（即“全会员”）开始，之后是近 30 天有登录网站的卖家（说明是“活”的卖家，这里经过了直观的业务思考），再到近 1 年有新发或重发 Offer 的卖家，然后是当前有效 Offer 的卖家，最后是当前有可在线交易 Offer 的卖家，这个分析过程其实是第一部分的思考，它们构成了金字塔的下半部分，基本上是基于业务背景的了解和顺理成章的逻辑来“进化”的，之所以在“全会员”与“当前有可在线交易 Offer”之间安插了另外 3 层逐步“进化”的指标，主要也是基于业务方需要门槛的进度和细分的考虑，但这不是主要的核心点。

接下来，从“当前有可在线交易 Offer 的卖家”开始，层层进化到最高端的“近 30 天有在线交易的卖家”，也就是找出影响卖家成交的核心因素，并将之提炼成具体的层级和门槛，这一部分则是本案例的重点和核心所在。

如何找出其中的核心要素以及重要性的先后顺序？在本课题中，使用了预测（分类、响应）模型的方法，即通过搭建预测（响应）模型（目标变量是“近 30 天是否在线成交”，输入变量由数据分析团队与业务团队共同讨论确定），并通过多种模型算法的比较，最后找出决定交易的几个最重要的输入变量及先后次序。

最终的分层模型大致如图 3-3 所示，限于企业商业隐私的考虑，针对该数据做了处理，请勿对号入座。

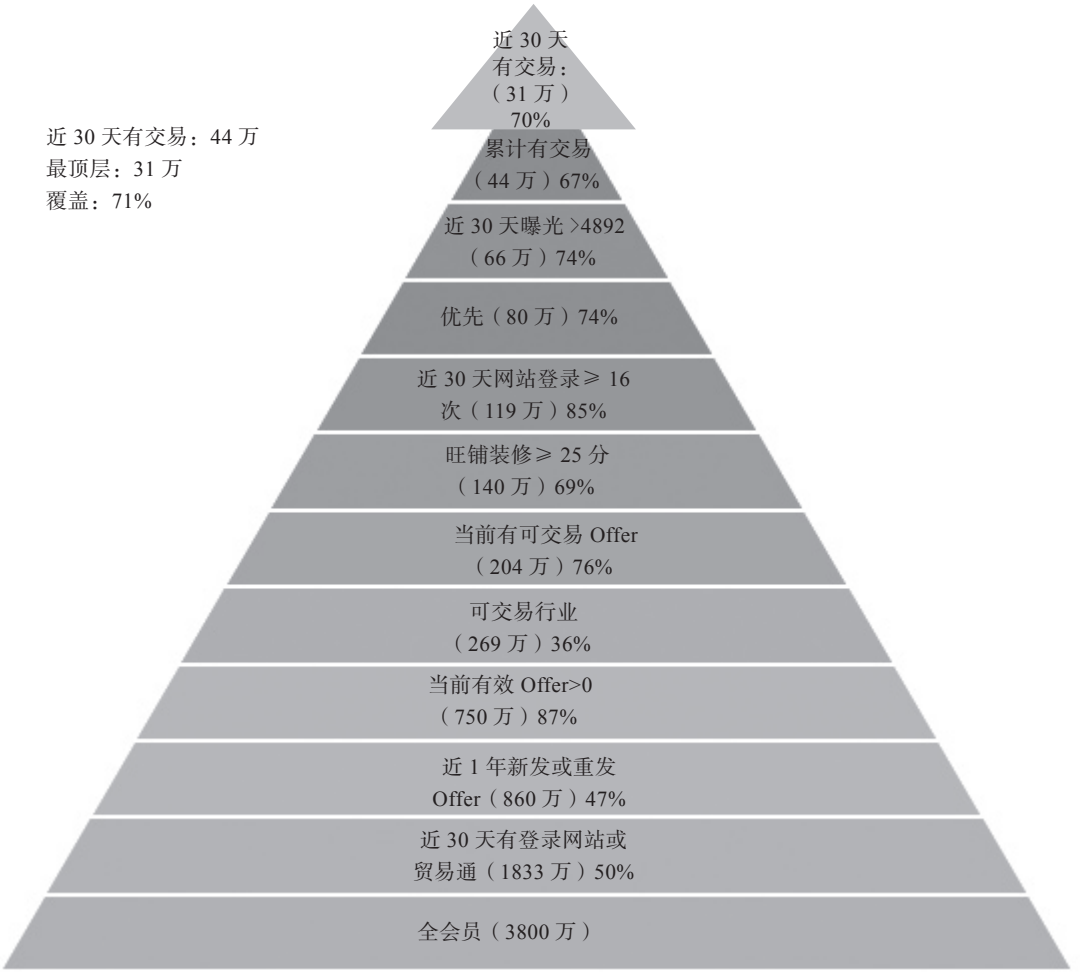


图 3-3　交易卖家分层示意图

该金字塔每一层里的数量代表满足该条件的会员（卖家）数量，而且各层之间的条件是连贯且兼容的，比如，从下往上数，第 6 层“当前有可交易 Offer”的用户有 204 万人，占其前一层“可交易行业卖家”269 万人的 76%，而且该层的用户必定是同时满足其下 5 层的所有条件的（包括来自可交易行业，当前有有效 Offer，近 1 年有新发或重发 Offer，近 30 天有登录网站或即时通信工具等）。

细心的读者可能会发现，最顶层的人数是 31 万，占近 30 天有交易卖家总数的 71%，为什么不能占近 30 天有交易卖家总数的 100%？这个差距正是由金字塔模型的本质所决定的，无论这个层层进化的金字塔模型多么完美，它还是无法完全圈定有交易卖家的总数，总是有一部分有交易的卖家不是满足上述金字塔上半部分的那些条件、门槛、阀值。这也是类似的分层模型只能

看大数、看主流的主要原因和特点，但是只要这个模型可以圈定大多数的人群（比如本项目实现的 71%，或者更高），那它就有相当的代表性，就可以作为相应的决策参考和业务参考。

当然，这个模型是否可以投入应用，还需要进一步检验，常规的检验方法就是通过不同时间段的数据，看是否有相似的规律、门槛、占比、漏斗，也就是看这个金字塔的结构是否具有一定时间长度的稳定性。在本项目中，我们通过前后各半年的数据分别进行了验证，发现这个金字塔的结果总体还是比较稳定的，确实可以作为决策参考和业务借鉴。

案例二：客户服务的分层模型

背景：A 产品是一个在线使用的付费产品，其主要功能就是让卖家实时获悉来自己网店的买家，可以让卖家通过主动对话促成双方的交谈，一旦对上话，卖家就可以得到由系统提供的买家联系方式等。很明显，该产品的核心功能（卖点）就是让卖家第一时间抓住来店铺的买家，并通过对话拿到买家的联系方式，方便后期的跟进，直至达成交易。现在该产品的客户服务团队正在负责付费用户的后期续费工作，该客服团队希望数据分析师帮他们制作一个付费用户的分层模型，在业务方的设想中该模型至少有 3 层，每一层可以对应相应的客服方案来帮助该层客户解决问题，模型的最终目的是促进付费客户的续费率稳步提升。具体来说，业务方希望根据业务敏感和客服资源储备，对付费用户进行 3 个群体的划分，每个群体有明确的业务诊断和客服方案（第一个群体，“体质差的客户群体”，比如访客数比较少，并且客户登录在线平台的次数也比较少（导致双方握手交谈可能性不高），这群客户被认为是最次要关注的；第二个群体，“问题客户群体”，比如对该产品的功能点使用都很少的客户，针对这群客户，客服团队可以对他们提供有针对性的产品功能教育；第三个群体，“生死线客户”，这群客户特点是有相对而言数量较多的访客，但是他们很少主动洽谈（以至无法拿到买家的联系方式，影响后期的成交），之所以称之为“生死线客户”，是因为客服团队希望作为重点关怀的群体，把他们从产品使用的“无效性”上拉回来，把他们从可能流失（续费）的生死线上拉回来（这群客户有理由从产品中获益（拿到买家联系方式），只是他们没有主动联系客户，如果他们能主动与买家洽谈，从而拿到联系方式，他们的成交业务有理由明显上升）。

该案例的分层模型用不上复杂的建模技术，只需要基于简单的统计技能就可实现。在深度把握产品价值和业务背景的前提下，我们与业务方一起基于他们设想的 3 个细分群体，根据实际数据找出了相应的具体阀值。具体来说，针对“体质差的客户群体”，基于访客数量和自身登录平台的天数和次数，进行两维数据透视，就可以找到满意的阀值和门槛定义；针对“问题客户群体”，只需要针对各功能点使用情况的 10 分位，找出最低的 20% ~ 30% 用户就可以了；针对“生死线客户群体”，同样是基于访客数量和自身主动洽谈的次数，进行两维数据透视，也可以找到满意的阀值和门槛定义，这样就能根据数据分布情况找到有很多访客，同时主动洽谈次数很少的客户群体。上述群体划分的方法主要是基于业务理解和客服

团队的资源配备的，事后的方案验证也表明，该种群体划分不仅能让业务方更容易产生理解和共鸣，也能很好地稳定并提升付费用户的续费率。

3.9 卖家（买家）交易模型

卖家（买家）交易模型的主要目的是为买卖双方服务，帮助卖家获得更多的买家反馈，促进卖家完成更多的交易、获得持续的商业利益，其中涉及主要的分析类型包括：自动匹配（预测）买家感兴趣的商品（即商品推荐模型）、交易漏斗分析（找出交易环节的流失漏斗，帮助提升交易效率）、买家细分（帮助提供个性化的商品和服务）、优化交易路径设计（提升买家消费体验）等。交易模型的很多分析类型其实已经在其他项目类型里出现过了，之所以把它们另外归入卖家（买家）交易模型的类型，主要是希望和读者一起换个角度（从促进交易的角度）来看待问题和项目。“横看成岭侧成峰”，同样的模型课题，其实有不同的主题应用场景和不一样的出发点，灵活、自如是一个合格的数据分析师应该具备的专业素养。

3.10 信用风险模型

这里的信用风险包括欺诈预警、纠纷预警、高危用户判断等。在互联网高度发达，互联网技术日新月异的今天，基于网络的信用风险管理显得尤其基础，尤其重要。

虽然目前信用风险已经作为一个独立的专题被越来越多的互联网企业所重视，并且有专门的数据分析团队和风控团队负责信用风险的分析和监控管理，但是从数据分析挖掘的角度来说，信用风险分析和模型的搭建跟常规的数据分析挖掘没有本质的区别，所采用的算法都是一样的，思路也是类似的。如果一定要找出这两者之间的区别，那就得从业务背景考虑了，从风险的业务背景来看，信用风险分析与模型相比于常规的数据分析挖掘有以下一些特点：

- 分析结论或者欺诈识别模型的时效更短，需要优化（更新）的频率更高。网络上骗子的行骗手法经常会变化，导致分析预警行骗欺诈的模型也要因此持续更新。
- 行骗手段的变化很大程度上是随机性的，所以这对欺诈预警模型的及时性和准确性提出了严重的挑战。
- 对根据预测模型提炼出的核心因子进行简单的规则梳理和罗列，这样就可在风控管理的初期阶段有效锁定潜在的目标群体。

3.11 商品推荐模型

鉴于商品推荐模型在互联网和电子商务领域已经成为一个独立的分析应用领域，并且正

在飞速发展并且得到了广泛应用。因此除本节以外，其他章节将不再对商品推荐模型做任何分析和探讨，至于本节，相对于其他的分析类型来说，会花费更多的笔墨和篇幅。希望能给读者提供足够的原理和案例[⊖]。

3.11.1　商品推荐介绍

电子商务推荐系统主要通过统计和数据挖掘技术，并根据用户在电子商务网站的行为，主动为用户提供推荐服务，从而来提高网站体验的。根据不同的商业需求，电子商务推荐系统需要满足不同的推荐粒度，主要以商品推荐为主，但是还有一些其他粒度推荐。譬如Query 推荐、商品类目推荐、商品标签推荐、店铺推荐等。目前，常用的商品推荐模型主要分为规则模型、协同过滤和基于内容的推荐模型。不同的推荐模型有不同的推荐算法，譬如对于规则模型，常用的算法有 Apriori 等；而协同过滤中则涉及 K 最近邻居算法、因子模型等。没有放之四海而皆准的算法，在不同的电子商务产品中，在不同的电子商务业务场景中，需要的算法也是不一样的。实际上，由于每种算法各有优缺点，因此往往需要混合多种算法，取长补短，从而提高算法的精准性。

3.11.2　关联规则

1. Apriori 算法

电子商务中常用的一种数据挖掘方法就是从用户交易数据集中寻找商品之间的关联规则。关联规则中常用的一种算法是 Apriori 算法。该算法主要包含两个步骤：首先找出数据集中所有的频繁项集，这些项集出现的频繁性要大于或等于最小支持度；然后根据频繁项集产生强关联规则，这些规则必须满足最小支持度和最小置信度。

上面提到了最小支持度和最小置信度，事实上，在关联规则中用于度量规则质量的两个主要指标即为支持度和置信度。那么，什么是支持度和置信度呢？接下来进行讲解。

给定关联规则 $X=>Y$，即根据 X 推出 Y。形式化定义为：

$$\text{支持度}(X=>Y)=\frac{\text{同时包含}X\text{和}Y\text{的记录数}}{\text{数据集记录总数}}$$

$$\text{置信度}(X=>Y)=\frac{\text{同时包含}X\text{和}Y\text{的记录数}}{\text{数据集中包含}X\text{的记录数}}$$

⊖　本节内容由淘宝网的商品推荐高级算法工程师陈凡负责编写，陈凡的微博地址为hppt://weibo.com/bicloud。

假设 D 表示交易数据集；K 为项集，即包含 k 个项的集合；L_k 表示满足最小支持度的 k 项集；C_k 表示候选 k 项集。Apriori 算法的参考文献[㊀]描述如下。

在该算法中，候选集的计算过程如下所示。

```
L1={ 满足最小支持度的 1 项集 }
for (k=2; Lk-1 ≠∅; k++)
   Ck=candicate_gen( Lk-1 )// 计算候选项集
   for all transactions t ∈ D do
          Ct=subset(Ck,t)// 候选集是否包含在 t 中
          for all candicates c ∈ Ct do
                 c.count++
end
   Lk={c ∈ Ck | c.count 大于等于最小支持度 }
end
合并所有的 Lk，得到频繁项集
```

首先进行连接运算如下：

```
insert into Ck
select p.item1, p.item2, p.itemk-1,…, q.itemk
from Lk-1 p, Lk-1 q
where p.item1=q.item1 and … and p.itemk-2=q.itemk-2 and p.itemk-1<q.itemk-1;
```

然后根据频繁项集定理（即频繁项集的子集必定是频繁项集）进行剪枝，过滤掉非频繁项集，过程如下所示：

```
forall itemset c ∈ Ck
   forall (k-1) 子集 s of c do
          if (s ∉ Lk-1 ) then
   delete c from Ck
```

从上述算法中可以看出，该算法存在一些困难点，譬如需要频繁扫描交易数据集，这样如果面临海量数据集，就难以满足实际应用需求；对于大型数据集，计算候选集算法的效率较低，这也是一个难以克服的问题。目前已经有一些优化的方法用于处理这些问题，譬如 FP-growth 算法[㊁]。在实际应用中，随着数据的不断增长，可能还需要通过分布式计算来提高算法性能，譬如机器学习算法包 Mahout [㊂]中实现了的并行版本 FP-growth 算法。

2. Apriori 算法实例

假设给定如下电子商务网站的用户交易数据集，其中，定义最小支持度为 2/9，即支持度计数为 2，最小置信度为 70%，现在要计算该数据集的关联规则，如表 3-1 所示。

㊀ Rakesh Agrawal , Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the 20th International Conference on Very Large Data Bases, p.487-499, September 12-15, 1994

㊁ Jiawei Han , Jian Pei , Yiwen Yin, Mining frequent patterns without candidate generation, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p.1-12, May 15-18, 2000, Dallas, Texas, United States

㊂ Mahout，http://mahout.apache.org/

表 3-1　用户交易数据集

交易标识	购买商品列表
2001	I1,I2,I5
2002	I2,I4
2003	I2,I3
2004	I1,I2,I4
2005	I1,I3
2006	I2,I3
2007	I1,I3
2008	I1,I2,I3,I5
2009	I1,I2,I3

计算步骤如下所示。

步骤 1，根据 Apriori 算法计算频繁项集。

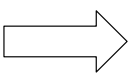
1）计算频繁 1 项集。扫描交易数据集，统计每种商品出现的次数，选取大于或等于最小支持度的商品，得到了候选项集，如表 3-2 所示。

表 3-2　频繁 1 项集

商品集	包含该商品集的记录数
I1	6
I2	7
I3	6
I4	2
I5	2

2）根据频繁 1 项集，计算频繁 2 项集。首先将频繁 1 项集和频繁 1 项集进行连接运算，得到 2 项集，如下所示：

商品集
I1
I2
I3
I4
I5



商品集
I1,I2
I1,I3
I1,I4
I1,I5
I2,I3
I2,I4
I2,I5
I3,I4
I3,I5
I4,I5

扫描用户交易数据集，计算包含每个候选 2 项集的记录数，如表 3-3 所示。

表 3-3 候选 2 项集

商品集	包含该商品集的记录数
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

根据最小支持度，得到频繁 2 项集，如表 3-4 所示。

表 3-4 频繁 2 项集

商品集	包含该商品集的记录数
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2

3）根据频繁 2 项集，计算频繁 3 项集。首先将频繁 2 项集进行连接，得到 {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}，然后根据频繁项集定理进行剪枝，即频繁项集的非空子集必须是频繁的，{I1, I2, I3} 的 2 项子集为 {I1,I2}，{I1,I3}，{I2,I3}，都在频繁 2 项集中，则保留；

{I1, I2, I5} 的 2 项子集为 {I1,I2}，{I1,I5}，{I2,I5}，都在频繁 2 项集中，则保留；

{I1, I3, I5} 的 2 项子集为 {I1,I3}，{I1,I5}，{I3,I5}，由于 {I3,I5} 不是频繁 2 项集，移除该候选集；

{I2, I3, I4} 的 2 项子集为 {I2,I3}，{I2,I4}，{I3,I4}，由于 {I3,I4} 不是频繁 2 项集，移除该候选集；

{I2, I3, I5} 的 2 项子集为 {I2,I3}，{I2,I5}，{I3,I5}，由于 {I3,I5} 不是频繁 2 项集，移除该候选集；

{I2, I4, I5} 的 2 项子集为 {I2,I4}，{I2,I5}，{I4,I5}，由于 {I4,I5} 不是频繁 2 项集，移除该候选集。通过剪枝，得到候选集 {{I1, I2, I3}, {I1, I2, I5}}，扫描交易数据库，计算包含候选 3 项集的记录数，得到表 3-5。

表 3-5　频繁 3 项集

商品集	包含该商品集的记录数
I1, I2, I3	2
I1, I2, I5	2

4）根据频繁 3 项集，计算频繁 4 项集。重复上述的思路，得到 {I1,I2,I3,I5}，根据频繁项集定理，它的子集 { I2,I3,I5} 为非频繁项集，所以移除该候选集。从而，频繁 4 项集为空，至此，计算频繁项集的步骤结束。

步骤 2，根据频繁项集，计算关联规则。

这里以频繁 3 项集 {I1, I2, I5} 为例，计算关联规则。{I1, I2, I5} 的非空子集为 {I1,I2}、{I1,I5}、{I2,I5}、{I1}、{I2} 和 {I5}。

规则 1，{I1,I2}=>{I5}，置信度为 {I1, I2, I5} 的支持度除以 {I1,I2} 的支持度，即 2/4=50%，因其小于最小置信度，所以删除该规则。

同理，最后可以得到 {I1,I5}=>{I2}，{I2,I5}=>{I1} 和 {I5}=>{I1,I2} 为 3 条强关联规则。

然而，在实际应用 Apriori 算法时，需要根据不同的粒度，譬如类目、商品等，结合不同的维度（浏览行为，购买行为等）进行考虑，从而构建符合业务需求的关联规则模型。在电子商务应用中，关联规则算法适用于交叉销售的场景。譬如，有人要出行（飞往北京），根据计算出的关联规则（如：机票 => 酒店）来考虑，那么，可以根据用户购买的机票，为用户推荐合适的北京酒店；再比如，在情人节，根据关联规则，将巧克力和玫瑰花进行捆绑销售等。

另外，关联规则还可以用来开发个性化电子商务推荐系统的 Top N 推荐。首先，根据用户的交易数据，计算用户在特定时序内购买过的商品；然后，根据关联规则算法，计算满足最小支持度和最小置信度的商品关联规则；再根据用户已经购买的商品和商品关联规则模型，预测用户感兴趣的商品，同时过滤掉用户已经购买过的商品，对于其他的商品，则按照置信度进行排序，从而为用户产生商品推荐。

3.11.3　协同过滤算法

协同过滤是迄今为止最成功的推荐系统技术，被应用在很多成功的推荐系统中。电子商

务推荐系统可根据其他用户的评论信息，采用协同过滤技术给目标用户推荐商品。协同过滤算法主要分为基于启发式和基于模型式两种。其中，基于启发式的协同过滤算法，又可以分为基于用户的协同过滤算法和基于项目的协同过滤算法。启发式协同过滤算法主要包含 3 个步骤：1）收集用户偏好信息；2）寻找相似的商品或者用户；3）产生推荐。

“巧妇难为无米之炊”，协同过滤的输入数据集主要是用户评论数据集或者行为数据集。这些数据集主要又分为显性数据和隐性数据两种类型。其中，显性数据主要是用户打分数据，譬如用户对商品的打分，如图 3-4 所示。

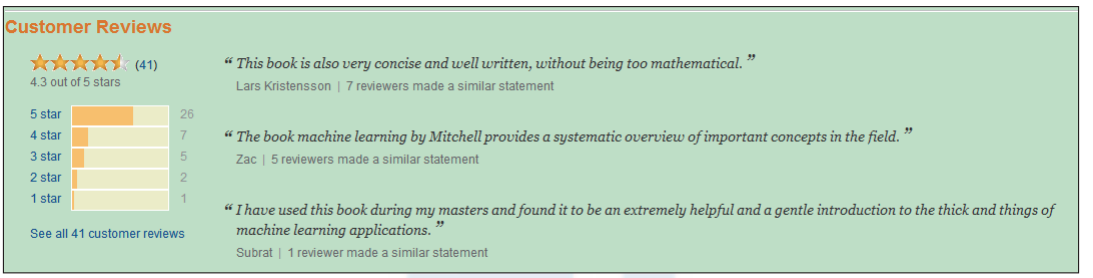


图 3-4　某电商网站用户对某商品的评分结果

但是，显性数据存在一定的问题，譬如用户很少参与评论，从而造成显性打分数据较为稀疏；用户可能存在欺诈嫌疑或者仅给定了部分信息；用户一旦评分，就不会去更新用户评分分值等。

而隐性数据主要是指用户点击行为、购买行为和搜索行为等，这些数据隐性地揭示了用户对商品的喜好，如图 3-5 所示。

隐性数据也存在一定的问题，譬如如何识别用户是为自己购买商品，还是作为礼物赠送给朋友等。

图 3-5　某用户最近在某电商网站的浏览商品记录（左侧的 3 本书）

1. 基于用户的协同过滤

基于用户（User-Based）的协同过滤算法首先要根据用户历史行为信息，寻找与新用户相似的其他用户；同时，根据这些相似用户对其他项的评价信息预测当前新用户可能喜欢的项。给定用户评分数据矩阵 R，基于用户的协同过滤算法需要定义相似度函数 $s:U\times U\rightarrow R$，以计算用户之间的相似度，然后根据评分数据和相似矩阵计算推荐结果。

在协同过滤中，一个重要的环节就是如何选择合适的相似度计算方法，常用的两种相似度计算方法包括皮尔逊相关系数和余弦相似度等。皮尔逊相关系数的计算公式如下所示：

$$s(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r}_u)(r_{v,i}-\overline{r}_v)}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r}_u)^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-\overline{r}_v)^2}}$$

其中，i 表示项，例如商品；I_u 表示用户 u 评价的项集；I_v 表示用户 v 评价的项集；$r_{u,i}$ 表示用户 u 对项 i 的评分；$r_{v,i}$ 表示用户 v 对项 i 的评分；$\overline{r}_u$ 表示用户 u 的平均评分；$\overline{r}_v$ 表示用户 v 的平均评分。

另外，余弦相似度的计算公式如下所示：

$$s(u,v)=\frac{r_u\cdot r_v}{\|r_u\|_2\|r_v\|_2}=\frac{\sum_i r_{u,i}r_{v,i}}{\sqrt{\sum_i r_{u,i}^2}\sqrt{\sum_i r_{v,i}^2}}$$

另一个重要的环节就是计算用户 u 对未评分商品的预测分值。首先根据上一步中的相似度计算，寻找用户 u 的邻居集 $N\in U$, 其中 N 表示邻居集，U 表示用户集。然后，结合用户评分数据集，预测用户 u 对项 i 的评分，计算公式如下所示：

$$p_{u,i}=\overline{r}_u+\frac{\sum_{u'\in N}s(u,u')(r_{u',i}-\overline{r}_{u'})}{\sum_{u'\in N}|s(u,u')|}$$

其中，$s(u,u')$ 表示用户 u 和用户 u' 的相似度。

假设有如下电子商务评分数据集，预测用户 C 对商品 4 的评分，见表 3-6。

表 3-6　电商网站用户评分数据集

用户	商品 1	商品 2	商品 3	商品 4
用户 A	4	?	3	5
用户 B	?	5	4	?
用户 C	5	4	2	?
用户 D	2	4	?	3
用户 E	3	4	5	?

表中？表示评分未知。根据基于用户的协同过滤算法步骤，计算用户 C 对商品 4 的评

分，其步骤如下所示。

（1）寻找用户 C 的邻居

从数据集中可以发现，只有用户 A 和用户 D 对商品 4 评过分，因此候选邻居只有 2 个，分别为用户 A 和用户 D。用户 A 的平均评分为 4，用户 C 的平均评分为 3.667，用户 D 的平均评分为 3。根据皮尔逊相关系数公式来看，用户 C 和用户 A 的相似度为：

$$s(C,A)=\frac{(5-3.667)(4-4)+(2-3.667)(3-4)}{\sqrt{(5-3.667)^2+(2-3.667)^2}\times\sqrt{(4-4)^2+(3-4)^2}}=0.781$$

同理，$s(C, D)=-0.515$。

（2）预测用户 C 对商品 4 的评分

根据上述评分预测公式，计算用户 C 对商品 4 的评分，如下所示：

$$p_{C,4}=3.667+\frac{0.781\times(5-4)+(-0.515)\times(2-3)}{0.781+0.515}=4.667$$

依此类推，可以计算出其他未知的评分。

2. 基于项目的协同过滤

基于项目（Item-Based）的协同过滤算法是常见的另一种算法。与 User-Based 协同过滤算法不一样的是，Item-Based 协同过滤算法计算 Item 之间的相似度，从而预测用户评分。也就是说该算法可以预先计算 Item 之间的相似度，这样就可提高性能。Item-Based 协同过滤算法是通过用户评分数据和计算的 Item 相似度矩阵，从而对目标 Item 进行预测的。

和 User-Based 协同过滤算法类似，需要先计算 Item 之间的相似度。并且，计算相似度的方法也可以采用皮尔逊关系系数或者余弦相似度，这里给出一种电子商务系统常用的相似度计算方法，即基于条件概率计算 Item 之间的相似度，计算公式如下所示：

$$s(i,j)=\frac{\text{freq}(i\cap j)}{\text{freq}(i)\cdot\text{freq}(j)^{\alpha}}$$

其中，$s(i, j)$ 表示项 i 和 j 之间的相似度；$\text{freq}(i\cap j)$ 表示 i 和 j 共同出现的频率；$\text{freq}(i)$ 表示 i 出现的频率；$\text{freq}(j)$ 表示 j 出现的频率；α 表示阻力因子，主要用于平衡控制流行和热门的 Item，譬如电子商务中的热销商品等。

接下来，根据上述计算的 Item 之间的相似度矩阵，结合用户的评分，预测未知评分。预测公式如下所示：

$$p_{u,i}=\frac{\sum_{j\in S}s(i,j)r_{u,j}}{\sum_{j\in S}|s(i,j)|}$$

其中，$p_{u,i}$ 表示用户 u 对项 i 的预测评分；S 表示和项 i 相似的项集；$s(i,j)$ 表示项 i 和 j 之间的相似度；$r_{u,j}$ 表示用户 u 对项 j 的评分。

3. Item-Based 协同过滤实例

在电子商务推荐系统中，商品相似度计算有着很重要的作用。它既可用于一些特定推荐场景，譬如直接根据当前的商品，为用户推荐相似度最高的 Top N 商品。同时，它还可以应用于个性化推荐，从而为用户推荐商品。电子商务网站收集了大量的用户日志，譬如用户点击日志等。

基于 Item-Based 协同过滤算法，笔者提出了一种增量式商品相似度的计算解决方案。该算法计算流程如图 3-6 所示。

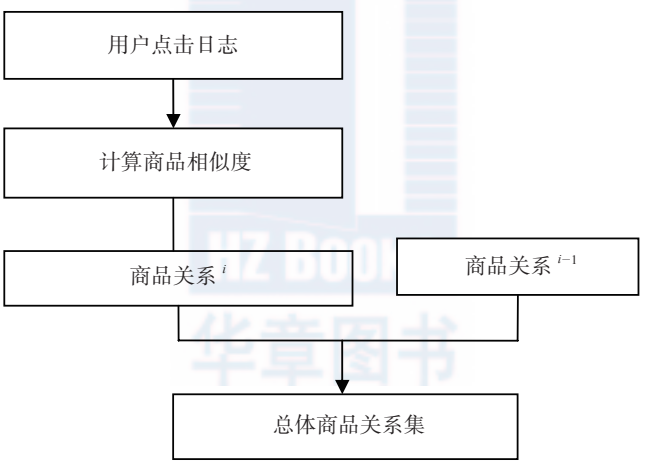


图 3-6　增量式商品相似度计算流程图

其中，商品关系 ^{i} 表示第 i 天的商品关系数据集。

具体计算步骤如下。

1）获取当天用户点击行为数据，过滤掉一些噪声数据，譬如商品信息缺失等。从而得到用户会话 sessionID、商品 ID（商品标识）、浏览时间等信息，如表 3-7 所示。

由于 A4 的浏览时间和 A1、A2、A3 相差较大，因此将其过滤掉，这里定义为 1800 秒，如表 3-8 所示。

表 3-7 用户点击行为日志表

用户会话 ID	浏览商品的时间	Item Pairs
100	A1, 20:12	A1, A2 A1, A3
	A2, 20:13	A2,A1 A2, A3
	A3, 20:15	A3,A1 A3, A2
	A4, 23:30	

表 3-8 过滤后的用户点击行为日志表

浏览商品的时间	Item Pairs
A1, 20:12	A1, A2 A1, A3
A2, 20:13	A2,A1 A2, A3
A3, 20:15	A3,A1 A3, A2

2）首先，计算任意两种商品之间的共同点击次数。然后，根据基于条件概率的商品相似度计算方法来计算商品的相似度。商品相似度公式如下。

$$s(i,j)\frac{\text{freq}(i\cap j)}{\text{freq}(i)\cdot\text{freq}(j)}$$

其中，$s(i,j)$ 表示项 i 和 j 之间的相似度；$\text{freq}(i\cap j)$ 表示 i 和 j 共同出现的频率；$\text{freq}(i)$ 表示 i 出现的频率；$\text{freq}(j)$ 表示 j 出现的频率。

3）合并前一天计算的商品相似度数据，进行投票判断，选择相似度较大的作为新的商品相似度，从而实现增量式商品相似度计算。

3.11.4 商品推荐模型总结

对于商品推荐模型，除了上述介绍的基于关联规则和基于协同过滤的算法外，还有其他一些常用的算法，譬如基于内容的算法，即根据商品标题、类目和属性等信息，计算商品之间的关系，然后结合用户行为特征，为用户提供商品推荐。商品推荐模型面临着许多重要问题，譬如特征提取问题，即如何从商品标题、类目和属性中提取商品的重要特征、新用户问题，即如何解决用户行为较少，提升推荐质量、新商品问题，即如何处理长尾商品问题，让更多的商品有推荐展现的机会、稀疏性问题，即对于庞大的用户和商品数据，用户评分数据往往会显得非常稀疏等。面对这些问题，在实际应用中，需要根据业务场景，充分利用各种算法的优点，从而设计出混合推荐算法，以便提升推荐质量。

3.12　数据产品

数据产品是指数据分析师为了响应数据化运营的号召，提高企业全员数据化运营的效率，以及提升企业全员使用数据、分析数据的能力而设计和开发的一系列有关数据分析应用的工具。有了这些数据产品工具，企业的非数据分析人员也能有效地进行一些特定的数据分析工作。因此可以这样理解，数据产品就是自动化、产品化了数据分析师的一部分常规工作，让系统部分取代数据分析师的劳动。

其实，我们每个人在日常生活中或多或少都使用过各种各样的数据产品，有的是收费的，有的是免费的。最常见的免费数据产品，就是我们登录自己的网上银行，来查看自己在过去任何时间段的账户交易明细。如果你有在当当网上的购物体验，那么对当当网账户里的操作应该比较熟悉，如图 3-7 所示，用户可以在“我的收藏”页面针对自己的所有收藏商品进行有效的管理，这也是一种免费的数据产品。

图 3-7　“我的收藏”页面

当然了，上面列举的这些产品更多的是方便用户进行个人财务、商品管理的，还不是专门针对用户进行数据分析支持的。下面这个例子，如图 3-8 所示则是跟数据分析功能相关的数据产品，量子恒道作为淘宝网的一个免费数据产品，可以帮助网商自我进行精准实时的数据统计、多维数据分析，从而为网商交易提供更强的数据驱动力。

图 3-8　量子恒道的分析展示

3.13　决策支持

决策支持是现代企业管理中大家耳熟能详的词汇。数据分析挖掘所承担的决策支持主要是指通过数据分析结论、数据模型对管理层的管理、决策提供响应和支持，从而帮助决策层提高决策水平和质量。

对于现代企业和事业单位的管理层来说，数据分析的决策支持一部分是通过计算机应用系统自动实现的，这部分就是所谓的决策支持系统（Decision Support System，DSS），最常见的输出物就是企业层面的核心日报、周报等。每天会由计算机应用系统自动生成这些报表，供管理层决策参考，另一部分是非常规的、特定的分析内容，包括特定的专题分析、专题调研等。

无论是报表还是专题分析，对于数据分析师来说，所涉及的承担决策支持的工作与支持业务部门的数据分析，在技术和方法上并没有本质的区别和差异。但是在以下方面会有一定的差别：

- 决策支持的数据分析工作要求数据分析师站在更高的角度，用更宽的视野进行数据分析。由于是供企业决策层参考的，所以数据分析师要站在企业全景、市场竞争的全局来考虑分析思路和结论。
- 服务的对象不同。这似乎是废话，但是在数据分析挖掘实践中，这的确也是数据分析师不能回避的问题。在实践中，因为是为决策层服务的，所以对分析的时间要求常会更严格，项目的优先级也会更高，而且对结论的准确性和精确性的要求也会相对比较苛刻。