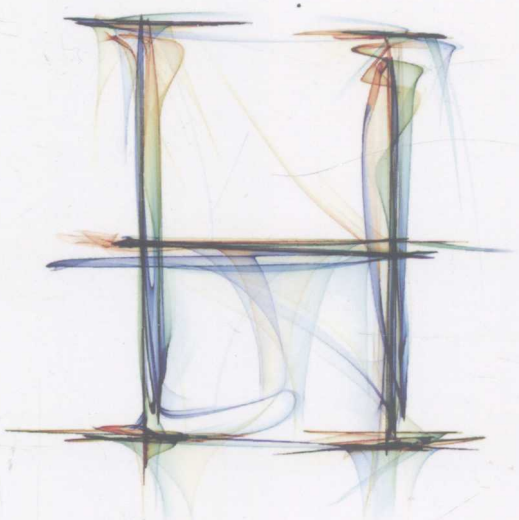
HZ BOOKS
华章IT

10余位大数据领域资深专家和科研人员，结合10余年大数据挖掘咨询与实施经验，手把手教你从海量数据中淘金。

从大数据挖掘的应用出发，以电力、航空、医疗、互联网、制造、电信等行业真实案例为主线，详细讲解了数据挖掘建模的过程和数据挖掘的二次开发。

大数据技术丛书



Hadoop Practice of Big Data Analysis and Mining

Hadoop大数据分析与挖掘实战

张良均 樊哲 赵云龙 李成华◎等著

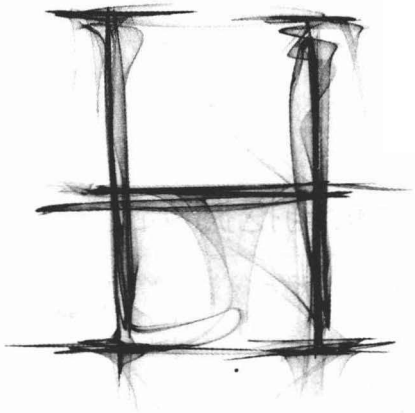
机械工业出版社
China Machine Press

大数据技术丛书

Hadoop Practice of Big Data Analysis and Mining

Hadoop大数据分析与挖掘实战

张良均 樊 哲 赵云龙 李成华 刘丽君 刘名军 肖 刚 著
云伟标 王 路 刘晓勇 薛 云 廖晓霞 徐英刚



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Hadoop 大数据分析与挖掘实战 / 张良均等著 . —北京：机械工业出版社，2015.12
（大数据技术丛书）

ISBN 978-7-111-52265-2

I. H… II. 张… III. 数据处理软件 IV. TP274

中国版本图书馆 CIP 数据核字（2015）第 281852 号

Hadoop 大数据分析与挖掘实战

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：高婧雅 责任校对：董纪丽
印 刷：中国电影出版社印刷厂 版 次：2016 年 1 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：19
书 号：ISBN 978-7-111-52265-2 定 价：69.00 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

Preface 前　言

为什么要写这本书

到2012年为止，Farecast系统用了将近十万亿条价格记录来帮助预测美国国内航班的票价。Farecast票价预测的准确度已经高达75%，使用Farecast票价预测工具购买机票的旅客，平均每张机票可节省50美元[⊖]。

Farecast是大数据公司的一个缩影，也代表了当今世界发展的趋势。但与国外相比，我国由于信息化程度不太高，企业内部信息不完整，零售业、银行、保险、证券等对大数据分析与挖掘的应用并不太理想。但随着市场竞争的加剧，各行业对大数据分析与挖掘技术的研究与应用意愿越来越强烈，可以预计，未来几年，各行业的数据分析一定都是大规模的数据挖掘与应用。在大数据时代，数据过剩、人才短缺，数据挖掘专业人才的培养又需要专业知识和职业经验积累。所以，本书在注重大数据时代数据挖掘理论的同时，也注意与大数据项目案例实践相结合，这样可以让读者体验真实的大数据挖掘学习与实践环境，更快、更好地学习大数据分析与挖掘知识以及积累职业经验。

总地来说，随着大数据时代的来临，大数据分析与挖掘技术将具有越来越重要的战略意义。大数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产要素，人们对于海量数据的运用将预示着新一轮生产率增长和消费者盈余浪潮的到来。大数据分析与挖掘技术将帮助企业用户在合理时间内攫取、管理、处理、整理海量数据，为企业经营决策提供积极的帮助。大数据分析与挖掘作为数据存储和挖掘分析的前沿技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。虽然大数据目前在国内还处于初级阶段，但是其商业价值已经显现出来，特别是有实践经验的大数据分析人才更是各企业争夺的热门资源。

大数据时代来临，风云变化，时不我待！

⊖ 维克托·迈尔·舍恩伯格. 大数据时代－生活、工作与思维的大变革. 2012

本书特色

本书作者从实践出发，结合大量大数据挖掘工程案例及教学经验，以真实案例为主线，深入浅出介绍大数据挖掘项目中针对数据分析的各个流程：数据探索、数据预处理、分类与预测、聚类分析、关联规则挖掘、智能推荐等。因此，图书的编排以解决某个大数据应用的挖掘目标为前提，先介绍案例背景提出挖掘目标，再阐述针对这个目标使用的大数据挖掘分析方法与过程，最后完成模型构建，在介绍建模过程中会针对每个大数据项目的特点进行分析，同时提供上机实验，把相关的建模操作提供给读者。在本书的高级篇中，介绍大数据挖掘的二次开发实例，方便有能力的读者进行相关二次开发。

根据读者对案例的理解，本书配套提供了真实的原始样本数据文件及建模仿真平台，读者可以从“泰迪杯”全国大学生数据挖掘竞赛网站（http://www.tipdm.org/ts/655.jhtml）免费下载。另外，为方便教师授课需要，图书还特意提供了建模阶段的过程数据文件、PPT 课件，读者可通过“勘误与支持”中的联系方式咨询或者获取文件。

本书适用对象

❑ 开设有大数据挖掘课程的高校教师和学生。

目前国内不少高校将数据挖掘引入本科教学中，在数学、计算机、自动化、电子信息、金融等专业开设了数据挖掘技术相关的课程，但目前这一课程的教学仍然主要限于理论介绍。因为单纯的理论教学过于抽象，学生理解起来往往比较困难，教学效果也不甚理想。本书提供的基于实战案例和建模实践的教学内容，能够使师生充分发挥互动性和创造性，理论联系实际，使师生获得最佳的教学效果。

❑ 大数据挖掘开发人员。

这类人员可以在理解大数据挖掘应用需求和设计方案的基础上，结合图书提供的基于第三方接口快速实现大数据挖掘应用的编程。

❑ 需求分析及系统设计人员。

这类人员可以在理解数据挖掘原理及建模过程的基础上，结合数据挖掘案例完成精确营销、客户分群、交叉销售、流失分析、客户信用记分、欺诈发现、智能推荐等数据挖掘应用的需求分析和设计。

❑ 进行大数据挖掘应用研究的科研人员。

许多科研院所为了更好地对科研工作进行管理，纷纷开发了适应自身特点的科研业务管理系统，并在使用过程中积累了大量的科研信息数据。但是，这些科研业务管理系统一般没有对这些数据进行深入分析，对数据所隐藏的价值并没有进行充分挖掘利用。科研人员需要大数据挖掘建模工具及有关方法论来深挖科研信息的价值，从而提高科研水平。

❑ 关注大数据分析的人员。

业务报告和商业智能解决方案对于了解过去和现在的状况可能是非常有用的。但是，数据挖掘的预测分析解决方案还能使这类人员预见未来的发展状况，让他们的机构能够先发制人，而不是处于被动。因为数据挖掘的预测分析解决方案将复杂的统计方法和机器学习技术应用到数据之中，通过预测分析技术来揭示隐藏在交易系统或企业资源计划（ERP）、结构数据库和普通文件中的模式和趋势，从而为决策提供科学依据。

如何阅读本书

本书共16章，分三个部分：基础篇、实战篇、高级篇。基础篇介绍了数据挖掘、Hadoop大数据的基本原理，实战篇通过对案例深入浅出的剖析，使读者在不知不觉中通过案例实践获得大数据项目挖掘分析经验，同时快速领悟看似难懂的大数据分析与挖掘理论知识。读者在阅读过程中，应充分利用随书配套的案例建模数据，借助TipDM-HB大数据挖掘建模平台，通过上机实验，快速理解相关知识与理论。

第一部分是基础篇（第1~6章），第1章的主要内容是数据挖掘概述、大数据餐饮行业应用；第2章针对大数据理论知识进行基础讲解，简明扼要地针对Hadoop安装、原理等做了介绍；第3章介绍了大数据仓库Hive的安装、原理等内容；第4章介绍了大数据数据库HBase的安装、原理等内容；第5章介绍了几种大数据挖掘建模平台，同时重点介绍了本书使用的开源TipDM-HB大数据挖掘平台；第6章介绍数据挖掘的建模过程、各种挖掘建模的常用算法与原理以及挖掘建模在大数据挖掘算法库Mahout中的实现原理。

第二部分是实战篇（第7~14章），重点分析大数据挖掘技术在法律咨询、电子商务、航空、移动通信、互联网、生产制造以及公共服务等行业的应用。在案例结构组织上，按照先介绍案例背景与挖掘目标，再阐述大数据时代针对大数据的分析方法与过程，最后完成模型构建的顺序进行的，详细分析了建模过程关键环节。最后通过上机实践，加深对大数据挖掘案例的认识以及分析流程。

第三部分是高级篇（第15~16章），介绍了基于Hadoop大数据开发的相关技术以及开发步骤，并使用实例来展示这些步骤，使读者可以自己动手实践，亲自体会开发的乐趣；还介绍了基于TipDM-HB大数据挖掘平台的二次开发实例，借助TipDM-HB大数据挖掘平台二次开发工具，可以更加快捷、高效地完成相关大数据应用的二次开发，降低开发难度，使读者更方便地体会到大数据分析与挖掘的强大魅力。

勘误和支持

除封面署名外，参加本书编写工作的还有刘名军、肖刚、云伟标、王路、刘晓勇、薛云、廖晓霞、徐英刚等。由于笔者的水平有限，编写时间仓促，书中难免会出现一些错误或者不

准确的地方，恳请读者批评指正。为此，读者可通过笔者微信公众号 TipDM（微信号：TipDataMining）、TipDM 官网（www. tipdm. com）反馈有关问题。也可通过热线电话（40068-40020）或企业 QQ（40068-40020）进行在线咨询或通过扫描以下微信公众号的二维码咨询获取。

读者可以将书中的错误及遇到的任何问题反馈给我们，我们将尽量在线上为读者提供最满意的解答。图书的全部建模数据文件及源程序，可以从全国大学生数据挖掘竞赛网站（www. tipdm. org）下载，我们会将相应内容的更新及时发布更正出来。如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@ qq. com，期待能够得到您的真挚反馈。

致谢

在本书编写过程中，得到了广大企事业单位科研人员的大力支持！在此谨向中国电力科学研究院、广东电力科学研究院、广西电力科学研究院、华南师范大学、广东工业大学、广东技术师范学院、南京中医药大学、华南理工大学、湖南师范大学、韩山师范学院、中山大学、广州泰迪智能科技有限公司、武汉泰迪智慧科技有限公司等单位给予支持的专家及师生致以深深的谢意。

在本书的编辑和出版过程中还得到了参与“泰迪杯”全国大学生数据挖掘建模竞赛（http://www. tipdm. org）的众多师生及机械工业出版社杨福川、高婧雅等无私的帮助与支持，在此一并表示感谢。

张良均

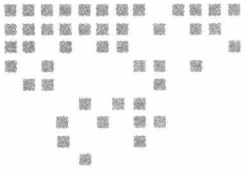
Contents 目　录

实 战 篇

高　级　篇

基 础 篇

- 第 1 章　数据挖掘基础
- 第 2 章　Hadoop基础
- 第 3 章　Hadoop生态系统：Hive
- 第 4 章　Hadoop生态系统：HBase
- 第 5 章　大数据挖掘建模平台
- 第 6 章　挖掘建模



Chapter 1 第 1 章

数据挖掘基础

1.1 某知名连锁餐饮企业的困惑

国内某餐饮连锁有限公司（以下简称 T 餐饮）成立于 1998 年，主要经营粤菜，兼顾湘菜、川菜、中餐等综合菜系。至今已经发展成为在国内具有一定知名度、美誉度，多品牌、立体化的大型餐饮连锁企业。属下员工 1000 多人，拥有 16 家直营分店，经营总面积近 13 000 平方米，年营业额近亿元。其旗下各分店均坐落在繁华市区主干道，雅致的装潢，配之以精致的饰品、灯具、器物，出品精美，服务规范。

近年来餐饮行业面临较为复杂的市场环境，与其他行业一样，餐饮企业都遇到了原材料成本升高、人力成本升高、房租成本升高等问题，这也使得整个行业的利润率急剧下降。人力成本和房租成本的上升是必然趋势，如何在保持产品质量的同时提高企业效率，成为了 T 餐饮急需面对的问题。从 2000 年开始，T 餐饮通过加强信息化管理来提高效率，目前已上线的管理系统包括以下几个。

（1）客户关系管理系统

该系统详细记录了每位客人的喜好，为顾客提供个性化服务，满足客户个性化需求。通过客户关怀，提高客户的忠诚度。比如企业能随时查询了解今天哪位客人过生日或其他纪念日，根据客人的价值分类进行相应关怀，如送鲜花、生日蛋糕、寿面等。通过本系统，还可对客户行为进行深入分析，包括客户价值分析、新客户分析与发展，并根据其价值情况提供给管理者，为企业提供决策支持。

（2）前厅管理系统

该系统通过掌上电脑无线点菜方式，改变了传统“饭店点菜、下单、结账一支笔、一张

纸，服务员来回跑的局面”，快速完成点菜过程。通过厨房自动送达信息，服务员的写菜速度加快不需要再通过手写，同时传菜部也轻松不少，菜单会通过电脑自动打印出来，差错率降低，也不存在厨房人员看不懂服务员字迹而搞错的问题。

（3）后厨管理系统

信息化技术可实现后厨与前厅沟通无障碍，客人菜单瞬间传到厨房。服务员只需点击掌上电脑的发送键，客人的菜单即被传送到收银管理系统中，由系统的电脑发出指令，设在厨房等处的打印机立即打印出相应的菜单，厨师按单做菜。与此同时，收银台也打印出一张同样的菜单放在客人桌上，以备客人查询以及作结账凭据，使客人明明白白消费。

（4）财务管理系统

该系统完成销售统计、销售分析、财务审计，实现对日常经营销售的管理。通过报表，企业管理者很容易掌握前台的销售情况，从而达到对财务的控制。通过表格和图形可以显示餐厅的销售情况，如菜品排行榜、日客户流量、日销售收入分析等；统计每天的出菜情况，可以了解哪些是滞销菜，哪些是畅销菜，从而了解顾客的品位，有针对性地制订出一套既适合餐饮企业发展又能迎合顾客品位的菜肴体系和定价策略。

（5）物资管理系统

该系统主要完成对物资的进销存，实际上就是一套融采购管理（入库、供应商管理、账款管理）、销售（通过配菜卡与前台销售联动）、盘存为一体的物流管理系统。对于连锁企业，还涉及统一配送管理等。

通过以上信息化的建设，T 餐饮已经积累了大量的历史数据，有没有一种方法可帮助企业从这些数据中洞察商机，提取价值？在同质化的市场竞争中，怎样找到一些市场以前并不存在的“捡漏”和“补缺”？

1.2　从餐饮服务到数据挖掘

企业经营最大的目的就是盈利，而餐饮业企业盈利的核心就是其菜品和顾客，也就是其提供的产品和服务对象。企业经营者每天都在想推出什么样的菜系和种类会吸引更多的顾客，究竟各种顾客各自的喜好是什么，在不同的时段是不是有不同的菜品畅销，当把几种不同的菜品组合在一起推出时是不是能够得到更好的效果，未来一段时间菜品原材料应该采购多少……

T 餐饮的经营者想尽快地解决这些疑问，使自己的企业更加符合现有顾客的口味，吸引更多的新顾客，又能根据不同的情况和环境转换自己的经营策略。T 餐饮在经营过程中，通过分析历史数据，总结出一些行之有效的经验：

- 在点餐过程中，由有经验的服务员根据顾客特点进行菜品推荐，一方面可提高菜品的销量，另外一方面可减少客户点餐的时间和频率，提高用户体验；
- 根据菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行预测，以便餐饮企业提前准备原材料；

- 定期对菜品销售情况进行统计，分类统计出好评菜和差评菜，为促销活动和新菜品推出提供支持；
- 根据就餐频率和金额对顾客的就餐行为进行评分，筛选出优质客户，定期回访和送去关怀。

上述措施的实施都依赖于企业已有业务系统中保存的数据，但是目前从这些数据中获得有关产品和客户的特点以及能够产生价值的规律更多依赖于管理人员的个人经验。如果有一套工具或系统，能够从业务数据中自动或半自动地发现相关的知识和解决方案，这将极大地提高企业的决策水平和竞争能力。这种从数据中“淘金”，从大量数据（包括文本）中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程，这就是**数据挖掘**；它是利用各种分析工具在大量数据中寻找其规律和发现模型与数据之间关系的过程，是统计学、数据库技术和人工智能技术的综合。

这种分析方法可避免“人治”的随意性，避免企业管理仅依赖个人领导力的风险和不确定性，实现精细化营销与经营管理。

1.3 数据挖掘的基本任务

数据挖掘的基本任务包括利用分类与预测、聚类分析、关联规则、时序模式、偏差检测、智能推荐等方法，帮助企业提取数据中蕴含的商业价值，提高企业的竞争力。

对餐饮企业而言，数据挖掘的基本任务是从餐饮企业采集各类菜品销量、成本单价、会员消费、促销活动等内部数据，以及天气、节假日、竞争对手以及周边商业氛围等外部数据；之后利用数据分析手段，实现菜品智能推荐、促销效果分析、客户价值分析、新店选点优化、热销/滞销菜品分析和销量趋势预测；最后将这些分析结果推送给餐饮企业管理者及有关服务人员，为餐饮企业降低运营成本，增加盈利能力，实现精准营销，策划促销活动等提供智能服务支持。

1.4 数据挖掘建模过程

从本节开始，将以餐饮行业的数据挖掘应用为例来详细介绍数据挖掘的建模过程，如图 1-1 所示。

1.4.1 定义挖掘目标

针对具体的数据挖掘应用需求，首先要明确本次的挖掘目标是什么？系统完成后能达到什么样的效果？因此必须分析应用领域，包括应用中的各种知识和应用目标，了解相关领域的有关情况，熟悉背景知识，弄清用户需求。要想充分发挥数据挖掘的价值，必须要对目标