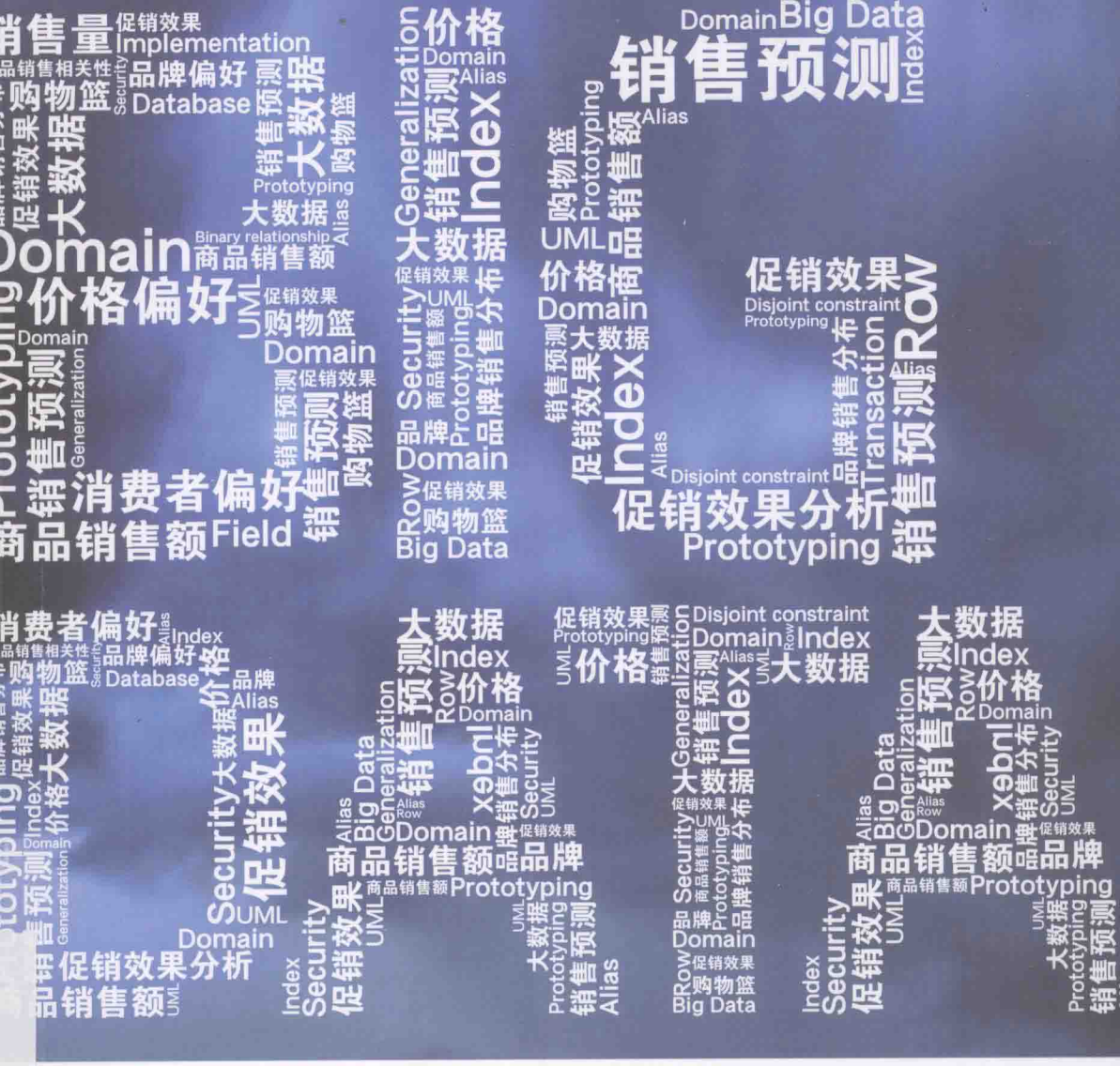

数海淘金

Analysis Practice on Sales Data of a Supermarket Chain

连锁超市销售数据分析实务

陈绛平 / 等著　靳雨涵 / 主审

本书对商品销售额、销售量、商品销售相关性、品牌销售分布、品牌偏好、价格偏好、商品规格偏好、价格弹性系数等进行分析的方法、过程及商业意义，并且在分析方法上有一定的创新。

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

数海淘金

Analysis Practice on Sales Data of a Supermarket Chain

连锁超市销售数据分析实务

陈绛平 / 等著　靳雨涵 / 主审

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

数海淘金：连锁超市销售数据分析实务 / 陈绛平等著. —杭州：浙江大学出版社，2014.4
ISBN 978-7-308-09979-0

Ⅰ.①数… Ⅱ.①陈… Ⅲ.①连锁超市－销售量－统计数据－分析 Ⅳ.①F717.6

中国版本图书馆 CIP 数据核字（2012）第 097679 号

数海淘金——连锁超市销售数据分析实务
陈绛平 等著
靳雨涵 主审 冯胜男 参与审校

责任编辑 周卫群
封面设计 续设计
出版发行 浙江大学出版社
（杭州市天目山路 148 号 邮政编码 310007）
（网址：http://www.zjupress.com）
排 版 杭州中大图文设计有限公司
印 刷 富阳市育才印刷有限公司
开 本 710mm×1000mm 1/16
印 张 11.5
字 数 207 千
版 印 次 2014 年 4 月第 1 版 2014 年 4 月第 1 次印刷
书 号 ISBN 978-7-308-09979-0
定 价 35.00 元

《数海淘金--连锁超市销售数据分析实务》研究和创作组成员

组　　长 陈绛平

副 组 长 靳雨涵

组　　员 张　宸、杨晓贤、吴　丰、陈爽婕、龚　群、项江帆、陈丽华、傅媛雁、董　理、马　晨、赵　健、郑琼雅、俞荣钧、陈　敏、何利明、夏昭天、严飞雪、林加辉、李欣欣、陈　晟、夏　婧、吴慧婷、唐　超、李　操、葛森森、徐　瑜、张峰铭、颜兴挺、吴丹君、吴　彬、冯彦婷、史姗姗、宋　薇、王晨晖、王　烁、郑丽丹、吴志龙、陈浩然、何李广、李修国、胡佳明、杜亚卡、王　迪、童刚刚、张振林、章　镇、罗治州、徐逸士、张冰雨、周群旻、卢　峥、杨秋实、叶　庆、石志豪、沈赓贤、陈金辉、金弘毅、张　帅、朱之博、徐敬业、邱　侠等

主　　审 靳雨涵

参与审校 冯胜男

序

《数海淘金——连锁超市销售数据分析实务》是浙江大学城市学院的学术团队较长时间潜心治学和开展科学研究的成果。

传统数据研究侧重于“白盒研究”，而对通用性较强的“黑盒模型”和普适规律系统研究较为缺乏。因此，面对当今社会产生的大量数据，就需要运用系统论的观点，对数量巨大的数据进行统计性的搜索、比较、聚类等手段以挖掘其内在规律。

“大数据”这一概念虽在20世纪80年代初提出，但对其真正进行系统研究始于21世纪初。随着社会经济的发展，其本身的丰富性和复杂性在为该领域的研究带来一定的挑战，同时也提供了新的发展空间。人们通过有效地组织和使用数据，进一步提升大数据研究的效率效益和时效性，这将对相关产业的升级与新产业诞生的研究产生巨大的推动效应。

《数海淘金——连锁超市销售数据分析实务》是以“购物篮分析”为切入点，介绍如何分析问题并运用相关工具对实际问题进行计算和分析结果，对零售行业的数据库营销进行研究。该书从零售行业数据库营销角度出发，以消费组合和消费偏好为主要研究对象，详细阐述了数据库营销分析问题的思路，说明了数据分析的过程，并对数据分析结果进行评估评价。在分析方法上，采用了许多数学分析工具，并提出了一些独到分析方法和研究思路。这些成果对超市等连锁零售行业的“大数据”营销研究作出了新的贡献，部分成果填补了国内的空白。

《数海淘金——连锁超市销售数据分析实务》是我院探索科研与教学相结合、教师科研和指导学生开展研究创新活动相结合的结晶。这些研究成果既可以作为营销理论研究者参考之用，也可以成为零售企业进行营销数据分析和商品管理的指南。相信这本书的出版能够对“大数据”营销的研究和应用起到非常好的促进作用。

我有幸成为较早阅读本书的读者，我将我的阅读体会作为本书的序，希望读者们也能够得到许多新的收获。

郑健壮　教授

2013年12月24日于杭州

前言

Preface

本书以系统学科思想为指导，以应用统计学、消费者行为学、数据挖掘技术、市场营销学和复杂网络等学科的理论、知识为研究基础，以零售行业数据库为对象，利用数据分析等工具，深层次挖掘消费者偏好、特征，掌握其内在规律，从而进行关系维护和销售预测。

“大数据”这一概念虽在20世纪80年代初提出，但是对其真正研究始于21世纪初。其社会、经济、科研价值以及本身的丰富性和复杂性在为该领域的研究带来一定挑战的同时，也提供了一定发展空间。倘若人们能够更有效地组织和使用数据，解决巨量数据所带来的数据技术处理瓶颈、大数据研究的效率效益和时效性等问题，对于相关产业的升级与新产业诞生将会有巨大的推动。

在知识经济发展和社会信息变革的大背景下，大数据已吸引了越来越多的关注。然而，由于其自身的数据体量巨大(Volume)、类型繁多(Variety)、价值密度低、商业价值高(Value)、处理速度快、时效性要求高(Velocity)等特点，增加了现实生活中大数据研究的复杂性，对人类的数据驾驭能力提出了新的挑战。

90年代，沃尔玛集团曾经对大量超市购物小票数据进行研究，分析其中的规律，得出了一个非常著名的结论：那就是购物篮中的啤酒与尿不湿表面上不相干，但是数据分析的结果却是它们具有非常大的相关性。这个研究成果发表以后，世界各国纷纷开展了对购物篮的研究。现代零售企业也日益重视对销售数据的分析和成果运用。大型零售企业一般都设有品类部，以开展相应的数据分析和商品管理工作。

本书聚焦于“购物篮分析”，通过对连锁超市销售数据库的研究，从海量数据分析着手把脉消费者特征，挖掘其中的潜在规律，从中发现具有商业价值的信息。

本书以消费组合和消费偏好为主要研究对象，详细阐述了数据库营销的分析问题的思路，说明了数据分析的过程，并对数据分析结果进行评估评价。主要内容包括研究背景、研究方式、工具介绍、数据仓库的建立过程、购物篮商品组合

分析、消费者偏好分析、价格弹性分析、促销效果分析、时间序列预测分析、商品季节性分析、品牌分布规律研究等。

通过本书介绍的案例分析，读者可以运用相关技能，结合需要研究的现实问题，完成数据库营销研究过程。

本书主要包括以下内容：

第一章是本书的基础理论部分，包括研究背景、研究方式、工具介绍等内容。

第二章介绍了数据仓库的建立过程，涉及具体数据库结构的分析，相关数据质量的检测与纠错，在此基础上描述了数据处理过程。

第三章提出购物篮分析和商品组合，旨在研究如何通过有效的数量方法，来发现各种商品在消费者购物篮中同时出现的规律，即各种商品之间的相关度。

第四章通过对购物篮数据进一步的分析，揭示不同消费者在购买过程中的各种偏好，并发掘出除了一般的口味、价格偏好以外的包装规格等偏好。

第五章侧重价格弹性和通过价格促销的效果分析，从产品的销售额最大化出发，制定出一个产品定价模型，为超市促销提供合理定价依据。

第六章通过时间序列分析，结合商品季节性销售特点，对商品进行销售预测。

第七章从品牌角度对牙膏、饼干、纸制品等商品进行销售额分析，提出销售额分布的幂次法则，商品的规模和其相应名次之间存在着幂次方的反比关系。

第八章介绍了运用 Apriori 算法对整个数据仓库的数据进行关联规则参数的计算过程，并对结果进行了列表分析。

第九章从实务操作角度对数据库营销进行展望，并对未来进一步开展实时或者准实时的数据分析工作提出建议。

本书在写作过程中，努力做到具有实用性、创新性和可操作性。实用性方面，能够结合当今知识经济和信息化研究背景，进行数据挖掘，实现由理论到实务分析。创新性方面，本书与现在热议的“大数据”研究相结合，并从零售行业角度对海量数据进行分析，得出规律。可操作性方面，本书以图文结合的方式，方便读者更直观地了解相关概念和工具。

本书由陈绛平总撰稿。参与本书数据分析课题研究和撰写的还有浙江大学城市学院信息管理、工商管理等专业的教师和学生，主要有靳雨涵、张宸、杨晓贤、吴丰、陈爽婕、龚群、项江帆、陈丽华、傅媛雁、董理、马晨、赵健、郑琼雅、俞荥钧、陈敏、何利明、夏昭天、严飞雪、林加辉、李欣欣、陈晟、夏婧、吴慧婷、唐超、李操、葛森森、徐瑜、张峰铭、颜兴挺、吴丹君、吴彬、冯彦婷、史姗姗、宋薇、王晨晖、王烁、郑丽丹、吴志龙、陈浩然、何李广、李修国、胡佳明、杜亚卡、王迪、童刚刚、张

振林、章镇、罗治州、徐逸士、张冰雨、周群旻、卢峥、杨秋实、叶庆、石志豪 沈赓贤、陈金辉、金弘毅、张帅、朱之博、徐敬业、邱侠等。这些研究成果有些已经发表在学术期刊上，有些发表在国际学术会议上，受到了各方的关注和好评。本书由靳雨涵负责主审，冯胜男帮助进行了部分修订工作。

业内专家叶耀庭、董伟、鲍慧群、倪煜珍等为本书的研究工作提供了大力支持和热情指导。虞镇国、范晓清、蔡颖、叶华平、叶志洪、项江帆等同事为本书的研究工作提供了许多帮助。郑健壮教授在百忙之中为本书题写了序。谨向他们表示衷心的感谢。

浙江大学城市学院企业管理学科（浙江省重点学科）为本书提供了部分资助，浙江大学城市学院商务信息分析实验室（杭州市重点实验室）为本书的研究提供了技术条件，在此表示衷心的感谢。

我们非常希望这些成果能够为零售行业的经营管理者所用，进而提高企业的经营效益，能够使顾客感到更加满意。同时，对于消费品的生产者也有一定的借鉴意义。也希望能够给有关机构的市场营销研究工作以及大专院校的教学科研提供参考作用。

希望本书能够给您带来启发和收获。

作　者

2013 年 12 月于杭州

目 录

Contents

第一章　数海淘金的由来

第二次世界大战以后，市场营销环境发生了深刻的变化，营销理念和营销方法也在不断地演进和创新。从市场环境来看，绝大多数商品已经从卖方市场变成了买方市场，产品越来越丰富，竞争不断加剧，企业面临着前所未有的挑战。营销者已经普遍从思想和实践两方面经历了从传统的销售观念到现代市场营销观念的转变。通过对消费者行为的分析，了解顾客需求的动向和特征，为消费者提供满足他们所需要的产品和服务，提高他们的消费体验满意度，成为营销成功的关键手段之一。

近年来，随着计算机及网络技术的普及，我们得以获得大量的营销数据。同时，借助于这些先进的软硬件技术，一些新的营销观念有了实践的机会。其中对销售数据进行深层次的挖掘，找出其中潜在的规律，进而在营销要素设计和营销实践过程中运用所发现的规律，能够发觉顾客的隐性需求和消费习惯，发现潜在的顾客群，获得更多的市场机会，更好地满足顾客的需要，以扩大和强化市场优势，最终实现企业经营效益的提高。我们称之为“数据库营销”。由于数据量非常庞大，动辄以“千万”、“亿”计数，而且信息量及其丰富，现在形象地称其为“大数据营销”。

数据库营销广泛应用于许多行业，在欧美已经得到了广泛的应用。首先，数据库营销可以用来搜集、整理顾客的数据资料，构建顾客数据库。包括顾客个人资料、交易记录以及网站访问记录、意见反馈等信息。厂商可以从中选择适当的消费者，有针对性地进行沟通或者推广，提高营销宣传率，增加销量。同时由于选取了部分对象，因而可以大大降低营销成本。

比如发布直邮、Email、短信广告，如果对受众群体不加区别滥发，则不仅成本高昂、效率低下，还容易引起消费者的反感。而如果对消费者进行细分研究，有针对性地发布信息，则不仅广告的成功率高，而且负面反应也小很多。

数据库营销可以使企业细分各种不同人群的消费特点，实现准确定位。据百度百科报道，目前美国已有56%的企业正在建立数据库，85%的企业认为他们需要数据库营销来加强竞争力。由于运用消费者数据库能够准确找出某种产品的目标消费者，企业就可以避免使用昂贵的大众传播媒体，可以运用更经济的促销方式，从而降低成本，增强企业的竞争力。据有关资料统计，运用数据库技术进行筛选消费者，其邮寄宣传品的反馈率，是没有运用数据库技术进行筛选而发送邮寄宣传品的反馈率的10倍以上。

很多读者可能都会遇到过“被”数据库营销的体验。如汽车保险到期前的一段时间，车主经常会接到推销车险的电话。这是最简单的运用车主数据库进行车辆保险营销推广的例子。

还有，小孩家长也会经常接到电话，推销针对孩子所在年龄层的教育培训项目。这是利用人口数据库进行营销的例子。

稍微复杂一点的情况也可能会碰到。如在银行办理业务时，银行柜员有时候会向顾客推销基金、理财产品等等。其实这过程的背后并不简单。银行的数据分析部门运用数据挖掘等技术，对客户数据进行了大量的分析，摸索出某些客户群对某类产品需求较高的规律。在银行柜员进行业务操作时，如果这位顾客属于某项业务的潜在顾客群时，电脑系统会有提示信息。于是，柜员就会根据电脑的提示有针对性地向顾客推荐金融产品。这是数据库营销比较高级的应用。

通过对业务数据库的深入分析和有效运用，降低了企业的经营成本，提高了企业的运营效率、经营效益和市场竞争能力。这就是“数海淘金”一词的含义。

在零售行业，通常把数据库营销聚焦于“购物篮分析”，就是通过对销售数据库中购物小票数据的分析，发现潜在的规律，并运用到营销实践活动中。购物篮分析中最重要的部分就是商品相关性分析，商品相关性分析就是要找出购物篮中商品之间的关联度分析，即利用关联规则发现购物篮隐藏的相互关系。最好的一个例子就是20世纪90年代沃尔玛超市的“啤酒与尿布”的案例。他们通过对数据库中大量销售数据的分析，发现啤酒与婴儿尿布经常同时出现在一个购物篮中，进一步调查得知这是年轻的父亲购买婴儿用品时顺便给自己带上一些啤酒。于是，沃尔玛就对商品的布置进行了调整，实现了比较好的效果。

超市销售数据的分析对于深入了解企业经营状况，探索消费者的偏好与消费倾向，进而研究如何改进商品促销方案效果，提高超市经营业绩是非常有帮助的。借助于当今先进的数据仓库和数据挖掘软件技术，还可以深入挖掘商品销售数据中潜在的规律，分析出商品之间的关联效应，以改进商品陈列布局，调整销售策略。

本书主要以某连锁超市6年共2亿多条销售记录为依据进行分析，基于市场营销学、应用统计学和人工智能（数据挖掘技术）等学科，描述了对该超市集团的销售记录进行数据清理、纠错并建立数据仓库的过程，详细介绍了运用多种数据挖掘和统计分析的方法，对商品销售额、销售量、商品销售相关性、品牌销售分布、品牌偏好、价格偏好、商品规格偏好、促销效果、销售预测、价格弹性系数等进行分析的方法和过程。由于数据量庞大，因此在分析过程中，并不是所有分析项目都采用了全部的数据。在此说明一下，以免读者困惑。

我们从多个角度对这些数据进行了长达3年的分析，从中发现了许多有趣的现象意义，其中有很多非常有价值的商业研究结果。

第二章　建立数据仓库

数据分析的前提之一是必须有大量的实时数据，其二是这些数据必须真实可靠。由于技术局限和人为因素，在超市的销售数据中不免有差错、遗漏、不匹配等各种现象存在，超市管理系统中原有的数据库结构也不尽合理，需要我们在数据仓库建设过程中一一加以解决。

我们采用 MS SQL SERVER 中的 Microsoft Business Intelligence Development Studio（以下称 BI）工具来完成 ETL（Extract，Transfer，Load）过程，包括了数据检测、清洗和数据仓库建立的过程。

一、数据库结构分析

要进行数据分析，首先必须了解系统和数据库的结构。

（一）系统结构分析

该超市集团公司总部信息中心设有数据库服务器，为总部运营提供服务。除便利店以外的各家门店都设有独立的数据库服务器，单独建库，基础数据全备份。便利店由于点多店小，合并成一个数据库，服务器设在总部信息中心。

这个模式的好处是各门店超市管理系统运营基本独立，相互之间互不干扰，也不用通过网络将大量数据传递到总部进行处理。缺点是经营数据各自为政，难以进行全面、深入、细致的分析。

（二）数据库结构分析

各门店数据库（包括合并的便利店）数据结构完全一样，内有 1000 多个表，基础数据完全相同，经营数据各自独立。

与销售经营相关的表有：BUY1（购物小票汇总表），BUY2（购物小票明细表），GOODS（商品信息表），BRAND（品牌名称表），SORT（商品分类表），STORE（门店信息表）等。

由于各个门店数据库中许多数据重复，占用了大量的磁盘空间，因此在建立数据仓库时必须考虑只提取销售分析时必要的数据。

(三)表结构分析

BUY1 表中，有流水号、收银机号、收银员编号、顾客购物品种数、购物小票总金额以及付款时间等信息。

BUY2 表中，有流水号、不同品种商品的结账顺序号、商品内部代号、商品进价、商品牌价、购买数量、实付金额及优惠金额等信息。

BUY1 和 BUY2 之间是一对一或者一对多的关系，通过流水号和收银机号进行关联。值得注意的是，BUY1 的流水号有大量重复的现象。不仅仅是不同的门店编号各自独立，有重复，同一门店中不同的收银机之间的编号也是独立进行的，也有大量的重复编号。

还有一个问题，由于各个门店独立建库，因此 BUY1 和 BUY2 中都没有门店信息的字段，也就是说与 STORE 表无法关联。

GOODS 表信息最丰富，有品名、产地、等级、分类号和品牌编号等。BUY2 表和 GOODS 表之间用商品内部编号来关联。

BRAND 中除了品牌名称，就是一个与 GOODS 表关联的品牌编号。

SORT 表与 GOODS 表通过品牌编号来关联，表中商品依次按照大类、中类、小类区分，但是这三个分类号全部在一个字段中，难以按照不同的分类层次进行查询。

数据库关系见图 2.1。

(四)数据结构调整

综合以上信息，我们发现从信息质量的角度来看，数据库结构需要进行一定的调整，以满足建立数据仓库和进行数据分析的需要。

1. BUY1 和 BUY2 表改造

首先，在各门店数据库的 BUY1 表和 BUY2 表中添加三个新的字段：新的流水号、门店编号和购物日期。

新的流水号由原流水号加上店号和收银台号组成，这样可以确保新流水号在 BUY1 中是唯一的。

2. SORT 表改造

将在 SORT 表中增加大类编号、大类名称、中类编号、中类名称、小类编号和小类名称等 6 个字段，将原有的分类编号和分类名称字段中的信息分别提取