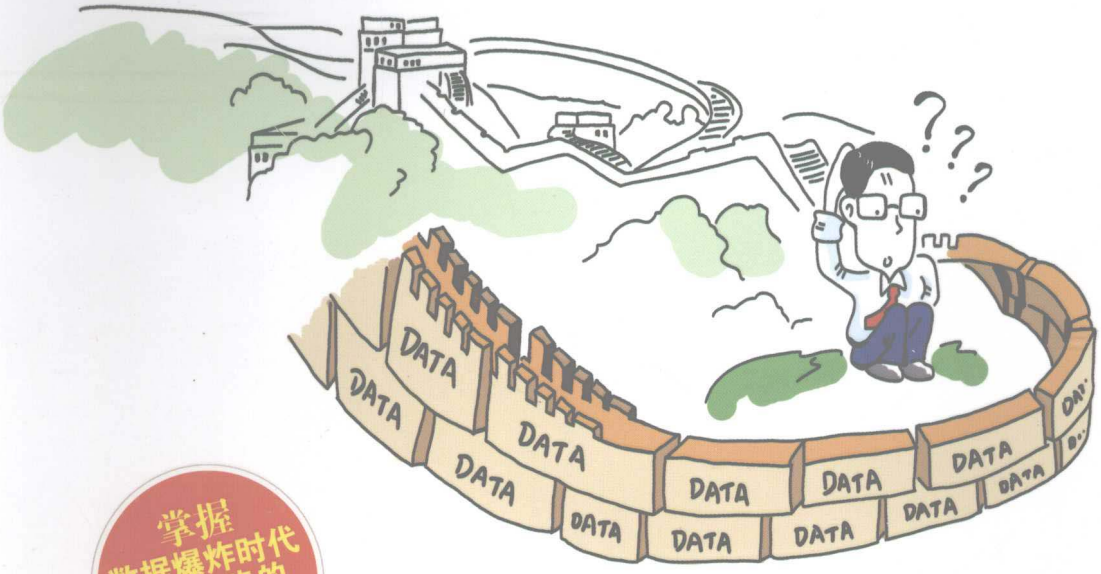
“Your track will be continued via this”

让世界更清晰

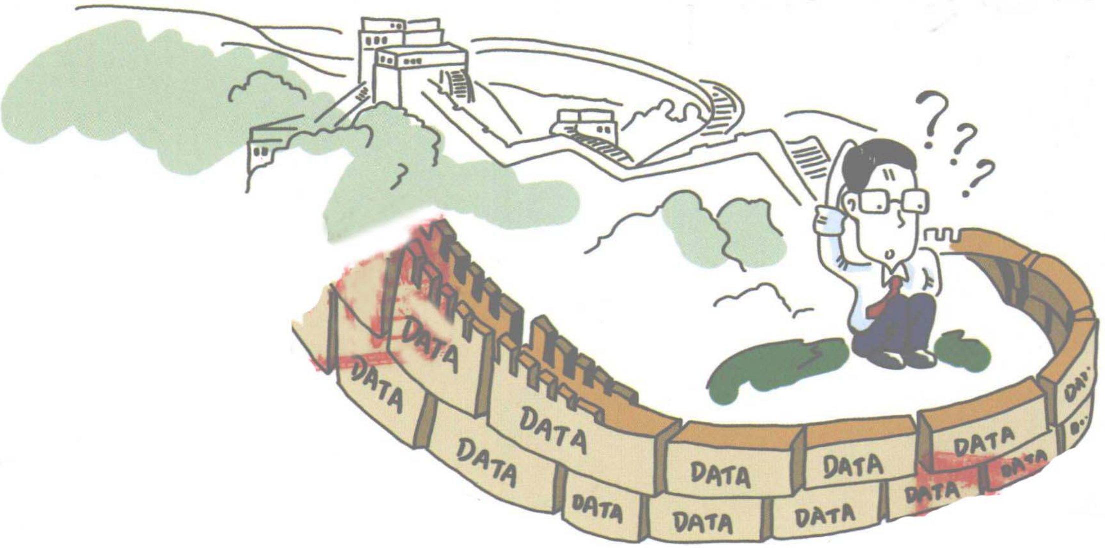


掌握
数据爆炸时代
先人一步的
新思维

大话数据挖掘

西安美林电子有限责任公司 编著

清华大学出版社



大话数据挖掘

西安美林电子有限责任公司 编著

清华大学出版社
北京

内 容 简 介

本书以 EMBA 班的“数据挖掘技术及其应用”教学为场景，带领读者步入数据挖掘的神秘殿堂，领略数据挖掘的神奇魅力。全书分为 9 章：第 1 章从三个真实故事开始数据挖掘之旅；第 2 章以某企业生产中遇到的质量控制难题的解决过程为线索，展现数据挖掘的实施过程；第 3 章到第 9 章以典型案例的形式分别介绍了数据挖掘技术在电力行业、交通航空领域、冶金行业、税务与金融行业、电信行业、故障诊断以及互联网行业的应用。

数据挖掘是一种专业性极强的技术，本书避开大量晦涩的概念和令人生畏的数学公式，以师生互动讨论的形式让读者走进数据挖掘殿堂，进而深入浅出、循序渐进地感知数据挖掘。随着阅读，读者会自然而然地身临课堂，“让数据说话，从数据中发现规律，科学决策”等新的理念会使读者对实际工作中面临的复杂问题浮想联翩、另辟新径。

本书适合企事业部门的领导、管理人员、生产一线的技术人员，另外，学生或者行业工作者，可以通过本书的阅读，为以后的学习奠定好基础。

图书在版编目(CIP)数据

大话数据挖掘 / 西安美林电子有限责任公司编著. —北京：清华大学出版社，2012
ISBN 978-7-302-29813-7
Ⅰ.①大… Ⅱ. ①西… Ⅲ. ①数据采集 Ⅳ. ①TP274
中国版本图书馆 CIP 数据核字(2012)第 190096 号

责任编辑：栾大成
封面设计：杨玉芳
责任校对：徐俊伟
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　**印　　张**：17.75　**插　　页**：1　**字　　数**：343 千字
版　　次：2013 年 1 月第 1 版　　**印　　次**：2013 年 1 月第2 次印刷
印　　数：5001～8000
定　　价：39.00 元

产品编号：048928-01

前　言

本书的萌发

上世纪80年代末到90年代初，国内外广泛流传着一句耐人寻味的话语：**我们沉浸在数据的海洋中，却渴望着知识的淡水。**这句话生动地描绘了当时人们面对海量数据的迷惘和无奈。就在这时，世界商业巨头沃尔玛从其庞大的交易数据库中演绎了一场“啤酒和尿布的故事”，揭示了一条隐藏在海量数据中的、美国人的一种行为规律：年龄在25～35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%～40%的人顺手为自己买几瓶啤酒。受这条简单的客户行为模式的启发，沃尔玛调整了商品布局，并策划了促销价格，结果销售量大增。这一现象引起了科学界的注意，他们将“啤酒和尿布的故事”引申为“关联规则获取”，进而将**“从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又潜在有用的信息和知识的过程”**定义为**“数据挖掘”**。

需求是成功之源，于是西方发达国家刮起了一场数据挖掘的风暴。商业界发现了沃尔玛迅猛发展的秘诀，纷纷效仿。电信行业也沸腾了，各公司纷纷争先恐后地利用数据挖掘这一锐利武器解决他们面临的最紧迫的问题（如客户分群、客户会流失原因及预测、业务套餐及响应、关联消费等）。工业界也行动了，他们从堆积如山的数据中，挖掘出指导生产和管理的决策规则。

上世纪90年代中期以后，基于数理统计、人工智能、机器学习、人工神经网络等多种技术的数据挖掘技术已经成为研究和应用的热点，数据挖掘在我国也开始推广应用。然而，从这么多年的情况来看，我国数据挖掘的应用与发达国家还有很大差距。我们仅在互联网、金融、电信和商业等领域有一些成功的应用，而在其他行业如制造、航空、医药、军工、化工、税务、反恐和刑侦等只有少量的尝试。为什么会这样呢？IT界、企业界和学术界的有识之士无不在思考着这样的问题。进行数据挖掘，数据是基础，难道是我国的信息化建设还未达到一定的程度，数据积累不够？

进入21世纪前可以这么说，可现在，显然不是。目前，我国的大中型企业，大多建立了先进的信息化系统，甚至相当多的企业构建了数据仓库，而且数据日复一日、爆炸式地增长，可谓堆积如山。然而，很多企业对数据挖掘的认识还不全面，甚至感

觉其神秘不可信，这样的话，生产管理中遇到了不能解决的问题，自然不会用数据挖掘的思想思考，甚至基层部门提出使用这样的方法，管理层却因对此不甚了解而无力推动。

为此，我们期望从领导层和生产一线的工作人员普及数据挖掘知识开始，唤起人们对数据新的认识：数据是客观实际的反映，它体现了营销规律、生产规律、经营规律和产品质量控制规律。更重要的是，使企业管理告别基于简单统计分析的“报表”决策时期，跨入数据挖掘的“知识”决策时代。

为了实现这一目标，迫切地需要一本使企业管理者和基层工作者喜闻乐见的读物。然而，市面上的数据挖掘书籍几乎全是教科书形式，理论性太强，满篇数学公式，让人望而却步，而且应用实例甚少，让人难以理解。在这种情况下，我们大胆地萌发出一种案例教学法编写思路，以课堂教学为线索，介绍数据挖掘的基本概念和应用过程，让读者轻松地走进数据挖掘，领略数据挖掘的神奇魅力。

本书的读者群

如果您是一位企业或政府部门的领导，您可以利用乘飞机的闲暇，与本书中的徐教授和各行各业的精英们一起，走进数据挖掘的世界，相信当您下飞机的时候，一定会浮想联翩，产生许多新的思路；

如果您是一位企业管理、生产一线的技术人员，利用一个周末的休息时间，通过本书，您会对数据挖掘有初步而较为系统的了解和认识，您会自觉地尝试利用数据挖掘的方法解决实际问题；

如果您是一位想系统学习数据挖掘知识的学生或科技工作者，亦可以通过本书的阅读，为以后的学习奠定好基础。

本书的内容

全书共 9 章。第 1 章，揭开数据挖掘的面纱，从三个真实而有趣的故事开始，让读者了解数据挖掘的概念、数据挖掘产生与发展、数据挖掘的功能和数据挖掘技术，本章深入浅出地介绍了关联规则、聚类分析、预测（分类和回归）、时间序列等数据挖掘方法及常用算法；第 2 章简述数据挖掘流程，以某冶金企业生产中遇到的质量控制技术攻关难题的解决过程为线索，活灵活现地展现了一个数据挖掘问题的项目立项

及其实施过程；第 3 章到第 9 章以典型案例的形式分别介绍了数据挖掘技术在电力行业、交通航空领域、冶金行业、税务与金融行业、故障诊断、电信行业、互联网行业方面的应用。

本书的特色

形式新颖

本书以 EMBA 班的“数据挖掘技术及其应用”教学为场景，通过教师与学员互动共鸣的形式，带领读者步入数据挖掘的神秘殿堂，领略数据挖掘的神奇魅力。这种写作方式，避免了传统教科书理论性太强，数学公式繁多，让非专业数据挖掘者望而却步的缺陷。

案例导读

本书通过数据挖掘的典型案例，引导读者领略如何利用数据挖掘技术解决各行各业生产和管理中的实际问题。摈弃了晦涩难懂的理论，在解决问题的过程中了解数据挖掘技术及其应用方法，学会**“让数据说话，以数据辅助决策”**的新理念。

创作团队

本书由西安交大美林数据挖掘研究中心策划，靖稳峰、卢耀宗等编写，程宏亮为本书审定了章节划分并精选了案例素材，王璐为本书审定了故事构思和语言风格，程宏斌、李炜、强劲和黄蓉等对本书提出了大量的建设性构想和修改意见，并参与了部分章节的编写。陈浩铭和王羽为本书制作了精美插图。

致谢

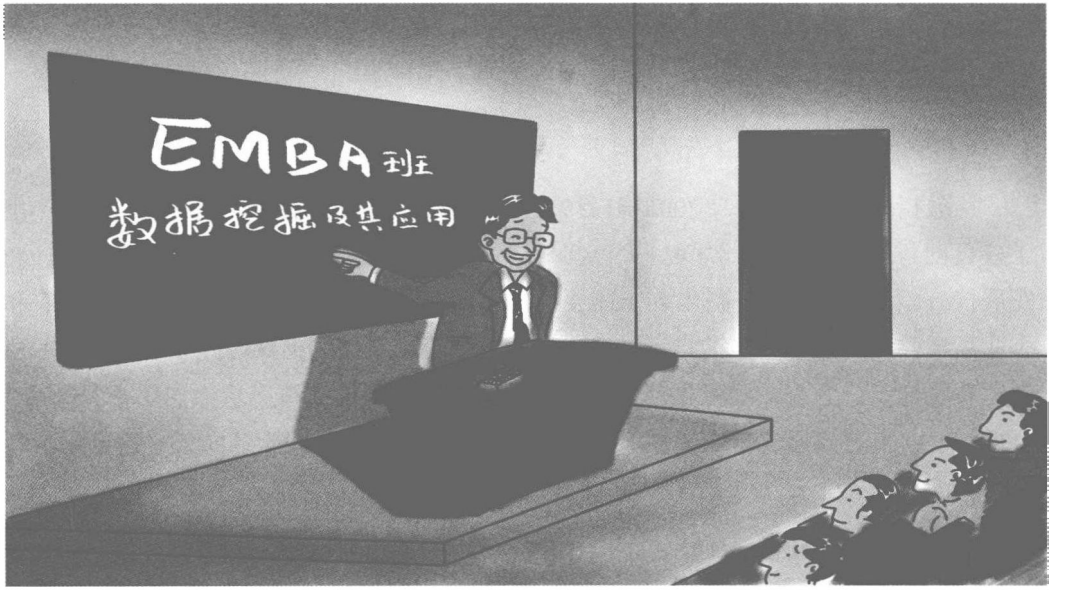
西安交通大学徐宗本院士在百忙中对本书的构思、写作给予了悉心指导，清华大学出版社栾大成编辑对本书原稿字斟句酌，使得本书增色不少，这里一并表示衷心感谢。西安交大美林数据挖掘研究中心还有许多同事为本书的出版付出了大量心血，在此表示诚挚的谢意。

编者

目 录

第 1 章　揭开数据挖掘的面纱

徐教授是某985院校的著名教授，国内数据挖掘专家、智能信息处理研究方向学术带头人，主持了20多项国家项目和国际合作项目，具有丰富的数据挖掘项目实施经验，获得过多项国家级大奖。数十年来，他潜心科研，除了给自己学院的本科生和研究生上课外，一直谢绝其他授课邀请。这次他破例了，欣然接受了本校管理学院第5届EMBA班的“数据挖掘及其应用”课程……



1.1 历史的使命

今天是第一节课，徐教授一跨进教室，迎接他的是学员们一阵热烈的掌声。他习惯性地扫视了一下学生，果然正像管理学院张院长介绍的那样，在座的学员不同寻常，年龄在35～50岁之间，个个西装革履，精神焕发，眼睛里放射出对新知识无比渴望的光芒。

徐教授走上讲台，先在黑板上写下了自己的名字和联系方式，然后微笑着说："同学们，今天我能站在这儿给大家上课，不是因为你们管院张院长有面子，也不是因为你们这些学员地位有多高，说实在的，是党中央、国务院让我来的。"学员们个个目瞪口呆。

有人嘀咕道："难道中央还关心我们这个EMBA班？。"

"关心，而且非常关心。"徐教授铿锵有力地回答。

大家更加疑惑了。

徐教授提高了嗓门："2006年1月9日，在全国科技大会上，党中央、国务院作出了建设创新型国家的重大决策。大家都知道，创新型国家是指以技术创新为经济社会发展核心驱动力的国家。技术创新需要科学家和科技工作者的努力，更离不开政府和企业高层领导和管理人员的推动。张院长在邀请我来给你们上课时介绍说，在座各位都在政府部门或者企业地位显赫，所以我欣然地、破天荒地答应了你们院长的邀请。不过，别以为是你们的乌纱帽吸引了我，而是你们每一个人身上肩负的'建设创新型国家'的历史使命召唤着我。"

徐教授越说越激动，喝了口水继续说："我为科学事业奋斗了一辈子，深知'象牙塔'里的发明、创造，需要与经济建设结合才更能体现出其价值，才更能为建设创新型国家做出贡献。理论创新的成果要真正转化为生产力，迫切需要一种推动力、催

化剂。而能起到这种作用的主体非你们这些人莫属，诚如是，你们就是建设创新型国家的排头兵。你们说，党中央能不关心你们吗？”

徐教授的话音刚落，教室里立刻响起长时间的掌声。

他双手从上向下慢慢挥动，示意大家停下，接着说：“近十年来数据挖掘技术飞速发展，在国外，数据挖掘正在变成整个信息技术的核心之一。尤其是世界500强企业均设立了数据挖掘研发与应用部门，数据挖掘技术已成为其业务成功的关键因素。2007年5月，《纽约时报》以‘**数据挖掘正在进入主流**’为题，介绍了数据挖掘技术，并指出这种新技术正在变成人们工作和生活中不可或缺的一个部分。”

徐教授停顿了一下，向大家问道：“在国内，数据挖掘应用的状况怎样？”

T钢铁公司的李部长抢先答道：“在我国，数据挖掘在互联网、金融、电信和商业等领域已经有一些成功的应用，而在其他行业如制造、航空、医药、反恐和刑侦等只有少量的尝试。”

“李部长的评价比较客观，但大家想过没有，为什么我们与发达国家的差距就这么大呢？”徐教授反问道。

教室里一阵沉默。

于是，徐教授坦率地表达了自己的看法：“其实我也一直在考虑这个问题，当然这里面的原因很多。直到你们管院张院长请我给你们上数据挖掘课时，我又发现了一个不可忽视的因素——政府和企业高层对数据挖掘不甚了解而导致他们对此不够重视或不能站在一定的高度提出有价值的需求。”

徐教授的一席话引起了李部长的共鸣，激动地说：“是的，徐教授讲得太对了。就拿我们钢铁公司来说吧，这几年，我们整天喊‘挺进世界500强’，忙于引进国外先进设备扩大生产规模，但却忽视与外界的技术交流而成为井底之蛙，就连数据挖掘这样在世界500强企业如雷贯耳的新技术我们却闻所未闻。由于自己不具备这方面的知识，生产管理中遇到了不能解决的问题，自然不会用数据挖掘的思想思考，甚至基层部门提出使用这样的方法，领导层却因对此不甚了解而不给力支持。”

李部长的话说到了其他学员的心坎上，他们个个首肯。

徐教授走下讲台，语重心长地说：“所以，我给你们上数据挖掘课来了，我期望从领导普及数据挖掘知识开始，唤起人们对数据的新认识，使你们告别基于简单统计分析的‘报表’决策时期，跨入使用数据挖掘技术的‘知识’决策时代。你们这些社会各界的精英们肩负的历史责任太大了，不管是政府部门的领导还是企业的老总，你们每天都在做各种各样的决策，稍有不慎就可能给国家和企业带来重大损失。我相信各位想为国家贡献自己的力量，但陷入‘心有余而力不足’的境地，正所谓‘我们沉浸在数据的海洋，渴望知识的淡水’！”

听完徐教授一席话，下面的各位老总感慨颇多，台下一片沉思。

徐教授鼓励大家道：“数据挖掘的最高境界就是‘**从数据中获取知识，辅助科学决策**’。希望通过我们的数据挖掘课程的学习，使你们了解到**什么是数据挖掘？它能够干什么？有哪些数据挖掘技术？怎么应用？**大家要认识到，数据挖掘不同于一般的管理软件，编好了拿来用就是了，数据挖掘在行业的成功应用也是一种创新。其实在数据挖掘算法方面，国内（也包括我）的研究团队也有一系列的国际水平的研究成果，但愿我们一起共同努力，推动数据挖掘技术在各行各业的应用，为建设创新型国家做出最大的贡献！”

教室里，又是一阵激动人心的掌声。

徐教授摆了摆手，接着说：“不过，给你们上这门课可让我费了不少脑筋，你们这些学员走向工作岗位都在10年以上了，大学所学的数学知识大都还给了老师，针对

研究生的讲法对你们不适用了。不过，我想出一种专门针对你们的案例教学法，通过典型的应用实例深入浅出地介绍数据挖掘的概念、功能、流程和算法。”

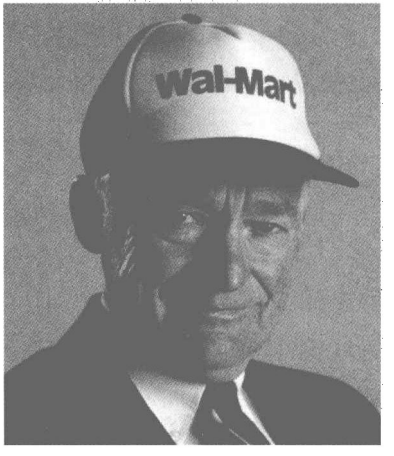
“太好了，徐老师。我曾经翻过几本数据挖掘的书籍，但理论性太强，满篇数学公式，真让人望而却步，而且应用实例甚少，让人难以理解。”李部长感慨地说。

徐教授接着说：“OK，言归正传，让我们开始数据挖掘之旅吧。我先给大家讲三个真实的故事，让你们感受一下数据挖掘到底是神马还是浮云？”

1.2 数据挖掘的故事

1.2.1 震撼业界的发现

“有一个人叫萨姆·沃尔顿的人，大家认识吧？”徐教授问道。



教室里鸦雀无声。

“那沃尔玛，谁没听说过？”徐教授接着问。

“连三岁小孩都知道。”一学员小声说。

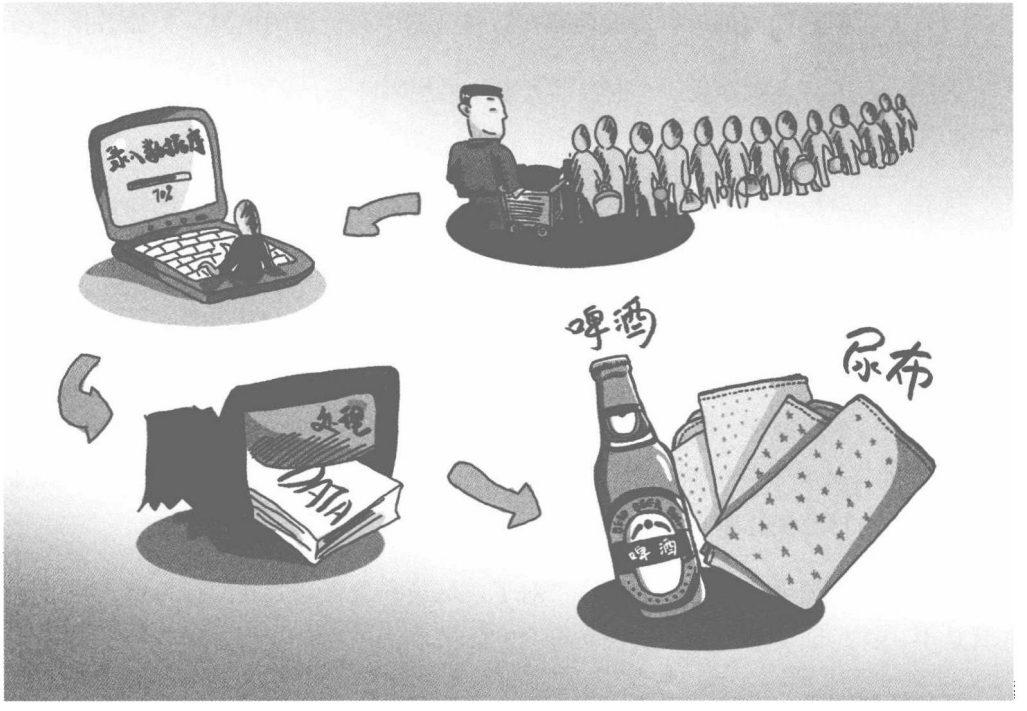
“哈哈，萨姆·沃尔顿是沃尔玛公司的创始人呀！”徐教授笑着说。

“对了，想起来了，萨姆·沃尔顿，是他将一个百货商店奇迹般地经营为全球最大的连锁零售企业，早在 1985 年 10 月就被《福布斯》杂志列为全美富豪排行榜的首位，连美国前总统布什都赞扬他是地道的美国人，展现了创业精神，是美国梦的缩影……”某超市的万总补充说。

“是的，勤奋、创新是这位智慧商人成功的法宝。他的‘日落原则’、‘十英尺

态度’和‘三米微笑’等服务理念以及营销策略‘女裤理论’和‘啤酒与尿布’至今在商业界令人津津乐道。更令人难忘的是，本世纪初**‘啤酒与尿布’**简直就成了‘数据挖掘’的代名词。”徐教授继续说。

“啤酒与尿布，这两个风马牛不相及的东西怎么与数据挖掘扯上了关系？徐老师，快给我们讲讲吧！”移动公司的梁总有点着急了。



“1983 年，当一般零售商还在进行信息化建设的时候，沃尔玛已经开始与休斯公司合作，花费 2400 万美元发射了一颗人造卫星，此后先后投入 6 亿多美元建起了电脑与卫星系统，还发明了条形码、无线扫描枪、计算机跟踪存货等新技术。借助于整套的高科技信息网络，沃尔玛的各部门沟通、各业务流程可迅速、准确地运行，数据库系统很快积累了海量的经营数据，包括大量的顾客消费行为记录。一年一度的圣诞节快要到了，沃尔玛人按照惯例又一次筹划节日的营销策略。这一次他们使用了一种新的‘购物篮分析’软件，对海量的顾客消费行为进行分析，一个意外地

发现让他们瞠目结舌，‘跟尿布一起购买最多的商品竟然是啤酒！’”

“这怎么可能呢？”有学员也感到疑惑不解。

“经过反复计算、核实，结论没有错。”徐教授答道。

“不过，这个故事告诉我们什么？”又有人问道。

“告诉我们数据挖掘可以发掘埋藏在海量数据中有价值的信息。”徐教授答道。

突然，后排有人大声说：“也告诉大家如果想喝啤酒，老婆不让买，就说去买尿布吧！”惹得大家哄堂大笑。

接着，徐教授问：“这是数据挖掘技术对历史数据进行分析得出的知识，这个结果符合现实情况吗？是否有利用价值？”

“还利用价值，真是六月里穿皮袄——反常！”有学员不以为然。

“紧接着，沃尔玛派出市场调查人员和分析师对这一结果进行了深入研究，证实它揭示了一条隐藏在‘尿布与啤酒’背后的美国人的一种行为模式：一些年龄在25～35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%～40%的人会顺手为自己买几瓶啤酒。”

刚才那位学员想通了，小声说：“对了，这是在美国，老外的行为模式与中国人就是不一样！证实了这样的发现是符合实际的，沃尔玛会怎么办呢？”

徐教授挥动了一下电子教鞭，大声说：“沃尔玛立即采取了行动，将卖场内原来相隔很远的妇婴用品区与酒类饮料区的空间距离拉近，使顾客更加方便。然后对本地区新生育家庭的消费能力进行了调查，对这两个产品的价格也做了调整，并向一次购买达到一定金额的顾客赠送婴儿奶嘴及其他小礼品，结果是尿布与啤酒的销售量双双大增。”

某超市的万总激动地站了起来，情不自禁地说：“不愧为全球零售业巨头啊，高招，值得借鉴！”