HZ BOOKS
华章教育

管理科学与工程精品教材

Mc Graw Hill Education

商业数据挖掘导论

（美）戴维·奥尔森（David Olson）
内布拉斯加大学（林肯）

（中）石 勇（Yong Shi）
中国科学院研究生院
内布拉斯加大学（奥马哈）

著

吕 巍 等译

Introduction to Business Data Mining

机械工业出版社
China Machine Press

管理科学与工程精品教材

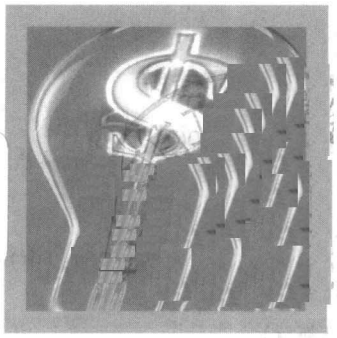
商业数据挖掘导论

（美）戴维·奥尔森（David Olson）
内布拉斯加大学（林肯）

（中）石 勇（Yong Shi）
中国科学院研究生院
内布拉斯加大学（奥马哈）

著

吕 巍 等译



Introduction to Business Data Mining

机械工业出版社
China Machine Press

本书综合商业专业知识和数据挖掘模型开发于一体，系统地介绍了数据挖掘商业环境、数据挖掘技术及其在商业中的应用。在注重对数据挖掘技术讲解的同时，强调了数据挖掘在商业决策领域中的应用，弥补了大多数数据仓库技术类书籍商业应用不足的缺点。本书主线清晰，案例丰富，语言精练。

本书既可以作为商业专业本科生、研究生的教材，也可以在MBA、EMBA教学和企业培训中使用。

David Olson, Yong Shi. Introduction to Business Data Mining.

ISBN 007-124470-0

本书版权登记号：图字：01-2006-3189

图书在版编目（CIP）数据

商业数据挖掘导论/（美）奥尔森（Olson, D.），（中）石勇著；吕巍等译. －北京：机械工业出版社，2007.8

（管理科学与工程精品教材）

书名原文：Introduction to Business Data Mining

ISBN 978-7-111-22017-6

Ⅰ.商…　Ⅱ.①奥…　②石…　③吕…　Ⅲ.数据采集－计算机应用－商业经营　Ⅳ.F715-39

中国版本图书馆CIP数据核字（2007）第114759号

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：程　琨　　版式设计：刘永青

北京慧美印刷有限公司印刷　·　新华书店北京发行所发行

2007年8月第1版第1次印刷

184mm×260mm　·　15.5印张

定价：38.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

本社购书热线：（010）68326294

投稿热线：（010）88379007

译 者 序

虽然我们已经进入了网络经济时代，数据库技术的应用也越来越普及，但是当企业面对大量数据时，仍然觉得无从下手，结果收集在大型数据库中的数据变成了“数据坟墓”，原因是：决策人员缺少从数据中发现知识的工具。数据挖掘技术于20世纪90年代开始迅速兴起，并广泛地运用于商业领域中。商业数据挖掘这一领域横跨商业、计算机科学、统计分析等学科，能够有效地帮助企业将巨大的数据资源转换为有用的知识与信息资源，提高企业在信息密集环境下的商业决策能力。

国内外的大多数数据挖掘类书籍，或者偏重于数据仓库技术的描述，而忽略了对相应的商业问题的描述；或者更多地讲述商业决策，而缺乏对数据挖掘技术的讲解。笔者庆幸能够有机会阅读这本《商业数据挖掘导论》，发现本书是一本既涵盖数据挖掘技术应用，又充分概括各类商业决策问题的书籍。本书深入浅出地描述了数据挖掘技术在商业领域的应用，是一本非常理想的读物。因此，笔者付出了半年的心血将此书译成中文，以飨读者。

笔者一直致力于基于数据挖掘技术的精确营销研究工作。引入本书，正值中国企业营销进入精确营销时代之际，引入的时机可谓恰到好处。消费者行为细分、客户流失预警、客户关系管理、交叉销售等理念已经成为商业研究的热点。现代企业一般都拥有庞大的数据资源，数据挖掘技术无疑能够提升现代企业的精确营销能力。

本书的作者戴维·奥尔森（David Olson）是内布拉斯加大学的商学博士，主要关注于多重目标决策，曾在80多家相关期刊上发表过其研究成果。石勇教授的研究领域主要是信息超载、多重标准决策和电信管理，曾为许多著名公司做过数据挖掘和知识管理等项目的咨询工作。两位作者都有丰富的商业数量分析实战经验和数据挖掘教学经验，这使得本书特色鲜明。

- **内容翔实。**本书综合商业专业知识和数据挖掘模型开发于一体，系统地介绍了数据挖掘商业环境、数据挖掘技术及其在商业中的应用。在注重对数据挖掘技术讲解的同时，强调了数据挖掘在商业决策领域中的应用，弥补了大多数数据仓库技术类书籍商业应用不足的缺点。
- **主线清晰。**全书以数据挖掘的商业应用为主线，这一主线实际上就是读者不断学习、

理解和掌握商业数据挖掘技术的过程。此外，随着最近几年数据挖掘技术的迅速发展，本书还介绍了商业数据挖掘的最新研究领域，如文本挖掘、Web挖掘以及数据挖掘中的道德问题等。

- **案例丰富。**数据挖掘技术的应用是为了能够更好地进行商业决策，通过案例，读者可以更轻松地了解商业数据挖掘的过程。本书详细阐述了数据挖掘技术在电信、保险、零售业以及人力资源等商业领域中的应用，初涉该领域的读者也可以较为全面地了解到数据挖掘技术的商业价值。
- **语言简练。**无论是对数据挖掘技术的讲解，还是对商业运用的描述，本书总是力图用最简练的方式贴近读者。

对于广大商学院的学生来说，这是一本很好的数据挖掘技术书籍。同样，对于从事商业数据挖掘的工作人员来说，这本书又可以使他们很好地理解商业问题。特别是对实践于银行业、保险业、电信业、零售业的行业人士来说，这本书具有现实的借鉴意义。所以，本书既可以作为商业专业本科生、研究生的教材，也可以在MBA、EMBA教学和企业培训中使用。

本书由我的研究生梯队主译，我的学生王雷参与了第1、2章，练叔凡参与了第3章，赵诚宁参与了第4章，柏佳洁参与了第5章，李玉峰参与了第6、7章，尚烁徽参与了第8、9章，赖敏参与了第10、11章，孟韫参与了第12、13章的翻译工作，并由李玉峰、王雷、孟韫对本书进行了初校，全书由我统稿并定稿。

本书的出版得到了机械工业出版社华章公司的吴颖洁女士和李玲女士的大力支持，在此一并表示感谢！由于译者水平有限，加之时间仓促，书中难免有翻译不当之处，恳请各位同仁指正！

吕　巍

上海交通大学

2007年5月

作者简介

戴维·奥尔森（David Olson）

戴维·奥尔森是内布拉斯加大学James & H. K. Stuart管理信息系统教授。1981年，获得内布拉斯加大学商学博士学位。曾在80多家相关期刊上发表过研究成果，主要关注于多重目标决策。他教授管理信息系统、管理科学和运营管理领域的课程。他的著作有《选择问题的决策援助》、《信息系统项目管理概论》、《企业资源规划系统管理手册》，合作编著的著作有《决策支持模型和期望系统》、《管理科学概论》、《激励和风险分析概论》、《商务统计学》、《决策分析中的质量信息》、《统计、决策分析和决策建模》、《战略定位问题的多重标准》和《商业数据挖掘导论》。他在该课题的国际、国内会议上做过100多次演讲，是信息系统联盟成员，决策科学委员会成员，运营研究和管理科学委员会成员，决策多重标准协会成员。他协调了决策科学委员会论文竞赛，创新教育竞赛，主持博士事务委员会，并三次被全民选举为副主席和国内项目主席。1981～2001年，他在得克萨斯州A & M大学度过，并在最后两年成为信息和运营管理学院商业学教授。在A & M大学，他获得了商学院和商业研究院的研究员奖励，并连续两年获得商业分析学院优秀分析员称号。同时，他也是决策科学委员会的研究员。

石勇（Yong Shi）

石勇教授目前是中国科学院数据技术和知识经济研究中心主任，中国科学院成人大学副主席。1999年，他成为内布拉斯加大学（奥马哈）信息科学和技术学院杰出的信息技术教授。石勇教授的研究领域包括数据挖掘、信息超载、多重标准决策和电信管理。他出版了7本书，在各种期刊和会议上发表文章60多篇，是期刊《国际信息技术与决策制定》的主编，期刊《国际运营与数量管理》的区域主编，并是多家学术期刊编委会成员。石勇教授获得了很多引人注目的奖励，包括杰出青年科学家、中国自然科学奠基人（2001年）、中国科学院海外技术顾问（2000年5月）以及IEEE计算机协会杰出访问者项目代言人（1997～2000年）。他还为许多著名公司做过数据挖掘和知识管理等项目的咨询工作。

前　言

本书的写作目的是服务于那些想要学习数据挖掘的高年级本科生和研究生。数据挖掘是一个非常有用的主题，通过最近开发的信息技术，使数量分析应用于大规模数据已经成为可能。我们两人都教授过这些内容，都有丰富的商业数量分析的经验，石勇还拥有丰富的商业数据挖掘分析实战经验。借此机会，我们要感谢内布拉斯加大学（奥马哈）的几位研究生——寇钢、燕念、庄伟，感谢他们利用计算机软件为我们提供了本书中的数据挖掘报告。

本书内容

我们的目标是介绍数据挖掘的基础概念，收集大量数据进行分析并最终获得商业理念的潜力。我们把材料分为四个部分。第一部分介绍概念。第二部分描述和解释数据挖掘基础理论。第三部分关注数据挖掘的商业应用。第四部分介绍更深入的研究领域，包括网络挖掘、文本挖掘和数据挖掘的伦理要素。第一部分是导论。第二部分是介绍大量不同的数据挖掘技术的章节。不是所有的章节都需要涵盖，这一系列可以根据教师的需求有所变动。第三部分讲述应用，我们觉得这些章节包含了最有趣和最重要的材料。关注技术的教师也许对这部分并不感兴趣。相反，对商业应用感兴趣的教师可以在阅读第二部分之前直接阅读第三部分。这种方式在可以应用数据挖掘软件建模的时候更为有用。第四部分，我们认为现在很重要，并在将来更加重要。然而，内容选择和阅读顺序还是要依赖于教师的需要的。

第一部分：导论

第1章是对数据挖掘的整体介绍，描述了数据挖掘的一般过程。对有用的商业应用进行概述。第2章详细介绍了数据挖掘的过程，用典型的数据解释这一过程，从数据挖掘软件中取得数据的可视化效果。第3章介绍了数据挖掘的支撑——数据库，并描述了多种不同的软件工具，从数据集市的数据仓库产品到在线分析过程，而且关注数据质量，通过原形数据解释不同的概念。

第二部分：数据挖掘工具

第4章描述了数据挖掘技术和功能。第5章描述和解释了聚类算法，并介绍了可以支持的软件产品。第6章介绍了回归分析的各种形式，来最优拟合给定的数据集。第7章讨论神经网络，它是人工智能的一个应用，适合许多数据挖掘应用。第8章介绍了决策树分析方法，

描述了基础算法，还有决策树结构、机器学习、决策树模糊集。提供相应软件产品的介绍，详细解释了See5。第9章介绍了基于线性程序的拟合数据的方法，描述并解释了实际数据挖掘的应用过程。

第三部分：商业应用

第10章描述了数据挖掘在商业中的应用，关注这些分析在商业决策中的价值。这里包括重要的客户关系管理。描述了增益这一概念，回顾了芬格侯（Fingerhut）公司市场细分的开发历程。第11章描述了市场购物篮分析，一个更为高质量的数据挖掘工具，并通过案例进行描述。本章还讲述了活跃性、同族定位、交叉销售等数据挖掘中的基础概念。

第四部分：发展中的问题

第12章介绍了文本挖掘和Web挖掘。第13章讨论了与数据挖掘相关的伦理问题。

戴维·奥尔森
内布拉斯加大学（林肯）
石　勇
内布拉斯加大学（奥马哈）

目 录

第一部分

导　　论

第 1 章

商业数据挖掘简介

学习目标

- 引入数据挖掘的概念
- 介绍数据挖掘的典型应用
- 解释关键概念
- 数据挖掘工具简介
- 介绍本书其他章节的轮廓

随着科学技术的进步，我们获取大量数据的能力大大增强。在吸纳数据方面，计算机系统要比人工快得多。更重要的是，网络的存在使得在全球范围内实时共享数据成为可能。

最近的一次政治事件再次表明数据能够预测未来这一事实。有些人把没能防范恐怖袭击的责任归咎于系统的不足，理由是，如果挖掘到足够深度，你总是能够找到一些数据或是一个备忘记录，这些信息能够预示即将发生的事件。然而，你还会发现，大量被预测的事情，事实上并没有发生。显然，很多组织都有一个明确的需求，这就是更快更合理地处理数据。数据挖掘应用分析用来检测样本并完成预测，数据挖掘不是一门完美的科学，它的目的是为了获取微小的优势，因为完美的预测是永远不可能的。这些微小的优势可以给商业带来巨大的收益。例如，零售组织已经开发了先进的客户细分模型。以前，零售组织的注意力往往集中于向那些购买意愿较小的顾客散发零售资料，现在，它们把注意力集中在那些购买意愿较高的客户细分上，从而节省了成本。银行及其他组织开发了先进的客户关系管理程序（由数据挖掘技术支撑），能够预测特殊类型客户对于该组织的价值，并且能够预测贷款回报率等。保险公司很早以前就开始使用统计分析工具了，保险公司使用一些工具预测保险欺诈，这些工具都是由数据挖掘工具拓展而来的。这仅仅是众多数据挖掘应用中的三个例子而已。

本书将描述数据挖掘的一些商业应用，也将讲述数据挖掘的一般过程、数据挖掘所需要的数据库工具以及数据挖掘应用技术。

1.1　介绍

数据挖掘是指对储存在电脑中的海量数据进行分析。例如，食品店通过客户购买而获取了大量数据，条形码使我们的付款变得非常便利，并且提供给零售公司大量数据。食品店和其他零售商店能够快速地掌握顾客的购买行为，并通过计算机对产品进行精确定价。同样，计算机可以通过即时检测现存产品数量，来帮助商店进行清单管理。他们还可以应用计算机技术联系供应商，使他们不会出现断货情况。计算机使商店的财务系统能够更加精确地进行成本控制，分配利润。所有的这些信息都是基于粘贴在产品上的条形码信息。从条形码中获取的数据和其他渠道获取的信息一起，被用来进行数据挖掘分析。

数据挖掘不仅仅局限于商业应用。2004年美国大选时，两大政党都应用了数据挖掘预测选票。[1] 数据挖掘在医学领域也得到了大量应用，他们收集并分析病人的诊断记录来帮助制定最佳的锻炼计划。[2] Mayo Clinic与IBM合作开发了一个在线计算机系统，监测过去的100个同性别、同年龄的Mayo病人对特殊治疗的反应。[3]

数据挖掘的商业应用也是相当广泛的。丰田公司应用数据挖掘的**数据仓库**（data warehouse）功能来选择更为有效率的运输路径，这给客户节约了平均19天的时间。数据仓库（将在第3章中讨论）是一个巨大的数据库系统，能够储存像沃尔玛这样的大型商业组织的交易数据。丰田公司也能够更快地发现销售动向，从而决定最佳的新交易地点，这使其在北美的销售利润增加到3000万美元。[4]

数据挖掘在银行争夺信用卡客户、[5] 保险和电信公司预测欺诈、[6] 制造企业的质量管理，[7] 以及其他很多方面都得到了广泛的应用。数据挖掘还被于来增强食品安全、[8] 犯罪侦破[9]和观光旅游[10]等。芬格侯（Fingerhut）公司把目标集中在具有高回应的小客户群，从而在**微观市场营销**（micromarketing）上做得非常成功。像纳利公司(R.R.Donnelly & Sons)这样的传媒公司，提供顾客及其生活方式数据，以及为其定制印刷品，主要针对的是那些应用数据挖掘进行目录营销的公司。

数据挖掘包括统计和/或人工智能分析，通常应用于大数据集。传统的统计分析通常比较直接，是因为存在一系列明确的预期结果。这种路径可以被称为受监督的路径。然而，数据挖掘不仅仅是技术工具的应用。数据挖掘的核心是**知识发现**（knowledge discovery）即学习新的有用的东西，这被称为是不受监督的路径。例如，在决策树分析中，大量的工作是通过自动方式完成的。当然，数据挖掘不会局限在自动分析。人与计算机合作完成的图标工具和突发事件识别都能强化人类的知识发现。

数据挖掘在商业中有多种应用方式，以下是三个例子：

- **顾客分析**（customer profiling）：识别那些有最大利益的顾客细分群体。
- **目标指定**（targeting）：发掘被对手获取的高利润客户的特性。
- **市场购物篮分析**（market-basket analysis）：发掘顾客购买的产品中，哪些可以用来做产品定位和交叉销售。

这些不是数据挖掘的所有应用，但它们是最重要的三个商业应用。顾客分析是客户关系管理中最为关键的部分，这将在第10章中进行详细论述。目标指定是进行客户流失管理和客户营业额管理的核心概念，这也将在第10章中讲述。市场购物篮分析是数据挖掘的一个很有趣的应用，我们将在第11章中讨论。

1.2　进行数据挖掘需要什么

数据挖掘需要界定目标问题，收集能够有助于更好地理解市场的数据，以及一个能够提

供统计分析或者其他分析方式的计算机模型。下面是两个通常的数据挖掘研究方式。**假设检验**（hypothesis testing）是指一个行动与结果之间关系的理论假设。很简单，我们假设广告可以产生更大的利润。长期以来，这种关系已经被零售企业在其各自的组织内部研究过。数据挖掘是基于大量数据基础上进行的关系识别研究，而这些数据可能包括为了促销或特定产品线利润率而做的各类广告反应率的调查。第二种数据挖掘的方式是知识发现，在这种分析方式中，可能不存在预想的论断，但其中的关系可以通过对数据的观察而获得。这可以通过形象化工具对数据的展示而获得，或者是通过相关性分析等基础统计分析来获得。

形式多样的计算机分析模型被应用到数据挖掘中。本书第5～9章将讨论各类模型。所有这些都需要对数据进行访问。通常，包含数据仓库和数据中心的系统用来管理大量数据（见第3章），其他的数据挖掘分析应用于较小的数据库，比如联机分析处理系统。

1.3 数据挖掘

数据挖掘在众多分析方法中，也被称为探测性数据分析。这些通过收银机、扫描器和公司专题讨论而获取的大量数据，要受到检测、分析、缩减和再利用。研究主要是在预测销售额、市场反应和利润的各种不同模型中进行，分类统计方法是数据挖掘的基础，自动人工智能方法也得到一定的应用。然而，通过分类统计方法而进行的系统探测仍是数据挖掘的基础。一些统计分析领域所开发的工具在自动化控制处理数据的过程中得到应用。

数据挖掘工具应该是通用的、可升级的、能够精确预测行为与结果之间的反应和能够自动执行的。**多种工具**（versatile）是指工具能够应用于各种模型；**可测度**（scalable）工具是指既能运用在小的数据集中，也能在大的数据集中应用的工具。自动化非常有用，但它的应用应该是相对的。一些分析功能是自动化的，但需要人工给执行程序设定初始值。事实上，分析员的判断对数据挖掘的成功执行起着关键作用。在研究中，选取合适的数据显得特别关键，这往往需要对数据进行转换。越多的不确定性就会产生越多的结果，同时，显示出的数据之间的关键关系就越少。Two Crows公司总裁Herb Edelstein说："如果你没有统计或挖掘方面的背景，要想完成数据挖掘工作，那将是令人难以置信的事。"[11]

数据挖掘迅速发展，也使商业受益匪浅。其中最受益的两个应用领域是：第一，市场营销组织应用客户细分来识别那些对不同形式营销传媒敏感的客户群；第二，银行应用数据挖掘来预测客户对提供不同服务的反应。许多公司正在应用这些技术来识别高价值客户，从而为其提供所需的服务以留住他们。

北达科他第一国民银行（First National Bank of North Dakota）发现仅仅10%的顾客为公司贡献了几乎全部的利润。[12] 旧金山的美国银行（Bank of America in San Francisco）同样发现，顾客中的小部分——大约20%，决定着银行的利润。美国银行找出了大客户的特征，并为其提供针对性服务。它们还能够通过电话业务术语评估哪些特定顾客会把生意交给竞争对手，或是流失的可能性。

娱乐行业对数据仓库和数据挖掘也有一定的应用。很久以前，娱乐场很想知道其顾客的每一个细节。[13] 哈拉斯娱乐有限公司(Harrah's Entertainment Inc.)是应用激励措施的赌场之一。[14] 大约有800万顾客持有金卡（total gold cards），它可以用来在娱乐场游戏、就餐、住宿，或是其他方式的消费。累计的点数可以用来补充就餐或住宿，如果有更多的点数则会受到奖励，因为它为公司创造了利润。这些信息被送到公司的核心数据库中，并保留多年。Trump公司的Taj卡也采取相似的做法。最近，高度竞争引发了对数据挖掘的广泛应用。

Bellagio酒店和Mandelay Bay酒店不再总是以广告来告知人们它们的地址，取而代之的战略是宣传其高档的场馆环境。数据挖掘用来识别那些比较空闲的人群，有价值客户将受到重点关注。数据仓库使娱乐场能够估算玩者的终身价值。旅游计划激励、室内促销、公司业务介绍、客户跟进是用来维系最有价值客户的工具。博彩业（Casino gaming）拥有现在最大的数据库之一，在那里，非常个性化的私人信息都可以挖掘，一些顾客被认为是可以鼓励其长期玩下去的，而其他一些顾客则被认为是不该受到鼓励的。Harrah公司发现26%的赌者贡献了82%的收益。它们还发现，最有价值的顾客不是那些很悠闲的人群，而是那些以前是职业玩家的中年人。Harrah公司开发了一个定量模型来预测长期的个人消费，并启动了一个项目邀请那些在3个月内没有光顾的每月消费达1000美元的顾客。如果顾客在光顾后流失，那么他将被邀请参加另一个特殊的活动。[15]

数据挖掘甚至被应用在艺术传播中。[16] 这包括为展览识别潜在顾客。管理展览的软件程序像飞机上的坐标软件一样高效地运行着，伴随着向电子办公的实现，产生了大量的数据，并被输入到数据仓库中。

1.4　集聚营销

芬格侯公司成立于1948年，目前已经成为集聚营销领域的领袖。最近几年，公司向6500万名客户邮寄了大约130份不同的产品目录，迅速建立起了数据量达6千兆的数据仓库。当时的报道称，数据挖掘分析应用了300多个预测模型，[17] 来关注企业的1200万最为活跃的客户的3000个不同特征。1999年2月，联合百货（Federated Department Stores）支付了17亿美元给芬格侯公司以购买其数据库。[18] 低收入家庭每年会带给芬格侯公司16亿～20亿美元的业务。[19] 在公司的顶峰时期，芬格侯公司每年邮寄3.4亿份产品目录给它们的700万名活跃客户。[20] 芬格侯公司邮寄分类目录（每年4亿份）[21] 给它们的数据挖掘客户，并期待他们能对芬格侯公司众多市场产品中的一款感兴趣。每一条生产线都有自己的产品目录。目标客户被认为是购买边际可能性较高的群体的一个小子集［营销术语中**增益**（lift）的概念——见第10章］。联合百货希望能够把芬格侯公司的技术应用到它们的Macy和Bloomingdale超市中。

芬格侯公司应用细分、决策树、回归分析、神经网络模型等工具，其中SAS和SPSS作为回归分析工具，还有神经网络工具。当芬格侯公司的700万名活跃客户中的一位订购了一件产品（玩具、游戏、家居用品或是其他产品）时，交易、统计和心理数据都被储存在公司的相关数据库中。每一位顾客都有高达3000个潜在数据条目。公司有一位员工专门负责数据仓库，[22] 他的另一个角色就是培训公司其他员工，使他们懂得使用数据仓库。

分类模型将订购信息、基本统计信息和芬格侯公司的产品供给信息结合在一起。这使得芬格侯公司能够针对那些购买潜力最大的顾客进行新的邮寄。芬格侯公司的分析员发现：刚刚搬迁的顾客在接下来的12周内的购买量会是原来的3倍。[23] 芬格侯公司因此制作了一个产品目录，涵盖了家具、电话、装饰品等新搬迁的客户最有可能购买的产品，而删除了珠宝、家用电器等产品。

另一个应用是邮件流优化。这个模型显示了那些最有可能对现有产品目录做出回应的顾客。芬格侯公司借助邮件流优化每年节省了大约300万美元的成本。[24] 1998年，清单销售行业出现滑坡，这个系统的应用使芬格侯公司摆脱了这一趋势，在邮寄量减少了20%的同时，净利润超过了3700万美元。[25]

神经网络模型是一个常用的数据挖掘工具（见第7章），被用来识别邮寄样品和电话订