Broadview
WWW.BROADVIEW.COM.CN

数据仓库与数据挖掘
技术应用丛书

Visual Data Mining

Techniques and Tools for Data Visualization and Mining

可视化数据挖掘

数据可视化和数据挖掘的技术与工具

[美] Tom Soukup Ian Davidson 著

朱建秋 蔡伟杰 译

WILEY 電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据仓库与数据挖掘技术应用丛书

Visual Data Mining

Techniques and Tools for Data Visualization and Mining

可视化数据挖掘

数据可视化和数据挖掘的技术与工具

[美] Tom Soukup Ian Davidson 著

朱建秋 蔡伟杰 译


B1291532

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书描述了可视化数据挖掘技术，以及可视化数据挖掘技术能够解决的商业问题。在介绍完业务问题和基本原理后，以一个完整的实例逐步讲解如何利用可视化数据挖掘技术实施商业智能项目的方法。利用可视化数据挖掘工具和技术，分析人员能够从全新的角度快速、轻松地检索信息解决常见的商业问题。可视化数据挖掘使数据挖掘变得简单，非技术出身的业务经理们利用它能够更好地了解市场并做出明智的决策。

另外，本书还介绍了可视化工具方面的知识，拓宽了读者的范围。本书适合于数据可视化和可视化数据挖掘商业智能解决方案实施单位的各层次人员，包括：数据分析员、业务分析员、领域专家和决策人员。

John Wiley & Sons, Inc.
Copyright © 2002 by Tom Soukup and Ian Davidson. All rights reserved.
All rights reserved. Authorized translation from the English language edition published by John Wiley & Sons, Inc.
本书简体中文专有翻译出版权由 John Wiley & Sons Inc. 授予电子工业出版社，未经许可，不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号　图字：01-2002-5362

图书在版编目（CIP）数据

可视化数据挖掘：数据可视化和数据挖掘的技术与工具/（美）苏克（Soukup,T.），（美）戴维森（Davidson,I.）著；朱建秋，蔡伟杰译. —北京：电子工业出版社，2004.1
（数据仓库与数据挖掘技术应用丛书）

书名原文：Visual Data Mining: Techniques and Tools for Data Visualization and Mining

ISBN 7-5053-9301-4

Ⅰ.可… Ⅱ. ①苏… ②戴… ③朱… ④蔡… 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（2003）第 100730 号

责任编辑：毕　宁　bn@phei.con.cn
印　　刷：北京增富印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
经　　销：各地新华书店
开　　本：787×980　1/16　印张：22.5　字数：370 千字　彩插：4
印　　次：2004 年 1 月第 1 次印刷
印　　数：6 000 册　　定价：58.00 元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系。联系电话：（010）68279077。质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

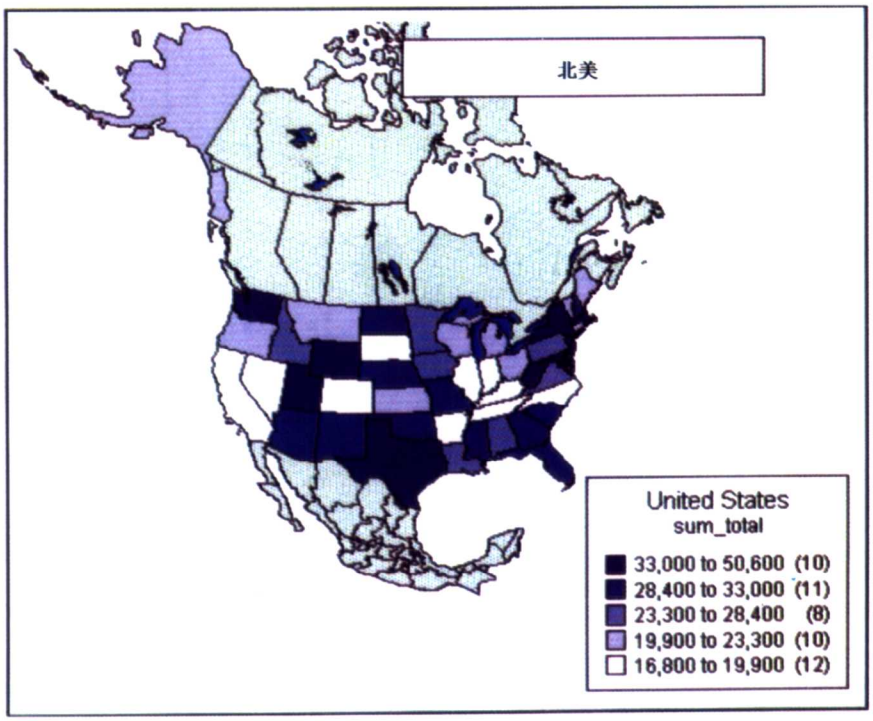


插图 1　每月各店平均利润随机抽样 50 000 条记录的地图可视化

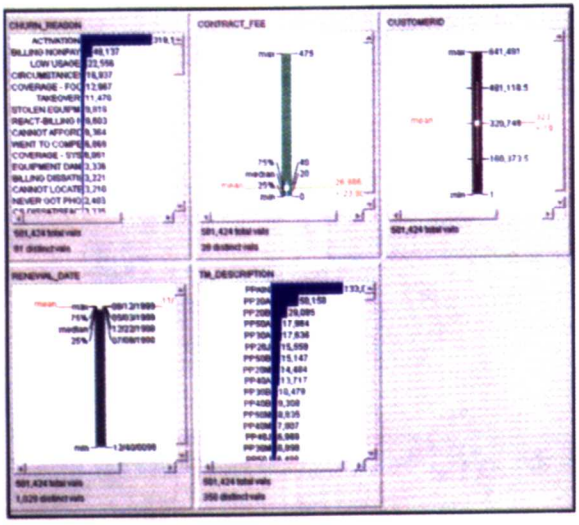


插图 2　转换后的合同表的分布和统计图

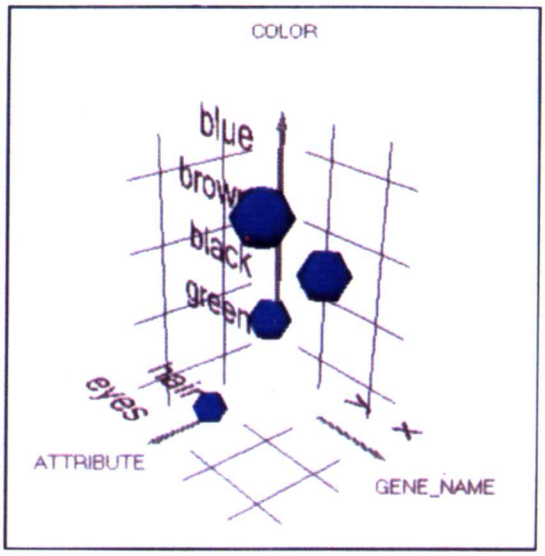


插图 3　使用记录权重作为实体尺寸的基因数据集的散点图

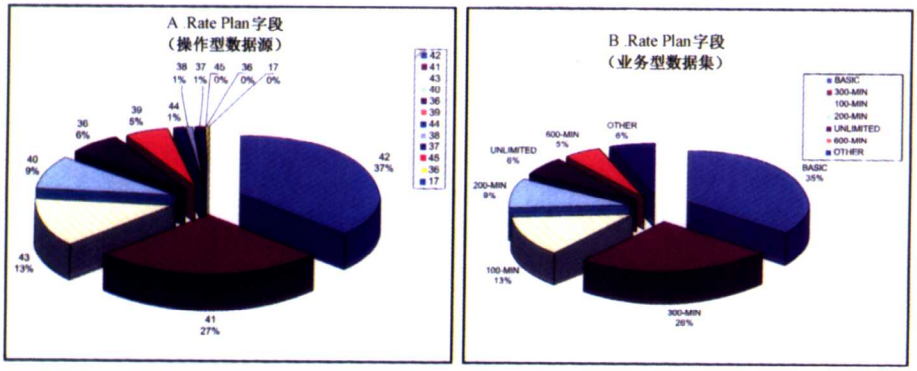


插图 4　使用分割饼图验证逻辑字段的分组操作

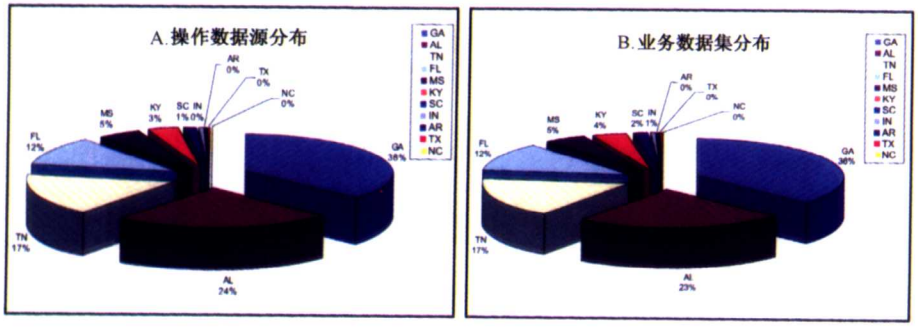


插图 5　使用分割饼图验证 STATE 字段的 ECTL 操作

此为试读，需要完整PDF请访问：www.ertongbook.com

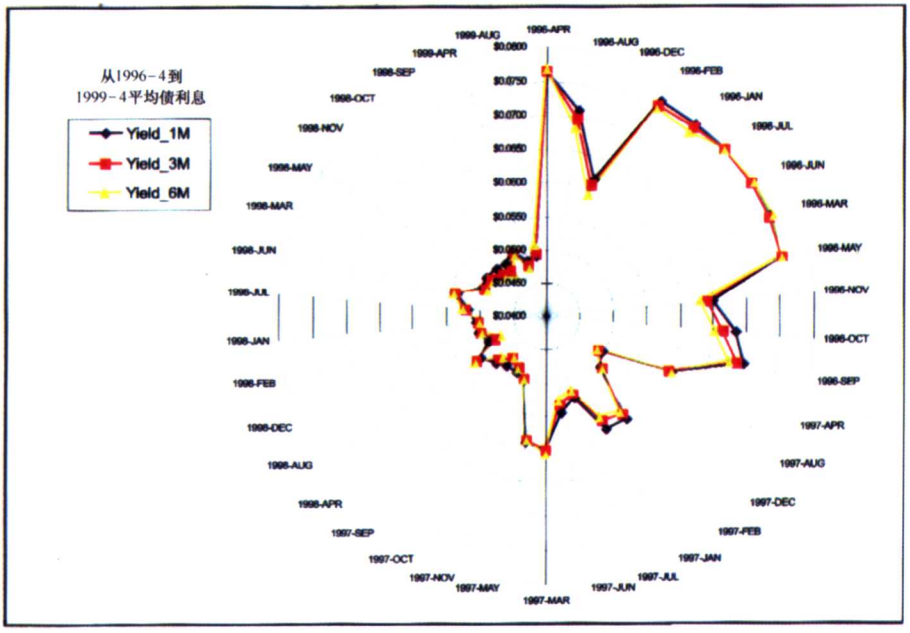


插图 6　平均公债利息雷达图

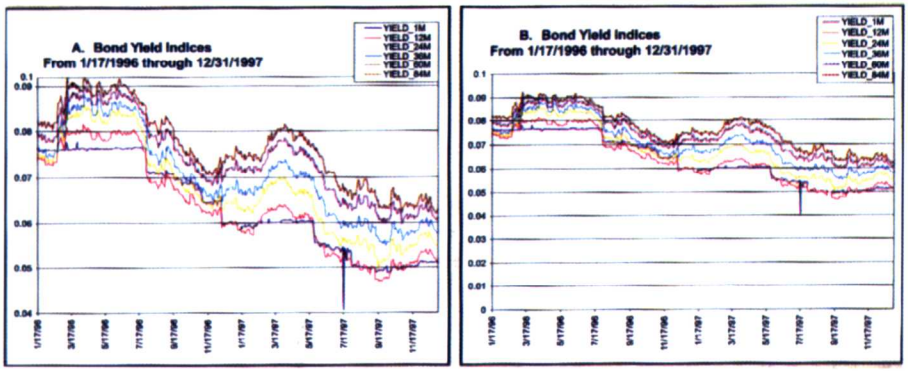
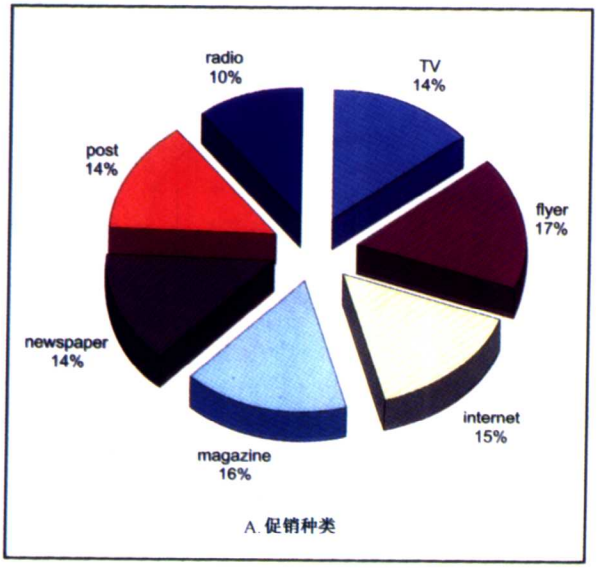


插图 7　使用不同的 y 轴区间的平均公债利息折线图



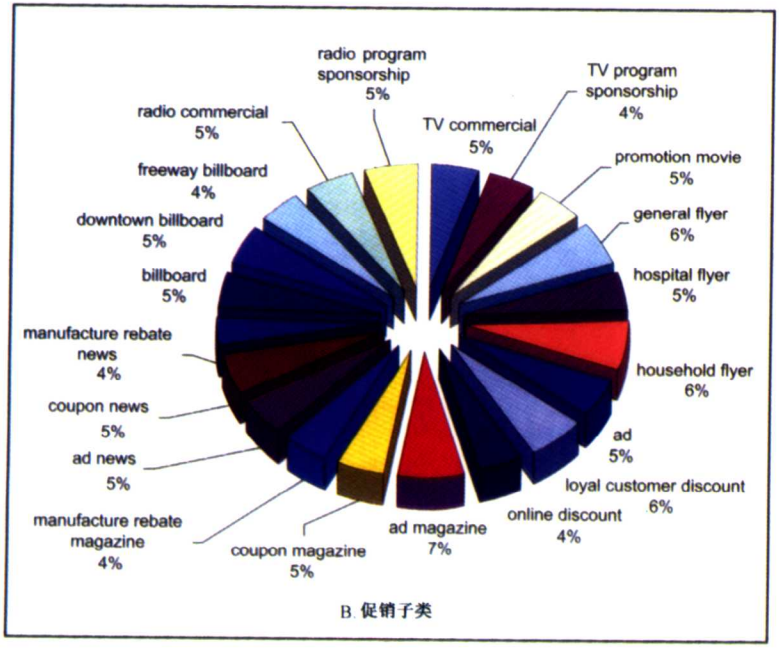


插图 8　促销和各促销种类平均成本的分割饼图

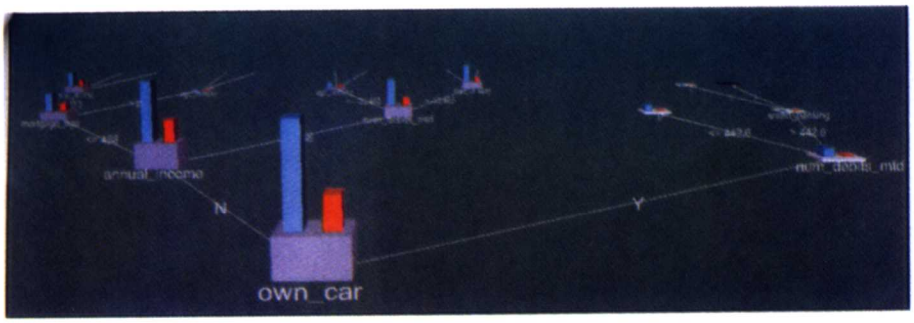


插图 9　决策树模型的树型图

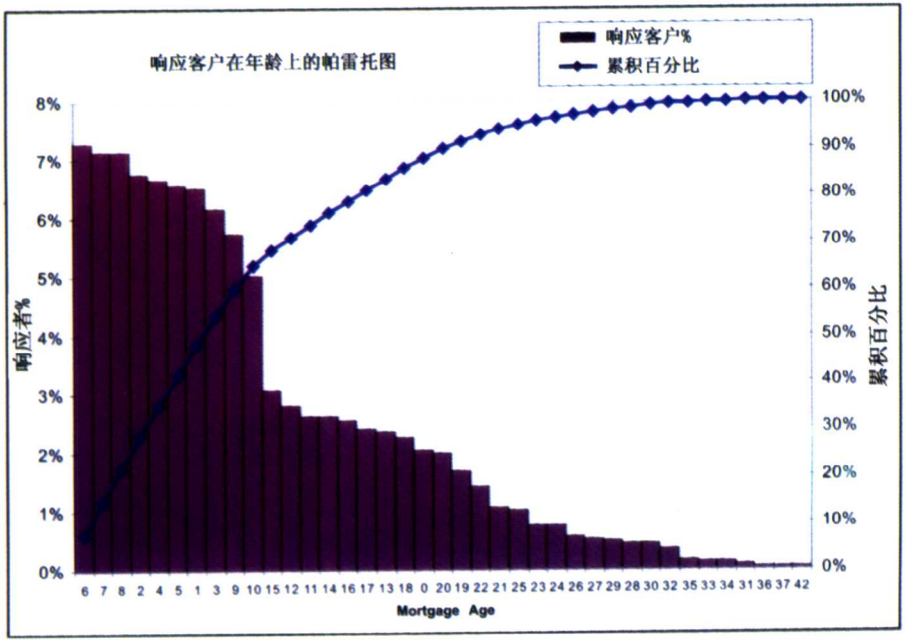


插图 10　响应者抵押期限的帕雷托图

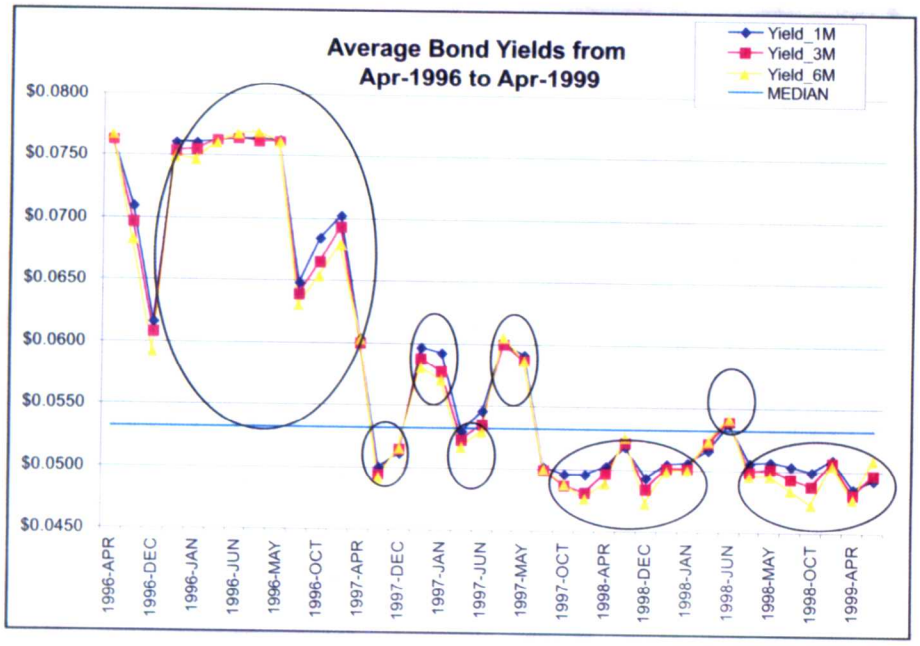


插图 11　平均公债指数折线图中发现的聚类

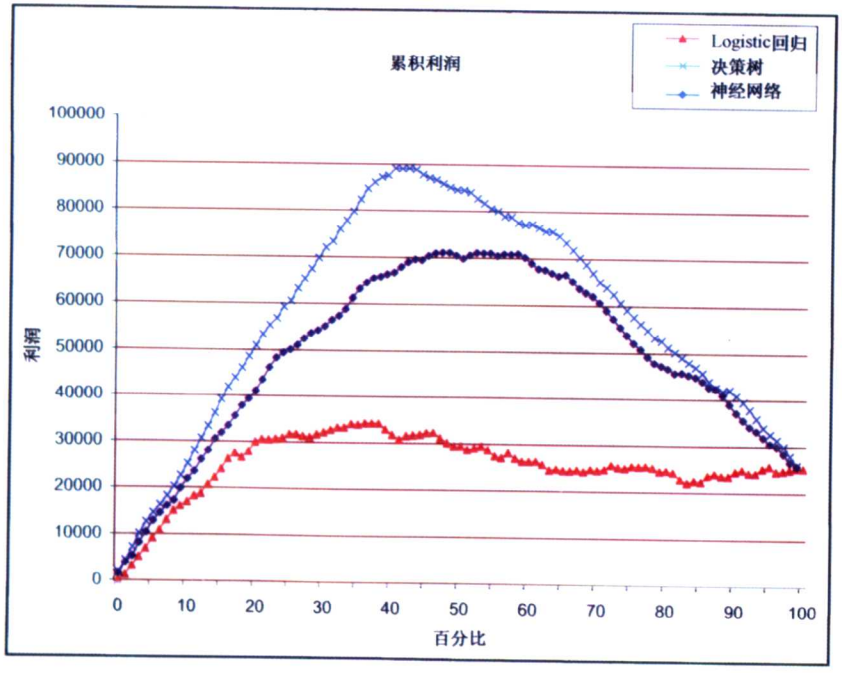


插图 12　累积利润图

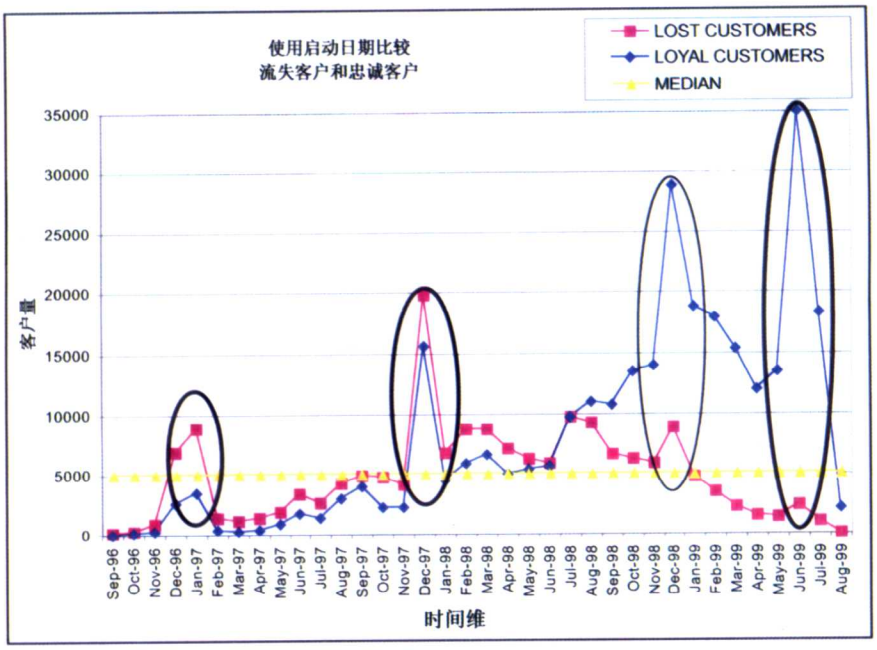
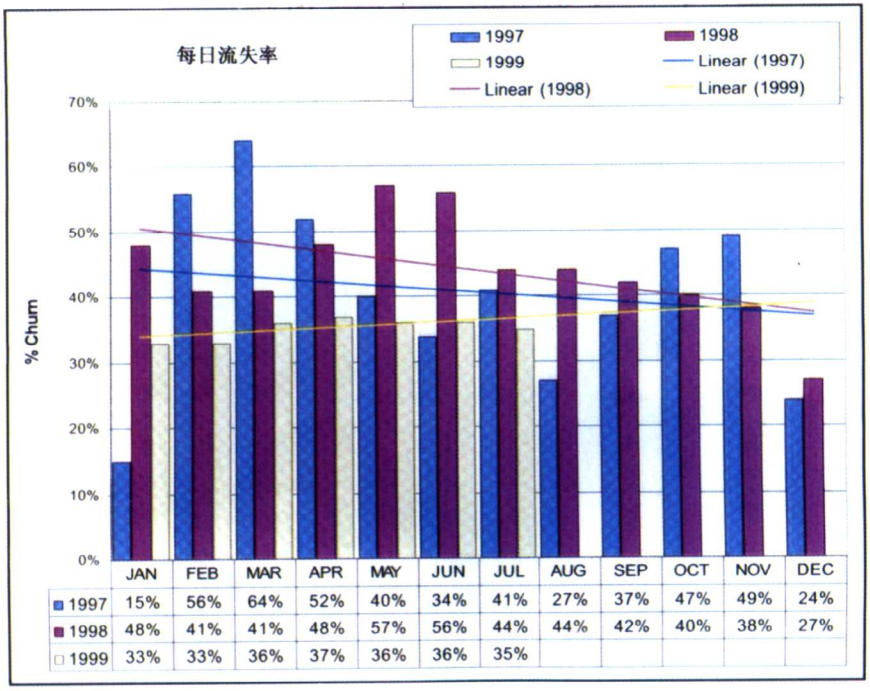


插图 13　使用启动日期比较流失客户和忠诚客户



	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1997	15%	56%	64%	52%	40%	34%	41%	27%	37%	47%	49%	24%
1998	48%	41%	41%	48%	57%	56%	44%	44%	42%	40%	38%	27%
1999	33%	33%	36%	37%	36%	36%	35%					

插图 14　每月流失率趋势对比

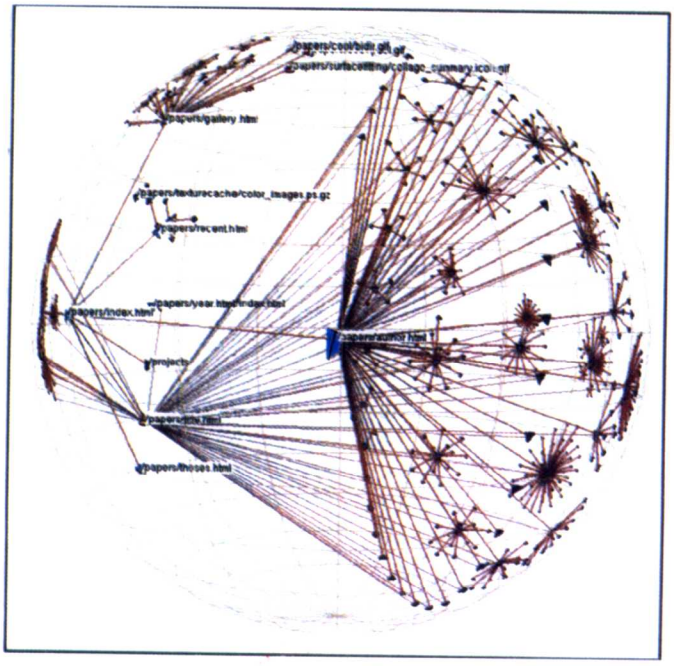


插图 15　Web 站点结构三维双曲线树型可视化（经过 T.Munzer 允许后使用）

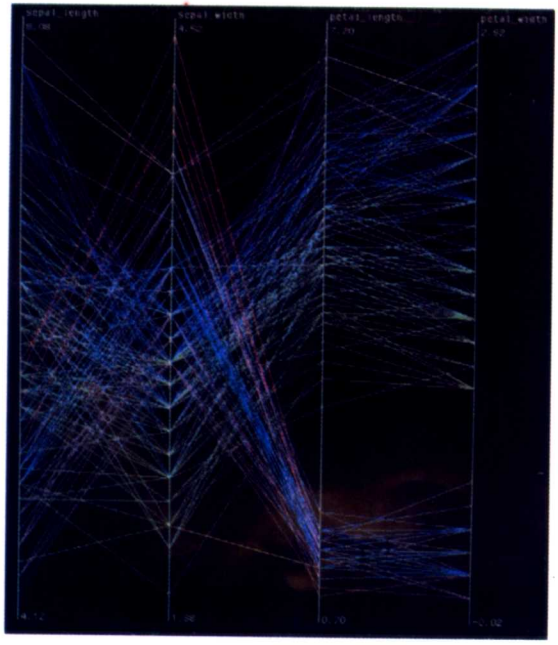


插图 16　平行坐标可视化（经 Matthew O.Ward 允许后使用）

此为试读，需要完整PDF请访问：www.ertongbook.com

对本书的一些赞誉

“对于需要或正在构建可视化数据挖掘系统的人来说，本书是非常好的材料和重要的资源。它将一些经过深思熟虑的示例和非常实用的建议紧密地结合起来。”

Stephen G. Eick
Visual Insights 公司技术总监

“当我接收到一个新的数据集时，我所做的第一件事就是用我喜爱的数据挖掘可视化工具检查其变量的分布。因为可视化信息思考的过程是如此快速和直观，以至于我能立刻分辨出哪些变量存在可疑的异常数据，哪些变量只有单个值，哪些变量存在很多空值，哪些变量可以作为连续值等，而且这些仅仅是刚开始！正如本书所展示的那样，可视化在数据挖掘过程的每一步中都起到非常重要的作用。Soukup 和 Davidson 每个例子提供了作为样本的 SQL 代码，从而带领读者穿越数据挖掘过程的每一个细节。实际上，本书中关于项目计划、数据抽取、转换和清洗的许多建议适用于所有的数据挖掘项目，而不仅仅局限于可视化。”

Michael J.A.Berry
Data Miners 公司创办人

“数据挖掘是一柄双刃剑。可视化使得数据挖掘工具的能力更加强大，分析人员使用如此强大的工具，有时能做出超越建模界限的工作。Soukup 和 Davidson 的书不仅极好地指导了如何协调这种关系，而且并没有滥用可视化的能力。他们对整个数据挖掘过程，从原始数据处理到得出最后的结论，所给出的全面的观点，是所有数据挖掘者值得遵循的典范。”

Thomas Warden
Allstate 研究和规划中心副总裁助理

献给 Ed 和我的家庭，感谢他们的鼓励。

——TOM

献给我的妻子和父母，感谢他们的支持。

——IAN

致　谢

如果缺少大家的帮助，本书是无法完成的。

我们首先要感谢评论家们对我们工作及时的批评，同时感谢我们的编辑Emilie Herman，他很有经验地带领我们完成了本书的写作。

我们要感谢Oracle网络和SPSS公司，他们分别提供了Oracle和Clementine的评估拷贝。这些产品帮助我们在本书中示范了关键的概念。

最后，我们都从Silicon图形公司的数据挖掘项目中学习到许多有用的知识。连同我们其他数据挖掘项目的经验，形成了本书中所提出的可视化数据挖掘方法。

Tom Soukup and Ian Davidson

我非常真诚地感谢和我一起从事数据挖掘项目的人员，是他们给我示范并教会我许多成功数据挖掘项目的方方面面。

Ian Davidson

我要感谢所有和我一起参加数据挖掘和商业智能项目的同伴，他们的聪明和见识帮助我形成了一套成功的可视化数据挖掘方法。

Tom Soukup

关 于 作 者

Tom Soukup 是数据挖掘和数据仓库专家，在数据管理和分析方面有 15 年多的丰富经验。目前效力于 Konami Gaming 系统公司，是商业智能主管和数据库管理员。

Ian Davidson，已经参与了多个商业数据挖掘项目，比如交叉销售、客户保留、汽车索赔及信用卡欺诈甄别。他最近加入了奥尔巴尼的纽约大学，目前是计算机科学系的助教。

商　　标

Microsoft，Microsoft Excel 及 PivotTable 是美国或其他国家 Microsoft 公司的注册商标或者商标。

Oracle 是 Oracle 公司的注册商标。

SPSS 是 SPSS 公司的注册商标，Clementine 和 Clementine Solution Publisher 也是 SPSS 公司的注册商标或商标。

MineSet 是 SGI 公司的注册商标。

出 版 说 明

如果没有对海量数据进行科学分析的能力，沃尔玛的老板再精明，也绝对想不到“啤酒与尿布”这两个风马牛不相及的东西之间还有着千丝万缕的联系。而将它们放在一起，竟然增加了啤酒销量，可见数据分析的巨大威力。

信息系统数年中收集了海量数据，且数据还正以指数级增长，企业迫切地需要高效、精确、科学地分析数据，以找出其背后的寓意，进而了解企业的经营状况和外部环境，做出科学的决断，在现代激烈的竞争中胜出。所以，如何将数据点石成金，更是摆在我们面前很现实也很诱人的一个问题。

现在，很多人已经意识到数据中潜在的大量商机，并踏踏实实地进行着从数据中沙里淘金的工作。特别是在信息化的大潮中，上至政府，下到企业，从银行到电信，再到网站、超市，人们都希望用数据分析这根魔杖赢得先机。与此同时，人们也在企盼着相关书籍，以便工作中学习参考。在广泛征询专家和用户的基础上，秉着选题全面、内容经典、译者严谨的原则，我们适时地推出了这套《数据仓库与数据挖掘技术应用丛书》，以飨读者。本丛书有如下几本：

- 数据仓库基础
- OLAP 解决方案：多维信息系统的构建技术
- 数据仓库工具箱：维度建模的完全指南（第二版）
- 数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法
- 数据仓库及其在电信领域中的应用
- 疑难数据仓库专家解决方案
- IBM 数据仓库和商业智能工具
- 可视化数据挖掘：数据可视化和挖掘的技术与工具
- 点击流数据仓库
- Web 数据挖掘：将客户数据转化为客户价值
- 企业信息工厂
- 机器学习与数据挖掘：方法和应用

本丛书既包括商业智能（BI）的基础——数据仓库（DW），也包括数据仓库上的两类不同目的的数据增值操作——联机分析处理（OLAP）和数据挖掘（DM）；既覆盖基础理论，如数据仓库基础，又提供不同领域的解决方案，

此为试读，需要完整PDF请访问：www.ertongbook.com