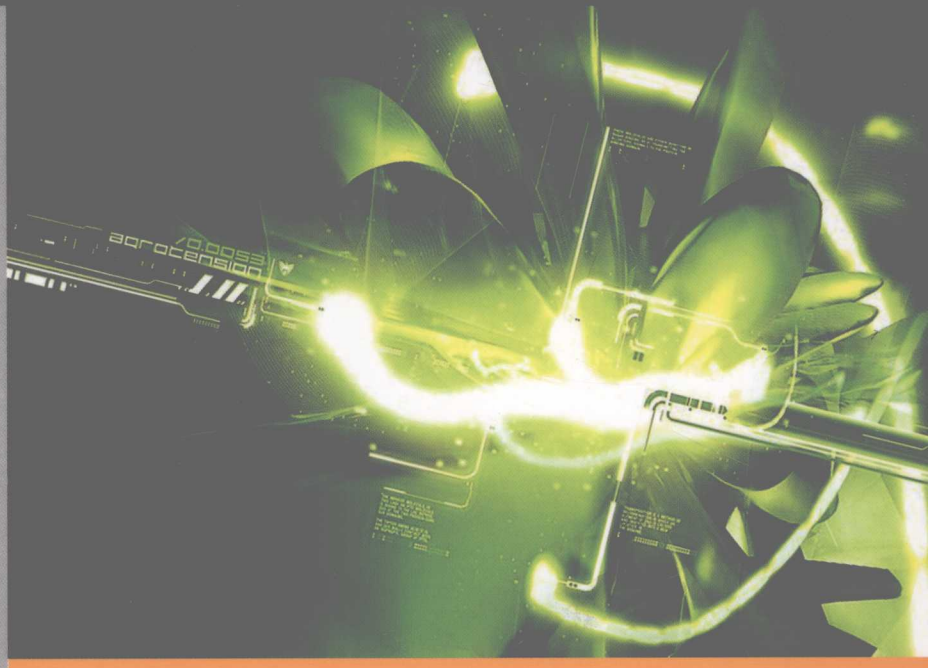
祖巧红 编著

面向物流企业数据在线分析挖掘及应用



科学出版社
www.sciencep.com

面向物流企业数据在线分析挖掘及应用

祖巧红　编著

科学出版社

北京

内 容 简 介

本书对数据挖掘及联机分析理论体系进行了概述，通过案例示范了数据挖掘的各个环节，并结合物流企业的三个综合案例进行了数据挖掘及联机分析理论的系统设计和应用。全书共分9章，第1～3章介绍了数据挖掘的基本理论体系，对数据挖掘常用算法及相关理论的发展过程进行了总体阐述；第4～6章针对数据挖掘过程的各个环节进行了理论阐述，并通过案例建立和检验数据挖掘过程；第7～9章介绍了三个综合案例，设计并实现了一个联机客户分析挖掘系统，构建了一个面向SOA的数据挖掘服务平台，研究数据挖掘算法、联机分析挖掘及其多维可视化技术在物流企业、制造业辅助决策方面的实际应用。

本书可供从事物流工程、物流管理、制造业信息化、计算机应用等领域的相关高校师生参考，也适合对复杂海量信息处理有兴趣的专业技术研究人员使用。

图书在版编目(CIP)数据

面向物流企业数据在线分析挖掘及应用／祖巧红编著．—北京：科学出版社，2009

ISBN 978-7-03-025096-4

Ⅰ．面… Ⅱ．祖… Ⅲ．物资企业－企业管理－数据采集 Ⅳ．F253

中国版本图书馆CIP数据核字(2009)第128813号

责任编辑：耿建业 王向珍／责任校对：陈丽珠

责任印制：赵 博／封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2009年7月第 一 版 开本：B5（720×1000）

2009年7月第一次印刷 印张：17

印数：1—2 500 字数：333 000

定价：50.00元

（如有印装质量问题，我社负责调换）

前　言

数据挖掘的目的在于使用所发现的模式帮助解释当前的行为或预测未来的结果。数据挖掘技术涉及下列各方面：数据仓库及在线分析处理技术、数据预处理、挖掘工具选择、关联规则、分类和预测、聚类分析及时序和序列数据的挖掘。

本书分为三个部分：

第一部分：数据挖掘理论体系研究

第一部分即本书前3章内容。介绍了数据挖掘的基本理论体系，对数据挖掘常用算法、数据仓库、在线分析处理等相关理论及应用进行了总体阐述，为本书的其余部分奠定理论基础。

第二部分：数据挖掘过程阐述

第二部分即第4～6章，针对数据挖掘过程的各个环节，如数据预处理、数据仓库的构建、报表的制作进行了必要的理论阐述，提供了详细的基于SQL Server 2005软件环境的操作实例，并通过案例建立和检验数据挖掘过程。

第三部分：数据挖掘实例设计与实现

本书侧重研究在线分析处理和数据挖掘结合的在线分析挖掘及其多维可视化技术，本书第三部分即第7～9章通过三个综合案例着重研究数据挖掘算法、在线分析挖掘及其多维可视化技术在物流企业、制造业辅助决策方面的实际应用。

第7章研究了基于关联规则的在线分析挖掘（OLAM）及其多维可视化的若干关键技术。基于某企业实际销售主题数据仓库，对产品销售的数据进行了序列关联分析并将其可视化，剖析了销售产品之间的关联规律，为企业的促销策略等决策支持提供依据。对产品销售额、订单量等交易数据通过多维度多层次的上卷、下钻、横切、纵切等在线分析，以可视化、可理解的方式剖析了深层的客户属性因素。

第8章结合某企业实际数据源，构建、设计并实现了一个基于客户分析主题数据仓库的联机客户分析挖掘系统，将基于粗糙集的模糊评价算法、马尔可夫预测模型、数据挖掘算法集成到系统中，实现客户的终生价值、资信计算及客户忠诚度预测，对客户进行细分，深入分析产品销售规律，并能实现OLAM的可视化分析。作为一个工具平台，该系统对客户的进一步管理及研究提供了有力的支

持，为管理分析及决策层提供了有力的决策支持。

第9章的综合案例构建了面向SOA的数据挖掘服务平台，以物流企业为例，通过分析数据挖掘技术在物流行业中的应用，探讨数据挖掘服务如何设计，并将数据挖掘服务划分为若干个粗粒度的子服务。基于SOA的思想，研究基于WCF技术架构数据挖掘服务，包括数据挖掘服务中各子服务的设计、架构及实现等技术，在该服务平台中综合实现数据（大文件）上传服务、数据清洗服务、在线分析处理服务及数据挖掘算法服务，并重点集成基于SOA的数据挖掘服务在物流运输中的调用，在基于Web网页的三维电子地图系统中，通过调用数据挖掘服务，实现物流运输中的路径选择、路径优化等决策支持功能。

本书能够撰写完成，得益于武汉理工大学陈定方教授的悉心指导，李文锋教授的热心帮助，非常感谢他们长期以来的鼓励和支持！另外，武汉理工大学物流信息与控制工程研究所研究生张波、贾爱华、高海耀、吴婷、王慧、黄雄伟、李宁、杜海涛等在案例实现与整理、文字校对等方面做了大量具体的工作，在此表示深深的感谢！

限于作者的水平，书中难免有不妥之处，敬请读者批评指正。

作　者

2009年3日

目　录

第1章　数据挖掘原理

1.1　知识发现与数据挖掘

数据挖掘和知识发现是随着数据库和机器学习的发展而兴起的。在20世纪80年代末出现了一个新的术语，它就是数据库中的知识发现（Knowledge Discovery in Databases，KDD）。KDD将信息变为知识，从数据矿山中找到蕴藏的知识金块，是从元数据中发掘模式的方法。人们接受了这个术语，并用KDD来描述整个数据挖掘的过程，包括最开始的制定业务目标到最终的结果分析，而用数据挖掘（Data Mining，DM）来描述使用挖掘算法进行数据挖掘的子过程。

数据挖掘的定义为：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。数据挖掘提取的知识可以表示为概念、规律、模式、约束和可视化。数据挖掘算法的好坏将直接影响到发现知识的好坏，数据挖掘的任务是从数据中发现模式。

KDD的定义为：从大量数据中提取出可信的、新颖的、有用的并能被人理解的模式的高级处理过程。模式可以看成是知识的雏形，经过验证、完善后形成的知识。KDD是一个高级的处理过程，它是从数据集中识别出用模式表示的知识。高级的处理过程是指一个多步骤的处理过程，多步骤之间相互影响、反复调整，形成一种螺旋式的上升过程。KDD是一门交叉学科，涉及人工智能、机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化和专家系统等多个领域。

严格地说，KDD被认为是从数据中发现有用知识的整个过程，而数据挖掘指的是KDD整个过程中的一个特定步骤，是KDD中最核心的部分。然而在通常情况下，许多人把数据挖掘与KDD广泛地认为是同一个概念。一般在科研领域中称为KDD，而在工程领域则称为数据挖掘。

1.2　数据挖掘概论

1.2.1　数据挖掘的对象和任务

1）数据挖掘的对象

数据挖掘可以在任何类型的数据上进行，既可以是来自社会科学的数据，又

可以是来自自然科学的数据，还可以是卫星观测得到的数据。数据形式和结构也各不相同，可以是传统的关系数据库、面向对象的高级数据库系统，也可以是面向特殊应用的数据库，如空间数据库、时序数据库、文本数据库和多媒体数据库等，还可以是 Web 数据信息。

2）数据挖掘的任务

数据挖掘的目标是从海量数据中发现隐含的、有意义的知识。它的任务主要是进行分类、预测，建立时间序列模式，进行聚类分析、关联分析预测和偏差分析等。

（1）分类。分类就是按照一定的标准把数据对象划分成不同类别的过程。

（2）预测。预测就是通过对历史数据的分析找出规律，并建立模型，通过模型对未来数据的种类和特征进行分析。

（3）时间序列模式。时间序列模式就是根据数据对象随时间变化的规律或趋势来预测将来的值。

（4）聚类分析。聚类分析是在没有给定划分类型的情况下，根据数据信息的相似度进行数据聚集的一种方法。

（5）关联分析预测。关联分析预测就是对大量的数据进行分析，从中发现满足一定支持度和可信度的数据项之间的联系规则。

（6）偏差分析。偏差分析就是通过对数据库中的孤立点数据进行分析，寻找有价值和意义的信息。

1.2.2 数据挖掘的应用

数据挖掘从一开始就是面向应用的技术。目前，在很多领域数据挖掘都是一个很流行的词。在政府管理决策、商业经营、科学研究、农业、工业企业决策、军事、医学、天文地理等各个领域都应用到了数据挖掘技术，尤其是在如银行、电信、保险、交通、零售等商业领域。

1）在科学研究中的应用

从科学研究方法学的角度看，科学研究可分为三类：理论科学、试验科学和计算科学。计算科学是现代科学的一个重要标志。计算科学工作者主要和数据打交道，每天要分析各种大量的试验或观测数据。随着先进的科学数据收集工具的使用，如观测卫星、遥感器、DNA 分子技术等，数据量非常大，传统的数据分析工具无能为力，因此必须有强大的智能型自动数据分析工具才行。

数据挖掘在天文学上有一个非常著名的应用系统：SKICAT（Sky Image Cataloging and Analysis Tool）。它是美国加利福尼亚理工学院喷气推进实验室（即设计火星探测器漫游者号的实验室）与天文科学家合作开发的用于帮助天文学家发现遥远的类星体的一个工具。SKICAT 既是第一个相当成功的数据挖掘应用，也

是人工智能技术在天文学和空间科学上第一批成功应用技术之一。利用 SKICAT，天文学家已经发现了 16 个新的极其遥远的类星体，该项发现能帮助天文工作者更好地研究类星体的形成以及早期宇宙的结构。

数据挖掘在生物学上的应用主要集中于分子生物学特别是基因工程的研究。在基因研究中，有一个著名的国际性研究课题——人类基因组计划。据报道，1997 年 3 月，科学家宣布已完成第一步计划：绘制人类染色体基因图。然而这仅仅是第一步，更重要的是对基因图进行解释从而发现各种蛋白质（有 10000 多种不同功能的蛋白质）和 RNA 分子的结构和功能。近几年，通过用计算生物分子系列分析的方法，尤其是通过基因数据库搜索技术人们已在基因研究上作出了很多重大发现。

2）在商业上的应用

在商业领域特别是零售业，数据挖掘的运用是比较成功的。由于 MIS 系统在商业的普遍使用，特别是条码技术的使用，可以收集到大量关于购买情况的数据，并且数据量在不断激增。数据挖掘技术可以为经营管理人员决策提供有力的依据，这样对促进销售及提高竞争力是大有帮助的。

3）在金融上的应用

在金融领域，数据量非常巨大，银行、证券公司等交易数据和存储量都很大。而由于信用卡欺诈行为，银行每年的损失非常大。因此，可以利用数据挖掘对客户信誉进行分析。典型的金融分析领域有投资评估和股票交易市场预测。

4）在医学上的应用

数据挖掘在医学上的应用十分广泛，从分子制药到医疗诊断都可以利用数据挖掘的手段来提高效率和效益。在药物合成方面，通过对药物分子化学结构的分析，可以确定药物中哪种原子或原子基因对什么病能够发挥作用，这样在合成新药时，可根据新药的分子结构确定该药将有可能治疗哪一种疾病。

5）在学校教育中的应用

数据挖掘可以帮助学校分析学生历史信息，决定哪些人愿意报考何种专业，发送手册给他们；分析教师的学历、年龄、职称等与授课效果的关联规则，制订教学方案，提高教学质量。

数据挖掘还可用于工业、农业、交通、电信、军事、IT 等其他行业。数据挖掘具有广泛的应用前景，它既可应用于决策支持，也可应用于数据库管理系统（DBMS）。数据挖掘作为决策支持和分析的工具，还可以用于构造知识库。在 DBMS 中，数据挖掘可以用于语义查询优化、完整性约束和不一致检验等。

1.2.3 在线分析数据挖掘系统、软件工具

数据挖掘技术已经从最开始的一个简单的算法包发展到通用挖掘平台和专业挖掘工具两大种类。其中，像 IBM、NCR、SAS、微软、SPSS、StatSoft 等厂商的数据挖掘产品（模块）基本都是通用型工具平台；而像美国的 Unica 公司、费尔艾萨克公司（Fair Isaac Corporation）则主要专注于营销自动化、信用卡积分等细分领域，其产品属于后一种工具。具体来看，目前在数据挖掘领域声势颇大的大多是通用型工具平台。

NCR 的数据挖掘工具是与其数据仓库整合在一起的。具体来说，其数据挖掘工具可以按照挖掘的步骤主要分成 Profiler（数据探查器）、ADS Generator（分析数据集成生成器）、Warehouse Miner（数据仓库挖掘器）和模型管理器四块。目前 Teradata 最新版的数据挖掘方案是 Teradata Warehouse Miner 5.1。

SAS 公司和 SPSS 公司作为两家从传统的统计分析技术发展而来的数据挖掘厂商，二者在业内的影响力可谓有目共睹。其中，SAS 公司提供了 SAS Enterprise Miner（企业数据挖掘）、SAS ETS（时间序列预测）、SAS OR（运筹学）、SAS STAT（统计分析）、SAS QC（质量控制）等一系列工具，SPSS 公司也提供了 Clementine 和 AnswerTree 两项产品。

微软 SQL Server 2005 在数据挖掘方面的突破与创新曾被人看做最惊艳的地方。Microsoft SQL Server 2005 Data Mining 平台引入了大量的数据挖掘功能，其本身就是一个开发智能应用程序的平台，而非一个独立应用程序。而且，这一平台与所有 SQL Server 产品实现了集成，包括 SQL Server、SQL Server Integration Services 和 Analysis Services。SQL Server 2005 中最重要的数据挖掘功能就是其处理大型数据集的能力，它允许模型对整个数据集运行，从而消除了采样方面的挑战。另外，微软在数据仓库市场倡导了另一个概念——数据集市。数据集市就是一个面向部门应用的、小型的数据仓库，它所采用的技术与数据仓库相似，但存储的内容更加专题化。对于数据集市这样的规模，微软的解决方案便可成为理想的选择。

总体来看，像 IBM、NCR、Oracle、微软这些平台工具厂商基本上都是以提供“整车”为己任。数据挖掘的工具经过近十年的发展已经得到相当广泛的应用，但是，目前多数数据挖掘软件采用的技术和功能比较有限，难以胜任复杂数据挖掘任务。

当前具有代表性的数据挖掘软件，一类是基于统计分析的软件，如 SAS、SPSS 等；另一类是应用新技术，如模糊逻辑、人工神经网络、决策树理论的工具，如 CBR Express、Esteen、Kate-CBR、FuzzyTECH for business、Aria、Neural network

Browser 等软件。这些软件并不是包罗万象地应用于任何数据挖掘技术的软件，而是有所侧重。

与国外成熟的软件相比，国内在数据挖掘技术方面起步相对较晚，数据挖掘的软件运用和开发还未全面展开，尤其是模糊逻辑、人工神经网络、决策树等，对数据挖掘工具的开发不足。因此，开拓数据挖掘工具的应用和实践是未来数据挖掘工作中亟待解决的问题。

目前，数据挖掘工具正以前所未有的速度发展，并且不断扩大用户群体。在未来愈加激烈的市场竞争中，拥有数据挖掘技术必将比别人获得更快速的反应，赢得更多的商业机会。

数据挖掘是一个交叉学科领域，受多个学科影响，包括数据库系统、统计学、机器学习、可视化和信息科学，因此，就产生了大量的、各种不同类型的数据挖掘系统。对数据挖掘系统分类可以帮助用户区分数据挖掘系统，确定最适合其需要的数据挖掘系统。

数据挖掘软件经历了或即将经历四个时代。第一代数据挖掘系统仍未发展完全，第二代、第三代数据挖掘系统已经出现。目前未见到任何第四代数据挖掘系统的报道。集成第二代、第三代以及第四代数据挖掘和预言模型系统，将其与数据仓库合并，以提供一个集成的系统来管理日常的商业过程。

从表 1-1 可以看出，普适计算时代的数据挖掘软件是数据挖掘软件发展的高级阶段。目前，大量数据挖掘软件仍处于第三代，而且有待继续发展。

表 1-1　数据挖掘软件经历的四个时代

代	特征	数据挖掘算法	集成	分布计算模型	数据模型
第一代	作为一个独立的应用	支持一个或者多个算法	独立的系统	单个机器	向量数据
第二代	和数据库以及数据仓库集成	多个算法，能够挖掘一次不能放进内存的数据	数据管理系统，包括数据库和数据仓库	同质、局部区域的计算机群集	有些系统支持对象、文本和连续的媒体数据
第三代	和预言模型系统集成	多个算法	数据管理和预言模型系统	Intranet/extranet 网络计算	支持半结构化数据和 Web 数据
第四代	和移动数据/各种计算设备的数据联合	多个算法	数据管理、预言模型、移动系统	移动和各种计算设备	普遍存在的计算模型

目前国内的数据挖掘系统仍然存在许多不足，主要表现在以下两个方面。

一方面，随着数据挖掘工具日益广泛的应用，人们发现有些工具只有精通数

据挖掘算法的专家才能熟练使用。如果对算法不了解，难以得出好的模型。所以迫切需要一类使用简单而又具有针对性、功能良好的数据挖掘软件。

另一方面，国内对数据挖掘方面的算法和理论研究较多，而对数据挖掘软件和工具的设计与实现方面的研究相对较少。

数据挖掘工具一共经历了三个阶段，分别如下所述。

（1）独立的数据挖掘软件（1995 年以前）。该阶段对应第一代数据挖掘系统，出现在数据挖掘技术发展早期，研究人员开发出一种新型的数据挖掘算法，就形成一个软件。这类软件要求用户对具体的算法和数据挖掘技术有相当的了解，还要负责大量的数据预处理工作。

（2）横向的数据挖掘工具集（1995 ~ 1999 年）。此类工具集的特点是，提供多种数据挖掘算法，包括数据的转换和可视化。由于此类工具并非面向特定的应用，是通用的算法集合，因此，称之为横向的数据挖掘工具。

（3）纵向的数据挖掘解决方案（1999 年开始）。此类工具的特点是，针对特定的应用提供完整的数据挖掘方案，因此，称之为纵向的数据挖掘解决方案。

在目前的数据挖掘应用中，企业往往都希望针对自身特点，定制与之相适用的数据挖掘软件，如在证券系统或电信系统中嵌入神经网络预测功能；在客户关系管理系统中嵌入客户分类功能或客户行为及客户流失分析功能；在电子商务中嵌入选择最可能购买产品的客户功能；在物流管理系统中嵌入运费预测、货运优化功能等。当前的数据挖掘软件多数都向着纵向的数据挖掘解决方案方向发展。

最近，Gartner Group 的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。根据最近 Gartner 的 HPC 研究表明：“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。”

当前，中国关于数据挖掘研究工作的机构也相继出现，如中国人民大学统计系数据挖掘中心、台湾辅仁大学管理学院创新育成中心，近几年均从事数据挖掘的研究工作。从事数据挖掘研究的代表性专家有谢邦昌教授、张尧庭教授、朱世武教授、韦端博士、陈江教授、赵民德教授等。除此以外，国内部分高校一些学者也相继加入到数据挖掘的理论和应用研究的队伍中来。从数据挖掘的研究队伍来看，大部分是统计方面的专家学者，而从数据挖掘的学科要求来看，数据挖掘是集统计学、计算机科学、金融学以及实际应用方面的理论为一体的综合学科，因此，数据挖掘的理论和应用研究也需要大批其他学科的专家学者加入，与统计学家共同为数据挖掘的理论研究和应用研究作出贡献。

与国外相比，国内对数据挖掘与知识发现（DMKD）的研究稍晚，没有形成整体力量。1993 年国家自然科学基金首次支持该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中国科学院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工程研究所对模糊方法在 KDD 中的应用进行了较深入的研究；北京大学也开展了对数据立方体代数的研究；华中科技大学、复旦大学、浙江大学、中国科技大学、中国科学院数学研究所、吉林大学等单位开展了对关联规则挖掘算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web 数据挖掘。

对数据挖掘的研究主要体现在以下几个方面：数据挖掘中的隐私保护，即在数据挖掘的过程中，在得到预想结果的前提下，又不透露任何不该泄漏的资料；对 KDD 方法的研究进一步发展，如近年来注重对贝叶斯（Bayes）方法以及 Boosting 方法的研究和提高；传统的统计学回归法在数据挖掘中的应用等。

最近，数据挖掘研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多个学科之间的相互渗透。国外很多计算机公司非常重视数据挖掘系统的开发应用，IBM 与微软都成立了相应的研究中心，进行这方面的工作。许多著名的计算机公司开始尝试着 KDD 软件的开发，比较典型的有 SAS 公司的 Enterprise Miner，IBM 公司的 Intelligent Miner 等。

当今信息技术正在从数据处理向数据使用方向转变，人们在拥有海量数据的同时却苦于信息的缺乏。能迅速地提供更准确、高质量的信息，已成为人们当前迫切需要解决的问题。在此情况下，在线分析处理（On-Line Analytical Processing，OLAP）与数据挖掘技术就成为当前信息领域的研究重点。OLAP 与数据挖掘具有各自的内在技术和应用范围，在决策分析中，如何将它们结合起来协调使用，发挥其最佳作用和价值是本研究的初衷。OLAP 的主要工作是将数据仓库的数据转换到多维数据结构中，并且调用多维数据集（Cube）来执行有效且非常复杂的查询。OLAP 是一种自上而下、不断深入的技术，用户提出问题或假设，OLAP 负责从上到下深入提取关于该问题的详细信息，并以可视化的方式提供给用户。

目前，我国出版有关数据挖掘方向的具有代表性的著作有：Han 和 Kamber 的《数据挖掘概念与技术》，2001 年机械工业出版社出版；张尧庭、谢邦昌、朱世武的《数据挖掘入门及应用———从统计技术看数据挖掘》（电子版）。具有代表性的论文主要有：朱世武等“据挖掘与其他技术比较”，《统计研究》2003 年第 7 期；赵黎明等“基于数据挖掘的专利引文研究与知识发现”，《预测》2002 年第 6 期；张喆等“数据挖掘技术在 CRM 中的应用”，《中国管理科学》2003 年第 1 期；《统计与信息论坛》2002 年刊登了中国人民大学统计学系数据挖