

<<海量数据挖掘技术研究>>

图书基本信息

书名：<<海量数据挖掘技术研究>>

13位ISBN编号：9787811402582

10位ISBN编号：7811402580

出版时间：2010-12

出版时间：浙江工商大学出版社

作者：刘君强

页数：176

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## &lt;&lt;海量数据挖掘技术研究&gt;&gt;

## 内容概要

随着信息技术特别是网络技术的飞速发展，人们收集、存贮、传输数据的能力不断提高。数据出现了爆炸性增长，与此形成鲜明对比的是：对决策有价值的知识却非常匮乏。

知识发现与数据挖掘技术正是在这一背景下诞生的一门新学科。

数据挖掘要在实际应用中发挥作用，高性能挖掘算法和数据挖掘软件平台是重要的技术基础。

本书以数据挖掘最基本问题、频繁模式与关联规则挖掘为切入点，研究高时间效率、高空间可伸缩性的挖掘算法和分布，异质、海量数据的协同挖掘软件模型，并探讨了数据挖掘过程中的隐私保护问题。

本书首先发现了基于树表示形式的虚拟投影方法，用于按深度优先挖掘密集型数据集；提出了稀疏型数据集表示形式及非过滤投影方法；进一步提出了基于伺机投影的思想，设计并实现了基于伺机投影的全新算法OpportuneProject，对比实验表明，该算法挖掘各种规模与特性数据库的效率与可伸缩性都是最佳的。

由于其内在的计算复杂性，挖掘密集型数据的频繁模式完全集非常困难，解决办法是挖掘频繁模式的闭合集或最大集。

本书提出了一种组织闭合模式集的复合型频繁模式树，支持搜索空间的高效剪裁，有效地平衡了树生成与树剪裁的代价，实现了闭合模式集挖掘算法CROP，其效率与可伸缩性大大优于CHARM等算法。在此基础上，本书提出了闭合性剪裁和一般性剪裁相结合，并能适时前窥的最大模式挖掘算法MOP，大大优于MaxMiner和MAFIA等算法。

本书进一步提出了逆字典树剪裁、层次标记等新技术，以及根据信息熵自动生成与人机交互相结合来确定数值型与类别型属性概念层次的新方法，不仅支持逐层挖掘，而且能进行跨层挖掘，并实现了多支持率剪裁，将所提出的挖掘频繁模式完全集、闭合集的新算法推广到无冗余关联规则、多维多层多数据类型关联规则、多支持率分类规则的挖掘问题。

本书在所取得的数据挖掘算法研究成果基础上，对数据挖掘软件模型作了深入研究。

首先提出了数据挖掘作业描述语言MDL和挖掘任务模型脚本语言，设计并实现了一个集成数据仓库管理功能、挖掘引擎具有一定智能、体系结构可扩展的数据挖掘工具。

本书在研究分布式问题求解技术和分析移动型智能代理技术的基础上，提出了从网络海量数据中发现有用知识的协同挖掘模型。

首先定义了黑板和知识源的描述语言以及知识交换格式，设计和实现了支持互联网上分布式问题求解的黑板系统，提出了分布式网络海量数据挖掘系统DistributedMiner。

接着在分析移动式智能代理技术的基础上，设计了一种移动式智能代理服务器，通过重构基础结构，提出了移动式网络海量数据挖掘系统模型MobileMiner。

最后，本书研究了挖掘事务型数据过程中的隐私保护问题。

由于事务型数据的极度稀疏性，任何单一技术难以有效发挥作用，或是导致过高的信息损失，或是处理结果难以解释，或是技术自身性能有缺陷。

本书提出了集成概化技术与消隐技术来降低信息损失。

然而，从技术上讲，集成并非易事。

本书提出了一种新颖的方法来解决效率与可伸缩性的问题。

采用此方法处理过的数据能够应用标准的数据挖掘工具进行分析。

## <<海量数据挖掘技术研究>>

### 作者简介

刘君强，男，教授，浙江省杭州市人。

毕业于加拿大Simon Fraser大学获哲学博士学位，浙江大学获工学博士学位和管理学硕士学位，北京大学获理学学士学位。

以第一作者在ACM SIGKDD、IEEE ICDE、IEEE ICDM、PAKDD、《电子学报》、《计算机学报》、《软件学报》、《计算机研究与发展》、《中国图形图象学报》、《系统工程理论与实践》等发表多篇论文，承担省部级研究课题多项。

研究兴趣涉及数据挖掘、网络信息安全、隐私保护、管理信息系统、软件工程。

## 书籍目录

前言摘要ABSTRACT第一章 概论第一节 数据挖掘技术的兴起第二节 数据挖掘的主要问题一、数据挖掘任务与知识类型二、数据挖掘的过程三、数据挖掘的对象四、数据挖掘的应用五、数据挖掘面临的挑战第三节 本书的工作第四节 本书的结构第二章 数据挖掘技术综述第一节 频繁模式与关联规则挖掘一、单层单维布尔型关联规则挖掘与Apriori算法二、对Apriori算法的改进三、频繁模式与关联规则挖掘研究的新发展第二节 闭合模式挖掘与A-Close算法一、闭合模式挖掘与A-Close算法二、其他闭合模式挖掘算法第三节 最大模式挖掘与Pincer-Search算法一、最大模式挖掘与Pincer-Search算法二、其他最大模式挖掘算法第四节 多层多维关联规则挖掘一、多层关联规则挖掘问题二、多维关联规则挖掘问题第五节 对关联规则挖掘的其他扩展一、顺序模式挖掘二、基于约束的关联规则挖掘三、并行挖掘问题四、复杂检索问题五、关联规则与相关性六、其他问题第六节 数据挖掘软件系统第七节 保护隐私的数据挖掘技术一、全局概化技术二、全消隐技术三、局部概化技术四、带宽矩阵方法五、其他相关工作第八节 数据挖掘技术的应用一、数据挖掘的应用领域二、企业营销应用数据挖掘技术第三章 伺机投影策略的挖掘算法第一节 引言第二节 问题的描述第三节 频繁模式树的构造第四节 模式支持集表示与投影一、稀疏型PTS的基于数组表示及其投影二、密集型PTS的基于树表示及虚拟投影第五节 伺机投影策略与OpponuneProject算法一、伺机投影的启发式原则二、估计TVLA和TTF的大小三、OpponuneProject算法第六节 性能评价一、数据集及其特性二、基本实验结果三、可伸缩性试验第七节 小结第四章 闭合模式与最大模式挖掘第一节 引言第二节 问题的描述第三节 复合型频繁模式树及其生成一、复合型频繁模式树CFIST二、CFIST结点的合并三、CFIST的生成算法第四节 CFIST的剪裁与包含关系的检查一、高效的CFIST局部剪裁二、分枝包容关系的快速检查三、快速杂凑法第五节 CROP：挖掘闭合模式的高性能算法一、平衡CFIST生成与剪裁效率二、CROP算法第六节 CROP性能测评一、CROP与CHARM效率对比二、CROP与CLOSET效率对比三、CROP与MAFIA效率对比四、可伸缩性实验第七节 挖掘最大频繁模式的新算法MOP一、最大频繁模式集及其剪裁二、MOP算法三、MOP的性能评价第八节 小结第五章 多维多层关联规则、分类规则与空间关联规则第一节 关联规则与无冗余关联规则第二节 多层频繁模式挖掘一、问题的描述二、逆字典树与多层频繁模式三、层次标记技术与模式支持集四、高性能多层频繁模式挖掘算法五、性能测评第三节 多维多层多数据类型关联规则挖掘一、多维多层多数据类型关联规则挖掘问题二、MDML-PP算法三、性能测评第四节 挖掘多支持率分类规则一、分类规则挖掘与TTF扩展二、多支持率剪裁三、分类规则及其单阶段挖掘算法四、对比实验第五节 空间关联规则的挖掘一、空间关联规则二、两阶段挖掘策略三、基于辅存分而治之的方法第六节 提高挖掘算法可伸缩性的技术一、海量数据挖掘策略二、缓冲管理技术三、挖掘算法改进及其性能分析第七节 小结第六章 智能型数据挖掘工具设计与实现第一节 引言第二节 数据仓库及其管理一、数据仓库模型与OLAP二、数据仓库的框架描述三、数据仓库管理器第三节 数据挖掘任务的描述、管理及执行机制一、数据挖掘作业Job的描述二、挖掘任务模型Scenario的定义三、挖掘任务模型的管理与执行第四节 智能型数据挖掘引擎一、算法描述库与算法模块二、知识库与引擎管理器第五节 SmartMiner体系结构第六节 关键技术与SmartMiner原型实现第七节 小结第七章 网络海量数据协同挖掘第一节 引言第二节 分布式黑板控制一、问题求解的黑板系统二、分布式问题求解与黑板控制第三节 形式化描述语言一、黑板的描述二、知识源的描述三、知识交换格式第四节 实现分布式黑板控制的一般智能代理一、智能代理GA的结构设计二、智能代理软件DBC-MA的实现第五节 分布式数据挖掘系统DistributedMiner一、分布式知识发现功能二、DistributedMiner的黑板设计三、挖掘平台体系结构四、DistributedMiner的实现与应用第六节 从分布计算到移动计算一、什么是智能代理二、智能代理的特征三、移动型智能代理四、典型mobile agent系统第七节 移动式数据挖掘系统模型一、移动型智能代理服务器二、DBC-MA变型三、MobileMiner工作流程第八节 小结第八章 挖掘事务型数据过程中的隐私保护第一节 引言第二节 隐私保护与匿名化模型第三节 集成概化与消隐技术的基本方法一、割集栅格的自顶向下贪婪法搜索二、为割集寻找一个好的消隐方案三、算法描述第四节 解决效率与可伸缩性瓶颈的关键技术一、最小隐私威胁二、多轮次求解策略第五节 信息损失与性能的实验评估一、信息损失评估二、效率评估三、可伸缩性评估第六节 小结参考文献后记



<<海量数据挖掘技术研究>>

版权说明

本站所提供下载的PDF图书仅提供预览和简介, 请支持正版图书。

更多资源请访问:<http://www.tushu007.com>