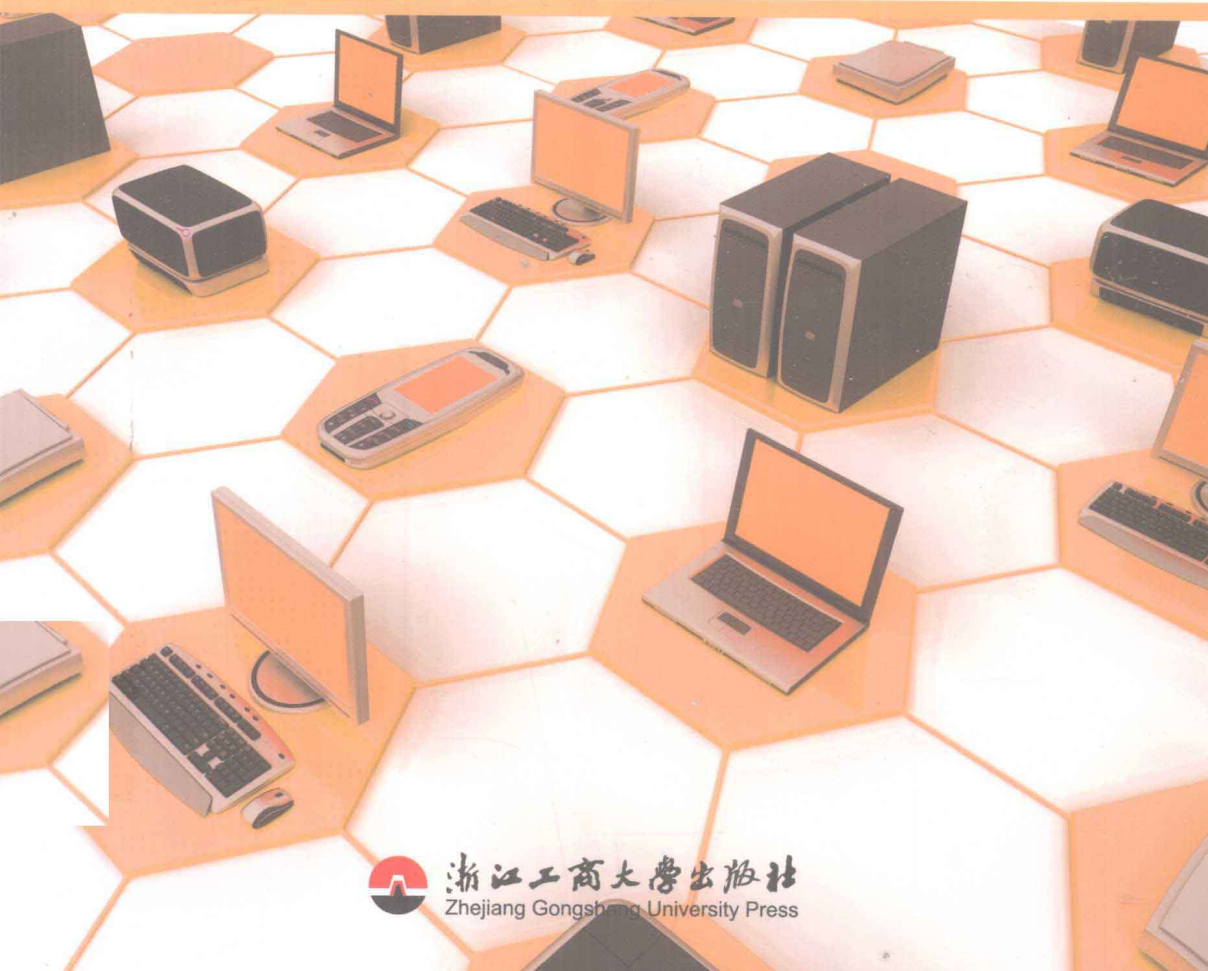
刘君强 著

海量数据挖掘技术研究

Research on Data Mining Technology for Very Large Databases



浙江工商大学出版社
Zhejiang Gongshang University Press

海量数据挖掘技术研究

刘君强 著

浙江工商大学出版社

图书在版编目(CIP)数据

海量数据挖掘技术研究 / 刘君强著.—杭州：浙江工商大学出版社,2010.12

ISBN 978-7-81140-258-2

Ⅰ.①海… Ⅱ.①刘… Ⅲ.①数据采集－研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2010)第 239406 号

海量数据挖掘技术研究

刘君强 著

责任编辑 陈维君
封面设计 刘 韵
责任印制 汪 俊
出版发行 浙江工商大学出版社
(杭州市教工路 198 号 邮政编码 310012)
(Email：zjgsupress@163.com)
(网址：http://www.zjgsupress.com)
电话：0571－88904980,88831806(传真)
排 版 杭州兴邦电子印务有限公司
印 刷 杭州广育多莉印刷有限公司
开 本 710mm×1000mm 1/16
印 张 12.25
字 数 214 千
版 印 次 2010 年 12 月第 1 版 2010 年 12 月第 1 次印刷
书 号 ISBN 978-7-81140-258-2
定 价 28.00 元

前　言

“知识就是力量”。我在1997年至1999年参加国防科技预研项目“通信对抗智能化信息融合技术”的研究期间，特别是在技术方案调研过程中，真正体会到了这句话的内涵。对数据挖掘技术的研究兴趣也正是从这时开始的。1999年至2000年，我到加拿大Simon Fraser University计算机学院做访问学者，在国际知名学者Jiawei Han的指导下，从事数据挖掘研究与开发。回国后，我的导师潘云鹤院士把高性能挖掘算法和网络协同挖掘模式研究确定为我攻读博士学位的论文主攻方向，也为本书的形成奠定了基础。

的确，要从浩瀚的数据海洋中及时发现有用的知识，数据挖掘算法的时间效率与空间可伸缩性直接决定了这项技术的有用性。然而，目前已经提出的很多挖掘算法只能挖掘小数据集和稀疏数据集，遇到海量数据集和密集型数据集往往崩溃。因此，本书把高性能的海量数据挖掘算法作为研究的第一个重点，以最基本知识类型——频繁模式与关联规则的挖掘为切入点。

本书创新性地提出了伺机挖掘的思想，指出任意数据集都不能简单地归入某个单一特性类别，其不同子集往往会有截然不同的特性，在挖掘过程中应根据局部数据子集的特性变化不断地调整挖掘方法。其具体体现是在频繁模式挖掘算法上取得了新突破。

·首次提出了密集型数据子集的虚拟投影方法，以及稀疏型数据子集的非过滤投影方法，巧妙地解决了提高时间效率与节省存贮开销相互矛盾的问题。

·提出了在挖掘过程中不断根据局部数据子集特性自动调整解空间搜索策略、决定数据子集表示形式、选择投影方法的启发式原则。

·设计并实现了全新算法OpportuneProject，其性能远优于Apriori、FP-Growth等典型算法，特别是挖掘海量数据时，其性能要高出若干数量级。

本书还创新地提出了一种表达解空间的复合型频繁模式树，以及与之相适应的解空间搜索效率与剪裁效率相平衡的原则。复合型频繁模式树与其他算法采用的字典树和概念格相比，结点数和层次少得多，且所提出的局部和全局剪裁方法具有较高的解空间剪裁效率。从而在闭合频繁模式与最大频繁模式的挖掘

算法上取得了新进展。大量实验表明，所设计与实现的算法CROP和MOP，性能分别比CHARM和MaxMiner等此前公开发表的最好算法要高5倍到2个数量级。

本书进一步提出了分解频繁模式树复合结点的方法，实现了无冗余关联规则的有效挖掘。提出逆字典树剪裁、层次标记等两项新技术，实现了直接挖掘跨层频繁模式的新算法mfpRLTL。将信息熵概念应用到确定属性的概念层次，提出了多维多层多数据类型关联规则挖掘算法MDML-PP。将多支持率剪裁引入到分类规则，提出了单阶段挖掘分类规则算法CRM-PP。这些工作拓展了伺机挖掘的思想，具有一定创新性。

随着网络技术的飞速发展，大量分散的、异质的数据在网络上以指数级的速度爆炸性地增长，要在网络中进行知识发现，唯有按照协同计算模式设计挖掘平台才有可能完成。因此，网络海量数据协同挖掘系统模型是本书第二个研究重点。

本书把知识发现系统自身的智能性、体系结构的开放性、平台的异质性作为基本标准，提出并设计实现了一个具有数据仓库管理功能的智能型数据挖掘工具，通过支持分布式问题求解的黑板控制机制集成该工具，形成了分布式数据挖掘系统模型，进一步引入移动型智能代理的思想，提出了移动式网络海量数据挖掘系统模型，主要创新点如下。

·提出了一种数据挖掘作业描述语言MDL和挖掘任务控制脚本。前者是一种类SQL语言，用于描述数据挖掘的基本作业。复杂问题通过挖掘任务控制脚本组合基本作业形成任务模型。

·提出了黑板和知识源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。提出了基于分布式黑板控制机制的分布式数据挖掘系统模型。

·提出了一种实现数据挖掘代理的生成、迁移、异地执行的移动型智能代理服务器，设计了移动式网络海量数据挖掘模型。

任何一项技术在造福人类的同时，也有可能带来一些负面的影响。数据挖掘技术广泛应用的同时，也带来了信息安全与隐私保护方面的隐忧。本书的第三个研究重点是数据挖掘过程中的隐私保护问题。本书总结了作者近年来在事务型数据的隐私保护研究中取得的部分新成果，主要创新点如下。

·提出集成全局概化技术与全消隐技术来进行隐私保护。文献已报道的研究工作均采用单一技术，因而会造成很高的信息损失。本书指出消隐技术能去除非

正常项目，从而避免由此造成的过度概化，概化技术能通过合并相似的项目来避免消隐大量项目。在任何单一技术难以奏效的情况下，技术集成能大大降低信息损失。

·集成技术求解问题时的解空间远远大于采用任何单一技术的解空间。因此，保证集成技术的时间效率与空间可伸缩性绝非易事。本书提出了多轮次、自顶向下搜索的求解策略来解决此挑战。

·提出的集成技术比目前最先进的单一技术具有更好的数据效用，并具有很高的效率和可伸缩性。集成技术保持了数据的域互斥性，并且保留了项目集的频度，因而处理后的数据能够直接采用标准的数据挖掘工具进行分析，挖掘结果在原始数据中必定也成立。

简而言之，本书最大的创新是提出了伺机挖掘思想，基于这一思想提出并实现了海量数据挖掘的若干高性能算法；将分布式问题求解和移动型智能代理技术应用于数据挖掘软件系统设计，提出了网络海量数据协同挖掘模型；研究了数据挖掘过程中的隐私保护问题，提出全新的基于技术集成的隐私保护方案。

刘君强

2010年9月于浙江工商大学

摘 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。

[关键词]：知识发现，数据挖掘，关联规则，分类规则，多维多层多数据类型关联规则，频繁模式，闭合频繁模式，最大频繁模式，黑板系统，分布式问题求解，智能代理，移动型智能代理，协同数据挖掘，分布式数据挖掘，移动式数据挖掘，智能型数据挖掘工具，算法，软件，海量数据库，保护隐私的数据挖掘技术

ABSTRACT

With the development of information technology, especially the emerging of the network technology, our abilities to collect, store and transfer data have been improved dramatically. Comparing to the explosive growth of data, our needs for decision relevant knowledge are not satisfied yet. Knowledge discovery and data mining technology is an important approach to address this problem. To be useful for real world applications, high performance mining algorithms and software platforms are in urgent need. This book focuses on the research into efficient and scalable mining algorithms, software platforms, and privacy protection techniques that support the knowledge discovery in distributed, heterogeneous, and very large databases.

This book presents a novel algorithm, called OpportuneProject, which is fundamentally different from those proposed in the past in that it proposes novel methods to build tree-based pseudo projections and array-based unfiltered projections for projected transaction subsets, which makes our algorithm both CPU time efficient and memory saving. It opportunistically chooses between different structures to represent projected transaction subsets, and heuristically decides projection methods to be employed. Basically, the algorithm grows the frequent item set tree by depth first search, whereas breadth first search is used to build the upper portion of the tree if necessary. The empirical results show that our algorithm is not only the most efficient on both sparse and dense databases at all levels of support threshold, but also highly scalable to very large databases.

Because of the inherent complexity, mining complete set of frequent patterns could be impractical. Alternatives are to mine closed set or maximal set of frequent patterns. A novel compound frequent itemset tree is proposed to enumerate closed set of frequent patterns, which facilitates fast growth, efficient local pruning, and global subsumption checking of search space. The fast hashing methods are developed. A new algorithm, called CROP to mine closed frequent patterns is designed whose performance is maximized by balancing tree growth and tree pruning overheads. Based on that, an efficient algorithm, called MOP is proposed to discover

maximal frequent patterns, which combines closure checking with inclusion checking, and employs lookaheads. CROP and MOP are more efficient and scalable than the counterparts.

This book further proposes some new techniques, namely the reverse lexicographic tree for organizing multi-level frequent patterns, which facilitates strong pruning of the search space and hence addresses the efficiency bottleneck, the taxonomic labeling that is applicable to various data representations, which resolves the scalability bottleneck, and information entropy based method to partition quantitative intervals and qualitative values. With these techniques, multi-dimension, multi-level, multi-data-type association rules can be mined by constrained single-dimension single-level boolean algorithms. Upon that, the design of new algorithms mfpRLTL and MDML-PP are presented. To mine classification rules, a new algorithm, called CRM-PP is also proposed, which pushes multiple minimum support thresholds into the discovery stage of frequent patterns, and generates rules in a single stage. Algorithms mfpRLTL, MDML-PP, and CRM-PP are one to three orders of magnitude efficient than algorithms derived from Apriori and FP-Growth.

The second part of this book dedicates to the research into data mining software systems. Such a system, called SmartMiner, is proposed based on the research fulfillments in data mining algorithms and expert systems achieved by the author. SmartMiner presents a mining definition language, called MDL, a script language that describes mining scenarios, and integrates data warehousing functionalities. Its mining engine has kind of intelligence in that it employs heuristics to select algorithms and to adjust environment settings.

This book also presents cooperative mining software platforms for knowledge discovery in distributed, heterogeneous, and very large databases. A formal language describing blackboard and knowledge source is proposed. A blackboard system model, called DBC-MA, based on a production system is designed and implemented, which is the major component for distributed problem solving. A distributed data mining platform, called DistributedMiner is designed by integration of DBC-MA and SmartMiner. Based on the analysis of mobile agent technology, a mobile agent server model and mobile data mining platform is proposed.

The third part of this book focuses on privacy protection problem in mining and publishing transaction data. A key feature of transaction data is the extreme sparsity, which renders any single technique ineffective in anonymizing such data. Among recent works, some incur high information loss, some result in data hard to interpret,

and some suffer from performance drawbacks. This book proposes to integrate generalization and suppression to reduce information loss. However, the integration is non-trivial. This book proposes novel techniques to address the efficiency and scalability challenges. Extensive experiments on real world datasets show that this approach outperforms the state-of-the-art methods, including global generalization, local generalization, and total suppression. In addition, transaction data anonymized by this approach can be analyzed by standard data mining tools, a property that local generalization fails to provide.

Keywords: knowledge discovery, data mining, association rules, classification rules, multi-level multi-dimension multi-data type rules, frequent patterns, closed frequent patterns, maximal frequent patterns, blackboard systems, distributed problem solving, intelligent agents, mobile agents, cooperative data mining, distributed data mining, mobile data mining, intelligent data mining tools, algorithms, software, very large databases, privacy preserving data mining

目 录