中国计算机学会学术著作丛书

知 识 发 现

史忠植摇著

清 华 大 学 出 版 社

（京）新登字 员源愿号

内 容 简 介

知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。知识发现将信息变为知识，从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展作出贡献。

本书全面而又系统地介绍了知识发现的方法和技术，反映了当前知识发现研究的最新成果。

全书共分 员源章。第 员章是绪论，介绍知识发现的重要概念和任务。第 圆章讨论决策树，它是归纳学习方法中最实用的一种技术。关联规则挖掘是近几年应用最为广泛的方法，第 猿章将对重要的关联规则挖掘算法进行讨论。第 源章讨论范例推理，它是一种有效的实用技术。第 缘章探讨模糊聚类法。第 远章讨论粗糙集。第 苑章是贝叶斯网络，贝叶斯网络可以处理不完整和带有噪声的数据集，它用概率测度的权重来描述数据间的相关性。第 愿章探讨支持向量机，它在近几年知识发现研究中是极其活跃的研究课题。第 怨章讨论隐马尔科夫模型。第 员园章是神经网络，书中着重介绍几种实用的算法。第 员员章讨论进化和遗传算法。第 员圆章介绍知识发现平台 [illegible] 接着，以 宰[illegible]知识发现、生物信息处理为例，介绍知识发现的应用。第 员猿章关于 宰[illegible]知识发现。第 员源章介绍生物信息处理中基因组模式的发现。

本书内容新颖，认真总结了作者的科研成果，取材国内外最新资料，反映了当前该领域的研究水平。论述力求概念清晰，表达准确，突出理论联系实际，通过实例说明原理，富有启发性。本书对从事知识发现、数据挖掘、机器学习、人工智能研究和知识管理的科技人员具有重要参考价值，可以用作计算机、信息技术等专业博士生、硕士生的教材。

书摇摇名：知识发现
作摇摇者：史忠植摇著
出 版 者：清华大学出版社（北京清华大学学研大厦，邮编 员[illegible]
[illegible]
责任编辑：薛慧
印 刷 者：清华大学印刷厂
发 行 者：新华书店总店北京发行所
开摇摇本：[illegible]印张：[illegible]字数：[illegible]千字
版摇摇次：[illegible]年 员月第 员版摇[illegible]年 员月第 员次印刷
书摇摇号：[illegible]
印摇摇数：[illegible]
定摇摇价：[illegible]元

前摇摇言

随着计算机应用及 [illegible] 的日益普及，“丰富的数据与贫乏的知识”问题也日见突出，世界上的数据正以惊人的速度增长，堆积如山。不同领域的人们都期待着从这些数据中得到自己想要的答案，将信息变为知识，从数据矿山中找到蕴藏的知识金块。

知识发现正是这样一种从数据中挖掘知识的工具，它集数据收集、数据清洁、降维、规则归纳、模式识别、数据 [illegible] 结果分析及评估、可视化输出等多种过程于一身，是统计学、计算机科学、模式识别、人工智能、机器学习及其他学科相结合的产物。它不仅被许多研究人员看作是数据库系统和机器学习方面一个重要的研究课题，而且被许多工商界人士看作是一个能带来巨大回报的重要领域。从数据库中发现出来的知识可以用在信息管理、查询响应、决策支持、过程控制等许多方面。从 [illegible] 世纪 [illegible] 年代中期的小范围研究到如今的蓬勃兴起，知识发现已经在企业界与科学界占据了一席之地。事实上，世界 [illegible] 强企业中的 [illegible] 都涉足知识发现的前瞻性研究或拥有一个或多个知识发现产品系统。它们帮助企业进行客户关系管理，减少不必要的投资，提高资金周转和回报，帮助人们迅速获取所需的知识和信息，提高工作效率，改进服务质量。

知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程，它与数据仓库有着密切的联系。数据仓库是源于可操作数据的一个数据存储中心。数据仓库中的信息是面向主题的、稳定的且具有历史数据属性的，因此数据仓库用于存储大规模的数据集。知识发现与数据仓库、决策支持的结合预示着信息和知识管理的一个全新的变革。

本书全面而又系统地介绍了知识发现的方法和技术。全书共分 [illegible] 章。第 [illegible] 章是绪论，介绍知识发现的重要概念和任务。第 [illegible] 章讨论决策树，它是归纳学习方法中最实用的一种技术。关联规则挖掘是近几年应用最为广泛的方法，第 [illegible] 章将对重要的关联规则挖掘算法进行讨论。第 [illegible] 章讨论范例推理，它是一种有效的实用技术。第 [illegible] 章探讨模糊聚类法。第 [illegible] 章讨论粗糙集。第 [illegible] 章是贝叶斯网络，贝叶斯网络可以处理不完整和带有噪声的数据集，它用概率测度的权重来描述数据间的相关性。第 [illegible] 章探讨支持向量机，它在近几年知识发现研究中是极其活跃的研究课题。第 [illegible] 章讨论隐马尔科夫模型。第 [illegible] 章是神经网络，书中着重介绍几种实用的算法。第 [illegible] 章讨论进化和遗传算法。第 [illegible] 章介绍知识发现平台 [illegible]。接着，以 [illegible] 知识发现、生物信息处理为例，介绍知识发现的应用。第 [illegible] 章关于 [illegible] 知识发现。第 [illegible] 章介绍生物信息处理中基因组模式的发现。

本书是中国科学院计算技术研究所智能信息处理重点实验室有关机器学习和知识发现研究工作的总结。涉及的研究项目得到国家自然科学基金、国家 [illegible] 高技术计划、北京市自然科学基金、国家重点科技攻关项目的资助。参加该项研究工作的人员有叶

世伟副教授、何清副教授、李晓黎博士、叶施仁博士、宫秀军博士生、刘少辉博士生、郑毅、郑金华教授、张建博士、王军博士、张颖博士、李云峰博士、潘谦红博士、吴斌博士生、贾自艳博士生、游湘涛、任力安、李宝东等。本书得到清华大学出版社计算机专著出版基金的资助。

史忠植

二〇〇五年 忽月

目录

第员章摇绪摇摇论

员援摇知摇摇识

摇摇人类从工业社会向知识社会演进时，政治经济中心正从“生产”转向“发现、发明和创新”。知识正在成为创新的核心，知识创新成为知识经济发展的最主要的动力源泉。知识经济对物质文明发展能够发挥巨大的推动作用，依靠无形资产的投入来实现可持续发展的，推动经济全球化发展。

在信息科学中，信息是根据表示数据所用的约定，赋予数据的意义。数据是事物、概念或指令的一种形式化的表示形式，以适合于用人工或自然方式进行通信、解释或处理。信息是数据所表达的客观事实。数据是信息的载体，与具体的介质和编码方法有关。圆园世纪 缘园年代，[illegible]对信息的数学本质进行过研究，提出了著名的 [illegible]信息论。他用熵的概念来研究信息的容量，采用比特作为度量信息的单位。

信息经过加工和改造形成知识。知识是人类在实践的基础上产生又经过实践检验的对客观实际的可靠的反映。知识是人脑创新的成果，是人类智慧的结晶。智慧是人类文明的源泉，是推动历史发展的永衡动力，是生产力诸要素中的核心。知识一般可分为陈述性知识、过程性知识和控制性知识。陈述性知识提供概念和事实，描述系统状态、环境和条件，使人们知道是什么。例如，在一个知识检索系统中，陈述性知识包括陈述具体事实的数据库内容。过程性知识提供有关状态的变化、问题求解过程的操作、演算和动作的知识。智能信息检索系统利用过程性知识处理陈述性知识。用控制策略表示问题的知识常称为控制性知识。控制性知识，即元知识，包含有关各种处理过程、策略和结构的知识，常用来协调整个问题求解的过程。

知识具有下列特性：

(员) 知识的客观性。虽然知识是人脑对信息加工的成果，但这些成果是客观的，人类对自然、社会、思维规律的认识是客观的，这些规律的运行是不以人的意志为转移的。

(圆) 知识的相对性。人类对自然、社会、思维规律的认识必须有一个过程。在一段时间内认为正确的东西，经过变革，可能发生变化。[illegible]年第 [illegible]届国际人工智能大会上，[illegible]的计算机和思维奖授予 [illegible]的 [illegible]。他提出基于行为的人工智能，认为智能不要知识表示，智能不要推理[illegible]。[illegible]年第 [illegible]届国际人工智能大会上，[illegible]的计算机和思维奖授予斯坦福大学的 [illegible]，以表彰她在概率推理、机器学习方面的贡献[illegible]。

(猿) 知识的进化性。人类在认识客观世界和主观世界的过程中，不断对真理的长河加入新的内容，知识不断更新，例如对物质结构的认识，对基因的认识等。

(源) 知识的依附性。知识有载体，载体分层次。离开载体的知识是没有的。随着载体的消失，知识也跟着消失。

(缘) 知识的可重用性。在使用过程中知识可以反复重用。当然，要根据具体情况作

具体分析，灵活应用知识。

(迈 知识的共享性。基础研究一般由政府进行投资，所得到的科学知识具有共享性；但最新的技术知识受到知识产权法保护，使用者只有支付一定的费用，才能获得这种知识的使用权。知识产权的保护对发展技术和知识经济是非常重要的国策。

员爱摇知 识 发 现

知识发现是从数据集中抽取和精化新的模式。知识发现的范围非常广泛，可以是经济、工业、农业、军事、社会、商业、科学的数据或卫星观测得到的数据。数据的形态有数字、符号、图形、图像、声音等。数据组织方式也各不相同，可以是有结构、半结构或非结构的。知识发现的结果可以表示成各种形式，包括规则、法则、科学规律、方程或概念网等。

目前，关系型数据库应用广泛，并且具有统一的组织结构，一体化的查询语言，关系之间及属性之间具有平等性等优点。因此，数据库知识发现 (噪[illegible]藻[illegible]运词词 的研究非常活跃。该术语于 员怨怨怨年出现，云[illegible]定义为“运词词是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程”[illegible]。在上面的定义中，涉及几个需要进一步解释的概念：“数据集”、“模式”、“过程”、“有效性”、“新颖性”、“潜在有用性”和“最终可理解性”。数据集是一组事实 云(如关系数据库中的记录)。模式是一个用语言 蕴来表示的一个表达式 耘，它可用来描述数据集 云的某个子集 云耘，耘作为一个模式要求它比对数据子集 云耘的枚举要简单 (所用的描述信息量要少)。过程在运词词中通常指多阶段的处理，涉及数据准备、模式搜索、知识评价以及反复的修改求精；该过程要求是非平凡的，意思是要有一定程度的智能性、自动性 (仅仅给出所有数据的总和不能算作是一个发现过程)。有效性是指发现的模式对于新的数据仍保持有一定的可信度。新颖性要求发现的模式应该是新的。潜在有用性是指发现的知识将来有实际效用，如用于决策支持系统里可提高经济效益。最终可理解性要求发现的模式能被用户理解，目前它主要是体现在简洁性上。有效性、新颖性、潜在有用性和最终可理解性综合在一起称为兴趣性。

由于知识发现是一门受到来自各种不同领域的研究者关注的交叉性学科，因此导致了很多不同的术语名称。除了 运词词外，主要还有如下若干种称法：“数据挖掘”(噪[illegible]蚤[illegible]孕，“知识抽取”(蚤[illegible]藻[illegible]、“信息发现”(蚤[illegible]藻[illegible]、“智能数据分析”(蚤[illegible]藻[illegible]、“探索式数据分析”(藻[illegible]藻[illegible]、“信息收获”(蚤[illegible]藻[illegible]和“数据考古”(噪[illegible]藻[illegible] 等等。其中，最常用的术语是“知识发现”和“数据挖掘”。相对来讲，数据挖掘主要流行于统计界 (最早出现于统计文献中)、数据分析、数据库和管理信息系统界，而知识发现则主要流行于人工智能和机器学习界。

知识发现过程可粗略地理解为三部曲：数据准备 (噪[illegible]藻[illegible]、数据开采以及结果的解释评估 (蚤[illegible]藻[illegible] (见图 员爱)。

·圆·

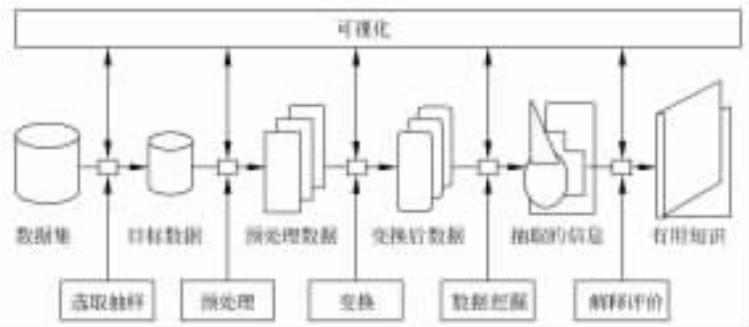

图 数据挖掘知识发现过程示意图

数据准备

数据准备又可分为三个子步骤：数据选取（确定操作对象）、数据预处理（消除噪声）和数据变换（减少维数）。数据选取的目的是确定发现任务的操作对象，即目标数据（通常是数据库的一个子集），它是根据用户的需要从原始数据库中抽取的一组数据。数据预处理一般可能包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换（如把连续值数据转换为离散型的数据，以便于符号归纳，或是把离散型的转换为连续值型的，以便于神经网络归纳）等。当数据开采的对象是数据仓库时，一般来说，数据预处理已经在生成数据仓库时完成了。数据变换的主要目的是消减数据维数或降维（减少所考虑的变量个数），即从初始特征中找出真正有用的特征以减少数据开采时要考虑的特征或变量个数。

数据挖掘阶段

数据挖掘阶段首先要确定开采的任务或目的是什么，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了开采任务后，就要决定使用什么样的开采算法。同样的任务可以用不同的算法来实现，选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来开采；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的（简单直观、容易理解的）知识（采用规则表示的开采方法显然要好于神经网络之类的方法），而有的用户或系统的目的是获取预测准确度尽可能高的预测型（黑箱）知识。

完成了上述准备工作后，就可以实施数据挖掘操作了。具体的数据挖掘方法将在后面章节中作较为详细的论述。需要指出的是，尽管数据挖掘算法是知识发现的核心，也是目前研究人员的主要努力方向，但要获得好的采掘效果，必须对各种采掘算法的要求或前提假设有充分的理解。

结果解释和评价

数据挖掘阶段发现出来的模式，经过用户或机器的评价，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要整个发现过程退回到发现阶段之前，如重新选取数据、采用新的数据变换方法、设定新的数据挖掘参数值，甚至换一种采掘算法（如当发现任务是分类时，有多种分类方法，不同的方法对不同的数据有不同的效果）。另外，运行时由于最终是面向人类用户的，因此可能要对发现的模式进行可