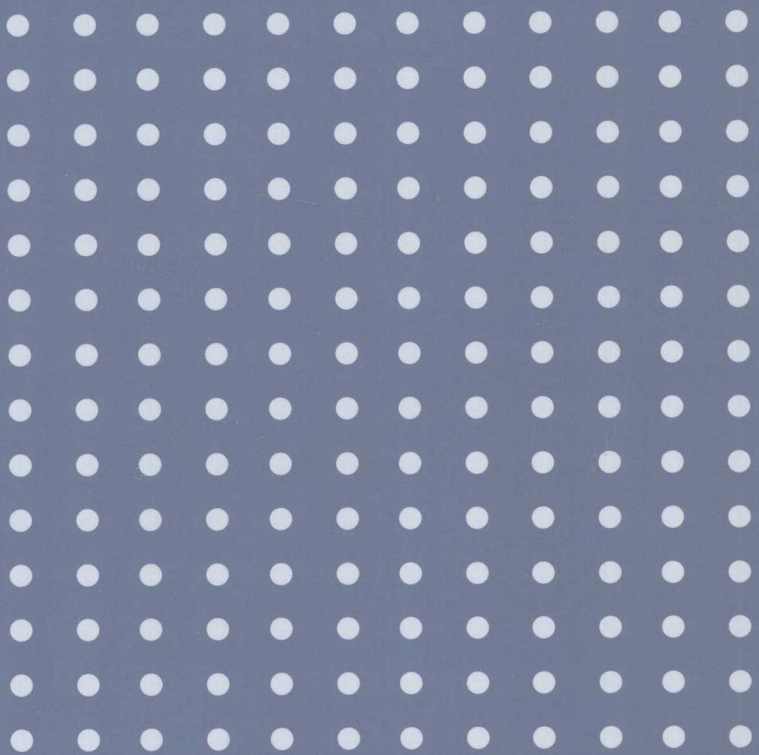
重点大学计算机专业系列教材

数据挖掘原理与算法(第二版)教师用书

毛国君 段立娟 编著

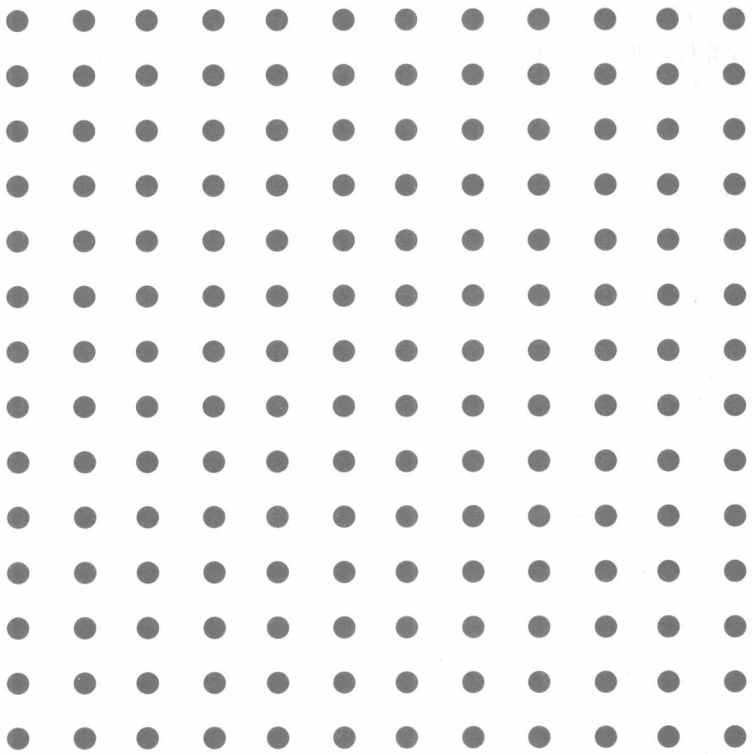


清华大学出版社

重点大学计算机专业系列教材

数据挖掘原理与算法（第二版）教师用书

毛国君 段立娟 编著



清华大学出版社
北京

内 容 简 介

《数据挖掘原理与算法》一书出版以来，被许多高校作为本科生或者研究生教材使用，是一本全面介绍数据挖掘和知识发现技术的专业书籍，具有内容系统、知识含量高等特点。为了让教师更好地使用教材《数据挖掘原理与算法》(第二版)，作者又编写了本书。本书分四个部分：一、对教材每章的部分习题给出了参考答案；二、介绍各章授课内容重点与课时分配；三、针对不同的授课学生对象给出了课时安排的建议；四、提供了两套样本试卷及其参考答案。

本书供使用《数据挖掘原理与算法》一书的教师作参考。

图书在版编目(CIP)数据

数据挖掘原理与算法(第二版)教师用书/毛国君，段立娟编著. —北京：清华大学出版社，2009.6
(重点大学计算机专业系列教材)
ISBN 978-7-302-19350-0

Ⅰ. 数…　Ⅱ. ①毛… ②段…　Ⅲ. 数据采集—高等学校—教学参考资料　Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2009)第 010846 号

责任编辑：丁　岭　张为民
责任校对：焦丽丽
责任印制：杨　艳
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京市清华园胶印厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：5.5　**字　数**：134 千字
版　　次：2009 年 6 月第 1 版　　**印　　次**：2009 年 6 月第1次印刷
印　　数：1～1500
定　　价：16.00 元

INTRODUCTION

出版说明

随着国家信息化步伐的加快和高等教育规模的扩大，社会对计算机专业人才的需求不仅体现在数量的增加上，而且体现在质量要求的提高上，培养具有研究和实践能力的高层次的计算机专业人才已成为许多重点大学计算机专业教育的主要目标。目前，我国共有16个国家重点学科、20个博士点一级学科、28个博士点二级学科集中在教育部部属重点大学，这些高校在计算机教学和科研方面具有一定优势，并且大多以国际著名大学计算机教育为参照系，具有系统完善的教学课程体系、教学实验体系、教学质量保证体系和人才培养评估体系等综合体系，形成了培养一流人才的教学和科研环境。

重点大学计算机学科的教学与科研氛围是培养一流计算机人才的基础，其中专业教材的使用和建设则是这种氛围的重要组成部分，一批具有学科方向特色优势的计算机专业教材作为各重点大学的重点建设项目成果得到肯定。为了展示和发扬各重点大学在计算机专业教育上的优势，特别是专业教材建设上的优势，同时配合各重点大学的计算机学科建设和专业课程教学需要，在教育部相关教学指导委员会专家的建议和各重点大学的大力支持下，清华大学出版社规划并出版本系列教材。本系列教材的建设旨在“汇聚学科精英、引领学科建设、培育专业英才”，同时以教材示范各重点大学的优秀教学理念、教学方法、教学手段和教学内容等。

本系列教材在规划过程中体现了如下一些基本组织原则和特点。

1. 面向学科发展的前沿，适应当前社会对计算机专业高级人才的培养需求。教材内容以基本理论为基础，反映基本理论和原理的综合应用，重视实践和应用环节。

2. 反映教学需要，促进教学发展。教材要能适应多样化的教学需要，正确把握教学内容和课程体系的改革方向。在选择教材内容和编写体系时注意体现素质教育、创新能力与实践能力的培养，为学生知识、能力、素质协调发展创造条件。

3. 实施精品战略，突出重点，保证质量。规划教材建设的重点依然是专业基础课和专业主干课；特别注意选择并安排了一部分原来基础比较好的

优秀教材或讲义修订再版，逐步形成精品教材；提倡并鼓励编写体现重点大学计算机专业教学内容和课程体系改革成果的教材。

4. 主张一纲多本，合理配套。专业基础课和专业主干课教材要配套，同一门课程可以有多本具有不同内容特点的教材。处理好教材统一性与多样化的关系；基本教材与辅助教材以及教学参考书的关系；文字教材与软件教材的关系，实现教材系列资源配套。

5. 依靠专家，择优落实。在制订教材规划时要依靠各课程专家在调查研究本课程教材建设现状的基础上提出规划选题。在落实主编人选时，要引入竞争机制，通过申报、评审确定主编。书稿完成后要认真实行审稿程序，确保出书质量。

繁荣教材出版事业，提高教材质量的关键是教师。建立一支高水平的以老带新的教材编写队伍才能保证教材的编写质量，希望有志于教材建设的教师能够加入到我们的编写队伍中来。

教材编委会

FOREWORD

前言

《数据挖掘原理与算法》一书出版以来，被许多高校作为本科生或者研究生的教材使用。几年来许多教师给出了很好的建议，因此我们在2007年针对相关问题进行了修订并出版了其第二版。该教材是一本全面介绍数据挖掘和知识发现技术的专业书籍，具有内容系统、知识含量高等特点。可能也正是因为这些特点，作为教材给教师带来了一些授课难点。特别是，由于教材使用的对象不同，对教材内容进行选择是必需的。为了让教师更好地使用《数据挖掘原理与算法》一书，减轻教师的负担，我们编写了本教师用书。

《数据挖掘原理与算法(第二版)教师用书》主要从四个部分为教师提供了参考：一、对教材每章的部分习题给出了参考答案；二、介绍各章授课内容重点与课时分配；三、针对不同的授课学生对象给出了课时安排的建议；四、提供了两套样本试卷及其参考答案。

目的是为了帮助教师提高讲课的效率，但不能代替教师的教学研究工作。特别考虑到教师用书也可能被学生使用，故对教材后面的习题并没有给出全部解答。

整体上说，数据挖掘技术包含概念与过程、原理与方法两个主要部分。对于有关概念与过程，主要集中在《数据挖掘原理与算法》(第二版)第1章和第2章，不论学生对象如何，教师都应该给予重视，力求全面而直观地进行介绍。数据挖掘中的原理与方法，分布在《数据挖掘原理与算法》(第二版)的第3~8章，涵盖关联规则、分类、聚类、序列、空间以及Web挖掘等分支。我们认为，关联规则、分类、聚类是经典内容，不论学生对象如何，教师都应该选择一些典型的理论和算法进行剖析。对于不同的教学对象，教师可以对第3~5章的内容进行合理选择。例如，如果准备给本科生开一个只有32课时的课程，那么最起码的要求是在对于关联规则、分类、聚类等基本概念和原理讲述清楚的前提下，能把Apriori、ID3和k-means算法剖析清楚即可。第6~8章的内容相对比较松散，对于研究生来说，我们认为需要进行选择性地介绍或讨论。这是因为这些内容属于数据挖掘的较前沿的课题，而且有着很广泛的研究和应用价值，因此对于研究生将来的研究工作可能会有很大的帮助。

《数据挖掘原理与算法》(第二版)共分8章，各章相对独立，而且每章的

内容都是从前往后难度逐渐增大的。因此,教师完全可以发挥自己的想象力和知识上的优势进行内容选择。此外,如果读者是从事计算机相关研究和开发的人员,本教师用书可能也能帮助读者节约宝贵时间,提高《数据挖掘原理与算法》(第二版)一书的利用效率。总之,作者希望通过本教师用书,提供一个很好地利用《数据挖掘原理与算法》(第二版)的辅助材料,促进数据挖掘技术的普及与提高。

作　者

2008 年 12 月于北京

CONTENTS

目录

PART I

第一部分 各章习题及部分参考答案

第1章　绪　　论

1. 给出下列英文缩写或短语的中文名称和简单的含义

(1) Data Mining

(2) Artificial Intelligence

(3) Machine Learning

(4) Knowledge Engineering

(5) Information Retrieval

(6) Data Visualization

参考答案：(1) 数据挖掘。简单地说就是从大型数据中挖掘所需要的知识。

(2) 人工智能。简单地说就是研究如何应用机器来模拟人类某些智能行为的基本理论、方法和技术的一门科学。

(3) 机器学习。简单地说就是研究如何使用机器来模拟人类学习活动的一门学科。

(4) 知识工程。简单地说就是研究知识信息处理并探讨开发知识系统的技术。

(5) 信息检索。简单地说就是研究合适的信息组织并根据用户需求快速而准确地查找信息的技术。通常指的是计算机信息检索，它以计算机技术为手段，完成电子信息的汇集、存储和查找等的相关技术。

(6) 数据可视化。简单地说就是运用计算机图形学和图像处理等技术，将数据换为图形或图像在屏幕上显示出来。它是进行人机交互处理、数据解释以及提高系统可用性的重要手段。

2. 给出下列英文缩写或短语的中文名称和简单的含义。

(1) OLTP(On-line Transaction Processing)

(2) OLAP(On-line Analytic Processing)

(3) Decision Support

(4) KDD(Knowledge Discovery in Databases)

(5) Transaction Database

(6) Distributed Database

参考答案：略。

3. 为什么说数据挖掘是未来信息处理的骨干技术之一？

参考答案：数据挖掘之所以被称为未来信息处理的骨干技术之一，主要在于它以一种全新的概念改变着人类利用数据的方式。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行简单地查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。

4. 从商业需求角度分析数据挖掘技术产生的合理性。

参考答案：略。

5. 支撑数据挖掘技术的主要研究基础学科有哪些？说明数据挖掘产生的技术背景。

参考答案：任何技术的产生总是有它的技术背景的。数据挖掘技术的提出和普遍接受

是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。普遍认为,对数据挖掘产生决定性作用的三个主要技术是:数据库技术、统计学和包括机器学习在内的人工智能技术。

在关系型数据库的研究和产品提升过程中,人们一直在探索组织大型数据和快速访问的相关技术。高性能关系数据库引擎以及相关的分布式查询、并发控制等技术的使用,已经提升了数据库的应用能力。在数据的快速访问、集成与抽取等问题的解决上积累了经验。数据仓库作为一种新型的数据存储和处理手段,被数据库厂商普遍接受并且相关辅助建模和管理工具快速推向市场,成为多数据源集成的一种有效的技术支撑环境。因此,人们已经具备利用多种方式存储海量数据的能力。这些丰富多彩的数据存储、管理以及访问技术的发展,为数据挖掘技术的研究和应用提供了丰富的土壤。

计算机芯片技术的发展,使计算机的处理和存储能力日益提高。随之而来的是硬盘、CPU 等关键部件的价格大幅度下降,使得人们收集、存储和处理数据的能力和欲望不断提高。经过几十年的发展,计算机的体系结构,特别是并行处理技术已经逐渐成熟和普遍应用,并成为支持大型数据处理应用的基础。计算机性能的提高和先进的体系结构的发展使数据挖掘技术的研究和应用成为可能。

历经了十几年的发展,包括基于统计学、人工智能等在内的理论与技术性成果已经被成功地应用到商业处理和分析中。这些应用从某种程度上为数据挖掘技术的提出和发展起到了极大地推动作用。数据挖掘系统的核心模块技术和算法都离不开这些理论和技术的支持。从某种意义上讲,这些理论本身的发展和应用为数据挖掘提供了有价值的理论和应用积累。

6. 数据挖掘技术是一个交叉研究分支,简述影响它产生和发展的主要研究学科或分支及其关系。

参考答案:略。

7. 数据(Data)、信息(Information)和知识(Knowledge)是人们认识和利用数据的三个不同阶段,数据挖掘技术是如何把它们有机的结合在一起的?

参考答案:从数据、信息和知识三个层面上看,数据是最原始的未经组织和处理的信息源。信息或称有效信息是指对人们在某些方面有价值的东西。知识是一种现实世界信息的抽象和浓缩,是一种概念、规则、模式和规律等。数据挖掘技术通过对原始数据进行微观、中观乃至宏观的统计、分析、综合和推理,发现数据间的关联性、未来趋势以及一般性的概括知识等,转变成可以用来指导人们某些高级商务活动的有用信息。

8. 从数据挖掘研究角度看,如何理解数据、信息和知识的不同和联系。

参考答案:略。

9. 简述数据挖掘技术将来的发展趋势。

参考答案:对于数据挖掘技术的发展趋势,应该分两方面辩证的理解。

(1) 数据挖掘技术已经存在相当大市场,将成为对工业产生重要影响的关键技术之一。同时,并行计算机体系结构研究和 KDD 也被列入今后 5 年内公司应该投资的 10 个新技术领域之一。这些资料都表明,数据挖掘技术在将来有很大的发展潜力及空间。

(2) 数据挖掘技术作为一门新技术,仍有许多问题需要研究、解决和探索。分析目前的研究和应用现状,对于数据挖掘技术将来的工作重点有:

① 数据挖掘技术与特定商业逻辑的平滑集成问题；
② 数据挖掘技术与特定数据存储类型的适应问题；
③ 大型数据的选择与规格化问题；
④ 数据挖掘系统的构架与交互式挖掘技术；
⑤ 数据挖掘语言与系统的可视化问题；
⑥ 数据挖掘理论与算法研究。

10. 按你对数据挖掘技术的了解，你认为它的研究将面临的主要挑战和对策是什么？

参考答案：略。

11. 你认为应该如何来理解 KDD 与 Data Mining 的关系？说明你的理由。

参考答案：关于 KDD 与 Data Mining 的关系有以下几种说法。

(1) KDD 看成数据挖掘的一个特例。这是早期比较流行的观点，在许多文献可以看到这种说法。因此，从这个意义上说，数据挖掘就是从数据库、数据仓库以及其他数据存储方式中挖掘有用知识的过程。这种描述强调了数据挖掘在源数据形式上的多样性。

(2) 数据挖掘是 KDD 过程的一个步骤(从狭义角度考虑)。这种观点得到大多数学者认同，有它的合理性。KDD 是一个广义的范畴，它包括数据清洗、数据集成、数据选择、数据转换、数据挖掘、模式生成及评估等一系列步骤。这样，可以把 KDD 看作是一些基本功能构件的系统化协同工作系统，而数据挖掘则是这个系统中的一个关键的部分。

(3) KDD 与 Data Mining 含义相同(从广义角度考虑)。有些人认为，KDD 与 Data Mining 只是叫法不一样，它们的含义基本相同。事实上，在现今文献的许多地方，这两个术语仍然不加区分地使用着。

从上面的描述中可以看出，数据挖掘概念可以在不同的技术层面上来理解，但是其核心仍然是从数据中挖掘知识。数据挖掘定义有广义和狭义之分。从广义的观点上，数据挖掘是从大型数据集中，挖掘隐含在其中的、人们事先不知道的、对决策有用的知识的过程。从狭义的观点上，可以定义数据挖掘是从特定形式的数据集中提炼知识的过程。

12. 解释将 Data Mining 理解为 KDD 整个过程的一个关键步骤地合理性。

参考答案：略。

13. 根据挖掘数据的对象不同，可以将数据挖掘技术进行分类，简述这些分类类型。

参考答案：根据挖掘数据的对象不同，数据挖掘技术可以分为关系型数据库挖掘、面向对象数据库挖掘、空间数据库挖掘、时态数据库挖掘、文本数据库挖掘、多媒体数据库挖掘、异质数据库挖掘、遗产数据库挖掘、Web 数据挖掘等。

14. 根据数据挖掘技术所依赖的主要技术来划分，数据挖掘技术有哪些主要的分类？简述这些类型的主要技术特点。

参考答案：略。

15. 粗糙集的知识形成主要是基于什么思想的？简述粗糙集理论中的信息系统、近似空间、下近似、上近似、约简等概念。

参考答案：粗糙集的知识形成思想可以概括为：一种类别对应于一个概念(类别一般表示为外延即集合，而概念常以如规则描述这样的内涵形式表示)，知识由概念组成；如果

某知识中含有不精确概念,则该知识不精确。粗糙集理论是一种研究不精确、不确定性知识的数学工具。

(1) 信息系统:一个信息系统 S 是一个四元组 $S=<U,A,V,f>$,其中 U 是对象(或事例)的有限集合,记为 $U=\{x_1,x_2,\cdots,x_n\}$; A 是属性的有限集合,记为 $A=\{A_1,A_2,\cdots,A_m\}$; V 是属性的值域集,记为 $V=\{V_1,V_2,\cdots,V_m\}$,其中 V_i 是属性 A_i 的值域; f 是信息函数(Information Function),即 $f: U\times A\rightarrow V, f(x_i,A_j)\in V_j$。

(2) 近似空间:近似空间有一个二元组 $<U,R(B)>$ 给出,其中 U 是对象(或事例)的有限集合,记为 $U=\{x_1,x_2,\cdots,x_n\}$; B 是 A 的属性子集,$R(B)$ 是 U 上的二元等价关系,即 $R(B)=\{(x_1,x_2)\mid f(x_1,b)=f(x_2,b),b\in B\}$。

(3) 下近似和上近似:对任意一个概念(或集合)O,B 是 U 的一个子集,O 的下近似定义为 $\underline{BO}=\{x\in U\mid [x]_{R(B)}\subset O\}$,其中 $[x]_{R(B)}$ 表示 x 在 $R(B)$ 上的等价类。O 的上近似定义为 $\overline{BO}=\{x\in U\mid [x]_{R(B)}\cap O\neq\varnothing\}$。一个概念(或集合)的下近似中的元素肯定属于该概念(或集合);而一个概念(或集合)的上近似概念(或集合)只是可能属于该概念。

(4) 约简:即极小属性集,也就是去掉约简中的任何一个属性,都将使得该属性集对应的规则覆盖反例,即导致规则与例子的不一致。

16. 简述粗糙集知识形成主要过程,为什么说它和数据挖掘技术在解决问题空间上有很大的重合性。

参考答案:略。

第 2 章　知识发现过程与应用结构

1. KDD 是一个多步骤的处理过程，它一般包含哪些基本阶段？简述各阶段的功能。

参考答案：KDD 是一个多步骤的处理过程，一般分为问题定义、数据抽取、数据预处理、数据挖掘以及模式评估等基本阶段。

(1) 问题定义阶段的功能：和领域专家以及最终用户紧密协作，一方面了解相关领域的有关情况，熟悉背景知识，弄清用户要求，确定挖掘的目标等要求；另一方面通过对各种学习算法的对比进而确定可用的学习算法。

(2) 数据抽取阶段的功能：选取相应的源数据库，并根据要求从数据库中提取相关的数据。

(3) 数据预处理阶段的功能：对前一阶段抽取的数据进行再加工，检查数据的完整性及数据的一致性。

(4) 数据挖掘阶段的功能：运用选定的数据挖掘算法，从数据中提取出用户所需要的知识。

(5) 模式评估阶段的功能：将 KDD 系统发现的知识以用户能了解的方式呈现，并且根据需要进行知识评价。如果发现知识和用户挖掘目标不一致，则重复以上阶段以最终获得可用的知识。

2. 为什么一个完整的知识发现要多种技术结合、多阶段集成。

参考答案：略

3. 简述在数据挖掘前要进行数据预处理的理由及其解决的主要问题。

参考答案：数据预处理包括：数据清洗、数据变换和数据归约等，是进行数据分析和挖掘的基础。如果所集成的数据不正确，数据挖掘算法输出的结果也必然不正确，这样形成的决策支持是不可靠的。因此，要提高挖掘结果的准确率，数据预处理是不可忽视的一步。

对数据进行预处理，一般需要对源数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行平滑，对丢失的数据进行填补，消除“脏”数据，消除重复记录等。

4. 为什么在知识发现过程中，要强调和用户交互的必要性？通常需要那些专长的技术人员支持？

参考答案：略

5. 阶梯处理过程模型是知识发现的基本模型，画出它的基本处理流程，并简要说明各阶段的任务。

参考答案：阶梯处理过程模型的基本处理流程如图 2-1 所示。

各阶段的主要任务是：

(1) 数据准备：了解相关领域的情况，弄清楚用户的要求，确定挖掘的总体目标和方法，并对原数据结构加以分析、确定数据选择原则等工作。

(2) 数据选择：从数据库中提取与 KDD 目标相关的数据。

(3) 数据预处理：主要是对上一阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行处理，对丢失的数据可以利用统计方法进行填补。对一

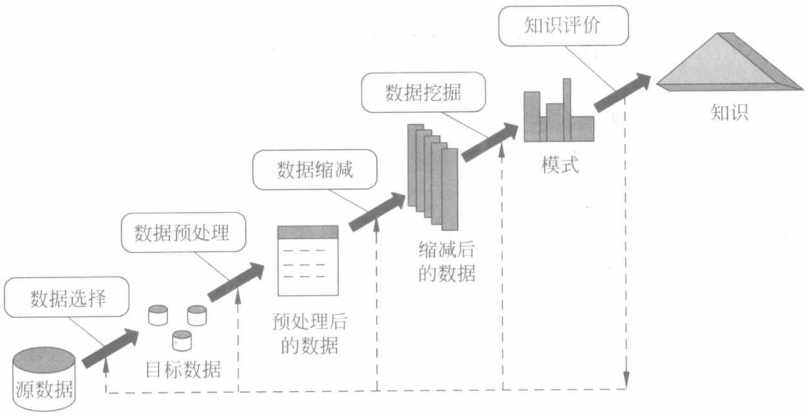


图 2-1　KDD 阶梯处理过程模型

些不适合于操作的数据进行必要的处理等。

(4) 数据缩减：对经过预处理的数据，根据知识发现的任务对数据进行抽取处理，使数据再次精简取其精华，更好地集中于用户挖掘目标上。

(5) 确定 KDD 的目标：根据挖掘的目标和用户的要求，确定 KDD 所发现的具体知识模式和类型(如分类、聚类、关联规则等)。

(6) 确定数据挖掘算法：根据上一阶段所确定的模式，选择合适的数据挖掘算法。(包括选取合适的参数、知识表示方式，并保证数据挖掘算法与整个 KDD 的评判标准相一致)。

(7) 数据挖掘：运用选定的算法，从数据中提取出用户所需要的知识。

(8) 模式解释：对发现的模式进行解释。在此过程中，为了取得更为有效的知识，可能会返回前面处理步骤中的某些步以改进结果，保证提取出的知识是有效和可用的。

(9) 知识评价：将发现的知识以用户能了解的方式呈现给用户。这期间也包含对知识的一致性的检查，以确信本次发现的知识不与以前发现的知识相抵触。

6. 简述螺旋处理过程模型相对于阶梯处理过程模型的优缺点。

参考答案：略

7. 简述以用户为中心的处理模型的基本思想。

参考答案：注重对用户与数据库交互的支持，用户根据数据库中的数据，提出一种假设模型，然后选择有关数据进行知识的挖掘，并不断对模型的数据进行调整优化，以提高数据挖掘的准确性和效率。因此，以用户为中心的处理模型的核心是将与用户的交互思想贯穿于数据挖掘的整个过程中。

8. 联机 KDD 模型需要解决哪些主要问题？

参考答案：略

9. 知识发现软件或工具的发展经历哪三个主要阶段？简述他们的主要特点。

参考答案：知识发现软件或工具的发展经历了独立的知识发现软件、横向的知识发现