21世纪高等学校计算机专业
核心课程规划教材

数据挖掘原理与算法（第3版）

教师用书

◎ 毛国君　编著

清華大學出版社

21世纪高等学校计算机专业
核心课程规划教材

数据挖掘原理与算法（第3版）
教师用书

◎ 毛国君　编著

清華大學出版社
北京

内容简介

《数据挖掘原理与算法(第3版)》全面介绍了数据挖掘和知识发现技术,具有内容系统、知识含量高等特点,被许多高校作为本科生或者研究生教材使用。为了让教师更好地使用教材《数据挖掘原理与算法(第3版)》,作者又编写了这本教师用书。本书分四个部分:(1)对教材每章的部分习题给出了参考答案;(2)介绍各章授课内容重点与课时分配;(3)针对不同的授课学生对象给出了课时安排的建议;(4)提供了六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(第3版)》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3版)教师用书/毛国君编著. —北京:清华大学出版社,2017
(21世纪高等学校计算机专业核心课程规划教材)
ISBN 978-7-302-45121-1

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

中国版本图书馆CIP数据核字(2016)第227228号

责任编辑:刘　星
封面设计:刘　键
责任校对:焦丽丽
责任印制:王静怡

出版发行:清华大学出版社
网　　址:http://www.tup.com.cn, http://www.wqbook.com
地　　址:北京清华大学学研大厦A座　　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
课件下载:http://www.tup.com.cn,010-62795954
印 装 者:三河市金元印装有限公司
经　　销:全国新华书店
开　　本:185mm×260mm　　印　　张:7　　字　　数:168千字
版　　次:2017年3月第1版　　印　　次:2017年3月第1次印刷
印　　数:1～2000
定　　价:25.00元

产品编号:067464-01

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2016年针对相关问题进行了修订并出版了其第3版。该教材是一种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，作为教材来说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，教师们必须对教材内容进行选择。为了让教师更好地使用《数据挖掘原理与算法(第3版)》一书，减轻教师的负担，我们编写了这本教师用书。

《数据挖掘原理与算法(第3版)教师用书》主要从四个部分为教师提供了参考：(1)对教材每章的部分习题给出了参考答案；(2)介绍各章授课内容重点与课时分配；(3)针对不同的授课学生对象给出了课时安排的建议；(4)提供了六套样本试卷及其参考答案。目的是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。有关概念与过程的内容，主要集中在《数据挖掘原理与算法(第3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。数据挖掘中原理与方法的内容，分布在《数据挖掘原理与算法(第3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只有32课时的课程，那么在对于关联规则、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，可以进行选择性的介绍或讨论，因为这些内容属于数据挖掘较前沿的课题，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3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，这本教师用书也能帮助读者节约宝贵时间，提高《数据挖掘原理与算法(第3版)》一书的利用效率。总之，作者希望通过这本教师用书，提供一个很好地利用《数据挖掘原理与算法(第3版)》的辅助材料，促进数据挖掘技术的普及与提高。

作　者

2016年12月于北京

目 录

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的查询，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：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的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的推动作用。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地结合在一起的？

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，应该分两方面辩证地理解。

(1) 数据挖掘技术已经存在相当大的市场，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) 把 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 看成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，在现今的许多文献中，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的合理性。

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？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：对于任意一个概念（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。

17. 说明以下说法的合理性：数据挖掘将是大数据分析的重要理论、方法和技术的支撑。

参考答案：

在理论上，数据挖掘和大数据分析的研究目标是一致的，都是期望从大型数据集中发现有价值的知识模式。

从方法上说，数据挖掘方法可以为大数据分析提供全面的分析方法的支撑。数据挖掘技术已经针对大容量、高速流动或者多数据类型并存的数据集的挖掘方法进行研究，取得不同程度的研究成果。

从技术上说，除了随着高性能计算机、网络技术以及云计算的发展使大数据分析的基础硬软件环境得到满足外，分析的技术就是一个从原始的大型数据集中发现可以利用的规律性知识的过程，因此数据挖掘技术的充分发展才能使大数据分析成为可能。

18. 从大数据的4V属性角度说明大数据时代对数据挖掘的主要技术需求。

参考答案：略。

19. 说明数据挖掘与社会网络研究的相同点和不同点。

参考答案：

相同点可以概括为：

在社会性应用方面，它们有共同的应用需求：从社会性数据中发现社会关系。在方法上，它们都是数据驱动的，即从数据出发来寻找数据隐藏的规律。

不相同点可以概括为：

社会网络研究的侧重点是发现社会关系相关的问题。数据挖掘则不限于社会性数据。社会网络来自社会学，数据挖掘来自计算机科学，当然它们都有自己的理论和方法。社会网络与数据挖掘的交叉研究是相互借鉴的关系，并不是期望被相互取代。

20. 说明数据挖掘技术在社会网络分析中的应用价值。

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，对其中的噪声数据进行平滑，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？通常需要哪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，

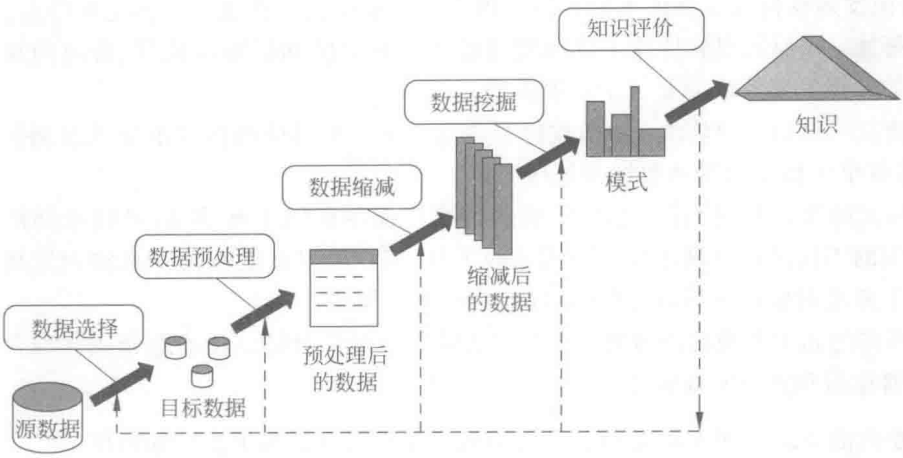

图 2-1　KDD 阶梯处理过程模型

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，对其中的噪声数据进行处理，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，可能会返回到前面的某些处理步骤中以改进结果，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，并不断地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了哪三个主要阶段？简述它们的主要特点。

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、横向的知识发现工具和纵向的知识发现解决方案三个主要阶段。

(1) 独立的知识发现软件：这类软件要求用户必须对具体的数据挖掘技术和算法有相当的了解，还要手工负责大量的数据预处理工作。

(2) 横向的知识发现工具：这些集成软件属于通用辅助工具范畴，可以帮助用户快速完成知识发现的不同阶段处理工作。使用这些工具，用户可以在数据挖掘和知识发现专家的指导和参与下开发对应的应用，起到了加速应用研制的作用。

(3) 纵向的知识发现解决方案：这种方法的核心是针对特定的商业领域和商业逻辑提供完整的数据挖掘和知识发现解决方案。

10. 横向的知识发现工具集和纵向的知识发现解决方案的主要区别是什么？

参考答案：略。

11. 什么是知识发现项目的过程化管理？它的意义如何？

参考答案：知识发现是一个包括数据抽取、数据选择、数据挖掘以及模式评估等在内的系统化挖掘知识的过程。由于数据挖掘项目规模庞大，进行过程管理可以使其更加规范化。有效的过程化管理是把实际问题分为若干子任务，在上一过程没有完成的情况下，下面的过程不能进行，以保证各个阶段的有序执行。

通过这样的模块化的管理过程，可以更好地完成数据挖掘任务，提高数据挖掘的效率和精度。

12. 简述强度挖掘的 I-MIN 过程模型的主要阶段和任务。

参考答案：略。

13. 简述数据挖掘语言的三种基本类型和特点。

参考答案：根据功能和侧重点不同，数据挖掘语言可以分为三种类型：数据挖掘查询语言、数据挖掘建模语言、通用数据挖掘语言。

(1) 数据挖掘查询语言：遵循类似 SQL 的语法，通过数据挖掘的任务、功能以及其他约束的指定、知识形成和展示等系列工作，以类似于查询的形式输入到数据挖掘系统中，通过数据挖掘系统产生对应的结果。

(2) 数据挖掘建模语言：这是对数据挖掘模型进行描述和定义的语言。数据挖掘系统在模型定义和描述方面有标准可以遵循，那么各系统之间可以共享模型，既可以解决目前各数据挖掘系统之间封闭性的问题，又可以在其他应用系统中间嵌入数据挖掘模型，解决统一的知识发现描述问题。

(3) 通用数据挖掘语言：通用数据挖掘语言合并了上述两种语言的特点，既具有定义模型的功能，又能作为查询语言与数据挖掘系统通信，进行交互式挖掘。

14. 为什么说数据挖掘语言研制对数据挖掘技术的发展是至关重要的？

参考答案：略。

第3章 关联规则挖掘理论和算法

1. 简单地描述下列英文缩写或短语的含义。

(1) Parallel Association Rule Mining

(2) Quantities Association Rule Mining

(3) Frequent Itemset

(4) Maximal Frequent Itemset

(5) Closed Itemset

参考答案：

(1) 并行关联规则挖掘。它是指利用并行处理技术、使用并行挖掘算法或在并行计算的环境下完成数据的高效挖掘工作。

(2) 数量关联规则挖掘。它是指对含有诸如工资、价钱等非离散的数值属性的数据进行挖掘的技术。数量关联规则挖掘需要解决连续属性的离散化等问题，有更广泛的商业应用。

(3) 频繁项目集。它是指出现频率高的项目对应的集合，反映交易数据中项目出现的频度信息。挖掘频繁项目集是关联规则挖掘的基础，许多关联规则挖掘方法是基于频繁项目集发现的。

(4) 最大频繁项目集。它是指在频繁项目集中不出现相互包含的项目子集。最大频繁项目集可以使用最少的信息来保证频度信息的不丢失。

(5) 关闭(或闭和)项目集。简单地说，对于一个关闭项目集的任何元素，要么不被任何元素所包含，要么只被小于它的支持度的元素所包含。

2. 解释下列概念

(1) 多层次关联规则

(2) 多维关联规则

(3) 事务数据库

(4) 购物篮分析

(5) 强关联规则

参考答案：略。

3. 给出一个项目集 I_1 在数据集 D 上的支持度(Support)的定义，并直观地解释它的含义。

参考答案：设 $I_1 \subseteq I$，项目集 I_1 在数据集 D 上的支持度是包含 I_1 的事务在 D 中所占的百分比。直观上说，一个项目集在一个数据集 D 上的支持度反映了这个项目集在数据集中出现的频率。

4. 从统计学的观点说明一个项目集 I_1 在数据集 D 上的支持度的含义。

参考答案：略。