武器装备数据挖掘技术

主编　张凤鸣　惠晓滨

国防工业出版社
National Defense Industry Press

武器装备数据挖掘技术

主　编　张凤鸣　惠晓滨

副主编　魏　靓　王焕彬　李正欣

编　委　黄　莺　车万方　宋志华

赵　罡　许建虹　黄　鹤

刘文杰　耿慧欣　李永宾

李姗姗　宋晓博　王树生

国防工业出版社

·北京·

内 容 简 介

本书以武器装备全寿命管理过程中的大量数据为研究对象，以数据挖掘方法技术为主线，对武器装备数据挖掘的理论、方法、算法进行系统的介绍，并以应用案例对其具体实践进行深入探索。本书注重系统性、突出实用性，既注重对数据挖掘方法的系统介绍，也突出不同挖掘模式在武器装备全寿命周期管理中的应用，方便读者系统学习和深刻理解武器装备全寿命管理过程中的各种数据挖掘技术。

本书面向各类武器装备管理和研究人员，重点探讨武器装备论证、研制、试验鉴定、作战和维修保障等阶段的数据挖掘方法和技术。本书可作为高等院校装备管理、信息管理、数据挖掘等专业本科生、研究生的教材，也可供从事武器装备管理工作的研究人员使用。

图书在版编目（CIP）数据

武器装备数据挖掘技术 / 张凤鸣，惠晓滨主编. —北京：国防工业出版社，2017.6
ISBN 978-7-118-11370-9

Ⅰ. ①武… Ⅱ. ①张… ②惠… Ⅲ. ①数据处理－应用－武器装备 Ⅳ. ①E92-39

中国版本图书馆 CIP 数据核字（2017）第 180500 号

※

国防工业出版社出版发行
（北京市海淀区紫竹院南路 23 号 邮政编码 100048）
三河市腾飞印务有限公司印刷
新华书店经售
*
开本 787×1092 1/16 印张 13 字数 292 千字
2017 年 6 月第 1 版第 1 次印刷 印数 1—2000 册 定价 78.00 元

（本书如有印装错误，我社负责调换）

国防书店：(010) 88540777 发行邮购：(010) 88540776
发行传真：(010) 88540755 发行业务：(010) 88540717

前　言

数据挖掘是研究从大量数据中自动提取有用信息、规律、模式的领域，自其出现后一直是数据处理、信息分析等专业方向的研究热点，目前，其理论、方法和技术还正在快速成熟中。

武器装备从论证、研制、生产、使用到维修保障是一个复杂的系统工程过程，在这个过程中会产生大量的各类数据，对这些数据目前还缺少比较有效的处理方法。本书结合作者所在团队长期在该领域进行的学术研究和科研实践，尝试从武器装备全寿命数据的深层次处理、分析和挖掘角度，对其数据挖掘方法、技术进行系统探讨。期望该书的编撰对武器装备全寿命数据分析人才的培养和武器装备数据挖掘技术的应用实践起到一定的推动作用。

本书以武器装备论证、研制、试验鉴定、作战和维修保障等阶段产生的海量数据为研究对象，以数据挖掘方法技术为主线，对武器装备数据挖掘的理论、方法、算法进行系统的介绍，并以应用案例对其具体实践进行深入探索。在编写思路、章节安排、内容撰写上，本书充分考虑了从事武器装备工作的管理人员和研究人员的使用需求。

本书在编著过程中，得到了空军装备部机关、西北工业大学、空军指挥学院、空军工程大学、空军装备研究院、国防工业出版社等单位领导和专家学者的大力支持和审读把关，对他们付出的辛勤劳动和贡献的卓越智慧我们表示诚挚的谢意。同时，本书吸纳了许多相关学科领域专家学者的理论研究成果，为提升本书的学术价值和理论高度发挥了重要的作用，在本书付梓之际，我们对这些成果的创造者表示深深的谢意；另外，本书的部分研究成果是在国家自然科学基金项目（61502521，多元时间序列相似模式挖掘中支持 DTW 距离度量的子序列搜索方法研究）的支持下取得的，在此对国家自然科学基金委员会的支持表示感谢。

本书由张凤鸣、惠晓滨担任主编。主要章节由张凤鸣、惠晓滨、魏靓、王焕彬、李正欣、黄莺、车万方、宋志华、赵罡、许建虹、黄鹤、刘文杰、耿慧欣、李永宾、李姗姗负责撰写。在全书的撰写、修改和统稿过程中，教员古清月、宋晓博、张磊，高工王树生，研究生郭庆、张贾奎、贾宇翔、刘炳琪等均做了大量工作。

武器装备的复杂性决定了其全寿命周期数据管理的复杂性。本书对武器装备寿命周期数据挖掘的概念、理论、方法进行了初步探索，由于学识有限，对很多问题的认识还需要进一步研究和完善，书中不妥之处在所难免，敬祈大家批评指正。

作　者

目　录

第1章 绪 论

武器装备管理既包括对军事需求和技术能力物化为军事装备过程的管理，还包括对军事装备经过训练和保障转化为部队战斗力过程的管理，管理活动贯穿于武器装备的全寿命周期。随着武器装备管理信息化建设的持续推进，在武器装备全寿命管理的链条上，涉及研制、设计、试验、生产、使用以及保障的众多单位都会成为海量武器装备数据的生产者，这些全寿命周期数据有效支撑着日趋复杂的武器装备采办和管理过程。但是，传统的信息处理方法无法发现数据中存在的关系和规则，无法根据现有的海量数据预测未来的发展趋势，缺乏挖掘数据背后隐藏知识的手段，导致了“数据爆炸但知识贫乏”现象。而数据挖掘正是一种知识获取的手段和工具，它可将武器装备全寿命管理过程中积累的海量数据转化为知识，为武器装备管理的科学决策提供技术支撑。

1.1 数据挖掘的概念及研究领域

数据挖掘（Data Mining）这个术语最早是在统计学领域、数据分析领域和 MIS 社区使用，在早期，它常常被贬义地用于由于无目标分析而导致数据爆炸的情况，随着统计分析、机器学习、人工智能、数据库等学科的发展，特别是人工智能的发展，数据挖掘逐渐成为了一门新兴的、涉及各种不同领域的交叉学科。

简单地说，数据挖掘就是从大量数据中提取或“挖掘”知识。有趣的是能够表达这一概念的术语，除了数据挖掘外，还有很多，如“信息抽取”（Information Extraction）、“信息发现”（Information Discovery）、“知识发现”（Knowledge Discovery）、“智能数据分析”（Intelligent Data Analysis）、“信息收获”（Information Harvesting）、“数据捕捞”（Data Dredging）、“数据/模式分析”（Data/Pattern Analysis）等。

术语的混乱是一门新兴学科走向成熟的常见现象，因此有必要给出术语内涵外延的精确表述，以下是数据挖掘几种有影响和代表性的定义：

1. Data Miners Inc

数据挖掘是对大量数据进行的自动或半自动的识别和分析，目的是从中寻求有意义的模式、规则。

2. SAS Institute Inc

数据挖掘是为寻求商业优势而对大量数据进行选择、探测和建模进而发现未知模式的处理过程。

这两种定义基本上都强调了数据挖掘应该具有以下几个特点：大量数据、模式发现、自动或半自动的数据处理形式。

3．Jorgensen M, Gentleman R.（Harvard College）

数据挖掘是统计分析技术和机器学习技术在大规模数据中的应用，这些应用往往以半自动方式进行。

4．David J. Hand（Professor of Statistics）

数据挖掘是一门基于统计学、数据技术、模式识别和机器学习并尝试通过二次分析得到未知有用关系的新学科。

这两种定义都偏重从技术支撑的角度来阐述数据挖掘，这种定义形式有利于数据挖掘过程的商业应用和工程化。

5．U.M. Fayyad, G. Piatetsky-Shapiro

数据挖掘是KDD（Knowledge Discovery in Databases，数据库中的知识发现）处理过程中的一个步骤，它使用数据分析和发现算法，在计算效率可以接受的情况下，得到特定的模式集合。其中KDD是从数据中提取有效、新颖、有潜在应用价值、最终能被理解模式的高级过程。

KDD这个术语是在第一届KDD Workshop（1989年）上被明确的，KDD的出现使U.M. Fayyad和G. Piatetsky-Shapiro重新思考数据挖掘和KDD之间的相互定位关系，并试图通过以上的定义把两者统一起来，他们认为，KDD是一个广义的知识发现过程（KDD Process），而数据挖掘是该过程中的一个阶段，该阶段的核心就是提供效率可以接受的方法（Methods）或算法（Algorithms）。

这些定义都从各自不同的视图描述了数据挖掘的概念，由于U.M. Fayyad和G. Piatetsky-Shapiro在数据挖掘基础领域所做的卓越工作，两位学者所做的定义想必会得到越来越多人的认可，本书的写作也将采用这种定义。以下是针对KDD及DM概念的一些讨论。

从如上数据挖掘的定义看，对数据挖掘概念边界的界定，学术界基本上有广义和狭义两种理解。广义的理解认为数据挖掘涵盖了知识发现的全过程，从处理过程生命周期上看，它与KDD是一样的；以U.M.Fayyad为代表的狭义理解则认为，数据挖掘仅仅是KDD过程中的一个阶段，也是最重要的一个阶段，它对整个过程提供方法和算法的支撑，完成从预处理后的数据中挖掘出相应的模式，并作为KDD下一个阶段模式评价、评估的依据。在数据挖掘的狭义理解中，如果把数据挖掘界定为KDD的一个阶段，那么许多数据挖掘的替代术语，如上文提到的“信息抽取”“信息发现”“知识发现”等，都将会变成KDD的替代语，同时，常说的数据挖掘系统或者原型系统的本质也都是一个知识发现系统，只是其核心采用了数据挖掘引擎。

KDD是基于数据库的知识发现，但它的数据可以是广义的数据，即可以是任何一个有关事实或观察数据的集合。KDD可以从数据中提取有效、新颖、有潜在应用价值、最终能被理解的模式，模式是在描述集合F中某子集时较之逐一列举集合元素更为简洁的描述，例如：描述“若成绩在81～90之间，则成绩优良”可称为一个模式，但描述“若成绩为81、82、83、84、85、86、87、88、89、90，则成绩优良”就不能称为一个模式。在KDD中发现的模式经过有效性、新颖性、潜在应用价值、可理解性评估后即可成为整个过程的最终产品即知识，通常可以用概念（Concepts）、规则（Rules）、规律（Regularities）等形式表示。

KDD是对数据进行更深层处理的过程，而不是仅仅对数据进行加减求和等简单运算或查询，因此认为它是一个高级过程（非平凡过程）。

KDD、DM 与统计学的差别：由于二者目标相似，一些人（尤其是统计学家）认为数据挖掘是统计学的分支，这是一个不切合实际的看法，因为数据挖掘还应用了其他领域的思想、工具和方法，尤其是计算机学科，例如数据库技术和机器学习，而且它所关注的某些领域和统计学家所关注的有很大不同。从总体上说，KDD、DM 更偏重于计算逻辑，它的处理过程可能是经验性的，而统计学更偏重于数学逻辑上的精确推理。

KDD、DM 与机器学习的不同：KDD、DM 是从现实世界中存在的一些具体数据中提取知识，这些知识在数据挖掘出现之前早已存在，但是人们往往没有意识到或没有明确这些知识的存在，而机器学习是一种有目的的“教”；同时，KDD、DM 的数据源往往规模庞大，而且数据不一定完整、一致，这就要求 KDD、DM 更加重视算法的效率和健壮；还有就是 KDD、DM 往往侧重于描述性任务，而机器学习往往侧重于预测性任务。

数据挖掘技术是统计理论、机器学习、人工智能、数据库的结合技术，具有较为广泛的应用前景。专家预测数据挖掘在未来十年内会有革命性进展，是个性化 Web、个人偏好分析、实时识别等的关键技术。

数据挖掘学科经过较长时间的发展，已经形成了丰富的方法体系，本书对数据挖掘当前的研究领域进行了总结，见图 1.1。

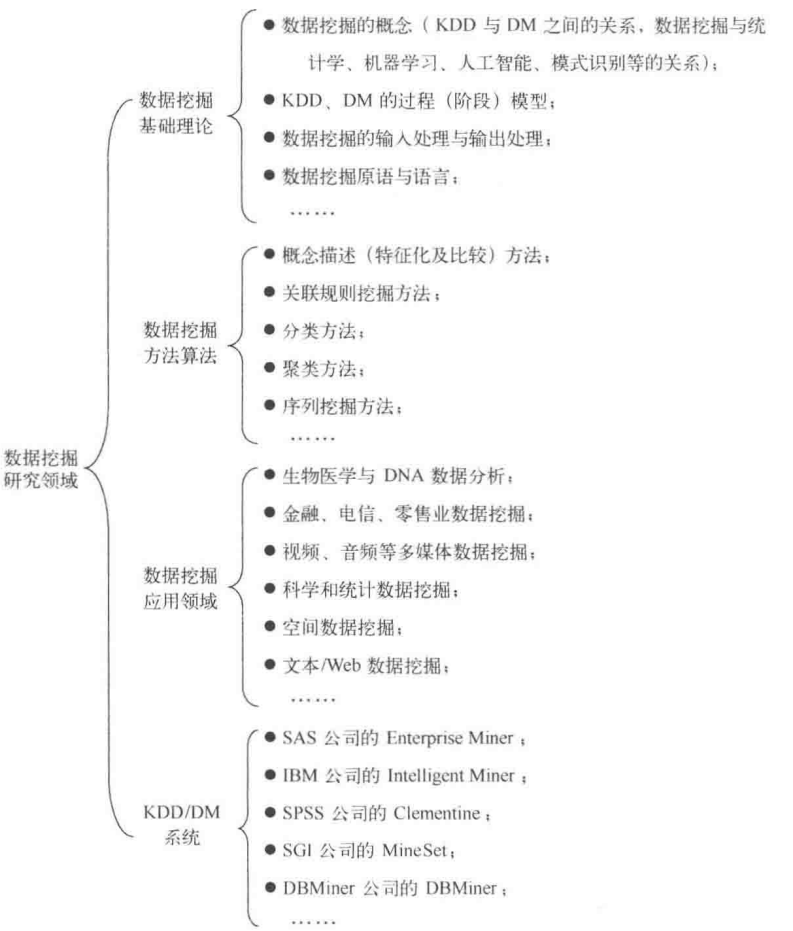

图 1.1　数据挖掘当前研究领域

1.1.1 分类方法

分类的目的是通过一个分类函数或分类模型（常称作分类器）将待分类数据集中的数据映射到某一个给定的类别中。分类在数据挖掘中是一项非常重要的任务，目前在商业上应用很多。

分类的基本目标就是构造分类器，要达到这个目标首先需要设计分类器。设计分类器的基本做法是用一定数量的样本（称为训练样本集）给出一套分类判别准则，使得按照这套分类判别准则对待分类数据进行分类所造成的识别错误率最小或引起的损失最小，训练完毕后对任何一个样本都可以利用分类器将它归到某一类别。

当前，能完成分类挖掘任务的算法比较多，如常见的决策树、Bayes、SVM、神经网络、粗糙集方法等。

1．决策树

构造一个决策树分类器通常分为两步：树的生成和剪枝。树的生成采用自上而下的递归分治法，如果当前训练实例集合中的所有实例是同类的，则构造一个叶节点，节点内容即是该类别，否则，根据某种策略选择一个属性，按照该属性的不同取值，把当前实例集合划分为若干子集合，对每个子集合重复此过程，直到当前集合中的实例是同类为止；剪枝就是剪去那些不会增大树的预测错误率的分支，经过剪枝，不仅能有效地克服噪声，还使树变得简单，容易理解。由于生成最优决策树是 NP-hard 问题，目前的决策树算法一般都采用启发式属性选择策略来解决最优决策树的生成问题。

由于决策树是一个有效的数据描述、分类、特征化工具，它的研究得到了很多学科的支持，如统计学、模式识别、决策理论、信号处理、机器学习、人工神经网络等。构造决策树的算法有很多，1986 年 J.Ross Quinlan 在 Machine Learning Journal 上发表了题为“Induction of Decision Trees”的论文，引入了一种新的 ID3 算法，随后，他对 ID3 算法进行了补充和改进，提出了后来非常流行的 C4.5 算法。ID3、C4.5 以及后来出现的 C4.5 的商业改进版本 C5.0 采用信息熵增益及其改进增益率进行属性选择，可以有效克服增益偏向于多值属性的缺点，这几种算法已经成为机器学习领域构造决策树的经典算法。

很多学者都从一定的角度提出了对 Quinlan 系列算法的扩充，如 S. Ruggieri 通过对 C4.5 算法决策树节点构造策略进行改进而提出的 EC4.5 算法、Dennis Shasha 发展的适于并行分布式分类的 PC4.5 等。

有些研究者对决策树在超大规模数据集中的应用作了研究，如 SLIQ，它采用预排序技术来克服需将所有数据放入内存的问题，从而能处理更大的数据库，并用 MDL 剪枝算法，使树更小和精度更高。

还有一种尝试是通过属性组合来构造多元决策树，一般的认识是多元测试形成的决策树较单元决策树的精度高，但是构造多元决策树的复杂度要远高于单元决策树。在实现途径上，属性组合可以采用逻辑组合，也可以采用代数组合。

2．Bayes 方法

贝叶斯统计分析起源于英国学者 T. Bayes 的一篇论文“An essay towards solving a problem in the doctrine of chances”，该文给出了著名的贝叶斯公式和一种归纳推理方法。其

后一些统计学家将其发展成一种系统的统计推断方法，到20世纪30年代形成了贝叶斯学派，20世纪五六十年代发展成了一个有影响的统计学派。

在数据挖掘中，主要有两种Bayes方法，即朴素Bayes方法和Bayes信念网络。前者直接利用Bayes公式进行预测，把从训练样本中计算出的各个属性值和类别频率比作为先验概率，并假定各个属性之间是独立的，这样就可以用Bayes公式和相应的概率公式计算出要预测实例对各类别的条件概率值。选取概率值最大的类别作为预测值。此方法简单易行并且具有较好的精度，缺点是不能刻画属性之间的依赖关系。

Bayes信念网络最早是由R.Howard和J.Matheson提出的，早期常见于专家系统，用于描述不确定的专家知识，Bayes信念网络是一个带有概率注释的有向无环图。这个图模型能有效地表示大的变量集合的联合概率分布，从而适合用来分析大量变量之间的相互关系，利用Bayes公式的学习和推理功能，实现预测、分类等数据挖掘任务。

训练Bayes信念网络需要进行网络结构和网络参数两部分的学习。学习过程和学习方法的推导可见W.Buntin、G.Cooper等人的文献。如果网络结构确定，Bayes信念网络的训练主要是条件概率表（CPT）的计算，方法与朴素Bayes分类的方法类似；如果网络结构不确定，则两部分的学习过程都要进行，但获得最优的结构和参数都是NP问题，因此在训练过程中可以采用启发式方法。

3．SVM

统计学习理论（SLT）是由Vapnik等人提出的一种小样本统计理论，着重研究在小样本情况下的统计规律及学习方法性质。SLT为机器学习问题建立了一个较好的理论框架，也发展了一种新的通用学习算法——支持向量机（SVM），能够较好地解决小样本学习问题，已初步表现出了很多优于已有方法的性能。一些学者认为，SLT和SVM正在成为继神经网络研究之后新的研究热点，并将推动机器学习理论和技术的快速发展。

支持向量机理论的最大特点是根据Vapnik结构风险最小化准则，尽量提高学习机的泛化能力，即由有限的训练集样本得到的小的误差能够保证对独立的测试集仍保持小的误差。另外由于支持向量机算法是一个凸优化问题，因此局部最优解一定是全局最优解，且SVM的复杂度和实例集的维数无关。

SVM的基本思想是通过某种事先选择的非线性映射将输入向量映射到一个高维特征空间，然后在这个空间中构造最优分类超平面。由于在高维特征空间中构造最优超平面，只需要计算特征向量与特征空间中向量的内积，然后使用某种核函数在原空间计算就可以了，从而克服了维数困难。通过选用不同的核函数，可以构造输入空间中不同类型的非线性决策面的学习机。

对于分类问题，支持向量机算法根据区域中的样本计算该区域的决策曲面，由此确定该区域中未知样本的类别，对于估值问题，支持向量机算法对区域中的样本进行回归，确定该区域的映射函数,从而得到该区域中未知样本的取值。

由于SVM方法较好的理论基础和它在一些领域（如手写数字识别）中表现出的优秀推广性能，近年来，许多关于SVM方法的研究，包括算法本身的改进和算法的实际应用，都被陆续提了出来。尽管SVM算法的性能在许多实际问题的应用中得到了验证，但是该算法在计算上存在着一些问题，包括训练算法速度慢、算法复杂而难以实现以及检测阶段

运算量大等。由于传统的利用标准二次型优化技术解决对偶问题的方法可能是训练算法慢的主要原因，近年来人们针对 SVM 方法本身的特点提出了许多算法来解决对偶寻优问题，这些算法的一个共同思想就是循环迭代：将原问题分解成为若干子问题，按照某种迭代策略，通过反复求解子问题，最终使结果收敛到原问题的最优解。根据子问题的划分和迭代策略的不同，又可以大致分为两类，一类是所谓的“块算法”（Chunking Algorithm），另一类是固定工作样本集的方法。块算法和固定工作样本集方法的主要区别在于：块算法的目标函数中仅包含当前工作样本集中的样本，而固定工作样本集方法虽然优化变量仅仅包含工作样本，但其目标函数却包含整个训练样本集。

4．神经网络

神经网络在过去十几年里取得了飞速的发展，发展出了很多模型及其改进型，目前，在应用和研究中采用的神经网络模型不下 30 种，其中较有代表性的大约有十几种，例如 BP、Hopfield、Kohonen、 ART 等，但人工神经网络在知识获取方面存在先天不足：由于神经网络的知识获取过程是一个“黑箱”系统，得到的知识也是以权值形式表现的隐式知识，因此难以被人理解，而在分析和决策领域，一个没有推理过程和决策依据的结论是很难被分析员和决策者所接受的。

要克服神经网络无法获取显式规则的先天不足，就要对神经网络在知识获取和知识表达两方面进行改进。

在知识获取（Knowledge Extraction）方面是给定一个已经训练好的神经网络，从中提取显式的知识(一般是符号形式)，提取的方法一般分为分解抽取方法和学习抽取方法两类。分解抽取方法的最大特点是对神经元网络内部的单个节点所表示的概念进行解释，从每个节点中抽取的规则是由与此节点相连的诸输入节点表示的，典型的分解算法有 SUBSET 及其改进 MOFN、KT 和 RULEX；学习抽取方法的基本思想是将训练后的神经元网络看成一个黑箱，而把规则抽取过程看成一个学习过程，其中所学的目标概念由神经元网络函数计算得到，而其输入变量则由神经元网络的输入特征组成。学习抽取方法主要利用符号学习算法作为学习工具，而利用神经元网络作为学习例子生成器，主要代表方法有 TREPAN 和 RL。

在知识表达（Knowledge Representation）方面是让神经网络中抽象的权值能够代表一定的知识，例如使权值代表规则的编码，这样在网络训练结束后，通过解码就可以得到规则。这种方法已经在 DNA 结构分析、自然语言处理的语法规则提取和化学反应的预测中得到了应用。

1.1.2 聚类方法

聚类是一种普遍使用的数据分析方法，在统计学、机器学习和数据挖掘上都有应用。在机器学习领域，聚类是无指导学习（Unsupervised Learning）的一个例子，它的主要特点是不依赖预先定义的类信息作为指导；在数据挖掘领域，聚类的研究主要集中在为大型数据库提供高效而实用的聚类方法支撑。

聚类是按照某个特定标准（通常是某种距离）把一个数据集分割成不同的类，使得类内相似性尽可能的大，同时类间的区别性也尽可能的大。直观地说，最终形成的每个聚类，

在空间上都是一个稠密的区域。

聚类方法主要分为如下几类：

（1）划分方法（Partitioning Method）。该方法首先得到一个初始划分，然后采用迭代重定位技术，试图通过将对象从一个簇移到另一个簇来改进划分的质量，具有代表性的包括 *k*-means、*k*-medoids、CLARANS 等启发式方法。

（2）层次方法（Hierarchical Method）。层次聚类方法可以分为分裂法（Divisive）和聚合法（Agglomerative），后者把实例集合看作单独的类，自下而上地合并；前者则相反，先把整个实例集作为一个类，逐渐分裂。层次聚类方法常见的有 BIRCH、CURE、Chameleon 等。

（3）基于密度的方法（Density-Based Method）。它根据密度的概念而不是通常使用的距离来聚类对象，常见的有 DBSCAN、DENCKUE、OPTICS 等。

（4）基于网格的方法（Grid-Based Method）。它首先将对象空间分划为有限数目单元形成的网格结构，然后在网格结构上进行聚类，典型的有 STING、CLIQUE、WaveCluster 等。

（5）基于模型的方法（Model-Based Method）。首先对每个簇假设一个模型，然后进行数据与模型的最佳匹配，有代表性的方法有 COBWEB、CLASSIT、AutoClass 等。

聚类是无指导的学习方法，其所研究的数据没有类别标签，于是就很难判断得到的聚类划分是否反映了事物的本质，Ada 等人对此问题作了初步探讨。

1.1.3 文本挖掘

文本挖掘与其他挖掘的不同在于文本挖掘的对象是所谓的半结构化数据（Semistructure Data），它既不是完全无结构化的也不是完全结构化的，例如，一个文档中可能包含：标题、作者、出版日期、长度、分类等结构字段，也可能有大量非结构化的成分，如摘要、文本正文等。文本挖掘的内容主要有文本检索、文档分类、自动摘要及自然语言处理等。

文本挖掘的一个重要基础研究是如何对半结构化数据进行建模和表示。目前文档的表示大都采用特征空间方法，即文档被表示成一个个独立的词及其出现频率，从中选出能够代表文档的词作为特征向量，每个文档由一个特征向量表示。对于一组文档可由这些特征向量组成一个词频矩阵。

对于文本检索来说，常用的方法有 TFIDF（Term Frequency/Inverse Document Frequency）、Bool 检索、语义网络等，检索的基本度量标准是查准率（Precision）和查全率（Recall）；文本的分类有许多方式，几乎传统的分类方法都可以用，常用的有 TFIDF、朴素 Bayes、ANN 等，其中 Wang 提出了基于关联挖掘的自动文档分类方法，Ipeirotis 提出了对只能通过查询进行访问的文本数据库的自动分类方法。

1.1.4 Web 挖掘

Web 挖掘可以看成是文本挖掘的扩展，它不但有文本挖掘的内容，即 Web 内容挖掘，同时又有 Web 链接结构和页面属性挖掘（路径分析、关联页面、页面 Ranking、Authoritative 页面等）、Web 使用挖掘（日志挖掘、用户模式挖掘等）。

Web 挖掘有如下特点：Web 数据量庞大；页面内容和组织结构复杂；动态性强；用户群体复杂，这些特点决定了 Web 挖掘必须有稳健而高效的挖掘算法作支撑。

利用挖掘 Web 链接结构来识别 Authoritative 页面的方法可参见 Chakrabarti 和 Kleinberg 等人的研究成果；利用 Hub 页面来寻找 Authoritative 页面可运用 HITS（Hyperlink-Induced Topic Search）算法；页面 Ranking 的排列可运用 Brin 和 Page 提出的算法。在 Web 日志挖掘方面，有 Perkowitz 提出的通过挖掘用户访问模式进而自动构造可适应 Web 站点的算法；有 Tauscher 提出的挖掘 Web 可用性的方法等。

1.2 数据挖掘的过程模型

如前所述，数据挖掘有广义和狭义两种理解，广义理解的数据挖掘和 KDD 都涵盖了知识发现的整个过程，下面对数据挖掘的基本过程和步骤进行建模。

1.2.1 知识发现的基本过程分析

从源数据中发现有用的知识是一项系统工程。一般地说，其过程可以简单地概括为：首先从数据源中抽取感兴趣的数据，并把它组织成适合挖掘的组织形式；然后，调用相应的算法产生所需的模式和规则；最后对生成的模式进行评估，确认生成的规律和知识，并把有价值的知识集成到已有的管理系统中。具体来说，一般应具有如下关键步骤。

1．数据的清洗和抽取

如前所述，在开始一个知识发现项目之前必须清晰地定义挖掘目标。虽然挖掘的最后结果是不可预测的，但是要解决或探索的问题应该是可预见的。盲目地挖掘是没有任何意义的。在弄清业务问题后就可以进行数据的准备，包括数据的清洗和抽取等环节。

数据清洗是指去除或修补源数据中的不完整、不一致、含噪声的数据。数据不完整是指由于人为疏忽、未及时登记、保密措施限制等原因使数据分析人员无法得到某些数据项。假如这个数据项正是知识发现系统所关心的，那么这类不完整的数据就需要修补。常见的不完整数据的修补办法有：

（1）使用一个全局值来填充（如“Unkown”、估计的最大数或最小数）。

（2）统计该属性的所有非空值，并用平均值来填充空缺项。

（3）只使用同类对象的属性平均值填充。

（4）利用回归或工具预测最可能的值，并用它来填充。

数据不一致可能是由于源数据库中对同样属性所使用的数据类型、度量单位等不同而导致的。因此需要定义它们的转换规则，并在挖掘前统一成一个形式；噪声数据是指那些明显不符合逻辑的偏差数据（如飞参系统在某些情况下采集到的量化参数值），这样的数据往往影响挖掘结果的正确性。目前讨论最多的处理噪声数据的方法是数据平滑（Data Smoothing）技术。主要有：①利用分箱（Binning）方法检测周围相应属性的值来进行局部数据平滑；②利用聚类技术检测孤立点数据，对它们进行修正；③利用回归函数探测和修正噪声数据。