SHUJU WAJUE JISHU

# 数据挖掘技术

陈文伟　黄金才　赵新昱　著

北京工业大学出版社

## 内 容 提 要

本书以数据挖掘技术的方法为线索，以数据挖掘的技术成果为重点，对知识发现与数掘挖掘的概念、数据挖掘的方法和技术、数据挖掘的知识表示、基于信息论的数据挖掘方法、关联规则挖掘、神经网络、遗传算法、公式发现、数据挖掘的应用等内容作了全面、系统的介绍。

本书是数据挖掘技术最新成果的论著，将对数据挖掘技术在我国的发展起到推动作用。

**图书在版编目(CIP)数据**

数据挖掘技术/陈文伟，黄金才，赵新昱著，—北京:北京工业大学出版社,2002.12

ISBN 7-5639-1205-3

Ⅰ.数... Ⅱ.①陈...②黄...③赵... Ⅲ.数据采集 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2002)第 104745 号

**数据挖掘技术**

陈文伟 黄金才 赵新昱 著

※

北京工业大学出版社出版发行

邮编:100022 电话:(010)67392308

各地新华书店经销

徐水宏远印刷厂印刷

※

2002年12月第1版 2002年12月第1次印刷

787 mm × 1092 mm 16开本 14印张 349千字

印数: 1～3000册

ISBN 7-5639-1205-3/T·195

定价: 25.00元

# 序　言

数据挖掘技术是人工智能中的机器学习和数据库技术结合而发展起来的新技术。当前数据挖掘已经从传统的关系型数据库的挖掘发展到文本、多媒体、Web的数据挖掘。数据挖掘已经成为数据仓库、决策支持系统的重要组成部分。数据挖掘技术的出现引起学者的广泛关注，在国内外已经形成研究热潮。

陈文伟等同志长期从事机器学习和智能决策技术研究。从国外1996年数据挖掘概念提出之后，作者就开始跟踪研究，并于1997年在《计算机世界》报专题版上最早向国内读者介绍了数据挖掘技术。作者及课题组成员对数据挖掘技术的研究得到了国家自然科学基金项目的资助，经过多年的研究，在数据挖掘领域取得了多项成果。

本书系统介绍了各类经典的数据挖掘技术，如决策树方法ID 3和C 4.5，集合论方法AQ，关联规则发现方法Apriori和FP算法，公式发现的BACON系统，神经网络和遗传算法等。本书还重点介绍了作者在数据挖掘领域研制的部分成果，具体有：基于信道容量的IBLE方法，超曲面神经网络Cover模型与CC模型，FDD公式发现系统，周期关联规则发现CCAR方法，遗传分类学习系统GCLS等。

相信本书能够为我国数据挖掘技术的发展起到一定的推动作用。

汪　浩

2002年10月2日

# 前　言

数据挖掘(data mining,DM)是20世纪90年代中期兴起的新技术,它是从数据库中发现知识(knowledge discovery in database,KDD)的主要步骤。随着时代的前进和技术的进步,各类数据库和数据在急剧地增长,如何从大量的数据中获取知识就成了人们关注的焦点。数据挖掘就是从数据中挖掘出知识的技术。

数据挖掘实质上是从人工智能中的机器学习演变过来的。机器学习是让计算机模拟人的学习方法获取知识。人工智能从20世纪50年代兴起时就开始了机器学习的研究,当时的典型成果是神经网络的感知机模型中的网络权值学习和下西洋跳棋的启发式函数。到90年代,出现了基于信息论的ID3、IBLE方法,基于集合论的粗糙集方法、关联规则发现方法,仿生物学的神经网络和遗传算法,以及公式发现的BACON方法和FDD方法等一大批的数据挖掘方法。这些方法在实践中都取得了显著的成果。

数据挖掘主要是在关系数据库中进行,随着90年代初多媒体技术的发展,产生了文本、图形、视频等不同数据类型的数据挖掘。

90年代中期兴起的数据仓库,已明确地把数据挖掘作为它重要的分析工具。数据仓库中的大量数据($10^9$字节～$10^{12}$字节)及多维数据组织模型为数据挖掘提出了新的要求,也促进了数据挖掘的发展。

90年代初Internet网络的迅速发展,使Web数据挖掘成了热门研究课题。

数据挖掘技术涉及人工智能的机器学习、数据库与数据仓库、统计学、可视化图形学等各个领域,吸引了大批学者从事学术研究和工具产品的开发。在20世纪90年代中后期,在国外数据挖掘已经形成高潮,我国研究数据挖掘的学者数量也在迅速增长。为了促进数据挖掘技术的发展,我们撰写本书,介绍数据挖掘的各种已成熟的方法和技术,同时也介绍我们在数据挖掘方面所做的工作。

我指导的博士生与硕士生完成的数据挖掘工作有:基于信道容量的IBLE方法(钟鸣)、基于归一化互信息的IBLE-R方法(赵东升)、公式发现系统FDD(赵新昱、张帅等)、超曲面神经网络CC模型和Cover模型(黄金才)、遗传分类学习系统GCLS(邹雯)、数据挖掘服务器(黄金才、赵新昱、何义等)以及粗糙集的研究(赛英、马建军等)、知识发现过程的研究(陈元)等。

我们在数据挖掘和数据仓库方面的研究工作,得到了国家自然科学基金的资助。

我们愿意和大家一道共同探讨数据挖掘技术,听取大家宝贵的意见,共同推动我国数据挖掘技术的发展。

陈文伟

2002年5月

# 目　录

# 第 1 章　知识发现与数据挖掘综述

哲学家培根的名言:知识就是力量。知识如何获得? 人类获得知识的方法是通过不断学习和反复实践。人们从小学到中学到大学,共学习 16 年左右。要获得更多知识还要攻读硕士、博士学位,再加大约 6 年。人们在参加工作后,用学到的知识解决问题,丰富、实践知识,同时创造新知识,解决新问题。人们就是这样在学习和实践中,不断丰富知识和强化知识。

用计算机帮助人们获取知识开始于 20 世纪 50 年代。神经网络感知机模型能对线性样本进行学习获得网络权值知识,它不但能识别训练样本,而且能鉴别新样本。Samuel 下棋程序能通过人机对弈自适应修改启发式函数中的系数,提高机器获胜的能力。最后,该下棋程序能胜过该程序的研制者。1980 年第一次国际机器学习研讨会召开,标志"机器学习"学科成立。它属于人工智能研究领域。我国也于 1987 年召开了第一届全国机器学习研讨会,并成立了"中国机器学习学会"。由于人工智能专家系统出现知识获取"瓶颈"后,"机器学习"引起了人们的重视并得到迅速的发展。

从数据库中发现知识(knowledge discovery in database,KDD)是 20 世纪 80 年代末开始的。KDD 一词是在 1989 年 8 月在美国底特律市召开的第一届 KDD 国际学术会议上正式形成的。KDD 研究的问题有:(1)定性知识和定量知识的发现;(2)知识发现方法;(3)知识发现的应用等。

1995 年在加拿大召开了第一届知识发现和数据挖掘(data mining,DM)国际学术会议。由于把数据库中的"数据"形象地比喻成矿床,"数据挖掘"一词很快流传开来。

数据挖掘是知识发现中的核心工作,主要研究发现知识的各种方法和技术。机器学习的很多方法都已转变为数据挖掘的方法。

## 1.1　知识发现和数据挖掘的概念

### 1.1.1　定义

知识发现(KDD)被认为是从数据中发现有用知识的整个过程。数据挖掘被认为是 KDD 过程中的一个特定步骤,它用专门算法从数据中抽取模式(pattern)。

KDD 过程定义为(Fayyad 等,1996):

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高级处理过程。其中,数据集:事实 $F$(数据库记录)的集合;模式:用语言 $L$ 表示的表达式 $E$,它所描述的数据是集合 $F$ 的一个子集 $F_E$,它比枚举所有 $F_E$ 中元素更简单,我们称 $E$ 为模式;有效、新颖、潜在有用、可被人理解:表示发现的模式有一定的可信度,应该是新的、将来有实用价值的、能被用户所理解的。

KDD 过程图如图 1.1 所示。

KDD 过程可以概括为三部分:数据准备(data preparation)、数据挖掘(data mining)及结果

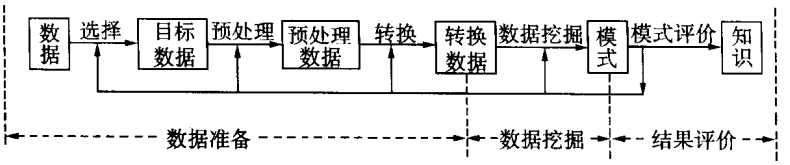


图 1.1 KDD过程图

的解释和评估(interpretation & evaluation)。

#### 1.1.1.1 数据准备

数据准备又可分为三个子步骤:数据选取(data selection)、数据预处理(data preprocessing)和数据变换(data transformation)。

数据选取的目的是确定发现任务的操作对象,即目标数据(target data),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数据,以便于符号归纳,或是把离散型数据转换为连续值型数据,以便于神经网络计算)等。数据变换的主要目的是消减数据维数或降维(dimension reduction),即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量个数。

#### 1.1.1.2 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的,如数据分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后,就要决定使用什么样的挖掘算法。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挖掘;二是要根据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(descriptive)、容易理解的知识(采用规则表示的挖掘方法显然要好于神经网络之类的方法),而有的用户只是希望获取预测准确度尽可能高的预测型(predictive)知识。选择了挖掘算法后,就可以实施数据挖掘操作,获取有用的模式。

#### 1.1.1.3 结果的解释和评估

数据挖掘阶段发现出来的模式,经过评估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,这时则需要回退到发现过程的前面阶段,如重新选取数据,采用新的数据变换方法,设定新的参数值,甚至换一种挖掘算法等等。另外,KDD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视化,或者把结果转换为用户易懂的另一种表示,如把分类决策树转换为"IF…THEN…"规则。

数据挖掘仅仅是整个过程中的一个步骤。数据挖掘质量的好坏有两个影响要素:一是所采用的数据挖掘技术的有效性,二是用于挖掘的数据的质量和数量(数据量的大小)。如果选择了错误的数据或不适当的属性,或对数据进行了不适当的转换,则挖掘的结果是不会好的。

整个挖掘过程是一个不断反馈的过程。比如,用户在挖掘途中发现选择的数据不太好,或使用的挖掘技术产生不了期望的结果。这时,用户需要重复先前的过程,甚至从头重新开始。

可视化技术在数据挖掘的各个阶段都扮演着重要的角色。特别是在数据准备阶段,用户

可能要使用散点图、直方图等统计可视化技术来显示有关数据，以期对数据有一个初步的了解，从而为更好地选取数据打下基础。在挖掘阶段，用户则要使用与领域问题有关的可视化工具。在表示结果阶段，则可能要用到可视化技术以使得发现的知识更易于理解。

## 1.1.2 数据挖掘任务

数据挖掘任务有6项：关联分析、时序模式、聚类、分类、偏差检测、预测。

### 1.1.2.1 关联分析

关联分析是从数据库中发现知识的一类重要方法。若两个或多个数据项的取值之间重复出现且概率很高时，就存在某种关联，可以建立起这些数据项的关联规则。

例如，买面包的顾客有90%的人还买牛奶，这是一条关联规则。若商店中将面包和牛奶放在一起销售，将会提高它们的销量。

在大型数据库中，这种关联规则是很多的，需要进行筛选。一般用“支持度”和“可信度”两个阈值来淘汰那些无用的关联规则。

“支持度”表示该规则所代表的事例(元组)占全部事例(元组)的百分比。如既买面包又买牛奶的顾客占全部顾客的百分比。

“可信度”表示该规则所代表事例占满足前提条件事例的百分比。如既买面包又买牛奶的顾客占买面包顾客中的90%，则可信度为90%。

### 1.1.2.2 时序模式

通过时间序列搜索出重复发生概率较高的模式。这里强调时间序列的影响。例如，在所有购买了激光打印机的人中，半年后80%的人再购买新硒鼓，20%的人用旧硒鼓装碳粉；在所有购买了彩色电视机的人中，有60%的人再购买VCD产品。

在时序模式中，需要找出在某个最小时间内出现比率一直高于某一最小百分比(阈值)的规则。这些规则会随着形式的变化做适当的调整。

时序模式中，一个有重要影响的方法是“相似时序”。用“相似时序”的方法，要按时间顺序查看时间事件数据库，从中找出另一个或多个相似的时序事件。例如在零售市场上，找到另一个有相似销售的部门，在股市中找到有相似波动的股票。

### 1.1.2.3 聚类

数据库中的数据可以划分为一系列有意义的子集，即类。在同一类别中，个体之间的距离较小，而不同类别的个体之间的距离偏大。聚类增强了人们对客观现实的认识，即通过聚类建立宏观概念。例如鸡、鸭、鹅等都属于家禽。

聚类方法包括统计分析方法、机器学习方法和神经网络方法等。

在统计分析方法中，聚类分析是基于距离的聚类，如欧氏距离、海明距离等。这种聚类分析方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分。

在机器学习方法中，聚类是无导师的学习。在这里距离是根据概念的描述来确定的，故聚类也称概念聚类，当聚类对象动态增加时，概念聚类则称为概念形成。

在神经网络中，自组织神经网络方法用于聚类。如ART模型、Kohonen模型等，这是一种无监督学习方法。当给定距离阈值后，各样本按阈值进行聚类。

#### 1.1.2.4 分类

分类是数据挖掘中应用得最多的任务。分类是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述。一般用规则或决策树模式表示。该模式能把数据库中的元组映射到给定类别中的某一个。

类的内涵描述分为：特征描述和辨别性描述。

特征描述是对类中对象的共同特征的描述。辨别性描述是对两个或多个类之间的区别的描述。特征描述允许不同类中具有共同特征。而辨别性描述中不同类不能有相同特征。辨别性描述用得更多。

分类是利用训练样本集(已知数据库元组和类别所组成的样本)通过有关算法而求得。

建立分类决策树的方法，典型的有 ID3，C4.5，IBLE 等方法。建立分类规则的方法，典型的有 AQ 方法、粗集方法、遗传分类器等。

目前，分类方法的研究成果较多，判别方法的好坏，可从三个方面进行：

(1)预测准确度(对非样本数据的判别准确度)；

(2)计算复杂度(方法实现时对时间和空间的复杂度)；

(3)模式的简洁度(在同样效果情况下，希望决策树小或规则少)。

在数据库中，往往存在噪声数据(错误数据)、缺损值和疏密不均匀等问题。它们对分类算法获取的知识将产生不良的影响。

#### 1.1.2.5 偏差检测

数据库中的数据存在很多异常情况。从数据分析中发现这些异常情况也是很重要的，应引起人们对它更多的注意。

偏差包括很多有用的知识，如：

(1)分类中的反常实例；

(2)模式的例外；

(3)观察结果对模型预测的偏差；

(4)量值随时间的变化。

偏差检测的基本方法是寻找观察结果与参照之间的差别。观察常常是某一个域的值或多个域值的汇总。参照是给定模型的预测、外界提供的标准或另一个观察。

#### 1.1.2.6 预测

预测是利用历史数据找出变化规律，建立模型，并用此模型来预测未来数据的种类、特征等。

典型的方法是回归分析，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。预测时，只要输入任意的时间值，通过回归方程就可求出该时间的状态。

近年来，发展起来的神经网络方法，如 BP 模型，实现了非线性样本的学习，能进行非线性函数的判别。

分类也能进行预测，但分类一般用于离散数值；回归预测用于连续数值；神经网络方法预测既可以用于连续数值，也可以用于离散数值。

### 1.1.3 数据挖掘分类

数据挖掘涉及多个学科方向，主要包括：数据库、统计学和机器学习等。

数据库技术经过20世纪80年代的大发展，除关系数据库外，又陆续出现面向对象数据库、多媒体数据库、分布式数据库以及Web数据库等。数据库的应用由一般查询到模糊查询和智能查询，数据库计算已趋向并行计算。从以上各类数据库中挖掘知识正在兴起并已得到迅速发展。

统计学是一门古老学科，现已逐渐走向社会。它已成为社会调查、了解民意以及制定决策的重要手段。

机器学习是人工智能的重要分支。它是在专家系统获取知识出现“瓶颈”后发展起来的。机器学习的大部分方法和技术已演变为数据挖掘方法和技术。

数据挖掘可按数据库类型、挖掘对象、挖掘任务、挖掘方法与技术以及应用等几方面进行分类。

#### 1.1.3.1 按数据库类型分类

数据挖掘主要是在关系数据库中挖掘知识。随数据库类型的不断增加，逐步出现了不同数据库的数据挖掘。现有：关系数据挖掘、模糊数据挖掘、历史数据挖掘、空间数据挖掘等多种不同数据库的数据挖掘类型。

#### 1.1.3.2 按数据挖掘对象分类

数据挖掘除对数据库这个主要对象进行挖掘外，还有文本数据挖掘、多媒体数据挖掘、Web数据挖掘。由于对象不同，挖掘的方法相差很大。文本、多媒体、Web数据均是非结构化数据，挖掘的难度很大。

目前Web数据挖掘已逐步引起人们的关注。

#### 1.1.3.3 按数据挖掘任务分类

数据挖掘的任务有：关联分析、时序模式、聚类、分类、偏差检测、预测等。按任务分类有：关联规则挖掘、序列模式挖掘、聚类数据挖掘、分类数据挖掘、偏差分析挖掘和预测数据挖掘等类型。

各类数据挖掘由于任务不同，将会采用不同的数据挖掘方法和技术。

#### 1.1.3.4 按数据挖掘方法和技术分类

数据挖掘方法和技术较多，在下一节中将详细讨论。在此对其分类进行说明。

**1. 归纳学习类**

该类又分为基于信息论方法挖掘类和基于集合论方法挖掘类。基于信息论方法是在数据库中寻找信息量大的属性来建立属性的决策树。基于集合论方法是对数据库中各属性的元组集合之间关系(上、下近似关系，覆盖或排斥关系，包含关系等)来建立属性间的规则。各类中又包括多种方法，主要用于分类问题。

**2. 仿生物技术类**

该类又分为神经网络方法类和遗传算法类。神经网络方法是在模拟人脑神经元而建立的MP数学模型和Hebb学习规则基础上，提出的一系列算法模型，用于识别、预测、联想、优化和

聚类等实际问题。遗传算法是在模拟生物遗传过程,对选择、交叉、变异过程建立的数学算子。主要用于问题的优化和规则的生成。

**3. 公式发现类**

在科学实验与工程数据库中,用人工智能方法寻找和发现连续属性(变量)之间关系,建立变量之间公式,已引起人们的关注,该类中有多种数据挖掘方法。

**4. 统计分析类**

统计分析是一门独立学科,由于能对数据库中数据求出各种不同的统计信息和知识,故它也构成了数据挖掘中一大类方法。

**5. 模糊数学类**

模糊数学是反映人们思维的一种方式。将模糊数学应用于数据挖掘各项任务中,形成了模糊数据挖掘类。如模糊聚类、模糊分类、模糊关联规则等。

**6. 可视化技术类**

可视化技术是一种图形显示技术。对数据的分布规律进行可视化显示或对数据挖掘过程进行可视化显示,会明显提高人们对数据挖掘的兴趣和挖掘效果。该技术已形成了可视化数据挖掘类的多种方法。

本书的内容将按照数据挖掘的方法和技术分类,对具体的方法进行详细和深入的介绍,以便读者学习和使用这些方法和技术,对实际问题完成数据挖掘任务。

## 1.1.4 数据挖掘对象

数据挖掘的对象主要是关系数据库,这是典型的结构化数据。随着技术的发展,数据挖掘对象逐步扩大到半结构化或非结构化数据,这主要是指文本数据、图像和视频数据,以及 Web 数据等。

### 1.1.4.1 关系数据库

目前,建立的数据库都是关系数据库。数据挖掘方法也主要是研究数据库中属性之间的关系,挖掘出多个属性取值之间的规则。由于关系数据库的特点,促使了数据挖掘方法的改善。数据库的特点如下。

**1. 数据动态性**

数据的动态变化是数据库的一个主要特点。由于数据的存取和修改,使数据的内容经常发生变化,这就要求数据挖掘方法能适应这种变化。渐增式数据挖掘方法就是针对数据变化后,挖掘的规则知识能满足变化后的数据库内容。

**2. 数据不完整性**

数据的不完整性主要反映在数据库中记录的域值丢失或不存在(空值)。这种不完整数据给数据挖掘带来了困难。为此,必须对数据进行预处理,填补该数据域的可能值。

**3. 数据噪声**

由于数据录入等原因,造成错误的数据,即数据噪声。含噪声的数据挖掘会影响抽取模式的准确性,并增加了数据挖掘的困难度。

在数据挖掘中要考虑噪声的影响,利用概率方法排除这些噪声。

**4. 数据冗余性**

数据冗余性表现为同一信息在多处重复出现。函数依赖是一个通常的冗余形式。冗余信

息可能造成错误的数据挖掘，至少有些挖掘的知识是用户不感兴趣的。为避免这种情况的发生，数据挖掘时，需要知道数据库中有哪些固有的依赖关系。

**5. 数据稀疏性**

数据稀疏性表现为实例空间中数据稀疏，数据稀疏会使数据挖掘丢失有用的模式。

**6. 海量数据**

数据库中数据在不断增长，已出现很多海量数据库。数据挖掘方法需要逐步适应这种海量数据挖掘，如建立有效的索引机制和快速查询方法等。

### 1.1.4.2 文本

文本是以文字串形式表示的数据文件。文本分析包括：关键词或特征提取、相似检索、文本聚类和文本分类等。

**1. 关键词或特征提取**

一篇文本中，标题是该文本的高度概括。标题中的关键词是标题的核心内容。关键词的提取对于掌握该文本的内容至关重要。

文本中的特征如人名、地名、组织名等是某些文本中的主体信息，特征提取对掌握该文本的内容很重要。

**2. 相似检索**

文本中的关键词的相似检索是了解文本内容的一种重要方法。例如“专家系统”与“人工智能”两个关键词是有一定联系的。研究专家系统的文本一定属于人工智能的研究领域。

**3. 文本聚类**

对于文本标题中关键词（主题字）的相似匹配是对文本聚类的一种简单方法。定义关键词的相似度，将便利文本的简单聚类，类中文本满足关键词的相似度，类间文本的关键词超过相似度。

**4. 文本分类**

将文本分类到各文本类中，一般需要采用一个算法。这些算法包括分类器算法、近邻算法等。这需要按文本中的关键字或特征的相似度来区分。

### 1.1.4.3 图像与视频数据

图像和视频数据是典型多媒体数据。数据以点阵信息及帧的形式存储，数据量很大。图像与视频的数据挖掘包括：图像与视频特征提取、基于内容的相似检索、视频镜头的编辑与组织等。

**1. 图像与视频特征提取**

图像与视频数据特征有颜色、纹理和形状等。这些特征提取用于基于内容的相似检索。海水蓝色、海滩黄色、房屋的形状及颜色等需要从大量图像和视频数据中提取。

**2. 基于内容的相似检索**

根据图像、视频特征的分布、比例等进行基于内容的相似检索，可以将图像和视频数据进行聚类以及分类，也能完成对新图像或视频的识别。如对遥感图像或视频的识别，这种应用非常广泛，例如森林火灾的发现与报警、河流水灾的预报等。

**3. 视频镜头的编辑与组织**

镜头代表一段连续动作（视频数据流）。典型的镜头编辑如足球赛的射门、某段新闻节目

等,需要在冗长的视频数据流中进行自动裁取。

经过编辑的镜头,按某种需要重新组织,将形成特定需求的新视频节目,如足球射门集锦、某个新闻事件的连续报道等。

#### 1.1.4.4 Web 数据

随着 Internet 网的发展和普及,网站数目迅速增长,以及入网人员剧烈增多,使网络可提供的数据量呈指数增长。Web 数据挖掘已成为新课题。Web 数据挖掘的特点如下。

**1. 异构数据集成和挖掘**

Web 上每一个站点是一个数据源,各数据源都是异构的,形成了一个巨大的异构数据库环境。将这些站点的异构数据进行集成,给用户提供一个统一的视图,才能在 Web 上进行数据挖掘。

**2. 半结构化数据模型抽取**

Web 上的数据非常复杂,没有特定的模型描述。虽然每个站点上的数据是结构化的,但各自的设计对整个网络是一个非完全结构化的数据,称为半结构化数据。

对半结构化数据模型的查询和集成,需要寻找一种半结构化模型抽取技术来自动抽取各站点的数据。

XML 是一种半结构化的数据模型,容易实现 Web 中信息共享与交换。

Net Perceotian 公司采用了"实时建议"技术,能够根据用户以往的浏览行为来预测该用户以后的浏览行为,从而为用户提供个性化的浏览建议。

Web 数据挖掘正在逐步形成热点。

## 1.2 数据挖掘方法和技术

数据挖掘方法是由人工智能、机器学习的方法发展而来,结合传统的统计分析方法、模糊数学方法以及科学计算可视化技术,以数据库为研究对象,形成的数据挖掘的方法和技术。

数据挖掘的方法和技术可以分为六大类。

### 1.2.1 归纳学习方法

归纳学习方法是目前重点研究的方向,研究成果较多。从采用的技术上看,分为两大类:信息论方法(这也是常说的决策树方法)和集合论方法。每类方法又包含多个具体方法。

#### 1.2.1.1 信息论方法(决策树方法)

信息论方法是利用信息论的原理建立决策树。由于该方法最后获得的知识表示形式是决策树,故一般文献中称它为决策树方法。该类方法的实用效果好,影响较大。

信息论方法中较有特色的方法以下几种。

**1. ID3 等方法**

Quinlan 研制的 ID3 方法是利用信息论中互信息(信息增益)寻找数据库中具有最大信息量的字段,建立决策树的一个结点,再根据字段的不同取值建立树的分枝,再由每个分枝的数据子集重复建树的下层结点和分枝的过程,这样就建立了决策树。数据库愈大这种方法效果愈好。ID3 方法在国际上影响很大。ID3 方法以后又陆续开发了 ID4、ID5、C4.5 等方法。

**2. IBLE(information-based learning from examples)方法**

我们研制的 IBLE 方法,是利用信息论中信道容量,寻找数据库中信息量从大到小的多个字段的取值建立决策规则树的一个结点,根据该结点中指定字段取值的权值之和与两个阈值比较,建立左、中、右三个分枝,在各分枝子集中重复建树结点和分枝的过程,这就建立了决策规则树。IBLE 方法比 ID3 方法在识别率上提高了 10 个百分点。

### 1.2.1.2 集合论方法

集合论方法是开展较早的方法。近年来,由于粗集理论的发展使集合论方法得到了迅速的发展。这类方法中包括:覆盖正例排斥反例的方法(典型的方法是 AQ 系列方法)、概念树方法和粗糙集(rough set)方法

**1. 覆盖正例排斥反例方法**

它是利用覆盖所有正例,排斥所有反例的思想来寻找规则。比较典型的有 Michalski 的 AQ11 方法,洪家荣改进的 AQ15 方法以及 AE5 方法。

AQ 系列的核心算法是在正例集中任选一个种子,它到反例集中逐个比较,对字段取值构成的选择子相容则舍去,相斥则保留。按此思想循环所有正例种子,将得到正例集的规则(选择子的合取式)。

AE 系列方法是在扩张矩阵中寻找覆盖正例排斥反例的字段值的公共路(规则)。

**2. 概念树方法**

将数据库中记录的属性字段按归类方式进行合并,建立起来的层次结构称为概念树。如对"城市"概念树的最下层是具体市名或县名(如长沙、南京等),它的直接上层是省名(如湖南、江苏等),省名的直接上层是国家行政区(如华南、华东等),再上层是国名(如中国等)。

利用概念树提升的方法可以大大浓缩数据库中的记录。对多个属性字段的概念树提升,将得到高度概括的知识基表,再将它转换成规则。

**3. 粗糙集方法**

在数据库中将行元素看成对象,列元素看成属性(分为条件属性和决策属性)。等价关系 $R$ 定义为不同对象在某个(或几个)属性上取值相同,这些满足等价关系的对象组成的集合称为该等价关系 $R$ 的等价类。条件属性上的等价类 $E$ 与决策属性上的等价类 $Y$ 之间有三种情况。

(1)下近似:$Y$ 包含 $E$;

(2)上近似:$Y$ 和 $E$ 的交非空;

(3)无关:$Y$ 和 $E$ 的交为空。对下近似建立确定性规则,对上近似建立不确定性规则(含可信度),无关情况不存在规则。

## 1.2.2 仿生物技术

仿生物技术典型的方法是神经网络方法和遗传算法。这两类方法已经形成了独立的研究体系。它们在数据挖掘中也发挥了巨大的作用,我们将它们归并为仿生物技术类。

### 1.2.2.1 神经网络方法

它是模拟了人脑神经元结构,以 MP 模型和 Hebb 学习规则为基础的,建立了三大类多种神经网络模型。