数据挖掘综述

王光宏，蒋 平

（同济大学 信息与控制工程系，上海 200092）

摘要：从人工智能、统计分析和数据库技术 3 个方面对数据挖掘技术进行了总结；从模式识别的角度讨论了数据挖掘技术的主要任务，包括分类、聚类、回归、关联、序列和偏差 6 种模式的识别. 详细介绍了数据挖掘技术的常用方法，包括模糊理论、粗糙集理论、云理论、证据理论、人工神经网络、遗传算法以及归纳学习. 列举了当前数据挖掘技术的实际应用场合，并指出其今后的发展趋势以及急需关注的问题.

关键词：数据挖掘；数据库中知识发现；人工智能；模式
中图分类号：TP 311.13　　**文献标识码**：A　　**文章编号**：0253－374X(2004)02－0246－07

Survey of Data Mining

WANG Guang-hong, *JIANG Ping*

(Department of Information and Control Engineering, Tongji University, Shanghai 200092, China)

Abstract: This paper provides a survey for data mining, which combines artificial intelligence, statistical analysis, and database management system attempting to extract knowledge from databases. From the point of view of pattern recognition, different data mining tasks are discussed, including classification, clustering, regression, association, sequential and deviation pattern recognition. Mostly used methods are introduced in detail, including fuzzy method, rough sets theory, cloud theory, evidence theory, artificial neural networks, genetic algorithms and induction learning. Applications, trends and attentions of data mining are also mentioned finally.

Key words: data mining; knowledge discovery in databases; artificial intelligence; pattern

早在 1982 年，趋势大师约翰·奈斯比(John Naisbitt)在他的首部著作《大趋势》(Megatrends)[1]中就提到："人类正被信息淹没，却饥渴于知识."计算机硬件技术的稳定进步为人类提供了大量的数据收集设备和存储介质；数据库技术的成熟和普及已使人类积累的数据量正在以指数方式增长；Internet 技术的出现和发展已将整个世界连接成一个地球村，人们可以穿越时空般地在网上交换信息和协同工作. 在这个信息爆炸的时代，面对着浩瀚无垠的信息海洋，人们呼唤着一个去粗取精、去伪存真的能将浩如烟海的数据转换成知识的技术. 数据挖掘(data mining, DM)就是在这个背景下产生的.

1 数据挖掘的概念

数据挖掘是通过仔细分析大量数据来揭示有意义的新的关系、趋势和模式的过程. 其出现于 20 世纪 80 年代后期，是数据库研究中一个很有应用价值

收稿日期：2003－01－16
基金项目：国家自然科学基金资助项目(60175028)
作者简介：王光宏(1978－)，男，福建霞浦人，博士生. E-mail: ghwangtongji@sina.com.cn

的新领域，是一门交叉性学科，融合了人工智能、数据库技术、模式识别、机器学习、统计学和数据可视化等多个领域的理论和技术. 数据挖掘作为一种技术，它的生命周期正处于沟坎(chasm)阶段，需要时间和精力去研究、开发和逐步成熟，并最终为人们所接受[2,3]. 由于数据挖掘是数据库中知识发现(knowledge discovery in databases，KDD)的核心步骤(如图 1 所示)，发现了隐藏的模式，所以从模式处理的角度，许多人认为两者是等同的[3]. 图 1 中的数据仓库(data warehouse)是整个数据挖掘技术的基础. 20 世纪 80 年代中期，数据仓库之父 W. H. Inmon 在《建立数据仓库》(Building the Data Warehouse)[4]一书中定义了数据仓库的概念，随后又给出了更为精确的定义：数据仓库是在企业管理和决策中面向主题的、集成的、时变的以及非易失的数据集合. 与其他数据库应用不同的是，数据仓库更像一种过程——对分布在企业内部各处的业务数据的整合、加工和分析的过程. 传统的数据库管理系统(database management system，DBMS)的主要任务是联机事务处理(on-line transaction processing，OLTP)；而数据仓库则是在数据分析和决策方面提供服务，这种系统被称为联机分析处理(on-line analytical processing，OLAP). OLAP 的概念最早是由关系数据库之父 E. F. Codd 于 1993 年提出的[5]. 当时，Codd 认为 OLTP 已不能满足终端用户对数据库查询分析的需要，结构化查询语言(structured query language，SQL)对数据库进行的简单查询也不能满足用户分析的需求. 用户的决策分析需要对关系数据库进行大量计算才能得到结果，因此 Codd 提出了多维数据库和多维分析的概念.

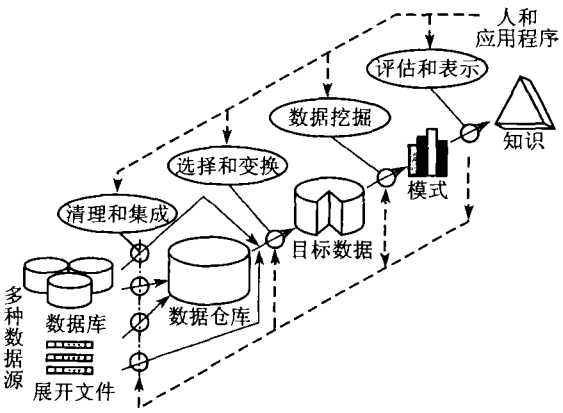

图 1　数据挖掘——数据库中知识发现的核心步骤

Fig. 1　Data mining as a major process of knowledge discovery in databases

数据挖掘产生于应用，且应面向于应用. 数据挖掘的交叉产业标准过程(cross-industry standard process for data mining，CRISP-DM)是当今数据挖掘业界通用流行的标准之一，是 SPSS(Statistical Product and Service Solutions，当时为 Integral Solutions Limited，ISL)、NCR(National Cash Register)和 Daimler Chrysler(当时为 Daimler-Benz)3 家公司在 1996 年制定的，它强调的是数据挖掘在商业中的应用，解决商业中存在的问题，而不是把数据挖掘局限在研究领域. CRISP-DM 参考模型中包括：商业理解、数据理解、数据准备、建立模型、模型评估和模型发布(如图 2 所示)[6].

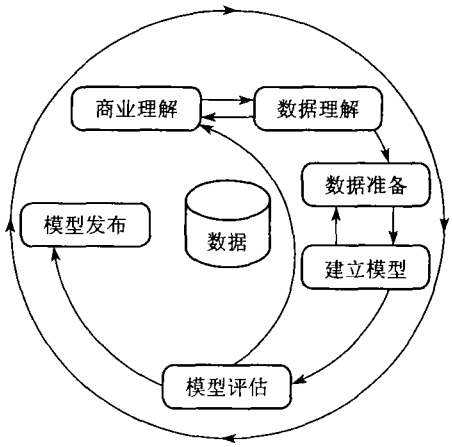

图 2　CRISP-DM 参考模型

Fig. 2　Phases of the CRISP-DM process model

2　主要研究内容

数据挖掘的任务就是发现隐藏在数据中的模式. 其可以发现的模式一般分为两大类：描述型(descriptive)模式和预测型(predictive)模式. 描述型模式是对当前数据中存在的事实做规范描述，刻画当前数据的一般特性；预测型模式则是以时间为关键参数，对于时间序列型数据，根据其历史和当前的值去预测其未来的值. 根据模式特征，可将模式大致细分如下.

2.1　分类模式(Classification)

分类就是构造一个分类函数(分类模型)，把具有某些特征的数据项映射到某个给定的类别上. 该过程由 2 步构成：模型创建和模型使用. 模型创建是指通过对训练数据集的学习来建立分类模型；模型使用是指使用分类模型对测试数据和新的数据进行分类. 其中的训练数据集是带有类标号的，也就是说

在分类之前,要划分的类别是已经确定的.通常分类模型是以分类规则、决策树或数学表达式的形式给出的.

2.2 聚类模式(Clustering)

聚类就是将数据项分组成多个类或簇,类之间的数据差别应尽可能大,类内的数据差别应尽可能小,即为"最小化类间的相似性,最大化类内的相似性"原则.与分类模式不同的是,聚类中要划分的类别是未知的,它是一种不依赖于预先定义的类和带类标号的训练数据集的非监督学习(unsupervised learning),无需背景知识,其中类的数量由系统按照某种性能指标自动确定.

2.3 回归模式(Regression)

回归模式的函数定义与分类模式相似,主要差别在于分类模式采用离散预测值(例如类标号),而回归模式采用连续的预测值.在这种观点下,分类和回归都是预测问题.但在数据挖掘业界,大家普遍认为:用预测法预测类标号为分类,预测连续值(例如使用回归方法)为预测[3].许多问题可以用线性回归解决,对于许多非线性问题可以通过对变量进行变换,从而转换为线性问题来解决.

2.4 关联模式(Association)

关联模式是数据项之间存在的关联规则,是在同一事件中出现的不同项之间的相关性,比如顾客在同一次购买活动中所购买的不同商品之间的相关性.

记项集 $I=\{i_1,i_2,\cdots,i_n\}$,其中 i_n 称为项(item);交易集 $D=\{T_1,T_2,\cdots,T_p\}$,其中 T_p 称为交易(transaction),也是项的集合,并且 $T\subseteq I$.一条关联规则是形如 $X\Rightarrow Y$ 的蕴涵关系式,其中 $X\subset I$,$Y\subset I$,并且 $X\cap Y=\varnothing$.规则 $X\Rightarrow Y$ 在交易集 D 中的支持度(support)是交易集 D 中包含 X 和 Y 的交易数与所有交易数之比,记为 support$(X\Rightarrow Y)$,即

$$\text{support}(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,\ T\in D\}|/|D|$$

规则 $X\Rightarrow Y$ 在交易集 D 中的可信度(confidence)是指交易集 D 中包含 X 和 Y 的交易数与包含 X 的交易数之比,记为 confidence$(X\Rightarrow Y)$,即

$$\text{confidence}(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|\{T:X\subseteq T,T\in D\}|$$

对于给定的一个交易集 D,挖掘关联模式问题就是产生支持度和可信度分别大于用户给定的最小支持度(minsupport)和最小可信度(minconfidence)的关联规则.

最著名的关联规则挖掘算法是由 Agrawal 等于1994年提出的 Apriori 算法[7,8].Apriori 算法的基本思想是:统计多种商品在一次购买中共同出现的频数,然后将出现频数多的搭配转换为关联规则.Apriori 算法的核心是:用前一次扫描数据库的结果产生本次扫描的候选项目集,从而提高搜索的效率.其后人们又提出了诸多关联规则挖掘算法,主要工作集中在如何提高项集的生成效率和降低计算代价上.

2.5 序列模式(Sequential)

序列模式是描述基于时间或其他序列的经常发生的规律或趋势,并对其建模.一个典型的例子就是:在购买 PC 机的顾客当中,70%的人会在半年内购买内存条.序列模式将关联模式和时间序列模式结合起来,重点考虑数据之间在时间维上的关联性.有3个参数的选择对序列模式挖掘的结果影响很大:① 序列的持续时间 t,也就是某个时间序列的有效时间或者是用户选择的一个时间段;② 时间折叠窗口 $w(w\leqslant t)$,在某段 w 时间内发生的事件可以被看作是同时发生的;③ 所发现模式的时间间隔.

2.6 偏差模式(Deviation)

偏差模式是对差异和极端特例的描述,如聚类外的离群值.大部分数据挖掘方法都将这种差异信息视为噪声而丢弃,然而在一些应用中,罕见的数据可能比正常的数据更有用[3].比如信用卡的欺骗检测(fraud detection),通过检测一个给定帐号与其历史上正常的付费相比,可以付款数额特别大这一异常数据为依据来发现信用卡被欺骗性使用.

3 数据挖掘的常用方法

3.1 模糊(Fuzzy)方法

美国控制论专家、数学家查德(Zadeh)于1965年发表的论文《模糊集合》(Fuzzy Sets)[9],标志着模糊数学这门学科的诞生.模糊集合和模糊推理是模糊方法的数学基础,模糊集理论以不确定性的事物为研究对象,是经典集合理论的扩展.隶属度函数是模糊集合的特征函数,是模糊概念的核心,它的取值范围从普通集合{0,1}的两个值扩充到[0,1]闭区间内连续值,隶属度函数的定义如下.

设给定论域 U,U 到[0,1]闭区间的任一映射 μ_A,

$$\mu_A:U\to[0,1],\quad u\to\mu_A(u)\quad u\in U$$

都确定 U 的一个模糊子集 A,μ_A 是 A 的隶属度函

数,$\mu_A(u)$是 u 对 A 的隶属度,表征 u 属于 A 的程度.其实模糊集合并不是一个集合,因为其核心是隶属度函数,如同概率中的随机变量,所以更确切地应该称其为模糊函数.

模糊推理就是由前提(A'),依规则(if A then B)推理,得到结论(B').规则是一种蕴涵关系,推理中将前提和规则结合是一种合成关系.在模糊方法中,关系就是模糊集合,找关系便是找隶属度函数.如何选择蕴涵关系和合成算法是模糊推理的关键,这里有许多经验的成分.模糊推理注重的是把握结论的趋势,是近似的而不是精确的结果.当然,模糊推理的结果也可能是错的,所以还要实践检验.

模糊逻辑系统已用于许多特别是基于规则的分类领域,包括医疗和财经.在基于规则的分类系统中引入模糊逻辑,就可以定义"模糊"阈值或边界,可以避免原系统固有的缺陷:对于连续值,有陡峭的截断[3].从而可能获得一个更合理的分类结果.

3.2 粗糙集(Rough Sets)理论

Rough集理论是由波兰华沙理工大学的 Z. Pawlak 教授于 1982 年提出的一种研究不完整、不确定知识和数据的表达、学习和归纳的理论方法[10],现已成为DM/KDD研究中的最有力工具,也最有发展前途.Rough 集理论采用上近似集合(upper approximations)R^* 和下近似集合(lower approximations)R_*来定义 Rough 集[11],即

$$R^*(X)=\bigcup\{Y\in U/R: Y\cap X\neq\varnothing\}$$

$$R_*(X)=\bigcup\{Y\in U/R: Y\subseteq X\}$$

其中:U 是全域;$X\subseteq U$ 是目标集合;R 是 U 上的等价关系;$Y\in U/R$ 表示 Y 是 U 上按等价关系 R 做成的等价类.上近似集合可理解为所有那些与 X 有交的等价类的并集;下近似集合可理解为所有那些被包含在 X 里面的等价类的并集.

上近似集合和下近似集合之间的差称为 X 的 R 边界线集(boundary region),表示为:$BN_R=R^*(X)-R_*(X)$.它是那些通过等价关系 R 既不能在 X 上分类,也不能在 $\tilde{X}$ 上分类的元素的集合.

从语义角度来看,$R_*(X)\subseteq X\subseteq R^*(X)$,介于 $R_*(X)$和 $R^*(X)$之间的集合簇均属 Rough 集,而上下近似集合则是该空间的两个极限.在 Rough 逻辑推理中,不必涉及隶属度函数,这样的软计算技术有助于知识获取瓶颈的突破.Rough 集理论的核心特点是无需提供问题所需处理的数据集合以外的任何先验信息,这也可能是因为其无法获得客观事实的足够支持.从这点看,Rough 集理论是对 Fuzzy 集理论的发展和重大拓展.

Rough 集理论可以用于分类,发现不准确数据或噪声数据内在的联系.找出可以描述给定数据集中所有概念的最小属性子集是个 NP-难问题.在给定的现实世界数据中,往往有些类不能被可用的属性区分,那么就可以用 Rough 集来近似地定义这些类.

3.3 云(Cloud)理论

云理论是李德毅教授于 1995 年提出的用于处理不确定性的一种新理论[12].作为处理模糊性问题的主要工具,Fuzzy 集理论提出了隶属度函数来刻画模糊事物的亦此亦彼性;然而,一旦用精确的隶属度函数来描述模糊集,之后的模糊推理等环节就不再有模糊性了.这就是传统模糊集理论的不彻底性.针对这一问题,李德毅教授在传统模糊集理论和概率统计的基础上提出了定性定量不确定性转换模型——云模型,把定性概念的模糊性和随机性完全集成到一起,构成定性和定量相互间的映射,作为知识表示的基础[13].

云是用语言值描述的某个定性概念与其数值表示之间的不确定性转换模型.设论域 $U=\{x_1,x_2,\cdots,x_n\}$,T 是与 U 相联系的语言值.U 中的元素 x 对于 T 所表达的定性概念的隶属度 $C_T(x)$(或称 x 对于 T 的相容度)是一个具有稳定倾向的在[0,1]中取值的随机数,隶属度在论域上的分布称为隶属云,简称为云.对于任意的 $x\in U$ 到区间[0,1]的映射是一对多的转换,x 对于 T 的隶属度是一个概率分布而非固定值,从而产生了云,而不是 Fuzzy 集理论中的一条明晰的隶属度曲线.云由云滴组成,每个云滴是定性概念在定量上的一次实现,单个云滴可能无足轻重,但云的整体形状反映了定性概念的基本特征.如果从 Fuzzy 集理论的观点来看,云的数学期望曲线是 Fuzzy 集理论中的隶属度曲线.

图 3 显示了语言值"约 60 kg"的隶属云,以及隶属云的 3 个数字特征值.

期望值 E_x(expected value):是概念在论域中的中心值,是最能代表这个定性概念的值,也就是说,它 100% 隶属于这个定性概念.

熵 E_n(entropy):是定性概念模糊度的度量,反映了在论域中可被这个概念所接受的数值范围,体现了定性概念亦此亦彼性的裕度.熵越大,概念所能接受的数值范围也越大,概念就越模糊.

超熵 H_e(hyper entropy):是熵的熵,反映了云滴的离散程度.超熵越大,云滴离散度越大,隶属度

的随机性越大,云的"厚度"也越大.云的"厚度"是不均匀的,腰部最为分散,顶部和底部汇聚性好.这表明,靠近概念中心或远离概念中心处隶属度的随机性较小,而离概念中心不近不远处隶属度的随机性较大.这与人的主观感受一致.

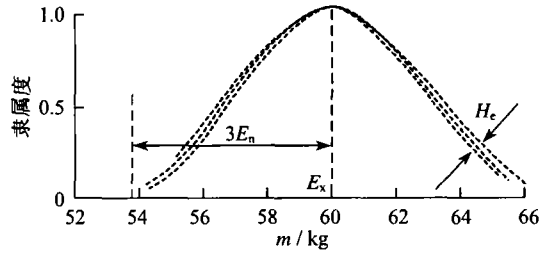

图3 "约60 kg"的隶属云的数字特征示意图

Fig.3 Cloud of "about 60 kg"

在数据挖掘中,云理论常和Rough集理论相结合.前者研究的是数据的模糊性和随机性,为定量定性间的不确定性转换提供模型;而后者强调的是数据的不完备性和不可分辨性,但其处理方法是确定性的,要求属性值都是定性值.所以两者具有某种互补性.

3.4 证据理论(Evidence Theory)

证据理论又称Dempster-Shafer理论,是经典概率论的扩充[14].首先由Dempster在20世纪60年代提出,在70年代中期由Shafer进一步发展,形成处理不确定信息的证据理论.该理论的一个重要贡献就是划清了不确定和不知的界限[13].

在证据理论中,一个样本空间称为一个识别框架,用Ω表示.Ω由一系列对象构成,对象之间两两互斥,且包含当前要识别的全体对象.Ω的所有子集的集合记为2^{Ω}.若Ω中有n个对象,则有2^n个子集.每个子集对应一个命题(证据或结论).证据理论的基本问题是:已知识别框架Ω,判明Ω中一个先验的未定位对象属于Ω的某个子集A的程度.

Anand等于1994年提出了一个基于证据理论的通用数据挖掘框架EDM(database mining based on evidence theory)[15],在EDM框架下DM/KDD的技巧主要是根据不同的任务选择和开发不同的算子.EDM已经用于从关系数据库发现强规则,以及从大量的金星表面图像数据中识别火山.同时,EDM的算法本质上是并行的,在处理并行、分布及异构数据库时有显著的优越性.由于证据理论在处理不确定性方面的优点,加之EDM的开发和应用,使得基于证据理论的方法在数据挖掘中具有潜在的应用性.

3.5 人工神经网络(artificial neural network,ANN)

人工神经网络由多个神经元按照某种方式相互连接形成,靠网络状态对外部输入信息的动态响应来处理信息,网络的信息分布式存储于连接权系数中,使网络具有很强的容错性和鲁棒性.神经网络的核心是结构和算法,例如以结构见长的Hopfield网和以算法见长的BP(back propagation)网[16].

同模糊逻辑系统相比,模糊逻辑系统是从宏观功能上"软"模拟人脑的逻辑思维机制,而神经网络是从微观结构上"硬"模拟人脑的经验思维机制;模糊逻辑系统的智能级别为推理级,而神经网络的智能级别为感知级.

在数据挖掘中,神经网络主要用于获取分类模式.但是由于神经网络分类方法获取的模式隐含在网络结构中,而不是显示地表达为规则,不容易被人们理解和解释;另外要多次扫描训练数据,网络的训练时间较长.因此与其他数据挖掘方法不同,神经网络用于数据挖掘,要解决好两个关键问题:一是降低训练时间,二是挖掘结果的可理解性.

3.6 遗传算法(genetic algorithms,GA)

遗传算法最先由John Holland于1975年提出[17].其模拟生物的进化和遗传,借助选择(selection)、交叉(crossover)和变异(mutation)操作,使要解决的问题从初始解逐步逼近最优解,解决了许多全局优化问题.可以说"优胜劣汰"原则和种群"多样性"是GA的灵魂."选择"保证了前者,"交叉"和"变异"保证了后者.但"选择"容易产生早熟个体,使"多样性"过早丧失,而"交叉"和"变异"的随机性太大,带有盲目性.

GA通过编码将优化等问题从问题空间映射到GA的操作空间,再通过译码将操作结果从操作空间映射回问题空间.对于具体的问题,常常有限制条件,即存在一个可行解空间.为了保证最后的解是可行的,可以采取两种方法:一是将可行解空间与GA的操作空间一一对应;二是将可行解空间包含于问题空间中.

GA是依据随机技术来保证其寻优方向的确定算法,从"最优个体在运动过程中越来越多"可推断出:GA只能保证全局寻优的趋势.已有理论证明,不改造的GA不能达到全局最优,只能寻到全局最优的邻域.在达到该邻域后可采用局部寻优(如梯度法)来最终达到全局最优点.目前关于GA的最好理论结果是:经过改进的GA能够依概率达到全局最优.由于小概率事件有可能发生,所以算法有可能不

收敛.这与理想的期望——依概率1收敛到全局最优——有一定的距离.

遗传算法易于并行,已广泛用于分类和优化问题.在数据挖掘中,还可用于评估其他挖掘算法的适合度[3].

3.7 归纳学习(Induction Learning)

归纳学习是从大量的经验数据中归纳抽取出一般的规则和模式,是一种重要的数据挖掘方法.归纳学习的算法大部分来自于机器学习(machine learning)领域,其中最著名的是C4.5(Quinlan,1993).C4.5是一种决策树算法,由ID3算法发展而来.它们采用熵来选择属性,分类速度快,适合于大数据库的学习.而C4.5在ID3的基础上增加了将决策树转换为等价的产生式规则的功能,并解决了连续取值的数据的学习问题.

针对数据概化(data generalization),韩家炜教授等提出了面向属性的归纳(attribute oriented induction,AOI)[18].AOI的基本思想是:考察与任务相关的数据中每个属性的不同值的个数,通过概念树(concept tree)提升对数据进行概化,归纳出高层次的模式[3].对于已准备好的数据,AOI的基本操作是数据概化,在初始工作关系上进行属性删除或属性概化.AOI需要背景知识,常以概念树的形式给出.

4 数据挖掘的实际应用

数据挖掘技术最初就是面向应用的,尤其是在银行、电信、保险、交通、零售(如超级市场)等商业领域.数据挖掘所能解决的典型商业问题有:客户关系管理(customer relation management,CRM)、数据库营销(database marketing)、客户群体划分(customer segmentation & classification)、交叉销售(cross-selling)等市场分析行为,以及客户流失性分析(churn analysis)、客户信用记分(credit scoring)及欺诈发现等等.加拿大Simon Fraser大学KDD研究组根据加拿大BC省电话公司拥有的十多年的客户数据,总结、分析并提出新的电话收费和管理办法,制定既有利于公司又有利于客户的优惠政策;美国著名的国家蓝球队NBA的教练,利用IBM公司提供的数据挖掘工具Advanced Scout临场决定替换队员;美国Firstar银行使用Marksman数据挖掘工具,根据客户的消费模式预测何时为客户提供何种产品;还有不少DM/KDD产品用来筛选Internet上的新闻,保护用户不受无聊电子邮件和商业推销的干扰.另外数据挖掘在生物学以及工业领域也有很广泛的应用.

5 数据挖掘的发展趋势

美国已经开始研发一套名为"信息全面感知"(total information awareness,TIA)的反恐怖主义信息监控系统,主要运用数据挖掘技术,搜集全球各地计算机使用者传递的信息,综合情报单位搜集能力,筛检可疑的线索与实证,及时发出预警信息.TIA系统将能提供诸如特定地区的旅行记录、可疑电子邮件来往、不寻常的资金转移、罕见的医疗行为(如炭疽热治疗)等信息.专家指出,这套系统极为庞大复杂,完成后将是人类有史以来最大规模的信息监控系统.

WWW为数据挖掘提供了丰富的数据资源,同时也是一个艰巨的挑战.Web Mining是一项复杂的技术,由于Web数据挖掘比单个数据仓库的挖掘要复杂得多,因而面向Web的数据挖掘成了一个难以解决的问题.而XML(eXtensible Markup Language)的出现为解决Web数据挖掘的难题带来了机会.由于XML能够使不同来源的结构化的数据很容易地结合在一起,因而使搜索多样的不兼容的数据库成为可能.另外还有空间数据挖掘(Spatial DM),多媒体数据挖掘(Multi-Media DM)(包括Text DM,Video DM和Audio DM),DNA数据挖掘和生物信息学(DNA DM and Bio-Informatics)等.

数据挖掘的理论基础和挖掘算法还有很大的空间有待发展和完善.

数据挖掘的个人稳私和信息安全问题也是急需关注和解决的.

6 结语

不管是研究领域还是商业应用,数据挖掘都是一个热门话题,正得到人们越来越多的关注,而且数据挖掘技术也正在逐渐地成熟.要想真正做好数据挖掘,应该从3个方面综合考虑:用数据挖掘来解决的商业问题的类型;为进行数据挖掘所做的数据准备,数据挖掘的各种算法和理论基础.

参考文献:

[1] Naisbitt J. Megatrends: Ten new directions transforming our lives

[M]. New York: Warner Books, 1982. 16 - 17.
[2] Agrawal R. Data mining: Crossing the chasm[R/OL]. http://www.almaden.ibm.com/cs/people/ragrawal/papers/kdd99_chasm.ppt, 2002 - 11 - 20.
[3] Han J W, Micheline K. 数据挖掘概念与技术[M]. 范 明, 孟晓峰译. 北京: 机械工业出版社, 2001.
[4] Inmon W H. Building the data warehouse[M]. 3rd ed. New York: John Wiley & Sons Inc, 2002. 31 - 145.
[5] Codd E F, Codd S B, Salley C T. Beyond decision support[J]. Computer World, 1993, 27(30): 87 - 89.
[6] Shearer C. The CRISP-DM model: The new blueprint for data mining[J]. Journal of Data Warehousing, 2000, 5(4): 13 - 22.
[7] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. Proceedings of the 20th International Conference on Very Large Databases[C]. Santiago: Morgan Kaufmann, 1994. 487 - 499.
[8] Agrawal R, Shafer J C. Parallel mining of association rules[J]. IEEE Transaction on Knowledge and Data Engineering, 1996, 8(6): 962 - 969.
[9] Zadeh L. Fuzzy sets[J]. IEEE Information and Control, 1965, 8(3): 338 - 353.
[10] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341 - 356.
[11] Zhai L Y, Khoo L P, Fok S C. Feature extraction using rough set theory and genetic algorithms—an application for the simplification of product quality evaluation[J]. Computers and Industrial Engineering, 2002, 43: 661 - 676.
[12] 李德毅, 史雪梅, 孟海军. 隶属云和隶属云发生器[J]. 计算机研究和发展, 1995, 32(6): 15 - 20.
[13] 邸凯昌. 空间数据发掘与知识发现[M]. 武汉: 武汉大学出版社, 2000.
[14] Yager R R, Kacprzyk J, Pedrizzi M. Advances in the dempster shafer theory of evidence[M]. New York: John Wiley & Sons Inc, 1994. 5 - 34.
[15] Anand S S, Bell D A, Hughes J G. A general framework for database mining based on evidence theory[R]. Antrim: University of Ulster, 1994.
[16] 孙增圻, 张再兴, 邓志东. 智能控制理论与技术[M]. 北京: 清华大学出版社, 1997.
[17] Holland J H. Adaptation in natural and artificial systems[M]. Ann Arbor: University of Michigan Press, 1975.
[18] Cai Y, Cercone N, Han J W. Attribute-oriented induction in relational databases[A]. Shapiro G P, Frawley W J. Knowledge Discovery in Databases[C]. Cambridge: AAAI/MIT Press, 1991. 213 - 228.

·下期文章摘要预报·

磁浮列车运行控制系统二维速度防护曲线仿真

江 亚, 吴汶麒, 刘 进

速度防护是轨道交通列车自动控制系统中最主要的控制策略. 结合高速磁浮列车的运行特点, 提出了二维速度防护曲线的概念和原理, 对高速磁浮列车在强制制动条件和惰性条件下的速度曲线进行了计算, 给出了高速磁浮列车二维速度防护曲线的一种实用算法, 并在C语言环境下对其进行了数值仿真.

机器 - 基础动力耦合系统功率流传递主动控制

孙玉国

针对柔性基础上马达振动的主动隔离问题, 建立了机器 - 基础动力耦合系统输出反馈控制的状态空间方程, 对系统的动力耦合特性进行了分析, 从功率流传递的观点对主动控制策略进行了评估. 数值计算表明, 速度反馈可有效地降低耦合系统共振区域的振动能量传递; 受增益系数选择范围的限制, 位移反馈不利于低频扰动的隔离; 反馈信号的拾取位置对耦合系统稳定性有重要影响.