基于关联规则的数据库知识发现及应用

■宋　丽　牡丹江师范学院　林　利　牡丹江医学院

[摘　要]　随着数据在日常决策中的重要性越来越显著，人们对数据处理技术的要求也不断提高，需要对数据进行更深层次的处理，以得到关于数据的总体特征以及对发展趋势的预测。本文介绍了数据库知识发现及关联规则,最后将二者结合应用于教学中，进而实现客观地、科学地教学评估与知识发现，指导学校的教学工作。

[关键词]　知识发现　数据库知识发现　关联规则

一、知识发现

随着数据在日常决策中的重要性越来越显著，人们对数据处理技术的要求也不断提高，需要能够对数据进行更深层次的处理，以得到关于数据的总体特征以及对发展趋势的预测。过去，人们依靠经验、大量的计算和人脑的智慧来处理这些深层次的信息，为决策提供技术支持。然而数据量爆炸性的增长使得传统的手工处理方法逐渐变得不切实际了,现在的用户很难再像从前那样，自己根据数据的分布找出规律，并根据此规律进行分析决策。而且对于超市商品的销售记录、保险公司的客户记录、医学上的成千上万份病历等等的这些天体数据来说，如果由手工处理的话需要几十个人几年时间,而且由于数据的繁杂，在由人工对数据进行处理过程中，很难找出关于数据较为全面的信息,这样许多有用的信息仍然隐含在数据中而不能被发现和利用，造成数据资源的浪费，更无法体现出信息的时间效应。由此便迫切需要采用自动化程度高、效率好的数据处理方法来帮助人们更高效地进行数据分析,自动发现数据中隐藏的规律或模式,为决策提供支持。知识发现(Knowledge Discovery in Databases,简称KDD)就是为迎合这种要求而产生并迅速发展起来的一门技术,它是用于开发信息资源的一种新的数据处理技术。

许多专家都给出了知识发现的定义，最新的、在KDD领域一致认可的描述性定义是Fayyad等人给出的：KDD是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非平凡过程。

知识发现过程可粗略的理解为三部曲:数据准备(data preparation)、数据挖掘,以及结果的解释评估(interpretation and evaluation)(如图1所示)。

数据准备阶段的工作包括3个方面的内容:

1.数据选取，主要是确定目标数据——根据用户需要从原始数据库中抽取一组感兴趣的数据,并将其组织成适合挖掘的数据组织形式。

2.数据预处理，也叫数据清洗，主要包括如下工作要做:消除噪音数据(这里提及的噪音数据是指那些明显不符合逻辑的偏差数据，如某职员200岁，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音数据的方法是数据平滑(Data smoothing)技术、推导计算缺值数据、消除重复记录、完成数据类型转换等。

3.数据变换,主要是指对数据进行降维处理。数据挖掘阶段是根据挖掘的任务或目的使用具体的挖掘算法对准备好的数据集进行知识发现。这些知识是隐含的、先前未知的、对决策有潜在价值的，提取的知识表示为概念(Concepts)、规则(Rules)、规律(Regularities)和模式(Patterns)等形式。这些规则蕴含了数据库中一组对象之间的特定关系，揭示出一些有用的信息，为经营决策、市场策划和金融预测等提供依据。例如，从超级商场的大量交易数据中发现，顾客购买牛奶时通常也同时会购买面包,如果将这两种食品放在同一货架上或同时进行广告宣传,肯定会大大提高销售量。通过数据挖掘技术，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，并从不同角度显示，从而使大型数据库作为一个丰富可靠的资源为知识归纳服务。

最后一阶段是对于挖掘出来的模式进行解释和评价,剔除冗余或无关的模式，将结果展现给用户。

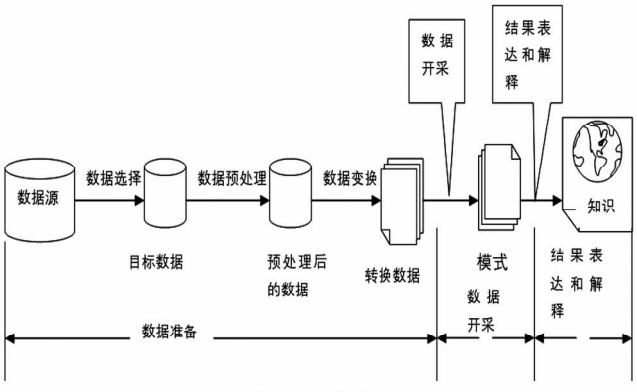

图1　KDD过程图

二、关联规则

关联规则也称为关联模式，是美国BIM Almaden Research Center的R.Agrawal等人于1993年提出的,是数据挖掘研究中的一个重要课题。关联规则是指大量数据中项集之间有趣的关联或相关联系。关联规则发现的对象主要是事务数据库,例如售货数据,也称为货篮数据。它是描述一个事务中物品之间同时出现的规律的知识模式。用D表示全体事务的集合。用I代表D中所有数据项(物品)的集合。假设有一个物品集A，一个事务T，如果A⊂T，则称事务T支持物品集A。关联规则是一种蕴含关系:A=>B，其中A，B是两组物品，A⊂I，B⊂I，且A∩B=ϕ。衡量规则优劣的指标有二:

1.支持度(Support)。它是对A=>B的重要性(或适用范围)的衡量，集合D中规则A=>B的支持度定义为物品集A，B同时出现的概率。支持度描述了A和B这两个物品集的并集在所有的事务D中出现的概率有多大。如果某天有1000个顾客到商场购买物品，其中有100个顾客同时购买了牛奶和面包，那么牛奶=>面包的支持度就是10%。

2.可信度(Confidence)。它是对关联规则的准确度的衡量，集合D中规则A=>B的可信度定义为在物品集A出现的前提下，B出现的概率。如上面所举的牛奶和面包的例子，该关联规则的可信度就回答了这样一个问题:如果一个顾客购买了牛奶，那么他同时也购买面包的可能性有多大呢?在上述的例子中，如果购买牛奶的顾客中有70%的人购买了面包，则该规则的可信度是70%。

关联规则的挖掘问题就是在事务数据库D中找出具有用户给定的最小支持度和最小可信度的关联规则。挖掘关联规则是指在数据库中挖掘出具有这种形式的规则:由于某些事件的发生而引起另外一些事件的发生。它在决策支持系统、专家系统和智能信息系统等各个方面起着重要的作用。并且，随着数据库应用的普及，数据挖掘的应用越来越广，包括零售商的货篮分析、销售分析、金融信贷风险分析、医学诊断和物流货源分析等其他领域。由于挖掘出的关联规则既可以检验行业内长期形成的知识模式，也能够发现隐藏的新规律，在近几年内这方面的研究就倍受人们的关注。

近年来，对关联规则的挖掘的研究主要集中在以下几个方面:(1)对由R.Agarwa等人提出的Apriori算法的改进，这方面的工作主要集中如何有效的生成最大项目集以及改善该算法的效率上面;(2)对于关联规则阀值的研究，这个方面的工作主要集中在如何调整阀值使得挖掘出来的规则具有更大的关联性与有用性以及更加符合人们的要求;(3)提出关联规则发现的并行算法;(4)扩展关联规则发现问题，如广义多层关联规则、定量关联规则、循环关联规则和具有利润约束关联规则等等。

三、基于关联规则的数据库知识发现应用

在教育教学中,学校教学主管部门需要对教师的教学情况进行评价,如何客观公正地评价教师的教学情况是摆在教育管理部门课题,如何在此基础上合理地安排师资和学时,全面提高学生的知识水平尤为重要。随着计算机技术和网络技术的发展,许多学校都为某些学科建立网上考试系统,由于计算机的高效处理和海量存储能力以及数据挖掘技术发展,使我们能够利用计算机排除人为因素,客观地进行教学评价与知识发现。

在教学评估系统中，首先建立星型结构（如图2所示），建立了六维数据库,当然也可以根据需要建立更多维的数据库。大多数据情况下,与数据挖掘任务有关的数据是存储在应用数据库中,这些数据往往是为应用目的而建立的,不能直接运行挖掘算法，而是要进行必要的抽取和格式的整理工作,对字符型的属性都要进行属性展开，需要对数据进行清理和约简,建立适合挖掘的关联数据。

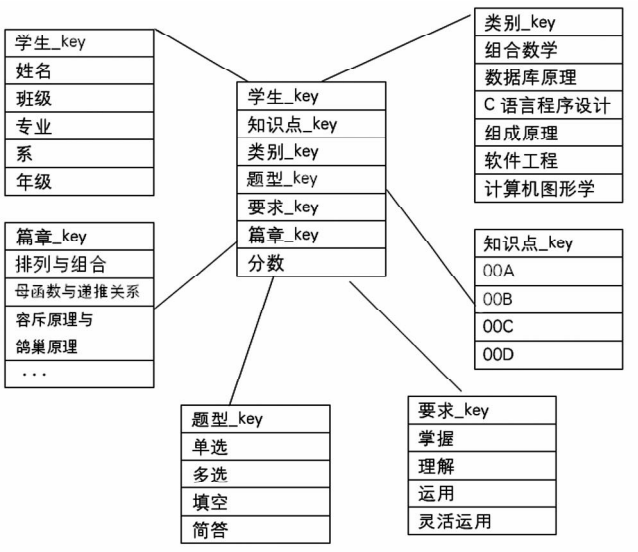

图2 六维数据库星型结构图

例如，在考试系统的学生答题库中，记录了学生对本学科各个知识点的掌握情况,利用关联规则中的适当算法可以求得各知识点之间的关联，提取某些新颖的关联为教学服务。

下面取得关联规则的一部分来举例说明知识点的关联关系。

关联关系	可信度	支持度
A=>B	48.09%	40.2%
B=>C	68.71%	37.62%
C=>D	24.86%	30.97%

若规定支持度大于20%，可信度大于40%,就可以得到,如果知识点A掌握的好,那么知识点B、C就掌握的好，这样就为教学管理部门提供客观依据加强知识点A的教学工作。进而实现客观地、科学地教学评估与知识发现，指导学校的教学工作。

四、结束语

数据挖掘或数据库知识发现，受到了当今国际人工智能与数据库界的广泛重视。关联规则是数据挖掘研究中的一个重要研究课题。在该方面的研究起步虽晚，其发展速度却非常惊人，其研究硕果也是层出不穷。但目前的关联规则挖掘技术也存在着明显的不足：对小数据集适用性较强，但对于海量数据而言却显现出明显的缺陷。在这个信息时代，数据量爆炸性地增长，关联信息每天都迭迭涌现、悄悄溜走，为了充分利用数据资源，研究适合于从大数据集中进行关联规则挖掘的新算法有待于进一步探索。

参考文献:

[1]史忠植著:知识发现.北京:清华大学出版社，2002

[2]杨炳儒:知识工程和知识发现.机械工程出版社，2003

[3]白石磊 毛雪岷 王儒敬等:基于数据库和知识库的知识发现研究综述[J].广西师范大学学报:自然科学版,2003(1):136～138

[4]杨 武 陈 庄:数据库知识发现技术及应用[J].重庆工学院学报:自然科学版，2001,15(2):32～34

[5]李雄飞 苑森淼 董立岩:基于相联规则的数据挖掘理论.吉林工业大学学报(自然科学版),2000,30(2):43～46