一个 KDD 应用系统的设计与实现

俞机运　黄上腾　（上海交通大学计算机系 200030）

摘要：本文介绍了一个基于高校科研管理的 KDD 应用系统的结构、功能和实现技术。

一、引言

如何使数据库中大量的数据真正发挥作用，使数据库能够为我们提供分析、决策的功能，使数据库的应用上升到智能化的高度，是当前面临的问题。在某高校科研管理信息系统的开发过程中，我们已经建起了一个包含项目、成果、经费、人员等在内的科研管理数据库。进一步的要求是对该数据库中大量的原始数据进行知识的抽取，为该校在科研课题的规划、导向和争取，科研资源的合理利用，科研体制与机制等方面提供宏观的决策信息。

数据库中的知识发现（Knowledge Discovery in Databases，简称 KDD），是 90 年代计算机科学中一个引人注目的新领域。短短几年中，KDD 的研究不仅在发现的知识种类（模式）、发现算法、系统模型等方面取得了长足的进展，而且开发了一些实验系统和实际应用系统。我们在吸收现有 KDD 技术的基础上，结合高校科研管理的特定应用背景，实现了一个具有多种知识发现能力的交互式集成 KDD 应用系统——科研管理知识获取系统（Knowledge Explore for Management of Scientific Research，简称 KER）。

二、系统结构

KER 并不是完全自动地从数据库存储的所有数据中发现可能存在的知识，这是不大可能也是不必要的。系统是在发现任务（用户感兴趣的问题）的驱动下，首先得到知识基表（从原始数据库中经数据汇集处理后，得到与发现任务相关的所有数据组成的二维表），其次是在与用户的交互过程中，对知识基表进行宏观和微观两个方向的操作，最终抽取出能在一定程度上满足用户兴趣的知识。KER 的系统结构框图如图 1 所示。

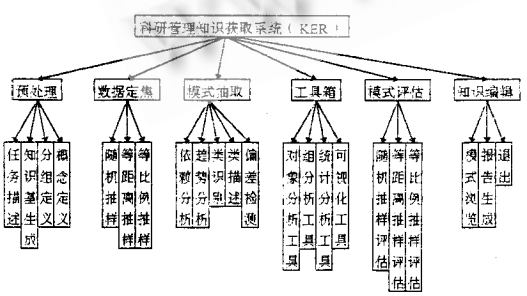

图 1　科研管理知识获取系统（KER）结构框图

三、系统功能描述

KER 是一个具有多种知识发现能力的交互式集成 KDD 应用系统。第一，基于高校科研管理的特定应用背景，系统提供了五类模式的抽取：依赖关系分析；趋势分析；类识别；类描述；偏差检测。第二，系统是交互的。KER 是实现十分强调发现过程的交互性，要求用户能够参与知识发现的全过程，使用户能够随时根据计算机所得到的中间结果，对发现过程加以监控和引导，从而使计算机的发现过程综合到用户自己的决策过程中去，直至得到满意的结论为止。第三，系统是集成的。横向上实现了用户输入（包括发现任务及领域知识的输入）、数据定焦、模式抽取、模式评估、知识编辑的集成；纵向上通过 ODBC，系统提供了与多种数据库的集成。下面我们对 KER 的各模块功能作一简要描述。

1. 预处理

预处理模块实现用户发现任务及特定领域知识的输入，并生成与发现任务相关的知识基表。

任务描述—KER 的发现任务描述界面提供了任务编号、任务说明、主题词、约束条件等。其中，主题词体现了发现任务所描述的具体内容。主题词可以是数据库中关系表的属性域，也可以是自己定义的专业术语。例如，对于发现任务“90 年以来学校在计算机方面的科研力量发展如何？”熟悉科研管理的用户挑选的主题词可能是：研究人员的学历、年龄、在项目中的分工；参与的科研项

目的来源、学科类别、课题性质、经费总额等等。约束条件是:参与项目的立项年份〉= 1990;研究人员的主要从事专业是计算机,并且参与过项目研究。在 KER 中,所有主题词与约束条件的输入均可以通过鼠标点取下拉菜单实现,非常直观。

知识基生成—知识基是与发现任务相关的一组数据。对于用户指定的每一个发现任务,系统首先将其转化成标准的 SQL 查询语句(其中发现任务的主题词对应查询语句的 SELECT 子句,约束条件对应 WHERE 子句)。通过 ODBC,系统查询源数据库,得到相应的初始知识基表。

分组定义—对某些属性值的一次性分组,将有助于知识的抽取。例如,主管项目经费的同志知道,本校科研项目的经费大部分集中在 30 万到 50 万。因此,该校的科研项目按经费可以分为三组:30 万以下;30 万～50 万;50 万以上。分组定义模块提供了这种领域知识的输入。

概念定义—概念实际上是一组类描述规则。类描述是 KER 的一个重要功能。例如,用户可能关心计算机系的科研力量,也可能关心电子信息学院的科研力量,KER 提供了这种对不同层次的概念的描述。为此,用户应该提供一个必要的背景概念体系。在概念定义模块中,用户是通过构造概念树的方法来实现这种领域知识的输入。图 2 就是用户定义的一棵概念树。

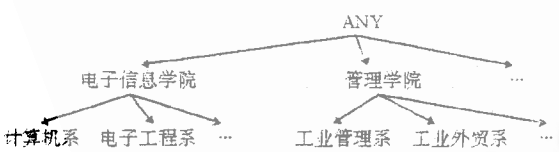

图 2　系(院)概念树

2.数据定焦

数据定焦发生在两种情况下:或者是与发现任务相关的初始知识基过于庞大,不能一下子调入内存;或者是用户为了便于模式抽取过程的执行,强行限制每次进行模式抽取所处理的数据记录。数据定焦是通过对初始知识基的统计抽样实现的。在 KER 中,我们提供了三种抽样方法:随机抽样、等距离抽样、等比例抽样。经过数据定焦后得到的知识基是真正要在其上进行模式抽取的数据。

3.模式抽取

从关系数据库发现的知识通常又被称为模式(PATTERN),术语模式是指数据库中元素(如记录、属性)之间的某种关系。模式抽取是系统的一个核心模块,各种面向属性的归纳算法、面向元组的归纳算法以及统计分析的算法嵌入在该模块中。基于高校科研管理的特定应用背景,KER 提供了五类模式的抽取:

依赖分析—数据依赖是数据库中存在的一类重要的可被发现的知识。在 KER 中,我们要发现的是隐含在数据库中数据之间的概率依赖关系。系统可以对用户指定的每一个"属性对"进行依赖分析,并计算出它们之间依赖的"可信度",如图 3 所示。KER 的这种依赖分析有时对用户有着直接的意义。例如,可以发现某些属性之间以前不知道的强依赖关系;当某一属性的值发生了变化,可以通过依赖图搜寻引起这种变化的原因。

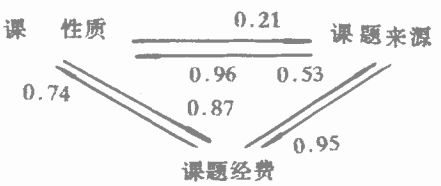

图 3　KER 发现的一个简单依赖图

趋势分析—对用户指定的一个或几个关键词,系统能够将相关数据的发展结果以可视化的形式呈现给用户,并预测今后一段时期的变化,使用户对这些数据的发展趋势有一个感性的认识。如通过趋势分析,我们发现:90 年以来从事计算机的研究人员的学历呈上升趋势,而研究人员的年龄则呈下降趋势。

类识别—数据库中的大量记录可以被分割成一些有特定意义的子类,对这些子类的识别可能对用户的决策也有着重要意义。定义分组、定义概念都是进行类识别的方法,但这些子类是用户已知的。通过对指定知识基表的特定属性或属性之间进行聚类、回归等统计分析,系统可以发现数据库中隐含的子类。

类描述—对用户指定的某一层次的概念类,系统依据用户定义的概念树,对知识基表的各个属性进行提升、归纳,最终得到描述该概念类的本质特征或分类特征的宏元组。这是一种面向属性的归纳方法。在类描述模块中,KER 还提供了一种面向元组的归纳方法。该方法对

特定知识基表的所有元组进行一次检索，通过分析面向元组的依赖性来发现该知识基表中所有元组都满足的特征规则。

偏差检测—数据的偏差往往代表了一类潜在的有意义的模式：某一时期的数据呈现出与其他时期有很大的出入；预测值与实际值的误差等等。在 KER 中，我们提供了两种偏差的检测：发现异常与发现变化。对用户指定的类别和关键词，系统可以在相关的知识的基表中检测出异常的记录；随着数据库中大量数据的不断积累，系统可以检测出已抽取的某些模式的变化，这些结果可能揭示出学校科研力量的变化。

4.工具箱

工具箱是系统进行模式抽取的辅助模块，主要有：

对象分析工具—所谓对象分析，就是指以用户指定的属性域中各个不同的取值为基本对象，将每个对象作为基本单位来分析。它既可以分析出总体中的每一个对象所占的比重，又可以比较多个对象内部的相同和不同。

组分析工具—它以用户指定的或系统内定的属性为中心属性，将每一个属性域与中心属性域构成一张子表，分别分析当前知识基表中其他属性对中心属性的关系。组相关分析使得用户能够从各个角度来分析中心属性的影响，有利于简化多个属性的域的复杂关系。

统计分析工具—实现一些基本的统计计算，包括求最大、求最小、求平均、多元回归、主成分分析、因子分析等工具。

可视化工具—该工具可以对用户指定的表格自动生成二维或三维的统计图。当表格的数据发生变化时，相关图形会自动发生相应的变化。

5.模式评估

如果模式抽取是在数据定焦(抽样)后的知识基上进行，为了验证模式的准确度，对抽取出来的模式应该回到初始知识基中重新抽样评估。评估模块的功能就是给抽取出的模式定一个相对数值(可信度)，然后根据这些值决定哪些模式以怎样的顺序排列提供给用户。与数据定焦相对应，模式评估模块也提供了三种抽样评估：随机抽样评估；等距离抽样评估；等比例抽样评估。

6.知识编辑

该模块向用户提供了两种对发现的知识的编辑环境：表格形式的模式浏览和文本方式的报告生成。

模式浏览－－将抽取出来的模式以二维表的形式提供给用户，其中每一个元组又称为宏元组。用户根据自己的领域知识或应用需要对表格进行浏览编辑。

报告生成－－对当前表格形式的发现结果自动生成相应的文字报告，同时也向用户提供了一个复合文件编辑环境，使用户能够自己组织出包含文字、图形和表格的知识发现报告。

退出－－结束整个系统的运行。

四、系统实现技术

数据库中的知识发现是一个涉及数据库、人工智能、统计学等多门学科的交叉研究领域，在发现系统的实现中对数据库管理、人机界面和交互形式、统计工具和发现工具、表格处理、图形生成以及文字表格图形的复合等均有较高的要求，每一方面的要求都可以找到一个相当强的软件来支持，但没有一个软件可以满足所有这些方面的要求。因此，软件集成是我们实现 KER 的关键途径。KER 是 WINDOWS 环境下的一个应用，主要开发工具是 POWERSOFT 的 POWERBUILDER 和 MICROSOFT 的 VISUAL C＋＋。利用 WINDOWS 提供的 DDE 和 OLE 技术，将 POWERBUILDER、VISUAL C＋＋、EXCEL 和 WORD 集成为一个可以同时提供完整的数据库管理、强有力的计算功能、优良的制表工具、灵活多样的绘图工具以及 C＋＋语言的 DLL 之间的自由调用。

五、结束语

KDD 是一门面向信息社会的新技术，是人们对现有的信息处理技术不满足的必然结果。KER 的实现是我们将 KDD 技术用于具体应用领域的一个尝试。目前该系统已正常运行，受到学校科研管理决策者的好评。

参考文献

[1] Rakesh Agrawal, "Database Mining: A performance Perspective", IEEE Trans. Knowi. Data Eng, Vol. 5, No. 6, Dec. 1993, P914－924.

[2] 孟海军，"数据库中发现知识的方法"，小型微型计算机系统，Vol. 17, No. 2, Feb. 1996, P12－16.

[3] C. J. Matheus, "Systems for knowledge Discovery in Databases", IEEE Trans. Knowl. DataEng, Vol. 5, NO. 6, Dec. 1993, P903－913.

(来稿时间：1996 年 11 月)