数据挖掘与决策支持系统

谢　榕　（武汉测绘科技大学城市建设学院　430070）

摘要：决策支持系统的建立是一项复杂的系统工程。本文分析传统决策支持系统开发中存在的问题，探讨数据挖掘技术在决策支持中的应用，讨论基于这一技术的决策支持系统建立中的有关问题，如数据挖掘的层次空间、数据仓库的组织、知识发现方法等，最后进一步提出决策支持系统的基本结构框架和系统的建立方法。

关键词：数据挖掘　数据仓库　知识发现　决策支持系统

一、概述

近年来，随着科学技术的不断发展，数据库规模日益扩大，复杂程度不断增长，从大量数据中及时地获取制定有利于社会发展的策略信息显得越来越重要。然而作为信息处理新发展阶段的决策支持系统尚处于初级阶段，投入应用的成功实例并不多。有些开发的系统仅为简单的查询系统或报表系统，并不能给决策者提供辅助决策信息。分析原因主要有以下几个方面：(1)决策支持系统需要以集成数据为基础。然而现实中的数据往往是分散管理的且大多分布于异构的数据平台，数据集成不易。(2)决策支持涉及大量历史数据和半结构化问题，传统的数据库管理系统因自身的局限性并不提供这些方面的支持。(3)决策支持系统的建立需要对数据、模型、知识和接口进行集成。数据库语言数值计算能力较低，因而采用数据库管理技术建立决策支持系统在辅助决策支持方面知识表达、知识综合和知识推理能力比较薄弱，难以满足人们日益提高的决策要求。

90 年代初，数据挖掘技术的发展给以上问题的解决带来了新的契机。下面本文着重探讨数据挖掘技术在决策支持系统中的应用。

二、数据挖掘技术

1. 数据挖掘(Data Mining)的基本特点

数据挖掘是在一些事实或观察数据的集合中寻找模式的决策支持过程[8]。它具有以下特点：(1)数据挖掘要处理大量的数据，待处理的数据规模可能达到 GB、TB，甚至更大；(2)由于用户不能形成精确的查询要求，依靠数据挖掘技术为用户寻找他可能感兴趣的信息；(3)它把大量的原始数据转换成有价值的知识，用于描述过去的趋势和预测未来的趋势；(4)数据量增长快速，许多数据来不及分析就过时了。数据挖掘能快速地作出响应，提供决策支持信息。

2. 数据挖掘的层次空间与知识发现

从应用深度上，我们将数据挖掘划分为三个层次空间(如图 1 所示)：(1)数据空间。它利用现有数据库管理系统的查询检索和报表功能，进行基于关键字的决策查询，实现联机事务处理(On - Line Transaction Processing，简称 OLTP)。(2)聚合空间。利用聚集运算(Sum、Ave、Max、Min)，结合多维分析和统计分析，实现在线分析处理(On - Line Analytical Processing，简称 OLAP)，以提供决策参考的统计分析数据。(3)影响空间。按照相似性的聚类、差异性的分类方法，发现关联性及结构模式、顺序模式、相似时序，建立预测模型，从数据库或大量数据记录中发现隐含的有用信息，这是在更深层次上的知识发现，是数据挖掘实质性内涵。

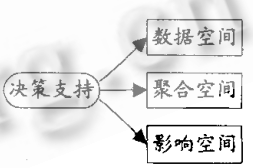

图 1　数据挖掘的层次空间

以上数据挖掘的各个层次空间反映了不同级别的查询请求，这种划分有利于知识的逐步提取，知识的提取过程即为决策支持过程。在传统决策支持系统中，知识库中的知识和规则是由专家或程序人员建立的，由外部输入，而数据挖掘是从系统内部自动获取知识的过程。同数据库管理系统查询检索的信息相比，数据挖掘的知识是隐含的、精练的和高水平的。

三、数据仓库的组织

数据挖掘在数据库上进行知识发现，为决策服务，需要新的数据管理支持平台。数据仓库从事物发展的角度

来组织和存储数据，供用户进行数据访问和分析[3,8]，因此它以其内在的对决策的支持能力，为数据挖掘提供了一种新型的数据存储地，同时它也为决策支持系统提供了可取的数据组织方式。

根据以上数据挖掘的层次空间，数据仓库可由五部分组成（如图 2 所示），即现状数据层、历史数据层、综合数据层和专题数据层、控制管理层。

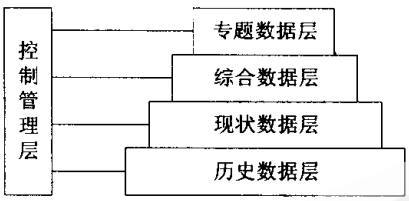

图 2 数据仓库的数据组织结构

1. **现状数据层**

它存储当前最新详细数据资料。数据从外部进入数据仓库时，首先被直接放入该层中。一般常规数据库可对其进行处理，这类数据又称为系统的基本数据。

2. **历史数据层和综合数据层**

决策所需要的信息是通过基本数据所体现的整体趋向或随时间变化而表现出来的变化趋势，必须对基本数据进行分类、析取、归纳、加工等处理才能得到这些信息。基本数据在时间控制机制下生成历史数据，放入历史数据层，供现状数据层、综合数据层和专题数据层调用。在综合机制下对基本数据进行综合、提取生成综合数据，放入综合数据层，它包括各种统计数据、指标、评价计算结果、预测分析数据等。

3. **专题数据层**

专题数据层存储对基本数据加工处理所形成专用数据，以及所建立的数学模型、预测模型、评价模型中各类参数等。

4. **控制管理层**

除了以上四个数据层外，决策支持还需要运用业务部门外部数据，它们一起共同构成数据仓库的信息来源。在控制管理层，通过建立提取器（Extractor），将来自信息源的、影响数据仓库信息的数据转化为数据仓库模式。当信息源中数据发生变化时，集成器（Integrator）对信息进行过滤、总结，并和其他信息合并，把新的信息集成到数据仓库中。

物理上，一个完整的数据仓库由以下四部分定义：（1）仓库设计部分。它负责数据仓库环境的定义和设置。（2）数据获取部分。它从外部数据源析取和变换数据，使这些数据以数据仓库的方式组织和存储。当数据来自不同数据源时，它还需解决不同数据源数据的重复和不一致性等问题。（3）数据管理部分。它完成数据更新、仓库例行维护以及分布数据的管理。（4）数据访问部分。它面向最终用户，在决策支持系统中，向决策者提供决策信息及分析报告。

四、知识发现方法研究

数据挖掘的知识通常表现为概念、规则、规律、模式、约束和可视化等形式。这些知识经过解释后可以直接在实际系统中应用，用以辅助决策过程，或者提供给领域专家，修正专家已有的知识体系，也可以作为新的知识转存到应用系统的知识库中。发现的过程是使数据挖掘利用各种知识发现算法从数据库中发现、表达、更新和解释有关知识。

1. **知识的发现**

数据关联是数据库中存在的一类重要的可被发现的知识[4]。若两个或多个变量的取值之间存在某种规律性，则称为关联。数据间的关联通过关联规则表示，其形式为：$A_1 \wedge A_2 \wedge \cdots \wedge A_i \rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_j$。如果 $B_1, B_2, \cdots, B_j$ 出现，则 $A_1, A_2, \cdots, A_i$ 一定出现，这表明数据 $A_1, A_2, \cdots, A_1$ 和数据 $B_1, B_2, \cdots, B_j$ 之间存在某种联系。关联分析采用关联规则归纳技术找出数据库中数据项（属性，变量）之间内在隐藏的关联网。

2. **不确定性知识的表达**

人们对事物的判断、预测和决策等是在问题域的信息不完全、不精确或者模糊的条件下进行的。粗集理论作为一种智能数据决策分析工具，被研究并应用于这种不确定性的知识获取和知识表达中。它通过构造描述领域知识的概念集 $A=\{a_1, a_2, (, a_n\}$，$a_i(\in A$，由系统的事实库形成对象集 $E=\{e_1, e_2, (, e_m\}$，$e_j(A$，基于分类集 P 和对象集 E，对 E 经过一定的操作产生核 $t^c(P,E)$ 和包络 $t^e(P,E)$，从而形成该问题的一个粗集 $[t^c(P,E), t^e(P,E)]$，构成不确定性区间，这样使用上限和下限两个量作为不确定性的测度。

3. **知识的更新和完善**

在人工智能和机器学习研究领域，神经网络（Neural Network）通过对大量样本模式的学习，得到从 n 维输入向量空间到 m 维输出向量空间的非线性映射 F：F：Rn? Rm。利用神经网络输出结果经专家认可后，将其作为新的样本实例存入系统中。因此，建立这种系统自学习

模型，可以不断地从样本模式中学习专家用于决策的定性的、经验性的知识，从而保证系统不断地获取新的知识以及对系统中拥有的网络知识进行更新和完善。

4. 知识的表达和解释

系统中最重要的应用是用户能够理解所发现的知识，这要求知识的展现不限于传统的数字或符号，而是更容易理解的方式，如表格、直方图、散点图或自然语言文本报告等。数据可视化采用直观的方式将信息模式、数据的关联或趋势多维地呈现给决策人员，决策人员通过可视化技术交互地分析复杂的空间数据关系，并能深入到数据的结构中了解数据的状况、内在本质及规律。

五、决策支持系统的建立

基于以上讨论，一种基于数据挖掘的决策支持系统基本结构框架如图 3 所示。它由数据库、数据仓库、数据仓库管理模块、数据挖掘工具、知识库、知识发现模块、人机交互模块组成。系统的主要输入是源于数据库的数据以及存储在知识库中的知识和经验。人机交互模块通过自然语言处理和语义查询在用户和系统之间提供相互联系的集成界面。数据仓库管理模块完成数据仓库的创建以及数据仓库中数据的综合、提取等各种操作，负责管理整个系统的运转。数据挖掘工具用于完成实际决策问题所需的各种查询检索工具、多维数据的 OLAP 分析工具和数据开采 DM 工具等，以实现决策支持系统的各种要求。知识发现模块控制并管理知识发现过程，它将数据的输入和知识库中的信息用于驱动数据选择过程、知识发现引擎过程和发现的评价过程。

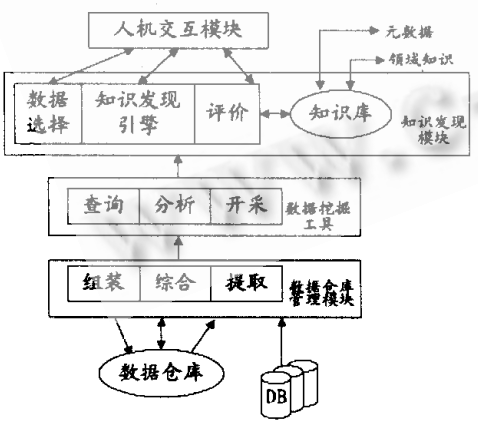

图 3　决策支持系统的基本结构框架

在图 3 中箭头方向为控制流。决策支持同数据仓库管理是密切联系的。用户发出决策请求命令后，通过数据挖掘工具触发数据仓库管理模块从数据仓库中获取与任务相关的数据。在知识发现模块中提供了大量知识发现引擎抽取算法，从数据仓库中选择的数据在知识发现引擎里得到处理，生成辅助模式和关系。在对这些模式和关系进行评价后，它们中的一些被认为感兴趣的数据将提供给决策部门应用。有些发现还可能加入到知识库中，以用于后继的知识发现过程和知识发现评价。

建立该决策支持系统的过程可描述如下：(1)分析决策需求，描述和表示决策的问题。这是一个分析过程，通过了解决策者的需求，确定决策主题、决策风格、流通信息及其传送方式等。(2)确定数据来源，建立数据仓库。从可操作的数据记录、数据库或文件系统中筛选所需的数据，对它们重新进行组织，存入数据仓库的不同信息层。然后综合并行技术、关系数据库系统和中间件，在现有异构环境基础上建立数据仓库。(3)针对所要发现任务的所属类别，如归类、回归分析、聚类、发现关联规则等，设计或选择有效的数据挖掘算法并加以实现。(4)数据挖掘，逐层综合。调用数据挖掘功能，从平凡的历史数据中提出综合数据，独立存储为库文件，作为更高一层数据挖掘的对象。与最终用户交互、协同，得到宏观性数据和趋势性知识。(5)测试与评价所发现的知识，对知识进行一致性、效用性处理。(6)应用开发。根据最终用户的要求，建立适用于决策支持的数据仓库的集成界面和应用程序，使用户能在决策支持中运用所发现的知识。

以上过程不是简单的线性流程，而是一个学习、发现和修改的过程，步骤之间包含了循环和反复，这样可以对所发现的知识不断求精、深化，并使其易于理解。

参 考 文 献

[1] R. Agrawal et al. Database Mining: A Performance Perspective. IEEE Trans. On Knowledge and Data Eng., 1993, 5(6)

[2] Sarabjob S. Anand, Bryan W. Scotney. Designing a Kernel for Data Mining. IEEE Expert Intelligent Systems and Their Applications, 1997, March - April

[3] 姚卿达，黄晓春，刘向民. 数据仓库和数据采掘应用研究. 计算机科学，1996, 23(6)

[4] 胡侃，夏绍玮. 基于大型数据仓库的数据采掘：研究综述. 软件学报，1998, 9(1)

(来稿时间：1999 年 2 月)