DOI:10.16526/j.cnki.11-4762/tp.2010.01.034

文章编号:1671-4598(2010)01-0161-04　　中图分类号:TP311.131　　文献标识码:A

综合决策支持系统中数据挖掘功能的设计与实现

周学全[1],张志杰[2],张笃行[3],赵　旭[4]

(1.南通航运职业技术学院,江苏 南通　226006;2.西南民族大学 计算机科学与技术学院,四川 成都　610041;
3.内江师范学院,四川 内江　641002;4.成都信息工程学院 电子工程系,四川 成都　610041)

摘要:通过对现有综合决策支持系统的架构分析,根据需要增加了一个中间件子系统部分,并进一步研究在该架构下如何利用数据挖掘技术来实现对决策主题数据的综合,即应该采取何种数据挖掘技术以及如何获取知识提供给综合决策支持系统;最后以灰色系统理论的建模法为例,给出了具体的理论、算法以及实现步骤与方法,可以为相关系统的设计与研究提供一定的借鉴和参考。

关键词:综合决策支持系统;数据挖掘

Design and Implementation of Data Mining Function for Synthetic Decision Support System

Zhou Xuequan[1], Zhang Zhijie[2], Zhang Duxing[3], Zhao Xu[4]

(1.Nantong Shipping College, Nantong　226006, China; 2.College of Computer Science & Technology, SouthWest University for Nationalities, Chengdu　610041, China; 3.NeiJiang Normal University, Neijiang　641002, China)
4.Department of Electronic Engineering, ChengDu College of Information Engineering, Chengdu　610041, China)

Abstract: Through the analysis of existing framework for Synthetic Decision Support System, according to the need for additional part of a middleware subsystem, and further research in the framework of how to use data mining techniques to achieve a comprehensive thematic data for decision-making, namely, what data mining techniques should be taken and how to obtain the knowledge available to the Synthetic Decision Support System and finally through the theory of gray system modeling method as an example of a specific theory, algorithms and implementation steps and methods for the design and research of related systems to provide some reference and reference.

Key words: synthetic decision support system; data mining

0　引言

综合决策支持系统(Synthetic Decision Support System,简称SDSS)是辅助决策者通过数据、模型和知识,以人机交互方式进行半结构化或非结构化决策的计算机应用系统。SDss是管理信息系统(MIS)向更高一级发展而产生的先进信息管理系统。SDss为决策者提供分析问题、建立模型、模拟决策过程和方案的环境,调用各种信息资源和分析工具,帮助决策者提高决策水平和质量。

综合决策支持系统能够把数据仓库、联机分析处理(On-Line Analysis Processing,简称OLAP)、数据开采、模型库结合起来,是一种更高级形式的决策支持系统。其中数据仓库能够实现对决策主题数据的存储和综合,OLAP实现多维数据分析,数据开采用以挖掘数据库和数据仓库中的知识[1],模型库实现多个广义模型的组合辅助决策,专家系统利用知识推理进行定性分析[2]。

本文按照现有综合决策支持系统的架构,分析并根据需要增加了一个中间件子系统部分,再进一步研究在该架构下如何利用数据挖掘技术来实现对决策主题数据的综合,即应该采取何种数据挖掘技术以及如何获取知识提供给综合决策支持系统,并且以灰色系统理论的建模法为例,给出具体的理论、算法以及实现步骤与方法。

收稿日期:2009-07-12;　修回日期:2009-08-30。

作者简介:周学全(1973-),男,山东烟台人,硕士研究生,主要从事计算机网络技术方向的研究。

1　综合决策支持系统的架构

目前,标准的决策支持系统基本结构主要由4个部分组成,即数据部分、模型部分、推理部分和人机交互部分。

其中,各个部分承担的主要功能如下:

(1)数据部分是一个数据库及其管理系统;

(2)模型部分包括模型库(MB)及其管理系统(MBMS);

(3)推理部分由知识库(KB)、知识库管理系统(KBMS)和推理机组成;

(4)人机交互部分是决策支持系统的人机交互界面,用以接收和检验用户请求,调用系统内部功能软件为决策服务,使模型运行、数据调用和知识推理达到有机地统一,有效地解决决策问题。

目前,数据格式日趋多种多样,需要处理的数据量越来越多,因此为统一数据格式、进行数据格式转换,增加一个提供统一数据接口的中间件子系统,因此该决策支持系统调整后的架构如图1所示。

经过调整后,该决策支持系统大体上由以下5个部分组成:

(1)对数据进行管理的决策数据管理子系统;

(2)模型、方法和知识管理子系统;

(3)智能决策支持子系统;

(4)提供统一数据接口的中间件子系统;

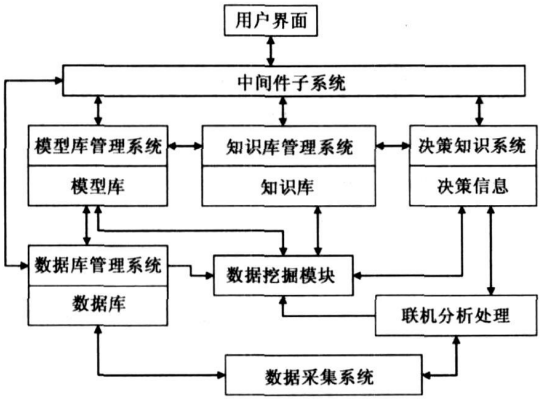

图1 决策支持系统示意图

（5）人机交互部分，即人机交互界面。

其中，各个组成部分的比较关键功能如下：

1）决策支持与数据管理系统

数据管理系统必须为决策支持的分析处理提供以下服务：

（1）根据主题需要，从 OLAP 数据库中抽取分析用的数据。为此在抽取过程中要对原始数据进行分类、求和与统计等处理，抽取的过程实际上是数据的再组织。

（2）在抽取过程中，完成数据净化，即去掉不合格的原始数据，必要时还必须对缺损的数据加以补充。

（3）在改变分析、决策的主题时，可以按主题进行数据查询与访问。

（4）采用脱机大容量存储、联机磁盘存储和内存存储的多级存储模式，解决数据量巨大及按照主题、粒度划分的数据组织问题。

2）模型、方法和知识管理系统

在决策支持系统中，模型、方法和知识的管理是核心，它对依问题建立的模型库、方法库和知识库进行管理。

模型、方法和知识管理系统的主要任务是：

（1）对模型库、方法库和知识库进行维护。模型、方法和知识管理系统必须有对数据统一的维护与操作，并保证使用数据的一致性：一是系统运行过程调用每个库时不发生矛盾，特别是对知识库的维护更为复杂；二是每种模型、方法和知识都能调用到。

（2）模型、方法和知识管理系统根据用户的要求和数据仓库提供的数据，能有效地选择模型、方法和知识，经系统运行得到相应的结果，并将结果送给交互环境进行输出。

3）智能决策支持系统

智能决策支持系统一般是在模型、方法和知识管理系统的基础上增加专家系统和数据采掘与知识发现技术[3]。随着数据量的增大，不确定因素的增多，专家系统技术和各种推理技术对提高决策支持的准确度十分必要，在人也无法描述出数据间的关系时，就需要采用数据采掘与知识发现技术。

4）提供统一数据接口的中间件子系统

由于决策的需要，要尽可能地收集各种数据，因此也就存在原始数据来源多、结构混乱、连接分析对不少数据格式不可行的问题。

中间件系统的目标旨在对原始数据进行一定的预处理，从而得到结构简单、格式规范、统一的规范化数据[4]，供 OLAP 数据库、专家系统进行分析、预测和报告生成等进一步操作使用。

其实，这部分如果要扩展成为一个主要部件，就可以采用建立一个简单的数据仓库的方法，将经过处理后形成规范化的数据放在数据库中，也就是产生一个数据仓库，数据仓库的具体实现形式与难易程度需视技术要求而定。

2 数据挖掘原理

数据挖掘（Data Mining，简称 DM）是一种崭新的信息处理技术，其主要特点是对数据库的大量业务数据进行抽取、转化、分析和模式化处理，从中提取辅助决策的关键知识，即从一个数据库中自动发现相关运作模式，换而言之就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[5]。

从知识提取的角度来看，就是将计算机系统直接获取的原始数据、原始信息进行加工处理，提取出概念、规则、模式、规律和约束等知识。

因此，数据挖掘把对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。

数据挖掘发现知识采用的技术涉及到许多不同领域：数据库技术、人工智能技术、数理统计、可视化技术、并行计算等，形成若干新的技术热点，同时，必须解释的是：发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。

经过数据挖掘处理后发现的知识可以被用于信息管理，查询优化，决策支持和过程控制等，还可以用于数据自身的维护等诸多方面。

数据挖掘与传统的数据分析（如查询、报表、联机应用分析）的本质区别是：数据挖掘是在没有明确假设的前提下去挖掘信息和发现知识，数据挖掘所得到的信息应具有未知、有效和实用 3 个特征。

目前，在综合决策支持系统中，实现数据挖掘技术主要有以下 4 个相关操作：预测性建模、数据库分段、连接分析与偏离检测[6]。

（1）预测性建模：有限的经验性知识无法覆盖可能出现的全部情况，因此，还需要从实际数据中发掘出知识中没有提到但有可能对企业行为产生影响的信息。

（2）数据库分段：分段存储主要是通过均衡磁盘 I/O，实现内部查询的并行操作、并行地扫描多个磁盘上的数据来提高查询效率，其使数据库性能的提高主要来自于 I/O 并行度的提高，而不是 I/O 性能的提高。

（3）连接分析：从一些用户的行为中分析出一些模式，并且将产生的模式概念应用于更广的用户群体中，这样可以充分利用数据间的固有关系，提供强大的可视化能力，并且容易创建衍生属性。

（4）偏离检测：数据偏离指的是与标准值不符，或在比对试验中，数据与其他经验数据得出的中位值偏离很大。一旦发生这样的情况，应对于数据进行分析，找出偏离原因，提供模型、方法和知识管理系统重新处理。

在预测性建模、数据库分段、连接分析与偏离检测这 4 个实现数据挖掘技术中，比较关键的是预测性建模，直接关系到

数据分析的速度与质量。

3　预测性建模

下面，就预测性建模技术，做一个简单的分析与应用案例。

数据挖掘中，建模技术已经有很多种，如：卡方自动交叉检验（CHAID：Chi-squared Automatic Interaction Detector）方法、决策树技术、有理函数模型（RFM：Rational Function Model）模型、Bayes 估计（Bayes estimation）、多层 Bayes 估计（Multiple Bayes estimation）、经验 Bayes 估计模型（Empirical estimation），线性和 logistic 回归预言模型，神经网络预言模型，基于分类回归树 CART 模型[7] 等。每种模型都有其优劣，应用场合也有所不同。

这里，以灰色系统理论（Grey System Theory）的建模为例，表明如何利用建模理论为数据挖掘应用建立预测分析的模型。

灰色系统理论就是一种研究少数据、贫信息不确定性问题的新方法，即运用一定的数学方法，使信息不完全明确的系统经数据处理后能得到较明确的、符合实际情况的结果的一种新兴数学预测系统，由邓聚龙教授创立。

该理论以“部分信息已知，部分信息未知”的“小样本”、“贫信息”不确定性系统为研究对象，主要通过对“部分”已知信息的生成和开发，提取有价值的信息，实现对系统运行行为、演化规律的正确描述和有效监控。该系统已在社会、经济、农业、生态、气象、环境、政法及管理等部门得以较为广泛的应用。

3.1　灰色预测模型建模原理与计算方法

灰色数列预测模型是以灰色系统概念为核心，通过将无规律的原始数据生成、建模、拟合后，进而推测未来的一种新兴数学预测模型系统。最常用的模型是含一个变量的一阶微分方程，称之为 GM（1,1）模型。GM（1,1）模型的建立步骤与方法：

（1）一次累加生成：设原始数列 $X(t)=\{x(1), x(2), \cdots, x(n)\}$，对其进行一次累加生成，以弱化其随机性，强化其规律性，得累加生成列

$$Y(t):\quad y(t)=\sum_{j=1}^{t}x(t) \tag{1}$$

（2）均值生成：对累加数据列按公式（2）作均值生成，得均值数据列

$$z(t)=\frac{1}{2}[y(t)+y(t-1)] \tag{2}$$

（3）建立 GM（1,1）模型：建立关于 $Y(t)$ 的一阶线性微分方程：

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t}+ay(t)=u \tag{3}$$

此式即为 GM（1,1）预测模型，解该变量分离型微分方程得其特解为：

$$y(t)=\left[x(1)-\frac{u}{a}\right]e^{-a(t-1)}+\frac{u}{a} \tag{4}$$

式中 a，u 为待定系数，根据最小二乘法估计参数向量，并由矩阵计算得其表达式为：

$$a=\frac{1}{D}\{(N-1)[-\sum_{t=2}^{N}x(t)z(t)]+(\sum_{t=2}^{N}z(t)\sum_{t=2}^{N}x(t))\} \tag{5}$$

$$u=\frac{1}{D}\{[\sum_{t=2}^{N}z(t)][-\sum_{t=2}^{N}x(t)z(t)]+[\sum_{t=2}^{N}Z^2(t)][\sum_{t=2}^{N}X(t)]\} \tag{6}$$

式中，$D=(N-1)[\sum_{t=2}^{N}z^2(t)]-[\sum_{t=2}^{N}Z(T)]^2$ (7)

由式（4）所得估计值（t）数列作累减还原生成，得原始数列 $X(t)$ 的估计值（t）数列：

$$\hat{x}(t)=\hat{y}(t)-\hat{y}(t-1)\,23^{34} \tag{8}$$

对数列（t）与 $X(t)$ 进行拟合效果检验（可靠性检验）若两者拟合精度好，则模型可用于外推预测；若两者拟合精度不合格，则不可直接用于外推预测，须经残差修正后，再进行外推预测。确定灰色数列模型的可靠性可用平均相对误差、后验差比值和小误差概率来检验。

平均相对误差：$\hat{e}$

$$\hat{e}=\frac{1}{\sum x(t)}(\sum\hat{x}(t)-x(t))\times 100\% \tag{9}$$

残差 $\varepsilon(t)$

$$\varepsilon(t)=x(t)-\hat{x}(t) \tag{10}$$

计算后验差比值 C 和小误差概率 P：

$$C=\frac{S_2}{S_1} \tag{11}$$

$$P=P(|\varepsilon(t)=x(t)-\hat{\varepsilon}(t)<0.6745S_1) \tag{12}$$

式中，$S_1^2=\frac{1}{n}\sum_{t=1}^{n}[x(t)-\hat{x}(t)]^2$；$S_2^2=\frac{1}{n}\sum_{t=1}^{n}[\varepsilon(t)-\hat{\varepsilon}(t)]^2$；根据精度检验等级参照表判断灰色数列的拟合优度。

外推预测如果拟合优度高，即模型预测效果满意，可按下式进行外推预测：

$$\hat{x}(t)=\hat{y}(t)-\hat{y}(t-1)\qquad t=n+1, n+2\cdots\cdots \tag{13}$$

3.2　灰色预测模型的应用

在完成基于灰色数列预测模型构建后，就可以进行预测。

预测按照以下步骤进行：

（1）收集整理数据，使其具有代表性（这里为提高数据精度，应该采取其他数据处理手段，如聚类算法等，防止数据偏移）；

（2）划分数据集为样本集与修正集：其中样本集可以得出模型的主要参数，而修正集可以检验并且校核模型的主要参数；

（3）利用模块以及校核后模型的主要参数进行预测；

（4）对于预测结果进行评价，并且调整或者修改模型的主要参数（这里进行评价是提高预测质量的关键，必须采用综合评价方法，仅仅使用数学方法或类似专家系统的知识方法都是不充分的）；

4　小结

本文按照现有综合决策支持系统的架构，分析并根据需要增加了一个中间件子系统部分，再进一步研究在该架构下如何利用数据挖掘技术来实现对决策主题数据的综合，即应该采取何种数据挖掘技术以及如何获取知识提供给综合决策支持系统，并且以灰色系统理论的建模法为例，给出具体的理论、算法以及实现步骤与方法。

目前该综合决策支持系统应用数据挖掘技术，还存在的主要问题为：复杂环境下的综合决策模型的提出、分布式综合决策模型的架构、海量数据存储预处理算法的实现、数据预处理方法评价指标等。这些还需要在今后的课题中进一步具体研究。

参考文献:
[1] AlSairafi S, Emmanouil F S, Ghanem M, et al. The design of discovery net: towards open grid services for knowledge discovery [J]. High-Performance Computing Applications, 2003, 17 (3): 297-315.
[2] Brezany P, Janciak I, W¨ohrer A, et al. GridMiner: A Framework for knowledge Discovery on the Grid-from a Vision to Design and Implementation [A]. Proceedings of the Cracow Grid Workshop [C]. Cracow, Poland, 2004.
[3] Cannataro M, Congiusta A, Mastroianni C, et al. Grid-based data mining and knowledge discovery [A]. Intelligent Technologies for Information Analysis [C]. Springer, Berlin, Germany, 2004.
[4] Cunningham H, Humphreys K, Gaizauskas R, et al. Software infrastructure for natural language processing [A]. Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97) [C]. San Francisco, CA, USA, 1997: 237-244.
[5] Daniel G, Dienstuhl J, Engell S, et al. Advances in computational intelligence-theory and practice, chapter Novel Learning Tasks, Optimization, and Their Application [C]. Springer 2002, 245-318.
[6] Euler T. Publishing operational models of data mining case studies [A]. In Proceedings of the Workshop on Data Mining Case Studies at the 5th IEEE International Conference on Data Mining (ICDM) [C]. Houston, Texas, USA, 2005: 99-106.
[7] Kietz J U, Vaduva A, Z¨ucker R. MiningMart: Metadata-Driven Preprocessing [A]. Proceedings of the ECML/PKDD Workshop on Database Support for KDD [C]. 2001.

(上接第 129 页)

能及读写使能引脚与 PCI9052 的 ISA 模式的地址线、数据线、读信号和写信号引脚相连实现上位机读写数据的功能。

本设计主要用到了 PCI9052 的 ISA 模式的功能，将以往的 ISA 接口的设计转换为 PCI 接口。PCI9052 的 MODE 引脚是模式选择信号，本设计中该引脚接地，使用 ISA 非复用模式。同时通过在配置芯片 93LC46B 中设置 PCI9052 的相应配置寄存器来将 PCI9052 设为 16 位 ISA 工作模式，配置芯片 93LC46B 的设置内容如图 2 所示（表中地址和内容均采用 16 进制表示）[2]。

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	29	84	B5	10	80	06	02	00	50	90	B5	10	00	00	00	01
10	F0	FF	00	00	FF	FF	F1	FF	FE	FF	00	00	FF	FF	F0	FF
20	00	00	00	00	00	00	01	00	00	00	01	00	00	01	01	00
30	00	02	01	00	00	00	00	00	40	00	22	00	00	00	22	00
40	40	00	01	00	40	00	01	00	00	00	00	00	08	00	01	00
50	00	00	09	00	01	01	01	00	00	02	09	00	00	00	5B	11
60	7C	00	92	4C	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF

图 2 93LC46B 0x00h ~ 0x6Fh 单元内容

4.4 FPGA 中的模块电路设计

在 FPGA 中实现了 F2812、双口 RAM 和 DEI1016 的外围控制电路，并提供 F2812 及 DEI1016 工作的时钟信号。同时，在 FPGA 中还实现了一个 32 位的实时时钟，用于为接收的 ARINC429 数据打时标，满足实际系统的需要。而且在 FPGA 中还实现了 8 个定时器，分别由 8 个定时器周期寄存器、8 个定时器周期个数寄存器和 8 个定时器控制寄存器来控制，通过分别设置这 3 类定时器寄存器，可以实现 ARINC429 数据的定时连续发送或定时发送一次等功能，满足工程应用上的不同需求。总的来说，FPGA 中的模块电路从功能上主要可分为 2 个部分，分别说明如下：

(1) 产生 F2812 及 DEI1016 工作时钟的模块

在本设计中，由 1 个 24 MHZ 的晶振给 FPGA 提供时钟信号源。由 FPGA 提供给 F28121 个 24MHz 的工作时钟。同时，通过分频电路 FPGA 可以分别提供给 DEI1016 的工作时钟分别有 1MHz、960kHz 和 480kHz，分别可以产生 100k、12 5k、96k、48k 4 种 429 通信速率，满足不同外部设备的要求。

(2) 数据通信及功能控制模块

该模块中用 VHDL 语言设计了两个双向三态的数据收发驱动器，分别用于 F2812 和双口 RAM、F2812 和 DEI1016 之间的数据通信。为了使 F2812 访问外部的双口 RAM、FPGA 及 DEI1016 协议芯片，把双口 RAM 映射到 F2812 的 Zone0 和 Zone2，把 DEI1016 映射到 F2812 的 Zone6。在 Zone0 上实现 7028 的 8 个字标志寄存器的访问，分配的地址范围为 0x2000 ~ 0x2007；在 Zone2 上实现 7028 的 64k 字双口 RAM 的访问，分配的地址范围为 0x80000 ~ 0x8FFFF；在 Zone6 上实现 8 个 ARINC429 数据发送通道、16 个 ARINC429 数据接收通道、DEI1016 芯片的复位及控制字寄存器、状态字寄存器、接收和发送 ARINC429 数据产生给 F2812 的中断及实时时钟和定时器等板卡功能，分配的地址范围为 0x100000 ~ 0x100078。其中接收和发送 ARINC429 数据产生给 F2812 的中断分别通过 DEI1016 芯片的接收器数据准备好信号 DR 及发送器准备好信号 TXR 的电平跳变作为 D 触发器的时钟输入来产生。实时时钟、定时器及 F2812 中断的屏蔽寄存器、状态寄存器等信息利用了 FPGA 上的存储器空间来实现。

5 结束语

本设计采用智能化设计能够满足工程现场不同计算机的接口需求，同时其 8 发 16 收的 ARINC429 通信资源可以满足多数用户的要求，而且根据不同应用的要求可以对收发通道进行裁减，变为 1 发 2 收、2 发 4 收、3 发 6 收、4 发 8 收、5 发 10 收、6 发 12 收、7 发 14 收、8 发 16 收，在为用户节省费用的同时可以减少通信接口板的数量。通过整个设计方案可以看出，用 DSP 加 FPGA 的电路设计，使得电路设计简单且调试容易，升级和维护也十分方便。根据其在不同工程中的应用，接口板工作稳定，能够满足不同应用的要求，有着广泛的应用前景。

参考文献:
[1] 邓智敏，张 军. 基于 HS3282 的 ARINC429 总线通讯卡的设计与应用 [J]. 计算机测量与控制，2004，12 (5)：476-479.
[2] 吴业进，刘 锋. PCI9052 总线接口芯片及其 ISA 模式应用 [J]. 测控技术与设备，2003，29 (4)：24-26.
[3] DEI 1016 ARINC 429 Transceiver datasheet [Z]. Device Engineering Inc.，2002. 9.
[4] BD429 ARINC 429/RS-422 Line Driver Integrated Circuit datasheet [Z]. Device Engineering Inc.，2002. 7.
[5] 苏奎峰，吕 强，等. TMS320F2812 原理与开发 [M]. 北京：电子工业出版社，2005.
[6] TMS320F2810，TMS320F2811，TMS320F2812，TMS320C2810，TMS320C2811，TMS320C2812 Digital Signal Processors Data Manual [Z]. Texas Instruments，2003. 12.