基于纺织企业纱线质量数据仓库的设计

吴程赟,李 奇
(东南大学,江苏 南京 210096)

摘 要:基于纺织企业纱线质量管理数据,结合数据仓库、数据挖掘、联机分析处理技术,探讨了纱线质量数据仓库整体设计过程,并且实施数据挖掘,分析数据挖掘结果,对企业生产提供依据。提出了基于纺织企业纱线质量数据仓库的组成与设计思路,以实例阐述了构建数据仓库的体系结构和数学模型以及数据挖掘、联机分析处理技术,结果表明基于纱线质量的数据仓库的建立结合数据挖掘和联机分析处理技术对企业质量管理层确实提供了有效的分析手段。

关键词:纱线;数据仓库;数据挖掘;联机分析处理
中图分类号:TP311.138 **文献标识码**:A **文章编号**:1673-629X(2006)06-0077-03

Design of Yarn Quality Data Warehouse Based on Textile Enterprise

WU Cheng-yun, LI Qi
(Southeast University, Nanjing 210096, China)

Abstract: Presents the process of integral design of yarn quality data warehouse based on textile enterprise, which combined data warehouse, data mining and on line analysis processing (OLAP) technology, and puts it in practice with data mining, then analyzes the result which can provide gist to manufacture of enterprise. The composition and design of yarn quality data warehouse are put forward based on textile enterprise, and then the system structure, data model, data mining and OLAP are given from examples, and result shows the effect of this means is satisfactory.

Key words: yarn; data warehouse; data mining; OLAP

0 引 言

随着企业计算机应用的不断深入,许多纺织企业已经投入了大量的时间和资源建立了庞大而复杂的管理信息系统,经过多年的运行,积累了大量的宝贵数据资源。面对日益激烈的市场竞争,这些企业迫切希望能有一个强而有力的分析工具来帮助他们从这些海量的数据中充分挖掘有意义的信息,以辅助高层领导者进行计划和指导决策活动。

数据仓库的目的是为了建立一种体系化的数据存储环境,将分析决策所需要的大量数据从传统的操作环境中分离出来,使分散、不一致的操作数据转成集成、统一的信息,进而支持决策。完整的数据仓库包括三个方面的技术内容:数据仓库技术、联机分析处理技术和数据挖掘技术[1]。该文以安徽华茂集团公司管理信息系统为例,针对技术质量数据建立数据仓库,并提出适用纺织行业的方案设计思想。

收稿日期:2005-09-06
作者简介:吴程赟(1981-),男,江苏常熟人,硕士研究生,研究方向为数据仓库与数据挖掘;李 奇,教授,博士研究生导师,研究方向为智能控制、综合自动化、计算机监控等。

1 数据仓库的概念

1.1 数据仓库(DW)

数据仓库概念创始人 W. H. Inmon 对数据仓库的定义:数据仓库是面向主题的(subject oriented)、集成的(integrated)、稳定的(nonvolatile)、不同时间的(time-variant)数据集合,用于支持经营管理中的决策制定过程。主题是与传统数据库的面向应用相对应的,是一个抽象的概念。集成性是指在数据进入数据仓库之前,必须经过数据加工和集成,这些数据可能来自不同的数据源,当这些数据进入 DW 时要进行必要的转换、重新格式化、重新排列以及汇总等操作[1]。如对不同的数据来源要统一数据结构和编码,将原始数据结构从面向应用过程到面向主题、面向支持决策的转换。DW 稳定性是指数据通常是以批量方式载入与访问的,在 DW 中的数据在进行转载时是以静态快照的格式进行的。当产生后续变化时,一个新的快照记录就会被写入 DW。不同时间的是指 DW 中的数据是历史数据的集合。

1.2 数据仓库的应用

建立数据仓库的目的是将可用的数据和信息按可用的形式和格式传送给用户。利用一系列决策支持工具增加用户的能力,以便更好地分析数据并做出决策。决策支

持过程通常使用的方法为分析处理信息、分析处理和数据挖掘[2]。信息处理包括查询、计算和报表等;分析处理包括在线分析处理(OLAP);数据挖掘包括统计分析、知识发现等。

2 实例介绍数据仓库的设计

以安徽华茂集团管理信息系统中纱线质量数据建立数据仓库,主要介绍面向主题即面向对象的设计方法。

2.1 数据仓库体系结构设计

企业数据仓库系统的技术体系结构通常包括后台数据预处理(数据获取)、数据仓库数据管理和数据仓库的前台查询服务(应用服务)三大部分(见图 1)。

(1) 安徽华茂集团信息管理系统运行多年,积累了大量的数据,涉及到纱线质量数据的数据源包括多个子系统:技质棉检系统、纺部试验系统、织部试验系统、技质成检系统、技质办公室系统以及公共库。

(2) 元数据:元数据是描述数据仓库内数据的结构和建立方法的数据,记录了数据的结构和数据仓库的任何变化,以支持数据仓库的开发和使用。

(3) 上面的数据源不可能直接载入数据仓库的数据库中,必须进行数据的预处理。数据的预处理包括数据源的定义,从数据源提取相关数据到预处理数据区(数据准备区),在数据准备区中对数据进行净化处理、转换,再将数据加载到数据仓库。

(4) 数据仓库的管理包括数据仓库的创建、数据仓库的维护、对数据仓库中数据的重整和数据仓库的元数据管理等。该部分的核心功能是完成数据仓库的建模、确定数据的粒度级别、指定数据仓库的物理存储模式、确保数据仓库的运行效率等。

(5) 在线分析处理(OLAP):在线分析处理以数据立方体或多维的方式来查看数据,允许用户进行钻取以获得更详细或更概括的数据,或者对不同的维进行切片等操作。

2.2 数据仓库数据模型设计

进行数据仓库的设计开发时,与设计传统数据库一样,通常进行三个层次的数据建模:即建立高层的概念建模、中层的逻辑建模和底层的物理模型[3]。数据模型设计首先是对用户需求的归纳,纱线质量是纺织厂的生命,因此在此设计基于纱线质量的数据仓库模型,数据仓库开发产品采用 Microsoft SQL Server 2000。

2.2.1 高层概念模型

数据仓库概念模型的设计主要是确定数据仓库中应该包含的数据类及其相互关系。传统的概念模型设计采用实体-关系(E-R)模型来建模,而数据仓库的概念模型一般采用多维数据模型来建模。多维数据模型是一种能够清楚地表达分析领域的数据模型,它非常直观,容易理解,与人们分析问题的思维方式一致。

在多维数据模型中,包含两种建模要素:观察事物的角度和观察得到的事实数据。前者被称为维度,后者被称为事实。一个分析领域或主题表达为由多个维度和一组事实数据构成的星型模型[1]。在此确定纱线质量作为主题域,建立维度事实星型模型(见图 2)。

2.2.2 中间数据模型

高层数据模型建好后,就要建立下一层中间层模型

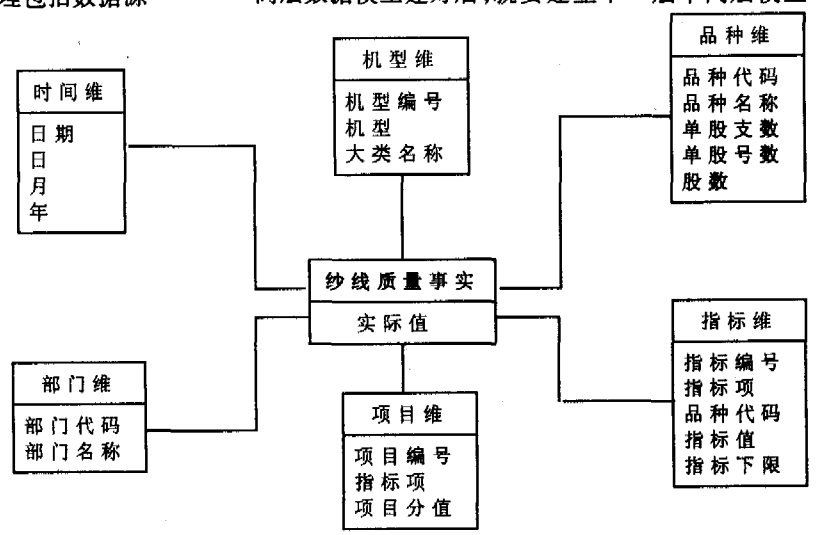

图 2 纱线质量的多维数据模型

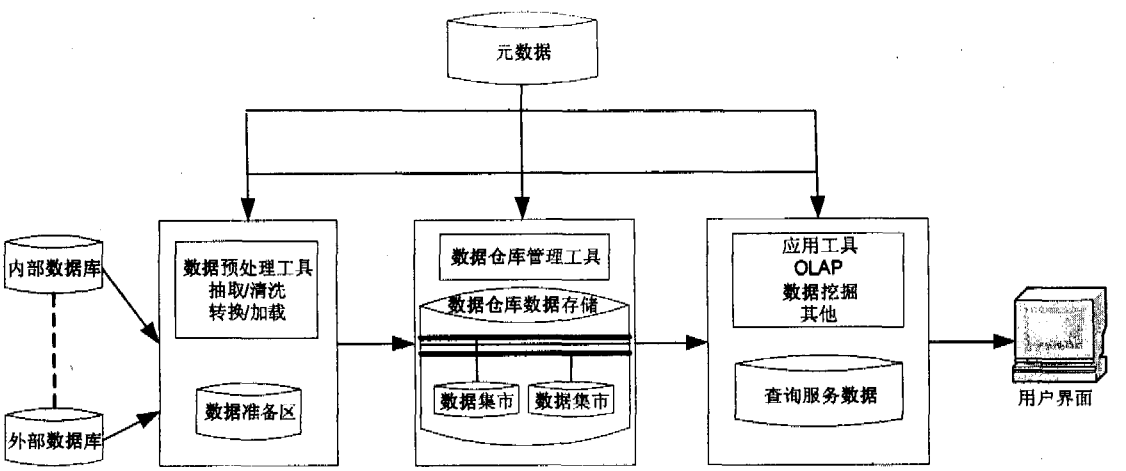

图 1 数据仓库的技术体系结构

(DIS)即逻辑数据模型。对高层模型中标识出的每个主要主题域、实体,都要建立一个中间层模型。数据仓库逻辑模型的设计主要包括以下工作:粒度层次划分,数据分割策略的确定,关系模式的定义,数据源及数据抽取模型的确定等等。

逻辑模型主要使用事实表和各维度表的关系模式来表达,而关系模式的确定与粒度层次的划分有关。表1与表2是对应上述纱线质量主题的纱线质量事实表和品种维度表的关系模式详细说明。

表1　纱线质量事实表的关系模式

列名	含义	列的码属性	取值范围	类型与大小
Time_ID	时间码	主码列,外码列	正整数	Integer
Department_ID	部门码	主码列,外码列	正整数	Integer
Machine_ID	机型码	主码列,外码列	正整数	Integer
Product_ID	品种码	主码列,外码列	正整数	Integer
Standand_ID	指标码	主码列,外码列	正整数	Integer
Project_ID	项目码	主码列,外码列	正整数	
Finished Values	完成值		正数	Float

表2　品种维度表的关系模式

列名	含义	列的码属性	取值范围	类型与大小
Product_ID	品种码	主码列	正整数	Integer
Product_Number	品种代码		实际品种代码	Varchar(20)
Product_Name	品种名称		实际品种名称	Varchar(50)
…	…	…	…	…

关系模式表中的码并不是具有原始含义的各种代码,例如品种维度表的码Product_ID并不是品种的真实代码Product_Number,它是从1开始的自然数,这样便于维度表与事实表之间的连接,并且在各种实际代码发生变化的情况下也不会影响连接应用。

数据抽取模型是逻辑模型的一部分,它包括对数据源的说明、数据抽取规则、数据源的列与数据仓库的对应关系。为了将数据载入数据仓库,必须首先从数据源抽取数据到数据准备区。为此,应该确定可以从哪些数据源抽取数据,这些数据源是基于什么系统平台的(见表3)。

表3　数据源抽取对象表

系统平台	数据库名	表名
Windows/SQL Server	公共	GG纱产品代码
Windows/SQL Server	纺部试验	FB质量考核明细表
Windows/SQL Server	纺部试验	FB质量考核指标表
Windows/SQL Server	公共	ZT部门
…	…	…

数据抽取到数据准备区后,并不能直接加载到数据仓库中去,还需要对数据进行各种清理工作,包括格式转换、类型转换、统一单位,或者将数据按照划分的粒度层次进行汇总、聚集等。经过抽取和清理的数据,才能从数据准备区加载到数据仓库中去。

2.2.3　物理数据模型

逻辑数据模型封装用户的需求,在体系结构设计阶段要将中间层数据模型转换成物理模型。换句话说,在体系结构设计阶段实际确定中层数据模型在磁盘驱动器上的结构和布局。这个任务包括将实体变成表以及进行必要的重建构造;定义所需要的索引;开发在磁盘上分割数据的计划。然而这一步是要考虑性能特性,意味着确定数据的粒度与分区是很重要的。粒度是数据仓库的数据单位中保存数据的细化或综合程度的级别。越是详细的数据,粒度级别就越小;越是概括的数据,粒度级别就越大[4]。由于用户查询的质量数据一般以月份为单位,但也可能查询某一天具体值,并且判断是否超下限。因此,可以建立两个粒度级别。第一个是低粒度级的,每天的质量数据都记录下来;第二个是高粒度级的,记录一个月的平均值,因此相对某一品种只有一条记录。采用高粒度存放在数据仓库中,供技术质量分析人员使用,可以提高查询效率,及时定出质量指标,便于考核,同时节省存储费用。

3　数据挖掘与联机分析处理

3.1　数据挖掘

数据挖掘(Data Mining),简单地讲,就是要从大量的数据中整理出或挖掘出有用的知识,这些知识是隐含的、事先未知的、具有潜在有用信息的,它们可表示为概念、规则、规律、模式等形式[1]。数据挖掘技术主要分为"关联规则"、"时间序列"、"聚集"、"分类"、"估值"等几类,研究和开发要涉及到多个领域的知识。建立在数据仓库基础上的数据挖掘,可以简化数据挖掘过程的某些步骤,大大提高数据挖掘效率。数据挖掘技术是数据仓库应用中比较重要也是相对独立的部分,涉及到数理统计、模糊理论、神经网络和人工智能等多种技术[5]。

3.2　联机分析处理

联机分析处理(OLAP),是分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互的存取,从而获得对数据的更深入了解的一类软件技术[3]。OLAP的目标是满足决策支持或者满足多维环境下特定的查询和报表需求,它的技术核心是"维"(dimension)这个概念,"维"是人们观察客观世界的角度。可把一个实体的多项重要属性定义为多个维,使用户能对不同维上的数据进行比较,因此OLAP也可以说是多维数据分析工具的集合。OLAP工具可通过多维的方式对数据进行分析、查询和报表。在纱线质量数据仓库系统中,通过Windows/SQL Server Analysis Services建立的多维数据模型,包括时间维、部门维、品种维等,而这些维的不同组合和所考察的纱线质量数据构成的多维组就是OLAP分析的基础,可形式化表示为(维1,维2,…,维 n ,纱线质量),多维分析是指对以多维形式组织起来的数据采取切片(Slice)、切块(Dice)、钻取(Drill－down和Roll－up)、旋转(Pivot)等各种分析动作,以求剖析数据,使用户能多角度、多侧面地观察数据仓库中的数据,从而深入理解包含在数据中的信息。

4　结束语

建立一个成功的数据仓库系统,需要有一个坚实的基

(下转第82页)

$F_i = 0$ 区间约束值上限 ≥ 试卷约束函数值 ≥ 区间约束值下限

F_i = (试卷约束函数值 − 区间约束值上限) × K_{i1} 试卷约束函数值 ≥ 区间约束值上限

F_i = (试卷约束函数值 − 区间约束值下限) × K_{i2} 试卷约束函数值 ≤ 区间约束值下限

点约束值、区间约束值上限、区间约束值下限值都是组卷蓝图规定的常数值,而试卷约束函数值是指进化学习过程中产生的某一个样本试卷相应的计算约束函数值。上述的 K_i, K_{i1}, K_{i2} 是根据经验确定的常数。如果某一约束比较重要,则相应的 K_i, K_{i1}, K_{i2} 常数就取比较大的值,反之,则取比较小的值。实际上,这些常数的具体数值仅作为优化学习方向的一个指导。计算表明,在组卷条件合适的情况下,优化目标最后都是收敛到零。如果组卷条件与试题库不是很匹配,最后优化目标大于零。而这个大于零的数据主要是由权重比较小的约束条件贡献的。

这里选择了一个包含 550 个试题、4 种类型(单选题、多选题、判断题、填空题)的小试题库作为例子进行计算。该试题库的选择题题分固定,每题 1 分。其他类型题目分值不固定。试题属性包括:试题类型、分值、答题时间、难易度、知识点等。分别选用 4 种组卷蓝图(表 1 为其中一个组卷蓝图)进行计算,均能得到比较好的结果。从这些计算过程可以得出以下几条经验:

表 1 组卷蓝图描述

约束类型	约束序号	内容	点约束		区间约束			
			点约束值	权重因子	约束下限	下限权重	约束上限	上限权重
试题类型	1	选择题 20 分	20	10				
	2	多选题 20~30 分			20	5	30	5
	3	判断题 20~30 分			20	5	30	5
	4	填空题 20~30 分			20	5	30	5
分值	5	试卷总分 100	100	10				
难度	6	试卷难度系数			2.4	100	2.6	100
知识点	7	第 1 章 20~25 分			20	10	25	10
	8	第 2 章 0 分	0	100				
	9	第 3 章 20~25 分			20	10	25	10
	10	第 4 章 25~30 分			25	10	30	10
	11	第 5 章 25~30 分			25	10	30	10
	12	第 6 章 0 分	0	100				
时间	13	考试时间 100~120 (min)			100	10	120	100

(1)初始概率的选取:取值最好在试卷试题总数估计值/试题库试题总数值附近。

(2)修正因子不能太大,一般选值在 0.01~0.1 之间,以 0.05 左右为好。

(3)使用信息熵来估计进化的程度。PBIL 算法最后收敛到一个点(信息熵为零)。如果收敛太快,可以减小修正因子;反之增大修正因子。

(4)组卷蓝图的制订一定要参考试题库的一些综合信息。蓝图制订不合理,就得不到满意解。

如果一次搜索得不到满意的试卷,可以重复多次,但如果多次搜索不到,则需要修改组卷蓝图,来调整搜索策略重新搜索。计算经验表明,PBIL 算法在试题库组卷问题中收敛速度较快且稳定,对组卷蓝图中约束条件增减的适应性非常好。

4 结 论

PBIL 作为一种概率指导的进化计算方法,用学习概率代替进化的人口(积累了一定知识的人口)在两个方面显示出其优越性。第一,它的后代产生更具方向性,因而往往能获得更快的收敛速度;第二,它产生后代的方法简单,免去了普通进化计算为选择好的生成后代的方法而进行的探索(不同的重组与变异方式产生的结果有时差异极大)。在上述组卷系统中的应用经验表明该算法快速、稳定,取得了良好的效果。

参考文献:

[1] 姚 新,陈国良,徐惠敏,等.进化计算研究进展[J].计算机学报,1995,18(9):694-706.

[2] Baluja S. Genetic Algorithms and Explicit Search Statistics [A]. Advances in Neural Information Processing System[C]. MA:MIT Press,1996.

[3] 金炳尧,蔚承建,何振亚.一个用于优化搜索的学习算法[J].软件学报,2001,12(3):448-453.

[4] 金炳尧,蔚承建,何振亚.进化算法 PBIL 在时间表问题中的应用[J].系统工程理论与实践,2000,20(5):104-108.

[5] 胡琨元,崔建江,郑秉霖,等.基于信息熵的自适应 PBIL 算法及其应用[J].系统仿真学报,2003(8):1175-1178.

[6] Sebag M,Ducoulombier A. Extending Population-Based Incremental Learning to Continuous Search Spaces[Z]. Lecture Notes in Computer Science,2003.

(上接第 79 页)

础,如何提取有效数据进行分析是关键,对原有数据源数据的提取是一个反复的过程,需要不断维护和更新。纱线质量数据仓库的建立,把品种、部门、机型等数据进行有效的集成,为企业的各层决策、分析人员使用,及时了解纱线生产状况,做出合适的决策。

参考文献:

[1] 周根贵.数据仓库与数据挖掘[M].杭州:浙江大学出版社,2004.

[2] 陈京民.数据仓库原理、设计与应用[M].北京:中国水利水电出版社,2004.

[3] 项 军,雷英杰.数据仓库技术与应用[J].计算机与现代化,2004(11):53-55.

[4] 张 宁,李强娇.基于 ERP 的企业数据仓库设计[J].计算机工程与设计,2005,26(2):351-353.

[5] 黄 容,党齐民,欧建雄.基于连锁超市的数据仓库开发模型[J].计算机与现代化,2003(2):21-23.