数据仓库技术在医院病情诊疗分析中的应用研究

周蓝桨,周肆清,杨 炼
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:分析了病案统计分析系统的现状及存在的不足。为了使病案得到充分的有效的利用,提出了将数据仓库技术应用其中的方法。以病情诊疗分析为主题,介绍了此方法的原理及主要功能,给出方法实现的具体步骤,对方法的关键部分进行了详细的解释说明。其核心思想是通过数据仓库来清洗纷繁芜杂的数据,然后利用联机分析系统独特的多维方式对数据进行分析,使用户从不同的维了解历史及现状,最后利用数据挖掘工具自动地挖掘潜在的模式,找到正确的决策。

关键词:数据仓库;病情诊疗分析;联机分析处理;数据挖掘
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2008)08-0230-03

Research and Application of Data Warehouse in Hospital for Analysis of Diagnosis of State of Illnesses

ZHOU Lan-zi, ZHOU Si-qing, YANG Lian
(School of Information Science and Engineering, Central South Univ., Changsha 410083, China)

Abstract: Analyzed the status quo and the shortage of the statistics and analysis of the disease case. In order to make disease case get fully and effectively use, a method was given which used the data warehouse technique in it. With the subject that the analysis of the diagnosis of the state of illnesses, also introduced the elements and chief function of the method, and the implementary process was given too; moreover, the pivotal parts of the method were explained amply. The central idea is to clean the complicated data through data warehouse, and then utilize uniquely multidimensional way of online analytical process to analyze data which makes user known the history and present situation from different dimensions, and finally make use of data mining to excavate potential pattern automatically and find right decision.

Key words: data warehouse; analysis of diagnosis of state of illnesses; online analytical processing; data mining

0 引 言

进入21世纪以来,随着科技的迅速发展和信息化的全球覆盖,数据库相关技术的应用已经越来越广泛。如今的病案统计分析系统也日趋成熟完善,但是大部分依旧停留在服务于OLTP的传统的数据库信息系统上,一般都只用于日常的信息操作处理,随着信息量日益丰富,使得系统存在以下几个问题。首先,系统数据量剧增,增加了查询分析复杂化,满足不了处理日常事务的速度,这样必然使得查询统计分析的速度减慢,同时历史数据也脱离了现有系统,得不到充分利用,这对统计分析是个极大的损失。其次,展现给用户的统计分析报表形式简单,如此用户只能从一个方面而不是不同的方面了解历史及现状。最后,通过统计分析的报表,用户只能了解表面现象,而对于数据之间的强大的联系却一无所知,没有给决策者提供充分有用的决策信息。

基于以上原因使得病案统计分析的利用度极其低微,冯嵩、张文君等提出了将数据仓库技术应用于医院信息管理系统中[1,2]。然而在国内数据仓库技术现正处于起步初始阶段,在病案统计分析中也并没有得到充分的应用和研究,最明显的就是在诊断、治疗中没有得到最大决策支持的应用[3,4]。所以,文中在此方面进行了一些探索。

收稿日期:2007-11-24
基金项目:湖南省自然科学基金资助项目(06JJ5131);省教育厅科研资助项目(07C388)
作者简介:周蓝桨(1985-),女,湖南长沙人,硕士研究生,研究方向为数据库技术;周肆清,副教授,研究方向为计算机应用技术及数据库技术。

1 数据仓库核心技术

1.1 数据仓库

数据仓库(Data Warehouse,DW)的概念是由数据仓库之父William H. Inmon提出的,是指一个面向主题的、集成的、时变的、非易失的数据集合,支持管理部门的决策过程[5]。数据仓库是一种管理技术,为决策

者提供各种类型的、有效的数据分析,起到决策支持的作用。数据仓库的要素包含[5]:数据的抽取、转换和装载(Extract Transform Load,ETL);数据仓库的存储;数据仓库的管理和维护。

1.2 联机分析处理

联机分析处理(Online Analytical Processing,OLAP)的概念最早是由关系数据库之父 E. F. Codd 于 1993 年提出。OLAP 主要通过多维的方式来对数据进行分析、查询和报表。它的目标是满足决策支持或多维环境特定的查询和报表需求,使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的并真实反映企业维特性的信息进行快速、一致、交互地存取,从而获得对数据的更深入了解[5]。

1.3 数据挖掘

公认的数据挖掘(Data Mining,DM)定义是指从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的知识和规则,为经营决策、市场策划、金融预测等提供依据,帮助企业的决策者调整市场策略,减少风险,做出正确的判断和决策[6]。数据挖掘是数据驱动的,系统能够根据数据本身的规律性,自动地挖掘数据潜在的模式,或通过联想,建立新的业务模型,找到正确的决策,是一种真正的知识发现方法。

2 数据仓库技术在医院病情诊疗分析中的应用方法具体实现

2.1 建立数据仓库及其模型

数据仓库软件运行环境选取:操作系统 Windows 2000 Server;数据仓库产品 Microsoft SQL Server 2005 企业版;开发程序语言 VB. NET。硬件实现环境:CPU1. 60GHz;内存 512MB;硬盘 60G。

建立数据仓库模型首先必须进行需求分析。数据仓库是按照决策分析的主题来组织数据,因此只需提取有关决策主题的那部分事务处理数据库中的数据即可。现确定主题:病情诊疗分析(根据病情所诊断出的病种分析、针对此病种的治疗方案分析以便医生实行最好的治疗手段),如此只提取与病情的诊断以及治疗手段相关的表和数据。

2.2 数据提取

做好数据仓库的核心数据库之后,下一步关键的问题就是如何根据数据仓库的要求收集并提取外界数据源中的数据,即数据的提取、净化和加载问题[7]。这一步的好坏直接影响着数据仓库中的数据质量。在此采用的方法是在 Microsoft SQL Server 2005 中用 VB. NET 编程实现 DTS 数据转换服务。首先,利用 DTS 图形向导创建 DTS 包。在目标数据源中创建表的过程中注意数据类型的转换。以 Illnesses_condition 表为例,部分代码如下:

```
create table Diagnosis_illnesses.dbo.Illnesses_condition
(Inhospital_num int null,
Charge_type char(10) null,
Illnesses_condition char(2) null,
Alergy_drugs char(100) null,
Medical_record_quality char(2) null)
```

然后,用 VB. NET 使用 DTS 包。部分代码如下:

```
Public Sub Task_Sub1(ByVal goPackage As DTS.Package2)
oTask = goPackage.Tasks.New("DTSDataPumpTask");
Public Sub oIllnesses_conditionTask1_Trans_Sub1(ByVal oIllnesses_conditionTask1 As DTS.DataPumpTask2)
oTransformation = oIllnesses_conditionTask1.Transformations.New("DT-SPump.DataPump TransformScript")
```

2.3 创建以及分析多维数据集

数据提取、净化和加载问题解决完之后,紧接的就是创建、分析多维数据集。管理关系数据以进行多维使用的最常用的方式是星型架构:将星型结构中的各个维表同事实表连接,形成一个多维数据表,然后再此基础上进行各角度的预计算,将计算结构存储形成多维数据库[8]。在此采用 SQL Server 2005 Analysis Service 中的多维数据集创建向导来完成:选择事实数据表、创建星型架构维度、创建雪花架构维度、存储多维数据集。文中以病情诊疗分析为主题的 OLAP 模型结构如图 1 所示,架构中的核心表是事实数据表,其他表是维度表。多维数据集主要是对用户当前及历史数据进行分析、辅助领导决策,它的分析方法主要有[6]:

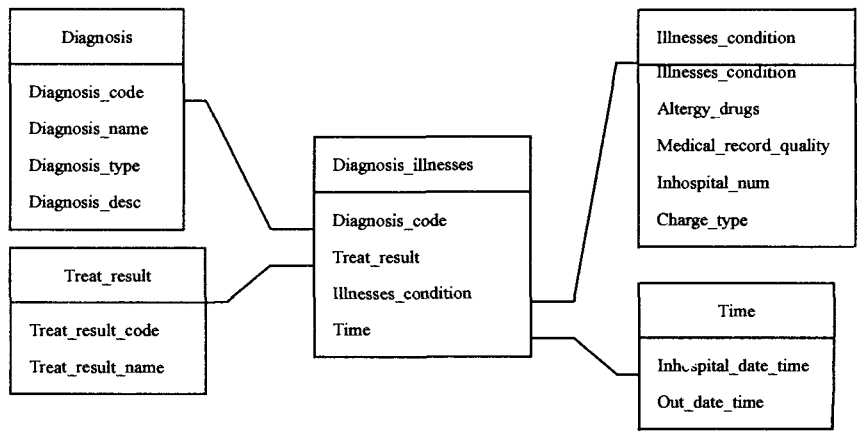

图 1　病情诊疗分析模型结构图

(1)旋转分析:旋转是一种目视操作,转动数据的视角。

(2)钻取分析:钻取是通过改变维的层次及变换分析的粒度来深化或浅化数据。它包括上钻和下钻。上钻是通过一个维的概念分层向上攀升或者通过维规约,在数据立方体上进行聚集;下钻是通过沿维的概念分层向下或引入新的维实现。

(3)切片分析:切片是在给定的数据立方体的一个维上进行选择,导致一个子方。

(4)切块分析:切块是通过对两个或者多个维进行选择,定义子方。

各种分析方法都以剖析数据为目标,这样分析人员能够从多个角度、多侧面地观察多维结构中的数据,从而深入了解包含在数据中的规则信息。图 2 给出了各种分析方法的操作。

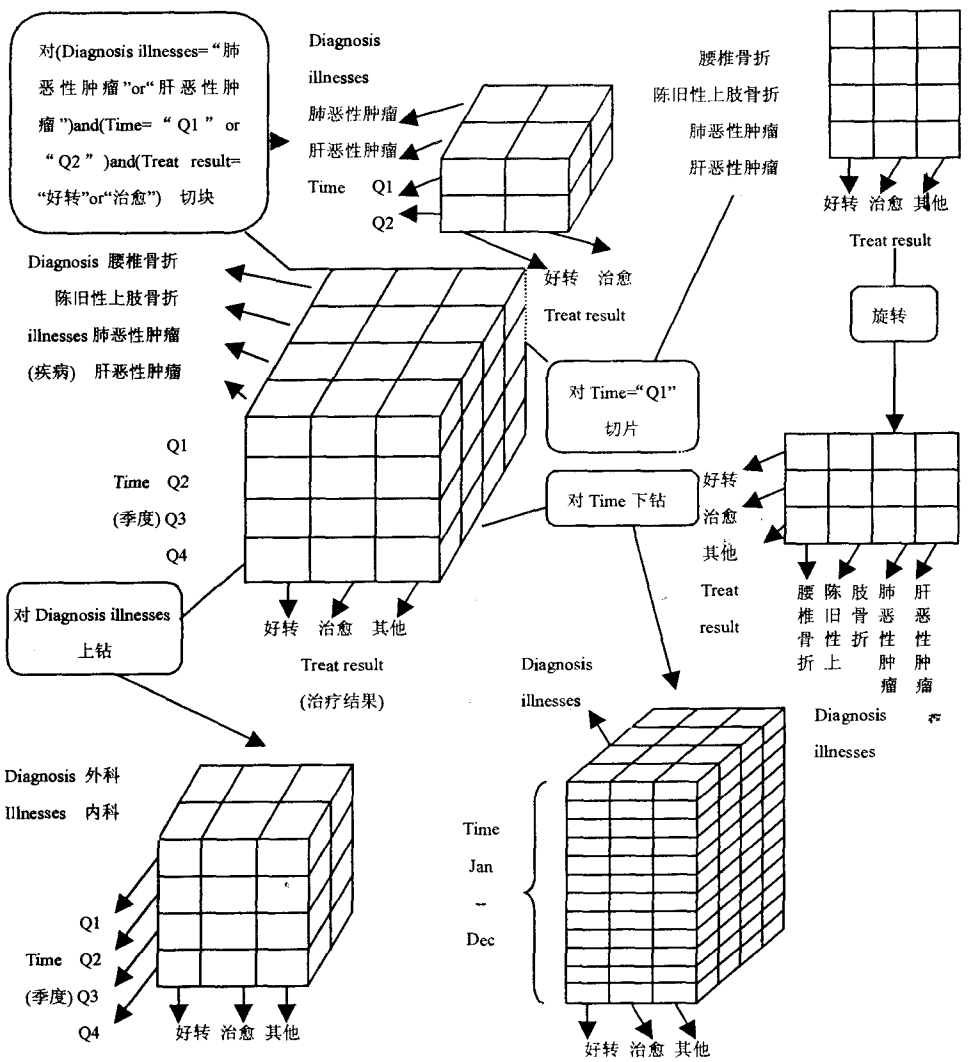

图 2 分析方法操作

通过多维数据集的生成和处理,分析人员知道了过去以及现在的情况,深入理解包含在数据中的信息及内涵。但是事务之间的潜在关系却需数据挖掘这一重要工具来完成。

2.4 数据挖掘

数据挖掘是 Analysis Services 最重要的一个新功能。数据挖掘可以筛选大量的数据,以便发现规律或者趋势。数据挖掘的方法有很多种,常用的算法有人工神经网络、遗传算法、决策树算法、最近邻技术、规则归纳和可视化等。针对文章的主题:病情诊疗分析,采用在 SQL Server 2005 中使用数据挖掘算法中的决策树算法[9,10]。决策树的基本算法是贪心算法,以自顶向下递归的方式构造决策树。该方法先根据训练子集形成决策树,它是由样本属性作为结点构成的一颗外向树,其中的非叶由判定对象属性构成,叶结点由分类属性构成。决策树自根开始按层构造,每次选取一个属性作为当时测试结点,结点选择通过信息论的信息增益的熵值作度量,选择熵值最大的属性作为当前结点。如果该树不能对所有对象给出正确的分类,那么选择一些例外加入到训练子集中,重复该建树与剪枝的过程一直到形成正确的决策树。对病情诊疗分析,叶子结点代表病情显示症状,中间一层结点代表疾病,中间二层结点代表疾病类,中间三层结点代表治疗手段,最上层结点代表治疗结果。通过对数据仓库中的病案数据进行开采,生成规则和决策树,紧接着对新的数据进行分析和预测。

3 结束语

目前在国内数据仓库技术的应用研究正处于初始阶段,随着信息化技术的发展,数据仓库将会起到越来越大的作用。文中提出了数据仓库技术在医院病情诊疗分析中的应用研究,而这只是数据仓库在医院病案资源的一方面,对其他的方面有待于更深更广的研究。

(下转第 236 页)

图 3 是重庆城市一卡通系统采用 Portal 技术的体系结构逻辑图。

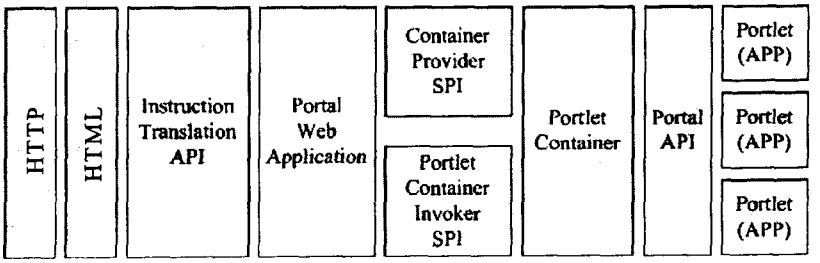

图 3 一卡通系统 Portal 体系结构的逻辑图

正是借助于 Portal 技术的这一优越特性，重庆城市一卡通系统既解决了已经建立起来的多个子系统的有效整合与集成难题，又为用户创造了一种享受“一站式服务”的全新概念，以城市一卡通系统提供的“彩票投注业务”为例，完整的业务流程所涉及的阶段为：充值→投注→支付。按照原来设计的一卡通系统，该过程需要分别访问三个不同的子系统，用户需要首先访问一卡通网站的充值页面为一卡通卡进行充值，然后返回到彩票网站页面进行投注，最后一步是彩票网调用支付子系统的支付接口完成此次投注金额的支付。而现在采用了 Portal 技术后，用户可以根据需要定制自己的客户化显示页面，即将这三个阶段通过定制访问不同业务的 Portlet 来集成到一个页面，从而在一个页面即可完成整个交易流程，真正得到“一站式服务”的初体验。而且由于在此过程中不需要用户访问不同的页面重复输入用户名和密码，从而能够有效地保证用户信息的安全性。

3 结束语

Portal 技术在重庆城市一卡通系统的成功应用，明显地缩短了新增业务的开发周期，降低了管理与维护的成本，通过相同业务的开发与管理在使用 Portal 技术前后的比较，开发周期和管理成本由原来的 5 人/月降低到现在的 2 人/月，运行维护的成本也由原来的 3 人/月减少到现在的 1 人/月。同时，Portal 技术所提倡的三大优越特性在重庆城市一卡通系统的应用中得到了更加充分的体现，为今后 Portal 技术在相似领域的应用奠定了基础。

参考文献：

[1] 王 辉. “一卡通”在数字社区中的应用[J]. 金卡工程，2003(3):14－16.

[2] Berard E V. Essay on Object－Oriented Software Engineering [M]. [s.l.]:Addison Wesley,1993.

[3] 梅尔斯. Java XML 编程指南[M]. 北京：电子工业出版社，2001.

[4] 格林沃尔德. OraclegiAS portal 宝典[M]. 北京：电子工业出版社，2002.

[5] 王 萍，李其均. 基于门户框策的资源整合系统的设计与实现[J]. 计算机应用研究，2005，22(6):162－164.

[6] 陈毓林，许舒人，朱靖宇，等. 一个 Portal 协作框架的分析与设计[J]. 计算机工程，2006，32(11):2－3.

[7] 吴 迪，陈 钢. 新一代的 Web Services 技术[J]. 计算机应用研究，2003，20(3):4－5.

(上接第 232 页)

随着医疗信息技术的不断发展和进步，每天病案数据成堆，为了更加有效地利用这些历史数据，更好地为病人服务，深层次地挖掘大量蕴藏有价值的信息，使医院决策者能及时掌握情况，制定发展目标，推动医学的科学研究，使得以决策支持为主要目的的病案统计分析系统的建设在医院病案管理系统中变得刻不容缓。

参考文献：

[1] 张文君. OLAP 技术在医院决策支持系统中的应用[J]. 医疗设备信息，2005，20(12):13－14.

[2] 杨文彬. 决策分析系统中的数据仓库技术及其应用[J]. 学术与研究，2006(2):18－19.

[3] 鹿晓明. 基于医院信息系统的多维数据分析的研究与应用[J]. 情报学报，2006，25(4):493－498.

[4] 陈雪峰，蔡 锋，钱宗才，等. 数据挖掘在恶性血液病数据库中的应用[J]. 中国血液流变杂志，2005，15(2):310－314.

[5] 王 珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1999.

[6] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001.

[7] Widom J. Research Problem in Data Warehousing[C]//Proc. of the 4th Int'l Conference on Information and Knowledge Management (CIKM). Baltimore, United States: [s. n.], 1995:25－30.

[8] Wong S T, Hoo K S Jr, Knowlton R C, et al. Design and applica－tions of a multimodality image data warehouse frame－work[J]. J Am Med Inform Assoc, 2002, 9(3):239－254.

[9] Mannila H. Data mining: machine learning, statistics and database[C]//In: Proceedings of the 8th International Conference on Scientific and Statistical Database Systems. Stockholm, Sweden: [s. n.], 1996:2－9.

[10] Milan Z, Gou M, Peter K, et al. Mining diabetes database with decision trees and association rules[C]//In: Proceedings of the 15th IEEE Symposium on Computer－Based Medical Systems(CBMS). [s.l.]:[s.n.], 2002:134－139.