基于数据库仓库与 OLAP 技术实现医院医疗数据质量监控

刘晓辉 彭传薇 李小华
（广州军区广州总医院 信息科 广东 广州 510010 flystone89@hotmail.com）

摘要：基于满足不断增长的大数据量处理的新需求，本文在医院医疗数据质量监控上引入了建立在多维数据库基础上的数据仓库技术和 OLAP 系统，大幅提升数据处理能力的同时，提供旧数据的分析、展示功能，为数据的预防性监控提供依据。本文首先详细论述了多维数据模型设计内容与建立过程， 然后以监控漏填“出生地”数据为例，通过 OLAP 系统详细分析了产生异常数据的原因。

关键词：数据仓库　OLAP 系统　多维数据模型　医疗数据质量监控

医院全面应用 HIS 系统后，医院内管理数据与医疗数据越来越多地实现了数字化。在医疗数据质量监控方面，数字化的扩大及发展，极大地提升了工作效率，并在数据完整性方面提供了快速的检测和解决方法，为医院实现大数据量、实时的数据质量监控提供了可能。然而，随着信息化的不断发展，数据量不断增加，进行监控时处理数据的时间越来越长，给信息的反馈提出了更高的要求。基于满足不断增长的大数据量处理的新需求，同是也是对现有监控方法的补充，我们在医院医疗数据质量监控上引入了建立在多维数据库基础上的数据仓库技术，通过建立多维数据模型，大幅提升数据的处理能力，建立与数据质量监控有关的 OLAP 系统，分析以往数据的变化规律，进一步探究产生问题的原因，为数据的预防性监控提供依据。

数据仓库处理数据时，数据模型的选取，直接影响到处理效率的高低。与关系型数据仓库的星形或雪花形数据模型相比，多维数据模型在多表连接数据查询方面有着较大的优势，因此我们采用具有多维处理能力的 cache 数据库作为数据处理中心，建立多维分析模式，对数据进行检测、分析和评价。

一、多维数据模型的建立

多维数据库模型在设计上包括三部分内容：一是维度确定；二是维度内层次划分；三是粒度大小的确定。

1、确定维度。

确定多维数据库中维的数目和维的内容，即多维数据库结构，是多维数据库设计的关键。本

文的设计思想是建立有限主题的多维数据库，即在同一个多维数据库内建立一定数量的主题域，根据某一个或几个相近功能的主题的自然管理方式设计结构，从数据库中抽取有关部分数据进行组织和汇总，以确保能以较小冗余满足多方面的分析需求。根据数据监控主要集中在病案首页和诊断分析上的特点，建立住院病人分析主题和病种分析等两个主题。这两个主题包含的维度主要有：科室、医生、身份、费别、职业、军种、性别、出生地、合同单位、号类、诊断类别、疾病大类、疾病中类、疾病小类等。下面以住院病人分析主题为例具体说明维度的划分情况和划分原因。

住院病人分析主题是指到医院就诊的住院病人的各种信息在一段时间内的变化和分布情况，是对医院病源情况的监控分析。它包含五个维度：科室、医生、职业、地区和时间。其中科室维又分为入院科室和出院科室两个细类。入院科室维描述了病人进入医院住院时所登记的第一个科室收治病人的监控变化情况，出院科室维描述了病人出院时所登记的科室收治病人的监控变化情况，医生维描述了全院所有医生收治病人的情况，职业维和地区维分别描述了来医院就诊病人的职业分布和地区分布的监控情况，时间维描述了医院收治病人的周期性变化情况。

2、划分维度层次

多维数据库中各维经过层次划分后，基本确定了每一个维的垂直汇总路径。当数据按照这些已经定义的汇总路径进行构造后，如果沿着其中任何一条路径自上而下地分析时，就实现了数据的掘进分析，如果将其中某一个路径与不同方向的其它不同的几条路径作任意组合时，则是对数据的面与块的切割的分析方法。

表 2 多维数据库中事实表、维度和层次结构

主题	事实表	维度	度量
住院病人分析	病人住院主记录 病人主索引 科室字典 职业字典 合同单位字典 入院方式字典	入院科室、出院科室、职业、身份、军种、费别、主治医师、性别、出生地	人数
病种分析	诊断分类记录 诊断记录 病人主记录	诊断类别、疾病大类、疾病中类、疾病小类、身份、军种、费别、	人数、费用

	疾病字典	出生地、民族	

3、划分粒度层次

多维数据库设计中要解决的一个重要问题是决定粒度层次划分，粒度层次划分适当与否直接影响到多维数据库中的数据量和所适合的查询类型。对数据进行一定程度的预综合，可以提高多维数据库的查询效率。同时为了避免多维数据库的空间膨胀，粒度的划分是必不可少的。

划分粒度的方法是:首先，估算数据仓库中数据的行数和DASD(Direct Access Storage Device)数；其次，由估算出的数据量和DASD数，决定如何划分粒度。但是，划分粒度的决定性因素并非是总的数据量，而是总的行数。这是因为对数据的存取通常是通过存储索引来实现的。

一般情况下，年数据量在超过 100, 000 行时就应该进行粒度划分，增加一个综合级别；超过一百万行时就应该考虑多重粒度的划分。从以上数据标准看，本系统采用双重或多重粒度划分效果最好。

每月月初系统将医院信息系统内的增量数据加载到多维数据库中。形成每月进行一次数据追加，年底以月为基本粒度单位进行综合汇总的数据流向。当前年度内的数据是以日为单位记录的，做为当前详细数据层，由源业务处理系统数据库中采集和集成后直接导入；一年以上数据以月为单位进行综合，成为以月为单位的数据层，用于纵向对比分析和预测；超过两年以上的数据按季度进行综合，成为季度数据层；超过五年以上的数据以年为单位进行综合，成为以年为单位的数据层。

二、数据质量监控实现

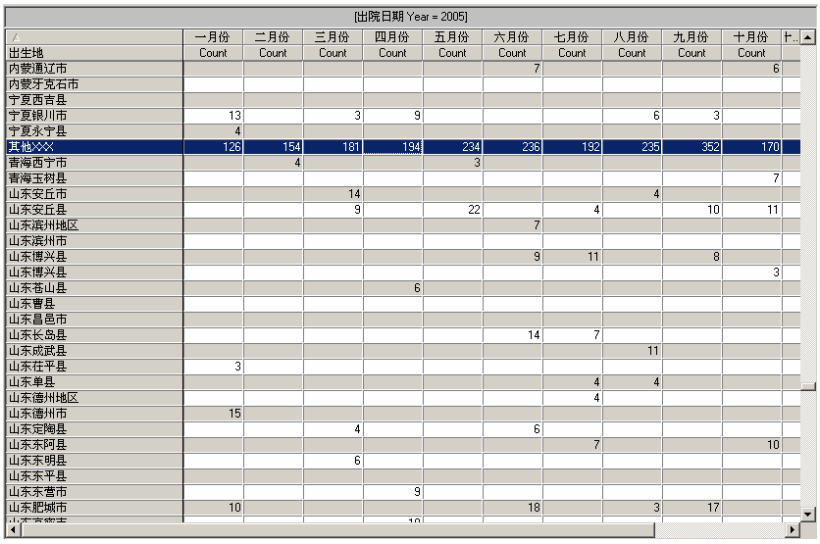
在数据质量监控的呈现上，本系统采用了数值加上图形的方式进行展示，力图使所要表达的数据既直观又准确。系统中大部分的数据都采用数值显示，有利于同一时期的排序和比较，小部分是汇总形式的数据，用图形的方式直观的展示出增减情况。

在数据的 OLAP 分析方面，本系统是一个高度开放的系统，既有事先已制定好的分析路线，又可以在现有的模型上自行设计维度组合，形成新的分析路线。

利用多维数据库技术的医院医疗数据质量监控，主要是对重要信息缺漏实现实时监控、查找缺漏信息产生的原因，并能在大样本数据的基础上实现相应项目的趋势分析。

1、实时监控重要信息的缺漏情况。我们对包含重要信息的维度按不同时间区间进行检索，把为空的项目归为“其他”类，先检索出每月或每天的漏填数量，再通过细节展示列出为空的具体项目。这种方法可以很快地查询出数据库中所有缺漏项目的具体记录，为数据实时监控和信息反馈提供了快捷有效的方法。图 1 是检索 2005 年各月漏填“出生地”数据的统计情况。图 2 列出了漏填数据的所有具体条目信息。通过导出操作将数据导出，分类汇总后，及时反馈给相应登

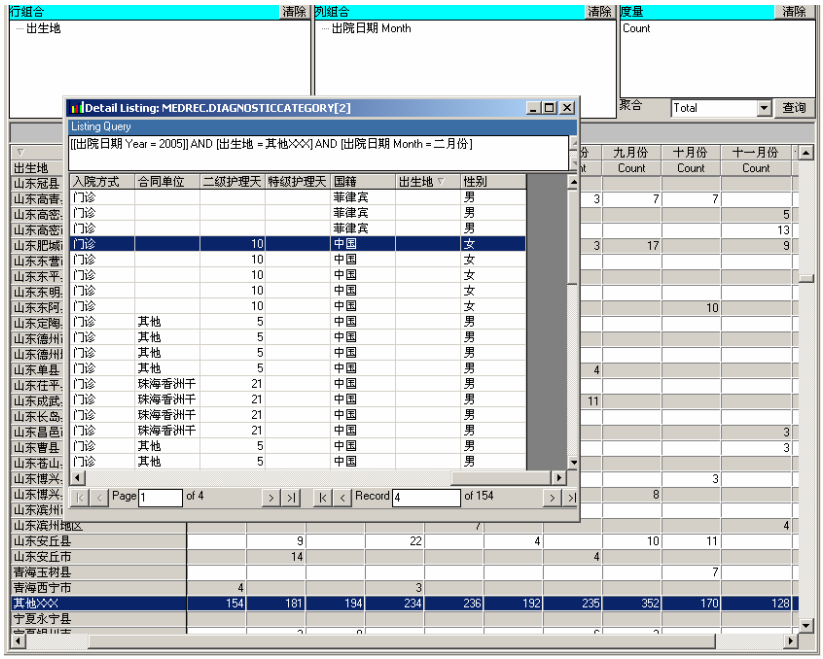
记部门补入漏填数据。



[出院日期 Year = 2005]

出生地	一月份 Count	二月份 Count	三月份 Count	四月份 Count	五月份 Count	六月份 Count	七月份 Count	八月份 Count	九月份 Count	十月份 Count
内蒙通辽市						7				6
内蒙牙克石市										
宁夏西吉县										
宁夏银川市	13		3	9				6	3	
宁夏永宁县	4									
其他XXX	126	154	181	194	234	236	192	235	352	170
青海西宁市		4			3					
青海玉树县										7
山东安丘市			14					4		
山东安丘县			9		22		4		10	11
山东滨州地区						7				
山东滨州市										
山东博兴县						9	11		8	
山东博兴县										3
山东苍山县				6						
山东曹县										
山东昌邑市										
山东长岛县						14	7			
山东成武县								11		
山东茌平县	3									
山东单县							4	4		
山东德州地区							4			
山东德州市	15									
山东定陶县			4			6				
山东东阿县							7			10
山东东明县			6							
山东东平县										
山东东营市				9						
山东肥城市	10					18		3	17	

图 1



行组合 清除 | 列组合 清除 | 度量 清除

出生地 | 出院日期 Month | Count

聚合 Total 查询

Detail Listing: MEDREC.DIAGNOSTICCATEGORY[2]

Listing Query

[(出院日期 Year = 2005)] AND [出生地 = 其他XXX] AND [出院日期 Month = 二月份]

入院方式	合同单位	二级护理天	特级护理天	国籍	出生地	性别
门诊				菲律宾		男
门诊				菲律宾		男
门诊				菲律宾		男
门诊		10		中国		女
门诊		10		中国		女
门诊		10		中国		女
门诊		10		中国		女
门诊		10		中国		女
门诊	其他	5		中国		男
门诊	其他	5		中国		男
门诊	其他	5		中国		男
门诊	其他	5		中国		男
门诊	珠海香洲干	21		中国		男
门诊	珠海香洲干	21		中国		男
门诊	珠海香洲干	21		中国		男
门诊	珠海香洲干	21		中国		男
门诊	其他	5		中国		男
门诊	其他	5		中国		男

Page 1 of 4 | Record 4 of 154

出生地	九月份 Count	十月份 Count	十一月份 Count
山东冠县			
山东高青	7	7	
山东高密			5
山东高密			13
山东肥城	17		9
山东东营			
山东东平			
山东东明			
山东东阿		10	
山东定陶			
山东德州			
山东德州			
山东单县			
山东茌平			
山东成武			
山东长岛			
山东昌邑			3
山东曹县			3
山东苍山			
山东博兴		3	
山东博兴	8		
山东滨州			
山东滨州地区			4
山东安丘县	10	11	
山东安丘市			
青海玉树县		7	
青海西宁市			
其他XXX	352	170	128
宁夏永宁县			

图 2

2、查找缺漏信息产生的原因以及趋势分析。OLAP 分析过程是一种启发式的分析过程，一个分析过程总是以提出问题而开始，随着这个问题的解决，又引发用户提出新的问题，一个接一个，用户的分析工作不断深入。OLAP 工具通过多维的方式对数据进行分析、查询和报表。多维分析是指以多维形式组织起来的数据采取切片、切块、钻取、旋转等各种分析动作，以求剖析数据，使用户能从多个角度、多侧面地观察数据库中的数据，从而深入理解包含在数据中的信息。

还是以“出生地”字段为例，把该字段为空的 2005 年各月数字画成曲线图，内容如图 3 所示。从图中可以看出“出生地”字段缺失是随着收容量的增减而变化。但九月份的缺失份数明显高于其他月份，为了查询产生的原因，我们运用 OLAP 工具作进一步的分析。首先把九月份“出生地”字段为空的病人选取出来，查询这些病人入院时间的分布情况，如图 4 所示。

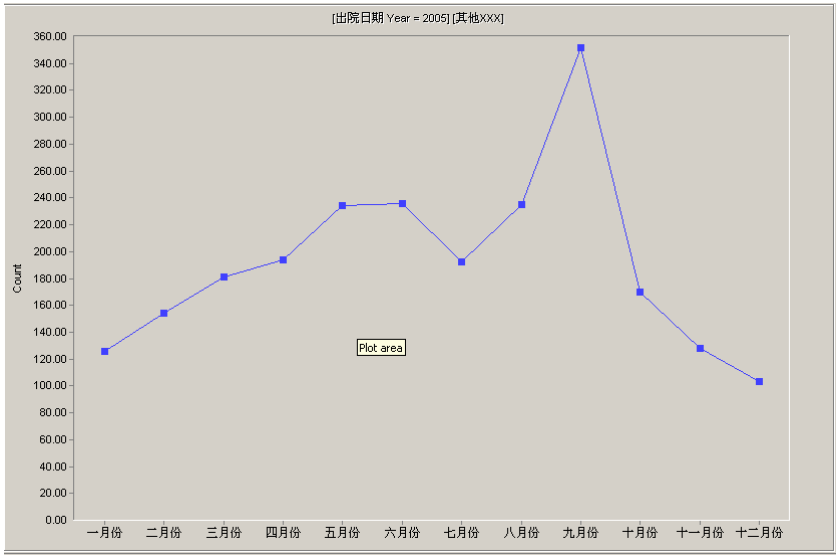


图 3

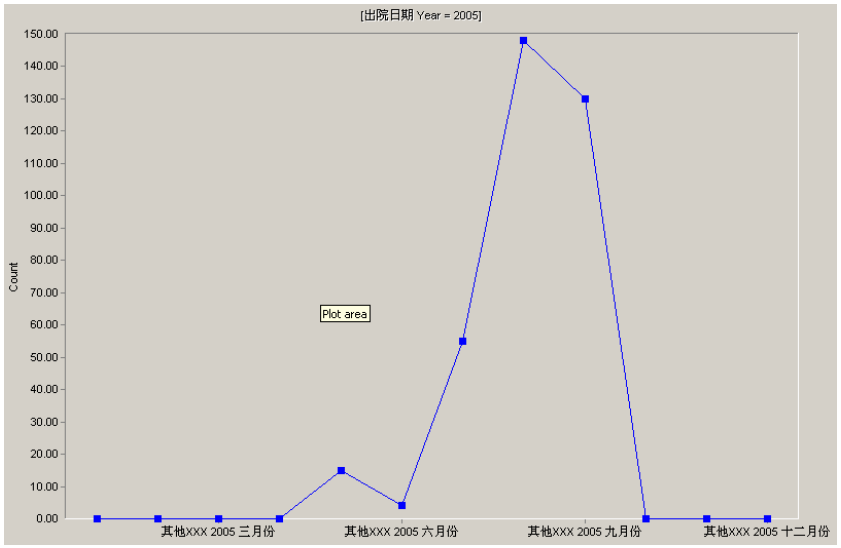


图 4

从图 4 中发现这部分病例入院时间分布在 7、8、9 三个月份，是否因为这三个月病人入院人数剧增带来漏填数据的增加呢。这三个月病人的总体入院情况分布，我们通过图 5 的数据分析了解到，7、8、9 三个月份并无大的起伏。通过其它数据也未发现其它异常情况。在现有的多维数据库里找不到数据异常增高的具体原因。后来通过讯问才知道，入院登记处 2005 年 7 月份来了一位新同志，因为不熟悉业务，造成了不少数据的漏填，从而造成了 7、8、9 这三个月数据漏填率的不正常升高。

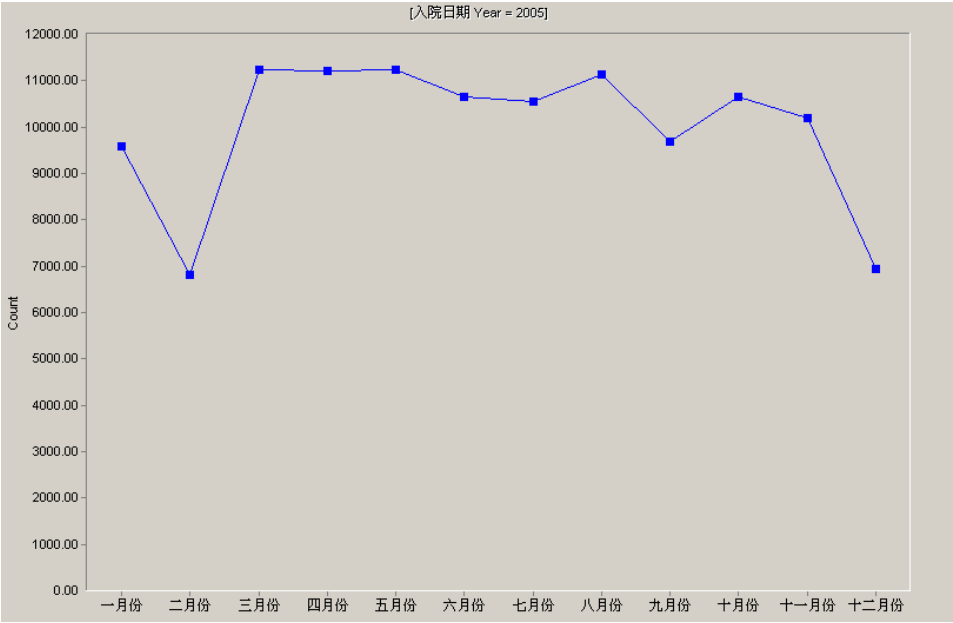


图 5

三、通过数据仓库和 OLAP 分析系统，我们既可以在多维数据库中快速查找和定位具有缺漏项的数据，还能同过各种分析手段，提前掌握各种数据的变化趋势，在数据出现错漏高峰期到来之前提早做好各方面的准备工作，预防性地进行医疗数据质量监控。建立在多维数据库技术之上的数据仓库技术以及 OLAP 等分析技术，丰富了数据质量监控的手段，为数据质量监控提供了强有力的支持。

参考文献

1. W. H. Inmon. Building the Data Warehouse，Third Edition. 北京：机械工业出版社，2003.3
2. Kimball,R. The data warehouse toolkit. John Wiley. 1996
3. 王珊著. 数据仓库技术和联机分析处理. 科学出版社，1998.5
4. 李包罗. 医院信息管理系统. http://www.chis.com.cn/show.asp?id=305
5. 陈金雄. 数据仓库与医院管理决策. 中国医院. 2004,8(10):12-4.
6. 杨光等.OLAP 技术及其发展.计算机应用研究.1999(7):7-10
7. 彭传薇,李小华,刘琛玺. 医院医疗数据质量现状和影响因素分析 中国医院管理, 2005, (09)
8. 彭传薇,刘琛玺,李小华. 浅谈医疗数据质量重要性及其影响 解放军医院管理杂志, 2005, (05) .
9. 马慧敏,王保义,许正伟. 异质数据仓库中数据质量管理研究及实现 微机发展 , 2004, (07)
10. 苏军霞. 计算机病案质量监控系统在病案质量管理中的作用 中国病案 , 2004, (08)