面向科学基金管理数据数据仓库概念模型设计

刘 权[1],肖智博[2],鲁明羽[2]
LIU Quan[1],XIAO Zhi-bo[2],LU Ming-yu[2]

1.国家自然科学基金委员会计划局,北京 100085
2.大连海事大学 信息科学技术学院,辽宁 大连 116026
1.National Natural Science Foundation of China,Beijing 100085,China
2.Department of Information Science Technology,Dalian Maritime University,Dalian,Liaoning 116026,China
E-mail:liuquan@nsfc.gov.cn

LIU Quan,XIAO Zhi-bo,LU Ming-yu.Data warehouse conceptual design toward science fund management data.Computer Engineering and Applications,2009,45(36):131-133.

Abstract: National Natural Science Foundation is playing a more and more important role to promote technological progress.For years,the NNSF has produced large amounts of data,in order to better manage and understand the data,building the data warehouse has become the necessity.This article uses object-oriented concepts model to establish a data warehouse toward NNSF management data,and through the establishment of the data warehouse and a visualization system toward NNSF data,the model is validated.

Key words: Science Fund Management Data;data warehouse;object-oriented conceptual design

摘 要:国家自然科学基金委员会在促进科技进步方面正发挥着越来越多的作用。多年来,国家自然科学基金委员会产生大量的数据,为便于更好地管理和理解数据,建立数据仓库就非常有必要。利用面向对象的概念模型建立了国家自然科学基金委员会管理数据数据仓库,并通过数据仓库的建立以及面向科学基金管理数据的展示系统的建立验证了模型。

关键词:科学基金管理数据;数据仓库;面向对象的概念结构设计
DOI:10.3778/j.issn.1002-8331.2009.36.039 **文章编号:**1002-8331(2009)36-0131-03 **文献标识码:**A **中图分类号:**TP311

1 引言

随着国家对科技扶持力度的不断加大,科学基金数据量与日剧增。在中国,国家自然科学基金委员会(NSF)平均每年接受超过 50 000 份的申请,8 000 到 10 000 个项目获批, 约有 5 000 个项目需要结题[1]。在 2008 年美国 NSF 收到 200 000 个申请,预计今年会增长到 275 000 个[2]。如此庞大的数据量对传统的科学基金数据管理(以数据库技术为核心)提出了一系列新的挑战。

数据仓库技术[3],作为一种有效的海量数据管理方法,在各个领域发挥着重要作用[4-6]。概念模型是数据仓库的灵魂。截止目前, 人们普遍采用四种概念模型来设计和开发数据仓库:StarER 模型[7]、多维 ER 模型、维度事实模型和面向对象多维模型[8-9]。

然而,科学基金数据不完全等同于其他领域的数据,有其自身特点:(1)时间跨度性:一项工作通常跨越了一定的时间,从几个小时到数年不等。在时间跨度里,某些项目的相关属性值可能会随着时间而改变, 从而可能会导致其他属性值的改变,进而为维护一个属性在其他维度的变化造成了困难。(2)生命周期性:一个项目通常包括若干个生命周期,而几个生命周期之间的数据便有了固定的流向。这些特性导致了当前主流的概念模型设计方法不适用于科学基金数据的管理。概念设计的主要目的是在用户需求的基础上对数据进行一个正规的、完整的抽象化设计。这种抽象表示还必须是正规的、完整的,对接下来的逻辑结构设计不产生任何歧义的[10]。而上述四种方法对于时间跨度性和生命周期性均不能进行很好的描述。

鉴于此,提出了一种新的概念模型设计方法,该模型在传统概念模型的基础上考虑了时间跨度性和生命周期性,以适应对科学基金数据更好的描述和建模。

2 面向科学基金管理数据的数据仓库概念结构设计

2.1 双重粒度

粒度级别的设定会影响到数据仓库中数据量的大小和数据仓库所能回答的查询类型。尽管大部分的数据仓库查询和数据挖掘可以在已经被压缩的、存取效率更高的轻度综合的数据上进行, 但是考虑到如果需要分析更低一级别的细节数据,可能会涉及到数据的真实档案层。在真实档案层上,访问数据是

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.J0724003)。
作者简介:刘权(1966-),男,博士,研究方向为 GIS 应用;鲁明羽(1963-),通讯作者,男,博士,博士生导师,研究方向为数据仓库与数据挖掘;肖智博(1984-),男,博士,研究方向为数据仓库与数据挖掘。
收稿日期:2009-09-02 **修回日期:**2009-11-05

代价昂贵的事情,但是考虑到未来数据挖掘的需求,如果必须进入这一级别,双重粒度是必须的选择[3]。双重粒度级别可以保证数据仓库中包含整个基金委活动数据和历史数据。

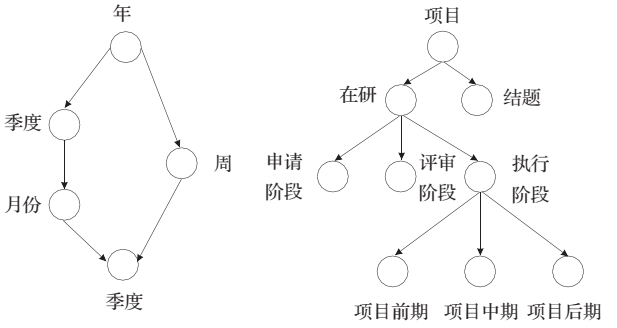

图1 双重粒度

2.2 多对多关系

图2中显示了一个事实表有 n 类维度,也体现了整个模型的思想。维度类的最小基数被定义为1,来表示一个事实对象实例总是和其他维度的事实对象实例相关联。事实类角色的基数被定义为*,来表示一个维度对象可以是一个、零个,或者是多个事实对象实例[9]。

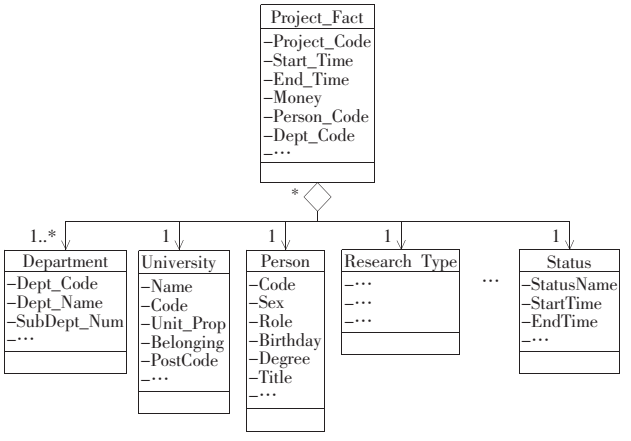

图2 多对多关系

对于图2来说,事实表和维表之间的关系可以通过数字和它们之间的连线体现。多对多关系中,在维表用1..*表明这个事实表能与一个或多个维表相关。例如,一个项目可以与一个或多个部门相关,例如一个项目具有双重信息既可以属于化学部也可以是信息部。

2.3 扩展度量

局和函数可以用来作为扩展度量的方法。比如,对于在科学基金管理中两个比较重要的参数指标:资助率和资助强度,可以定义如下:

$$资助率=\frac{COUNT(批准项目)}{COUNT(申请项目)}*100\%$$

$$资助强度=\frac{SUM(批准金额)}{COUNT(批准项目)}$$

对于OLAP操作,扩展度量和原始定义的度量值是没有差别的,同样可以对其进行累加等操作。

2.4 累加性和OLAP操作

在模型中所定义的度量都是可以累加的,也就是说所有的SUM操作可以应用在聚合操作的所有维表中。但是上述已经定义好的扩展度量由于本身已经使用了聚合函数它们就是非累加性的。而在现实中,累加不同学校级别的支持强度是没有意义的。

众所周知,OLAP能够提高数据仓库的分析能力。考虑到用户的需要,可以采用一系列OLAP操作对多维数据视图进行数据分析。

2.5 约束性和完整性

约束性是指一个叶子节点仅仅属于一个根节点。完整性是指根节点仅包括叶子节点。图3中显示的是具有约束性和完整性的分类等级。沿着中间线可以看出,一个省份拥有多座城市,一座城市拥有多所大学,而一所大学仅属于一座城市,一座城市仅属于一个省份。通过这种方式,建立约束性和完整性的等级。为了避免数据库框架自动从UML类表到关系OLAP工具的混乱,给每个等级一个属性来区分。

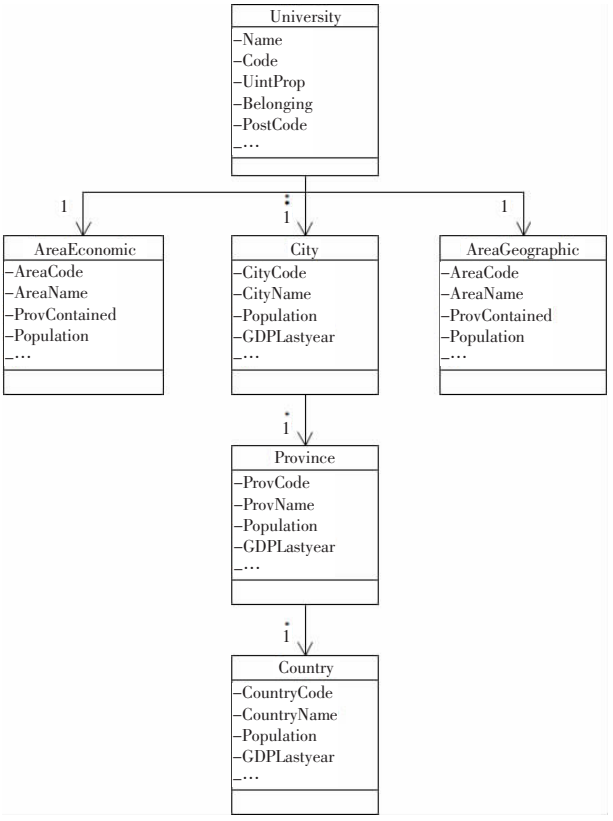

图3 包含完整性、约束性和多类路径等级的多分类层次

2.6 分类等级

图4中定义了多类路径等级。多类路径等级被用作一个对象用不同的分类标准分到不同的类别的情形,就像是中国可以被认为是东亚的国家,也可以认为是北半球的国家。

图中考虑的是维度的类和附属类。项目中,部门维表有7个不同的附属维表,为了简化,图中仅仅画了其中的2个。对于每个附属维表,它们都代表了不同的学科。

3 结果与讨论

ADOMD.NET是对ADO.NET在多维数据集上的扩展,是专门面向Analysis Services数据库的高级编程接口。该文采用了ADOMD.NET作为多维数据集与前段展示系统的通信桥梁,通过MDX表达式进行数据查询、读取多维数据集中的数据,将数据传送到DevExpress展示控件进行展示。

相比较于传统的建模方法而言,维度事实模型不能支持维度间的衍生度量,分类层次以及维度的泛化/特化;StarER模型

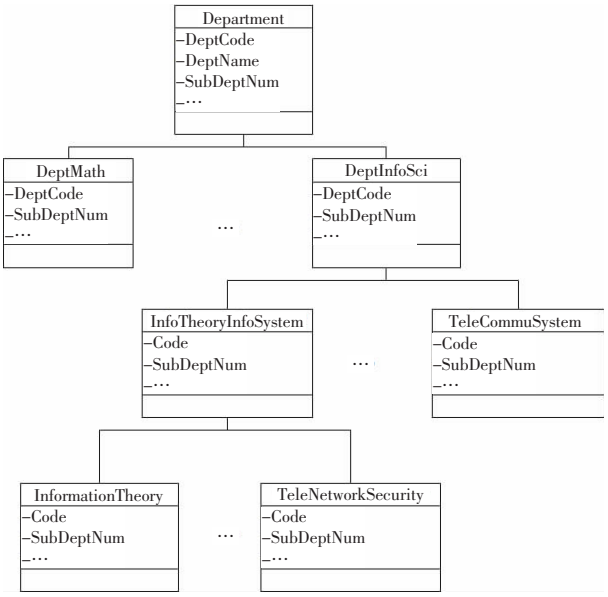

图 4　学部维度及其不同的子维度

除了不支持衍生度量之外，也不支持建模工具；多维 ER 模型不支持度量的可加性，对于衍生维度和多对多关系支持也不够好。最为重要的是，包括面向对象多维模型，以上的模型都不能很好地描述面向科学基金管理数据的时间跨度性和生命周期性。

在第二部分的基础上，应用展示控件得到了一些新颖直观的展示图，可以看到，图 5~图 7 直观地对自然科学基金的各种情况进行分析。

图 5 为饼形图，展示了全国各个省市自然基金申请金额占全国总申请金额的百分比中前十位的省市。可以看出北京市和上海市在申请金额上一枝独秀，申请金额约占全国总量的一半。

图 5　柱形图比较申请金额

图 6 为堆叠面积图选取了不同地域的几个比较有代表性的省市进行比较，并且将不同学部的信息放在一起进行比较。从图中不仅可以看出不同省市的批准金额上的差别，也可以看出不同学部在这个省份的科研分布状况。

图 7 为柱形图和线形图的叠加图。图中在批准项目数和批准金额两个项目上，将全国前十位的省份进行比较，同时又将这十个省份的数额同除掉这十个省市之外的全国其他地区总和进行比较。从两张图中可以看出全国各省份中，批准项目数和批准金额前十位省份的位置都没有发生变化。而且无论是柱形图还是线形图，都可以看出北京市都遥遥领先于其他省份，甚至超过了其他地区的总和。

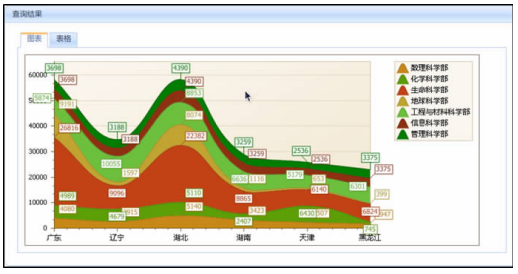

图 6　堆叠面积图比较省份在学部的申请状况

图 7　柱形图与线形图叠加比较

4　总结

该文面向科学基金管理数据，针对科学基金管理数据的特征，改进了面向对象多维概念模型，具体阐述了对概念结构的设计。最后通过面向科学基金管理数据的数据仓库建立，并使用了 DevExpress 控件作为展示媒介，建立了一套基于数据仓库的面向科学基金数据的展示系统，验证了模型的有效性。

参考文献：

[1] 国家自然科学基金委员会 2009 统计数据[EB/OL].[2009-09-10]. http://www.nsfc.gov.cn/nsfc2009/index.htm.

[2] Kwok R.US agencies brace for flood of grant applications[EB/OL]. [2009-03-16].http://www.nature.com/news/2009/090316/full/news.2009.167.html.

[3] Inmon W.Building the data warehouse[M].[S.l.]:John Wiley & Sons, 1996.

[4] Rai A,Dubey V,Chaturvedi K,et al.Design and development of datamart for animal resources[J].Computers and Electronics in Agriculture,2008,64(2):111-119.

[5] 侯亚荣，万雅奇，张书杰.教育考试数据挖掘的研究与实现[J].计算机工程与应用，2008，44(16)：132-134.

[6] 丁建华，彭政，王飞.生物数据仓库研究及应用[J].计算机工程与应用，2005，41(12)：192-194.

[7] Tryfona N,Busborg F,Christiansen J G.StarER:A conceptual model for data warehouse design[C]//ACM 2nd International Workshop Data Warehousing and OLAP.Kansas City,USA:ACM,1999:3-8.

[8] Gaede V,Gnther O.Multidimensional access methods[R].ACM Computing Surveys,1997.

[9] Carreira P,Galhardas H,Lopes A,et al.One-to-many data transformations through data mappers[J].Data & Knowledge Engineering, 2007,62(3):483-503.

[10] Phipps C,Davis K.Automating data warehouse conceptual schema design and evaluation[C]//DMDW,Toronto,Canada,2002:23-32.