基于语义分析的统计报表多维数据建模方法

沈 晖 陆育锋 郭以东 杨 勇
(中国石油勘探开发研究院西北分院计算机技术研究所 兰州 730020)

摘 要 在统计类应用系统实施过程中,依据前台报表的样式和填报要求设计后台数据库表结构是一项重要的基础性工作,但往往凭借的是系统设计人员的经验,被看作是一项艺术而不是有章可循的技术,直接导致的结果是:针对同一套业务统计报表,不同的系统设计人员可能设计出千差万别的后台数据库表,带来系统运行性能与日后运维升级等方面的问题。为了改变这种状况,对现实中基于业务统计报表设计后台数据模型的过程进行了深入研究,总结提出一种基于语义分析的从统计报表集合抽取统一多维数据模型的规范化、程序化方法,并在某大型国企信息系统项目中进行了应用验证。

关键词 统计报表,多维数据建模,语义分析,计算机辅助软件工程(CASE)
中图法分类号 TP391 **文献标识码** A

Multiple Dimension Data Modeling Method for Statistical Data Sheets Based on Semantic Analysis

SHEN Hui LU Yu-feng GUO Yi-dong YANG Yong
(Computer Technology R&D Unit, Research Institute of Petroleum E&P-Northwest, Petrochina, Lanzhou 730020, China)

Abstract For implementing statistical analysis applications, it is the most fundamental task to design back-end database to support front-end data sheets processing, including data capturing, calculation and presentation. It always depends on technical designer's personal experiences instead of rule-based step-by-step procedure and has been regarded as "art" instead of "scientific technology". A risky result is that different technical designers might create different data models for one same set of statistical data sheets pre-defined by business end-users, which can cause big troubles both for system performance and on-going application maintenance. Through in-depth study on real case to summarize business rules, a formal method based on semantic analysis was proposed to support creating unified multiple dimension data model from pre-defined statistical data sheets and applied in a large SOE information system project as PoC(Proof-of-Concept).

Keywords Statistical data sheets, Multiple dimension data modeling, Semantic analysis, Computer aided software engineering(CASE)

1 引言

统计是一切决策分析的基础,而统计的结果是通过一系列具备业务含义的统计报表体现的。从技术上来看,各企业的统计类应用系统基本都按照"原始数据采集→KPI指标计算→统计报表生成"的数据流来实现,看起来再简单不过。然而在现实应用中,不同系统尽管表面上看起来在前台功能与报表展示效果方面相差无几,但在操作便捷性与运行效率等方面的用户体验可能差别巨大。这方面的差别主要体现在后台数据模型的设计上。统计报表多维数据建模的主要困难在于业务与技术的脱节。在国企,统计分析需求往往由业务人员自上而下提出,即统计报表由业务人员设计并作为标准发布,而业务人员并不熟悉多维数据模型的概念,也就不可能按照指标和维度的分类原则科学设计业务报表,导致现行报表体系先天具有复杂多样、灵活多变的特点;而技术人员又往往只关注技术实现,对业务理解并不深入,也就很难在设计后台数据库表时有效地将现行各类复杂统计报表翻译为一套不重不漏的事实表和维度表集合。为改善这种局面,有必要对后台数据模型进行重构,形成一套支持各类统计报表填报与统计数据综合分析的统一数据模型。归纳起来,统一数据模型的设计目标包括:(1)"不重、不漏"地体现所有统计报表(包括不同层级定义的多套报表体系、同一套报表体系中包含的各类报表)所需填报的数据项;(2)支持面向基层填报人员的统计数据填报优化(基于统一数据模型设计基础数据采集表),确保基层填报人员对于所需填报的所有统计数据项"一次填报";(3)减少不同技术设计人员的个人经验对于设计结果的影响,确保设计出的数据模型是"统一"的,即模型完全基于业务规则生成,一旦业务规则确定并进行了标准化定义,不同技

到稿日期:2012-11-04 返修日期:2013-02-23 本文受中国石油天然气集团公司"十二五"信息化重点项目(E7)资助。
沈 晖(1974—),男,博士,主要研究方向为企业信息化规划与系统实施、商务智能与决策支持系统,E-mail:shenhui@tsinghua.org.cn;陆育锋(1962—),男,高级工程师,主要研究方向为企业信息化系统集成与实施、计算机应用;郭以东(1969—),男,高级工程师,主要研究方向为集团公司信息化建设实施、网络管理及业务决策支持系统;杨 勇(1982—),男,博士,工程师,主要研究方向为软件设计开发。

术设计人员即可按照一套标准化方法流程设计出一样的后台数据模型。

本文对现实中基于业务统计报表设计后台数据模型的过程进行了深入研究，总结提出了一种基于业务语义分析的从企业统计报表集合中抽象形成统一多维数据模型的规范化、程序化方法，并在某大型国企能源管理系统统计模块的需求分析与系统设计工作中进行了应用验证。

2 面向多维数据建模的统计报表数据项分析

2.1 相关研究

不同于事务处理类应用系统(OLTP)基于简单业务表单设计的后台关系型数据建模，统计分析类应用系统(OLAP)的后台数据模型设计(对应于数据仓库 DW 或统计数据库 SDB)要灵活、复杂得多。在多维数据建模和语义分析方面，理论上已提出基于事实表与维度表组合的星型-雪花型多维数据模型，并将其用于 OLAP 数据建模[1,2]。Egozi 等人研究了基于语义分析的信息抽取[3]。陈竞波等人提出一种基于语义的报表系统模型，灵活地分离并绑定报表数据和报表结构[4]。刘庆伟等人在关系数据库业务模型分析的基础上建立多维模型定义，使数据具有多维的概念[5]。杨小献等人提出了基于规则的柔性综合统计报表[6]。但是这些研究的核心是统计报表本身或是最终的数据模型，而现实的情况是企业的报表表样已经确定，这些研究无法指导从企业统计报表集合中抽象形成统一多维数据模型。

2.2 业务统计 KPI 指标与多维数据模型指标在概念上的联系与区别

一般来说，业务上定义的统计 KPI 指标是个业务概念，而多维数据模型是个技术概念，两者同样涉及指标，但含义不尽相同。举例来说，在能源统计 KPI 指标体系中，“原油替代量(原煤)”是一个指标，其业务含义是“以原煤替代原油的量(将替代原油用的原煤折合成原油的量)”；而将其翻译成严格的多维数据模型，则将“能源替代量”作为指标，“原油”和“原煤”可作为两个不同维度“被替代能源种类”和“替代能源种类”所取的维值，将这两个维度与指标进行组合，即构成业务意义上的统计 KPI 指标。在这一例子中，两者的对应关系如图 1 所示。

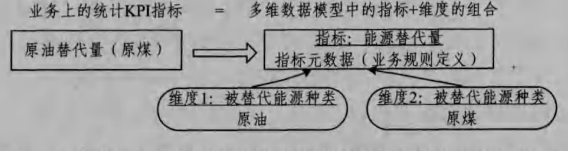

图 1 业务统计 KPI 指标与多维数据模型指标之间的对应关系示例

对于统计类应用系统来说，将业务上的统计 KPI 指标翻译为技术上由指标和维度组合形成的多维数据模型是极具现实意义的，这意味着前台应用可以和后台数据库表结构相分离，从而为日后系统运维与升级带来方便[7]。以图 1 为例，业务上未来可能会增加新的统计 KPI 指标如“原油替代量(电)”、“原煤替代量(天然气)”等，如果后台数据模型按照图 1 右所示方式设计，则在支持这些新的业务需求时不需要对后台数据库表结构进行任何调整；而如果当初设计时没有对其进行指标和维度的拆分，则势必需要对相应的数据表结构进行修改，日后维护工作量大并且极易出现错误。

因此，按照多维数据模型概念设计统计报表后台统一数据模型是本文立论的原则，而基于多维数据模型概念对统计报表中的各数据项(即报表中主栏行和宾栏列对应的空格)进行指标和维度的拆分是其中最为基础性的工作。

2.3 基于语义分析的统计报表数据项分析过程与结果示例

我们以一张统计报表作为研究对象，并选取表中的一个数据项进行面向多维数据建模的深入分析，其过程与结果如图 2 所示。

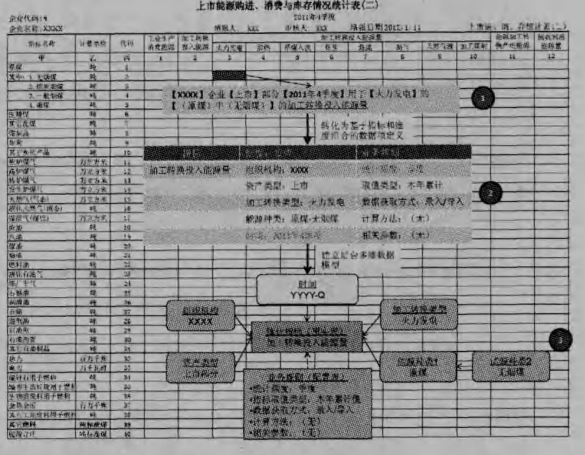

图 2 统计报表数据项分析的过程与结果示例

整个分析过程分为 3 步：

(1)数据项语义分析：对数据项的业务含义按照严格的名词偏正词组语法规则进行精确定义，即描述成多个定语＋名词性主语的组合。如图 2 所示，对该数据项的精确定义为：“XXXX 企业上市部分 2011 年 4 季度用于火力发电的原煤中无烟煤的加工转换投入能源量”，对这个名词偏正词组进行语义分析，可分解为 5 个定语修饰 1 个名词性主语，其中：

a)定语：【XXXX】企业，【上市】部分，【2011 年 4 季度】，用于【火力发电】，【原煤中无烟煤】。其中，有 1 个定语自身又存在嵌套分类关系——【原煤中无烟煤】：原煤-无烟煤(无烟煤属于原煤的一种)。

b)主语：加工转换投入能源量(“加工转换”和“投入”可理解为“能源量”的定语，但考虑到“加工转换投入”实际上体现了此数据项的统计主题，因此不再拆分)。

(2)数据项形式化表述：根据数据项语义分析的结果，将主语定义为指标，而将 5 个定语定义为 5 个维度的维值(时间被看作一个特殊维度)，同时从该数据项的填报说明中提取相应的业务规则——元数据(包括统计频度、取值类型、数据获取方式、计算方法、相关参数等)，则形成图 2 所示的以结构化表格形式定义的该数据项数据字典。

(3)多维数据模型生成：根据步骤(2)的结果，可直接转换成为后台星型-雪花型结构的多维数据库表，并在维度表中插入维值记录，包括：

a)事实表(1 张)：加工转换投入能源量；

b)维度表(4 张，时间是一种维度，但不作为维度表)：组织机构、资产类型、加工转换类型、能源种类，其中，能源种类这张维度表存在维值嵌套情况，需要对维度表进一步拆分，或

者通过设定维值 ID 的编码规则体现出多个维值间的分类隶属关系(推荐采用后一种方案,例如:设定原煤的 ID 为"01",无烟煤的 ID 为"0101",并从编码规则上定义"0101"隶属于"01");

c)配置表(元数据):定义事实表指标取值的统计频度、取值类型、数据获取方式、计算方法、相关参数等规则类信息。

从上述分析可以得出结论:以多维数据模型的视角来看,任何一个统计数据项都可以规范化定义为指标+维度+业务规则(元数据)的组合。

2.4 针对数据项分析过程的形式化算法描述

进一步将上述从数据项的语义分析入手在业务人员的协助下推导生成多维数据模型的过程进行形式化描述,有助于将这一有章可循的操作流程在计算机中进行程序化实现。

这一过程翻译成伪代码的算法描述如下:

算法:DataItemAnalysis. 通过数据项语义分析生成多维数据库表

预定义:#Predefined(m,r)

- m(元数据表 metadata):预定义表结构及数据记录,包括 5 张表:
  m1(统计频率):m1 [m1_ID as PRIMARY KEY,m1_value]
  m2(取值类型):m2 [m2_ID as PRIMARY KEY,m2_value]
  m3(数据获取方式):m3 [m3_ID as PRIMARY KEY,m3_value]
  m4(计算方法):m4 [m4_ID as PRIMARY KEY,m4_value]
  m5(相关参数):m5 [m5_ID as PRIMARY KEY,m5_value]
- r(配置表 rule):仅预定义表结构,不插入数据记录(空表)
  r [f_ID,m1_ID,m2_ID,m3_ID,m4_ID,m5_ID as PRIMARY KEY];
  f_ID:表示指标 ID(指标英文标识)
  m1_ID,m2_ID,m3_ID,m4_ID,m5_ID:表示对应元数据表 m1,…,m5 的业务规则 ID

输入:

- f:指标定义
- d[i]:维度定义,i=1,…,N,N 为数据项对应的维度总数
- dv[i][j]:维度 i 的维值,j=1,…,P,P 为维度 i 对应的维值数
- r[k]:业务规则定义,k=1,…,5,对应元数据表 m1,…,m5

输出:多维数据库表结构及数据记录

- 维度表 d[i]:新建表结构,插入维值记录 d[i][j]
- 事实表 f:新建/修改表结构
- 配置表 r:插入对应事实表的业务规则记录

方法:

```
Procedure DataItemAnalysis(f,d[i],dv[i][j],r[k])
{
(1)input f [指标中文名,指标英文标识,指标取值数据类型],d[i](i=1,…,N)[维度 i 中文名,维度 i 英文标识,维度 i 取值数据类型],dv[i][j](i=1,…,N;j=1,…,P)[对应维度 i 的维值],r[k](k=1,…,5)[统计频度,取值类型,数据获取方式,计算方法,相关参数] & Submit;//业务人员通过人机交互界面在前台录入数据项的形式化表述(英文标识默认为中文名的拼音缩写,可在前台修改以避免重名)
   //后台处理维度表及相应维值(新建维度表/插入新维值)
(2)for(i=1 to N)
(3)  {create table d[i]_temp(ID AUTO_INCREMENT PRIMARY KEY,d[i]_value);//在 DB 中创建维度表 d[i]_temp(临时表):d[i]为维度 i 英文标识
(4)  if table d[i] exist in DB then //在 DB 中判断维度表 d[i]是否已经存在
(5)    drop table d[i]_temp;//如果 d[i]在 DB 中已存在,删除临时表
(6)  else
(7)    alter table rename d[i]_temp d[i];//如果 d[i]在 DB 中不存在,将临时表改名为正式表
(8)  for(j=1 to P)
(9)    insert table d[i](ID,dv[i][j]);//在维度表 d[i]中插入维值:ID 根据编码规则自动编号
(10) }
   //后台处理事实表(新建事实表/修改事实表——增加新维度)
(11)if table f not exist in DB then //在 DB 中判断事实表 f(指标英文标识)是否已经存在
(12)   create table f(ID AUTO_INCREMENT PRIMARY KEY,f_value);//如果 f 在 DB 中不存在,创建新的事实表 f
(13)else //如果 f 在 DB 中已存在,判断是否需要为 f 增加新维度
(14)  {for(i=1 to N)
(15)    {if table f not exist column d[i]_ID then //在 DB 中判断事实表 f 中是否存在列 d[i]_ID(维度 d[i]的英文标识)
(16)      alter table f add d[i]_ID PRIMARY KEY;//如果 f 中不存在 d[i]_ID,添加维度 d[i]_ID 并作为 f 的主键
(17)    }
(18)  }
   //后台处理配置表(插入新规则——事实表对应业务规则)
(19)for(k=1 to 5)
(20)  Insert table r(f_ID,r[k]_value);//在配置表中添加业务规则(事实表 ID、对应业务规则 ID 的业务规则取值)
}
```

为了验证算法的可行性,我们在.NET+MySQL 环境下进行了计算机辅助软件工程(CASE)工具[5]原型的编码实现,并以 2.2 节的示例内容作为算例进行了功能测试。

3 完整流程和应用验证

3.1 完整流程

2.3 节算法的主要作用是对单张统计报表的单一数据项进行语义分析,从而提炼出该数据项的后台多维数据模型。在实际的统计分析工作中涉及到的报表不止一张,数据项更是成百上千乃至上万,只有对所有报表中的所有数据项都依照上述算法进行操作后,才能形成一套完整的统一多维数据模型。

我们将对从企业统计报表全集(多张报表)中抽取统一多维数据模型的完整过程加以描述,即形成方法的端到端流程图,如图 3 所示。

作为端到端方法应用的入口,业务人员首先需要对现行统计制度所要求的各类统计报表及报告进行汇整,形成一套结构化的统计报表集合(Excel 表格);然后在人机交互下,依次对每张统计报表中的每个数据项调用 2.3 节给出的算法程序进行操作;最终在遍历完所有统计报表中的所有数据项后,形成针对所需报送的每项统计数据严格定义的标准化数据字典,并自动生成基于多维数据模型的"不重不漏"、结构优化的后台统一数据库表结构。

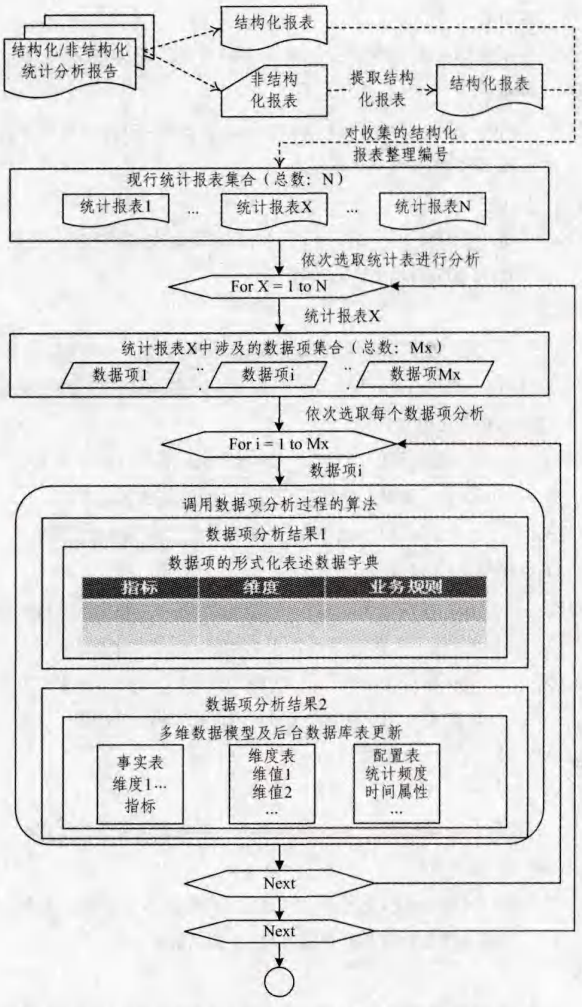

图 3　从统计报表集合抽取统一多维数据模型的端到端流程

3.2　实例验证

下面以某大型国企能源管理系统项目为例对上述流程进行验证。该项目旨在对该企业的节能降耗业务进行统一管理，关、停、并、转现行各类相关应用系统，形成全集团范围内使用“一套系统”管理节能降耗业务的局面。其中统计分析是最为基础、用户量最大并且使用频率最高的功能模块。为了有效验证此方法的应用效果，需要为项目设定一些假设条件，以便以共同的基准进行方法应用前后的对比分析。在本项目启动之前，已经有一个现行系统在运行，其统计子系统前、后台设计的依据就是集团统一定义的 56 张统计报表；本文将现行统计报表体系作为事实标准，假定其在项目实施过程中不作任何修订。

在项目中严格按照图 3 所示的方法流程并依托开发的 CASE 工具原型，由业务人员参与对统计模块进行了从前台业务报表到后台数据模型的分析设计，过程和主要结果归纳如下：

• 现行统计报表体系的数据项统计结果（报表中主栏行与宾栏列相交的空格数）：共计 7259 个需要报送的数据项，其中：录入项 2920 个，计算项 2141 个，其他为参数项（计算所需的单价、折算系数等参数）和带入项（为进行同比、环比等比较分析而从历史记录或其他表中直接取得数据，无需录入和计算）。

• 针对统计报表数据项的分析结果（指标＋维度＋业务规则）：所有数据项共涉及：

指标：98 个，其中：录入指标 61 个，计算指标 37 个；

维度（体现统计对象的分类）：13 类共 26 个（有 3 个大类维度存在嵌套维度，可派生出多个维度）；

业务规则（元数据）：(1) 统计频度 3 种，(2) 取值类型 3 类，(3) 数据获取方式 3 类，(4) 计算方法 8 种，(5) 相关参数 4 类。

• 依据数据项分析结果，设计后台多维数据模型：

事实表：58 张；

维度表：26 张；

配置表：1 张（合并 5 类元数据）。

以多维数据模型的设计结果对现行统计报表中的所有录入项进行填报优化，进一步设计出 37 张数据采集表，可以覆盖现行 56 张统计报表中所有 2920 个录入项，并且确保相同数据仅录入一次。

与现行系统相比较，在新系统中所需填报的录入项共计 757 个，这意味着基层填报人员的录入工作量能够减少 70% 以上。同时，现行系统中的后台数据模型采用的是与前台统计报表完全一致的设计，而新系统中的后台数据模型则是结构优化的多维数据模型，这意味着系统性能、可维护性与可扩展性方面质的飞跃。

结束语　通过探究企业统计类应用系统实施项目中根据业务统计报表的形式和内容设计后台数据模型这一过程的内在规律，提出了一种基于语义分析的从统计报表集合建立统一多维数据模型的规范化、程序化方法。这一方法的主要意义在于能够弥合业务人员与技术人员之间的鸿沟，真正将统计数据建模工作从依靠专业经验的无规则“艺术”变为有章可循的通用“技术”。在实际项目中的应用实践表明，通过使用该方法，能够为企业统计类应用系统的分析设计工作带来极大的方便。

另外，该方法的形式化描述为开发出相应的统计数据建模 CASE 工具软件奠定了基础，通过应用该工具，可实现业务人员与技术人员协同开展统计报表的数据建模工作，从而在降低此类工作进入门槛的同时有效提高完成质量。

参 考 文 献

[1] Han Jia-wei, Kamber M. Data Mining: Concepts and Techniques, second edition [M]. Elsevier Inc., 2006

[2] 严金贵. 基于 ER 模型的多维数据建模方法研究 [D]. 重庆：重庆大学，2006

[3] Egozi O, Markovitch S, Gabrilovich E. Concept-Based Information Retrieval Using Explicit Semantic Analysis [J]. ACM Transactions on Information Systems, 2011, 29(2): 87

[4] 陈竞波，王永贵. 基于语义的报表系统模型 [J]. 计算机工程，2010, 36(10): 259

[5] 刘庆伟，孙静. 关系数据库中多维数据分析及展现的研究 [J]. 计算机工程与设计，2008, 29(9): 2262

[6] 杨小献，赵云娣，谢自美. 基于规则的柔性综合统计报表技术 [J]. 计算机应用研究，2005, 22(12): 54

[7] Fowler M. Patterns of Enterprise Application Architecture [M]. Addison Wesley, Inc., 2003