利用数据仓库技术开发文化稽查统计分析系统

李 山

（华东交通大学 经济管理学院，江西 南昌 330013）

摘 要：提出统计分析系统不应该归入普通管理信息系统，而应该根据用户具体需求，充分分析其本质，利用数据仓库技术进行开发和实现，并阐述了如何利用数据仓库技术从需求分析到最终表现的开发全过程。

关键词：数据仓库;统计分析;需求分析;工作流

中图分类号：TP311.13 **文献标识码**：B

Developing the statistic analysis system of culture checking by using data warehouse

LI Shan

(School of Economics and Management, East China Jiaotong University, Nanchang 330013, China)

Abstract: It is suggested that the statistic analysis system is not fallen down into common MIS as a module. According to the requirements of users, analyze the special features. Using the technology of data warehouse, realize the system. While, illustrate the whole process from requirement analysis to final representation.

Key words: data warehouse; statistic analysis; requirement analysis; workflow

统计分析系统（Statistic Analysis System）不是归入到普通管理信息系统 MIS 中的模块或插件，而是建立在 MIS 基础之上，具有一定辅助决策能力的独立系统。往往在传统 MIS 中嵌入统计分析系统，会造成 MIS 运行的数据吞吐瓶颈，给客户带来 MIS 运行缓慢的错觉。尤其是当业务数据量很大的时候，这种情况会突显出来。为此，使用有效的技术手段构造独立的统计分析系统是很有必要的。在开发“文化稽查统计分析系统”项目的时候，采用了数据仓库技术，构建起运行在“文化稽查管理信息系统”之上的统计分析系统。本文介绍了相关的构建过程和关键技术的实施。

1 需求分析

1.1 需求特点

建立统计分析系统依然要经过严格的需求分析阶段，只有在明确的需求指导下，才能开发出满足客户真正需要的系统。MIS 系统是建立在非信息化的原始手工平台上的全新系统，而该系统则是在原有的 MIS 系统开放平台上构造上层系统，因此具两大特点：(1)业务过程信息化。在需求分析阶段不需要重新分析整个业务过程，因为这些复杂的业务流程已经整理并实现在良构的 MIS 中，需关注的应该是对于领导决策层关心的业务数据及其表现形式上。(2)无需采集数据。由于数据的采集过程已经由 MIS 完成，因此，只需要去分析现有的数据集即可。

1.2 关键业务需求

正因为上述需求特点，可以将工作重心从整理业务流程上转移到数据分析上。通过与客户的交流，建立起共性需求。对于任何统计分析系统，都有对数据进行归并和分类的过程，并且提供给决策层的数据往往是在某个层面上的汇总结果。因此，将“文化稽查统计分析系统”的需求归纳成：(1)建立分项统计功能。即对决策层面临的“举报”、“稽查”、“立案”、“处罚”等业务主题建立各自独立的统计模块。(2)确立统计方式为：汇总与分类，同时要多维度表现。即可以在任何统计分项上，考核各统计指标，建立起按照时间、地点、任务划分的统计过程。(3)同时要采用灵活的表现方式。即可以以表格和图形的方式展现给最终用户。

对整个统计过程简单建模如图1所示。这在需求上就确立了该系统的特点符合构造数据仓库的特点，即面向主题，用于决策支持，与时间刻度相关的系统。

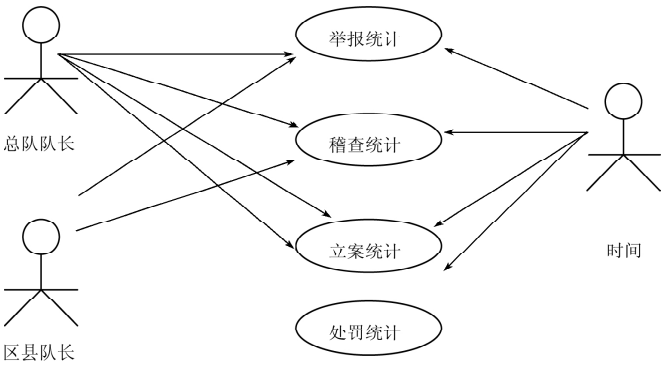

图 1 用户统计用例分析

2 数据预处理

采用基于工作流（Workflow）方式的数据预处理过程。在原有的MIS系统上很容易总结工作流。例如在该系统中，从原有的MIS中截获的基本过程是：举报、稽查、立案和处罚，但是这些只是基本工作过程，在它们之间还有一定的关联关系，这就要通过对业务过程进行分析（Business Process Analysis），以便更好地建立数据集。

2.1 工作流分析

对于整个文化稽查业务基本上划分出上述的5个过程（Process），在各过程之间是判断与选择的关联关系。基本工作流程描述如图2所示。

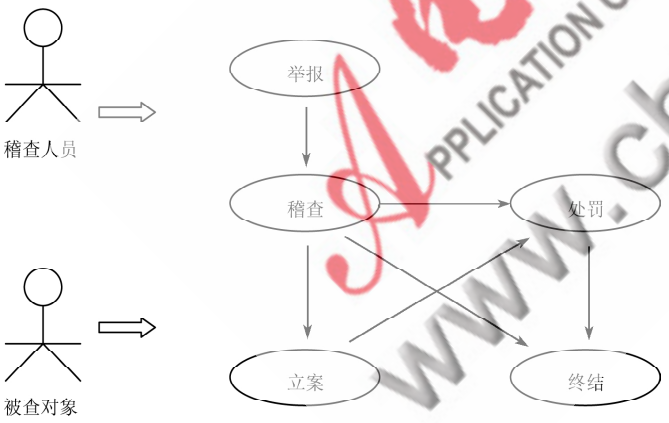

图 2 稽查业务基本流程分析

对于一般的系统，可以从定义过程开始进行分析。

定义 1：

$P=\{p_i \mid i \in N\}$

$R=\{(<p_i, p_j>, ct) \mid i,j,t \in N \wedge i \neq j \wedge ct \in C\}$

P是定义在业务过程上的集合；R是定义在P上的关系对与条件判断C的有序对集合。通过给定这样两组集合，可以在确立主题统计指标之间关系的时候进行直接关联。

这样上述过程可以更加精确的描述：

$P=\{p_1$: 举报, p_2: 稽查, p_3: 立案, p_4: 处罚, p_5: 终结 $\}$

$R=\{(<p_1, p_2>, c_1), (<p_2, p_3>, c_2), (<p_2, p_4>, c_3), (<p_2, p_5>, c_4), (<p_3, p_4>, c_5), (<p_4, p_5>, c_6)\}$

$C=\{c_1$: 接受, c_2: 待处理, c_3: 现场裁决, c_4: 正常, c_5: 裁决, c_6: 结案$\}$

2.2 数据准备

基于上述定义的工作流过程，可以确定需要数据的范畴，并且建立指标集。在数据预处理阶段，将原有业务数据库中的数据按照上述过程进行了划分，确立了分别反映前4个过程的4个关键数据表，并且在它们之间建立了以集合C为条件的关联关系。

JuBao(ID#, …)

JiCha(ID#, JuBaoID, LiAnID…)

ChuFa(ID#, JiChaID, …)

JieAn(ID#, ChuFaID, JiChaID)

按照这4个表中的主外键确立过程关系，同时根据具体情况去除一些异常数据，如图3所示。

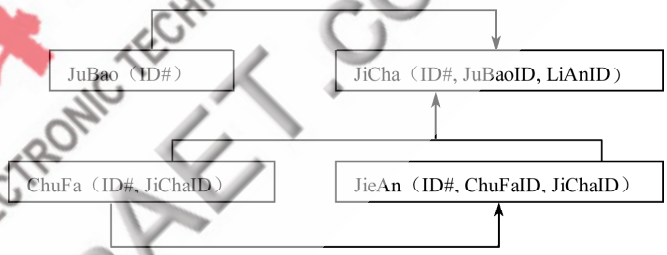

图 3 用于构建数据仓库的基础数据模型

3 数据仓库建模

3.1 确立主题

依照工作流总结的4个基本过程，可以定义出4个主题，如图4所示，按照它们在需求阶段确定的内容，划分数据间的粒度大小。

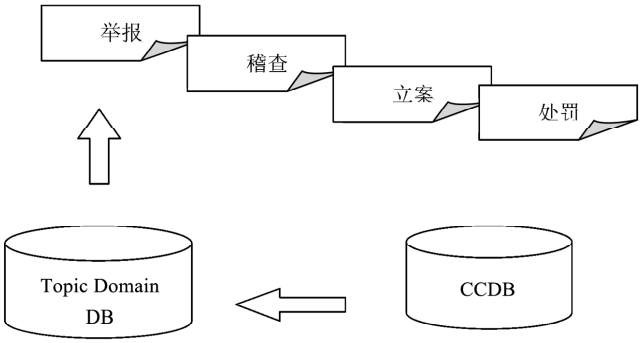

图 4 根据工作流程定义的数据仓库的主题

在粒度划分上要遵循客户实用性原则，即依照客户需求将各维度（Dimension）划分成不同的类别，以便于用户识别。例如：时间维度，可以划分成按年、季度、月份、周和日期的不同粒度。地区维度，可以划分为市、区（县）、街道等。

3.2 建立信息包

确立主题之后，在主题的作用域内确立维度、事实（Facts），并建立起信息包（Information Package）。

例如：对于“稽查”主题，在用户看来需要了解的信息包括，稽查单位数、处罚数量、代立案数量等一些业务指标，而这些正好构成了我们要求解的事实。同时关心在不同时间片断，不同地区，以及考量各业务部门之间的这些指标的变化情况，这样就构成了统计时需要的维度。依次，建立如图5所示的信息包。

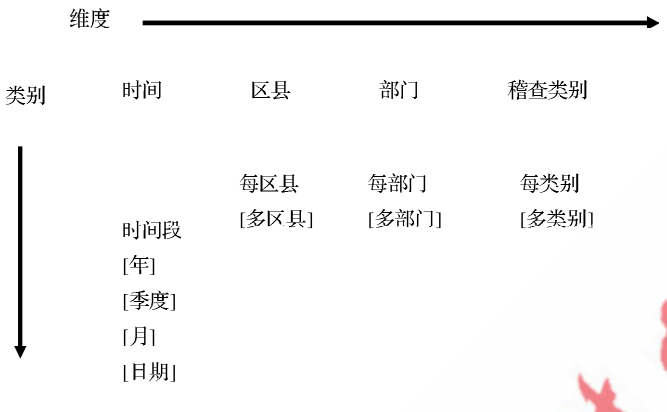

图 5 稽查信息包的建立

3.3 建立星型模型

信息包的确立是建立数据集合的基础，但是需要将这种二维表现模型转换成具有多维度表现的星型模型，如图6所示。

图 6 稽查信息包的星型模型

4 实现数据仓库并开发系统

4.1 基本过程

星型模型指导我们去发现和抽取维度信息、事实数据，最终建立数据仓库，为统计分析系统的开发奠定基础。由模型到物理实现需要经历如图7所示的基本过程。

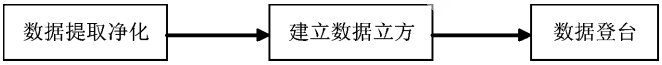

图 7 数据仓库的构建基本过程

建立数据仓库前期需要对业务数据进行净化，消除异常数据，提炼符合要求的基础数据集，并在此之上依照星型模型构建各个主题的数据立方（Data Cube），最后将数据立方登台到物理数据库中，实现统计分析的进一步处理。

例如对于“稽查”主题，我们首先寻找和构建维度表。一般地，可以将维度表描述为：$D = \{di \mid i \in N \wedge di \in R\}$。同时发现事实数据提取的业务表。在这里的事实业务表为上述4个基本表中的JiCha。在清理完上述事实表和构建好维度表之后，需要利用这些表格建立数据立方，计算出各项指标值。

续上过程，一般在构建数据立方过程，可以采用标准SQL完成。一般可以描述为：

$d_i \times d_j$（$0< i,j \leq$ Count(维度表) $\wedge i \neq j$）即各维度的笛卡尔积。或：

```
SELECT COUNT(*), Date, District, …
FROM JICHA
GROUP BY Date, District, …
```

最后将此结果集记录在专门用于统计分析使用的物理数据库中。

4.2 构建前端统计分析系统

在完成数据仓库的物理实现后，可以在此基础上开发相应的统计分析系统，并且需要利用到很多表现丰富的前端处理技术。在此系统中，基本采用以下过程来建造这个前端，如图8所示。

图 8 前端统计分析系统的构建过程

在对统计结果进行展现的时候往往需要满足客户适时调整展现结果的需要，这就需要采用数据钻取（Data Drill）技术，而这个技术在很多商业化的开发工具中都作为包的形势提供给开发人员，因此，开发过程会相对方便和快捷。

数据仓库技术自提出到现在，具体在工程界的应用并不是十分到位，其中一个重要的原因在于客户与开发组织在实现与之相关的项目时，往往不区分传统业务系统和数据仓库系统，这样就会在概念和技术实现上受到阻碍，从而不能满足最终用户的需要。本文从建立统计分析系统在需求上的本质特征，提出两者分离并形成层次关系，利用数据仓库技术，从而很好地解决了上述不足。但是在实现过程中发现，对于实现这种统计分析系统，并非只限于采用数据仓库技术的直接结果，项目的实施还要受到开发成本、用户概念接受程度、现有MIS的完备程度等诸多因素影响，因此在实际开发过程中要权衡考虑。

参考文献

[1] KANTARDZI M. Data mining Concepts, Model, Methods and Algorithms[M]. Tsinghua University Publisher, 2003.

[2] HAMMERGREN T. Data Warehouse Technology[M]. Ventana Communications Group, Inc., 1997.

（收稿日期：2008-12-17）