文章编号:1001 - 9081(2006)04 - 0867 - 03

用层次分析发现多维数据模型的主题域

刘 烨,洪 佳,季石磊,李万勇
(天津工业大学 管理学院, 天津 300160)
(liuye331@eyou.com)

摘 要:为满足组织的决策需求和战略目标,提出在实施目标—问题—度量(GQM)方法中,使用层次分析发现多维数据模型的主题域,通过关联属性挖掘算法发现数据库中与主题关联的关系属性。分析结果对科学合理地布置多维数据的呈现结构有指导作用。

关键词:主题域;层次分析;多维数据模型
中图分类号: TP311.13 **文献标识码**:A

Hierarchical analysis for discovering subject field of multidimensional data model

LIU Ye, HONG Jia, JI Shi-lei, LI Wan-yong
(*School of Management, Tianjin Polytechnic University, Tianjin* 300160, *China*)

Abstract: To meet an organization's decision requirements and strategy goals, the subject field of multidimensional data model could be discovered by hierarchical analysis in the process of implementing Goal-Question-Metric(GQM). The attributes associated with a subject could be found by an mining algorithm in a database. The result has guide to show the multidimensional exhibition structure in data warehouse scientifically.

Key words: subject field; hierarchical analysis; multidimensional data model

0 引言

数据仓库和联机分析处理技术在业务数据处理和决策分析中正变得越来越重要。构建数据仓库多维数据模型的方法和技术已有论述[1],其中最重要的是分析和设计出主题域模型,并在有效的主题域内确定主要的实体(待分析的数据项——维)。可是,一个有效的主题域模型很大程度上取决于业务分析的目标,即多维数据模型产生的信息类型很大程度上取决于一个组织的分析需求。而数据库中的信息和组织的分析需求之间存在差距,由管理者确立的系统目标与工程师设计的数据库模式有冲突。从数据库和组织目标中发现用于分析的数据和信息的新方法是一个挑战。本文提出用层次分析方法建立多维数据模型的主题域,并将此方法用于政府的行政许可审批事项(Online Analysis Processing,OLAP)系统的多维数据建模。

1 研究的背景与思路

1.1 研究案例的描述

本研究对某市的行政许可审批事项 OLAP 系统建模,其中该市的技术质量监督局的行政审批事项分布在其管辖的二十个区县局和部门办理,提供 60 多项审批服务。各办理部门受理的审批事项既有同类型的,也有不同类型的。据统计 2005 年上半年审批事务处理记录已超过 18 万条。目前,部分事项审批受理点已建立了办理事项的事务型数据库系统。

本案例的目标是为政府提高其执政能力提供有效的决策依据,所以决策系统提供的决策信息必须能反映服务对象(企业或个人)申请事项审批全过程的服务需求信息及其相互关系。我们能从已有数据库中表的属性和数据分析的目标确定待分析的主题以及相关的维,但经与政府部门的中、高层管理人员交流,发现现有的事务型数据库中记录的信息无法满足他们所要呈现的分析主题及其相关的维信息的需求,如数据库中表的某些属性对决策分析毫无用处(如审批方式等),而同时又缺乏用于决策分析的属性(如企业类型、企业所属行业等)。因此目前的数据仓库多维数据模型的构建方法无法直接采用。由此需要我们寻求新的方法确定待分析的主题域,以满足决策者对数据分析的需求。

1.2 研究的思路

数据仓库应面向主题,而主题是分割数据的一种机制。提取主题的一般方法是:利用与主题的相关性把数据仓库中的数据分为多个主题域(主题域是与组织相关的重要的物理项、概念、人、地点和事件的主要分组[1])。本研究案例面临的问题是:2004 年 7 月 1 日我国刚刚实施《中华人民共和国行政许可法》,支持行政许可事项受理的电子审批系统尚不成熟,且原事务型数据库中的信息又不能完全满足分析需求。考虑采用目标—问题—度量(Goal-Question-Metric,GQM)方法[2],用于确定组织的重要目标、达到目标需解决的问题以及满足用于分析问题的度量变量。本案例研究的基本思路是:采用 GQM 模式对组织(政府)的相关中、高层管理决策人员进行采访,将获得的信息分组和整理,确定组织的目标及影响目标的相关问题(影响因素)。本文提出用关系矩阵描述组织目标、影响因素及关系属性(度量变量)的相关性;用层次分析确定与主题相关性强的属性或属性组合[3,4],由此确定多维数据模型的待分析主题域。本文把每个主题域作为一

收稿日期:2005 - 10 - 20;**修订日期**:2005 - 12 - 28 **基金项目**:天津市高等学校科技发展基金项目(520126)

作者简介:刘烨(1962 -),女,天津人,副教授,硕士,主要研究方向:企业建模、信息集成; 洪佳(1982 -),女,天津人,硕士研究生,主要研究方向:数据仓库、电子政务; 季石磊(1984 -),男,硕士研究生,主要研究方向:信息集成、数据仓库; 李万勇(1983 -),男,四川人,硕士研究生,主要研究方向:信息集成、商务智能.

个独立的数据集来考虑,即不考虑主题域之间的关系。图1为构建多维数据模型的层次分析过程。其中:

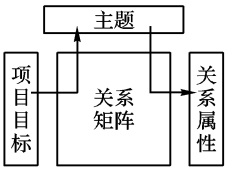

图1 层次分析过程

(1)项目目标:组织(如:政府)的战略和决策目标。

(2)关系矩阵:描述了组织的目标及其影响因素之间,影响因素与其度量变量之间的关系。

(3)主题:通过对关系矩阵的分析,确定待分析主题。

(4)关系属性:数据库中用于度量影响因素的关系属性。

2 层次分析

2.1 建立关系矩阵

关系矩阵用于决定组织目标及其影响因素之间的关系[5](图2)。关系矩阵中的每个元素分为左上和右下两部分:左上部分表示目标和影响因素之间的关系;右下部分表示影响因素和关系属性之间的关系。单圈表示有关联的关系;双圈表示有强关联关系;空白表示没有明显关系。

目标	影响因素					
	企业类型	企业所属行业	企业所属区域	行政事项类别	企业规模	事项受理日期
经济发展	◎	◎	○	○	◎	
政府价值	○	○	○	◎	○	○
群众满意度	○			◎		○
公共安全				○	○	○

图2 案例政府的关系矩阵(元素的上三角关系)

目标	影响因素						关系属性
	企业类型	企业所属行业	企业所属区域	行政事项类别	企业规模	事项受理日期	
经济发展	◎ ◎	◎	○	○	◎		企业类型
政府价值	○	○ ◎	○	◎ ◎	○	○	所属行业
群众满意度	○ ○			◎	◎	○	企业规模
公共安全			◎	○	○	○	所属区域
		◎		◎			事项类别
						◎	受理日期
						○	受理数

图3 案例政府的关系矩阵(元素的下三角关系)

对案例政府的部分公务员及中、高级管理人员进行采访,将他们所描述的信息及信息之间的关系分组,确定目标及其影响因素之间的关系,图2为这一阶段的关系矩阵的分析结果。图2中的强关联关系有:

经济发展—企业类型:民营企业对社会发展的作用越来越大;中介机构是市场经济是否完善的标志;国有企业是改革的方向。

经济发展—企业所属行业:高新技术产业的发展标志着社会发展的水平。

经济发展—企业规模:大企业是地区经济发展的支柱。

政府价值—行政事项类别:政府要抓住关键问题。

群众满意度—行政事项类别:群众关心,办事方便。

2.2 建立主题域模型

构建性能良好的主题域模型是数据仓库数据组织成功的关键。上述关系矩阵可用于确定与项目目标和影响因素有关的关系属性。可是,组织的决策目标不同,影响决策的因素对决策的影响程度就不同。数据库中的关系属性有些是影响某项决策的关键因素,有些是决策分析的可变因素。为了加以区别,给出定义如下。

定义1 关键属性

决策域 P 是数据库 D 中用于分析的数据,如要分析的数据是2003年至2004年的数据,假设 集合 $A = \{ai \mid ai$ 是数据库 D 中的一个属性,$i = 1,2,\cdots,n\}$,如果 $aj \subset A$(aj 是 A 的真子集),且选择的一些数据 $d \subset D$,$d(aj) \subset P$,则称 aj 是一个关键属性。例如:要分析一年中申请事项超过50项的企业,则“受理数”就是关键属性。

关键属性是数据库中影响某个决策的重要因素。

定义2 关联属性

决策域 P 是数据库 D 中用于分析的数据。假设 集合 $A = \{ai \mid ai$ 是数据库 D 中的一个属性,$i = 1,2,\cdots,n\}$,而集合 $A' = \{aj \mid aj$ 是数据库 D 中的一个关键属性,$j = 1,2,\cdots,m\}$。如果 $ak \subset A$ 且 $ak \subset A'$,对一些数据集 $D(ak1),D(ak2),\cdots,D(akn)$,其中 $ak1,ak2,\cdots,akn$ 是 ak 的值,且 $\bigcup_{i=1}^{n} D(aki) = D$,则 ak 是一个关联属性。例如:要分析一年中申请“重要工业产品许可证”事项超过50项的企业,则“事项类别”就是一个关联属性。

关联属性是一个变量,用于将数据库中的数据分为具有不同特征的数据簇,每个数据簇将驱动一个不同的决策,且它们将影响系统的目标。

分析和区分数据库中属性的迭代挖掘算法[3,4,6]如下:

```
Integer m,n
Real v                          /* v 是一个可接受的支持度[6,7] */
                                /* m 是数据库中关联属性的个数 */
n = 1
WHILE ( n < m )
    DO N = m C n ,
    /* N 是关联属性的组合数; n 是未分析的关联属性组合 */
    WHILE 存在未参与分析的关联属性组合时,选择一个未分析
        的关联属性组合;
        挖掘 D = {d | d 是数据,d(a1,a2,…,aj) ⊂ P),(a1,a2,
          …,aj) 是一个关键属性集,P 是决策域。}
        挖掘 D' = {d | d 是数据,d(ak ⊂ (a1,a2,…,aj)) ⊂ P,
          (a1,a2,…,aj) 是一个关键属性集,ak 是一个关联属
          性,P 是决策域。}
        计算 ratio = D'/D
        IF ratio > = v , 则输出 D'
    ENDWHILE
```

```
    n = n + 1
ENDWHILE
END
```

由 GQM 方法和图 2 中的关系矩阵,数据库中有 7 个关系属性可用于度量决策因素。影响因素与关系属性之间的相关性强弱关系标注在矩阵元素的右下角,如图 3 所示。

本案例中,目前政府关注的主要目标是推动经济发展的执政能力。从图 2 可知,由此目标引出的主题——事项应以“受理日期”和“受理数”为关键属性,由上述关联属性迭代挖掘算法,案例政府的数据库由企业类型(Business)、所属行业(Industry)、企业规模(Scale)和事项类别(Class)关联属性分为四个数据簇(即决策准则)。

运用上述迭代挖掘算法,四个关联属性的所有组合见图 4 ,这是一个完全相关性的层次结构。为消除和过滤那些与此目标主题关联关系不大的数据元素,假设给定可接受的 v 值为 0.5,对图 4 中的各属性及属性组,由 2003 年、2004 年、2005 年数据作决策域。第一遍扫描仅考虑含有单个属性的项集,第二遍扫描考虑有两个属性组合的项集,依次类推。一次扫描中,每取 1 条数据库中的业务记录,将含有与该记录相关的属性或属性组的计数器加 1。一遍扫描结束时,小于 v 值支持度的属性或其属性组被删除。一旦某个属性或属性组被删除了,就不需要再考虑它的任何超集了。本案例数据分析表明,在关键属性“受理日期”和“受理数”的基础上,关联属性“企业类型”、“所属行业”和“事项类别”及其组合对所选主题的分析有明显的相关性,它们是决策分析的关键因素。由此可得,以推动经济发展的执政能力为目标的事项主题域模型(图 5)。

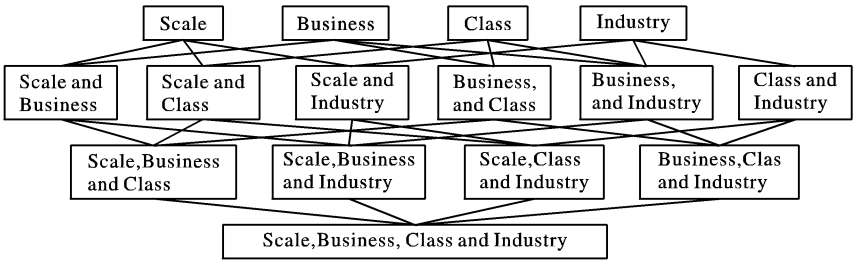

图 4　关联属性的挖掘和层次分析过程

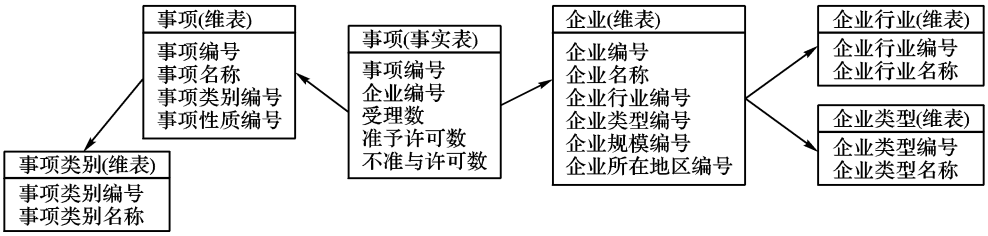

图 5　事项主题域模型

为规范数据维和减少数据冗余,本案例的主题域模型采用雪花型架构。一个主题域应具有为组织的尽可能多的目标服务的信息类型,根据本项目对象的特征,事项主题域涉及的对象有服务实体(包括企业和事项)及服务主体(包括办件和审批),图 5 中仅给出了含有服务实体维的事项主题域模型。

主题域模型只是描述了组织的高层数据模型,构建数据仓库的多维数据模型还需要在此基础上根据业务数据模型建立系统的多维数据模型[8],还需要考虑增加时间维、各种派生维等对主题进行全面考核的指标及其汇总计算等维属性的设计。

3　多维数据的呈现结构设计

多维数据的呈现结构设计是表现多维数据模型分析功效的关键。由于上述对于主题域的某一目标相关的关联属性进行了层次分析,给出了与主题目标相关性强的关联属性“企业类型”、“所属行业”和“事项类别”及其组合,它们是决策分析最本质的关键因素,这对科学合理地布置多维数据的最终呈现结构有指导性作用。多维数据的展现结构应围绕相关性强的关联属性及其它们的组合来设计,而不应盲目设计。

4　结语

利用 GQM 方法构建的层次分析矩阵反映了项目目标与数据库中属性的密切关联关系,由此构建的多维数据模型能更好地支持组织的决策目标,迭代关联属性挖掘算法有助于系统开发者更有效地抽取和组织与主题密切相关的数据维。本案例构建的主题域模型保证了系统目标与政府的分析需求一致。

本文提出了一个有用的关联属性挖掘技术,这一技术也能用于传统的关联规则和其他的管理决策问题中,可以更有效地帮助组织和信息部门生产更高质量的信息。

参考文献:

[1] GALEMMO N, GEIGER JG. 数据库仓库设计[M]. 于戈, 等译. 北京: 机械工业出版社, 2004. 9 - 96.

[2] BASILI VR, ROMBACH HD. The TAME Project: Towards Improvement-Oriented Software Environments[J]. IEEE Transactions on Software Engineering, 1988, 14(6): 758 - 773.

[3] FUKUDA T, MORIMOTO Y. Data Mining with Optimized Two-Dimensional Association Rules[J]. ACM Transactions on Data System, 2001, 26(2): 179 - 213.

[4] CARTER CL, HAMILTION HJ. Efficient Attribute-Oriented Algorithms for Knowledge from Large Database[J]. IEEE Transactions on Knowledge and Data Engineering, 1998, 10(2): 193 - 208.

[5] YEN SJ, CHEN ALP. A Graph-Based Approach for Discovering Various Types of Association Rules[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(5): 839 - 845.

[6] 赖邦传, 陈晓红, 周辉. 基于数据仓库的高效关联规则的挖掘[J]. 计算机工程, 2004, 30(5): 6 - 8.

[7] SILBERSCHATZ A, KORTH HF, SUDARSHAN S. Database System Concepts[M]. 4th Edition. Beijing: China Machine Press, 2003. 584 - 586.

[8] SILVRRSTON L. Data Model Resource Book, Volume 1[M]. Beijing: China Machine Press, 2004. 243 - 263.