文章编号: 1001- 2486(2003) 01- 0031- 04

ER2MD: 一种 ER 模型到多维模型的转换算法*

唐九阳, 杨 强, 沙基昌, 邓 苏, 陆昌辉
(国防科技大学人文与管理学院, 湖南 长沙 410073)

摘 要: 多维模型是数据仓库概念设计不可缺少的工具, 将传统数据库实体关系模型直接转换到多维模型, 势必在很大程度上缩短数据仓库系统的开发周期。针对数据仓库概念设计需求, 在传统数据库基础上, 提出一种面向多维模型的转换算法 ER2MD。该算法可将满足一定条件的实体关系模型转换到多维模型, 产生的多维模型符合广义多维范式, 能够确保多维数据库分析计算的有效性, 进而有利于物理数据库的设计。

关键词: ER 模型; 多维模型; 转换; 广义多维范式
中图分类号: TP311.13 文献标识码: A

ER2MD: An Algorithm of the Transformation from ER Model to Multidimensional Model

TANG Jiu-yang, YANG Qiang, SHA Ji-chang, DENG Su, LU Chang-hui
(College of Humanities and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Multidimensional modeling is one of the chief techniques in conceptual design of data warehouse. If ER models of traditional databases can be directly transformed into multidimensional models, it would efficiently shorten the life circle of data warehouse design. Based on the traditional databases, an algorithm of the transformation from ER models to multidimensional models is presented. Those ER models under some conditions can be transformed, and the resulting multidimensional models conform to the standard of the generalized multidimensional normal form, ensuring the validity of analytical computations on multidimensional databases. Furthermore, the resulting multidimensional models favor an efficient physical database design.

Key words: ER model; multidimensional model; transformation; generalized multidimensional normal form

作为概念建模工具的实体关系(ER)模型和多维(MD)模型具有各自的特点。ER 模型是直接从现实世界中抽象出实体类型及实体间联系, 然后用 ER 图表示的数据模型, 通过消除数据之间的冗余, 提高联机事务处理(OLTP)系统中数据库操作的速度和精确性。MD 模型则为了简化终端用户的操作, 提供灵活丰富的多维分析与查询, 采用旋转、嵌套、切片、钻取和高维可视化技术, 展示多维视图的结构, 支持用户从不同的角度去直观地理解、分析数据, 进行决策支持[1~3]。

文献[4, 5] 在这方面做了初步的探讨, 提出只要符合一定的条件, ER 模型就能派生出 MD 模型的初始结构。文献[6, 7] 提出了基于逻辑数据模型设计 MD 模型的方法, 由于这些方法都是基于具体应用的 ER 模型, 因此没能给出通用的设计框架。文献[8~10] 从企业操作数据库的概念模型出发, 给出了数据仓库概念设计的框架, 但是给出的模型或是对 ER 模型的扩展[8], 或是对 MD 模型的扩展[9,10]。事实上, 并不是每个 ER 模型都能够被描述成包含等价信息的 MD 模型的集合[4], 文献[6~10] 都没能给出两种模型可以进行转换的条件。而且由于缺乏统一的 MD 模型设计标准, 上述方法都没能针对具体的 MD 模型设计标准, 对建好的 MD 模型进行分析。

针对上述方法的不足, 在形式化模型描述的基础上, 首先明确 ER 模型和 MD 模型可进行转换的条件, 然后给出转换的 ER2MD 算法, 最后证明转换后的 MD 模型符合文献[11] 提出的广义多维范式。

* 收稿日期: 2002- 07- 02
基金项目: 国家自然科学基金资助项目(60172012)
作者简介: 唐九阳(1978—), 男, 博士生。

1 模型描述

要建立 ER 模型和 MD 模型的转换, 有必要先对这两种模型进行形式化的描述。目前, 已经提出了多个形式化的 MD 模型[6, 10, 12~ 15], 这里在文献[10, 12]的基础上, 建立一个能有效支持 ER 模型转换的 MD 模型。为此, 首先给出 ER 模型和 MD 模型的定义。

定义 1 ER 模型是一个三元组 $ER = < E, A, R>$, 其中: $E = \{e_1, e_2, \cdots, e_n\}$, e_i 是实体名; $A = \{a_1, a_2, \cdots, a_k\}$, a_i 是属性名; $R = \{r_1, r_2, \cdots, r_p\}$, r_i 是关系名。

假定 r_i 与 e_m、e_p 和 e_r 关联, 如图 1 所示, 则有 $r_i \subseteq Dom(e_m) \times Dom(e_r) \times Dom(e_p)$, 那么关系 r_i 与实体 e_r 所对应的基数 n 可以表示成 $\#(r_i \cap (e_m^{\tau} \times Dom(e_r) \times e_p^{\tau})) = n$, 其中 Dom 表示实体和属性所在的值域, e_m^{τ} 表示 $Dom(e_m)$ 中的一个取值, # 表示集合中元素的个数。

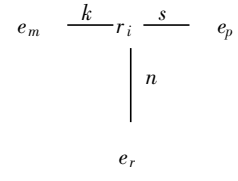

图 1 ER 图

Fig. 1 ER diagram

定义 2 MD 模型是一个五元组 $MD = < F, Dim, M, Att, FDs>$, 其中: $F = \{f_1, f_2, \cdots, f_r\}$, f_i 是事实名; $Dim = \{\dim_1, \dim_2, \cdots, \dim_p\}$, $\dim_i$ 是维名; $M = \{m_1, m_2, \cdots, m_k\}$, m_i 是度量名; $Att = \{d_1, d_2, \cdots, d_t\}$, d_i 是维属性名; FDs 是函数依赖集。

维属性集 $Att = Att_t \cup Att_c \cup Att_p$, 其中 $Att_t \cap Att_c = \Phi$, $att_t \cap Att_p = \Phi$, $Att_c \cap Att_p = \Phi$。Att_t 是惟一标识各个维的标识属性集, 如{产品_ID, 时间_ID}, 其中产品_ID 是产品维的标识属性, 时间_ID 是时间维的标识属性; Att_c 是反映维层次结构的维层次属性集, 如{日, 月, 年, 县, 市, 省}, 标识属性和维层次属性统称作维级; Att_p 是描述维级相关信息的非维层次属性集, 可以用于减少查询结果, 如{客户名, 客户年龄}。

函数依赖集 FDs 中的函数依赖 FD 用 $\rightarrow$ 表示, $FD: Att \rightarrow Att$。函数依赖分全函数依赖和部分函数依赖。对 $A, B \subset Att$, 如果 B 全函数依赖于 A, 则 B 中的属性称作必要属性; 若 B 部分函数依赖于 A, 则 B 中的属性称作可选属性。如对客户_ID 来说, 客户名是必要属性, 客户年龄是可选属性(对每个客户_ID, 有惟一的客户名与之对应, 而相应的客户年龄可能取值为空)。显然, 对 $\forall d_i \in Att_t$, 不存在 $d_j \in Att \setminus \{d_i\}$, 使得 $d_j \rightarrow d_i$。

由定义 2 可知, 本文提出的 MD 模型可以刻画多重维层次结构。如函数依赖集:

{客户_ID $\rightarrow$ 职业, 客户_ID $\rightarrow$ 分支机构, 职业 $\rightarrow$ 客户类型, 分支机构 $\rightarrow$ 客户类型} 存在:

{客户_ID $\rightarrow$ 职业 $\rightarrow$ 客户类型}

{客户_ID $\rightarrow$ 分支机构 $\rightarrow$ 客户类型}

两重维层次, 其中客户_ID 称作分裂维级, 客户类型称作连接维级。

2 ER2MD 算法

2.1 转换条件

MD 模型描述了事实和维之间多对多的关系(多对一的关系是多对多关系的特例), 这种关系通过它们之间的连接体现出来。但在另一方面, 不管是否存在多对多的关系, 总能定义 ER 模型。

在实际应用中, 往往使用一个大规模的 ER 模型描述可能发生的各种事务处理。这个主 ER 模型可能包含了订货、购买、顾客付账、运货、产品退货和其它事务, 这使得它自身变得异常复杂。与 ER 模型对整个商业过程建模相比, MD 模型只围绕特定的商业过程或主题展开, 每个 MD 模型只对一个商业过程建模。因此, 从这里可以看出, 并不是每一个 ER 模型都能被描述成包含等价信息的 MD 模型的集合。如果要分析的主题所在的 ER 模型中不包含多对多关系, 就无法确定事实, 也就无法定义一系列的 MD 模型(即 MD 模型的存在与多对多关系的存在紧紧联系在一起[4])。

对 ER 模型中的 n 元关系 r_i 引入标记

$$R(i, n, j) = \#(r_i \cap (\underbrace{e_l^{\tau} \times \cdots \times Dom(e_x^{\tau}) \times \cdots \times e_p^{\tau}}_{n\text{个}})),\ 1 \leqslant j \leqslant n,\ n \geqslant 2$$

其中 e_x 是关系 r_i 的 n 元序偶中处于第 j 位的实体, 那么整个式子代表关系 r_i 与实体 e_x 所对应的基数。如当 $i=1, n=3, j=2$ 时, $R(i, n, j)$ 可以表示成

$$R(1,3,2)=\#(r_1 \cap (el^{\tau} \times Dom(e_x) \times e_p^{\tau}))$$

设 $ARik$ 为关系 ri 的属性, 有 $ARik \in A$, 且 $1 \leqslant k \leqslant \#$ (R_i 的属性集), 记

$$Numeric(ARik)=Boolean(Dom(ARik) \subseteq Number)$$

即判断属性 $ARik$ 是否为数值类型, 其中 $Number$ 是所有数值类型的值的集合。

根据以上描述, 可以得到转换条件为: $\exists j(R(i, n, j) \geqslant 2) \wedge \exists k Numeric(ARik)$。

2.2 转换算法

从ER 模型中挑选出与主题相关的商业过程, 设包含的实体为 $E=\{e_1, e_2, \cdots, e_n\}$, 属性为 $A=\{a_1, a_2, \cdots, a_l\}$, 关系 $R=\{r_1\}$(在这里只讨论单重事实的情况)。根据分析的需求, 给出度量集 $M=\{m_1, m_2, \cdots, m_p\}$。

ER2MD 算法的基本思路是以事实为中心, 查找能够通过函数决定度量的最小维级集合, 逐个建立 ER 模型中实体所对应的维; 然后, 基于 ER 模型的属性集区分维层次属性和非维层次属性; 最后借助在每个维所确定的函数依赖集中找非传递的函数依赖, 确定维层次。

转换算法描述如下:

$Dim=\Phi$; $Att=\Phi$; $FDs=\Phi$;
$f_1=r_1$; //确定事实
for 每个 $m_i \in M$ //确定通过函数决定度量的最小维级集合
找最小的集合$\{a_j, \cdots, a_r\} \subseteq A$, 使得$(a_j, \cdots, a_r) \rightarrow m_i$;
$Att_t=Att_t \cup \{a_j, \cdots, a_r\}$; //确定维标识属性集
for 每个 $a_i \in Att_t$
$\dim_j=a_i$ 所在的e_j; //维标识属性惟一确定各个维
$Att_{tj}=a_i$ //确定每个维对应的维标识属性, Att_{tj}表示 $\dim_j$ 的标识属性
$Dim=Dim \cup \{\dim_j\}$;
for 每个 $\dim_i \in Dim$
for 每个 $a_j \in e_i$ 的属性集 \ Att_t // \ 表示集合中的差运算
if a_j 是必要属性且有 $a_k \in A \setminus Att_t \setminus \{a_j\}$, 使得 $a_k \rightarrow a_j$ or a_k 是必要属性且有 $a_j \rightarrow a_k$
$Att_c=Att_c \cup a_j$ //确定维层次属性集
else
$Att_p=Att_p \cup a_j$ //确定非维层次属性集
for 每个 $a_i, a_j \in Att_c$
if $a_i \rightarrow a_j$ and $Att_{tk} \rightarrow a_i (1 \leqslant k \leqslant \#(Att_t))$
//确定每个维的函数依赖, $FD_{\dim_k}$是维 $\dim_k$ 的函数依赖集
$FD_{\dim_k}=FD_{\dim_k} \cup \{FD \mid a_i \rightarrow a_j\}$
for 每个 $\dim_i \in Dim$
$FDs=FDs \cup FD_{\dim_i}$;
for 每个 $a_i \in Att_c$ //找非传递的函数依赖, 确定维层次
在每个维的函数依赖集中挑选出满足下列条件的函数依赖: $a_j \neq a_i$, $a_i \rightarrow a_j$, 且不存在 $a_k \neq a_j \neq a_i$, 使得 $a_i \rightarrow a_k \rightarrow a_j$

2.3 算法分析

本文给出的 ER2MD 算法可以由广义多维范式的标准验证。首先, 回顾文献[11]中提到的与广义多维范式相关的一些定义。单维模型 D 是满足以下条件的维属性集合 Att: 对 $\forall d_i \in Att$, 存在 $d_j \in Att \setminus \{d_i\}$, 使得 $d_i \rightarrow d_j$ 或 $d_j \rightarrow d_i$。MD 模型 $Mul=(\{D_1, \cdots, D_K\}, m_i)$, 其中$\{D_1, \cdots, D_K\}$ 是单维模型

集合, 度量 m_i 由$\{D_1, \cdots, D_K\}$中出现的属性函数决定。设 d_t 是一维标识属性, d_c 是一维层次属性, d_p 是一非维层次属性, 那么 d_p 关于d_c 的元素c 上下文有效是指: 对于元素 c 所在维层次上的d_t 的每个元素, d_p 都有一确定值; 若 d_t 的元素不在 c 的维层次路径上, 则 d_p 值无定义。

单维模型 D 符合维范式: 只有一个维标识属性 d_t; 标识属性的元素是完全的; 所有维属性都是必要的。MD 模型 $Mul = (\{D_1, \cdots, D_K\}, m_i)$ 符合广义多维范式: 对每个非维层次属性 $d_p \in D_i$, 都有一个维层次属性 $d_c \in D_i$ 的元素指明了 d_p 的上下文有效性; 只包含维级的单维模型都符合维范式; 维互相正交; 度量 m_i 全函数依赖于相应的最小标识属性集。根据 ER2MD 算法有以下引理:

引理 1 (1) 算法确定的各个维层次都是只依赖于一个维标识属性的单维模型; (2) 各个维互相正交; (3) 维中如果不存在多重维层次, 则所有的维级都是必要的; (4) 维中如果包含了多重维层次, 则连接维级的元素指明了非维层次属性的上下文有效性; (5) 每个度量由相应的最小维标识属性集全函数决定。

定理 1 ER2MD 算法产生的 MD 模型符合广义多维范式。

证明 使用以上引理, 对照文献[11]中的定义, 可以很容易验证 ER2MD 算法产生的 MD 模型符合广义多维范式。

3 结论

本文以传统数据库为基础, 提出了一种将 ER 模型转换到 MD 模型的 ER2MD 算法, 并借助广义多维范式对转换后得到的 MD 模型进行分析, 验证了算法的正确性。本文提出的 ER 模型和 MD 模型具有较强的语义表达能力, 能够方便、直接地表达转换中的各种语义知识, 从而有效地支持从操作数据源中导出数据仓库的初始结构。与其他转换方法相比, 本文提出的 ER2MD 算法独立于数据库和数据仓库的逻辑结构不考虑具体技术条件的限制, 是纯粹概念意义上的转换, 进而为数据仓库概念设计提供了一个框架。

参考文献:

[1] Batini C, Ceri S, Navathe S B. Conceptual Database Design: An Entity-Relationship Approach[M]. California(USA): The Benjamin/Cummings Publishing Company, 1992.

[2] White Paper: The Data Modeling Problem[J]. http://citm.utdallas.edu/pub-whitepapers.html.

[3] Inmon W H. Building the Data Warehouse[M]. Chichester(USA): John Wiley & Sons, second edition, 1996.

[4] Kimball R. A Dimensional Modeling Manifesto[J]. DBMS, August 1997, 10(9).

[5] Firestone J. Dimensional Modeling and E-R Modeling in the Data Warehouse[J]. http://www.dkms.com/DMERDW.html.

[6] Cabibbo L, Torlone R. A Logical Approach to Multidimensional Databases[J]. In: Proc. 6th EDBT, Valencia, Spain, 1998: 183-197.

[7] Moody D, Kortink M. From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design[J]. In: Proc. of Int'l Workshop on Design and Management of Data Warehouses, Stockholm, Sweden, 2000: 5.

[8] Boehnlein M, Ende A. Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems[J]. DOLAP 1999: 15-21.

[9] Golfarelli M, Maio D, Rizzi S. Conceptual Design of Data Warehouses from E/R Schemes[J]. In: Proceedings of the Hawaii International Conference on System Sciences, Hawaii, 1998: 334-343.

[10] Golfarelli M, Maio D, Rizzi S. The Dimensional Fact Model: A Conceptual Model for Data Warehouses[J]. International Journal of Cooperative Information Systems, 1998, 7(2&3): 215-247.

[11] Lehner W, Albrecht J, Wedekind H. Normal Forms for Multidimensional Databases[J]. In: Proc. 10th SSDBM, Capri, Italy, 1998: 63-72.

[12] Datta A, Thomas H. The Cube Data Model: A Conceptual Model and Algebra for On-line Analytical Processing in Data Warehouses[J]. Decision Support Systems, 1999, 27(3): 289-301.

[13] Gyssens M, Lakshmanan V S. A Foundation for Multi-dimensional Database[J]. In: Proceeding of the 23rd VLDB Conference, Athens, Greece, 1997: 106-115.

[14] Husemann B, Lechtenborger J, Vossen G. Conceptual Data Warehouse Design[J]. In: Proceedings of International Workshop on Design and Management of Data Warehouses, Stockholm, 2000: 6.

[15] Blaschka M, Sapia C, Hofling G. On Schema Evolution in Multidimensional Databases[J]. In: Proceedings of the DaWak99 Conference, Florence, Italy, 1999: 153-164.