

基于依赖关系的大规模主题数据库的分解模式^{*}

刘文远¹ 于家新¹ 徐丽娜¹ 陈国鹰^{1,2}

(燕山大学信息科学与工程学院 秦皇岛 066004)¹ (河北工业大学信息工程学院微电子所 天津 300130)²

摘要 本文对主题数据库之间的依赖关系进行了定义和说明,并细分其为实体依赖和操作依赖。在此基础上,对主题数据库之间的依赖关系进行了度量,提出了主题数据库之间的相对依赖度和相对被依赖度,并以此确定主题数据库的规模大小。最后给出了分解主题数据库算法的描述。通过这个算法可以把大规模的主题数据库科学地分解成若干个子主题数据库,降低主题数据库之间的依赖度,同时又保证了主题数据库内部的联系是紧密的,大大提高了检索数据库的效率,并且有利于对主题数据库的管理和维护。

关键词 主题数据库,依赖关系,实体依赖,依赖度

Mode of Large-scale Subject Database's Subdivision Based on Dependency Relations

LIU Wen-yuan¹ YU Jia-xin¹ XU Li-na¹ CHEN Guo-ying^{1,2}

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)¹

(Department of Information Engineering, Hebei University of Technology, Tianjin 300130, China)²

Abstract The dependency among subject databases is defined and explained in this paper, and it is fractionated into entity dependency and operation dependency, then the subject database dependency relation is measured on this basic. The relative dependency and opposite among subject databases are proposed by which subject database scale can be affirmed. Finally, the description of subdivision arithmetic is proposed through which large-scale subject database can be subdivided into certain subject databases. The depending degree among subject databases can be reduced and the internal relation of subject database will be kept close simultaneously. So the retrieval efficiency to database will be enhanced greatly and it is advantageous to the subject database's management and maintenance simultaneously.

Keywords Subject database, Dependence relation, Entity dependency, Depending degree

1 引言

在信息资源规划(Information Resource Planning,简称IRP)中,主题数据库的划分是一个重要的核心环节,有些规模过大的主题数据库需要进一步细分,以便于管理和维护。确定主题数据库规模的大小没有一个严格的标准,实际设计过程中往往受系统分析员本人的水平和管理经验的影响,人为因素造成主题数据库设计的质量良莠不齐,也会影响到企业日后应用项目的开发。

主题数据库之间是存在一定的依赖关系的,它们之间的依赖关系是主题数据库之间联系中最重要、最普遍的一种联系,这种关系广泛地存在于信息资源规划的整个过程当中。国内外在对主题数据库划分技术的研究中,对主题数据库的划分只提出了定性的原则,缺乏定量研究。2001年,方菲提出了数据依赖的概念^[1],2004年欧阳和胡顺仁在数据依赖的基础上对类之间的依赖关系进行了度量^[2,3]。本文提出了主题数据库之间实体依赖和操作依赖的定义和度量,并以此来分解大规模的主题数据库。

2 主题数据库之间依赖关系的定义

2.1 依赖关系的定义

在IRP系统数据建模的主题数据库划分过程当中,S为

所有主题数据库的集合,记为:

$$S = \{D_1, D_2, \dots, D_n\}$$

其中, n 为集合 S 中主题数据库的个数。

由于每个主题数据库都是由实体组成的,我们可以将主题数据库看作为实体的集合。主题数据库 D_i 记为 $D_i = E_i$, $E_i = \{e_{i1}, e_{i2}, \dots, e_{ip}\}$ 。对主题数据库 D_i 进行的操作全体记为 O_i , $O_i = \{o_{i1}, o_{i2}, \dots, o_{iq}\}$,其中 $i=1, \dots, n$ 。

E_i 为主题数据库 D_i 中所有实体的集合, O_i 为对主题数据库进行的所有操作的集合, p, q 为主题数据库 D_i 中实体的数目和对主题数据库 D_i 的操作的数目。

那么,主题数据库 D_i 和 D_j 之间所有的各种关系都是 $((E_i \cup O_i) \times (E_j \cup O_j)) \cup ((E_j \cup O_j) \times (E_i \cup O_i))$ 的子集($\times$ 为笛卡儿乘积)。本文研究的主题数据库之间的依赖关系 R 也是其中的一个子集,定义如下:

定义1 设主题数据库 D_i 和 D_j 之间的依赖关系 R 为定义在集合 $(E_i \cup O_i)$ 到 $(E_j \cup O_j)$ 上的二元关系,记为

$$R = \{\langle a, b \rangle \mid a \in (E_i \cup O_i) \wedge b \in (E_j \cup O_j)\}$$

其中,序偶 $\langle a, b \rangle$ 表示元素 a 依赖于元素 b ,即元素 b 的变化会引起元素 a 的变化。

2.2 依赖关系的基本分类

如前所述,和主题数据库相关的两个要素是实体和操作,主题数据库之间的依赖关系也是这两个要素之间的联系的外

^{*} 国家科技部高新技术计划项目(2005EJ000017)、国家电子信息发展基金及河北省信息产业发展计划项目(2005035025)。刘文远 教授,博士后,研究方向为信息资源规划、虚拟企业、电子商务和数据挖掘;于家新 助教,硕士,研究方向为智能计算,信息安全;徐丽娜 硕士研究生,研究方向为信息资源规划;陈国鹰 教授,博士,研究方向为信息资源规划、电子政务;

部表现。主题数据库之间的依赖关系,正是主题数据库之间的相互依赖的表现。由实体引起的依赖称为实体依赖,记为 RE;由操作引起的依赖称为操作依赖,记为 RO。这两种依赖类型实际上是主题数据库之间依赖关系的细分,其定义和描述如下。

2.2.1 实体依赖

定义 2 设两个实体 $e_u \in E_i, e_{ju} \in E_j$, 若主题数据库 D_i 中实体 e_u 的变化会引起主题数据库 D_j 中实体 e_{ju} 的变化, 则称实体 e_{ju} 实体依赖于实体 e_u , 记为 $e_{ju} \text{ RE } e_u$; 同时称实体 e_u 被实体 e_{ju} 实体依赖, 记为 $e_u \text{ REb } e_{ju}$ 。

定义 3 设实体 $e_{iu} \in E_i$, 操作 $o_{ju} \in O_j$, 若存在对主题数据库 D_i 的操作 o_{ju} 对主题数据库 D_i 中的实体 e_{iu} 的调用、修改等情况, 则称操作 o_{ju} 实体依赖于实体 e_{iu} , 记为 $o_{ju} \text{ RE } e_{iu}$; 同时称实体 e_{iu} 被操作 o_{ju} 实体依赖, 记为 $e_{iu} \text{ REb } o_{ju}$ 。

定义 4 若存在实体 $e_u \in E_i, e_{ju} \in E_j$, 且 $e_{ju} \text{ RE } e_u$ 或存在实体 $e_{iu} \in E_i$, 操作 $o_{ju} \in O_j$, 且 $o_{ju} \text{ RE } e_{iu}$, 则称主题数据库 D_j 实体依赖于主题数据库 D_i , 记为 RE_{ji} ; 同时称主题数据库 D_i 被主题数据库 D_j 实体依赖, 记为 REb_{ij} 。

2.2.2 操作依赖

定义 5 设两操作 $o_{ix} \in O_i, o_{jy} \in O_j$, 如果操作 o_{ix} 的执行当且仅当执行操作 o_{jy} 之后才能进行, 或操作 o_{jy} 的执行会触发操作 o_{ix} 的执行, 则称操作 o_{ix} 操作依赖于操作 o_{jy} , 记为 $o_{ix} \text{ RO } o_{jy}$; 同时称操作 o_{jy} 被操作 o_{ix} 操作依赖, 记为 $o_{jy} \text{ ROb } o_{ix}$ 。

定义 6 若存在操作 $o_{ix} \in O_i, o_{jy} \in O_j$, 且 $o_{ix} \text{ RO } o_{jy}$, 则称主题数据库 D_i 操作依赖于主题数据库 D_j , 记为 RO_{ij} ; 同时称主题数据库 D_j 被主题数据库 D_i 操作依赖, 记为 ROb_{ji} 。

3 依赖关系度量的理论分析

仿照文献[4-6]给出的类的依赖度和被依赖度的概念, 本文给出了主题数据库绝对依赖度和绝对被依赖度的定义。

定义 7 主题数据库 D_i 对主题数据库 D_j 的绝对依赖度 ad_{ij} 是主题数据库 D_i 对主题数据库 D_j 的实体依赖和操作依赖的依赖数目之和。

定义 8 主题数据库 D_i 对主题数据库 D_j 的绝对被依赖度 ab_{ji} 是主题数据库 D_j 对主题数据库 D_i 的实体依赖和操作依赖的依赖数目之和, 即 ad_{ji} 。

设主题数据库 D_j 中的实体个数为 r , 对主题数据库 D_i 的操作的个数为 s , 根据以上定义, 得到如下性质:

定理 1 设 $A = \{(a_i, a_j) | (a_i \in E_i, a_j \in E_j \text{ 或 } a_i \in O_i, a_j \in E_j) \text{ 且 } a_i \text{ RE } a_j\}$, $B = \{(b_i, b_j) | b_i \in O_i, b_j \in O_j \text{ 且 } b_i \text{ RO } b_j\}$, 则主题数据库 D_i 对主题数据库 D_j 的绝对依赖度, 即主题数据库 D_j 对主题数据库 D_i 的绝对被依赖度为

$$ad_{ij} = ab_{ji} = |A \cup B|$$

其有效取值范围为: $0 \leq ad_{ij} \leq (p+q) * (r+s)$ 。

证明: A 为主题数据库 D_i 对主题数据库 D_j 的所有实体依赖, B 为主题数据库 D_i 对主题数据库 D_j 的所有操作依赖。实体依赖是实体对实体和操作对实体的实体依赖, 操作依赖是操作对操作的操作依赖, 所以 A 与 B 的交集为空, $|A \cup B| = |A| + |B|$, 即为主题数据库 D_i 对主题数据库 D_j 的所有实体依赖和所有操作依赖的依赖数目之和。由定义 7 和定义 8 可知, $ad_{ij} = ab_{ji} = |A \cup B|$ 成立。

定理 1 表明这样一个性质: 对于存在依赖关系的两个主题数据库, 一个主题数据库对另一个主题数据库的绝对依赖度即为相应主题数据库对该主题数据库的绝对被依赖度。

当主题数据库 D_i 对主题数据库 D_j 的绝对依赖度 $ad_{ij} = 0$ 时, 表示主题数据库 D_i 不依赖于主题数据库 D_j , 称这种关系为零关系, 即主题数据库 D_i 零依赖于主题数据库 D_j ; 当 $ad_{ij} + ab_{ji} = (p+q) * (r+s)$ 时, 表示主题数据库 D_i 的所有实体和对主题数据库 D_i 的所有操作对主题数据库 D_j 的所有实体和对主题数据库 D_j 的所有操作都存在依赖关系, 我们称这种依赖关系为满关系, 又称主题数据库 D_i 和主题数据库 D_j 互为满依赖; 当主题数据库 D_i 对主题数据库 D_j 的绝对依赖度 $ad_{ij} = 0$ 并且绝对被依赖度 $ab_{ji} = 0$ 时, 表明主题数据库 D_i 和主题数据库 D_j 之间不存在依赖关系, 称这种关系为独立关系。

定义 9 主题数据库 D_i 对主题数据库 D_j 的相对依赖度 rd_{ij} 为主题数据库 D_i 对主题数据库 D_j 的实体依赖和操作依赖的依赖数目之和与主题数据库 D_i 到 D_j 之间全关系的数目的比值, 即 $rd_{ij} = \frac{|ad_{ij}|}{|R|}$ 。

定义 10 主题数据库 D_i 对主题数据库 D_j 的相对被依赖度 rb_{ji} 为主题数据库 D_j 对主题数据库 D_i 的实体依赖和操作依赖的依赖数目之和与主题数据库 D_i 和 D_j 之间全关系的数目的比值, 即 $rb_{ji} = \frac{|ab_{ji}|}{|R|}$ 。

根据上述定义, 显然可得到如下性质:

定理 2 $rd_{ij} = rb_{ji}, rb_{ij} = rd_{ji}$ 。

证明略。

当主题数据库 D_i 对主题数据库 D_j 的相对依赖度 $rd_{ij} = 0$ 时, 表示主题数据库 D_i 不依赖于主题数据库 D_j , 这种关系为零关系, 即主题数据库 D_i 零依赖主题数据库 D_j ; 当 $rd_{ij} + rb_{ji} = 1$ 时, 表示主题数据库 D_i 的所有实体和对 D_i 的所有操作对主题数据库 D_j 的所有实体和对 D_j 的所有操作都存在依赖关系, 这时的关系为满依赖; 当 $rd_{ij} = 0$ 且 $rb_{ji} = 0$ 时, 表明主题数据库 D_i 和主题数据库 D_j 之间不存在依赖关系, 这时的关系为独立关系。

相对度量比较科学、客观地反映了两个主题数据库之间相互依赖的程度, 它不仅考虑依赖关系的数目, 而且用笛卡尔乘积来度量两个主题数据库之间的总关系, 包含了两个主题数据库之间关系的规模大小。

4 依赖性度量的细分

在前面依赖性分类中, 将主题数据库的依赖关系细分为实体依赖和操作依赖。这两种类型的依赖可以单独存在, 也可以同时存在两个主题数据库之间或同一个主题数据库中。两个主题数据库之间或同一个主题数据库中也可以存在一个或多个实体依赖、操作依赖。在前述基础上, 可以对主题数据库之间的依赖关系度量进行细分, 有利于在 IRP 系统数据建模中更全面、更详细地把握主题数据库之间的依赖特性。

定义 11(实体依赖度) 描述一个主题数据库 D_i 对另一个主题数据库 D_j 实体依赖的程度, 即主题数据库 D_i 对主题数据库 D_j 的实体依赖的依赖数目与主题数据库 D_i 和 D_j 之间全关系数目的比值, 记为 de_{ij} 。设 $A = \{(a_i, a_j) | (a_i \in E_i, a_j \in E_j \text{ 或 } a_i \in O_i, a_j \in E_j) \text{ 且 } a_i \text{ RE } a_j\}$, 则有

$$de_{ij} = \frac{|A|}{|R|}$$

其有效取值范围为: $0 \leq de_{ij} \leq 1$ 。

定义 12(操作依赖度) 描述一个主题数据库 D_i 对另一个主题数据库 D_j 操作依赖的程度, 即主题数据库 D_i 对主题

数据库 D_j 的操作依赖的依赖数目与主题数据库 D_i 和 D_j 之间总关系数目的比值, 记为 do_{ij} 。设 $B = \{(b_i, b_j) | b_i \in O_i, b_j \in O_j \text{ 且 } b_i \text{ RO } b_j\}$, 则有:

$$do_{ij} = \frac{|B|}{|R|}$$

其有效取值范围为: $0 \leq do_{ij} \leq 1$ 。

实体依赖度、操作依赖度是主题数据库之间的依赖程度的进一步细化描述, 它比依赖度、被依赖度对主题数据库依赖程度的描述要具体。很显然, 从其定义上可以得到如下性质:

定理 3 $de_{ij} + do_{ij} = rd_{ij}$ 。

证明略。

主题数据库 D_i 对主题数据库 D_j 的实体依赖度 $de_{ij} = 0$ 时, 表示主题数据库 D_i 不实体依赖于主题数据库 D_j ; 同样, 主题数据库 D_i 对主题数据库 D_j 的操作依赖 $do_{ij} = 0$ 时, 表示主题数据库 D_i 不操作依赖于主题数据库 D_j ; 当 $de_{ij} = 0$ 且 $do_{ij} = 0$ 时, 得 $rd_{ij} = 0$, 这时的关系为零关系; 当 $de_{ij} + do_{ij} = 1$ 时, 得 $rd_{ij} = 1$, 这时的关系为满关系。

5 大规模主题数据库的分解

在信息资源规划过程中, 系统分析员和业务员在业务需求分析的基础上通过一些算法生成最初的主题数据库, 这些主题数据库只是具有一定的逻辑意义, 但是对主题数据库的规模大小基本没有考虑。下面将对分解大规模主题数据库的算法进行描述。

5.1 大规模主题数据库的分解原理

主题数据库的划分是对一个企业的所有数据资源进行划分。如果是一个大规模的企业, 只是根据一定的形式化方法一味地进行划分而不考虑主题数据库的规模, 结果往往会出现一些规模过大的主题数据库。大规模的主题数据库不易管理, 维护起来也存在一定的困难, 系统检索数据时要花费大量的时间, 会影响后期的应用软件开发的成本和周期。所以, 期待一种方法来减小主题数据库的规模, 使其规模大小合适。

通过以上定义可知, 主题数据库之间的联系紧密程度可以通过实体依赖和操作依赖来体现。那么, 对于大规模的主题数据库, 可以通过这两种度量方法来对其进行分解, 先计算主题数据库的自身的依赖度, 根据具体情况给出一个依赖度值, 以这个值为依据来判断是否进一步划分主题数据库。划分之后的各个子主题数据库之间的依赖度应该是小于某个值的, 否则划分不成功。下面将给出分解算法的描述。

5.2 大规模主题数据库的分解算法的描述

所谓的分解就是将规模较大、主题数据库自身的依赖度低的主题数据库逻辑分解成若干个主题数据库。分解后的每个主题数据库都应该具有一定的逻辑意义, 并且每个主题数据库内部依赖度较高。分解算法描述如下:

1) 给定 $rd_0 = \epsilon_0$ /* 如果主题数据库自身的相对依赖度小于 ϵ_0 , 则对主题数据库进行分解, 否则不对主题数据库进行分解 */

2) 给定 $rd_1 = \epsilon_1$ /* 如果分解后的子主题数据库之间的依赖度都小于 ϵ_1 , 则分解成功, 否则重新划分主题数据库 */

3) 计算待分解数据库 D 自身的相对依赖度 rd :

$$rd = \frac{|ad|}{|R|}$$

4) IF $rd \leq rd_0$ THEN

L: 按照某种聚类方法将主题数据库 D 分解为主题数据库 $D_1, D_2, \dots, D_n$

FOR $i = 1$ to n

FOR $j = 1$ to n

IF $rd_{ij} > rd_1$ 重新选取聚类方法, 跳到 L

END FOR

END FOR

5) 结束

如图 1 所示, 主题数据库 D 规模较大, 自身的依赖度较低, 通过分解算法将主题数据库 D 分解成子主题数据库 $D_1, D_2, \dots, D_n$ 。分解后的各个子主题数据库内部的自身依赖度还是很高的。

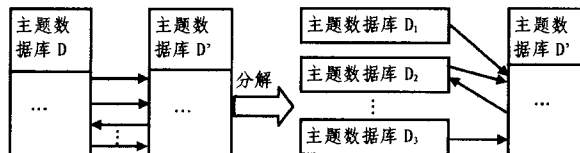


图 1 将主题数据库 D 分解并降低其依赖度

结束语 依据主题数据库的依赖关系来分解大规模主题数据库的方法, 只是为我们在主题数据库的分析和设计阶段提供指导方法, 不可能也没有必要指定一个主题数据库的依赖度为多少才合适, 或者给出依赖度合理的范围。设计者可以在不同的情况下, 针对不同的主题数据库, 利用这种方法来分解大规模的主题数据库, 使其规模大小、依赖度都比较合适。其中划分的方法有很多种, 但是哪种方法比较适合于划分主题数据库, 还没有定论, 有待分析和研究。

参考文献

- [1] 方菲, 等. 面向对象软件回归测试技术研究. 软件学报, 2001, 12(3): 372-376
- [2] 欧阳, 等. 面向对象系统的类之间依赖关系的度量研究. 计算机科学, 2004, 31(2)
- [3] 胡顺仁, 等. 基于类之间的依赖关系确定类的规模. 计算机科学, 2004, 31(3)
- [4] Jones C. Applied Software Measurement. McGraw-hill, 1986
- [5] Fenton N E, Metrics S. A Rigorous Approach. New York: Chapman & Hall, 1991
- [6] Chidamber R, Kemerer F. A metrics suite for object-oriented design. IEEE Trans on Software Engineering, 1994, 20: 45-66