数据源集成系统中全局模板的增量维护策略*

王 宁[1] 徐宏炳[2] 王能斌[2]

[1](电力自动化研究院系统研究所 南京 210003)

[2](东南大学计算机系 南京 210096)

E-mail: bbxyiqih@public1.ptt.js.cn

摘要 异构数据源集成系统需要集成包括WWW在内的各种数据源,一些半结构化数据对应的模式不仅量大且修改频繁,致使元数据的生成十分耗时.该文提出全局模板的增量维护策略.当局部模板发生变化时,利用局部模板的改变量(即模板增量)来计算全局模板的增量,进而修改全局模板.模板增量是传统实视图维护技术中增量概念的扩展,它不仅能描述对象结构特征的改变量,还能描述对象行为特征的改变量,易于表达各种数据源的模式变化.

关键词 异构数据源,数据集成,半结构化数据,数据模式,一致性,增量维护.

中图法分类号 TP311

计算机网络的普及和WWW(world-wide-web)的出现给数据集成系统的研究带来新的挑战.WWW上普遍存在的是半结构化或称自描述的数据,这些数据与传统的关系数据库或面向对象数据库中数据最大的区别在于,它们不遵循某个固定的模式[1].

Lore[2]是斯坦福大学研制的一个管理半结构化数据的数据库系统,它使用数据导则(DataGuide)[1]描述数据的模式.数据导则通过扫描数据库中的数据得到,它仅描述单个无模式数据源的结构,并不考虑多数据源集成的语义.在一个异构数据源集成系统中,全局结点集成数据的模式描述同样重要.由于全局结点往往并不真正存储对象,而是存储全局视图的定义[3],像Lore那样,利用扫描数据库的方法取得全局数据的模式显然不切实际.Versatile[4]是东南大学研制的一个基于CORBA[5]的可扩展的分布式数据集成系统.Versatile系统采用模板统一描述各种数据源数据的模式,不通过扫描数据库,而是利用局部模板之间的操作来构造集成系统的全局模板.

Yannis Papakonstantinou提出,作为一个异构数据源集成系统,其全局视图定义机制应充分考虑没有规则模式或模式经常变动数据源的要求,不仅能处理模式的不规则性,而且能在不修改视图定义的情况下处理某些数据源的模式变化[6].由于Versatile的视图定义机制能满足上述要求,因此,系统除了像Lore那样考虑局部模板与数据之间的一致性维护外,还需考虑当局部模板发生变化时,全局模板的维护策略.

在传统的数据库系统中,模式信息相对于数据少得多,而且模式修改也不频繁.然而,正如Serge Abiteboul在文献[7]中所举例说明的那样,半结构化数据的模式与传统数据库系统中数据的模式不同,它不仅量大,而且修改频繁,在有些情况下,模式信息量甚至超过数据量.在数据仓库实视图维护技术中,由于增量维护技术[8,9]大大降低了实视图刷新的代价,因而得到普遍使用.异构数据源集成系统要集成各种数据源的数据,有些局部数据源,例如WWW,其模板不仅大,而且修改频繁,为减少全局模板的刷新代价,宜选用增量维护策略.

为集成各种数据源的数据,Versatile中的全局对象及模板均基于带根连通有向图,然而,已有研究[10,11]仅给出包代数下实视图增量表达式的生成方法.本文从数据源集成系统全局模板维护的角度出发,引入模板增量的

---

* 本文研究得到国家自然科学基金资助.作者王宁,女,1967年生,博士,工程师,主要研究领域为数据库系统,分布对象技术.徐宏炳,1947年生,副教授,主要研究领域为数据库应用.王能斌,1929年生,教授,博士生导师,主要研究领域为数据库及信息系统.

本文通讯联系人:王宁,南京 210003,江苏省南京市蔡家巷24号,电力自动化研究院系统研究所

本文1998-01-19收到原稿,1998-04-14收到修改稿

概念来描述不同数据库状态对应模板之间的差异,并且在模板操作的基础上讨论由局部模板增量维护全局模板的方法.

1　OIM 对象模板及模板操作

1.1　OIM(object integration model)对象模板

Versatile 是一个基于 CORBA 的可扩展的分布式数据集成系统,在 IONA 公司的 OrbixWeb 产品上,Versatile 目前正对微软公司的 SQL Server、面向对象数据库系统 Versant、文件系统、超文本数据(即 WWW 中的数据)进行包装和集成.

Versatile 采用基于带根连通有向图的对象集成模型[12](简称 OIM 对象模型)作为公共数据模型.一个 OIM 对象 O 是一个带根连通有向图,表示成 $O(r,V,E)$.图中每个结点表示对象,边表示对象与其成员之间的关系.根结点 r 是一个聚集对象,它是引用类型的,V 是该聚集对象及其所有成员对象的集合,E 是对象与其成员之间关系的集合.OIM 对象模型将元数据附在数据上,便于集成来自各种异构数据源的异构数据,特别是自描述数据.与其他基于图结构的数据模型[2,13]不同,它将数据和方法统一表示成对象,称做常规对象和方法对象.用其作为公共数据模型,集成系统不仅能集成来自各种异构数据源的异构数据,而且能集成已有的方法和应用程序.

Versatile 采用模板 $M(O)$表示 OIM 对象 O 的结构,模板本身也是 OIM 对象,其所有从根出发路径的集合中不存在同类路径*.一致模板 $EM(O)$是一种特定的模板,它能精确地反映 OIM 对象 O 的结构.然而,对于没有固定模式结构的 OIM 对象来说,一致模板或者难以确定,或者确定所需的计算量太大(一般通过扫描数据库取得),一般用模板取代.模板中有的对象,数据库中不一定存在.但是,模板不会"屏蔽"数据库中的对象,也就是说,数据库中有的对象,模板中必有沿同类路径的同类对象.因此,虽然模板可能与数据库中对象的准确结构不完全一致,但作为元数据使用仍然是合理的.

在一个异构数据源集成系统中,参与集成的数据源可能是多种多样的.有的具有显式的模式结构,如数据库系统,其模式结构可直接转化为模板;有的没有显式的模式结构,如文件系统,Versatile 系统提供模式描述语言描述其结构,它的模板可从模式描述文件取得;还有的数据源包含自描述数据,如 WWW 上的 HTML 文件,其模板可用类似于数据导则的获取方法,通过扫描数据库取得.

例 1:图 1 是反映研究生和教师情况的 OIM 对象 O_1,它的一致模板如图 2 所示.

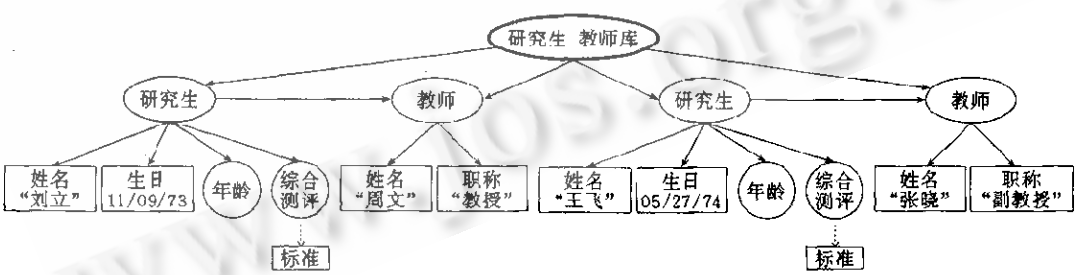

图1　研究生_教师库的OIM对象 O_1 的结构

为简单起见,图中每个结点表示一个对象.椭圆形结点和矩形结点表示常规对象,其中,椭圆形结点是复杂对象,矩形结点是原子对象,加粗的椭圆形结点表示根对象.圆形结点表示方法对象.实有向边表示复杂对象与其成员之间的引用关系,虚有向边表示方法对象与其参数之间的关系.

---

* 两条路径属同类路径是指,它们经过的结点数相等,且除第 1 个结点外,对应结点是同类结点.同类结点一般指两个结点具有相同的对象名和对象类型.如果两个方法结点是同类结点,它们除了具有上述条件之外,参数之间还需存在一一对应关系,且对应的参数结点均是同类结点.

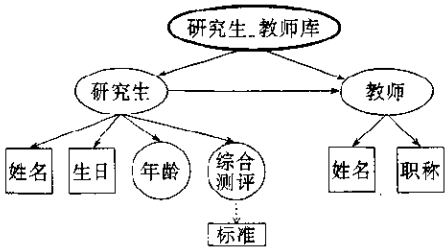

图2 研究生_教师库的OIM对象 O_i 的一致模板

1.2 模板操作

在 Versatile 系统中,全局对象由局部对象通过 OIM 对象代数[12](包括对象并、差、选择、投影、粘贴及切削 6 种操作)构造而成,全局模板由局部模板通过模板操作构造而成,为方便以后叙述,这里简单介绍模板的几个基本操作.

设有 OIM 对象 O_1,O_2 及其模板 $M(O_1),M(O_2)$,$P'_1=\{p'_1,p'_2,\ldots,p'_n\}$,$p'_1,p'_2,\ldots,p'_n,p_1,p_2$ 均是从根出发的路径.模板操作共有 5 种,它们分别是:

(1) $M(O_1)$和 $M(O_2)$的模板并,记作 $M(O_1)\oplus_m M(O_2)$,它是 O_1,O_2 对象并$(O_1\oplus O_2)$的模板.

(2) $M(O_1)$的同类模板,记作$\equiv M(O_1)$,它与 $M(O_1)$相等,既是 O_1、O_2 对象差$(O_1\Theta O_2)$的模板,也是 O_1 在某个条件 fs 下对象选择$(\sigma[fs](O_1))$的模板.

(3) $M(O_1)$在 P'_1 上的模板投影,记作 $\prod_m[p'_1,p'_2,\ldots,p'_n](M(O_1))$,它是 O_1 在 P'_1 的同类路径集上对象投影$(\prod[TLP(P'_1,O_1)](O_1))$的模板.

(4) $M(O_1)$在 P'_1 上的模板切削,记作 $\overline{\prod}_m[p'_1,p'_2,\ldots,p'_n](M(O_1))$,它是 O_1 在 P'_1 的同类路径集上对象切削 $\overline{\prod}([TLP(P'_1,O_1)](O_1))$的模板.

(5) $M(O_1)$在被粘点 p_1 对于 $M(O_2)$在粘贴点 p_2 的模板粘贴,记作 $M(O_1)\otimes_m[p_1,p_2]M(O_2)$,它是在某个粘贴条件 fp 下,O_1,O_2 对象粘贴$(O_1\otimes[TLP(p_1,O_1),TLP(p_2,O_2),fp]O_2)$的模板.

$TLP(P,O)$表示路径或路径集 P 在 OIM 对象 O 的所有从根出发路径集中的同类路径集.

在以上 5 个模板操作中,单目操作的优先级大于双目操作,优先级相同的操作按从左至右的顺序计算,但括号可改变计算顺序.

2 模板增量

Versatile 系统的全局模板是由局部模板通过操作构造而成的,局部模板发生变化时,全局模板理应作相应改变.全局模板的维护策略有重新计算(recomputation)和增量维护(incremental maintenance)两种.为减少全局模板的刷新代价,Versatile 选用增量维护策略,利用局部模板变化的部分(称做模板增量)来构造全局模板的增量,从而修改全局模板.

定义 1. 设有 OIM 对象 O_1,O_2 及其模板 $M(O_1),M(O_2)$;P_1,P_2 分别表示 $M(O_1),M(O_2)$中所有从根出发的路径的集合.如果 P_1 中任一路径都能在 P_2 中找到其同类路径,则称 $M(O_1)$包容于 $M(O_2)$,或称 $M(O_2)$包容 $M(O_1)$,记作 $M(O_1)\subseteq M(O_2)$.如果存在 $P_1\to P_2$ 的一一对应映射 Ψ,对于任一 $p_i\in P_1$,$\Psi(p_i)$与 p_i 属同类路径,则称 $M(O_1)$与 $M(O_2)$相等,记作 $M(O_1)=M(O_2)$.如果 $M(O_1)$既不包容于 $M(O_2)$,也不包容 $M(O_2)$,则称 $M(O_1)$与 $M(O_2)$互不包容.

定义 2. 设有 OIM 对象 O_1,O_2 及其模板 $M(O_1),M(O_2)$;P_1,P_2 分别表示 $M(O_1),M(O_2)$中所有从根出发的路径的集合,P_1 除 P_2 外的路径集 $P_1(\bar{p}_2)=\{p_i\mid p_i\in P_1\wedge\neg\exists\, p_j(p_j\in P_2\wedge p_i$ 与 p_j 是同类路径$)\}$.

定义 3. 设有 OIM 对象 O_1,O_2 及其模板 $M(O_1),M(O_2)$;P_1,P_2 分别表示 $M(O_1),M(O_2)$中所有从根出发的路径的集合.

(1) 如果 $M(O_1)$不包容于 $M(O_2)$,存在一个 OIM 对象 MB,P 是 MB 中所有从根出发的路径的集合,P 中不含同类路径,且存在 $P_1(\bar{p}_2)\Downarrow\to P$ 的一一对应映射 f,对于任一 $P_i\in P_1(\bar{p}_2)\Downarrow$,$f(P_i)$与 P_i 属同类路径,则称 MB 是 $M(O_1)$和 $M(O_2)$的模板差,记作 $M(O_1)\Theta_m M(O_2)$.

(2) 如果 $M(O_1)$包容于 $M(O_2)$,$M(O_1)$和 $M(O_2)$的模板差为空,记作$\varnothing$.

需要说明的是,$P\downarrow$ 表示路径集 P 中所有路径内涵*的并集,称作 P 的内涵.空模板是一个空 OIM 对象,该 OIM 对象只有一个根结点,记作$\varnothing$,不同于空集符号$\varnothing$.

随着用户的增、删、改操作,数据库中数据会发生变化.如果将 OIM 对象看成时间的函数,OIM 对象 O 在时刻 t 的值记作 $O(t)$,OIM 对象 O 在时刻 t 的模板反映该对象在时刻 t 的结构,记作 $M(O,t)$.

定义 4. 设有 OIM 对象 O 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O,t_1),M(O,t_2)$.如果存在模板 MB_1,MB_2,满足以下两个条件:

(1) $MB_1\Theta_m MB_2=MB_1$,

(2) $M(O,t_1)\oplus_m MB_1\Theta_m MB_2=M(O,t_2)$,

则称 MB_1 为 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板正增量,记作 $+\underset{t_1\rightarrow t_2}{\delta}M(O)$;$MB_2$ 为 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板负增量,记作 $-\underset{t_1\rightarrow t_2}{\delta}M(O)$.

定义 4 的条件(1)保证,在同一个时间段,不允许 OIM 对象的模板正增量和负增量之间存在相同量,以免在增量维护时引起冲突.

例 2:假设 OIM 对象 O_1 在时刻 t_1,t_2 的一致模板分别如图 2、图 3 所示,根据定义 4,图 4、图 5 中的模板分别是 O_1 在 $t_1\rightarrow t_2$ 时间段的模板正增量和模板负增量.

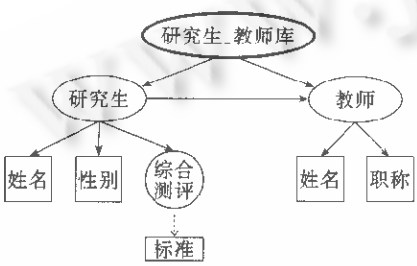

图3　OIM对象O_1在时刻 t_2的一致模板

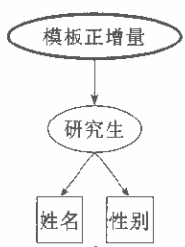

图4　O_1 在时间段 $t_1\rightarrow t_2$ 的模板正增量

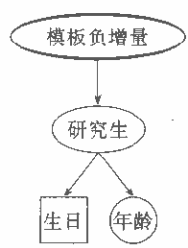

图5　O_1 在时间段 $t_1\rightarrow t_2$ 的模板负增量

定理 1. 设有 OIM 对象 O 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O,t_1),M(O,t_2)$.$M(O,t_2)\Theta_m M(O,t_1)$ 是 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板正增量,$M(O,t_1)\Theta_m M(O,t_2)$ 是 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板负增量.

证明:根据定义 4,必须证明以下两个结论成立:

(1) $(M(O,t_2)\Theta_m M(O,t_1))\Theta_m(M(O,t_1)\Theta_m M(O,t_2))=M(O,t_2)\Theta_m M(O,t_1)$,

(2) $M(O,t_1)\textcircled{+}_m(M(O,t_2)\Theta_m M(O,t_1))\Theta_m(M(O,t_1)\Theta_m M(O,t_2))=M(O,t_2)$.

首先证明结论(1).

假设 P_1 表示 $(M(O,t_2)\Theta_m M(O,t_1))\Theta_m(M(O,t_1)\Theta_m M(O,t_2))$ 中所有从根出发至叶结点的路径的集合,P_2 表示 $M(O,t_2)\Theta_m M(O,t_1)$ 中所有从根出发至叶结点的路径的集合.除此以外,其他模板 MB 的所有从根出发至叶结点的路径的集合用 $P(MB)$ 表示.

根据模板差的定义,对于任一 $p_i\in P_1$,一定存在唯一的 $p_j\in P_2$,p_i 与 p_j 属于同类路径.也即存在 $P_1\rightarrow P_2$ 的映射 f,对于任一 $p_i\in P_1$,$f(p_i)$ 与 p_i 属于同类路径.由于任何模板从根出发的路径集中不存在同类路径,映射 f 是单射无疑.下面用反证法证明 f 是满射的.

如果 f 不是满射的,则至少存在一个 $p_k\in P_2$,找不到任何 $p_m\in P_1$,使 p_k 与 p_m 属于同类路径.因此,p_k 一定与 $P(M(O,t_1)\Theta_m M(O,t_2))$ 中某一路径同类,即 p_k 与 $P(M(O,t_1))$ 中某一路径同类,而不与 $P(M(O,t_2))$ 中任何路径同类,这与 $p_k\in P_2$ 矛盾.所以,f 是满射的.

从以上证明可知:存在 $P_1\rightarrow P_2$ 的一一对应映射 f,对于任一 $p_i\in P_1$,$f(p_i)$ 与 p_i 属于同类路径.因此,结论(1)成立.

---

* 一条路径 p 的内涵 $p\downarrow$ 是指所有从 p 的始点出发且包含于 p 的路径集合.

同理可证结论(2). □

定义 5. 设有 OIM 对象 O 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O,t_1),M(O,t_2)$,称 $M(O,t_2)\Theta_m M(O,t_1)$ 为 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板净正增量,记作 $+n\underset{t_1\rightarrow t_2}{\delta}M(O)$,称 $M(O,t_1)\Theta_m M(O,t_2)$ 为 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板净负增量,记作 $-n\underset{t_1\rightarrow t_2}{\delta}M(O)$.

根据定理 1,模板净正增量是一种特定的模板正增量,模板净负增量是一种特定的模板负增量.一个 OIM 对象在某时间段的模板正增量和负增量可能有冗余,即正增量中包含已有的量或负增量中包含原先没有的量,模板净正增量与净负增量不会存在这样的冗余.图 5 给出的是 OIM 对象 O_1 在 $t_1\rightarrow t_2$ 时间段的模板净负增量,图 4 仅给出 O_1 在同样时间段的模板正增量而非净正增量.

根据上述定义,下面 4 个结论是显然的.

(1) 当且仅当 $M(O,t_2)\subseteq M(O,t_1)$,$+n\underset{t_1\rightarrow t_2}{\delta}M(O)=\varnothing$,也即 OIM 对象 O 的模板在 $t_1\rightarrow t_2$ 时间段没有净正增量.

(2) 当且仅当 $M(O,t_1)\subseteq M(O,t_2)$,$-n\underset{t_1\rightarrow t_2}{\delta}M(O)=\varnothing$,也即 OIM 对象 O 的模板在 $t_1\rightarrow t_2$ 时间段没有净负增量.

(3) 当且仅当 $M(O,t_1)=M(O,t_2)$,OIM 对象 O 的模板在 $t_1\rightarrow t_2$ 时间段既没有净正增量也没有净负增量.

(4) 当且仅当 $M(O,t_1)$ 与 $M(O,t_2)$ 互不包容,OIM 对象 O 的模板在 $t_1\rightarrow t_2$ 时间段既有净正增量又有净负增量.

定义 6. 设有 OIM 对象 O 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O,t_1),M(O,t_2)$. 二元组 $\langle+\underset{t_1\rightarrow t_2}{\delta}M(O),-\underset{t_1\rightarrow t_2}{\delta}M(O)\rangle$ 称做 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板增量,记作 $\underset{t_1\rightarrow t_2}{\Delta}M(O)$. 二元组 $\langle+n\underset{t_1\rightarrow t_2}{\delta}M(O),-n\underset{t_1\rightarrow t_2}{\delta}M(O)\rangle$ 称做 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板净增量,记作 $n\underset{t_1\rightarrow t_2}{\Delta}M(O)$. 如果 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段既没有模板正增量也没有模板负增量,则称 OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板增量为空,即 $\underset{t_1\rightarrow t_2}{\Delta}M(O)=\varnothing$.

3 全局模板的增量维护策略

Versatile 的全局模板是由局部模板通过模板操作构造而成,构造全局模板的模板操作共有 5 种,这一节主要在模板操作的基础上讨论由局部模板增量维护全局模板的方法.

定义 7. 设有 OIM 对象 O 及其模板 $M(O)$,P 是 $M(O)$ 中所有从根出发至叶结点的路径集合,p_1,p_2 分别是两条路径,P 的路径 p_2 对于 p_1 的替换 P_1(记作 $\overset{p_1}{p_2}[P]$)按如下步骤生成:

(1) $P_1=P$;

(2) 构造一个集合 P',$P_1=P_1-P'$,其中

$$P'=\{p_i\mid p_i\in P\wedge p_i\Downarrow \text{中包含 } p_2 \text{ 的同类路径}\};$$

(3) 构造一个集合 P'',$P_1=P_1\cup P''$,其中

$P''=\{p_1p_i\mid p_i$ 是 $M(O)$ 中从 p_k 的终点出发至叶结点的路径 $\wedge p_k\in P\Downarrow\wedge p_k$ 与 p_2 是同类路径$\}$.

上述定义中,p_1p_i 称做路径 p_1 与 p_i 的粘连,是由 p_1 的终点与 p_i 的始点相连而成的路径.

定义 8. 设有 OIM 对象 O 及其模板 $M(O)$,P 是 $M(O)$ 中所有从根出发至叶结点的路径集合,p_1,p_2 分别是两条路径.对于模板 MB 以及 MB 中所有从根出发至叶结点的路径集合 P_1,如果存在 $\overset{p_1}{p_2}[P]\rightarrow P_1$ 的一一对应映射 f,对于任一 $P_i\in\overset{p_1}{p_2}[P]$,$f(P_i)$ 与 P_i 属于同类路径,则称 MB 是 $M(O)$ 的路径 p_2 对于 p_1 的替换,记作 $\overset{p_1}{p_2}[M(O)]$.

定理 2. 设有 OIM 对象 O_1 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O_1,t_1),M(O_1,t_2)$,OIM 对象 O_2 及其在时刻 t_1,t_2 的模板 $M(O_2,t_1),M(O_2,t_2)$,$P'_1=\{p'_1,p'_2,\ldots,p'_n\}$,$p'_1,p'_2,\ldots,p'_n,p_1,p_2$ 均是从根出发的路径,fs 和 fp 分别表示选择条件和粘贴条件,OIM 对象 O 在 $t_1\rightarrow t_2$ 时间段的模板增量 $\underset{t_1\rightarrow t_2}{\Delta}M(O)$ 用 $\langle+\underset{t_1\rightarrow t_2}{\delta}M(O),$

$-\underset{t_1\to t_2}{\delta} M(O)$)表示.

(1) 当 $O=O_1\oplus O_2$ 时,

$$+\underset{t_1\to t_2}{\delta} M(O)=(+\underset{t_1\to t_2}{\delta} M(O_1)\oplus_m+\underset{t_1\to t_2}{\delta} M(O_2)),$$

$$-\underset{t_1\to t_2}{\delta} M(O)=(-\underset{t_1\to t_2}{\delta} M(O_1)\Theta_m M(O_2,t_2))\oplus_m(-\underset{t_1\to t_2}{\delta} M(O_2)\Theta_m M(O_1,t_2)).$$

(2) 当 $O=O_1\Theta O_2$ 或 $O=\sigma[fs](O_1)$ 时,

$$+\underset{t_1\to t_2}{\delta} M(O)=-\underset{t_1\to t_2}{\delta} M(O_1),$$

$$-\underset{t_1\to t_2}{\delta} M(O)=-\underset{t_1\to t_2}{\delta} M(O_1).$$

(3) 当 $O=\prod[p'_1,p'_2,\ldots,p'_n](O_1)$ 时,

$$+\underset{t_1\to t_2}{\delta} M(O)=\prod\nolimits_m[TLP(P'_1,+\underset{t_1\to t_2}{\delta} M(O_1))](+\underset{t_1\to t_2}{\delta} M(O_1)),$$

$$-\underset{t_1\to t_2}{\delta} M(O)=\prod\nolimits_m[TLP(P'_1,-\underset{t_1\to t_2}{\delta} M(O_1))](-\underset{t_1\to t_2}{\delta} M(O_1)).$$

(4) 当 $O=\overline{\prod}[p'_1,p'_2,\ldots,p'_n](O_1)$ 时,

$$+\underset{t_1\to t_2}{\delta} M(O)=\overline{\prod}\nolimits_m[TLP(P'_1,+\underset{t_1\to t_2}{\delta} M(O_1))](+\underset{t_1\to t_2}{\delta} M(O_1)),$$

$$-\underset{t_1\to t_2}{\delta} M(O)=\overline{\prod}\nolimits_m[TLP(P'_1,-\underset{t_1\to t_2}{\delta} M(O_1))](-\underset{t_1\to t_2}{\delta} M(O_1)).$$

(5) 当 $O=O_1\otimes[p_1,p_2,fp]O_2$ 时,假设

$$ps=TLP(p_2,\prod\nolimits_m[TLP(p_2,M(O_2,t_2))]M(O_2,t_2)),$$

$$pt=TLP(p_1,-\underset{t_1\to t_2}{\delta} M(O_1)),$$

$$px=TLP(p_1,M(O_1,t_2)),$$

$$py=TLP(p_2,-\underset{t_1\to t_2}{\delta} M(O_2)),$$

$$pm=TLP(p_1,(-\underset{t_1\to t_2}{\delta} M(O_1)\Theta_{m_{pt}^{px}}[\prod\nolimits_m[TLP(p_2,M(O_2,t_2))]M(O_2,t_2)])),$$

$$pn=TLP(p_2,(-\underset{t_1\to t_2}{\delta} M(O_2)\Theta_{m_{py}^{ps}}[M(O_1,t_2)])),$$

$$+\underset{t_1\to t_2}{\delta} M(O)=+\underset{t_1\to t_2}{\delta} M(O_1)\otimes_m[TLP(p_1,+\underset{t_1\to t_2}{\delta} M(O_1)),TLP(p_2,+\underset{t_1\to t_2}{\delta} M(O_2))]+\underset{t_1\to t_2}{\delta} M(O_2),$$

$$-\underset{t_1\to t_2}{\delta} M(O)=(-\underset{t_1\to t_2}{\delta} M(O_1)\Theta_{m_{pt}^{px}}[\prod\nolimits_m[TLP(p_2,M(O_2,t_2))]M(O_2,t_2)])$$

$$\otimes_m[pm,pn](-\underset{t_1\to t_2}{\delta} M(O_2)\Theta_{m_{px}^{py}}[M(O_1,t_2)]).$$

定理 2 的证明方法与定理 1 类似,限于篇幅,不一一赘述.

定理 2 实际上给出了由局部模板增量构造全局模板增量的方法.其中使用同类路径集只是为了描述方便.在 Versatile 中,对象(模板)投影、切削、粘贴操作使用的路径是由一组对象名通过“:”连接而成的路径表达式[13],无需真正求同类路径集.

全局对象在取得 $t_1\to t_2$ 时间段的模板增量后,根据定义 4 的(2),可以求得全局对象模板在 t_2 时刻的值.而且,由定义 4 很容易推得以下定理.

定理 3. 设有 OIM 对象 O 及其在时刻 $t_1,t_2(t_1<t_2)$ 的模板 $M(O,t_1),M(O,t_2)$,如果 O 在 $t_1\to t_2$ 时间段的模板正增量和模板负增量分别是 MB_1,MB_2,则 $M(O,t_2)=M(O,t_1)\Theta_m MB_2\oplus_m MB_1=M(O,t_1)\oplus_m MB_1\Theta_m MB_2$.

定理 3 说明,通过对 OIM 对象 O 在时刻 t_1 的模板添加 $t_1\to t_2$ 时间段的模板正增量(用模板并操作)以及去除模板负增量(用模板差操作),可以得到 O 在时刻 t_2 的模板,其结果与添加和去除的次序无关.

由定理 2,全局对象在 $t_1\to t_2$ 时间段模板正增量的计算较为容易,而模板负增量的计算相对复杂.除设计较

优的算法外，还可利用模板和一致模板的区别来减少全局模板维护的工作量. Versatile 系统区分模板和一致模板的概念，在对模板一致性要求不是很高的情况下，可采用放弃部分局部模板负增量(假设该局部模板负增量为空)的方法取得修改过的全局模板. 这样得到的模板可能不是一致模板，这种不一致是指模板中有的对象在数据库中不一定存在. 事实上，一个可扩展的异构数据源集成系统要集成包括 WWW 在内的各种数据源，有些数据源模式量大且变化频繁，一味追求模板一致性显然不切实际. Versatile 系统的模板保证不会"屏蔽"数据库中的对象，数据库中有的对象，模板中必有沿同类路径的同类对象. 采用放弃部分模板负增量的维护方法虽然在某种程度上牺牲了模板的一致性，却能大大减少维护工作量，而且非一致模板仅影响查询效率，不会造成严重错误，仍然可以作为集成系统的元数据使用. 当然，哪些局部模板负增量可以放弃要视具体情况而定，一般放弃 WWW 等模式量大且变化频繁的数据源的模板负增量.

4 结束语

Versatile 是一个可扩展的异构数据源集成系统，需要集成包括数据库系统以及 WWW 在内的各种数据源，有些局部数据源的模板不仅量大，而且修改频繁. 为减少全局模板的刷新代价，增量维护策略是必然的选择.

本文从数据源集成系统全局模板维护的角度出发，引入模板增量的概念描述不同数据库状态对应模板之间的差异. 模板增量是传统实视图维护技术中"增量"(delta)概念的扩展，由模板正增量和模板负增量两部分组成. 它不仅能描述对象结构特征的"增量"，还能描述对象行为特征的"增量"，易于表达各种数据源的模式改变量. 除此以外，本文还在模板操作的基础上讨论由局部模板增量维护全局模板的方法. 进一步的研究将给出计算全局模板增量的优化算法，在整个系统中快速计算出所有全局模板的增量数据.

参考文献

1 Goldman Roy, Widom Jenniferc. DataGuides: enabling query formulation and optimization in semistructured databases. In: Bocca Jorge, Jarke Matthias, Zaniolo Carlo eds. Proceedings of the 23rd International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1997. 436～445

2 McHugh J, Abiteboul S, Goldman R *et al*. Lore: a database management system for semistructured data. Sigmod Record, 1997, 26(3): 39～53

3 Motro Amihai. Superviews: virtual integration of multiple databases. IEEE Transactions on Software Engineering, 1987, SE-13(7): 785～798

4 Wang Ning, Chen Yin, Yu Ben-quan *et al*. Versatile: a scaleable CORBA-based system for integrating distributed data. In: Zhou Li-zhu ed. Proceedings of the 1997 IEEE International Conference on Intelligent Processing Systems. Beijing: Tsinghua University Press, 1997. 1589～1593

5 Otte Randy, Pafrik Paul, Roy Mark. Understanding CORBA. Englewood Cliffs, NJ: Prentice Hall, Inc., 1996

6 Papakonstantinou Yannis, Garcia-Molina Hector, Ullman Jeffrey. MedMaker: a mediator system based on declarative specification. In: Su Stanley Y W ed. Proceedings of the 1996 International Conference on Data Engineering. Los Alamitos, CA: IEEE Computer Society Press, 1996. 132～141

7 Abiteboul Serge. Querying semi-structured data. In: Afrati Foto, Kolaitis Phokion eds. Lecture Notes in Computer Science, Proceedings of the 6th International Conference on Database Theory. Heidelberg: Springer-Verlag, 1997. 1～18

8 Griffin T, Libkin L. Incremental maintenance of views with duplicates. In: Jagadish H V, Mumick Inderpal Singh eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: the Association for Computing Machinery, Inc., 1995. 340～351

9 Gupta A, Mumick I S, Subrahmanian V S. Maintaining views incrementally. In: Jagadish H V, Mumick Inderpal Singh eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: the Association for Computing Machinery, Inc., 1993. 157～166

10 Hull Richard, Jacobs Dean. Language constructs for programming active databases. In: Lohman Guy M, Sernadas Amilcar, Camps Rafael eds. Proceedings of the 17th International Conference on Very Large Data Bases. San Mateo: Morgan Kaufmann Publishers, 1991. 455～468

11 Ross Kenneth A, Srivaatava Diversh, Sudarshan S. Materialized view maintenance and integrity constraint checking: trading space for time. In: Jagadish H V, Mumick Inderpal Singh eds. Proceedings of the ACM SIGMOD International

Conference on Management of Data. New York: the Association for Computing Machinery, Inc., 1996. 447～458

12 王宁,徐宏炳,王能斌. 基于带根连通有向图的对象集成模型及代数. 软件学报,1998,9(12):894～898
(Wang Ning, Xu Hong-bing, Wang Neng-bin. A data model and algebra for object integration based on a rooted connected directed graph. Journal of Software, 1998,9(12):894～898)

13 Buneman P, Davidson S, Hillebrand G *et al*. A query language and optimization techniques for unstructured data. In: Jagadish H V, Mumick Inderpal Singh eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: the Association for Computing Machinery, Inc., 1996. 505～516

Incremental Maintenance of Global Templates in Heterogeneous Data Integration System

WANG Ning[1]　XU Hong-bing[2]　WANG Neng-bin[2]

[1](*Nanjing Automation Research Institute　System Control Corporation　Nanjing*　210003)
[2](*Department of Computer Science　Southeast University　Nanjing*　210096)

Abstract　A heterogeneous data integration system can integrate semi-structured data which usually have large and changeable metadata, and generation of metadata is very time consuming. The metadata of Versatile are expressed in the form of templates. An incremental strategy for maintenance of templates is proposed, which can update the templates based on the changes (or "delta" in this paper) of the template rather than regeneration of all the templates from scratch. Different from deltas used for traditional incremental maintenance of materialized views, template deltas are able to describe the differences between not only the structures but also the behaviors of objects. It can express the differences between templates in various data sources more easily.

Key words　Heterogeneous data sources, data integration, semistructured data, data schemata, consistency, incremental maintenance.