

多数据库系统中基于 XIDM 的模式映射方法研究

李瑞轩¹ 卢正鼎¹ 肖卫军¹ 李兵²

¹(华中科技大学计算机科学与技术学院 武汉 430074)

²(武汉大学计算机软件工程国家重点实验室 武汉 430072)

(rxli @public1wh1hb1cn)

摘要 建立多数据库系统中多种模式间的映射是多数据库解决模式异构性的关键,而选用一种适用于这些映射的集成数据模型又是模式映射的基础。为了集成数据库、文件系统、Web 文档等多种异构系统中的数据,提出了一种基于 XML 的集成数据模型(XIDM)作为多数据库系统的公共数据模型。基于 XIDM 模型,给出了全局模式与输出模式、输出模式与局部模式之间的模式映射方法,并对全局映射和局部映射在多数据库原型系统 Panorama 中的实现进行了示例。最后从数据模型的集成能力和模式映射的可操作性方面对基于 XIDM 的模式映射方法进行了性能比较和分析。

关键词 多数据库系统;模式映射;集成数据模型;可扩展标记语言

中图法分类号 TP311.113

Research on XIDM2Based Schema Mapping in Multidatabase Systems

LI Rui2Xuan¹, LU Zheng2Ding¹, XIAO Wei2Jun¹, and LI Bing²

¹(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074)

²(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072)

Abstract Establishing schema mappings among different schemas is a key to solve the problem of schema heterogeneity in multidatabase systems. Since an integrated data model suitable for the transformation is the basis of schema mappings, an XML2based integrated data model (XIDM) is presented as the common data model to integrate multiple heterogeneous data sources such as database tables, files and web documents in a multidatabase system. An XIDM2based approach of schema mapping among global schemas, export schemas, and local schemas is also given. Then, the illustration and implementation of global and local mappings in multidatabase prototype —Panorama system are shown, and finally, the performance evaluation and analysis of XIDM2based schema mapping are also discussed.

Key words multidatabase systems; schema mapping; integrated data model; extensible markup language (XML)

1 引言

多数据库系统是多个预先存在的、异构的、分布的数据库系统的联合,它主要解决的问题是如何在多个局部数据库系统之间实现数据的共享和集成^[1]。在多数据库系统中,由于各局部数据库模式

是由不同的用户,在不同的时间和地点,基于不同的数据模型独立地设计的,它们之间可能存在着各种差异和冲突。为了实现用户对多数据库系统的透明访问,需要研究一种方法能在多数据库全局层屏蔽这些差异和冲突。但是,不能简单地通过修改局部数据库模式来解决这一问题,因为多数据库系统的一个基本特征是要保证各局部数据库的自治性,必

收稿日期:2002-11-29;修回日期:2003-06-17

基金项目:国家高性能计算基金项目(99319);国家科技攻关计划项目(2002BA103A04)

须保证那些建立在原有各自数据库之上的应用程序(开发这些应用往往已投入了巨额的资金)仍然能够继续运行。通常采用的解决办法是在多数据库系统中构造一个全局模式,这一全局模式是由各参与的成员数据库中的局部模式经过一定的模式变换得到的^[2]。

多数据库系统大多采用4级或者类似的模式结构,包括局部模式、输出模式(也叫中间模式)、全局模式(也叫联邦模式)和外模式。数据库模式一般都是根据实际应用的需要建立的,多数据库全局模式的建立也是如此。这就是说只需要提取那些对多数据库全局模式有用的局部模式信息,并进行相应的模式变换,从而建立全局模式。在多数据库系统中,全局模式构成一个虚拟库。多数据库用户只能访问全局模式,它的实际数据必须从各局部数据库系统中获得^[3]。这里,使用模式映射来表示局部模式通过输出模式集成为全局模式的相应转换。经过模式集成后,全局模式中存储的信息实际上是全局模式到输出模式、输出模式到局部模式的映射。本文提出了一种基于XML的集成数据模型(XIDM)作为多数据库的公共数据模型,给出了多数据库全局模式到输出模式及局部模式的映射方法,并给出了模式映射在原型系统Panorama中的实现示例。

2 基于XML的集成数据模型

随着计算机和网络技术的发展,需要共享和集成的信息可能不仅包括存储在数据库中的结构化数据,而且还有存储于文件系统、HTML/XML文档等系统中的半结构化和无结构化数据。它们有的具有不同的数据模型,有的则没有规范的数据模型,因此需要采用某种公共数据模型,它可以方便地表示没有预知模式的数据,以及从各种数据源集成的数据^[4]。XML是一种元标记语言,它可以描述线性表、树、图等多种数据结构,基于XML的数据模型可以集成包括数据库系统、文件系统、Web信息系统等多种异构系统中的数据。因此,我们提出了一种基于XML的集成数据模型(XIDM)作为多数据库系统的公共数据模型,并在自行设计的多数据库原型系统Panorama中进行了实现^[5]。下面对其进行描述。

定义 11 具有相同文档类型定义(DTD)或XML模式定义(XML schema)的所有元素的集合称为元素簇(element cluster)。

元素簇的概念类似于面向对象方法中类的概念,它对一类元素进行了范型化,而一个具体的元素则类似于类的一个实例(或对象)。例如, !ELEMENT Employee (# PCDATA | Name | Title) 3 和 !ATTLIST Employee No ID # REQUIRED Dept IDREF # IMPLIED Director IDREF # IMPLIED 一起给出了元素簇 Employee 的定义。

在这里引入元素簇的概念,是因为多数据库全局模式仅包含使用DTD或XML Schema定义的一系列元素的模式,而不包含实际的元素内容。XIDM模型也只对元素的模式进行定义,而不涉及元素的内容模型,下面将以元素簇为基础对XIDM模型进行描述。

定义 21 一个多数据库全局模式可以表示为一个图 $G = (Vertex, Edge)$, 其中, G 是一个置标的有向连通图, $Vertex$ 是结点(元素簇) $ECluster$ 的集合, $Edge$ 是边 $ClusterB$ 的集合。

在XIDM模型中,图 G 的结点可分为根结点、中间结点和叶结点3种。入度为零的结点是根结点,出度为零的结点是叶结点,其余是中间结点。 G 中有且仅有一个根结点,它所对应的元素簇称为根元素簇。叶结点对应的元素簇称为叶元素簇。下面给出结点的定义。

定义 31 在XIDM模型中,图 G 的一个结点(元素簇)可以表示为一个五元组 $ECluster = (Key, Attributes, SubEClusters, Qualifications, Mappings)$, 其中, Key 是元素簇 $ECluster$ 的惟一性约束定义属性(关键属性)列表; $Attributes$ 是元素簇 $ECluster$ 的属性的有序列表,且 $Key \neq Attributes$; $SubEClusters$ 是元素簇 $ECluster$ 的子元素簇的有序列表; $Qualifications$ 是元素簇 $ECluster$ 的元素所满足的限定条件集; $Mappings$ 是元素簇 $ECluster$ 的模式映射信息集合。

在多数据库系统中,全局模式中的元素簇又称为全局元素簇,输出模式中的元素簇又称为输出元素簇,它们都是虚拟的。为了讨论方便,元素簇又可简化表示为 $EC(K, A, S, Q, M)$, 其中, EC 表示 $ECluster$, K 表示 Key , A 表示 $Attributes$, S 表示 $SubEClusters$, Q 表示 $Qualifications$, M 表示 $Mappings$ 。与此对应的边的定义如下:

定义 41 在XIDM模型中,图 G 的每一条边是有向且置标的,它可以表示为一个三元组 $ClusterB = (EC_1, EC_2, Label)$, 其中, EC_1 是起始元素簇, EC_2 是结尾元素簇, EC_1, EC_2 可以分别使用元素簇

EC_1, EC_2 的 Key 表示, $Label$ 是边上的标记

在 XIDM 模型中有两类边:标记类边和引用类边。标记类边表示元素簇与其子元素簇的联系,由元素簇指向其子元素簇,采用子元素簇的标记标识。引用类边表示不同元素簇之间的引用关系,由引用元素簇指向被引用元素簇,一般采用引用元素簇的引用属性 ($Keyref$ 属性) 标识。边有时也简化表示为 $CB(EC_1, EC_2, L)$ 。

需要指出的是,公共数据模型定义了多数据库的输出概念模式,与此对应,全局数据模型则定义了多数据库的全局概念模式。公共数据模型和全局数据模型可以是相同的,也可以是不同的。为了简单起见, Panorama 系统中的公共数据模型和全局数据模型采用了相同的数据模型。

3 模式映射

在多数据库系统的 4 级模式中,如果不考虑针对不同应用的外模式,全局用户访问的是多数据库全局模式,而要获得全局访问的结果,就必须通过输出模式再到局部模式,将针对全局模式的查询转换为针对局部模式的局部查询,才能得到具体的数据。因此,在全局模式、输出模式和局部模式之间必定存在一种映射机制将它们维系起来,这就是模式映射。全局模式与输出模式之间的映射称为全局映射 (GM),输出模式与局部模式之间的映射称为局部映射 (LM)。

3.1 全局映射

在 Panorama 系统中,全局模式和输出模式均是使用 XIDM 模型进行描述的,因此,全局模式和输出模式之间的映射不涉及不同数据模型之间的转换,而只需针对不同局部模式翻译得到的输出模式进行映射。在多数据库系统中,全局映射主要分为两类:水平映射 (HM) 和垂直映射 (VM)。

定义 51 全局元素簇 $GEC(K, A, S, Q, M)$ 上的水平映射 HM 是一操作,它按照一组给定的条件 P_1, \dots, P_n 将 GEC 映射成一组输出元素簇 $EEC_1(K_1, A_1, S_1, Q_1, M_1), \dots, EEC_n(K_n, A_n, S_n, Q_n, M_n)$, 满足:

- (1) $K_1 = K_2 = \dots = K_n = K$;
- (2) $A_1 = A_2 = \dots = A_n = A$;
- (3) $S_1 = S_2 = \dots = S_n = S$;
- (4) $Q_i = (Q - P_i)$;

$$(5) M_i = LM_i$$

其中, K, K_1, \dots, K_n 分别为 GEC, EEC_1, \dots, EEC_n 的关键属性 (下同); $P_1 \dots P_n = True, i \in \{1, \dots, n\}$ 。

水平映射定义了全局元素簇所描述的元素按给定的一组条件横向映射成输出元素簇所描述的元素。

定义 61 全局元素簇 $GEC(K, A, S, Q, M)$ 上的垂直映射 VM 是一操作,它按照一组给定的属性列表,元素簇列表序偶: $U_1, V_1, \dots, U_n, V_n$ 将 GEC 映射成一组输出元素簇 $EEC_1(K_1, A_1, S_1, Q_1, M_1), \dots, EEC_n(K_n, A_n, S_n, Q_n, M_n)$, 满足:

- (1) $K_1 = K_2 = \dots = K_n = K$;
- (2) $A_i = U_i \cap K_i$;
- (3) $S_i = V_i$;
- (4) $Q_1 = Q_2 = \dots = Q_n = Q$;
- (5) $M_i = LM_i$ 。

其中, $U_i \cap A, V_i \cap S, U_1 \dots U_n \cap K = A, V_1 \dots V_n \cap S, i \in \{1, \dots, n\}$ 。

垂直映射定义了全局元素簇所描述的元素按元素簇的属性和子元素簇纵向映射成输出元素簇所描述的元素。此时,为了保证全局元素簇的可重构性,应将关键属性映射到各输出元素簇中。

多数据库全局映射可以从根元素簇开始,采用深度优先或广度优先方法遍历整个图。对于既有水平映射又有垂直映射的情况,可先水平映射后垂直映射,也可先垂直映射后水平映射。此外,边的映射有如下定理:

定理 11 如果全局元素簇 $GEC_1(K_1, A_1, S_1, Q_1, M_1)$ 和 $GEC_2(K_2, A_2, S_2, Q_2, M_2)$ 之间有边 $CB(GEC_1, GEC_2, L)$, 且它们在同一输出模式上的映射分别为 $EEC_{1i}(K_{1i}, A_{1i}, S_{1i}, Q_{1i}, M_{1i})$ 和 $EEC_{2i}(K_{2i}, A_{2i}, S_{2i}, Q_{2i}, M_{2i})$, 则输出元素簇 EEC_{1i} 和 EEC_{2i} 之间也有边 $CB_i(EEC_{1i}, EEC_{2i}, L)$ 。

证明: 略。

多数据库系统中的全局映射与分布式数据库中的分片操作类似,但并不完全相同。例如在水平映射中,不同的输出元素簇的限定条件可能有相交的部分,即不同的输出元素簇所描述的元素允许存在重复,因为这些元素对应应在局部模式中的概念 (对象、记录或元素等) 可能在集成到多数据库系统之前就早已存在;在垂直映射中,不仅要在全局元素簇的属性进行映射,而且还要对它的子元素簇进行映射。

312 局部映射

在进行模式集成时,需要将局部模式翻译为输出模式,在此过程中,建立了输出模式到局部模式的映射。下面主要就实际应用中常见的面向对象数据库、关系数据库和 HTML/XML 文档进行讨论。

31211 XIDM 模型与面向对象模型的映射

定义 71 输出元素簇 $EEC(K, A, S, Q, M)$ 在面向对象模型上的局部映射 LM 是一操作,它按照一组转换函数 f_A, f_R, f_B, f_Q 将 EEC 映射成一局部类 $LC(K, A, R, B, Q, M)$, 满足:

- (1) $K = f_A(K)$;
- (2) $A = f_A(A) \quad f_A(S)$;
- (3) $R = f_R(A)$;
- (4) $B = f_B(S)$;
- (5) $Q = f_Q(Q)$;
- (6) $M = \S 1$

其中, K 为 LC 的关键属性或 OID, A 为 LC 的属性列表, R 为 LC 的联系列表, B 为 LC 的方法列表, Q 为 LC 的限定条件集, M 为模式映射信息。函数 f_A, f_R, f_B, f_Q 分别将元素簇的属性、子元素簇、限定条件转换为类的属性、联系、方法及限定条件。

和全局映射一样,对输出元素簇进行局部映射时,也可以从根元素簇开始进行映射,采用深度优先或广度优先方法遍历整个图。根元素簇被映射为类的聚集。对任意输出元素簇 $EEC(K, A, S, Q, M)$, 如果 S 中某个子元素簇已经由函数 f_A 转换为类的属性或由 f_B 转换为类的方法,则下次搜索到该子元素簇时不再对其进行映射。此时边的映射定理如下:

定理 21 在 XIDM 模型映射到面向对象模型时,对于输出元素簇 $EEC_1(K_1, A_1, S_1, Q_1, M_1)$ 和 $EEC_2(K_2, A_2, S_2, Q_2, M_2)$, 如果 EEC_1 和 EEC_2 之间有引用类边 $CB(EEC_1, EEC_2, L)$, 则它们分别被映射为类 LC_1 和 LC_2 , 且 LC_1 和 LC_2 之间存在联系(双向的)或继承关系(LC_1 为 LC_2 的子类); 如果 EEC_1 和 EEC_2 之间有标记类边 $CB(EEC_1, EEC_2, L)$, 则 EEC_1 被映射为局部类 LC_1 , EEC_2 被映射为 LC_1 的属性或方法。

31212 XIDM 模型与关系模型的映射

定义 81 输出元素簇 $EEC(K, A, S, Q, M)$ 在关系模型上的局部映射 LM 是一操作,它按照一组转换函数 f_A, f_Q 将 EEC 映射成一局部关系 $LR(K, A, Q, M)$, 满足:

- (1) $K = f_A(K)$;

$$(2) A = f_A(A) \quad f_A(S);$$

$$(3) Q = f_Q(Q);$$

$$(4) M = \S 1$$

其中, K 为 LR 的关键字段, A 为 LR 的字段列表, Q 为 LR 的限定条件集, M 为模式映射信息。函数 f_A, f_Q 分别将元素簇的属性、子元素簇、限定条件转换为关系的字段及限定条件。

根元素簇被映射为一个数据库实例。同样地,对任意输出元素簇 $EEC(K, A, S, Q, M)$, 如果 S 中某个子元素簇已经由函数 f_A 转换为关系的属性,则下次搜索到该子元素簇时不再进行映射。此时边的映射有如下定理:

定理 31 在 XIDM 模型映射到关系模型时,对于输出元素簇 $EEC_1(K_1, A_1, S_1, Q_1, M_1)$ 和 $EEC_2(K_2, A_2, S_2, Q_2, M_2)$, 如果 EEC_1 和 EEC_2 之间有引用类边 $CB(EEC_1, EEC_2, L)$, 则它们分别映射为关系 LR_1 和 LR_2 , LR_1 和 LR_2 之间存在引用完整性约束,且 LR_1 引用 LR_2 ; 如果 EEC_1 和 EEC_2 之间有标记类边 $CB(EEC_1, EEC_2, L)$, 若 EEC_2 为叶元素簇,则 EEC_1 被映射为关系 LR_1 , EEC_2 被映射为 LR_1 的字段,否则, EEC_1 被映射为关系 LR_1 , EEC_2 被映射为另一关系 LR_2 , 且 LR_1 引用 LR_2 。

31213 XIDM 模型与 HTML/XML 文档模型的映射

由于 XIDM 模型是基于 XML 建立的,所以 XIDM 模型与 HTML/XML 文档模型之间不存在模型间的转换,只需要进行命名、精度等方面的转换。我们仍然使用 XIDM 模型来描述 HTML/XML 文档。

定义 91 输出元素簇 $EEC(K, A, S, Q, M)$ 在 HTML/XML 文档模型上的局部映射 LM 是一操作,它按照一组转换函数 f_A, f_S, f_Q 将 EEC 映射成一局部元素簇 $LEC(K, A, S, Q, M)$, 满足:

$$(1) K = f_A(K);$$

$$(2) A = f_A(A) \quad f_A(S);$$

$$(3) S = f_S(A) \quad f_S(S);$$

$$(4) Q = f_Q(Q);$$

$$(5) M = \S 1$$

其中, K 为 LEC 的关键属性, A 为 LEC 的属性列表, S 为 LEC 的子元素簇列表, Q 为 LEC 的限定条件集, M 为模式映射信息。函数 f_A, f_S, f_Q 分别将输出元素簇的属性、子元素簇、限定条件转换为局部元素簇的属性、子元素簇及限定条件。在 XIDM

模型映射到 HTML/ XML 文档模型时,边的映射也可通过相应的命名转换等来处理

对于其他半结构化数据,一般必须经过预处理才能进行集成和映射。这种预处理主要是抽取关键的、用户感兴趣并且对于生成模式结构必需的数据。一些“骨架”数据,如 CAD/ CAM/ CAPP 系统中的设计物品名、物品的组件、组件间的关系以及设计的先前版本等信息,工作流管理系统中的业务流程、过程定义、活动、状态变迁等信息,多媒体文件系统中文件名、媒体类型、文件大小等,都可以作为数据抽取的指示信息。这些抽取出的结构数据称为模式信息,将它们用 XIDM 模型范型化,形成一个便于集成的局部模式^[6]。

4 实现示例

在我们自行设计实现的多数据库原型系统 Panorama 中,采用 XIDM 模型作为其公共数据模型,它可以集成面向对象数据库、关系数据库以及 HTML/ XML 文档等局部系统,给全局用户提供一个统一的数据描述方法^[7]。

图 1 给出了 Panorama 系统中 Company 的部分全局模式,它包含根元素簇 Company 以及全局元素簇 Employee, Department, Address 等。

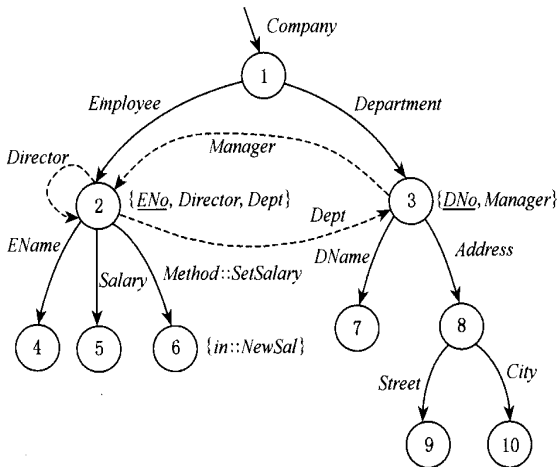


图 1 Panorama 系统中 Company 的部分全局模式

图 1 中,每个结点(元素簇)分别由 1, 2, 3, ... 表示,标记类边用实线表示,引用类边用虚线表示,每个元素簇的子元素簇由标记类边表示,元素簇的属性由一个“{}”对给出。元素簇的模式映射信息可用双虚线边表示,它连接全局模式与输出模式、输出模式与局部模式中对应的概念,边上的符号为元素簇的限定条件,由一个“[]”对给出。为简化图示,这里仅给出图 1 中结点 3 和 7 的示例,如图 2 所示:

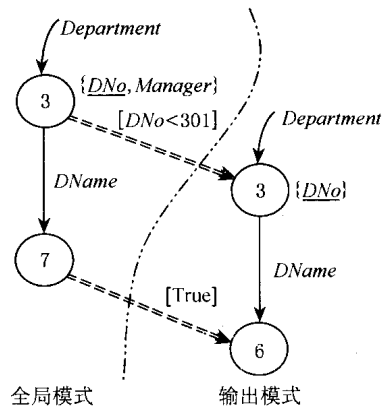


图 2 图 1 中结点 3, 7 的全局映射

在对 Company 进行全局映射时,设若全局元素簇 Employee 在 { ENo, Dept }, { EName, Method SetSalary} 和 { ENo, Director, Dept }, { EName, Salary} 上垂直映射,全局元素簇 Department 以条件 [DNo < 301] 和 [DNo > 210] 水平映射,并在 { DNo }, { DName, Address} 和 { DNo, Manager }, { DName, Address} 上垂直映射,全局元素簇 Address 以条件 [True] 和 [True] 水平映射,可得图 3 和图 4 所示的输出模式 ES₁ 和输出模式 ES₂。

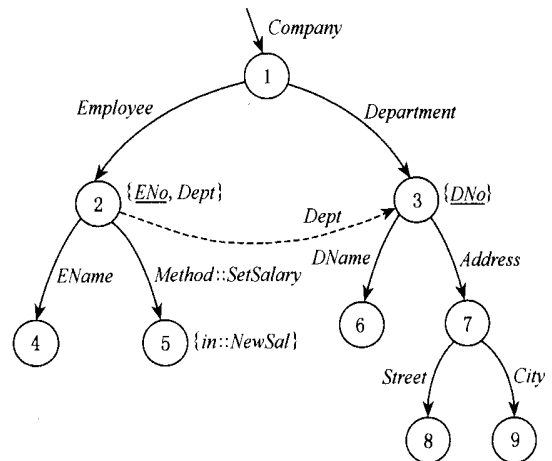


图 3 Company 映射的输出模式 ES₁

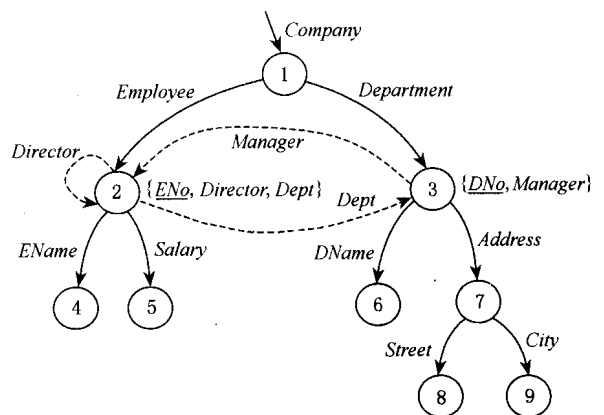


图 4 Company 映射的输出模式 ES₂

设若输出模式 ES_1 和 ES_2 分别由局部类模式和关系模式翻译得来,则可分别将其映射为如下所示的 LS_1 和图 5 所示的 LS_2

输出模式 ES_1 映射为局部类模式 LS_1 :

```
module Company {
  interface Staff (key No){
    attribute integer EmpNo;
    attribute string Name;
    relationship Department Dept
      inverse Department Contain;
    boolean SetSalary(in float NewSal);
  };
  interface Department (key Number){
    attribute string Number;
    attribute string Name;
    attribute string Struct Addr
      {string Street, string City} Address;
    relationship Set Staff Contain
      inverse Staff Dept;
  };
}
```

数据库实例: Corporation

Person				
No*	Name	Salary	Director^	Branch^

Branch			
Number*	Name	AddrID^	Manager^

Address		
AddrID*	Street	City

其中*表示关键字段, ^表示外键字段

图 5 输出模式 ES_2 映射为局部关系模式 LS_2

值得指出的是,在将输出模式映射为局部模式时,可由转换函数对元素簇、属性名等进行相应的转换,如输出模式 ES_1 中元素簇 *Employee* 映射为局部类 *Staff*, *Employee* 的属性 *EName* 映射为 *Staff* 的属性 *Name*, 输出模式 ES_2 中元素簇 *Department* 映射为局部关系 *Branch*, *Department* 的属性 *DNo* 映射为 *Branch* 的属性 *Number*

数据库实例: Corporation

由于 XIDM 模型本身是以 XML Schema 为基础建立的,所以对于 HTML/XML 文档类型,基本上不做什么转换就可以与 XIDM 模型建立映射,这里不做示例

5 性能比较

这里从数据模型和模式映射两个方面进行比较和分析

(1) 数据模型

传统的多数据库系统大多采用面向对象模型作为公共数据模型^[8],面向对象模型有丰富的语义表示,具有很强的描述能力,它可以同时集成来自多个异构数据源中的数据,新的数据源只要经过包装并描述其检索能力后,就可插入到集成系统中,系统的对象集成查询语言可以完成对象之间导航式查询以及对一些集合类型对象的查询,但面向对象模型能够管理的主要还是结构化数据,对于存储于 HTML/XML 文档、多媒体数据、集成制造等系统中的半结构化和无结构数据,使用面向对象模型来描述则比较困难

XIDM 模型是一种基于 XML 的数据模型,XML 是一种元标记语言,它允许用户为各类数据创建自己的标记,制定各种各样的置标语言,以此来创建不依赖于平台、语言的开放数据,XML 完全面向数据内容,是自描述语言,可以描述线性表、树、图等多种数据结构,从而能成为不同应用系统之间的一种通用数据接口标准,便于不同系统之间信息的交换, XIDM 模型是一种图结构的模型,可以表达更加丰富的语义,它可以表示没有预知模式的数据,更加方便地集成包括数据库系统、文件系统、Web 信息系统等多种异构系统中的数据

(2) 模式映射

多数据库系统对模式的处理通常采用模式集成的方法,它是一种自底向上的方法,包括模式翻译、模式集成等步骤,模式映射是一种自顶向下的方法,它根据应用需求建立多数据库全局模式,通过模式分析和模式变换,确立从哪些局部模式提取哪些数据,从而建立全局模式到输出模式、输出模式到局部模式的映射,与传统的模式集成方法相比,模式映射的方法更灵活、更符合实际应用的需要、更具可操作性

由于传统的多数据库系统多采用面向对象数据模型,因而研究面向对象模型与其他模型之间映射的比较多^[8,9],但要建立面向对象模型与 HTML/XML 文档、文件系统等半结构化模型之间的映射则比较困难,基于 XIDM 的模式映射方法,将模式映射分成全局映射和局部映射两个方面,全局映射主

要解决模式之间的重构问题,局部映射主要解决模式之间的冲突问题,通过形式化的方法比较全面地解决了多数据库模式集成和模式映射中必将遇到的关键问题。同时,上述模式映射方法不仅可以建立 XIDM 模型与对象模型、关系模型之间的映射,而且由于 XIDM 模型本身具有天然的自描述特性,使得公共数据模型与半结构化数据模型之间的映射更为方便。

6 结 论

模式映射是多数据库全局模式信息的一部分,它一方面给出了局部系统中的概念如何通过模式翻译和集成转换为全局模式中的概念,另一方面也给出了多数据库全局查询所要求的数据如何从局部系统中去获取。模式映射是多数据库全局查询分解处理的基础,模式映射方法的选取与多数据库系统所采用的公共数据模型紧密相关。

目前,国内外有不少关于多数据库公共数据模型和模式集成的研究,但随着新的协议和方法的不断涌现,新的应用需求不断增多,急需提出新的数据描述方法和模式集成策略。本文首先给出了一种基于 XML 的公共数据模型(XIDM),在此基础上给出了基于 XIDM 的模式映射方法,并在 Panorama 原型系统中进行了实验验证。

环境的分布性、数据描述和数据操作的异构性以及各局部成员系统的自治性,给多数据库系统的集成和处理带来了许多新的问题。比如,如何更好地提供统一的数据描述方法,如何提供更有用的查询处理办法。此外,分布异构环境下多数据库事务正确性标准、并发控制机制、安全访问控制策略等,都是值得重新研究的。

参 考 文 献

- 1 A P Sheth, J A Larson. Federated database system for managing distributed, heterogeneous, and autonomous database. ACM Computing Surveys, 1990, 22(3): 183 ~ 236
- 2 P Martin, W Powley. Database integration using multidatabase views. IBM Center for Advanced Studies 1993 Conf (CASCON '93), Toronto, 1993
- 3 S A Becker, R Gibson, N L Leist. A study of a generic schema for management of multidatabase systems. Journal of Database Management, 1996, 7(4): 14 ~ 20

- 4 M P Reddy, B E Prasad, P G Reddy *et al*. A methodology for integration of heterogeneous databases. IEEE Trans on knowledge and Data Engineering, 1994, 6(6): 920 ~ 933
- 5 Li Bing, Lu Zhengding, Xiao Weijun *et al*. An architecture for multidatabase systems based on CORBA and XML. In: Proc of the 12th Int'l Workshop on Database and Expert Systems Applications (DEXA '2001). California: IEEE Computer Society Press, 2001. 32 ~ 37
- 6 卢正鼎,李瑞轩,肖卫军. 多数据库事务处理模型中局部代理的设计与实现. 计算机研究与发展, 1998, 35(12): 1130 ~ 1134 (Lu Zhengding, Li Ruixuan, Xiao Weijun. Design and implementation of the local agent in a multidatabase transaction model. Journal of Computer Research and Development (in Chinese), 1998, 35(12): 1130 ~ 1134)
- 7 卢正鼎,李兵,肖卫军等. 基于 CORBA/XML 的多数据库系统研究与实现. 计算机研究与发展, 2002, 39(4): 443 ~ 449 (Lu Zhengding, Li Bing, Xiao Weijun *et al*. Research and implementation of a multidatabase system based on CORBA and XML. Journal of Computer Research and Development (in Chinese), 2002, 39(4): 443 ~ 449)
- 8 J L Koh, A L P Chen. An approach to querying multiple object databases. Journal of Information Science and Engineering, 2002, 18(2): 281 ~ 310
- 9 T Hader, G Sauter, J Thomas. The intrinsic problems of structural heterogeneity and an approach to their solution. VLDB Journal, 1999, 8(1): 25 ~ 43



李瑞轩 男,1974 年生,讲师,博士研究生,主要研究方向为 Web 与数据库、分布式异构系统集成。



卢正鼎 男,1944 年生,教授,博士生导师,主要研究方向为 CIMS、分布式异构系统集成、信息安全。



肖卫军 男,1974 年生,博士,讲师,主要研究方向为数据库技术、分布式异构系统集成。



李兵 男,1969 年生,博士后,副教授,主要研究方向为系统集成、计算机软件工程。