第 6 卷第 8 期
1995 年 8 月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 6, No. 8
Aug. 1995

IMDBS 系统中的多源数据模型、语言及查询处理*

王国仁 于 戈 张 斌 单吉第 郑怀远

(东北大学计算机系,沈阳 110006)

摘要 IMDBS 是一个集成式的多数据库系统,该系统是通过对已有的局部异构数据库进行信息和数据的集成,为全局用户提供共享的集成信息;同时每个局部应用仍能在被集成后的数据上运行,这样就保证了局部数据库的自治性. IMDBS 系统主要是基于源标签集成机制,并与非一范式(NF^2)模型相结合作为多源数据模型. 多源数据语言为 PSQL/NF,它是一个类 SQL 语言,非常方便于用户的使用. 该系统具有自下而上集成关系、网状和层次数据库以及自上而下地进行分布数据管理的功能,较好地解决了异构分布库系统的全局分布透明性和局部自治性的矛盾. 本文主要介绍 IMDBS 系统所采用的多源数据库模型、多源数据语言以及多源查询处理的实现技术.

关键词 源标签,非一范式数据模型,模式集成,多数据库系统,多源查询处理.

近年来,随着数据库应用技术的飞速发展,尤其是计算机集成制造系统 CIMS、计算机辅助工程设计和制造(CAD/CAM)、工程信息系统、多媒体应用等新的数据库应用领域的出现,人们已经广泛地认识到许多重要的应用系统需要存取和集成多个异构数据库系统[1-3]. 如果一个企业已在 CAD、CAM、生产管理与控制等应用中开发了大量的数据库应用软件并建立了有大量数据的数据库,而且在这些应用子系统中数据库系统是异构的和分散的[2,4],但它们之间又有许多信息需要共享与交换. 因此,如何把这些系统中的异构数据库系统集成起来以实现信息集成与共享是当前许多应用系统所迫切需要解决的实际问题[2,4,5]. 人们对异构数据库系统的信息集成已经进行了很多深入的研究工作[1,6-13],但它们大多数仅仅解决了同构数据库系统之间的信息集成问题.

为了彻底解决传统数据库系统之间的集成问题,本文提出了一种基于非一范式(NF^2)的源标签集成机制的异构数据库系统集成方法. 源标签概念是由 MIT 在文献[1]中提出的,但他们将源标签概念和关系模型相结合,仅仅解决了关系数据库系统之间的集成问题. 为了集成非关系数据源,本文对 MIT 提出的源标签概念作了必要的扩充,并与非一范式模型相

* 本文 1994-01-10 收到,1994-06-02 定稿

本研究得到 863 高技术的资助.作者**王国仁**,1966 年生,讲师,主要研究领域为分布式数据库. **于戈**,1962 年生,副教授,主要研究领域为分布式数据库. **张斌**,1964 年生,讲师,主要研究领域为数据库. **单吉第**,1936 年生,副教授,主要研究领域为数据库,网络. **郑怀远**,1931 年生,教授,主要研究领域为数据库.

本文通讯联系人:王国仁,沈阳 110006,东北大学计算机系

© 中国科学院软件研究所 http://www.jos.org.cn

结合,有效地解决了传统数据库的集成问题.对于一个分散的数据库系统来说,在集成过程中必然会遇到各类冲突问题,如命名冲突,数据表达冲突,数据量纲冲突和数据结构冲突等[2-4],使用源标签集成机制能够很好地而且很容易地解决各类冲突问题[1,14].而且源标签集成机制对自上而下地管理分布集成数据是非常方便的.在前面提到的新的数据库应用领域中,其应用系统是非常复杂的[15],既有结构化数据,又有非结构化数据,如图形、工艺规程、NC代码等.对于这些非常复杂的应用环境,采用非一范式(NF^2)数据模型作为集成数据模型是非常有效的[5,14,16-18].

多数据库集成系统依据有无集成模式可分为两大类:带全局模式的集成系统和不带全局集成模式的集成系统.由于前者透明性和信息共享程度较高,故IMDBS系统采用带全局集成模式的系统结构.在多库集成系统中有许多问题需要深入研究,如系统结构、集成机制、查询处理、多库事务管理、并发控制、数据模式和语言的转换和系统安全性等.本文主要讨论IMDBS系统中的集成模型、集成语言和查询处理技术,其它问题将有其它文章讨论.

IMBDS是一个集成式的多数据库系统.该系统是通过对已有的局部异构数据库系统进行自下而上的信息集成和自上而下的分布管理,为全局用户提供共享的集成信息,同时各个局部数据库应用程序仍能在被集成后的数据上运行,这样就保证了局部自治性.在这一系统中主要是基于源标签的集成机制,采用非一范式(NF^2)的数据模型作为多源集成数据模型,多源数据语言采用便于用户操作和使用的类SQL语言—PSQL/NF.IMDBS系统具有自下而上集成已有的关系,层次和网状数据库以及自上而下地进行分布数据管理的功能,基本上实现了关系、网状和层次数据库的集成,解决了全局分布透明性和局部自治性的矛盾.

本文第1节着重讨论IMDBS系统所采用的多源数据模型(包括源标签和非一范式的有机结合)、多源数据语言(包括多源数据定义语言和多源数据操纵语言)等内容;第2节讨论基于多源数据模型的多源关系代数;有关多源查询处理将在第3节中讨论;最后是总结.

1　IMDBS系统的多源数据模型和多源数据语言

本节我们主要讨论在IMDBS系统中所采用的多源数据模型PDM和多源数据语言PDL,并给出PDM的形式化定义和PDL的BNF形式化描述.

1.1　多源数据模型

在IMDBS系统中所采用的数据模型称为多源数据模型,多源数据模型是源标签和非一范式数据模型的有机结合.源标签在文献[1]中已作了较详细的讨论,但它仅适合于集成关系数据库系统.为了能够使用源标签来集成非关系数据源,我们对源标签的概念作了必要的扩充,并将其与非一范式数据模型有机地结合起来,以适应于集成非关系数据源的需要.

采用非一范式与源标签相结合作为集成模型是基于以下几点考虑:(1)采用源标签集成机制比较容易解决多库集成过程中的各类冲突问题;(2)非一范式作为关系模型的扩展,它本身有着简单、易懂等优点,与之相应的语言也容易掌握;(3)异构数据库的集成问题是非常复杂的.因此,针对各种数据库系统所支持的数据模型的特征,采用分层集成的方法在实际应用中是行之有效的.对集成关系数据库采用关系模型加上源标签集成机制作为集成模型;对传统数据库的集成宜采用非一范式加上源标签集成机制作为集成模型;而对O—O数据

© 中国科学院软件研究所 http://www.jos.org.cn

库的集成则应采用O—O模型加上源标签机制作为集成模型;(4)对存取路径透明的关系数据库和对存取路径不透明的层次和网状数据库的存取路径可在源标签机制得到有机的统一;(5)非一范式数据模型可表示工程数据应用中的复杂数据,这样从结构上来说非一范式模型具有O—O模型的结构特征.非一范式的另一个重要特征是它具有很强的伸缩可扩性,即非一范式很容易退化为关系模型,在其基础上加上对象管理机制即可扩展为O—O模型.

定义 1. 多源模式

一个多源模式 P 被定义为一组规则的集合:$P=\{R_1,R_2,\cdots,R_m\}$. 其中 $R_j(1\leqslant j\leqslant m)$ 是一个形如 $R_j=(R_{j1},R_{j2},\cdots,R_{jn})$ 的规则,R_j 和 R_{ji} 称为属性. 一个属性既可以出现在规则的左部,也可以出现在规则的右部. 仅出现在规则右边的属性称为零阶多源属性;出现在规则左边的属性称为高阶多源属性;仅出现在规则左边的属性称为广义高阶多源属性. 在多源模式 P 中,如果 R_{ji} 是一个高阶多源属性,则它又是一组规则的集合,这是一种嵌套递归结构;如果 R_{ji} 是一个零阶多源属性,则用一个二元组:$R_{ji}=\langle PA_j,MA_j\rangle$ 来定义,其中 PA_j 为零阶多源属性名,MA_j 为 PA_j 的源标签,源标签是一个局部属性描述集合,描述一个零阶多源属性的数据来源. 下面给出源标签的定义.

定义 2. 源标签

一个源标签被定义为一个四元组的集合:$MA_j=\{\langle LDT,LD,LS,LA\rangle\}$. 其中 LDT 为局部数据库类型,R:表示关系数据库,H:表示层次数据库,N:表示网状数据库;LD 为局部数据库名,LS 为局部模式名,LA 的内容随数据库类型的不同而不同,对 R 类而言,LA 表示局部属性名;对 H 类而言,LA 的格式为:父片段记录名.字段名,若父片段记录名为 $ROOT$ 时,字段名为二级树中根片段记录的字段,否则字段名为二级树中子女片段的字段;对 N 类而言,LA 的格式为:系名.系首记录名.字段名,当系首记录名为 $SYSTEM$ 时,表示该系类型为奇异系,且字段名为该奇异系属记录的字段,当系首记录名为 $ROOT$ 时,字段名为系首记录中的字段,否则字段名为系属记录中的字段. 一个源标签可以清楚地描述一个零阶多源属性的数据来源,可以来自于一个局部数据库,也可以来自于多个局部数据库.

下面用一个例子来说明多源数据模型是如何来表示集成模式的. 假设在某个关系数据库系统中有两个局部关系模式:CORP(cno,cname,business),branch(cn,bno,blocation)分别是数据库CD和BD中的两个关系,在某个层次数据库系统中有一个层次局部模式DEPT,在DEPT中有两个片段(记录类型)dept和emp,如图1所示. 它们在数据库ED中. 下面把局部关系模式和局部层次模式集成为下面的非一范式多源模式(图2).

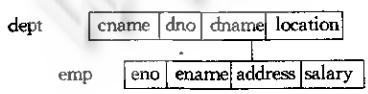

图1 一个层次数据模式

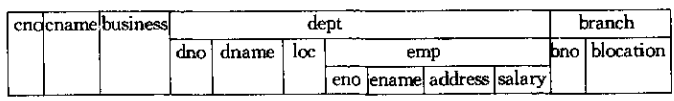

图2 一个集成后的非一范式多源数据模式

对于每一个零阶多源属性都应该有相应的源标签,每个零阶多源属性对应的源标签模式如下:

cno {⟨R,CD,corp,cno⟩}

cname {⟨R,CD,corp,cname⟩,
⟨R,BD,branch,cn⟩,

business {⟨R,CD,corp,business⟩}

dno {⟨H,ED,DEPT,ROOT.dno⟩}

dname {⟨H,ED,DEPT,ROOT.dname⟩}

© 中国科学院软件研究所 http://www.jos.org.cn

〈H,ED,DEPT,ROOT.cname〉} Location {〈H,ED,DEPT,ROOT.Location〉}
eno {〈H,ED,DEPT,dept.eno〉}　ename {〈H,ED,DEPT,dept.ename〉}
address {〈H,ED,DEPT,dept.address〉}　salary {〈H,ED,DEPT,dept.salary〉}
bno {〈R,BD,branch,bno〉}　blocation {〈R,BD,branch,blocation〉}

1.2　多源数据语言 PSQL/NF

多源数据语言主要包括两大类:多源数据定义语言和多源数据操纵语言.多源数据定义语言用来把若干个局部数据库模式集成为一个或多个多源数据模式,各个局部数据库模式之间发生的各类冲突均能在这里得到较好的解决.多源数据定义语言的 BNF 描述见附录 A.对于上面的例子其多源数据定义语句如下所示:

```
Intetgration pscheme
  corp ( item cno { (R,CD,corp,cno) }
       item cname { (R,CD,corp,cname),(R,BD,branch,cn),(H,ED,DEPT,ROOT.cname) }
       item business { (R,CD,corp,business) }
       item pscheme
            dept ( item dno { (H,ED,DEPT,ROOT.dno) }
                   item dname { (H,ED,DEPT,ROOT.dname) }
                   item Location { (H,ED,DEPT,ROOT.Location) }
                   item pscheme
                        emp (item eno { (H,ED,DEPT,dept.eno) }
                             item ename { (H,ED,DEPT,dept.ename) }
                             item address { (H,ED,DEPT,dept.address) }
                             item salary { (H,ED,DEPT,dept.salary) }
                            ) dept.emp
                   item pscheme
                    branch (item bno { (R,BD,branch,bno) }
                            item blocation { (R,BD,branch,blocation) }
                           ) corp.cname
       )
```

从上面的集成语句可以看出,存取路径透明的关系数据库和存取路径不透明的网状和层次数据库通过源标签导航描述得到了有机的统一,因为对于网状和层次类型的存取必须通过源标签和导航描述来给出其存取路径.对网状数据库而言,主要通过系名和主记录名来导航;对层次库则通过层次路径来导航.通过集成语句集成之后,网状和层次的存取路径和关系一样,对全局用户是透明的,但它们的存取路径被记录在集成字典当中,有关集成字典的讨论见文献[4,14].

当局部数据库模式被集成为多源集成模式后,就可以使用多源数据语言来操纵多源集成模式中的多源数据.实际上多源数据并不存在,它是一个虚拟概念,多源数据实际上仍保留在各个局部数据库当中.全局用户通过多源集成模式为操作基础,而局部用户则仍以局部模式为操作基础,但它们实际是在一套物理数据之上进行操作,这样就保证了局部场地自治性.在这种系统中数据的完整性和一致性问题是非常重要的,这些问题将有另文专门讨论.多源数据操纵语言采用 SQL 语言的基本结构,它支持多源数据模型,由于它引进了层次(嵌套)的概念,所以它不再是单纯的说明性语言,而成为半说明性半过程性的数据语言.多源数

© 中国科学院软件研究所 http://www.jos.org.cn

据操纵语言主要由查询语句、插入语句、更新语句和删除语句等几部分组成.多源数据操纵语言的具体语句格式见附录A.下面看一个查询的例子,在CORP这一非一范式结构中查找公司号为100的公司的名称、业务范围、分公司号以及该公司中部门号为101的部门名称、地址和该部门所有的雇员工资总和.完成这一查询的查询语句如下:

```
select cname,business, (select dname,location,(select sum(sal) from emp)
                        from dept where dno = 101),(select bno from branch)
from corp where cno = 100;
```

2 多源关系代数

在IMDBS系统中,一条多源查询请求的处理与转换是以多源关系代数为基础的.在讨论多源查询处理之前,先定义一些必要的多源关系代数运算:

(1)多源并操作($\cup^P$)

设r_1,r_2为多源模式R上的两个多源关系,X为R上所有零阶多源属性的集合,Y为X上所有高阶多源属性的集合,则:

$$r_1 \cup^P r_2=\{t\mid((\exists\, t_1)\in r_1\wedge(\exists\, t_2)\in r_2(t[X]=t_1[X]=t_2[X]\wedge t[Y]=(t_1[Y]\cup^P t_2[Y]))\vee(t\in r_1\wedge(\forall\, t')\in r_2(t[X]\neq t'[X]))\vee(t\in r_2\wedge(\forall\, t')\in r_1(t[X]\neq t'[X])))\}$$

(2)多源交操作$\cap^P$

设r_1,r_2为多源模式R上的两个多源关系,X为R上所有零阶多源属性的集合,Y为R上所有高阶多源属性的集合,则

$$r_1\cap^P r_2=\{t\mid(\exists\, t_1)\in r_1\wedge(\exists\, t_2)\in r_2(t[X]=t_1[X]=t_2[X]\wedge t[Y]=(t_1[Y]\cap^P t_2[Y])\wedge t[Y]\neq\Phi)\}$$

(3)多源差操作$-^P$

设r_1,r_2为多源模式R上的两个多源关系,X为R上所有零阶多源属性的集合,Y为R上所有高阶多源属性的集合,则

$$r_1-^P r_2=\{t\mid(\exists\, t_1)\in r_1\wedge(\exists\, t_2)\in r_2((t[X]=t_1[X]=t_2[X]\wedge t[Y]=(t_1[Y]-^P t_2[Y])\wedge t[Y]\neq\Phi)\vee(t\in r_1\wedge(\forall\, t')\in r_2(t[X]\neq t'[X])))\}$$

(4)多源连接操作∞^P

设R_1,R_2分别表示多源关系模式R_1和R_2上的属性集合,r_1,r_2分别表示多源关系模式R_1和R_2上的任意两个多源关系,X为$R_1\cap R_2$上的高阶多源属性,$A=R_1-X$,$B=R_2-X$,则$R=r_1\infty^P r_2$产生一个R上的关系r,有$R=(A,X,B)$,且:

$$r=\{t\mid((\exists\, u)\in r_1\wedge(\exists\, v)\in r_2(t[A]=u[A]\wedge t[B]=v[B]\wedge t[X]=(u[X]\cap^P v[X])\wedge t[X]\neq\Phi)\}$$

(5)多源投影操作$\prod^P$

设r是多源模式R上的一个多源关系,$X=\{x_1,x_2,\cdots,x_n\}$是R的多源属性子集,则rr在X上的投影操作定义为:

$$\prod{}^P_X(r)=\{t[x_1,x_2,\cdots,x_n]\mid\exists\, t'\in r\wedge t=t'[x_1,x_2,\cdots,x_n]\}$$

© 中国科学院软件研究所 http://www.jos.org.cn

(6)多源合并操作 Σ^P

设 r_1,r_2多源模式 R 上的两个多源关系,$X=R_1\cap R_2$,$A=R_1-X$,$C=R_2-X$,则:

$$r_1\Sigma^P r_2=\{t\mid(\forall u)\in r_1\wedge(\forall v)\in r_2((t[A]=u[A]\wedge t[B]=u[B]=v[B]\wedge t[C]=v[C])\vee(\forall w)\in r_1(u[B]\neq w[B])\wedge t[A]=u[A]\wedge t[B]=u[B]\wedge t[C]=nil)\vee(\forall w)\in r_2(v[B]\neq w[B])\wedge t[A]=nil\wedge t[B]=v[B]\wedge t[C]=v[C]))\}$$

(7)多源选择操作 δ^P

设 r 是多源模式 R 的多源关系,则:

$$\delta_F^P(r)=\{t\mid t\in r\wedge F(t)\ is\ true\}$$

(8)*Retrieve* 操作

Retrieve 操作是一个不带任何条件的选择操作.

3 IMDBS系统中的多源查询处理

在IMDBS系统中,一个多源查询处理过程如图3所示.

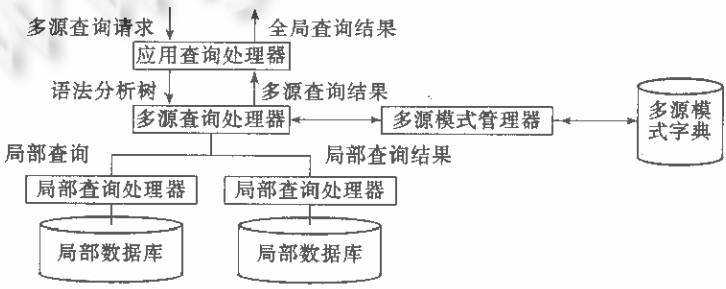

图3 多源查询处理过程

一条多源查询请求经过应用查询处理器的处理后产生一语法分析树,语法分析树经过多源查询处理器的分析与置换、优化与分解后产生一查询执行计划,并将各个局部子查询发送到相应的局部场地执行之.将各个局部查询结果收集到多源查询处理器所在的场地,以便于多源查询处理器的使用.应用查询处理器最后将多源查询处理器的执行结果组装成全局结果后返回给用户.

3.1 应用查询处理器

应用查询处理器的处理流程如图4所示.应用操作处理器负责对用户查询请求进行合法性检查,词法分析和语法分析,根据多源模式定义,将多源查询请求转换为语法分析树,语法分析树是用C语言结构来表示的.

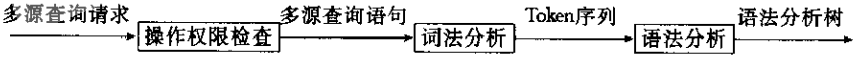

图4 应用操作处理器的处理流程

3.2 多源查询处理器

一个多源查询请求经过应用操作处理器的处理后产生一棵语法分析树,语法分析树经多源查询处理器PQP的分析与置换、优化与分解后,产生一个操作执行计划.多源查询处理器的处理流程如图5所示.

© 中国科学院软件研究所 http://www.jos.org.cn

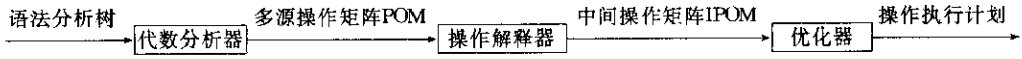

图5 多源操作处理器的处理流程

代数分析器的任务是将一棵语法分析树转换为多源操作矩阵POM,POM经代数操作解释器被置换为中间操作矩阵IPOM,IPOM最后被优化为操作执行计划.多源操作矩阵POM是以多源代数操作为基础的7元操作矩阵:〈操作序号,操作码,左多源关系,左多源属性,比较符,右多源属性,右多源关系〉;POM中的每一行对应一个或两个多源关系之间的操作.操作解释器利用集成信息,将POM转换为中间操作矩阵IPOM,转换过程采用两遍扫描算法:第一遍扫描POM的左半部,第二遍扫描POM的右半部,并根据集成信息将修正执行场地.两边扫描过程中需要对POM中的多源代数操作进行解释,并产生IPOM,IPOM只比POM增加一列:执行场地.两边扫描算法请参见文献[4,14].优化器的作用是以涉网分析优化为主进行中间操作矩阵的优化,并产生一个单点操作序列集合即操作执行计划.

4 结束语

本文详细讨论了一个集成式多数据库系统IMDBS的多源数据模型、多源数据语言、多源关系代数及多源查询处理技术.提出了基于非一范式数据模型的源标签集成机制和集成方法,对异构数据库系统的信息集成来说,它具有一定的普遍意义.该系统的主要贡献在于对MIT[1]提出的源标签机制进行了必要的扩充,以适应于集成非关系数据源,并提出了相应的实现策略与实现算法.

IMDBS系统是以CIMS重点应用工厂为研究背景,其实现环境是SUN4工作站,支持TCP/IP协议的通讯网络,UNIX操作系统,ORACLE和SYBASE关系数据库系以及MU/FO网状数据库系统.该系统已通过863CIMS主题的验收,通过大量的测试表明该系统(如系统性能方面)取得了良好的效果.

参考文献

1 Richard Y W, Madnick S E. A polygen model for heterogeneous database system: the source tagging perspective. Proc. of The 16th VLDB Conf., Brisbane, Australia, 1990. 519—538.

2 王国仁,郑怀远.基于EER数据库集成方法的研究.计算机研究与发展,1993,**30(12)**:36—40.

3 王国仁,张霞,周云凤等.面向对象和关系数据库系统的集成方法.第九届全国数据库学术会议论文集,上海,1990. 291—299.

4 寥卫东.一个基于NF^2的多源数据库集成系统的研究与实践[硕士论文].东北大学,1993.

5 王国仁,于戈,周云凤等.CIMS环境下多数据库的集成技术.第二届中国计算机集成制造系统学术会议论文集,深圳,1992.

6 Barrett S. Strategic alternatives and interorganizational systems implementations: an overview. Journal of Management Information System, 1986,**3(3)**:12—25.

7 Day V. View definition and generalization for database integration in multidatabase System. IEEE Trans. on Software Enigeering. 1984,**10(6)**:15—24.

8 Elmasri R. Schema integration algorithms for federated database and logical database design. Submitted for publication, 1987.

9 Litwin W. A multidatabase interoperability. IEEE Computer, December 1986.

© 中国科学院软件研究所 http://www.jos.org.cn

10 Deen S M. Data integrated in distributed databases. IEEE Transtractions on Software Engineering. 1987,13(7): 38—71.

11 Frank W. A conceptual model for integrated autonomous processing: an international bank's experience with large databases. Proc. of the 18th Int'l Conf. on Information Systems, 1987. 153—161.

12 Godes D B. Use of heterogeneous data sources: three cases studies. WP#CIS—89—02, Sloan School of Management, M17, Cambridge, MA, 1989.

13 Heimbigner D. A federated architecture for information management. ACM Transactions on Office Information Systems, 1985,3(3):11—32.

14 张迎红. 多源数据库集成系统中集成机制的研究与实现[硕士论文]. 东北大学,1992.

15 郑怀远,于戈. 计算机集成制造系统(CIMS)中的数据库技术. 计算机世界,1992第49期.

16 Roth M A. Theory of non—first—normal from relational database. Dissertation, The University of Texas at Austin, 1986.

17 Roth M A. SQL/NF: a query language for NF relational language. Information Systems, 1991,12(1):23—44.

18 Zheng Huaiyuan, Yu Ge, Wang Guoren *et al*. Extending polygen paradigm with NF^2 data model to integrate multidatabase in CIMS environments. Proceeding of ICCIM, Beijing, 1993. 278—281.

附录 A　多源数据定义语言和多源数据操作语言的 BNF 描述

多源数据定义语言的 BNF 描述:

〈集成语句〉::=INTEGRATION PSCHEME〈多源模式名〉(〈多源属性描述表〉)

〈多源属性描述表〉::=〈多源属性描述表〉,ITEM〈多源属性〉

〈多源属性〉::=〈零序多源属性名〉〈源标签描述〉|〈PSCHEME〈高序多源属性名〉(〈多源属性描述表〉)〈导航描述〉

〈源标签描述〉::={〈源标签描述表〉}

〈源标签描述表〉::=〈源标签〉|〈源标签描述表〉,〈源标签〉

〈源标签〉::=(〈局部数据库类型〉,〈局部数据库名〉,〈局部模式名〉,〈局部属性描述〉)

〈局部数据库类型〉::=R|H|N

〈局部属性描述〉::=〈局部属性名〉|〈父记录名〉.〈字段名〉|〈系名〉.〈系首记录名〉.〈字段名〉

〈导航描述〉::=〈多源属性名〉.〈多源属性名〉|〈层次路径〉|〈系名〉

〈层次路径〉::=〈父片段名〉.〈子女片段名〉

多源数据操纵语言 BNF 描述:

〈查询语句〉::=SELECT〈多源属性表〉FROM〈多源模式名〉[〈WHERE 子句〉]

〈多源属性表〉::=〈多源属性〉|〈多源属性表〉,〈多源属性〉

〈多源属性〉::=〈零序多源属性名〉|〈查询语句〉

〈插入语句〉::=INSERT INTO〈多源属性名〉(〈多源属性名表〉) VALUES (〈多源属性值表〉)

〈多源属性名表〉::=〈多源属性名〉|〈多源属性名表〉,〈多源属性名〉

〈多源属性名〉::=〈零序多源属性名〉|〈高序多源属性名〉(〈多源属性名表〉)

〈多源属性值表〉::=〈多源属性值〉|〈多源属性值表〉,〈多源属性值〉

〈多源属性值〉::=〈原子值〉| (〈多源属性值表〉)

〈删除语句〉::=DELETE FROM〈多源模式名〉[〈WHERE 子句〉]

〈修改语句〉::=UPDATE〈多源模式名〉SET〈修改描述表〉

〈修改描述表〉::=〈零序属性名〉=〈原子值〉|〈高序多源属性名〉=〈高序多源属性修改〉

〈高序多源属性修改〉::=〈插入语句〉|〈删除语句〉|〈修改语句〉

© 中国科学院软件研究所 http://www.jos.org.cn

POLYGEN DATA MODEL, LANGUAGE AND QUERY PROCESSING IN IMDBS SYSTEM

Wang Guoren Yu Ge Zhang Bin Shan Jidi Zheng Huaiyuan

(*Department of Computer Science, Northeast University, Shenyang* 110006)

Abstract IMDBS is an integrated multi—databases system which can integrate information and data in existing local heterogeneous database systems and supply global users with shared integrated information. And all local database applications can still execute on the integrated database, thus local autonomy might be preserved. IMDBS based on the source tag integration mechanism adopted the non—first—normal form(NF^2) data model as its polygen data model. PSQL/NF, a SQL—like polygen data language, is defined for users to use it to manipulate database conveniently. The system can integrate traditional databases such as relational, hierarchical and network databases bottom—up and manage distributed data top—down, so it fundamentally resolves the conflict between global distribution transparency and local autonomy in heterogeneous distributed database systems. This paper mainly describes the polygen data model, the polygen data language and the implementation techniques of query processing of the IMDBS system.

Key words Source tag, NF^2 data model, schema integration, multi—database system, query processing.

© 中国科学院软件研究所 http://www.jos.org.cn