一种基于中间库的数据库间数据转换技术

姚领众　（北京理工大学计算中心）

1. 引言

近年来，以 ORACLE、SYBASE 为代表的各大数据库公司都纷纷行动起来，把连接其他公司的 DBMS 产品，实现异构环境下各数据库间的互操作做为提高其产品性能，增加产品竞争力的一种重要手段，并取得了重大成果。如 ORACLE 通过一些网络协议的支持，可与 DB2、IMS、RMS、Rdb、SQL/400 等 DBMS 连接。SYBASE 的 API 可以连接 Rdb、ORACLE、Ingres 及 DB2，实现了异构环境下建立具有较高性能的分布式数据库系统。然而，由于我国近几年流行的数据库主要是 ORACLE，Informic、dBASE Ⅲ、FOXBASE 及 FoxPro 等。ORACLE 与 SYBASE 提供的新技术对这几种数据库间实现信息互用及联合操作仍然是无能为力的。因此，解决这几种数据库间数据转换与联合使用仍具有重要现实意义。

目前，异种数据库间数据转换尚停留在一对一直接转换阶段。在这种方式下，几个数据库系统之间的相互转换可用 n 个结点的完全图边表示两个系统之间的双向数据转换。n 个结点的完全图共有 $Cn^2 = \frac{1}{2}n(n-1)$ 条边。因此，如果用 D_n 表示实现几个数据库系统间相互转换时要解决的数据类型匹配个数及相应数目的双向数据转换程序，则 $D_2 = 1, D_3 = 3, D_4 = 6 \ldots\ldots, D_n$ 随 n 的增长上升很快，当 n = 10 时，达到 45。图 1 是一个与结点完全图，表示 5 个数据库系统间的转换。

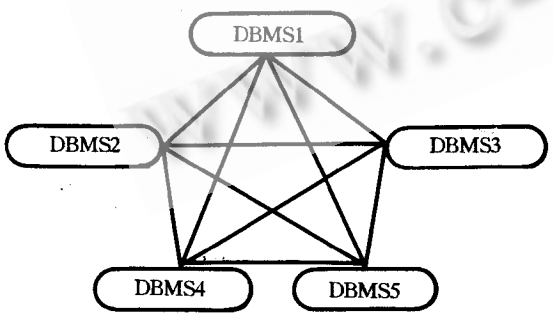

图 1　5 个数据库间相互转换构成的完全图

如果要将 n+1 个数据库系统加入前面 n 个异构库实现相互转换，那么就要解决该数据库系统同前面 n 个异构库的数据类型匹配问题，并设计 n 个转换程序。

这种方法的缺点是显而易见的。工作量大、效率低、难以集成在一起、难以实现分布式的需要和网络要求，构成的系统可扩充性差。数据库种类繁多，解决数据类型匹配就非常麻烦，难以适应飞速发展的数据库技术的需要。本文介绍通过构造中间数据库来实现多个异构数据库之间的相互转换。

2. 引入中间库的基本思想

中间数据库是一个结构特殊的数据库，任何两个异构数据库间实现数据转换都要经过下面两步：将源数据库转换成中间数据库；将中间库转换成目的库。

n 个异构数据库系统通过中间库作为媒介实现数据的相互转换。可表示为如图 2 所示的一棵树 。

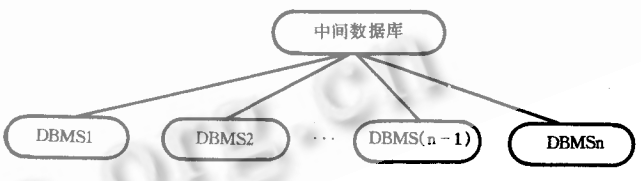

图 2　n 个数据库系统通过中间库实现

该树实际上是一个以中间库为根，以 n 个数据库系统为叶结点，高度为 2 的树。父结点与子结点之间的连线（树枝）表示某数据库系统与中间库之间的双向数据转换，图中共有 n 条树枝。这样要实现多个异构数据库之间的数据转换只要描述出它们的数据类型和中间数据库数据类型的对应关系，通过中间数据库就可以实现。如果一个新的数据库要同前面 n 个异构数据库之间实现数据互换，只需解决它同中间数据库的数据类型匹配，就可以方便实现双向数据转换。而且这种方法把 n 个异构数据库之间的数据互相转换集成在一起，有利于实现分布式管理和异构网络通信。

3. 实现中的关键技术

(1)中间库的定义。中间数据库的数据类型定义应该容易实现同各种常用数据库之间的数据类型匹配。选择最流行的关系数据库作为定义其结构的参考模型是一种合理的方案。ORACLE 是目前唯一可以通用于近七十种大型机、小型机和微型机上的关系数据库系统。因此在建立中间数据库时,我们主要参考了 ORACLE 的数据类型定义,并增加了复合型。

中间数据库定义的数据类型如下:

①CHAR(字符型)。最大长度为 256。字符型数据在记录中以左对齐的 ASCII 码形式存入字段内容,长度不够右边添空格。

②NUMBER(数据型)。其字段长度为 m,小数点右边长度为 n。在记录中以右边对齐的 BCD 码形式存放,长度不够左边添空格。

③DATE(日期型)。定长 8 个字节。

以年、月、日形式存放,如 1991 年 12 月 12 日在记录中的存放形式为:1991 12 12。

④LONG(长字符串型)。定长为 4K,主要用来存储一段文字,以串结束符"\0"来表示字符串的结束,在记录中以左对齐的 ASCII 码形式存入字段内容,长度不满 4K 在右边添"空格",长度超过 4K 则超过部分自动删除。

⑤COMP(复合型)。复合字段把两个或多个(最多 8 个)前面定义好的字段合在一起。在记录中复合型字段不占用空间。

在设计上述数据类型的存储格式时,主机参考 ORACLE 数据类型的外部显示方式,同时也兼顾同其他各种关系数据库存储形式转换的方便。

(2)中间数据库的组成。在数据转换过程中,中间数据库仅仅起个解释性作用或媒质作用。中间数据库的数据字典仅仅包含关系定义和字段定义,对任何数据库来说有了上述信息就足以定义一个关系数据库。中间数据库的数据字典我们用一个结构数据来表示,数组的每一项包含了一个字段的必要信息。

用 C 语言描述如下:

```
stuct Field{
Char * Field-name;/* 指向字段名字符串的指针 */
int Field-len1;/* 字段长度 */
int Field-len2; /* 小数点后的长度 */
char * Field-type;} /* 指向字段类型名字符串的指针 */
```

数据部分始终包含一个正在进行转换的记录。

(3)转换过程。以中间数据库作为媒介,将源数据库转换成目的数据库的过程分为模式转换和数据转换两部分。模式转换实际上就是将源数据库的数据字典换成目的的数据库的数据字典,过程如图 3 所示。数据转换就是将源数据库记录读出来经一系列转换,最后同目的库模式结合,便产生目的数据库,如图 4 所示。

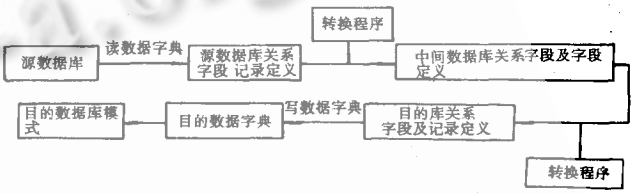

图 3　生成目的数据库模式过程

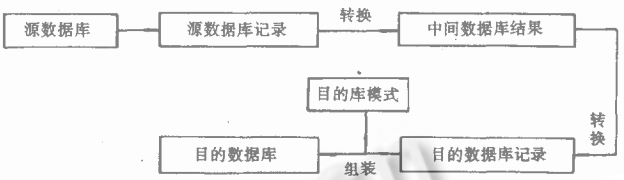

图 4　生成目的库

(4)数据格式转换方法。我们采用以下四种方法存取相应数据库中的关系定义、字段定义及记录内容,实现数据格式转换:①利用数据库本身提供的数据操纵语言,存取所需的内容;②利用数据库提供的数据转换工具;③利用高级语言提供的预编译接口(即宿主语言接口)及高级语言存取数据库;④从数据库的物理结构入手,从物理结构中存取所需内容。

4. 例子

在本课题中,我们已经实现了:①在单个结点上(DOS 或 XENIX)ORACLE、dBASE、Intormix 及 FoxBASE 之间的相互转换;②在 Ethernet 网络环境下(含有 Dos/Windows, XENIX, UNIX 三类结点),不同结点间转换。

由于篇幅所限，不能一一列举，仅选其中一例“dBASEⅢ与中间数据库的双向数据转换”介绍如下。

从分析物理库结构入手，从物理结构中存取所需内容，以提高转换效率。

(1)dBASEⅢ库结构分析。在dBASEⅢ中，与数据转换有关的有两种文件：·DBF文件与·DBT文件。·DBF文件包含数据库文件的结构和数据信息；·DBT文件存放memo字段信息。·DBF文件的结构部分由文件信息与字段两部分组成，分别如图5、图6所示。

标志	文件更新日期	记录数	结构区长度	记录长度	备用
1	2	5	9	11	13 32

图5　文件信息

字段名	00	类型	存储位移	未用	字段宽度	小数位数	内部使用
1	10	12	13	14	16	18	19 32

图6　字段表结构

·DBT文件按块组织，每块512字节，第0块前4个字节存放了该文件所含的块数。当记录的memo字段有值时，在memo字段位置存放了该字段在·DBT文件中的起始块号。库记录部分以ASCII码形式存储，每个记录以一个字节的删除标志开始，各记录依次存放。

(2)dBASEⅢ和中间数据库的数据对应关系列表如下：

dBASEⅢ			中间数据库		
类型	长度	小数	类型	长度	小数
char	m		char	m	
number	m	n	mumber	m	n
Logical	1		char	1	
date	8		date	8	
memo	10		long	4000	

对于表中char，date，Logical三个字段，数据的存储格式完全一致，因此在转换时直接把一种数据库中抽出的数据装入另一数据库相同类型的字段中。对于number和memo字段，对一种数据库中抽出的数值型数据需先进行格式转换预处理后才能装入另一库中。

(3)dBASEⅢ到中间数据库的数据转换。先进行库文件结构的转换，如下图所示：

图7　dBASEⅢ库结构转换为中间库结构

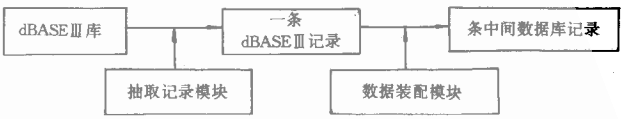

图8　dBASEⅢ到中间库的记录转换

(4)中间数据库到dBASEⅢ的转换。该转换也分为库结构与数据记录转换两步。转换过程可看作是(3)的逆过程，原理相似，不再赘述。

5. 结束语

本文所述的数据转换技术是北京理工大学承担的国防科工委“八五”重点预研课题“异构型数据库的转换与联合使用”(编号为7A.3.4.4)的一部分，该课题已通过部级鉴定。

·投稿须知·

·内容开门见山，直接进入主题；
·文稿尽量用打印稿，行距不宜过小，插图必须描绘清晰；
·程序不宜过长，若超过150行请指出重要段落及可删略部分，一律上机调试通过，并注明软、硬件运行环境；
·参考文献只指明主要2～3篇。