64-67 数据库 协调器 源描述 信息源查询平台 (16)

计算机科学1998 Vol. 25 №. 4

基于协调器和源描述的集成化多信息源查询平台*)

An Integrated Multiple-Information-Source Query Platform Based on Mediator and Source Descripton

许学标 熊新阶 顾 宁 施伯乐

(复旦大学计算机科学系 上海 200433)

TP311.13

摘 要 In this paper, the Open World Assumption is adopted as the fundamental of integration, a data model mixed with object and relation model is given as the foundation of integrated query, then the description form of both contents and query expression are discussed, which will facilitate the expansion of information sources. Finally, a framework based on Mediator and sources description is presented, which will make well use of the description information in individual sources too.

关键词 Mediator, Multiple-information-source, Relation-object data model, Source description, Integrated query

一、引言

进入九十年代以来，数据库技术和网络技术的飞速发展使信息产业突飞猛进。随着 Internet/Intranet 上的 WWW 日渐流行，越来越多地出现分布式的、自治的、包括各类数据库、正文、声音、图像、HTML 等异构形式的信息源。设计灵活高效、准确方便的多信息源集成化查询处理工具和查询处理机制平台，对它们进行有效地组织管理和应用，使其成为一个大的"信息宝库"而不是一些闲散无用的"信息垃圾"，已成为众多研究者追求的目标。过去的数据库方法包括多数据库系统、联邦数据库、分布式数据库和松耦合信息集成系统等，其实际效果并不能完全满足应用的要求。

1992年5月，斯坦福大学的 Wiederhold 教授在 IEEE Computer 上发表的《将来信息系统结构中的协调器》一文中，提出了区别于上述诸方法的一种信息系统集成的体系结构，即基于协调器(Mediator)的多信息源集成的体系结构，备受关注。此后，不少信息系统的集成也都采用这一种体系作为开发的基础，研究能处理具有关联异构的多信息源的查询机制，提供统一的查询接口(视图)，以查询浏览相关信息。这些信息系统有著名的美国国防部(DARPA)的13项目、IBM 的 Garlic、Stanford 的 TSIMMIS、Maryland 的 HERMES、以及 Colorado 大学的 Squirrel、Mermaid、Whips、CARNOT、DISCO 等。这些系统多把全局模式看作是从局部模式得到的一个大视图，而将各个数据源使用的不同数据模型翻译成统一的数据模型，在客户端的查询和应用将针对虚的全局模式数据构造查询和获得答案。这些系统在模式集成中很少根据各个信息源所具有的不同能力、查询要求和具体的模式功能信息进行查询处理、优化等；同时，它们多是基于封闭世界假说(CWA)，也就是说，假设所有下层信息源中的数据就是所有保证答案的全部数据，所有数据源中数据的并集等于全部的数据。实际上这一假设很难成立，因为在多信息源存取系统中，由于信息源的开放性，根本就无法保证下层源的数据就是全部相关的数据。为此，与封闭世界假说不同，本文建立查询基于的是更符合多信息源查询实际的开放世界假说(OWA、Open world Assumption)，即认为所有数据源中的信息加在一起只是全部应该得到数据的一部分，源中所有数据的并集只是全部数据的子集。同时作者给出了利用信息源的描述信息进行多信息源集成查询的体系结构和方法，主要包括基于面向对象和关系模型相结合的数据模型，各个信息源的功能、查询要求和具体约束的信息描述以及如何有效地利用这些源描述进行查询处理和优化等。最后给出了实现的结构和一个图书多信息源查询的例子说明。

二、数据模型

这里采用的数据模型是将关系和面向对象结合起来，在关系模型上扩充了某些面向对象的特征，用于直接简便地描述信息源的内容和查询表达。这样，

*)本研究得到国家八六三计划、国家九五科技攻关和上海市科委发展基金的支持。许学标 博士生，主要研究领域为 OODB 和多信息源集成互操作。熊新阶 副教授，目前研究领域为 OODB、DBMS 应用系统。顾 宁 副教授，博士，主要研究领域为 OODB、CSCW、工程数据库等。施伯乐 教授，博士导师，主要研究领域为 KDB、OODB 和多媒体数据库。

一方面可以利用关系上的强大方便的查询功能和成熟的查询技术，同时可以用对象方法屏蔽许多信息源中的独特内容，用对象接口方便实现异构信息源上的简单查询。模型中包含关系、类、类层次、属性、对象等概念，其中关系的概念和关系模型中的关系一样；而对象模型则加以扩充，即，类是由一组相关联的属性构成的公共模板；类中包含对象，每个对象也有一个唯一的标识符(oid)。每个属性可以是单值的，也可以是多值的；类和类之间可以是不相交的，也可以是存在超子类关系的，即在类和类之间存在偏序关系(<)和不相交关系，其定义具体如下：

当类 C 为 D 的子类时，称 C、D 间存在偏序关系'<'，记 C<D。如果两个类之间没有共同的对象，则这两个类是不相交的。

为了将一元关系和类进行统一处理，有如下定义：

①当 x 为类 C 的一个对象 o 的标识时，X 为对象属性名，称(X)∈C；这样，一个一元关系 X 和一个类 C 联系在一起；这样，oid 就可以作为关系属性值使用了。

②设 A 为类 C 的属性，当(X)∈C 且 x.A=y(y 为在属性 Y 上的值，y 称 x 的 A 填充)，称(X,Y)∈C；这样，一个二元关系(X,Y)和每个属性 X,Y 联系起来，类中的属性也可作为二元关系进行处理了。

③对于单值属性 A，用 A(X)表示 A(x,y)保持的唯一值。

为使这些关系能完全捕获类层次的语义概念，模型中包括了如下的完整性约束，用包含依赖和函数依赖形式表示如下：

①设类 C、D 有偏序关系 C<D，则当把 C、D 作为关系处理时有包含关系 C⊆D；

②对应于单值属性 A 的任一辅助关系 A(X,Y)，有函数依赖 A：X→Y。对任一对不相交类(C、D)，当把 C、D 作为关系时，有 C∩D=Φ。

同样地，关系包含元组，类中包含对象，关系和类的属性值可以是原子值(串或整型数)或对象标识。对象可以属于多个类，即使这些类没有'<'关系，例如，一个对象可以是两个类的实例，而这两个类可能并没有超子类关系。这样类和对象之间的捆绑不如 ODMG'93那样紧密了，这主要是针对多信息源的特点而定义的，因为这样可以将视图功能很好地包含在模型中。

三、世界视图和查询描述

在所有的查询过程中，用户都处于一个包含所有信息源内容的全局性视图上，称为世界视图。它和一个模式一样，是模型中的类和虚关系的大集合。这里，我们没有用模式表示而是用世界视图，原因是：世界视图是作为用户提出查询的模式而存在的，这样，用户只需和世界视图打交道，而不用和单个源模式打交道，在这一点上世界视图和关系模型中的视图相一致；另一方面，关系模型中视图的数据可以存储于临时关系中，而世界视图则不然，它是并不存在的，仅用于信息源的内容描述。在世界视图上进行查询的一般表示形式为：

$$Q(\underline{X})\leftarrow R_1(Z_1),\cdots,R_n(Z_n),C_Q$$

其中，$R_1(Z_1),\cdots,R_n(Z_n)$为世界视图上的关系；C_Q 为形如 $\mu\theta\upsilon$ 的有序子目标的合取；$\theta\in\{<,>,\geqslant,\leqslant\}$，$\mu,\upsilon\in\bigcup_{1\leqslant i\leqslant n}Z_i$；$X\subseteq\bigcup_{1\leqslant i\leqslant n}Z_i$。

四、信息源内容和能力的描述

这里在世界视图上提出所使用的查询语句，而回答查询语句的数据实际存储在外部的多个信息源上。因此，要回答查询就必须描述每个信息源的内容与世界视图上的类、属性和关系的联系。由于相对于包含全部信息的世界视图而言，任何信息源都无法回答关于其内容的任意查询，这是由异构信息源的形式和功能所决定的，因为信息源中数据管理的模式不同，信息源的查询能力就会各不相同，例如关系库上的查询，就可以针对所有的属性进行查询，而 OODB 上的查询功能就弱一点，基于 WWW 网络查询工具的功能就仅限于对关键字上的查询了。此外，例如许多包含声音、图形/像、正文和大的二进制对象等的多媒体信息源的查询能力往往根据设计者的水平和能力而定，更是千差万异。为了生成可实际执行的查询计划，就需要带有信息源处理能力的描述。为此，文中给出信息源内容和能力的描述方法，通过将各个信息源所能表达的查询方式、满足的信息约束等用类信息的形式提供给协调器上的世界视图，对于用户提出的查询目标的生成、查询的优化等至关重要。这里先给出底层信息源的内容描述，然后再给出信息源查询能力的描述形式。

4.1 信息源的内容

针对每个具体的信息源，需要对其表达的内容进行描述，这基于如下考虑：

(1)局部化原则 信息源的数目可能很大，且经常变化，集成化的多信息源查询工具应能在不改变世界视图和不影响其他信息源描述的前提下增加新的信息源。

(2)便捷有利原则 许多信息源包含了彼此相关的信息，通过描述能够更细微地模型化其内容的差异，使查询相关的多个源能尽可能紧密地确定下来。信息源的内容，相当于库中元数据的内容，能够直接将元数据利用到查询中，扩大了查询的灵活性和查询能力。同时，直接加以描述，而不以层次化的方式体现，可以简明直接，定义数据源时只需在描述中增加信息源上的内容描述信息即可，从而可以避免 OODB 层次化中输入复杂路径所带来的麻烦。

采用前面的数据模型，把信息源的内容模型化成一或多个关系上的元组。这样可以进一步把源中的这些关系描述成比较谓词和世界视图关系之上的查询，从而使查询的定义和描述更加灵活方便。

为此，我们把每个源模型化成包括元组的一或多个关系，称之为"源关系"。为了保证世界视图上的关系和源关系间能方便地定义映射关系，要求源关系的名称和世界视图关系中的名称不相交。对于每一个源关系，必须指定描述源关系中元组所要满足条件的合取查询。根据 OWA 假说，底层的信息源关系都是不完全的，它们不能包含所有满足查询的元组，并且也不是封闭的，新源的增加将会产生新的满足条件的元组。例如大部分的图书馆信息源都不能保证收录了全部出版的书籍，用户只能得到图书馆中现有情况下的查询结果。基于 OWA 假设就可以保证，信息源中不必也不一定包含所有满足查询的所有元组。这也是在下节的例中用包含("⊆")而不是等价导出("←")的原因。

所以，利用前面数据模型的类层次、类的不相交性和内置谓词等概念，对信息源内容使用查询描述的方法基本上可以满足前面的两点要求。

4.2 信息源的处理能力

上面的内容描述部分仅仅给出了在信息源中有什么，但要将描述应用于多信息源的集成化查询中，还必须给出每个信息源的能力描述，即该信息源中包含的内容能回答什么样的查询。其中查询的输入、输出项和条件表示等内容都是必须的。对具不同查询能力的信息源进行查询，必须充分利用它们的查询能力和约束信息。同时，在查询计划生成时，如果保证查询计划子目标和信息源的(查询)能力相一致，并尽可能多地利用信息源所提供的查询处理能力描述，将会极大地提高查询的求解速度和执行效率。这里对信息源的查询处理能力的描述，我们将用记录的方式表达。关于查询的处理能力包含两方面的内容：(1)限制信息源中所能使用的选择能力；(2)限制信息源中所能接受变量绑定(Binding)的能力。即能力记录将指定能够给信息源什么输入，允许的最大和最小输入量，信息源可以使用的选择以及源的可能输出。通俗地讲，能力记录描述的是可以给每个信息源什么样的参数。具体地，设 X 是由源关系 R 中的变量和常量构成的元组集合，记为 R(X)，如果 x 是属于 X 的一个变量，或如 A 为一个属性名，x=A(X)且有 x 属于 X，称 x 是 R 的一个参数。对应于每个信息源中的各个关系，我们用形如(S_m,S_{out},S_{sel},min,max)的能力记录来表达它的查询处理能力。其中，S_m,S_{out},S_{sel}为 R 的参数集合，min,max 为整数，X 中的每一个变量必须在 S_m或 S_{out}的参数中出现。查询能力记录(S_m,S_{out},S_{sel},min,max)所表达的具体含义如下：

S_m是在查询中需要输入的所有参数的集合，S_{out}是所有可以从信息源中返回的参数集，min 表示的是为了进行查询，至少要进行绑定的变量个数，max 是进行查询时，所能输入的变量的最多数目。能力记录是指为了从信息源中获得 R 关系的一个元组，信息源必须给出 S_m中的至少 min 个变量进行绑定。S_{sel}是查询时所能使用的选择变量的集合，即信息源中可使用的选择原子的表达形式为 α op c(c 为常量，op∈{≠,≥,≤,=})的变量参数 α 的集合，且 S_{sel}中的元素必须是 $S_m \cup S_{out}$的子集。

4.3 信息源描述用于查询上的扩充描述

给定一源关系为 R，设有 S_m中参数 $\alpha_1,\cdots,\alpha_n$ 的对应输入值分别为 $a_1,\cdots,a_n$，加在 R 上的选择形式为 $\gamma_1,\cdots,\gamma_k$。其中 γ_i(i=1,…,k)的形式是 α op c(c 为常量，op∈{≠,≥,≤,=})，经过上述选择操作后求得的 S_{out}中的参数 $\beta_1,\cdots,\beta_n$ 值的结果元组表示为 $(Y_1,\cdots,Y_f)$，即为结果关系 $R'(Y_1\cdots Y_f)$，它应满足如下条件：

$R'(Y_1\cdots Y_f):-R(X_1\cdots X_m),\alpha_1=a_1,\cdots,\alpha_n=a_n,\beta_1=Y_1\cdots\beta_f=Y_f,\gamma_1,\cdots\gamma_k$。

给定形如 $R\subseteq Q_R$ 的内容描述和上面的输入输出特性说明，称

$R'(Y_1\cdots Y_f)\subseteq Q_R,\alpha_1=a_1,\cdots,\alpha_n=a_n,\beta_1=Y_1\cdots\beta_f=Y_f,\gamma_1\cdots\gamma_k$

为对应输入/输出特性描述下关系 R 的扩充描述，用以简化查询计划的表达式，并有利于查询执行时可执行查询计划的生成。

五、信息源查询的应用例子及实现结构

这里先给出图书信息查询中常见的四个信息源的情形及其信息描述和查询能力描述的例子。其中四个信息源具体叙述如图1，为此而定义的类层次如表1所示，各个信息源的内容和查询能力描述见图2。

信息源1：综合图书信息源 接受输入为书的类型和种类，可选输入项为价格范围和出版范围；对每个满足条件的书，给出其类型，出版年份，价格及出版社地址。
信息源2：计算机类书籍信息源 接受输入为书的类型，可选项为作者名；对每个满足条件的项，输出年份，作者名，种类和价格，摘要。
信息源3：古书信息源　解放前出版的书籍 输入类型和年代范围，对每个满足条件的项，输出书的类型，作者，年份和价格。
信息源4：外文原版书　价格在20.0$以上 输入语言，类型，可选输入为年份，对每个满足条件的项，输出其类型，书名，年份和价格。

图1　多信息源的例子

表1 模型使用的类层次信息

（无关类'人'和'录音机'略）

类名	超类名	属性	不相交类名
出版物		类型	人
书	出版物	类型,年份,种类	录音机
古书	书	类型,书名,年份,种类,作者	新书
新书	书	类型,书名,年份,种类,价格	古书
外文原版书	书	类型,书名,年份,价格,语言	古书
计算机类书	新书	类型,年份,种类,作者名,摘要	古书

信息源3：古书信息源 解放前出版的书籍
内容描述：$V_2(b)\subseteq Book(b), Year(b,y), y<=1949$
能力描述：({Type(b), Year(b)}, {Type, Name, Year, Author}, {}2, 2)

信息源4：外文原版书信息源
内容描述：V4(b)⊆Book(b), Language(b,1), 1<>'中文'; Price(b,p), P>20.0
能力描述：({Language(b), Type(b),}, {Type, Name, Year, Author, Price}, {Year}, 2, 3)

图2 与图1多信息源相对应的源描述

（这里给出的是信息源3、4的源描述，信息源1、2可用同样方法获得）

最后，基于上面的数据模型和源信息描述，这里给出了基于协调器的多信息源集成化查询平台的体系结构，具体见图3。

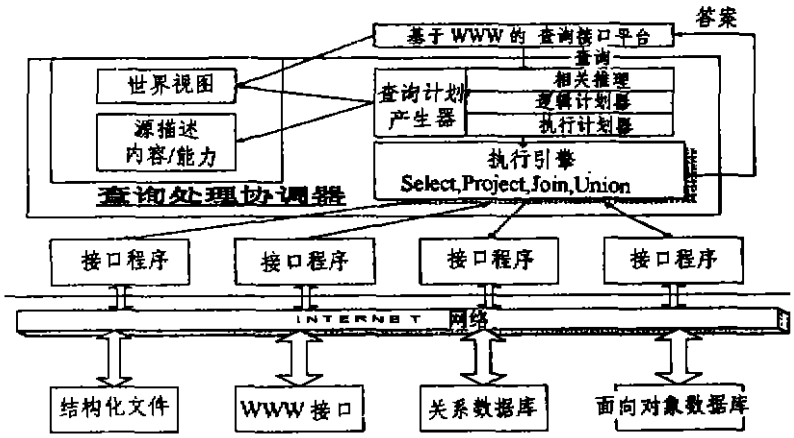

图3 基于协调器的多信息源集成化查询平台的体系结构

结束语 多信息源的集成化查询是当今网络和信息技术相结合的必然要求。灵活、便捷、实用、高效的查询平台工具对于充分发挥信息的作用具有重要的意义。本文给出了基于协调器和源描述的多信息源集成化查询平台，介绍了它使用的数据模型、信息源的内容和查询能力描述，以及如何将这些描述应用到多信息源的集成化查询过程去的方法。该方法使得信息源的扩充只需通过增加查询就可以实现，易于扩充，对于构建新型查询平台具有实用性和指导意义，进一步的工作包括：如何获取各个信息源的描述内容，如何根据信息源的更新，维护信息查询的正确性，以及如何更有效地将信息源描述应用到查询目标的生成计划中，去除查询无用信息，以提高多信息源上的集成化查询的效能等。

参 考 文 献

[1]G. Wiederhold, Mediators in the Architecture of future Information systems, IEEE Computer, March, 1992

[2]R. Hull, Managing Semantic Heterogeneity in Databases: A Theoretical Perspective, In Proc. 16th ACM Symp. Principles of Database Sys., May, 1997

[3]A. Levy et al., Querying Hetereogeneous Information Sources Using Source Descriptions, In Proc. of the 22nd VLDB Conf., 1996

[4]J. D. Ullman, Information Integration Using Logical Views, In Proc. of Intl. Conf. On Database Theory, May, 1997

[5]多信息源的面向对象互操作平台技术报告，复旦大学计算机科学系数据库中心 OODB 课题组，1997.9