面向异构数据集成的数据源能力描述框架*)

唐九阳　张维明　修保新　肖卫东
(国防科技大学管理科学与工程系　长沙410073)

摘　要　异构数据源集成系统旨在为用户提供一个一致的访问接口，由于参与集成的各数据源不仅高度自治、模式各异、更新频繁，而且查询功能有各自特殊的限制，给查询处理过程中数据源定位和查询优化造成一定的困难。本文在分析异构集成系统特征和功能需求的基础上，提出一种基于KQML的数据源能力描述框架，为各数据源灵活动态的发布自身能力提供保证。进而通过形式化的规范描述刻画数据源的结构特征和行为特征，为定位查询相关数据源奠定基础，并有助于全局查询处理器对查询计划进行优化，缩减查询的搜索空间，提高查询效率。

关键词　能力描述，异构数据源集成，查询模式，KQML

A Data Source Capability Description Framework towards Heterogeneous Data Sources Integration

TANG Jiu-Yang　ZHANG Wei-Ming　XIU Bao-Xin　XIAO Wei-Dong
(Department of Management Science and Engineering, National University of Defense Technology, Changsha 410073)

Abstract　The heterogeneous data sources integration systems aim at providing a uniform interface for applications. The integrated data sources can be full-fledged databases, simple files, HTML pages or specialized data sources that posses diverse query processing capabilities, which make it difficult for the location of data sources and query optimization. This paper investigates the issues in data sources description in terms of their contents and processing capabilities, and proposes a framework to describe the capabilities of the data sources in fine detail. When a data source is registered with the mediator, the mediator uploads the capabilities of the wrapper, and smoothly integrates these capabilities into query processing, making a well foundation for the discovery of data sources and query optimization.

Keywords　Capability description, Heterogeneous data sources integration, Query planning, KQML

1　引言

计算机网络的普及和WWW的出现使得可访问的非结构化与结构化的数据源的数量迅速增加，从大量分布式的异构数据源进行数据集成变得日趋重要[1]。异构数据源集成旨在为这些异构的数据源提供一个完整的数据源模式和一致的访问接口，消除异构，实现分布透明。目前，开发异构数据源集成系统通常采用数据仓库法和虚拟法[2]。相比而言，虚拟法适合数据源数目多，数据更新频繁，而且无法预知用户的查询种类的情况，是当前异构集成研究的热点。本文的讨论针对虚拟法，它基于一个中介模式(Mediator Schema)，数据仍保存在局部数据源中，通过各数据源的封装器(Wrapper)将数据虚拟成中介模式(图1)。

虚拟法集成结构中为用户提供数据源的统一访问存在以下困难。首先，用户的查询基于中介模式，而查询引擎需要与封装器交互，将针对中介模式的查询转换为基于各数据源局部模式的查询，即确定与查询相关的数据源。因此，为了回答用户查询，必须建立局部模式和中介模式之间的语义映射，其中语义映射是一组称为语义映射关系式的规则集，由领域专家和系统设计员在系统设计时共同预定义。表达能力丰富的语义映射有助于全局查询处理器在查询重写时剪除不相关及冗余的数据源。随后，为了辅助全局查询处理器确定向各数据源传送执行的操作以及执行的顺序，要求各数据源能精确描述自身支持的查询模式，将通过查询模式反映出来的查询能力以通用的方式注册到中介器(Mediator)。

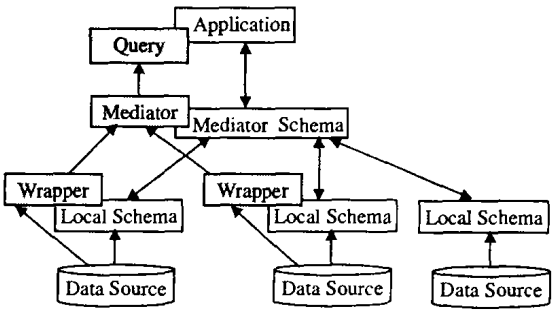

图1　异构数据源集成系统结构图

本文在分析异构数据源集成系统的特征和功能需求的基础上，提出一种基于KQML的数据源能力描述框架(本文中，数据源能力为一泛化概念，包括数据源的模式信息和查询处理能力，借以刻画参与集成的各数据源可以提供的内容和服务)，为各数据源灵活动态地发布自身能力提供保证。进而通过形式化的规范描述刻画数据源的结构特征(数据源的模式和语义映射)和行为特征(数据源支持的查询处理)，为定位查询相关数据源奠定基础，并有助于全局查询处理器对查询计

*)国家自然科学基金资助项目(60172012)。**唐九阳**　博士研究生，主要研究领域为信息系统集成、智能决策支持技术；**张维明**　教授，博士生导师，主要研究领域为信息系统工程、智能决策支持技术；**修保新**　博士研究生，主要研究领域为粒度计算；**肖卫东**　博士研究生，副教授，主要研究领域为信息系统工程、信息管理。

划进行优化，缩减查询的搜索空间，提高查询效率。

2 相关工作

针对异构数据集成问题，国外一些机构进行了相应研究。现有大部分工作从辅助全局查询处理器考虑各数据源对局部数据访问模式限制的角度出发，侧重于数据源查询能力的描述。TSIMMS[3]借助关系查询描述语言(RQDL)构造表达连接查询能力的查询模板。Garlic[4,5]和 DISCO[6]系统对目标查询表达式没有限制，容许数据源描述连接和非连接的查询能力。文[7]中使用上下文无关文法的语言描述数据源封装器的能力。然而，上述方法都没有涉及数据源的模式信息描述。

文[8～10]的工作和本文的研究相近，通过外部声明描述数据源的模式信息和查询能力，但还存在相应不足。NAIL![8]系统首次提出面向异构集成的数据源能力描述，可是局限于属性集的有限约束表达。Information Manifold[9]中基于五元组的能力记录对查询能力的描述粒度较粗：局限于满足一定数目的数据集合的绑定操作，不能对任意必选属性、可选属性应用约束的情况进行刻画。文[10]虽然对数据源查询处理能力的表示进行了深入探讨，但由于查询模式中没有指定与输入相对应的输出属性集，加之该方法的数据源模式信息描述中缺少局部模式和中介模式的语义映射，无法辅助异构集成系统中查询计划的生成。

3 数据源能力描述框架

异构集成系统中数据源更新频繁的特点要求描述机制能动态增加、更改、注销数据源能力描述而不改变中介模式以及其它数据源的能力描述。KQML 是一种用于交换信息和知识的语言以及协议，适合自主和异步主体间的知识共享，其目的在基于知识的异质系统间实现互操作和集成。因此，针对集成系统中数据源的异构性和动态性，提出基于 KQML 的数据源能力描述框架：

```
(advertise,
   :sender (wrapper)
   :receiver mediator
   :language Advertise
   :ontology (ontology)
   :domain (domainExpr)
   :operation (operationExpr)
   :content (ask-*,
                :sender (anyone)
                :receiver (wrapper)
                :language (language)
                :ontology (ontology)
                :domain (domainExpr)
                :operation (operationExpr)
                :content (patternExpr)))
```

在 KQML advertisement 中，外层 :content 槽描述该数据源可以处理的消息请求格式，内层 :content 槽则包含了数据源可以处理的查询模式。

由于 KQML 语言的 performative 集具有可扩展性，我们在 advertisement 中添加了两个槽。:domain 槽指定数据源的模式信息规范；:operation 槽指定 language 支持的操作。如内层的 language 是 SQL 语言，那么相应的 operation 为 selection，projection，join 等。:operation 槽只限于对数据源数据集合操作的描述，在没有进一步约束的情况下，缺省认为数据源完全支持 :operation 槽中的操作。否则，需要通过在内层 :content 槽中定义查询模式细致的刻画数据源支持的属性级的操作。

每个数据源对应一个这样的框架，数据源能力的修改，都通过该框架中槽值的修改来体现。参照文[11]提出的多 Agent 系统的实现，本文提出的能力描述框架可以作为各个数据源发布的广告(Advertisement)，存储在一个广告数据库中，以后由中介器取出，与基于中介模式的查询请求进行匹配，完成数据源的定位以及辅助查询优化。

4 数据源能力描述规范

可以采用说明法(Declarative approach)和过程法(Procedure approach)描述数据源的能力。比较而言，说明法涉及细节少，抽象程度高，而且不依赖特定语言的语法，有更强的适应性和可扩展性。因此，这一节里采用说明法给出数据源模式信息规范(domainExpr)和查询能力规范(patternExpr)的形式化定义。

4.1 数据源模式信息规范

为描述的方便，本文使用一个或多个包含元组的关系描述数据源的局部模式，这些包含元组的关系称为源关系。从异构集成系统的可扩展性和可成长性出发，采用 LAV 方法[12]表达局部模式和中介模式的语义映射。即对于每一个源关系，都指定一个基于中介模式的查询描述源关系中元组必须满足的条件。那么，在新增数据源时只涉及对该数据源的建模和语义映射描述，该特点对动态的环境特别有用。先给出语义映射形式化定义。

定义1 语义映射为 $V_i(\overline{X})\subseteq R_1(\overline{Z_1}),\mathrm{K},R_n(\overline{Z_n}),C_z$[1]，其中：$V_i$ 是数据源模式中的关系，R_1,K,R_n 是中介模式中的关系；C_z 是中介模式中变量的条件约束：$u\theta c,\theta\in\{<,>,\leqslant,\geqslant,=\},u\in \mathrm{Y}_{1\leqslant i\leqslant n}\overline{Z_i},c$ 为常数；$\overline{X}\subseteq \mathrm{Y}_{1\leqslant i\leqslant n}\overline{Z_i}$。

定义2 数据源模式信息规范为 $domainExpr=\langle Schema,Mapping\rangle$。其中，$Schema$ 是数据源的局部模式，$Mapping$ 是数据源局部模式与中介模式的语义映射。

例1 中介模式中包含两个关系：汽车〈汽车号，样式，颜色，生产年份〉，车评〈样式，评价信息，评分〉。关系汽车描述待销售的汽车信息，关系车评则针对给定的汽车样式，提供相应的评价信息并给出1－10的数值评分。如果数据源1保存所有红色汽车的信息，数据源2保存1999年以后生产、评分为10的待销售的汽车信息，数据源3保存评分低于9的车评信息，则对应的语义映射为：

$V1$〈汽车号，样式，生产年份〉$\subseteq$汽车〈汽车号，样式，颜色，生产年份〉，颜色＝红色

$V2$〈汽车号，样式，颜色，生产年份〉$\subseteq$汽车〈汽车号，样式，颜色，生产年份〉，车评〈样式，评价信息，评分〉，生产年份$\geqslant$2000，评分＝10

$V3$〈样式，评价信息，评分〉$\subseteq$车评〈样式，评价信息，评分〉，评分$\leqslant$8。

在例1中，假定用户提出以下问题：查找1999年以前生产的待销售的汽车信息，则根据各数据源的语义映射，直接判定数据源2与查询不相关(数据源2中“生产年份$\geqslant$2000”与查询条件中“生产年份$\leqslant$1998”矛盾)，数据源3也为查询无关数据源(不包含汽车信息模式)，这对于匹配处理和生成查询计划过程中有关数据源的“剪枝”有很大的帮助，通过减少查询计划生成的个数，提高查询效率。

4.2 查询能力规范

[1]本文基于开放世界假设(Open world assumption)[13]，即模式中不一定存储现实世界中满足定义的所有元组，也就是说，各数据源模式的实例为模式内涵的子集。因此，我们使用连接符$\subseteq$连接描述头和描述体以说明数据源局部模式不用包含满足查询的所有元组。

为了辅助全局查询处理器指定向各数据源执行的操作以及执行的顺序，要求中介器预先了解各个数据源的处理能力：如选择、投影、连接、聚集、分组等。可以采用查询模式表示数据源支持的查询，即查询处理能力。其中每个数据源可能接收多种查询输入，那么相应对应多个查询模式。由于查询模式中的属性值可以由全局查询中的约束条件直接提供，也可以由其它数据源完成查询后提供，因此，本文借助查询模式以及查询模式对应的输出属性集刻画数据源的查询能力。

由于各数据源对查询功能有各自特殊的限制，主要体现在：对特定属性、部分属性或任意属性应用约束的说明；有限或无限数目的原子条件的表达；and 或 or 在查询条件中的应用；属性间进行比较的运算操作。为精确描述数据源的查询模式，先引入下列符号集：

A：domain 槽中定义的属性集合；Q：输出属性集合，其中，$Q\subseteq A$；O：比较运算符集合（=，>等）和集合元素的运算符集合（In，Not In 等）；B：布尔运算符集合（and、or）；S：集合运算符（Union，Intersect）；G：集函数集合；P：查询模式集合；C：定义的条件模式集合。

4.2.1 *条件规范* 根据上面定义的符号集，接下来定义原子条件表达式：

·$\langle a,o,l\rangle$；

·$\langle a,o,a'\rangle$，连接条件；

·$\langle g,o,l\rangle$，应用诸如聚集的全局函数；

·$\langle a,o,p\rangle$，嵌套查询；

其中 l 代表任意值，$a,a'\in A,g\in G,p\in P$。如$\langle author,=,l\rangle$说明查询条件中可以对属性 author 应用“=”运算符。

复合条件表达式定义如下：

·$c[bc]^*$，包含任意数目的原子条件，其中查询条件没有先后次序；

·$c[b(c)]^*$，包含任意数目的原子条件，条件有先后次序；

·$c[bc]^{n-m}$，包含至少 n 至多 m 个原子条件，条件没有先后次序；

·$c[b(c)]^{n-m}$，包含至少 n 至多 m 个原子条件，条件有先后次序。

其中 $c\in C,b\in B$，* 表示任意数目。在条件表达式中需要显式指定可以应用约束的特定属性和特定运算符，如$\langle city,=,l\rangle and\langle arrival,=,l\rangle$说明查询条件必须同时作用在 city 和 arrival 属性上，其中运算符为“=”操作。此外，我们使用“$a”代表条件适用于 domain 槽中的任意属性。如需要强调属性的唯一性，则使用“&a”表示其它条件中没有使用的任意属性。例如$\langle\&a,o,l\rangle b\langle\&a,o,l\rangle$意味着条件表达式必须同时应用到两个不同的属性上。

4.2.2 *查询模式* 根据以上定义，查询模式 $p\in P$ 有以下几种方式：

·$c|r$，原子条件中没有集合运算符；

·$p[sp]^*$，有集合运算符并且查询模式没有先后次序；

·$p[s(p)]^*$，有集合运算符并且查询模式有先后次序。

其中 $c\in C,s\in S,r\in A\cup G,p\in P$。

定义3 查询能力规范为 $patternExpr=\{\langle p_i,q_i\rangle|p_i\in P,q_i\in Q,1\leqslant i\leqslant\#P,\#P=\#Q\}$，其中 p_i 是查询模式，q_i 是该查询模式对应的输出属性集合。

例2 数据源1的查询能力规范为：{⟨p1，q1⟩，⟨p2，q2⟩}，其中 p1：电影公司=l，q1：{影片名}；p2：影片名=l，q2：{电影公司}。数据源2的查询能力规范为：{⟨p1，q1⟩}，其中 p1：影片名=l，q1：{影片名，发行年份}。数据源3的查询能力规范为：{⟨p1，q1⟩}，其中 p1：影片名=l，q1：{影片名，主角}。

假定用户提出以下问题：查找1999年潇湘电影制片厂发行的、由姜文主演的影片。中介器首先确定数据源1、数据源2和数据源3为查询相关数据源；接下来，中介器需要搜索匹配满足查询条件的查询模式，发现数据源1的 p1可以执行（电影公司=潇湘电影制片厂），由于其对应的输出属性集为{影片名}，从而数据源2和数据源3的查询条件得到满足，可以并发执行。从本例中看出，数据源通过查询模式和输出属性集反映出来的查询能力，成为异构集成系统中确定查询子目标次序的关键。

4.3 示例

例3 旅游门户网站是集成了航班查询、火车车次查询、住宿、旅行线路规划等多个数据源的集成系统。其中对应的部分中介模式为：

住宿⟨宾馆，城市，入住日期，离开日期⟩；
地理位置⟨宾馆，城市，地址⟩；
宾馆信息⟨宾馆，城市，星级，价格，推荐信息⟩。

数据源1为北京一家提供三星级以上宾馆预订服务的网站，它的局部模式为 $V1$⟨宾馆，入住日期，离开日期⟩；数据源2提供全国五星级宾馆的浏览信息：$V2$⟨宾馆，城市，价格，推荐信息⟩。其中数据源1提供的服务如下：

·用户可以直接浏览北京所有宾馆的信息；

·通过指定“入住日期”和“离开日期”属性值检索可以入住的宾馆，这两个属性必选，其它为可选属性；

·要缩小搜索范围，用户可以进一步指定“星级”或“价格范围”属性值；

·只有“价格范围”属性可以应用“≤”运算符，其它的属性只能应用“=”运算符；

·所有条件都是可连接的。

则数据源1的语义映射为：

$Mapping1$：$V1$⟨宾馆，入住日期，离开日期⟩⊆住宿⟨宾馆，城市，入住日期，离开日期⟩，宾馆信息⟨宾馆，城市，星级，价格，推荐信息⟩，城市=北京，星级≥3

数据源2的语义映射为：

$Mapping2$：$V2$⟨宾馆，城市，价格，推荐信息⟩⊆宾馆信息⟨宾馆，城市，星级，价格，推荐信息⟩，星级=5

数据源1、数据源2分别对应的模式信息规范为⟨$V1$，$Mapping1$⟩、⟨$V2$，$Mapping2$⟩。

数据源1的查询能力描述如下（数据源2略）：

p1：|&a，q1：{宾馆，地址，星级，价格}；
c1：⟨入住日期，=，l⟩and⟨离开日期，=，l⟩；
p2：c1|&a，q2：{宾馆，地址，星级，价格}；
p3：c1 and ⟨星级，=，l⟩|&a，q3：{宾馆，地址，价格}；
p4：c1 and ⟨价格，≤，l⟩|&a，q4：{宾馆，地址，星级，价格}；
p5：c1 and ⟨星级，=，l⟩ and ⟨价格，≤，l⟩|&a，q5：{宾馆，地址，价格}。

则数据源1的查询能力规范为 $Pattern\ Expr=\{\langle p_i,q_i\rangle,1\leqslant i\leqslant 5\}$。

结论 随着互联网的进一步发展，异构数据源集成成为具有极大潜力的研究方向。本文通过引入数据源能力的概念，提出一种基于 KQML 面向异构集成的数据源能力描述框架，该框架以数据源模式信息以及数据源查询能力的描述为支撑，较好地满足了异构集成系统的功能需求，并为查询优化

（下转第188页）

4，这费时为 $O((k_v+k_l)^2)$。

如在第3节讨论的，如果应用情形1中的任一分枝，分枝搜索过程并满足递推公式(1)与(2)；如果应用情形2中的 T_1 至 T_5 分枝，分枝搜索过程并满足递推公式(3)与(1)；最后动态规范列表所花时间最多为 $O((k_v+k_l)^3)$。

根据文[5]中的 Master 定理，容易验证：$F(k_v+k_l)=1.19^{k_v+k_l}$ 满足所有递推关系(1)、(2)与(3)，由此我们知道解决 Min-CVCB 问题的时间为 $O(1.19^{k_v+k_l}+(k_v+k_l)^3+m\sqrt{n})$，这里 m 为图 G 的边数，即相当于矩阵列 A 的错误点数，n 为图 G 的点数，即相当于矩阵列 A 的行数与列数之和 $m+n$，通常情况下，k_v+k_l 比 n 小得多，因此我们有如下定理。

定理5 Min-CVCB 问题能在时间 $O(1.19^{k_v+k_l}+m\sqrt{n})$ 内解决，其中 k_v+k_l 为给定的参数，n，m 分别为二分图的点与边数，也即 Min-FCRA 问题能在时间 $O(1.19^k+kn)$ 内解决，这里 k 为可替换的行与列数，n 为矩阵的行与列数。

结论与进一步的研究 参数计算定理在工业应用中设计实际的算法是相当有用的，特别是对于许多 NP-难的最优化问题。在本文中，研究了在 VLSI 制造领域得到了广泛关注的可重构阵列的最小瑕点覆盖的复杂性，并提出了简单而有意义的改善算法。本文提出的算法概念简单、容易实现，且很好地组合了最近在研究参数计算中提出的技术与经典几何构造的结果。

本文的改进是基于减少指数运行时间中的底数，在研究 NP-难的优化问题中具有理论与实际的重要性。如在参数化的点覆盖算法中减少底数0.01，则在生物化学的相关应用程序中可以达到改善60%的运行时间[6]。算法相对以前关于该问题的算法减少了底数0.07(从 1.26^k 到 1.19^k)。

进一步的研究是再详细分析本文第3节中的各种情形，将算法的运行时间指数部分改善为 $(1+\varepsilon)^k$。

参 考 文 献

1 Hasan N, Liu C L. Minimum Fault Coverage in Reconfigurable Arrays. In: Proc. 18th Int. Symp. on Fault-Tolerant Computing (FTCS'88), 1988. 348～353

2 Nilsson N J. Principles of Artificial Intellegence. Tioga Publishing Co., 1980

3 Chen J, Kanj I. An Constrained Minimum Vertex Covers of Bipartite Graphs: Improved Algorithms. In: Proc. 27th Intl. Workshop on Graph-Theoretic Concepts in Computer Science (WG'2001), Lecture Notes in Computer Science 2204, 2001. 55～65

4 Lovasz L, Plummer M D. Matching Theory, Annals of Discrete Mathematics 29, North-Holland, 1986

5 Cormen T H, Leiserson C E, Rovest R L. Introduction to Algorithms. McGraw-Hill Book Company, New York, 1992

6 Bodlaender H L, Downey R G, Fellows M R, et al. Parameterized complexity analysis in computational biology. Computer Applications in the Biosciences, 1995, 11: 49～57

(上接第172页)

奠定了基础。相对于传统的"源描述"，该框架中的模式信息规范清晰描述了参与集成的各异构数据源局部模式与中介模式的语义映射，为选择查询相关数据源提供依据；其次，借助查询能力的精确刻画，为确定查询子目标的次序并向参与查询的数据源传送最小的查询提供保证。

我们下一步要进行的工作是利用已有的大量传统数据，如关系数据、结构文档等作为数据基础，采用虚拟法建立基于数据源能力描述的异构集成原型系统。

参 考 文 献

1 Navathe S, Donahoo J. Towards Intelligent Integration of Heterogeneous Information Sources. In: Proc. of the 6th Intl. Workshop on Database Re-engineering and Interoperability, March 1995

2 Jakobovits R. Integrating Heterogeneous Autonomous Information Sources: [Univ. of Washington Technical Report, UW-CSE-971205]. July 1997

3 Li C, Yerneni, Vassalos V, Papakonstantinou Y, Garcia-Molina H, Ullman J. Capability based mediation in tsimmis. In: Proc. of ACM SIGMOD Conf. 1998

4 Haas L, Kossman D, Wimmers E, Yang J. Optimizing queries across diverse data sources. In: Proc. of VLDB Conf. 1997

5 Carey M, et al. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach. In: Proc. RIDE-DOM Workshop, 1995. 124～131

6 Kapitskaia O, Tomasic A, Valduriez P. Scaling Heterogeneous Databases and the Design of Disco: [INRIA Technical Report]. 1997

7 Garca-Molina H, Labio W, Yerneni R. Capability sensitive query processing on internet sources. In: Proc. of the 15th Intl. Conf. on Data Engineering, Sydney, Australia, Mar. 1999

8 Morris K, Ullman J, Gelder A. Design Overview of the NAIL! System. In: Proc. ICLP, 1986

9 Kirk T, Levy A, Sagiv Y, Srivastava D. The Information Manifold. AAAI Symposium on Information Gathering in Distributed Heterogeneous Environment, 1995

10 Yang J, Xu L. Describing capabilities of internet data sources for information discovery and sharing. ADC 2001. 21～28

11 Sycara K, Widoff S, Klusch M, Lu J. LARKS: Dynamic Matchmaking Among Heterogeneous Software Agents in Cyberspace. Kluwer, 2002

12 Calì A, Calvanese D, Giacomo G D. Maurizio Lenzerini: On the Expressive Power of Data Integration Systems. In: Proc. of the 21st Int. Conf. on Conceptual Modeling (ER 2002). 338～350

13 Halevy A Y. Theory of Answering Queries Using Views. SIGMOD Record, 2000, 29(4): 40～47