信息集成研究综述*)

杨先娣[1] 彭智勇[2] 刘君强[2] 李旭辉[2]
(武汉大学计算机学院 武汉大学计算中心 武汉 430072)[1]
(武汉大学软件工程国家重点实验室 武汉 430072)[2]

摘 要 信息集成所要解决的问题是把位于不同的异构信息源上的数据合并起来,以便为用户提供一个这些数据的统一视图。在当前的实际应用中,设计信息集成系统很重要,并且已经成为数据库领域的研究热点。本文对这一领域的研究做了综述,包括信息集成的方法、逻辑框架、查询处理,以及Web上半结构化数据的集成。最后,对将来的研究主题进行了展望。

关键词 信息集成,异构,LAV,GAV,半结构化数据

An Overview of Information Integration

YANG Xian-Di[1] PENG Zhi-Yong[2] LIU Jun-Qiang[2] LI Xu-Hui[2]
(Computer Center, School of Computer, Wuhan University, Wuhan 430072)[1]
(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072)[2]

Abstract Information integration is the problem of combining the data residing at different heterogeneous sources to provid the user with a unified view of these data. The problem of designing information integration systems is important in current real world applications, and becomes a hot topic of the database research. This paper tries to present an overview of this topic and provides some research problems in the future. The important aspects include methods of information integration, information integration framework, query processing, semi-structured data and Web information integration.

Keywords Information integration, Heterogeneous, LAV, GAV, Semi-structured data

1 引言

随着计算机应用需求的不断增强,分布式系统和网络环境日益普及,大量的异构信息源系统被分散在各个网络节点中,而它们之间往往是相互独立的。为了使这些孤立的数据能够更好地实现资源共享,迫切地需要建立一个公共的集成环境,对用户提供一个统一的、透明的访问界面,信息集成的研究因此而起。信息集成所要解决的问题是把位于不同的异构信息源上的数据合并起来,以便为用户提供一个这些数据的统一视图,称为全局模式[1]。信息集成屏蔽了各种异构数据间的差异,通过异构数据集成系统进行统一操作,因此集成后的异构信息对用户来说是统一的、无差异的。

信息集成自被提出以来就引起了国内外众多科研人员的关注,已成为当前数据库领域中的重要研究方向。很多研究项目是针对信息集成问题而开展的,如TSIMMIS[2]、Information Manifold[3]、Garlic[4]等系统都对不同的信息源提供了一种统一的集成存取。信息集成也涉及到了很多基础问题的研究,如集成系统的逻辑框架、全局模式与局部模式的映射方法、异构信息源上的查询推理和优化、中间件/包装器技术、半结构化数据等研究。本文综合了国内外的相关研究文献,对信息集成及相关的应用和研究进行了综述,进而对将来的研究主题进行了展望。

文章第2部分介绍了两种信息集成的方法:物化方法和虚拟方法;第3部分对信息集成的逻辑框架进行了形式化,并在此基础上讨论了两种基本的建模方式:LAV和GAV;第4部分概述了信息集成中的查询处理问题;第5部分讨论了Web上半结构化数据的集成问题;最后是本文的结论,对信息集成领域进行了展望。

2 信息集成的方法

目前,在开发信息集成系统时采用的方法虽各不相同,但其基本的方法可分为两类:物化方法(Materialized,也称数据仓库法)和虚拟方法(Virtual,也称中间件法)。

(1)物化方法:在客户端与数据源(服务器)之间增加一层,称为数据仓库,用于存储来自各数据源的待集成数据,系统提供对这个数据仓库的查询机制。这种方法的优点是既可用于信息集成,又可用于决策支持查询。该方法存在的问题是,当信息源的数据发生变化时,数据仓库中的数据也要做相应的修改。因此,这种间接访问方式的最大缺点是数据更新不及时、数据重复存储。这种方法通常需要一些新的技术,如有效的数据加载和增量更新维护等。

(2)虚拟方法:该方法使用了与数据仓库法完全不同的结构。数据仍保存在各数据源上,集成系统仅提供一个虚拟的集成视图(即全局模式)和对该集成视图查询的处理机制。系

*)国家自然科学基金项目(60273072,60473076)资助。**杨先娣** 讲师、在职博士研究生,主要研究方向:数据库理论与技术、信息集成、数据挖掘;**彭智勇** 教授、博士、博士生导师,主要研究方向:数据模型、高级数据库管理系统、多媒体数据库、安全数据库、网上信息集成等;**刘君强** 博士研究生,主要研究方向:生物信息集成、数据挖掘;**李旭辉** 讲师、博士,主要研究方向:分布式数据库、信息集成。

统能自动地将用户对全局模式的查询请求转换成对各异构数据源的查询，它依赖于两类软件组件：包装器(wrappers)和中间件(mediators)。包装器包装数据源，把底层的数据对象转换为统一的数据模型；在某种程度上，中间件是信息源中数据的一个视图，其中并没有数据。用户可以对中间件进行查询，对于每一个用户的查询模式需要一个中间件，不同中间件结果之间一般没有一致性约束。中间件从包装器或其他中间件获取信息，通过集成不同数据源信息，并解决它们之间的冲突来提炼信息，然后把信息或者提供给用户，或者提供给其他的中间件。由于该方法不需要重复存储大量数据，并能保证查询到最新的数据，因此比较适合于高度自治、集成数量多且更新变化快的异构信息源集成。该方法中的技术涉及到更多的查询上的代数操作。首先，中间件应当确定出哪个信息源对给定的查询有用，当需要集成的信息源巨大时，这一问题是非常重要的；其次，一旦确定了有关的信息源，中间件应当执行源到源的查询变换，该过程有时称为查询重写(Query rewriting)。当从两个或多个信息源抽取数据时，中间件还需要生成一个全局的执行计划，以确定以何种顺序对信息源进行查询。

比较上述两种方法，物化方法中，中间层备份全局模式中的数据，系统需要维护一个与信息源中数据一致的视图副本，全局查询直接在集成系统本地执行，查询不需要访问源数据，所以响应查询一般比较快捷，但维护具体的视图代价也高；特别当数据源更新时必须相应地更新视图；并且存储空间需求大，比较适合于数据仓库这类实时性要求不高的应用。在虚拟方法中，中间层不备份任何数据实例，只作为用户和信息源之间的接口，通过查询规划将全局查询转换成信息源上的查询命令；在处理查询时，由于需要访问信息源，所以响应查询一般比较费时，其查询代价较高。有些系统采用混合方法，同时是虚拟的也是物化的，中间层只保存至关重要、变化较少的数据，或者不常在线信息源中的数据，而其它数据仍直接从信息源本身实时获取，例如 TSIMMIS。

3 信息集成的逻辑框架

异构信息源集成的首要任务就是要为集成系统设计一个公共的逻辑框架，以对全局模式和来自不同信息源的各种数据进行形式化描述，从而便于统一处理。逻辑上，信息集成系统的特性是由全局模式和一系列源模式来描述的。不同的信息源中包含着实际数据，全局模式在这些信息源上提供了一个和谐的、集成的、统一的虚拟视图，可以被用户查询。显然，设计信息集成系统的关键就是在这些源模式和全局模式之间建立映射，这就有必要对信息集成系统进行形式化。目前，被大家所普遍认可的形式化方法是意大利的 Maurizio Lenzerini 等倡导的三元组方法，文[1,7,10,29]都是建立在该逻辑框架基础上的。下面将讨论基于此方法的信息集成的逻辑框架。

由于信息集成系统主要由全局模式、一系列源模式和它们之间的映射组成，那么一个信息集成系统 I 可以形式化为一个三元组〈G,S,M〉，其中，

• G 是全局模式，用语言 L_G 来表示，L_G 是基于字母表 A_G 上的，A_G 包含了 G 中每个元素的符号(即如果 G 是关系型的，则为关系；如果 G 是面向对象型的，则为类，等等)。

• S 是源模式，用语言 L_S 来表示，L_S 是基于字母表 A_S 上的，A_S 包含了 S 中每个元素的符号。

• M 是 G 和 S 之间的映射，由一系列如下形式的断言组成：

$q_S \rightsquigarrow q_G$，$q_G \rightsquigarrow q_S$

其中，q_S 和 q_G 分别是源模式 S 上和全局模式 G 上的查询，它们有相同的算子。查询 q_S 由基于字母表 A_S 上的查询语言 $L_{M,S}$ 来表示，查询 q_G 由基于字母表 A_G 上的查询语言 $L_{M,G}$ 来表示。断言 $q_S \rightsquigarrow q_G$ 表示：源上的查询 q_S 表示的概念对应于全局模式中查询 q_G 表示的概念(断言 $q_G \rightsquigarrow q_S$ 的含义，依此类推)。

这里，信息集成系统 I 的查询根据全局模式 G 提出，用基于字母表 A_G 上的查询语言 L_Q 来表示。从集成系统所呈现的虚拟数据库中抽取哪些数据，这是查询企图提供的说明。根据映射方式的不同，提出了两种基本的建模方式：LAV 和 GAV，下面将分别讨论。

3.1 LAV(Local-as-View)

基于 LAV 方式的信息集成系统 I=〈G,S,M〉中，映射 M 把源模式 S 的每一个元素 s 和 G 上的查询 q_G 联系起来，查询语言 $L_{M,S}$ 只允许由字母表 A_S 上的符号来表示。这样，LAV 中表示映射的一系列断言具有如下形式：$s \rightsquigarrow q_G$

LAV 方式的基本思想是：每个源 s 的内容按照全局模式上的视图 q_G 来定义，即它要求为每一个信息源 S 给出一个针对集成视图的查询，说明集成视图中的哪些元组(或对象)可在 S 中找到。这样，系统必须通过推理演绎才能将全局查询转换成信息源上的子查询，计算复杂度高。但添加一个新信息源只意味着为映射增加一条新的断言，其他不变。Information Manifold[3] 和 Agora[11] 信息集成系统就是采用的 LAV 方式。Information Manifold 用描述逻辑表示全局模式，并且采用合取查询语言作为查询语言 L_Q 和 $L_{M,G}$。Agora 使用 XML 全局模式，用户查询和映射中的查询都采用基于 XML 的查询语言。

3.2 GAV(Global-as-View)

基于 GAV 方式的信息集成系统 I=〈G,S,M〉中，映射 M 把全局模式 G 中的每一个元素 g 和 S 上的查询 q_S 联系起来，查询语言 $L_{M,G}$ 只允许由字母表 A_G 上的符号来表示。这样，GAV 中表示映射的一系列断言具有如下形式：$g \rightsquigarrow q_S$

GAV 方式的基本思想是：全局模式的每个元素 g 的内容按照信息源局部模式上的视图 q_S 来定义，即它要求为集成视图中的每一个虚拟关系(或虚拟对象类)R 写出一个查询，说明如何从信息源得到 R 的元组(或对象)。这样，对于全局查询，只要简单将该查询中的全局视图展开，即可得到相应的子查询，查询处理简单。但添加一个新信息源时，会影响全局模式中的定义，相关的视图必须重新定义。很多信息集成系统都是 GAV 方式的，如 TSIMMIS[2]、Garlic[4]、MOMIS[12] 和 Squirrel[13]。

一些研究学者对这两种基本方式进行了更深入的研究，并提出了其他值得关注的建模方法，例如文[6]介绍了一种 GLAV 方式，该方式综合应用了 LAV 和 GAV 的映射方法，断言是 $q_S \rightsquigarrow q_G$ 的形式，其中 q_S 是源模式上的合取查询，q_G 是全局模式上的合取查询；文[7]对 LAV 和 GAV 进行了详细的比较，并且讨论了 GLAV 到 GAV 的转换。文[8]从视图的查询处理角度对 LAV 和 GAV 两种方式进行了比较；文[5]介绍了一种 BAV(both-as-view)方式，模式之间的映射采用了一系列双向的转换路径，通过这些路径就可能从局部模式上提取以全局模式作为视图的定义(即 GAV)，也可能从全局模式上提取以局部模式作为视图的定义(即 LAV)。BAV

方式优于GAV和LAV方式，在于它真正支持全局和局部模式的演变，包括局部模式的添加和删除，这个特点使其非常适合P2P的信息集成。

上述对信息集成的形式化是建立在一阶逻辑基础上的，即用一阶逻辑来解释映射断言，因此不能用于直接处理不一致的数据源。如果在信息集成系统 $I=\langle G,S,M\rangle$ 中，从各个源中提取的数据不满足G的完整性约束，则I的全局数据库不存在，那么查询应答也将毫无意义。当各个源中的数据相互不一致时，会出现这种情况。实际上，可以对各个源中抽取的数据进行合适的转换和清洗来解决这个问题。

4 信息集成中的查询处理

查询处理是信息集成系统中最重要的问题之一，它是用户和信息集成系统之间进行交互的重要手段。用户在全局视图上以一定的查询语言提出查询，以此获取所需的信息。而系统要正确地响应用户查询，就必须进行一系列复杂的过程：(1)查询分析。中间件首先对全局查询进行分析和合法性检验，以确保其语法、语义正确；(2)查询规划。中间件为合法的全局查询选择信息源，并将其分解为一系列信息源子查询和一个全局查询余项——即包装器不能完成而必须由中间件完成的处理；(3)局部子查询执行。中间件将分解得到的子查询分派到合适的包装器执行，包装器进而将子查询转换成可由信息源完成的查询请求，并将结果返回中间件；(4)返回结果。中间件聚集子查询结果并完成剩余的处理，得到全局查询结果并响应用户。该过程涉及到很多理论问题的探讨(如查询规划、查询优化、查询应答、查询重写、完整性约束条件、查询的包容性、计算复杂性等等)，因而查询处理成为信息集成研究领域中的热点，有关此方向的文献也最多。限于篇幅，在此只做简单的介绍。

先介绍几个通用的假设条件。

• CDA(Closed Domain Assumption)：数据库中的对象集正好和所有扩展视图中出现的对象集一致，即 $D=ext(V_1)\cup\cdots\cup ext(V_k)$。

• ODA(Open Domain Assumption)：所有扩展视图中出现的对象集仅是数据库中的对象集的一个子集，即 $D(ext(V_1)\cup\cdots\cup ext(V_k)$。

• SVA(Sound View Assumption)：如果 $ext(V_i)\subseteq ans(def(V_i),D)$，则称视图 V_i 关于数据库D是Sound的(满足SVA)，记为 $as(V_i)=SVA$。也就是说，如果元组(a,b)出现在 $ext(V_i)$ 中，则可以肯定(a,b)在 $ans(def(V_i),D)$ 中；但如果元组(a,b)没有在 $ext(V_i)$ 中出现，则(a,b)不一定就不在 $ans(def(V_i),D)$ 中。

• CVA(Complete View Assumption)：如果 $ext(V_i)\supseteq ans(def(V_i),D)$，则称视图 V_i 关于数据库D是Complete的(满足CVA)，记为 $as(V_i)=CVA$。也就是说，如果元组(a,b)出现在 $ext(V_i)$ 中，则不能肯定(a,b)在 $ans(def(V_i),D)$ 中；如果元组(a,b)没有在 $ext(V_i)$ 中出现，则(a,b)肯定就不在 $ans(def(V_i),D)$ 中。

• EVA(Exact View Assumption)：如果 $ext(V_i)=ans(def(V_i),D)$，则称视图 V_i 关于数据库D是Exact的(满足EVA)，记为 $as(V_i)=EVA$。也就是说，视图 V_i 的扩展正好就是满足该视图的元组集，EVA既是SVA又是CVA。

4.1 LAV中的查询处理

LAV系统中的查询应答实质是在不完全信息上进行推理的一种扩展形式[9]。在LAV映射的基础上，当在全局模式上响应一个查询时，我们仅知道与源相关的视图的扩展，而这些在全局数据库上只提供了部分信息。在LAV中，信息源被建模为全局模式上的视图，因此其中的查询处理问题也被称为基于视图的查询处理。也就是说，计算一个查询的应答是基于一系列视图，而不是数据库中的原始数据[14]。

基于视图的查询处理有两种方式：基于视图的查询重写和基于视图的查询应答。在前一种方式中，已知一个查询q和一系列视图定义，目标是用一种固定的语言 L_R 来重组该查询，这里的 L_R 只与这些视图相关并为q提供应答。查询重写分两步：首先按照给定的查询语言重新表示该查询，第二步是在视图扩展上对重写的查询求值。这里的关键在于，想要用来重写的语言通常和表示原查询的语言是一致的，查询重写的目的是重组，它独立于当前的源数据库。显然，在目标语言 L_R 中可能不存在与原查询等价的重写。在这种情况下，应求出最大包含重写(maximally contained rewriting)来解决问题，即最佳的表达原查询的方式。一些文献研究了不同类型查询下的重写问题，如合取查询(带或不带算术比较)[17,18]、析取查询[19]、聚集查询[20~22]、递归查询和非递归查询[23]、用描述逻辑(Description Logics)表示的查询[24]、正规路径查询及其扩展形式[25,26]。

在基于视图的查询应答中，除了已知查询q和视图定义以外，还给出了这些视图的扩展，目标是通过视图扩展的知识计算出元组t，t就是q的应答。在查询应答中，对查询以怎样的形式处理并不作任何限制，唯一的目的就是利用所有可能的信息，尤其是一些视图扩展来计算查询的结果。近年来对查询应答的研究也有很多，文[15]给出了一个基于视图查询应答的综合框架和一些有意义的结果，这个框架考虑到视图定义及其扩展的各种不同假设(CDA，ODA，EVA等)；在文[16]中，假设视图和查询按照不同的语言(不带和带不等式的合取查询、正向查询、Datalog、一阶查询)来表示，在这种不同假设下来分析复杂性问题，复杂性的度量与视图扩展(数据复杂性)的大小相关。对于文中所考虑的各种查询语言，EVA使问题复杂化了。例如，对于合取查询的查询应答的数据复杂度，在SVA下是PTIME的，而在EVA下是coNP-complete的，原因是EVA下引进了否定形式，因此它可促使用存储在视图中的对象进行推理，这就使复杂度提高了。文[27]讨论了基于视图的查询应答与约束补偿之间的密切联系，这是一个比较复杂的问题。

4.2 GAV中的查询处理

在很多GAV的信息集成系统中，全局模式中不允许完整性约束条件，并且假设视图是EAV的。在这种条件下，整个信息集成系统就具有单一数据库的特点，因此查询处理可以用简单的展开策略来完成。但是，当表示全局模式的语言允许完整性约束条件，并且视图是SVA时，GAV系统中的查询处理就复杂得多。在这种情况下，完整性约束条件将被用以克服源中数据的不完整性。下面的例子给出了外键约束条件，如果在处理查询时只是简单地展开查询，则会得出错误的结果。

假设信息集成系统 $I=\langle G,S,M\rangle$ 中的G由以下关系组成：

employee(Ecode,Ename,Ecity)；

company(Ccode,Cname)；

employed(Ecode,Ccode).

约束条件为：

key(employee)={Ecode}；

key(company)={Ccode}；

employed[Ecode]⊆employee[Ecode]；

employed[Ccode]⊆company[Ccode].

源模式 S 由 3 个源组成：源 s1 有 4 个属性，即雇员的工号、姓名、城市和年龄；源 s2 有 2 个属性，即公司的代码和名称；源 s3 有 2 个属性，即雇员的工号和公司的代码，表示公司的雇用信息。映射 M 定义如下：

employee ⇝{x,y,z|s1(x,y,z,w)}；

company ⇝{x,y|s2(x,y)}；

employ ⇝{x,w|s3(x,w)}。

假设源数据库 D 中的数据如下所示：

S_1^D：

12	Calvin	Rome	21
15	Alice	HongKong	24

S_2^D：

AF	IBM
BN	Microsoft

S_3^D：

12	AF
16	BN

查询 q 要查找雇员的工号：{x|employee(x,y,z)}，通过简单地展开 q 得的地结果是{12}。但是，考虑到完整性约束条件 employed[Ecode]⊆employee[Ecode]，工号为 16 的雇员也应该是查询的结果，即使他并没出现在 s1 中。所以 q 的正确结果应该是{12,16}。

上述问题的解决途径就是，在全局数据库中增添合适的元组，以满足外键约束的要求，同时要遵从映射关系。具体的做法各不相同，文[10,28]就是围绕这个问题展开的。

5 半结构化数据与 Web 信息集成

WWW 应用的迅速发展产生着海量的 Web 信息，所以半结构化数据的集成是当前的研究热点。近来越来越多的系统使用 XML 作为数据交换的中间模式，XML 是 WWW 上的半结构化数据。XML 除了可以表示关系型数据，还可以表示处理层次与图形结构等其它的数据形式，将成为互联网上数据交换的标准，已经有一些异构信息源集成采用了 XML。因为传统的关系数据模型和对象模型已不能很好地表示 Web 上的信息，所以数据库界出现了半结构化数据模型的研究领域。根据用户的需求，要生成一个全局 XML 视图，用全局 DTD (也称 GDTD)视图表示。

5.1 半结构化数据的模式描述

对于半结构数据的模式，目前已经提出了多种描述形式。比较有代表性的有两种：

(1)基于逻辑的描述形式。如一阶逻辑、描述逻辑以及 Datalog 等，这些描述非常类似，但在表达能力等方面则有所差别。一阶逻辑对于半结构数据来说过于简单，从而不能很好地适应半结构的需求，另外一阶逻辑很可能导致不可判定性或难处理性；描述逻辑是上世纪 80 年代早期提出的，在人工智能、软件工程和数据库领域，已被用来作为知识表示的工具，同样可用于描述半结构化数据模式；Datalog 是一种数据库语言，也可以看作是一种基于逻辑的数据模型。采用 Datalog 规则描述半结构数据模式的主要思想是通过指明应有的入边和出边来对对象的类型进行定义，而这种模式定义即为一组 Datalog 规则。

(2)基于图的描述形式。最典型的就是斯坦福大学提出的 OEM 模型。由于半结构数据一般采用带标记的有向图表示，所以这种描述形式的一个显著特点是模式与数据采用同一种图模型。模型图通常是一个有根的、边上带有标记的有向图。这种边标记图可与数据图相同，也可以对其进行扩充。模型图中的结点也可以加以一定的注释，表明其代表的语义或其他特定的含义等。要注意的是，基于图的模式描述形式中有两个待研究的问题：(1)对于给定的数据图，如何判定该图是否与一个给定的模型相符合；(2)若该给定的数据图与一个给定的模式图的确相符合，则如何得到数据图中的对象与模式图中的类型之间的对应关系。这两个问题都需要更深入地研究，目前对这两类问题的研究主要是使用仿真的概念来解决。

5.2 半结构化数据的抽取

数据抽取通常采用 Wrapper 技术，Wrapper 是根据特定的生成规则从 Web 数据源中执行抽取的程序。目前，Web 站点上的数据信息一般采用 HTML 描述，抽取规则是基于 HTML 文档格式的。有两种方式：(1)将文档看作字符流，抽取规则基于分界符，作为分界符的可以是 HTML 标签、特征字符串和标点符号等，根据这些分界符就可以将所需数据抽取出来；(2)将文档看作树结构，抽取规则基于树路径，首先根据 HTML 标签将文档分析成树结构，再通过规则中的路径在树中搜索相应的结点，最终得到所需数据。

Wrapper 与数据源的格式密切相关，对不同格式 HTML 文档的抽取就需要使用不同的抽取规则，因而每个数据源都需要有各自的 Wrapper。Wrapper 的生成方法可以分为 4 类：(1)手工编写 Wrapper 程序语言的方式，典型例子是 TSIMMIS[30]，抽取过程是基于过程化的程序，但是抽取结果要依赖于文档的结构。(2)机器学习的方法，根据用户提供的一组例子以及用户标记的信息，从大量的 Web 页面中的正例和反例中学习。(3)受指导的交互式 Wrapper 生成方法，提供可视化的向导方式进行 Wrapper 生成。用户可以通过浏览的方式来标记文档，提示例子映射关系。系统经过归纳生成抽取规则，用户可以查看所抽取的数据是否合适以判断抽取规则是否准确。如果不合适，用户可以提供新的页面，重新生成抽取规则，如 Lixto[31] 和 SG-WRAP[32]。(4)自动抽取的 Wrapper，文[33]的 EXALG 是采用公共模板的形式来抽取 Web 信息，它实现了在给定页面集的情况下，自动地抽取结构化数据。

上述方法也各有弊端：所有手工的 Wrapper 生成方法对生手来说都很难使用；机器学习方法的缺点在于需要大量的例子页面，Wrapper 表达能力有限；受指导的交互式 Wrapper 生成方法也需要与用户的交互时间；自动抽取的 Wrapper，缺点主要存在于几个重要的假设：它假设了大量的在模板上产生的标记有唯一的角色、等价类都是有效的、每个类型构造器都已经被实例化等。另外，所有 Wrapper 生成方法都存在一个共同的问题：当 Web 页面格式发生变化时，Wrapper 就会失效，这就提出了 Wrapper 维护的问题，Wrapper 的维护主要涉及两个问题：变化数据项的识别和抽取实例的获取。因此，需要新的研究方法来简化 Wrapper 生成过程，提高 Wrapper 对动态页面变化的自适应能力。研究快速有效地自动生成

Wrapper 的方法，有助于减小维护的代价。

5.3 半结构化数据的查询

由于半结构数据的结构不完全及不规整等特点，直接使用传统查询语言是不合适的。半结构数据查询语言应当自动地使用户从严格的类型约束中解脱出来。也就是说，当在不清楚数据的结构类型时，能够使用路径表达，通过"导航"的方式来遍历数据图。半结构数据查询语言还应当具有对半结构数据的重构能力。目前已经开发出了几种用于半结构数据的查询语言，如 LOREL[34]、WebSQL[35]、WebOQL[36] 及 StruQL[37] 等。这些查询语言的共同特点是，通过使用正规路径表达式，可以遍历数据中任意长度的路径。因此，这种查询语言是递归的。由 AT&T 实验室开发的 XML-QL 是一种可以对 XML 数据进行查询的语言，并且利用 XML-QL 的查询方式可以实现 XML 数据的抽取、转换和集成。

5.4 Web 信息集成

Web 信息源分散、动态的特点使得 Web 上的信息集成比基于数据库的信息集成更复杂，应用更广泛。文[41]提出了 Web 信息集成的新方案 ARAIADNE，它是以层次化的方式对 Web 数据建模，并对其创建索引，从而解决 Web 页面的定位问题，但基于此的查询和优化处理还要进一步改善和提高。InfoSleuth[39] 提出了基于内容的数据分发技术，它通过构建信息代理(Agent)来实现。ARAIADNE 和 InfoSleuth 都通过领域模型(Domain model)或本体论(Ontology)来描述数据和资源的特性，获取数据的模式，从而实现更好的语义共享，达到信息集成统一的数据表示。但它们还没有实现动态的信息集成。鉴于上面的缺点，文[38,40]提出了通过 Web 服务组合进行信息集成的技术，它提出了对一些不是很复杂的服务组件进行服务重用，并规定了服务的粒度。文[42]提出了下一代电子商务上复杂 Web 服务的部分解决方法，包括共享上下文及 Web 服务组合。基于共享上下文，服务组件及基本服务间的关联问题能对服务的组合进行监督和指导，但不能实现复杂 Web 服务执行的自动化，也不能实现服务组件和基本服务的动态交互、协调及状态保持。这些都是今后 Web 信息集成研究的重点。另外，怎样用 Ontology 很好地定义语义信息，怎样验证和测试服务，怎样把基于服务的过程转换为基于数据的集成，怎样保证服务的有序性等，都是 Web 信息集成技术所面临的重大挑战。

结论和展望 随着网络技术的发展和商务处理的全球化，信息集成技术成为下一代 Internet 网中的信息融合、信息处理、信息发布等的关键技术。Web services 的不断研究和发展给信息集成技术提供了更广阔的发展空间。利用本体描述服务的结构、类型和语义，从而使 Web services 语义表示模型化、统一化，从语义层就解决了不同数据源或系统的异构问题；P2P 技术是新兴的基于对等网的架构，在 P2P 计算平台上建立信息集成系统是一种较为理想的实现方案，可以有效地利用 P2P 本身的优势，高效实现信息集成系统资源的自治；网格是一种集成的资源和服务的环境，经常被形容为"虚拟的超级计算机"，为信息集成技术带来了更好的发展契机。用超级计算能力来处理信息集成的海量数据是最好的解决方案之一。所以，下一代信息集成技术将是充分利用传统的信息集成、Web services、P2P、网格等技术，构造一个虚拟的、实现更加高效、准确服务的、具有超级计算能力的、能更好分析数据并获得丰富知识的集成系统。

信息集成的涉及面非常广，是一个具有远大前景和巨大挑战的研究领域。本文仅对该领域的主要理论研究做了简单的综述，这些研究方向也都存在一些有待解决的问题。限于篇幅，一些相关的重要问题在文中没有讨论，列举如下：

1)查询推理，即查询的包容性。

2)信息集成中数据质量、数据清洗、数据协调的问题。

3)在信息集成系统的设计中，怎样建立一个恰当的全局模式，以及逻辑映射断言的建立。

4)在不同的数据源中，怎样建立自动映射数据项的规则。

5)怎样使信息集成系统提供最优化的查询求值。

我们相信，通过对该领域的研究和探讨，能对信息集成所涉及到的关键技术有一个较为全面的认识和理解，从而为今后开发实际的系统奠定坚实的理论基础，也希望这些工作可以为同行们提供有价值的借鉴和参考。

参 考 文 献

1 Calvanese D, De Giacomo G, Lenzerini M. Description logics for information integration. In: Kakas A, Sadri F, editors. Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski, Vol 2408 of Lecture Notes in Computer Science, Springer, 2002. 41～60

2 Garcia-Molina H, Papakonstantinou Y, Quass D, et al. The TSIMMIS project: integration of heterogeneous information sources. Journal of Intelligent Information Systems, 1997, 8(2): 117～132

3 Kirk T, Levy A Y, Sagiv Y, et al. The Information Manifold. In: Proceedings of the AAAI 1995 Spring Symp on Information Gathering from Heterogeneous, Distributed Enviroments, 1995. 85～91

4 Carey M J, Haas L M, Schwarz P M, et al. Towords heterogeneous multimedia information systems: the Garlic approach. In: Proc. Of the 5th Int Workshop on Research Issues in Data Engineering-Distributed Object Management (RIDE-DOM'95). IEEE Computer Society Press, 1995. 124～131

5 McBrien P J, Poulovassilis A. Data integration by bi-directional schema transformation rules. In: Proc. of ICDE03. IEEE, 2003

6 Friedman M, Levy A, Millstein T. Navigational plans for data integration. In: Proc. Of the 16th Nat Conf on Artificial Intelligence (AAAI'99), 1999. 67～73

7 Cali A, Calvanese D, De Giacomo G. On the expressive power of data integration systems. In: Proc. of the 21st Int Conf on Conceptual Modeling (ER2002), Vol 2503 of Lecture Notes in Computer Science. Springer, 2002. 338～350

8 Ullman J D. Information integration using logical views. In: Proc of the 6th Int Conf on Database Theory (ICDT'97), Vol 1186 of Lecture Notes in Computer Science. Springer, 1997. 19～40

9 van der Meyden R. Logical approaches to incomplete information. In: Chomicki J, Saake G, editors. Logics for Databases and Information Systems. Kluwer Academic Publisher, 1998. 307～356

10 Cali A, Calvanese D, De Giacomo G, et al. Data integration under integrity constraints. Information System, 2004, 29: 147～163

11 Manolescu I, Florescu D, Kossmann D. Answering XML queries on heterogeneous data sources. In: Proc. of the 27th Int Conf. on Very Large Data Bases (VLDB 2001), 2001. 241～250

12 Beneventano D, Bergamaschi S, Castano S, et al. Information integration: the MOMIS project demonstration. In: Proc. of the 26th Int Conf. on Very Large Data Bases (VLDB 2000)

13 Zhou G, Hull R, King R, et al. Using object matching and materialization to integrate heterogeneous databases. In : Proc. of the 3rd Int Conf on Cooperative Information Systems (CoopIS'95), 1995. 4～18

14 Halevy A Y. Answering queries using views: a survey. Very Large Database J, 2001, 10(4): 1995. 270～294

15 Grahne G, Mendelzon A O. Tableau techniques for querying information sources through global schemas. In: Proc. of the 7th Int Conf on Database Theory (ICDT'99), Vol 1540 of Lecture Notes in Computer Science, Springer. 1999. 332～347

16 Abiteboul S, Duschka O. Complexity of answering queries using materialized views. In: Proc. of the 17th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems (PODS'98), 1998. 254～265

17 Levy A Y, Mendelzon A O, Sagiv Y, et al. Answering queries using views. In: Proc. of the 14th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems (PODS'95), 1995. 95～104

18 Rajaraman A, Sagiv Y, Ullman J D. Answering queries using templates with binding patterns. In: Proc. of the 14th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems (PODS'95), 1995

(下转第 80 页)

象的视频细粒度可分级编码作为与 MPEG-4 相适应的编码技术，具有重要的理论研究价值和实际应用背景。

(3)空间和质量混合细粒度可分级视频编码技术

所谓空间和质量混合细粒度可分级是指在空间细粒度可分级确定前提下的数率细粒度可分级，或在数率细粒度可分级确定前提下的空间细粒度可分级。Internet 的发展将对视频的可分级编码技术提出更高的要求，信这种混合的细粒度可分级技术一定会得到很好的发展。

参 考 文 献

1 钟玉琢，向哲，沈洪. 流媒体和视频服务器. 北京：清华大学出版社，2003

2 ISO/IEC JTC1/SC29/WG11. Overview of the MPEG-4 standard. N3536, Beijing, 2000

3 Sikora T. MPEG digital video-coding standards. IEEE Signal Processing Magazine, 1997, 14(5): 82～110

4 王相海. 基于小波的图像和视频可分级编码研究：[南京大学博士后研究工作报告]. 2001

5 Li W. Overview of fine granularity scalability in MPEG-4 video standard. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11(3): 301～317

6 ISO/IEC 14496-2/ PDAM4. Coding of Audio-Visual Objects, Part-2 Visual, Amendment 4: Streaming Video Profile, 2000

7 Jiang H, Thayer G M. Using frequency weighting in FGS bitplane coding for natural video. ISO/IEC JTC1/SC29/WG11, MPEG99/M5489, 1999

8 Li W. Frequency weighting for FGS. ISO/IEC JTC1/ SC29/ WG11, MPEG99/M5589, 1999

9 王相海，张福炎. 静态图像编码研究进展. 计算机研究与发展，2001, 38(11): 1315～1326

10 Yan R, Wu F, Li S, et al. Error resilence methods in the FGS enhancement bitstream. ISO/IEC JTC1/SC29/ WG11, MPEG00/ M6207, July 2000

11 van der Schaar M, Radha H, Chen Y. An all FGS solution for hybrid temporal-SNR scalability. ISO/IEC JTC1/ SC29/ WG11, MPEG99/M5552, Dec. 1999

12 Macnicol J, Frater M, Arnold J. Results on fine granularity scalability. Melbourne, Australia: ISO/IEC JTC1/SC29/ WG11, MPEG99/m5122, Oct. 1999

13 Li S, Wu F, Zhang Y Q. Study of a new approach to improve FGS video coding efficiency. ISO/IEC JTC1/ SC29/ WG11, MPEG99/ M5583, Dec. 1999

14 Wu F, Li S, Zhang Y Q. A framework for efficient progressive fine granularity scalable video coding. IEEE Trans on Circuits and Systems for Video Technology, 2001, 11(3): 332～344

15 Li S, Wu F, Zhang Y Q. Experimental results with progressive fine granularity scalable (PFGS) coding. ISO/IEC JTC1/SC29/ WG11, MPEG99/ M5742, 2000

16 Sun X, Wu F, Li S, et al. Macroblock-based progressive fine granularity scalable video coding. IEEE International Conference on Multimediaand Expo(ICME), Tokyo, August, 2001

17 孙晓艳，高文，吴枫，等. 基于宏块的渐进、精细可伸缩的视频编码. 软件学报，2002, 13(11): 2134～2141

18 孙晓艳，高文，吴枫，等. 基于宏块的具有时域和 SNR 精细可伸缩的视频编码. 计算机学报，2003, 26(3): 345～352

19 Wang Q, Wu F, Li S, et al. Fine-granularity spatially scalable video coding. In: IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Salt Lake City, 2001, 3: 1801～1804

20 Horn U, Stuhlmuller K W, Link M, et al. Robust internet video transmission based on scalable coding and unequal error protection. Signal Processing Image Commun, 1999, 15: 77～94

21 Rose K, Regunathan S L. Toward optimality in scalable predictive coding. IEEE Trans Image Processing, 2001, 10(7): 965～976

22 Wilson D, Ghanbari M. Transmission of SNR scalable two layer MPEG-2 coded video through ATM networks. In: Proc. 7th Int Workshop Packet Video, Mar 1996. 185～189

23 Ghanbari M, Seferidis V. Efficient H. 261-based two-layer video codecs for ATM networks. IEEE Trans Circuits Syst Video Technol, 1995, 5: 171～175

24 Ding G G, Guo B L. Macroblock-based fine granularity scalable video coding with leaky prediction. http://www. compscip-reprints. com/Computer_vision/0306001

25 Wang Z, Lu L, Bovik A C. Rate scalable video coding using a foveation-based human visual system model. In: IEEE International Conference on Acoustics, Speech, & Signal Processing, 2001, 3: 1785～1789

(上接第 59 页)

19 Afrati F N, Gergatsoulis M, Kavalieros T. Answering queries using materialized views with disjunction. In: Proc. of the 7th Int Conf on Database Theory(ICDT'99), Vol 1540 of Lecture Notes in Computer Science, Springer, 1999. 435～452

20 Cohen S, Nutt W, Serebrenik A. Rewriting aggregate queries using views. In: Proc. of the 18th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems (PODS' 99), 1999. 155～166

21 Grumbach S, Rafanelli M, Tininini L. Querying aggregate data. In: Proc. of the 18th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems(PODS'99), 1999. 174～184

22 Srivastava D, Dar S, Jagadish H V, et al. Answering queries with aggregation using views. In: Proc of the 26th Int Conf on Very Large Data Bases(VLDB'96), 1996. 318～329

23 Duschka O M, Genesereth M R. Answering recursive queries using views. In: Proc. of the 16th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems (PODS' 97), 1997. 109～116

24 Beeri C, Levy A Y, Rousset M-C. Rewriting queries using views in description logics. In: Proc. of the 16th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems(PODS' 97), 1997. 99～108

25 Calvanese D, De Giacomo G, Lenzerini M, et al. Rewriting of regular expressions and regular path queries. In: Proc of the 18th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems(PODS'99), 1999. 194～204

26 Calvanese D, De Giacomo G, Lenzerini M, et al. Query processing using views for regular path queries with inverse. In : Proc. of the 19th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems(PODS 2000), 2000. 58～66

27 Calvanese D, De Giacomo G, Lenzerini M, et al. View-based query processing and constraint satisfaction. In: Proc. of the 15th IEEE Symp on Logic in Computer Science(LICS 2000), 2000. 361～371

28 De Giacomo G. Intensional query answering by partial evaluation. J of Intelligent Information Systems, 1996, 7(3): 205～233

29 Lenzerini M. Data integration: a theoretical perspective. In: Proc. of the 21st ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS 2002), Madison, WI, USA, 2002. 233～246

30 Hammer J, Brenning M, Garcia-Molina H, et al. Template-based wrappers in the TSIMMIS system. In: Proc. of ACM SIGMOD Conference, Tucson, USA, 1997. 532～535

31 Sahuguet A, Azavant F. Building Light-Weight Wrappers for Legacy web Data-Sources Using W4F. In: Proc. of the Very Large Data Bases(VLDB), 1999. 738～741

32 Meng X F, Lu H J, Wang H Y, et al. SGWRAP: A Schema-Guided Wrapper Generator. In: Proc. of the 18th International Conference on Data Engineering (ICDE' 02), San Jose, USA, 2002. 331～332

33 Arasu A, Garcis-Molina H. Extracting Structured Data from Web Pages. SIGMOD 2003, June 9～12

34 Abiteboul S, Quass D, McHugh J, et al. The lorel query language for semistructured Data. International Journal on digital libraries, 1997, 1(1): 68～88

35 Buneman P. Semistructured data. In: Proc. of the 16th ACM SIGACT SIGMOD SIGART Symp on Principles of Database Systems(PODS'97), 1997. 117～121

36 Arocena G O, Mendelzon A O. WebOQL: Restructuring documents, Databases and Webs. In: Proc. of ICDE. Orlando, FL, 1998. 24～33

37 Fernandez M, et al. A query language for a web-site management system. SIGMOD Record, 1997, 26(3): 4～11

38 Hansen M, Madnick S. Data integration using Web Services. In: Proc. of the VLDB, 2002

39 Bayardo J, Bohrer M. InfoSleuth: Agent-Based semantic integration of information in open and dynamic environment. In: Proc. of the ACM SIGMOD, 1997

40 Cabrera F, Copeland G. Web services coordination. BEA System & IBM Corporation & Microsoft Corporation. 2002

41 Knoblock C A, Minton S. The ariadne approach to web-based information integration. Int' l Journal on Intelligent Cooperative Information Systems(IJCIS), 2001, 10(1-2): 145～169

42 Tsur S. Are web services the next revolution in E-Commerce. In: Apers P. ed. Proc. of the 27th Int' l Conf on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 614～617