

Web-based 多数据源集成系统的研究

Research on Web-based Multiple Data Source Integration System

戴青云 樊沛 李刚 董逸生

(东南大学计算机系 南京210096)

Abstract With the development of Internet, network data increase drastically, and the application requirements according to the network information are becoming more and more, thus traditional methods of browsing by text linking and keyword searching can't satisfy more complex applications. Multiple Data Source Integration System is to set up an environment so that all sorts of sources can cooperate to provide better services. This paper first introduces technological background of multiple data source integration system, then gives the key technology in constructing such a system, and analyzes probable deficiencies in current systems, at last brings forward the future work.

Keywords Information integration, Heterogeneous database integration, Wrapper, Middle model, Semi-structure data, Query optimization

1 引言

多数据源信息集成系统是近年来数据库界和人工智能界研究的热点。在复杂的数据库应用环境中,如 CIMS、DSS、IQ&AS、DW 等,各种数据往往分布在多个不同源系统上,如数据库系统、文件系统,以及 Web 系统。随着时间的推移,这些源系统的数据量将越来越大,尤其是 Web 上的数据可以说是无序性的。如果让用户直接到杂乱无章的数据海洋中查找有用的信息,是不切实际的,但这样的应用需求却日益紧迫。在这种形势下,多数据源系统应运而生,它的主要目的是为了不同的数据源协同工作,为用户提供比较自然的查询界面,使他们能方便地访问所需要的信息。这种形式的信息获取与通常的搜索引擎不同,后者仅能让用户按关键字或其它特征值从多数据源中找到相关的信息,而前者则能接受用户比较复杂的查询,系统通过对各个数据源集成处理,并将各个数据源的查询结果加以综合,返回给用户。

2 多数据源信息集成系统的发展

2.1 异构数据库集成

多数据源信息集成系统的概念最早是源于80年代中期提出的异构数据库的集成。他的典型环境是企业的信息集成。一个企业内部有很多个部门,各个部门又分为若干工作组。各个子系统在最初设计应用系统时并没有考虑到以后的集成问题,因而根据条件采取不同数据库,如 Access、SQL Server、Oracle 等。而随着新的应用需求出现,比如 CIMS 决策支持系统、数据仓库,均需要将各个子系统的信息集成起来,这样,异构数据库集成就成为不可避免的问题。比如在决策支持系统中,企业的决策者要决定某种产品是否继续投放以及产量多少,他必须考虑这种产品的原料成本、生产成本、设备成本、产品的质量、销售情况、利润多少,以及其它厂家的竞争状况等等很多的信息。这些信息常常存放在多个数据库中,很可能是异构的。那么从这些异构数据库中如何将数据集成、查询、分析是非常复杂的问题。多数据源信息集成系统就是完成这一工作的,他可以屏蔽各个数据库间数据结构的冲突,数据语义的冲突,和识别不同的命名方式等,为最终用户提供透明性访问。

2.2 Web-based 多数据信息集成系统

WWW 的发展使 Web 上的数据成为最大的信息来源,它具有全面、及时、多样化的特点,所以基于 Web 的信息集成系统是网络时代的产物。但是 Web 上的数据往往是高度自治的,许多的半结构化数据(包括物理层的半结构和逻辑层的半结构),不可能按传统数据库中结构化数据来处理,这是一个新的挑战。因此基于 Web 的多数据源信息集成系统应该能对终端用户提供自然的访问界面,由多数据源信息集成系统内部机制进行查询重组、优化、数据转换等工作,最终向用户提交回答。这种多数据源信息集成系统应该屏蔽多数据源中数据命名不一致,半结构化和结构化数据的不一致,以及网络上各数据源查询能力的不同等因素。

3 多数据源信息集成系统的技术问题

3.1 技术背景

多数据源信息集成系统的提出和实现有它的技术背景。首先是数据库技术的日益成熟和广泛应用。在数据库中存放着大量的结构化数据,有固定的数据模型,这为多数据源信息集成系统提供了基本的数据来源;SQL 语言提供了强大的对结构化数据的查询功能;分布式数据库的查询优化和执行能力也为多数据源信息集成系统的查询提供了技术基础。其二是 Web 技术的迅猛发展。OSI 网络体系结构的七层模型和支持文本交换的一组协议以及 HTML 语言的广泛应用, B/S 三层 Web 应用的体系结构等技术的成熟为从 Web 上抽取文档提供了必要的基础。近年来数据交换的新的格式标准 XML 以其自身的特点,为 Web 页面的管理、搜索,以及 Web 上的各种应用的连接提供了便利,从而为信息集成提供了另一个思路。所有这些都是多数据信息集成系统开发的技术基础,但是在现存的环境中建立多数据源信息集成系统还必须重点考虑以下几个方面的问题。目前多数据源信息集成系统的体系结构如图1所示。

3.2 数据建模

Web 信息集成系统的目标是支持对 Web 上多个数据源的查询。它除与异构数据库集成系统相同外,还要处理大量的、数目递增的 Web 数据源,描述 Web 数据源特征的元数据很少,各种数据源有很强的自治性。

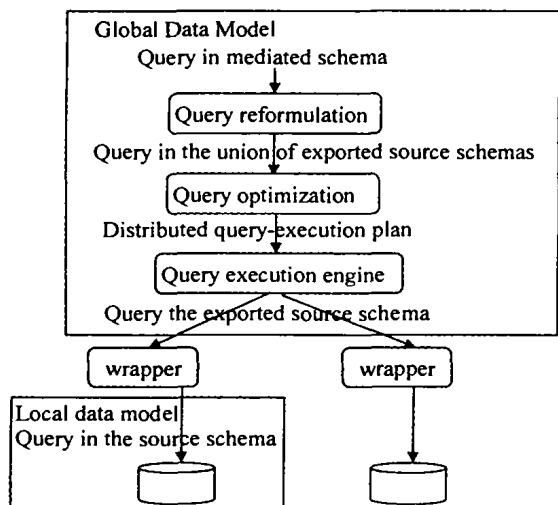


图1 信息集成系统的体系结构

建立 Web 信息集成系统有两种方法：数据仓库方法和虚拟方法。前者是将各数据源的数据装载到数据仓库中，用户的查询基于数据仓库的数据；后一种方法如图1所示，它基于一个“中间模式”(Mediated Schema)，数据仍保存在局部数据源中，通过各数据源的“包装程序”(Wrappers)将数据虚拟成中间模式，用户的查询基于中间模式，不必知道每个专门的数据源的特点，查询执行引擎直接与 Wrappers 打交道，将基于中间模式的查询转换为基于各局部数据源的模式。集中模式和各个数据源的关系一般用“源描述”来表示，源描述包括数据源的内容、属性、内容的约束，以及查询能力等。虚拟方法更适应于数据源数目多、各局部数据源的自治性很高且局部数据经常变化的 Web 环境。

3.3 半结构化数据的处理

半结构化数据有两层含义，一种是指在物理层上缺少结构的数据，另一种是指在逻辑层上缺少结构的数据。有一些结构化数据，如元组(tuple)，为用于 Web 页面的显示而与 HTML 语言的标记符号嵌在一起，构成了物理上的半结构化数据。Web 中有大量丰富的数据：文本、图片、声音、图像等，这些数据多存在于 HTML 文件中，没有严格的结构及类型定义，这些都是逻辑层半结构化的(Semi-structured)数据。

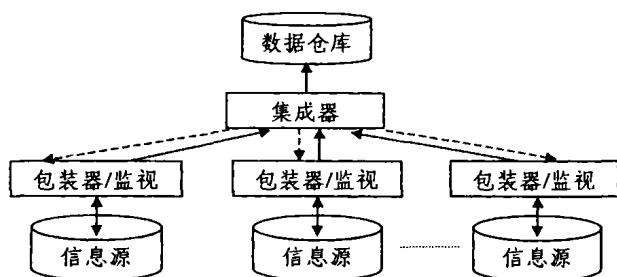


图2 数据仓库的体系结构

在多数数据源信息集成系统中对于半结构化数据的提取、表示和查询，一般通过包装器来处理。图2是多数数据源信息集成系统的典型应用数据仓库体系结构。图中底部是信息源。这不仅指那些常见的数据库，也包括文件、HTML 文件、知识库、Legacy 系统等信息源。连接着每个信息源的是包装器/监视器(Wrapper/Monitor)。包装器负责把源信息的信息格式转换成仓库系统使用的的数据格式和数据模型。监视器部分负责自动检测信息源中数据的变化并把这些变化上报给集成器

(Integrator)。每当有新的信息源挂上仓库系统，或每当信息源中相关信息发生变化时，新的或改变的数据就传送给集成器(Integrator)，集成器对这些信息进行过滤、总结或和其他信息源进行合并处理，并安置在仓库中。为把新信息准确地集成到仓库中，集成器可能还要从原来或相关的其他信息源中获取进一步的信息。

包装器在进行信息转换时面临的主要问题就是如何将半结构化的数据转换成结构化的信息。由于包装器是对应于数据源的，而如此庞大的 Web 数据源使得人工构造包装器的代价太大，因此需要有能快速建造并自动维护包装器的工具。南加州大学开发的 Stalker inductive-learning 系统，能够通过预训练学习到从页面上提取信息的规则。目前这类工具还不是很完善，因为在多数数据源信息集成系统的上层应用中，用户提交查询后才能进行训练，在响应速度和处理效率上似乎不是很实用。

近年来 XML 及其相关技术的迅猛发展，为多数数据源信息集成系统中对半结构化数据的处理提供了解决的方法。XML 文档的优点之一是它的可阅读性，无论是机器还是人都可以将它读懂。它既适合于表示存储于数据库中的数据，又能存储非结构化文档。随着 XML 的广泛应用，Web 上大量的信息都将通过 XML 文档来进行存储、交换，或者通过各种应用接口用 XML 来表现。目前，已有很多处于试运行阶段的 XML 工具，比如：FrameMaker+SGML(用于编辑和排版)、ADEPT Editor(用于内容管理的编辑器)、Xmetal(友好的 XML 编辑器)、XML Styler(图形化的 XSL 样式表编辑器)等等都可以将 HTML 文档根据你自己定义的 XML 样式转化为 XML 文档。对于一些特殊的行业中的二进制数据，如印刷行业，医疗行业的 X 光图像、CT 图像，多媒体行业的视频图像数据和音频数据等，还可以使用 CDATA、或编码转换，或使用 MIME 等方法将这些二进制数据直接嵌进 XML 文档中。

3.4 查询优化及查询处理

传统的关系数据库系统的查询处理过程一般都是由系统对查询语句解析，将结果递交给查询优化器，再由优化器生成查询执行计划，包括 join, selection, projection 等操作的执行顺序，sort-merge join 或 hash join 的方法等。优化器通过评估比较查询执行的代价，确定最优计划，提交给查询执行引擎。

而在多数数据源信息集成系统中，数据分布在远程的自治结点上，优化器知道的信息很少。另外，各数据源的查询处理能力不同，优化器需要考虑各数据源对局部数据访问模式的限制和其特有的抽取数据的演算能力，尽量把各数据源能胜任的工作推到相应的数据源，减少网络数据传输。还要考虑 Web 数据源的完备性问题，对不同的数据源的数据完备性进行评价，分析数据源之间的信息重叠，对回答查询具有重要意义。

查询算法的主要困难在于如何将用户对中间模式(mediated schema)提交的查询重新形式化为对各个数据源的查询，如何选择最少数目且关联度最大的数据源完成查询，以及如何向参与查询的数据源传送最小的查询等问题。对于查询重新形式化，在知识表示系统上用推理机制来进行查询重构是目前比较好的一种方法。对于选取数目最小且关联度最大的数据源，一般是采用匹配源数据与查询，提取。例如，在华盛顿大学开发的 Information Manifold project 中，当大量的数

(下转封四页)

(上接第 120 页)

数据源与给定的查询相关时,多数据源信息集成系统使用 probabilistic information 依赖与选取查询信息和源信息之间交叉的多少,对数据源的访问进行排序,从而在少量关联度较大的数据源中进行查询。再如,AT&T 研究实验室研制的 whirl system,对各个数据源与查询的 textual similarity 排序,从中选取 k 个 highest-scoring answers,其中 k 可以由用户指定。

4 多数据源信息集成系统面临的挑战

多数据源信息集成系统涉及的领域很广,尽管目前在系统的体系结构、多数据源的建模、查询优化及执行、Web 信息的抽取表示和存储等方面的研究已经取得了很大的成果,但是仍然有很多问题没有得到解决,还有的是理论上证明有效但实用性仍值得怀疑的,都急需进一步研究。

4.1 半结构化数据

对半结构化数据的处理,XML 的出现并不可能完全取代包装器的方法,主要有以下几个原因:

(1)将大量异构系统的 Web 应用转换成 XML 的 Web 应用需要较长时间的建立、冲突处理和和维护工作。

(2)用 XML 表示的半结构化数据对于领域内的数据源访问是有利的,但对于跨领域知识的表示在数据的粒度上不可能达到一致。而多数据源信息集成系统面临的事务很可能是面向多领域的。

4.2 语义集成

语义的问题对于 Web-based 多数据源信息集成系统是不可避免的。数据集成主要解决同义词、同名异义词、取值范围冲突、数据丢失、值冲突、语义冲突、结构冲突、不同类型对象之间的映射与转换等问题。其中语义冲突主要由语义异构引起的冲突,语义异构是指对相同或相关信息的不同理解、使用和表达。由于信息语义和行为语义表达了多信息源之间的依赖关系,多信息源系统的集成模型必须能够描述原信息、中间信息、集成信息、局部数据库系统和应用系统的语义,多信息源系统的模式转换和集成必须是带有语义的。

这一点上,前面所述具有中间模式的体系结构中的包装器应该有解决语义冲突的功能。

另外,用对象交换(Object Exchange)方法也是在多数据源信息集成系统中加进语义信息的一种途径。还有用元数据来管理语义的方法也已受到广泛的关注和应用,W3C 提出的

用于描述 Web 资源的 RDF(资源描述框架)就是一个用元数据来描述网络资源的标准,RDF 实际上给出了 Web 数据集成元数据解决方案。但是在具体的实现中,需要涉及到人工智能方面的知识。比如,如果用知识表示语言来描述各个数据源,那么对用户提交的问题和数据源之间的匹配,以及各个数据源之间的语义匹配应会更准确,而不至于产生脏数据。因此,将 AI 与 DB 的方法相结合是今后多数据源信息集成系统研究的主要趋势,否则很难取得突破。如果面向编程,面向机器阅读的语义网能普及,那么语义集成的问题也许能得到更好的解决,但这也面临着和用 XML 解决半结构化数据表示的同样的问题。

结语 多数据源信息集成系统的应用非常广泛,从企业中的 CIMS、DSS,到智能问答系统、数据仓库,都存在着从海量异构数据源中快速有效地提取信息,进行集成,并将结果反馈给用户。一个成功的多数据源信息集成系统应该是独立于特殊的系统和环境的,如果将它做成一个即插即用的插件,能够灵活地与各种应用连接,那么还可以应用于处理日常事务的系统、文件流系统等应用环境中去。

参考文献

- 1 Levy AY. The Information Manifold approach to data integration. IEEE Intelligent Systems, September/October 1998
- 2 Knoblock C A, Minton S. The Ariadne approach to Web-based information integration. IEEE Intelligent Systems, September/October 1998
- 3 Cohen W W. The Whirl approach to information integration. IEEE Intelligent Systems, September/October 1998
- 4 Papakonstantinou Y, et al. Object Exchange Across Heterogeneous Information Sources. IEEE Computer. 251~260
- 5 Jing S, et al. An Object-wrapping Technique for Integrating Non-traditional Database Systems. IEEE International Conference on Intelligent Process Systems
- 6 Levy A Y, Rajaraman A, Ordille J J. Query Heterogeneous Information Sources Using Source Descriptions. In: Proc. 22nd Int'l Conf. Very Large Databases (VLDB-96), Morgan Kaufmann, San Francisco, 1996. 251~262
- 7 Florescu D, Koller D, Levy AY. Using Probabilistic Information in Data Integration. In: Proc. 23rd Int'l Conf. Very Large Databases, Morgan Kaufmann, 1997. 216~225
- 8 Ambite J L, Knoblock C A. Flexible and Scalable Query Planning in Distributed and Heterogeneous Environments. In: Proc. Fourth Int'l Conf. Artificial Intelligence Planning Systems, AAAI Press, 1998. 3~10
- 9 Cohen W W. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. In: Proc. ACM Sigmod-98, ACM Press, New York, 1998. 201~212
- 10 Knoblock C A, et al. Modeling Web Sources for Information Integration. In: Proc. 15th Nat'l Conf. AI (AAAI-98), AAAI Press, 1998. 211~218

计算机科学

(1974年1月创刊)

第29卷第9期(月刊)

2002年9月25日出版

中国标准刊号: ISSN 1002-137X
CN50-1075/TP

定价: 18.00元 国外定价: 5美元

邮发代号: 78-68

发行范围: 国内外公开

主管单位: 国家科学技术部

主办单位: 国家科技部西南信息中心

编辑出版: 《计算机科学》杂志社

重庆市渝中区胜利路132号 邮政编码: 400013

电话: (023) 63500828 E-mail: jsjxx@swic.ac.cn

社长: 牟炳林

主编: 朱宗元

印刷者: 重庆科情印务有限公司

总发行处: 重庆市邮政局

订购处: 全国各地邮政局

国外总发行: 中国国际图书贸易总公司(北京399信箱)

国外代号: 6210-MO