

数据集成综述

陈跃国 王京春

(清华大学自动化系 北京100084)

摘要 介绍了数据集成的基本概念及其难点,讨论并比较了模式集成方法、数据复制方法以及综合数据集成方法,详细阐述了数据集成的主要难点——数据源的异构性,最后对数据集成的研究前景做出了展望。

关键词 数据集成,模式集成,数据复制,异构性

A Review of Data Integration

CHEN Yue-Guo WANG Jing-Chun

(Department of Automation, Tsinghua University, Beijing 100084)

Abstract Firstly, the concepts and difficulties of data integration are introduced. Secondly, three typical data integration approaches: Schema Integration, Data Replication and Hybrid Approach are discussed and compared. Then, heterogeneity problem, a main difficulty of data integration, is explained in detail. Finally, further research fields are suggested.

Keywords Data integration, Schema integration, Data replication, Heterogeneity

1 数据集成概述

1.1 数据集成的基本概念

就大型企业和政府部门的信息化而言,信息系统建设通常具有阶段性和分布性的特点,这就导致“信息孤岛”现象的存在。“信息孤岛”造成系统中存在大量冗余数据、垃圾数据,无法保证数据的一致性,从而降低信息的利用效率和利用率。为解决这一问题,人们开始关注数据集成研究。数据集成的核心任务是要将互相关联的分布式异构数据源集成到一起,使用户能够以透明的方式访问这些数据源^[1]。集成是指维护数据源整体上的数据一致性、提高信息共享利用的效率;透明的方式是指用户无需关心如何实现对异构数据源数据的访问,只关心以何种方式访问何种数据。实现数据集成的系统称作数据集成系统(图1),它为用户提供统一的数据源访问接口,执行用户对数据源的访问请求。

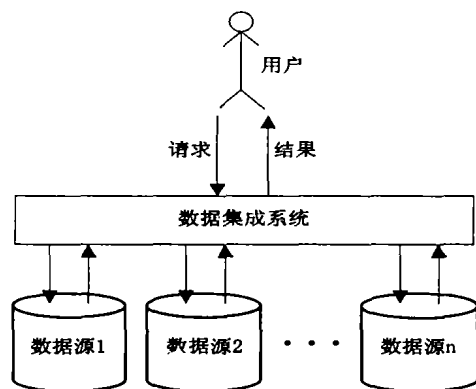


图1 数据集成系统模型

数据集成的数据源主要指 DBMS,广义上也包括各类 XML 文档、HTML 文档、电子邮件、普通文件等结构化、半结

构化信息。

数据集成是信息系统集成的基础和关键。好的数据集成系统要保证用户以低代价、高效率使用异构的数据。要实现这个目标,必须解决数据集成中的一些难题。

1.2 数据集成难点

数据集成的难点^[2,3]可以归纳为以下主要方面:

(1) 异构性:被集成的数据源通常是独立开发的,数据模型异构,给集成带来很大困难。这些异构性主要表现在:数据语义、相同语义数据的表达形式、数据源的使用环境等。

(2) 分布性:数据源是异地分布的,依赖网络传输数据,这就存在网络传输的性能和安全性等问题。

(3) 自治性:各个数据源有很强的自治性,它们可以在不通知集成系统的前提下改变自身的结构和数据,给数据集成系统的鲁棒性提出挑战。

为了解决这些难题,人们尝试了很多方法。但还没有完全解决数据集成中的一些难题,这也是人们一直关注数据集成研究的原因。本文综述了目前已有的几种典型数据集成方法:模式集成方法、数据复制方法以及在这两种方法基础上的综合方法。

2 数据集成方法

2.1 模式集成方法

模式集成是人们最早采用的数据集成方法^[4,5]。其基本思想是,在构建集成系统时将各数据源的数据视图集成成为全局模式,使用户能够按照全局模式透明地访问各数据源的数据^[6]。全局模式描述了数据源共享数据的结构、语义及操作等。用户直接在全局模式的基础上提交请求,由数据集成系统处理这些请求,转换成各个数据源在本地数据视图基础上能够执行的请求。模式集成方法的特点是直接为用户提供透明的数据访问方法。由于用户使用的全局模式是虚拟的数据源

陈跃国 硕士研究生,主要从事电子政务,数据集成,知识管理等研究。王京春 副教授,主要研究方向为流程行业综合自动化系统理论及应用,包括多变量预测控制、CIMS 总体设计、供需链系统等。

视图,一些学者也把模式集成方法称作虚拟视图集成方法^[3]。

模式集成要解决两个基本问题^[7]:构建全局模式与数据源数据视图间的映射关系;处理用户在全局模式基础上的查询请求。

模式集成过程需要将原来异构的数据模式作适当的转换,消除数据源间的异构性,映射成全局模式。全局模式与数据源数据视图间映射的构建方法有两种^[6]:全局视图法(Global-as-View,也有称作 global-centric)和局部视图法(Local-as-View,也有称作 source-centric)。全局视图法^[8]中的全局模式是在数据源数据视图基础上建立的,它由一系列元素组成,每个元素对应一个数据源,表示相应数据源的数据结构和操作;局部视图法^[9]先构建全局模式,数据源的数据视图则是在全局模式基础上定义,由全局模式按一定的规则推理得到。文[10]给出了两种方法的比较。

用户在全局模式基础上查询请求需要被映射成各个数据源能够执行的查询请求,这一过程有很多算法^[6,10,11],其中基于局部视图法的映射算法^[10,11]比较复杂,而基于全局视图法的映射算法^[6]则要简单许多。

联邦数据库和中间件集成方法是现有的两种典型的模式集成方法。

2.1.1 联邦数据库 是早期人们采用的一种模式集成方法^[12]。联邦数据库中数据源之间共享自己的一部分数据模式,形成一个联邦模式。

联邦数据库系统按集成度可分为两类^[3]:采用紧密耦合联邦数据库系统和采用松散耦合联邦数据库系统。紧密耦合联邦数据库系统使用统一的全局模式,将各数据源的数据模式映射到全局数据模式上,解决了数据源间的异构性。这种方法集成度较高,用户参与少;缺点是构建一个全局数据模式的算法复杂,扩展性差。松散耦合联邦数据库系统比较特殊,没有全局模式,采用联邦模式。该方法提供统一的查询语言,将很多异构性问题交给用户自己去解决。松散耦合方法对数据的集成度不高,但其数据源的自治性强、动态性能好,集成系统不需要维护一个全局模式。

2.1.2 中间件集成方法 此集成方法是另一种典型的模式集成方法,它同样使用全局数据模式。G. Wiederhold 最早给出了基于中间件的集成方法的构架^[13]。与联邦数据库不同,中间件系统不仅能够集成结构化的数据源信息,还可以集成半结构化或非结构化数据源中的信息,如 Web 信息。斯坦福大学 Garcia-Molina 等人在1994年开发了 TSIMMIS 系统^[14],就是一个典型的中间件集成系统。

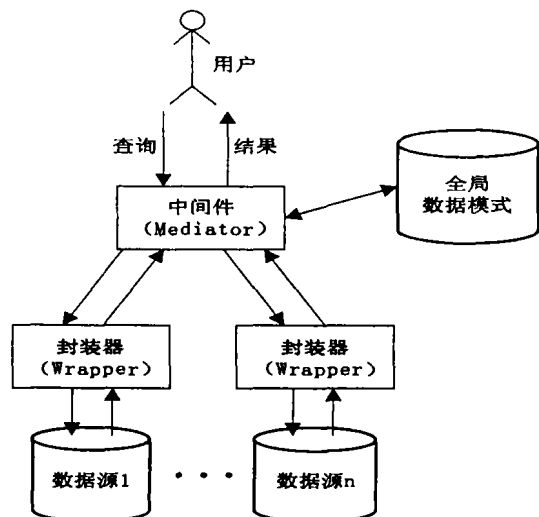


图2 基于中间件的数据集成模型^[3]

典型的基于中间件的数据集成系统(图2)主要包括中间件和包装器,其中每个数据源对应一个包装器,中间件通过包装器和各个数据源交互。用户在全局数据模式的基础上向中间件发出查询请求。中间件处理用户请求,将其转换成各个数据源能够处理的子查询请求,并对此过程进行优化,以提高查询处理的并发性,减少响应时间。包装器对特定数据源进行了封装,将其数据模型转换为系统所采用的通用模型,并提供一致的访问机制。中间件将各个子查询请求发送给包装器,由包装器来和其封装的数据源交互,执行子查询请求,并将结果返回给中间件。

中间件注重于全局查询的处理和优化,相对于联邦数据库系统的优势在于:它能够集成非数据库形式的数据源,有很好的查询性能,自治性强;中间件集成的缺点在于它通常是只读的,而联邦数据库对读写都支持。

2.2 数据复制方法

数据复制方法^[15]将各个数据源的数据复制到与其相关的其它数据源上,并维护数据源整体上的数据一致性、提高信息共享利用的效率。数据复制可以是整个数据源的复制,也可以是仅对变化数据的传播与复制。数据复制方法可以减少用户使用数据集成系统时对异构数据源的数据访问量,从而提高数据集成系统的性能。

最常见的数据复制方法就是数据仓库方法^[16]。该方法将各个数据源的数据复制到同一处——数据仓库。用户则像访问普通数据库一样直接访问数据仓库。

数据复制方法可以从数据传输方式和数据复制触发方式两个方面来划分。

数据传输方式是指数据在发布数据的源数据源和订阅数据的目的数据源间的传输形式,可分为数据推送和数据拉取。数据推送是指源数据源主动将数据推送到目的数据源上。而数据拉取则是目的数据源主动向源数据源发出数据请求,从源数据源获取数据到本地。在有些情况下,数据发布端传送到数据订阅端的数据并不直接存储到目的数据源中,需要经过数据订阅端的本地化处理,这时通常采用缓存来协调数据发布端和数据订阅端的异步。在数据推送的方式下,数据缓存要构建在数据订阅端;而在数据拉取的方式下,数据缓存则要构建在数据发布端。

数据复制触发方式是指集成系统调用数据复制的方式。集成系统通常预先定义了一些事件,这些事件可以包括:对数据发布端引起的数据变化的某个操作、数据发布端数据缓存累积到一定批量、用户对某个数据源发送访问请求、具有一定间隔的时间点等。当这些事件被触发时执行相应的数据复制。因此,数据复制触发方式按事件定义的不同可以分为:数据变化触发、批量触发、客户调用触发、定时触发等。

数据复制通常直接采用端到端方式,也有一些数据集成系统使用专为数据周转服务的数据平台。数据复制时,数据发布者先将数据传送到这个数据平台上,由数据平台处理后转发给数据订阅者。数据平台要处理好网络负担和并发控制问题。使用数据平台的好处是单点控制、便于管理。但数据平台增加了系统的复杂性,降低了系统的可靠性。

2.3 综合型集成方法

以上两种数据集成方法各有优缺点及适用范围。模式集成方法为用户提供了全局数据视图及统一的访问接口,透明度高;但该方法并没实现数据源间的数据交互,用户使用时经常需要访问多个数据源,因此该方法需要系统有很好的网络

性能。数据复制方法在用户使用某个数据源之前,将用户可能用到的其它数据源的数据预先复制过来,用户使用时仅需访问某个数据源或少量的几个数据源,这会大大提高系统处理用户请求的效率;但数据复制通常存在延时,使用该方法时,很难保障数据源之间数据的实时一致性。

表1 两种数据集成方法的比较

集成方法	优点	缺点
模式集成方法	实时一致性好 透明度高	执行效率低 网络依赖性强 算法复杂
数据复制方法	执行效率高 网络依赖性弱	实时一致性差

模式集成方法适用于被集成的系统规模大、数据更新频繁、数据实时一致性要求高的情况。当很难预测用户的查询需求时,也适合采用这种方法。在模式集成方法中,人们通常采用中间件方法。由于联邦数据库在集成时需要为每个数据源单独编写大量的通讯接口,因此现在单纯的联邦数据库方法已很少被采用。

数据复制则适用于数据源相对稳定、用户查询模式已知或有限的情况。当数据分布性比较广,网络延迟较大,同时又需要有很短的处理时间时,也可考虑采用数据复制集成方法。有些应用场合需要对数据进行备份,这时通常采用数据复制方法;还有一些场合,出于机密性的考虑,数据不允许复制,这时就要使用模式集成方法了。

为了突破两种方法的局限性,人们通常将这两种方法混合在一起使用,即所谓的综合方法。综合方法通常是想办法提高基于中间件系统的性能^[3],该方法仍有虚拟的数据模式视图供用户使用,同时能够对数据源间常用的数据进行复制。对于用户简单的访问请求,综合方法总是尽力通过数据复制方式,在本地数据源或单一数据源上实现用户的访问需求;而对那些复杂的用户请求,无法通过数据复制方式实现时,才使用虚拟视图方法。文[17]给出了一种综合集成方法。

3 数据集成的数据源异构性问题

数据源的异构性一直是困扰很多数据集成系统的核心问题,也是最近人们在数据集成方面研究的热点^[18,19]。异构性的难点主要表现在语法异构和语义异构上。

语法异构一般指源数据和目的数据之间命名规则及数据类型存在不同。对数据库而言,命名规则指表名和字段名。语法异构相对简单,只要实现字段到字段、记录到记录的映射,解决其中的名字冲突和数据类型冲突。这种映射都很直接,比较容易实现。因此,语法异构无需关心数据的内容和含义,只要知道数据结构信息,完成源数据结构到目的数据结构之间的映射就可以了。

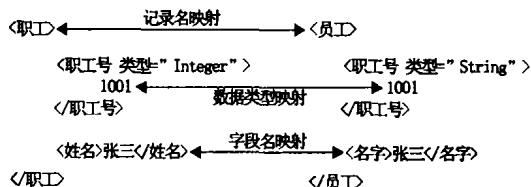


图3 语法映射实例

从图3中我们可以看出语法异构的特点,字段数据内容

(“1001”、“张三”)在映射过程中没有发生变化,我们称之为保持了字段的原子性。

当数据集成要考虑数据的内容和含义时,就进入到语义异构的层次上。语义异构要比语法异构复杂得多,它往往是需要破坏字段的原子性,即需要直接处理数据内容。常见的语义异构包括以下一些方式:字段拆分、字段合并、字段数据格式变换、记间字段转移等(图4~7)。

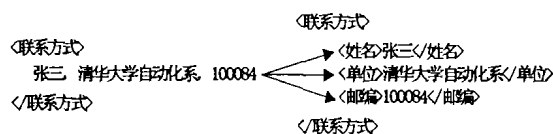


图4 字段拆分

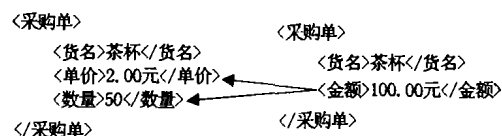


图5 字段合并

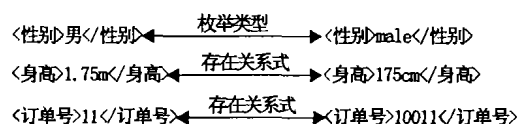


图6 字段数据格式变换

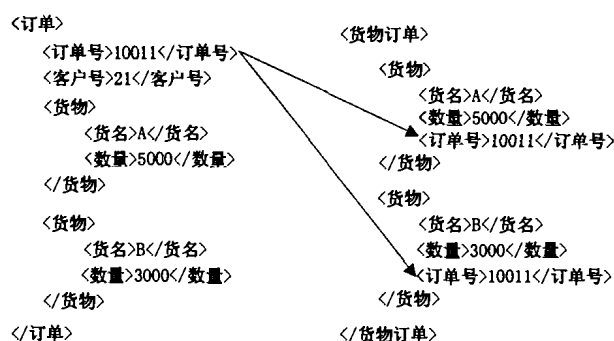


图7 记录间字段转移

语法异构和语义异构的区别可以追溯到数据源建模时的差异:当数据源的实体关系模型相同,只是命名规则不同时,造成的只是数据源之间的语法异构;当数据源构建实体模型时,若采用不同的粒度划分、不同的实体间关系以及不同的字段数据语义表示,必然会造成数据源间的语义异构,给数据集成带来很大麻烦。

事实上,现实中数据集成系统的语法异构现象是普遍存在的。上面提到的几种语法异构属于较为规则的语法异构,可以用特定的映射方法解决这些问题。还有一些不常见或不易被发现的语法异构,例如数据源在构建时隐含了一些约束信息,在数据集成时,这些约束不易被发现,往往会造成错误的产生。如某个数据项用来定义月份,隐含着其值只能在1到12之间,而集成时如果忽略了这一约束,很可能造成荒谬的结果。此外,复杂的关系模型也会造成很多语义异构现象。对于这些复杂问题,还有待人们进一步去研究。

展望 数据集成经过多年的研究,已形成了几种典型的集成方法。虽然各种方法还存在一定的缺陷,但随着一些关键

问题的解决,依赖这些方法的数据集成方案被越来越多地应用到各个领域。我们认为,今后数据集成的研究应该注重以下方面:

(1)关系数据模型与基于XML的半结构化数据间的映射。要保证映射前后数据的完整性及一致性约束。

(2)半结构化数据全局模式的构建方法和映射方法。同样要保证数据的完整性和一致性约束能够在半结构化的数据间传递。

(3)数据集成过程中安全可靠的数据传输。

(4)依赖网格计算技术构建的数据集成解决方案。

我们相信,在XML、Web Services及网格计算等技术规范的推动下,数据集成中的一些难题将会得到很好的解决,数据集成的应用也会更加广泛。

参考文献

- Maurizio L. Data Integration: a Theoretical Perspective. In: Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 2002
- Hasselbring W. Information System Integration. Communications of the ACM, 2000, 43(6):33~38
- Convey C, Karpenko O, Tatbul N. Data Integration Services. <http://www.cs.brown.edu/people/tatbul/cs227/chapter.pdf>, 2001
- Batini C, Lenzerini M. A comparative analysis of methodologies for database schema integration. ACM Computing Surveys, 1986, 18(4):323~364
- Spaccapietra S, Parent C, Dupont Y. Model Independent Assertions for Integration of Heterogeneous Schemas. VLDB Journal, 1992, 1(1): 81~126
- Halevy A Y. Theory of answering queries using views. SIGMOD Record, 2000, 29(4):40~47
- Cali A, et al. Accessing Data Integration Systems through Conceptual Schemas. In: Proc. of ER2001, 2001. 270~284
- Goh C H, et al. Context interchange: New features and formalisms for the intelligent integration of information. ACM Trans. on Information System, 1999, 17(3):270~293
- 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
- Ullman J D. Information integration using logical views. In: Proc. of ICDT'97, volume 1186 of LNCS, 1997. 19~40
- Abiteboul S, Duschka O. Complexity of answering queries using materialized views. In: Proc. of PODS'98, 1998. 254~265
- Sheth A P, Larson J A. Federated databases for managing distributed, heterogeneous, and autonomous databases. Computing Surveys, 1990, 22(3): 183~236
- Wiederhold G. Mediators in the architecture of future information systems. IEEE Computer, 1992, 25(1): 38~49
- Chawathe S, Garcia-Molina H, Hammer J, et al. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In: 10th Meeting of the Information Processing Society of Japan (IPSJ), 1994. 7~18
- Benefits of Transformational Data Integration. <http://www.grcdi.nl/benefits.pdf>
- Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. SIGMOD Record, 1997, 26(1): 65~74
- Ashish N. Optimizing Information Mediators By Selectively Materializing Data: [PhD thesis]. USC, March 2000
- Cali A, et al. Data Integration under Integrity Constraints. CAISE, 2002. 262~279
- Fagin R, et al. Data Exchange: Semantics and Query Answering. In: Intl. Conf. on Database Theory (ICDT), 2003

(上接第14页)

参考文献

- 都志辉,陈渝,刘鹏. 网格计算. 清华大学出版社, 2002
- Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. <http://www.globus.org/research/papers/anatomy.pdf>
- Foster I, Kesselman C, Nick J M, Tuecke S. The Physiology of the Grid: An Open Grid Service Architecture for Distributed Systems Integration. <http://www.globus.org/research/papers/ogsa.pdf>
- Foster I, Kesselman C, Nick J M, Tuecke S. Grid Services for Distributed System Integration. IEEE June 2002. 37~46. <http://www.globus.org/research/papers/ieee-cs-2.pdf>
- Jackson M, Antonioletti M, Krause M. Grid Data Service Factory. <http://www.ogsa-dai.org/docs/current/OGSA-DAI-USER-UG-GDSF.pdf>, June. 2003
- Antonioletti M, Jackson M. OGSA-DAI Product Overview. <http://www.ogsa-dai.org/docs/current/OGSA-DAI-USER-UG-PR-ODUCT-OVERVIEW.pdf>, May. 2003
- Jackson M. DAI Service Group Registry. <http://www.ogsa-dai.org/docs/current/OGSA-DAI-USER-UG-DAISGR.pdf>, June. 2003
- Hong N P C, et al. Grid Database Service Specification. <http://www.gridforum.org/Meetings/ggf7/drafts/DAIS-GGF7S-tatementSpec.pdf>, Feb. 2003
- Atkinson M P, et al. Grid Database Access and Integration: Requirements and Functionalities. <http://www.gridforum.org/Meetings/ggf7/drafts/DAIS-GGF7RF.pdf>, 2003
- Raman V, et al. Chaitan Baru Services for Data Access and Data Processing on Grids. <http://www.gridforum.org/Meetings/ggf7/drafts/DAIS-DataServices.pdf>, Feb. 2003
- Oracle and the Grid. <http://otn.oracle.com/products/oracle9i/grid-computing/OracleGridWP.pdf>, Nov. 2003
- Jacob B, et al. Enabling Applications for Grid Computing with Globus. <http://www.redbooks.ibm.com/redpieces/pdfs/sg246936.pdf>, May. 2003
- <http://www.ogsa-dai.org.uk>