

数据交换研究

李亚楠, 刘连忠, 贾熹星

(北京航空航天大学 电子政务研究所, 北京 100083)

摘 要:随着信息技术的不断发展,数据交换逐渐成为实现信息共享与高效利用的关键技术之一。归纳了数据交换技术发展过程及其主要研究方向,分析了典型的数据交换概念。通过总结数据交换领域的研究状况,归纳出了两种数据交换模式,探讨了现有的数据交换系统存在的不足之处。在分析了策略的优点后,提出了策略驱动的数据交换机制的想法。作为总结给出了数据交换领域的发展趋势。

关键词:数据交换;XML;策略;综述

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2008)02-0005-04

Survey on Data Exchange

LI Ya-nan, LIU Lian-zhong, JIA Yi-xing

(Institute of E-Government, Beihang University, Beijing 100083, China)

Abstract: With the development of the information technology, data exchange has generally become critical technology of achieving information sharing and highly efficiently use. In the paper, the evolvement and main research fields of data exchange were summarized, the classical definition of data exchange was analyzed and expressed in formal language. Through summarizing the current research activities, two patterns of data exchange were brought up. Additionally, some disadvantages of existing data exchange systems were discussed and the policy-driven data exchange mechanism was proposed. Finally, it was concluded with some promising tendency about research on data exchange.

Key words: data exchange; XML; policy; survey

1 数据交换起源与发展

1.1 数据交换研究的必要性和重要意义

20世纪60年代末,人们就认识到了数据交换的重要性。早期的一个比较好的数据交换系统是1977年由IBM开发的EXPRESS^[1],其主要功能是在异构的模式之间实现转换。如今,随着网上不同格式数据的迅猛增长及其信息共享需求的增加,数据交换变的愈加重要。

数据交换问题解决后,才能对其它诸如OLAP, OLTP等提供数据基础;数据交换质量的好坏直接影响在交换后的数据上其它应用能否有效进行;数据交换可以避免不同数据在结构和语义上的差异造成的数据转换引起的错误。因此,数据交换对政府、企业信息化的发展意义重大。

1.2 数据交换研究发展史

对数据交换的研究始于20世纪60年代末,那时的研究还很少。最显著的贡献是CODASYL工作组成员的工作。他们试图开发出一种定义数据结构、结构间关系和实现结构转换的通用方法^[2]。同时Michigan大学也进行了相应研究,Sibley和Taylor在文献[3]中通过对数据定义和映射语言需求进行分析,提出了一种设计数据表示、转换语言的通用方法。此外S. Lin等针对人文学科中特定的自定义数据的转换进行研究,提出了一种图形驱动的数据转换方法。

然而CODASYL工作组以及Sibley和Taylor的工作太广泛,使得具有经济可行性方案的实现就需要太多的研究时间。而S. Lin和J. Heller针对特定应用提出的方案又太窄,没有真正解决数据转换的主要问题。

于是在1976年V. Y. Lum等人提出一种通用的数据转换和重构方法,它是上面两种方法的折中,能在较短的时间内为较广大的用户所用^[2]。并于1997年在原型系统EXPRESS中得到了验证。在EXPRESS中,为了完成数据的表示和转换设计了DEFINE和CONVERT两种语言^[2]。

收稿日期:2007-05-20

基金项目:国家“863”计划资助项目(2005AA113040)

作者简介:李亚楠(1982-),女,河北唐山人,硕士研究生,研究方向为数据交换、信息集成;刘连忠,教授,研究方向为网络安全、数据库技术。

上述研究大多是基于如何定义数据表示和转换语言以实现数据的转换开展的,如 IBM1976 年定义的 DEFINE 和 CONVERT,1968 年 10 月由 ANSI X3 工作组建立的 Ad Hoc 委员会对数据表示语言的研究。W3C 组织于 2000 年 10 月 6 日发布的 XML1.0 版本后,XML 的特点使其逐渐成为数据表示和转换的标准,同时由于大多数数据存储在关系数据库 RDB 中,从此对 XML 及其与 RDB 转换的研究成为热点。

美国 Stanford 大学的 Lore 项目是最早开展 XML 与 RDB 相关研究的项目^[4]。

A.Deutsch 等人在 1998 年进行的 STORED 项目中采用数据模式发现技术对 DTD 到关系模式转换算法进行了首次尝试^[5]。J. Shanmugasundaram 提出了三个著名的内嵌算法^[6]。Lee Dongwon 博士第一个在数据结构正确转换的基础上,提出了数据语义的约束保留机制^[7]。IBM 的 Ronald Fagin, Phokion G. Kolaitis 等人结合 Clio 数据交换原型系统,对数据交换的概念进行了形式化定义,给出了数据交换环境下的查询响应相关算法^[8]。同时提出一种方法,能从所有的数据交换的解决方案中,确定一个特定的解决方案类,称之为“universal”,并证明了 universal 解决方案正好代表了数据交换问题的解空间。并对 universal 解决方案进一步研究,提出了 core 的概念,core 同样是 universal solution,但是具有最小性和惟一性的特点,这使它成为数据交换最理想的解决方案^[9]。

纵观数据交换研究的发展历史可以看出,数据交换研究的发展离不开数据表示和模式间转换语言的发展,当 XML 逐渐成为数据表示和交换的标准后,对基于 XML 的数据交换的研究成为热点,也是以后数据交换研究发展的趋势。

2 数据交换的研究现状

2.1 数据交换的定义

一个权威的数据交换定义是 IBM Almaden 研究所的 Ronald Fagin 等人对数据交换的定义^[8,9]:数据交换可以表示为四元组 $(S, T, \sum_{st}, \sum_t)$ 。其中 S 表示源模式, T 表示目标模式, \sum_{st} 表示源模式 S 和目标模式 T 之间的依赖关系集, \sum_t 表示 T 中存在的依赖关系集。数据交换描述为:对于一个给定的满足源模式的实例 I ,找到一个 J ,并且 J 需要满足如下条件: J 既符合目标格式又满足目标关系集 \sum_t ; I 和 J 之间的关系满足 S 和 T 之间的依赖关系集 \sum_{st} 。

2.2 数据交换的主要研究方向

归纳现有数据交换研究活动,主要包括如下四个

方面:

(1) XML 模式语义建模理论的研究。

目前有十几种 XML 模式描述语言,包括 DTD, XDR, SOX, XML - Schema, Schematron, DSD 等等。其中 DTD, XML - Schema, XDR 和 SOX 属于语法模式, Schematron 属于类型模式, DSD 则介于二者之间。不同的模式描述语言对语义约束机制的支持不同,对 DTD 和 XML - Schema 的语义研究较多。其中 Dongwon Lee 博士基于 UCLA 大学的 EXPRESS 项目提出了基于 XML - Schema 的语义建模理论^[10]。Jennifer Widow 教授及 J. Shanmugasundaram 博士对 XML 的语义建模理论也有研究。

(2) 关系模式到 XML 模式的映射研究(见表 1)。

表 1 关系模式到 XML 模式映射的研究现状

方法	关系模式与 DTD	关系模式与 XML - Schema
面向数据结构	Deutsch ^[5] J. Shanmugasundaram ^[6] V. Turau ^[11]	Lee Dongwon ^[13]
保留语义约束	Lee Dongwon ^[7] Joseph Fong ^[12]	Jia Bei ^[14] Sun Hongwei ^[15]

(3) XML 模式到关系模式的映射研究。

在现有的研究中没有给出(2)和(3),是可逆过程的证明。对于(2)和(3),根据研究的 XML 模式文件是 DTD 还是 XML Schema 及其是只注重数据结构的转换还是同时做到了语义约束的保留这两个方面,可以将研究分成四类。

在面向结构的关系模式与 DTD 的映射算法中:A.Deutsch 等人进行了首次尝试,采用数据模式发现技术,根据 XML 实例文件获得一个合理的 XML 模式描述 DTD,然后研究 DTD 到关系模式的映射算法^[5]。J. Shanmugasundaram 提出了三个著名的内嵌算法,是面向数据结构转换方面最著名的转换算法^[6]。Turau 在 1999 年对关系数据库的单表到 DTD 的转换进行了研究^[11]。

在保留语义约束的关系模式与 DTD 的映射算法中:Lee Dongwon 博士第一个在数据结构正确转换的基础上,提出了数据语义的约束保留机制。他在 J. Shanmugasundaram 内嵌算法的基础上,对 DTD 的语义约束进行了分析,对数据语义约束转换意义、方法、实现算法进行了完整的描述^[7]。Joseph Fong 通过使用 EER(Extended Entity Relationship)模型实现关系数据的概念模型到 XML 模式的转换。

在面向数据结构的关系模式与 XML - Schema 的映射算法中:Lee Dongwon 博士提出了面向数据结构的统一 XML Schema 到关系数据模式的映射算法^[13]。

在保留语义约束的关系模式与 XML - Schema 的映射算法中:Jia Bei 等人提出的方法通过灵活地制定提取和载入规则限制了生成的 XML 结构和将 XML 文件载入关系数据库的具体形式^[14]。孙宏伟等人建立了基于正则树的 XML 形式化描述方法,提出了保留语义约束的 XML - Schema 到关系模式的映射算法^[15]。

(4)源 XML 文档和目标 XML 文档之间的转换的研究。

在源 XML 文档与目标 XML 文档之间的转换方面,目前有多种方法,将现有研究方法分类列在表 2 中。其中的一个主流是使用 XSLT 完成 XML 文档之间的转换,比如微软的 BizTalk Server 和 IBM 的交换原型 Clio^[3],但这种方法存在语法繁杂,只能对单一源 XML 进行映射操作等不足之处。为此王瑜等提出基于映射函数的计算方式,将复杂的 XML 数据抽取、加工、组合等操作分解为若干相对简单的映射函数,通过将映射函数进行组合,完成任意复杂的映射操作^[8]。此外柴小路等人提出了以关系模式为中间标准来实现的方法,但这种方法的不足之处是仅支持同一路径中的实体的联系,对支持转换规则的能力还不足。

表 2 源 XML 文档和目标 XML 文档之间转换研究现状

方法	实例
XSLT 技术	微软的 BizTalk Server ^[15] IBM 的交换原型 Clio ^[1,2]
基于映射函数	王瑜等 ^[16]
以关系模式为中间标准	柴小路 ^[17]

2.3 数据交换的主要解决方法

数据交换解决方案就其本质来说可以分成两种数据交换模式:

1)点对点交换模式。

点对点数据交换模式主要用于松散耦合的系统之间,这些系统一般是不同行业的系统,有数据交换的需求,但是数据交换并不是经常发生的事件,具有数据交换不频繁、交换数据量小的特点。

这种数据交换通常又叫做“部落式的交换”。因为在这种情况下,不同的行业有不同的行业要求和安全措施等,通常没有或者很难形成一个统一的数据交换标准,导致了相同的数据分析处理模块在很多应用中被重复地撰写,可能只是为了将某一数据源的数据转换到各个不同的目标数据源中去。由于没有实现标准,各个系统的实现人员也几乎没有可能将代码重用。代价是 $n * (n - 1)$,其中 n 代表应用系统数。

2)星状交换模式。

星状交换模式主要用于一些紧耦合的组织内部,或者某一行业的系统中。这些系统因为是同一组织或者行业内部的系统,比较容易形成一个统一的数据交换规范,尤其当 XML 成为一种数据格式描述的元语言标准以后,使用 XML 制定的应用领域的交换标准的出现,使得在各个应用领域中都形成了星状交换模式。其中每个系统将其内部的数据转换成行业标准的基于 XML 的数据格式用于系统间的交换,这样和点对点交换模式相比就带来了线形(n)的交换代价的好处。

从中可以得出:交换的精髓在于集中和标准,集中的星状交换模式带来了线形的交换代价,而交换标准的确立又使得集中的交换真正成为可能。然而现实问题是确定一个领域乃至是全球范围内统一的标准是非常困难的,因此虽然星状的交换模式有显然的线形代价的好处,但是在不能确定一个数据交换标准的情况下是不能使用这种模式的。

3 数据交换系统存在的不足

上面的研究成果提供了理论基础,各个公司和研究机构也开发出了很多数据交换系统。这些数据交换系统虽然功能强大,但是配置复杂,主要存在如下问题:

1)针对具体的应用制定。

只能满足用户在某一种或某一类应用上的交换,当用户想在其他方面交换的时候,必须为它专门做新的设计,通用性低。

2)配置复杂。

大多数的数据交换方案需要手工配置,并且配置复杂,容易出错,降低效率。

3)数据交换需求变化改动大。

同一应用的数据交换需求也是随着业务不断变化的,现有的数据交换机制对业务需求变化的适应性低,不够灵活。

理想的数据交换方案应该是在不降低企业安全措施的前提下,对参与交换的数据格式没有限制,能够实现数据交换类型的多样性,配置简单,手工配置少,是一个开放的、低成本的解决方案。

策略独立于具体的实现,定义了各种行为要做什么,而没有定义具体怎么做,这使得当业务需求改变时不需要重新编程,只需改变相应策略。通过定义策略,管理者能够采用面向应用的业务规则,将高层的规则自动转化为低层的行为,具有减少手工配置,降低配置错误,独立于实现,便于实现自动化等优点。在策略的

研究方面, IETF 提出了两个著名的关于策略的模型: 基于策略的网络管理 PBNM (Policy Based Network Management) 和策略核心信息模型 PCIM (Policy Core Information Model)^[18], 为特定领域对策略的使用提供了支持。目前对策略应用最多的领域是网络安全和网络管理领域, 其次策略在访问控制、QoS 中也有应用。这些领域通过应用策略解决了应用层出不穷, 业务不断变化, 配置复杂等问题。可以预见将策略应用于数据交换领域, 能够解决数据交换类型多样, 交换需求多变, 手工配置多等问题, 将成为解决数据交换问题的一个发展趋势。

4 总结和展望

数据交换问题的提出是因为存在着大量的信息孤岛, 同时这些信息孤岛之间交换与共享数据的需求不断提高。数据交换问题的解决能大大提高效率, 改善决策和投资环境, 增加效益, 对政府和企业的发展都有着重大意义。

目前, 对数据交换问题的研究依然非常活跃。业界许多著名企业的研究中心将数据交换作为主要研究内容, 如 IBM, 微软, Sybase 等。国内也有不少的机构在从事数据交换方面的研究, 如清华大学在基于 XML 的企业信息集成, 东南大学对基于 RDB 的 XML 模式映射等技术, 西北工业大学在 XML 与 RDB 之间的多层次双向集成, 中国科学院计算技术研究所 XML 到关系数据库的形式化映射方面的研究。

针对现有的数据交换产品的不足以及使用策略的优点, 预测将策略运用到数据交换领域, 实现一种策略驱动的数据交换机制将是数据交换研究的一个重要方向。此外对数据交换标准的研究也将是一个重要的方向。

参考文献:

- [1] Shu N C, Housel B C, Taylor R W, et al. EXPRESS: A Data Extraction, Processing, and Restructuring System[J]. ACM Transactions on Database Systems, 1977, 2(2): 134 - 137.
- [2] Lum V Y, Shu N C, Housel B C. A General Methodology for Data Conversion and Restructuring[J]. IBM J. Res and Develop, 1976, 20(5): 483 - 497.
- [3] Sibley E H, Taylor R W. A Data Definition and Mapping Language[J]. Communications of the ACM, 1973, 16(12): 750 - 759.
- [4] McHugh J, Abiteboul S, Goldman R, et al. Lore: A Database Management System for Semistructured Data[J]. SIGMOD

Record, 1997, 26(3): 54 - 66.

- [5] Deutsch A, Fernandez M F, Suciu D. Storing Semistructured Data with STORED[C]//Proc. the 1999 SIGMOD Conference. Philadelphia, USA: [s. n.], 1999: 431 - 442.
- [6] Shanmugasundaram J, Tufte K, Zhang C, et al. Lore: Relational database for querying XML documents: Limitations and opportunities[J]. The VLDB Journal, 1999, 26(3): 302 - 314.
- [7] Lee D, Chu W W. Constraints - preserving Transformation from XML Document Type Definition to Relational Schema [C]//Proc. 19th Int's Conf. on Conceptual Modeling, 2000. Salt Lake City: [s. n.], 2000.
- [8] Fagin R, Kolaitis P G, Miller R J, et al. Data Exchange: semantics and query answering[J]. ACM Theoretical Computer Science, 2005, 336: 89 - 124.
- [9] Fagin R, Kolaitis P G, Popa L. Data Exchange: Getting to the core[J]. ACM Transactions on Database Systems, 2005, 30(1): 174 - 210.
- [10] Mani M, Lee D, Muntz R R. Semantic Data Modeling using XML Schemas[C]//Proc. 20th Int'l Conf. on Conceptual Modeling, 2001. UK: Springer Verlag, 2001: 149 - 163.
- [11] Turau V. Making legacy data accessible for XML applications[J]. Informatik Spektrum, 1999, 22(2): 3 - 12.
- [12] Fong J, Pang F, Bloor C. Converting Relational Database into XML Document[C]//Proceedings of the 12th International Workshop. Washington: IEEE Computer Society, 2001.
- [13] Mani M, Lee D. XML to Relational Conversion using Theory of Regular Tree Grammars[C]//VLDB Workshop on Efficiency and Effectiveness of XML Tools and Techniques. [s. l.]: [s. n.], 2002: 134 - 137.
- [14] Bei Jia, Cai Fei, Tao Lie - Jun. A Direct Method of Data Exchange between XML and Relational Database[C]//26th Int. Conf. Information Technology Interfaces. [s. l.]: [s. n.], 2004: 127 - 132.
- [15] Sun Hongwei, Zhang Shusheng. Constraints - preserving Mapping Algorithm from XML - Schema to Relational Schema[J]. Lecture Note in Computer Science, 2002, 9(2): 193 - 207.
- [16] 王 瑜, 金 峰, 张 凯, 等. 基于多 XML 文档的政务数据交换平台构造及实现[J]. 计算机工程, 2004, 30(5): 52 - 54.
- [17] 柴小路. XML 数据环境下基于关系模式的数据交换方法 [EB/OL]. 2001 - 06 - 01. <http://www.gis8.com/zhaishow.asp?id=342>.
- [18] Ylitalo K. Policy core Information model[EB/OL]. 2000 - 11 - 15. <http://www.cs.helsinki.fi/u/kraatika/Courses/QoS00a/ylitalo.pdf>.