网络环境下异构数字印刷数据集成系统架构设计①

赵鹏飞[1]，王晓红[2]

[1](上海理工大学 光电信息与计算机工程学院，上海 200093)
[2](上海理工大学 出版印刷与艺术设计学院，上海 200093)

摘 要：为了解决网络环境下异构数字印刷数据的访问和集成问题，在对数字印刷数据特点进行分析的基础上，提出了异构数字印刷数据集成系统的整体框架。对框架结构所涉及的实现方法进行了分析研究，重点讨论了异构印刷数据的访问和提取的实现过程。最后详细研究了系统架构中的虚拟数据中心和查询引擎关键技术，为实现数字印刷企业各连锁店之间或企业各部门之间数据的透明化访问提供了有效的解决方案。

关键词：异构数字印刷数据；虚拟数据中心；查询引擎；缓存机制

Frame Design of Heterogeneous Digital-Printing Data Integration System in the Network Environment

ZHAO Peng-Fei[1], WANG Xiao-Hong[2]

[1](School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)
[2](College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstrac: In order to resolve the problem of the heterogeneous integration of digital-printing data in network environment, on the basis of analysis digital-printing data characteristics, an overall framework of heterogeneous digital-printing data integration system is put forward in this paper. After the methods of framework development being studied, the implementation process of heterogeneous data access and extraction are mainly discussed. Finally, the virtual data center and query engine of integration system is studied, and an effective solution is provided in order to realize transparent access to data between the chains or the business sectors in digital-printing company.

Keywords: heterogeneous data of digital-printing; virtual data center; query engine; cache mechanism

随着计算机技术的发展，网络的出现彻底改变了人们的生活，同时促进了各行业的发展，而这种改变和促进更多的是体现在服务上。对于印刷行业来说，将 IT 技术和数码技术相结合，构建商务印刷平台，促进数字印刷企业从分散式到集中式经营管理的变革，可消除在个人个性化和商务类业务交接的瓶颈。因此就需要将数字印刷与网络进一步结合，实现印刷活件的网上提交、生产过程的数字网络化和活件或内容的数字化管理。由于在数字印刷的实现过程中涉及到在线接单、在线计算、个性化定制等数字印刷基础服务，同时包括数字印刷远程打样服务、可变数据服务等特殊服务功能。所以有大量数据存在于各种异构的状态下，而网上数据资源共享是实现商务印刷平台的核心，其中异构数字印刷数据集成又是实现印刷数据资源共享的关键。

因此，本文以构建上海数字印刷在线集成管理与服务平台项目为基础，重点研究异构数字印刷数据集成过程，提出了数据集成系统的架构体系。最终实现

① 基金项目:上海市科学技术委员会科研计划项目(09220502700)
收稿时间:2010-06-01;收到修改稿时间:2010-07-10

数字印刷企业各个连锁店之间或企业各部门之间的数据资源共享。提高数字印刷的生产效率，提升数字印刷企业的核心竞争力。

1　系统架构设计

1.1 基本介绍

数字印刷数据涉及到数字印刷企业信息数据、印刷数字资产管理数据、生产流程控制数据、统计信息数据以及用户自定义数据等。为了既能实现各地数据的访问、传输和存储，又能确保系统的松散耦合性、良好的扩展性。将整个集成系统的架构设计为基于 Web Service 技术的分布式 Web 应用系统[1,2]，系统架构分为应用层、虚拟视图层、数据适配层，如图 1 所示。

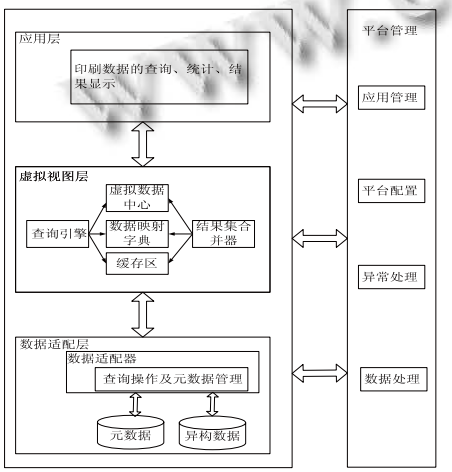

图 1　异构数字印刷数据集成系统体系结构

该体系结构基于虚拟视图法，使得印刷数据仍然保留在各个数据库中，集成系统只是提供一个虚拟的集成视图和对该集成视图查询的一种处理机制。使得底层数据库具有高度自治性并能保证查询结果是实时的，以满足数字印刷实时性的需要。

系统工作原理：用户根据实际需要输入 SQL 查询某个虚拟视图，查询引擎在缓冲区中查询，如果存在缓存则直接返回查询结果；否则根据虚拟视图和映射字典将该 SQL 分解成对应各个底层数据库的子查询，最后传输到数据适配器。数据适配器又将子查询分发到各个对应数据库并负责把查询结果返回结果集合并器。结果集合并器根据实际条件合并结果集，最后返回到应用层。

2　系统架构实现

异构数字印刷数据集成系统所实现的业务功能主要包括为异构印刷数据的访问、提取、存储 3 个要点。

2.1 异构印刷数据的访问

针对数字印刷数据分散、结构异构的特性，本系统利用索引图及预定义词表定义元数据，构建多样化可配置元数据库[3]，以实现数据/元数据的一体化管理、分布式数据存储、远程异构数据直接访问，以及自动启动指定的应用程序来显示和处理印刷数据等。

针对本系统中分布异构的各个印刷数据库，各个数据库用户通过元数据注册服务，利用索引图及预定义词表，提交自己本地所用的印刷数据库元数据信息(如：服务器名，IP 地址，用户名，密码)，实现各个数据库结构之间的映射，如图 2 所示。

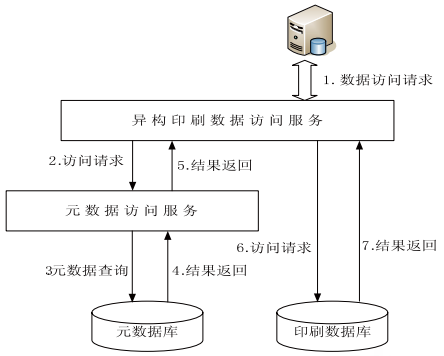

图 2　异构数字印刷数据访问流程图

2.2 异构数字印刷数据的提取

各个印刷数据库的数据信息量巨大，如果全部数据采用采用集中式管理，不但传输量非常大，增加网络负荷，而且效率很低。通过仔细分析现有数字印刷数据，可以从数据地位置角度将数据分为本地数据和异地数据两大类[4]。

本地数据的特点是数据量小、访问频度高；而异地数据的数据量较大，访问频度相对较低。因此，异构数字印刷数据集成系统的数据提取方式设计为：每次系统初始化时，首次提取全部的现状印刷数据，以后每次只提取变化的数据，通过系统提供的数据集成服务，与各个印刷数据库的数据同步。对本地数据，系统不做集中处理，各个印刷数据库保留本地的印刷数据。通过系统的本地数据分布式查询服务，提供各个数据用户所需的数据。

系统所面对的是一个分布式的异构的印刷数据体

系，在印刷企业一级的印刷数据的存放方式有着极大的多样性。所以为了解决从这些具有多样性印刷数据存放方式中提取数据的问题，本系统设计采用在印刷企业部署客户端程序的方式来解决此问题。

2.3 异构数字印刷数据的存储

本地异构印刷数据具有数据量小、访问频度高的特点。采用在客户端保留本地数据的同时，通过本系统定时集中到平台管理者的数据库中，实现异地数据库的集中式管理[5]。这样既可以满足本地印刷数据的使用需求，又不会对平台管理者的服务器存储能力、运行性能以及网络传输负荷造成压力。

这种数据的既集中又分布的体系结构不但加快了系统的运行速度，而且极大缩短了数据查询、提取所需的时间，提高了系统的效率及安全性。

3 系统架构关键技术

以上提出的异构印刷数据集成系统架构在实现过程中需要解决四个方面的关键技术，即是虚拟数据中心设计、查询引擎模块设计、缓存设计。

3.1 虚拟数据中心设计

虚拟数据中心由虚拟数据视图构建并具有数据采集和数据管理的虚拟中心，它并不实际的保存大量数据在“中心”，只是以“虚拟数据中心”的形式为数字印刷过程提供数据采集的代理功能[6]。虚拟数据中心的处理流程如图 3 所示。

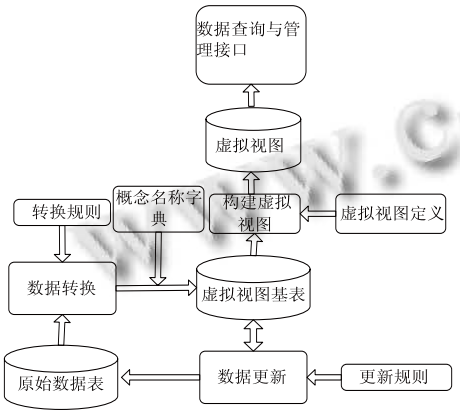

图 3 虚拟数据中心数据处理流程图

3.1.1 虚拟数据中心主要部件及其功能介绍

a) 虚拟视图基表：由原始数据表通过数据净化处理得到，一张原始数据表对应一张虚拟视图基表，把原始数据表中的属性名等转换成系统可以识别的统一名称。

b) 虚拟视图：由虚拟视图基表构成，其生成规则使用 DATALOG 形式描述。如下所示：

stu_info(stuname,departname):student(stuno,stuname,deptno),(dept,deptname,deptman),deptno=dept;

stu_info(stuname,departname) 基于 student (stuno,stuname,dep tno)和(dept,deptname，deptman)两个虚拟视图基表并通过 deptno＝dept 关系构成。

虚拟视图以 DATALOG 的形式以字符串的物理形式存储在系统中，使用时取出解析生成的数据库视图供查询使用。

c) 概念名称字典：在本系统应用中，概念名称词典在物理形式上为一张关系表，关系模式为 concept_dic(dialogna-me, uniname), 其中两个属性 dialogname 和 uniname 分别对应原始数据表中的属性名和映射到虚拟视图基表的属性名称。

d) 转换规则:在对原始数据表中属性值进行转换，形式上设计为一个二元组(attrname，funcname)的集合，attrname 为待转换属性名称，funcname 为转换的功能函数，例如常用的 trim、midtrim、format 等。在物理上表示为一张关系表，模式为 filter_rule(rawtablenme, attrname, funcname)其中 rawtablename 为原始数据表名。

e) 更新规则:形式上设计为一个四元组(soutable, destable,soufield,desfield)的集合。其中 soutable 和 destable 是原始数据表和虚拟视图基表。Soufield 和 desfield 是原始数据表和虚拟视图基表中的对应属性对。

3.1.2 虚拟数据中心处理流程

原始数据表中的数据与虚拟视图基表中的数据在属性值、规格类型和字段名称方面存在差异性，因此必须首先根据转换规则将属性值规格化再进入虚拟视图基表，并且根据概念名称字典将属性名称统一，最后形成虚拟视图基表。

另外在原始数据表和虚拟视图基表的属性关联方面也存在差异性，有针对的建立了一套更新规则，以据此实时更新数据。这种虚拟视图表结构关系定义了虚拟视图的构建过程，在虚拟视图基表建立好以后就可据此完善虚拟数据视图。

3.2 查询引擎模块设计

查询引擎模块设计为三部分组成：

a) 查询解析。对用户提出的查询请求进行简单的

语法分析，对不合理的查询进行修改或者返回给用户，并把初步处理的查询进行重写[7]。

b) 查询重写。给定一个数据库 D 及其在数据库 D 上定义的试图集合 V{V1,V2,…,Vn}。对数据库 D 的查询 q，如果查询 q 至少查询了视图集合中的一个视图，而又存在另一个查询 q1，并且 q1 的查询结果与 q 在数据库中的查询结果一致，则 q1 是 q 的查询重写。如果 q1 只对视图集合 V 的视图进行查询，则称上述查询重写 q1 为 q 的完全重写。

c) 查询分解。将重写后的查询分解为对各个数据源的查询请求。

3.3 系统缓存设计

在异构印刷数据集成系统架构中，元数据的提取、SQL 查询语句的解析和结果集合并都是及其耗时的操作。因此设计一个合理的缓存机制将极大地提高查询效率[8]。在本系统中设计两块缓存区。一个是元数据，在印刷数据在第一次添加进来时，数据适配器将其元数据提取出来随同数据对象一起写入 XML 文件。以后再次启动系统时首先搜寻缓存区，不必要重新提取元数据。缓存默认为系统定时更新。另一个是查询解析和结果合并集，SQL 语句经过解析后，其分解出的子查询和连接条件将被缓存到 XML 文件中，同时最后的结果集也会被缓存。下次再分解同样的 SQL 语句时，将先查询缓存。用户也可以分别设置 SQL 解析与结果集缓存的有效期。如果结果集的缓存期超过有效期，则使用 SQL 解析缓存；如果 SQL 解析缓存也超过了有效期，则重新分解，否则只需重新执行查询操作即可。

4 结束语

本文从实际出发，目的旨在解决数字印刷在线集成管理与服务平台中的异构数据访问和集成问题，设计了一个基于 Web Service 技术的分布式 Web 应用系统。重点介绍了异构数字印刷数据集成系统的整体框架模型，采用该模型可以实现数据源的“即插即用”，允许数据源的动态集成。对实现框架结构所涉及的实现方法和关键技术进行了系统研究。另外本文创新地设计了一个基于 XML 的缓存机制，提高了系统响应效率，有效实现了查询优化。但是由于印刷数据来源的多样性和不确定性，以及印刷数据的特殊性，对数据的更新是通过在视图中查询其提交时间来实现的，不允许在提取的过程中对其进行写操作。所以在系统中存在着数据提取速度较慢的问题，还有待于进一步研究。

参考文献

1 刘威,杨丹.基于虚拟视图的异构数据库集成平台的研究.计算机技术与发展,2009,19(6):91－94.

2 辜寄蓉,陈先伟,曾铭.异构国土地籍数据库网上汇交系统架构设计.计算机工程,2008,34(12):262－267.

3 Jurgen G, Thomas J, Boris. A generic visualization and editing facility for heterogeneous metadata. Computer Science Research and Development, 2009,3(24):119－135.

4 辜寄蓉,窦智,陈先伟.基于 XML 的异构地籍数据映射技术及其实现.计算机与数字工程,2009,37(2):72－75.

5 Baghaeri R, Nasiri R, Peyravi MH. Toward an elastic service based framework for enterprise application integration. Proc. of the 5th International Conference on Software Engeering Research. Management and Applications, 2007. 711－719.

6 王操,许云才,张晓初,顾晓忠.基于 XML 和虚拟数据中心的网上城市数据集成.计算机工程,2003,29(21):61－63.

7 袁景凌,徐丽丽,苗连超.基于 XML 的虚拟法异构数据集成方法研究.计算机应用研究,2009,26(1):172－174.

8 赵洁,张鹏,齐德昱.多数据库中间件中分布异构数据缓冲区系统的实现.计算机应用研究,2008,25(1):215－219.