报表管理系统中策略器的设计与实现①

韩宗营[1,2], 王 宁[2], 臧 辉[3]

[1](中国科学院大学, 北京 100049)
[2](中国科学院沈阳计算技术研究所, 沈阳 110168)
[3](鞍山市自来水总公司, 鞍山 114002)

摘 要: 在“辽河流域水环境管理技术综合示范”项目中, 业务系统中存在 Oracle、SqlServer、MySql 等多种异构数据库, 一个通用的面向服务的报表系统, 必须能从这些异构的数据库中提取数据、处理数据进而生成报表. 因此, 结合报表管理系统的实际需求, 通过在报表管理系统实现一个策略器完成异构数据库的数据提取、数据处理. 本文首先介绍了报表管理系统的结构和基本原理, 然后详细介绍了策略器总体结构、基本功能和具体实现方式.

关键词: 异构数据库; 数据提取; 策略器

Design and Implementation of Strategy in Report Management System

HAN Zong-Ying[1,2], WANG Ning[2], ZANG Hui[3]

[1](Chinese Academy of Sciences University, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology of Chinese Academy of Sciences, Shenyang 110168, China)
[3](Anshan City Water Corporation, Anshan 114002, China)

Abstract: In the project of “Comprehensive Demonstration of Water Environment Management about LiaoHe River Basin”, there are many heterogeneous databases in business systems, like Oracle, SqlServer, MySql, and so on. A generic service oriented reporting system, must be able to extract data from the heterogeneous database, processing data and generate reports. Therefore, combined with the actual needs of report management system, we build a strategy to complete data extraction, data processing in report management system. This paper firstly introduces the basic principle of the structure of the report management system, and then introduces the strategy overall structure, basic function and Realization way.

Key words: heterogeneous database; data extraction; strategy

在辽河流域水环境管理中一般存在着大量的报表服务, 像水质统计表、在线数据表、手工监测统计表、环境统计表等, 这些报表与水环境管理的核心业务密切相关, 管理和业务快速变化给的报表系统提出了一个很高的要求: 灵活、高效、可配置. 然而现存的报表系统把报表服务“硬编码”在程序代码中, 因而无法有效地管理报表服务, 当报表需求发生变化时, 很难灵活做出变化. 修改报表的样式必须大动干戈修改程序源代码, 带来了极大的风险, 频繁的让技术人员修改系统, 也难以保证信息安全性[1].

针对以上问题, 试想能否像提取数据信息一样,把报表需求从具体的业务系统的程序代码中剥离出来. 报表管理系统的出现使这一问题的解决成为可能, 报表管理系统提供统一的报表需求. 分离了报表与业务系统之间的耦合关系. 当用户报表需求发生变化时, 只需要修改报表注册服务的信息即可完成. 本文介绍了报表管理系统中策略器的设计与与实现, 主要是解决报表管理系统从异构数据库中提取数据、处理数据的问题. 业务系统中, 存在各种各样的数据库, 因此报表管理系统是否支持从这些数据库中提取数据决定了报表管理系统的适应性. 在一个面向服务的报表管理系统中, 提供对 Oracle、SqlServer、MySql 等主流数

① 基金项目:国家水体污染控制与治理科技重大专项(2012ZX07505004)
收稿时间:2013-07-24;收到修改稿时间:2013-08-26

据库的支持是必须的.

1 报表管理系统

1.1 报表管理系统的基本原理

报表管理系统它是根据报表格式定义, 报表内容的各种算法生成报表的一种思想. 报表管理系统从业务系统数据库获取数据, 在原始数据基础上, 定义报表的格式, 报表的算法, 并且根据自定义的算法完成数据的计算, 报表管理系统输出的数据信息, 经报表解释接口实现它的解释, 并生成相应的报表展示给用户. 用户也可以根据实际需求, 随时调整报表主题及算法的定义语言, 再重新运行报表时, 报表管理系统立即根据定义后的内容进行处理, 产生经过改变后的报表数据, 最后, 根据定义的用户定义的报表格式模板显示报表的内容. 报表管理系统对外提供一个通用的报表服务, 所有的业务系统可以通过 web 注册报表服务, 这样业务系统可以不必考虑本系统中的报表开发详细过程.

1.2 报表管理系统的结构

报表管理系统从整体上划分为: 请求器、格式定义器、报表定义器、运行器、策略器、输出器. 每一个部分都是一个独立运行的模块, 他们之间相互协作, 共同完成报表请求参数封装、报表格式设计、报表算法及参数等定义、报表管理系统运行、报表数据处理、报表最终显示功能[2].

一次报表服务的过程如下:

请求器收到业务系统的报表服务请求, 进行参数封装; 运行器接受请求器的参数信息, 从报表定义器中获取报表数据源等定义信息, 通知策略器提取报表数据; 策略器将抽取的数据按照报表定义的数据处理算法经计算处理后生成填充报表的最终数据, 并将数据交给运行器, 再由运行器转交给输出器, 输出器结合该报表定义好的模板, 将数据填充到报表模板, 最后把以 pdf、html、Excle 等格式将报表输出.

2 策略器的设计与实现

2.1 策略器总体设计

策略器主要和运行器进行交互, 策略器总体结构图如图 2 所示.

策略器主要功能:

① sql 语句翻译模块: 根据报表数据配置信息,

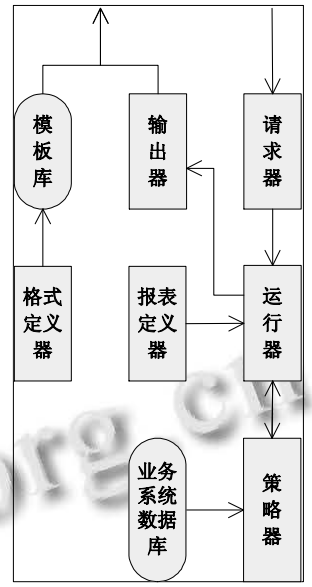

图 1 报表管理系统整体结构图

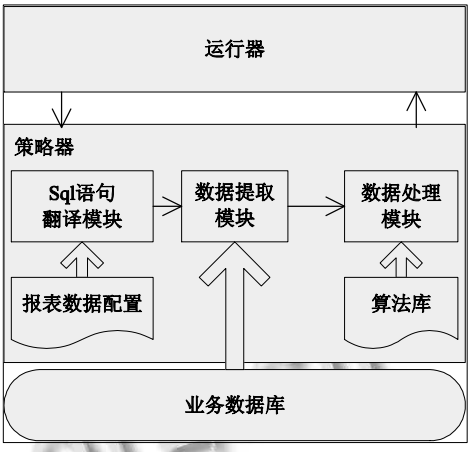

图 2 策略器总体结构图

翻译 sql 语句,生成查询策略.

② 数据提取模块: 完成业务系统数据库的连接和数据提取.

③ 数据处理模块: 根据用户选择的数据过滤策略过滤数据, 然后进行数据计算和统计等相关数据处理, 生成填充报表的数据.

2.2 各模块的详细设计与实现

2.2.1 sql 语句翻译模块

不同数据库所支持的 sql 语句基本相同, 但还是有差别的. 例如对分页查询, SqlServer 使用 top 关键字进行分页, Oracle 使用 rownum 关键字进行分页, 而 MySql 使用 limit 关键字进行分页, 并且他们的分页查询 sql 语句结构完全不一样. 其他数据库在分页查询上也不同.

因此, 在将参数翻译成标准 sql 后需要进行适配.

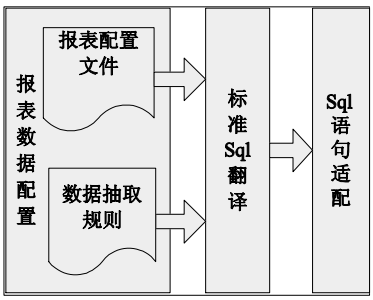

图 3　sql 语句翻译模块结构图

标准 sql 翻译模块根据报表服务的 ID 读取报表配置文件, 根据配置文件完成标准 sql 语句中列的选择和多表连接部分.

报表数据抽取模板结构如表 1.

表 1　报表数据抽取模板结构

ServiceID			
数据表1	列名	类型	
数据表1	列名	类型	
数据表2	列名	类型	
数据表2	列名	类型	
表连接	连接类型	数据表1列名	数据表2列名

采用简单规则策略进行数据抽取, 根据用户选择的数据抽取规则翻译 sql 中 where 子句的查询约束条件和数据分页条件[3].

用户自定义的抽取规则如表 2.

表 2　用户自定义的抽取规则

规则ID	规则类别		
规则1	列约束	列名	约束条件
规则2	列约束	列名	约束条件
规则3	分组约束	列名	groupby
规则4	Having	列名	having
规则5	排序约束	列名	desc
规则6	分页约束	初始行号	结束行号

报表配置文件和报表的抽取规则以 xml 文件的形式存储. 系统中使用 sax 解析 xml 文件, 他不需要像 dom 解析那样在内存中建立一个 dom 对象, sax 解析是基于事件驱动的, 占用内存小, 速度快效率高.

2.2.2 数据提取模块

该模块的主要功能是与业务系统数据库的连接, 抽取数据.

连接池管理: 获取与业务系统数据库的 jdbc 连接. 一个面向服务的报表系统, 如果每一个报表服务的到来都重新执行申请连接提取数据关闭连接的过程是非常消耗时间的, 因此我们采用连接复用的策略, 引入连接池. 连接池基本的思想是在系统初始化的时候, 将数据库连接作为对象存储在内存中, 当用户需要访问数据库时, 并非建立一个新的连接, 而是从连接池中取出一个已建立的空闲连接对象. 使用完毕后, 用户也并非将连接关闭, 而是将连接放回连接池中, 以供下一个请求访问使用. 而连接的建立、断开都由连接池自身来管理. 同时, 还可以通过设置连接池的参数来控制连接池中的初始连接数、连接的上下限数以及每个连接的最大使用次数、最大空闲时间等等.

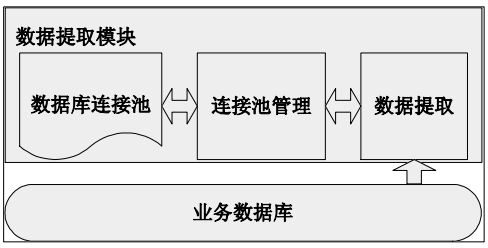

图 4　数据提取模块结构图

数据提取: 根据 sql 语句翻译模块生成的 sql 语句和从连接池中获取的连接, 执行 sql 语句, 从业务系统数据库中提取相关数据, 数据存放在 resultset 结果集中.

2.2.3 数据处理模块

该模块的主要功能是把 resultset 结果集转换为可读写的数据对象, 依照用户定制的数据过滤策略完成数据对象的过滤, 然后对数据依照用户定制的算法进行处理, 最终生成填充报表的最终数据.

该模块的结构图如图 5 所示.

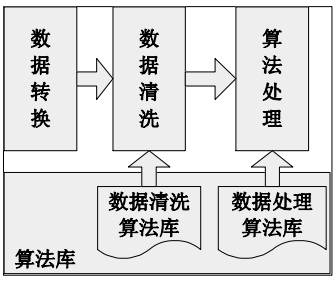

图 5　数据处理模块结构图

数据转换: 在一个报表中, 许多业务数据是通过

表中的多个行列计算得来的, 然而对 resultset 来说直接进行数据的计算和组织是非常麻烦的, 现在系统中普遍使用 Java 类映射 resultset 表结构, 将 resultset 转化为 arriaylist, 然后再进行数据处理. 然而对于一个面向服务的报表系统, 面向的数据库有成千上万个, 并且由于报表的可配置性, 生成报表的数据信息也在变化, 因此我们无法利用 Java 类映射数据库结构. 为了方便数据处理, 以及尽早的释放数据库连接, 我们把 resultset 进行分解, 一条记录是 meta 对象的 arraylist, 一个 resultset 是 record 对象的 arraylist.

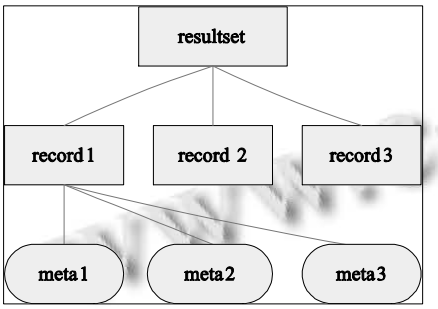

图 6 resultset 的转换结构图

数据清洗: 为了能生成准确的报表, 有效的反应实际情况, 我们采用数据清洗策略提高生成报表的数据质量. 数据清洗原理是利用现有的技术和方法清洗脏数据, 将原有不合理要求的数据转化为满足数据质量要求的数据[4]. 在报表管理系统中, 所使用的数据清洗策略有以下几种:

① 空值清洗方法: 指数值实际存在, 但是数据不完整. 针对空值问题, 处理策略有两种: 可以用最大最小值、中间值、平均值或更复杂的概率统计函数代替; 可以由用户手动输入一个值.

② 错误值检测及清洗: 错误值也成为噪声数据或孤立节点, 指一条记录中某些属性值异常. 对此类数据, 我们采用简单规则库检测、修正数据的错误. 规则类文件由使用报表的业务系统注册时根据用户的选择生成, 采用 xml 的形式存储.

③ 重复记录清洗: 在报表管理系统中, 重复数据是由用户建立规则指定两条记录中哪几个属性相同, 那么这两条数据即为重复数据. 我们采用 Sorted Neighborhood Method(排序邻居法)方法[5], 该算法的主要思想是: 将记录按指定用户指定的一个属性(该属性相同两条记录才有可能重复)进行排序, 在排序后的数据记录中维持一个包含 W 个记录的窗口, 只检测窗口内的记录是否重复, 不断的向前推进窗口, 直到所有数据都已过滤.

算法处理: 依据用户定义的算法进行数据的处理. 其中包括, 增加一行, 或增加一列, 统计分析等, 主要是完成数据的计算, 生成填充报表的最终数据, 例如工资表中工资属性是根据多个属性通过计算得到的, 饼状图的最终数据是对数据记录统计后得到的.

3 结语

本文从辽河流域水环境项目实际需求出发, 在报表管理系统中实现策略器, 完成查询语句的翻译、数据抽取和数据处理, 最终生成填充报表的数据.

系统采用 java 语言, myeclipse 平台开发.

参考文献

1 何仁杰,梁冰.用规则引擎替换代码.计算机世界,2004,14 B4、B5.

2 裴晓华.报表引擎系统研究开发[学位论文].西安:西安理工大学,2009.

3 刘如九,张振山,柴天佑.一种通用的多数据库间数据抽取方法及应用.北京交通大学学报,2008,32(4):14–18.

4 周芝芬.基于数据仓库的数据清洗方法研究[学位论文].上海:东华大学, 2004.

5 Hernández MA, Stolfo SJ. Real world data is dirty: Data cleansing and the merge/ purge problem. Data Mining and Knowledge Discovery, 1998,(2): 127–138.