利用 UML 开发基于 J2EE 的跨源搜索引擎研究

唐海军，黄春毅

（四川大学 公共管理学院，四川 成都 610064）

摘 要：本文针对目前难以有效整合多源异构数据检索的问题提出了一种跨源搜索引擎（CSE）设计思想，并基于 J2EE 和 UML 技术对 CSE 进行建模，文中对 CSE 的设计思想、系统结构、设计模式和系统执行进行了理论上的描述。

关键词：跨源搜索引擎；CSE；UML；J2EE；建模

中图分类号：TP311 **文献标识码**：A **文章编号**：1007－7634(2007)04－0579－04

Study on the CSE Based on J2EE and UML

TANG Hai－jun，*HUANG Chun－yi*

（*School of Public Administration*，*Sichuan University*，*Chengdu* 610064，*China*）

Abstract: This paper proposes a kind of CSE design thought which aims at resolving theproblem of how to effectively conform multi－source isomeruous data searching, and constructs models based on J2EE and UML technology. Theoretical description is given to CSE' sdesign thought, systematic structure, and design mode as well as system administration.

Key words: cross search engine; CSE; UML; J2EE; modeling

1 引 言

数据源共享技术研究的目标之一是支持通过网络对多个异构数据源的查询[1-3]。网络异构数据源既包括规则结构的数据，又包括半结构化的数据，甚至还有非结构化的数据，这些数据源不仅数据模型不同，而且查询能力各异。如何在用户层屏蔽这些差异性，给用户提供一个满足需求的资源整合解决方案[4]，就成了跨源搜索的难题之一。

本文提出了一种跨源搜索引擎（Cross Search Engine，CSE）设计思想，并基于 J2EE 面向对象技术，利用 UML 对 CSE 进行建模，同时对 CSE 系统的逻辑执行进行了描述。

2 CSE 的设计思想

相对于目前大多数单一数据源搜索而言，用户更需要一个能从统一搜索点搜索到多源数据库的搜索引擎。CSE 的设计采用基于 J2EE 和 XML 技术，向用户提供统一的检索接口，将用户的检索要求通过自身设计的独立索引组件用统一的检索表达式并发地检索本地和广域网上的多个分布式异构数据源，并对检索结果加以整合，在经去重和排序等操作后，以统一的格式将结果呈现给用户。

CSE 与目前一些汇总式搜索引擎不同[5]，它有

收稿日期：2006－09－28

作者简介：唐海军（1981－），男，四川大英县人，在读硕士研究生，从事信息系统分析与设计研究；黄春毅（1957－），女，湖南平江县人，主任，硕士生导师，从事信息检索技术与系统、信息系统分析与设计、电子政务、竞争情报等研究.

自己的独立索引组件。这意味着 CSE 在处理过程中不需要触发其他的搜索引擎，避免了被不同搜索引擎强加在查询语句上的语法构造规则问题，具有很好的跨平台特性。

CSE 具有以下优点：①提供单一检索入口，检索多源异构数据；②平台透明，可移植性和扩展性强；③结果整合，最终结果按模板输出，并且按统一标准排序，方便用户的浏览和操作；④并发检索，节省检索时间。

3 CSE 的系统结构

3.1 功能模块

(1) URL 服务器。将要抓取的 URL 清单发送给网络蜘蛛，并将抓取到的 web 页面送到贮存服务器。

(2) 贮存服务器。将抓取到的 web 页压缩并且储存到数据仓库中，每一个 web 页都有一个相关联的 ID。

(3) XML 解析器。对 XML 格式数据在其被加入索引表之前进行处理。

(4) 索引器。读取数据仓库中的记录并解析。然后将记录按词频分别索引进不同的数据库。

(5) URL 核对器。读取锚点文件并且转换关联 URL，同时使被选中的 URL 具有有效的占位符。

(6) 检索器。通过使用索引器生成的词典和反相索引数据回应检索需求。当用户提交查询时，检索器通过查询数据字典得到索引器生成的数据，并检索出一系列记录。

(7) HTML 生成器。用标准 XSLT 处理机利用 JSP 将 XML 格式数据转换成 HTML 格式输出。

3.2 J2EE 层次

本系统采用 struts[6] + JavaBean 封装 + Hibernate 的整体开发框架，即 Web 层用 Struts 框架，检索逻辑层采用 JavaBean 封装的类，数据持久层采用 Hibernate 框架。如图 1 所示。其中，BO 表示检索对象，PO 表示数据持久对象，DTO 表示数据传输对象。

Struts 主要提供了视图、控制器的解决方法，并没有关注模型的实现。检索逻辑层是由一些检索逻辑类组成，与检索逻辑相关的操作都封装在检索逻辑类里，每一个检索逻辑类都是一个 JavaBean 类；数据持久层利用数据库连接池，Hibernate 使用

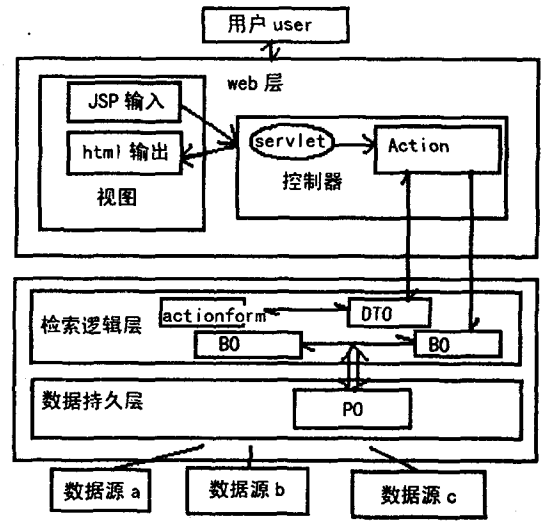

图 1 J2EE 层次图

大量的配置来代替硬编码，同时对值、对象和不同数据源数据进行映射；在持久层，系统编写了一个数据访问类，把与数据库有关的所有操作都封装起来，检索逻辑层只需要调用数据访问类的相关方法就可以实现所有的数据库操作。

4 CSE 的设计模式

4.1 工厂模式

由于需求数据来自于不同的数据源，CSE 的跨源性索引组件类在实现时就需要描述不同数据源的特征。工厂模式消除了 JAVA 类中把特殊应用类嵌入代码的需要，从而提高了松散耦合，其功能静态产生，不需要调用 new 方法去构造对象。

工厂模式能很好地封装基于 JSP 的搜索组件，它有三个属性。首先，处理器工厂提供一系列不同数据库的入口，这对于跨源数据库数据检索非常重要；其次，在给定区域中，它的功能是与数据库类独立的。我们可以改变数据库类的执行而不影响处理器工厂；第三，其他需要搜索组件的应用软件可以重用处理器代码。通过简单地改变基于 XML 的属性文件的配置，就可以实现跨源数据库选择，很好地解决了资源元数据获取技术难题[5]。

4.2 策略模式

策略模式由分解主机的运算法则和将运算法则封装为一个独立类两部分组成，对象具有不同的行为，那么将这个对象的每个行为封装到不同的类而

不是嵌入到对象的方法体中，跟踪行为就会变得简单得多。

策略模式允许我们任何时候都可以转换算法。例如 JSP 策略对象封装基于 JSP 的检索组件，从而来自于 JSP 策略的新型处理能被执行。每一个筛选对象在一个 JSP 查询过程中都提供一个标签。查询处理器和数据库管理器对象分别传递查询请求，并且连接不同组织结构的数据源。

5　系统设计

CSE 系统的 UML 图包括用例图，包图和交互图。

5.1　用例图

图 2 对 CSE 的用例进行描述，其中有四个主要参与者：①网络蜘蛛—搜索普通 web 数据；②索引器—索引普通 web 数据和特定数据库数据；③搜索引擎—通过 JSP 传递查询请求，搜索相关文档，并显示查询结果；④用户 - 提出查询请求。

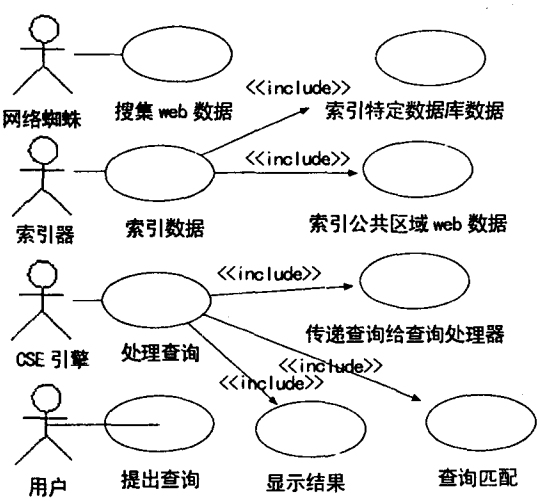

图 2　CSE 系统用例图

5.2　数据包图

图 3 描述了 CSE 的数据包功能组件，JSP 处理器是用户与系统的交互场所。索引器、数据库管理器和网络蜘蛛处于构建和维护数据库的保护层。检索处理器是与用户的检索请求和检索结果显示相关联的组件。

5.3　系统交互图

图 4 描述了 CSE 的交互模型，当 JSP 查询对象

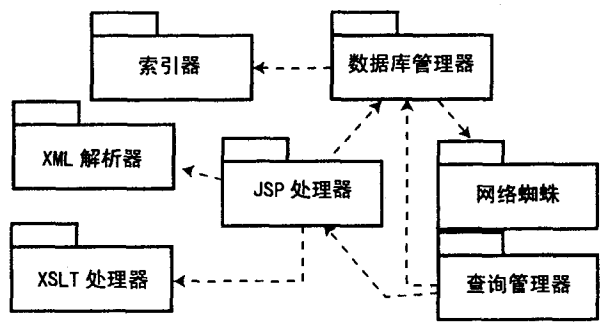

图 3　CSE 系统包图

被创建的时候，它可以向工厂对象申请其 JSP 策略。

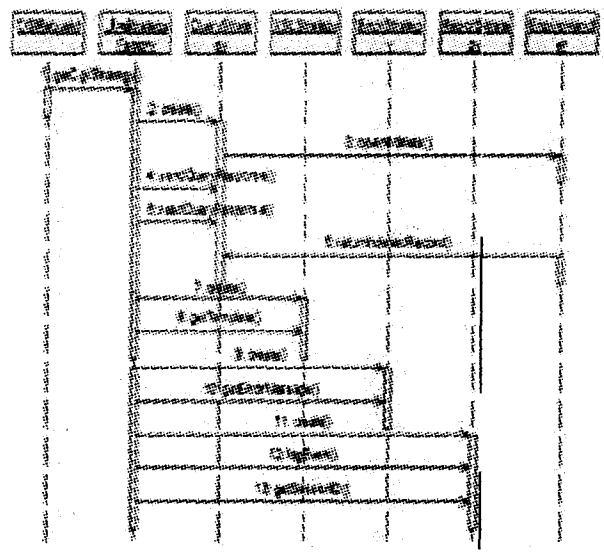

图 4　CSE 系统交互图

一个 JSP 策略工厂负责创建 JSP 程序需要的所有的策略，包括：①查询（第二步），②XSL（第七步），③错误（第九步），④报告（第十一步）等策略。

逻辑上，查询策略对象将用户查询通过固定的 daemon 传递进入引擎，该引擎由处理器工厂对象产生的引擎处理器管理。引擎处理器创建一个引擎配置对象，通过一个 XML 配置文件描述的一系列检索规则和选项来解析，并且处理每一个管理器。

6　系统执行描述

6.1　JSP 组件

浏览器向 Web 服务器发出请求，JSP 组件将请求封装为 XML 的格式。如果一个 JSP 程序在每次请求的时候都执行，并且在一个过载的服务器上每

秒执行很多次，那么就应该考虑如何最小化 JSP 的启动时间。因此，在连接的时候，筛选的清单应当尽可能静态初始化。系统通过创建一个二维数组 String [] {name, value} 来实现这种功能，其中 name 是一个筛选名，value 是一个可以被调用来创建一个筛选器的函数。

6.2 中转服务器

中转服务器针对某一特殊查询的选择将每个查询请求都用 XML 查询语法详细标定，再将其分解为对指定数据源的子查询，然后对 JDBC 进行操作。JDBC 将结果集返回中转服务器，中转服务器将结果集封装为 XML 格式返回 JSP，最后 JSP 将结果解析得到查询结果。中转服务器是查询管理器中最重要的组件，它和客户机之间所有的查询和回应都是 XML 文档。

6.3 XSL 组件

XSLT 将一个 XSL 模板文件和一个 XML 数据联合起来生成一个 HTML 输出文件。

JSP 维护一个包含 XSLT 模板文件的目录，JSP 能够产生的每个可能的反馈页面都有一个对应的模板文件，一旦查询被执行，CSE 就显示一系列包含被检索到的信息的 HTML 格式摘要页。

7 结 论

基于 J2EE 的跨源搜索引擎（CSE），可以解决当前难以有效检索多源异构数据的问题。CSE 系统利用 XML 和 J2EE 的跨平台特性，可以检索到网络上不同数据源不同格式的数据，而且不需要触发其他的搜索引擎。其功能特点表现为：①在选择查询期内查询多源数据库索引，②用 XML 和 XSL 动态生成交互界面，③编辑和执行布尔检索和模糊检索，④通过数据库类型筛选查询结果。基于 J2EE 的 CSE 的跨数据源检索能力可以大大减少用户在查找信息的时候转换数据库所带来的不便，提高了检索效率。因此，CSE 系统对第二代搜索引擎设计的研究提供了一定的参考价值。

参考文献

1 Knoblock C A, Minton S, Ambite J, et al. The Ariadne Approach to Web based Information Integration[J]. Journal on Intelligent Cooperative Information Systems (IJCIS), 2001, 10(1 - 2): 145 - 169.

2 Lieming Huang, Matthias Hemmje, Erich J Neuhold. ADM IRE: An adaptive data model for meta search engines[J]. Computer Networks, 2000, 33(1): 431 - 448.

3 李广建，张智雄. 国外跨库检索系统研究项目及其特点[J]. 情报理论与实践，2004，27(4)：444 - 447.

4 张 淼，智能检索及跨库检索技术在数据库建设中的应用研究[J]. 科技情报开发与经济，2005，15(12)：233 - 235.

5 陈 珉，喻丹丹，涂国庆. 分布式数据库系统中数据一致性维护方法研究[J]. 国防科技大学学报，2002，(3)：76 - 80.

6 徐贵水，王 彤，霍好田，等. 跨库检索系统 MDBSS 的设计与实现[J]. 计算机应用，2003，23(1)：121 - 123.

（责任编辑：刘凤勤）