分布异构信息资源集成系统实现研究

齐惠颖[1], 王 欣[2]

(1. 北京大学医学部 计算机教研室, 北京 100191, qhy@bjmu.edu.cn;
2. 哈尔滨工业大学 科学与工业技术研究院, 哈尔滨 150001)

摘 要: 为了实现分布、异构网络信息资源的统一访问,提出了一种信息资源集成系统实现方案.该方案针对异构资源采用 Web Service、SRU 和元搜索技术实现虚拟集成,通过开放链接 OpenURL 实现服务集成,使用基于情景敏感的知识库提供用户的个性化服务.实验结果证明提出的信息资源集成方案可以实现资源和服务的有效集成,使网络服务机构的用户能够一站式获取各种信息及服务,进一步提高了信息服务的质量.

关键词: 资源集成; 服务集成; 知识库
中图分类号: TP39 G250 **文献标志码:** A **文章编号:** 0367-6234(2010)11-1838-04

Research on the realization of distributed and heterogeneous information resources integrated system

QI Hui-ying[1], WANG Xin[2]

(1. Computer Division, Peking University Health Science Center, Beijing 100191, China, qhy@bjmu.edu.cn;
2. Institute of Science and Industrial Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: An integrated system scheme is proposed to realize uniform access of distributed and heterogeneous network resources. The scheme achieves virtual integration using Web Service、SRU and meta search technology, realizes service integration by OpenURL, and provides personalized service through constructing context-based knowledge base. The experimental result demonstrates that the scheme obtains effective integration of different kinds of resources and service, and one-stop access of the network resources and service realizes efficiently as well.

Key words: resources integration; service integration; knowledge base

随着数字资源建设的不断深入,我国各类信息服务机构购买了越来越多的信息资源,由于这些资源具有不同的信息检索界面,用户在利用信息资源时需要切换到不同的环境分别进行检索,这个过程客观上增加了信息利用的复杂性.由于不同类别的用户对各类期刊的使用权限不同,目前的各类检索系统无法感知用户环境信息,对不同用户提供个性化的服务.针对上述问题,本文提出了一种分布、异构资源的集成方案,本方案利用标准互操作协议和互操作技术将多种分布、异构的资源有机地集成到统一的环境中,同时,基于开放链接机制实现服务集成,将信息服务机构提供的各种扩展服务无缝地连接起来,通过基于情景敏感的知识库提供用户的个性化服务,使用户对各类信息资源和服务实现一站式获取.

1 信息资源集成系统的体系结构

信息资源集成系统是集成了多种信息资源,为终端用户提供了统一检索平台.集成检索结果,提供与上下文相关的全文在线获取、原文传递、参考咨询等服务.系统的体系结构如图 1 所示.

数据层由知识库和各类资源组成,知识库是整个系统框架的核心部分,提供一系列的规则,是为用户提供恰当服务的依据.各类资源包括订购

收稿日期: 2009-11-30.
基金项目: 国家科技图书文献中心资助项目.
作者简介: 齐惠颖(1973—),女,博士,讲师.

或免费的数据库资源、本地加工资源和网络资源等；功能层由前台检索界面和后台管理系统组成，前台检索界面是多种数据库资源的统一检索界面，后台系统主要负责知识库的初始化、更新以及多种维护功能；服务层为用户提供了一个集成的多样化的服务，包括多种信息服务（例如全文获取、参考咨询服务、馆际互借和 OPAC 等）以及多种 Internet 服务（例如搜索引擎等）. 由上述信息资源集成系统的体系结构可以看出，其实现的关键是如何实现异构资源集成和多种扩展服务集成.

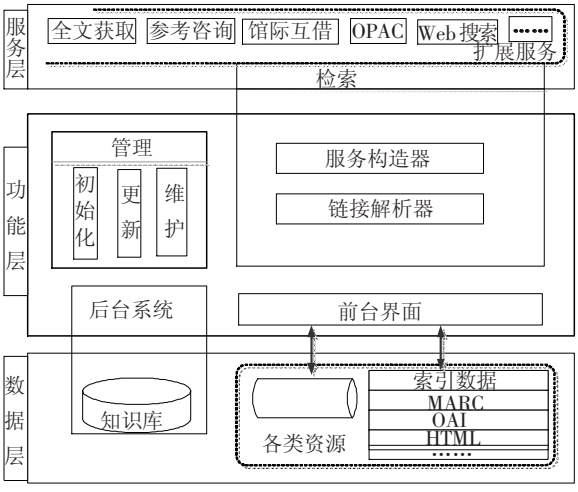

图 1　信息资源集成系统的体系结构

2　基于虚拟集成的资源集成

实现多种异构资源的集成涉及到异构系统之间的通信和互操作，由于大多数商业性学术资源的提供商不支持遵循 OAI 的数据收集，因此要集成这些分布异构资源就需要利用各种标准协议和技术实现虚拟集成，针对不同资源拥有者提供的不同接口采用相应的协议和技术实现检索结果的获取. 此外，对异构资源的检索结果的呈现，保证集成检索结果的条理性也是需要考虑的问题.

2.1　检索结果获取

针对异构资源最常见的 Web Services 接口和 SRU 接口，集成检索均支持对应的互操作. Web Services 是一个采用 XML，通过 URL 来发布接口和应用绑定的软件体系结构[1-2]. 这些定义接口可以被其他软件系统发现，并通过基于 XML 和 IP 协议的消息通讯机制集成到不同的应用系统中. Web Services 通过 SOAP 传输消息，系统将用户的查询请求进行 SOAP 封装，通过 HTTP 的 POST 方法发送到 Web 服务器，Web 服务器再把请求转发给相应的 Web Services 请求处理器，请求处理器对 SOAP 进行解析后将应答返回给 Web 服务器，Web 服务器再将应答返回给系统[3-4].

SRU（Search/Retrieve URI Service）是一种基于 XML 的网络搜索协议[5]，它定义了一个通用、抽象的模型，各个异构系统可以将其具体实现映射到该抽象模型上，从而实现不同网络资源、分布式数据库的统一检索功能. 对于支持 SRU 或一些系统的访问方式经过简单分析，转换成类似 SRU 格式的都采用基于 SRU 协议集成. SRU 的请求信息是通过 HTTP 的 GET 方法发送的，其信息检索请求利用 URL 格式编码[6]. 例如，对 Calis 系统构造一个 SRU 格式的检索请求式，题名检索词为 Nano 时查询式为

http://ccc.calis.edu.cn/result.php?op=simple&pos=0&at=Nano&from=1&date_from=2003&date_to=2008&per=50&max=10000&order=1&p=2.

其中，op 为搜索方式，at 为搜索的关键词，from 为检索域（值 0 ~ 4 分别表示不同的检索字段，如题名、关键词等），date_from 和 date_to 分别为查询时的间范围，per 为每页显示记录条数，max 为最大查询记录数，p = 2 为取出查询结果的第 2 页.

对于一些不提供标准接口的异构资源，虽然这些系统各自具有不同的结构，但都基于 HTTP，以 B/S 方式进行服务，对这类资源采用元搜索技术实现检索结果的获取[7]. 元搜索技术的本质是一种模拟技术，资源集成系统来模拟浏览器或者客户端，通过 HTTP 协议向数据提供方发送检索请求，由数据提供方返回 HTML 格式的检索结果页面，然后由资源集成系统对该页进行分析，生成汇总结果集传送给用户.

2.2　检索结果呈现

资源集成系统中检索的数据源物理上分布在不同的位置，需要将多个结果集进行合并、重新排序处理后形成最终的结果集呈现给用户. 在系统向多个数据源发送检索请求到返回给用户结果的过程中，所需时间的长短是系统响应速度的关键所在，为避免由于结果集占有服务器内存过大而引起宕机现象，在用户检索时，每次只取固定数目的检索结果集存储在服务器端，用户在这些结果集中翻页时不用二次查询，直接从缓存中提取结果提高了翻页速度，当翻页达到指定的数目时，系统将再次进行检索. 这样既避免了将检索结果全部放到服务器端，随着用户增多而导致检索结果集不断增大，造成内存不够的问题，同时也避免了每次翻页都要发送检索请求造成时间开销过大.

3 基于 OpenURL 的服务集成

服务集成的关键是为一个链接源提供与链接源上下文相关的多种扩展链接服务,而 OpenURL 是一个基于上下文的、开放的信息资源与查询服务之间的通信协议标准[8]. OpenURL 协议提升了在开放链接环境下,提供定位服务的服务组件和信息资源之间的互用性[9-10]. 基于开放链接的集成服务的基本工作流程如图 2 所示.

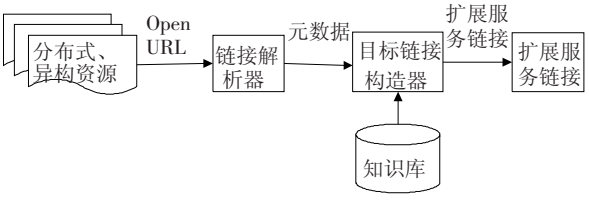

图 2 基于开放链接服务集成的工作流程

3.1 链接解析器

链接解析器的主要功能是从检索请求中获取元数据,根据知识库中的知识分析元数据,为用户选择相关资源和服务提供目标解析机制. 以 MetaPress 数据库为例来说明 OpenURL 的链接机制. 在 MetaPress 数据库中查找一篇文献得到的 URL 页面为

http://iospress. metapress. com/openurl. asp? genre = article&issn = 0169 - 2968&volume = 80&issue = 1&spage = 199.

可以看出这个地址是符合 OpenURL 语法的. 若获取全文,基地址为

http://iospress. metapress. com/openurl. asp?

解析规则为

genre = article&issn = @ get (issn) &volume = @ get (volume) &issue = @ get (issue) &spage = @ get (spage) .

对于没有采用 OpenURL 标准的数据库,需要对其链接机制进行分析,将文献页面的 URL 转换为符合 OpenURL 语法的 URL. 例如在 Mary Ann Liebert 数据库中查找到两篇文献的 URL 分别为:

1) http://www. liebertonline. com/action/doSearch? target = article&volume = 23&year = 2009&issue = 6&journalCode = apc&journal = apc&searchText = + Rapid + Development + of + Advanced + Liver + Fibrosis + after + Acquisition + of + Hepatitis + C + Infection + during + Primary + HIV + Infection&filter = issue&x = 28&y = 6.

2) http://www. liebertonline. com/action/doSearch? target = article&volume = 23&year = 2009&issue = 6&journalCode = apc&journal = apc&searchText = Autoimmune + Hepatitis + in + the + HIV - Infected + Patient% 3A + A + Therapeutic + Dilemma&filter = issue&x = 12&y = 11.

可以看到 URL 中 searchText 字段后面是文献的标题,journalCode 和 journal 字段表示文献所在期刊. 为实现 URL 地址的转换,首先将一篇文献的元数据替换到另一篇文献的 URL 仍能定位到该文献,说明此链接是有效的. 接着分别将元数据字段为空,验证是否还能定位到该文献,对上述 URL,去掉 journalCode 和 journal 字段也能唯一定位到文献. 因此,通过上述转换过程可以得知在数据库 Mary Ann Liebert 获取全文,基地址为

http://www. liebertonline. com/action/doSearch?

解析规则为

target = article&volume = @ get (volume) &year = @ get (year) &issue = @ get (issue) &journalCode = @ get (journalcode) &journal = &searchText = @ get (atitle) &filter = issue&x = @ get (x) &y = @ get (y) .

3.2 知识库

知识库反映的是信息服务机构的网络数据库订购情况和用户的权限信息,其内容包括目标链接、链接解析规则、潜在服务的推理规则、各种资源的映射规则、扩展服务规则等. 它是提供扩展服务链接时所用到的元数据的集合,为构造扩展服务链接提供一系列规则,根据访问资源或服务的用户标识(如 IP 地址),利用给定元数据为用户提供最多的、恰当的服务.

由于期刊数据库将收录的期刊信息在其网站上全部列出,而 Web 页面通过结构化的 HTML 标记来控制页面的布局和显示. 通过对国内外的网络数据库页面特征分析,分析 HTML 文档的结构归纳出每个提取域统一的定界符,找到其排列的规律,构造一个抽取器,提取需要的元数据内容,实现知识库的构建和更新.

3.3 目标链接构造器

目标链接构造器的主要功能是根据知识库中期刊数据库提供的信息、各种服务规则和用户的权限标识为用户构造恰当的服务链接,并提供所有的服务链接的页面. 实现机制是根据链接解析器析取出的参数和知识库里的访问规则判定传来 OpenURL 的链接源所在的机构是否订购了该期刊的这些卷期,如果订购了,判断此期刊中的基地址和指向规则是否为空. 若不为空则直接用此期刊中的规则来构造定位到文章的链接;若为空则寻找哪些数据库中收录了此期刊,再将这些数据库中可通过规则定位到文章的链接都构造出来,并提供给用户. 例如获得的参数为 issn = 0169 -

2968，volume = 80，issue = 1，spage = 199. 根据其中的 issn 和知识库中的存储信息可知期刊是 Processes of Petri Nets with Range Testing Processes of Petri Nets with Range Testing，然后判断链接源所在机构购买了这个期刊对应的卷期. 接着查看知识库中此期刊的基本信息和权限信息，得出获取全文服务的基地址为

http://iospress. metapress. com/openurl. asp?

指向规则为

genre = article&issn = 0169 - 2968&volume = 80&issue = 1&spage = 199.

最后根据获得的基地址和指向规则，构造了指向所找寻文章的全文目标链接为

http://iospress. metapress. com/openurl. asp? genre = article&issn = 0169 - 2968&volume = 80&issue = 1&spage = 199.

4　系统性能实验

集成系统解决的是多个源系统的整合，其关键技术是互操作和结果融合. 系统运行在第三方信息服务机构，当用户使用集成系统时，集成系统是以同步的方式并发检索多个源系统. 因此，多用户、多关键词并发检索时的性能是系统健壮性的重要体现. 实验是通过模拟第三方机构的最终用户并发操作的测试. 系统的测试环境为 CPU 为 PentimIV1. 7G GHz、内存为 512 M、操作系统为 WindowsXP SP2、网络带宽为 100 M.

实验 1　在 0 ~ 5 s 时间内生成 600 个用户，并发进行检索操作. 图 3 显示的是负载测试期间相对任何指定时间点用户的数目的平均事务响应时间. 结果显示系统响应检索操作的平均时间是 2. 142 s；再随机分配用户查看检索到的资源，平均每个用户点击 1 ~ 2 次，整个检索事务的响应时间为 10 ~ 20 s 之间，系统响应操作时间较理想.

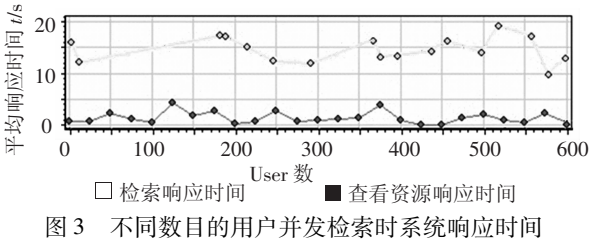

图 3　不同数目的用户并发检索时系统响应时间

实验 2　为用户随机分配检索词，模拟不同数量用户进行检索. 图 4 是相对任何指定数目的用户并发检索时平均事务响应的时间，检索事务的使用时间为 19 ~ 25 s. 可以看出，并发用户数量在 500 个以内时，系统整个事务的响应时间没有随用户增多而延迟.

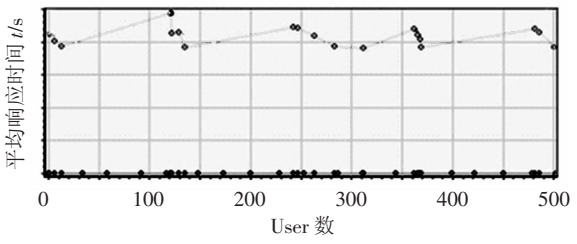

图 4　多用户并发检索时系统整个事务的响应时间

5　结　论

1）针对网络环境下不同信息资源呈现分布式、异构的特点，分别使用相应的标准互操作技术可以有效地实现资源的集成.

2）基于 OpenURL 的开放链接框架提供了一种开放的、可扩展的描述元数据的标准，它描述链接源的上下文和服务，实现了服务的有效集成.

3）通过知识库提供的一系列规则，根据访问资源或服务的用户标识为用户提供恰当的服务，进一步提高了信息服务的质量.

参考文献：

[1] CURBERA F, DUFTLER, KHALAF M R, *et al.* Unraveling the Web services Web: An introduction to SOAP, WSDL, and UDDI[J]. IEEE Internet Computing, 2002, 6(2): 86 - 93.

[2] YANG J. Web service componentization[J]. Communications of the ACM, 2003, 46(10): 35 - 40.

[3] NIOLOUDIS N, MINGINS C. XML Web services automation: A software engineering approach[C]//Proceedings of the Ninth Asia-Pacific Software Engineering Conference. Washington, DC: IEEE Computer Society, 2002: 417 - 424.

[4] CURBERA F, KHALAF R, MUKHI N, *et al.* The next step in Web services[J]. Service-oriented computing, 2003, 46(10): 29 - 34.

[5] SRU: Search and Retrieve via URL [EB/OL]. [2010 - 03 - 15]. http://www. loc. gov/standards/sru/index. html.

[6] 李春旺，王小梅，王昉，等. 基于 SRU 的集成服务平台设计与实现[J]. 现代图书情报技术，2007(10)：12 - 15.

[7] SELBERG E, ETZIONI O. The MetaCrawler architecture for resource aggregation on the Web[J]. IEEE Expert, 1997, 12(1): 11 - 14.

[8] HODGSON C. Understanding the OpenURL framework[J]. NISO Information Standards Quarterly, 2005, 17(3): 1 - 4.

[9] COLLINS M D D, FERGUSON C L. Context-sensitive linking: It's a small world after all[J]. Serials Review, 2002, 28(4): 267 - 282.

[10] WALKER J. Open linking for libraries: The OpenURL framework[J]. New Library World, 2001, 102(4/5): 127 - 134.

（编辑　张　红）