●黄如花 陈 朋

基于网络的集成化信息检索*

摘 要 集成化信息检索是以信息集成与服务集成为显著特征,以达到知识共享的最大化为目的,实现对由互联网连接起来的数字资源库群的分布式管理及跨平台、跨语种的网络化存取。参考文献 7。

关键词 集成化信息检索 跨库检索 异构平台 网络环境
分类号 G354

ABSTRACT Integrated information retrieval is to realize the distributed management and the cross-platform and cross-language online access of Internet-linked digital resources, with the characteristics of information integration and service integration and the objective of maximum sharing of knowledge. 7 refs.

KEY WORDS Integrated information retrieval. Cross search. Heterogeneous platform. Network environment.
CLASS NUMBER G354

集成化信息检索是以信息集成与服务集成为显著特征,以达到知识共享的最大化为目的,实现对由互联网连接起来的数字资源库群的分布式管理及跨平台、跨语种的网络化存取。集成化信息检索是顺应用户的需求,本着界面无缝化、集成化、统一化的检索理念,为解决异构数据库的统一检索问题而提出的。目前,集成化信息检索主要集中于异构资源的统一检索,尤其是跨库、跨平台检索系统的研究与开发。

1 开发集成化信息检索系统的必要性

1.1 用户的需求

在因特网上,多种硬件系统、平台与系统软件由不同的网络协议和网络体系结构连接着。用户要面临检索前对多种超大规模数据库、多媒体和分布式体系结构的选择,检索过程中的权限要求与 IP 验证,检索结果中出现的死链、大量无用或重复信息等障碍,用户难以在一个集成化信息检索平台上一站式获取和处理自己所需信息。而用户真正需要的是在一个站点、通过一个步骤、进行一次检索就能获得所有的资源与服务。用户的需求是推动信息服务产业发展的根本动力,以满足用户的信息需求为目的的集成化信息检索的开发已成必然之势。

1.2 资源共享的需要

目前,国内外许多信息服务机构已建立了自己的信息平台,但这些数据库与数据库之间、系统与系统之间没有互联互通,无法共享和综合利用,实际上形成了一个个信息孤岛。互联网的迅猛发展将信息化推到一个崭新高度,为实现资源共享提供了有利条件。而实现资源共享的前提,就是要消除信息孤岛,加强互联互通的网络体系建设,保障信息快速有效传递。正是出于这样的考虑,信息产业部将集中力量抓好信息资源的开发利用,鼓励发展各类公共数据库,加快信息资源共享体系建设。

1.3 图书馆学五法则在网络环境下的运用与发展

早在 1931 年,印度著名图书馆学家阮冈纳赞就提出了"图书馆学五法则"。他认为图书馆生存之道只能建立在"读者至上"的服务机制与"以用户为中心"的服务理念之上,要顺应时代的需要而不断实现工作的创新[1]。为了顺应信息技术带来的冲击与影响,美国研究图书馆组织集成化信息服务分部的高级分析师 Walt Crawford 和美国图书馆协会图书馆与信息技术分会的现任主任暨加州大学图书馆馆长 Michael Gorman 在其合著的 *Future libraries: Dreams, Madness & Reality* 一书中提出了"图书馆学新五律",明确指明了信息时代数字图书馆服务的方向:图书馆为全人类服务、知识传递多元化、利用科技提升服务、保障知识自由取得和承前启后再创新[2]。网络环境为这些原则的实现提供了更大的便利。基于网

* 本文是国家社会科学基金项目的成果(项目编号:03BTQ021)。

络的信息检索系统要体现出以人为本的人性化服务观念，必须尽可能地节省读者的时间与精力，减轻其认知负担，将用户需要的信息资源集中于一个界面，屏蔽不同数据库之间的各种差异。

1.4 现行检索方式存在种种弊端

现行数据库的单库检索机制对用户的要求很高：熟悉每个数据库的资源范围、收录年限与检索界面；掌握其独有的检索指令、检索步骤与逻辑算符；具备一定的计算机操作水平和信息素养，对繁杂的检索结果进行多次检索以便去重及过滤。

面向专业用户的 DIALOG，STN，Lexis-Nexis 等著名的国际联机信息检索系统虽然可以实现跨库检索或跨文档检索，但是这些数据库或文档在一个主机上，必须具有相似的结构、字段与索引，且主机负担相当重。没有受过专门训练的用户习惯使用的搜索引擎只能查询静态的、公开发布的页面，对灰色文献、动态文献和隐性网站（invisible web）却无能为力，搜索到资源的范围与覆盖率也相当有限；而且搜索引擎采用关键词的字面匹配，缺乏对关键词所处的语言环境以及对知识的处理与理解能力，无法满足用户的个性化需求，查准率远远低于商用数据库。

现行的数据库和搜索引擎的检索方式大大加重了用户的认知负担，用户对一个课题的检索需要在不同的数据库或搜索引擎之间切换，众多的平台与数据库已经开始让读者无所适从。用户还有可能因为过多的并发用户而不得不忍受漫长的等待方能登录，检索的时间与费用成本大大超出了用户的承受能力。

基于网络的集成化信息检索可以实现各种类型信息的跨来源的检索，理解不同的结构、分类和术语，有时还能利用用户的知识或智能，甚至允许用户进行个性化定制。

2 开发集成化信息检索系统的可行性

2.1 标准与协议的支持

基于网络的集成化信息检索系统的开发和运行得益于通用的网络协议，更依赖于和信息处理、检索与传输等有关的标准与协议。

2.1.1 标准标记语言

在资源加工软件系统中采用标准标记语言对资源内容进行标记，是实现高效跨库检索的重要基础。当用户查询某一特定内容时，借助于支持这些语言的查询引擎，就可以将多个平台、站点中的相关信息一并呈现在用户计算机屏幕上。

可扩展标记语言（XML）是有关信息处理的一系列国际标准之一，用来定义具有特殊目的的标记语言。它在集成化信息检索中突出的优势是可以实现不同来源数据的集成。XML 能够使不同来源的结构化的数据很容易地结合在一起，软件代理商可以在中间层的服务器上对从后端数据库和其他应用处得来的数据进行集成，然后，数据就能被发送到客户或其他服务器做进一步集合、处理和分发。

除了 SGML/XML，被万维网联盟（W3C）推荐为标准标记语言的还有：标准通用标记语言（SGML）、可扩展层叠语言（The Extensible Stylesheet Language，XSL）、可扩展层叠语言转换（XSL Transformations，XSLT）、可扩展标记语言路径语言（XML Path Language，Xpath）和可扩展标记语言链接语言（XML Linking Language，XLink）等。

2.1.2 元数据标准

元数据是因特网上组织信息与资源发现的重要工具，由一组有关资源的各个方面的属性组成，每个属性包括一个属性类型的一个或多个属性值。目前出现了很多种元数据规范，如：都柏林核心（Dublin Core，DC）、因特网内容选择平台（Platform for Internet Content Selection，PICS）、资源描述框架（Resources Description Frame，RDF）、编码档案描述（Encoding Archival Description，EAD）、更结构化的文本输入创始计划（Text Encoding Initiative，TEI）和政府信息定位器服务（Government Information Location Service，GILS）等。都柏林核心元数据已成为国际上最通用的元数据，也是 W3C 推荐的元数据标准。

2.1.3 互操作协议

元数据采集的开放文档先导协议（Open Archives Initiative Protocol for Metadata Harvesting OAI）是建立在 HTTP 协议基础上的应用协议，旨在实现分散的、不同系统平台之间元数据的交换与共享，简化电子资源的有效传递，提高系统的互操作能力。OAI 定义了能够从存储器获取含有元数据记录的机制，从而使资源提供者响应服务提供者的请求，并以 OAI 要求的格式（XML）提供元数据；服务提供者采集元数据，并基于元数据提供增殖服务。由于 OAI 协议的简单性、灵活性和平台独立性，许多数字图书馆项目都提供了 OAI 接口，如著名的“美国的记忆”、“网络化学位论文数字图书馆”和“伦敦档案”等。

2.1.4 Z39.50信息检索协议

Z39.50协议已经发展为美国国家信息标准和基于网络的信息检索国际标准。它支持计算机使用一种标准的、相互可理解的方式进行通讯,并支持不同数据结构、内容、格式的系统之间的数据传输,实现异构平台、异构系统之间的互联与查询。凡是支持Z39.50标准的检索系统都可以直接检索其他支持Z39.50标准的检索系统中的数据库。源系统还可同时对多个支持Z39.50协议的目标系统进行广播式检索,即使用同一个检索表达式,同时对多个目标系统进行检索,并将结果整合。

2.1.5 开放的统一资源定位器(OpenURL)

它是在不同的信息源之间进行数据传递的一个标准。它通过统一资源定位器将信息传递给SFX解析服务器,解析服务器受理URL语法并提供上下文的链接。它引入SFX解析服务器,拓宽用户与所求资源之间的通道,实现不同供应商的或不同平台上内容关联的数据库间的相互链接;可统一检索不同网址上的多个数据库或信息资源,并避免因网址改变或网络阻塞等故障导致的"死链",还可根据用户需求提供匹配的资源服务[3]。图书馆、数据库供应商和信息服务商等机构的资源只要遵守OpenURL,就能被SFX服务器识别并传递给服务对象。如果资源提供商在其数据库中内置SFX按钮,各个图书馆就能通过SFX服务器实现与这些电子资源的动态链接。目前,SFX服务链接的对象包括文章的全文、摘要、发表该文章的期刊的目录、引文标题内关键词网页检索、馆际互借请求、局域网书目自动检索及数据库内同一作者的多文章检索。

其他相关的协议还有馆际互借协议(InterLibrary Loan)等。

2.2 数据库技术的发展

2.2.1 面向对象的技术与CORBA

面向对象的方法吸取了结构化的主要思想与优点,综合功能抽象和数据抽象,将数据与操作放在一起,作为一个相互依存不可分割的整体来处理。公共对象请求代理体系结构(Common Object Request Broker Architecture, CORBA)是基于面向对象的应用软件体系结构和对象技术规范,其核心是一套标准的语言、接口和协议,以支持异构分布应用程序间的互操作性及独立于平台和编程语言的对象重用。它解决了分布式计算环境中不同硬件设备和软件系统的互联,增强网络间软件的互操作性,使构造灵活的分布式应用系统成为可能。

2.2.2 动态数据库访问技术

开放数据库互连(Open Data Base Connectivity, ODBC)实际上是一个数据库访问库,包含访问不同数据库所要求的ODBC驱动程序。应用程序要操作不同类型的数据库,只需调用ODBC所支持的函数,并动态链接到不同的驱动程序。目前,PowerBuilder利用ODBC可实现对多个数据库间的动态切换。

Java数据库连接(Java Data Base Connectivity, JDBC)技术则可为各种常用数据库提供无缝联接。它定义了Java语言同SQL数据之间的程序设计接口,支持不同的关系数据库,使得程序的可移植性大大加强,而且用户可以使用JDBC-ODBC桥驱动器将JDBC函数调用转换为ODBC。JavaSoft公司开发的标准统一的SQL数据存取接口——JDBC API为Java程序提供了一个统一、无缝地操作各种数据库的接口。

2.2.3 通用服务中间件技术

中间件(middleware)是位于平台(硬件和操作系统)和应用之间的通用服务,这些服务具有标准的程序接口和协议。针对不同平台,它们可以有符合接口和协议规范的多种实现。中间件可运行于多种硬件和操作平台,支持分布计算,提供跨网络、硬件和操作平台的透明应用或交互服务;支持标准的协议与标准的接口。中间件提供的程序接口定义了一个相对稳定的高层应用环境,不管底层的计算机硬件和系统软件怎样更新换代,只要将中间件升级更新,并保持中间件对外的接口定义不变,应用软件几乎不需作任何修改。中间件能够屏蔽操作系统和网络协议的差异,为应用程序提供多种通讯机制;并提供相应的平台以满足不同领域的需要。中间件为应用程序提供了一个相对稳定的高层应用环境。

2.3 网络检索工具的启示

网上的独立搜索引擎的覆盖面有限,各搜索引擎的用户接口又是异构的,且有其特定而复杂的界面和查询语法,这非常类似于异构数据库的单库检索。为了弥补独立搜索引擎的不足,增加一次可调用的搜索引擎的数量,提高查全率,元搜索引擎便应运而生。在用户看来,元搜索引擎提供的是一个能够同时查询多个搜索引擎的集成界面,屏蔽了各个搜索引擎的位置、接口等细节。而在后台,它将用户的检索请求同时提交给不同的独立搜索引擎执行检索,并将来自于不同搜索引擎的检索结果进行去重、统一排序后以统一的格式提供给用户。

网上的检索代理比元搜索引擎调用的资源更多。如 Copernic 可指定调用某些搜索引擎或调用某些数据库，还可以进行限制性的定制，Webseeker 还允许用户增加自己的搜索引擎或增加自己的数据库，因此，可以借鉴和引入元搜索引擎与网络检索代理的思想，将一个数据库看成是一个独立的搜索引擎，从而设计一个集成化检索系统，实现数据库的跨库检索。

2.4 国内外若干个集成化检索系统的范例

2.4.1 国外的集成化检索系统范例

典型的代表是“英国电子图书馆计划”的复合图书馆(hybrid library)项目。复合图书馆就是要在一个现实的图书馆中，采用数字图书馆的各种技术，跨越不同载体，实现对信息资源存取的无缝化与检索集成化。作为 eLib 的 5 个复合图书馆原型建设项目之一，AGORA 基于 MODELS 理念建立复合图书馆管理系统，以实现资源的发现、定位、检索与传递。AGORA 高级检索允许用户一次同时检索多个数据库，无论其结构是 MARC，还是 EAD、DC、GILS 及 CIMI；利用 URL 能确定并传递可通过 Web 检索的电子资源；用户能定义检索结果的显示方式或保存检索结果，并可根据查看的目标进行去重与过滤；将 Z39.50 资源和非 Z39.50 资源整合在一起等[4]。

英国的 DNER 系统结合 AGORA 复合图书馆与资源发现网(Resources Discovery Network, RDN)平台功能，借鉴模式化信息构建(MODELS Information Architecure, MIA)的思想，构建了由终端用户客户机、资源表示层、协调层、中间层、交流层和资源与服务提供层构成的信息平台，运用多种应用协议(如 LDAP, Whois++, Z39.50, ILL, FTP)有效地整合了多种资源与服务。最具特色的是其中间层能识别终端用户，为他们建立个人档案，根据档案为该用户提供个性化的检索与查询服务及使用环境[5]。

2.4.2 国内的集成化检索系统范例

我国集成化信息检索的开发也取得了一定的进展，各种集成化信息系统与平台相继诞生。中国国家科学数字图书馆(Chinese Science Digital Library, CSDL)的跨库集成检索系统可以实现对 9 个全文数据库、1 个文摘数据库和 19 个 OPAC 数据库的集成化检索。为了帮助用户使用电子资源时实现跨平台的一站式检索，中国高等教育文献保障体系和北京大学图书馆共同开发了资源统一检索平台，目前正在试运行。清华大学图书馆在其主页推出了集成化检索功能，但有权限限制。清华同方的异构统一检索平台(Uniform Search Platform, USP)通过一个统一用户界面，帮助用户在多个网络数据库搜索平台上实现信息检索操作。用户向 USP 发出检索请求，USP 根据配置信息，把检索请求转换成对应于不同搜索引擎的实际检索请求，并向多个搜索引擎发出实际检索请求，搜索引擎执行检索请求后将检索结果传送回 USP，USP 再把检索结果进行智能化整合，最后把检索结果传送给用户。USP 为用户提供各种信息资源，目前已经拥有中国知识创新工程数据库(CNKI)、科技新刊报道数据库(STARTS)、CALIS 高校学位论文数据库、(香港)中国资讯行数据库(China InfoBank)、工程索引(EI)、科学引文索引(SCI)、OCLC 数据库和商业经济文摘数据库(ABI/INFORM)等 60 多个专业数据库引擎，随时为用户服务，不同搜索引擎检索结果在同一窗口内显示。

西安交通大学图书馆跨库检索系统可在统一界面下检索 19 个中外文全文数据库、二次文献数据库、专题数据库和特种文献数据库。并提供简单检索与高级检索两种选择，前者可从字段和时间范围进行选择，结果按时间和相关度排列，后者支持布尔逻辑检索。南京大学图书馆网络数据库一站式检索系统也投入了使用，但全文资源只限南京大学校园网内部用户访问。上海交通大学图书馆也正在开发集成化检索平台。

3 开发集成化信息检索系统的问题

集成化信息检索是用户、研究人员与开发者们共同的愿望，但在当前，要真正实现集成化检索还有一些困难。

(1)中间件技术的问题。许多集成化检索系统的开发都依靠中间件技术，但中间件服务也并非万能药。多数流行的中间件服务使用专有的 API 和专有的协议，使得应用某一个中间件实现对来自不同厂家数据库的互操作很难。有些中间件服务只提供一些平台的实现，从而限制了应用在异构系统之间的移植。应用开发者在这些中间件服务之上建立自己的应用还要承担相当大的风险，随着技术的发展他们往往还需要不断地对系统进行升级换代[6]。

(2)集成检索系统的智能化问题。集成化信息检索走上智能化的道路是信息检索系统发展的必然趋势，但不会一蹴而就。集成化信息检索系统包括检索方法的智能化、

(下转第 60 页)

并努力开发兼容性较好的软件系统，以保证机构间信息传输的稳定性与正确性。在服务上应注意人才的吸纳和培养，使咨询人员能够对用户的提问尽快作出反应，并尽快通过自己的各种知识帮助用户解决问题，及时反馈给用户。

参考文献

1 肖冬梅.合作数字参考咨询服务研究.图书情报工作，2003(5)
2 朱丽东.数字参考服务形式与问题研究.图书馆论坛，2003(8)
3 赵乃瑄.国内图书馆实行数字化参考咨询服务的探讨.新世纪图书馆，2003(4)
4 白广思，李朝明.省级数字参考咨询服务系统研究.河南图书馆学刊，2002(10)
5 张晓林.数字化参考咨询服务.四川图书馆学报，2001(1)
6 李晓芸.数字参考咨询服务新进展.图书馆学研究，2002(3)
7 http://www.lib.stju.edu.cn/chinese/cirtual reference desk/websites.htm
8 http://www.zslib.com.cn

焦玉英 武汉大学信息管理学院教授，博士生导师。通信地址：武汉。邮编430072。
王娜 武汉大学信息管理学院情报学硕士，研究生。通信地址同上。 (来稿时间：2004－05－21)

(上接第49页) 检索中各服务的个性化和人机接口的简易化三个主要方面，是基于语音识别、信息抽取和自然语言理解的检索形式，其核心是对用户的查询计划、意图、兴趣等进行跟踪而又不会泄露用户隐私。这必然涉及到信息的挖掘、信息抽取、信息过滤、知识发现和信息推送等多种技术。而集成化信息检索系统的开发才开始不久，还远远没有成熟到能集各种优秀的信息资源于一身，更难以实现深层的知识挖掘与推理。

(3)检索速度与质量协调的问题。网络系统中的服务器主要负责元数据的索引和查询，将查询结果通知对象服务器(可在本地，也可在异地)，并由对象服务器取出最后结果，这就要求服务器具有集中管理性、可扩展性、高速传输性等优越性能。而集成化的检索系统处在超大容量的分布式资源库、超大规模并发访问的用户群、甚至7/24全天候的在线服务环境中，必须配置具有快速准确的检索能力和简单检索界面的全文检索软件、性能优良的操作系统、信息安全系统，以及数据库管理系统和调度系统等，方能保障系统的容量、平均反应时间、吞吐能力和重载情况(负载很重时)的稳定性等。

(4)数据库厂商出于本能，对跨平台会有强烈的排斥心理，除非这个跨平台由他自己开发。由于利益的驱动，数据库厂商通常会开发自己独特的检索和阅读平台，而且还会极力对自己数据库的库结构、软件设计方法等进行保密，而这些恰恰是对开发集成化检索系统具有重要作用的信息[7]。

集成化检索中还需要解决知识产权问题、跨语种检索和机器翻译等技术问题。图书情报界对资源使用的观念也有待更新。而且，集成化信息检索不是最终目的，用户希望借助于网络超媒体、超链接和检索不受时空限制等优势，在集成化信息检索的基础上获得全文链接、文献传递和虚拟参考服务等。

因此，我们指的集成化信息检索系统的集成程度是有限度的，而要建设一个能够包容全世界所有数据库的"超级大平台"是不现实的，也没有任何一个机构或用户能够订购所有的数据库。

参考文献

1 Ranganathan, SR. Five laws of library science. Madras, Madras Library Association; London, G Blunt and Sons, 1931
2 Crawford, Walt and Michael Gorman. Future libraries: dreams, madness & reality. Chicago and London: American Library Association, 1995
3 王善平.万维网资源整合工具——Open URL.上海交通大学学报，2003(增刊)
4 Agora Demonstrator. http://hosted.ukoln.ac.uk/agora/demonstrator.html(访问时间：2004－04－18)
5 Andy Powell. DNER Portal Architecture. http//www.rdn.ac.uk/publications/mia/(访问时间：2004－04－18)
6 中间件. http://www.huihoo.com/middleware/index1.html(访问时间：2004－04－19)
7 刘锦山.跨平台神话的破灭. http://www.chinalibs.net(访问时间：2004－04－19)

黄如花 武汉大学信息管理学院副教授，博士。通信地址：武汉。邮编430072。
陈 朋 武汉大学信息管理学院2002研。通信地址同上。 (来稿时间：2004－04－23)