# 文献资源统一检索系统原理

李俊敏 刘 军 陈良强
(浙江大学，杭州 310027)

〔摘 要〕 互联网环境下文献资源数据库快速发展，为统一检索系统提出了需求。本文结合当前文献资源统一检索的发展情况，阐述了统一检索系统的原理和实现方法，比较了几种用于实现统一检索的协议与技术，探讨了各自的优缺点，分析了统一检索系统的用户需求和发展方向。

〔关键词〕 文献资源；统一检索；系统原理；信息检索

〔Abstract〕 In this paper, we expatiated on the principle and method for designing or developing a document resource unitied retrieval system, discussed their advantages and disadvantages, compared several protocols and technologies used for unitied retrieval, and then demonstrated the developmental direction of unitied retrieval system for future.

〔Key words〕 documents resource; unified retrieval; system principle; information retrieval
〔中图分类号〕G253 〔文献标识码〕A 〔文章编号〕1008－0821 (2007) 06－0120－03

随着互联网技术的广泛应用，大量产生的数据库电子文献资源为读者利用文献提供了很大的便利。同时，为了应对读者需要，多个系统、多个数据库的统一检索系统理论和应用逐步发展起来，为用户提供了更好的电子文献资源获取途径。

## 1 统一检索是文献资源检索的发展趋势

数据库电子文献的大量产生和广泛应用，大量数据库提供商参与市场竞争，形成了各个文献资源数据库都具有分布式、异构性、访问方法和检索界面多样化等特点。图书馆系统涉及到对多种类型的数字资源信息检索和获取，导致了检索的复杂性和不规范性。各个电子文献数据库分别有各自的检索界面，读者为了找到特定的某篇文献资源，往往要登录到多个系统在多个数据库中检索。而且对于某些特定数据库才收藏的文献，如果读者不了解该数据库，就无从检索。

Web搜索引擎为解决以上问题提供了一个可能的选择。但是，目前Web搜索工具主要搜索那些在网上公开的并且可以自由访问的信息资源，很难搜索到访问受到限制的资源，如出版物、书籍数据库中的资源等。目前Web搜索引擎很难满足学术科研需要。

统一检索系统理论与应用实例为解决文献资源检索问题提供了更好的解决方案。统一检索系统能够对分布在本地和异地的各种异构资源包括原文、图片、引文、文摘、馆藏、相关文献等提供统一的检索界面和统一的检索语言，从而提高资源的利用率。统一检索系统的目标是为用户提供统一检索、一次性用户认证、不同系统之间的无缝链接和完整的服务体系，从而使读者一次检索，就可以将馆藏各类文献查找完毕，输出全部检索结果。

数字图书馆是当前快速发展的领域，已经积累了大量电子文献和电子化书籍。但是跨库跨平台的统一检索发展相对滞后。在解决异构平台的信息资源检索技术上，国内图书馆间存在较大差异，有的可以实现部分数据库间的跨库检索，大部分不能向用户提供方便检索的统一界面，不能提供异构数字资源间的互操作。

中国高等教育文献保障系统管理中心（CALIS）的统一检索平台已推出3.5版，仍在发展中。维普、万方、清华同方等公司的统一检索或跨库检索系统已经初步进入实用阶段。

## 2 统一检索系统的实现原理

当前文献资源统一检索主要是基于对元数据的统一检索。元数据是关于原始数据的信息，例如，某篇文献或者某本书籍的题名、作者、出版者、摘要等信息是关于该文献或者该书籍的元数据。另外一种实现统一检索的技术是基于语义的检索，目前研究较多的是Ontology技术，是一个很有潜力的智能化统一检索技术，但是还没有实际使用。

文献资源统一检索涉及到统一检索用户界面和检索功能设计、异构数据库的互操作、查询结果处理中的信息融合等问题。基于元数据的统一检索系统要充分发现各个源数据库的共性，同时要兼顾不同数据库的差异性，让用户使用统一检索的同时，也可以选择使用特定数据库的独特检索功能。

### 2.1 直接整合文献资源检索接口的方法

各文献资源提供商一般都提供了Web检索页面，利用这些文献资源数据库系统提供的Web客户端访问接口如CGI、asp、jsp、aspx等检索页面，提取共性部分，构建统一检索页面，针对用户在统一检索页面中输入的查询条件，利用多线程技术同时构造针对各个数据库系统的查询表单数据，用HTTP协议Get或者Post方法提交，获取并分析返

收稿日期：2006－08－07
作者简介：李俊敏，现在浙江大学图书馆工作。
刘 军，现在浙江大学图书馆工作。
陈良强，现在浙江大学图书馆工作。

回的结果数据，返回的结果数据一般都是 HTML 或者 XML 格式，根据 HTML 或者 XML 特定标签或者标识来分析处理和合并目标数据，这些目标数据包括题名、作者、出版日期、摘要等元数据以及获得的元数据条目数目、全文超连接等，最后将处理结果呈现给最终用户，实现了统一检索的目的。

这种方法适用于所有提供了 Web 查询检索访问的数据库，不需要源数据库系统做出任何修改，具有广泛的适应性，并且实现起来难度不大。但是这种方法需要针对不同的信息源开发相对应的检索接口，即配置数据源。在当前文献资源种类繁多、数据库结构各异、查询结果页面常有变化的情况下，需要大量的人力来配置数据源，系统维护成本很大。尤其是目前多数数据库 Web 检索返回的是 HTML 格式的数据，某些返回的 HTML 页面格式经常变动，导致提取目标数据很困难，需要经常地检测和更新数据源配置参数。另外，为了统一检索界面设计的需要，不得不舍弃了一些文献资源数据库系统的特定检索功能，这样就不能满足读者的某些特殊检索需要。

Calis 统一检索系统实现了这种方法，能够对分布在本地和异地的各种异构资源提供统一的检索界面和检索语言，用户可按学科、按数据库名称、按语言种类同时检索多个平台上的多种资源，输入一个检索式，便可以看到多个数据库的查询结果，对多个库结果合并和排序，并可进一步得到详细记录和下载全文。

### 2.2　可以用于实现统一检索的协议与技术

如果各个文献资源数据库系统都开发符合某种标准的接口，并且制订了通过网络访问这种接口的标准，那么，统一检索系统就可以方便地访问这些数据库系统，从而为用户提供统一的检索界面。为此，可以利用现有协议以支持统一检索，也出现了专门用于实现统一检索的协议。以下分析了几种常用协议的原理和优缺点。

#### 2.2.1　Z39.50 协议

Z39.50 协议是一种客户机/服务器体系结构中的机器间信息检索的应用层协议，定义了一种通用的语言以执行信息选择、检索和获取，使客户端和服务器之间的通讯和互操作标准化，为位于客户端的用户界面与文献资源数据库服务器相分离提供了一种解决方案。Z39.50 协议定义了多种资源格式、检索人口、检索属性集，支持不同数据结构、内容、格式的系统之间的数据传输，支持分布式环境下的全文文档、书目目录数据、图片、多媒体等信息的检索，可以满足信息资源创建者、提供商和用户的不断变化的需求，实现异构系统之间的互联与查询，对各种不同资源提供统一检索界面，并进行资源整合。

Z39.50 协议的运行机制是，客户机提供用户界面，它向服务器发送查询请求，服务器软件负责管理信息、执行检索、返回结果。具体过程是，客户端建立一个到服务器的连接，初始化一个 Z39.50 协议会话，磋商查询等操作的条件、边界值如最大返回记录数等；然后客户端就可以提交经过标准化的查询，查询语句指定要查询的数据库、查询目的参数以决定记录是否需要返回；服务器解析查询语句，在指定数据库中执行检索、创建结果集，并根据指定的参数返回记录；继而客户端可以获取结果或者继续进行更多查询；客户端可以进一步使用服务器端提供的检索服务、扫描服务、分类排序服务等，可以在必要时删除结果集；最后客户端处理、合并结果并显示结果给用户。Z39.50 协议也提供了访问控制机制和记账/资源控制机制、解释机制、扩展服务群机制和终止服务机制等。

Z39.50 协议的实现与计算机硬件、操作系统、数据库、搜索引擎等无关，已经广泛地应用在分布式检索系统中，现在大多数图书馆自动化系统、联合目录系统、统一检索系统等都实现了 Z39.50 协议，并且与之相关的开放源码也很多，方便了系统开发。国内外有许多利用 Z39.50 协议实现跨库统一检索的实例。但是，Z39.50 协议比较复杂，学习难度较大，协议的开发则更难，开发成本高。Z39.50 协议采用有状态的、安全的网络连接，使得运行成本较高。同时，Z39.50 协议是客户机/服务器模式下的协议，不适合在因特网中推广使用。这些缺点限制了它在统一检索系统中的使用。

#### 2.2.2　OpenURL 协议

OpenURL 协议也即 Z39.88 协议是一种开放的信息资源与查询服务之间的通信协议标准，其核心是定义了一类用于描述上下文的对象和一种传输机制。同时规定了一套完整的 OpenURL 框架，用来规范对上下文对象的描述和传输，它提供了一种在信息服务者之间传递对象元数据的格式。不同机构、不同领域的异构资源可通过对框架中各组件元素进行注册来实现对 OpenURL 的支持。OpenURL 协议用 HTTP 的 POST 或者 GET 方式发送查询参数，POST 和 GET 方式共用相同的语法结构。OpenURL 定义了一种传输信息的元数据或资源信息的语法标准，将查询字符串用 URL 编码，生成传输元数据及访问元数据的可执行 URL，传送给支持 OpenURL 协议的数据库系统。

SFX 系统中的 SFX 连接采用 OpenURL 标准实现不同信息资源和服务组件之间的互操作，允许在一个开放连接的环境下进行本地化。SFX 系统在图书馆领域有一定应用。

#### 2.2.3　OAI－PMH 协议

OAI－PMH 协议的原理是通过元数据收获的方式从数据提供者中获取元数据，并存储在本地的元数据库中，然后在本地数据库基础上向用户提供基于元数据的统一检索服务。OAI－PMH 协议定位在轻量级别的互操作，由于这种收获与仓储的具体实现无关，仓储只需要提供符合该协议的元数据，并不用开放其本地资源，在目前条件下，这种互操作框架是比较现实和可行的，因而逐步受到重视和应用。

#### 2.2.4　Dublin Core 规范

Dublin Core 规范是用于标识电子资源的一种简要目录模式。它参照图书馆卡片目录的模式，制定了 15 项广义的元数据（Metadata），包括名称（Title）、创作者（Creator）、主题及关键词（Subject and Keywords）、说明（Description）、出版者（Publisher）、发行者（Contributor）、时间（Date）、类型（Type）、格式（Format）、标识（Identifier）、来源（Source）、语言（Language）、相关资源（Relation）、范围（Coverage）、版权（Rights）等。比较全面简洁地概括了电子资源的主要特征，涵盖了资源的重要检索点、辅助检索点或关联检索点，以及有价值的说明性信息。这些元数据

不仅适用于电子文献目录，也适用于各类电子化的公务文档目录，产品、商品、藏品目录，具有广泛的实用性。符合Dublin Core协议规范的数据可以用HTML、XML或者RDF格式表示，这为它在Web环境下的传输提供了很大的方便，可以使用HTTP、SOAP或者其他传输协议来交换数据。

2.2.5 基于WebService标准的协议

通过Google因特网搜索API，应用程序可以实现在线使用Google的搜索服务。Google因特网搜索服务API按照WebService标准构建，应用程序可以通过WSDL发现其SOAP接口规范，然后构建查询字符串和查询参数集，以SOAP协议格式提交，并接收返回结果。Google Web APIs规范定义了搜索查询格式、搜索参数、过滤器、限制条件、输入输出编码等，返回的结果是xml格式的，并且有其一定的包装规范。随着Google学术搜索的推出，利用或者整合Google学术搜索服务作为实现统一检索的一种途径具有一定的价值，并且对于开发统一检索系统具有参考价值。

中国华中科技大学开发的信息检索服务协议（Information Retrieval Webservice Protocol，IRWP协议）已经作为草案提交到RFC组织。IRWP协议利用目前正在快速发展的WebService技术，为开发基于开放的公用网络协议实现统一检索，提供了一种新的选择。

Fedora开放源代码项目为数字内容的管理和发布提供了一个可扩展的面向服务的架构。其核心是一个支持多视图和关系的数字对象模型，数字对象用来包装本地管理的内容或者对远程内容的索引，而多视图支持为数字对象的Web服务提供了可能。Fedora的所有功能都以Web服务的形式发布，并受一定的访问控制策略保护。Fedora可以用于图书馆丛书管理、多媒体出版系统、知识库和数字图书馆。

### 2.3 建立统一检索数据库

整合已有的数据库，合并为一个全新的统一的数据库。可以利用现有数据库检索平台将目标数据检索出来添加到一个全新的统一的数据库。也可以建立元数据联合仓库，用统一的数据库结构和统一的查询语言来实现统一检索。例如，对于传统的光盘数据库，可以考虑用此方法实现统一检索，并在Web环境下使用。这种方法需要建立一个庞大的数据库，数据库建立、存储、维护的成本就会随之变得很大，所以只适合小型的统一检索系统。

中国工程物理研究院科技信息中心西文数据库统一检索系统就是以建立一个全新的数据库的方法来实现的。

## 3 结束语

随着数字化书目信息数据库、文献索引数据库、引用数据库、全文数据库等快速发展，统一检索系统理论与应用在快速发展，期望能够为用户提供了统一检索、一次性用户认证、不同系统之间的无缝链接和完整的服务体系。目前正是统一检索理论、系统及配套软件快速发展的时期，国内外开发出了多个用于统一检索的协议、语言或者规范，但是各文献数据库厂商对各种规范的支持不同，很难统一起来。而建立统一检索数据库或者配置数据源的方法，不是统一检索发展的主流，随着文献资源数据库的增多、数据库规模的增大，这两种方法就突显了其缺点。所以，必须建立Web环境下的统一检索标准，推动一种或者几种能在Web环境下使用的统一检索语言获得各个文献资源数据库的广泛支持，才能实现真正的统一检索局面，提升文献资源检索的服务质量。

## 参 考 文 献

[1] 中国高等教育文献保障系统 [EB]. http://www.calis.edu.cn/, 2004.

[2] Information Retrieval: Application Service Definition & Protocol Specification [EB]. http://www.niso.org/standards/standard-detail.cfm? std-id=465, 2003.

[3] The OpenURL Framework for Context－Sensitive Services [EB]. http://www.niso.org/standards/standard-detail.cfm? std-id=783, 2004.

[4] SFX System [EB]. http://www.exlibrisgroup.com/, 2006.

[5] The Open Archives Initiative Protocol for Metadata Harvesting [EB]. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, 2004.

[6] RFC2413 [EB]. http://www.ietf.org/rfc/rfc2413.txt, 1998.

[7] Dublin Core Metadata Initiative [EB]. http://www.dublincore.org/, 2006.

[8] Develop Your Own Applications Using Google [EB]. http://www.google.com/apis/, 2006.

[9] Information retrieval protocol for digital resources [EB]. http://dris.hust.edu.cn/Chinese/Docs/draft－ietf－liang－irpdl－03.txt, 2004.

[10] Fedora Project [EB]. http://www.fedora.info/, 2005.

[11] 王永川. 科技信息港西文数据库统一检索平台 [J]. 信息与电子工程, 2004, 2 (2).

(上接第116页)

## 参 考 文 献

[1] 中共广东省委、广东省人民政府关于加快建设文化大省的决定 [J]. 广东省人民政府公报, 2003, (16): 2－22.

[2] 薛涌. 公共图书馆：文明的见证 [N]. 南方都市报, 2004－06－19, (A03).

[3] 李国新. 论图书馆的法治环境 [J]. 中国图书馆学报, 2000, (3): 25－29.

[4] 刘小琴. 关于中国图书馆法的思考. 见：吴建中主编. 战略思考——图书馆发展十大热门话题 [M]. 上海：上海科学技术文献出版社, 2002: 226－230.

[5] 汤旭岩, 欧阳军, 颜学勤. 地方图书馆立法述论. 见：吴建中主编. 战略思考——图书馆发展十大热门话题 [M]. 上海：上海科学技术文献出版社, 2002: 231－236.

[6] 刘昆雄. 论我国21世纪图书馆“两极”发展 [J]. 中国图书馆学报, 2002, (3): 25－28.

[7] 黄俊贵. 办馆效益探微 [J]. 中国图书馆学报, 2000, (1): 8－13.