现代情报
2008年1月 第1期
January.2008 No. 1

基于本体的跨库检索集成系统框架研究

王轶珺
（南开大学泰达学院，天津 300457）

〔摘 要〕 本文综合考虑外文网络电子资源基于语义的著录、组织、检索和保存等问题，将本体、语义的概念和理论应用于图书馆分布式网络数据库检索领域，设计了基于本体的跨库检索集成系统框架。

〔关键词〕 元数据；本体；跨库检索

〔Abstract〕 This paper considers synthetically serval problems of electronic resources, and applys ontology's theory in library distributed network databases retrieval, then brings forward the frame of multi - database retrieval integration system.

〔Key words〕 Metadata; Ontology; Multi - database retrieval
〔中图分类号〕G250.76 〔文献标识码〕B 〔文章编号〕1008 - 0821 (2008) 01 - 0170 - 06

目前，各高校图书馆网络电子资源增长迅速，电子资源综合利用的问题也日显突出，尤其是外文网络数据库资源，对它的利用现存的主要问题是：缺乏统一检索；检索缺乏语义，不够精确；无法保存较多的数据资源；资源描述和组织方面，缺乏统一、完善的著录和组织体系等。国外研究跨库检索技术比较成熟，但基本为商用型产品，国内在这方面主要集中在中文数据库方面，对于外文数据库的跨库检索也多处于起步或试验阶段。

为了解决外文网络数据库利用中存在的问题，并能综合考虑资源的著录、组织、跨库检索、语义性及保存等多个方面，本人认为，设计一个统一的集语义著录、资源导航、跨库检索、数字资源存储等在内的语义集成系统框架，很有必要，不仅符合未来的网络发展趋势，同时也是将语义网的新技术应用到图书馆电子资源检索和集成方面的一点尝试。

1 设计思想与系统层次结构

基于本体的跨库检索集成框架的主要思路：根据外文网络数据库资源的特点，通过语义层建立概念、概念间的语义关系等，将其用于跨库检索中用户查询语句的语义分析、推理与优化，从而改善跨库检索的准确性，同时兼顾电子资源知识管理和保存问题。在这个思路的基础上，提出本系统框架的层次结构模型图，如图1所示。

<table>
<tr><td>应用层</td><td colspan="2">跨库浏览与检索</td></tr>
<tr><td rowspan="2">语义层</td><td colspan="2">语义分析与推理</td></tr>
<tr><td>元数据库(RDFS)</td><td>本体库(OWL)</td></tr>
<tr><td>数据层</td><td>用户信息库</td><td>电子资源保存库</td></tr>
<tr><td>操作系统层</td><td colspan="2">Windows</td></tr>
</table>

图1 基于本体的跨库检索集成系统层次结构模型

该层次结构模型可以看成是结合语义网模型的一个应用模型。

2 功能设计与模块划分

系统框架的整体功能目标：系统应该具有针对分布、异构的网络数据库的跨库检索功能；跨库检索应该是基于本体的，这样才能更好的满足用户检索需要；系统应该提供用户统一、友好的登录、导航和检索等界面；系统应该符合资源组织的标准和规范，具有良好的兼容性和扩展性；系统应该具有电子资源保存的功能。进一步将系统模块划分为：

2.1 Web用户界面，主要完成用户统一登录、浏览、检索等功能。

2.2 用户查询模块，主要功能是配合系统元数据库和本体库实现查询的语义优化和处理，解决语义查询问题。

2.3 浏览与导航模块，主要是利用基于元数据库和本体库建立的各种索引，满足用户有目的浏览的需要。

2.4 跨库检索模块，主要是解决异构数据库的统一检索问题。

2.5 结果整合模块，主要是对检索返回的结果进行归并、整合等处理，以便将有序的结果提供给用户。

2.6 文档处理模块，主要是为电子资源保存而设的，对于检索结果查重，库中没有则将相关信息保存到元数据库和电子资源库中。

2.7 用户统计模块，主要是记录用户相关信息，如注册信息、检索记录等，以便统计用户对各资源的利用率、用户信息需求等。

3 系统框架、工作原理及流程

在层次结构和功能模块划分的基础上，进而设计了基于本体的跨库检索集成系统框架，如图2所示。

收稿日期：2007—11—26
作者简介：王轶珺，现在南开大学泰达学院图书资料中心工作。

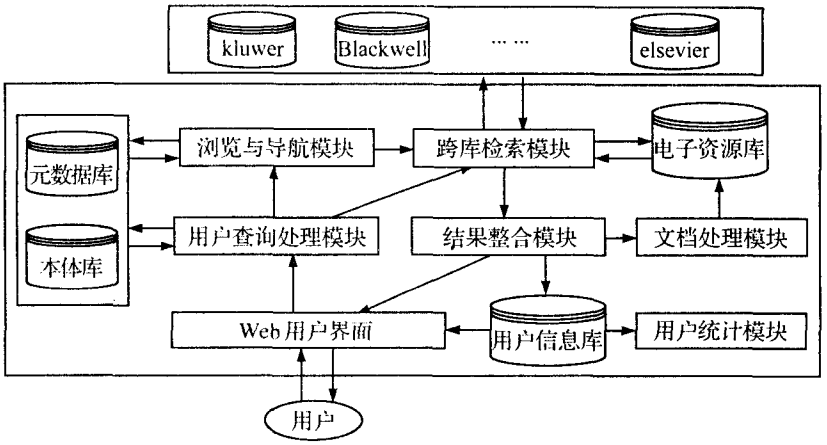


图 2　基于本体的跨库检索集成系统框架图

工作原理如下：

3.1　首先用 DC 和 RDFS 构建外文网络资源的元数据库，用 Protégé 构建分类主题本体库，作为本系统框架的基础，为基于本体的跨库检索集成提供条件和可能。

3.2　用户通过系统的 Web 用户界面的统一认证模块登录，统一认证模块查询用户信息库以确认用户身份。

3.3　身份通过验证后，有效用户就可以开始浏览、检索、使用个性化服务等，无效用户被拒绝使用本系统。

3.4　用户登录成功后，本系统就开始跟踪用户的浏览和检索行为，将所获得的数据经过筛选、提取后存入用户信息库中，以备个性化服务和用户统计之用。

3.5　用户输入的查询条件，先被送入用户查询处理模块，该模块主要是结合本体库和元数据库，对用户的查询语句进行分析、语义处理（如规约、推理、重构等），包括采用过滤手段去除用户查询语句中的禁用词，通过同近义词匹配等语义扩展检索、相关词提示等手段，可以辅助用户正确表述检索需求，使查询结果更准确。语义查询、解析、推理等的实现主要依靠通过 Jena 技术的接口访问元数据库和本体库实现。这是本系统进行语义优化的重要部分，处理后的查询被分别送入浏览与导航模块或跨库检索模块。

3.6　浏览与导航模块是建立在本体库和元数据库基础上的，该模块主要解决依赖索引的库、刊两级的用户浏览和简单检索请求，全文则需要从本地 SQL 数据库或通过跨库检索模块获取。

3.7　跨库检索模块接到处理过的查询语句后，按照用户选择的不同网络数据库源对重组后的查询进行分解，然后用多线程技术实现多个数据库的并发检索。

3.8　检索结果返回后，先不直接返回用户，而是转入结果整合模块进行过滤、排序、去重及合并等处理，处理后的结果数据在返回用户 Web 界面的同时，也提交给文档处理模块、用户信息库（这里只保留简单的检索记录）。

3.9　文档处理模块接到整合的检索结果后，对元数据库和本地电子资源库进行查重、添加等操作，用于电子资源的本地保存。

以上简要概括如下：

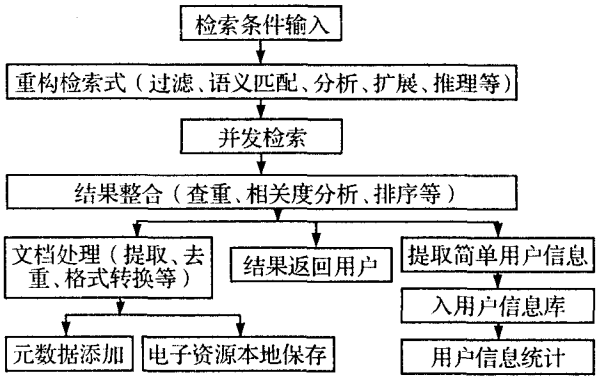


图 3　主要检索过程示意图

4　关键技术问题分析与实现思路

4.1　系统开发语言、软件架构

由于 Java 语言具有语法简单、可移植性好、面向对象、安全性好、支持多线程、开放源代码、扩展性好、方便网络编程与分布式开放等众多优点，并且非常适合解决跨库并发检索及数据库联接等问题。因此，可选择 Java 作为系统的开发语言。

由于本系统框架涉及网络检索、用户 Web 界面设计、数据库操作等多方面，因此，整体软件架构可以选择 Java 平台的 MVC 模式，采用“JSP + Java Servlet + JavaBean”的方式。其中，与用户的交互可以使用 JSP；Java Servlet 配合 JSP 一起实现内容与显示逻辑的分离，比如：Java Servlet 负责处理用户在 Web 界面的输入，并转发给 Model 的用户查询处理模块；主要的业务逻辑部分由 JavaBean 来完成，比如负责执行跨库并发检索、实现结果整合等。

4.2　系统数据库设计

由于系统是基于本体的，并涉及资源组织、保存、用户认证等功能，因此设计 4 个数据库：用户信息库、电子资源保存库、元数据库和本体库。其中，元数据库和本体库是基于本体的系统框架的重要基础，必不可少，没有它们，就不存在应用层的语义检索集成的可能。元数据库自身也具有一定的语义关系，但由于元数据与本体所描述的内容性质不同，在系统中的作用不同，使用的方法和工具也不同，因此将两者分开建库，能提高效率。

4.2.1　用户信息数据库，主要用于存放用户的注册信息、

个性化设置数据及用户统计模块所需的数据，如用户的密码、检索或浏览兴趣偏好、个人收藏夹等。

4.2.2 电子资源保存数据库，主要是保存跨库检索后经过文档处理过的全文数据和一些必要的提供索引用的或便于文献识别的数据，如DOI或本地数据库的唯一编号等。

4.2.3 元数据库，主要功能：资源的知识组织功能；为索引与导航提供支持；为用户查询处理模块提供语义支持；为文档处理模块提供查重基础等。

4.2.4 系统本体库，主要是对抽象的概念、概念间的关系进行描述，主要的功能是为用户查询处理模块提供语义支持，优化用户查询；为索引和导航模块的学科分类导航和主题导航提供支持。它为整个框架提供本体语义支持。

4.3 数据库接口与存储

考虑用户信息库、电子资源库的功能及它们的数据类型情况，适合用表的结构来保存数据，因此这两个库都选择关系数据库来实现。具体可选择Microsoft SQL Server2000关系数据库来实现，它在兼容性、安全性、可扩展性、分布式事务支持等方面都有很多优势。这两个SQL数据库都通过JDBC接口与本系统的Java程序相联接，系统通过JDBC接口来访问数据库。

元数据主要是RDF文档，而系统本体数据使用OWL描述，选择OWL或OWL DB方式保存。系统元数据库和本体库统一通过Jena技术与系统相联接，存储时，一方面保存所有的RDF和OWL文档，以便满足特殊处理或交换需要；另一方面，同时也可将数据存储于MySQL中，通过Jena来访问和转换。另外，由于系统本体库是通过protégé创建的，也可以直接将本体项目存成Protégé Database或OWL/RDF Database，并与MySQL数据库相连接，而系统管理员可以利用Protégé访问和修改MySQL中的本体数据，修改过的数据也可同时被系统通过Jena的MySQL数据库接口所访问和使用。

4.4 元数据库构建

合理、标准化的元数据库，相当于为电子资源做了一定的语义标注，有助于提高检索系统的联想能力和精确性。元数据库主要采用DC都柏林元数据和RDF(S)资源描述框架及模式两种技术来描述电子资源的元数据信息。

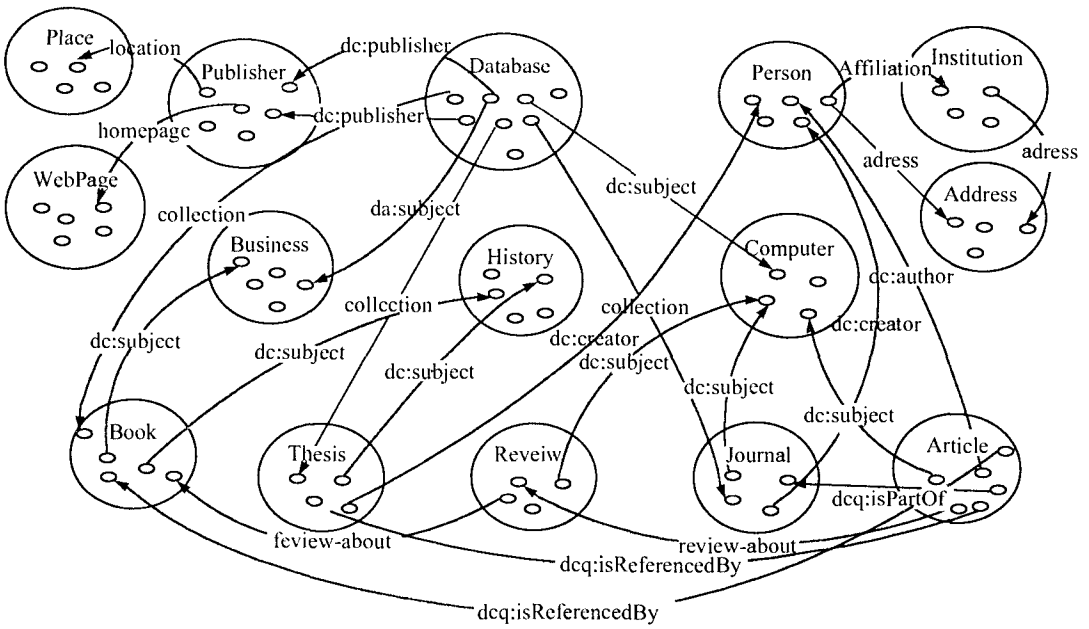
由于网络数据库实际包含的文献类型很多，元数据描述时，首先要针对不同的资源类型，选择适合的DC核心元素，并分别补充适当的自定义元素，然后结合RDF(S)进行元数据描述。

下面是本人以ACM数据库中一篇文章“The CAN microcluster: Parallel processing over the controller area network”为例的元数据描述文档。

```
<? xml version="1.0"? >
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:dcq="http://purl.org/dc/terms/" >
  <rdf:Description >
    <dc:title >
      The CAN microcluster: Parallel processing over the controller area network </dc:title >
    <dc:identifier >
      <rdf:Alt >
        <rdf:li
          rdf:resource =
          "http://portal.acm.org/citation.cfm? id = 1101670.1101672&coll = portal&dl = ACM…" >
        <rdf:li
          rdf:resource = "http://portal.acm.org/citation.cfm? doid = 1101670.1101672" >
      </rdf:Alt >
    </dc:identifier >
    <dcq:author >
      <rdf:Bag >
        <rdf:li > Paul A. Kuban </rdf:li >
          <rdfs:label > author's organization </rdfs:label >
          <rdf:value > University of Southern Indiana, Evansville, IN
          </rdf:value >
        <rdf:li > Rammohan K. Ragade </rdf:li >
          <rdfs:label > author's organization </rdfs:label >
          <rdf:value > University of Louisville, Louisville, KY
          </rdf:value >
      </rdf:Bag >
    </dcq:author >
    <dc:source > Journal on Educational Resources in Computing(JERIC) </dc:source >
      <dcq:vol > Volume 5 </dcq:vol >
    <dcq:issue > Issue 1(March 2005) </dcq:issue >
    <dcq:tableOfContents >
      <rdf:Description
        rdf:about =
          "http://portal.acm.org/toc.cfm? id = 1101670&type = issue&coll = Portal&…" >
      </rdf:Description >
    </dcq:tableOfContents >
    <dcq:page > Pages:1 - 12 </dcq:page >
    <dc:identifier > ISSN 1531 - 4278 </dc:identifier >
  <dcq:date > 2005 </dcq:date >
  <dc:identifier > DOI 10.1145/1101670.1101672 </dc:identifier >
  <dcq:keywords > CAN, Cluster, controller area network, distributed, embedded systems, microcontrollers, parallel </dcq:keywords >
  <dcq:abstract >
    Most electrical engineering and computer science undergraduate programs include…
</dcq:abstract >
<dc:type > journal/article </dc:type >
```

```
    < dc:format > PDF < /dc:format >
    < dc:format > 99K < /dc:format >
    < dc:language > en < /dc:language >
    …
  < /rdf:Description >
< /rdf:RDF >
```

其它文献类型的描述过程类似，只是选取 DC 元素有所区别，所以不再一一叙述。而众多的元数据信息汇集在一起时，最终形成一个语义信息网，对于索引、导航、语义分析和检索等都非常有用。为说明元数据库中各种类、属性与实例之间的语义关系，本人做了一个局部的示意图如下：



（注：其中大圆圈代表类；小圆圈代表实例；弧线箭头代表属性。）

图 4　元数据库中类、属性与实例语义关系图

本小节仅是对构建元数据库的基本思路和描述过程的一点展示，构建完整的元数据库还需大量的工作去完成。

4.5　本体库构建

本体构建思路：结合传统文献组织技术（分类法、主题词表等），利用 Protégé 编辑工具、Racer 推理机来创建本体和本体推理尝试。系统本体库主要采用 OWL 语言、Protégé 编辑工具等来创建。

由于本文是以外文网络数据库资源为对象，涉及概念广泛，无法个人自建，因此考虑结合比较通用的经过专家长期实践总结的传统文献组织方法来解决本体库的概念确定问题。LCC 和 LCSH 在西文文献著录和组织中应用的相当广泛，并具有描述特定学科领域知识、有等级结构、存在概念及概念之间的关系等特点，因此可选择 LCC 和 LCSH 参考构建本体。当然真正完整意义上的本体库应该在参考很多专业领域词表、同义词表、反义词表及文档语义分析等的基础上来构建，是一项相当复杂的工程，由于个人能力与篇幅所限，适当简化，下面仅选择 LCSH 为代表，举例阐述本体的创建。

国会图书馆主题词表 LCSH 是至今为止最常用并被广为接受的通用主题词汇表。首先分析表中的关系后，简单定义四种本体属性关系：hasSynonymousTerm、hasBorderTerm、hasNarrowTerm、hasOtherRelationTerm。其中，hasSynonymousTerm 主要用于表示词表中具有同义和近义词性质的 USE、UF 关系，该关系具有对称性；hasBorderTerm 和 hasNarrowTerm 表示 BT 和 NT 关系，具有传递性；hasOtherRelationTerm 表示除了同义、近义、分级关系之外的相关关系。下面是以主题词表中的 34 个 management 有关的主题词为例，在 Protégé 环境下，先后创建了主题本体概念和它们的属性关系，如图 5 所示。

Racer 支持 Protégé 本体开发环境，具有支持 RDF 查询语言、OWL 推理等功能。当 Protégé 与 Racer 连接上，可使用 Protégé 中 OWL 下的“check consistency”、“classify taxonomy”、“compute inferred types”功能，对概念的一致性、同推理机的同步等进行检查，并计算等价类、推断层次等。

对于本体的查询和推理，不仅可在 Protégé 环境下，使用 SPARQL 语言进行查询、连接 Racer 进行推理，同时还可以直接在 Racer 中使用 nRQL（New RacerPro Query Language）实现 RDF、OWL 的查询和推理。

本小节仅是对本体构建的一点尝试，完整的本体库构建还需完成大量工作，对概念、类、属性、关系等有更精确的划分。只有建立了比较完整或完善的本体库，才更有利于使用 Protégé 和 Racer 来实现系统的语义查询和推理。

4.6　跨库检索

本框架选择目前比较成熟的 HTTP + Java 的方式来设计跨库检索部分，比较符合目前外文网络数据库现状。跨库检索模块主要功能是将经过语义处理后的用户查询进行分解，利用 HTTP 协议和 Java 多线程技术，并发向各网络数据库发出检索、查询请求，并将返回的检索结果转发给结果整合模块。

在 Java 语言中，与网络操作有关的类和接口封装在 Java.Net 包中，其中与 HTTP 协议相关的一些类和接口，封装了 HTTP 协议的一些细节。由于外文网络数据库具有异构、

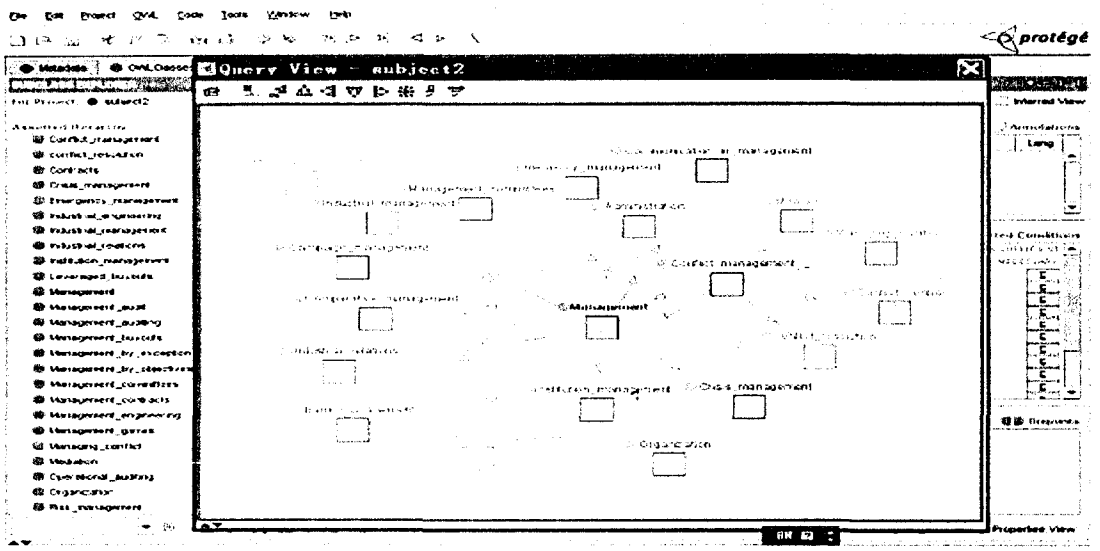


图5 Management 相关概念之间的关系图

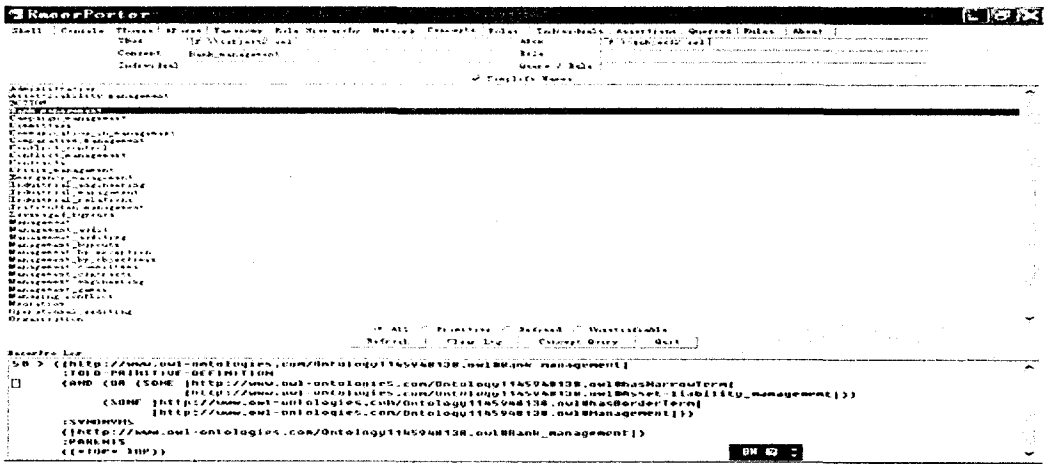


图6 Racer 中“Bank Management”的概念查询图

商业性等特点，要想实现跨库检索，首先要了解各数据库的检索格式等信息。Java.net.URL 类为使用 URL 从 Internet 上获取信息提供了一个简洁的用户编程接口。因此，系统可以利用Java.net.URL 类来获取网络数据库的元数据及检索参数等信息。首先系统通过向网络数据库的检索系统发送 HTTP 请求，然后从结果分析得到其检索格式或参数，再按照该网络数据库的检索格式发送新的 HTTP 请求，实现检索。

为访问 URL 指向的网络资源，首先要构造一个 URL 对象，如：

URL url1 = new URL(“http://”);

然后通过 URLConnection 类中的 openConnection ()方法尝试连接 URL 指向的网络资源，如连接成功，会返回一个 URLConnection 对象，失败会产生 IOException。如果连接建立，则可使用 getInputStream ()和 getOutputStream ()方法来得到连接的输入和输出数据流，从而可以进一步对获得的数据进行分析，得到系统需要的格式和参数。

以上的分析是系统对单个网络数据库实现检索的过程，跨库检索要实现同时检索多个数据库，需要应用 Java 的多线程技术来实现，具体可以使用Java.lang.Runnable 接口，先是为用户选择的要检索的每个数据库创建线程，然后使用加锁和同步控制，实现多线程检索。

5 结 语

本文针对外文网络数据库综合利用问题，提出基于本体的跨库检索系统的层次结构和集成框架，并阐述了系统框架的功能设计与模块划分、工作原理与流程、关键技术分析与实现思路。本文的系统框架有自身的特点，与普通的检索系统有所区别：

5.1 跨库检索系统，针对的是分布的、独立的、异构的、商业网络数据库资源，要解决的是它们的统一检索问题；

5.2 基于本体的，不同于一般的跨库检索系统，它为检索增加了语义性，增加了检索系统的联想空间，有利于提高检索精确性，从而更好地满足读者的检索需求。同时由于基于本体，系统中资源组织、索引和导航功能也能得到更好地发挥。

5.3 综合考虑电子资源利用问题的、功能集成的系统框架，不仅仅是单纯的检索系统。本框架包含了资源的组织、检索、保存等几个方面。

5.4 语义网在数字图书馆方面的一个应用尝试，采用了 RDF、OWL、Protégé、racer 等技术、语言和工具。

由于本文的系统还仅是一个框架，尚有很多细节需要更深入地研究，而且有些想法也还不成熟，还需要进一步探讨和改进。

参 考 文 献

[1] 张秋. 基于 Web 的数字图书馆跨库检索系统的比较研究 [J]. 图书情报工作，2005，(4)：88－91.
[2] 王效岳，王志玲. 国内外异构数据库统一检索系统的比较研究 [J]. 情报杂志，2005，(12)：116－118.
[3] 张艳华. 统一检索系统的对比分析 [J]. 情报杂志，2005，(6)：71－72.
[4] 朱虎明. 数字图书馆中统一检索系统的研究与开发 [学位论文] [D]. 西安：西安电子科技大学，2004.
[5] W3C [EB]. http://www.w3.org (Accessed Sept. 4, 2006)
[6] Dublin Core Metadata Initiative [EB]. http://dublincore.org (Accessed Mar. 2, 2006)
[7] 中国万维网联盟 [EB]. http://bbs.w3china.org/index.asp (Accessed Sept. 4, 2006)
[8] 宋炜，张铭. 语义网简明教程 [M]. 北京：高等教育出版社，2004.
[9] 胡娟. 数据库统一检索平台的功能比较 [J]. 现代情报，2005，(4)：174－177.
[10] 韩炳黎. 数字图书馆统一检索研究及开发 [J]. 现代情报，2006，(3)：25－27.
[11] 金国强，豆洪青. 基于 CORBA 技术的图书馆多库统一检索系统 [J]. 情报科学，2005，(9)：1372－1375.
[12] 李海军. 跨库检索系统的研究与开发 [学位论文] [D]. 西安：西安电子科技大学，2005.
[13] Jena－A Semantic Web Framework for Java [EB]. http://jena.sourceforger.net/index.html (Accessed Mar. 2, 2006)
[14] Jena 简介 [EB]. http://www.ibm.com/developerworks/cn/java/j－jena/index.html (Accessed Mar. 2, 2006)
[15] 陈琮. 基于 Jena 的本体检索模型设计与实现 [学位论文] [D]. 武汉：武汉大学，2005.
[16] 秦春秀. 基于本体的 Web 信息检索系统及其关键技术研究 [学位论文] [D]. 西安：电子科技大学，2005.
[17] 殷兆麟，周智仁，范宝德，等. Java 网络应用编程 [M]. 北京：高等教育出版社，2004.
[18] 廖义奎. Java Web 开发之 Struts 编程基础与实例精讲 [M]. 北京：中国电力出版社，2006.
[19] 郭瑞华，张玉莉. 语义 Web 上 DC 元数据的描述及抽取技术 [J]. 现代情报，2005，(6)：212－214.
[20] 陶兰，杨睿，陈冲，等. 面向语义 Web 的 RDF 数据处理和应用 [J]. 深圳大学学报：理工版，2005，(4)：330－333.
[21] 李玲，唐胜群. 知识网格中基于 RDF 的知识表示技术和应用 [J]. 计算机应用研究，2005，(12)：223－225，229.
[22] 张蓉，申德荣，于戈. Ontology 在异构数据库集成中的应用 [J]. 计算机工程，2004，(12)：29－31.
[23] 刘海滨，李冠宇，刘发军. 基于 Ontology 的信息集成研究综述 [J]. 计算机工程与应用，2005，(25)：159－161.
[24] 艾伟. 本体的构造及其应用研究 [学位论文] [D]. 武汉：武汉理工大学，2005.
[25] 美国国会图书馆 [EB]. http://www.loc.gov (Accessed Mar. 2, 2006).
[26] Protégé [EB]. http://protege.stanford.edu (Accessed Apr. 8, 2006)
[27] Racer [EB]. https://www.racer－systems.com (Accessed Apr. 8, 2006)

(上接第 169 页)

度慢，软硬件故障多，易死机。读者经常恶意修改系统配置，破解还原卡等。管理维护困难。

经过比较分析，广东省立中山图书馆决定使用选择升腾公司的 C－3000 型终端机代替原来使用的 PC 机。该机采用 WINCE 3.0 平台，支持 RDP、ICA、TELNET 等多种协议。自带终端客户端与服务器端的终端服务配合良好。

现有 15 台电子阅读用机，使用一台终端服务器。实际应用中，服务器系统资源占用率都比较低。终端产生的网络流量低，大约为 PC 机所需流量的1/10，给网络带宽带来的影响几乎可以忽略不计。

投入运行后的使用效果很好。画面显示流畅，操作感觉不出延迟，色彩效果好。运行也十分稳定，管理人员的维护工作量大大降低。

5 注意事项

5.1 系统安全设置与某些浏览器的特殊要求可能会有矛盾。例如对 C 盘根目录要求有写权限，程序安装目录要求写权限和取消错误报告功能等，需要适当的进行调整。

5.2 终端机用户访问范围应有所限定（限于局域网），避免读者使用终端机进行一些不必要的操作，占用机位，影响正常使用的读者。

5.3 各类电子资源的查询往往基于 web 页面，我们需要在 IE 浏览器的选项中选中“INTERNET 属性——高级——强制屏幕外合成”一项，以减少终端机浏览网页时产生的网页闪烁现象。

6 结束语

Windows 终端服务还在不断发展当中，微软新一代服务器操作系统 Windows 2008 也将要推出市场，它所带来的进步必将为 WBT 在图书馆信息化利用方面带来更广阔的应用前景。

参 考 文 献

[1] 王晓东，王硕. Windows 终端在图书馆中的应用 [J]. 大学图书馆学报，2002，(5)：34－35，39.
[2] 周东. 配置安全、高效的图书馆终端服务器 [J]. 现代情报，2004，(6)：73－76.
[3] 陈丕庞，方岩雄. Windows 终端服务系统在高校图书馆应用方案的设计 [J]. 图书馆论坛，2006，(2)：130－132.
[4] 杨子伍，陈如好. 终端技术在图书馆应用系统中的优势 [J]. 图书情报工作，2004，(4)：76－77.