

基于公开Web Services API的外文图书采访系统的设计与实现*

□ 俞小怡 刘凡儒 金玉玲 / 大连理工大学图书馆 大连 116085

摘要: 利用乐天、谷歌和亚马逊等知名网站所提供的公开Web Services API接口, 可以开发基于Web Services API的外文图书采访系统, 提供采集处理以日文、英文为主的多语种外文图书书目信息。文章以日本乐天网上商城提供的API为例, 研究基于网上外文图书的采访流程和策略, 介绍数据采集转换的工作原理和元数据获取的核心技术。

关键词: WEB Services API, 外文图书, 乐天, 元数据, JSON数据接口

DOI: 10.3772/j.issn.1673-2286.2011.10.006

1 引言

近年来, 各大图书馆采用的图书采购方式是通过公开招标方式, 确定固定的图书供应商作为合作伙伴。图书供应商提供的有针对性图书可以涵盖图书馆所需的大部分学科的采集范围, 并且必须支持MARC标准格式的编目信息。图书采访馆员从供应商提供的图书征订目录中进一步精选出本馆预算范围内的有价值图书。虽然图书征订目录中的信息量很大, 但是由于有些特殊专业的需求存在信息不对称的缘故, 采访馆员受限于图书征订目录的筛选, 还不能满足读者的需要。对于大量的国外出版物, 由于出版发行流程的关系, 还存在着不同程度的滞后。特别是我校的基础研究一直处于国际领先地位的化工、环生等学科的出版物, 对专业性和实时性的要求更高, 对服务的可用性与期望值也提出更高要求。

随着互联网及电子商务的蓬勃发展, 读者的荐购书籍中有相当一部分是从互联网上获取的图书信息。为此, 我们需要在网上书店中迅速地找到读者所需要的图书。而且, 网上书店的基本图书信息中除了征订目录中基本的出版信息外, 经常包含内容概要、简明目录、章节选读和作者延伸等更详尽的资料, 保证采访馆员更准确地选取符合馆藏原则和更好地为读者服

务的图书, 并且避免新书的漏订和重订等缺失。利用乐天、谷歌和亚马逊等国际知名网站上图书供应商公布的公开Web Services API接口, 我们设计开发了基于Web Services API的外文图书采访系统, 实现了对日文、英文等多语种外文图书书目信息的及时采集和MARC格式的转换, 为外文图书的征订工作提供一种新型的检索编目途径, 提升了图书馆采访的质量, 提高了为读者服务的响应速度。

2 基于网上外文图书的采访流程和策略

在乐天、谷歌和亚马逊等网上书店中, 收藏有数以百万计的不断更新的图书信息, 完全动态收割所有的书目信息在技术上是可行的, 但在图书馆自身的存储容量和采访馆员的工作强度上看都是不可行的。所以, 我们的网上外文图书采访系统是作为现行书商指定书目系统的补充系统并行存在的。其设计的主要功能有: (1) 采集购入读者荐购的非征订目录里的书籍; (2) 按学科分类定向收集指定专业的出版物; (3) 按权威作者群分类收集指定科研工作者的出版物; (4) 和指定图书有一定关联程度的出版物。采访流程和策略如下:

* 本文系教育部人文社会科学研究规划基金项目“期刊文献知识元挖掘及知识库构建研究”(项目编号: 09YJA870005)研究成果之一; “中央高校基本科研业务费专项资金资助”项目“研究型大学图书馆馆藏文献资源结构建设研究”(项目编号: DUT10RW409)研究成果之一。

2.1 建立外文图书信息库

这部分主要完成从网上书店提取图书信息并和既存的书商征订书目进行比照,只对比照不一致的图书信息进行收录。收录图书信息时对书名和摘要目录中出现的关键词进行提取。关键词的提取过程中要剔除一些常见的助词、介词,并统计每个关键词出现的频度,对于外语中出现的一些新的合成词、缩写等随时添加到系统词典中以扩展补充现有词库,这种动态生成的词典在查找新书方面有重要的意义,新词的大量和反复出现往往预示着一个全新领域的形成。定期检查和统计出现的新词,以及一定时间内出版书籍的新关键词出现频度,可以很大程度上为采访工作提供热门书籍和热门学科的方向性。

对于图书的重复判定除了使用最常用的ISBN判定外,还结合关键字地匹配法进行判定。对于虽然ISBN号不同,但书名作者相同或高度相似,出现的关键词也高度匹配的情况下,也认为是相同图书,存入别库,生成提示信息,必要时可以人工确认。

图书信息库的建立过程是全部由网络爬虫自动完成的,不需要人工干预,可以让整个系统每日夜间运行,动态更新信息库,有效解决图书查重问题,并保证信息的新鲜度和有效性。

2.2 图书的甄别选择

本系统作为传统图书采访方式的补充和延伸,在互联网上可以有效采访到某些全新领域的图书、查找和指定图书有一定相关度的图书以及科研工作者的指定出版物。除了提供传统的按中图法、科图法、ISBN、作者、关键字等检索图书的途径外,还提供按学科分类、图书的关联度、权威作者群、科研合作关系等进行得检索。使用雅虎的分词技术提取关键词,对于共同作者、关键字一致程度等进行评估,并给出其关联度,借助关联度可以检索有高度关联性的图书资料;根据系统词典的增量来查找最近出现的一些新词和新的研究领域。

通过上述无须人工干预的信息采集策略,从浩瀚的书海中提取精选的书籍,再通过采访馆员的少量确认工作,可以及时快速地把这些新领域、权威性、具备馆藏价值和符合馆藏标准的图书采购入馆。

2.3 网上外文图书采访系统与现行图书采访系统的对接

目前,国内外文图书采购方式主要有:根据各合作书商提供的书目数据选择订购;根据各合作书商来本馆搞样书书展现场看样选购;教师推荐图书的零星订购等。这种采购方式对书商特定供货的依赖性和局限性较大,选购水准制约于书商的道德水准和技术水准,直接影响到外文图书采购的质量和速度。而通过本系统筛选甄别的书目信息可以作为外文图书采购人员辅助订购的重要书目依据。

为进一步降低采访馆员的工作强度,本系统将甄选出的已查重书目信息导入现行采访系统中。实现网上外文图书采访系统与现行图书采访系统的数据无缝对接后,可以有效改善外文图书订购书目数据对供货商的依赖关系,大幅提高图书馆外文图书藏书质量。将甄选后的最新书目信息利用数据接口导入现行采访系统当中(笔者所在馆使用的是“汇文文献信息服务系统Libsys4.5-采访”),生成MARC标准格式的编目信息,使得采访人员无须增加额外的工作量,就可以利用本系统方便地进行批查重、选购和生成订单。利用相似的数据接口,同样也可以将搜集到的书目信息方便地导入国内其他采访系统中,为外文图书采访人员提供方便、快捷的书目信息服务,精确有效地采访到急需的外文图书,准确定位扩大外文图书采购的书目范围。

通过本系统确定待采购的图书后,可以通过网上书店购买,有一部分网上书店提供完整的Web API支持接口,可以充分利用该功能,通过本系统的相关模块直接进行图书订购,以简化工作流程和提高效率。而对于只能选择传统的人机交互方式订购图书的书店以及图书订购后的物流配送付款流程等在以往的论文中已有详细讨论,在此不再赘述。

3 基于Web Services API采集图书元数据的原理

日前,乐天、谷歌和亚马逊等各大互联网巨头纷纷公布API接口,允许传统的图书馆书目管理系统能够与网络图书搜索书目记录之间实现无缝深度对接,使得数据的传递和转换更加灵活。我们以近邻的日本乐天网上商城提供的Web Services API为例来说明采集图

书元数据信息的原理。乐天为了方便广大分销商和读者检索图书元数据,提供REST和JSON两种形式的接口,其中JSON形式接口的URL链接如下:

`http://api.rakuten.co.jp/rws/3.0/json?[parameter]=[value]...`

其中parameter用来有选择地提取所采集图书的元数据信息,既可以同时提取多项,也可以只提取其中一项,如书名、作者、出版社名(来源)、ISBN和书目分类等。按照采编的需要,除了使用书目分类作为关键字来分门别类地获取该类的全部图书元数据信息外,也能够使用ISBN编码对既存图书元数据进行校验更新。parameter中还必须包含一个Developer ID,即开发者ID,这个ID是到乐天的开发网注册后申请的一个32位的文字列,主要为了防止一些攻击性为目的非法访问。另外operation和version两个参数也是必选项,前者是指定海量Web Services API中的具体访问对象,后者是指所访问Web Services API的版本号。

把上面的URL中的parameter和value部分导入适当的值后,向乐天的服务器提交一个请求,Web Services服务器会以JSON格式返回请求的图书元数据信息。

JSON格式是一种由JavaScript派生而来的用于数据交换的数据记述形式,2006年7月被制定为RFC 4627标准。JSON格式可以描述为一个集合,这个集合的每个元素都由一个项目名和它的值配对组成。下面是一个简单的JSON对象的例子:

```
{books [{"title": "Web 安全管理技术", "price": 35}, {"title": "网站构筑入门", "price": 20},]}
```

JSON对象的特征是直观易读,接近自然语言,并能够被C、C++、C#、Java、Perl、Python等大多数编程语言所处理。尤其是随着Web2.0和Ajax的兴起,很多Web Services API都放弃了传统的XML形式,而采用这种更简单明了的JSON形式,以便于使用JavaScript进行数据处理。

乐天的Web Services API可返回的检索项有50多项

内容,其中有一些用在网上销售的字段,比如在库本数、送货费用等可弃之不用。选择时保留标题、副标题、系列标题、作者、出版社、ISBN编号、定价、内容概要、封面照片、目次页、章节选读和其他相关情报即可。

4 元数据获取的核心技术剖析

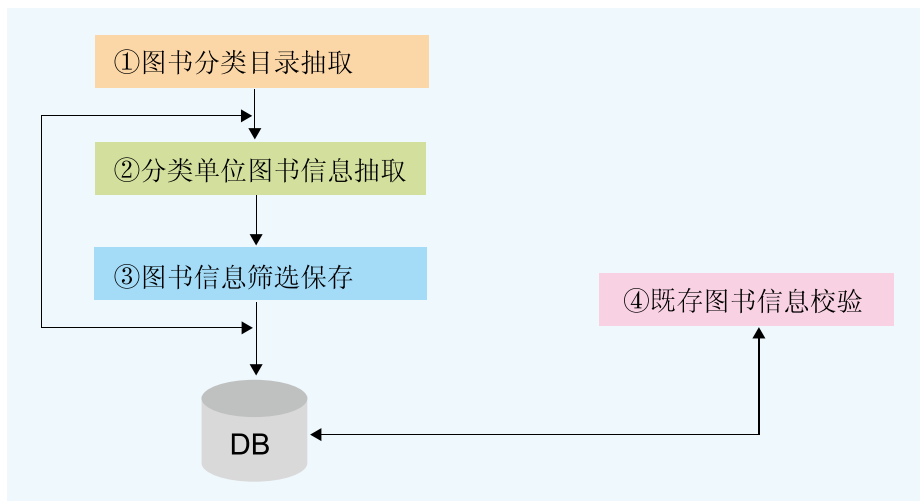


图1 数据获取子系统的核心模块构造图

数据获取子系统的核心模块构造如图1所示。

(1) 乐天的书目分类编码方式是规则树形结构,共由四级组成。每一级由固定三位的数字来表示。由此,9位数字可以完整表示第三级的一个分类,12位数字可以完整表示第四级的一个分类。例如位于第三级的分类“001004008”代表日本小说,位于第四级的分类“001004008005”代表作者名是“な行”开始的日本小说等。作为检索参数的分类编码可以是第一到第四级的任意一种分类,对应的检索到的结果是该分类下的所有图书。详细的图书分类编码可以通过乐天提供的“乐天图书分类检索API(Books Genre Search)”来取得。如图2所示。

图书分类目录通过Web Services API(Books Genre Search)从乐天取得,取得的深度可以通过参数来控制。取得了指定深度的目录后,删减掉一些收录对象外的部分,保存必要的内容,在后续的②、③处理中以分类为单位循环取得相应的分类内图书元数据。

(2) 图书元数据抽取部分的核心代码:

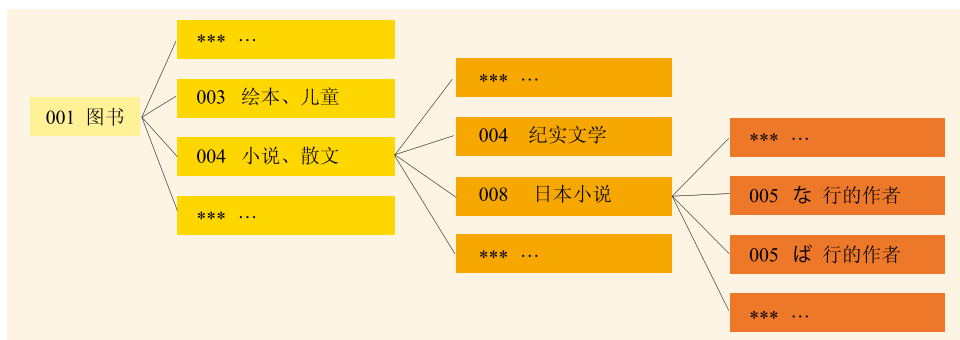


图2 乐天图书分类编码示意图

//Web Services 参数指定, 001004008在这里是示例, 实际程序中用变量替换

```
String book_url = "http://API.rakuten.co.jp/rws/3.0/
json?";
```

```
book_url = book_url + "operation=BooksBookSearch&
version=2010-03-18";
```

```
book_url = book_url + "&developerId=208dadff0c123
45678901234569095d7";
```

```
book_url = book_url + "&booksGenreId=001004008&
sort=%2BitemPrice";
```

```
//生成Web Services url
```

```
URL url = new URL(API_url);
```

```
//建立流输入, 以UTF8字符集取得元数据
```

```
Object content = url.getContent();
```

```
if (content instanceof InputStream) {
```

```
    BufferedReader reader = new BufferedReader(
        new InputStreamReader( (InputStream)
content, "UTF8"));
```

```
    while ( (line = reader.readLine()) != null) {
        strcontent = strcontent + line;
    }
```

```
    reader.close();
```

```
}
```

// 使用jsonic把取得的图书元数据从JSON转变成POJO (预先定义好的JAVA类)

```
RakutenBookAPI rakutenAPI =JSON.
decode(strcontent,
```

```
RakutenBookAPI.class);
```

(3) 对于取得的图书元数据一般不能直接用来供读者查询, 而是需要根据预先设定的一些关键字进行必要的

过滤, 如果存在一些不适宜公开或者不必要公开的内容时, 要把这些书存放在另外的数据库表中, 留待人工确认或者直接删除。

④图书元数据校验部分负责对书库内既存的图书信息进行定期扫描, 随时更新既存的图书信息,

这部分也是通过乐天的这个图书检索API来完成的, 不过这里指定的参数不是分类ID, 而是ISBN编码, 可以直接定位到特定的图书上。

除了元数据取得之外, 本系统还应用了关键字提取技术。这里仍然以日文图书为例, 使用雅虎的分词技术提取核心代码如下:

//通过yahoo!的日文分词API提取日文关键词, input-text是被解析的文本

```
String web_base_service_str =
```

```
" http://jlp.yahooapis.jp/MAService/V1/parse?";
```

```
web_base_service_str += "appid=myapi-
id&results=ma&sentence=";
```

```
String web_service_str = web_base_service_str +
java.net.URLEncoder.encode(input-text, "UTF8");
```

```
DocumentBuilderFactory dbfactory =
DocumentBuilderFactory.newInstance();
```

```
DocumentBuilder builder = dbfactory.
newDocumentBuilder();
```

```
Document doc = builder.parse(web_service_str);
```

```
//分解后的关键词列表被放入par_list中
```

```
Element root = doc.getDocumentElement();
```

```
NodeList par_list = root.getElementsByTagName("wor
d");
```

输入的文本input-text被分解成一个个关键词放在par_list中, 除了关键词本身外, 词性、读音等信息也包含在其中, 可以根据这些信息来把一些常见的助词、介词等剔除, 并统计每个关键词的出现频度。

5 Web Services API元数据获取的延伸

上文介绍了通过Web Services API获取外文图书书目信息的方法。随着Web2.0和Ajax的兴起,网上公开了大量的Web Services API,除了大家熟悉的一些关键字搜索、地图搜索和一些网上商城提供的商品搜索外,还有其他一些非常实用的API可以被图书馆所利用。例如,ASIAL公司提供的用于把HTML文件和WEB网页转换成PDF文件的API,通常的WEB网页中由于存在各种动态元素,随着时间的变迁WEB网页会变得面目全非,而转换成PDF文件就可以把某一时间点的网页非常简单地保存下来,这个对于图书馆信息采集是一个非常实用的功能。日本EAST公司提供的词典API也是一项非常实用的功能,该API提供了英日、日英、WIKI等三种辞典,共计约40万词汇可供查询,利用这个API完全可以构建自己所需要的各种不同范围的辞典和词库等。另外像谷歌、雅虎等提供的网页翻译、文档分析等API对于信息采集也都非常实用。

6 结语

互联网的开放性把海量的网络资源和超载的数据信息推送到我们面前,但是对于大多数非计算机专业的图书馆员来说,其中包含的许多图书信息难以采集和利用,也不能兼容和顺利导入到我们现有的图书集成管理系统。其中,外文图书书目信息的获取工作有时会遇到数据更新不及时和征订目录数据不能满足科研人员的实际工作需要的瓶颈。通过我们开发的基于Web Services API的外文图书采访系统作为现有图书馆管理采访系统的有力辅助和参考系统,可以把从海量网络图书资源中甄选出的有价值数据及时提取并融合到现在图书馆集成信息平台上,更好地满足各院系师生的前沿科研工作的需要,这也是走向图书馆2.0的发展方向之一。

参考文献

- [1] 王雄,褚伟.一种面向对象的Web Service API框架设计[J].计算机技术与发展,2009(8):54-57,61.
- [2] 乐天ブックス书籍检索API[EB/OL]. [2011-05-24]. <http://webservice.rakuten.co.jp/API/booksbooksearch/>.
- [3] Amazon[EB/OL]. [2011-05-24]. <http://www.amazon.co.jp/gp/feature.html?ie=UTF8&docId=451209>.
- [4] Google Book Search APIs[EB/OL]. [2010-05-24]. <http://code.google.com/intl/en/APIs/books/docs/getting-started.html>.
- [5] 吴广印.基于Web Service构架的资源共享技术研究与实现[J].情报学报,2007,26(6):851-857.
- [6] 顾宁,刘家茂,柴晓路.Web Service原理与研发实践[M].北京:机械工业出版社,2007.
- [7] 和泉宪明,幸岛明夫,车谷浩一,等.多粒度ロジトリに基づくWeb サービスのビジネスモデル驱动連携[J].信学技报,2003,2002(32):43-48.
- [8] 串户洋平,石井健一,山内宽己,等.Web サービスアプリケーションのソフトウェアメトリクスに関する考察[J].信学技报,2004,2003(316),113-118.
- [9] BEYER D, CHAKRABARTI A, HENZINGER T A. Web service interfaces [C]//Proceedings of the Fifteenth International World Wide Web Conference (WWW 2005). Chiba: ACM Press,2005.
- [10] 王雄.一种面向对象的Web Service API框架设计与实现[D].合肥工业大学,2009.
- [11] ASIAL HTML2PDF[EB/OL]. [2011-05-24]. <http://www.html2pdf.biz/api.php>.
- [12] イースト辞书Webサービス[EB/OL]. [2011-05-24]. <http://www.est.co.jp/>.

作者简介

俞小怡 (1971-), 女, 馆员, 研究方向: 数字图书馆。E-mail: yuxiaoyi@dlut.edu.cn
刘凡儒 (1971-), 女, 副研究馆员, 研究方向: 图书采购。
金玉玲 (1961-), 女, 副馆长, 研究馆员, 研究方向: 数字图书馆。

Design and Implementation of Foreign Language Books Collection and Mining System Based on Web Services API

Yu Xiaoyi, Liu Fanru, Jin Yuling / Dalian University of Technology Library, Dalian, 116085

Abstract: Using the open Web Services interfaces provided by Rakuten, Google and Amazon, etc., we developed Foreign Books Collection and Mining System based on Web Services API, and supplied a new way of collection and processing in Japanese, English-based multilingual bibliography of foreign language books. Taking Japan Rakuten online mall as an example, we designed the operation processes and policies of online foreign language books collection and mining, and described the working principle and the core technology of metadata mining.

Keywords: Web Services API, Foreign language books, Rakuten, Metadata, JSON Data interface

(收稿日期: 2011-05-19)