基于 CURL 的网络学习资源聚合系统开发①

唐四薪, 谭晓兰, 李 浪

(衡阳师范学院 计算机科学与技术学院, 衡阳 421002)

摘 要: 设计并实现了一个基于 CURL 的网络学习资源聚合系统. 利用 CURL 多线程函数将数据使用 GET 或 POST 方式同时发送给所有资源网站, 对资源网站返回的 HTML 代码进行统一编码, 使用正则表达式提取返回代码的搜索结果区域, 使用 PHP DOM 操作类修正代码中的图像和链接的 URL 地址, 再将所有返回代码载入到同一个页面中, 并使用瀑布流模型实现逐段加载.

关键词: CURL; 学习资源; 聚合

Development of Web Learning Resources Polymerization System Based on CURL

TANG Si-Xin, TAN Xiao-Lan, LI Lang

(School of Computer Science and Technology, Hengyang Normal University, Hengyang 421002, China)

Abstract: We design and implement a polymerization system for web learning resources based on CURL. Using the CURL multi-threading functions, the system can send data to all resource websites at the same time through GET or POST method. And then the system can unify all returned HTML codes from resources websites, and use regular expressions to extract the search results area of the returned codes. It uses PHP DOM manipulation class to fix the image and the URL addresses of links in the code, and then loads all returned code into the same page. Thus it can realize piecewise loading by using waterfall flow model.

Key words: CURL; learning resources; polymerization

随着 Internet 上教学资源的丰富, 很多教师(尤其是高校教师)都喜欢在网上搜索课件等教学资源, 以帮助自己备课, 很多学生也需要搜索教学资源进行自学或拓展性学习. 但如果使用百度等通用搜索引擎搜索教学课件, 一般只能搜索到百度文库(或其他文库)及一些零散的课件, 这些课件大都是由网友上传, 其权威性、系统性和数量有时难以满足备课或学习的需要.

另一方面, 近年来我国的各大出版社都注重加强教材书籍的配套资源建设, 许多出版社网站上已经积累了丰富的教材配套教学资源. 这类教学资源一般由书籍作者提供, 比较系统, 但出版社网站上海量的教学资源在百度等搜索引擎上却几乎搜索不到.

如果用户分别去访问每家出版社网站, 再逐个搜索每个网站上的教学资源, 那将是一件非常繁琐的事情. 为了让用户方便地同时搜索很多家出版社、精品课等网站上的学习资源, 我们开发了《多线程网络学习资源聚合系统 V1.0》(网址为 www.keso8.top), 该系统采用 CURL 多线程访问模式, 能同时搜索很多家网站上的学习资源, 并将所有资源聚合到一个界面中供用户浏览下载.

1 统一查询技术理论与实现

WWW 是一个庞大、异构、分布式的文档集合, 里面存在着各种各样的文档, 主要有以 XML、JSON 为代表的半结构化数据, 和以 HTML、文本文件为代表的无结构化数据. 对异构数字资源进行统一查询的目的是为了克服各个资源子空间之间的障碍, 使得分布式资源通过统一查询系统处理成异构虚拟资源实体[1]. 尽管目前有少量网站提供了输出 JSON 或 XML

① 基金项目:湖南省教育科学“十二五”规划一般资助课题(XJK013BXX005);衡阳市科技计划(2016KJ02)

收稿时间:2016-09-19;收到修改稿时间:2016-10-31 [doi: 10.15888/j.cnki.csa.005788]

半结构化数据的功能，但大多数网站仍只提供了 HTML 文档.

对于只提供了 HTML 文档的出版社网站群进行统一资源查询有三种方案：①开发一种接口软件直接查询各个出版社网站的数据库，这种方式开发难度最低，但需要出版社提供目标数据库的接口，出于安全性和数据保密性等考虑，出版社一般是不愿意提供的；②利用出版社的搜索接口，进行统一查询，将查询得到的数据进行结构化处理之后存储到本地数据库中，再对本地数据库进行查询，这种方式的查询速度较快，但由于复制了出版社网站中的内容到本地存储中，会引起版权争议问题；③进行统一查询，将查询得到的数据进行结构化处理后直接显示在页面中. 这种方式由于没有复制内容到本地，不会引起版权问题. 本系统采用第③种方式开发.

CURL(Client URL)是由瑞典 curl 组织开发的用于获取远程文件信息或传输文件的工具[2]，它支持很多协议，如 HTTP、FTP 和 Telnet 等，PHP 也支持 cURL 库——libcurl, CURL 支持命令行方式和 PHP 脚本代码两种工作模式. 一般利用 cURL 来抓取远程网页，与 Ajax 等技术相比，其优势在于：①能发送 GET 或 POST 数据给远程网页；②能实现多线程任务式抓取网页(即同时抓取多个网页). ③ 抓取速度快，对于 2000 个 HTTP 请求文件, CURL 每分钟可打开 2000 次[3].

2 系统的设计与实现

目前，大多数出版社网站都提供了搜索书籍配套资源的表单. 当用户单击搜索按钮后，就会转到处理表单的页面，将搜索结果呈现给用户. 本系统提供一个表单供用户输入搜索关键词，然后利用 CURL 直接发送关键词数据给各个网站的搜索处理页，最后获取返回的搜索结果页面代码并将其聚合到本系统搜索结果页面中，系统的工作流程如图 1 所示.

2.1 数据库的设计

为了使采集的资源网站群具有可扩展性，本系统将所有要采集的出版社网站的网址等信息保存在一个表中，该表的结构如下：

sites(id, name, url, charset, pregmatch, valid, postdata, imgsrcp, asrcp, sort, descp)

各字段的含义对应如下：

sites(序号，网站名，网站搜索入口 URL，网站的编码类型，内容区域的正则表达式匹配码，是否有效，存储 post 数据，图像 URL 地址前缀，超链接 URL 地址前缀，排序，描述).

如果要新增采集的网站，只需将新网站的信息添加到 sites 表中即可.

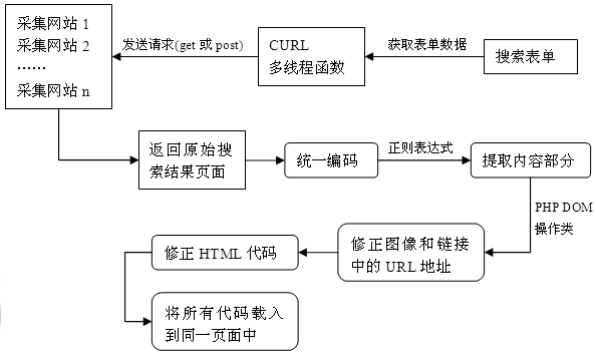

图 1　系统工作流程

2.2 关键词编码转换

在发送关键词数据给表单查询页之前，需要先进行 URL 编码以保证数据被正确接收，一般使用 urlencode()函数将关键词转换为 URL 编码. 另外，由于各个网站网页采用的编码不同(本系统使用的页面编码是 gb2312)，因此如果采集的网站网页编码是 utf-8 类型的，还需要用 iconv 函数先将关键词转换为 utf-8 编码. 代码如下：

```
$qstrs[$i]= urlencode(iconv("gb2312","utf-8",$key))
```

2.3 发送 CURL 请求

CURL 支持多线程技术，即能同时发送多个请求给不同的网站，并返回结果数据[4]. 与 Ajax 技术相比，CURL 的多线程技术同时访问大量网站所耗费的时间只是其中访问最慢的网站所耗费的时间，而 Ajax 等其他技术是所有网站打开时间之和[5].

对于不同的网站来说，其搜索表单的提交方式不同，有些是 GET，有些是 POST，本系统将 post 方式提交的数据存储在 postdata 字段中，然后判断 postdata 字段是否为空，如果为空则以 GET 方式提交 CURL 请求，否则以 POST 方式提交 CURL 请求. 使用 CURL 多线程函数同步发送多个 HTTP 请求的代码如下：

```
$mh = curl_multi_init();
foreach ($urls as $i => $url) {        //对每个采集的 URL 网址
 $conn[$i]=curl_init($url);
```

```
if($postd[$i]!=''){          //如果 post 数据不为空,则发送 post 请求
$key1 = iconv("gb2312", "utf-8",$key);
$postdd=$postd[$i].$key1;
$fields=passtr($postdd);
curl_setopt ($conn[$i], CURLOPT_URL, $url );
curl_setopt ($conn[$i], CURLOPT_POST, 2);
curl_setopt ($conn[$i], CURLOPT_HEADER, 0 );
curl_setopt ($conn[$i], CURLOPT_POSTFIELDS, $fields);  }
curl_setopt($conn[$i],CURLOPT_RETURNTRANSFER,1);
curl_multi_add_handle($mh,$conn[$i]); // 添加 cURL 资源句柄
}
do { $n=curl_multi_exec($mh,$active); }
while ($active);
foreach ($urls as $i => $url) {
$res[$i]=curl_multi_getcontent($conn[$i]);    // 将返回的内容保存在数组元素中
……//统一编码为 gb2312
$time[$i]=curl_getinfo($conn[$i],CURLINFO_CONNECT_TIME);
curl_close($conn[$i]);          }
```

2.4 对返回代码进行统一编码

CURL 发送关键词数据给各网站后, 各个网站会返回搜索结果页面的 HTML 代码, 由于各个网站的编码不同, 返回的 HTML 代码编码也不同, 必须统一编码才不会出现乱码现象. 本系统页面采用 gb2312 编码, 因此将所有返回的 HTML 代码统一为 gb2312 编码, 关键代码如下:

```
if($charset[$i]!='gb2312')
$res[$i]=iconv("utf-8","gb2312.IGNORE",$res[$i]);
//统一编码为 gb2312
```

2.5 提取内容部分

由于各个网站返回的搜索结果是整个网页的 HTML 代码, 而本系统需要载入的只是其中的搜索结果列表区域. 因此必须提取其中的结果内容部分. 本系统首先人工找到搜索结果区域起始和结尾部分的特征代码, 然后再根据首尾代码人工写出匹配整个区域的正则表达式代码, 例如:

```
/<div\s+class="main_content">\s*<divclass="Sear(.*?)>.*?End\s+-->\s*<\/div>/is
```

将其保存在数据表的 pregmatch 字段中, 最后使用正则表达式匹配函数 preg_match 提取其中的内容部分. 代码如下:

```
for($i=0;$i<$count;$i++){          //分别提取每个返回的网页
preg_match($preg[$i],$res[$i],$mes);          // 提取内容部分
}
```

2.6 修正图像和链接中的 URL 地址

由于各个网站返回的 HTML 代码中的图像和链接的 URL 地址一般是相对 URL 形式, 载入到本系统以后 URL 地址中的域名会发生变化, 因此必须修正图像和链接中的 URL 地址, 比如在代码中的相对 URL 前添加原网站的域名和路径. 这需要使用 PHP 的 DOM 操作类(simple_html_dom)更改 a 标记或 img 标记的 src 属性. 代码如下:

```
include('simple_html_dom.php');          // 引入 DOM 操作类
$html->load($mes[0]);                    //载入提取的内容部分
$imgs = $html->find('img');
foreach($imgs as $v){                    //对每个 img 标记
    if($v->src[0]!='/' && $v->src[0]!='.' && $v->src[0]!='h')
$v->src='/'.$v->src;
$v->src = $imgsrc[$i].$v->src;          //修正 img 标记的 src 属性
}
$as = $html->find('a');
foreach($as as $u)                       //对每个 a 标记
if($u->href[0]!='h'&& $u->href[1]!='t')
    $u->href = $asrc[$i].$u->href;       //修正 a 标记的 src 属性
```

2.7 载入修正后的内容区域

对各网站返回的 HTML 代码经过上述一系列处理后, 接下来就是将处理后的代码载入到本系统的搜索结果列表页面中. 本系统将每个网站返回的 HTML 代码分别放置到一个类名为 cbs 的 div 中, 并在其上方添

加一个 h2 标记用来放置出版社名, 以及一个“more”链接, 用于链接到采集网站的原始搜索列表页. 具体代码如下:

```
<?  for($i=0;$i<$count;$i++){ ?>
<div class="cbs" style="margin:0px 0 0 20px;">
 <h2><b> 耗时: <? echo  $time[$i] ?></b><a class="top" href="#"></a> <a id="m<? echo $id[$i] ?>" href="<? echo  $urls[$i] ?>" target="_blank"></a> <? echo $name[$i];  ?></h2>
<? echo  $html->save();  ?></div>
<? }
$html->clear();
}  ?>
```

由于本系统只提取了每个网页的搜索结果列表区域的 HTML 代码, 这些 HTML 代码引用的 CSS、JS 代码全部丢失. 导致载入到搜索结果页后的页面很不美观. 因此, 还需为这些 HTML 代码添加适当的 CSS 代码来进行美化和布局. 并利用 PHP 的 DOM 操作类适当添加一些 HTML 元素来清除浮动.

最后, 利用 Ajax 技术对搜索结果页添加瀑布流效果. 即开始只载入搜索结果的一部分, 当用户向下滚动网页时再继续载入下一部分内容. 这样, 既加快了搜索结果页的显示, 又能使搜索结果页在最开始时不至于很长.

3 结语

网络上的学习资源丰富, 但仍存在检索不方便的问题. 目前, 各所高校都在积极建设网络教学资源系统, 但这些系统一般都没有将出版社等网站上的教学资源包含利用起来, 很大程度上造成了教学资源的重复建设. 为了方便广大师生更好地获得各种网站上的学习资源, 本文采用 CURL 对各种网站上的学习资源进行聚合查询, 有效利用了出版社等网站上的学习资源. 本系统仍可进一步完善, 比如增加根据用户的查询结果向用户智能推荐资源的功能.

参考文献

1 何芸.基于 CURL 的大学异构数字资源统一查询系统研究.图书馆工作与研究,2013,(6):32.

2 唐四薪.PHP 动态网站开发.北京:清华大学出版社,2015.

3 杨思炜,高东怀,宁玉文.基于云计算的网络学习资源共享研究.中国教育信息化,2012,(5):124.

4 唐四薪.Web 标准网页设计与 PHP.北京:清华大学出版社,2016.

5 唐四薪.PHP Web 程序设计与 Ajax 技术.北京:清华大学出版社,2014.