利用 VBA 程序和 HTTPS 协议获取参考文献的 DOI 信息*

侯修洲† 黄延红
中国科学杂志社,100717,北京

摘 要 为了丰富文章信息,提升读者的阅读体验,本文作者编写了 VBA 程序,并利用 HTTPS 协议自动获取参考文献的 DOI 信息和超链接。经测试,VBA 程序运行后,每 50 条参考文献大约需要 1 min 即可解析完成,对注册过 DOI 的期刊文献的命中成功率几乎达到 100%。对参考文献的体例结构进行精准分析和拆分是 VBA 程序运行成功的基础,向 CrossRef 机构申请成为会员使得 DOI 数据解析不受数据条目的限制,并对今后的数据挖掘提供方便。

关键词 VBA 程序;HTTPS 协议;DOI

Obtaining references DOI information automatically by VBA program and HTTPS protocol // HOU Xiuzhou, HUANG Yanhong

Abstract In order to enrich the article information and enhance the readers' experience, we write a VBA program, and use the HTTPS protocol to automatically obtain references DOI information and hyperlinks. The test shows that, when we run the VBA program, 50 references need to be resolved about one minute, and the shooting success rate is almost 100 percent for registered journal literature in CrossRef database. An accurate analysis and splitting the structure of references are key points when running the VBA program. Becoming a member of the CrossRef institution will make no limit for CrossRef DOI data query, and facilitate future data mining in CrossRef database.

Keywords VBA program;HTTPS;DOI

Authors' address Science China Press,100717,Beijing,China

DOI:10.16811/j.cnki.1001-4314.2016.05.018

数字对象唯一标识符 DOI (digital object identifier)是辨识文献关联信息资源的关键字段信息,通过 DOI 可以快速链接到出版商网站发布的原文网页,也可以获取到该文献的完整 Metadata 元数据,方便读者下载和文献管理,在网络信息资源利用及文本挖掘方面有不可替代的地位和作用[1-2]。近年来,出版商为了丰富文章信息,提升读者的阅读体验,往往会在参考文献中列出各个数据库的链接目标源,DOI 及其相关链接便是其中的一种。同时,我国已于 2015 年 12 月公布了 GB /T 7714—2015《信息与文献 参考文献著录规则》[3],该标准明确将 DOI 作为必备著录项目[4];但是在实际编辑出版中,出版商如何快速获取参考文献的 DOI 信息呢? 在搜索引擎逐篇查询显然不可取,效率、时间和质量都难以保证。VBA 程序(Visual Basic for Applications)在编辑出版工作中的应用已有多人进行了尝试[5-6]。本文充分利用 CrossRef 机构的会员身份,尝试借助 VBA 程序和 HTTPS 协议自动获取 Word 文档中的参考文献 DOI。

1 DOI 查询方法

1.1 简单查询方法 DOI 作为 CrossRef 网站(www.crossref.org)的核心产品,该网站提供了查询 DOI 的几种方式:首先可以在其官方首页搜索关键词或主题检索文献信息,可以间接查询到 DOI 信息;其次,该网站提供了一个简单查询页面(http://www.crossref.org/simpleTextQuery),任何人都可以在这个页面注册 Email 账户,然后将本地的文献信息拷贝到"Enter text in the box below"搜索框查询 DOI 信息。每次查询时允许多条批量查询,并且可以附带查询 PubMed IDs 信息;但是该方法限制每个月最多只能查询 1 000 条,并且还需要将返回的结果网页的数据复制到 Word 文档中。由于复制的仅仅是文本信息,原文档的格式也难以保留,并且每一个步骤都是手动的,效率也不是很高。王玥等[7]利用 VBA 编程在 Word 中实现了自动调用简单查询网页地址、自动填充 Email 和参考文献等功能;但是每个月1 000条的文献查询限制并阻碍了该方法的推广,并且由于每个刊的参考文献体例格式不一样,实际上,CrossRef 网站无法与待查的参考文献做出精确匹配,而这个简单查询网页的核心只是一个模糊查询算法,难以精准锁定返回每篇文献的 DOI。

1.2 利用 API 接口查询 DOI 的方式 CrossRef 网站已经认识到简单查询页面不够精准的问题,并且已开发出一套通过 API(Application Programming Interface,应用程序编程接口)精准锁定 DOI 的方法,主要包含 2 类,即 OpenUrl 开放链接源和 HTTPS 协议,这 2 种方式都允许会员和普通 Email 用户查询,唯一区别是会员查询没有条数限制。需要说明的是,由于注册 DOI 信息的主要是科技期刊文献, CrossRef 网站并不提供基于 API 接口的书籍、专利、学位论文、会议文集等其他

*2015 年文化产业专项资金项目"中国科技类学术期刊国际传播平台"支持
† 通信作者

形式的文献查询,所以本文讨论的也主要是如何获取期刊文献的 DOI 信息。

1.2.1 OpenUrl 开放链接源查询 DOI　通过 OpenUrl 开放链接源查询 DOI(http://help.crossref.org/using_the_open_url_resolver),对于会员,其查询格式如下(其中黑色字体部分是用户的用户名和密码):

https://DOI.crossref.org/openurl?pid=**username**:**password**&aulast=Maas%20LRM&title=JOURNAL%20OF%20PHYSICAL%20OCEANOGRAPHY&volume=32&issue=3&spage=870&date=2002

对于普通 Email 账户,只需将上面网址中的"username:password"替换为注册过的 Email 账户名即可。

不过,无论是会员还是普通用户,对于 OpenUrl 开放链接源查询方式[8-9],上述地址最后返回的网页只是出版商的文章详情页,至于用户怎么从出版商的网页上获取 DOI 信息,CrossRef 是不负责查询的。由于各个出版商的网页格式各种各样,如果想通过网页抓取 DOI 文本信息,这对程序设计是一个很大的挑战。

1.2.2 HTTPS 协议查询 DOI　通过 HTTPS 协议查询 DOI(http://help.crossref.org/using_http),对于会员,其查询格式如下(其中黑色字体部分是用户的用户名和密码):

https://DOI.crossref.org/servlet/query?usr=<**USERNAME**>&**pwd**=<**PASSWORD**>&qdata=|%20Natl%20Acad.%20Sci.%20USA|Zhou|94|24|13215|1997|||

上述网址中"qdata="之后每一个竖线间隔对应一个字段,分别为刊名、第一作者的姓、卷、期、首页码、出版年。需要提醒注意的是,该网址末尾最后 3 个竖线不能省略,否则不能获取 DOI 信息。

对于普通 Email 账户,需要将上面网址中的"usr=<USERNAME>&pwd=<PASSWORD>"替换为注册过的 Email 账户名即可,其代码为"pid=email 账户名"。

相比 OpenUrl 开放链接源查询方式,HTTPS 协议更简单快捷。比如通过 HTTPS 协议,网页只显示一串简单文本,由于无关信息少,所以打开网页速度非常快,并且可以直接在返回的网页文本中显示 DOI,该字段信息被固定在最后一条竖线的末尾,方便后续程序来获取该文献的 DOI 信息,如图 1 所示。

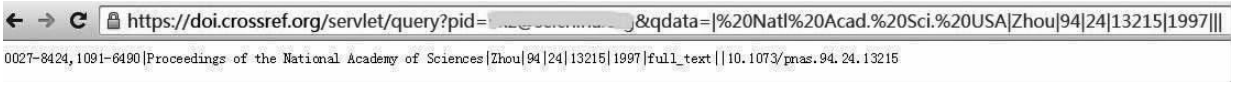

图 1　利用 HTTPS 协议获取文献 DOI 信息的示意

2　利用 VBA 程序和 HTTPS 协议

如果我们想利用 HTTPS 协议自动获取每条参考文献的 DOI,首先面临的是 HTTPS 协议网址如何获取到文献的基本元数据,并且通过文献的元数据能够唯一确定这篇文献。一般来说,只要解析出文献的刊名、年、卷、首页码即可唯一确定文献,在 Word 文档中,解析文献的刊名-出版年-卷-首页码元数据、启动 HTTPS 协议网址,获取到返回网页的文本并解析出 DOI 信息,以及最后决定性的一步,即将该 DOI 信息按照一定的格式填写到文献末尾。所有这些任务都可以交给 VBA 程序来完成。

2.1　解析文献元数据　《以 Science China Physics, Mechanics & Astronomy》一篇参考文献为例,我们尝试分析一下这篇文献的体例格式:"Roberts P H, Glatzmaier G A. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. Nature, 1995,377:203-209"。经过分析,我们得出结论:这篇文献的体例为【作者.题名.刊名,年,卷:首页码-尾页码】,那么转换为程序语言则需要首先识别出【*.*.*,*,*:*】这样的段落,然后再做数据解析拆分,即能得到该参考文献的刊名、年、卷、首页码信息,其中*为通配符,句点、逗号和冒号则是拆分段落中各元素时的标记位置,拆分函数代码如下:

```
n1=InStr(x,".")          第 1 个句点出现的位置
n2=InStr(n1+1,x,".")     第 2 个句点出现的位置
n3=InStr(n2+1,x,",")     第 2 个句点之后第 1 个逗号的位置
n4=InStr(n3+1,x,",")     第 2 个句点之后第 2 个逗号的位置
n5=InStr(n4+1,x,":")     第 2 个句点之后第 2 个逗号之后第 1 个冒号的位置
```

上述代码中,Instr 为 VBA 中的字符串函数,主要功能是获取指定字符在字符串首次出现的位置,以"n2=InStr(n1+1,x,'.')"为例,其中,n1+1 是寻找的起始位置,x 为寻找的字符串,"."是寻找的字符。后面"'"代表注释部分。

接下来我们要根据上述拆分点解析出具体数据,代码如下:

```
j=Trim(Mid(x,n2+1,n3-n2-1))          期刊名
year=Trim(Mid(x,n3+1,n4-n3-1))       出版年
vol=Trim(Mid(x,n4+1,n5-n4-1))        卷
If InStr(vol,"(")>0 Then
vol=Trim(Left(vol,InStr(vol,"(")-1))
```

```
End If
ref. SetRange ref. Start+n5+1,ref. End
fpage=Trim(ref. Words(1). Text)
```

其中,Trim 函数功能为删掉字段首尾空格,Mid、Left 函数为取值函数,其意义和 Excel 软件的同名函数一致。

2.2 启动并发送 HTTPS 协议 当我们获得了文献的关键数据之后,就可以启动 HTTPS 协议了。具体过程如下:

```
link="https://DOI. crossref. org/servlet/query? usr=用户名 &pwd=密码 &qdata=" & "|" & j & "|" & "|" & vol & "|" & "|" & fpage & "|" & year & "|" & "|" & "|" 组合为 HTTPS 协议网址
Set http=CreateObject("Microsoft. XMLHTTP")
http. Open "POST",link,False
http. send""                    '发送 http 协议请求网址
```

2.3 获取返回数据并解析 DOI 信息

```
If http. Status=200 Then
re=http. responsetext           '获取 http 协议返回文本
DOI=Trim(Mid(re,InStrRev(re,"|")+1,Len(re)-InStrRev(re,"|")-1)) '解析 DOI 数据
ref. SetRange ref. End-1,ref. End-1          '创建写入位置
ref. Select
ActiveDocument. Hyperlinks. Add
Anchor:=Selection. Range,Address:="http://dx. doi. org/" & DOI,
    TextToDisplay:="[Cro
End If
```

3 主函数运行过程

综合上述代码,主函数运行过程如下:

```
Sub DOI()
Set myrange = Selection. Range      '对选定的段落进行操作
For Each i In myrange. Paragraphs
  Set ref = i. Range
    If ref Like"*. *. *,*,*:*"Then
      x = ref. Text                 '提取识别段落的文本
       "2.1 解析文献元数据相关代码"
       "2.2 启动并发送 HTTPS 协议相关代码"
       "2.3 获取返回数据并解析 DOI 信息相关代码"
    End If
Next
End Sub
```

需要说明的是,必须先选中参考文献,才可以运行 VBA 程序。有关 VBA 程序的函数解释、界面介绍、录制宏及运行程序等方面可参考文献[7]。经测试,运行 50 条参考文献查询大概需要 1 min,对注册过 DOI 的期刊文献的成功率几乎达到 100%。最终运行程序后的结果如图 2 所示,图中 CrossRef 已经自动带上了 DOI 的超链接。

1 Brandenburg A, Subramanian K. Astrophysical magnetic fields and nonlinear dynamo theory. Phys Rep, 2005, 417: 1–209, [CrossRef]
2 Roberts P H, Glatzmaier G A. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. Nature, 1995, 377: 203–209, [CrossRef]
3 Takahashi F, Matsushima M, Honkura Y. Simulations of a quasi- Taylor state geomagnetic field including polarity reversals on the Earth simulator. Science, 2005, 309: 459–461, [CrossRef]
4 Kageyama A, Miyagoshi T, Sato T. Formation of current coils in geodynamo simulations. Nature, 2008, 454: 1106–1109, [CrossRef]
5 Gailitis A, Lielausis O, Dement'ev S, et al. Detection of a flow induced magnetic field eigenmode in the Riga dynamo facility. Phys Rev Lett, 2000, 84: 4365–4368, [CrossRef]
6 Stieglitz R, Müller U. Experimental demonstration of a homogeneous two-scale dynamo. Phys Fluids, 2001, 13: 561–564, [CrossRef]
7 Monchaux R, Berhanu M, Bourgoin M, et al. Generation of a magnetic field by dynamo action in a turbulent flow of a liquid sodium. Phys Rev Lett, 2007, 98: 044502, [CrossRef]
8 Spence E J, Reuter K, Forest C B. A spherical plasma dynamo experiment. Astrophys J, 2009, 700: 470–478, [CrossRef]
9 Zimmerman D S, Triana S A, Lathrop D P. Bi-stability in turbulent, rotating spherical Couette flow. Phys Fluids, 2011, 23: 065104, [CrossRef]
10 Roberts G O. Dynamo action of fluid motions with two-dimensional periodicity. Philos Trans R Soc Lond A, 1972, 271: 411–454, [CrossRef]

图 2 利用 VBA 程序和 HTTPS 协议获取文献 DOI 信息的示意

4 结束语

本文利用 VBA 程序和 HTTPS 协议成功自动解析出参考文献的 DOI 信息。需要说明的是,由于期刊均有自己特有的文献体例格式,所以本文中的拆分规则不可能要去适应所有期刊的体例格式。这里只是提供一个解决问题的思路,具体问题还需具体分析。比如还是上面那篇文献,有的期刊体例格式可能是如下形式:"P. H. Roberts, and G. A. Glatzmaier. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal, Nature 377,203 (1995)",经分析,此体例可以拆分为【*and 作者,题名,刊名 卷,首页码 (年)】,那么转换为程序语言则需要首先识别出【*and*, *, * (*)】这样的段落,然后再按照这个数据特点进行进一步的拆分。

此外,注册为 CrossRef 网站的会员非常必要,相比普通 Email 用户,会员用户获取该网站的引文信息更加便捷,并且可以免费获得 CrossRef 的技术支持。

5 参考文献

[1] 任瑞娟,孙玲玲,赵然,等. DOI 在网络信息资源管理中的应用价值分析[J]. 情报科学,2010,28(8):1143
[2] 张欣欣,缪弈洲,张月红. CrossRef 文本和数据挖掘服务: