·技术与应用·

基于 WOS API 的论文自动查收查引程序设计与实现

高　营
（深圳大学城图书馆　广东深圳　518055）

摘　要：论文查收查引是图书馆的一项重要服务内容。工作中他引次数的查询手工检索过程步骤复杂，在整个工作中占用较多时间。文章利用Python语言编写程序通过WOS API接口获取数据，进行自动他引次数的查询，简化了工作流程，起到了较好的效果。

关键词：查收查引；程序设计；Python；WOS API
中图分类号：G254.97;G252.62　　**文献标识码：**A

Design of an Automatic Program Based on WOS API for Paper Coverage and Cited Reference Retrieving

Abstract Paper coverage and cited reference retrieving are important service items in libraries. Because the manual cited reference retrieval process is complicated and time-consuming, the author of this paper employs Python to write a program to obtain data through WOS API interface, so as to automatically query the number of cited references. As a result, more simplified workflow and higher efficiency can be achieved.

Key words cited reference retrieving; program design; Python; WOS API

1　引言

科技论文是科研人员研究的结晶，其被各类数据库收录及被引用的次数在一定程度上能反映学者的科研能力和水平，因此论文查收查引证明被广泛应用于职称评定、课题申报、奖励申报、人才选拔、科研评估以及成果鉴定等工作中[1-3]。由于其可计量以及具有相对客观性等，查收查引服务目前仍然是高校图书馆信息检索工作的一项重要内容。

随着我国科研论文数量的增加，论文收录引用的需求随之增加，深圳大学城图书馆（以下简称“我馆”）近年来一直呈上升趋势，2017年全年完成查收查引报告1 414份，检索文献30 018篇。论文查收查引服务是工作量大、重复度高的劳动，以人工为主的工作方式已难以满足及时、高质量的服务要求[4-6]。同时委托人一般会在项目申报截止前较短时间集中申请论文查收查引证明，导致同时收到大量委托，更增加了工作难度。图书馆员们从各种方面研究如何提高工作效率，目前已有许多关于手工检索技巧、流程规范等方面的研究文献[1-3, 7]。中国科学技术大学樊亚芳、陈锴等提出利用Excel的筛选功能、Endnotes Web以及EndNote、NoteExpress等文献管理软件辅助检索，改进检索和统计流程[1-2, 7]。华南理工大学图书馆涂颖哲利用工具软件进行论文查收查引，但其他引查询仍需要人工进行二次检索[8]。也有一些机构开发了论文查收查引工具，如中国科学院系统[9-10]以及CALIS技术中心与北京大学图书馆联合开发的CALIS论文收录及引用系统[11]。查收查引自动化工作有较好的设计和实现，但系统较为复杂，需要进行购买且每个图书馆有其各自不同的需求，有一些功能不能完全满足。

论文查收查引工作中他引次数的查询手工检索步骤复杂，在整个工作中占用较多工作量。我馆查收查引工作中他引次数采用的严格排自引，即引用文献和被引用文献中，只要有一个作者相同即为自引。通常情况下，论文查收查引中的去除自引论文工作是利

用WOS（Web of Science）数据库的作者分析功能及精炼检索功能来完成的，这种方法适用于对于他人引用次数不多或者需检索引用情况的文献合作者不多时，当用户查询的论文引用过多或者原论文合作者过多时，如果检索人员仍使用精炼检索将无法完成多合作者引用的排除[2, 12]，该方法工作量大且容易出错。上述作者排除法需要对每一篇论文的被引论文列表进行一次操作，存在大量的重复劳动，工作效率低下。中国科学技术大学樊亚芳等采用EndNote、NoteExpress等文献管理软件中的检索功能来解决这个问题，利用检索功能在被引论文Library中查找作者含有被检索人的论文，即可批量去除自引论文，利用Label 功能分类统计总他引频次[1-2]。该方法能减少一定的劳动强度，但需要下载所有施引文献。哈尔滨工业大学图书馆李莘等采用的他人引用查询方法在当用户查询的论文引用过多或者原论文合作者过多时，如果检索人员仍使用精炼检索将无法完成多合作者引用的排除，巧用WOS数据库的高级检索功能，将高级检索功能与被引参考文献检索结合完成 SCI/SSCI/A&HCI多合作者自引的排除，检索方式为：#1 NOT AU=（被引文章所有作者的姓名），其中#1为被引文献标题[12]。高级检索的方法存在一个问题，如果论文作者非常多，手工编写检索并不是一个简单的方法。华南理工大学图书馆涂颖哲的论文查收查引工具软件的他引查询过程可以软件提取作者列表，然后在Word软件里利用查找替换生成检索式来进行查询[8]。

我馆采用的他引查询方法是利用检索式排除作者的方法，在工作中采用Word宏来辅助生成WOS平台排除自引检索式来辅助检索，这种方法可以在一定程度上减少手工工作，但未减少在线检索的步骤，仍需要花费较多的时间。

电子科技大学蔺梅芳应用Python语言开发的SCI引文检索自动化软件能够实现4种他引标准下的引文检索，该软件他引判断方法通过程序读入施引文献，进行作者字段比较来排除自引，仍需要下载所有施引文献记录，且检索过程采用模拟浏览器访问方式来获取数据[13]。

我们在工作中结合工作实际，利用Python语言开发了WOS他引查询软件，利用WOS平台提供的APIs（Web of Science Web Services）获取数据，采用SOAP（Simple Object Access Protocol）可以直接获取WOS平台提供的格式化数据。程序可以提取输入内容中的WOS入藏号，自动生成相关检索式进行检索，直接获得他引次数，减少人工操作的步骤，大大减轻了工作强度，提高了准确度，并且结果具有可重复性。

2 设计与实现

2.1 设计思路

软件模拟人工查询他引次数的步骤，实现自动获得他引次数，具体流程如下（见图1）。

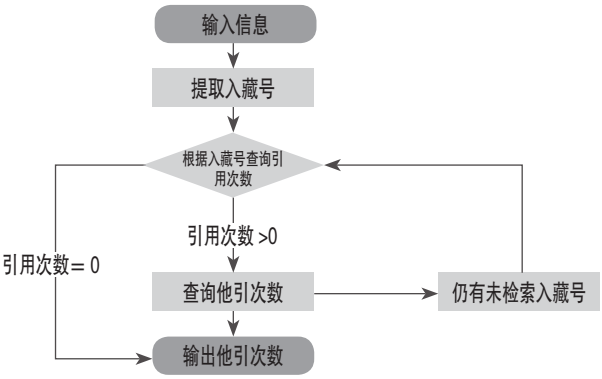

图 1 他引查询程序流程图

输入信息可以直接从Word文件或网页信息复制粘贴含有入藏号的信息，也可以输入WOS平台导出记录文件，软件从输入信息中利用正则匹配提取所有入藏号信息。软件按照入藏号通过WOS API查询引用次数，如果引次数为0次则不进行他引次数查询，如果引用次数大于0次，进入他引查询步骤，直至查询完成所有入藏号的查询，显示输出结果。

2.2 开发语言选择

Python语言诞生于20世纪90年代初，现今已经成为最受欢迎的程序设计语言之一[14]，在最近的Git Hub排行榜中名列前茅。Python是一种解释型、面向对象、动态数据类型的高级程序设计语言，有丰富的标准库和其他一些扩展库，可以用较少的代码完成一些复杂的工作。并且Python具有简单、易学、免费、开源等诸多优点。GitHub中有非常多的代码实例可供参考，是一种易于阅读和方便编写代码的语言，因此我们选择Python作为开发语言。

2.3 WOS 平台 API

蔺梅芳等开发的引文自动化检索软件[13]和涂颖哲开发的论文查收查引工具软件[8]等均采用模拟浏览器访问的方式来获取数据，通过分析网页结构来提

取数据。这种方式有可能受到网页读取速度的影响[13]，如果网站升级网页结构变化，则需要修改程序来适应相应的变化，利用WOS API可以避免这些影响。

Web of Science Web Services[15]是基于SOAP 1.1（Simple Object Access Protocol，简单对象访问协议）和WSDL 1.1（Web Services Description Language，网络服务描述语言）的API，用于访问和搜索Web of Science数据库订阅内容。该API有两个服务接口WOKMWSAuthenticate和WokSearch。其中WOKMWSAuthenticate是身份验证和会话管理服务，WokSearch提供数据检索服务。可以在学校或机构IP范围内通过WOKMWSAuthenticate接口获取授权信息，然后通过WokSearch接口来进行检索和获取数据。可以通过该API可获取到格式规范的XML数据（见图2），由于仅获取所需要的数据，不需要打开整个网页，可避免受到网页读取速度或网站改版的影响，且所获取数据具有稳定、规范的数据结构。

```
<?xml version="1.0" ?>
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
  <soap:Body>
    <ns2:citingArticlesResponse xmlns:ns2="http://woksearch.v3.wokmws.thomsonreuters.com">
      <return>
        <queryId>1</queryId>
        <recordsFound>50</recordsFound>
        <recordsSearched>58309696</recordsSearched>
        <parent>
          <REC r_id_disclaimer="ResearcherID data provided by Clarivate Analytics"
xmlns="http://scientific.thomsonreuters.com/schema/wok5.4/public/FullRecord">
            <UID>WOS:000301273800025</UID>
            <static_data>
              <summary>
                <EWUID>
                  <WUID coll_id="WOS"/>
                  <edition value="WOS.SCI"/>
                </EWUID>
                ...

          </REC>
        </return>
    </ns2:citingArticlesResponse>
  </soap:Body>
</soap:Envelope>
```

图2　通过WOS API获取数据结构

2.4　实现过程

WOS数据库收录的每一篇文章都有一个唯一的入藏号，入藏号是以“WOS:”开头，后面是15—16位数字或字母组成的编号，我们可以用正则表达式（WOS:\w{15,16}）来匹配提取所有入藏号。

通过Python的suds库访问WOS API来获取数据（见图3），通过WOKMWSAuthenticate接口的authenticate操作来获取授权会话session SID，并将获得的授权信息加入搜索进程。WokSearch服务有7个检索操作[15]，包括search、citedReference、citingArticles、relatedRecords、retrieveById、citedReferencesRetrieve、retrieve。其中Search查询可以提交查询并返回结果，与网页界面高级检索查询功能返回结果一致；citingArticles查询可以获取引用查询文章的所有文章，我们可以通过这个功能检索引用次数；可以结合search和citingArticles进行他引次数查询。首先通过citingArticles查询文章的所有引用次数，即总引次数，同时通过这个查询可以获得所查询文章的详细信息以及该查询的查询序号（queryId），从所得详细信息中提取出所有作者列表，生成检索式#queryId NOT AU=（被引文章所有作者的姓名），通过search查询获得他引次数及他引文章列表信息。

```
# -*- coding:utf-8 -*-

from suds.client import Client
import xml.etree.ElementTree as ET

AUTH_URL = 'http://search.webofknowledge.com/esti/wokmws/ws/WOKMWSAuthenticate?wsdl'
SEARCH_URL = 'http://search.webofknowledge.com/esti/wokmws/ws/WokSearch?wsdl'

#获取授权信息
auth_client = Client(AUTH_URL)
SID = auth_client.service.authenticate()

#获取数据
databaseId = "WOS"
uid = 'WOS:000301273800025'
editions = [{'collection':'WOS', 'edition':'SCI'},
            {'collection':'WOS', 'edition':'SSCI'}] #None
queryLanguage = "en"
timeSpan = {'begin': "1900-01-01", 'end': "2018-12-31"}
rparams = {'firstRecord' : 1, 'count': 1}
search_client = Client(SEARCH_URL, retxml=True)
search_client.set_options(headers={'Cookie': 'SID="%s"' % SID})

response = search_client.service.citingArticles(databaseId,
                                                uid,
                                                editions,
                                                timeSpan,
                                                queryLanguage,
                                                rparams)

#获取引用次数
root = ET.fromstring(response)
cite_num = root.findall(".//recordsFound")[0].text
print(cite_num)
```

图3　程序实现过程

为了方便其他同事利用该程序，利用wxPython编写可视化界面GUI（见图4），并用pyinstaller打包成独立的exe文件，可以拷贝到任何电脑使用，不需要电脑安装Python环境的步骤。程序可以通过导入文本格式文件来提取入藏号，检索完成后可以导出csv格式的文件，方便后续工作。

图4　他引次数查询程序界面

3　效果分析

软件开发以来已经使用两年有余，期间经历WOS从汤森路透到科睿唯安的转变，均能稳定使用。由于使用步骤简便，输出结果稳定可靠，可以节约大量时间，减少重复性劳动，提高工作效率，得到了查收查引工作人员的认可。

表1比较了本文程序及涂颖哲[8]、王学勤[9]和蔺梅芳[13]所发表论文的引文检索用时，本程序同蔺梅芳文章中的程序用时相当，在引用查询方面优于其他两篇文章中的程序。因为均为Python语言编写，均采用多线程，考虑样本选择、网络环境等可能存在差异，效率基本相当，但其结果有部分还需要人工复核。本程序通过两次检索可以得到他引查询结果。

表1 引文查询用时相关文献比较

被检索论文数/篇	本软件用时/分钟	蔺梅芳文献用时/分钟	涂颖哲文献用时/分钟	王学勤文献用时/分钟
10	0.71	1.88	1.67	4
50	1.76	2.13	8.33	42
100	2.55	2.53	20.83	80

此外，本程序具有非常好的扩展性，除了可以检索WOS总引、他引外，通过设置检索数据库范围还可以检索SCI总引和他引。通过设置检索时间范围，可以满足一些特殊的需要。比如某项目需要检索近5年的SCI引用情况，通过修改检索时间范围timespan开始和结束时间，就可以完成相应他引次数的查询。比如某大型科技公司集团有近千篇SCI收录论文，其科技办工作人员想知道每个月其所有论文的被引次数变化，如果通过人工检索工作量可想而知，经常需要几天的时间，对本程序进行简单修改，只需要一个小时左右即可完成检索，省时省力，而且还能保证数据准确可靠。

4 结语

本文利用程序来进行他人引用次数的查询，减轻了工作强度，并且减少了人为操作带来的误差和错误，有很好的重复性和稳定性。程序为单文件形式，方便传播及使用，得到了检索人员的好评。

在后续工作中可以在软件的易用性、美观性等方面进行改进。例如，可建立数据库保存相关人员的论文记录，下次再进行检索时仅需要检索新增加记录即可；除了检索总引次数和他引次数外，可同时导出检索论文列表及引文列表，更进一步可生成出具报告的相关内容和格式，最大程度地降低人工工作强度。

参考文献：

[1] 张雪娟,樊亚芳.Note Express在论文查收查引工作中的应用[J].情报探索,2017(6):45-49.

[2] 樊亚芳.利用文献管理软件提高论文查收查引工作效率的实践与应用[J].高校图书馆工作,2017(2):63-66.

[3] 宋成方.查收查引服务质量提高路径及其延伸服务探析[J].山东图书馆学刊,2012(5):47-50.

[4] 杨华,杜如坤.SCI收录论文证明自助打印系统的设计与应用[J].中华医学图书情报杂志,2017,26(10):53-55.

[5] 王洪军,张玉,李焱,等.基于Web的中文期刊查收查引跨库检索系统研发[J].中华医学图书情报杂志,2016(6):24-28.

[6] 张勇,李爽.公共图书馆查收查引服务实践研究[J].图书馆学研究,2015(24):63-66.

[7] 樊亚芳,陈锴.利用Excel和End Note Web提高论文查收查引工作效率[J].图书馆杂志,2013(1):32-34.

[8] 涂颖哲.论文查收查引工具软件的设计与应用实践[J].农业图书情报学刊,2015(8):34-38.

[9] 王学勤,郝丹,郑菲,等.“查收查引报告自动生成系统”应用实践研究[J].图书情报工作,2014,(16):131-137.

[10] 马海收,刘媛媛,郑菲,等.基于ISI Webof Know ledge引证检索服务统计软件设计与实现[J].情报杂志,2012(2):148-152.

[11] 马芳珍,李峰,季梵,等.对CALIS查收查引系统的测试和应用效果评价[J].大学图书馆学报,2016(2):97-102.

[12] 李莘,李雪婷.查收查引常见问题及解决技巧探讨[J].图书馆建设,2015(9):78-80.

[13] 蔺梅芳,翟燕,张宇娥.应用Python语言的引文检索自动化软件设计与实践[J].四川图书馆学报,2016(3):42-45.

[14] Interactive: The Top Programming Languages 2018[EB/OL].[2018-12-25].https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2018.

[15] Web of Science Web Services Expanded HELP[EB/OL].[2018-12-25].http://ipscience-help.thomsonreuters.com/wosWebServicesExpanded/WebServicesExpandedOverviewGroup/Introduction.html.

作者简介：高营（1979—），男，硕士，深圳大学城图书馆助理研究员，研究方向为学科服务。

收稿日期：2018-12-25