基于 CrossRef 数据库的参考文献自动加工及 XML 标引方法

侯修洲 黄延红
《中国科学》杂志社,100717,北京

摘 要 为了提升参考文献的自动化加工水平和准确率,通过编写 VBA(Visual Basic for Applications)程序,利用 HTTPS 协议自动获取参考文献的 DOI 信息,进一步利用获取的 DOI 信息从 CrossRef、PubMed 和 ADS 数据库挖掘出文献的元数据信息,并按照期刊的具体格式规范文后参考文献的加工和 XML 信息标引。经测试,运行 VBA 程序后,每 50 条参考文献只需 5 min 即可完成解析和文献加工,大大提升了编辑效率和准确率。认为对参考文献的体例和各数据库的 API 接口熟练掌握和使用是程序运行成功的关键。

关键词 VBA 程序;HTTPS 协议;DOI;参考文献;自动化;XML

Method of references processing and XML marking automa-tically by CrossRef Database// HOU Xiuzhou, HUANG Yanhong

Abstract In order to improve the automation level and accuracy of references processing, a series of VBA (Visual Basic for Applications) programs are compiled, which can use the HTTPS protocol to automatically obtain references DOI information, and further make use of the acquired DOI to mine the metadata information from CrossRef, PubMed and ADS database. Then we also complete the references processing and XML marking with a given reference format in journals by the acquired metadata information. It is much effective that copyediting process of 50 references can be completed within five minutes. The experiences in references style and database API interface are the key of the program running.

Keywords VBA program; HTTPS; DOI; reference; automation; XML

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

DOI:10.16811/j.cnki.1001-4314.2017.01.023

在科技论文中,参考文献是文章的重要组成部分,既具有索引作用,也方便读者进行二次文献扩展阅读[1]。为了规范参考文献著录体例格式,我国已于 2015 年 12 月公布了新的参考文献著录标准 GB/T 7714—2015《信息与文献 参考文献著录规则》[2],如何在稿件加工中严格执行新的国家标准,是一项非常复杂而又烦琐的工作,比如引文作者的姓名、题名、刊名、卷、期、页码等信息,纯手工编辑校对出错率相当高[3-4]。本文尝试提出一种新的快速、高效而又相对准确的参考文献自动加工方法,该方法基于 CrossRef 数据库的基础信息,按照期刊要求可以输出 TXT 文本格式文献,也可以输出拆分完整的 XML 标记型格式文献。

曾经有人尝试利用 Google 学术网站的引用工具来快速加工文献,这种方法最大的弊端是需要逐条加工,并且 Google 学术搜索网站提供的查询并不是精确查询,其查询数据也有很多缺项和错误[5]。本文作者在之前的一篇论文中介绍了利用 VBA(Visual Basic for Applications)程序和 HTTPS 协议获取参考文献的 DOI 信息[6],那么是否可以利用已经解析出的 DOI 信息来对参考文献进行辅助加工和校对呢?答案是肯定的,并且所有的文献解析、数据挖掘以及后期的文献自动加工和数据输出均由 VBA 程序完成。该方法的优点是不需要文献的作者、题名的格式这些信息,只需大致确定文献的刊名、年、卷、页码信息即可获取文献的 DOI 信息,并利用获取的 DOI 到 CrossRef 数据库进行数据挖掘,进而获取文献的全部元数据,对这些元数据进行程序化自动校正和修改,即可完成该文献的编辑加工。

1 方法

由于注册 DOI 信息的主要是期刊文献,CrossRef 网站并不提供基于 API 接口的书籍、专利、学位论文、会议文集等其他形式的文献查询;所以本文讨论的主要是如何利用 VBA 程序获取期刊文献的 DOI 信息,以及由此 DOI 信息进一步挖掘 CrossRef 的元数据信息,然后利用这些元数据信息来对文献进行编辑加工。

1.1 分析文献样式 按照文献[6]介绍的方法,如果想提取出文献的 DOI,则必须知道文献的结构类型,并解析出文献的刊名、年、卷、页码等元数据。对于作者提交的各式各样的参考文献格式,我们需要预见一些可能的格式,以便不论原始文献是什么结构类型,都能准确解析元数据。参考文献样式一般分为顺序编码制和著者-出版年制 2 大类,围绕这 2 大类会衍生出若干分支类型,以下是本文总结划分的现行的基本类型。

1.1.1 顺序编码制

1)Gailitis A, Lielausis O, Dement'ev S, et al. Detection

of a flow induced ***. Phys Rev Lett,2000,84:4365-4368

2) Gailitis A, Lielausis O, Dement'ev S, et al. Phys Rev Lett,2000,84:4365-4368

3) M.Aspelmeyer, T. J. Kippenberg, and F. Marquardt, Rev.Mod.Phys.86,1391(2014).

1.1.2 著者-出版年制

1) Zhuang W, Feng M, Du Y, 2013. Low-frequency***.J Geophys Res-Oceans,118:1302-1315

2) Zhuang W, Feng M, Du Y.2013. Low-frequency***. J Geophys Res-Oceans 118,1302-1315

3) Zhuang W, Feng M, Du Y. 2013. J Geophys Res-Oceans,118:1302-1315

1.1.3 DOI 识别方法　Paci I, Johnson C J, Chen X D, et al. Singlet**.J Am Chem Soc,2006,128:16546-16553,DOI:10.1103/PhysRevB.92.041104

1.2 解析文献元数据　当我们按照 1.1 节总结的文献类型识别出文献结构后,就可以对文献进行拆分,并解析出元数据[6],为挖掘出 DOI 信息做数据准备。

1.3 从 CrossRef 数据库挖掘 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,其中刊名、卷、首页码、年为解析的文献元数据:

https://DOI.crossref.org/servlet/query? usr=<USERNAME>&pwd=<PASSWORD>&qdata=|刊名||卷||首页码|年|||

1.4 利用 DOI 进行数据挖掘　当 DOI 信息成功获取后,我们就可以利用 DOI 到 CrossRef 数据库挖掘出标准的元数据信息,以及进一步挖掘出 Pubmed、ADS 和 arXiv 等编码信息。从 CrossRef 数据库挖掘出 XML 元数据的 http 接口协议如下,可以看到,只要提供用户名、密码和 DOI 信息就可以获取该文献的完整信息:https://doi.crossref.org/search/doi? pid=<USERNAME:PASSWORD>&format=unixsd&doi=<DOI>。图 1 是 API 接口返回的参考文献的 XML 信息。

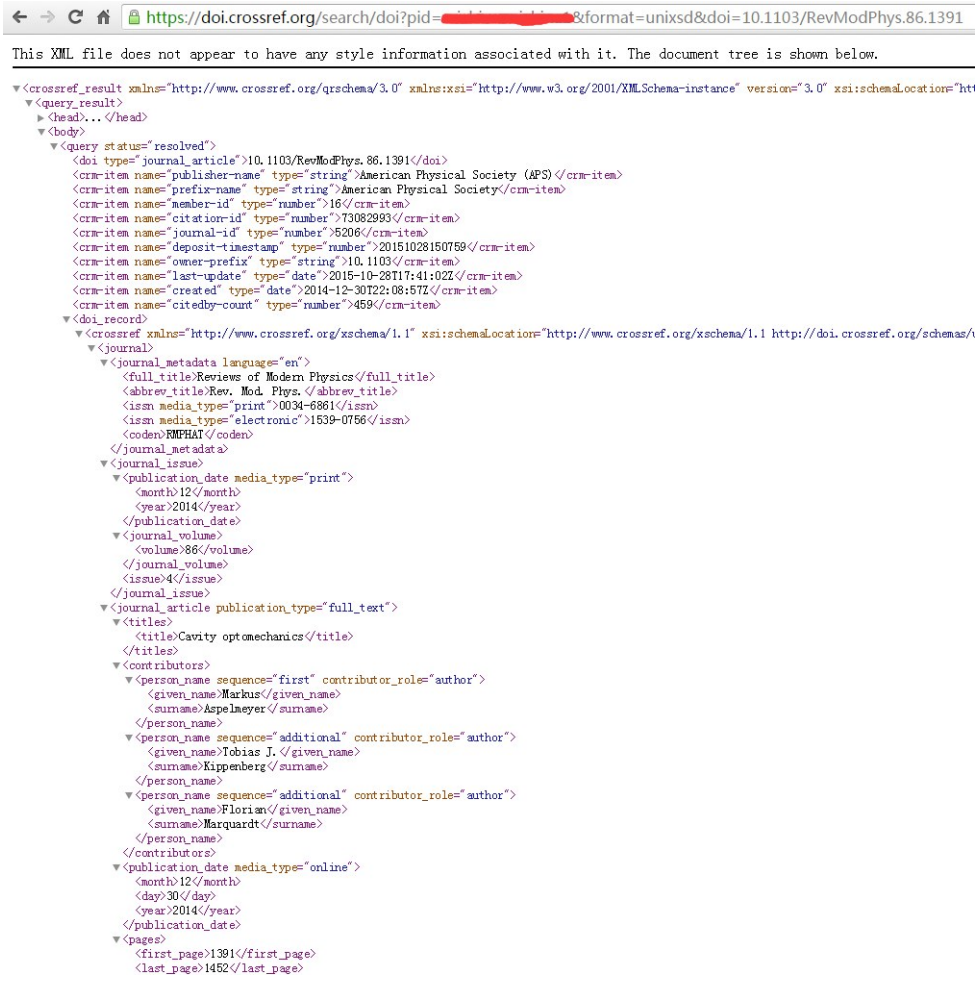

图 1　CrossRef 数据挖掘 XML 页面

同理,我们也发现了 Pubmed、ADS 和 arXiv 的 http 协议,其中 Pubmed 文摘数据库是美国国立卫生研究院主办的网站,用户可以免费获取每篇文章的 XML 信息,并且只需要 DOI 信息即可,其 XML 信息的 API 接口为:

http://www.ncbi.nlm.nih.gov/pubmed/? term=<DOI>&report=xml&format=text

ADS 数据库是哈佛大学主办的数学物理力学天文材料类数据库网站，并且也记录文章的 arXiv 号，信息权威，更新快，深受专业人士喜爱。ADS 开放的是 bibtex 文本接口，也是只需要 DOI 信息即可，其文本的 API 接口为：http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=<DOI>&data_type=BIBTEX

1.5 利用挖掘出的元数据进行文献再加工 当我们最终获取到文献的元数据后，就可以用这些信息按照事先设定的格式对文献进行加工。比如作者的姓和名的顺序，是否有缩写点，名字中间是否有空格，是否需要题名，刊名是否需要缩写，年卷期页码的位置等等。按照刊物生产的要求，可以输出文本格式的文献，也可以输出拆分好的便于 XML 生产的标记性文献。图 2 是 XML 标记型文献输出样例，其中\author{ }代表作者信息，里面每一组<a><\/a>代表一个作者信息，<g>代表作者名，<s>代表作者姓。此外，\title{ }、\journal{ }、\year{ }、\vol{ }、\fpage{ }、\lpage{ }，分别为题名、刊名、年、卷、首页、尾页信息，\doi{ }、\pubmed{ }、\ads{ }、\arxiv{ }分别为 CrossRef、Pubmed、ADS 和 arXiv 数据库的文章 id 信息。

```
1  <r>\author{<a><g>A</g><s>Gailitis</s></a><a><g>O</g><s>Lielausis</s></a><a><g>S</g><s>Dement'ev</s></a><etal>et al.</etal>}, \title{Detection of a Flow Induced Magnetic Field Eigenmode in the Riga Dynamo Facility}, \journal{Phys Rev Lett}, \year{2000}, \vol{84}, \fpage{4365}, \lpage{4368}, \doi{10.1103/PhysRevLett.84.4365}, \pubmed{10990687}, \ads{2000PhRvL..84.4365G}</r>

2  <r>\author{<a><g>M</g><s>Aspelmeyer</s></a><a><g>T J</g><s>Kippenberg</s></a><a><g>F</g><s>Marquardt</s></a>}, \title{Cavity optomechanics}, \journal{Rev Mod Phys}, \year{2014}, \vol{86}, \fpage{1391}, \lpage{1452}, \doi{10.1103/RevModPhys.86.1391}, \ads{2014RvMP...86.1391A}, \arxiv{1303.0733}</r>
```

图 2 XML 标记型参考文献输出样例

因为解析后的文献数据是完全拆分好的，程序能够输出为 XML 标记型格式；当然，也可以按照给定的刊物文献体例进行任意组合输出。比如作者姓和名的先后顺序，名是否为缩写、是否含缩写点、是否含空格，是否需要输出题名，以及刊名、年、卷、期、页码的先后顺序和具体展现格式，并且程序还为刊名字段专门建立了 ISO 缩写单词词库，可以保证输出的刊名符合 ISO 缩写标准格式。

图 3 是原始文献从分析、解析、信息挖掘、信息加工及信息输出的全流程 VBA 程序设计示意图。当解析出 DOI 并成功挖掘出元数据后，还需要与 Pubmed 元数据、ADS 元数据以及原文献的题名、刊名、年、卷、页码等信息进行交叉校对，以验证信息的正误或者缺失。需要说明的是，如果没有解析出 DOI 信息，或者解析出的文献第一作者不在原文中，则该条文献保持原样不变，这时需要编辑进一步核对文献格式或进行手动加工。

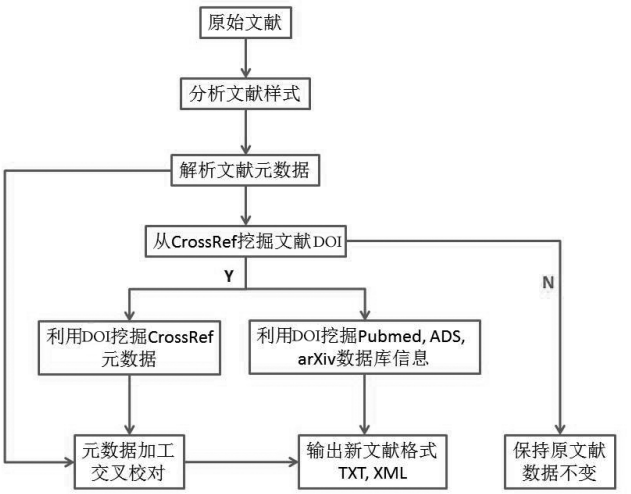

图 3 参考文献自动加工及 XML 标引流程示意图

2 结论

本文结果适应参考文献自动加工的需求，并且能满足 XML 生产转换的需求，VBA 程序安装简单[7]，操作易于上手，有利于程序的推广和使用。经测试，运行 VBA 程序后，每 50 条参考文献只需 5 min 即可完成解析和文献加工，大大提升编辑效率。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，同时对各数据库的 API 接口熟练掌握和使用是程序成功的关键。

本软件可以与 Word 软件紧密结合，只需一键即可完成文献加工，不需要到外部网页或软件中进行信息校对或采集，自动化程度高，并且可以将参考文献解析为 XML 标记性语言输出，适合各种刊物文献格式的编辑加工和 XML 排版生产。

3 参考文献

[1] 李丽，张凤莲.应重视参考文献表的编辑加工[J].编辑学报，2004，16(6)：412

[2] 信息与文献 参考文献著录规则：GB/T 7714—2015[S].北京：中国标准出版社，2015

[3] 宋春燕，王菊香.科技期刊论文参考文献核查与校对方法[J].编辑学报，2012，24(3)：249

[4] 朱建新. 科技论文参考文献勘错以及查漏补缺的一些技巧[J]. 学报编辑论丛，2011：103

[5] 李万会，张晶. 利用“谷歌学术搜索”快捷地编辑加工参考文献[J]. 学报编辑论丛，2013：228

[6] 侯修洲，黄延红.利用 VBA 程序和 HTTPS 协议获取参考文献的 doi 信息[J].编辑学报，2016，28(5)：466

[7] 王玥，毛善锋，刘谦.Word 文档中通过 CrossRef 自动查询与整合英文参考文献 DOI 的实践[J].中国科技期刊研究，2013，24(2)：333

(2016-08-22 收稿；2016-10-31 修回)