期刊编辑发表论文情况的文本挖掘与分析*

谭春林[1)] 刘清海[2)]

1)《华南师范大学学报(自然科学版)》编辑部,510631;2)《中山大学学报(医学科学版)》编辑部,510080:广州

摘 要 基于万方数据库"高级检索"功能,通过Python编程对期刊编辑发表学术论文的文本信息进行挖掘。对全国各省市期刊编辑发表论文的总篇数、关键词的频次等进行了统计分析。结果表明:全国各省市中,北京的期刊编辑发文量遥遥领先,编辑的学术氛围浓厚;全国期刊编辑发文量在2011年以后出现几次断崖式下降;编辑发文量的下降或与2011年出版业转制、新媒体和自媒体引起的出版业3次"离职潮"有关。文中提出的Python编程挖掘文本的方法,为期刊编辑从事相关学术研究提供了新思路。

关键词 期刊编辑;编辑学;论文信息;Python;文本挖掘;数据分析

Text mining and analysis of papers published by journal editors //TAN Chunlin, LIU Qinghai

Abstract Based on the advanced searching function of Wanfang database, the text information of papers published by journal editors is mined by Python programming. The total amount of papers published by editors of journals in various provinces and the frequency of keywords are analyzed. The results show that: among the provinces and municipalities in China, Beijing is far ahead of the others in the total amount of editorial papers, and the academic atmosphere of editors is strong; the total number of papers published by journal editors has declined several times since 2011. The decline in editorial output may be related to three "turnover tides" in the publishing industry caused by the transformation of publishing industry, the development of new media and Self-Media in 2011. The Python programming method proposed in this paper provides a new way for journal editors to engage in relevant academic research.

Keywords journal editor; editorial science; paper information; Python; text mining; data analysis

First-author's address Editorial Office of Journal of South China Normal University(Natural Science Edition), 510631, Guangzhou, China

DOI:10.16811/j.cnki.1001-4314.2019.04.015

随着互联网的飞速发展,数据挖掘与数据分析手段已经渗透到各个领域。近几年来,基于数据挖掘与数据分析在期刊编辑出版领域的研究逐渐兴起。2013年张晓倩首次提出数据挖掘在网络在线投稿系统中的应用[1];2015年以来,各种文本数据挖掘与分析工具(如HADOOP[2]、CrossRef Metadata API[3]、ROST News Analysis Tool 4.5[4]、Citespace和SCI2[5])相继被介绍或被应用于编辑与出版工作实践。已有编辑同人开始利用数据挖掘研究精准化办刊策略[6]、尝试学术期刊选题策划创新路径[4]、基于题录信息分析的期刊数据研究[7],以及将数据挖掘与分析技术应用于微信文章栏目选题策划[8]等。可见,利用出版数据的分析和解读,指导期刊的选题策划、设计与经营[9],已开展得比较深入与普遍。

在前期研究中,笔者采用传统手动获取方式,从QQ群及电商平台挖掘论文交易信息来分析学术不端诱因[10],其研究工作量大,效率低。本文尝试利用Python编程对期刊编辑发表的论文元数据进行挖掘与分析,为期刊编辑从事学术研究提供一种新途径。

1 研究方法

1.1 数据来源 利用万方数据知识服务平台V2.0(http://www.wanfangdata.com.cn)的"高级检索"功能对编辑部的编辑发文情况进行大数据挖掘和数据分析。本文仅对论文发表情况进行研究,不考虑学位论文、专利、标准等检索范围;因此,在"高级检索"中将"文献类型"设置为"期刊论文"和"会议论文",检索模式默认为"精确",根据需要选择"检索信息"中的字段(作者单位),输入不同关键词,同时考察与(and)、或(or)、非(not)条件,实现多个关键词、多种条件检索。

1.2 关键词筛选 采用手动检索,筛选题名关键词,检索的文献量如表1所示。以"科技期刊""中文期刊""期刊""编辑""编辑学""期刊出版""出版"等关键词的检索结果包含了公告、征稿启事等;因此,为了确保检索结果准确,以"作者单位"检索"编辑部"的方式进行检索。

1.3 开发环境 基于Python 3.6编程语言环境,采用PyCharm 2017.3.3(Professional Edition)开发工具设计网页文本数据挖掘程序,代替人工检索模式,快速抓取所需文本数据。需要导入(import)Python标准库:正则表达式操作(re)、字符串操作(string)、时间的访问与转换(time)、URL处理模块(urllib)、超文本标记语言支持(html),以及可以从HTML或XML文件中提取数据的第三方扩展库——BeautifulSoup。

* 2015年广东省公益研究与能力建设专项资金项目(2015A030302074);广东省科学技术期刊编辑学会面上项目(201801)

表1 不同检索关键词及发表时段对期刊论文、会议论文的搜索结果

序号	搜索词	发表时段	搜索表达式	结果/条
1	编辑部	—2018	作者单位:(“编辑部”) * Date:—2018	310 414
2	编辑部;广东	—2018	(作者单位:(“编辑部”) * 作者单位:(“广东”)) * Date:—2018	4 624
3	编辑部;广州	—2018	作者单位:(“编辑部,广州”) * Date:—2018	2 929
4	编辑部	2017	作者单位:(“编辑部”) * Date:2017—2017	14 551
5	编辑部;广东	2017	(作者单位:(“编辑部”) * 作者单位:(“广东”)) * Date:2017—2017	270
6	编辑部;广州	2017	(作者单位:(“编辑部”) * 作者单位:(“广州”)) * Date:2017—2017	233

注:检索采用 http://www.wanfangdata.com.cn 的“高级检索”功能;统计日期为2018-12-16。

理论上,不考虑反机器人机制和网速等因素时,从数以万计的检索结果中抓取文本数据,耗时仅需数十分钟到数小时。这相比手动检索和复制(通常需要数十天)网页信息具有很大的优势。

传统文献调研方法通常采用手动检索数据库(中国知网、万方数据等)和搜索引擎(百度学术、谷歌学术等),对于研究少量文献的需求基本能满足;但在大数据时代,数据的获取至关重要,采用手动检索效率低且远不能满足要求。下文以2017年广东编辑发表学术论文为例,提出通过Python编程挖掘数据的方法。

1.4 文本挖掘算法 从网页挖掘文献元数据的算法流程图如图1所示。文本挖掘程序包括初始化、一级检索、二级检索。

第1步:初始化。输入检索条件queryString,设置检索结果的分页标准pageSize=50(50篇/页),利用字符串拼接方法构造文献检索查询的网址。例如:url='http://www.wanfangdata.com.cn/search/searchList.do?searchType=all&pageSize=50&searchWord=((作者单位:(“编辑部”) * 作者单位:(“广东”)))起始年:2017 结束年:2017。调用预查询prePicker(url)函数,获得检索结果中的总篇数n,计算检索总页数m。构造检索页的网址,并存入urls数据集:

$$\text{urls}=[u_0,u_1,\cdots,u_m]。$$

第2步:一级检索。包含文献编号集(art_ids)的获取、文献详细页面网址集(art_urls)的构造2部分。

在文献art_ids的获取程序中,利用for循环,从初始化过程获得的检索页面网址集urls中遍历列表,将每一个检索页网址url传参数给linkPicker()函数。利用正则表达式re.findall()的非贪婪模式“(.*?)”精准查找并抓取每条文献的art_id,并将其追加到art_ids列表中。为了方便对程序运行过程的监控,按相同方式获取文献的标题集(titles)。

$$\text{art\_ids}=[I_0,I_1,\cdots,I_n]$$
$$\text{titles}=[t_0,t_1,\cdots,t_n]$$

在文献详细页面网址集的构造过程中,利用for循环遍历art_ids列表,读取某篇文献的编号art_id(即I_i),构造该文献详细页的网址(art_url)为:

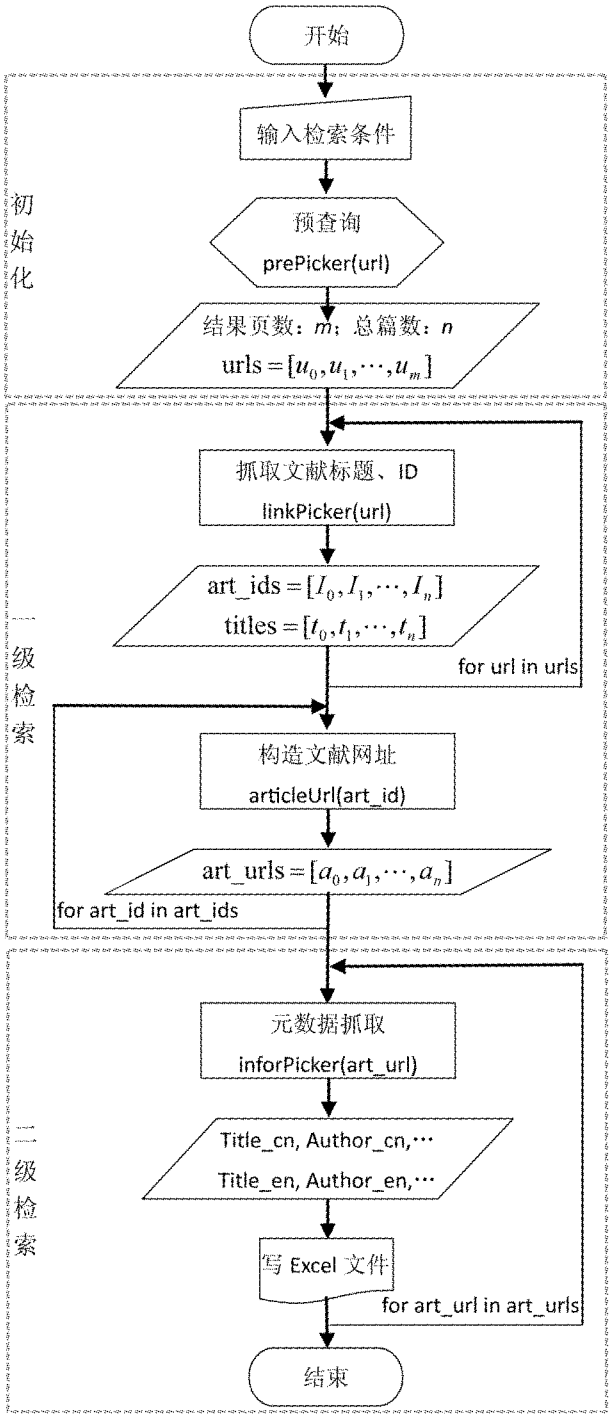

图1 一级检索的程序流程

art_url = ‘http://www.wanfangdata.com.cn/details/detail.do?_type=perio&id={}’.format(art_id)
将 art_url 追加到 art_urls 网址集中。

第 3 步:二级检索。文献的详细页面(二级网页)包括题名、作者、单位、摘要、关键词、刊名、年卷期、页码、栏目、基金项目、在线出版日期等中英文信息。二级网页是本文数据挖掘需要访问的重要页面。

以作者(Author)、标题(Title)信息的获取为例,用于二级页面文本挖掘的函数包括 infoPicker()和 file.write()2 个函数,其中 infoPicker()是提取文本的核心程序。

infoPicker()函数只有 1 个参数 art_url,可以将 1 个网址传参给该函数,以实现对该网址相应网页上指定内容的提取,该函数的代码如下:

```
def infoPicker(art_url):
    res = requests.get(art_url, headers = headers)
    soup = BeautifulSoup(res.text, ‘lxml’)
    content = soup.select(‘#div_a > div > div.left_con > div.left_con_top’)[0]
    Author = re.findall(‘ < div class = “info_right” > (.*?) </div > ’, res.text,re.S)
    if Author! =[]:
        Author = filter_tag(Author[0]).strip().replace(‘’, ‘,’
        Title = re.findall(‘ < div class = “title” > (.*?) </div > ’, res.text, re.S)
if Title! =[]:
    Title = filter_tag(Title[0]).strip().replace(‘*’, ‘’)
```

如果遇到非 utf-8 编码的特殊字符时,在后续输出文件操作的过程中会出错,因此,在 infoPicker()函数中需要去除文本数据中的特殊字符。例如“*”是方正排版引入的“*”,需要利用 replace()替换将其去除。

file.write()函数包含 2 个参数(file_path, datas),其中 file_path 是输出文件的地址(包括文件名),datas 是需要输出的数据(文本信息)。将二级检索挖掘的文本数据采用 file.write()函数输出到 Excel 文件中。该函数的代码如下:

```
def xls_write(file_path, datas):
    f = xlwt.Workbook()
    Sheet1 = f.add_sheet(u‘sheet1’,cell_overwrite_ok = True)
    i = 0
    for data in datas:
        for j in range(len(data)):
            sheet1.write(I, j, data[j])
        i + +
    f.save(file_path)
```

2　结果与讨论

2.1　文本挖掘为学术研究提高效率　利用 Python 编程进行文本数据挖掘,具有速度快、挖掘精准,可大幅提高编辑从事学术研究的效率。从文献网页挖掘的文献元数据信息具有准确、清晰等特点,它完美地过滤了网页的 html 代码和特殊字符,方便后期的数据分析。从挖掘的 270 篇广东省期刊编辑发表的论文元数据中,统计出 1 124 个关键词及其频次,绘制出关键词云图(图 2),可以直观地显示 2017 年广东省期刊编辑的研究热点:科技期刊、学术期刊、广东、青年编辑、大数据、编辑加工、高校学报、互联网 + 等。

图 2　2017 年广东省期刊编辑发文的关键词云图

2.2　各省市期刊编辑的发文量比较　通过数据挖掘统计出近 20 年来以及 2017 年度期刊编辑部编辑发表学术论文的年度总发文量(图 3),图中按照 20 年来的总发文量降序排列。

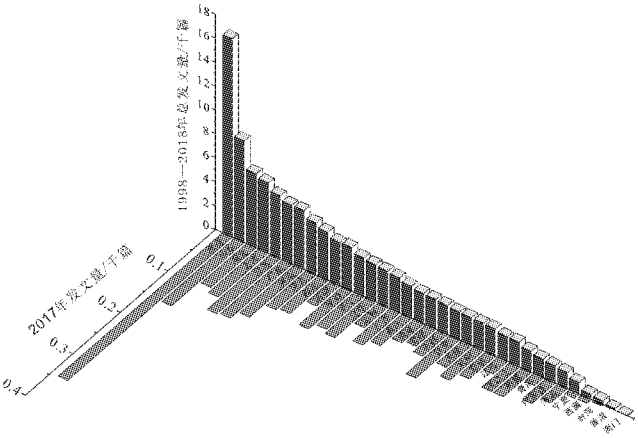

图 3　近 20 年及 2017 年度各省市期刊编辑的总发文量

期刊编辑的总发文量排名前10位的地区为:北京(1万6 548篇)、山东(8 470篇)、湖北(6 176篇)、上海(5 862篇)、江苏(5 228篇)、广东(4 900篇)、河南(4 897篇)、河北(4 325篇)、四川(3 987篇)、湖南(3 600篇);2017年度期刊编辑发文量排名前10位的地区分别为:北京(346篇)、山东(153篇)、江苏(133篇)、广东(130篇)、河南(115篇)、上海(105篇)、湖北(95篇)、陕西(91篇)、辽宁(86篇)、河北(84篇)。

2.3 各省市期刊编辑的总发文量变化趋势 通过对近20年来全国各省市期刊编辑的年度发文量(图4)分析发现,无论是全国总发文量还是各省市的期刊编辑发文情况,均表现为以2011年为分水岭分为2个阶段:上升期、回落期。全国期刊编辑年度发文量在2009年飞跃式增加,在2011年达到峰值(9 229篇),随后在2012年、2015年出现2次断崖式下降。北京地区期刊编辑的年度发文量下降趋势类似。

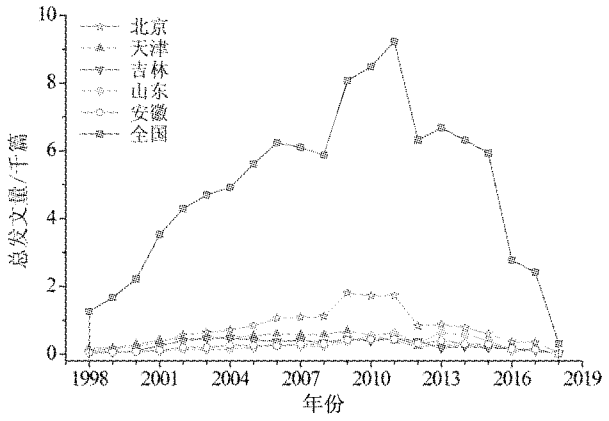

图4 近20年来全国及前5位省市期刊编辑的发文量

为了分析期刊编辑发文量回落的原因,笔者查阅了国家新闻出版管理部门在2009—2017年发布的新闻出版产业分析报告数据,绘制期刊出版从业人数的年度变化趋势如图5所示。

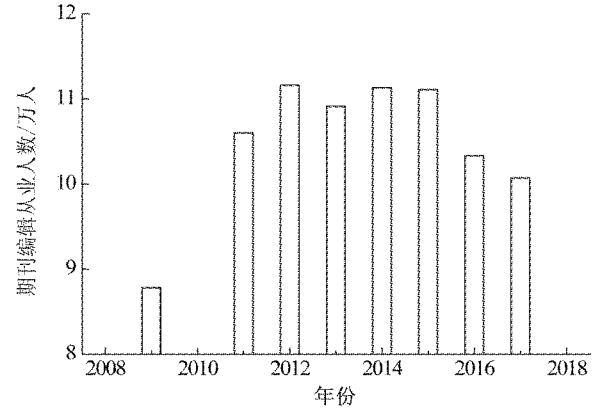

图5 近10年来期刊出版从业总人数

期刊出版从业人数在2012年前逐年上升达到峰值并在2012—2015年趋于稳定,在2016年以后开始回落。从业人数的减少或与2011年出版业转制、新媒体和自媒体的强大吸引力等因素引起的出版业3次"离职潮"有关;因此,期刊编辑发文量在2011年的迅速回落主要是由期刊编辑从业人数大幅减少引起的。

3 结论

本文基于万方数据库,利用Python编程对近20年来全国各省市期刊编辑发表论文的文本大数据进行挖掘和分析。期刊编辑利用大数据思维和技术,从事编辑与出版学领域的研究,告别传统的原始的人工检索获取方式,一方面可以大幅提高调研效率,另一方面可以大幅度增加研究的样本量。本文首次将文本挖掘技术应用于编辑出版学研究领域,挖掘的文献元数据可被用于文献计量学、统计学相关学术研究和后期参考文献格式的自动修订软件的开发。本文的研究方法为期刊编辑从事编辑与出版学研究提供了新技术、新方法和新思路。

4 参考文献

[1] 张晓倩. 数据挖掘在网络在线投稿系统中的应用[J]. 办公自动化(学术版), 2013(8): 36

[2] 欧阳柏成. 基于HADOOP的数据挖掘技术研究[J]. 信息与电脑, 2015(16): 80

[3] 张欣欣, 缪弈洲, 张月红. CrossRef文本和数据挖掘服务:《浙江大学学报》(英文版)的实践[J]. 中国科技期刊研究, 2015, 26(6): 594

[4] 王秀芝, 宋迎法. 基于文本数据挖掘的学术期刊选题策划研究[J]. 煤炭高等教育, 2016, 34(5): 123

[5] 杨静, 程昌秀. 文献"大数据"分析软件Citespace和SCI2的对比分析研究[J]. 计算机科学与应用, 2017, 7(6): 580

[6] 侯丽珊. 基于数据挖掘的精准化办刊策略[J]. 中国科技期刊研究, 2018, 29(5): 515

[7] 车尧, 宋扬, 李兵. 基于题录信息分析的期刊数据研究:以《情报学报》为例[J]. 中国科技期刊研究, 2018, 29(4): 406

[8] 王志鸿, 杨松迎, 郭敏, 等. 基于微信平台的科技期刊内容服务策略及实现[J]. 编辑学报, 2018, 30(5): 522

[9] 李雪, 王占坤, 崔晓健, 等. 科技期刊编辑新媒体出版能力的培育[J]. 编辑学报, 2016, 28(6): 602

[10] 谭春林. QQ群消息及"代笔"交易的挖掘与学术不端诱因分析[J]. 中国科技期刊研究, 2019, 30(7):721

(2018-12-26收稿;2019-04-16修回)