doi:10.3969/j.issn.1672-1292.2022.01.006

基于爬虫和统计技术的智能科研辅助系统

隆征帆[1],杨　柳[1,2],张　星[1]

(1.湘潭大学数学与计算科学学院,湖南 湘潭 411105)
(2.湘潭大学科学工程计算与数值仿真湖南省重点实验室,湖南 湘潭 411105)

[摘要]　设计了一种宏观与微观相结合的文献分析基本框架. 首先,基于统计学的平均思想提出了一种文献质量指标评价体系. 然后,基于爬虫和统计技术并借助于 Python 编程语言丰富而强大的标准库和第三方库,构建并编程实现了一个能完成文献自动收集和分析的智能科研辅助系统. 实验结果表明,用户输入检索条件后,系统能自动收集中国知网上相关文献信息并快速有效地向用户呈现一份图文并茂的文献分析报告.

[关键词]　科研辅助系统,Python,爬虫,文献收集和分析,指标评价体系
[中图分类号]TP274;TP391　[文献标志码]A　[文章编号]1672-1292(2022)01-0039-07

Intelligent Research Support System Based on Crawler and Statistical Technology

Long Zhengfan[1], Yang Liu[1,2], Zhang Xing[1]

(1.School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China)
(2.Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan 411105, China)

Abstract: This paper designs a basic framework of literature analysis which combines macro-analysis and micro-analysis. Firstly, it puts forward a literature quality index evaluation system based on the average idea of statistics. Based on crawler and statistical technology and with the help of rich and powerful standard library and third party library of Python programming language, an intelligent research support system which can automatically collect and analyze documents is constructed and programmed. After the user enters the search conditions in the graphical user interface provided by the system, the system will automatically collect the relevant literature information on CNKI, and quickly and effectively present a literature analysis report with both pictures and texts to the user.

Key words: research support system, Python, crawler, literature collection and analysis, index evaluation system

文献的收集和分析是许多科学研究工作的先导步骤,但经常会耗费大家很多时间和精力. 构建一个智能化的计算机辅助系统来完成文献的自动收集和分析具有十分重要的实际意义.

1　相关工作

随着计算机技术的高速发展,近十多年来人们在文献分析系统方面做了不少研究工作. 王曰芬等[1]开发了一个可对期刊发文量、关键词等作统计分析的软件. 张满年等[2]提出了构建科技期刊评价分析系统的思路. 姜春林等[3]基于 CSSCI 文献构建了一个文献数据共现矩阵生成软件. 谭淑琴[4]提出了构建基于自建数据库的文献自动计量分析系统的设计原则和功能描述. 龙海燕[5]结合复杂网络分析构建了一个用于文献分析与可视化的 RIA. 赵斌[6]构建了一个基于 GraphOLAP 的文献分析与可视化系统. 邢美凤等[7]设计并实现了一个基于多种文献数据库的共词聚类可视化分析系统. 李国俊等[8]实现了一个包含数据上传、基本计量信息可视化和词频分析可视化 3 个模块的文献计量可视化软件. 满俊麟[9]设计了一个

收稿日期:2021-08-31.
基金项目:国家自然科学基金面上项目(12071399);国家社科基金年度项目(20BTQ105);湖南省教育厅重点项目(18A048);湖南省学位与研究生教育改革研究项目(2019JGYB109).
通讯作者:杨柳,博士,教授,研究方向:大数据与优化. E-mail:yangl410@xtu.edu.cn

实现了文献管理、词频统计、共词分析、聚类分析和文献查询等功能的文献分析系统;佘玉轩等[10]开发了一个基于 Medline 的生物医学文献分析系统;周超峰[11]系统比较了多个国内外文献分析软件;李信等[12]考虑文献中词汇的语义功能,设计和实现一个基于词汇功能识别的科研文献分析系统;祝文君[13]构建了一个科研文档主题分析与主题推荐系统. 但在已有的研究中,我们还没找到一个可以实现文献自动收集和分析的智能化科研辅助系统. 为此,我们设计了以下全自动文献收集和分析系统.

2　系统的构建

文献分析可分为宏观分析和微观分析两个层面:在宏观层面,系统帮助用户了解检索主题下的研究概貌;在微观层面,系统帮助用户了解检索主题下的具体研究内容. 我们从文献时间分布、作者分布、学科分布和关键词分布这 4 个角度来刻画研究概貌. 其中时间分布帮助用户了解研究热度的变化趋势,作者分布帮助用户了解哪些人在做相关研究,学科分布帮助用户了解研究属于哪些学科领域,关键词分布帮助用户了解研究涉及的主题和方法. 由于许多检索主题下文献数量成千上万,向用户展示所有文献的具体研究内容是不合适的,应滤过质量较低的文献而只向用户展示质量较高的文献. 然而文献质量的评价并没有统一的标准. 引用率高的文献质量不一定高,引用率低的文献质量不一定低. 为此我们使用统计学中最基本的却有着广泛应用的平均思想,提出了一种综合考虑多个方面的文献质量指标评价体系. 设 C 代表文献引用量,D 代表文献下载量,Y_s 代表文献发表年份,Y_e 代表明年年份,C_r 代表文献刊物总引用量,D_r 代表文献刊物总下载量,R_n 代表刊物文献总量,N_a 代表文献作者量,M_a 代表作者文献量,N_r 代表文献参考文献量,N_c 代表文献引证文献量. 表 1 给出了评价指标体系的构成和计算公式.

表 1　评价指标体系的构成和计算公式

Table 1　Composition and calculation formula of index evaluation system

一级指标	二级指标	计算公式
文献影响力	文献年均引用量	$\bar{C}=C/(Y_e-Y_s)$
	文献年均下载量	$\bar{D}=D/(Y_e-Y_s)$
文献作者影响力	文献作者篇均、年均引用量	$\frac{1}{N_a}\sum_{i=1}^{N_a}\frac{1}{M_a^{(i)}}\sum_{j=1}^{M_a^{(i)}}\bar{C}_j^{(i)}$
	文献作者篇均、年均下载量	$\frac{1}{N_a}\sum_{i=1}^{N_a}\frac{1}{M_a^{(i)}}\sum_{j=1}^{M_a^{(i)}}\bar{D}_j^{(i)}$
文献刊物影响力	文献刊物篇均引用量	C_r/R_n
	文献刊物篇均下载量	D_r/R_n
文献参考文献影响力	文献参考文献篇均、年均引用量	$\frac{1}{N_r}\sum_{i=1}^{N_r}\bar{C}^{(i)}$
	文献参考文献篇均、年均下载量	$\frac{1}{N_r}\sum_{i=1}^{N_r}\bar{D}^{(i)}$
文献引证文献影响力	文献引证文献篇均、年均引用量	$\frac{1}{N_c}\sum_{i=1}^{N_c}\bar{C}^{(i)}$
	文献引证文献篇均、年均下载量	$\frac{1}{N_c}\sum_{i=1}^{N_c}\bar{D}^{(i)}$

从表 1 可以看到,我们设计的文献质量评价指标体系综合考虑了文献影响力、文献作者影响力、文献刊物影响力、文献参考文献影响力和文献引证文献影响力. 表中一级指标是其对应二级指标的平均. 为了消除量纲差异,所有二级指标均需作标准化处理. 我们把一级指标平均值排名靠前的文献推荐给用户,向用户展示这些文献的发表时间、作者、发表刊物、关键词和摘要信息. 考虑到定量评价的局限性,我们还会从未被定量推荐的文献中随机选择一定比例的文献展示给用户. 基于上述的文献分析基本框架以及文献质量评价指标体系,我们构建的系统由爬虫(crawler)、预处理(preprocessor)、分析(analyzer)、图形化界面(gui)和日志(logger)5 大模块构成. 图 1 展示了系统的架构.

其中 crawler 模块负责文献信息的收集,包括文献自身、文献作者、文献刊物、文献参考文献和文献引证文献的信息,这些信息将服务于文献分析. 经 crawler 模块获取的信息将通过 preprocessor 模块进行数据

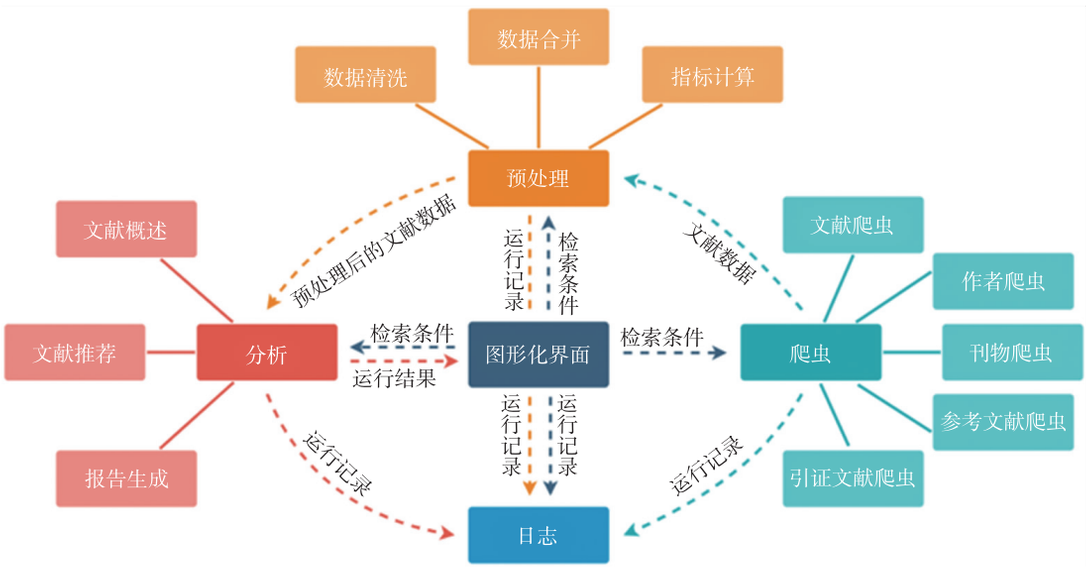


图 1　系统的架构

Fig. 1　Architecture of the system

清洗、数据合并和指标计算. 预处理后的文献数据将流经 analyzer 模块,该模块由 3 个子模块构成,其中文献概述子模块负责生成宏观层面分析结果,文献推荐子模块负责生成微观层面分析结果,报告生成子模块用来生成文献分析报告. Gui 模块用来向用户提供方便访问系统的接口,用户通过图形化界面提供的检索条件将作为整个系统的输入. Logger 模块则会记录其他 4 大模块的运行状况,为系统维护提供便利.

系统的所有模块均由 Python 实现. 其中 crawler 模块主要使用第三方异步网络请求库 aiohttp、多进程标准库 multiprocessing 和 HTML 页面第三方解析库 lxml,借助于多进程异步爬虫,可实现 50-200 篇/秒的文献数据获取速度. Preprocessor 模块主要使用第三方数据处理库 pandas 和提供正则表达式支持的文本处理标准库 re. Analyzer 模块主要使用 pandas、re、提供 Echarts 统计图形绘制 Python 接口的第三方库 pyecharts、提供图像处理的标准库 pillow、提供多维数组运算的第三方库 numpy、提供基本数学运算的标准库 math 以及提供 Word 文档生成的第三方库 python-docx. Gui 模块主要使用提供图形化界面绘制的标准库 tkinter. Logger 模块主要使用提供时间处理功能的标准库 datetime 和提供操作系统相关功能的标准库 system. 图 2、图 3 分别展示了系统的目录结构和图形化界面.

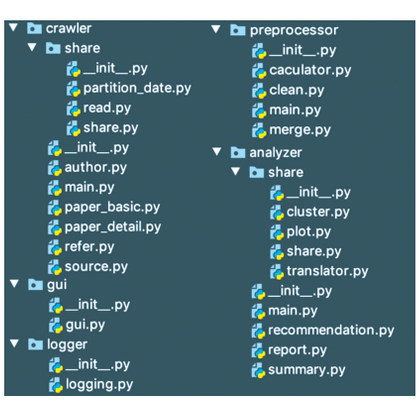


图 2　系统的目录结构

Fig. 2　Directory structure of the system

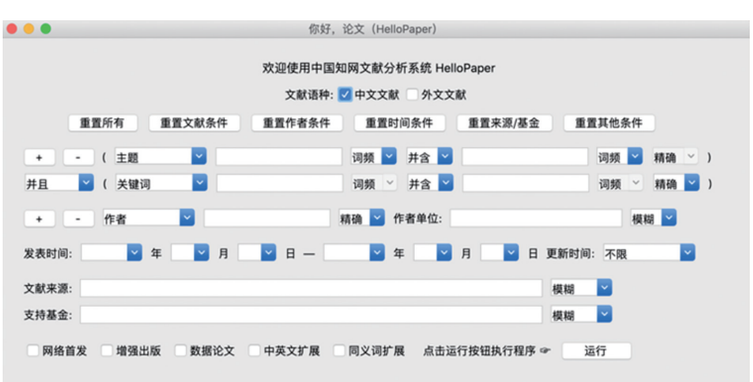


图 3　系统的图形化界面

Fig. 3　Graphic user interface of system

用户在图 3 给出的图形化界面输入检索条件后,系统会将检索条件自动传入 crawler 模块,crawler 模块随后开始自动获取中国知网上相关文献数据,获取结束后的文献数据将自动流经 preprocessor 模块和 analyzer 模块的文献概述子模块和文献推荐子模块作预处理和分析,最终由 analyzer 模块报告生成子模块自动汇总文献概述子模块得到的宏观分析结果和文献推荐子模块得到的微观分析结果,并将这些结果输出为一份图文并茂的 Word 文献分析报告.

3　系统的应用

本文以中国知网收录的新型冠状病毒肺炎相关中外文献为对象来阐述系统的应用. 针对中文文献的检索条件为:(主题="新冠肺炎"或含"新型冠状病毒肺炎")或者(主题="新冠病毒"或含"新型冠状病毒")、(发表时间="2019-11-01"—"2020-05-11"). 针对外文文献的检索条件为:(主题="2019-nCoV"或含"SARS-CoV-2")或者(主题="COVID-19"或含"Corona Virus Disease 2019")、(发表时间="2019-11-01"—"2020-05-11").

系统给出的文献分析报告由引言、文献概述和文献推荐部分组成. 图 4 是引言部分截图. 可见,系统会列出用户使用的检索条件并告知用户与检索条件匹配的文献总量以及这些文献的来源分布. 此外,系统还会从总体上向用户介绍文献分析报告的结构.

图 5 是文献概述部分截图. 可见,系统会首先介绍文献概述部分的基本构成并简要说明不同部分对于把握研究现状的意义,接着系统会依次介绍文献的时间、作者、学科和关键词分布并进一步说明这些分布对于把握研究现状的意义.

图 6 是文献推荐部分截图. 可见,系统会从检索获得的文献中依据指标评价结果筛选出排名靠前的文献推荐给用户,同时也会随机筛选出一部分文献推荐给用户,所有文献均添加了到中国知网对应文献详情页面的超链接.

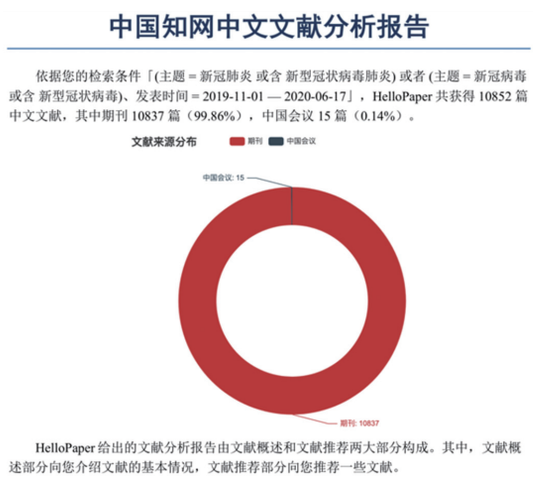


图 4　新型冠状病毒肺炎文献分析报告引言部分截图

Fig. 4　Screenshot of introduction section in the literature analysis report of COVID-19

通过阅读系统给出的新型冠状病毒肺炎中外文文献分析报告(以下简称分析报告),可以对新型冠状病毒肺炎的国内外研究现状形成一个整体性的认识. 图 7 是分析报告给出的时间分布图. 新型冠状病毒肺炎中文文献从 2 月开始出现爆炸性增长,在 3 月达到峰值,在 4 月开始下降,到 5 月已下降到 2 月水平之下. 与之不同的是,新型冠状病毒肺炎外文文献一直持续增长,并且增速越来越快. 国内外月度发文量变化趋势的显著差异,也从一个侧面反映了国内外疫情发展态势的迥异.

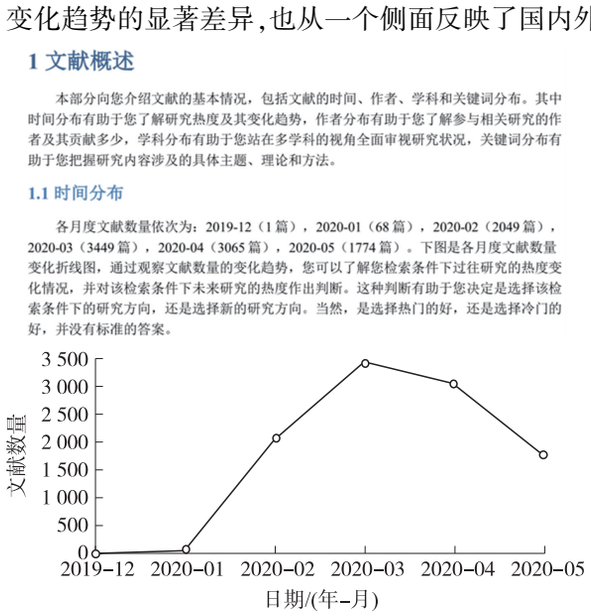


图 5　新型冠状病毒肺炎文献分析报告文献概述部分截图

Fig. 5　Screenshot of summary section in the literature analysis report of COVID-19

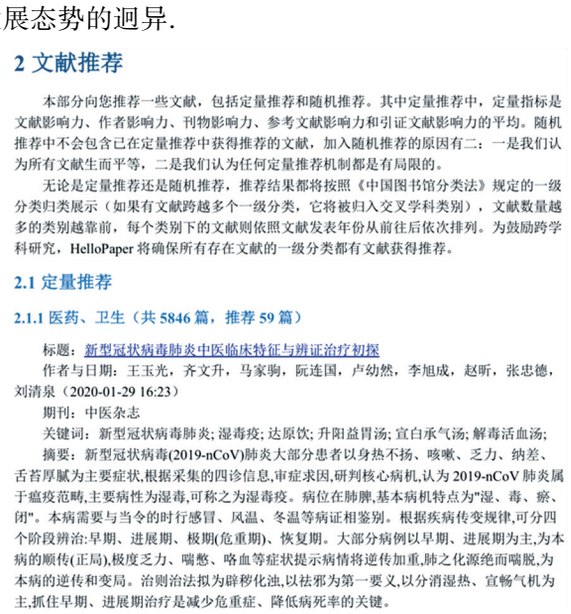
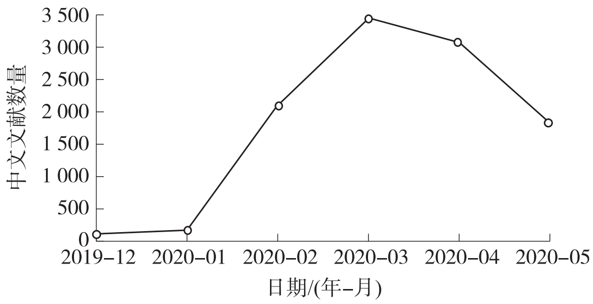
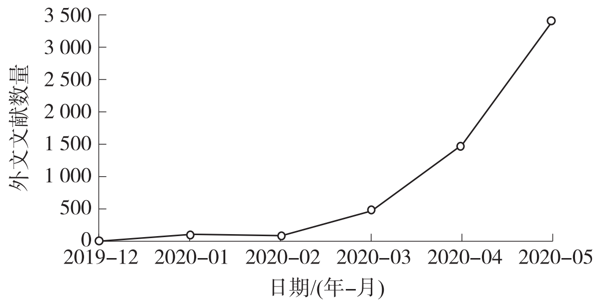


图 6　新型冠状病毒肺炎文献分析报告文献推荐部分截图

Fig. 6　Screenshot of recommendation section in the literature analysis report of COVID-19



(a) 中文文献数量



(b) 外文文献数量

图 7　新型冠状病毒肺炎中文文献与外文文献时间分布

Fig. 7　Distribution of publish time of Chineses literatue and foreign literatue about COVID-19

图 8 是分析报告给出的作者分布图. 汪晖、张铭、刘蕊是新型冠状病毒肺炎中文文献量最多的 3 位作者. 通过分析报告给出的文字描述可以知道他们的发文量达到或超过 20 篇,通过点击分析报告内置的超链接可知,汪晖来自于华中科技大学同济医学院附属同济医院,张铭、刘蕊来自于第四军医大学口腔医院. Donny Jackson、Brandi Vincent、Viroj Wiwanitkits 是新型冠状病毒肺炎外文文献量最多的 3 位作者,通过分析报告给出的文字描述可知,他们的发文量达到或超过 20 篇,通过查询作者信息可知,Donny Jackson 为无线通信领域期刊的编辑,Brandi Vincent 为 Nextgov 的通讯作者之一,Viroj Wiwanitkits 为海南医学院教授.

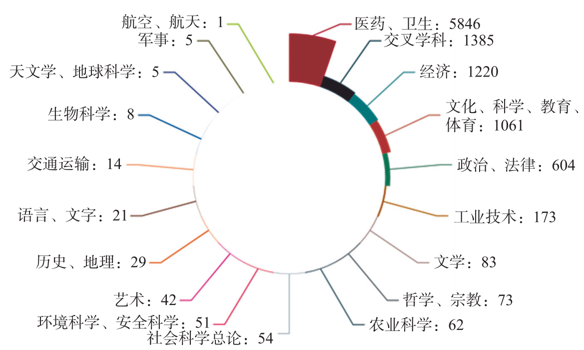
图 9 是分析报告给出的学科/期刊分布图. 结合分析报告给出的文字描述可以知道新型冠状病毒肺炎中文文献覆盖《中国图书馆分类法》规定的 22 个一级分类中的 18 个,另有多篇文献覆盖多个学科. 其中医药、卫生(5846 篇,54.45%),交叉学科(1385 篇,12.90%),经济(1220 篇,11.36%)是占比前三的领域,总占比 78.71%. 新型冠状病毒肺炎外文文献同样覆盖多个学科并主要集中在医药卫生领域,发文排名前三的是世界顶级医学期刊 The Lancet(139 篇)、传染病领域期刊 Journal of Infection(85 篇)和 International Journal of Infectious Diseases(77 篇).

(a) 中文文献作者分布

(b) 外文文献作者分布

图 8　新型冠状病毒肺炎中文与外文文献作者分布

Fig. 8　Distribution of author of Chineses literatue and foreign literatue about COVID-19



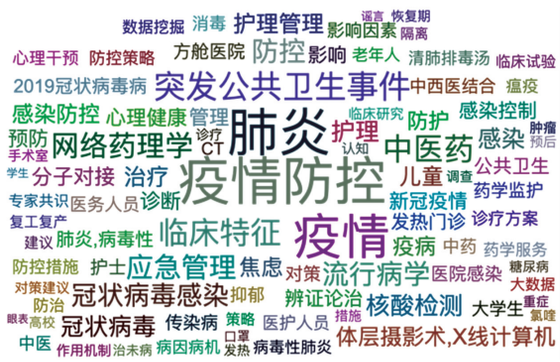
(a) 中文文献学科/期刊分布

(b) 外文文献学科/期刊分布

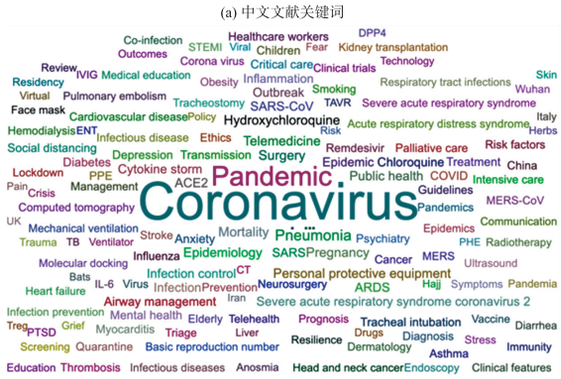
图 9　新型冠状病毒肺炎中文与外文文献学科/期刊分布

Fig. 9　Distribution of subject/journal of Chineses literatue and foreign literatue about COVID-19

图 10 是分析报告给出的关键词分布图. 结合分析报告给出的文字描述可知,除去与“新冠肺炎”“新冠病毒”相近或相同的关键词后,疫情防控(280 次)、肺炎(279 次)、疫情(255 次)、突发公共卫生事件(146 次)、临床特征(137 次)、中医药(136 次)、网络药理学(122 次)、应急管理(116 次)、流行病学(113 次)和防控(113 次)成为中文文献排名前十的关键词, Coronavirus (339 次)、Pandemic (136 次)、Pneumonia(39 次)、Mortality(25 次)、Surgery(25 次)、Pregnancy(24 次)、Telemedicine(24 次)、Epidemiology(23 次)、Hydroxychloroquine (23 次) 和 Personal protective equipment(23 次)成为外文文献排名前十的关键词. 可见在新型冠状病毒肺炎中文文献中,如何做好疫情防控、中医药在新型冠状病毒肺炎治疗中的应用、如何应用网络药理学提供的新的药物研发模式快速研制新药以及新型冠状病毒肺炎的临床特征等是研究者们最为关心的问题,而在新型冠状病毒肺炎外文文献中,作者们重点从疾病传播、疾病特征、疾病治疗、个人防护等角度展开研究.



(a) 中文文献关键词



(b) 外文文献关键词

图 10　新型冠状病毒肺炎中文与外文文献关键词分布

Fig. 10　Distribution of keywords of Chineses literatue and foreign literatue about COVID-19

从分析报告推荐的具体文献来看,研究者们从多个角度围绕“新型冠状病毒肺炎”疫情展开了研究. 例如:在医药卫生领域:夏文广等[14]通过分析新型冠状病毒肺炎患者的临床资料,认为中西医结合治疗新型冠状病毒肺炎的效果优于单纯的西药治疗. 在生物学领域:陈嘉源等[15]对新型冠状病毒作了生物信息学分析,在国际上首次从分子水平解释了 BB 冠状病毒变异快、宿主多且具有较强宿主适应性的原因. 在心理学领域:Consolo 等[16]研究了新型冠状病毒肺炎疫情暴发后意大利摩德纳和雷吉奥埃米利亚地区的牙医的心理反应,研究发现疫情给牙医带来了非常负面的影响,超过 80%的牙医对自己的职业前途表示担忧. 在经济领域:李柳颖等[17]以家庭为对象研究了新型冠状病毒肺炎疫情对居民消费行为的影响及这种影响的形成机制. 在教育领域:徐瑾劼[18]结合 OECD 关于全球教育系统疫情应对的调研,PISA 2018 和 TALIS 2018 的数据,比较了受疫情影响严重国家开展在线教育的资源及能力. 在数学领域:Zhang 等[19]利用七国集团中 6 个国家的数据构建了分段泊森模型,该模型能够预测这些国家的疫情转折点、疫情持续时间和发病率.

4　结论

本文构建了一个实现了文献自动收集和分析的科学研究辅助系统,设计了一种宏观与微观相结合的文献分析基本框架,提出了一种新的综合考虑多个方面的文献质量评价指标体系. 实际应用表明本文构建的系统可以辅助科研人员进行科学研究工作.

[参考文献](References)

[1]　王曰芬,颜端武,路菲. 文献计量与内容分析综合应用软件的开发与实验[J]. 图书情报工作,2005,49(6):24-28.

[2]　张满年,曾建勋. 基于网络的科技期刊评价分析系统的构建[J]. 中国科技期刊研究,2008,19(5):729-732.

[3]　姜春林,杜维滨,李江波. CSSCI 文献数据共现矩阵的软件实现[J]. 情报理论与实践,2008,31(6):937-940.

[4]　谭淑琴. 基于自建数据库的文献自动计量分析系统研究[J]. 现代情报,2009,29(8):164-165.

[5]　龙海燕. 基于 RIA 的科技文献分析与可视化系统的研究与实现[D]. 北京:北京邮电大学,2010.

[6]　赵斌. 基于 GraphOLAP 的文献分析与可视化系统的研究与实现[D]. 北京:北京邮电大学,2011.

[7] 邢美凤,许德山. 可视化的共词聚类系统分析及实现[J]. 现代图书情报技术,2011(Z1):62-67.
[8] 李国俊,刘恩涛,肖明. 文献计量可视化软件的分析与实现[J]. 图书馆杂志,2011,30(10):72-78.
[9] 满俊麟. 文献信息分析系统的设计与实现[D]. 大连:大连理工大学,2013.
[10] 佘玉轩,熊赟. Medas:一个基于 Medline 的生物医学文献分析系统[J]. 计算机研究与发展,2015,52(Suppl 1):102-106.
[11] 周超峰. 文献计量常用软件比较研究[D]. 武汉:华中师范大学,2017.
[12] 李信,程齐凯,刘兴帮. 基于词汇功能识别的科研文献分析系统设计与实现[J]. 图书情报工作,2017,61(1):109-116.
[13] 祝文君. 基于科研文档的主题分析与推荐系统[D]. 武汉:华中科技大学,2019.
[14] 夏文广,安长青,郑婵娟,等. 中西医结合治疗新型冠状病毒肺炎 34 例临床研究[J]. 中医杂志,2020,61(5):375-382.
[15] 陈嘉源,施劲松,丘栋安,等. 2019 新型冠状病毒基因组的生物信息学分析[J]. 生物信息学,2020,18(2):96-102.
[16] CONSOLO U,BELLINI P,BENCIVENNI D,et al. Epidemiological aspects and psychological reactions to COVID-19 of dental practitioners in the northern Italy districts of modena and reggio emilia[J]. International Journal of Environmental Research and Public Health,2020,17(10):3459.
[17] 李柳颖,武佳藤. 新冠肺炎疫情对居民消费行为的影响及形成机制分析[J]. 消费经济,2020,36(3):19-26.
[18] 徐瑾劼. 新冠肺炎疫情下全球教育体系的应对与在线教育的挑战——基于 OECD 全球调研结果的发现与反思[J]. 比较教育研究,2020,42(6):3-10.
[19] ZHANG X,MA R,WANG L. Predicting turning point,duration and attack rate of COVID-19 outbreaks in major Western countries[J]. Chaos,Solitons and Fractals,2020,135:109829.

[责任编辑:陈　庆]