·基础理论与方法·

# Python爬虫技术在信息流行病学中的应用

周江杰　王胜锋　李立明
北京大学公共卫生学院流行病与卫生统计学系 100191
周江杰和王胜锋对本文有同等贡献
通信作者：李立明，Email：lmlee@bjmu.edu.cn

【摘要】 Python网络爬虫技术是一种通过模拟用户的网络浏览行为以实现从网络中自动、大量提取信息的技术，是信息流行病学研究收集并整合多源异构信息数据的关键基础。Python网络爬虫可分为简单爬虫与大型爬虫，集数据采集与数据库构建于一体，语法简洁、灵活性高、学习成本低、维护成本低。它适用于信息流行病学的各种应用场景，通过对互联网中健康相关信息的分析，实现多种公共卫生监测、健康干预实施及效果评价、智慧寻医方略优化等目标。近年，我国政府开始鼓励对含互联网信息在内的多源大数据的整合利用，在此背景下，Python爬虫技术的应用场景势必会越来越多，相应的人才培养、技术革新建议纳入到公共卫生教育和科研体系之中。

【关键词】 Python爬虫技术；信息流行病学；公共卫生监测；健康干预；智慧寻医
DOI：10.3760/cma.j.cn112338-20190901-00643

**Application of Python web crawler technology in infodemiology**

*Zhou Jiangjie, Wang Shengfeng, Li Liming*
*Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China*
*Zhou Jiangjie and Wang Shengfeng contributed equally to the article*
*Corresponding author: Li Liming, Email: lmlee@bjmu.edu.cn*

**【Abstract】** Python web crawler technology, which automatically and massively getting information from the Internet by mimicking net users' browsing behavior, is a basic supporting technique to extract and integrate multi-source heterogeneous data in the field of Infodemiology. There are two types of Python web crawler: simple and massive-scale, both collect information simultaneously from the database establishment. Advantages of this technique are characterized as: being simple syntax, in high flexibility and low cost in learning and maintenance. Contents of the current application scenarios include surveillance, implementation and evaluation of health intervention programs on public health issues, as well as on smart doctor seeking. For the last two years, the Chinese government started to encourage the integration and utilization of multi-source heterogeneous data including internet information. Hence, the number of application scenarios for Python web crawler technology are bound to increase in the foreseeable future. Corresponding matched talent cultivations and technical innovations are suggested to add to the current education and research systems on public health issues.

**【Key words】** Python web crawler technology; Infodemiology; Public health surveillance; Health intervention; Smart doctor seeking
DOI:10.3760/cma.j.cn112338-20190901-00643

近年来，随着现代通讯技术的发展和通讯设备的广泛普及，互联网信息规模呈现爆炸式增长。医学领域研究者尝试利用互联网中与健康相关的信息，研究其发生、分布及影响因素，以期为防治疾病、促进健康的策略制定提供科学依据，进而形成了新的流行病学分支学科——信息流行病学（infodemiology）[1-4]。但互联网信息往往是零散且非结构化的，如何将研究所关注的目标信息完整、及时、高效的收集并整合为结构化可利用的数据，是信息流行病学研究开展中至关重要的环节。Python网络爬虫技术是一种通过模拟用户的网络浏览行为以实现从网络中自动、大量提取信息的技术[5]，可满足上述环节要求的同时又兼具轻量易用等特点，已得到广泛运用。为更好推动信息流行病学研究在我国健康领域的应用和发展，本文对Python网络爬虫技术的技术实现及应用领域进行概述介绍。

一、Python网络爬虫的概念和特点

网络爬虫（web spider或web crawler）指通过编写程序模拟浏览器上网，在互联网上抓取数据的过程。常用搜索引擎如Baidu、Google、Yahoo等，都是依托网络爬虫完成每次检索任务，把结果从每个符合要求的服务器端抓到本地，呈现给用户。网络爬虫可以用Java、C＋＋和Python等语言实现。和其他语言相比，Python语法更为简单，代码优美，支持模块多，学习成本低，并且已有强大的爬虫框架（scrapy），更适合进行爬虫开发。

二、Python网络爬虫的技术实现

根据适用场景，Python网络爬虫可分为简单爬虫与大型爬虫。一般来讲，当目标网站待爬取网页数≤100万时推荐使用简单爬虫，≥1 000万时推荐使用大型爬虫。

简单爬虫的基本流程：①根据需求分析目标网站和网页的结构；②编写地址提取程序，提取待爬取的网页地址以备后续网页下载程序调用；③网页下载程序获取地址，完成目标网页的下载；④网页解析程序调用所有目标网页，抓取每个网页中所包含的目标数据；⑤数据存储程序完成数据的写入存储。简单爬虫的核心为网页解析程序（对应lxml、BeautifulSoup包）、网页下载程序（对应requests包）和数据存储程序（对应pymysql包）。上述流程最好使用Python的多线程功能（对应threading包）实现异步执行，以提高效率。

大型爬虫在简单爬虫的基础上，采用分布式集群方法，将多个运行简单爬虫的计算机进行整合，共同完成爬取任务。集群通常由一台主服务器（master）和若干从服务器（slave）组成。主服务器负责维护任务队列并向从服务器分配抓取任务，从服务器负责运行简单爬虫，并将新提取出的网页地址反馈给主服务器，从而指导主服务器的进一步任务分配，如此循环自动进行（图1）[6]。与简单爬虫相比，大型爬虫的学习成本相对较高，且需要有服务器集群支持，二者在单台计算机上的运行速度没有区别。

三、Python网络爬虫在公共卫生实践中的应用

信息流行病学通过分析互联网中健康相关信息，可以实现多种公共卫生监测、健康干预实施及效果评价、智慧寻医方略优化等目标，而Python网络爬虫可满足上述应用场景。

1. 公共卫生监测：可分为疾病、症状、行为及危险因素和其他监测，Python网络爬虫可用于全部监测类型之中。相比传统监测受限于规模、时间、成本等因素，基于网络的监测具备对象范围广、内容隐私保护好、效率实时性强、信息获取更容易等特点[7]。目前，Python网络爬虫在公共卫生监测中多用在传染病领域，主要探讨目标关键词的搜索量变化趋势能否真实反映传染病的波动趋势[8]。国外自2006年起，陆续有多位研究者尝试借助网络平台进行流感、流感样症状、埃博拉病毒病、中东呼吸综合征等疾病或症状预测[2,7,9-15]，我国自2013年起也有研究者探讨利用搜索引擎进行流感监测的可能性[16]。

也有研究尝试将Python网络爬虫用于慢性病监测，如肿瘤、心脏病、系统性红斑狼疮、风湿性关节炎等[4,17]。利用Python网络爬虫技术开展的监测研究，其数据来源一般为搜索引擎和社交网络。搜索引擎使用的是网络用户所需要的信息，涉及的数据总量可达数千亿条[9]，且搜索引擎用户的搜索请求数据属于内部数据，通常并不对外开放。但进行简单估

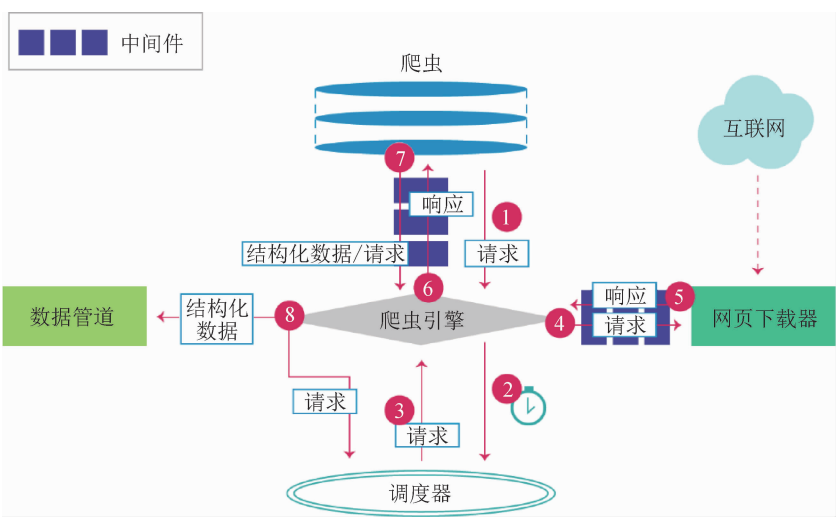


**图1** 大型爬虫的工作流程示意图

计时，可直接利用搜索引擎提供的“百度指数”“Google Trends”等功能[4,14,16,18-19]。社交网络使用的是网络用户产生的信息，涉及的数据总量相对前者较少，但也达到数十万到数百万条文本，但公开可及容易获取。无论哪种数据来源，研究思路都是抓取搜索记录或目标文本、确定关键词、分析关键词频率。此外，社交媒体相关研究还可开展内容分析。

目前研究人员基于网络爬虫技术构建了多种预测不同疾病发病或死亡趋势的模型，一定程度拓宽了公共卫生监测的视野。但由于数据来源及方法学的局限，导致这类研究仍处于探索层面，无法替代传统监测。从数据来源看，互联网中疾病或症状相关信息的产生或需求与疾病或症状的发生并非绝对映射关系，易受媒体引导和宣传的影响[14,20-21]。从方法学角度看，由于网络用户所使用的关键词常常比已知的疾病或症状的专业术语更为复杂多样[22]，导致搜索关键词时难以穷尽，从而难以避免检索不完整的问题[11,14]。从推广角度看，此类研究依托于信息化的高度发展，导致部分发展中国家应用受限[14,23]。同时，已有研究主要采用回顾性设计，尚未前瞻性地检验模型的真实预测能力也会影响相关研究的关注度[10,13,20,24-25]。但相关领域的方法学在不断完善与发展，如从数据来源角度尝试不同层面的整合[13,26]，从分析方法上更多地尝试机器学习技术等[10-11]。

此外，也有研究者利用Python网络爬虫探索信息流行病学对药物利用、药品不良反应的监测价值。Rastegar-Mojarad等[27]基于WebMD网站64 616条与180种药物相关评论，利用2种数据挖掘技术，发现了5种药物新的使用领域。Freifeld等[28]围绕23种选定药物，证实Twitter上相关不良反应事件数量与美国食品药品监督管理局不良事件报告系统对应数量明显正相关，后续研究在此基础上发展出了使用更少训练样本取得更高不良反应识别率的递归神经网络模型[29]，为此类研究提供了新的技术手段支持。上述探索为研究药品的超适应症使用、非法药物使用及扩适应症研究提供了新的思路，但因为这些研究几乎都局限在某类媒体数据基础上，其外推性缺少对各网络平台效果的对比证据，且存在选择性偏倚和关键信息缺失等问题[30]。

Python网络爬虫还可以用于卫生相关舆情的监测，其思路与其他舆情监测领域一致，有利于保证卫生管理机构对公众关心的热点问题做出及时的、有针对性的响应[18,31-33]。但解读这类研究结果时，需要注意样本的代表性，即网络用户并不能代表全部公众，例如2018年我国网络普及率为59.6%，尚有部分经济欠发达地区无法接触网络[34]。

2. 健康干预实施及评价：互联网信息不仅可作为信息流行病学的数据来源，也可作为健康干预手段开展干预性研究。对互联网相关信息进行分析，不仅可以描述健康信息的传播模式及影响因素，同时也能够评价不同干预措施的效果，甚至实现疾病预测预警的作用。如今越来越多的医疗机构或医生个人甚至媒体都在启用不同的网络信息平台，网络信息也对公众决策发挥着不可忽视的影响[7,35]，但如何最大化发挥互联网健康信息的健康宣教效果也成为研究者关注的话题。已有研究发现，利用恐惧诉求的心理现象、使用长消息文本等都能增加传播效果，并提出卫生机构可与社交媒体中的意见领袖进行合作[36]。寨卡疫情和H7N9流感期间，网络媒体在健康信息的传播中都发挥了显著作用，且用户需求的健康信息种类随疫情发展而变化[18,37]，而网络谣言是影响群体接受正确健康信息的主要因素，政府应当加强与意见领袖的沟通、加快更新突发事件信息的速度，避免谣言滋生[18]。这类研究的局限性是难以验证传播信息的真实性[24]，且受限于自然语言处理等文本处理技术的水平[37]。

此外，由于公众习惯借助网络表达想法，其中就可能包含了疾病相关的信息。国外研究者尝试利用这部分信息，结合定位功能，开展线上（如定向推送）/线下干预，从而实现疾病防治或预警的目的[38-39]，但国内则尚未见此类研究。这些研究一般建立在前述监测工作构建模型的基础上，立足于个性化医疗理念，更加贴近公众生活，真正指导个体健康。由于依托于网络，研究对象参与度高、时效性好、覆盖范围广。但局限性也与依赖网络有关，如网络的匿名性常会导致难以定位到真实个体，当前与电子病历、死因监测、肿瘤登记等其他来源数据库的链接屏障导致很难验证预测结局的真实性。随着医疗资料信息化水平的提高、不同数据资料的链接打通以及网络用户信息隐私政策的针对性调整，上述问题会逐渐改善，该领域的应用也会逐渐增加。

3. 智慧寻医方略优化：我国的在线医疗行业发展迅速，市场规模持续扩大，2018年已达177亿元[40]。在线医疗服务主要包括健康管理、就医推荐（智慧寻医）、在线问诊、患者论坛、在线挂号服务等类别。智慧寻医模式的发展一定程度上满足了公众的寻医就医需求，也随之发展出了对不同在线医疗平台的服务评价。在线医疗平台本身拥有庞大的用户规模，

每天产生大量的反馈数据，有研究利用Python爬虫技术将这些文本信息结构化，从平台、医院、医生或诊疗手段等维度进行评价，实现智能推荐，还可挖掘患者对疾病隐藏的关注点[41]。国外研究结果表明，医生的评分主要受医生行为举止和诊疗技术的影响[42]，而基于NHS Choices、Yelp或Facebook的研究进一步发现，用户的推荐率、评分或点赞数与医院的标化死亡率、高危死亡率、或再入院率呈显著负相关，与医院实际质量调查评分呈正相关[43-45]。类似研究在我国也已出现，但主要集中在定性描述，如负面评价的理由、患者最关注的话题等[41,46]。这类研究的优势除了数据公开可及、调查规模庞大外，还因为网络的匿名性可促使公众更加愿意反馈真实感受[42]，但由于每个在线平台的注册用户都有一定偏性且很难穷尽所有在线平台等原因，也存在选择性偏倚的问题。

四、方法和应用展望

自21世纪初信息流行病学萌芽至今，伴随互联网信息技术和Python技术的发展，国外对于此类研究的范式和领域也在不断丰富，国内也开始出现不同形式的探索。我国拥有庞大的网络信息资源，特别是近年来移动互联网的高速发展让网络覆盖面进一步扩大，我国的社交网络用户数目已接近美国的3倍[47-48]，但国内对于数据的利用却不及美国。同时，信息流行病学研究在数据分析和结果解读方面仍存在挑战。首先，尽管信息流行病学可以使用传统统计学方法，但由于其数据来源常为非结构化文本，因此预处理阶段不可避免的需要涉及自然语言处理和机器学习，而目前这些技术在中文文本处理领域的应用仍处于发展阶段[49]。目前信息流行病学领域常用的支持向量机（support vector machine）、隐式迪利克雷分布（latent dirichlet allocation）、条件随机场（conditional random fields）、循环神经网络（recurrent neural networks）等模型也仍是学术探讨的热点。研究者往往同时测试多种模型，从中选择出最适合其特定研究场景的模型[29,41,50]。其次，信息流行病学的数据来源依托于互联网，决定了产生数据的人群年龄较轻，更多来自经济发达地区，不可避免的影响研究对象的代表性和结论的外推性[16,33]。实际运用时，对已经相对成熟的公共卫生监测，建议将Python网络爬虫定位为传统监测方式的补充。但对本身处于关注焦点的健康干预实施及评价、智慧寻医方略优化等方面，Python网络爬虫技术势必拓宽研究思维，提供新的视角，但如何与传统的流行病学研究完美结合，也是亟需解决的问题[4]。

我国政府从2018年起鼓励医疗卫生机构与互联网企业合作以整合医疗卫生信息资源，提出要用大数据分析手段预测疾病的流行趋势[51]，相关政策也明确提出要加强人口健康信息化建设，推进公共卫生大数据应用[52]。在此背景下，Python爬虫技术作为一种语法简洁、灵活性高、学习成本低、维护成本低、集数据采集与数据库构建于一体的关键技术，应用范围和场景势必会越来越广泛，相应的人才培养、技术革新也应纳入到公共卫生教育和科研体系之中。

**利益冲突** 所有作者均声明不存在利益冲突

## 参 考 文 献


[1] Eysenbach G. Infodemiology: the epidemiology of (mis) information[J]. Am J Med, 2002, 113(9): 763-765. DOI: 10.1016/S0002-9343(2)01473-0.

[2] Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance[J]. AMIA Annu Symp Proc, 2006, 2006: 244-248.

[3] Eysenbach G. Infodemiology and infoveillance-tracking online health information and cyberbehavior for public health[J]. Am J Prev Med, 2011, 40(5 Suppl 2): S154-158. DOI: 10.1016/j.amepre.2011.02.006.

[4] 潘海峰，赵婵娜，叶冬青. 信息流行病学研究进展[J]. 中华疾病控制杂志，2019，23(5)：497-500. DOI：10.16462/j.cnki.zhjbkz.2019.05.001.
Pan HF, Zhao CN, Ye DQ. Research progress in infodemiology study[J]. Chin J Dis Control Prev, 2019, 23(5): 497-500. DOI: 10.16462/j.cnki.zhjbkz.2019.05.001.

[5] Mahto DK, Singh L. A dive into Web Scraper world[C]//2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). New Delhi, India: IEEE, 2016: 689-693.

[6] Scrapy Developers. Architecture overview[EB/OL]. (2009-05-07)[2019-03-18]. https://doc.scrapy.org/en/latest/topics/architecture.html.

[7] Merchant RM, Elmer S, Lurie N. Integrating social media into emergency-preparedness efforts[J]. N Engl J Med, 2011, 365: 289-291. DOI: 10.1056/NEJMp1103591.

[8] Guy S, Ratzki-Leewing A, Bahati R, et al. Social media: a systematic review to understand the evidence and application in infodemiology [M]//Kostkova P, Szomszor M, Fowler D. Electronic Healthcare. Berlin, Heidelberg: Springer, 2012, 91: 1-8. DOI: 10.1007/978-3-642-29262-0_1.

[9] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2008, 457(7232): 1012-1014. DOI: 10.1038/nature07634.

[10] Lamb A, Paul MJ, Dredze M. Separating fact from fear: tracking flu infections on twitter[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: Association for Computational Linguistics, 2013.

[11] Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic[J]. PLoS One, 2013, 8(12): e83672. DOI: 10.1371/journal.pone.0083672.

[12] Mciver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time[J]. PLoS Comput Biol, 2014, 10(4): e1003581. DOI: 10.1371/journal.pcbi.1003581.

[13] Broniatowski AD, Dredze M, Paul MJ, et al. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study[J]. JMIR Public Health Surveill, 2015, 1(1): e5. DOI: 10.2196/publichealth.4472.

[14] Alicino C, Bragazzi NL, Faccio V, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes[J]. Infect Dis Poverty, 2015, 4: 54. DOI: 10.1186/s40249-015-0090-9.

[15] Song J, Song TM, Seo DC, et al. Social big data analysis of information spread and perceived infection risk during the 2015 middle east respiratory syndrome outbreak in South Korea[J]. Cyberpsychol Behav Soc Netw, 2017, 20(1): 22-29. DOI: 10.1089/cyber.2016.0126.

[16] Yuan QY, Nsoesie EO, Lv BF, et al. Monitoring influenza epidemics in China with search query from baidu[J]. PLoS One, 2013, 8(5): e64323. DOI: 10.1371/journal.pone.0064323.

[17] Eichstaedt JC, Schwartz HA, Kern ML, et al. Psychological language on Twitter predicts county-level heart disease mortality [J]. Psychol Sci, 2015, 26(2): 159-169. DOI: 10.1177/0956797614557867.
[18] 王心瑶,郝艳华,吴群红,等. 社交媒体环境下H7N9事件网络舆情演变与比较分析[J]. 中国公共卫生,2018,34(9):1232-1236. DOI:10.11847/zgggws1117115.
Wang XY, Hao YH, Wu QH, et al. Evolution trajectory of network public opinion about H7N9 epidemic under social media environment: a comparative analysis[J]. Chin J Public Health, 2018,34(9):1232-1236. DOI:10.11847/zgggws1117115.
[19] 白宁,郁磊,靳祯. 基于百度指数的人感染H7N9禽流感疫情预测[J]. 公共卫生与预防医学,2018,29(6):8-12. DOI:10.3969/j.issn.1006-2483.2018.06.002.
Bai N, Yu L, Jin Z. Prediction of human infection with avian influenza H7N9 based on Baidu index[J]. J Pub Health Prev Med, 2018, 29(6): 8-12. DOI: 10.3969/j.issn.1006-2483.2018.06.002.
[20] Cook S, Conrad C, Fowlkes AL, et al. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic[J]. PLoS One, 2011, 6(8): e23610. DOI:10.1371/journal.pone.0023610.
[21] Brown NJL, Coyne JC. Does Twitter language reliably predict heart disease? A commentary on Eichstaedt et al. (2015a)[J]. Peer J, 2018, 6: e5656. DOI: 10.7717/peerj.5656.
[22] Milinovich GJ, Williams GM, Clements ACA, et al. Internet-based surveillance systems for monitoring emerging infectious diseases[J]. Lancet Infect Dis, 2014, 14(2): 160-168. DOI: 10.1016/S1473-3099(13)70244-5.
[23] Bernardo MT, Rajic A, Young I, et al. Scoping review on search queries and social media for disease surveillance: a chronology of innovation[J]. J Med Internet Res, 2013, 15(7): e147. DOI: 10.2196/jmir.2740.
[24] Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak[J]. PLoS One, 2010, 5(11): e14118. DOI: 10.1371/journal.pone.0014118.
[25] Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control[J]. PLoS Comput Biol, 2011, 7(10): e1002199. DOI: 10.1371/journal.pcbi.1002199.
[26] Lazer D, Kennedy R, King G, et al. The parable of Google flu: traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205. DOI: 10.1126/science.1248506.
[27] Rastegar-Mojarad M, Liu HF, Nambisan P. Using social media data to identify potential candidates for drug repurposing: a feasibility study[J]. JMIR Res Protoc, 2016, 5(2): e121. DOI: 10.2196/resprot.5621.
[28] Freifeld CC, Brownstein JS, Menone CM, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter [J]. Drug Saf, 2014, 37(5): 343-350. DOI: 10.1007/s40264-014-0155-x.
[29] Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts[J]. J Am Med Inform Assoc, 2017, 24(4): 813-821. DOI: 10.1093/jamia/ocw180.
[30] Kazemi DM, Borsari B, Levine MJ, et al. Systematic review of surveillance by social media platforms for illicit drug use[J]. J Public Health (Oxf), 2017, 39(4): 763-776. DOI: 10.1093/pubmed/fdx020.
[31] Kagashe I, Yan ZJ, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data[J]. J Med Internet Res, 2017, 19(9): e315. DOI: 10.2196/jmir.7393.
[32] 尹孔阳,丁一磊,朱大伟,等. 基于2015—2017年舆情监测数据的基层医疗卫生改革舆情评价[J]. 中华医学图书情报杂志,2017,26(8):28-33,40. DOI:10.3969/j.issn.1671-3982.2017.08.006.
Yin KY, Ding YL, Zhu DW, et al. Assessment of public sentiment on medical and health reform at grass-root level based on 2015-2017 public sentiment monitoring data[J]. Chin J Med Libr Inf Sci, 2017, 26(8): 28-33, 40. DOI: 10.3969/j.issn.1671-3982.2017.08.006.
[33] Fung ICH, Hao Y, Cai JX, et al. Chinese social media reaction to information about 42 notifiable infectious diseases[J]. PLoS One, 2015, 10(5): e0126092. DOI: 10.1371/journal.pone.0126092.
[34] 中国互联网网络信息中心. 中国互联网络发展状况统计报告 2019 [EB/OL]. (2019-02-28) [2019-04-14]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/.
China Internet Network Information Center. The 43rd China Statistical Report on Internet Development[EB/OL]. (2019-02-28)[2019-04-14]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/.
[35] Webb TL, Joseph J, Yardley L, et al. Using the internet to promote health behavior change: a systematic review and Meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy[J]. J Med Internet Res, 2010, 12(1): e4. DOI: 10.2196/jmir.1376.
[36] Lohmann S, White BX, Zuo Z, et al. HIV messaging on Twitter: an analysis of current practice and data-driven recommendations [J]. AIDS, 2018, 32(18): 2799-2805. DOI: 10.1097/Qad.0000000000002018.
[37] Vijaykumar S, Nowak G, Himelboim I, et al. Virtual Zika transmission after the first U.S. case: who said what and how it spread on Twitter[J]. Am J Infect Contr, 2018, 46(5): 549-557. DOI: 10.1016/j.ajic.2017.10.015.
[38] Lee H, Mcauley JH, Hübscher M, et al. Tweeting back: predicting new cases of back pain with mass social media data [J]. J Am Med Inform Assoc, 2016, 23(3): 644-648. DOI: 10.1093/jamia/ocv168.
[39] Yom-Tov E, White RW, Horvitz E. Seeking insights about cycling mood disorders via anonymized search logs[J]. J Med Internet Res, 2014, 16(2): e65. DOI: 10.2196/jmir.2664.
[40] Statista. Online healthcare market size in China from 2012 to 2020 [EB/OL]. (2019-09-23) [2019-10-14]. https://www.statista.com/statistics/941888/china-online-healthcare-market-size/.
[41] Hao HJ, Zhang KP. The voice of Chinese health consumers: a text mining approach to web-based physician reviews[J]. J Med Internet Res, 2016, 18(5): e108. DOI: 10.2196/jmir.4430.
[42] López A, Detz A, Ratanawongsa N, et al. What patients say about their doctors online: a qualitative content analysis[J]. J Gen Intern Med, 2012, 27(6): 685-692. DOI: 10.1007/s11606-011-1958-4.
[43] Greaves F, Pape UJ, King D, et al. Associations between web-based patient ratings and objective measures of hospital quality[J]. Arch Intern Med, 2012, 172(5): 435-436. DOI: 10.1001/archinternmed.2011.1675.
[44] Timian A, Rupcic S, Kachnowski S, et al. Do patients "like" good care? Measuring hospital quality via Facebook[J]. Am J Med Qual, 2013, 28(5): 374-382. DOI: 10.1177/1062860612474839.
[45] Bardach NS, Asteria-Peñaloza R, Boscardin WJ, et al. The relationship between commercial website ratings and traditional hospital performance measures in the USA[J]. BMJ Qual Saf, 2013, 22(3): 194-202. DOI: 10.1136/bmjqs-2012-001360.
[46] Zhang W, Deng ZH, Hong ZY, et al. Unhappy patients are not alike: content analysis of the negative comments from China's good doctor website[J]. J Med Internet Res, 2018, 20(1): e35. DOI: 10.2196/jmir.8223.
[47] Statista. Number of social network users in the United States from 2017 to 2023 (in millions)[EB/OL]. (2019-02-18)[2019-10-14]. https://www.statista.com/statistics/278409/number-of-social-network-users-in-the-united-states/.
[48] Statista. Number of social network users in China from 2017 to 2023 (in millions)[EB/OL]. (2019-09-23)[2019-10-14]. https://www.statista.com/statistics/277586/number-of-social-network-users-in-china/.
[49] Xiong Y, Wang ZM, Jiang DH, et al. A fine-grained Chinese word segmentation and part-of-speech tagging corpus for clinical text[J]. BMC Med Inform Decis Mak, 2019, 19 Suppl 2: 66. DOI: 10.1186/s12911-019-0770-7.
[50] Alessa A, Faezipour M. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: prediction framework study[J]. JMIR Public Health Surveill, 2019, 5(2): e12383. DOI: 10.2196/12383.
[51] 国务院办公厅. 国务院办公厅关于促进"互联网+医疗健康"发展的意见[EB/OL]. (2018-04-28)[2019-04-14]. http://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm.
General Office of the State Council, PRC. Opinions of the general office of the state council on promoting the development of "Internet plus Health Care"[EB/OL]. (2018-04-28)[2019-04-14]. http://www.gov.cn/zhengce/content/2018-04/28/content_5286645.htm.
[52] 国务院办公厅. 国务院关于印发"十三五"卫生与健康规划的通知[EB/OL]. (2017-01-10)[2019-04-14]. http://www.gov.cn/zhengce/content/2017-01/10/content_5158488.htm.
General Office of the State Council, PRC. Notice of the State Council on Issuing the Plan for Medicine and Health during the 13th Five-Year Plan Period[EB/OL]. (2017-01-10)[2019-04-14]. http://www.gov.cn/zhengce/content/2017-01/10/content_5158488.htm.

(收稿日期:2019-09-01)
(本文编辑:李银鸽)