

移动互联网 APP 医疗舆情系统信息采集技术探讨

邵 新 卫 玮 唐春霞 郑 萍

(新疆军区总医院信息科 乌鲁木齐 830000)

〔摘要〕 阐述医疗舆情系统架构,包括信息采集、过滤、分析、报告和预警 5 个模块,分析医疗舆情系统信息采集技术现状、问题及改进措施,针对移动互联网 APP 信息采集技术难点问题进行探讨。

〔关键词〕 医疗舆情系统;信息采集;移动互联网

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.07.014

Discussion on Information Collecting Technology of Medical Public Opinion System on Mobile Internet SHAO Xin, WEI Wei, TANG Chunxia, ZHENG Ping, *Information Department of General Hospital of Xinjiang Military Region, Urumqi 830000, China*

〔Abstract〕 The paper expounds the architecture of medical public opinion system, including five modules of information collecting, filtering, analysis, reporting and early warning. It analyzes the status, problems and improvement measures of information collecting technology of medical public opinion system, and discusses the difficult problems of mobile Internet APP information collecting technology.

〔Keywords〕 medical public opinion system; information collecting; mobile Internet

1 引言

现阶段我国医患纠纷日益增多,医患矛盾尖锐^[1],经互联网环境和媒体发酵形成舆情事件。医疗卫生领域互联网舆情具有触点多、燃点低、热度高的特点^[2],较易在全国范围扩散和暴发^[3-4]。针对上述问题医疗卫生相关部门不断完

善制度和管理,利用先进信息技术设计实现医疗舆情系统^[5]用于监控医疗舆情事件,及早发现、处理并控制影响。信息采集模块为医疗舆情监控系统最基础功能,随着互联网迅速发展信息采集技术更新面临挑战。本文从医疗舆情系统架构及系统需求出发,分析总结医疗舆情系统信息采集技术现状,针对移动互联网信息采集的新需求及其应对措施进行探索。

2 医疗舆情系统架构

2.1 概述

医疗舆情系统主要包含 5 个功能模块:信息采集、过滤、分析、报告和预警模块,见图 1。

〔收稿日期〕 2020-01-03

〔作者简介〕 邵新,硕士,高级工程师,发表论文 16 篇,参编论著 5 部;通讯作者:郑萍,主任医师。

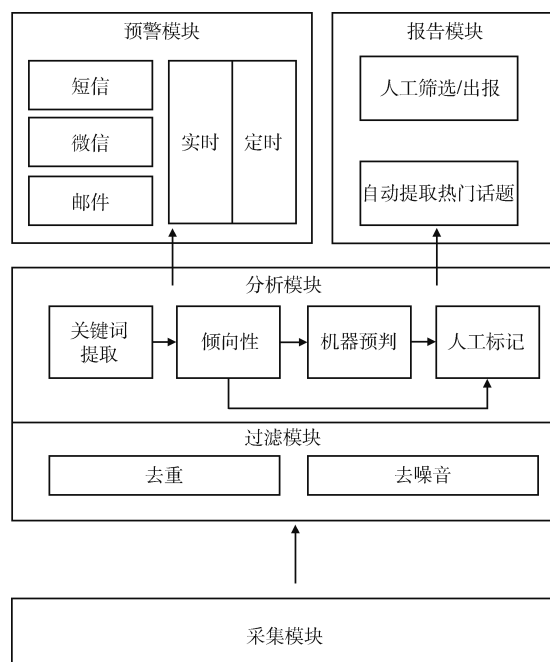


图 1 医疗舆情系统架构

2.2 信息采集

具有全网采集跟踪目标医疗机构相关信息功能。微博和微信等主流舆论媒体通过传统网站、移动互联网形式进行舆论信息传播。主要流量通道采取信息反爬取策略，对信息爬取采集任务增加难度、提出挑战。

2.3 过滤

数据抓取后需经过滤才能被舆情系统有效利用，主要流程有去噪音、去重等，其中包括广告数据等噪音处理。

2.4 分析

过滤后的数据将进行分析，主要操作包括提取关键词、倾向性分析。分析操作最初采用人工方式，进一步通过关键词过滤方式，即分析一组关键词的与或逻辑过滤判断信息的情感评分。随着信息技术发展，机器学习可更准确和高效地预判信息倾向性，正在逐渐替代原有分析方式，特别是人工方式。

2.5 报告

主要包括两种操作方式：一是预先设置关键

词，过滤数据进行统计分析，然后形成报告；二是系统自动发现热点问题，聚类信息后形成热门话题，用户从中挑选关注话题进行统计分析，由舆情系统自动生成报告。

2.6 预警

可根据用户设置进行实时预警和定时预警，用户可指定预警方式和预警接受人。预警方式主要包括邮件、短信、微信等。

3 医疗舆情信息采集现状

3.1 概述

信息采集模块是医疗舆情系统功能实现的基础，信息采集不全或者不够及时直接影响系统对舆情事件的监控管理效果。目前各大舆论媒体和平台采取多种反爬策略^[6]，以保护信息价值，维护系统稳定免受恶劣爬虫影响。舆情信息数据分布在各种网站和平台上，包括信息门户网站、贴吧、论坛、微博、微信、博客评论等。随着移动互联网发展，移动端 APP 逐渐承载较大信息流量。根据网站结构和信息采集方法可将传统网站分为静态、动态、账号关联网站 3 类。

3.2 静态网站

一般为新闻门户及自媒体博客网站，信息较简单、数据量不大、更新不频繁，舆情事件反应相对较滞后且防爬措施较弱，通过简单爬虫技术即可抓取信息，即模拟浏览器操作访问网页进行信息提取。开源爬虫框架有 Python 的 Scrapy、Java 的 Webmagic 等。

3.3 动态网站

随着网站前端框架的发展动态网站已普遍存在，其数据不以静态网页形式直接展现，而是通过 JavaScript 脚本动态生成页面。数据抓取方式分为两种：一是进行抓包分析，模拟 JavaScript 的 post 请求获取其中数据（一般是 json 格式）再进行解析；二是采用内置浏览器爬虫框架，爬虫启动浏览器进程后通过脚本进行操作，由内置浏览器加载解析动

态 JavaScript 脚本后分析 html 静态页面内容从而抓取目标信息数据。这类开源框架有 Python 的 Selenium + PhantomJS、Node.js 的 Puppeteer。

3.4 账号关联网站

随着舆论通道和平台壮大，网站一般会设置会员、账号等功能，以发展用户规模、提高用户黏性并进行定制信息推送。抓取此类平台信息需要模拟账号登陆，获取帐号 Session，以类似动态网站的抓取方式读取舆论通道数据。需要注册多个账号以应对封号情况。

3.5 防爬策略及应对

网站防御爬虫措施较多，增加了舆情系统信息采集的难度。常见防爬策略包括以下 3 种方式^[7]：一是封锁互联网协议（Internet Protocol，IP）地址。常用手段为后台日志监测到封锁异常频繁的 web 请求后封锁其 IP。对此可增加代理服务器以更换访问 IP 数量，增加信息采集稳定性。二是基于 Headers 区分浏览器行为和爬虫行为。对此爬虫脚本需要设置完整的 Header 参数以尽量模拟浏览器请求行为，可一定程度避免阻拦。三是基于用户代理（User - Agent）的防爬策略。可准备多个 User - Agent 参数进行动态更换以躲避拦截。此外针对根据访问频率增加验证码的反爬策略，可通过打码平台收费服务应对；针对简单验证码可通过机器学习方式识别。

4 移动互联网 APP 医疗舆情信息采集的改进

4.1 APP 采集现状

移动端数据采集方法较多，例如可通过抓包工具分析并模拟 APP 网络请求以抓取数据^[8]；采用 Python 爬虫结合 Hook 的方式^[6]获取微信公众号等 APP 数据。上述方式仍延续传统互联网爬虫思路，即以爬虫程序模拟网络请求，可满足较简单的 APP 信息采集需求。

4.2 APP 采集难点

移动互联网 APP 请求所需参数较传统 Web 网页更难提取，部分加密参数需反编译 APP 后解析源码获取。而部分请求不能完全模拟，采集功能受防爬策略拦截，增加移动互联网信息采集难度。

4.3 改进策略

采用沙盒模拟 APP 运行环境，结合代理抓包方式进行采集。通过 Appium 脚本在后台控制安卓模拟器模拟用户操作行为，在请求目标信息页面前发送应用程序接口（Application Programming Interface，API）请求，通知 AnyProxy 代理后端进行抓包，以缩小抓包筛选范围、提高效率，通过解析抓包数据获取所需信息，见图 2。此外可利用 Appium 脚本直接读取模拟器中 APP 界面内容，绕开抓包解析瓶颈，使采集流程更具通用性，可针对大多数 APP 采集信息。上述流程可以完整模拟用户请求，由 APP 产生请求的所有参数并自行解析，保证 APP 各种形式页面数据的采集需要。

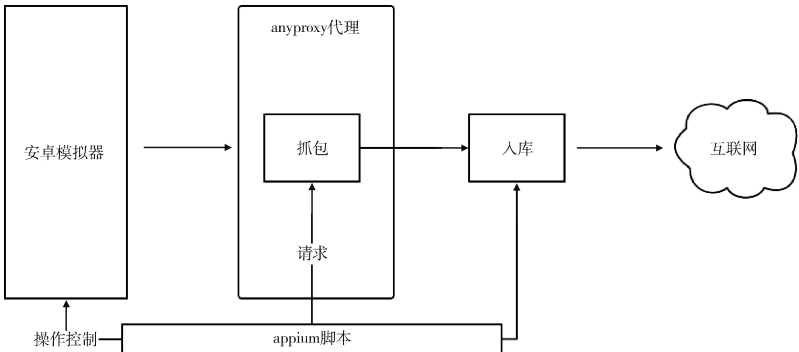


图 2 APP 信息采集流程

单病种付费、DRGS 等^[14]，并且理论实施尚存在较多适应性问题。下一步应关注与探讨如何构建更具普适性的指标体系与智能模型，为不同类型、不同规模医院提供无差别支持与帮助。

参考文献

- 1 凌志, 洪迎玉. 基于商业智能的医院决策支持系统研究与实现 [J]. 中国卫生信息管理杂志, 2013, 10 (6): 529-533.
- 2 徐晨昊. 探究关系型数据库 [J]. 通讯世界, 2019, 26 (1): 208-209.
- 3 熊瑶, 陈敏. 人工智能在医疗领域应用现状探讨 [J]. 医学信息学杂志, 2018, 39 (4): 24-28.
- 4 李军莲, 陈颖, 邓盼盼, 等. 国外基于人工智能的临床决策支持系统发展及启示 [J]. 医学信息学杂志, 2018, 39 (6): 2-6.
- 5 白玲. TABLEAU 在医疗卫生数据可视化分析中的应用 [J]. 中国数字医学, 2018, 13 (10): 72-74, 77.
- 6 周勇, 焦杨, 刘春玲, 等. 基于 ECHARTS 的北京市儿童住院患者来源可视化实现 [J]. 中国数字医学, 2017, 12 (10): 112-114.

- 7 马宏标, 任宏伟, 张弛. 数据可视化设计在医疗卫生领域中的应用探讨 [J]. 河北医学, 2018, 24 (6): 1052-1055.
- 8 胡晓梅, 陈迎春, 周福祥, 等. 基于单病种付费的分级诊疗实施效果及影响因素研究 [J]. 中国卫生政策研究, 2019, 12 (10): 51-57.
- 9 方建生, 王福民. 应用 HADOOP 提升数据库审计系统检索性能的研究 [J]. 计算机系统应用, 2016, 25 (12): 30-34.
- 10 李金苗, 李鹏, 刘庆金, 等. 智能语音识别技术在临床医疗的应用研究与实践 [J]. 中国卫生信息管理杂志, 2019, 16 (2): 218-221.
- 11 孟晓宇, 王忠民, 景慎旗, 等. 医疗人工智能的发展与挑战 [J]. 中国数字医学, 2019, 14 (3): 15-17.
- 12 陈少君. DRGS 医疗服务评价指标在基层医院医疗质量管理中的应用研究 [J]. 中国卫生产业, 2019, 16 (34): 9-11.
- 13 饶正寿, 饶楠杰, 孙晖, 等. 人工智能改变医院管理之探讨 [J]. 中国数字医学, 2019, 14 (1): 24-25, 32.
- 14 王沛陵, 郑杰, 贾方红, 等. 北京市按病种付费支付方式现状及对策研究——基于专家问卷调查结果分析 [J]. 中国社会医学杂志, 2019, 36 (5): 528-531.

(上接第 63 页)

5 结语

本文提出的移动互联网 APP 采集技术可提高信息采集可行性、通用性, 降低移动互联网 APP 信息采集门槛。但采集效率相对普通爬虫较低, 可通过模拟器并行部署提高效率。医疗舆情系统作为医疗机构重要的舆论监测控制工具, 其功能的可用性离不开信息采集技术支持, 随着网络媒体及各种舆论平台和信息渠道的不断发展, 新的防爬策略也会根据信息采集技术的更新而不断改善, 同时也会不断涌现新的舆论信息载体。信息采集技术随着网络环境变化而不断改进, 需要进一步研究和探索。

参考文献

- 1 郭永松. 不同国家医患纠纷处理方式研究 [J]. 中国医

- 院管理, 2010, 30 (5): 14-15.

- 2 杨军. 畅销书出版: 跨越“注意力经济”, 迈向“影响力经济” [J]. 中国图书评论, 2008, 33 (9): 84-88.
- 3 张雷. 注意力经济学 [M]. 杭州: 浙江大学出版社, 2002: 4-6.
- 4 孔庆波, 熊禄全. 大型体育赛事运营之影响力经济形成与挖掘 [J]. 体育文化刊, 2014, 37 (7): 99-125.
- 5 虞颖映, 吴立仁, 胡天天, 等. 医疗卫生行业互联网舆情监测预警系统设计及实现 [J]. 中华医学图书情报杂志, 2017, 26 (3): 37-40.
- 6 柳淑婷. 基于社交网络的舆情关键技术研究 [D]. 长春: 吉林大学, 2017.
- 7 陈利婷. 大数据时代的反爬虫技术 [J]. 电脑与信息技术, 2016, 24 (6): 60-61.
- 8 吕茜. 移动 APP 内容爬虫方法研究 [J]. 广播电视信息, 2017, 26 (3): 61-64.