DOI:10.3969/j.issn.1005-202X.2021.11.022

医学人工智能

基于Web爬虫技术的电子病历信息聚合工具的开发及验证

刘宏嘉[1,2],王静[3],黄宇亮[1,2],李晨光[1,2],吴昊[1,2],马文君[4],曹文田[4],张艺宝[1,2]
1.北京大学医学部医学技术研究院,北京 100191;2.北京大学肿瘤医院暨北京市肿瘤防治研究所放疗科/恶性肿瘤发病机制及转化研究教育部重点实验室,北京 100142;3.浙江大学医学院附属邵逸夫医院,浙江 杭州 310202;4.北京大学物理学院,北京 100871

【摘要】目的:以放疗科的临床和科研工作需求为导向,开发和验证一种基于Web爬虫技术的电子病历信息聚合工具。**方法:**基于Selenium框架和Python编程语言,设计一种基于Web爬虫的病历信息聚合工具,并列举两个实际应用场景:回顾性研究中的数据准备工作以及聚合报告新入院患者的常规临床检查结果作为例子进行说明。测试该工具对临床工作流程的益处,比较自动化方法和手动方法的效率和准确性。**结果:**与人工方法相比,自动信息聚合工具表现出优秀的效率和准确性。对于第一个场景,自动化工具从3 541例患者中提取出110例放射性肺炎的病例,平均每例患者耗时54 s;而人工方法提取出相同数量的病例,平均每例患者耗时90 s。对于另一个例子,自动化方法平均每例患者耗时10 s,而人工方法平均每例患者耗时75 s。**结论:**本工作开发的工具可以在较低访问权限下实现临床和科研工作所需的数据检索、分类汇总等特殊功能,具有安全、高效、准确、跨平台、易拓展等优势。

【关键词】网络爬虫;自动分析;大数据;医院信息系统
【中图分类号】R318 **【文献标志码】**A **【文章编号】**1005-202X(2021)11-1444-05

Development and validation of a Web-crawler-based medical records information aggregation tool

LIU Hongjia[1,2], WANG Jing[3], HUANG Yuliang[1,2], LI Chenguang[1,2], WU Hao[1,2], MA Wenjun[4], CAO Wentian[4], ZHANG Yibao[1,2]
1. Institute of Medical Technology, Peking University Health Science Center, Beijing 100191, China; 2. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Radiation Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China; 3. Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310202, China; 4. School of Physics, Peking University, Beijing 100871, China

Abstract: Objective To develop and validate a Web-crawler-based medical records information aggregation tool for the clinical and research practices in Department of Radiation Oncology. **Methods** Based on Selenium framework and Python programming language, a Web-crawler-based medical records information aggregation tool was designed. The benefits of the proposed tool for clinical workflow were analyzed under two scenarios, namely the data preparation for retrospective study and the aggregate report of routine clinical examination results of newly-admitted patients. Moreover, the efficiency and accuracy between automatic and manual methods were also compared. **Results** Compared with manual method, the automatic information aggregation tool had superior efficiency and accuracy. For the first scenario, automatic method identified 110 radiation pneumonia cases out of 3 541 cases at the time cost of about 54 s per case, while manual methods also identified the same number of radiation pneumonia cases but cost about 90 s per case. For the other scenario, automatic method cost about 10 s per case and manual method cost 75 s per case. **Conclusion** The Web-crawler-based medical records information aggregation tool can implement special functions such as data retrieval and subtotal for clinical and scientific researches under low access rights, with the advantages of security, efficiency, accuracy, cross-platform and easy-to-extend application.

Keywords: Web crawler; automatic analysis; big data; hospital information system

【收稿日期】2021-05-19
【基金项目】北京市自然科学基金(Z210008);国家重点研发计划(2019YFF01014405);国家自然科学基金(11505012);北京大学肿瘤医院科学研究基金学科骨干项目(2021-1);中央高校基本科研业务费/北京大学临床医学+X青年专项(PKU2021LCXQ027);北大百度基金资助项目(2020BD029);北京大学医学部教育教学研究立项项目(2020YB34)
【作者简介】刘宏嘉,硕士,研究方向:医学物理,E-mail: stevendeliu@foxmail.com;王静,博士,研究方向:医学物理,E-mail: fwjing@zju.edu.cn(刘宏嘉和王静为共同第一作者)
【通信作者】张艺宝,博士,高级工程师,硕士生导师,研究方向:医学物理,E-mail: ybzhang66@163.com

前 言

人工智能技术在医学科研和临床工作中得到越来越多的应用[1-4]，不仅提高工作效率[5]，而且减少了由于主观因素和经验水平导致的质量波动和地域差异[6-7]。对于放射治疗专业而言，除了标准DICOM格式的医学图像[8-9]、放疗计划[10-11]、剂量分布[12]等常用信息外，病历中记录的其他多模态信息也具有不可替代的数据价值[13-14]。

病历是病人在医院诊断治疗全过程的原始记录，包含首页、病程记录、检查化验结果、医嘱、手术记录、护理记录等。随着医学信息化的逐步推进，病历也从曾经的纸质、光盘等粗放存储演变成医院信息系统（Hospital Information System, HIS）集中管理。但是，相比DICOM格式的电子数据，HIS系统中存放的病历信息缺乏统一标准，对大数据应用背景下的自动挖掘和分析整理带来挑战[15]。系统提供的有限功能已不能满足科研对于批量查询和聚合分析等“定制化”需求，而市面上又缺乏针对HIS系统开发的数据自动挖掘服务工具，进行数据整合时往往使用传统人工方法进行整理。传统人工整理方法不仅效率低下，而且容易发生遗漏或错误，影响样本量的扩充和数据质量的保障[16]。作为访问权限较低的终端用户，如何在没有医院信息管理部门的配合下，安全、便捷地实现对批量病历信息的自动汇总和分析，在大数据时代背景下对于临床和科研工作具有重要意义。

本研究开发了一种基于Web爬虫技术的病历信息聚合工具，并以大量病历中的诊断报告信息筛选和血常规检查结果的特定指标信息聚合为例进行验证，同时与手工提取相关信息的过程作为参照进行对比验证分析，结果表明相对于手工提取的过程，自动方法在提取效率和准确性相对于人工方法均有较大提升。

1 材料与方法

1.1 需求背景和操作环境

本工作基于北京大学肿瘤医院的HIS病历系统，在内部网络环境下可通过患者身份编号、姓名等信息调取病历文书、诊断报告等信息。该病历查询系统是一个ASP.NET的Web应用程序，普通用户无法获取病历系统数据库的直接访问权限，也无法“定制”批量查询和整理等系统尚未提供的特殊功能。而直接获取信息的爬虫方法难以绕过Web应用程序的安全设计，同时也存在一定程度的安全风险。同时，为了节约计算资源，也为了探索日后本工作在利用小型计算设备的可能性，本工作在搭载了Raspbian（一种基于Debian的Linux操作系统）的树莓派4ModelB上执行（硬件上使用树莓派4B原生套件）。

1.2 程序架构

为了解决访问安全、自动化和实现特殊定制功能等问题，本工作采用Selenium技术来聚合信息。程序基于Web架构设计，采用名为Selenium的Web应用程序测试的框架和Python编程语言，通过浏览器来实现所需要的信息搜集。Selenium作为Web应用程序测试工具，能够在近乎真实的浏览器中执行测试，模拟人类用户操作。而Python作为一种优秀的胶水语言，在利用Python执行Selenium工具时，可以对浏览器所能展示的数据进行方便的分析聚合。整体的程序架构图如图1所示，工作流程如图2所示。

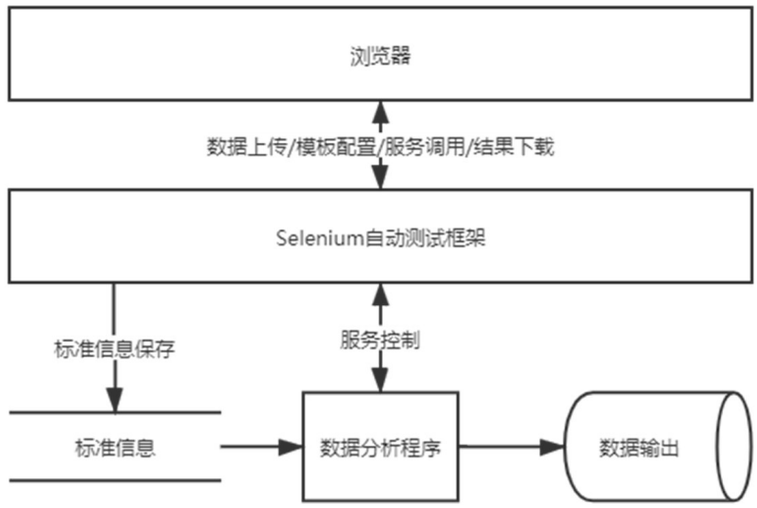

图1 基于Web爬虫的工具程序架构图

Fig.1 Program framework of Web-crawler-based tool

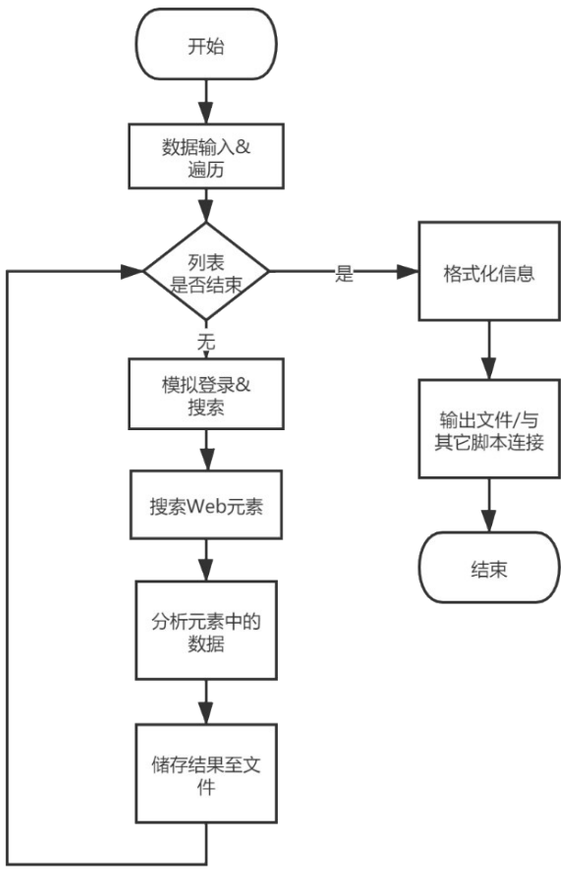

图2 聚合工具的工作流程图

Fig.2 Workflow chart of aggregation tool

该聚合工具接收研究人员可定位的患者基本信息(病历号、姓名等信息)以及所需信息项目名称列表,模拟实际登录过程,依次定位到患者信息所在页面。其中根据页面的DOM树结构,分析整个页面是否包含所需信息以及哪些信息可以被访问,返回所需的DOM树节点以及Xpath路径。之后根据返回的节点和路径来提取数据并根据预先定义的数据字典来判断数据准确性,并以结构化方式(Python数据字典)来进行储存。最后按照研究人员的需求,或者将结果存储为结构化文件,或者将结果接口与其它数据分析的脚本相连接,直接供研究使用。

1.3 测试目标和工作流程

本程序通过模拟浏览器操作来实现癌症患者相关数据的批量查询以及定位获取,利用服务器后台程序对获取到的数据进行信息聚合,并按照指定的格式输出所需数据的文件。验证测试目标包括:(1)从既往肺癌、食管癌、乳腺癌等胸部放疗患者病历数据库自动查询并汇总发生放射性肺炎的病例,以及距离放射性肺炎发生的最近一次放疗的时间;(2)对新入院患者的临床血常规检查报告的结果进行聚合。具体工作流程包括:(1)记录并模拟人工查询1例患者从授权进入到访问包含所需数据页面等全过程需要在Web界面上执行的操作;(2)将上述操作映射为Selenium操作语句(在本例下,执行的操作会进入到包含放射报告以及病历文书的界面),同时封装起来,暴露数据输入为唯一变量患者号;(3)分析数据所在页面网页源代码,找到目标数据所在的元素,映射为Selenium操作语句以获取所需数据(在本例下,需要获取放射诊断报告栏目下的所有子报告以及病历文书下的所有文档);(4)利用Python分析得到的数据。对于放射性肺炎测试目标,程序在得到的所有放射诊断报告中模糊搜索有关放射性肺炎的描述,判断具体患者是否被诊断为放射性肺炎。如有,则继续查询从无到有的变化日期,并检索该日期最近的若干病例文书,返回其中内容最多的文书(病历文书是一个长期积累,通常新的文书会包含之前旧文书的内容),以备日后科研工作分析以及错误核验。对于聚合新入院患者的临床血常规报告,该工具将遍历检查表,从中按名称判断是否是所需要查询的检查,然后记录该次检查时间、负责医生等信息,搜索表格中是否有所需的血常规子项目并记录对应数据储存到预设的数据结构中。最后将得到的数据按照需要的格式输出,如以患者号命名的文本研究件形式,或者是便于导出导入的Numpy数组格式等。

2 结 果

本工作成功开发了满足图1架构以及图2所示工作流程的基于Web爬虫技术的病历信息聚合工具。使用上述系统开发架构和工作流程建立的信息聚合工具成功自动登录了HIS系统,并在HIS系统上的不同区域和路径成功获得指定患者所需的准确信息。

放射性肺炎病例的获取:基于Web爬虫技术的病历信息聚合工具从3 541例患者中识别出110例放射性肺炎病例。使用手动方法对这3 541例患者进行识别与自动化方法对比,自动化方法每例患者耗时约为54 s,手动方法每例患者耗时约为90 s。

新入院患者的临床血常规检查报告的结果聚合:分析110例患者的血常规检测报告,使用自动化方法平均约10 s完成单例患者血常规分析的数据收集,而如果使用人工方法则需要75 s或者更多的时间。

此外,由于自动化方法是从HIS系统中直接获取结构化的病人信息,或是直接利用该信息进行判断,从而准确性更高。相比之下,上述两个验证测试目标在人工实现的过程中均发现了错误,比如填写报告数据遗漏和错误等。

3 讨 论

测试验证的结果表明，基于Web爬虫技术的病历信息聚合工具在效率和准确性方面均远胜于人工提取操作。值得注意的是，在不同需求背景下，自动工具的特点和优势也不尽相同。在程序设计时需要充分考虑具体任务的特点，个性化优化流程和代码。比如放射性肺炎病例的排除相比检出相对复杂：前者需要遍历所有的放射诊断报告，而后者只需在任一报告中发现有用信息即可停止检索。本例中放射性肺炎的检出发病率约为3.11%，而如果本程序用于搜索发病率更低、样本量更大的病例，自动工具与人工方法的效率差异会被进一步放大。如果涉及不同页面的切换，也会对结果产生秒级的影响。相比之下，第二个提取和整理血常规信息的例子相对简单。原因之一是血常规的报告相对规范且客观，没有放射诊断报告中涉及的人工描述和主观差异；原因之二是提取的信息主要是同一页面上的检验报告，不涉及到切换页面等耗时操作。如果报告内容涉及更多更复杂的数据，人工整理的效率和准确性会进一步降低，而对于自动化方法的影响仅在毫秒量级。

本工作的创新之处在于：(1)除了大幅缩短数据整理时间外，还实现了全程无人工干预的自动化流程；(2)由于使用了相对简洁的Python+Chromedriver+Selenium套装方案，本工具即使是在树莓派这样的微型机器上也能够流畅运行，降低对计算设备的要求，节省算力成本；(3)该工具同时兼容多种图形化工作界面，具有良好的跨平台特性；(4)具有很高的查准率和查全率，降低复核数据的工作压力；(5)兼具良好的安全性和灵活性，可以在没有访问原始数据库权限的情况下，实现低权限场景下获取数据，“定制”实现HIS系统中尚未提供的特殊功能。低权限场景下可以防止对HIS系统进行篡改以及降低暴露数据漏洞的可能，体现出良好的安全性。

在未来更复杂的应用场景中，本工作可以进一步拓展和改进的方面包括：(1)在分析更大数据量以实现规模效益时，可以利用分布式计算技术来进一步压缩执行时间。比如从该工具的系统架构入手，设计一个C-S(客户端-服务端)结构的工具，在若干个设备或虚拟机上使用实际执行代码的客户端，而服务端则把执行逻辑以及需要查找的患者根据客户端返回的运行状况进行动态的调度。预计采用此方法可以进一步大幅降低规模化查找的时间，另外也可以实现不同设备之间的负载均衡，从而使得算力和网络访问资源被更高效的利用[17-18]；(2)利用标准化接口降低该工具的耦合度[19-20]。以本研究中涉及的两个应用为例，虽然系统架构是一致的，但是在检索部分和分析部分分别使用了结构不尽相同的代码。这一方面体现了该工具良好的可拓展性，但另一方面也对非信息科学专业的人员提出更高要求，增加了根据不同背景环境来更改、调试代码所需的精力。因此，我们计划在使用该工具进行更多场景的测试之后，总结出更普适性的数据搜集办法。尝试在一套工具里面标准化提取不同种类HIS数据的接口，使得工具具有更好的用户友好度，同时进一步降低该工具的耦合度，提升其可拓展性，以及与其他科研代码的可整合性。

4 结 论

本工作成功开发并验证了一套基于Web爬虫技术的病历信息自动聚合工具，该工具具有安全、高效、准确、成本低、跨平台、易拓展等特点，可以在较低访问权限的情况下，“定制”实现临床和科研工作所需的数据检索、分类汇总等特殊功能，使得自动化技术在医学中得到进一步发展和应用。

【参考文献】

[1] ZHENG S, JABBOUR S K, O'REILLY S E, et al. Automated information extraction on treatment and prognosis for non-small cell lung cancer radiotherapy patients: clinical study[J]. JMIR Med Inform, 2018, 6(1): e8.

[2] MCNUTT T R, BOWERS M, CHENG Z, et al. Practical data collection and extraction for big data applications in radiotherapy[J]. Med Phys, 2018, 45(10): e863-e869.

[3] 严舒, 陈娟, 欧阳昭连. 基于NIH项目的美国医学人工智能发展态势分析[J]. 中国医疗设备, 2019, 34(12): 101-105.
YAN S, CHEN J, OUYANG Z L. Analysis on the development situation and trend of american medical artificial intelligence based on NIH project[J]. China Medical Devices, 2019, 34(12): 101-105.

[4] 金昌晓, 计虹, 席韩旭, 等. 大数据科研分析平台在临床医学研究中的应用探讨[J]. 中国数字医学, 2019, 14(2): 37-39.
JIN C X, JI H, XI H X, et al. The discussion about the application of the scientific research platform of big data in clinical medical research [J]. China Digital Medicine, 2019, 14(2): 37-39.

[5] HUANG Y, YUE H, WANG M, et al. Fully automated searching for the optimal VMAT jaw settings based on eclipse scripting application programming interface (ESAPI) and RapidPlan knowledge-based planning[J]. J Appl Clin Med Phys, 2018, 19(3): 177-182.

[6] WU H, JIANG F, YUE H, et al. Applying a RapidPlan model trained on a technique and orientation to another: a feasibility and dosimetric evaluation[J]. Radiat Oncol, 2016, 11(1): 108.

[7] CAROLIN S, OLIVER W, CHRISTIAN W, et al. Intercenter validation of a knowledge based model for automated planning of volumetric modulated arc therapy for prostate cancer. The experience of the German RapidPlan consortium[J]. PLoS One, 2017, 12(5): e0178034.

[8] CHANG Y, LAFATA K, SUN W, et al. An investigation of machine learning methods in delta-radiomics feature analysis[J]. PLoS One, 2019, 14(12): e0226348.

[9] VRBIK I, VAN NEST S J, MEKSIARUN P, et al. Haralick texture feature analysis for quantifying radiation response heterogeneity in murine models observed using Raman spectroscopic mapping[J]. PLoS One, 2019, 14(2): e0212225.

[10] 李正贤, 闫相东, 王博生, 等. 利用医科达病例训练瓦里安RapidPlan

模型的可行性和剂量学评估[J]. 中华生物医学工程杂志, 2020, 26(1): 9-14.
LI Z X, YAN X D, WANG B S, et al. A feasibility and dosimetric study of configuring a Varian RapidPlan model based on Elekta cases[J]. Chinese Journal of Biomedical Engineering, 2020, 26(1): 9-14.
[11] WU H, JIANG F, YUE H, et al. A dosimetric evaluation of knowledge-based VMAT planning with simultaneous integrated boosting for rectal cancer patients[J]. J Appl Clin Med Phys, 2016, 17(6): 78-85.
[12] MOORE C S, WOOD T J, SAUNDERSON J R, et al. The usefulness of large sample size patient dose audits for optimisation of CT automatic exposure control (AEC) settings[J]. J Radiol Prot, 2019, 39(3): 938.
[13] KAVULURU R, RIOS A, LU Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records[J]. Artif Intell Med, 2015, 65(2): 155-166.
[14] 孙凤英, 于修义. 基于胸外科结构化电子病历的人工智能诊断系统构建实践[J]. 中国数字医学, 2018, 13(11): 29-31.
SUN F Y, YU X Y. The construction and application of structured electronic medical record and artificial intelligence diagnosis system of thoracic surgery[J]. China Digital Medicine, 2018, 13(11): 29-31.
[15] 朱晓勃. 我国医院信息化建设现状与发展对策研究[J]. 现代仪器与医疗, 2015, 21(1): 76-79.
ZHU X B. Status of hospital informatization construction in China and research of development measures[J]. Modern Instruments & Medical Treatment, 2015, 21(1): 76-79.
[16] XIE C, YANG P, YANG Y. Open knowledge accessing method in IoT-based hospital information system for medical record enrichment[J]. IEEE Acc, 2018, 6: 15202-15211.
[17] 杨际祥, 谭国真, 王荣生. 并行与分布式计算动态负载均衡策略综述[J]. 电子学报, 2010, 38(5): 1122-1130.
YANG J X, TAN G Z, WANG R S. Survey of dynamic load balancing strategies for parallel and distributed computing[J]. Acta Electronica Sinica, 2010, 38(5): 1122-1130.
[18] CARDELLINI V, COLAJANNI M, YU P S. Dynamic load balancing on web-server systems[J]. IEEE Internet Compt, 1999, 3(3): 28-39.
[19] 黄光芳. 面向接口编程在三层架构系统中的设计及应用[J]. 计算机应用与软件, 2009, 26(6): 133-135.
HUANG G F. Designing and applying interface-oriented programming in three layer architecture system[J]. Computer Applications and Software, 2009, 26(6): 133-135.
[20] SAXENA V, KUMAR S. Impact of coupling and cohesion in object-oriented technology[J]. Int J Softw Eng Appl, 2012, 5(9): 671-676.

（编辑：陈丽霞）