兰坤, 吴琼, 耿艳兵. 基于 Python 的社交网站用户行为数据采集方法[J]. 智能计算机与应用,2024,14(6):219-223. DOI:10.20169/j.issn.2095-2163.240633

基于 Python 的社交网站用户行为数据采集方法

兰 坤[1], 吴 琼[1], 耿艳兵[2]
(1 长治医学院 计算机教学部, 山西 长治 046000; 2 中北大学 大数据学院, 太原 030051)

摘 要: 传统数据采集方法存在适用范围较窄、重复性工作量较大等问题,导致社交网站用户行为数据采集效率较差,提出基于 Python 的社交网站用户行为数据采集方法。采用情境标记法确定社交网站用户行为数据采集时机,基于 Python 语言搭建一个以 MJU 采样算法为 URL 地址管理中心的 Scrapy 爬虫框架,执行 Scrapy 爬虫框架完成社交网站用户行为数据的采集流程。实验结果表明,本文方法在采集社交网站的用户行为数据时,采集速率为 830 个/h,验证了该方法具有快速性。

关键词: Python; 社交网络; 用户行为数据; 数据采集方法
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0219-05

A Python-based method for collecting user behavior data on social media sites

LAN Kun[1], WU Qiong[1], GENG Yanbing[2]
(1 Department of Computer Teaching, Changzhi Medical College, Changzhi 046000, Shanxi, China;
2 School of Data Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: Traditional data collection methods have problems such as narrow applicability and large repetitive workload, resulting in poor efficiency in collecting user behavior data on social media sites. Therefore, a Python-based method for collecting user behavior data on social media sites is proposed. Using situational tagging to determine the timing of collecting user behavior data on social media sites, a Scrapy crawler framework based on Python language with MJU sampling algorithm as the URL address management center is constructed, and the Scrapy crawler framework is executed to complete the process of collecting user behavior data on social media sites. The experimental results show that the method proposed in this paper has a collection rate of 830 user behavior data per hour when collecting user behavior data from social media sites, which verifies the speed of the method.

Key words: Python; social networks; user behavior data; data collection methods

0 引 言

自 21 世纪以来,计算机技术的发展速度飞快,互联网已经成为社会生活和行业工作中不可或缺的组成部分。当下,人们进行信息传播和资源共享无不需要社交网络。社交网络正在人们生活中发挥出更大的优势作用。如果可以采集社交网站上用户行为数据,对于社交网站的商业应用以及可持续发展都具有重要的现实意义。研究可知,数据采集是一个研究热度不断攀升的前沿领域,在金融数据分析、电力、零售等多个领域都有着重要的作用,因此学术界对其给予了高度关注。

陈辉等学者[1]通过信任模型设计一种分簇 WSNs 数据可靠数据采集方法,利用信任模型过滤非正常节点数据,可解决无线传感网络监测数据可靠性较低的问题。杨杉等学者[2]从压缩感知角度出发,设计一种数据采集新方法,通过压缩感知提取能源数据特征,实现采集方法的数据稀疏化处理,具有实时性、同步性与准确性。谢蓉蓉等学者[3]提出基于网络爬虫的网页大数据抓取方法,通过计算获取数据的关键特征,采用广度优先的策略进行数据信息的抓取,采用相位重构相空间的方法获取爬行维数,通过引入关联维数,实现对页面大数据的抓取,具有较高的准确率。赵瑞丹等学者[4]提出基于爬虫技术和语义分析的网络舆情采集方法,采用主题网络爬虫,对收集到的 Web 信息进行深度挖掘,并采用向量空间模型对收集到的 Web 信息进行二次过滤,以确保收集到的 Web 信息质量。

基金项目: 山西省自然科学基金面上基金(202103021224192)。
作者简介: 兰 坤(1977-),女,工程硕士,副教授,主要研究方向:信息处理与分析。Email:lk@czmc.edu.cn
收稿日期: 2023-04-28

然而,社交网站上用户行为数据具有规模大、增长速度快等特点,为数据采集带来了一定困难,因此社交网站用户行为数据采集方法仍是一个值得深入研究的课题。本文研究基于 Python 的社交网站用户行为数据采集方法,通过情境标记法确定数据采集时机,需要考虑到不同社交网站的特点和使用情况,制定相应的采集策略,可以更加精准地确定用户行为数据的采集时间,提高数据采集效率。采用了 Python 中的 Scrapy 爬虫框架,具有高效、稳定、扩展性强等特点,可以实现多线程、分布式等功能,利用 MJU 采样算法,有效地管理 URL 地址,避免重复采集数据,提高数据采集效率和准确性。

1 确定社交网站用户行为数据采集时机

进入信息时代以来,国内的社交网站数据规模不断扩大。但是,由于这些网站用户的行为数据往往十分复杂且庞大,传统的数据采集方法已经无法高效且精准地进行有价值数据的采集,所以本文针对社交网站用户行为数据采集方法展开深入研究。在实际的社交网站用户行为数据采集过程中,一般需要请求用户配合来获取行为数据,如果在不合时宜的情况下进行数据采集任务(如会议、学习期间),那么不仅用户可能会拒绝采集请求,而且所采集的很多行为数据会呈现出相同且重复性的内容,所以本文在进行社交网站用户行为数据采集时,首先需要确定采集时机。简单来说,就是根据社交网站中用户行为习惯数据掌握用户的使用情境,从使用情境角度出发,划分出用户可打扰情境,作为最佳用户行为数据采集时机[5]。情境就是用户访问社交网站时的一些基本属性,如环境信息、位置信息、社会信息等,对此可用式(1)来描述:

$$Q = R(q_1, q_2, \cdots, q_i, \cdots, q_n) \quad (1)$$

其中,Q 表示用户访问社交网站的高级情境,R 表示用户低级情境 $q_1, q_2, \cdots, q_i, \cdots, q_n$ 的组合。一般高级情境较为复杂,无法通过社交网站直接确定,所以本文利用式(1)使用低级情境组合形式来得到高级情境。这里,低级情境的数学表达满足式(2):

$$q_i \in \{q_y, q_h, q_r, q_w, q_s, q_t, q_b, q_c\};\ i \geqslant 2 \quad (2)$$

其中,q_y 表示用户情境;q_h 表示环境情境;q_r 表示任务情境;q_w 表示位置情境;q_s 表示社会情境;q_t 表示时间情境;q_b 表示设备情境;q_c 表示基础设施情境。在此基础上即可得到相关情境信息,也就是描述用户访问社交网络场景及状态的信息[6]。针对不同的社交网站,由于用户访问情境不同,所以进行数据采集的时机也各不相同。由文中上述内容可知,确定用户访问的低级情境较为容易,根据相关经验以及相似研究数据即可确定其数据采集时机。但关于高级情景下的最佳数据采集时机,需要以情境感知机理角度出发,在上述情境信息的基础上,通过情境标记法来确定最佳数据采集时机,也就是让社交网站用户实时标记可打扰情境,然后采用情境推理规则来自动提取并处理社交网站用户可打扰情境规则,并以此作为参考确定最佳数据采集时机,为后续的数据采集奠定基础。

情境标记法在实现过程中可能具有用户拒绝标记、标记不准确、操作复杂度高与隐私问题,为了解决以上问题,可以考虑以下方案:

(1)提供激励措施:为用户提供积分、奖励等激励措施,鼓励用户参与情境标记,提高数据采集的效率和准确性。

(2)优化标记流程:通过简化标记流程、提供标记指南和示例等方式,降低用户标记的难度和复杂度,提高标记准确性。

(3)加强用户隐私保护:在采集用户数据时,应遵循相关法律法规,明确数据使用目的和范围,保护用户的隐私权利。

2 基于 Python 搭建 Scrapy 爬虫框架

Python 是一门面向对象的编程语言,具有强大的运算能力以及丰富的开发语言资源库,所以本文引入 Python 语言来设计社交网站用户行为数据采集方法[7]。基于 Python 进行社交网站用户行为数据采集,搭配 Numpy 等工具库,可有效减少社交网站用户行为数据采集流程中多个环节的计算量,从而提升用户行为数据采集效率,利用 Python 语言实现社交网站用户行为数据的精准、高效采集是本次研究中的重点内容[8]。由于社交网站自身特性原因,在采集用户行为数据时呈现出以下特征:社交网站用户行为数据具有多元化、复杂化的特点,不仅内容杂乱,而且数据存在形式也各不相同;用户行为数据具有动态性的特点,也就是说每时每刻社交网站上均会产生大量用户行为数据,所以数据的更新迭代是数据采集过程中不可忽视的重要因素;由于用户访问社交网站的设备不同,产生的行为数据具有异构性,也成为数据采集的难点之一。而 Python 语言有着良好的可解释性、交互性等特性,完全可以适应独特的社交网站用户行为数据采集任务,因此本文基于 Python 语言搭建一个 Scrapy 爬虫框架[9],根

据提前设定好的规则，在社交网站上进行用户行为数据脚本及程序的抓取，进而实现数据采集。Scrapy 爬虫框架是一种自动爬虫框架，采用 Python 语言开发和封装。可以提升爬虫的运行效率，只需简单设置爬虫规则，就能实现社交网站用户行为数据的采集。该框架由控制器、HTML 下载器、解析器和资源库构成。其中，控制器是核心，可协调整个框架的各个模块协调运行。主要任务是管理社交网站网页数据的抓取流程和 URL 地址。在本文的 Scrapy 爬虫框架中，使用 MJU 采样算法进行 URL 地址管理。MJU 算法是一种多重跳跃性采样算法，可通过减小节点自循环概率，来保障全局采样节点的公平性。在利用 MJU 采样算法进行 URL 地址管理时，首先需要确定多重跳跃的参数 η 与 μ。一般来说，这 2 个参数互相影响，可共同实现数据采样的 URL 地址管理工作，所以在进行社交网站用户行为数据采集时，需要得到 η 与 μ 的最佳取值。这里，η 可以根据式(3)来确定：

$$\eta = \frac{D \times N}{L} \tag{3}$$

其中，D 表示样本节点平均长度；N 表示需要采集的数据节点数量；L 表示节点链。在式(3)基础上，根据 η 与数据样本节点平均度分布图即可求出多重跳跃的另一个参数 μ 的最优值。在确定了 MJU 采样算法的多重跳跃参数后，需要计算出样本节点 i 到下一节点 j 的概率，由于 MJU 采样算法进行 URL 地址管理时可以建模为马尔科夫链，所以本文建立的节点跳跃概率计算如式(4)所示：

$$P_{(j,i)} = \begin{cases} \min\left(\dfrac{1}{d_j}, \dfrac{1}{d_i}\right) + \dfrac{\eta}{|Z|}, & \text{if } i \text{ is } j\text{'s neighbor} \\ \left(1 - \eta - \displaystyle\sum_{i=j}\left(\frac{1}{d_j}, \frac{1}{d_i}\right)\right)\mu + \dfrac{\eta}{|Z|}, & \text{if } j = i \\ \left(1 - \eta - \displaystyle\sum_{i\neq j}\left(\frac{1}{d_j}, \frac{1}{d_i}\right)\right)(1-\mu) + \dfrac{\eta}{|Z|}, & \text{if } d_j = d_i \\ \dfrac{\eta}{|Z|}, & \text{otherwise} \end{cases} \tag{4}$$

其中，$P_{(j,i)}$ 表示社交网站用户行为数据采样节点 i 随机选择下一节点 j 的概率；d_j、d_i 分别表示节点 j 与节点 i 的度数；$|Z|$ 表示原社交图中采样节点的总数。按照式(4)的用户节点转移概率即可实现 URL 地址的选择管理，从而确保 Scrapy 爬虫框架可以提取到有效的用户行为数据。HTML 下载器主要负责将 Scrapy 爬虫框架抓取的社交网站用户行为数据下载到本地，解析器主要负责用户行为数据的解析任务。资源库是 Scrapy 爬虫框架的基础，主要负责存储解析后的社交网站用户行为数据。综上，也就是本文基于 Python 语言搭建的 Scrapy 爬虫框架核心部分的设计构成。

3　执行 Scrapy 爬虫采集用户行为数据

一般来说，国内社交网站中用户个人信息为所有人可见，所以本文通过 Scrapy 爬虫框架来采集社交网站用户行为数据是可行的[10]，但是在利用 Scrapy 爬虫框架采集社交网站用户行为数据时，社交网站往往会因为要维持自身运行的稳定性与安全性，对爬虫框架的爬虫频率做出相应限制，从而导致数据采集失败。所以为解决该问题，本文设计了 Scrapy 爬虫框架执行流程来采集社交网站用户行为数据[11]如图 1 所示。

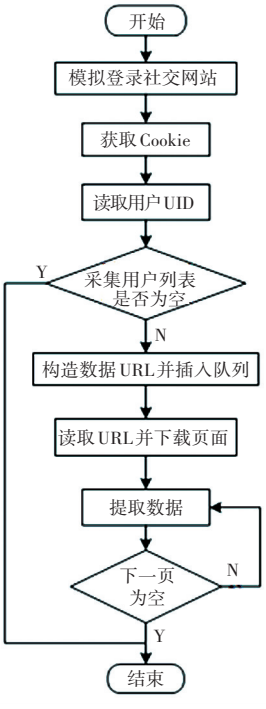

图 1　社交网站用户行为数据采集流程图

Fig. 1　Flow chart for collecting user behavior data on social networking websites

由图 1 可知,Scrapy 爬虫框架执行流程可以针对批量社交网站用户行为数据进行采集。如果想采集长期的社交网站用户行为数据,还需在图 1 中 Scrapy 爬虫框架执行流程的基础上考虑数据去重与信息更新问题,这里本文通过 Scrapy 爬虫框架的资源库实现用户行为数据的更新。该资源库是使用 Python 语言所编写的 MySQL 客户端数据库[12],存储性能稳定且功能完整,可以利用已存储的用户行为数据信息字段来判断最新存入的数据是否为重复数据,并进行相应的去重、更新等操作[13]。综上,本文利用 Python 语言设计一个 Scrapy 爬虫框架,然后基于执行框架爬虫流程实现对社交网站网页数据的提取,以供用户行为数据的采集[14]。

4 实验分析

4.1 实验准备

为验证本文设计社交网站用户行为数据采集方法的有效性,本节搭建局域网模拟网络环境对采集方法进行测试,如图 2 所示。

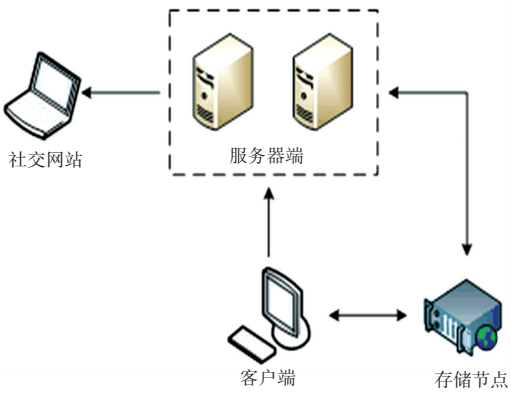

图 2 实验环境部署图

Fig. 2 Deployment diagram of experimental environment

由图 2 可知,在本次部署仿真实验环境中,当用户向社交网站下发数据采集请求时,服务器端会访问社交网站进行用户行为数据采集,并将采集数据存储至存储节点中,便于用户在客户端查看采集数据。进一步,该实验环境中各节点的硬件配置见表 1。

表 1 实验环境节点硬件配置

Table 1 Hardware configuration of experimental environment nodes

GB

节点	CPU	内存	硬盘
服务器端	Intel i5 9400f	8	240
客户端	Intel i5 9400f	4	240
存储节点	Intel i7 7700k	4	1 024

在此实验环境背景条件下,本文选用文献[3]提出的基于网络爬虫的网页大数据抓取方法、文献[4]提出基于爬虫技术和语义分析的网络舆情采集方法作为实验对照组,通过实验对比结果来判断本文设计方法的性能。由于国内社交网站种类较多,本文主要以 QQ 空间、新浪微博、人人网、知乎、百度贴吧、豆瓣这六大主要社交网站作为实验对象。与此同时,由于数据采集方法主要功能是从海量社交网站用户行为数据中采集有用信息,所以快速性是数据采集方法的重要评价标准。针对该评价指标,本节设计 2 个实验场景:一是社交网站给出一定量的用户行为数据,分别使用上述 3 种方法进行采集,统计并对比这 3 种方法所需的采集时间;二是用户给出一定的采集时间,分别使用上述 3 种方法对社交网站用户行为数据进行采集,统计并对比这 3 种方法采集的数据量。

4.2 实验结果

实验中,从上述六大社交网站中随机挑选 200 位用户作为实验目标,本节选择用户关注、评论、转发、点赞等行为数据作为采集对象。在场景 1 中,每一个社交网站给出不同数量的网页信息,采用这 3 种方法分别进行数据采集,得到各方法的采集时间对比结果见表 2。

表 2 场景 1 的采集速率对比结果

Table 2 Comparison results of collection rates for scenario 1

社交网站	采集内容/个	采集时间/h		
		本文方法	文献[3]方法	文献[4]方法
QQ 空间	1 200	1.4	1.8	2.5
新浪微博	1 500	1.8	2.4	3.7
人人网	1 800	2.1	3.3	4.7
知乎	2 100	2.5	4.3	5.9
百度贴吧	2 400	3.2	6.1	7.1
豆瓣	3 000	3.6	8.2	9.8

由表 2 可知,本文设计方法的平均采集速率为 828 个/h,较对照组方法提升了 314 个/h、450 个/h,而且本文设计方法的采集速率较为稳定,不会随着采集内容规模的变化而变化,但对照组方法的采集速率会随着采集内容的增加而降低,由此可以说明本文设计方法更加可行可靠。在场景 2 中,当采集各个社交网站用户行为数据时,控制这 3 种方法的采集时间相同,此时得到各方法的采集数量对比结果见表 3。

表 3　场景 2 的采集速率对比结果

Table 3　Comparison results of collection rates in scenario 2

社交网站	采集时间/h	采集数量/个		
		本文方法	文献[3]方法	文献[4]方法
QQ 空间	8	6 680	4 108	3 009
新浪微博	6	4 859	2 998	2 307
人人网	5	4 200	2 601	1 916
知乎	4	3 405	2 113	1 483
百度贴吧	3	2 409	1 527	1 112
豆瓣	2	1 703	1 036	750

由表 3 可知，本文设计方法的平均采集速率为 832 个/h，较对照组方法提升了 317 个/h、455 个/h，从数值上体现了本文设计方法的快速性，该方法通过 Python 语言进行社交网站用户行为数据的采集，可以充分利用网络爬虫框架的使用性能，从而提升数据采集速率。综上，本文设计数据采集方法的平均采集速率为 830 个/h，与实验对照组方法相比具有显著优势，采集速率越大，说明在同等时间内所采集的数据量越多。因此本文设计数据采集方法性能优越，可以满足社交网站用户行为数据的采集需求。

5　结束语

互联网时代的到来，国内各大社交网站发展迅速，为满足人们的使用需求，如何快速准确地采集社交网站用户行为数据具有重要意义，所以本文引入 Python 设计一种社交网站用户行为数据采集方法。该方法针对社交网站结构的独特性及用户行为数据的多样性，采用 Python 语言搭建一个网络爬虫框架，并通过该框架完成社交网站用户行为数据的采集任务。在前述研究基础上，本文还通过仿真实验验证了本文设计方法具有较高的采集效率，且在稳定性上具有一定优势，可以很好地应对社交网站用户行为数据采集任务，为推动国内社交网站的健康发展奠定理论基础。

参考文献

[1] 陈辉，张春雨. 基于信任模型的分簇 WSNs 可靠数据采集方法[J]. 传感技术学报，2021，34(11)：1530-1536.

[2] 杨杉，谭博，郭静波. 基于压缩感知的新一代能源互联网的数据采集方法[J]. 可再生能源，2022，40(7)：952-958.

[3] 谢蓉蓉，徐慧，郑帅位，等. 基于网络爬虫的网页大数据抓取方法仿真[J]. 计算机仿真，2021，38(6)：439-443.

[4] 赵瑞丹，朱旭. 基于爬虫技术和语义分析的网络舆情采集系统设计[J]. 电子设计工程，2021，29(14)：56-60.

[5] 慕爽，刘正捷，苏光华. 即时临场用户主观数据采集方法研究[J]. 包装工程，2021，42(14)：155-163.

[6] 刘多林，吕苗. 基于网络爬虫结合关联大数据的用户信息提取[J]. 计算机仿真，2021，38(8)：482-486.

[7] 戴礼灿，代翔，崔莹，等. 基于深度集成学习的社交网络异常数据挖掘算法[J]. 吉林大学学报(工学版)，2022，52(11)：2712-2717.

[8] 温明章，钱国富. 基于不同客户端软件采集技术的高校图书馆电子资源用户行为数据比较研究[J]. 图书情报工作，2022，66(6)：67-77.

[9] 沈承放，莫达隆，黄文韬. 网页数据采集算法及在住户调查中的应用[J]. 统计与决策，2021，37(7)：52-56.

[10] 张合生，焦鹏，胡琪睿，等. 基于 Jaya 优化标定的高精度数据采集方法[J]. 上海大学学报(自然科学版)，2022，28(3)：361-371.

[11] DE-CÓRDOBA G F，MOLINARI B，TORRES J L. Public debt frontier：A python toolkit for analyzing public debt sustainability[J]. Sustainability，2021，13(23)：13260.

[12] 陈欣，曹朝金，叶春森，等.社会科学数据的创建和使用研究—二次匹配数据采集规则的应用[J].图书情报工作，2021，65(10)：90-104.

[13] BUSTAMANTE G R，NELSON E J，AMES D P，et al. Water data explorer：An open-source Web application and Python library for water resources data discovery[J]. Water，2021，13(13)：1850.

[14] 沈承放，莫达隆，黄文韬. 网页数据采集算法及在住户调查中的应用[J]. 统计与决策，2021，37(7)：52-56.