基于定制主题网络爬虫技术的不良信息检测系统设计

沈冰　周亮　李娟　冯平　刘瑾
（成都市成华区网络信息中心　四川省成都市　610051）

摘　要：本文从营造绿色网络环境出发，在 5G 网络和大数据时代背景下，大量不良信息在网络环境上以几何速度广泛传播，污染了网络环境，面对庞大的网络信息资源，为加强互联网不良信息监控管理和收集，设计了一种基于定制主题网络爬虫技术的不良信息检测系统，实现高效化筛选不良信息。

关键词：检测系统设计；主题网络爬虫；数据采集

随着 5G 网络和大数据信息化的飞速发展，互联网中的信息数量以指数级速度增长。互联网中一方面蕴含着大量权威、真实、科学有益的信息，拓宽了我们的知识结构和交往渠道，但另一方面，社交平台中同时也充斥着大量宣染暴力、淫秽、赌博、邪教等不良信息，造成了网络信息安全危机。面对网络舆情风险日驱高度复杂化、常态化，如何从海量的互联网信息中，快速高效地筛选出某个主题不良信息成为舆情工作人员的重大挑战和亟待解决的现实难题。

1 研究背景

网络信息作为网络社会的体温计与晴雨表，是维护社会稳定的重要依据之一，也是防范与化解意识形态安全重要支撑。当前，我国各级政府部门对网络不良信息的治理逐渐从早期的“随意性”“人治性”“经验性”过渡到“制度性”“规范性”“科学性”。但是面对网络海量信息，人工搜索采集数据耗费时间，爬虫技术可以利用计算机自动地采集大规模数据。为此，在海量互联网数据和专用信息采集间需要构建一个特殊的信息筛选机制，提高专用信息获取效率。网络爬虫（Cralwer）正可以在数据采集和分析上发挥有效作用 [1]。与通用网络爬虫不同，主题网络爬虫可以根据特定算法按照预先设定的主题抓取与主题相关页面，它不是抓取整个互联网的网页，而是专门用于对某个主题的网页进行数据采集。为提高专用信息采集的精确性，本文设计一种基于定制主题网络爬虫技术的不良信息检测系统，旨在为相关人员提供有效的借鉴和参考。

2 爬虫原理与过程

网络爬虫是指一种从互联网上爬取信息的程序或者脚本，是加强互联网不良信息监控管理的基础。世界上最早的网络爬虫程序是 Matthew Gray 于 1993 年编写的，被后人称为“万维网漫游者”，是革命性的创新。网络爬行程序从一个或若干初始网页的 URL（Uniform Resource Locator）开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止，所有被爬虫抓取的网页将会被进行一定的分析、过滤，并建立索引，以便之后的查询和检索，同时网页信息储存到数据库中 [2]。既定爬取策略有广度优先、深度优先等。其框架结构流程如图 1 所示。

3 主题网络爬虫

通用爬虫又称全网爬虫（Scalable Web Crawler），爬行对象从一些种子 URL 扩充到整个 Web，能尽最大可能的采集信息数据。这类网络爬虫的爬行范围和数量巨大，对于爬行速度和存储空间要求较高，占用较多的储存空间和网络带宽资源，爬取网页的效率低下。在实际工作实践中，不希望收集到一些无关信息，迫切需要高速、高效地爬取与主题关联度比较高的网页，这就可以用到主题爬虫技术 [3]，提高信息采集率的有效性。通过对定制主题网络爬虫技术进行探析，根据不良信息特点，遇到相关敏感词语，如淫秽、赌博、色情、邪教、暴力、政治领导人等关键词，判断是否为违规网站，

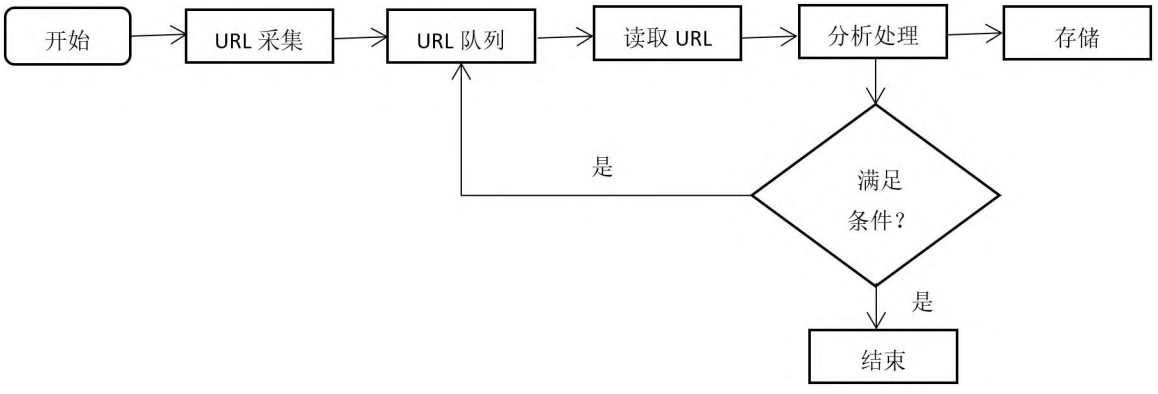

图 1：爬虫基本框架结构流程

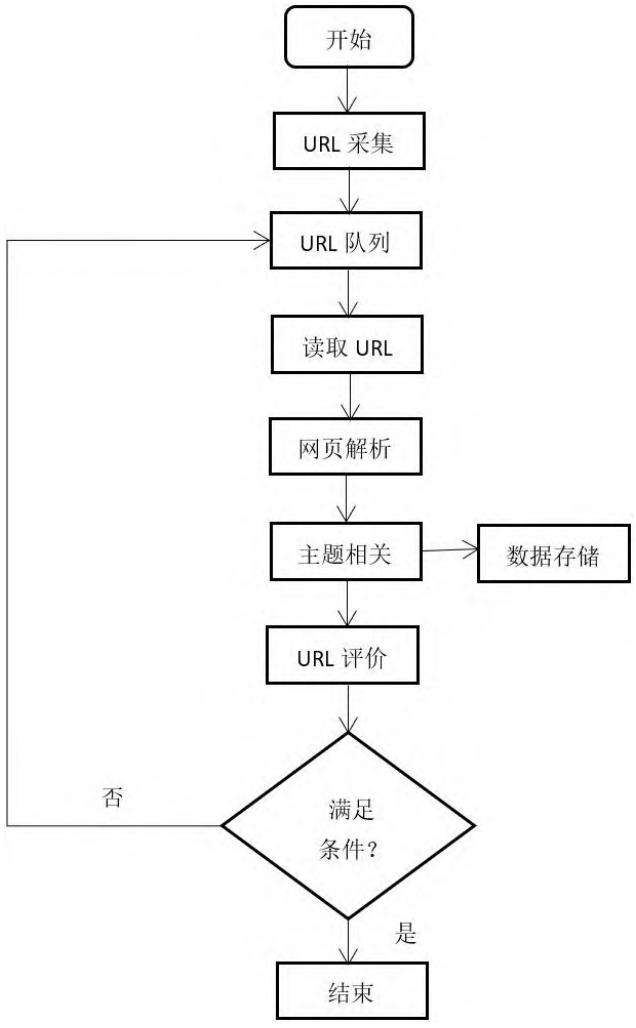

图 2：主题爬虫流程结构

从而提高信息筛选精确度，减少系统负载，提高运行爬行效率。主题爬虫和通用爬虫相比，就是在通用爬虫的基础上，增加了 URL 评价和主题相关度判断，按照系统设定的主题，从既定的初始地址种子集 URL 开始，按照特定的计算方法，分析爬行网页的主题相关度，只需要爬行与主题相关的页面，把与主题不相关的网页过滤去掉，将相关的主题网页储存到数据库中，相应的链接放进待爬行的 URL 队列中，循环往复上述过程，直到符合系统设定的条件为止。主题爬虫框架流程如图 2 所示。主题爬虫能快速的爬取与主题相关度高的网页，提高网页采集的覆盖率和网页的利用率 [4]。

4 系统设计与实现

4.1 架构设计

主题网络爬虫技术的关键就是对互联网信息的有效采集。采集范围主要包括当下热门社交平台，如：知乎、搜狗、新浪新闻、微博、微信公众号、网易新闻、搜狐新闻、360、凤凰、百度等网站。借鉴 OSI（Open System Interconnection Reference Model）网络模型，本系统构架从下到上可具体分为数据采集层、数据分析层、数据存储层和应用层 [5]。上层使用下层的各种服务，每一层负责解决一部分问题，通过各层的协作提供整体解决方案。从逻辑上看，数据源信息经过数据采集、数据清洗、深度分析、任务调度、主题识别、数据存储等环节后送与系统运用层进行调用显示。整个系统采用模块化设计，兼容 Linux、Windows 等多平台操作系统和 Mysql、DB2 等主流 SQL 数据库。

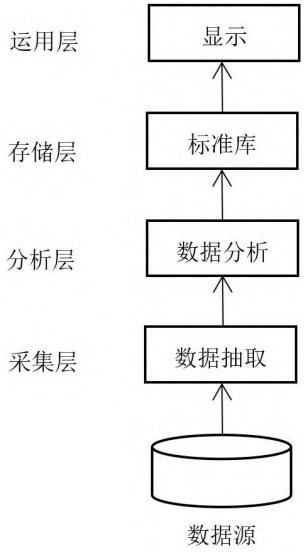

图 3：系统架构图

4.1.1 采集层

采集层位于平台的底部，主要完成域名解析、数据装载、数据转换、队列管理等功能。为了提高采集效率，可通过搭建主题网络爬虫集群服务器来实现。首先构建自己的 IP 地址池，之后模拟用户访问行为，最后不间断扫描和监控网络平台，将采集数据、用户行为和动作日志进行存储，再经上下层数据接口传送给分析层进行数据挖掘、分析和识别。

4.1.2 分析层

数据分析层的作用是对数据进行关联、挖掘、交互、识别分析等。经批处理、流式计算等方式对获取的数据信息进行二次加工，而后将处理过的数据传给存储层，再供运用层调用输出结果。

4.1.3 存储层

数据存储管理模块包含存储、发布以及检索功能。存储层将下层（分析层）送来的数据归类为元数据、标准库数据和其他数据等。

4.1.4 运用层

将分析处理好的信息通过系统具体的业务运用进行输出。系统架构如图 3 所示。

时间	名称	关键词	平台
09:33:08	test	赌博/色情/走私/军火/邪教	百度新闻
09:33:04	test	赌博/色情/走私/军火/邪教	头条
09:33:00	test	赌博/色情/走私/军火/邪教	知乎
09:32:57	test	赌博/色情/走私/军火/邪教	新浪新闻

图 4-1：爬虫结果图

	链接	详情
本地保存	https://www.sohu.com/a/490941030_1...	获取网页成功【中山一市民在色情网站上参与赌博“被洗钱”了！】
远程预览	https://www.sohu.com/a/490941030_1...	开始获取网页【中山一市民在色情网站上参与赌博“被洗钱”了！】
本地保存	https://www.toutiao.com/a7018750157...	获取网页成功【性交易，枪支贩卖，绝赌博，这片解开全球最...
远程预览	https://www.toutiao.com/a7018750157...	开始获取网页【性交易，枪支贩卖，绝赌博，这片解开全球最...
本地保存	https://www.sogou.com/link?url=hedJiaC291MB...	获取网页成功【日本黑帮合法吗？_为什么_】
远程预览	https://www.sogou.com/link?url=hedJiaC291MB...	开始获取网页【日本黑帮合法吗？_为什么_】
本地保存	https://k.sina.com.cn/article_3177450665	获取网页成功【“网恋”+“投资_赌博”=“杀猪盘”】
远程预览	https://k.sina.com.cn/article_317745...	开始获取网页【“网恋”+“投资_赌博”=“杀猪盘”】

图 4-2：爬虫结果图

4.2 功能和数据结构设计

根据上面的架构设计，本文基于主题网络爬虫技术的不良信息检测系统具体功能实现设计可分为：信息采集模块、信息监测模块、信息分析模块、信息显示模块。

4.2.1 信息采集模块

完成数据抽取、数据转换、数据装载等，包括采集时间管理、地址库管理、关键词管理。

4.2.2 信息分析模块

信息分析模块是指将信息采集后通过流式处理对数据进行数据清洗、数据统计、主题识别等，经过分析，找出某网页中不良信息并将其 URL、网站名称等关键信息记录到数据库中。

4.2.3 信息显示模块

经分析模块处理过含有不良信息的 URL、网站名称，会在系统的醒目位置显示，以页面闪动的形式进行重点提醒。

4.2.4 数据结构设计

系统数据结构设计如表 1 所示，包括时间、任务名称、关键词、平台、链接、网页详情等。通过这些设计使主题网络爬虫系统在运用层可直观清晰地看出筛选后含有的不良关键字的网页信息 [6]。

4.3 系统实现和仿真

对比 Python、C、C++、Java 各种主流语言特点，Python 比 C 更加简单，比 C++ 更容易上手，比 Java 更加简洁，而且代码易读性强，因此本系统基于 Python 技术构建 [7]，部分关键代码块如下：

（1）获取网页详细信息；

表 1：数据结构表

名称	记录信息
Time	记录采集的时间
Name	记录任务名称
Key	爬虫关键词
Social	平台信息
Url	网址链接
Details	详细信息

（2）分析页面内容；

（3）储存显示信息；

（4）用户代理 User-Agent 请求头。

```
import requests # 引入 requests 等各类模块
import pymysql
import time
import json
import traceback
# 获取网页详细信息 def getHTMLText(url):
url="https://www.baidu.com/"# 设置要请求的 url 值，以百度为例
Headers={'User-Agent':'Mozilla/5.0(Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/64.0.3282.119 Safari/537.36'}：# 为了更好的伪造自己是浏览器访问的，需要加一个头，让自己看起来是通过浏览器访问
html_code=getHTMLText(url)
# 分析页面内容 def parsePage(ilt, html):
reg_img=re.compile(reg) # 编译为了运行速度更快
linklist=reg.findall(html_code) # 进行匹配查找
# 储存显示信息 def printgetinfo(ilt):
printgetinfo(linklist) # 分析下载并保存数据
```

```
# 随机获取一个用户代理 User-Agent 的请求头
def get_request_headers():
  headers = {
     'User-Agent':random.choice(USER_AGENTS),
     'Accept':'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3',
     'Accept-language':'zh-CN,zh;q=0.9',
     'Accept-Encoding':'gzip, deflate,br',
     'Connection':'keep-alive',
     }
  return headers
# 主函数 def main():
# 程序入口 if __name__=="__main__":
     main()
```

选择特定数据源平台如搜狗知乎、新浪新闻、百度等，经模拟器仿真测试，部分含有敏感词和不良信息爬虫结果网页如图 4-1（含时间、名称、关键词、平台信息）和图 4-2（含网址链接、详细信息）所示。

4.4 关键问题

4.4.1 爬虫被封

在实际运行爬虫时发现，个别网站 Robots 协议限制较严格 [8]，遇到某些网站爬取过多就会出现异常甚至可能会封掉 IP，原因是网站进行了来源审查。即平台检测出爬虫脚本程序后，服务器便自动配置禁止某个 IP 访问。此时要用到反爬虫技术，通过修改请求头 User-Agent 来伪装浏览器进行请求，恢复正常访问网站功能。另外，还可以添加几秒延时，使访问浏览器的行为更接近于自然人正常式访问 [9]。

4.4.2 多线程爬虫

默认情况下，一个程序只有一个线程，代码是依次线性执行的，单线程爬虫每次只访问一个页面，不能并发执行。为了提高信息的采集效率，充分利用资源，爬虫程序应进行并发采集。多进程适合 CPU 密集运算型程序，每个进程都有单独的 GIL, 能实现真正意义上的并行执行 , 所以 Python 中的多进程的执行效率优于多线程。Python 提供多进程库 multiprocessing，用来处理与多进程相关的操作 [10]。Multiprocessing 内有 dummy 模块，利用其中的 Pool 类来实现线程池，从而有效提高信息采集效率。

5 结语

随着互联网、大数据等信息技术的飞速发展，Python 爬虫技术也越来越成熟，被广泛应用在 Web 开发、人工智能和嵌入式等很多领域。基于定制主题网络爬虫技术的不良信息检测系统，利用 Python 标准库和第三方常用库，对不良信息信息进行了定向爬取，能快速高效发现不良信息，方便管理部门及时处置有害信息的传播。经过仿真实验证明，爬虫可以根据用户的需求快速抓取目标数据信息，能够有选择性的进行网页访问，有助于用户快速精准地获取所需信息，具有一定的现实意义。后期需要完善的部分还有许多，例如 Robots 一旦发生了变化 , 就需要及时更新爬虫的解析规则 , 以确保爬虫的正常运行。

参考文献

[1] 孟宪颖 , 毛应爽 . 基于 Python 爬虫技术的商品信息采集与分析 [J]. 软件 , 2021, 42(11): 128-130.
[2] 齐晓霞 , 丁黄法 , 王琦进 . 基于特征集合的 XSS 漏洞安全研究 [J]. 西华大学学报 (自然科学版), 2018, 37(6): 37-41.
[3] 张芳 , 王培进 . 主题网络爬虫技术在高速公路信息采集中的应用 [J]. 烟台大学学报 (自然科学与工程版), 2017, 30(3): 255-257.
[4] 孙冰 . 基于 Python 的多线程网络爬虫的设计与实现 [J]. 网络安全技术与应用 , 2018(4): 38-39.
[5] 韩瑞昕 . 面向分布式的通用网络爬虫系统关键技术研究与实现 [C]. 北京工业大学 , 2020.
[6] 薛丽敏 , 吴琦 , 李骏 . 面向专用信息获取的用户定制主题网络爬虫技术研究 [J]. 信息网络安全 , 2017(02): 12-21.
[7] [德]Katharine Jarmul, [澳]Richard Lawson 著 , 李斌译 . 用 Python 写网络爬虫 (第 2 版)[M]. 人民邮电出版社 , 2018.
[8] 李培 . 基于 Python 的网络爬虫与反爬虫技术研究 [J]. 计算机与数字工程 , 2019, 47(6): 1415-1420, 1496.
[9] 池毓森 . 基于 Python 的网页爬虫技术研究 [J]. 信息与电脑 , 2021(21): 41-44.
[10] 苏国新 , 苏聿 . 基于 Python 的可配置网络爬虫 [J]. 宁德师范学院学报 (自然科学版), 2018, 30(04): 364-368.

作者简介

沈冰（1983-），男，河南省开封市人。硕士研究生，工程师。研究方向为通讯电子技术、网络安全。
周亮（1987-），男，重庆市人。大学本科学历。研究方向为电子政务。
李娟（1982-），女，四川省成都市人。硕士学位。研究方向为网络传播。
冯平（1981-），男，四川省南充市人。硕士学位。研究方向为政务信息化。
刘瑾（1988-），女，四川省成都市人。大学本科学历。研究方向为网络舆情。