面向多数据源的网络爬虫实现技术及应用

曾健荣 张仰森 郑 佳 黄改娟 陈若愚
(北京信息科技大学智能信息研究所 北京 100101)

摘 要 基于大数据技术的社会计算方法是目前学术界研究的热点,如何从网络上快速获取相应的数据资源是相关研究的关键。网络爬虫技术是目前进行网络数据采集的主要手段,针对现有爬虫技术不便于采集多源数据的问题,提出了一种面向多数据源的网络爬虫数据采集技术,在研究新浪微博、人民日报、百度百科、百度贴吧、微信公众号、东方财富股吧等6类媒体平台的数据采集爬虫的基础上,采用Servlet后台调度技术,将面向多数据源的网络爬虫进行融合,解决了面向不同媒体平台的数据采集问题。在实现过程中,首先借助Web应用程序测试工具包selenium实现模拟登录等人工操作,然后采用Xpath元素查询技术来解析网页源码,并提取出数据信息存入数据库,最后将爬取到的数据从数据库中读取出来并展示在前端页面中。实验表明,爬虫在保证数据完整性的前提下实现了采集效率的最大化。

关键词 数据采集,网络爬虫,多数据源,数据展示,信息处理
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.05.047

Implementation Technology and Application of Web Crawler for Multi-data Sources

ZENG Jian-rong ZHANG Yang-sen ZHENG Jia HUANG Gai-juan CHEN Ruo-yu
(Institute of Intelligent Information, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract The research of social computing method based on big data technology is the hot spot in the academic circle, and how to obtain the corresponding data resources from the network is the key to the research. At present, network crawler techno-logy is the main method to collect network data. In light of the problem that the existing crawler technology is not easy to collect multi-source data, this paper proposed a network-crawler data-acquisition technology facing multi-data sources. On the basis of six data collection crawlers on media platforms including Sina micro-blog, People's Daily, Baidu Baike, Baidu Tieba, wechat public account and Easter Wealth Stock Bar, the Web crawlers for multiple data sources are fused to solve the problem of data collection for different media platforms by backstage scheduling technology Servlet. During the implementation process, firstly, the Web application test kit selenium is used to simulate the artificial actions like logining, then the element query technology Xpath is used to analyze the source code of the Web page and extract the data information and put them into the database, finally the data crawled from multi sources are read out from database and displayed on front webpages. Experiments show that the crawler achieves the maximization of acquisition efficiency under the premise of ensuring data integrity.

Keywords Data acquisition, Network crawler, Multiple data source, Data display, Information processing

1 引言

随着历史的车轮步入21世纪的轨道,互联网获得了近20年的黄金发展时期,已然成为了人们日常生活中不可或缺的重要组成部分。如今互联网上的内容浩如烟海,人们几乎可以从互联网上获取任何想要的媒体信息资源,包括各种各样的新闻资讯、娱乐活动、市场行情、学术资料等。作为人类科技文明的重要载体,互联网极大地促进了人类社会的发展,给人们的生活带来了无尽的便利。

但是,互联网在给人们带来便利的同时,也带来了信息选择与整合上的困扰。我国互联网的迅猛发展催生了一大批媒体平台,例如新浪微博、人民日报网、百度贴吧、微信公众号、股票论坛等,在面对如此众多的媒体平台发布的海量信息时人们容易产生眼花缭乱的感觉,从而影响自己原有的判断。

到稿日期:2018-04-19 返修日期:2018-07-28 本文受国家自然科学基金项目(61772081,61602044),北京市教委科研计划项目(KM201711232014)资助。

曾健荣(1995—),男,硕士生,主要研究方向为中文信息处理;**张仰森**(1962—),男,博士后,教授,CCF高级会员,主要研究方向为中文信息处理、人工智能,E-mail:zhangyangsen@163.com(通信作者);**郑 佳**(1991—),男,硕士生,主要研究方向为中文信息处理、情感分析;**黄改娟**(1964—),女,高级实验师,主要研究方向为智能信息处理;**陈若愚**(1982—),男,博士,讲师,主要研究方向为自然语言处理。

此外，这些媒体平台都是分散而独立存在的，当人们想要收集这些媒体平台上的数据资源时必须手动访问对应的网站，这不便于人们对数据信息的收集与整合。因此，基于上述问题，本文创造性地提出了一种面向多数据源的网络爬虫技术，并结合 Web 前后端技术将几种具有典型性与代表性的媒体平台聚合到一起，实现了在媒体平台间的快速切换和自动化采集数据，从而节约了时间、提高了效率、降低了获取信息的成本，有利于人们对互联网上多数据源的信息收集与整合。

2　研究现状

目前网络数据采集主要是综合运用垂直搜索引擎技术的网络蜘蛛（或数据采集机器人）等技术来完成[1]。现阶段在国内从事“海量数据采集”的企业很多，它们大多是利用这种垂直搜索引擎技术来实现。还有一些企业实现了多种技术的综合运用，例如：“火车采集器”采用了垂直搜索引擎＋网络雷达＋信息追踪与自动分拣＋自动索引技术，将海量数据采集与后期处理进行了结合；深圳视界信息技术有限公司的“八爪鱼采集器”以完全自主研发的分布式云计算平台为核心，能在短时间内从不同的网站或者网页获取大量的规范化数据，帮助客户实现数据自动化采集、编辑、规范化，从而削弱对人工搜索及收集数据的依赖。

除了以上提及的商业应用外，国内不少学者也在进行数据采集方面的研究。电子科技大学的刘鹏飞在“京东商城舆情分析系统”中设计了一种面向多数据源，可对采集源、采集深度、采集类别进行详细配置，并按采集频率和并行多线程调度两种方式进行控制的数据采集方法[2]。郑州大学的赵俊采用调用腾讯微博 API 接口的方法，在通过 OAuth 授权认证获得访问令牌 Access Token 后，利用多账号复用和多线程来控制 API 接口的调用频率，从而达到尽可能高效地采集数据的目的[3]。罗咪使用 scrapy 多线程爬虫框架实现了模拟登录、动态网页抓取和克服微博反爬虫机制等功能[4]。在国外，Madhusudan 等为解决搜索引擎的深层网页缺乏索引的问题，提出了一种聚焦语义网络爬虫的方法，该方法首先获取相关网站，然后进一步通过深度搜索重新获取相关网站[5]。Hyo-Jung 等设计并实现了一种能够从深网中自动采集动态生成的网页的爬虫算法，这种方法使用脚本而不是关键字来作为链接[6]。Kumar 等提出了一种基于查询的爬虫方法，使用与用户感兴趣的主题相关的一组关键词，将关键词传递到 URL 对应网站的搜索查询界面，从而获取最相关的链接[7]。

在将跨媒体平台与多源数据采集相结合方面，国内外的相关研究都很少。因此，本文的工作是具有开创性和引领性的。

3　基于多数据源的数据采集方法

多数据源的数据采集涉及多种媒体平台，如微博、贴吧、百科等，多种媒体平台之间的互相切换、数据采集和数据展示是本文所要实现的系统的基础功能，系统的整体架构如图 1 所示。其中，数据采集是整个系统的核心，也是本文研究的重点。

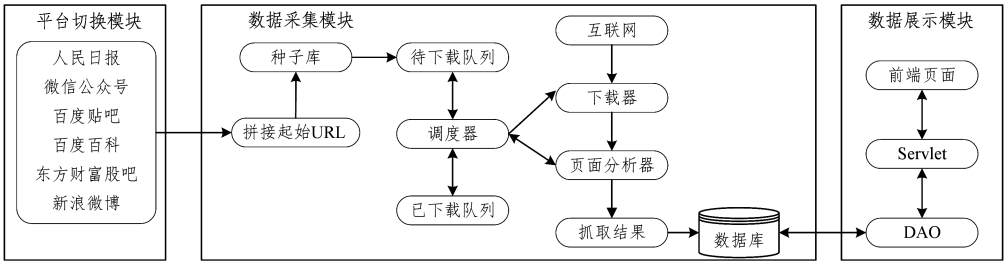

图 1　系统整体架构

Fig. 1　System architecture

在研究面向多数据源的数据采集方法时，由于媒体网站都有各自的特点，网页结构不尽相同，因此需要针对特定的媒体网站编写与其网页结构相适应的爬虫策略。

3.1　网络爬虫算法设计

从本质上来说，爬虫是一种互联网信息采集工具。1993 年初 Matthew Gray 实现了首个互联网爬虫 Wanders。在互联网会议上也陆续出现了越来越多的关于网络爬虫的学术论文[8-10]。

网络爬虫按照系统结构和实现技术可以分为以下几种类型：通用型网络爬虫[11]（General Purpose Web Crawler）、聚焦型网络爬虫[12]（Focused Web Crawler）、增量式网络爬虫[13-14]（Incremental Web Crawler）、深层网络爬虫[15]（Deep Web Crawler）。由于不同媒体平台的网页结构复杂、形式多样，无法使用单一类型的网络爬虫，因此本文将通用型爬虫和深层网络爬虫两类技术相结合来实现数据的采集。

本文采用广度优先遍历算法，设计了图 2 所示的网络爬虫。

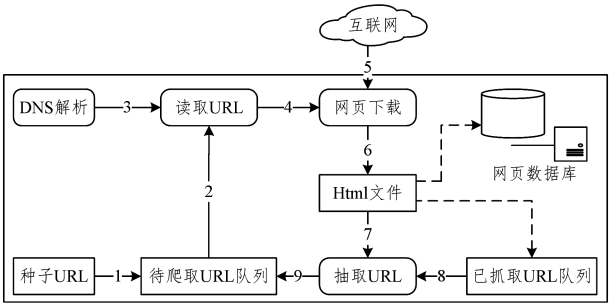

图 2　网络爬虫的基本原理

Fig. 2　Basic principle of Web crawlers

使用广度优先遍历策略的原因主要有以下 3 点：

1）重要的网页往往离种子比较近，例如打开人民日报、新浪微博网页时往往是最新、最热门的新闻，随着不断地浏览，

所看到的网页的重要性越来越低。

2)广度优先有利于多爬虫的合作抓取,多爬虫合作通常先抓取站内链接,抓取的封闭性很强。

3)深度优先搜索的优点是能遍历一个 Web 站点或深层嵌套的文档集合,缺点是 Web 结构相当深,有可能出现一旦"进去",再也"出不来"的情况。而本文要抓取的 6 类媒体平台网站的深度都较浅,深度优先搜索在这种情况下不大适用。

本文算法使用了两个 Map〈String,Boolean〉数据结构,键值对分别是链接和是否被访问的标志,这两个 Map 分别存放种子链接的 oldMap 和新链接的 newMap。算法描述如算法 1 所示。

算法 1 数据爬取算法

输入:种子 url

输出:各字段数据信息

Step1 将种子链接放在 oldMap 中,访问标志值置为 false;

Step2 若 oldMap 中无标志值为 false 的链接,转 Step9,否则取出其中首个 false 值的链接,向其发送请求,获得页面源码;

Step3 解析页面源码,提取各字段的数据信息和〈a〉标签下所有符合目标网站的链接;

Step4 把数据信息写入数据库中;

Step5 对于提取到的每一个链接,若其在 oldMap 和 newMap 中,则进行下一步,否则将其添加到 newMap 中并将标志值置为 false,表明未被访问过;

Step6 把 oldMap 中当前页面的链接的标志值改为 true,表明已经访问过;

Step7 若 newMap 不为空,将其中的链接全部添加到 oldMap 中;

Step8 转 Step2;

Step9 若 newMap 为空,则进行下一步,否则将 newMap 中的新链接添加到 oldMap 中,访问标志值置为 false,转 Step2;

Step10 任务结束,返回链接集合 oldMap。

此外为了提高数据采集的效率,采用多线程设计爬虫。Java 语言自身提供对多线程的支持,应用程序继承或者实现对象的不同,多线程有两种方式:一种是并发运行的对象直接继承 Java 线程类 Thread;另一种是定义并发执行对象实现 Runnable 接口。算法 1 在程序的具体实现上采用了第一种方式,实现了爬虫线程类 CrawlerThread。CrawlerThread 类继承自对多线程控制的 ThreadController 类,当待爬取队列中存在 URL 或未到达指定的层数时,创建一个新的线程,并且通过参数限定了爬取页面的层数和最大线程数。当没有需要爬取的 URL 时,CrawlerController 自行终止,通过消息系统通知 ThreadController,由 ThreadController 进行队列的转换工作。

3.2 基于新浪微博平台的数据采集方法

新浪微博需要用户登录验证才能进行无限制访问,有两种新浪微博服务器可供选择,一种是 weibo.cn 服务器(手机版微博),另一种是 weibo.com 服务器(电脑版微博)。手机版微博的页面相对电脑版的页面更加简洁,网页源码更少,登录账号和密码都不加密,登录也不需要填写验证码,而且所需信息全面。采用电脑版模拟登录时需要填写验证码,增加了自动爬取微博数据的难度,而且 JavaScript 和广告图代码很多,会降低 html 源码的分析效率,从而增加网络传输压力[16]。因此,采用手机版服务器进行模拟登录和解析源码、提取正文内容。

虽然手机版微博相比于电脑版微博要更方便采集数据,但是新浪微博的页面内容丰富,即便是手机版也使用了 AJAX(Asynchronous Javascript And XML,异步 JavaScript 和 XML)技术来动态加载数据,因此必须借助 selenium 工具包模拟用户的操作,从而在加载数据到页面中。selenium 框架底层使用 JavaScript 模拟真实用户对浏览器进行操作,执行其测试脚本时,浏览器自动按照脚本代码做出点击、输入、确定、验证等操作。

借助 selenium 工具包来采集新浪微博数据的基本步骤为:模拟登录、爬取用户页面的网页源码、页面解析和提取各字段内容并保存到数据库中。其中模拟登录是前提,解析网页源代码以提取正文是关键。

3.2.1 手机版微博的模拟登录

用 Chrome 浏览器的开发者模式分析手机版的登录方式的步骤为:

1)打开手机版微博登录 URL:passport.weibo.cn/signin/login,服务器返回一个带有用户名输入框和密码输入框的页面;

2)模拟输入用户名和密码,向微博服务器登录 URL 发送一个请求,该请求包含明文形式的用户名和密码;

3)微博服务器对收到的登录请求进行验证,登录成功后向客户端返回一个重定向 URL,并且 cookie 中包含 gsid_CTandWM 字段,浏览器解析该跳转 URL 进入登录成功页面并把所有 cookie 字段写入本地 Cookies 中。

基于以上分析,在程序中先加载浏览器驱动(以 chrome 浏览器为例),实例化一个浏览器对象,并根据此浏览器对象模拟出对应的登录步骤:

1)获取用户名输入框,输入登录用户名:

```
# hud 即是浏览器对象
WebElement mobile = hud.findElementByCssSelector ("input[name=mobile]");
mobile.sendKeys(new CharSequence[]{username});
```

2)获取密码输入框,输入登录密码:

```
WebElement pwd = hud.findElementByCssSelector ("input[name=password]");
pwd.sendKeys(new CharSequence[]{passwor d});
```

3)点击按钮提交输入:

```
submit.click();
```

3.2.2 爬取手机版微博网页

登录成功后注入 cookie 就能获取网页源码。

1)浏览器中注入 cookie

在请求微博网页方面,当启动 HttpClient 浏览器代理时,把通过 html 获取到的 cookieSet 注入进去。

2)获取微博页面源代码

通过 HttpClient 获取微博 html 源代码的具体流程是:

①把要访问的 URL 传给要执行的 Get 请求(这是因为单纯访问简短 URL 时，用 Get 请求更合适)；

②执行 Get 请求，服务器返回一个响应对象，通过该对象获取具体的 html 源码。

至此就得到了 AJAX 生成的动态页面信息，用层叠样式表(Cascading Style Sheets，CSS)选择器结合正则表达式就能够定位网页文档对象模型(Document Object Model，DOM)树中的节点，可抽取相关信息，包括用户 ID、微博数量、关注数量、粉丝数量、微博内容、点赞数量、转发数量、评论数量、微博发布时间等数据。

3.3　基于其他媒体平台的数据采集方法

百度贴吧也有小部分数据是动态加载的。如果以抓取静态页面的方法抓取百度贴吧的数据，虽然能抓取到部分数据，例如帖子标题、发帖人信息等，但是包含帖子每层楼内容的 html 源码却无法加载，只有等到浏览器显示这个页面时，JavaScript 脚本才会运行，从而显示贴子每层楼的具体内容。这时有两种方法可供选择：一种方法是像抓取新浪微博那样，分析 AJAX 请求，找到对应的加载数据的 JavaScript 脚本，并分析其逻辑，生成一个 http 请求，从而通过代码模拟该请求来获取数据；另一种方法是采取其他页面解析方式，用 Xpath 替代 CSS 选择器来抽取页面节点，进而获取数据信息。第一种方法需要研究 JavaScript 代码逻辑，还要依赖 selenium 自动化测试工具包，过程相对来说更繁琐冗杂。本着奥卡姆剃刀原理的“就简原则”，在此情形下采用第二种方法，用 Xpath 从网页源码中定位指定的元素[17]。事实证明，这种方法简单有效，能顺利抓取到贴吧名称、帖子 ID 和标题、每层楼的主要回帖以及对应的用户信息(包括用户 ID、名称、性别、账号等级、个人主页)等内容。

对于人民日报、百度百科、微信公众号、东方财富股吧，由于这 4 类网站没有涉及 AJAX 请求，因此完全可以将其当作静态页面来爬取，只须分析每一类网站的网页链接和网页源码的规律，找到包含待爬取信息的节点，就能获得该数据信息。以人民日报为例，2018 年 2 月 3 日发布的内容链接为 http://paper.people.com.cn/rmrb/html/2018-02/03/nbs.D110000renmrb_01.htm，只需要将“2018-02/03”替换为待抓取日期，就得到了该日期内容的网页链接。因此在前端页面设置一个日历框来选择待抓取日期，后台就能根据这个日期拼接出一个完整有效的 URL 作为种子链接，进而开始抓取该日期的文章。而百度百科的 URL 链接形式为 https://baike.baidu.com/item/＋词条的 URL 编码，因此只须输入要爬取的关键词，后台对获取的关键词进行 URL 编码，就能得到完整的 URL 作为种子链接。这样就能够灵活地确定爬虫的入口地址，而不需要手动地输入一个完整链接。

综上所述，除了新浪微博因游客身份访问受限，只能靠模拟登录的方式才能正常抓取页面以外，其他 5 类媒体平台都能在不登录的情况下无限制访问，并且都能以 CSS 选择器或 Xpath 方式解析页面[18]，抽取相关节点，提取想要的数据信息并保存到数据库中。

3.4　多数据源之间的切换方法

一般的网络爬虫系统多是单数据源，即只针对单一的一个网站平台爬取数据，而本文提出的面向多数据源的网络爬虫实现技术融合了多个网站平台，能采集这些多数据源的文本信息。本文的多数据源包括新浪微博、人民日报网站、百度贴吧、百度百科、“传送门”微信公众号网站、东方财富股吧 6 种媒体平台的数据源，在采集不同数据源的信息时，要先切换到待采集数据源上，图 1 中的平台切换模块即实现在这些数据源之间切换的功能。

多数据源切换方法的主要思想是：通过下拉菜单栏选择某一种媒体平台，进入到该平台的数据采集页面，实现“前端换源”；然后通过 AJAX 技术向后台发送附带参数的请求，根据参数内容确定究竟应该调用哪一种数据源对应的 Servlet，进而调用相应的爬虫程序，从而实现真正的换源。

以百度贴吧为例，在其他数据源采集页面下要切换到百度贴吧，先从“数据采集”下拉菜单栏中选择百度百科，进入到数据采集页面，如图 3、图 4 所示。

图 3　平台切换界面

Fig. 3　Platforms switching interface

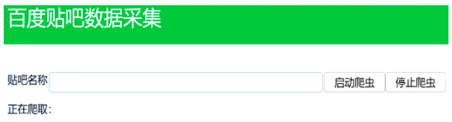

图 4　数据采集界面

Fig. 4　Data acquisition interface

然后输入要爬取的贴吧名称，点击“启动爬虫”按钮，即可触发 JavaScript 脚本，并向后台发送 AJAX 请求。AJAX 请求的“URL”参数指明了此请求要被提交到的 Servlet，在 Servlet 中将执行贴吧的采集程序。以上过程就是多数据源切换方法的执行流程。

4　多数据源的展示方法

爬取到的数据存储在数据库中。事先为各媒体平台设计相应的数据表，展示数据时就从这些表中读取数据并返回给前端页面。

4.1　数据表的设计

数据库选用 MySQL，为各类媒体平台创建对应的数据表，各数据表中包含要爬取的信息字段。媒体平台、数据表和字段信息的对应关系如表 1 所列。

表 1 各数据表的设计

Table 1 Design of each data table

媒体平台	数据表	字段
新浪微博	weibocontent	contentid,userid,content,likes,transfers,comment,time,platform
	weibouser	userid,nickname,birthday,gender,marriage,address,signature,weibonum,tags,followsnum,fansnum
微信公众号	wechataccounts	wechatid,name,description
	wechatarticle	Wechatid,time,title,content
人民日报	rmrb	id,title,author,content,date,categorytitle
百度贴吧	tiebapost	titlid,postid,content,username,isanonymous,date,commentnum
	tiebatitle	tiebaname,titleid,title,replynum,href
	tiebauserinfo	userid,username,usersex,levelid,levelname,currentscore,homepage
百度百科	baike	keywordid,keyword,content
东方财富股吧	stock	namenum,time,commentcontent,commentauthor,commentage,commentinfluence

在爬虫程序运行过程中,同一张表中的一条记录都采集下来后就执行 SQL 插入语句,并把数据写入数据库中。

4.2 数据展示方法

数据展示功能是在前端页面上展示抓取到的数据信息,主要采用了 AJAX 技术和 Bootstrap 框架来实现。数据信息在前端页面中按每一种媒体平台分类,再以表格的形式呈现出来,如表 2 所列。

表 2 数据展示界面内容

Table 2 Content of data display interface

微信号	公众号名称	功能介绍
xmwb1929	新民晚报	新民晚报是一张有温度、有力度、有维度、有尺度的报纸……
dichan360	房地产经理人联盟	房地产经理人公益平台,每天用实战案例的形式分享名企…
jccuimian	京城催眠师	向心理爱好者介绍心理学知识,以便更好地让自己处于……
lonelybrain	孤独大脑	关于思考的思考
ishengtaiwang	田野观察 AgriReview	田野观察(原爱生态网),做互联网十新农业最有深度……
Queenzhuyi	Queen 主义	为数万女孩提供服务的女神修炼顾问和互动平台;原创……
zhoumoqunawan	周末去哪玩	每日发布各城市最新最全最干货活动,告诉你周末去哪儿玩……
Tianya-China	天涯社区	【天涯社区】唯一官方账号
cctvnewscenter	央视新闻	中央电视台新闻中心公众号,提供时政、社会、财经……

表 2 中并没有展示所有字段,这是由于数据表中的多字段是为了保证信息的完整性,以便于后续的数据分析和挖掘工作,但是用户可能对某些边缘字段的信息并不关注,而只对一些关键、核心的数据感兴趣,因此在前端页面中只对用户关心的数据进行选择性展示。例如百度贴吧的 tiebapost 表中,舍去 titleid,postid,isanonymous 3 个边缘字段,只展示 content,username,date,commentnum 4 个核心字段的信息。

这里表格和分页效果通过应用 Bootstrap 框架的 table 插件来实现。当选择每页显示 m 条记录或者点击第 n 页时都会向后台发送 AJAX 请求,该请求包含了每页显示的记录数 pageSize、当前页数 pageNumber 和请求要提交到的 Servlet 名称 3 个参数,Servlet 根据前两个参数从数据库中读取指定的数据并返回给前端页面,页面局部刷新表格部分,将数据显示出来。由于数据量多,这里的分页采用的是服务器端分页,即在后台程序中取得当前页面需要加载的数据,否则采用客户端分页一次性将所有分页的数据加载到浏览器缓存中,容易卡顿,从而影响用户体验。

5 实验及结果分析

测试机器硬件配置为英特尔第四代酷睿 i5-4200H @ 2.80GHz 双核、8GB 内存,操作系统为 Windows 10 专业版,开发和测试平台为 Eclipse Neon. 3 Release (4. 6. 3)搭配 Apache Tomcat 8.5.31 版本 Web 服务器,浏览器选择 64 位版本的 Google Chrome 66.0.3359,数据库选择 64 位版本的 MySql 5.6.36,全部程序由 Java 开发实现。

当网络通信情况良好时,将 cn 版微博与文献[19]中的新浪微博 API、传统爬虫、com 版微博爬虫的爬取效率进行比较,结果如图 5 所示。从图 5 可以看到,cn 版微博爬虫的效率高于其他方式的微博数据采集方法。6 类数据源的平均数据采集效率如图 6 所示,可以看出,每一类媒体平台的数据采集效率都接近于 42000 条记录/h,说明此网络爬虫具有良好的性能。

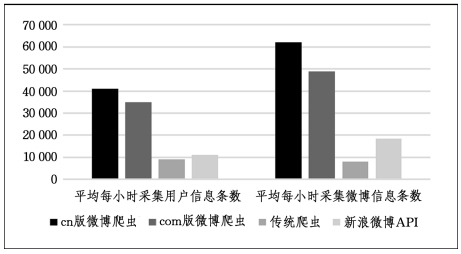

图 5 cn 版微博爬虫与其他微博爬虫的效率比较

Fig. 5 Comparison of efficiency between cn version of Weibo crawler and other Weibo crawlers

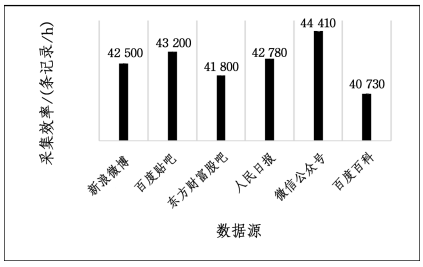

图 6 各数据源的平均数据采集效率

Fig. 6 Average data collection efficiency of each data source

结束语 本文以新浪微博、百度贴吧、百度百科、人民日报、微信公众号、东方财富股吧 6 种具有代表性的媒体平台为研究对象,提出了一种面向多数据源的网络爬虫技术实现方法,并介绍了其应用,旨在解决爬虫不便于采集多数据源信息的问题。

在前端页面中将这 6 类数据源融合到一个下拉菜单里。在切换数据源方面，利用附带参数的 AJAX 技术将采集请求提交到指定的 Servlet 类，实现了在不同数据源之间的切换。在数据采集方面，针对动态加载页面数据的新浪微博采用模拟登录、注入 cookie 的方式实现采集效率的最大化，对其余的静态网站直接使用 Xpath 技术定位页面元素，实现数据采集功能。在数据展示方面，采用服务器端分页的方式，将 AJAX 请求提交到指定的 Servlet，该 Servlet 根据 AJAX 参数从数据库中读取当前页面所需的数据并返回给前端页面。实验数据表明该网络爬虫技术效率较高，表现出了良好的性能。

现阶段网络爬虫的相关技术和算法还有待改善，本文提出的面向多数据源的网络爬虫技术也存在很多不足。比如爬取网页数据是通过 html 元素的路径特征来定位，这需要人工分析每类网页中节点的规律。如何将元素路径特征与文本块密度特征相结合，从而实现一种快速、通用的，适用于大数据环境下多语言、多风格的异构网页爬取方法，是下一步需要研究的工作。

参 考 文 献

[1] GAO F，LIU Z，GAO H. A universal vertical crawler with supervised breadth first search strategy[J]. Computer Engineering，2018，44(11)：295-305. (in Chinese)
高峰，刘震，高辉. 结合有监督广度优先搜索策略的通用垂直爬虫[J]. 计算机工程，2018，44(11)：295-305.

[2] LIU P F. Design and Implementation of Jingdong Mall Public Opinion Analysis System[D]. Chengdu：University of Electronic Science and Technology of China，2015. (in Chinese)
刘鹏飞. 京东商城舆情分析系统的设计与实现[D]. 成都：电子科技大学，2015.

[3] ZHAO J. Research on data acquisition and analysis method of social network[D]. Zhengzhou：Zhengzhou University，2015. (in Chinese)
赵俊. 社交网络的数据采集与分析方法研究[D]. 郑州：郑州大学，2015.

[4] LUO M. User data acquisition technology of sina micro-blog based on Python [J]. Electronics World，2018(5)：138-139. (in Chinese)
罗咪. 基于 Python 的新浪微博用户数据获取技术[J]. 电子世界，2018(5)：138-139.

[5] MADHUSUDAN P A，POONAM D. Deep web crawling efficiently using dynamic focused web crawler[J]. International Research Journal of Engineering and Technology，2017，4(6)：3303-3306.

[6] OH H J，WON D H，KIM C，et al. Design and implementation of crawling algorithm to collect deep web information for web archiving[J]. Data Technologies and Applications，2018，52(2)：266-277.

[7] KUMAR M，BINDAL A，GAUTAM R，et al. Keyword query based focused Web crawler[J]. Procedia Computer Science，2018，125：584-590.

[8] PRUTHI J，MONIKA. Implementation of Category - Wise Focused Web Crawler[M]// Big Data Analytics. Advances in Intelligen Systems and Computing. Springer，Singapore，2017：565-574.

[9] KIM T J，KIM H J. Machine Learning-Based Topical Web Crawler：An Ensemble Approach Incorporating Meta-Features [J]. Journal of Engineering and Applied Sciences，2017，12(18)：4651-4656.

[10] AGRE G H，MAHAJAN N V. Keyword focused web crawler [C]// Proceedings of IEEE International Conference on Electronics and Communication Systems (ICECS). New York：IEEE Press，2015：1089-1092.

[11] SUN B. Design and implementation of multi thread crawler based on Python [J]. Network Security Technology & Application，2018(4)：38-39. (in Chinese)
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.

[12] HU P R，LI S J. Focused crawler based on URL patterns [J]. Application Research of Computers，2018，35(3)：694-699，726. (in Chinese)
胡萍瑞，李石君. 基于 URL 模式集的主题爬虫[J]. 计算机应用研究，2018，35(3)：694-699，726.

[13] MENG Q H. Design and Implementation of Internet Data Incremental Acquisition System [D]. Beijing：Beijing University of Posts and Telecommunications，2015. (in Chinese)
孟庆浩. 互联网数据增量采集系统的设计与实现[D]. 北京：北京邮电大学，2015.

[14] WANG L Y. Research On Hot News Topic Detection Of Incremental Clustering [D]. Nanning：Guangxi University for Nationalities，2017. (in Chinese)
王丽颖. 增量式聚类的新闻热点话题发现研究[D]. 南宁：广西民族大学，2017.

[15] LIU Y，ZHENG C H. Research on Deep Network Crawler Based on Scrapy [J]. Computer Engineering & Software，2017，38(7)：111-114. (in Chinese)
刘宇，郑成焕. 基于 Scrapy 的深层网络爬虫研究[J]. 软件，2017，38(7)：111-114.

[16] SUN Q Y，WANG J F，ZHAO Z Q，et al. A micro-blog data acquisition scheme based On simulated login [J]. Computer Technology and Development，2014，24(3)：6-10. (in Chinese)
孙青云，王俊峰，赵宗渠，等. 一种基于模拟登录的微博数据采集方案[J]. 计算机技术与发展，2014，24(3)：6-10.

[17] WANG S Y，WANG X，SUN J Z. Web application test script repair based on XPath [J]. Application Research of Computers，2017，34(5)：1393-1396. (in Chinese)
王曙燕，王璇，孙家泽. 基于 XPath 路径的 Web 应用测试脚本修复[J]. 计算机应用研究，2017，34(5)：1393-1396.

[18] SHI S S. Research on key technology and system of accurate Web information extraction [D]. Nanjing：Nanjing University，2017. (in Chinese)
施生生. 精确 Web 信息抽取关键技术与系统研究[D]. 南京：南京大学，2017.

[19] YU Y. Web crawler data collection for Weibo [J]. China CIO News，2017(12)：36-37. (in Chinese)
于营. 面向微博的网络爬虫数据采集[J]. 信息系统工程，2017(12)：36-37.