“互联网 + 教育”环境下网站信息采集系统开发研究

况富强
（济源职业技术学院，河南 济源 459000）

摘要：高校上级管理部门繁多，新的要求、动态不能及时获取。经常出现新的要求、动态在上级主管部门网站的新闻、公告已发布多日，通过公文流转到高校后，学校准备申报或上报资料时间所剩无几，工作上造成被动，失去先机。本文开发一套数据采集系统，主要用来抓取主管部门网站的新闻、公告等内容，定时给学校相关各级领导分类推送上级要求和指示，以便及早谋划，助力高校高速发展。

关键词：高校；动态获取；数据采集

随着互联网项目的发展，互联网信息的渠道越来越多，如何有效地获取所需信息并将信息进行整合成为我们当前面对的首要问题。针对高校而言，上级管理部门繁多，政策公告、通知要求等随时下发，若不能及时获取会导致反应滞后，工作被动，不利于领导层决策的展开。为了能够及时了解到互联网平台的所需信息，利用爬虫技术研发一套内容分析平台是很有必要的，它能第一时间将热点信息自动采集、分析、推送给用户。针对高校获取通知公告信息滞后的痛点，拟开发一套网站信息采集系统，以达到能够灵活、迅速地抓取网页中大量非结构化的文本、图片、视频等资源信息，并对信息进行智能化分析研判的目的。本文将从数据采集方面阐述爬虫在网络内容采集方面的应用。

一、系统功能及设计思路

网站信息采集系统的功能需求为：可跨站点、多栏目采集；可指定时间段不间断爬取；要求配置简单，采集配置通用性要好；采集到的内容方便二次加工；有容错机制。基于以上要求，系统采用 Java 实现信息采集。用户可通过配置待采集站点信息和栏目信息，后台自动生成采集任务，按照配置信息采集需要的文章列表；支持分页采集和全量 / 增量采集配置；可对采集结果进行过滤和批处理；为实现多任务异步采集，项目引入线程池，提高采集效率同时降低资源消耗；同时为降低用户配置难度，提高用户体验率，系统还应增加配置校验功能，方便用户对配置参数进行自查以及采集效果的预览；当采集失败时，要有异常记录，用户可针对性地对这些记录进行重新采集等操作；最后针对五花八门的站点，采集功能通用性一定要做好，采集规则的实现上要兼顾多种内容格式。

二、系统架构设计

项目采用 BS 架构，前后端不分离方便部署实施，前端采用 VUE+ANTD 进行页面展示，包含采集配置、内容展示、操作管理等内容；后端采用 Java 多线程实现异步采集，采用多任务机制实现定时不间断采集，在采集功能基础上增加异常处理、信息推送等功能。出于安全性考虑，系统不会过多地依赖外部框架，在保证功能完备的前提下尽可能优化代码，以提高性能。

三、系统实现方案

采集方案：网站数据采集本质上是通过网站 URL 拿到数据后对数据进行分析筛选，获取到所需数据的过程。此项目用户先配置要采集的站点信息（包括开始结束时间、采集间隔等），再配置该站点下栏目，包括栏目地址、栏目 dom 位置、文章 dom 位置等，配置期间可通过 dom 校验入口检查输出采集结果是否符合预期。配置完毕后系统会根据配置自动生成任务，根据站点及栏目配置情况从线程池中抽取线程执行采集任务，项目会按照“站点 -->栏目”的顺序根据 dom 内容匹配到栏目下的文章 URL 列表，然后遍历文章详情页面获取所需要的属性信息经过后台处理后入库，然后重复此过程，直到达到站点配置的结束条件。配置了全量采集的会对全部内容进行采集，筛选掉重复部分后入库，增量采集则每次只采集最新页面新增内容。

DOM 配置方案：dom 配置分为栏目 dom 和文章 dom。栏目 dom 主要负责将栏目页面的内容元素指定到文章列表这一层，通过栏目 dom 的配置，系统能够采集到当前栏目的文章列表（标题、URL 等），为后面采集正文内容打基础。文章 dom 主要负责定位文章详情页中各个元素的位置，通过文章 dom 配置，系统能过获取到文章详情页需要的元素（标题、日期、作者、正文等），这些元素是采集系统最终需要的采集结果。dom 采用 json 格式存取，用户仅需配置所需各元素标记，系统自动拼接处理。对于采集结果需要排除的内容可于 dom 标记中以“排除内容”的方式提交。dom 配置支持自定义字段。

采集结果处理方案：用户可配置将采集结果采用企业微信或邮件通知的方式进行提醒。对于采集异常的文章资源，用户可在日志管理中进行重新采集。

扩展功能：后续二次开发可采用同样的“模块开发 + 内嵌”的方式融入项目，减少模块之前的耦合，方便扩展。接口扩展方面，项目采用 Restful 编程风格，提供丰富的 API 接口与测试接口，方便在此基础上进行二次开发。

四、网站信息采集系统实现

（一）信息采集器线程设计

采集器线程主要涉及轮训执行时间，创建可缓存线程池对象，启动采集任务，统计采集日志以及系统日志清理周期等，部分实现代码如下：

```
public static void spiderThread() {
    final long timeInterval = 10;
    Runnable runnable = new Runnable() {
      public void run() {
        RasLog raslog = new RasLog();
        ExecutorService cachedThreadPool = Executors.newCachedThreadPool();
        ThreadPoolExecutor threadPoolExecutor = (ThreadPoolExecutor) cachedThreadPool;
        threadPoolExecutor.setMaximumPoolSize(20);
        threadPoolExecutor.setKeepAliveTime(5, TimeUnit.SECONDS);
        threadPoolExecutor.allowCoreThreadTimeOut(true);
        SiteService.setDefaultSpiderStatus();
        while (true) {
          try {
             SpiderService.start(cachedThreadPool);
            StatisticService.Start();
            TimeUnit.SECONDS.sleep(timeInterval);
            LogService.clearLog(0, 7);
                } catch (InterruptedException e) {
            e.printStackTrace();
            raslog.load("[ spider采集器 ]: " + e.getMessage());
          }}}};
    Thread thread = new Thread(runnable);
    thread.start();
  }
```

（二）采集站点设计

```
public static void gatherSite(String siteid, String is_public, String site_name) {
    SiteService.updateSpiderStatus(siteid, 1);
    List<Map<String, Object>> list = ChannelService.getChannels(siteid);
    raslog.info("【" + site_name + "】[ " + RasUtil.getCurrentDate() + "] 开始采集: ,栏目数: " + list.size());
    if (list != null && list.size() > 0) {
      for (Map<String, Object> chnMap : list) {
        chnMap.put("is_public", is_public);
        SpiderChannelService.gatherChannel(chnMap);
      }
    }
    SiteService.setNext_time(siteid);
    String info = site_name + "采集结束，共" + list.size() + "个栏目。";
    LogService.saveSystemLog(siteid, 1, info);
    SiteService.updateSpiderStatus(siteid, 0);
}
```

五、采集系统实现及测试

网站信息采集系统的部署：应用系统采用两台服务器，采用一台数据库服务器，服务器操作系统均采用 CentOS7.9 版本。系统服务由系统软件（war 包）+ 运行容器（Tomcat）+ 数据库（MySQL）组成，系统软件放在容器目录下，跟随容器启动。

采集系统分三大模块，五个菜单，十个子项 . 三大模块分别是：采集模块、检索模块、推送模块，其中采集模块包含：待采集站点、栏目配置，采集结果展示、采集日志记录等；检索模块包含索引、分词等全文检索相关的配置，推送模块包含信息发送人及接收人信息配置。网站信息采集系统界面如下：

（一）信息管理

用于查看采集结果和检索结果。查看信息：主要市查看采集结文章的标题、站点名称、栏目名称、发布时间及原文链接等信息。信息检索：主要是检索采集的信息，用户可以根据标题检索，也可以检索信息来源，或者根据作者检索，也能通过内容进行检索。

（二）采集设置

采集相关配置入口。站点：采集站点配置，主要是采集站点名称的设置，采集开始时间和结束时间，采集状态，下次采集时间等。栏目：采集栏目、文章配置，主要是采集站点下具体的采集栏目的设置，采集间隔时间为 10 秒。

（三）推送设置

邮件推送相关配置。全局配置：邮件推送人信息配置，主要包括邮件配置键值，名称，邮箱账户、密码邮箱及 SMTP 信息，同时还需要配置发送人名称、邮件标题、内容模板及发送时间。用户订阅：邮件接收人信息配置，主要包括邮箱接收人姓名、用户名、电子信箱、订阅站点、推送日报、发送邮箱和企业微信状态按钮等信息。群机器人：主要配置企业微信群机器人 API 接口信息。标签：用于对采集站点信息做分类标记。

（四）日志管理

用于查看各栏目文章采集情况。采集日志：采集日志记录结果，对于采集失败的内容，可以做一个记录，有利于分析采集失败的原因。

六、结束语

本文基于 java 语言对上级部门网站信息进行了定向爬取。根据系统运行情况证明，数据采集系统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，给学校领导及时获取上级指示提供了极大的便利。但是需要完善的部分还有许多，在数据量较大时，采集系统的速度会减慢，可以尝试使用分布式爬虫进行采集，有利于采集效率的进一步提高。

参考文献：

[1] 肖新凤，张绛丽 . 基于 Python 的爬虫技术的网站设计与实现 [J]. 现代信息科技，2020，4（14）：73-75 + 78.

[2] 刘硕 . 精通 scrapy 网络爬虫 [M]. 北京：清华大学出版社，2017.

[3] 洪伟 . 分布式网络爬虫系统设计与实现 [D]. 沈阳理工大学，2020.

[4] 赵北庚 . 基于 Flask 与爬虫技术的可视化深度学习数据标注系统 [J]. 电子制作，2020（20）：36-37.

基金项目：济源市社科联项目（项目编号：JYZY-2022-46）

作者简介：况富强（1985-），男，硕士，研究方向：计算机软件、射频电路与微波天线。