基于 python 的聚焦网络爬虫数据采集系统设计与实现

杨国志 江业峰
(辽宁科技大学,辽宁 鞍山 114000)

摘 要:人类社会已经进入大数据时代了,随着互联网的迅猛发展,种类繁多,数量庞大的数据随之产生,作为辅助人们检索信息工具的搜索引擎也存在着一定的局限性,如:不同领域,背景的用户往往具有不同的检索目的和需求,通用搜索引擎所返回的结果包含大量用户不关心的网页。 为了解决这个问题,网络爬虫系统应运而生。众所周知,搜索引擎从互联网中靶向性筛选出有用信息,而网络爬虫又是搜索引擎的基础构件之一。本文实现了一个基于 python 语言的聚焦网络爬虫,利用关键字匹配技术对目标网站进行扫描,得到所需数据并抓取。

关键词:搜索引擎网络爬虫;python;网页分析算法
中图分类号:TP393 文献标识码:A 文章编号:2096-4390(2018)27-0073-02

1 概述

网络爬虫(Crawler)是搜索引擎(search engine SE)的基本构件之一,其直接面向互联网底层,它是搜索引擎的数据发源地,决定着整个系统的内容是否丰富、信息能否得到及时更新[1]。如果我们把互联网比作一张大网的话,那么爬虫技术这网上的蜘蛛,将网络节点比作网页的话,这个“蜘蛛”爬到何处就相当于访问了哪个网页,获得了相应的信息。而后我们可以顺着这些节点继续爬到下一个节点,这样整个网的所有节点,所有信息便会被这个“小蜘蛛”全部爬到。而搜索引擎就是将“小蜘蛛”所爬取的信息一定的策略在互联网中对信息进行处理,并为用户提供服务,从而起到信息导航的目的。我们经常看到不同的网站发布着同样的新闻,很多就是通过网络爬虫的技术从其它的网站爬取信息,然后放在自己的网站发布。同样,这样的爬虫技术也可以用来帮助我们做安全扫描分析等工作,这也是本文研究的重点。

2 概况

2.1 现状

90 年代出现了最早的搜索引擎,也就产生了网络爬虫。此时的爬虫在爬取整个网络的时候采取深度或广度优先的遍历方式。作为搜索引擎的信息资源采集的重要角色,网络爬虫的性能将直接影响整个搜索引擎索引网页的数量、质量和更新周期[2]。于是出现后面的分布式网络爬虫。分布式网络爬虫 (转下页)

化机,也会有空气净化机的网商推荐,把各种空气净化机的价格,参数,测评展示出来,为用户作参考,如果用户需要,可以直接点击链接进行询问或购买。

用户可以使用 qq 或微信等账号登录软件,手机通过蓝牙设备获取的用户自身周围的空气质量情况就可以保存到这个账号中,软件会对这些数据进行分析并打分,分为“综合打分”和“单项打分”,综合代表空气质量综合下来的好坏,单项是分别对“PM2.5”、“PM10”等项目的打分,分数就代表用户所处环境的质量好坏程度。用户可以在这个软件中添加好友,可以是通讯录里的好友,也可以是 qq、微信的好友,这些好友可以将自己的得分晒出来,与大家进行比拼,看谁周围的空气质量比较好。在比拼列表里,会有 top10 的排名,空气质量较好的前三名可自愿分享其周围空气质量比较好的原因,或者采取了哪些措施,把经验分享给大家,让家人远离较差的空气,生活更健康。

3 实现意义

3.1 监测区域空气质量,记录出行者出行后的污染物沾染量,提醒人们何时出行

软件会持续记录相关数据,一天之后会对人们关心的各种信息进行汇总,如当天在什么时候接触的污染源最严重,并以条形图,折线图等方式展现。软件还可以汇集不同地方出行者测出的数据,进而共享到云端,做成一张全国污染地图,更加方便人们的出行。

3.2 增设关于雾霾的防护小知识,便于了解其形成与危害

还可以增加一些关于雾霾的认知与防控的小栏目,这些栏目可以增加一些生活小常识,并实时提醒人们各个时节可以多吃什么食物,比如什么食物清肺润肺,并提供该种食物的营养价值,引导人们关注健康。

3.3 科普微运动,倡导健身生活方式

人们会因雾霾污染严重而选择不出门,面临雾霾几日不除的情况市民无法锻炼,我们可以在软件上及时科普一些微运动,在家中就可以做。对于健身爱好者,软件可以提供一些关于健身的信息,使其在雾霾天也能找到一种有效的锻炼方法,促进人们离健康更近了一步。

结束语

针对当前雾霾带给我们的严重危害,我们应当趋利避害,多方下手,从居民到政府到国家都需要做到及时预防及时监测。雾霾对于我们身体系统、心理健康以及交通安全的影响是不容小觑的,我们应当从根源入手,做到监测与防治相结合。同时,我们也会加强对这款软件的后期开发,完善硬件设施,从提醒到检测到互动到科普全方位进行完善,使这款软件的功能日益人性化。

参考文献

[1]吴兑.大城市区域霾与雾的区别和灰霾天气预警信号发布[J].环境科学与技术,2008(9).
[2]赵俊平,李亚军.如何积极有效地防御雾霾天气[J].科学之友,2009(2).
[3]潘铭.浅谈雾霾对人体健康的影响[J].微量元素与健康研究,2013(6).
[4]高凌云.实时监测全球雾霾流向趋势的软件[J].现代物理知识,2017(1).

作者简介:王艺婷(1997,12-),女,汉族,山东省青岛市人,青岛理工大学商学院财务管理专业在读学生。

可以看成是由多个集中式网络爬虫构成，分布式系统中的每个节点都可以看作一个集中式网络爬虫。分布式网络爬虫大大提高了爬取效率，目前分布式网络爬虫已经有了不少的应用，例如现在著名的 Google 和 Alta Vista 搜索引擎所采用的网络爬虫系统。

由于爬虫的重要性，Twisted 使用 python 语言写了一个广受欢迎的爬虫事件驱动网络框架：scrapy，scrapy 使用的是非堵塞的异步处理方式。scrapy 能够爬取 web 页面，并从页面中提取结构化的数据。它可以用来数据挖掘、监测、和自动化测试。

2.2 语言

本系统用 python 为脚本语言开发，python 脚本语言与其它编程语言相比优势在于它的语法简单、系统库强大、实现功能容易、高效率的高层数据结构、简单的面向对象编程、代码结构清晰易懂[3]。Python 相对于大多数计算机编程语言来说，在运行之前不需要将源码编译为操作系统可以执行的二进制格式(0110 格式的)，从而减少大型项目编译过程非常消耗时间，你可以直接从源代码运行程序。在计算机内部，Python 解释器把源代码转换成称为字节码的中间形式，然后再把它翻译成计算机使用的机器语言并运行。

2.3 聚焦爬虫技术

相比于通用网络爬虫，聚焦网络爬虫的工作流程比较复杂，由于 WEB 网站群结构层次多，目录深度广，数据量很大，单进程的爬虫很难满足快速抓取大量数据的要求，因此它需要实现通过一些网页分析算法过滤掉与搜索主题无关的链接，确保留下来的链接和内容与所要搜索的主题相关度更高，然后按照搜索策略，从相关队列中选择下一个要爬取的内容，并重复以上操作，直到满足用户的检索条件时程序停止。

3 系统设计与试验分析

3.1 聚焦搜索策略

程序主模块的主要功能是 web 爬取。通过用户提供的初始 URL 开始爬取[4]。聚焦爬虫会给它所下载的页面排序，然后根据这个顺序，插入到一个队列中，通过对弹出队列中的第一个页面进行分析和执行，接着下一个页面弹出。这种策略保证爬虫能有限跟踪那些被用户最需要的页面。另一方面线程池决定着整个程序的执行效率，创建太多的线程，有些线程又有可能未被充分的利用，程序将会浪费一定的资源。销毁太多线程，将导致之后浪费时间再次创建它们。创建线程太慢，将会导致长时间的等待，性能变差。销毁线程太慢，将导致其它线程资源饥饿。所以在程序的开发中，线程池相当的重要，应做到合理有效的利用。线程池模块主要是用来创建线程，加载爬虫模块中的爬取 url 任务到任务队列，每个线程从任务队列中获取任务并执行任务。

3.2 爬取范围

在我们能看到的网页中，存在各种各样的信息，但即便是最简单的常规网页，它们也会有对应的 HTML 代码，而我们的爬虫就是来抓取这些个 HTML 源代码的。

当然并不是所有的网页通过爬取返回的都是 HTML 代码，也会存在一些 JSON 字符串 (API 接口通常会是这样)，那么我们就需要对所传回的数据进行分析和整理，此外我们所接收的数据中还存在大量的二进制数据，我们可以通过文件后缀来区分图片，视频，音频。并对应的进行保存。

最后还有一种数据叫做配置文件和 CSS，JavaScript 等，这些也是最普通的文件 (编写网页时常会用到这些文件) 我们的爬虫只要在浏览器中发现他们就会直接访问并爬取。

3.3 数据解析

当我们用“小蜘蛛”将网页数据爬取下来之后，就需要将这些用户无法识别的代码转换成用户所需的文本，图片等信息。那么本文介绍的网络爬虫运用的就是 Xpath，即 XML 路径语言，他是利用 XML 文档中查找信息的语言，同样他也可以用于 HTML 文档搜索。

过程如下：首先导入 lxml 库的 etree 模块，然后声明一段 HTML 文本，并将其初始化，这就构成了一个 XPath 解析对象。

这里存在一个问题，那就是我们通常会用//开头的 XPath 规则来选取符合要求的所有节点，这时筛选出来的节点元素 Element 后所跟的名称就是文件类型。接下来就是进行节点轴选择了，XPath 提供了很多节点轴选择方法，包括兄弟元素，获取子元素等等。通过多次选择我们可以将各种类型的文件分类存储，比如第一次选择，我们调用 ancestor 轴，可以获得相应的祖先节点，如第一个节点 li 的祖先节点就包括 himl,body,div 和 ul。接着我们进行第二次选择，选取 div 的祖先节点，就这样通过一遍遍的选择就能后续所以的同级节点了。这样，用 XPath 的内置函数越多，HTML 信息的提取效率也就越高。

结束语

本文讲述利用 python 编写聚焦式网络爬虫的核心部分以及重点问题，阐述了网络爬虫的现状，利用 python 语言的优势以及聚焦式网络爬虫涉及到的相关算法，论证了聚焦式网络爬虫相比于其他网络爬虫的优势以及可行性，从理论上解释了聚焦网络爬虫的工作原理以及使用方法。[5]

参考文献

[1]winter.中文搜索引擎技术解密：网络蜘蛛[M].北京：人民邮电出版社，2004.

[2]罗刚王振东.自己动手写网络爬虫[M].北京：清华大学出版社，2010，10.

[3]郭涛，黄铭均.社区网络爬虫的设计与实现[J].智能计算机与应用，2012 (4) .

[4]刘晶晶. 面向微博的网络爬虫研究与实现[D].上海：复旦大学，2012.

[5]崔庆才.python3 网络爬虫开发实战上[M].北京：中国工信出版集团，人民邮电出版社，2017.