DOI:10.19695/j.cnki.cn12-1369.2017.05.065

网络爬虫在信息检索中的研究与应用

刘鑫

(神华准能集团培训中心,内蒙古鄂尔多斯 010300)

摘要:本文从网络爬虫的基本概念、网络爬虫的主要作用、网络爬虫的各种类型和网络爬虫的发展方向着手进行研究。各种爬虫的研究对于我们现如今的意义也相当重大,通过网络爬虫爬取的特定信息通过专业分析,可以影响着我们的生活,如经济、健康、工作效率等方面,本文主要运用Python编写网络爬虫,实现网络爬虫的功能。了解网络爬虫如何爬取信息,如何存储所爬取的信息,从而认识其在信息检索中的关键作用。

关键词:网络爬虫;信息检索;Python

中图分类号:TP391.3 **文献标识码**:A **文章编号**:1007-9416(2017)05-0095-03

1 绪论

当今社会,数据显得越来越重要,以往人们也意识到数据的重要性,但是以前的情况面对浩如烟海的数据,人们往往望洋兴叹。因为以以前的数据处理能力,很难对大量的数据信息进行处理分析。随着计算机技术的发展,数据的处理能力得到了极大的提高,尤其是近几年开启的云时代,让人们迎来了大数据时代,人们在处理数据的能力得到提高的时候,生产数据的能力也得到了极大的提升,因此获得数据,处理数据是人们提高对数据的利用的关键。

网络爬虫可以很容易的获取互联网上的信息,是我们获取大量网络上信息的高效工具,现如今有各种各样的网络爬虫在以不同的方式获取网络上的数据,抓取网络上有用的数据,方便人们对数据进行分析和利用。本课题对于网络爬虫进行研究,了解其在信息检索中的应用,并设计简单的网络爬虫,实现其功能。

2 网络爬虫的基本概念

网络爬虫(Web Crawler),又称为网络蜘蛛(Web Spider)或Web信息采集器,是一个自动下载网页的计算机程序或自动化脚本,是搜索引擎的重要组成部分。网络爬虫通常从一个称为种子集的URL集合开始运行,它首先将这些URL全部放入到一个有序的待爬行队列里,按照一定的顺序从中取出URL并下载所指向的页面,分析页面内容,提取新的URL并存入待爬行URL队列中,如此重复上面的过程,直到URL队列为空或满足某个爬行终止条件,从而遍历Web。该过程称为网络爬行(Web Crawling)[1]。

对于网络爬虫,基本的工作流程首先要有一个初始的URL,这个URL可以是一开始自己确定好,也可以是由用户输入获得,然后通过URL获取到网页的信息,接着抓取网页内的相关URL,对于满足条件的信息进行抓取,直到所有的条件满足才结束爬取的过程。当然,这只是对于网络爬虫大概的一个爬取过程,对于不同的网络爬虫来说,爬取的过程是不同的,但是他们都需要有一个URL,然后还有过滤的条件,以及存储所抓取到的信息的过程。

3 网络爬虫的类型

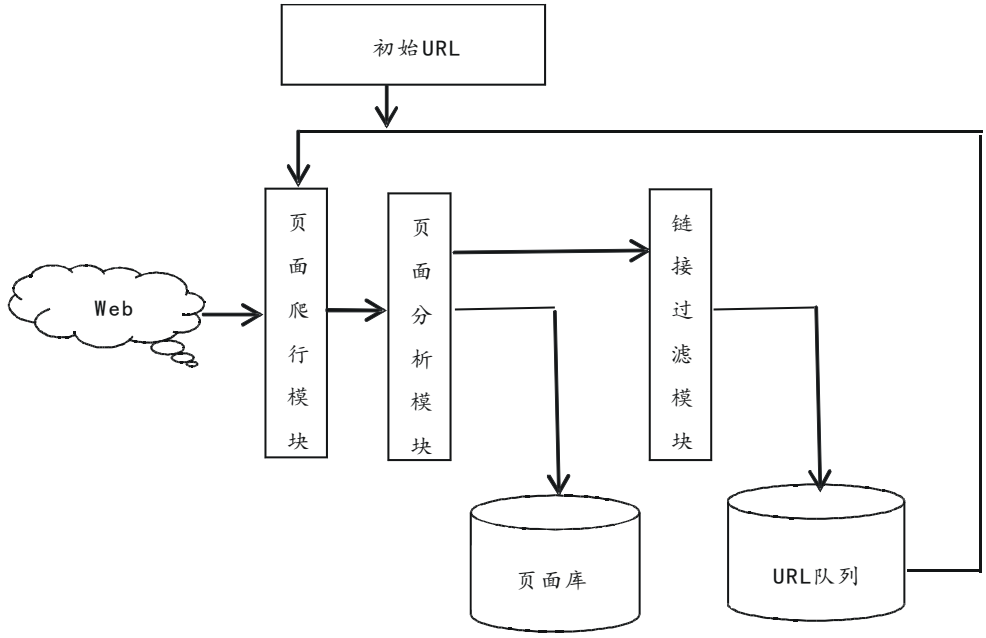

图1 通用网络爬虫体系结构

收稿日期:2017-05-16

作者简介:刘鑫(1984—),男,内蒙古包头人,硕士,工程师,研究方向:网络通信。

网络爬虫根据不同的应用,爬虫系统在许多方面也存在着不同的差异,按照系统结构和实现技术,我们可以将网络爬虫分为以下几类:通用型网络爬虫(General Purpose Web Crawler)、聚焦网络爬虫(Focused Web Crawler)、增量式网络爬虫(Incremental Web Crawler)、深层网络爬虫(Deep Web Crawler)。当然,实际上的网络爬虫不会是单一的技术实现,通常是由多种网络爬虫技术结合而成[2]。

3.1 通用网络爬虫

通用网络爬虫通常用于搜索引擎,它能从一些种子URL爬取大量网站,甚至是整个Web,仅仅受限于时间或者其他方面的限制,它的逻辑相比于其他提取规则的复杂的网络爬虫较为简单,但是其作用不可小觑。它主要用于门户站点搜索引擎和为大型的Web服务提供商采集数据。由于商业原因,这方面性能优秀爬虫的技术细节很少披露,但是此类的网络爬虫爬行的范围和数量巨大,并且其爬行的速度快,存储空间比较大。它们通常采用并行的方式,对爬行页面的顺序要求比较低,但是由于要刷新的页面很多,很长时间页面才能刷新一次。虽然其有一定缺陷,但是通用爬虫适用于搜索引擎,有比较强的应用价值,其结构如图1所示。

通用网络爬虫通常会采取一些爬行策略来提高爬行效率,如:深度优先策略、广度优先策略、最佳优先策略等。

(1)深度优先策略。深度优先策略所采取主要方法是按照由低到高的顺序,它首先从起始网页中的URL选择一个进入,然后对这个网页中的URL进行分析,接着再选择其中的一个URL进入,就像这样不断的层层深入,一个接着一个链接抓取,直到没有链接,不能深入为止。当一个分支爬取完后爬虫会返回上个分支继续爬取未爬取的链接,直到所有的链接遍历完成后,这时的爬行任务才算结束。其实,深度优先策略设计较为简单,但是若其爬取较深的站点时会造成资源的大量浪费,而且随着链接的深入,链接自身的价值往往较低,所有相较于其它两种策略,通常这种策略很少被用到。(2)广度优先策略。广度优先策略是指在爬虫爬取的过程中先完成当前页的所有爬取工作再进入下一层进行爬取,等下一层的爬取工作结束后再逐步深入进行爬取。此策略能控制爬行深度,避免了遇到一个无穷深的分支无法结束爬取浪费资源的情况。广度优先策略通常和网页过滤技术结合使用,先通过广度优先策略抓取网页,然后过滤掉无关的网页。但是此策略也有缺点,当抓取的网页过多时有许多无关的网页也会被下载并且过滤,影响效率,并且此策略要爬取目录较深的网页时,需要耗费大量时间。(3)最佳优先策略。最佳优先策略是先按照一定的网页分析算法进行分析,预选出几个和需求相似度高或者主题相近的URL进行爬取,它只爬取经过网页分析算法认为“有用”的URL,这种算法可以节约大量资源,改善了前两种策略的不足,但是此策略也有所不足,由于网页分析算法不够精确,所以有可能忽略大量相关的网页,所以需要结合具体的应用改善此策略。

3.2 聚焦网络爬虫

传统的网络爬虫通常被设计成尽可能多的覆盖网络,对于需要爬取页面的顺序和爬取网页主题是否相关关注度不是很大。聚焦网络爬虫解决了这一问题,它定向爬取与主题相关的页面,有选择的访问互联网上的网页和相关链接,极大的节约了资源和时间,适用于特定人员对特定特定领域信息获取的需求。

聚焦网络爬虫需要解决几个问题,如:如何对目标网页进行描述,如何对网页和数据进行分析和过滤,还有就是对URL的搜索和排序策略,为了解决以上问题,聚焦网络爬虫有以下几种实现策略:

(1)基于内容评价的爬行策略。它将用户输入的查询词作为主题,将文本相似度的计算方法引入到网络爬虫中,它爬取包含用户输入查询词的页面,但是不能判断所抓取的页面与主题关系的相关程度的高低后来有人利用空间向量模型计算和页面主题的相关程度,改善了这一缺点;(2)基于链接结构评价的爬行策略。此爬行策略访问网页中的链接,它通过HITS方法对网页中的链接进行评估,并按照一定的方法决定链接的访问顺序,还有的一种评估方法是通过PageRank算法,比较PageRank的值对网页中的链接进行排序访问;(3)基于增强学习的爬行策略。此策略是将增强学习引入聚焦爬虫,利用贝叶斯分类器讲网页中的超链接按文本和超链接文本进行分类,计算出链接的重要性,根据这个决定链接的访问顺序;(4)基于语境图的爬行策略。M.Diligenti.[3]等人提出了通过建立语境图(Context Graphs)学习网页之间的相关度,从而训练一个机器学习系统,通过这个机器学习系统可以计算当前页面到相关的网页距离,距离越近的的网页链接优先。

3.3 增量式网络爬虫

增量式网络爬虫(Incremental Web Crawler)是指对已下载网页采取增量式更新,为保证爬取的网页都是尽可能新的网页,所以此爬虫只爬取新产生或者已经发生变化的网页。与周期性爬行和刷新页面的网络爬虫对比,增量式爬虫只在网页新产生或者发生变化的页面才进行爬取,因而减少了数据的下载,节约了时间和空间等各种资源,但是其爬行的算法相较而言更复杂而且实现难度大大增加[4]。

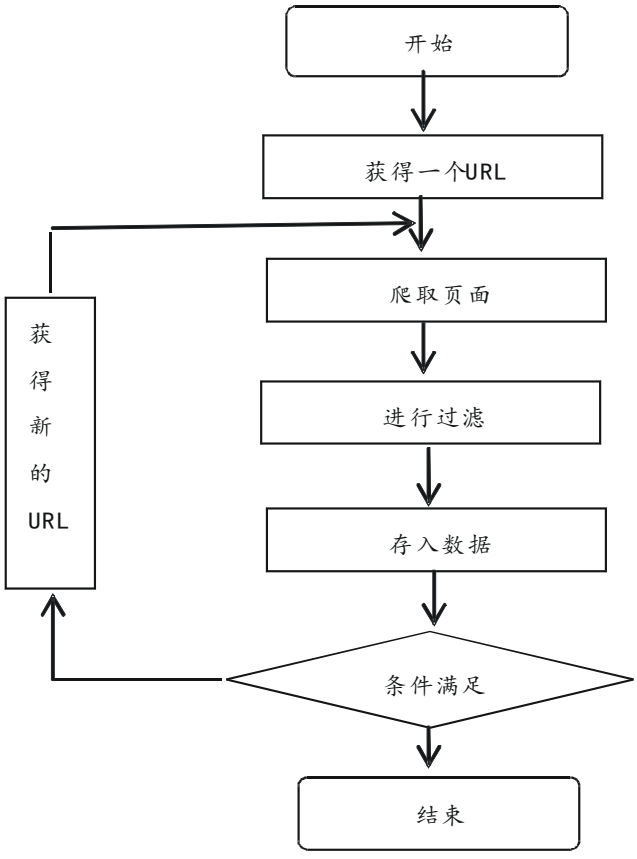

图2 贴吧爬虫流程

增量式爬虫主要是为了保证本地爬取到的网页为最新网页，并提高本地页面的质量，为了达到这两点要求，我们要做到以下几种方法：

(1)统一更新：爬虫需以相同的频率访问所有网页，且不考虑网页的改变频率；(2)个体更新：爬虫需根据个体网页的改变频率来重新访问各页面；(3)基于分类的更新：爬虫根据网页改变频率将网页分为更新较快和较慢的网页子集，然后以不同的访问频率对这两类网页进行访问。

4 系统分析

本课题所做的是一个爬取百度贴吧内容的网络爬虫，主要用urllib2模块编写爬虫，还有其他爬虫用requests模块编写，进行对比，现在主要谈爬取百度贴吧的网络爬虫，通过获得一个初始URL，页面爬行模块在互联网上抓取网页，经过分析模块将抓取的页面放入页面库中，并且过滤页面中的超链接放入URL队列，从而进行下一轮的页面抓取。对于所需要编写的网络爬虫先要得到一个URL，然后通过URL抓取页面，接着对页面内容进行分析，满足过滤条件的(即正则表达式)的存入页面库，不满足的接着爬取，直到相关的URL爬取完毕，如流程图2所示。

对于网络爬虫爬取贴吧的内容需要有爬虫自动抓取，但是初始的URL要由人为定义，并且对于抓取到的内容要进行过滤，获得自己需要的内容，剔除那些不需要的内容，将过滤后所得的需要的内容存储起来。这些只是最基本的功能，如果能够进一步完善人机交互的话，会使得本课题更加完善，如做出一个界面友好的前端等。

由于贴吧的信息容量大小不确定，所以此网络爬虫的工作量也挺难确定，而且还要对爬取的内容进行过滤，取得想要爬取的信息，而且只是在个人的笔记本电脑上运行网络爬虫的程序，这样有可能降低爬行的效率。但是对于本课题来说，我们编写的网络爬虫比较小，在普通的个人笔记本电脑足以运行，完全不影响运行的效率。此外，我们还可以通过改变网络爬虫的爬行策略来提高网络爬虫的运行效率。对于贴吧的网络爬虫来说，通常贴吧的基础架构变动不是很大，所以不用担心因为网站的改变，网络爬虫不能使用。此外这个爬虫经过很多人试验过，功能是能够实现的，还有，这个网络爬虫是由Python语言编写的，Python语言具有强大而且丰富的库，对代码的可用性和可靠性提供了强有力的保障，所以此网络爬虫在可靠性和可用性的需求也是满足的[5]。

对于本课题，主要研究了贴吧爬虫的爬取过程，对于以后可能还要研究爬取其他信息的网络爬虫，此外对于网络爬虫也要求有良好的用户体验，最好能增加一点用户界面的友好性，还有对于不同爬行策略的网络爬虫的性能进行对比，从而编写出目的性更强，性能更强的高效的网络爬虫，此外我们还应当考虑对于网络爬虫爬取到的信息处理问题等。

5 系统设计

5.1 抓取贴吧信息爬虫

网络爬虫主要是爬取互联网上的网页信息，获得需求的目标信息。对于这个贴吧的网络爬虫要包括的模块有页面爬行模块、页面分析模块、页面过滤模块、还有页面数据的保存模块。

5.2 抓取贴吧图片爬虫

对于贴吧的图片爬取，我设计的比较简单，功能模块因为爬取贴吧信息的网络爬虫都有，也包括页面爬行模块、页面分析模块、页面过滤模块、还有页面数据的保存模块

5.3 详细设计

对于贴吧的网络爬虫来说，首先要有一个URL，此次做的网络爬虫初始的URL前半部分是http://tieba.baidu.com/p/，这部分内容针对的是百度贴吧，由于百度贴吧的内容比较多，分类也烦杂，此时我们要求用户输入贴吧的页码，如3138733512，接着网络爬虫开始运行。对于百度贴吧，如果我们只爬取楼主发表的信息所以我们需要把楼主的信息放入类的初始化上，即init方法，除此之外我们还要将贴吧中比较重要的帖子页码这一参数放入此方法中。通过对网页源代码的分析，我们发现百度贴吧的每一层的主要内容都在标签 里面，因此我们采用的正则表达式来实现。查看运行结果我们发现除了我们需要爬取的楼层内容之外，还包含了大量的换行符和图片符，因此我们需要对所抓取的内容进行处理，将这些没用的标签去除掉，从而得到我们真正想要的纯正的信息[6]。

针对此类情况，我们编写一个Tool类，在它的里面定义一个替换各种标签的方法replace，其中也定义了几个正则表达式，re.sub方法对抓取的信息进行匹配替换处理，在代码运行后，我们需要写入需要爬行贴吧的具体URL，然后会让我们选择是否只爬取楼主的发表信息，接着会选择是否写入楼层信息，当选择完毕后，爬虫开始爬取所需要的信息并写入文件，运行完成后会生成一个TXT文件，上面有所需要抓取的信息。

6 结语

对于网络爬虫的发展主要要看信息检索的发展方向，这些可以从国内外的搜索引擎哪里看出一些端倪，伴随着大数据，云计算的浪潮，网络爬虫肯定会得到进一步的发展，对于信息爬取的效率也越来越高，不仅仅是信息爬取的速提高，而且信息爬取的准确性也会提高，而人们通过对于这些爬取下来的信息进行分析，会让这些信息充分发挥其作用。网络爬虫的设计将来会越来越智能化，不仅仅能高效的爬取需要爬取的信息，还能智能化的预测爬取相关需要的信息，如果这一功能将来得到实现，运用到智能机器人的身上，将会使机器人更趋于人类的思考方式。

参考文献

[1]张海藩，袁勤勇，李晔.软件工程导论[M].北京：清华大学出版社，2010.

[2][美]Justin Seitz 著.孙松柏，李聪，润秋译.Python黑帽子—黑客与渗透测试编程之道[M].北京：电子工业出版社，2015.

[3]M.Diligenti.,F.Coetzee, S.Lawtence, et al.FOcused crawling using context graphs[C].In Proceedings of 26th International Con-fer-ence on Very Large Database,Cairo,Egypt.2000.

[4][美]Justin Seitz 著.丁赟卿，译.Python灰帽子－黑客与逆向工程师的Python编程之道[M].北京：电子工业出版社，2011.

[5]Mark Lutz著.邹晓，瞿乔.任发科译.Python编程(上下两册)[M].北京：中国电力出版社，2015.

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