

基于网络爬虫的电力营商环境 大数据信息检索技术研究

文/燕跃豪 宋建辉 鲍 薇 孙晨光 原 征
国网河南省电力公司郑州供电公司

电力部门面对如此海量的信息应快速筛选出有用信息，并根据这些有用信息进行相应决策。论述了网络爬虫的基本原理及构建了网络爬虫系统的几个比较关键的技术，介绍了一种适用于电力部门信息采集的网络爬虫系统，对提高电力部门信息处理能力具有实际意义。

随着信息技术的快速发展，人类已经步入了大数据、云信息时代。如何在海量的信息里面找出有价值的数据像大海捞针一样困难。而网络爬虫技术的出现为解决这一困难创造了新的途径。在我国的电网行业中，同样需要处理海量的信息，需要从万维网上筛选出和电力营商环境相关的信息，并对这些信息进行处理，根据处理结果来简化电网工作人员的工作流程，提高工作效率，时刻把握电网相关的资讯。

网络爬虫的原理

网络爬虫技术是指一种按照一定的规则自动地抓取互联网信息的程序或脚本。在海量网络数据中，网络爬虫技术既能够自动下载网页的程序，又能够从网页中提取所需要的信息。因此在各大搜索引擎服务企业中，实现网络爬虫技术的优化创新成为当务之急。网络爬虫按照系统结构和实现技术，大致可以分为：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫及深层网络爬虫。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

1. 通用网络爬虫

通用网络爬虫俗称全网爬虫，爬取对象由点及面，从一些种子URL遍历至整个Web，主要服务于大型Web服务和门户网站搜索引擎。因为通用网络

爬虫需要爬取大量的网页，所以通用网络爬虫一般会采取一定的爬取策略，比如深度优先策略、广度优先策略。

2. 聚焦网络爬虫

聚焦网络爬虫俗称主题爬虫，需要操作人员提前预设所需要的爬取主题，爬虫将自动地筛选出与主题相关的内容。聚焦网络爬虫不像通用网络爬虫那样全面爬取整个Web上面的内容，它能极大地节省硬件和网络资源，所爬取的网页数量虽然较少，但是都是操作者所需要的内容。他能很好地满足一些特定人群对特定领域信息的需求。

3. 增量式网络爬虫

增量式网络爬虫是一种能不断更新爬取页面信息的爬取方法，它只爬取新产生的网页或信息发生更替的网页，不会重新下载没有发生变化的网页，能有效地节约数据下载量，保证数据的更新，但其爬取算法较为复杂，极难实现。

4. 深层网络爬虫

在互联网中，网页一般分为表层网页和深层网页。表层网页是指用户可通过超链接到达的静态网页构成的Web页面；深层网页大部分内容不可以通过静态链接获取，只有操作者提供一些关键词才能获得的网页。深层网页所能访问的信息容量是表层

网页的几百倍，是互联网上的一种新型信息资源。

在这选取的搜索策略是基于内容评价的爬行策略。这种爬取策略将用户输入的关键词作为主题，将文本相似度的计算方法引入网络爬虫中，其爬取结果包含了输入的关键词。主题网络爬虫工作流程如图1所示。

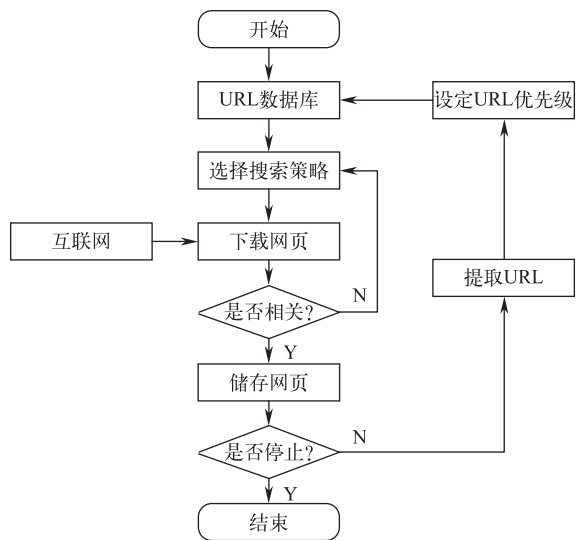


图1 主题网络爬虫工作流程

从图1可知，其基本的工作流程为从URL库中提取一个URL，基于内容评价的爬行策略判断该网页是否和爬取的内容相关，把相关的内容储存起来，不断重复这个过程，直到满足设置的爬取条件时终止爬取。

网页爬虫关键技术

1. 页面去重

在进行爬虫的过程中，往往会出现许多内容重复的页面，假如不加处理的话，这些重复的内容会极大占用计算机内存，限制爬虫速度，在查阅爬取结果时也会带来极大不便，需要去重操作保证爬取结果的唯一性。

2. 去噪

网站运营商为了维持网站的正常运行，往往会在网页上悬挂许多与爬取主题不相关的内容，这些无用信息会影响后期的阅读和存储，所以需要去除这些无用信息，保证爬取内容简洁。这些无用信息会带有一些相同的标志符号，比如“广告”等词语。去噪操作将在爬取过程中跳过异常特征的数

据，保留网页主干内容。

3. 统一格式

对于爬取到的网页一般都是各式各样的，不利于后期的查阅和展示。需要把网页的主干内容爬取下来，再用相同的格式进行重新排版。爬取的主干内容主要有标题、摘要及正文等。

4. 代理

爬虫过程是一个不断重复访问服务器的过程，而每一个网页一般都拥有自己的反爬机制，如果被网页监测到存在异常访问，就会中断操作者的访问，需要给爬虫披上一件“外套”，躲过网页的反爬机制。在进行爬虫时，由代理服务器将响应数据转发给爬虫服务器，这样所爬取的服务器就无法找到爬虫服务器的地址。爬虫之所以可以重复不断地访问网页是因为它能不断地切换代理服务器。因此，代理对于爬虫是必不可少的。

网络爬虫的电网信息检索设计

适用于电力部信息采集的网络爬虫系统结构如图2所示，该网络爬虫主要分为三大模块，分别为信息采集模块、网页处理及下载模块和数据展示模块。

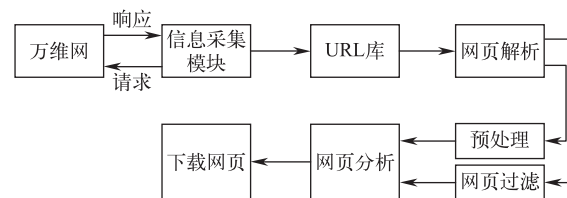


图2 网络爬虫系统结构图

1. 信息采集模块

信息采集模块的原理类似于搜索引擎。输入搜索的关键词，就会向万维网发送HTTP请求，检索包含这些关键字的信息，并下载包含这些信息的网页URL和标题，在这个过程中可以通过设置让其只检索在某一个时间段更新的网页，可以提高检索的效率并且保证检索信息时效性，当满足要求的网页都检索完成后，再将下载的结果储存在一个Excel里，采集到的结果如图3所示。

2. 网页处理及下载

信息采集模块得到的结果仅包含所需要采集网页的标题和链接，为了方便查看数据，还需要对页面进

标题	标题链接	
国瑞科技股东苏州开瑞减持297万股 套现6198.39万	https://baijiahao.baidu.com/s?id=1702912200140668309&wfr=spider&for=pc	
宁夏石嘴山红果子实现智能分布式FA功能百分百全覆盖	https://baijiahao.baidu.com/s?id=1702692764706037847&wfr=spider&for=pc	
湖南省首次“SG+配网分布式故障自愈实验室测试成功	https://baijiahao.baidu.com/s?id=1702622273111365463&wfr=spider&for=pc	
山东科汇电力自动化股份有限公司 首次公开发行股票科创板上市公告	https://baijiahao.baidu.com/s?id=1702564769455066458&wfr=spider&for=pc	
杭州南瑞仪保科技有限公司贵州黎平机场电力自动化改造项目	https://baijiahao.baidu.com/s?id=1702147016570525858&wfr=spider&for=pc	
2022年第十八届巴基斯坦国际工业自动化及机器人展览会ITIF	https://baijiahao.baidu.com/s?id=1702075335805324397&wfr=spider&for=pc	
打造能源与自动化领域开放交流平台 施耐德电气上线开发者社区	https://baijiahao.baidu.com/s?id=1701966586347495220&wfr=spider&for=pc	
高新区 竞逐“新”风口 释放创新动能	https://baijiahao.baidu.com/s?id=1701961552467051210&wfr=spider&for=pc	
清华“学霸”做一线电力工人有多牛?这位80后给出答案	https://baijiahao.baidu.com/s?id=1701945901513430047&wfr=spider&for=pc	
...深圳供电局投运国内首个超大城市电网主配一体化调度自动化系统	https://baijiahao.baidu.com/s?id=1701908153187947941&wfr=spider&for=pc	
从“清华博士”到一线电力工人,他说“这里有我成长的土壤”	https://baijiahao.baidu.com/s?id=1701875350802127920&wfr=spider&for=pc	
岳阳首条全自动化配电架空线路正式投运	https://baijiahao.baidu.com/s?id=1701869955859546951&wfr=spider&for=pc	
施耐德电气:“三步走”实现配电数字化,制胜可持续未来	https://baijiahao.baidu.com/s?id=1701869249722288702&wfr=spider&for=pc	
电力技术类有哪些专业?毕业后的就业方向是什么?	https://baijiahao.baidu.com/s?id=1701606557174791490&wfr=spider&for=pc	
金冠股份(300510.SZ)收到国家电网相关项目的6份中标通知书	https://baijiahao.baidu.com/s?id=1701449755993460139&wfr=spider&for=pc	
金冠电气以技术支撑高毛利 深耕行业获可观市场份额	https://baijiahao.baidu.com/s?id=1701449520636991041&wfr=spider&for=pc	
【岳阳好网民】崔彦伟——点燃国网科技的“星星之火”	https://baijiahao.baidu.com/s?id=1701429038093982715&wfr=spider&for=pc	
以物联网为核心的智能配电管理	https://baijiahao.baidu.com/s?id=1701354892080947228&wfr=spider&for=pc	
争当增量配电业务改革先行军	https://baijiahao.baidu.com/s?id=1701348577328952277&wfr=spider&for=pc	
许继电气:配网系列产品创新成果顺利通过中国电机工程学会权威鉴定	https://baijiahao.baidu.com/s?id=1701063901304129205&wfr=spider&for=pc	
配网自动化无线传输系统方案	https://baijiahao.baidu.com/s?id=1700993916524931192&wfr=spider&for=pc	
电网数字化坡长雪厚 科汇股份开启申购	https://baijiahao.baidu.com/s?id=1700988103194454542&wfr=spider&for=pc	
东莞供电局:数字赋能推动配网智能化,运维助力提升供电可靠性	https://baijiahao.baidu.com/s?id=1700970819020523532&wfr=spider&for=pc	
国网湘潭供电公司高考保电早行动	https://baijiahao.baidu.com/s?id=1700969132674876744&wfr=spider&for=pc	
电力标准配电自动化终端运维技术规范	https://baijiahao.baidu.com/s?id=1700865504494865486&wfr=spider&for=pc	
科汇股份25日起招股 深耕电气自动化领域	https://baijiahao.baidu.com/s?id=1700731580687134436&wfr=spider&for=pc	

图3 信息采集URL结果

行处理和下载。该模块包含了四个小的模块，分别为网页的解析、网页的过滤、预处理以及网页分析。

1) 网页解析模块。提取收集到的HTML网页中的重要信息链接和文本，利用获得的内容信息为后续网页的主题相关度计算铺平道路。

2) 网页过滤模块。该模块用于筛选与主题相关的URL，并通过过滤与主题相关的页面来确保主题爬虫系统的准确性。

3) 预处理模块。该模块对网页分析模块获取的网页内容信息进行处理，将文本内容转化为数学模型，通过分词、停词及干词等预处理，使计算机能够识别，从而为后续主题爬虫中网页分析模块的主题相似度计算做准备。

4) 网页分析模块。该模块主要进行主题相关度判断，并通过搜索策略优先访问与主题相关的URL。

根据以上几个步骤的处理之后得到的网页就是一个HTML文件，在不联网的情况下依然可以查看里面的内容，而且这些网页里面包含的异常信息也已经被删除掉了，里面只包含所需要的信息。

3. 数据储存和展示

为了方便操作人员的查看和修改，需要把下载下来的HTML文件装换为PDF文件，文件的转换可以使用Adobe Acrobat DC软件完成。另外需要保存网络爬虫爬取得到的数据。爬虫抓取后的数据一般

选择两种存储方式：本地保存CSV、Excel格式或者直接存储到数据库。以“电力营商环境”为关键词所采集到的最终成果以PDF格式储存，方便长时间存储和查阅。

结束语

大数据时代的到来给信息检索系统带来了变化。将网络爬虫技术与电网商业环境相结合，可以从根本上提高整体运行结构的质量和效率，保证管理措施和管理效果更符合市场需求，顺利优化仿真系统的有效性，从而保证企业的工作效率，从根本上实现技术工作按规范流程的优化运行，为电力系统智能电网的建设和电力部门工作人员的智能办公奠定坚实的基础。EA

参考文献

[1] 顾勤. 网络爬虫技术原理及其应用研究[J]. 信息与电脑(理论版), 2021, 33(4): 174-176.
[2] 潘晓英, 陈柳, 余慧敏, 等. 主题爬虫技术研究综述[J]. 计算机应用研究, 2020, 37(4): 961-965, 972.
[3] 贺洲强, 李应时, 王乃宁, 等. 考虑多源数据交互的配电网数字化“两票”研究与应用[J]. 电气应用, 2021, 40(8): 56-60.