西安航空学院

本科毕业设计（论文）开题报告

题　　目： 网络爬虫信息采集系统的设计与实现

学生姓名： 黄莉萍

学院（部）： 计算机学院

专业班级： 计算机科学 1912

指导教师： 张晓丽

完成时间： 2019　年 1 月　14 日

要　求

1、开题报告是毕业设计（论文）的总体构想，由学生在毕业设计（论文）工作前期独立完成。

2、开题报告正文用 A4 纸打印，各级标题用 4 号宋体字加黑，正文用小 4 号宋体字，20 磅行距。

3、参考文献不少于 10 篇（不包括辞典、手册），著录格式应符合 GB7714-87《文后参考文献著录规则》要求。

4、年月日等的填写，用阿拉伯数字书写。要符合《关于出版物上数字用法的试行规定》，如“2005 年 2 月 26 日”。

5、所有签名必须手写，不得打印。

1 研究目的及意义

随着互联网技术的飞速发展，当今社会已经成为了一个被信息包裹的社会。互联网上分布着各种各样类型或形式的信息资源，其中包括图文、视频、软件等。搜索引擎是信息检索的一种工具，已经成为了人们在生活和日常工作中获取信息的主要方式，为人们提供了巨大的便利，但是这种方式也存在一定的局限性。传统的搜索引擎存在着网页索引规模大、更新速度慢以及查询结果精度低等缺点，而且由于互联网资源太过丰富且信息内容太杂乱，因此如何从这浩瀚的信息库中快速、准确、安全地获取到人们感兴趣的信息目前已经成为了亟待解决的问题[1]。

目前网络爬虫是公认的完成网络信息自动或半自动采集任务最为有效的工具。网络爬虫的本质是一种可以分析和追踪网络超链接结构，并按照特定的策略持续进行资源发掘和收集的功能模块。网络爬虫技术是随着搜索引擎发展而伴随产生并普及的一种通用的信息采集技术，其最为成功且广泛的应用就是作为搜索引擎网络信息的前沿，负责完成网页信息的采集任务，为搜索引擎提供检索信息的数据来源。所以，可以说网络爬虫是搜索引擎信息提供者，其信息采集的性能和策略将直接影响到搜索引擎提供的网页质量以及信息更新的时效性[2]。

因此，本课题研究的网络爬虫信息采集系统不仅可以从庞大的信息库中提取用户所需的信息，并且对所提取出的信息进行必要的处理，让信息资源更具有准确性和安全性。而传统的搜索方式的信息采集结果包含了大量可能相似度并不高或者说用处并不是很大的网页信息，用户还得在其中去人工筛选所需信息，这样效率太低。为了解决这个问题，网络爬虫系统应运而生，它可针对性的根据需要精确的采集所需信息，且网络爬虫应用范围很广，所以对网络爬虫信息采集系统的深入研究具有很深远的意义。

2 国内外研究现状

目前国内外对于网络信息采集处理的研究多集中于网络信息采集和网页信息抽取两项关键技术上。

网络信息采集方面，网络爬虫是研究的重点。近年来，由于人们对网络信息的采集需求向具体化发展，传统的通用爬虫已经不能满足人们信息获取的需要，在传统的采集技术基础上，涌现出许多新型的各具特色的信息采集技术。主要是将网络信息采集技术的发展方向分为了以下几种：基于整个 Web 的信息采集方法（Scalable Web Crawling），基于用户个性化的信息采集方法（Customized Web Crawling），基于主题的信息采集方法（Focused Web Crawling），增量式信息采集方法（Incremental Web Crawling），基于 Agent 的信息采集方法（Agent Based Web Crawling）[3]，迁移的信息采集方法（Relocatable Web Crawling），基于元搜索的信

息采集方法(Metasearch Web Crawling)[4]。而对于面向主体的聚焦爬虫的研究成为这方面研究的热点。

国外的研究起步较早，已经形成了一些理论，也产生了一些成果：有学者使用基于文本图的聚焦爬虫实现面向主题的信息抓取，也有学者建立了聚焦爬虫的概率模型,还有学者通过“相关反馈”加速聚焦爬虫效率，这是在聚焦爬虫效率方面进行的有益探索。国内关于聚焦爬虫的研究起步较晚，但是近两年随着网络搜索的发展和国外这一领域研究的兴起，国内学者也对聚焦爬虫的研究产生兴趣，进行基于 Web 结构特征挖掘的网页类型自动识别方法的研究[5]，自动化的面向领域/主题的网页搜集系统就是聚焦爬虫的一种应用。

网页信息抽取方面,学术界关于网页信息抽取技术的研究也已经取得了一些成果。流行的网页信息抽取技术包括：基于智能模板的 BBS 网页文本提取方法[5]基于本体论的信息抽取方法[6]，基于文本标签属性的网页信息抽取方法研究[7]等等，以上方法在信息抽取上都取得了成功，但大多基于复杂的数学模型,实施较困难。工程上，已有的网络信息采集系统中，对于网页信息的抽取多基于网页结构的分析，利用模板进行网页信息的抽取,这方面的研究包括：模板化网页主题信息提取的研究，以及针对模板生成网页的自动信息提取的研究。

虽然目前对于互联网信息采集系统的研究方法很多,但是每种研究方法都或多或少的存在着很多不足之处。比如，越来越庞大的信息量，用户对于采集速度的要求越来越高，用户对于信息准确性的要求越来越大，以及对数据安全性越来越重视等。所以为了满足人们对于互联网信息采集系统提出的更新更高的要求，在网络信息采集系统的设计中可以结合爬虫理论进行研究与创新。

3 本课题要研究或解决的问题和拟采用的研究手段（途径）

3.1 研究或解决的问题

根据对本课题要实现的系统进行需求分析,本课题主要研究的是针对网络大量数据资源，开发一套爬虫系统，目的是用于采集图文、视频、软件等类型的信息资源，并且能对采集的结果进行保存以及对数据进行分析统计，研究或解决的问题主要包括：

（1）网页的获取

获取网页就是给一个网址发送请求，该网址会返回整个网页的数据。类似于在浏览器中键入网址，并按回车键，然后可以看到网站的整个页面。

在进行爬虫抓取页面操作时，首先要保证该页面是可被抓取的。对于一些网页访问用户只有在处于登录状态时，才能向服务器发出资源获得需求的网页，对其进行抓取时需要首先建立一个模拟目标网页登录方式的站点,这一站点能够有

效避开被抓取网页对于登录状态的有关要求及检测[8]。

（2）页面的解析

解析网页就是从整个网页的数据中提取想要的数据。类似于在浏览器中看到网站的整个页面，但是你想找的是产品的价格，价格就是你想要的数据。

Web 页面通常由 HTML（Hypertext Markup Language，超文本标记语言）形成页面框架，再融入其他技术丰富显示功能。HTML 作为标记语言，本身就是一组规范和标准，它就是利用标记符号来标记 Web 页面中的各部分内容和显示方式，浏览器也正是通过解析 HTML 的标记来呈现内容[9]。

Web 页面的不同内容通过不同类型的 HTML 标签及其属性呈现，例如文本、图像和超链接等。通过分析 Web 页面中的 HTML 标签，能检索和定位我们感兴趣的内容。HTML 标签主要有结构布局和显示特性两种功能。

在处理页面时，通常采用的是 re 正则表达式[10]，但这一表达式在 HTML 源代码信息量较多的情况下应用则相对困难，因而可以采用其他方法，比如，BeautifulSoup。

（3）链接去重

爬虫的爬行过程会访问大量的 URL 链接，也会有重复的 URL 出现。所以需要对 URL 的重复性进行检查[11]。

建立一个 URL 存储库,在搜索引擎中建立 URL 检测机制，如果 URL 被爬取过就记录下来，在爬取新 URL 之前先 URL 库中的资源进行对比，如果没有该记录，则正常解析爬取资源，如果有则忽略该 URL。

（4）数据的存储

使用 Mysql 数据库，保存采集处理之后的数据。熟悉数据库的连接及其基本的增删改查的使用。

（5）数据的分析统计

采集到所需数据之后，将其保存在数据库，通过获取数据库中保存的数据长度来分析统计各类数据的数量以及采集所耗时间。

3.2 拟采用的研究手段

本课题拟采用 c#或 Python 开发语言编程去实现各种信息资源的采集以及数据的处理，采用 Mysql 数据库去保存数据。开发过程中拟采用的研究手段是：

（1）查阅相关文献资料，学习设计思想并对需设计实现的功能进行分析；

（2）构建该系统的大致框架，搭建开发环境，完成总体设计；

（3）分别实现新闻类、视频类、图片、软件类网站的信息爬取；

（4）对爬取的各类网站信息数据进行分析处理；

（5）进行可视化界面的设计与实现；

（6）代码基本编写完成后，进行调试；

（7）测试各个模块功能使用情况，修改相关问题直到系统稳定。

4 工作进度安排

	设计（论文）各阶段名称	起 止 日 期
1	查阅相关资料，完成开题报告	2018 年 12 月 24 日~2019 年 1 月 10 日
2	分析系统功能需求，完成系统总体设计	2019 年 1 月 11 日~2019 年 2 月 21 日
3	设计系统库，编码实现系统功能	2019 年 2 月 22 日~2019 年 4 月 30 日
4	编码，测试，撰写论文	2019 年 5 月 1 日~2019 年 5 月 20 日
5	测试，完善系统功能，完成论文，答辩	2019 年 5 月 21 日~2019 年 6 月 13 日

5 参考文献

[1] 王子豪. 基于网络爬虫的信息采集技术研究[D]. 西北师范大学, 2018.

[2] 孙骏雄. 基于网络爬虫的网站信息采集技术研究[D]. 大连海事大学, 2014.

[3] 赵彦松. 基于网络爬虫的数据采集系统设计与实现[D]. 东北大学, 2015.

[4] 张婧, 刘彦君, 范漪萍, 贾明慧. 国内网络信息采集研究现状述评[J]. 科技管理研究, 2017,37(09):260-266.

[5] 文友栃. 网页分类与信息采集方法研究[D]. 电子科技大学, 2017.

[6] 刘丽娟, 张胤, 杨一. 基于本体思想的网页信息抽取方法[J]. 计算机与现代化, 2015(09):90-94.

[7] 沈娜. 基于文本标签属性的网页信息抽取方法研究[J]. 武汉职业技术学院学报, 2016,15(01):62-65+73.

[8] 魏冬梅, 何忠秀, 唐建梅. 基于 Python 的 Web 信息获取方法研究[J]. 软件导刊, 2018,17(01):41-43.

[9] Yuan Xiaohong,Zhou Sisi.Research and implementation of the technology supporting MicroBlog data collection based on web crawler[P],2012.

[10] William A.Barnett,Mingzhi Hu,Xue Wang.Does the utilization of information communication technology promote entrepreneurship: Evidence from rural China[J]. Technological Forecasting & Social Change,2019,141.

[11] 成功, 李小正, 赵全军. 一种网络爬虫系统中 URL 去重方法的研究[J]. 中国新技术新产品, 2014(12):23.

指导教师意见：

该论文选题较好，具有一定的研究意义。前期准备充分，查阅了大量相关信息系统相关文献资料，对国内外现状进行了分析研究，研究内容充实，方法合理，重点明确，进度安排合理，符合论文开题报告要求。

同意开题。

指导教师签名：张晓丽

2019 年 1 月 16 日

教研室意见：

选题符合要求，研究思路明确，方案切实可行，同意开题。

主任签字：王宇辛

2019 年 1 月 17 日