

甘肃省大学生创新创业训练计划 项目申报表 (创新训练项目)

推 荐 学 校 :	西北师范大学
项 目 名 称 :	基于 Scrapy 的网络爬虫软件的设计 与实现
所属一级学科名称:	计算机科学与技术
项 目 负 责 人 :	李娜
联 系 电 话 :	18809497802
指 导 教 师 :	马慧芳
联 系 电 话 :	18993133843
申 报 日 期 :	2018 年 4 月 25 日

甘肃省教育厅 制
二〇一八年四月

项目名称		基于 Scrapy 的网络爬虫软件的设计与实现					
项目所属一级学科		计算机科学与技术					
项目实施时间		起始时间：2017 年 12 月			完成时间：2018 年 10 月		
项目简介 (100 字以内)	随着互联网技术的发展，互联网中的信息呈现指数增长的趋势。这些信息广泛影响着人们的工作、学习、科研、社交以及生活。 为此，本项目致力于实现一种集成了多种策略的数据爬虫软件。用于各类特定数据采集任务。						
申请人或申请团队		姓名	年级	学号	所在院系/专业	联系电话	E-mail
	主持人	李娜	2016 级	201671020112	软件工程	18809497802	1347874109@qq.com
	成员	孙玉卓	2016 级	201671020223	计算机科学与技术	17361631173	228682774@qq.com
		周侗	2017 级	2017211643	计算机科学与技术	15693472627	1083993141@qq.com
		伊力亚·木拉提	2015 级	201571030330	计算机科学与技术	18794810040	ilyar1015@hotmail.com
指导教师	第一指导教师	姓名	马慧芳		单位	西北师范大学计算机科学与工程学院	
		年龄	37		专业技术职务	副教授	
	主要成果		主持在研国家自然科学基金 2 项、参与国家自然科学基金 3 项。在《Information sciences》、《Neurocomputing》、《Knowledge and Information Systems》、《软件学报》、《电子学报》、IJCNN、PAKDD、PRICAI 等国内外重要期刊和国际知名会议上发表论文 40 余篇，其中 SCI 收录 4 篇，EI 收录 18 篇。担任《Artificial Intelligence Review》、《IEEE Transactions on Cybernetics》、《IEEE Transactions on Systems、Man and Cybernetics》等期刊的审稿人。主持的主要科研项目有“基于图聚集技术的微博用户重叠社区发现方法研究”，国家自然科学基金项目，编号：61762078，2018-2021；“基于主题建模的微博语义理解与热点话题识别研究”，国家自然科学基金项目，编号：61363058，2014-2017 等。				

一、申请理由

本人所学专业是软件工程，有一定的专业基础，并且在 2017 年加入科研创新团队，有一定的基础知识的积累。

本项目组的人员包括不同年级结构层次合理，每个人都有相应的优点，在探讨中能够从不同的角度思考，提供不同的解决问题的思路。更重要的是，团队成员都具有良好的知识搜集以及学习能力，也做了一定的前期的工作。对爬虫软件设计实现的 python 语言的学习，基础简单的爬虫学习。相信前期的基础工作会给本团队的研究带来一些帮助。

二、项目方案

具体内容包括：

1、项目研究背景

网络信息资源的迅猛增长使得传统搜索引擎已经无法满足人们对有用信息获取的要求,作为搜索引擎的基础和重要组成部分,网络爬虫的作用显得尤为重要。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)。然而随着网络资源的爆炸式膨胀和信息量的飞速增长, 通用的爬虫技术在索引规模、更新速度和个性化等很多方面都已不能满足用户的需求。

因特网的迅速发展对万维网信息的查找与发现提出了巨大的挑战。面对这些挑战, 适应特定主题和个性化搜索的主题网络爬虫应运而生。基于主题网络爬虫的搜索引擎已经成为当前搜索引擎和信息挖掘中的一个研究热点和难点。

传统网络爬虫的目标是尽可能多地采集信息页面,而缺少考虑采集页面的准确性。主题网络爬虫则可以对整个 web 页面按主题采集,提高所关注主题的采集覆盖率。对于传统的网络爬虫来说,刷新一遍需要数周到一个月的时间,使得页面的失效率非常大。一个好的缓解办法就是采用主题网络爬虫,通过减小采集页面的数量来减小刷新时间,进而减小已采集页面的失效率。传统的网络爬虫要消耗很多系统和网络资源,而它们中大部分利用率很低,主题网络爬虫有效地提高了采集到页面的利用率。

为提高爬行效率, 爬虫需要在单位时间内尽可能多的获取高质量页面, 为了更高效率的进行数据的下载获取, 我们对主题网络爬虫技术的策略进行优化研究。

2、项目研究目标及主要内容

(1) 研究目标:

根据一定的网页分析算法过滤与主题无关的链接,保留主题相关的链接。存储所有被网络爬虫抓取的网页,进行一定的分析、过滤,并建立索引,这一过程所得到的分析结果还可能对后续的抓取过程进行反馈和指导,并且对策略进行优化,提高爬虫的爬行策略,使爬虫能够在单位时间内尽可能多的获取高质量页面。

(2) 研究内容:

网络爬虫按照系统结构和实现技术,大致可以分为以下几种类型:通用网络爬虫 (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)。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

在网络爬虫抓取 Web 信息的过程中,无论是宽度优先策略还是广度优先策略,都可能产生重复抓取的现象,也难以处理好 Web 页的优先级顺序和对相同 Web 站点的不定期抓取等问题,因此,网络爬虫的设计还要进行适当、合理的改进,提升并优化网络爬虫的抓取策略。网络爬虫被定义为一个程序或软件,它通过一个系统的、自动化的方式遍历 Web 和下载 Web 文档。基于知识的类型,网络爬虫通常分为三种爬行技术:一般用途爬行、聚焦爬行和分布式爬行。

(3) 解决的关键问题:

- 如何描述或定义感兴趣的主题
- 怎样决定待爬行 URL 的访问次序
- 如何判断一个网页是否与主题相关
- 怎样提高主题网络爬虫的覆盖度

3、项目创新特色概述

使用主题网络爬虫,进行算法优化。主题网络爬虫是一种下载特定主题网页的程序。利用在采集页面过程获得的特定信息,主题网络爬虫抓取的页面都是与主题相关的。这样可以更加简便快速的为用户抓取相关数据,节省时间。

4、项目研究技术路线

- 基于 python, java, c++等语言设计合理的爬虫软件,内嵌爬虫算法,用以实现爬取数据的地址定位,采集和存储等功能。
- 通过 Spider 进行主爬虫文件的编写,利用 Scrapy 根据网站的自身特性构建 URL,将抓取下来的网页信息进行信息提取,并将有用信息存储到 mongo 数据库,监测数据库插入情况。
- 基于 mongodb 和 bootstrap 设计实现数据可视化显示。
- 通过 Scrapy 实现基于 python 的网页爬取框架,Bootstrap 实现 HTML 的前端展示框架。

5、研究进度安排

(文献查阅)：2017 年 11 月至 2018 年 01 月
(方案设计)：2018 年 01 月至 2018 年 03 月
(实验研究)：2018 年 03 月至 2018 年 04 月
(数据处理)：2018 年 04 月至 2018 年 07 月
(研制开发)：2018 年 04 月至 2018 年 07 月
(撰写论文或研究报告)：2018 年 07 月至 2018 年 08 月
(结题和答辩)：2018 年 07 月至 2018 年 08 月
(项目鉴定)：2018 年 08 月至 2018 年 09 月
(成果推广或论文发表)：2018 年 09 月至 2018 年 10 月

6、项目组成员分工

李娜：资料搜集、python 编程语言学习研究；
孙玉卓：Bootstrap (HTML 的前端展示框架) 学习研究；
周侗：python 实现爬虫软件；
伊力亚·木拉提：Laravel (PHP web 的编辑框架)、python 实现爬虫软件；

三、学校提供条件

西北师范大学计算机科学与工程学院始终重视实验室建设工作，为教师、研究生和优秀本科生提供良好的科研环境，创造良好的研究氛围。依托这些实验室，教师和学生可以开展各类科研项目，创造科研成果，使学生有机会和条件在具体的科研项目中提高能力，培养科研素质。学院现有实验室总使用面积 4000 多平方米，设备数量达 3000 多台套，其中服务器和微型计算机 2000 余台，设备总值 1800 多万元。

本次项目的申报来自于学校创新创业学院以及计算机学院领导大力推行的“本科生科研能力提升计划”，共获得学校拨款 20 万元；同时，项目的开展依托于优秀的老师细心指导，及物联网实验室，来进行传感器等物联网工程关键技术研究、相关应用系统开发，为项目的开展提供了有力的设备基础及技术保障。

四、预期成果

基于 scrapy 的爬虫软件的设计实现

五、经费预算					
总经费（元）	6000	财政拨款（元）	5000	学校拨款（元）	1000
注：总经费、财政拨款、学校拨款由学校按照有关规定核定数目进行填写					
具体包括： 1、资料购置、打印、复印、印刷等费用：500 元 2、学生撰写与项目有关的论文版面费、申请专利费等：300 元 3、实验费用及其他：200 元。					
六、导师推荐意见 本项目针对数据爬虫软件的设计实现，采用了主题优化策略，尽可能对算法进行研究改进。研究思路清晰可行，团队成员包括不同年级结构层次合理，每个人都有相应的优点，且成员已开始进行知识储备学习，已完成一定的前期工作，同意推荐，建议优先资助。 签名：马慧芳 2018 年 4 月 26 日					
七、院系推荐意见  院系负责人签名：  学院盖章： 2018 年 4 月 27 日					
八、学校推荐意见：  学校负责人签名：  学校公章 2018 年 4 月 28 日					