科研课题清单

序号	名称	链接
1	软件测试工具在超星教学系统改进中的应用研究	点击查阅
2	Python 网络爬虫技术的研究与探索	点击查阅
3	基于深度学习的网络爬虫算法研究与优化	点击查阅

广 东 省 教 育 厅

粤教科函〔2021〕7号

广东省教育厅关于公布2021年度普通高校认定类科研项目立项名单的通知

各有关高校:

为深入实施创新驱动发展战略，落实《广东省教育厅 广东省科学技术厅关于印发科教融合协同推进高校科技创新能力提升工作计划的通知》（粤教科函〔2019〕57号），省教育厅组织开展了2021年度科研项目认定工作。经学校推荐、省教育厅组织形式审查，现将批准立项的2021年度高校认定类科研项目立项名单（见附件）下达各高校。

请各高校按照国家和省相关科研平台项目管理办法，统筹安排项目资金，加强资金管理，督促项目承担人按照项目申请书开展建设工作，协助解决项目实施过程中遇到的困难和问题，确保研究项目如期完成目标任务。

附件: 1.2021年度广东省普通高校特色创新类项目立项名单

2.2021年度广东省普通高校青年创新人才类项目立项

名单

广东省教育厅

2021年8月25日

（联系人及电话：曾俊伟，020-37627742）

公开方式：主动公开

校对人：曾俊伟

附件1

2021年度广东省普通高校特色创新类项目立项名单

1. 自然科学类				
序号	项目编号	项目名称	负责人姓名	所属学校
1	2021KTSCX001	音圈电机与偏磁电机（本体及驱动）设计与开发	卢少锋	华南理工大学
2	2021KTSCX002	老年人防跌倒外骨骼助行产品系统设计研究	熊志勇	华南理工大学
3	2021KTSCX003	新型高效呈味肽制备关键技术研究	崔春	华南理工大学
4	2021KTSCX004	“双碳”目标下基于计算性设计思维的低碳绿色校园规划智能优化研究	刘骁	华南理工大学
5	2021KTSCX005	多品种产品混流生产过程动态模式表征及智能调控方法	王世勇	华南理工大学
6	2021KTSCX006	基于注意力机制的安全性图像识别模型研究与应用	李海良	暨南大学
7	2021KTSCX007	中药来源的新型HDC抑制剂的发现与抗骨质疏松作用机制研究	邱佐成	暨南大学
8	2021KTSCX008	应用新型蓝莓综合开发技术推动乡村振兴	蒋鑫炜	暨南大学
9	2021KTSCX009	富硒富岩藻黄素微藻用于类风湿关节炎治疗及其作用机制探究	汪翔	暨南大学
10	2021KTSCX010	鸡柔嫩艾美耳球虫MIC3基因重组株构建及生物学特性研究	林瑞庆	华南农业大学
11	2021KTSCX011	生物质化学链气化中铁基载氧体的失活机理	胡志锋	华南农业大学
12	2021KTSCX012	Nrf2/GPX4介导的铁死亡在ATO致肉鸡肝损伤中的作用机制研究	胡莲美	华南农业大学
13	2021KTSCX013	木麻黄青枯病菌关键致病基因鉴定和功能研究	周筱帆	华南农业大学

序号	项目编号	项目名称	负责人姓名	所属学校
334	2021KTSCX334	基于大数据的数字教育资源促进乡村教学质量提升的策略研究——以云浮市为例	谭玉玲	罗定职业技术学院
335	2021KTSCX335	基于缓释功能设计的半互穿网络有机-无机纳米复合微凝胶的制备	练翠霞	顺德职业技术学院
336	2021KTSCX336	涂料用多机制耦合高效阻燃体系的制备及其应用性能研究	姜佳丽	顺德职业技术学院
337	2021KTSCX337	相变蓄冷材料的研制及其主-被动耦合系统的建筑节能应用研究	孙婉纯	顺德职业技术学院
338	2021KTSCX338	基于区块链的职业教育学生实践管理系统的研究与应用	李冠楠	顺德职业技术学院
339	2021KTSCX339	基于AR技术的农产品包装可视化研究与实践	赵江平	广东岭南职业技术学院
340	2021KTSCX340	课堂教学质量的两极定性评价WSR-可拓云模型及求解	耿江涛	广州涉外经济职业技术学院
341	2021KTSCX341	培养高职学生计算思维的Euclidean示范算法研究与实践	熊晓波	广州涉外经济职业技术学院
342	2021KTSCX342	5G时代基于现代学徒制的数字媒体专业职业本科教育创新实践研究	黄红林	广州涉外经济职业技术学院
343	2021KTSCX343	微型电窑烧制釉下五彩陶瓷作品实验及其教学作品研发	尚香	广州南洋理工职业学院
344	2021KTSCX344	基于大数据的用户个性化推荐系统研究与实践	薛慧丽	广州南洋理工职业学院
345	2021KTSCX345	基于机器视觉和人工智能深度学习技术的金属表面缺陷检测研究	马静	惠州经济职业技术学院
346	2021KTSCX346	基于花样小图技术的飞织3D针织鞋面产品设计与开发	陈文焰	惠州经济职业技术学院
347	2021KTSCX347	软件测试工具在超星教学系统改进中的应用研究	于明清	广州华南商贸职业学院
348	2021KTSCX348	Python网络爬虫技术的研究与探索	黄仁宏	广州华南商贸职业学院
349	2021KTSCX349	智能助老爬楼机器人轻量化设计研究	陈运胜	广州华立科技职业学院

广东省普通高校特色创新项目

申报书(自然科学)

项 目 类 别: 特色创新项目(自然科学)

项 目 名 称: 软件测试工具在超星教学系统改进中的应用研究

学 科 分 类: 计算机科学技术

项目负责人: 于明清

负责人手机: 13610015918

所 在 学 校: 广州华南商贸职业学院(盖章)

广东省教育厅制

二〇二一年五月

基本信息

项目信息	项目名称	软件测试工具在超星教学系统改进中的应用研究		
	项目类别	特色创新项目(自然科学)		
	研究类型	应用研究	申请金额	2(万元)
	学科一	计算机科学技术 - 计算机软件		
	学科二			
	学科三			
	计划开始日期	2021.7	计划完成日期	2023.7
	所属学校	广州华南商贸职业学院	学校类型	民办高职高专院校
	预期成果形式	论文、研究报告		

合作单位	合作单位名称	联系人	联系电话	通讯地址

负责人信息	姓　名	于明清	性　别	男	民　族	汉族
	出生年月	1969.4	学　历	大学本科	学　位	学士
	职　称	副高级		职　务	副教授	
	办公电话	13610015918		手　机	13610015918	
	一级学科	信息科学与系统科学		二级学科	控制理论	
	电子邮件	513752361@qq.com		身份证号	432902196904050336	
	人才层次	副教授				
	研究专长	软件系统研发				

摘要	通过本课题的研究，对基于 Web 应用的超星教学系统的用户登录和云盘上传两个模块，从界面、功能、接口和性能方面进行测试，综合运用自动化测试技术，使用多种主流的测试工具、尽可能对系统进行深入的测试，试图提出该系统的改进优化方案，提高超星教学系统的软件质量，以保障系统的兼容性、稳定性、完整性、易用性，保证用户使用的持续性；并且试图将对该系统的测试工具和方法推广到其他 Web 应用系统。
关键字	Web 应用　软件系统　集成测试 接口测试 性能测试

项目组成员

总数（含负责人）		高级		中级	初级	博士	硕士	学士
7		3		3	1	1	0	3
姓名	性别	出生年月	学位	职称	项目分工	工作单位		研究领域
于明清	男	1969.4	学士	副高级	制定测试计划和测试方案	广州华南商贸职业学院		系统开发
王芬	女	1982.8	硕士	中级	实施测试	广州华南商贸职业学院		软件测试
江东梅	女	1980.10	硕士	副高级	撰写研究报告	广东机电职业技术学院		高职教育
毛振宁	男	1986.5	硕士	中级	撰写论文	广州华南商贸职业学院		数据库管理
李锡炼	男	1980.6	学士	中级	提出优化方案	广州华南商贸职业学院		计算机应用
何达齐	男	1996.6	学士	初级及以下	撰写测试报告	广州华南商贸职业学院		计算机应用

经费申请表

（金额单位：万元）

预算科目	创新强校工程经费	备注（计算依据与说明）
一、科研业务费	1.3000 万元	
1、测试、计算、分析	万元	
2、会议费、差旅费	0.5000 万元	参加学术会议、外出交流
3、出版物、文献、信息传播	0.8000 万元	论文版面费
4、其他	万元	
二、试验材料费	0.2000 万元	
1、原材料、试剂、药品购置费	0.2000 万元	文献材料复印、项目材料打印
2、其他	万元	
三、仪器设备费	0.5000 万元	
1、购置	0.5000 万元	测试环境建设
2、试制	万元	
四、劳务费	0.0000 万元	
五、其他费用	0.0000 万元	
1、	万元	
2、	万元	
3、	万元	
4、	万元	
合计	2.0000	
与本项目有关的其他经费来源	其他计划资助经费	万元
	其他经费资助	万元
	其他经费合计	0.0000 万元

进度计划

序号	起止时间	阶段性研究工作进展	阶段性目标
1	2021.7-2022.1	调研被测系统存在的问题、分析需求、制定测试计划、设计测试用例	完成系统需求调研、编写开题报告
2	2022.1-2022.7	对被测系统的功能、界面、接口、性能实施测试	编写研究报告、发表论文 1 篇
3	2022.7-2023.1	分析测试结果	完成测试报告、制定优化方案
4	2023.1-2023.7	进行项目收尾，整理中期成果，公开发表，撰写结项报告、申请验收、结项	结题的总结报告、发表论文 1 篇

预期成果

<table>
<tr><td rowspan="3">论文（篇）</td><td>总数</td><td colspan="2">2</td></tr>
<tr><td>其中：CSCD 核心期刊</td><td colspan="2"></td></tr>
<tr><td>三大索引收录</td><td colspan="2"></td></tr>
<tr><td colspan="2">专著（部）</td><td colspan="2"></td></tr>
<tr><td colspan="2">研究报告（篇）</td><td colspan="2">1</td></tr>
<tr><td rowspan="4">专利（件）</td><td rowspan="2">数量（件）</td><td>申请</td><td></td></tr>
<tr><td>授权</td><td></td></tr>
<tr><td rowspan="2">其中发明专利</td><td>申请</td><td></td></tr>
<tr><td>授权</td><td></td></tr>
<tr><td colspan="2">鉴定成果（项）</td><td colspan="2"></td></tr>
<tr><td colspan="2">软件登记（项）</td><td colspan="2"></td></tr>
<tr><td colspan="2">新产品（种）（或新装备、新药等）</td><td colspan="2"></td></tr>
<tr><td colspan="2">新技术（项）（或新工艺等）</td><td colspan="2"></td></tr>
<tr><td colspan="2">其他</td><td colspan="2"></td></tr>
</table>

申请书正文

一、立项依据

1．立项的必要性及需求分析（立项背景、目的、意义）

随着 Internet 与电子商务的迅速发展，Web 应用系统大量出现，而Web 服务由于其自身无需安装、跨平台、分布式、动态交互的特点，已经给人们的日常生活和工作带来了很大的改变。由于目前软件复杂性和规模的变大，导致了对于系统测试难度的提高,因此Web 应用系统的功能以及性能问题，越发受到开发人员的关注。 现如今Web 的流量在 Internet 总的流量中所占有的百分比已经日渐增高，因此对于Web 应用所提供的功能，以及Web 服务器所要求的性能也逐渐变得很高。然而，实际上目前大多数的Web 应用系统根本不能够支持大量用户的同时访问，主要是因为这些系统并没有经过严格完整的性能测试。在系统对外提供服务前假如没有经过全面的测试，将会致使在系统正式上线后，一旦受到过多的用户同时并发请求访问时，可能会导致系统响应过慢或者系统崩溃的情况，给访问系统的用户带来非常糟糕的体验。 在Web 系统投入使用前，对系统进行全面的功能与性能分析测试，能够帮助开发者确定系统功能的正确性和影响系统性能的关键因素，从而有针对性地对系统进行分析和改进。相对于传统的软件测试而言，Web 应用的测试除了要对系统功能进行测试，同时需要关注的测试点还有很多。因此，Web 应用的测试相比一般应用程序的测试来说更加的复杂，同时也使软件测试迎来了新的机遇和挑战。

以往的软件测试一直采用手工测试，但随着软件日益复杂和庞大，手工软件测试设计的大量的重复性的工作，将耗费更大量的时间和人力，软件测试的开销将不断增大，如何更有效的进行测试就成为一个新的讨论热点，因而诞生了软件测试工具。现在，运用软件测试工具进行软件自动化测试已成为人们日益关注的一个焦点。

软件测试工具是用某种程序设计语言编制的自动测试程序，控制被测软件的执行，模拟手动测试步骤完成全自动或半自动测试。全自动测试过程中，不需要人工干预，由程序自动完成测试的全过程；而半自动测试就是指在测试过程中，需要由人工手动输入测试用例或选择测试路径，再由自动测试程序按照人工指定的要求完成测试。软件测试工具适用范围为软件需求变动不频繁，项目周期较长，编写的自动化测试脚本可重复使用等场景，主要适用于系统级的测试，而不适用于单元测试。典型的系统级的测试主要包括：集成测试、回归测试、系统测试和性能测试。

系统测试是指在系统已成为一个相对稳定的可测试版本以后，对系统进行的大规模的、多周期的、全面的功能测试。自动测试程序在这一阶段中，可以完成对全部功能或部分功能的测试。

性能测试是通过对被测系统进行长时间、多用户、大数据量等压力负载的测试，以验证软件系统是否能够达到用户提出的性能指标，同时发现软件系统中存在的性能瓶颈，优化软件，最后起到优化系统的目的。性能测试类型包括负载测试，强度测试，容量测试等。

时代不断发展，科学技术逐渐壮大人工智能以及大数据已迅速成长。随着测试越来越多地朝着更自动化的方向发展，人们将目光转向了人工智能（Artificial Intelligence），作者认为这将会

是软件测试的研究新方向。

本文结合目前正在全国高校广泛应用的Web 应用系统超星教学系统使用过程中发现的部分问题：（1）已注册的用户登录不成功（2）用户登录成功后出现界面不正确（3）高峰期用户无法成功登录（4）云盘上传资料失败（5）学生重复加入班级（6）学生测验自动改分不正确等。因此有必要运用软件测试工具来进行系统测试、性能测试以重现这些问题，分析测试结果，并提出系统或者模块优化方案，以提高超星教学系统的软件质量。

2．国内外研究现状、水平和发展趋势分析

自从二十世纪七十年代以来，作为软件这门学科的重要分支，软件测试也得到了相应的发展。国外的许多研究机构（如 National Software Testing Lab）、大学（如 George Mosan、Carnegie Mellon 等），以及公司等,对软件测试方面做了长时间的研究。其中 George Mosan 大学主要偏重面向对象的测试相关技术。国外对软件自动化测试的研究，已形成了一套较为成熟的软件测试流程，并产生了一批实用性的自动化测试工具，并在近几年提出了自动化测试框架的概念。IEEE、ACM 等美国专业机构制定了一系列软件测试规范。卡内基梅隆大学、华盛顿大学、美国国家软件测试实验室、Mercury Interactive、Rational Corporation等公司进行了大量的软件测试研究和应用工作。卡内基梅隆大学专注于回归测试和C/S测试技术，而乔治梅森大学专注于自动测试生成和面向对象测试技术。现如今，在欧美国家也已经有了一些特意针对 Web 系统进行性能测试的工具。其中比较常用的商用方面的测试工具有Mercury Interactive 的 LoadRunner，LoadRunner 能够预测系统的行为与性能，是一款负载测试工具，并且支持多种系统架构和协议，性能稳定，测试效果。Rational 公司的 Robot，在功能测试工具的领域较为突出，同样支持多种网络协议，而且还能够对各种协议过滤处理，但是在操作上对用户的要求较高。微软公司的 WAS（Web Application Stress Tool）适用于 B/S 架构，主要用来对网站系统执行压力测试]。另外还有一些免费的开源性能测试工具，如 Apache 的 JMeter，Open STA 组织的 Open STA 等。自2017 年开始出现基于AI的测试平台或测试工具，如Facebook发布名为Sapienz的工具，Bugdojo创建基于AI的测试平台，DiffBlue发布三款AI软件测试产品，微软发布“AI安全风险检测”工具等。

在国内对于软件测试工具方面的研究不是很多，北京航空航天大学与北京大学比较重视测试工具的研发和软件系统的分析，研制出了一系列以 SafePro C /C++ 和 SafePro / javao 为首的测试工具与程序理解工具。南京大学与航空计算机研究所在嵌入式的系统测试方面，研发了几种能够自动生成相关测试用例的工具，和系统资源的静态分析工具。国内内蒙古大学的刘亮等人设计并实现了针对 Web 项目的性能测试工具，并在有关操作的录制与重新播放、系统性能的监控、压力测试、以及测试结果的输出方面进行了一些研究，但是这个测试工具也有其不足之处，不能够支持多台测试服务器的协同工作。西安理工大学的刘苗苗等人，同样设计并研发了一个对 Web 系统执行性能测试的工具，能够模拟多用户同时访问 Web 系统，支持多主机多线程同时运行测试脚本，但是对于多种协议的支持还不够，例如 FTP、SMTP 等。

手工测试一直以来在软件测试项目中占据主导地位，近年来自动化测试技术在国内的发展速

度很快，从以前的不重视自动化测试技术到现在开始致力于自己测试部门的自动化测试。一些大中型企业成功的例子，更是加强了自动化测试技术的信心，部分知名软件公司都经有了自己的自动化测试平台，并带来了效益。自动化测试是对现有软件测试技术一个补充，它不可能完全淘汰手工测试。在测试项目中软件测试用例需要大量执行，测试环境具有多样性的情况下，自动化测试可以在没有人为干预的情况下完成测试工作。现在很多公司招聘测试人员的要求越来越高，很多好公司招Senior QA,都要求5年工作经验以上，精通软件测试工具，掌握多种编程语言，有丰富的自动化测试经验。运用软件测试工具进行软件自动化测试是趋势所向。

目前人工智能在测试领域的应用可以说是凤毛麟角，如阿里Ripper，中国信息通信研究院泰尔终端实验室的智测云测试平台引入AI技术等。

二、研究方案

1. 主要研究目标与研究内容

1.1 研究目标：

由于Web 应用系统的复杂性和用户行为的不可预见性，使得对该系统运用软件测试工具进行自动化测试比较困难。近年来关于Web应用系统的自动化测试的相关论文只是将重点放在优化功能测试结果和性能测试的负载、强度和持久度测试上，并没有详细介绍自动化功能测试和性能测试脚本的编写。在自动化测试中，合理的测试用例设计和编写自动化测试脚本是提高测试效率和节约测试时间的重要保障。在性能测试中，已有的性能指标不能综合反映的Web 应用系统的实际运行情况。综上所述，Web 应用系统正在迅速的普及，但是系统的质量保证，尤其是功能和性能方面的测试工作并未随着技术的发展而改进，为了保证系统功能和性能方面的质量，需要对系统的测试流程、方法、技术等进行深入的研究和改进。

本研究以基于Web 应用的软件超星教学系统作为实例，对系统的用户登录和云盘上传两个模块进行功能测试和性能测试，并进行了详尽的测试用例设计和测试脚本的编写，分析测试结果，给出系统和模块调优的方案，并试图找出Web 应用系统测试的改良方法。

1.2 研究内容：

超星教学系统是全国高校使用的系统，使用的人数有4000万人左右，同时高校经常采用在线课程学习，高峰时期会达到1000万人左右同时使用系统，因此对系统的功能和性能要求较高。重点选取首页登录与上传云盘的高频率使用模块进行测试。对该系统的两个模块进行了详尽的测试用例设计，并将软件测试工具应用于该系统中。主要内容如下：（1）分析了目前国内外关于软件测试技术与测试工具的研究现状与不足，提出一种结合自动化功能测试与性能测试的超星教学系统测试方案；（2）将软件测试工具应用于该超星教学系统，为系统两个主要模块设计了完备的测试方案，设计了测试用例并编写了对应的测试脚本。（3）为系统模块提出改进措施。

2. 技术关键（创新点与技术难点）

2.1 拟解决的关键问题

（1）Web 自动化测试工具对超星教学系统的用户登录模块和云盘上传模块进行界面测试，给出

调优方案，解决登录进入的界面不正确或云盘上传资料页面无反映的问题，保证模块界面正常。

（2）接口自动化测试工具对超星教学系统的用户登录接口和云盘上传接口进行功能测试，解决用户登录不成功、云盘上传资料失败的问题。根据历史数据模型推算，底层的1个bug大约会引发上层8个bug，所以对底层的接口进行测试变得很重要。

（3）性能自动化测试工具找出系统登录和上传资料的瓶颈，解决高峰时大量用户登录失败或者超星云盘用户一次上传多个资料失败的问题。高峰时期一般指周一至周五上午8：00左右；大量用户一般指超过1000万用户。每个超星云盘用户都具有100GB免费存储空间，上传文件不限大小。寻找系统登录和上传模块的瓶颈，发现模块对整个系统的性能影响。

2.2 创新点

（1）提出了基于 Python+Selenium+Web driver 的Web 自动化测试工具和Jmeter接口测试工具对系统两个模块的前端页面和重要接口进行自动化测试。前者实现前端页面的界面、功能自动化测试，后者在接口功能测试的基础上增加了性能测试，重现超星教学系统的实际运行情况，试图给出两个被测模块的调优方案。

（2）综合运用界面、功能、接口和性能的自动化测试工具，尽可能对Web 应用系统进行深入的测试，并试图改良测试方法。

2.3技术难点

（1）采用Web 自动化测试开源框架—Selenium3.0，脚本语言选用Python，进行用户登录、云盘上传页面测试。Selenium+Web Driver 能够提供支持动态网页。Web Driver提供设计良好的面象对象的API，因此采用Selenium+Python+Web driver 编写测试脚本对页面的功能和界面测试

（2）使用Jmeter+Fiddler对登录、云盘上传资料接口进行功能测试。一方面模拟用户使用流程，保证功能、逻辑正确；另一方面考虑接口调用的易用性。接口测试需要指定请求调用页面的URL、参数去调用接口，检验返回值是否符合期望。

（3）使用Jmeter对多用户同时登录或上传云盘接口进行性能测试。性能测试的衡量指标主要是平均响应时间、吞吐量、并发数。通过使用JMeter提供的功能，制定可视化的测试计划：包括规定使用多少负载、测试什么类型的请求、传入的参数以及测试结果显示方式。

3. 拟采取的研究方法、技术路线、实验方案及可行性分析

3.1 采取的研究方法有文献分析法、科学实验法、行动研究、质的研究等方法。各种方法的具体运用方法如下：

（1）文献分析法：查阅Web 应用系统自动化测试的国内外研究现状及研究成果，为本研究提供理论依据。

（2）科学实验法：设置前置条件，观察、研究Web 应用系统自动化测试的结果数据。

（3）行动研究：本课题采用的行动研究法主要应用于研究与探索自动化测试技术在Web 应用系统中的应用，提供具有指导价值的理论及方法。

（4）质的研究：在本研究中，针对自动化测试技术在Web 应用系统中的应用的实际情况，采用

参观、考察等手段收集国内和省内相关研究资料，对此进行描述性的分析和解释，吸取优秀的经验，设计更优越的测试方案等。

3.2 技术路线

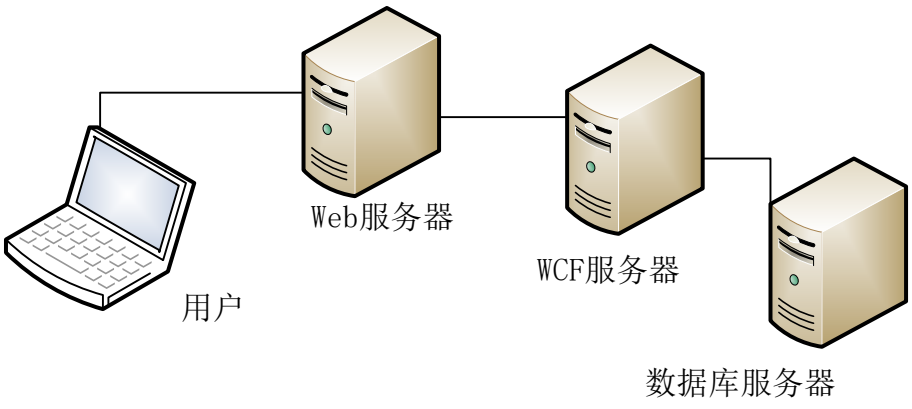
配置硬件环境：

服务器名称	配置/详细信息	数量	IP
Web 服务器			
数据库服务器			
客户端	内存 4GB	1	192.168.1.100

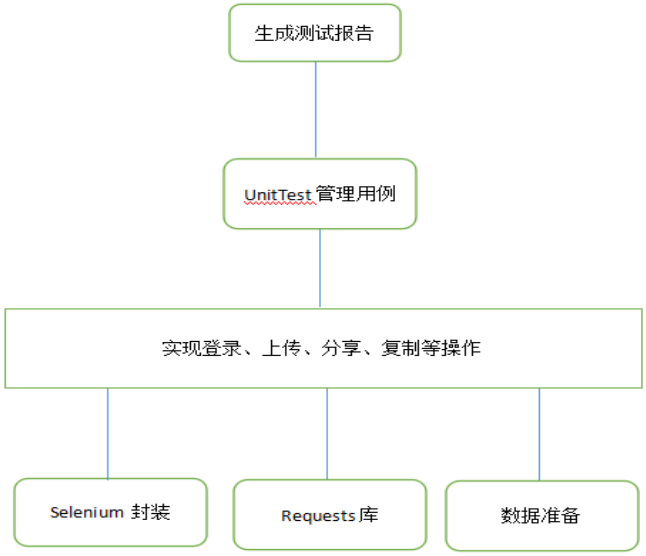
配置软件测试环境：

序号	软件名称	Web 服务器	数据库服务器	测试 PC
1	Windows 10			Windows10
2	Firefox 浏览器			
3	Pycharm2019			
4	Python 3.7			
5	Selenium3.0			
6	Web driver/geckodriver			
7	Jmeter4.0			
8	Fiddler			
9	其他			

测试环境组网图：



测试系统流程图



3.3 实施方案

3.3.1 登录模块测试

用例名称	登录测试	**用例编号**	001
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	输入正确用户名密码和验证码进行登录。		
预期结果	登录成功		
实际结果	登录成功		

用例名称	登录测试	**用例编号**	002
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	用户名为空，密码正确，验证码正确。		
预期结果	登录失败		
实际结果	账号不能为空		

用例名称	登录测试	用例编号	003
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	用户名正确，密码为空，验证码错误。		
预期结果	登录失败		
实际结果	验证码错误		

用例名称	登录测试	用例编号	004
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	用户名正确，密码正确，验证码为空		
预期结果	登录失败		
实际结果	验证码错误		

用例名称	登录测试	用例编号	005
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	用户名随机输入，密码随机输入，验证码正确。		
预期结果	登录失败		
实际结果	用户名和密码错误		

用例名称	登录测试	用例编号	006
测试步骤	1、打开 http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal 2、编写测试程序 3、观察反馈信息		
场景设计	用户名速记输入，密码随机输入，验证码随机输入。		
预期结果	登录失败		
实际结果	验证码错误		

登录功能自动化测试脚本编写，如下：

```
# 登录测试
from selenium import Web driver
driver = Web driver.Firefox()
driver.implicitly_wait(30)
# 打开登录网页
driver.get("http://passport2.chaoxing.com/login?fid=5758&refer=http://hnsmxy.fanya.chaoxing.com/portal")

# 账号
id = driver.find_element_by_id("unameId")
```

```
# 密码
pw = driver.find_element_by_id("passwordId")
# 登录按钮
sub = driver.find_element_by_class_name("zl_btn_right")
# 验证码
qrbtn = driver.find_element_by_id("numcode")
# 15914238993
Uid = "112314551"
# 18819181348asd
Upw = "223559422"
# 测试：
driver.implicitly_wait(30)
id.send_keys(Uid)
driver.implicitly_wait(30)
pw.send_keys(Upw)
driver.implicitly_wait(30)
print("验证码:",end="")
qrbtn.send_keys(input())
driver.implicitly_wait(30)
sub.click()
# 错误信息
massage = driver.find_element_by_id("show_error")
if(massage):
    print(massage.text)
driver.implicitly_wait(300)
# driver.quit()
```

用例名称	上传云盘测试	**用例编号**	001
测试步骤	1、用户登录 2、遍历生成的文件所在的文件夹，得到所有的文件路径 3、通过给 FileUpload 对象循环发送文件路径完成上传操作		
场景设计	用户登录，指定生成测试文件（txt,csv、word、pdf、jpg 等）		
预期结果	所有文件上传成功，查询文件信息显示正确		
实际结果			

3.3.3 性能测试

重点选取访问量大，对性能要求较高的首页登录页面并发访问能力、云盘上传的并发能力进行压力测试，了解系统的响应时间、吞吐量、并发数与系统可靠性，即请求正确率。

3.3.4 测试结果分析

根据测性能测试结果，分别对首页登录页面、云盘上传页面绘制并发用户量与平均响应时间、吞吐量的关系图。

3.3.5 系统优化措施

通过对超星教学系统首页登录与云盘上传两个模块进行功能和性能测试，以及根据测试过程中出现的系统问题，提出改进建议，以后还需要对优化后的系统进行回归测试，使系统在功能和性能方面能满足用户需求。

3.3.6 软件测试技术改善措施

（1）提高测试速度，加大测试准确性。由于软件测试的速度较慢以及准确性模糊不清，软件数据的测试效果和实用性就比较低。我们需要通过反复的实验找到降低测试速度和准确性的根本，做出进一步的完善和改进，使出错率降到最低运算速度提到更高。

（2）利用人工智能使数据处理更加全面。智能化技术的输入数据范围广泛而且具有人工智能特效。我们可以利用智能化改变软件测试的原始设计，使测试取长补短，在原来的基础上解决问题并且让测试数据更加方便快捷。

（3）组建大数据测试环境。完好的测试环境可以提高数据处理速度，保证数据信息的完整，使信息利用率提高。

3.4 可行性分析

（1）项目组成员年龄、学历、职称、知识结构合理，管理协调，科研能力强，并且具有丰富的企业实践的经验，这些都是本项目如期高质量完成的有效保证。

（2）项目负责人曾经完成多个系统的研发，具备丰富的系统研发经验和项目经验；项目组主要成员曾经负责多个系统的测试，包括 Web 应用程序、移动客户端软件、小程序，具有丰富的测试经验，并已撰写软件自动化测试方法技术类的论文有《接口测试中数据关联技术的运用》和《Fiddler 工具在接口测试中的应用》，负责的项目《自动化测试技术在 B/S 架构系统中的应用研究》在 2020 年度通过校级立项，并指导学生在 2021 年度广东省高职院校职业技能软件测试赛项竞赛中获得三等奖第二名。

（3）学校实训室具备完成项目的所需要的实验室硬件配置和软件环境。

4. 技术、经济效益及风险分析

4.1 技术可行性分析

（1）项目的技术路线合理、成熟，关键技术先进，软件测试工具采用主流开源工具，能够很好地实施软件系统测试并提出改进方案。

（2）在软件测试的规范、技术与相关意识上，我国与其他一些发达国家相比还有着较大的需要提升的空间。

（3）项目承担单位具有来自企业一线研发、测试、项目经验丰富的教师队伍，并具备实验室软硬件环境。

4.2 经济效益分析

项目的实施能够大幅度提高超星教学系统的软件质量，保证了软件的功能性、可靠性、可用性、效率、可维护性和可移植性。首先在本校实施推广，并逐步对省内其它高职院校甚至全国高职院校辐射和示范。项目所获得的成果可借鉴于中小型企业软件研发部门的测试工作实施方案。

受益面为同类高职或企业从事研发测试工作的人员。

4.3 风险分析

项目组已经充分做好风险分析，包括需求风险、技术风险、团队风险、关键人员风险、预算风险和范围风险，并准备好了风险应对措施。

5. 要达到的主要经济、技术指标

保证项目及产品符合质量要求，测试方法得当、测试充分，软件的性能指标满足用户需求。

6. 将提供的研究开发成果及形式

6.1完成基于Web 应用系统的软件系统超星教学系统中登录和云盘上传两个模块的测试方案、测试计划、测试用例、测试脚本、测试结果及分析报告，并提出系统模块调优方案。

6.2成果形式：

(1) 项目的研究报告：《自动化测试技术在WEB 应用系统中的应用》课题的研究报告1份。预计项目研究中期完成。

(2) 相关论文：在国内外刊物上发表相关学术论文2篇。预计项目研究中后期完成。

三、研究基础

1. 与本项目相关的研究工作积累和已取得的研究工作成绩

（1）查阅了一定数量的相关论文

（2）对系统的需求做了初步的调研。

（3）对系统使用中出现的一些问题进行了收集。

（4）搭建了测试环境，软硬件条件基本备好。

（5）项目组成员已确立，并根据各自的优势进行了分工。

（6）项目组组长曾经是从事一线系统研发的高级工程师，具备多个项目完成的经验；项目组主要成员曾经在企业从事相关的系统自动化测试工作，现今又在学校担任软件测试技术系列课程的教学工作，并且查阅多篇硕博论文或期刊论文进行相关技术的学习，相对深入地对所申请项目进行了研究。同时还在教学期间发表了两篇关于软件自动化测试工具的省级核心期刊论文。

2. 必要的场地、设备等支撑条件、组织措施及实施方案

（1）我院高度重视教研活动，非常注重教师的项目研究，尽全力给予项目研究者指导和帮忙。

（2）我校图书馆购买部分知网权限为项目研究者提供很多关于自动化测试技术的资料，能为本项目研究供给优质的资源。

（3）项目组组长负责开展活动，各组员之间结成研究对子，并且定期汇报、交流。每月月底再举行项目组组长例会，对项目进行阶段性的分析，并指导下阶段的工作。

（4）与上级科研处领导多联系，及时汇报项目研究阶段进展，确保项目扎实地有成效地开展。

3. 项目组负责人学术水平和管理能力情况，项目组主要成员的研究工作情况及在本课题中的工作分工

3.1 项目组负责人学术水平和管理能力情况

（1）参与国家 863 工程科研项目《JZCIMS-厂长信息管理系统》 研发（独立完成财务、销售、车间成本子系统）

(2)独立研发《MVS下ISPF系统汉化》

(3)独立研发《江汉石油管理局闲置低效信息管理系统》

(4)独立研发《建筑工程概预算信息系统》

(5)译著《IBM DATABAS DB2 INTRODUCTION》 （10万字）

(6)组织并审译《IBM DATABAS DB2 UTILITIES》 （20万字）

(7)在中文核心期刊《电子测量与仪器学报》2020第9期发表论文《部分传输系列的遗传模拟退火搜索方法》

3.2 项目组主要成员的研究工作情况及在本课题中的工作分工

项目组主要成员参与多个系统的测试，包括Web 应用程序、移动客户端软件、小程序，具有比较丰富的测试经验，已撰写两篇关于软件测试技术方面的论文，如《接口测试中数据关联技术的运用》和《Fiddler工具在接口测试中的应用》，负责的项目《自动化测试技术在WEB 应用系统中的应用研究》在2020年度通过校级立项，并指导学生在2021年度广东省高职院校职业技能软件测试赛项竞赛中获得三等奖第二名，在项目中担任测试管理和测试实施的工作；项目参与人主持或参与学校多个教学科研项目的实施，具备了项目实施的经验，在项目中担任数据收集、测试报告整理、提出系统优化方案等工作。

签字和盖章页(此页自动生成，打印后签字盖章，上传扫描件)

申请者：　　于明清　　依托单位：　　广州华南商贸职业学院

项目名称：　　软件测试工具在超星教学系统改进中的应用研究

申请者承诺：

本人符合各项申报条件。本表各项内容真实、数据准确，不涉密，没有知识产权争议。如果获准立项，承诺以本表为有约束力协议，遵守有关规定，按计划认真开展研究工作，取得预期研究成果，并按时报送有关材料。若填报失实和违反规定，本人将承担全部责任。

签字：于明清

项目组主要成员承诺：

本人保证有关申报内容的真实性。本人将严格遵守广东省教育厅的有关规定，切实保证研究工作时间，加强合作、信息资源共享，认真开展工作，及时向负责人报送有关材料。若个人信息失实、执行项目中违反规定，本人将承担相关责任。

编号	姓名	工作单位	分工	签名
1	于明清	广州华南商贸职业学院	制定测试计划和测试方案	于明清
2	王芬	广州华南商贸职业学院	实施测试	王芬
3	江东梅	广东机电职业技术学院	撰写研究报告	江东梅
4	毛振宁	广州华南商贸职业学院	撰写论文	毛振宁
5	李锡炼	广州华南商贸职业学院	提出优化方案	李锡炼
6	何达齐	广州华南商贸职业学院	撰写测试报告	何达齐

依托单位和合作单位承诺

已按填报说明对申请人的资格和申请书内容进行了审核。本单位保证对研究计划实施所需要的人力、物力和工作时间等条件给予保障，严格遵守广东省教育厅有关规定，督促负责人和主要成员以及本单位科研管理部门按照广东省教育厅的规定及时报送有关材料。

	依托单位	合作单位 1	合作单位 2
单位名称	广州华南商贸职业学院(公章)	(公章)	(公章)
承诺经费	2(万元)	0(万元)	(万元)
日期：	年　月　日	年　月　日	年　月　日

广东省教育厅科研项目重要事项变更申请表

项目名称	软件测试工具在超星教学系统改进中的应用研究	批准号	2021KTSCX347
		联系方式	15918580256
项目责任人	于明清	工作单位	广州华南商贸职业学院
批准立项时间	2021年 8 月	原项目成果形式	研究报告、论文
原完成时间	2023年8月	延期完成时间	

变更内容（请在方框内打“√”）：

☑变更项目责任人　☐变更项目管理单位　☐改变成果形式

☐更改项目名称　☐研究内容有重大调整　☐第一次延期

☐第二次延期　☐申请撤项　☑变更课题组成员　☐其他

变更事由：

（变更项目负责人须写明新项目负责人的性别、出生时间、职称、工作单位、联系电话、专业、研究方向及主要工作简历等情况，新项目负责人尽量为原课题组成员，并在下框中签名确认；变更课题组成员须写明在课题组中的排位，附上新课题组成员的简历，并附上原全体项目组成员签名；变更项目管理单位须由调出、调入单位签署意见。）

新项目负责人王玉山，男，1963年3月出生，计算机副教授，广州华南商贸职业学院，15918580256，计算机科学与技术专业，研究方向为软件工程，从事高校教学工作30余年。

同时，因项目组成员工作变动，将江东梅、毛振宁、李锡炼、何达齐，从本项目组成员名单中移除，并添加一位项目组成员：王珂（广州华南商贸职业学院云智信息技术学院副院长，1988年5月，男），调整后项目组成员排序如下：王玉山，王芳，王珂。

	项目负责人签章：王玉山 2022 年 12月 20日
项目依托单位意见	科研管理部门负责人签章： 22 年 12月 20日
转出单位意见及签章： 年 月 日	转入单位意见及签章： 年 月 日
教育厅项目管理单位意见	教育厅项目管理单位盖章： 年 月 日

注：申请延期一次最多不得超过 1 年，一个项目申请延期最多不得超过 2 次。

结项证书

项目类别：广东省高等学校特色创新项目（自然科学）

项目编号：**2021KTSCX347**

项目名称：软件测试工具在超星教学系统改进中的应用研究

负 责 人：王玉山

课题组成员：王芬、王珂

证书编号：**2021KTSCX347_230987**

所在单位：广州华南商贸职业学院

该项目经审核，符合结题条件，准予结项。

广东省教育厅科研处

2023 年 12 月 30 日

广 东 省 教 育 厅

粤教科函〔2021〕7号

广东省教育厅关于公布2021年度普通高校认定类科研项目立项名单的通知

各有关高校:

为深入实施创新驱动发展战略，落实《广东省教育厅 广东省科学技术厅关于印发科教融合协同推进高校科技创新能力提升工作计划的通知》（粤教科函〔2019〕57号），省教育厅组织开展了2021年度科研项目认定工作。经学校推荐、省教育厅组织形式审查，现将批准立项的2021年度高校认定类科研项目立项名单（见附件）下达各高校。

请各高校按照国家和省相关科研平台项目管理办法，统筹安排项目资金，加强资金管理，督促项目承担人按照项目申请书开展建设工作，协助解决项目实施过程中遇到的困难和问题，确保研究项目如期完成目标任务。

附件: 1.2021年度广东省普通高校特色创新类项目立项名单

2.2021年度广东省普通高校青年创新人才类项目立项

名单

广东省教育厅

2021年8月25日

（联系人及电话：曾俊伟，020-37627742）

公开方式：主动公开

校对人：曾俊伟

附件1

2021年度广东省普通高校特色创新类项目立项名单

1. 自然科学类				
序号	项目编号	项目名称	负责人姓名	所属学校
1	2021KTSCX001	音圈电机与偏磁电机（本体及驱动）设计与开发	卢少锋	华南理工大学
2	2021KTSCX002	老年人防跌倒外骨骼助行产品系统设计研究	熊志勇	华南理工大学
3	2021KTSCX003	新型高效呈味肽制备关键技术研究	崔春	华南理工大学
4	2021KTSCX004	“双碳”目标下基于计算性设计思维的低碳绿色校园规划智能优化研究	刘骁	华南理工大学
5	2021KTSCX005	多品种产品混流生产过程动态模式表征及智能调控方法	王世勇	华南理工大学
6	2021KTSCX006	基于注意力机制的安全性图像识别模型研究与应用	李海良	暨南大学
7	2021KTSCX007	中药来源的新型HDC抑制剂的发现与抗骨质疏松作用机制研究	邱佐成	暨南大学
8	2021KTSCX008	应用新型蓝莓综合开发技术推动乡村振兴	蒋鑫炜	暨南大学
9	2021KTSCX009	富硒富岩藻黄素微藻用于类风湿关节炎治疗及其作用机制探究	汪翔	暨南大学
10	2021KTSCX010	鸡柔嫩艾美耳球虫MIC3基因重组株构建及生物学特性研究	林瑞庆	华南农业大学
11	2021KTSCX011	生物质化学链气化中铁基载氧体的失活机理	胡志锋	华南农业大学
12	2021KTSCX012	Nrf2/GPX4介导的铁死亡在ATO致肉鸡肝损伤中的作用机制研究	胡莲美	华南农业大学
13	2021KTSCX013	木麻黄青枯病菌关键致病基因鉴定和功能研究	周筱帆	华南农业大学

序号	项目编号	项目名称	负责人姓名	所属学校
334	2021KTSCX334	基于大数据的数字教育资源促进乡村教学质量提升的策略研究——以云浮市为例	谭玉玲	罗定职业技术学院
335	2021KTSCX335	基于缓释功能设计的半互穿网络有机-无机纳米复合微凝胶的制备	练翠霞	顺德职业技术学院
336	2021KTSCX336	涂料用多机制耦合高效阻燃体系的制备及其应用性能研究	姜佳丽	顺德职业技术学院
337	2021KTSCX337	相变蓄冷材料的研制及其主-被动耦合系统的建筑节能应用研究	孙婉纯	顺德职业技术学院
338	2021KTSCX338	基于区块链的职业教育学生实践管理系统的研究与应用	李冠楠	顺德职业技术学院
339	2021KTSCX339	基于AR技术的农产品包装可视化研究与实践	赵江平	广东岭南职业技术学院
340	2021KTSCX340	课堂教学质量的两极定性评价WSR-可拓云模型及求解	耿江涛	广州涉外经济职业技术学院
341	2021KTSCX341	培养高职学生计算思维的Euclidean示范算法研究与实践	熊晓波	广州涉外经济职业技术学院
342	2021KTSCX342	5G时代基于现代学徒制的数字媒体专业职业本科教育创新实践研究	黄红林	广州涉外经济职业技术学院
343	2021KTSCX343	微型电窑烧制釉下五彩陶瓷作品实验及其教学作品研发	尚香	广州南洋理工职业学院
344	2021KTSCX344	基于大数据的用户个性化推荐系统研究与实践	薛慧丽	广州南洋理工职业学院
345	2021KTSCX345	基于机器视觉和人工智能深度学习技术的金属表面缺陷检测研究	马静	惠州经济职业技术学院
346	2021KTSCX346	基于花样小图技术的飞织3D针织鞋面产品设计与开发	陈文焰	惠州经济职业技术学院
347	2021KTSCX347	软件测试工具在超星教学系统改进中的应用研究	于明清	广州华南商贸职业学院
348	2021KTSCX348	Python网络爬虫技术的研究与探索	黄仁宏	广州华南商贸职业学院
349	2021KTSCX349	智能助老爬楼机器人轻量化设计研究	陈运胜	广州华立科技职业学院

广东省普通高校特色创新项目

申报书(自然科学)

项 目 类 别: 特色创新项目(自然科学)

项 目 名 称: Python网络爬虫技术的研究与探索

学 科 分 类: 计算机科学技术

项目负责人: 黄仁宏

负责人手机: 15809438881

所 在 学 校: 广州华南商贸职业学院(盖章)

广东省教育厅制

二〇二一年五月

基本信息

项目信息	项目名称	Python 网络爬虫技术的研究与探索		
	项目类别	特色创新项目(自然科学)		
	研究类型	应用研究	申请金额	2(万元)
	学科一	计算机科学技术 - 计算机软件		
	学科二			
	学科三			
	计划开始日期	2021.6	计划完成日期	2023.6
	所属学校	广州华南商贸职业学院	学校类型	民办高职高专院校
	预期成果形式	论文、研究报告、专利		

合作单位	合作单位名称	联系人	联系电话	通讯地址

负责人信息	姓　名	黄仁宏	性　别	男	民　族	汉族
	出生年月	1965.10	学　历	大学本科	学　位	学士
	职　称	副高级		职　务	教研室主任	
	办公电话	020-28388100		手　机	15809438881	
	一级学科	计算机科学技术		二级学科	计算机软件	
	电子邮件	yuxin8024@qq.com		身份证号	230103196510235251	
	人才层次					
	研究专长	网络管理与网络编程				

摘要	搜索引擎(Search Engine)作为一个辅助人们检索信息的工具极大方便了用户获取信息资源，但是这些通用性搜索引擎也存在：Web 网页数目庞大，增长迅速查全率不高，过期信息较多，经常有死链接，索引更新较慢，难于找到最新信息，同义词的大量存在，查准率不高等局限性。 基于 Python 网络爬虫技术研究与实现就是为了解决上述问题，利用网络爬虫自动下载，抓取目标，获取所需要的信息的特点，对数据的收集、分析、挖掘，使企业产品更符合消费者需求，帮助企业对客户资源进行精准锁定，提供更好的推广方案，提高有效转化率，为面向主题的用户查询准备数据资源，在数据收集、数据分析、聚焦爬虫、数据聚合等多个领域有着广泛的应用。
关键字	数据分析;数据采集; Beautiful Soup 技术;信息查询

项目组成员

总数（含负责人）		高级		中级	初级	博士	硕士	学士
5		1		4	0	0	2	3
姓名	性别	出生年月	学位	职称	项目分工	工作单位		研究领域
张海霞	男	1979.12	硕士	中级	项目的研究，课题的研究	广州华南商贸职业学院		计算机软件
于平	女	1977.5	学士	中级	数据收集，报告撰写	广州华南商贸职业学院		计算机软件
蔡选强	男	1983.8	学士	中级	调研分析，论文撰写	广州华南商贸职业学院		计算机软件
张亿军	女	1983.12	硕士	中级	材料收集，调研分析	广州华南商贸职业学院		计算机软件

经费申请表

（金额单位：万元）

预算科目	创新强校工程经费	备注（计算依据与说明）
一、科研业务费	1.0000 万元	
1、测试、计算、分析	0.1000 万元	
2、会议费、差旅费	0.1000 万元	
3、出版物、文献、信息传播	0.8000 万元	
4、其他	万元	
二、试验材料费	0.0000 万元	
1、原材料、试剂、药品购置费	万元	
2、其他	万元	
三、仪器设备费	0.0000 万元	
1、购置	万元	
2、试制	万元	
四、劳务费	万元	
五、其他费用	1.0000 万元	
1、课程建设	0.5000 万元	
2、资料打印	0.1000 万元	
3、实用新型专利	0.4000 万元	
4、	万元	
合计	2.0000	
与本项目有关的其他经费来源	其他计划资助经费	万元
	其他经费资助	万元
	其他经费合计	0.0000 万元

进度计划

序号	起止时间	阶段性研究工作进展	阶段性目标
1	2021.7-2021.9	1.本课题前期调研分析和可行分析。 2.成立项目工作组，制定详细研究方案 3.举行开题报告会，展开项目研究	1.课题前期调研分析和可行分析; 2.编写基于《基于 Python 的网络爬虫技术研究与应用》开题报告。
2	2021.10-2022.9	1.对种子页面的挑选与格式进行分析 2. 抓取策略的研究与选择 3.urllib，urllib2，requests 库的研究与使用 4.常用爬与反爬机制研究（ip 限制，UA 限制 ，基于 Headers ，cookie 限制 ，验证码模拟登陆） 5.对 Beautiful Soup 解析数据工具进行深入研究 6.实现对文档元素的查找与获取	1.爬取目标的选定和分析; 2.爬取效率与质量的提升与策略的研究 3.完成《Python 的网络爬虫技术研究与探索》-获取数据部分 4. 深入研究爬与反爬机制研究。 5. Beautiful Soup 解析数据工具和正则表达式的应用; 6.完成《Python 的网络爬虫技术研究与探索》-解析页面数据。 对数据进行采集、 分析、过滤、索引，并进行存储以便之后的查询和检索。
3	2022.10-2023.3	1.对项目整体测试 2.项目相关案例整理，完善课程资源建设	1. 完成《Python 的网络爬虫技术研究与探索》--项目测试部分; 2. 补充教学内容与课程资源，完善课程教学课件，提炼企业真实项目进行教学改造等。
4	2023.4-2023.7	进行项目收尾工作 ，整理终期成果，公开出版、发行，撰写结项报告，申请验收、结项	1. 课题的研究报告 1 篇; 2. 发表相关研究论文 2 篇; 3.《Python 的网络爬虫技术研究与探索》项目源码一份; 4. 课程资源拓展资源的建设更新和整合，课程教学相关基础实训教学案例 50 个; 5. 申请实用新型专利授权 1 项。

预期成果

论文（篇）	总数	2
	其中：CSCD 核心期刊	
	三大索引收录	
专著（部）		

研究报告（篇）		1	
专利（件）	数量（件）	申请	1
		授权	
	其中发明专利	申请	
		授权	
鉴定成果（项）			
软件登记（项）			
新产品（种）（或新装备、新药等）			
新技术（项）（或新工艺等）			
其他			

申请书正文

一、立项依据

（一）. 立项的必要性及需求分析（立项背景、目的、意义）

1.立项背景

随着网络的迅速发展，万维网成为大量信息的载体，传统的通用搜索引擎作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是，这些通用性搜索引擎也存在着一定的局限性：

（1）通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询，不能完全满足用户的要求，且不能实现自动化。

（2）不同领域、不同背景的用户往往具有不同的检索目的和需求，通过搜索引擎所返回的结果包含大量用户不关心的网页，过期信息较多，经常有死链接，信息不精准

（3）纯手工搜索、存在着工作量大、重复率高、效率低、时效性差和且广告众多等弊端；

（4）通用搜索引擎图片、数据库、音频、视频多媒体等不能很好地发现和获取，通用搜索引擎难以为用户提供有效的个性化服务。

为了解决上述问题，定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，它根据既定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息，利用相关网络爬虫技术与算法实现数据自动化采集与结构化存储，从商业价值为说，可为面向主题的用户查询准备数据，使用户得到的数据更精准，更加多样化，为用户的提供有效的个性化服务，体现其商业价值资源，从就业的角度来说，爬虫工程师目前来说属于紧缺人才，并且薪资待遇普遍较高所以，深层次地掌握这门技术，对于学生就业来说，是非常有利的。

2.研究目的与意义

网络蜘蛛即Web Spider，是一个很形象的名字。把互联网比喻成一个蜘蛛网，那么Spider就是在网上爬来爬去的蜘蛛。网络蜘蛛是通过网页的链接地址来寻找网页，从网站某一个页面（通常是首页）开始，读取网页的内容，找到在网页中的其它链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去，直到把这个网 站所有的网页都抓取完为止。如果把整个互联网当成一个网站，那么网络蜘蛛就可以用这个原理把互联网上所有的网页都抓取下来 。

因此，网络爬虫具有以下特点：指按照一定的规则，自动地抓取，自动下载，实现自动化。并模拟浏览器打开html网页，然后获取相关的数据信息，过滤分析这些代码从而得到我们要的资源 同时技持全网搜索，可以爬文献，爬图片、爬音音频，爬视频，只要你想爬，说无所不能。

因此自从大数据的概念被提出后，互联网数据成为了越来越多的科研单位进行数据挖掘的对象，利用相关网络爬虫技术与算法实现数据自动化采集与结构化存储，并利用算法进行一些归纳整理在人工智能、大数据、自动化运维、人类行为模拟、计算机视觉等领域大有可为，对企业来说商业价值是巨大的，因此，本课题研究意义：

（1）**研究爬虫技术研究,提取价值数据，体现其商业价值**

本课题研究爬虫相关策略和算法对数据据进行采集，抽取，将数据做成标准化的数据，然后进行数据分析、挖掘，使用户得到的数据更精准，更加多样化，为用户的提供有效的个性化服务）得到数据，体现其商业价值。

比如说有一家电器售卖公司，为了生存下去，它需要实时了解对手的状况，改进自己的产品，然而我们不可能从对手的网站上进行一遍一遍的复制黏贴，且不说耗费时间之多，而且还极可能一不小心复制错一个数字或是一个数据,导致极大的错误,因此我们利用爬虫技术可以帮助企业：

（a）了解市场信息，使企业产品更符合消费者需求;

（b）帮助企业降低生产成本，提高经济效益，增强市场竞争力;

（c）提供更好的推广方案，提高有效转化率;

（d）数据来分析用户行为，来分析自己产品的不足之处;

并IT、金融、交通、零售、证券制造等多个领域都有广泛的应用。

零售行业	股权研究
• 监控竞争对手和市场价格，以完善定价策略，在竞争中获取优势地位 • 从制造网站批量下载产品图像和说明，以不断补充在线市场 • 监控客户对产品、服务和品牌的看法 • 对比竞争对手的产品，完善自我产品	• 利用来自行业博客、社交媒体、新闻聚合网站等非传统来源的数据，改善投资模式 • 从各种市场信息源收集数据，了解当前股市情况 • 从公开可用的财务报表中提取数据，以简化研究
数据科学 利用海量的网络数据收集文本和图像，训练机器学习模型，例如： • 自动驾驶汽车 • 疾病诊断 • 确定贷款风险 • 评估招聘或合作的风险	**销售和营销** • 优化目标客户定位工作、生成潜在客户线索、创建数据驱动的内容、SEO优化 等 • 挖掘 Web 数据，以开发差异化内容营销资产 • 为营销工作制定高质量的潜在客户清单

（2）为大数据研究提供重要的数据源

本科题研究爬虫相关策略和算法，对爬取的数据进行收集存储，可为后续的数据分析、数据挖掘、人工智能、机器学习等提供重要的数据源。

（二）国内外研究现状、水平和发展趋势分析

1.国内外研究现状、水平

网络快速发展的今天，互联网承载着海量的信息，能够准确快速的提取我们所需要的信息是现在的挑战。传统的搜索引擎有Yahoo，Google，百度等，这些检索信息的工具是人们每天访问互联网的必经之路。但是，这些传统性搜索引擎存在着局限性，它不能全面的准确的找到所需要的信息，也会使一些和需求无关的内容一起搜索到。严重的降低了使用这些信息的效率，所以说提高检索信息的速度和质量是一个专业搜索引擎主要的研究内容。

搜索引擎主要是对用户要求的信息进行自动信息搜集，这个功能共分为两种：一种是定期搜索，即每隔一段时间搜索引擎主动派出“Spider”程序，目的是对一定IP地址范围内的互联网站进行检索，如果一旦发现新的网站，它会自动提取网站的信息和网址加入自己的数据库；另一种是提交网站搜索，即网站拥有者主动向搜索引擎提交网址，搜索引擎在一定时间内定向向你的网站派出蜘蛛程序，扫描你的网站并将有关信息存入数据库，以备用户查询。

如果用户以关键词查询所需要的信息时，搜索引擎会在数据库中进行搜寻，如果找到与用户要求内容相匹配的网站时，搜索引擎通常根据网页中关键词的匹配程度，出现的位置/频次，链接质量等特殊的算法计算出各网页的相关度及排名等级，然后根据关联度高低，按顺序将用户所需要的内容反馈给用户。

（1）搜索引擎分类

搜索引擎按其工作方式可分为三种，分别是全文搜索引擎，目录索引类搜索引擎[1]和元搜索引擎。

（a）全文搜索引擎

全文搜索引擎是最常用搜索引擎，大家最熟悉的就是国外的代表Google，和国内的代表

百度。它们通常都是提取各个网站的网页文字存放在建立的数据库中，检索与用户查询条件匹配的相关记录，然后按其自己设定的排列顺序将结果返回给用户。

从搜索结果来源的角度，全文搜索引擎又可细分为两种，一种是拥有自己的检索程序，它们拥有自己的网页数据库，搜索到得内容直接从自身的数据库中调用，如Google和百度；另一种则是租用其他引擎的数据库，但是，是按自定的格式排列搜索结果，如Lycos引擎。

（b）目录索引型搜索引擎

目录索引，就是将网站分类，然后存放在相应的目录里，用户在查询所需要的内容时有两种选择一种是关键词搜索，另一种是按分类目录一层一层的查找。据信息关联程度排列，只不过其中人为因素要多一些。如果按分层目录查找，某一目录中网站的排名则是由标题字母的先后以关键词搜索，返回的结果跟搜索引擎一样，也是按自定顺序决定。目录索引只能说有搜索功能，但仅仅是按目录分类的网站链接列表。用户完全可以不用进行关键词查询，仅靠分类目录也可找到需要的信息。目录索引型搜索引擎中最具代表性的是Yahoo（雅虎）。其他著名的还有Look Smart、About等。国内的搜狐、新浪、网易搜索也都属于这一类。

（c）元搜索引擎

当用户在进行查询时，元搜索引擎可以同时在其他多个引擎上进行搜索，将检索结果进行统一处理，并将结果以统一的格式返回给用户。正因为如此，这类搜索引擎的优点是返回结果的信息量更全面，但是缺点就是无用的信息太多不能准确的找到用户需要的结果。具有代表性的元搜索引擎有Dogpile、InfoSpace、Vivisimo等，中文元搜索引擎中著名的有搜星搜索引擎。在搜索结果排列方面，不同的元搜索引擎有不同的结果排列的方式。如Dogpile，就直接按来源引擎排列搜索结果，如Vivisimo，是按自定的规则将结果重新进行排列。

（2）通用网络爬虫和聚焦网络爬虫的工作原理

网络爬虫是搜索引擎的重要组成部分，它是一个自动提取网页的程序，为搜索引擎从网上下载网页。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。论文发表。与传统爬虫相比，聚焦爬虫的工作流程则较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存起来，进行一定的分析、过滤，并建立索引，为了方便之后的查询和检索。

（3）网络爬虫的搜索策略

（a）IP 地址搜索策略

IP地址搜索策略是先给爬虫一个起始的IP地址,然后根据IP地址以递增的方式搜索本IP地址段后的每一个地址中的文档，它完全不考虑各文档中指向其它Web 站点的超级链接地址。这种搜索策略的优点是搜索比较全面，因此能够发现那些没被其它文档引用的新文档的信息源；但是缺点是不适合大规模搜索。

（b）深度优先搜索策略

深度优先搜索是一种在开发爬虫早期使用较多的方法。它的目的是要达到被搜索结构的叶结点(即那些不包含任何超链的HTML文件）。例如，在一个HTML文件中，当一个超链被选择后，被链接的HTML文件将执行深度优先搜索，也就是说在搜索其余的超链结果之前必须先完整地搜索单独的一条链。深度优先搜索沿着HTML文件上的超链走到不能再深入为止，然后返回到某一个HTML文件，再继续选择该HTML文件中的其他超链。当不再有其他超链可选择时，说明搜索已经结束。

（c）宽度优先搜索策略

宽度优先搜索的过程是先搜索完一个Web 页面中所有的超级链接，然后再继续搜索下一层，直到底层为止。例如，一个HTML 文件中有三个超链,选择其中之一并处理相应的HTML文件，然后不再选择第二个HTML文件中的任何超链，而是返回并选择第二个超链，处理相应的HTML文件，再返回，选择第三个超链并处理相应的HTML文件。当一层上的所有超链都已被选择过，就可以开始在刚才处理过的HIML 文件中搜索其余的超链。宽度优先搜索策略的优点：一个是保证了对浅层的优先处理，当遇到一个无穷尽的深层分支时，不会导致陷进WWW 中的深层文档中出现出不来的情况发生；另一个是它能在两个HTML文件之间找到最短路径。宽度优先搜索策略通常是实现爬虫的最佳策略,因为它容易实现，而且具备大多数期望的功能。论文发表。但是如果要遍历一个指定的站点或者深层嵌套的HTML文件集，用宽度优先搜索策略则需要花费比较长的时间才能到达深层的HTML文件。

2.发展趋势分析

自从大数据的概念被提出后，互联网数据成为了越来越多的科研单位进行数据挖掘的对象，利用相关网络爬虫技术与算法实现数据自动化采集与结构化存储，并利用算法进行一些归纳整理在人工智能、大数据、自动化运维、人类行为模拟、计算机视觉等领域大有可为，对企业来说价值是巨大的。

二、研究方案

（一）主要研究目标与研究内容

1.主要研究目标

（1）爬虫技术的研究

通用搜索引擎也存在着搜索面广，内容不精准、图片，数据库，音频，视频不能很好地发现和获取、难以支持根据语义信息提出的查询及难以为用户提供有效的个性化服务，不能自能化的局限性,因此本课题主要研究利用爬虫技术及相关策略和算法对数据据进行采集，抽取，将数据做成标准化的数据，然后进行数据分析，得到数据的商业价值，其具体技术研究目标为：

（a）研究爬虫的【爬取策略与算法】

（b）突破爬虫【爬与反爬机制】

（c）对采集来的数据的分析、加工、存储，提取价值数据

（d）对价值数据进行分析挖掘，并使其产品化，商业业。

(2) 课程资源建设

目前《Python程序设计》虽然建设了一些课程资源，但是还不能满足教学的需要，尤其是随着技术的法则与进步，有些陈旧的资源需要更换，新技术资源要补充，本项目研究目的就是完成原有课程资源的补充和更新，资源包括课程介绍、教学大纲、教案或演示文稿、重点难点指导、和该课程实训指导书等教学活动资源；完成拓展资源的建设更新和整合，包括多媒体课件、课程章节源码等电子文档及完整的信息化教学设计方案等，促进教育教学改革，全面提升教学质量。

(3) 以市场需求为导向，紧跟产业办学

爬虫工程师目前来说属于紧缺人才，并且薪资待遇普遍较高所以，因此我们引入行业前沿技术应用，使学生的学习内容与目标工作岗位能力要求无缝对接，让毕业生满足产业需求，推动学生就业。

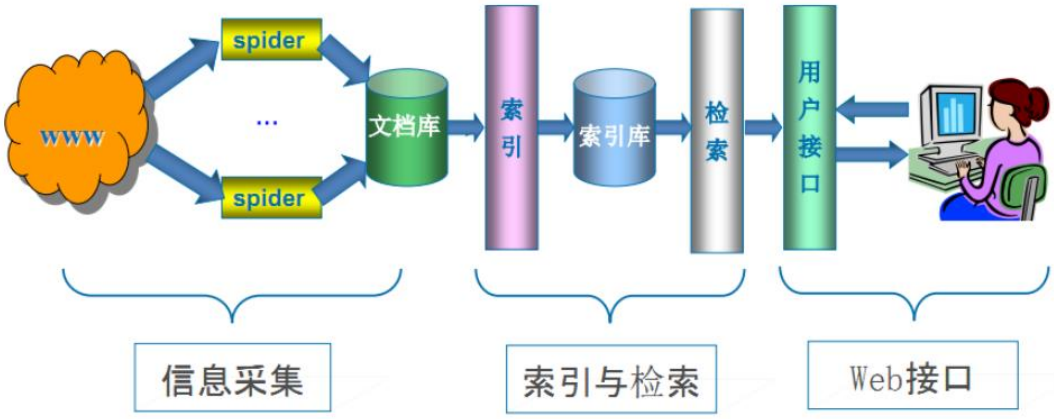
2.研究内容

（1）爬取《百度百科python词条相关词条页面--标题和简介相关1000个页面》为对象，研究网络爬虫技术和相关策略和算法，对并将获取到的数据存在本地数据库中，然后通过对收集的数据进行分析、挖掘，筛选出有价值的信息，最后结果信息可视化展示。

（2）研究提高爬虫的爬行速度，扩大数据下载量以及提升抓取信息的准确率，改进网络爬虫自身结构设计和调整策略选择来提高爬虫系统的效率，消除目前爬虫工作效率低的瓶颈，消除影响爬虫爬行效率的障碍，令爬虫达到高效且准确无误。

（3）研究在不影响服务器执行效率和不造成致命冲击的前提下提高爬行效率，对数据进行清洗去除数据重复性，提升数据的高质量，并对爬与反爬，验证码模拟登录等技术进行深入研究。

研究基本内容与工作流程如下：



其研究基本内容与工作流程如下：

（1）信息采集

（a）首先选取一部分的种子URL，将这些URL放入待抓取URL队列；

（b）取出待抓取URL，解析DNS得到主机的IP，并将URL对应的网页下载下来，存储进已下载网页库中，并且将这些URL放进已抓取URL队列。

（c）分析已抓取URL队列中的URL，分析其中的其他URL，并且将URL放入待抓取URL队列，从而进入下一个循环...

（2）索引与检索

（a）数据存储

搜索引擎通过爬虫爬取到的网页，将数据存入原始页面数据库，从而对数据进行数据清理，从而将冗余的重复的无用信息排查出去，并且对数据进行分类整理，聚类分析。

（b）数据分析与挖掘

搜索引擎将爬虫抓取回来的页面，进行各种步骤的预处理，提取文字，中文分词，消除

噪音（比如版权声明文字、导航条、广告等……），索引处理，链接关系计算，特殊文件处理等

（3）WEB接口

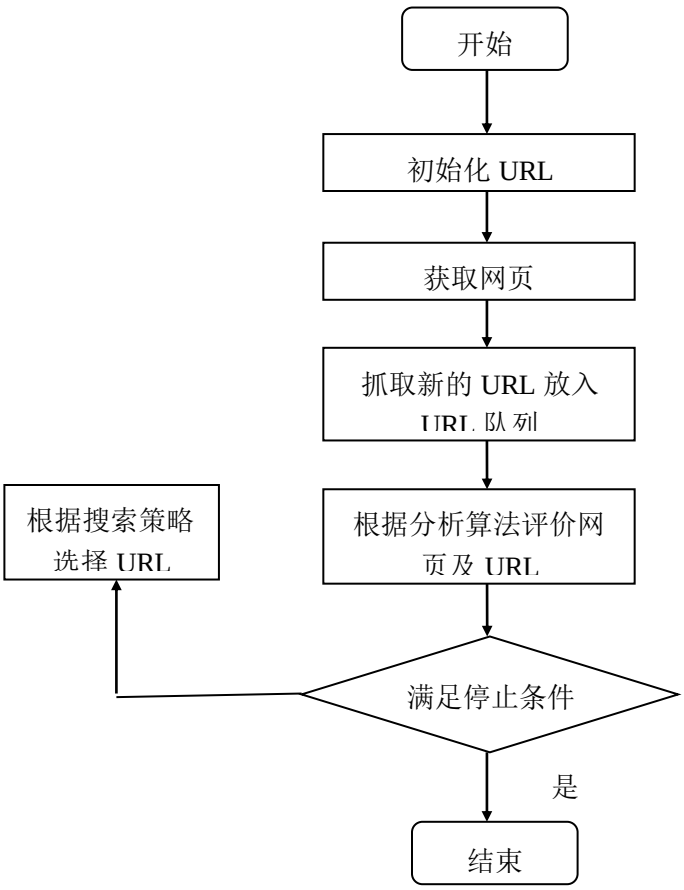
（a）对信息经过数据清洗后，筛选出来的可用的信息在网页或者APP上显示，可通过关键字等进行查询。

（b）对数据的收集、分析、挖掘，使用户得到的数据更精准，更加多样化，可为后续的大数据分析、挖掘、机器学习等提供重要的数据源。

（二）技术关键（创新点与技术难点）

1.技术关键点

聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索，如下图：



聚焦爬虫工程流程

因此聚焦爬虫研究，还需要解决三个主要关键问题：

（1）对抓取目标的描述或定义；（url,数据格式分析）

（2）对URL的搜索策略。（深度优先，广度优先）

（3）对网页或数据的分析与过滤；（数数清洗、去重、索引、提取文字、中文分词、消除噪音【比如版权声明文字、导航条、广告等…… 】）

2.技术难点

爬虫目的是自动化的从目标网页获取数据，我们对获取的数据如何剔除重复、无用的数据、如何应对反爬虫技术、如何进行数据清洗、大数据分析是本课题研究的技术难点:

(1)数据如何剔除重复无用数据

（a）将访问过的ur保存到数据库中, 把页面上爬取到的每个url存储到数据库，为了避免重复，每次存储前都要遍历查询数据库中是否已经存在当前url（即是否已经爬取过了）,若存在，则不保存，否则，保存当前url,继续保存下一条，直至结束。

（b）将访问过的ur保存到set(集合)中,只需要o(1)的代价就可以查询url

10000000*2byte*50个字符/1024/1024/1024=9G。

（c）url经过md5等方法哈希后保存到set中。

（d）用bitmap方法,将访问过的ur通过hash函数映射到某一位。

（e）bloomfilter方法对 bitmap进行改进,多重hash函数降低冲突。

(2)如何应对反爬虫技术

（a）Headers and referer 反爬机制

解决措施：通过审查元素或者开发者工具获取相应的headers 然后把相应的headers传输给python的requests（请求头），发送模拟User-Agent：通过发送模拟User-Agent来通过检验，将要发送至网站服务器的请求的 User-Agent 值伪装 成一般用户登录网站时使用 User-Agent 值这样就能很好地绕过。

（b）限制ip访问频率和次数进行反爬

解决措施：构造自己的 IP 代理池，然后每次访问时随机选择代理。

（c）UA限制(用户访问网站时的浏览器标识)

解决措施：构造自己的UA池，每次python做requests访问时随机挂上UA标识，更好地模拟浏览器行为。

（d）验证码反爬虫或者模拟登陆

解决措施：验证码识别,模拟登陆。

（e）cookie限制

网页在打开时会随机生成一个cookie，如果再次打开网页时这个cookie不存在，第三次打开仍然不存在，这就非常有可能是爬虫在工作了。

解决措施：在headers挂上相应的cookie或者根据其方法进行构造（例如从中选取几个字母进行构造）。如果过于复杂，可以考虑使用selenium模块（可以完全模拟浏览器行为）。

反爬技术线路图

(3)如何对数据进行清洗

无论是做机器学习还是做数据分析，都离不开获取数据后的第一步-数据清洗工作。据统计，数据清洗工作占据整个工作时间百分之50左右，有的甚至能达到百分之70。下面是进行数据清洗得思路流程:

数据清洗流程图

(a)数据缺失值判定

(I)热力图显示数据的缺失

```
sns.heatmap(data.isnull(),cmap="YlGnBu")
plt.show()
```

能够清楚看到哪些地方有缺失，缺失程度。

(II) 使用info()查看缺失值

```
print(data.info())
```

操作方便，执行更快，能立刻发现哪个属性存在缺失值。

(III) 使用apply()统计缺失率

```
count_missing =
```

```
data.apply(lambda :'{}%'.format(round(100*sum(x.isnull())/len(x),2)))
print(count_missing)
  plt.show()
```

通过这样的方法，可以统计出每一个属性的缺失率，百分比显示缺失率更加直观，对于缺失率高的属性，可以考虑删除。

(b)缺失值处理

(I)单行数据删除

data.dropna(inplace=True)，将存在缺失值的数据全部删除。

(II)整列属性删除

data.dropna(inplace=True,axis=1)，将存在缺失值的属性删除。

(III)均值、众数、0填充缺失值

```
#单列填充
data['下次计划还款利息'].fillna(value=data['下次计划还款利息。
'].mean(),inplace=True)
 #多列同时填充
 data1 = data[['下次计划还款本金','下次计划还款利息']].apply(lambda
 x:x.fillna(value=x.mean()))。
```

(c) 异常值检测

(I) 均值标准差异常值检测

(II)上下四中位和中位差异常值检测

(d) 异常值处理

(I)异常值删除

异常值删除操作需要两步，第一步是判断，第二步删除，当发现某一列异常值特别多的时候，我们会选择删除改属性。

(II)异常值重写

检测完异常值之后，除了删除数据之外，我们做的最多的就是重写异常值。

(4)如何进行大数据分析

要进行三种类型的数据分析：描述性分析、探索性分析以及预测性分析：

(a)描述性分析

主要是有目的去描述数据，这就要借助统计学的知识，比如基本的统计量、总体样本、各种分布等等。通过这些信息，我们可以获得对数据的初步感知，也能够得到

很多简单观察得不到的结论。所以其实描述性的分析主要需要两个部分的知识，其一是统计学的基础，其二是实现描述性的工具，用上述 Numpy 和 Pandas 的知识即可实现。

(b)探索性分析

探索性分析通常需要借助可视化的手段，利用图形化的方式，更进一步地去观看数据的分布规律，发现数据里的知识，得到更深入的结论。所谓“探索”，事实上有很多结论我们是无法提前预知的，图形则弥补了观察数据和简单统计的不足。

(c)预测性的数据分析

预测性的数据分析主要用于预测未来的数据，比如根据历史销售数据预测未来某段时间的销售情况，比如通过用户数据预测未来用户的行为.....

3.创新点

爬虫一般都是一个URL爬取完成再进行下一个，有多个URL的时候是用for循环实现对多个URL的爬取。几十个上百个URL勉强还能凑活需获取的数据为百万级别数据效率低，因此提升爬取效率本课题做了如下创新：

（1）协程。采用协程，让多个爬虫一起工作，可以大幅度提高效率。

（2）多进程。使用CPU的多个核，使用几个核就能提高几倍。

（3）多线程。将任务分成多个，并发（交替）的执行。

（4）分布式爬虫。让多个设备去跑同一个项目，效率也能大幅提升。

（5）打包技术。可以将python文件打包成可执行的exe文件，让其在后台执行即可。

（三）拟采取的研究方法、技术路线、实验方案及可行性分析

1. 研究方法

（1）项目的初期研究采用调查法、文献资料法、定性分析法

（a）调查法：系统地搜集本课题的相关材料并对其分析研究，为本课题提供技术支持。

（b）文献研究：对本课题进行搜集分析相关文献资料，并进行的研究，为本研究提供理论依据。

（c）定性分析法：对本课题中获得的各种材料运用归纳和总结，进行思维加工，从而能去粗取精、去伪存真、使之系统化、理论化。

（2）项目实施过程中采用实验研究法、测验法

（a）实验研究法：在本课题中对爬取策略和算法从已有的理论和经验出发，提出设计，然后通过在实践中实施、验证、修正，从而得到研究结果。

（b）测验法：对系统进行黑盒测试、白盒测试，发现软件程序中的错误、减少程序BUG，使程序符合设计要求。

2．技术路线

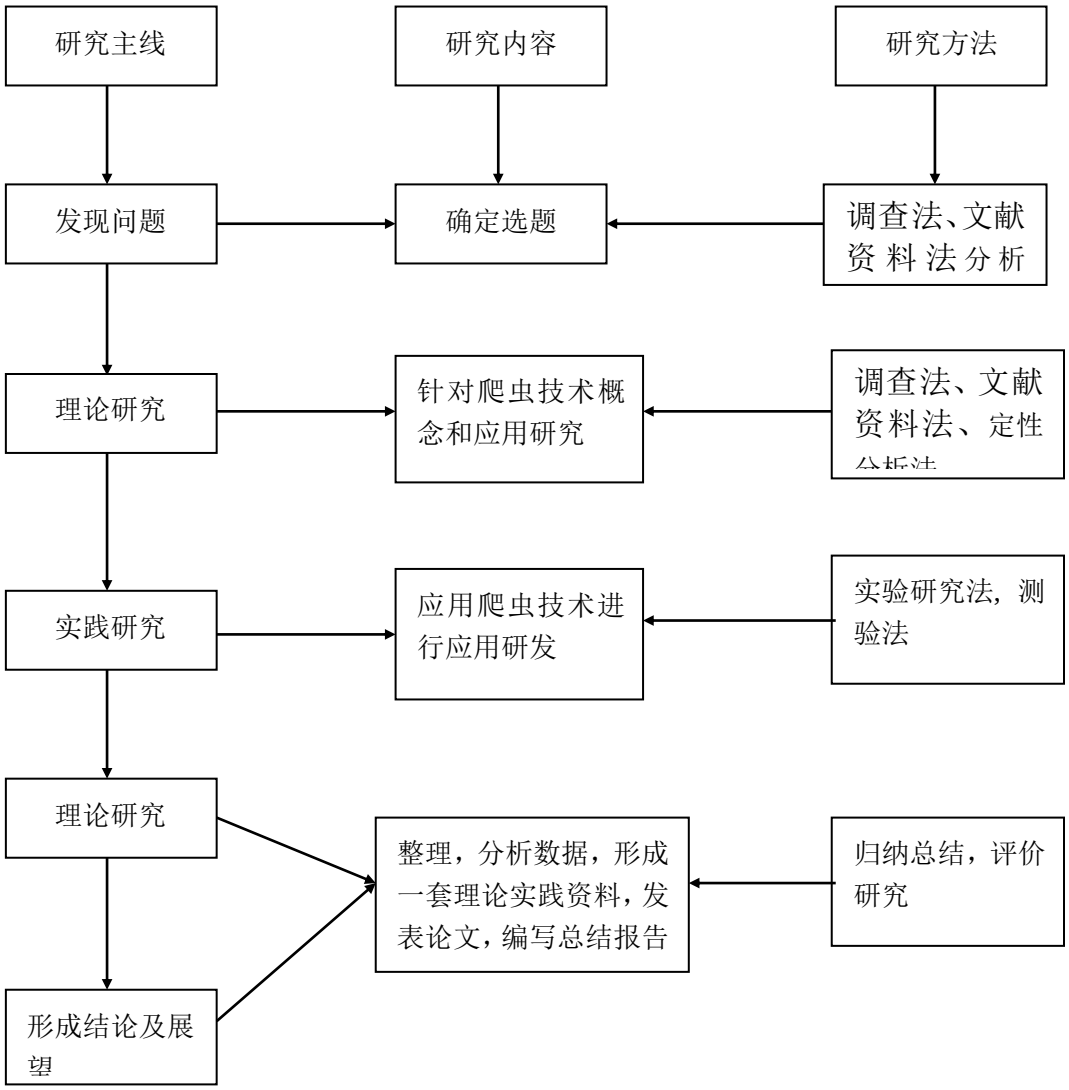
本项目研究的基本思路是：

（1）首先提出课题要解决的问题为突破口

（2）以解决问题为主线进行理论和实践研究

（3）在研究过程中根据研究的进度采取不同的研究方法来指导研究内容

（4）最后将研究成果发表论文，编写研究总结报告。



技术线路图

3．实验方案

（1）实验对象：爬取百度百科python词条相关词条页面--标题和简介相关1000个页面数据，

其入口面https://baike.baidu.com/view/21087.html

（2）实验目的：对指定的目标进行定向爬取，研究爬取策略和算法，对爬取与反爬机制进行研究，对所有被爬虫抓取的网页，进行一定的分析、过滤，并建立索引，并存储到数据，以便之后的查询和检索

（3）实验步骤

（a）获取初始的URL

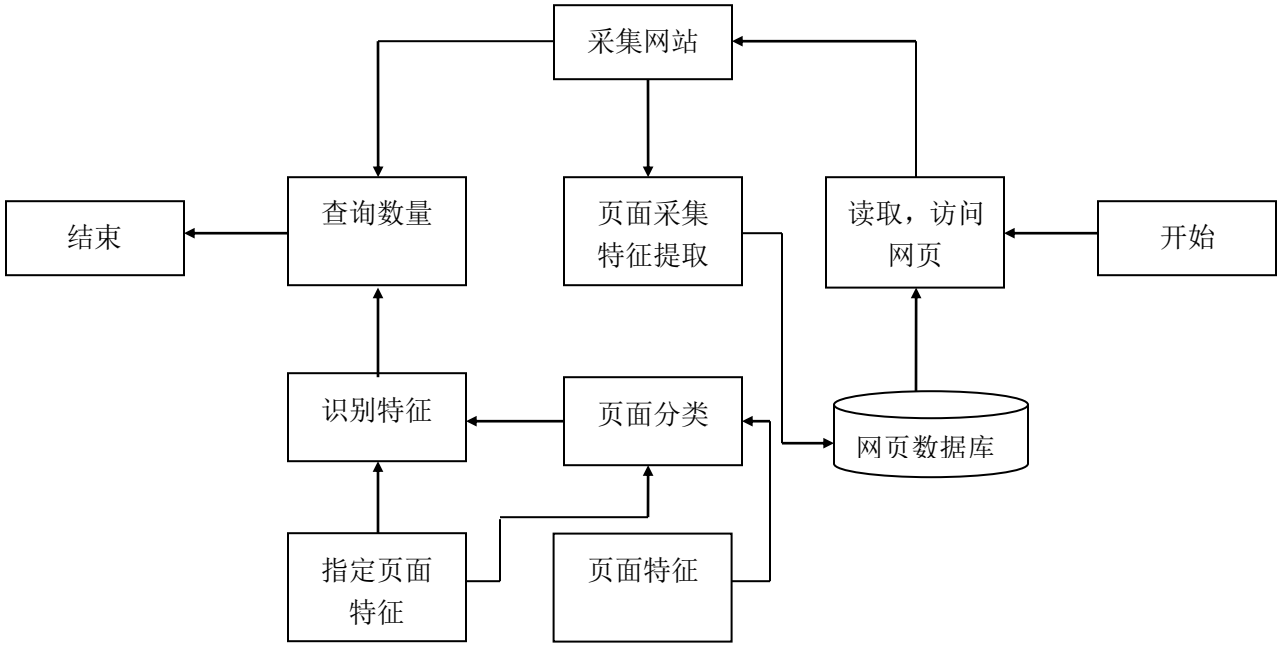
（b）根据初始的URL爬取网页，并获得新的URL

（c）从新的URL中过滤掉与爬取目标无关的链接。因为聚焦网络爬虫对网页的抓取是有目的性，所以与目标无关的网页将会被过滤掉。同时，也需要将已爬取的URL地址存放到一个列表中，用于去重和判断爬取的进程

（d）将过滤后的链接放到URL队列中

从URL队列中，根据搜索算法 ，确当URL的优先级，并确定下一步要抓取的URL地址。在通用网络爬虫中，下一步爬取那些URL，是不太重要的，但是在聚焦网络爬虫中，由于其具有目的性，故而下一步爬取哪些URL地址相对来说是比较重要的。对于聚焦网络爬虫来说，不同的爬取顺序，可能导致爬虫的执行效率不同，所以，我们需要依据搜索策略来确定下一步需要爬取那些URL地址

（e）从下一步要爬取的URL地址中，读取新的URL，然后依据新的URL地址爬取网页，并重复上述爬取的过程满足系统中设置的停止条件时，或无法获取新的URL地址时，停止爬行



实验流程图

4.可行性分析

（1）研究设计合理：

本课题研究小组在文献分析和前期调研的基础上，对该课题进行充分了论证与可行性分析及大量前期工作基础，最后确定主题，因此课题设计合理。

（2）研究方法适当

本课题的研究主要是通过对文献的研究和项目实现，并没有太大的经济花消，学校领导对本项目的研究特别重视，如果项目申报成功，学校将在研究经费上给予全力支持。项目组所有成员平时工作认真负责，有着强烈的事业心和责任心。在理论和实践教学上都具有丰富的经验，科研水平高在经济上可行。

（3）研究基础扎实

我们对文献资料和技术资料进行了深入的研究，为课题理论和技术提供支持。研究路线合理，关键技术成熟，可在规定的期限内能结项。

（四）技术、经济效益及风险分析

1．技术分析

网络爬虫是目前搜索引擎的重要组成部分，它的基本原则是在不影响服务器执行效率和不造成致命冲击的前提下提高爬行效率，爬虫需要在单位时间内获得尽可能多的高质量页面，这是它面临的困难之一，为了提高爬行速度，Web爬虫通常以"并行爬行"的方式工作，这也带来了新的问题：

（1）可重复性(并行运行的爬虫或爬行线程同时运行时，增加了重复页面)；

（2）质量问题(并行运行时，每个爬虫或爬行线程只能获取部分页面，导致页面质量下降)；

（3）通信带宽的成本(并行运行时，各个爬虫或爬行线程之间不可避免要进行一些通信，需要耗费一定的带宽资源)。

本课题旨在研究提高爬虫的爬行速度，扩大数据下载量以及提升抓取信息的准确率，改进网络爬虫自身结构设计和调整策略选择来提高爬虫系统的效率，消除目前爬虫工作效率低的瓶颈消除影响爬虫爬行效率的障碍，令爬虫达到高效且准确无误。

2.经济效益

（1）研究爬虫技术可以对搜索引擎的工作原理进行更深层次地了解。

（2）大数据时代需要进行数据分析，而学习爬虫之后，可以让我们方便地获取更多的数据源，从而进行更深层次更有效的数据分析，获得更多的价值。

（3）通过对爬虫的学习，可以让很多SEO从业者针对搜索引擎进行更好的优化。既然是针对搜索引擎进行优化，那么就必须了解搜索引擎的工作原理，这样在进行搜索引擎的优化时就可以有更好的针对性。

（4）目前来看，爬虫工程师还属于紧缺型人才，所以就业前景较为乐观，薪资待遇普遍较高，因此学习爬虫对于未来的发展是很有好处的。

3．风险分析

作为一种数据获取工具，网络爬虫的使用可以提升使用者的数据收集效率。但是技术的无限制使用必然带来混乱和网络秩序的崩溃，因此需要通过技术规范和法律规范的双重约束，进一步规制爬虫技术的使用范围和法律边界，防止爬虫技术被滥用侵害网络信息权利人的合法利益，因此本项目研究中拟从**技术规范和法律规范**两方面规避各种法律风险：

（1）网络数据爬取相关的技术规范

在技术规范方面，当前的网络爬取技术主要遵循“robots协议，爬取行为是否经过授权，获得对方的许可，Robots协议已经被认定构成互联网行业搜索领域内工人的商业道德，无视网站设置的robots协议而随意抓取网站内容的行为将涉嫌构成对违反诚实信用原则和商业道德的不正当竞争行为。

（2）网络数据爬取相关的法律规范

（a）数据爬取行为导致的民事侵权问题:

数据爬取行为中涉及的民事权益至少包括个人权益的个人信息权、财产权、知识产权，竞争法权益中的经营者利益、竞争秩序等，因此项目研究中需要综合民法典侵权责任编、著作权法、反不正当竞争法等法律法规对行为进行综合规制。遵守《民法典》人格权编第1038条对自然人的个人信息保护做出如下规定：“信息处理者不得泄露或者篡改其收集、存储的个人信息；未经自然人同意，不得向他人非法提供其个人信息，……信息处理者应当采取技术措施和其他必要措施，确保其收集、存储的个人信息安全，防止信息泄露、篡改、丢失”，对个人信息保护进行规制。

（b）数据爬取行为引发的刑事责任问题：

近十年来，由于数据采集规模快速增长，所采集的领域也逐渐从开放数据向商业数据、个人信息数据等敏感领域扩展，此类行为不可避免地从民事侵权行为逐渐转向犯罪行为。由于爬虫工具快速收集信息的特性，一旦开始自动运行，很容易超过相关标准，造成“情节严重”的后果。因此在项目研究应当在法律、行政法规规定的目的和范围内收集、使用数据，不得超过必要的限度。

爬虫作为一种计算机技术，具有技术中立性，爬虫技术在法律上从来没有被禁止。由于部分数据存在敏感性，如果不能甄别哪些数据是可以爬取，哪些会触及红线，就会涉及刑事处罚的风险，因此本项目中所有网络行为都要遵循《中华人民共和国网络安全法》。

（五）要达到的主要经济、技术指标

1.经济指标

本科题就是研究网络爬虫技术和相关策略和算法，对收集来的数据进行分析、挖掘，使用

户得到的数据更精准，更加多样化，可为后续的大数据分析、挖掘、机器学习等提供重要的数据源。

对于企业来说：

（1）了解市场信息，使企业产品更符合消费者需求。

（2）帮助企业降低生产成本，提高经济效益，增强市场竞争力。

（3）提供更好的推广方案，提高有效转化率 。

（4）数据来分析用户行为，来分析自己产品的不足之处。

对于教学来说：

（1）以拓展学生的编程素养，提升学生的编程能力，为后续课程打下坚实基础。

（2）对原有课程资源的补充和更新，丰富教学资源，促进教育教学改革，全面提升教学质量。

（3）以市场需求为导向，努力使学生的学习内容与目标工作岗位能力要求无缝对接，让毕业生满足产业需求，推动学生就业。

2.技术指标：

（1）是针对大数据应用，通过对海量词汇的对比，使用爬虫技术获取到目标客户关注的内容，下载到本地存储，再通过程序分析，将所需的数据提取分离出来，提供给目标客户;

（2）在不影响服务器执行效率和不造成致命冲击的前提下，提高爬虫的爬行速度，扩大数据下载量以及提升抓取信息的准确率;

（3）深入研究爬虫技术，突破爬与反爬机制;

（4）积极加强和完善课程资源建设，主要包括补充教学内容与课程资源，完善课程教学课件，提炼企业真实项目进行教学改造等;

（5）充分利用现代信息技术，创新教学管理及教学的方法与手段，提高教学管理水平和课堂教育教学质量。

（六）将提供的研究开发成果及形式

1．成果形式

（1）《Python的网络爬虫技术研究与探索》开题报告。预计项目研究初期完成。

（2）《Python的网络爬虫技术研究与探索》项目源码一份，预计项目研究中期完成。

（3）《Python的网络爬虫技术研究与探索》研究报告，预计项目研究中期完成。

（4）课程教学相关基础实训教学案例50个，预计项目研究中期完成。

（5）相关论文：

在国内外刊物上发表相关学术论文2篇。预计项目研究中后期完成。

（6）知识产权专利：

获得实用新型专利授权1项，预计项目研究中后期完成。

2．预期成果

（1）项目组成员本人或指导所任教课程班级学生参加专业竞赛获校级一等奖或获市厅级二等奖、省部级三等奖共 3 项以上

（2）开展横向课题的计划，并且与相关企业进行交流，确定意向。

（3）以项目建设为契机，促进《python 程序设计》课程改革，带动信息化教学过程探索、教育教学改革项目建设、青年教师培养等，提升专业人才培养质量。

三、研究基础

（一）与本项目相关的研究工作积累和已取得的研究工作成绩

1.已完成本课题前期调研分析和可以行分析。

2.研究相关文献，为项目奠定理论基础。

3.通过技术文章，资料的研究为项目提供了技术支持。

4.目前已对种子页面格式进行分析，制定了相应的抓取策略。

5.通过技术文章，资料的研究，对爬取与反爬有一定的个人见解。

（二）必要的场地、设备等支撑条件、组织措施及实施方案

1．必要的场地、设备等支撑条件

（1）项目组成员年龄、学历、职称、知识结构合理，管理协调，科研能力强，并且具有丰富的高等教育教学管理的经验，团队本身实力强劲，成绩斐然。

（2）项目组负责人具有深厚的专业学术背景，丰富的专业知识和敏锐的学术洞察力，超前意识和创新能力，能率领团队成员进行卓有成效的开展科研工作克服科研难题。

（3）课题组研究时间充足，本项目负责人是学校项目建设的核心成员和管理人员，研究项目建设和课程改革就是本职工作，所以能够全身心投入到本课题的研究工作中去，其他同志均为实践教学的骨干教师，有充裕的时间进行调查研究工作。

（4）资金的保障：学校领导对本项目的研究特别重视，如果项目申报成功，学校将在研究经费上给予全力支持。项目组所有成员平时工作认真负责，有着强烈的事业心和责任心。在理论和实践教学上都具有丰富的经验，科研水平高。

（5）依托学校和企业的支持，目前专业实训条件得到不断地改善，校内外实训室齐全，功能完备，能满足课题的资料设备等科研条件，完全能胜满足科研硬件条件。

2．组织措施及实施方案

成立项目领导小组，由组长统筹分配整个项目，确定目标，细化任务，并对项目的技术关键点，技术难点进行指导，新工艺改造，专利申请等，并对项目统筹整个项并对资源整合和有效管理。团队成员根据分工进行资料的收集与整理，细化各项指标，推进项目进度。具体措施如下：

（1）强化职责，提高认识，保证课题研究的实效性。

（2）加强学习，更新观念，增强教育科研意识

（3）健全科研制度，保障科研工作的顺利开展。

（4）加强师资队伍建设，不断提高教师业务素质。

（三）项目组负责人学术水平和管理能力情况，项目组主要成员的研究工作情况及在本课题中的工作分工

1.项目组负责人学术水平和管理能力情况

项目负责人具体统筹项目，省级教改项目申报，专利及软件著作权申报。具有深厚的专业学术背景，丰富的专业知识和敏锐的学术洞察力，超前意识和创新能力，在从事科学研究中能发现问题，寻找出解决问题的有效途径和方案，带领团队成员产出具亦影响的科研成果，同时具有较强的交际能力，科研团队组织、协调、管理和领导能力，在科研团队建设和管理中，能以自己的学术能力和人格魅力在团队中树立较高的威信，产生亲和力和凝聚力，率领团队成员进行卓有成效的开展科研工作克服科研难题

2.项目组主要成员的研究工作情况及在本项目中的工作分工

序号	姓名	性别	年龄	职务/职称	学历/学位	教学领域	在团队中的分工
1	黄仁宏	男	56	高级工程师	本科/学士	计算机网络	统筹项目，省级教改项目申报，专利及软件著作权申报
2	张海霞	男	41	讲师	研究生/硕士	软件工程	项目的研究，课题的研究
3	于平	女	43	讲师	本科/硕士	软件工程	数据收集，报告撰写
4	蔡选强	男	40	讲师	本科/学士	软件工程	调研分析，论文撰写
5	张亿军	女	37	讲师	研究生/硕士	软件工程	材料收集,调研分析

签字和盖章页(此页自动生成，打印后签字盖章，上传扫描件)

申请者：　　黄仁宏　　依托单位：　　广州华南商贸职业学院

项目名称：　　Python 网络爬虫技术的研究与探索

申请者承诺：

本人符合各项申报条件。本表各项内容真实、数据准确，不涉密，没有知识产权争议。如果获准立项，承诺以本表为有约束力协议，遵守有关规定，按计划认真开展研究工作，取得预期研究成果，并按时报送有关材料。若填报失实和违反规定，本人将承担全部责任。

签字：黄仁宏

项目组主要成员承诺：

本人保证有关申报内容的真实性。本人将严格遵守广东省教育厅的有关规定，切实保证研究工作时间，加强合作、信息资源共享，认真开展工作，及时向负责人报送有关材料。若个人信息失实、执行项目中违反规定，本人将承担相关责任。

编号	姓名	工作单位	分工	签名
1	张海霞	广州华南商贸职业学院	项目的研究，课题的研究	张海霞
2	于平	广州华南商贸职业学院	数据收集，报告撰写	于平
3	蔡选强	广州华南商贸职业学院	调研分析，论文撰写	蔡选强
4	张亿军	广州华南商贸职业学院	材料收集，调研分析	张亿军

依托单位和合作单位承诺

已按填报说明对申请人的资格和申请书内容进行了审核。本单位保证对研究计划实施所需要的人力、物力和工作时间等条件给予保障，严格遵守广东省教育厅有关规定，督促负责人和主要成员以及本单位科研管理部门按照广东省教育厅的规定及时报送有关材料。

	依托单位	合作单位 1	合作单位 2
单位名称	广州华南商贸职业学院（公章）	（公章）	（公章）
承诺经费	2（万元）	（万元）	（万元）
日期：	年　月　日	年　月　日	年　月　日

结项证书

项目类别：广东省高等学校特色创新项目（自然科学）

项目编号：**2021KTSCX348**

项目名称：Python 网络爬虫技术的研究与探索

负 责 人：黄仁宏

课题组成员：张海霞、于平、蔡选强、张亿军

证书编号：**2021KTSCX348_230988**

所在单位：广州华南商贸职业学院

该项目经审核，符合结题条件，准予结项。

广东省教育厅科研处

2023 年 12 月 30 日

广 东 省 教 育 厅

粤教科函〔2023〕8号

广东省教育厅关于公布2023年度普通高校认定类科研项目立项名单的通知

各有关高校:

为深入贯彻党的二十大精神，进一步提升全省高校科研创新能力，省教育厅组织开展了2023年度普通高校科研项目认定工作。经学校推荐、省教育厅组织审核，现将批准立项的2023年度普通高校认定类科研项目立项名单（见附件）下达各高校。

请各高校按照国家和省相关科研平台项目管理办法，统筹安排项目资金，督促项目承担人按照项目申请书开展研究工作，协助解决项目实施过程中遇到的困难和问题，加强项目管理和经费使用管理，确保研究项目如期完成目标任务。

附件：1.2023年度广东省普通高校特色创新类项目立项名单

2.2023年度广东省普通高校青年创新人才类项目

立项名单

广东省教育厅

2023年9月21日

（自然科学类联系人及电话：钟振原、王朕，020-37628043、020-37629319；人文社科类联系人及电话：曾俊伟、马思思，020-37627742、020-37628271）

公开方式：主动公开

校对人：马思思

2023年广东省普通高校特色创新类项目立项名单

1.自然科学类

序号	项目编号	项目名称	所属学校	负责人姓名
1	2023KTSCX001	模块化上转换基纳米颗粒自组装探究及其一体化肿瘤诊疗	中山大学	张振
2	2023KTSCX002	可见光无线通信与定位感知融合的基础理论研究	中山大学	周炳朋
3	2023KTSCX003	全球变暖和城市化下华南洪涝旱复合灾害演变机理与风险调控研究	中山大学	谭学志
4	2023KTSCX004	零功耗随机不确定网络的鲁棒通信理论与方法研究	中山大学	李兰花
5	2023KTSCX005	光滑粒子流体动力学及高性能船海数值水池技术研究	中山大学	孙鹏楠
6	2023KTSCX006	智能体复杂技能的自主学习	华南理工大学	齐雯
7	2023KTSCX007	动态光散射粒度检测方法开发与数据库建设	华南理工大学	柳青
8	2023KTSCX008	碳化硅基自适应变流器阻抗结构的设计、控制及应用	华南理工大学	邓文扬
9	2023KTSCX009	声响应电活性植入材料动态抗菌成骨研究	华南理工大学	于鹏
10	2023KTSCX010	面向高密度电子电路板的超精微缺陷检测技术研究	华南理工大学	刘艳霞
11	2023KTSCX011	甘油二酯胶体颗粒基皮克林乳液共负载体系构建与控释特性研究	暨南大学	仇超颖
12	2023KTSCX012	功能型个性化组织工程骨修复重度颌骨缺损研究	暨南大学	石海山
13	2023KTSCX013	玻纤复材固废粗纤维化回收及其增强混凝土的高值化利用机理研究	暨南大学	付兵
14	2023KTSCX014	考虑冠层叶面湿润时间异质性分布的柑橘溃疡病预警系统	华南农业大学	胡洁
15	2023KTSCX015	MCT4胞膜转位介导的乳酸外排对急性心梗后心肌损伤的保护机制	南方医科大学	李进晶
16	2023KTSCX016	基于心脏平扫的冠状动脉周围脂肪影像组学特征模型对低钙化积分患者冠状动脉斑块的临床价值	南方医科大学	梁健华
17	2023KTSCX017	关节腔注射SM04690阻断颞下颌关节骨关节炎进展的分子机制研究	南方医科大学	刘显文

398	2023KTSCX398	有限元仿真技术在铝合金细晶材料制备中的应用研究	顺德职业技术学院	皮云云
399	2023KTSCX399	基于新一代信息技术的高职实习岗位管理研究与应用	广东新安职业技术学院	杨崇
400	2023KTSCX400	增材制造合金丝材电磁高频加热熔盖成型技术研究	广东岭南职业技术学院	郑钢
401	2023KTSCX401	响应面法优化白芨多糖的提取工艺研究	广东岭南职业技术学院	李岩
402	2023KTSCX402	矮塔斜拉桥单箱多室宽幅箱梁剪力滞效应的研究	广东岭南职业技术学院	赵春齐
403	2023KTSCX403	方形茶饼自动定型软包装设备关键技术研究	广东岭南职业技术学院	叶立清
404	2023KTSCX404	5G+人工智能环境下高职新工科专业人才培养模式创新研究	广州涉外经济职业技术学院	黄勇
405	2023KTSCX405	益生菌发酵肉苁蓉多糖工艺优化及应用	广州涉外经济职业技术学院	胡明华
406	2023KTSCX406	基于数字化仿真的蓝牙耳机装配点胶保压治具设计策略与实证	广州南洋理工职业学院	刘卫东
407	2023KTSCX407	基于深度学习的网络爬虫算法研究与优化	广州华南商贸职业学院	王威
408	2023KTSCX408	荔枝树附生铁皮石斛活性成分评价	广州华立科技职业学院	蔡莉莉
409	2023KTSCX409	基于ChatGPT类人工智能技术对教学影响的研究	广州华立科技职业学院	张创基
410	2023KTSCX410	云计算环境下的可信计算技术研究	广州现代信息工程职业技术学院	黄毅
411	2023KTSCX411	新能源汽车热管理系统泵-阀联合控制研究与设计	广州松田职业学院	魏超
412	2023KTSCX412	面向智能制造领域的基于云、边、端协同应用机制研究	广州城建职业学院	苗晓培
413	2023KTSCX413	门锁自动组装设备设计与分析	广东南方职业学院	苏锡焕
414	2023KTSCX414	犹豫模糊集新的测度范式及决策应用	广东创新科技职业学院	郭志敏
415	2023KTSCX415	人工智能技术在5G直流电源电弧故障检测中的应用研究	广东创新科技职业学院	詹宝容
416	2023KTSCX416	基于深度学习的智能体育场馆灯光管理系统的设计	广东碧桂园职业学院	李国平
417	2023KTSCX417	粤港澳大湾区背景下火龙果深加工技术的研究与开发助力农业现代化发展	广东酒店管理职业技术学院	陆慧玲

广东省普通高校特色创新项目

申报书(自然科学)

项 目 类 别：特色创新项目(自然科学)

项 目 名 称：基于深度学习的网络爬虫算法研究与优化

学 科 分 类：工学 - 计算机科学与技术

项目负责人：王威

负责人手机：13952189588

所 在 学 校：广州华南商贸职业学院(盖章)

广东省教育厅制

二〇二三年四月

签字和盖章页（此页自动生成，打印后签字盖章，上传扫描件）

申请者：　　王威　　依托单位：　　广州华南商贸职业学院

项目名称：　　基于深度学习的网络爬虫算法研究与优化

申请者承诺：

本人符合各项申报条件。本表各项内容真实、数据准确，不涉密，没有知识产权争议。如果获准立项，承诺以本表为有约束力协议，遵守有关规定，按计划认真开展研究工作，取得预期研究成果，并按时报送有关材料。若填报失实和违反规定，本人将承担全部责任。

签字：王威

项目组主要成员承诺：

本人保证有关申报内容的真实性。本人将严格遵守广东省教育厅的有关规定，切实保证研究工作时间，加强合作、信息资源共享，认真开展工作，及时向负责人报送有关材料。若个人信息失实、执行项目中违反规定，本人将承担相关责任。

编号	姓名	工作单位	分工	签名
1	于平	广州华南商贸职业学院	统筹项目，论文撰写	于平
2	罗春	广州华南商贸职业学院	调研分析，报告撰写	罗春
3	张海霞	广州华南商贸职业学院	调研分析，数据收集	张海霞
4	徐胜东	广州华南商贸职业学院	材料收集，调研分析	徐胜东
5	王珂	广州华南商贸职业学院	项目研究，课题论证	王珂
6	刘永贤	广州华南商贸职业学院	调研分析，数据收集	刘永贤
7	廖莉	广州华南商贸职业学院	项目实验，报告撰写	廖莉

依托单位和合作单位承诺

已按填报说明对申请人的资格和申请书内容进行了审核。本单位保证对研究计划实施所需要的人力、物力和工作时间等条件给予保障，严格遵守广东省教育厅有关规定，督促负责人和主要成员以及本单位科研管理部门按照广东省教育厅的规定及时报送有关材料。

	依托单位	合作单位 1	合作单位 2
单位名称	广州华南商贸职业学院（公章）	（公章）	（公章）
承诺经费	2（万元）	（万元）	（万元）
日期：	2023年6月9日	年　月　日	年　月　日

课题编号
2023KTSCX407

广东省普通高校自然科学项目

开 题 报 告

课 题 名 称	基于深度学习的网络爬虫算法研究与优化
课 题 类 别	特色创新项目(自然科学)
所 属 学 科	计算机科学与技术
课题承担人	姚蔚芳
所 在 单 位	广州华南商贸职业学院

广东省教育厅科研处 制

一、开题活动简况（开题时间、地点、评议专家、参与人员等）

开题时间：2023 年 12 月 20 日

开题地点：广州华南商贸职业学院实训楼 3-810 会议室

参与人员：评审专家、相关项目负责人及重要成员、教学科研部相关人员

评议专家：

序号	姓名	工作单位	职称	组长/组员
1	张涛	广东职业技术学院	教授	组长
2	窦志铭	深圳职业技术大学	教授	组员
3	李振斌	广州华南商贸职业学院	教授	组员

二、开题报告要点（题目、内容、方法、组织、分工、进度、经费分配、预期成果等，限 5000 字，可加页）

（一）题目

基于深度学习的网络爬虫算法研究与优化

（二）内容

1、研究背景

旨在解决网络爬虫过程中，存在网络主题信息特征不明显，难以有效爬取特定领域主题信息的问题。针对这些问题，着重研究引入深度学习方法，在主题爬虫策略中对目标网页主题的判别，从技术角度将大量高维度的目标主题的网页特征进行提取，构建网页主题的判别器，进而改进主题爬虫策略，以提高主题网络爬虫算法的效率和精度，提升主题网页信息的质量，帮助特定领域的研究人员和分析人员降低信息获取成本，具体研究价值表现在如下方面：

（1）深度分析网络主题信息数据的特点以及主题爬虫的研究发展现状，提出深度学习与主题爬虫相结合的解决方法，提高对不断演变的网页信息的主题判别能力。

（2）针对已有的主题样本信息，将深度学习的优势，应用于网页的主题特征提取上，根据提取的特征，训练主题网页的判别模型，达到快速判别目标网页主题的任务。基于该判别模型，改进现有的主题爬虫策略路线，优化主题爬虫算法的效率和主题信息采集的精度。最终让用户在较短时间内，获取到更多主题相关的网页信息。

2、目标和拟解决的问题

随着网络技术的快速发展，网络信息的载体多种多样，促使互联网信息呈指数增长，给信息的发送、传递与收集带来了巨大的便利。因此针对海量的网络信息，如何提供一种精准、高效、便捷的主题爬虫算法，对网页信息实现精准采集，让需要研究和搜集相关领域信息的用户获取对自己有价值的信息，成为一个重要且有意义的研究工作。

本项目在对国内外相关研究分析基础上，基于深度学习神经网络，构建网页主题判别器，判断目标网页的主题，提出改进的主题爬虫策略，优化网络爬虫的效率。

（1）基于深度学习的主题识别研究

①现有爬虫分析

传统主题爬虫想要获取大量主题相关的网络信息有如下困难：主题识别难，主题判别难，主题词无法学习传统主题爬虫根据确定的主题词和特征文本从网络中进行目标网页的爬取，而无法从已经证实的主题爬虫相关资料中，自动提取与主题强相关的特征，再根据提取的主题特征对新一论的网络主题爬虫进行主题判别。

②主题网页判别流程

针对中文文本的主题判别，其理论核心是通过基于神经网络的语言建模方法，将已经获取的与主题相关的网页对象特征进行向量化。采用 Word2Vec 的 Skip-gram 模型和负采样，提取主题网页中与主题相关的特征，形成主题特征梯形，结合改进的 TF-IDF 计算的特征权重，作为改进神经网络判别器的初始输入。通过主题相关网页和非主题相关网页进行判别器模型的训练，最终实现对主题网页判别的过程。

③基于改进的 TF-IDF 权重计算

TF-IDF 算法是一种对主题词语在文本内容中的重要程度进行加权统计的一种方法，采用网页标签权重和词向量权重乘积的形式，统计网页文本特征与主题的相关性。考虑到待爬取的网页锚文本和标签对主题特征词的影响程度不同，可以根据不同标类型和其它网页文本结构特征，赋予网页特征词不同的权重，提升被标记特征在全文中的权重比。又考虑到使用单一标签权重，计算特征的主题贡

献程度，容易出现权重偏移的现象，采用标签权重的加权累积进行计算。

④基于 Word2Vec 的网页特征提取

主题网页正特征提取：在对网页文本的主题进行判别时，网页特征提取的结果作为深度神经网络判别模型的输入。因此从网页中提取出反映网页主题的关键特征，才能使得网页主题判别的效果较好。

主题网页负特征过滤：由于原始数据提供的主题网页样本数量太少，且都是主题相关的网页，无法通过足够的网页信息，区分与主题无关的网页特征。不可避免的将主题网页中的一部分无效特征预测为主题关联特征的现象。为了解决该问题，需要优化生成的正特征树，减少主题无关的特征数量，降低特征的数量，提高特征的精度。

⑤改进的神经网络判别器

采样改进的神经网络判别器，是基于循环神经网络进行的改进，引入 TF-IDF 权重作为输入特征的初始权重，改善特征被遗忘的问题，引入神经元边权重，改善反向传播过程中梯度消失的问题。

（2）基于深度学习的改进爬虫策略研究

①爬虫策略

现有的爬虫面向静态页面和动态页面两种类型，随着网页技术的不断发展，前端为了减少资源消耗，优化浏览器加载网页的速度，按需加载网络资源成为前端技术的重要手段。动态的网页中并不能直接从返回的 html 页面中获取到网页信息，通常需要模拟用户的操作，或模拟调用后端接口才能获取到完整的网页文本信息。模拟用户操作是指，爬虫系统在爬取网页信息时，模拟用户与网页的交互操作，直到完整的获取网页的文本信息。

②深度学习下的爬虫策略

动态页面分析：针对动态网站，在对动态的网页做爬虫设计时，分为两种思路，第一种是，模拟用户浏览网页操作，如登录，点击，翻页，输入，滚动，扫码等操作。通过分析，模拟用户的操作，将这些操作提前编排成爬虫动作，将一个个爬虫动作串接成爬虫线路，向爬虫路线中输入初始 URL 种子，设置爬虫调度策略，进行动态网页的爬虫。动态爬虫的成熟商业软件代表有集搜客、八爪鱼、火车头采集器等，相关的动态爬虫框架和工具代表有，Scrapy、Selenium 等。

爬虫动作：针对动态网站信息获取的问题，主题爬虫为了获取到更多的主题信息， 模拟人的动作行为，来访问网站，加载更多的动态页面，获取更多的主题信息。模拟人的行为，是模拟人工在浏览器上浏览网页的操作，模拟点击、翻页、滚动、输入、提交等动作。通过模拟这些用户动作，让爬虫无需获取更多的URL，就能够从网页中提取到更多的网页文本信息和详细内容。

③改进主题爬虫遍历策略

改进主题 URL 爬虫策略：针对上述问题，采用先局部广度遍历，获取一定数量的 URL，再进行下一层深度遍历，判别下一层的主题相关性。若主题相关，返回父级页面的 URL，进一步获取更多父级的 URL。否则从剩余的 URL 中进行深度遍历，获取更多的 URL，重复这样的过程，直到所有的 URL 队列被爬取消耗完毕。

主题爬虫算法流程设计：基于深度学习的主题判别算法，对优化的主题爬虫策略进行算法流程设计。算法主要分为三个关键步骤：网页爬取、主题判别和 URL 提取。网页爬取是根据待爬取的 URL，通过 http 协议获取到远程服务端的响应，客户端收到对应的网页文本信息的过程；主题判别是深度学习判别模型，用来判别客户端网页与主题的相关性；URL 提取是，从主题相关的网页中解析新的 URL，加入待爬取队列中；URL 提取和爬取的先后顺序正是爬虫策略的关键所在。

（三）方法

1、项目的初期研究采用调查法、文献资料法、定性分析法

（1）调查法：系统地搜集本课题的相关材料并对其分析研究，为本课题提供技术支持。

（2）文献研究：对本课题进行搜集分析相关文献资料，并进行的研究，为本研究提供理论依据。

（3）定性分析法：对本课题中获得的各种材料运用归纳和总结，进行思维加工，从而能去粗取精、去伪存真、使之系统化、理论化。

2、项目实施过程中采用实验研究法、测验法

（1）实验研究法：在本课题中对爬取策略和算法从已有的理论和经验出发，提出设计，然后通过在实践中实施、验证、修正，从而得到研究结果。

（2）测验法：基于深度学习神经网络，构建网页主题判别器，判断目标网

页的主题，提出改进的主题爬虫策略，优化网络爬虫的效率。

（四）组织和分工

姓名	性别	出生年月	学位	职称	项目分工
姚蔚芳	女	1964.11	学士	高级	统筹项目，报告撰写
于平	女	1976.8	学士	中级	统筹项目，论文撰写
罗春	女	1988.2	学士	中级	调研分析，报告撰写
张海霞	男	1979.12	硕士	中级	调研分析，数据收集
徐胜东	男	1994.8	硕士	中级	材料收集，调研分析
王珂	男	1988.5	硕士	副高级	项目研究，课题论证
刘永贤	男	1993.9	学士	中级	调研分析，数据收集
廖莉	女	1995.12	学士	副高级	项目实验，报告撰写

（五）进度安排

序号	起止时间	阶段性研究工作进展	阶段性目标
1	2023.10-2024.3	课题前期调研分析和可行性分析；成立项目工作组，制定详细研究方案；收集整理相关资料，展开项目研究。	编写《基于深度学习的网络爬虫算法研究与优化》开题报告
2	2024.4-2024.10	基本深度学习的主题识别研究；基本深度学习的改进爬虫策略研究。	编写研究报告，发表论文 1 篇
3	2024.10-2025.3	主题识别算法实验验证；改进爬虫策略验证。	完成试验验证，完善研究报告
4	2025.4-2025.10	进行项目收尾，整理终期成果，公开发表，撰写结项报告，申请验收结项。	结题的总结报告、发表论文 1 篇

（六）经费分配

预算科目	**支持经费（万元）**	**备注（计算依据与说明）**
一、直接经费	0.5000 万元	
业务费	0.5000 万元	会议费、差旅费、办公费

业务费	0.5000 万元	会议费、差旅费、办公费
设备费	0.0000 万元	
劳务费	0.0000 万元	
二、间接经费	1.0000 万元	资源建设、技术服务费等
三、其他	0.5000 万元	论文版面费、材料费等
合计	2.0000 万元	
与本项目有关的经费来源	“冲补强”专项资助经费	0.0000 万元
	其他政府资助	0.0000 万元
	学校支持经费	2.0000 万元
	企业支持经费	0.0000 万元
	其他（含自筹）	0.0000 万元
	合计	2.0000 万元

（七）预期成果

论文（篇）	总数	2
	其中：CSCD 和北大核心期刊	0
	三大索引收录	0
专著（部）		0
研究报告（篇）		1

课题主持人签名 [illegible]

2023 年12月26日

三、专家评议要点（侧重于对课题组汇报要点逐项进行可行性评估，并提出建议，限 800 字）

校外评审专家 1：张涛

评审意见：该项目有一定的研究基础，研究内容以互联网时代，面对海量的网络信息，如何为用户获取有价值的信息，并提出改进的主题爬虫策略。项目进度计划安排合理，科学划分建设内容，预期成果切实可行，为确保项目建设的系统性和实效性提供了有力支持；经费预算综合考虑项目需求，预算合理，保证项目顺利研究。建议进一步阐述项目研究的应用推广的价值。

该项目充分考虑了研究内容、研究方法、组织分工、进度计划、经费分配、预期成果等多个方面，具备科学性和可操作性。

校外评审专家 2：窦志铭

评审意见：项目组已经对研究背景、研究框架、研究内容，目标和拟解决的问题、基于深度学习的改进爬虫策略等研内容进行了初步的设计，覆盖了申报书的要求。项目组对研究方法、组织分工、研究进度、经费预算进行了合理安排和选择。项目的前期研究有一定研究基础。

鉴于课题研究有理论研究、有实践探索。建议课题组关注设计后的验证和知识产权取得等，争取在成果方式上更丰富。同意开题。

校内评审专家 3：李振斌

评审意见：课题组根据爬虫面向静态页面和动态页面两种类型的不同特点，瞄准深度学习下的爬虫策略、如何改进主题爬虫遍历策略、主题爬虫算法流程设计等关键技术进行研究和实践，前期准备工作较充分，研究的组织管理工作扎实，分工协作开展各项准备及后期研究规划活动，对课题要突破的重点问题和拟解决的关键问题分析比较准确，团队结构合理，实力较强，任务设计成员参与度高。建议准予开题。

评议专家组签名：

姓名	工作单位	职称	组长/组员	专家签名
张涛	广东职业技术学院	教授	组长	
窦志铭	深圳职业技术大学	教授	组员	
李振斌	广州华南商贸职业学院	教授	组员	

2023 年 12 月 20 日

四、重要变更（侧重说明对照课题申请书、根据评议专家意见所作的研究计划调整，限 1000 字，可加页）

课题主持人签名

年 月 日

五、所在单位科研管理部门意见

同意开题

科研管理部门盖章

2024年 1月5日

广东高校省级重点平台和重大科研项目

中期检查报告书

课 题 名 称	基于深度学习的网络爬虫算法研究与优化
课 题 类 别	特色创新项目(自然科学)
项 目 编 号	2023KTSCX407
课题承担人	姚蔚芳
所 在 单 位	广州华南商贸职业学院

一、研究工作进展情况（工作方案、调研计划、实施情况、拟开展的工作、存在的问题，能否按时完成研究计划、经费使用情况等）

（一）工作方案：

1.项目的初期研究采用调查法、文献资料法、定性分析法

（1）调查法：系统地搜集本课题的相关材料并对其分析研究，为本课题提供技术支持。

（2）文献研究：对本课题进行搜集分析相关文献资料，并进行的研究，为本研究提供理论依据。

（3）定性分析法：对本课题中获得的各种材料运用归纳和总结，进行思维加工，从而能去粗取精、去伪存真、使之系统化、理论化。

2.项目实施过程中采用实验研究法、测验法

（1）实验研究法：在本课题中对爬取策略和算法从已有的理论和经验出发，提出设计，然后通过在实践中实施、验证、修正，从而得到研究结果。

（2）测验法：基于深度学习神经网络，构建网页主题判别器，判断目标网页的主题，提出改进的主题爬虫策略，优化网络爬虫的效率。

3.总结阶段

（1）课题组成员总结归纳研究心得，积极撰写经验论文。

（2）对课题研究工作进行分类整理，归纳资料，总结成果，撰写研究报告。

（3）团队研究成员总结研究成果，撰写结题报告。

（二）调研计划

1.第一阶段:2023.10-2024.03

课题前期调研分析和可行性分析；成立项目工作组，制定详细研究方案；收集整理相关资料，展开项目研究。编写《基于深度学习的网络爬虫算法研究与优化》开题报告。

2.第二阶段:2024.04-2024.10

基本深度学习主题识别研究；基本深度学习改进爬虫策略研究。编写研究报告，发表论文 1 篇。

3.第三阶段:2024.10-2025.03

主题识别算法实验验证；改进爬虫策略验证。完成试验验证，完善研究报告。

4.第四阶段:2025.4-2025.10

进行项目收尾，整理终期成果，发表论文 1 篇，撰写结项报告，申请验收结项。

（三）实施情况

1.项目前期准备阶段

（1）已完成本课题前期调研分析和可行分析。

（2）已完成成立项目工作组，制定详细研究方案。

（3）已完成举行开题报告会，展开项目研究。

（4）已完成编写《基于深度学习的网络爬虫算法研究与优化》开题报告。

（5）已完成编写《基于深度学习的网络爬虫算法研究与优化》调研报告。

2.项目研究阶段

（1）实验设置

主要针对网络爬虫方法进行实验，选取了 web 的主题数据资源作为实验对象，所以数据均来源于 web。数据统计选了 100 种不同主题进行抓取，总共随机抓取了 50 万条数据。在数据抓取时，分别采用关键词匹配、多模式匹配、K 最近邻+TextRank、朴素贝叶斯和深度学习这几种算法对同一种主题进行抓取。网络抓取的目的是为了信息的快速识别和筛选，此时系统的查全率和查准率是衡量抓取方法是否有效的重要指标。为了适应网络抓取的性能要求，本文配置了表 1 的软硬件进行实操：

表 1. 实验环境参数

Hardware environment	Details
Central processing unit	Intel i7-13500P
Memory	16G
Motherboard	H87 i945
Hard drive	Maxtor
Graphics card	Intelb iris Xe
Operating system	Windows11
Compilation tools	PyCharm 2022.3
Compilation language	Python3.8
Deep learning framework	Tensorflow3.0

关键词匹配是一种简单直接的文本匹配方法，根据指定的关键词进行匹配，从而找出相关的内容。多模式匹配是指在一个文本中同时匹配多个模式，可以是多个关键词或者是复杂的正则表达式。最近邻+TextRank 是一种将 K 最近邻算法与 TextRank 算法结合起来的文本摘要方法，它首先使用 K 最近邻算法获取与指定问题相关的文本片段，然后使用 TextRank 算法对这些文本片段进行权重计算，得到最具代表性的摘要信息。朴素贝叶斯是一种基于贝叶斯定理的分类算法，它假设特征之间

相互独立，通过已知的特征来计算分类的概率。深度学习是一种机器学习方法，通过构建深层神经网络模型，学习大规模数据的表示和特征，并通过反向传播算法来进行模型参数的优化和训练。

（2）实验验证

主题识别验证分别进行权重计算、正特征提取、负特征过滤，对主题特征向量进行收敛，生成带权重的主题特征梯形。将主题相关的网页和主题特征梯形作为输入，建立网页主题识别模型，用训练集网页对主题识别模型进行训练，最终用训练好的模型判别一个新网页的主题相关程度。

对 web 的数据进行预处理主要包括，分词处理，特征提取，词频统计，去停用词等，实现降低噪声值，提高数据质量，进一步过滤特征词语，优化主题特征的表达。本文验证经过训练和调整参数后的基于深度学习的主题判别模型与其它主题爬虫在主题判别模型之间的差异性。共选取了 2000 条数据进行测试。

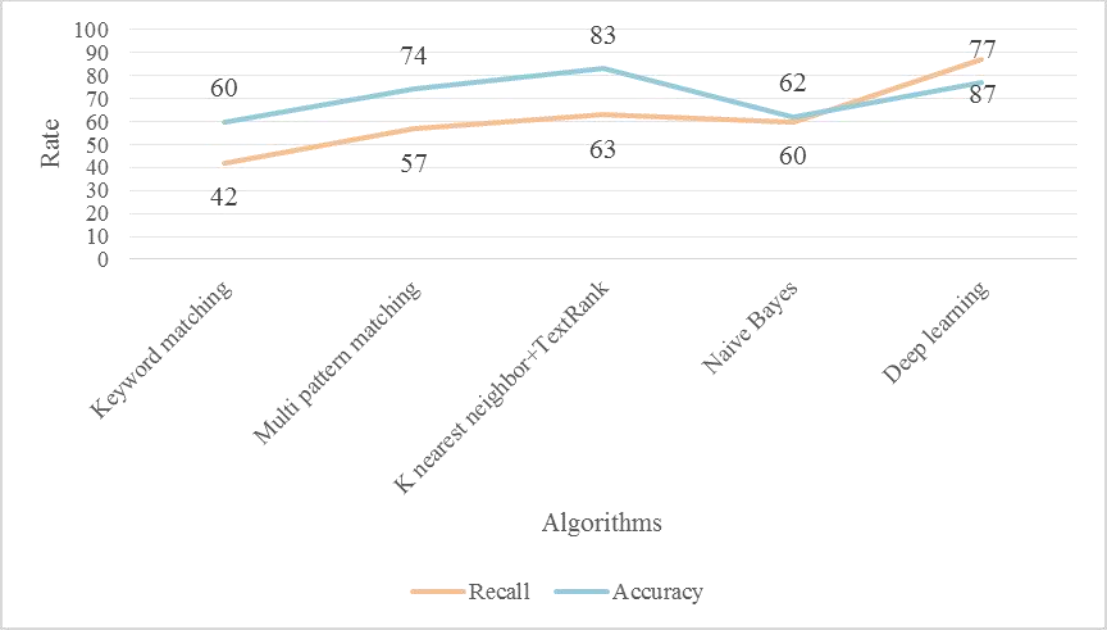


图 1. 不同算法的查全率和查准率

如图 1 所示，可以发现，在网络爬取数据时，除了深度学习外，其他几种算法查全率总是比查准率的性能差。我们可以看到，在关键词匹配算法中，其查全率和查准率数值都比较小，其中查全率时 42%，查准率时 60%。在多模式匹配中，查全率只有 57%，而查准率是 74%。在 K 最近邻+TextRank 方法下，网络抓取的查全率高于 60%，查准率达到 83%。朴素贝叶斯算法的查全率与查准率差别较小，分别为 60%和 62%。

（3）爬虫策略结果分析

广度优先级搜索从主页开始，并在每个级别按层次顺序访问页面。它可以快速识别网站的整体结构，但可能会在不太重要的页面上浪费时间。低优先级搜索跟随一条路径直到无法继续，然后跟

踪并选择另一条路径。尽管它可以更快地检测分支结构，但它可能缺少其他分支。PageRank 是谷歌提出的一种衡量网站重要性的算法。该算法将网站的含义与其他网站的链接关系相结合，对不同的网站进行排序。PageRank 更多地关注其他重要页面所引用的页面，以确定其含义。最高优先级搜索基于特定的评估函数来选择具有最有价值访问时间的页面。深度学习策略使用深度学习技术来训练神经网络模型，以预测页面的含义或相关性。该策略基于大量的训练数据和复杂的函数表示进行精确排序。

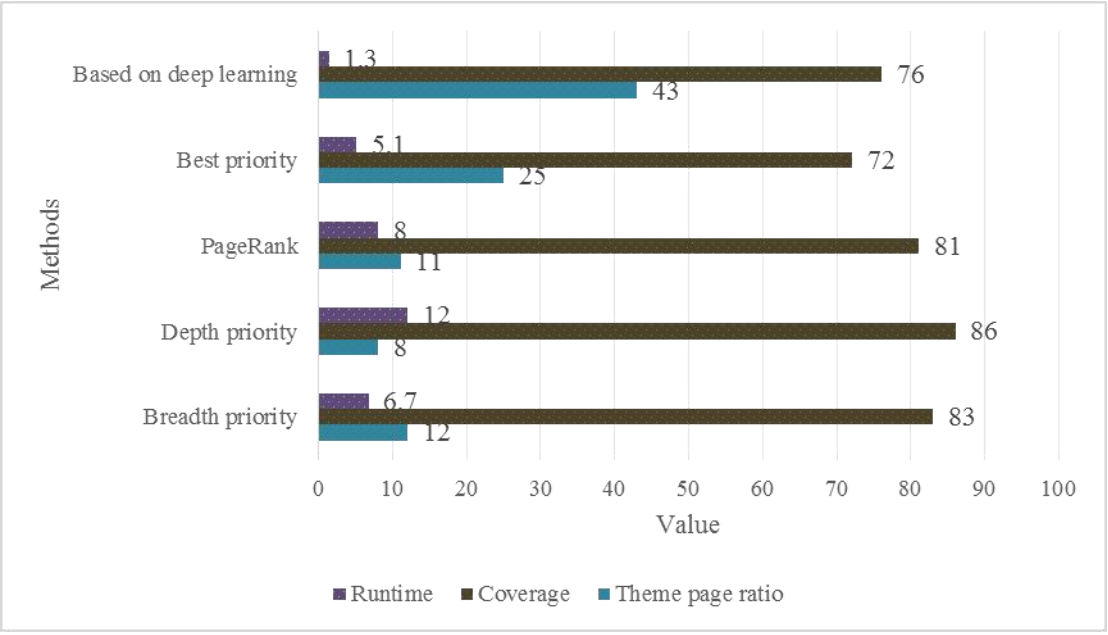


图 2. 不同爬虫策略的性能分析

如图 2 所示，可以发现，在广度优先的策略中，对主题抓取的运行时间为 6.7 小时，其覆盖率高达 83%，主题页比率为 12%。在深度优先策略中，其覆盖率最高，达到 86%，但是它的运行时间是最长的，达到 12 小时，而主题页比率最低，只有 8%。PageRank 策略的运行时间比广度优先策略长，但覆盖率较小，主题页比率也更小。最佳优先策略的主题抓取覆盖率最低，只有 72%，但主题页的比率不小，占有 25%，运行时间也缩减到 5.1 小时。而深度学习策略的覆盖率虽然不高，但是该方法下的网络主题抓取运行时间最少，只需要 1.3 小时，且其主题页比率占有 43%。

为了进一步对比不同爬虫策略，爬取目标主题网页的耗时情况，分别让五种爬虫分别爬取主题网页数到 500、1000、2000、5000 个为止。

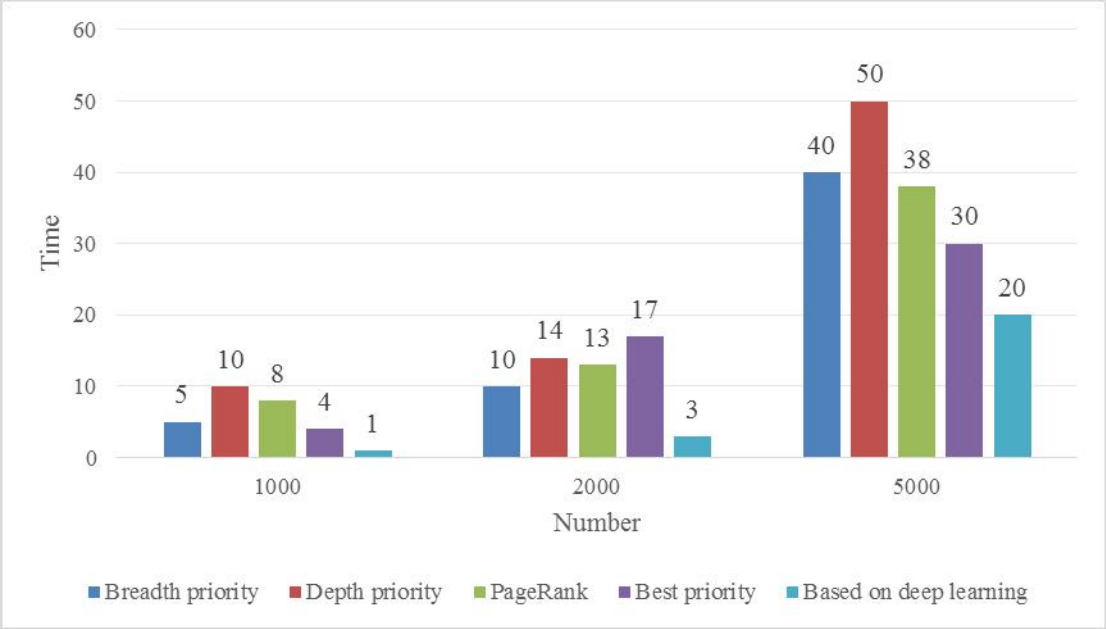


图 3. 不同爬虫策略抓取不同网页数的时间比较

如图 3 所示，在广度优先策略中，当网页数为 1000 时的运行时间需要 5 小时，网页数为 2000 时的时间需要 10 小时，当网页数为 5000 时需要 40 小时。深度优先策略比广度优先策略耗时更长，其爬取主题网页数到 1000、2000、5000 个的运行时间分别为 10h、14h 和 50h。PageRank 策略的运行时间相比于深度优先来说减少了，但幅度不大。其中，网页数为 1000 时的运行时间为 8 小时，网页数为 2000 时的时间为 13h，网页数为 5000 的耗时是 38 小时。最佳优先策略的耗时在网页数为 1000 和 5000 时比前面提到的三个策略都更低，但是在网页数达到 2000 时的用时最长。基于深度学习策略的网络爬取在不同网页数中耗时都最低，当网页数达到 5000 时其运行时间是广度优先策略的一半。

（四）拟开展的工作

1. 算法优化与创新

（1）深度学习模型改进

深度神经网络结构优化：进一步探索和优化深度神经网络的结构，如采用更深的卷积神经网络（CNN）、循环神经网络（RNN）或其变种（如 LSTM、GRU）等，以提高模型的表达能力和泛化能力。

注意力机制引入：在网络爬虫算法中引入注意力机制，使模型能够更准确地关注网页中的关键信息，提高信息提取的准确性和效率。

（2）特征提取与融合

多模态特征提取：除了传统的文本特征外，还可以探索图像、视频等多模态特征的提取和融合方法，以更全面地理解网页内容。

特征融合策略优化：研究如何有效地融合不同来源和类型的特征，提高特征表示的全面性和鲁棒性。

2.爬虫策略与效率提升

（1）智能化爬虫策略

基于强化学习的爬虫策略：利用强化学习技术，让爬虫根据历史经验和当前环境动态调整爬取策略，以实现更高效的爬取。

主题优先爬取：结合深度学习模型的主题判别能力，优先爬取与主题相关的网页，提高爬取效率和质量。

（2）并发与分布式爬虫

并发控制：优化爬虫的并发机制，合理控制并发请求的数量和频率，以减轻对目标网站的压力并提高爬取效率。

分布式部署：将爬虫系统部署在多个节点上，实现分布式爬取，进一步提高系统的可扩展性和稳定性。

3.应对反爬虫机制

（1）反爬虫策略识别

行为模拟：通过模拟人类浏览网页的行为（如点击、滚动、等待等），减少被识别为爬虫的风险。

动态网页处理：针对动态网页的反爬虫机制，研究如何有效地解析和提取动态加载的数据。

（2）加密与隐私保护

数据加密：对爬取的数据进行加密处理，确保数据在传输和存储过程中的安全性。

隐私保护：在爬取过程中尊重用户隐私和数据保护法规，避免泄露敏感信息。

4.实际应用与验证

（1）跨领域应用

将基于深度学习的网络爬虫算法应用于不同领域（如电子商务、金融、医疗等），验证其普适性和实用性。

（2）性能评估与优化

对爬虫系统的性能进行全面评估，包括爬取速度、数据质量、资源消耗等方面。

根据评估结果对算法和策略进行优化调整，以进一步提升系统的性能和稳定性。

（五）存在问题

1. 算法优化难题

模型复杂度与性能平衡：随着深度学习模型的复杂化，虽然可以提高模型的准确性和泛化能力，但同时也会增加计算复杂度和训练时间，如何在保证性能的前提下降低模型复杂度是一个挑战。

特征提取与融合：如何有效地从网页中提取出具有代表性的特征，并将这些特征有效地融合到深度学习模型中，以提高模型的判别能力，是一个持续需要优化的问题。

2. 爬虫效率问题

并发与分布式处理：随着爬取任务量的增加，如何高效地管理并发请求、合理分配资源，以及实现分布式爬取以提高整体效率，是亟待解决的问题。

动态网页处理：现代网页中越来越多的内容是通过JavaScript等脚本动态加载的，如何有效地解析和提取这些动态内容，对爬虫的效率提出了更高要求。

3. 数据质量问题

数据清洗与去重：爬取的数据中往往包含大量重复、无效或错误的信息，如何有效地进行数据清洗和去重，提高数据质量，是后续分析和利用数据的前提。

标签与标注：对于监督学习的深度学习模型，需要大量标注数据进行训练。然而，在实际应用中，高质量的标注数据往往难以获取，如何降低标注成本并提高标注质量是一个难题。

4. 网络环境适应性

反爬虫策略应对：随着网站反爬虫技术的不断发展，如何使爬虫能够有效应对各种反爬虫策略，如验证码、IP封锁等，是保持爬虫稳定运行的关键。

多源异构数据处理：互联网上的数据来源广泛、格式多样，如何设计灵活的数据处理框架以适应不同的数据源和数据格式，是爬虫系统需要解决的问题。

5. 伦理与法律问题

隐私保护：在爬取网页数据时，必须严格遵守相关法律法规和伦理规范，尊重用户隐私和数据保护要求。如何在合法合规的前提下进行数据采集和分析，是爬虫研究必须面对的问题。

知识产权：爬取的数据可能涉及版权、商标等知识产权问题，如何确保爬虫行为不侵犯他人的合法权益，是爬虫研究必须重视的方面。

针对以上问题，研究团队需要不断探索新的算法和技术手段，优化爬虫系统的设计和实现，同时加强与相关法律法规和伦理规范的衔接，确保爬虫研究的合法性和可持续性。

（六）能否按时完成研究计划

本课题研究小组在文献分析和前期调研的基础上，对该课题进行充分了论证与可行性分析及大量前期工作基础，最后确定主题，课题设计合理。学校领导对本项目的研究特别重视，在研究经费上给予全力支持。项目组所有成员平时工作认真负责，有着强烈的事业心和责任心。课题成员在理论和实践教学上都具有丰富的经验，科研水平高，为课题理论和技术提供支持。研究路线合理，关键技术成熟，因此可在规定的期限内结项。

（七）经费使用情况

本项目总投入经费为2万元，目前项目已经投入0.9150万元，以课题申报表的经费开支科目范围为依据，课题根据开支说明实报实销，报销总额在资助之内原则，具体使用情况如下表：

预算科目	**支持经费（万元）**	**备注（计算依据与说明）**
一、设备费	0.1000 万元	小额办公用品费
二、间接经费	0.5000 万元	资源建设、技术服务费等
三、其他	0.3150 万元	论文版面费
合计	0.9150 万元	

二、1—2 项代表性成果简介（基本内容、学术价值、社会影响等）

（一）课题研究基本内容及成果

1.课题研究基本内容

（1）针对网页标签结构不同，研究深度学习的网络爬虫算法

其中包含的文本权重不同的原理，提出改进的 TF-IDF 算法，加权计算不同标签中文本的权重，最终成为深度学习模型特征的权重输入。通过 Word2Vec 的 Skip-gram 正采样构建哈夫曼特征树，利用改进的 Skip-gram 进行负采样，对网页文本特征词树进行清洗，处理和归一化后，生成主题网页的词特征梯形。将词特征梯形和 TF-IDF 特征权重，作为循环神经网络的输入，构建特征之间的关联关系，训练调整模型，使模型通过识别网页特征，达到区分网页主题的目的。

（2）针对提高爬虫效率，减少对主题无关网页 URL 获取和解析时间

在基于深度学习的主题判别模型的基础上，优化爬虫策略，采用改进的主题 URL 爬虫策略。通过判断父页面的子页面的主题相关性，来决定是否扩大广度遍历，以获取父页面的更多同级 URL，减少对无效同级 URL 的获取，最大程度上采集与主题相关的 URL，提升单位时间内爬虫效率。

（3）针对不同的主题网页进行模型训练

对网页的主题进行判别和网页的爬取，将对模型参数和主题特征存入数据库中。从网页中解析，提取待采集的 URL 队列，将采集的网页存储到队列中，形成良好的爬虫流程，保障数据流的扭转和爬虫程序的高效运作。

2.课题研究成果

（1）《基于深度学习的 web 网络爬虫算法优化研究》录稿通知。

（2）《基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用》用稿通知书。

（3）《Python 网络爬虫技术的研究与探索》项目结题证书。

（4）《软件技术专业教学资源库》项目结题证书。

（5）《粤嵌通信-计算机应用技术专业大学生校外实践教学基地》项目结题证书。

（6）《基于深度学习的网络爬虫算法研究与优化》调研报告。

（二）学术价值

1.提升爬虫效率与准确性

精准采集：传统爬虫在网页主题判别方面存在局限性，如网页主题判别与文本上下文关联性差、少量固定特征词难以适应主题随时间变化等问题。通过深度学习算法，如结合 TF-IDF 和 Word2Vec 特征提取，构建网页主题判别模型，可以显著提升网页主题判别的准确性和效率，从而实现更精准

的信息采集。

优化策略：基于深度学习的网络爬虫可以模拟人对主题网页的发现行为，结合主题判别模型，实现改进的爬虫策略，如广度遍历爬虫策略和深度遍历爬虫策略的结合，优先爬取主题相关的 URL，从而提升网络爬虫的整体效率。

2.应对复杂网络环境

自适应能力：随着网络技术的快速发展，网页结构和内容日益复杂多样。基于深度学习的网络爬虫算法能够自适应地学习和识别不同网页的特征，从而有效应对复杂多变的网络环境。

处理海量数据：深度学习算法在处理海量数据方面具有显著优势。网络爬虫在爬取过程中需要处理大量的网页数据，基于深度学习的算法能够高效地处理这些数据，提高爬虫的稳定性和可靠性。

3.推动相关学科发展

促进交叉学科研究：网络爬虫算法的研究与优化涉及计算机科学、人工智能、数据挖掘等多个学科领域。基于深度学习的网络爬虫算法研究不仅推动了这些学科的发展，还促进了它们之间的交叉融合。

丰富研究内容：深度学习技术的引入为网络爬虫算法的研究提供了新的思路和方法，丰富了研究内容，推动了相关理论的完善和创新。

4.实际应用价值

助力信息检索与数据分析：基于深度学习的网络爬虫算法能够更精准地采集互联网上的信息，为信息检索和数据分析提供丰富的数据源，从而推动相关领域的发展。

支持决策制定：在商业、金融、医疗等领域，基于深度学习的网络爬虫算法可以实时采集和分析大量数据，为决策者提供有力的数据支持，帮助他们做出更加明智的决策。

综上所述，基于深度学习的网络爬虫算法研究与优化研究具有重要的学术价值和实践意义。它不仅提升了网络爬虫的性能和效率，还推动了相关学科的发展和创新，为信息检索、数据分析等领域的发展提供了有力支持。

（三）社会影响

1.信息获取与利用的高效性

精准信息采集：深度学习技术的应用使得网络爬虫能够更精准地识别并采集目标信息，减少了无效数据的抓取，提高了数据的质量和利用率。这对于企业和研究机构来说，意味着能够更快地获取到有价值的信息，支持决策制定和业务发展。

自动化处理：深度学习算法能够自动处理和分析海量数据，减轻了人工处理的负担，提高了工

作效率。这不仅降低了人力成本，还减少了人为错误的可能性。

2.推动技术创新与发展

算法优化与创新：基于深度学习的网络爬虫算法研究促进了算法的优化和创新，推动了人工智能技术的进一步发展。通过不断的研究和实践，研究人员能够发现新的算法模型和技术方法，提高网络爬虫的性能和效率。

跨学科融合：网络爬虫算法的研究涉及计算机科学、人工智能、数据挖掘等多个学科领域，推动了这些学科的交叉融合和创新发展。同时，也为其他领域的研究提供了有力的技术支持和参考。

3.社会经济效益

商业应用：在电子商务、金融、医疗等领域，基于深度学习的网络爬虫算法能够实时抓取和分析市场数据、用户行为等信息，为商家提供精准的市场分析和用户画像，支持精准营销和个性化推荐。这有助于提升企业的竞争力和盈利能力，促进经济发展。

公共服务：在政府部门、教育机构等公共服务领域，网络爬虫算法可以用于舆情监测、信息公开等方面。通过实时抓取和分析互联网上的信息，为政府部门提供决策支持，提升公共服务的效率和质量。

4.面临的挑战与应对策略

数据隐私与安全：随着网络爬虫技术的广泛应用，数据隐私和安全问题日益凸显。为了保障用户隐私和数据安全，需要采取数据加密、访问控制等安全措施，并加强法律法规的监管和约束。

反爬虫技术：一些网站为了防止数据被恶意抓取，会采用反爬虫技术。为了应对这一挑战，网络爬虫算法需要不断优化和创新，提高识别和绕过反爬虫技术的能力。

综上所述，基于深度学习的网络爬虫算法研究与优化研究对社会产生了深远的影响。它提高了信息获取与利用的效率，推动了技术创新与发展，促进了社会经济效益的提升。然而，也面临着数据隐私与安全、反爬虫技术等挑战。因此，在未来的研究中，需要更加注重算法的安全性和隐私保护能力，同时不断优化和创新算法模型和技术方法。

科研管理部门审核意见：

同意通过中期检查。

科研管理部门（签章）

2024年9月18日

注：如项目研究工作需推迟结项时间、调整研究方向、变更重要课题组成员等重大变更事项，需另填报《广东省教育科研项目重要事项变更申请表》。

项目批准号：2023KTSCX407

广东省普通高校人文社会科学研究项目

结 项 审 批 表

项目类别(在相应序号上划O)
1．省教育厅人文社会科学研究一般项目
2．省教育厅人文社会科学研究重点项目
3．省教育厅人文社会科学重点研究基地重大项目
4．省教育厅人文社会科学研究创新团队项目
5．省优秀青年创新人才培育项目
⑥ 特色创新类项目
7．青年创新人才类项目

学 科 门 类 软件工程

课 题 名 称 基于深度学习的网络爬虫算法研究与优化

批 准 日 期 2023 年 9 月 21 日

项目负责人 姚蔚芳

所 在 学 校 广州华南商贸职业学院

填 表 日 期 2025 年 11 月 19 日

广东省教育厅制

2008 年 2 月

一、 数据表

<table>
<tr><td rowspan="3">鉴定结项成果名称</td><td colspan="8">1.论文：基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用</td></tr>
<tr><td colspan="8">2.论文：基于深度学习的动态网页上下文内容识别与搜索</td></tr>
<tr><td colspan="8">3.研究报告：基于深度学习的网络爬虫算法研究与优化研究报告</td></tr>
<tr><td>主　题　词</td><td colspan="2">主题爬虫</td><td colspan="2">深度学习</td><td colspan="2">主题判别</td><td colspan="2">爬虫策略</td></tr>
<tr><td>预期成果形式</td><td colspan="3">论文、研究报告</td><td colspan="2">最终成果形式</td><td colspan="3">论文、研究报告</td></tr>
<tr><td>计划完成时间</td><td colspan="2">2025.10 .01</td><td colspan="2">实际完成时间</td><td colspan="2">2025.10.01</td><td colspan="2">申请鉴定时间</td><td>2025.11.19</td></tr>
<tr><td>成　果　字　数</td><td colspan="3">10 千字</td><td colspan="2">报送成果套数</td><td colspan="2">2</td><td colspan="2">是否出版</td><td>是</td></tr>
<tr><td>(计划)出版时间、单位</td><td colspan="9">1.2024.08 科技资讯
2.2025.05 黑龙江科学</td></tr>
<tr><td>获　奖　情　况</td><td colspan="9">2024-2025 学年广东省职业院校技能大赛（高职组）二等奖
2024 一带一路暨金砖国家技能发展与技术创新大赛全国总决赛三等奖</td></tr>
<tr><td>转摘、引用情况</td><td colspan="9"></td></tr>
<tr><td>结　项　种　类</td><td colspan="9">A. 正常　B. 提前　C. 延期　D. 免于鉴定　E. 申请中止或撤销</td></tr>
<tr><td colspan="10">项目负责人及课题组主要成员简况</td></tr>
<tr><td rowspan="5">项目负责人</td><td>姓　名</td><td>姚蔚芳</td><td>性别</td><td>女</td><td>民　族</td><td>汉</td><td>出生日期</td><td colspan="2">1964.11</td></tr>
<tr><td>所在单位</td><td colspan="3">广州华南商贸职业学院</td><td>行政职务</td><td>无</td><td>专业职务</td><td colspan="2">教师</td></tr>
<tr><td>研究专长</td><td colspan="3">计算机应用、前端开发</td><td>学　历</td><td>本科</td><td>学　位</td><td colspan="2">学士</td></tr>
<tr><td>通讯地址</td><td colspan="5">广州市白云区钟落潭镇长腰岭长学路 300 号</td><td>邮政编码</td><td colspan="2">510650</td></tr>
<tr><td>联系电话</td><td colspan="3">18898533053 (宅)
(办)</td><td>E-mail</td><td colspan="4">1003011792@qq.com</td></tr>
<tr><td rowspan="9">课题组主要成员</td><td>姓　名</td><td colspan="4">单　位</td><td>职称</td><td colspan="3">承担任务</td></tr>
<tr><td>于平</td><td colspan="4">广州华南商贸职业学院</td><td>副教授</td><td colspan="3">统筹项目，论文撰写</td></tr>
<tr><td>罗春</td><td colspan="4">广州华南商贸职业学院</td><td>讲师</td><td colspan="3">调研分析，报告撰写</td></tr>
<tr><td>张海霞</td><td colspan="4">广州华南商贸职业学院</td><td>副教授</td><td colspan="3">调研分析，数据收集</td></tr>
<tr><td>徐胜东</td><td colspan="4">广州华南商贸职业学院</td><td>讲师</td><td colspan="3">材料收集，调研分析</td></tr>
<tr><td>王珂</td><td colspan="4">广州华南商贸职业学院</td><td>副教授</td><td colspan="3">项目研究，课题论证</td></tr>
<tr><td>刘永贤</td><td colspan="4">广州华南商贸职业学院</td><td>讲师</td><td colspan="3">调研分析，数据收集</td></tr>
<tr><td>廖莉</td><td colspan="4">广州华南商贸职业学院</td><td>讲师</td><td colspan="3">项目实验，报告撰写</td></tr>
</table>

二、项目阶段性成果一览表

阶段性成果　注：成果形式为“研究报告”者填“使用单位”					
序号	成果名单	成果形式	作者	刊物年期、出版社和出版时间、使用单位	索引情况
1	基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用	论文	于平	科技资讯 2024.08	已收录
2	基于深度学习的动态网页上下文内容识别与搜索	论文	罗春	黑龙江科学 2025.05	已收录
3	人工智能算法开发软件	软件著作权	张海霞	国家版权局 2025.05	已获得
4	软件开发项目管理系统	软件著作权	张海霞	国家版权局 2025.05	已获得
5	数据防护采集管理系统	软件著作权	张海霞	国家版权局 2025.05	已获得
6	信息智能识别采集软件	软件著作权	张海霞	国家版权局 2025.05	已获得
7					
8					

注：（1）本表只填写标注“广东省教育厅××项目资助”字样的成果；

（2）主要阶段性成果的重要转摘、引用和应用情况可加页说明。

三、在该项目研究期间申报及承担其它项目情况

承担其它项目情况（2023 年—2025 年）				
	立项时间	项目名称	项目来源	批准经费
1	2023.09	学生增值评价视角下高职软件技术专业课程思政评价指标体系研究	广东省教育科学规划领导小组办公室	3 万
2	2023.11	《HTML5+CSS3 WEB 前端设计》课程思政示范课程	广东省教育厅	5 万
3	2023.12	基于混合式教学模式的高职课程数字化转型的实践研究	广东省教育评估协会	0.1 万
4	2023.12	“互联网+教育”背景下高职院校提升教师信息技术素养研究	广东省教育评估协会	0.1 万
5	2024.09	基于自动化测试的大数据应用能力评价系统	广东省教育评估协会	0.1 万
6	2025.11	面向多模态交互的鸿蒙分布式数据同步机制研究	广东省教育厅	2 万

四、总结报告

主要内容提示：预期计划执行情况；成果内容以及研究方法的突出特色、主要建树、创新和突破；学术价值和应用价值、社会效益和经济效益；不足和问题；尚需深入研究的问题。（3000字）

（一）预期计划执行情况

1.项目前期准备阶段

（1）已完成本课题前期调研分析和可行分析。

（2）已完成成立项目工作组，制定详细研究方案。

（3）已完成举行开题报告会，展开项目研究。

（4）已完成编写《基于深度学习的网络爬虫算法研究与优化》开题报告。

（5）已完成编写《基于深度学习的网络爬虫算法研究与优化》调研报告。

2.项目研究阶段

（1）爬取目标的选定和分析。（已完成）

（2）爬取效率与质量的提升与策略的研究。(已完成)

（3）完成实验设置-获取数据部分。(已完成)

（4）完成实验验证-进行权重计算、正特征提取、负特征过滤。(已完成)

（5）爬虫策略结果分析-广度优先级搜索与深度优先策略等对比。(已完成)

（6）算法优化与创新-深度学习模型改进、特征提取与融合。(已完成)

（7）爬虫策略与效率提升-智能化爬虫策略、并发与分布式爬虫。(已完成)

（8）应对反爬虫机制-反爬虫策略识别、加密与隐私保护。(已完成)

3.总结阶段

（1）课题组成员总结归纳研究心得，积极撰写研究论文。(已完成)

（2）对课题研究工作进行分类整理，归纳资料，整理文件，总结成果。(已完成)

（3）团队研究成员总结研究成果，采用逻辑的方法与经验筛选的方法进行总结，撰写结题报告(已完成)

（二）成果内容以及研究方法的突出特色、主要建树、创新和突破

1.成果内容

立足于当前主题爬虫信息采集的现实需求，针对网页信息难以获取，网页主题难以判别，公共网络资源信息难以采集加工、分析、利用等问题，提出一种基于深度学习的网络爬虫算法，判别网页信息的主题相关度，并对主题爬虫的策略进行优化，提高了爬虫获取主题相关网页的

效率，实现了爬取总量尽可能少的 URL 的情况下，爬取尽可能多的主题相关网页的目的。具体完成了以下工作:

（1）分析主题爬虫的网页结构，研究网页标签结构与主题特征的权重之间的关联关系，通过改进的 TT-IDF 算法对网页标签权重和对主题的贡献程度，进行加权计算，为主题判别模型，提供重要的初始化权重参数。

（2）提出用改进的 Word2Vec 特征提取方法，以 Skip-gram 模型构建哈夫曼特征树，预测主题中心词上下文特征，以负采样的方法过滤主题无关的特征。通过 TT-IDF和 Word2Vec 相结合所提取的主题特征，作为循环神经网络的带权输入特征向量，训练主题判别模型，直到能够对样本网页进行主题判别。

（3）基于强化学习和历史存储的 html 动作标签，利用正则表达式和 html 选择器，对网页常见的爬虫动作进行识别，实现对网页的自动“翻页”，“点击”，“滚动”等模拟人浏览网页的行为操作。最后为了优化爬虫策略，提高爬虫效率，在广度遍历爬虫策略和深度遍历爬虫策略的基础上，引入主题判别模型，实验最终表明该策略能提高爬虫的性能和效率。

2.研究方法的突出特色

（1）项目的初期研究采用调查法、文献资料法、定性分析法

①调查法：系统地搜集本课题的相关材料并对其分析研究，为本课题提供技术支持。

②文献研究：对本课题进行搜集分析相关文献资料，并进行的研究，为本研究提供理论依据。

③定性分析法：对本课题中获得的各种材料运用归纳和总结，进行思维加工，从而能去粗取精、去伪存真、使之系统化、理论化。

（2）项目实施过程中采用实验研究法、测验法

①实验研究法：在本课题中对爬取策略和算法从已有的理论和经验出发，提出设计，然后通过在实践中实施、验证、修正，从而得到研究结果。

②测验法：基于深度学习神经网络，构建网页主题判别器，判断目标网页的主题，提出改进的主题爬虫策略，优化网络爬虫的效率。

3.主要建树

（1）对网络爬虫技术和相关策略和算法进行了深入研究，并对收集来的数据进行分析、挖掘，使用户得到的数据更精准，更加多样化，为后续的大数据分析、挖掘、机器学习等提供重要的数据源。

（2）教师的实践能力和创新能力的提高,通过本项目研究，不仅提升了自身实践能力和创新能力，也促进了彼此的团队合作能力，教师队伍整体氛围温馨、和谐。由于项目主题主要来源于日常教学，教师们主动研究课纲、学生、教材，努力寻找课程与学生能力的契合点，使项目的成果能为我所用，真正服务于融合教学的课堂,从而推动教育教学的不断发展和进步。

（3）以市场需求为导向，紧跟产业办学,努力使学生的学习内容与目标工作岗位能力要求无缝对接，让毕业生满足产业需求，推动学生就业。

（4）课题研究成果

①基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用，科技资讯，2024.08

②基于深度学习的动态网页上下文内容识别与搜索，黑龙江科学，2025.05

③面向软件开发信息库的多源异构数据深层次挖掘方法，武汉工程职业技术学院学报，2024.03

④基于概率主题模型的软件开发数据库隐私数据泄露识别方法研究，河北软件职业技术学院学报2024.06

⑤《人工智能算法开发软件》,软件著作权,国家版权局,2025.05，登记号：2025SR0747699

⑥《软件开发项目管理系统》,软件著作权,国家版权局,2025.05，登记号：2025SR0747535

⑦ 《数据防护采集管理系统》,软件著作权,国家版权局,2025.05，登记号：2025SR0748617

⑧ 《信息智能识别采集软件》,软件著作权,国家版权局,2025.05，登记号：2025SR0748513

⑨2024-2025学年广东省职业院校技能大赛（高职组）二等奖

⑩2024一带一路暨金砖国家技能发展与技术创新大赛全国总决赛三等奖

⑪第十七届“挑战杯”广东大学生课外学术科技作品竞赛三等奖

⑫2025年广东省大学生计算机创新作品赛三等奖2项

⑬基于深度学习的网络爬虫算法研究与优化调研报告

⑭ 基于深度学习的网络爬虫算法研究与优化-研究报告

4.创新和突破

（1）深度分析网络主题信息数据的特点以及主题爬虫的研究发展现状，提出深度学习与主题爬虫相结合的解决方法，提高对不断演变的网页信息的主题判别能力。

（2）针对已有的主题样本信息，将深度学习的优势，应用于网页的主题特征提取上，根据提取的特征，训练主题网页的判别模型，达到快速判别目标网页主题的任务。基于该判别模型，改进现有的主题爬虫策略路线，优化主题爬虫算法的效率和主题信息采集的精度。最终让用户

在较短时间内，获取到更多主题相关的网页信息。

（3）应用深度学习的算法，收集各个主题特征，优化爬虫策略，以获取主题网页的数据是十分可行的，深度学习方法引入主题爬虫，使程序能够像人一样，从网页中分析、获取对自身有利的信息，并不断学习和思考网络信息特征演变的过程，并从中提取出有价值的网络开源信息，减少用户和数据研究人员从网络上搜集和获取主题网页信息的时间。

（三）学术价值和应用价值、社会效益和经济效益

1. 学术价值和应用价值

（1）提升爬虫效率与准确性

构建网页主题判别模型，可以显著提升网页主题判别的准确性和效率，从而实现更精准的信息采集。

（2）应对复杂网络环境

基于深度学习的网络爬虫算法能够自适应地学习和识别不同网页的特征，从而有效应对复杂多变的网络环境。深度学习算法在处理海量数据方面具有显著优势。

（3）推动相关学科发展

网络爬虫算法的研究与优化促进计算机科学、人工智能、数据挖掘等多个学科领域交叉学科研究。提供了新的思路和方法，丰富了研究内容，推动了相关理论的完善和创新。

（4）实际应用价值

助力信息检索与数据分析，从而推动相关领域的发展。为决策者提供有力的数据支持，帮助他们做出更加明智的决策。

2. 社会效益和经济效益

（1）信息获取与利用的高效性

深度学习技术能够更精准地识别并采集目标信息，能够更快地获取到有价值的信息，支持决策制定和业务发展。能够自动处理和分析海量数据，减轻了人工处理负担，提高了工作效率。

（2）推动技术创新与发展

网络爬虫算法的研究提高了网络爬虫的性能和效率，推动了计算机科学、人工智能、数据挖掘等学科的交叉融合和创新发展。同时，也为其他领域的研究提供了有力的技术支持和参考。

（3）社会经济效益

基于深度学习的网络爬虫算法能够实时抓取和分析市场数据、用户行为等信息，为商家提供精准的市场分析和用户画像，支持精准营销和个性化推荐。网络爬虫算法用于舆情监测、信

息公开等方面，为政府部门提供决策支持，提升公共服务的效率和质量。

（四）不足和问题

1.算法优化难题

如何保证模型复杂度与性能平衡，如何有效提取特征并融合到深度学习模型中，是持续需要优化的问题。

2.爬虫效率问题

并发与分布式处理，以及动态网页处理，对爬虫的效率提出了更高要求。

3.数据质量问题

数据清洗与去重，标签与标注，需要大量标注数据进行训练。

4.网络环境适应性

反爬虫策略应对，多源异构数据处理，是爬虫系统需要解决的问题。

5.伦理与法律问题

隐私保护，知识产权，是爬虫研究必须重视的方面。

（五）尚需深入研究的问题

基于深度学习的主题爬虫算法涉及多个领域的知识体系，许多技术还需要更深入的研究和探索，在主题判别模型训练过程中存在以下不足，需要进一步完善:

1.通过共享特征的形式，将其它用户或者机构训练好的主题认知模型进行共享，用户只用导入模型参数和特征，即可用来对相关的领域进行主题判别，减少主题爬虫，每次爬取不同主题网页时，对判别模型的训练时间。

2.当新的概念出现的时候，主题相关的样本数量过少，很难在小样本中挖掘主题的特征，在对网页进行判别和学习主题特征提取时，容易出现主题偏移，可以通过人工矫正的方式，对新出现的主题进行特征描述。

3.用高质量的主题网页或者相关文本，对模型进行训练，提高主题判别模型对主题的判别能力。

五、项目最终成果简介

主要内容与要求提示：

1.“最终成果简介”是结项的必需材料，供介绍、宣传、推广成果使用，同时要在学校学术网站公示。

2. 简介内容包括：该项目研究的目的和意义（略写）；研究成果的主要内容和重要观点、创新之处或对策建议（详写）；成果的学术价值、应用价值，以及社会影响和效益（略写）。

3. 简介内容应由项目负责人撰写；文章内容要层次清楚、观点明晰、用语准确、文风朴实，要有实质性内容，并具有整体性和系统性，不得简单排列篇章目录；成果形式为专著的5000字左右，调研报告、论文（集）等3000字左右。

（一）项目研究的目的和意义

1. 项目研究目的

（1）针对网页标签结构不同，研究深度学习的网络爬虫算法

其中包含的文本权重不同的原理，提出改进的TF-IDF算法，加权计算不同标签中文本的权重，最终成为深度学习模型特征的权重输入。通过Word2Vec的Skip-gram正采样构建哈夫曼特征树，利用改进的Skip-gram进行负采样，对网页文本特征词树进行清洗，处理和归一化后，生成主题网页的词特征梯形。将词特征梯形和TF-IDF特征权重，作为循环神经网络的输入，构建特征之间的关联关系，训练调整模型，使模型通过识别网页特征，达到区分网页主题的目的。

（2）针对提高爬虫效率，减少对主题无关网页URL获取和解析时间

在基于深度学习的主题判别模型的基础上，优化爬虫策略，采用改进的主题URL爬虫策略。通过判断父页面的子页面的主题相关性，来决定是否扩大广度遍历，以获取父页面的更多同级URL，减少对无效同级URL的获取，最大程度上采集与主题相关的URL，提升单位时间内爬虫效率。

（3）针对不同的主题网页进行模型训练

对网页的主题进行判别和网页的爬取，将对模型参数和主题特征存入数据库中。从网页中解析，提取待采集的URL队列，将采集的网页存储到队列中，形成良好的爬虫流程，保障数据流的扭转和爬虫程序的高效运作。

2. 项目研究意义

（1）深度分析网络主题信息数据的特点以及主题爬虫的研究发展现状，提出深度学习与主题爬虫相结合的解决方法，提高对不断演变的网页信息的主题判别能力。

（2）针对已有的主题样本信息，将深度学习的优势，应用于网页的主题特征提取上，根据提取的特征，训练主题网页的判别模型，达到快速判别目标网页主题的任务。基于该判别模型，改进现有的主题爬虫策略路线，优化主题爬虫算法的效率和主题信息采集的精度。最终让用户在较短时间内，获取到更多主题相关的网页信息。

（二）研究成果的主要内容和重要观点、创新之处或对策建议

1.研究成果的主要内容

（1）爬虫技术的深入研究

旨在解决网络爬虫过程中，存在网络主题信息特征不明显，难以有效爬取特定领域主题信息的问题。针对这些问题，着重研究引入深度学习方法，在主题爬虫策略中对目标网页主题的判别，从技术角度将大量高维度的目标主题的网页特征进行提取，构建网页主题的判别器，进而改进主题爬虫策略，以提高主题网络爬虫算法的效率和精度，提升主题网页信息的质量，帮助特定领域的研究人员和分析人员降低信息获取成本。

（2）教师的实践能力和创新能力的提高

通过本项目研究，不仅提升了自身实践能力和创新能力，也促进了彼此的团队合作能力，教师队伍整体氛围温馨、和谐。由于项目主题主要来源于日常教学，教师们主动研究课纲、学生、教材，努力寻找课程与学生能力的契合点，充分利用现代信息技术，创新教学管理及教学的方法与手段，提高教学管理水平和课堂教育教学质量，使项目的成果能为我所用,真正服务于融合教学的课堂,从而推动教育教学的不断发展和进步。

（3）以市场需求为导向，紧跟产业办学

努力使学生的学习内容与目标工作岗位能力要求无缝对接，让毕业生满足产业需求，推动学生就业。爬虫工程师目前来说属于紧缺人才，并且薪资待遇普遍较高所以，因此我们引入行业前沿技术应用，使学生的学习内容与目标工作岗位能力要求无缝对接，让毕业生满足产业需求，推动学生就业。

（4）课题研究成果

①基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用，科技资讯，2024.08

②基于深度学习的动态网页上下文内容识别与搜索，黑龙江科学，2025.05

③面向软件开发信息库的多源异构数据深层次挖掘方法，武汉工程职业技术学院学报，2024.03

④基于概率主题模型的软件开发数据库隐私数据泄露识别方法研究，河北软件职业技术学院学报2024.06

⑤《人工智能算法开发软件》,软件著作权,国家版权局,2025.05，登记号:2025SR0747699

⑥《软件开发项目管理系统》,软件著作权,国家版权局,2025.05，登记号:2025SR0747535

⑦《数据防护采集管理系统》,软件著作权,国家版权局,2025.05，登记号:2025SR0748617

⑧《信息智能识别采集软件》,软件著作权,国家版权局,2025.05，登记号:2025SR0748513

⑨2024-2025学年广东省职业院校技能大赛（高职组）二等奖

⑩2024一带一路暨金砖国家技能发展与技术创新大赛全国总决赛三等奖

⑪第十七届“挑战杯”广东大学生课外学术科技作品竞赛三等奖

⑫2025年广东省大学生计算机创新作品赛三等奖2项

⑬基于深度学习的网络爬虫算法研究与优化调研报告

⑭ 基于深度学习的网络爬虫算法研究与优化-研究报告

2.重要观点

（1）分析主题爬虫的网页结构，研究网页标签结构与主题特征的权重之间的关联关系，通过改进的 TT-IDF 算法对网页标签权重和对主题的贡献程度，进行加权计算，为主题判别模型，提供重要的初始化权重参数。

（2）提出用改进的 Word2Vec 特征提取方法，以 Skip-gram 模型构建哈夫曼特征树,预测主题中心词上下文特征,以负采样的方法过滤主题无关的特征。通过 TT-IDF 和 Word2Vec 相结合所提取的主题特征，作为循环神经网络的带权输入特征向量，训

练主题判别模型，直到能够对样本网页进行主题判别。

（3）基于强化学习和历史存储的 html 动作标签，利用正则表达式和 html 选择器，对网页常见的爬虫动作进行识别，实现对网页的自动“翻页”，“点击”，“滚动”等模拟人浏览网页的行为操作。最后为了优化爬虫策略，提高爬虫效率，在广度遍历爬虫策略和深度遍历爬虫策略的基础上，引入主题判别模型，实验最终表明该策略能提高爬虫的性能和效率。

3.创新之处

（1）深度分析网络主题信息数据的特点以及主题爬虫的研究发展现状，提出深度学习与主题爬虫相结合的解决方法，提高对不断演变的网页信息的主题判别能力。

（2）针对已有的主题样本信息，将深度学习的优势，应用于网页的主题特征提取上，根据提取的特征，训练主题网页的判别模型，达到快速判别目标网页主题的任务。基于该判别模型，改进现有的主题爬虫策略路线，优化主题爬虫算法的效率和主题信息采集的精度。最终让用户在较短时间内，获取到更多主题相关的网页信息。

（3）应用深度学习的算法，收集各个主题特征，优化爬虫策略，以获取主题网页的数据是十分可行的，深度学习方法引入主题爬虫，使程序能够像人一样，从网页中分析、获取对自身有利的信息，并不断学习和思考网络信息特征演变的过程，并从中提取出有价值的网络开源信息，减少用户和数据研究人员从网络上搜集和获取主题网页信息的时间。

（三）成果的学术价值、应用价值

1.提升爬虫效率与准确性

构建网页主题判别模型，可以显著提升网页主题判别的准确性和效率，从而实现更精准的信息采集。

2.应对复杂网络环境

基于深度学习的网络爬虫算法能够自适应地学习和识别不同网页的特征，从而有效应对复杂多变的网络环境。深度学习算法在处理海量数据方面具有显著优势。

3.推动相关学科发展

网络爬虫算法的研究与优化促进计算机科学、人工智能、数据挖掘等多个学科领域交叉学科研究。提供了新的思路和方法，丰富了研究内容，推动了相关理论的完善和创新。

4.实际应用价值

助力信息检索与数据分析，从而推动相关领域的发展。为决策者提供有力的数据支持，帮助他们做出更加明智的决策。

（四）社会影响和效益

1.信息获取与利用的高效性

深度学习技术能够更精准地识别并采集目标信息，能够更快地获取到有价值的信息，支持决策制定和业务发展。能够自动处理和分析海量数据，减轻了人工处理负担，提高了工作效率。

2.推动技术创新与发展

网络爬虫算法的研究提高了网络爬虫的性能和效率，推动了计算机科学、人工智能、数据挖掘等学科的交叉融合和创新发展。同时，也为其他领域的研究提供了有力的技术支持和参考。

3.社会经济效益

基于深度学习的网络爬虫算法能够实时抓取和分析市场数据、用户行为等信息，为商家提供精准的市场分析和用户画像，支持精准营销和个性化推荐。网络爬虫算法用于舆情监测、信息公开等方面，为政府部门提供决策支持，提升公共服务的效率和质量。

六、项目经费决算（单位：万元）

批准经费	2	实际到位经费	2
实际开支	2	结余经费	0
经费开支明细表			
1.资料费	0.1		
2.调研差旅费	0.2		
3.小型会议费	0.1		
4.设备费	0.1		
5.咨询费	0		
6.印刷费	0.5		
7.其他	1		
未支出经费用途： **无**			
单位财务部门意见 公 章　　负责人（签章）：何琳 2025年12月26日		**单位审计部门意见** 公 章　　负责人（签章）： 年　月　日	

注：资助经费 20 万及以上的人文社科项目需学校审计部门盖章

七、学校科研管理部门意见

主要内容提示：成果质量是否符合项目申请书（合同书）的要求，课题组的研究工作和自我管理是否符合项目管理的有关规定；对于经费决算是否同意财务意见。

成果质量符合申请书要求，同意结题。

公　章　广州华南商贸职业学院 科研管理专用章

负责人（签章）

2015年12月26日

八、教育厅科研处意见

公　章　　负责人（签章）

年　月　日

广东省教育厅科研项目重要事项变更申请表

项目名称	基于深度学习的网络爬虫算法研究与优化	批准号	2023KTSCX407
		联系方式	13952189588
项目责任人	王威	工作单位	广州华南商贸职业学院
批准立项时间	2023年 9月	原项目成果形式	论文2篇、研究报告1份
原完成时间	2025年10月	延期完成时间	

变更内容（请在方框内打“√”）：

☑变更项目责任人　☐变更项目管理单位　☐改变成果形式

☐更改项目名称　☐研究内容有重大调整　☐第一次延期

☐第二次延期　☐申请撤项　☐变更课题组成员　☐其他

变更事由：

（变更项目负责人须写明新项目负责人的性别、出生时间、职称、工作单位、联系电话、专业、研究方向及主要工作简历等情况，新项目负责人尽量为原课题组成员，并在下框中签名确认；变更课题组成员须写明在课题组中的排位，附上新课题组成员的简历，并附上原全体项目组成员签名；变更项目管理单位须由调出、调入单位签署意见。）

因原负责人个人原因离职，变更项目负责人为：姚蔚芳。

新项目负责人：姚蔚芳，女，1964 年 11 月生，高级工程师职称，省部级科技进步二等奖获得者，全国职业院校技能大赛裁判员。1989 年毕业于 211 大学合肥工业大学光电子技术专业，现任广州华南商贸职业学院专任教师，擅长计算机应用（硬件设计和软件编程）、Web 前端开发、图形图像处理等技术领域。长期在省部级央企中国兵器装备集团公司从事产品研发工作，主持了多项计算机应用项目。在从事高职教育期间，主编并出版规划教材 2 部，承担了《Web 前端开发》等多门课程的教学工作。

原全体项目组成员签名：

<table>
<tr><td colspan="3">项目负责人签章：王成
2023年10月17日</td></tr>
<tr><td>项目依托单位意见</td><td colspan="2">科研管理部门负责人签章：单常海
2023年10月17日</td></tr>
<tr><td colspan="2">转出单位意见及签章：

年 月 日</td><td>转入单位意见及签章：

年 月 日</td></tr>
<tr><td>教育厅项目管理单位意见</td><td colspan="2">教育厅项目管理单位盖章：
年 月 日</td></tr>
</table>

注：申请延期一次最多不得超过1年，一个项目申请延期最多不得超过2次。