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。

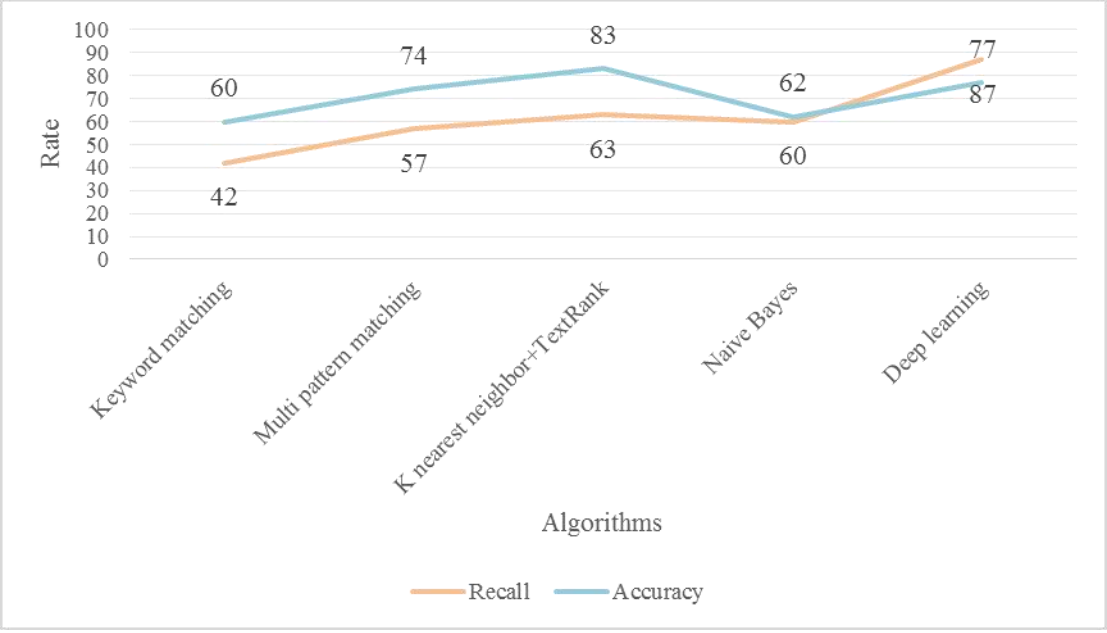

图 1. 不同算法的查全率和查准率

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。

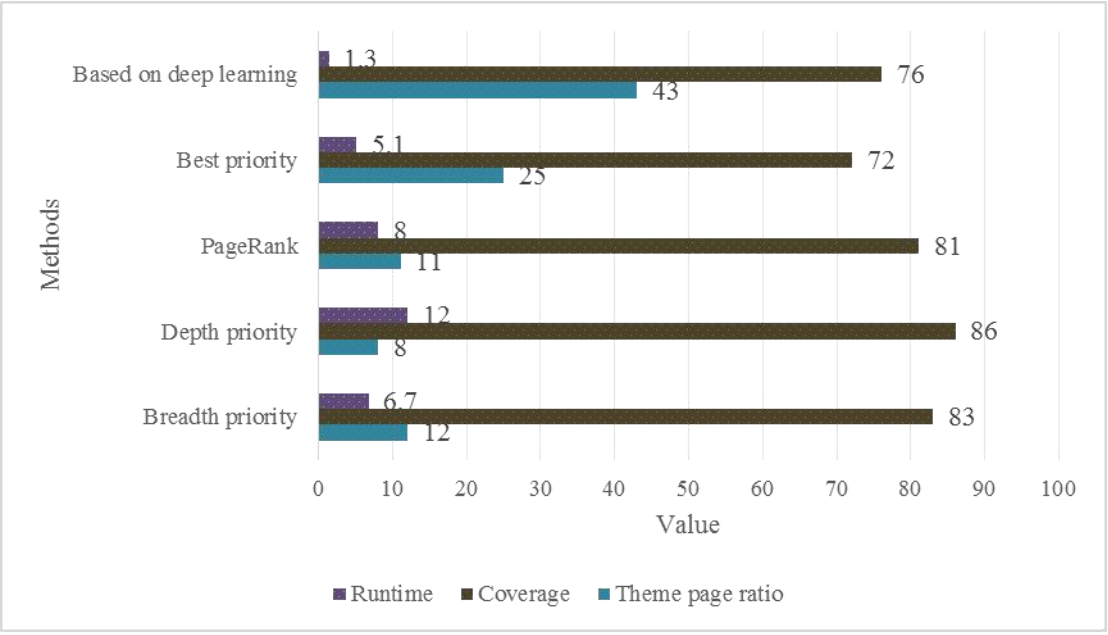

图 2. 不同爬虫策略的性能分析

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。

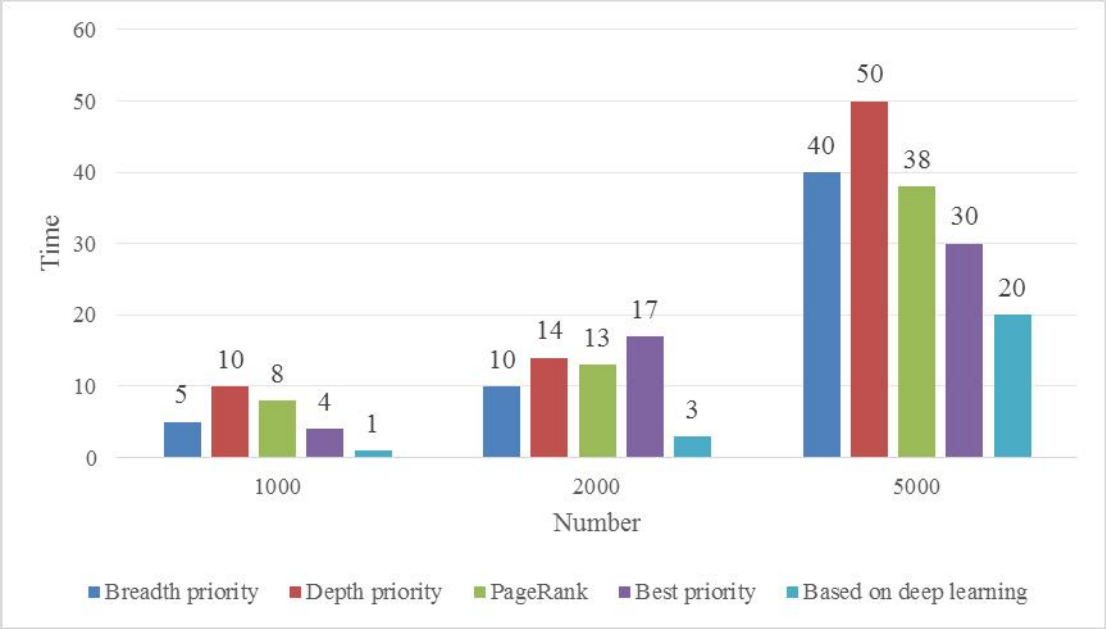

图 3. 不同爬虫策略抓取不同网页数的时间比较

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。