毕业论文（设计）任务书

一、任务说明

网络爬虫是一个具有一定的规则形成的脚本并可以自动的抓取互联网上的数据信息。如今，随着互联网大数据领域的蓬勃发展，网络爬虫的需求越来越大、应用场景越来越丰富，相关行业产值的增长率年年攀升。如何去完善爬虫的抓取的效率与高质量是目前和以后需要去解决的一个重点内容。现根据实际需要，设计一个分布式爬虫系统。

主要任务与目标：

1. 实现爬虫系统的分布式设计，使爬虫工作效率得到质的飞跃
2. 完善系统的反反爬虫设计，提高系统工作稳定性
3. 实现靶网站目的信息的爬虫脚本编写
4. 完成爬虫系统的部署，保证系统运维的高效性
5. 完成本爬虫系统与其他产品的性能对比。

二、任务要求

主要内容：

1) 分布式爬虫系统的架构设计

分布式爬虫是一种多机联合进行信息爬取的爬虫模式，根据机器之间的关系，分布式爬虫又分为主从式、对等式以及主从混合模式。分布式爬虫系统由多个模块构成，设计何种爬虫模式以及怎样设计系统各模块间的工作逻辑是研究的重点。

2) URL 的去重、调度算法设计实现

多机爬取目标网站 url 时不可避免会出现 url 重复的情况，重复的 url 会使系统进入循环，造成资源浪费。对于去重后的 url 进行任务分发，此时需要利用调度算法实现各个机器之间的负载均衡。

3) 反反爬虫功能的设计实现

有些网站针对短时间的大量请求，会启动反爬虫功能，造成爬取效率的降低。对此要考虑设计反反爬虫功能。

4) 目标网站信息抓取规则设计实现

爬虫程序是针对性很强的一类程序，若要实现抓取网站特定信息，则要根据网站结构量身制定相应的爬虫规则。

5) 爬虫系统的测试与部署

采用合适的部署方式可以大大提高爬取效率与系统管理，主流的部署方式有 Docker 与 VM 虚拟机

解决的问题

基于分布式架构的爬虫系统相较于传统爬虫程序，在信息爬取的规模和速率上都有了质的飞跃。引入反反爬虫策略，可以大大提高分布式爬虫系统的稳定性与生命力，以保证爬虫系统维持较高的性能。借助主流的项目部署平台 docker，可以对分布式爬虫系统实现更好的项目管理和运行效率。解决了传统爬虫爬取速率低，稳定性差，管理不方便的问题。

计划进度：

1-2 周	查阅相关资料，了解功能需求以及相关技术框架，准备开题答辩工作	7-8 周	设计分布式爬虫系统，并准备毕业论文的撰写工作
3-4 周	对网络爬虫及其分布式架构进行学习，开题答辩	9-10 周	根据设计思路完成系统具体的编程工作，完成爬虫系统的测试、部署、运行
5-6 周	对分布式爬虫架构及其各模块工作逻辑进行学习	11-12 周	完善毕业论文，准备毕业答辩

三、参考文献（规定阅读的文献不得少于 10 篇，其中外文文献不得少于 3 篇）

[1]洪伟. 分布式网络爬虫系统设计与实现[D].沈阳理工大学,2020.DOI:10.27323/d.cnki.gsgyc.2020.000101.
[2]马蕾. 分布式爬虫技术研究与实现[D].辽宁石油化工大学,2019.DOI:10.27023/d.cnki.gfssc.2019.000272.
[3]方奇洲. 基于 Docker 集群的分布式爬虫系统的设计与实现[D].武汉邮电科学研究院,2020.DOI:10.27386/d.cnki.gwyky.2020.000002.

[4]刘芳云,张志勇,李玉祥.基于 Hadoop 的分布式并行增量爬虫技术研究[J].计算机测量与控制,2018,26(10):269-275+308.DOI:10.16526/j.cnki.11-4762/tp.2018.10.058.
[5]刘星辰. 基于 Hadoop 的分布式网络爬虫的研究与实现[D].西安理工大学,2019.
[6]卢照,师军,张耀午,王琦.基于双缓冲的分布式爬虫调度策略的设计与研究[J].计算机与数字工程,2022,50(08):1686-1690.
[7]葛又嘉. 基于微服务架构的分布式爬虫系统设计与应用[D].南京邮电大学,2020.DOI:10.27251/d.cnki.gnjdc.2020.001058.
[8] Boldi P， Codenotti B， Santini M , et al. UbiCrawler: a scalable fully distributed Web crawler[J]. Software: Practice and Experience, 2004.
[9] Mirtaheri S M， Di Z， Bochmann G V , et al. Dist-RIA Crawler: A Distributed Crawler for Rich Internet Applications[C]// Eighth International Conference on P2p. IEEE, 2013.
[10] Zhong S， Deng Z . A Web Crawler System Design Based on Distributed Technology[J]. Journal of Networks, 2011, 6(12):1682-1689.

四、起止日期及进度安排

起止日期： 2022 年 10 月 9 日 至 2023 年 5 月 27 日

进度安排：

序号	时间	内容
1	2022.10.9-2022.10.20	开题与双向选题
3	2022.11.27-2022.12.24	搜集资料，准备开题报告
4	2022.12.25-2022.12.31	开题答辩
5	2023.1.1-2023.3.15	提交文献综述、外文翻译、开题报告
6	2023.3.15-2023.3.18	学院中期检查
7	2023.4.18-2023.4.22	系统验收
8	2023.4.23-2023.5.3	论文第一次定稿
9	2023.5.4-2023.5.7	学院集体查重、盲审
10	2023.5.9-2023.5.13	答辩
11	2023.5.16-2023.5.20	二次答辩
12	2023.5.21-2023.5.27	学院集中查重，论文最终定稿