关于论文选题，我根据实习的经验，对新闻类网页的网络爬虫比较熟悉，并且根据已有的论文，针对特定领域的爬虫研究并不多，我考虑做一个针对新闻的网络爬虫系统，并且提出一个优化的分布式爬虫系统解决方案。

题目暂定为《面向新闻类网页的网络爬虫系统设计与实现》

**背景**：随着互联网技术的的发展，网络中信息量快速增长，在新闻领域中体现非常突出。搜索引擎以及新闻推荐系统都非常依赖于网络爬虫，但是由于新闻类信息更新速度很快，传统的网络爬虫在抓取的实时性和并行速度上并不能满足要求。网络爬虫高度依赖于外部的Web服务器，存在很多不可控的因素，这样对系统的可靠性提出了更高的要求。现在的网络爬虫系统普遍采用分布式架构，但是分布式系统在业务逻辑复杂时存在数据同步和负载均衡的问题。

**解决问题：**新闻类网页的实时抓取，减少抓取的延迟。挖掘高质量新闻列表网页，减少爬虫的深度。可扩展的，健壮的，分布式网络爬虫系统。

**现状研究：**对于爬虫效率的提高，已经有大量的研究，比如URL去重和URL分配可能造成负载不均衡[1]。网络部分DNS解析会产生性能瓶颈[2]。分布式是提高爬虫效率最直接的方法，其中[3]提出了用分布式数据库解决存储问题。[4]实现了多节点并行的爬虫程序，使用MapReduce进行并行计算。

在评测标准方面，基于主题的网络爬虫侧重于准确率和查全率。分布式爬虫系统侧重于抓取速度和可扩展性。

**本论文的意义：**

上述研究提出了通用的网络爬虫效率问题的解决方法，但是在新闻类网页爬虫领域，可以有针对性的优化。比如新闻类网页通常由某些列表网页更新，那么爬虫的深度不需要太深，只需要挖掘出高质量的列表，对这些列表进行浅层的抓取。这部分主要为了提高抓取的准确率和查全率。

已有的分布式爬虫系统[5]采用将整个程序部署在单机，每个机器是相同的镜像的方式，这样容易产生数据同步和负载均衡问题，并且不利于模块解耦。[4]提出使用MapReduce进行并行计算，但是不适合新闻类的实时抓取。本文提出将爬虫分解成数据流模式，使用基于文件的分布式消息系统作为各模块之间通讯方式，解决分布式系统中数据同步和负载均衡问题。

**解决方案：**

主要功能是两个部分，新闻网页实时抓取，列表网页挖掘。

新闻实时抓取，将互联网复杂的有向图网络结构进行简化，使用数据流表示的线性结构，抓取时对种子页面做浅层的抓取。大概分为种子调度，列表网页抓取，详情网页抓取，文章去重这四个模块。各个模块之间解耦，使用分布式消息队列进行通信，这样各个模块可以部署在不同的机器上。将模块解耦的优点是提高系统的健壮性和可扩展性。部分机器的崩溃不会影响整体系统。并且每个模块的计算量和数据量并不相同，可以灵活地添加和删除某个模块的部署数量，满足负载变化的需求。

列表网页挖掘，使用传统的深层的爬虫，大规模挖掘可能的列表网页，通过列表网页识别得到候选种子，并且根据更新率和重复率评估质量。种子调度时根据评级设置不同的抓取频率和优先级。

**初步架构**



新闻实时抓取流程图：



列表挖掘流程图：



1.高性能网络爬虫:研究综述

2.网络爬虫效率瓶颈的分析与解决方案

3.付志超.基于Map／Reduce的分布式智能搜索引擎框架研究．硕士研究生论文. 武汉理工大学，2008．

4.苏旋.分布式网络爬虫技术的研究与实现

5.么士宇 基于分布式计算的网络爬虫技术研究