**可配置化网络爬虫项目总结说明书**

**小组名称：**可配置化网络平台小组

**小组口号：**我们爱学习

**指导教师：**陈博老师

**文档撰写人：**吴祥龙 关玉婷

**文档撰写时间：**2017年5月20日

团队分工记录表

|  |  |  |  |
| --- | --- | --- | --- |
| **项目名称** | **学号** | **姓名** | **分工** |
| 可配置化网络爬虫 | SA17225405 | 吴祥龙 | 开发总结 |
| SA17225094 | 关玉婷 | 开发总结 |
| SA17225362 | 王润 | 开发总结 |
| SA17225362 | 唐冬鑫 | 开发总结 |

**项目总结**

通过对搜索引擎和网络爬虫相关技术的学习，完成了可配置网络爬虫的实现，本系统是基于WebMagic利用java和scala语言实现的一个分布式的垂直爬虫框架。并且将开发爬虫的通用的操作封装了起来，例如html页面的下载，dom树的解析数据的存储等，同时还为用户屏蔽了开发分布式爬虫程序的很多细节，例如任务的同步，任务的调度，资源的管理等。利用本系统可以很方便的实现一个功能强大的分布式爬虫。

所实现的网络爬虫的结构和主要功能有：

1. 设计Master／work 架构的并行网络爬虫
2. 对普通网页的访问抓取
3. 能够对页面进行分析，提取出相应的网址
4. 能够对URL连接进行过滤，高效的URL消重算法
5. 可以进行网址的管理，分辨未抓取的网址，提高抓取效率

但是本系统还存在以下几点缺陷：

1.资源的隔离不彻底，本系统爬虫节点与爬虫节点之间的隔离方式是进程隔离，限制进程的内存，在同一台机器上如果一个爬虫节点的CPU，网络等资源占用比较高会影响到其他的爬虫节点。

2.整个爬虫集群，RM是管理者，所有爬虫都向他汇报状态，如果RM出现异常，会导致整个爬虫集群瘫痪。

3爬虫的下载器不支持js的渲染功能。

下一步我们会围绕着这几个点对这个爬虫系统进行进一步优化