**网络数据获取与分析**

项目目标：对互联网上各类网站信息进行高效率高可靠高实时的获取

技术指标：

1) 支持大规模数据源。预计达到10万以上

2）支持复杂数据源。包括网页、论坛、微博等不同数据格式，

ASP, JSP, PHP, AJAX等不同编程技术

3）支持反反爬虫技术和各项验证技术

4）支持对网络发布数据的低延迟获取

5）支持对网络发布数据的可验证、可追踪的爬取

当前需要进行的技术课题：

1. 对基于Scrapy框架的单个网站标准爬虫程序的确定以及验证其他程序是否符合此标准的验证过程
2. 对基于Scrapy框架的scheduler原理解析
3. 对基于Scrapy框架下多爬虫消耗带宽的计算或监控
4. 对基于多机系统的多爬虫程序设计
5. 对基于Scrapy框架下多爬虫消耗带宽的分配算法
6. 对redis 原理理解和拥有本项目可行性研究
7. 基于阿里云环境下的爬虫程序设计

当前需要进行的团队建设任务

1. 数据库设计文档（包括建立数据库SQL文件）
2. 对当前已爬取网站及其具体板块的描述文件
3. 对还需要进行爬取网站及其具体板块的描述文件
4. 对环境安装的优化以及无网络连接条件下环境安装