**三月爬虫模块总结**

1. **最初设计目标**
2. 实现各种主流社区网站和社交平台的分层爬取和内容解析功能。
3. 实现增量爬取，既保存原有爬取基数，又保证爬取任务不与原有的爬取重复。
4. 高效快速，能保证一定的信息时效性。
5. 运行稳定，行程良好的容错机制和预警机制，建立爬虫运行日志。
6. **爬虫设计思路**
7. 配置一定量的识别参数，以站点为单位，对每一层的页面按需提取需要的内容和通往下一层的链接
8. 在最底层的页面以同样的方式抽取所需的语料，直接形成原始情报库
9. **这样做的优缺点**

1.优点：

直接命中需要的目标，略去了爬去过程中对于各层垃圾链接的过滤和网页内容的去重工作。

爬虫有针对性，效率良好。

爬虫部分和解析器部分，下载的内容无需另外解析可直接入库。

2.缺点：

需要配置的参数较多，开始阶段，程序员会有一定的工作量。

爬虫生命力较脆弱，参数失效容易导致爬虫运行异常，且配置参数的技术要求较高。

1. **爬虫整体结构**

模块分成三层爬取结构：a.种子——主题，b.主题——文章，帖子底层节点，c.底层节点内容解析



1. **各时间段模块进展情况**

1．2.23~3.5

模块整体结构设计，各层解析内容实验（尝试7~8ge'zhan'di'a个站点）

2．3.6~3.15

确定数据库采用mysql+mongodb

编写各站点爬取程序（初步）

设计并建立爬虫所用的初期数据库，编写配套的持久层和服务层代码

3．3.16~3.25

实现增量爬取，建立单例缓存池

解析器部分编写，完善以配合其他部分

mysql数据库大调整，形成较完善的数据关系，实现对应的逻辑操作

实现爬虫部分多线程运行

mongodb语料库存储结构初步设计

日志系统及容错预警机制实现中

1. **后续工作**

1.日志系统及容错预警机制尽快完成

2.mongodb情报库存/取功能实现（整合入解析器）

3.模块检测和扩展更多类型的站点

4.压力测试

\*5.对后续分析统计的结果进行学习，指导爬虫运行

\*6.尝试使用代理服务器连接网络