关于爬虫程序运行流程及数据运转流程说明文档

一：整体架构图。



二：整体说明:

任务的详情和数据展示的管理界面,不在此文档赘述，此文档只介绍核心功能模块。

整体项目可分为三个部分：

1：数据的采集

2：数据的清洗

3：数据的使用

数据的采集：（请结合架构图查看描述）

编辑人员制定爬取任务后，该任务会进入任务队列，一个采集客户端有6个通道也就是说同时最多会有6个任务被执行，通道数可结合CPU核数调整。通道会获取到待执行的任务,根据编辑人员设置的url规则过滤掉不符合需求的url，将符合的url全部整合在一起并发去采集数据。然后程序根据智能算法获取爬取数据中的TDK和正文内容。将符合要求的数据写到存储原始数据的数据库。

数据的清洗：

清洗程序从存储原始数据的数据库中，提取未被清洗过的数据，提取数据中的高频词，去除敏感词，进行栏目分类。提取高频词是为了进行数据库的剔重，防止重复数据的写入。如果数据库中没有该条数据则将清洗过后的数据写入清洗完成数据库中。

数据的使用：

清洗完成数据库中的数据为有效数据，有效数据将被用作站群的展示文章。该数据对应的标题，还会通过百度搜索进行关键字搜索，将百度搜索得到的文章提取其TDK和正文内容，并提取其高频词，并将高频词所在的语句提炼出来，并写入数据库中为后续编辑人员和随即生成文章使用。