## 爬虫机制需求

爬虫实现基本原理：

通过浏览器控件，由系统驱动浏览器模拟人为的页面操作，进入到目标页面，对页面Html中的XPath来定位需要采集的数据，结合简单的数据清洗方式，从而实现互联网的分布式数据进行集中管理。

采集的内容：

1. 网页上可以看到的任何文字
2. 网页上图片的URL
3. 网页本身的URL
4. 网页上的超链接URL
5. 网页元素的innerHtml，outerHtml源码
6. 系统常量：当前日期时间，序号，任务ID，规则ID等

网页操作：

1. 打开URL
2. 点击页面上元素
3. 输入内容
4. 切换下拉内容
5. 鼠标悬停在元素上
6. 滚屏
7. 提取数据

逻辑操作：

1. 判断
2. 循环
3. 等待
4. 结束循环
5. 结束任务

数据清洗方式：

1. 替换字符串
2. 正则匹配
3. 正则替换
4. 去空格
5. 加前缀
6. 加后缀
7. 日期格式化
8. 转码
9. 任务编辑器
   1. 通过界面化操作方式，可以快速配置页面采集规则（无需像八爪鱼那样拖拽）。
   2. 配置的规则文件化，可导出导入。
   3. 自定义规则执行计划，指定日期，有效期，周期，频次，超时等
   4. 任务计划于核心管理系统中的任务管理中心保持同步。
   5. 任务日志查看，同步采集结果，系统中的任务执行日志。
   6. 规则保存后上传核心管理系统，也可以从核心管理系统下载。
   7. 用户登陆
10. 核心管理系统(规则，任务，终端管理)
    1. 规则存储，所有任务编辑器编写的规则信息集中管理，编辑器用户划分权限
    2. 用户管理：开发者用户，终端用户，注册，验证，注销
    3. 任务管理：任务状态，任务日志
    4. 终端资源管理：终端登记，终端状态，终端日志
    5. 任务分派管理：自动根据当前终端状态，为任务分派终端
    6. 异常处理：终端无响应，任务超时未完成等
11. 终端任务执行工具
    1. 终端心跳机制，终端用户登录
    2. 接收核心管理系统分派的任务
    3. 规则解析，执行，支持多任务同时执行，自定义任务数
    4. 简单的数据清洗功能
    5. 采集结果本地缓存，导出
    6. 采集结果传送给结果管理系统，支持断点续传
    7. 任务执行日志
    8. 一键安装，自动更新
12. 采集结果管理
    1. 接收各终端送来的数据
    2. 甄别任务完成状态，通知核心管理系统
    3. 保存终端，任务的日志数据
    4. 采集结果数据存储优化，分表，分库，读写分离等

具体采集规则的编写可参考 八爪鱼采集器

<http://www.bazhuayu.com/>

下载地址：

<http://update.bazhuayu.com/clientsetup/%E5%85%AB%E7%88%AA%E9%B1%BC%E9%87%87%E9%9B%86%E5%99%A8V5.0%E7%89%88%E5%AE%89%E8%A3%85%E5%8C%85.zip>