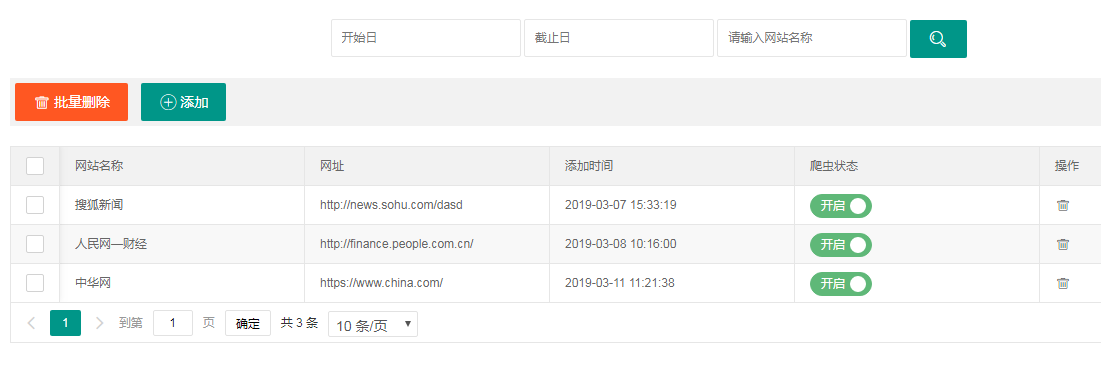
# 爬虫后台管理系统功能文档

## 一 网址管理

### 1.1在采集网址管理



显示网站名称，网站域名，添加时间，采集时间，采集状态，该网站今日被采集信息量，在采集网址数量，网址更新频率，所属爬虫引擎等信息。

根据开始采集的起止时间查询在此时间范围内开始采集的网站，显示列表信息，或输入网站名称查询网站信息。

可以对正在采集的网站进行关闭处理，关闭之后的网址会被剔除到被采集网站列表之外，可以根据网站被开启采集的时间段或网站名称查询某个或某些网站信息，对网站状态进行处理。可以批量关闭一些网站。

### **1.2待采集网址管理**

1.2.1 待采集网址列表信息

显示网站名称，网站域名，添加时间，采集状态，该网站国内是否可访问，网站类型，待采集网站数量等信息。

1.22 添加网址信息

存在添加按钮，点击按钮跳转添加网站页面，输入网站名称，网站类型，国内是否可访问等信息，采集状态默认为0，添加时间默认为系统时间，网站域名需要验证，不合法的网站需要提示，已存在域名有提示功能不予添加。

1.23 批量导入网址信息

存在批量导入按钮，点击按钮跳出选择文件窗口，（目前只读excel,后期扩展）读取excel表格添加上述信息，对不存在的字段设置为空，通过域名检验数据是否存在，已存在的数据跳过处理，事后跳出窗口给予提示添加网站数量，重复数据数量等信息，显示导入进度条。

1.24 删除/批量删除网址信息

存在批量删除按钮，点击删除按钮，弹出确认框，确认后，选中数据全部删除，每条数据后都有单个删除数据操作，此功能为从数据库中删除网址信息（谨慎操作）

1.25 条件查询网址信息

通过设置添加的起止时间或者模糊匹配网站名称查询网址信息，存在查询按钮，点击之后通过配置信息查询所有符合数据并做展示。

1.26 结果

显示该网站被采集了多少数据，点击结果显示详细信息，在某个时间开始到结束时期被采集了多少条数据，这些信息通过读日志的形式存储在mongodb中。

### **1.3问题网址管理**

查询在被采集过程中出现异常的网站，所在爬虫服务器，异常信息，发生异常的时间，异常是否解决等等信息，有提交反馈信息功能，以便运维人员后期维护，查询某时间段发生异常的信息，批量提交等等。

## **二 爬虫管理**

### **2.1在启动爬虫任务列表**

查询服务器中正在运行的爬虫程序，显示该爬虫的类型（通用爬虫，定向爬虫，高效率实时爬虫等），爬虫所采集的网站数量，采集的数据量，运行时间，采集效率等信息，可暂停/批量暂停爬虫（暂停的爬虫保持先前的配置信息）。

2.12 获取采集结果

通过读取日志获取采集过程中的信息，点击采集结果显示，已采集的网站，等待中的网站，每个网站被采集了多少条信息,还有多少条信息未采集。

### **2.2待启动爬虫任务列表**

2.21 新增爬虫任务

存在创建任务按钮，点击创建任务，弹出创建任务界面，加载未启用网站信息，勾选需要采集的网站。添加到采集池，设置采集深度，设置定时时间（可选）等配置信息，配置信息保存mongodb数据库，任务状态为未启动，任务名称，爬虫类型二进制文件，创建人，创建时间等信息保存在mysql数据中。

2.22 查询爬虫任务

查询服务器中未启动的爬虫程序，显示任务名创建时间，爬虫状态，添加者，配置信息，爬虫启动按钮，一键启动，点击配置信息，通过配置字段名查询mongodb相应的配置信息，配置信息在弹出框中，可对配置信息进行更改。

2.23 启动爬虫任务

读取采集池中的网址，依次将webinfo表中的status字段更新为1，将spider\_task中的status字段更新为1，配置爬虫开始采集，采集完毕之后再将webinfo表中的各字段改回去，采集结果统计写在日志系统中。

2.24 删除爬虫任务

存在批量删除按钮，点击删除按钮，弹出确认框，确认后，选中数据全部删除，每条数据后都有单个删除数据操作，此功能为从数据库中删除网址信息（谨慎操作）

## **三 数据统计**

### **3.1服务器监控**

### **3.2爬虫监控**

## **四 权限管理**

### **4.1成员管理列表**

管理员创建账号登录密码可供其他成员登录，没有注册功能，成员对爬虫，网站的操作信息都会展示在权限管理模块中（这一块还在构思中）

## **五 开发设计**

### **5.1数据库设计**

#### **4.11 mysql 数据库**

**4.111 数据表**

1. **User表**

字段

ID # id

Username #用户名

Password #密码

1. **Webinfo 表**

字段

ID #url的md5加密

Url #网站网址

Web\_name #网站名称

Add\_time #添加时间

Status #采集状态

Agaent #是否需要代理

Sort #网站类型

Info #备用字段

1. **Spider\_task表**

字段

ID #爬虫任务名的md5加密

task\_name #爬虫任务名

task\_job #爬虫任务的二进制文件

create\_time #创建时间

Status #爬虫状态

creater #创建者

#### **4.21 mongodb 数据库**

**4.211 数据表**

1. **config表**

{ "uuid" : "mysql中spider\_task中的uuid,用于定位该爬虫的参数",

"urls" : [用于采集的url池，这里是个列表],

"start\_time" : "开始采集的时间（注意，不是配置的时间，是爬虫启动的时间）",

"limit" : "限制采集深度",

"fixed\_time" : "设置定爬虫启动时间",

"agent" : "是否需要代理”

}

#### **4.31 redis 数据库**

### **5.2 接口数据对接**

#### **5.21下发任务**

#### **5.22 数据反馈**

### **5.3 分布式部署**

#### **5.21 主从机配置**