**数据检查插件开发文档**

## 一．基本思路

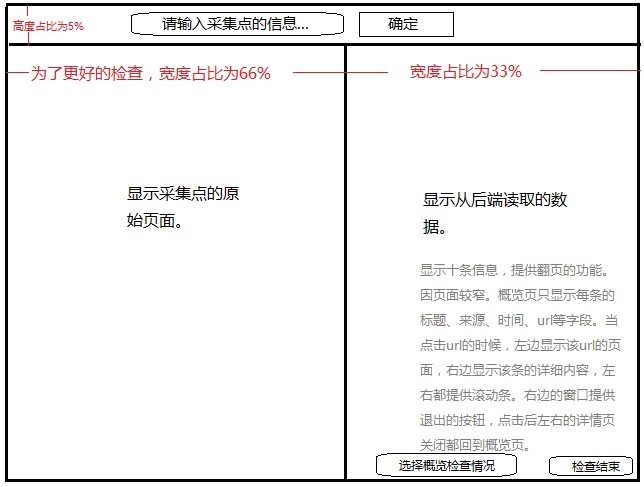
通过开发一个本地的页面，把待检查的采集点的数据和采集点原网页放在同一个页面中，这样能够使检查数据更加直观，可以减少很多窗口间的切换和一些不必要的重复操作。

把待检查的excel表放到程序的工作目录，检查的情况维护这个excel表，每次检查采集点更新检查情况到后端的excel表。

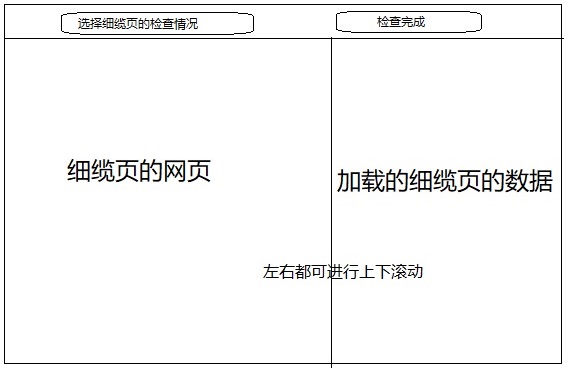
## 工具结构

**前端部分：**浏览器的满窗口。顶部为采集点的键入栏，左边显示采集点的原网页，右边显示后端读取的采集点数据，初始左右都是概览页，点击一条数据后左右都显示该条数据的详细内容。

**图1-概览页**

****

**图2.详情页**

****

1. 采集点输入：顶部的下拉框选择采集点的信息（网站名，栏目名，网站url），提供去重功能，检查过的采集点会从下拉框移除。
2. 左侧iframe窗口：通过采集点的url加载采集点的概览页的页面。因左侧的窗口较窄无法100%的展示网页的宽度，目前的解决方法chrome可通过ctrl+鼠标滚轮的方式缩小网页达到宽度100%的适配，使其无须使用左右的滚动条。
3. 右侧窗口：通过采集点（网站名，栏目名）调取后端的接口，一次获取10条新闻（按发布时间排序）展示出概览页，提供翻页的功能。
4. 概览检查：通过对比左右的概览页显示的新闻标题，对比是否有新闻漏采。检查结束进行勾选（正常，漏数据，数据重复，自定义错误）
5. 细缆检查：点击右侧概览页上对应新闻的url，点击后左右两侧都显示出该条新闻的细缆信息，进行该条的细缆内容检查。检查结束进行勾选（正常，内容错误，来源错误，时间错误，作者错误，自定义错误）。细缆内容检查完毕，点击右侧页面的关闭按钮，左右都回到概览页。
6. 采集点检查结束：概览和所有细缆检查结束。通过调用后端接口的方式，把采集的结果输出到excel表格。所有为正常时输出正常，有错误时，输出错误并贴出错误对应的url。
7. 进行下一个采集点的检查。

**后端部分：**

a.服务：架设本地的服务接口供前端调用。接口接受前端传递来的采集点参数，通过采集点参数驱动爬虫脚本爬取Hybase，返回加工好的数据给前端。维护一个要检查的excel表，与前端交互待检查的excel表的检查情况。

b.爬虫：爬取Hybase的数据页面。脚本携带采集点参数获取采集点的若干条新闻（至少10条）。解析整理每条新闻的content，authour，srcname，urltime，url，title字段。

## 开发分配

**前端部分：**何林，日期暂定3天。

**后端部分**：黄莉