IE 采集引擎

文档修改记录

修改内容	作者	时间
分析阶段-难点分析，技术架构，原型设计	刘楚门	2018.3.28 - 2018.4.13
验证阶段-工作计划	刘楚门	2018.4.16
验证阶段-驱动工具原型设计，定位工具原型设计修改	刘楚门	2018.4.28
实现阶段-工作计划	刘楚门	2018.5.21

Table of Contents

1. 难点分析

1.1. 网络请求拦截

要求：通过任何技术，去以阻塞/非阻塞的形式拦截 WebBrowser 的 HTTP 协议网络请求，并输出该请求的数据信息。

输入：进程名称/进程 ID，协议类型

输出：发送者全部信息，接受者全部信息，请求数据，响应数据

其他：控制该请求的成功或者失败。

解决 1：通过 FiddlerCore 拦截 WebBrowser 的 HTTP/HTTPS 请求。

解决 2：通过二次开发 Squid 开源代理服务器拦截 WebBrowser 的 HTTP/HTTPS 请求。

1.2. 通过鼠标定位并获取网页元素

要求：通过制作一个浏览器，用鼠标去定位任一网页中的标签元素，并输出其结构。

输入：鼠标点击，鼠标移动，鼠标划出一个区域

输出：标签元素，高亮显示元素，区域内的多个标签元素

其他：实时修改标签元素的属性与内容。

解决 1：通过在 WebBrowser API 层添加鼠标事件来实现与标签元素的交互。

1.3. 获取网页中 JS 运行时所有定时器

要求：访问一个存在多个定时器的网页，通过任何技术去得到 JS 运行时中的所有设置的定时器。

输入：多个定时器代码

输出：定时器代码执行状态

其他：管理所有定时器，控制定时器内的执行代码的顺序。

解决 1：通过重写 window.setTimeout 与 window.setInterval 系统函数来设计一套用户定时器控制模块。

1.4. 并发采集

要求：同时采集多个页面中的数据。

解决：通过新架构解决。

1.5. 脚本同步

要求：在本地可供单步调试。

解决：通过新架构解决。

1.6. DOM 结构发生局部变化的定位

要求：当 Ajax 请求结束后，局部 DOM 节点变化时，可以通过配置进行位置定位。

解决 1：尝试通过配置多个定位，定位 1 失效，采用定位 2。

解决 2：尝试通过正则表达式的方式进行提取（不依赖标签）。

1.7. 页面 JS 执行之前执行自定义的 JS

要求：在页面载入结束后，开始执行页面 JS 前。

输入：任何 JS 代码。

输出：JS 执行成功。

解决：通过在代理服务器中拦截 HTTP 响应，动态插入自定义 JS 在 Header 中，保证优先执行。

1.8. 页面弹出窗口的同步处理

要求：当用驱动脚本执行页面操作时，当弹出窗口时，可以同步处理窗口。

解决 1：通过重写 alert，confirm 根据驱动脚本同步点击。

1.9. 通过 FiddlerCore 拦截 HTTPS 协议的怪问题

要求：似乎跟证书有关。

解决：

1.10. 通过 FiddlerCore 修改 HTTP 响应体变成乱码的问题

要求：http://s.163.com

解决：修正格式。

1.11. WebBrowser 访问页面失败的事件处理

要求：在运行过程中因为网络问题导致页面无法显示，需要有处理事件。

解决：重试。

1.12. 断点续跑

要求：如果出现了关机，启动后能够重新接着上次的进度运行。

解决：保存网址，页数。

1.13. 获取页面元素的显示图像

要求：通过一个 HtmlElement 类型的节点得到该节点的显示图像。

输入：HtmlElement 元素

输出：图像的 Base64 编码

解决：html2canvas

1.14. JS 定时器接管模块

要求：通过设置 true 或者 false，决定定时器可以被执行或者被忽略。

输出：定时器接管模块代码。

解决：

1.15. WebDriver

1.15.1. 模拟 MouseHover 事件时，出现弹出菜单立刻收回的情况

解决：

1.15.2. 在页面进行异步加载时出现 WebElement 过期错误的情况

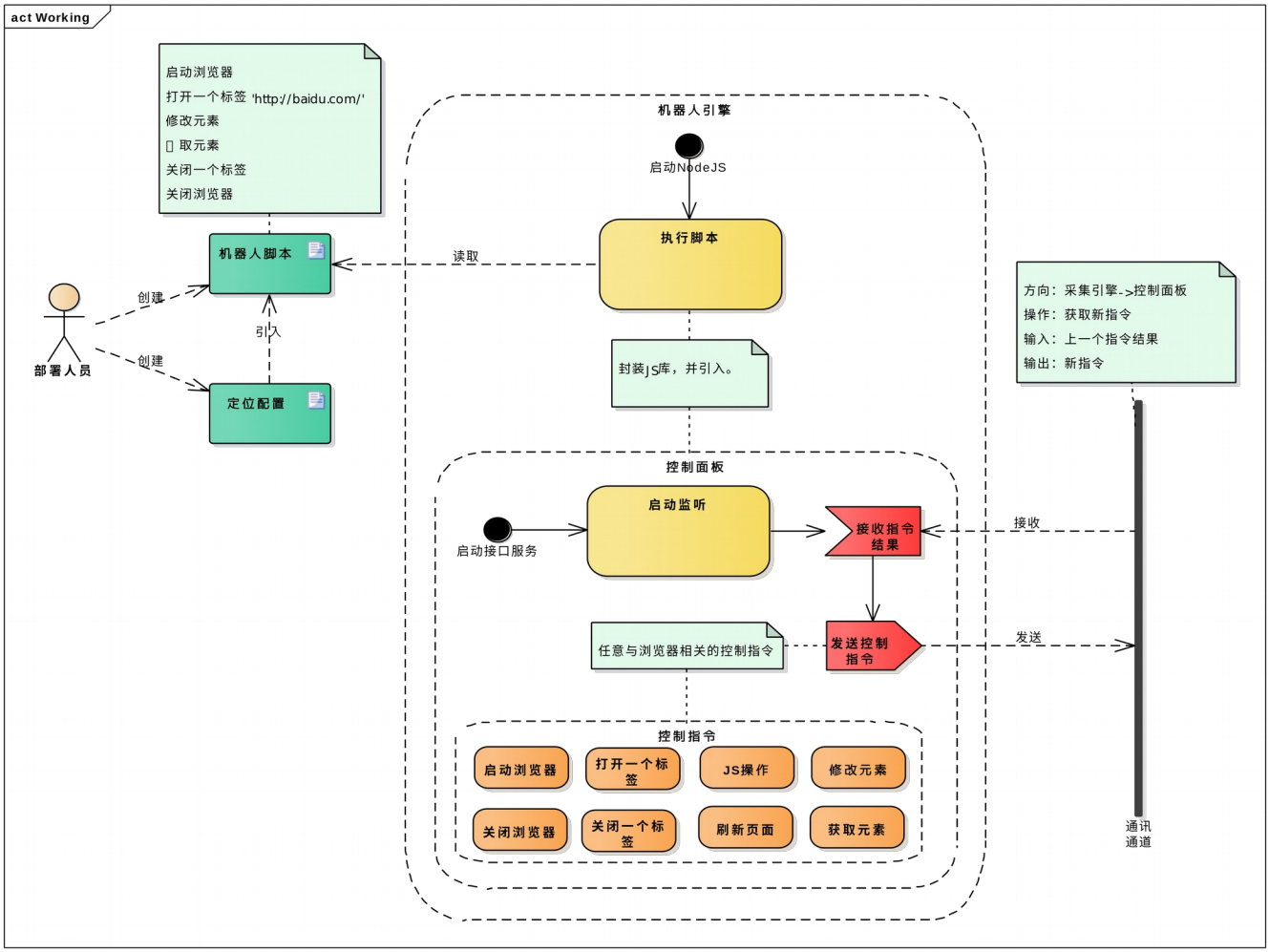
解决：重新获取。

1.15.3. 在运行过程中执行 excuteScript 脚本过程中阻塞

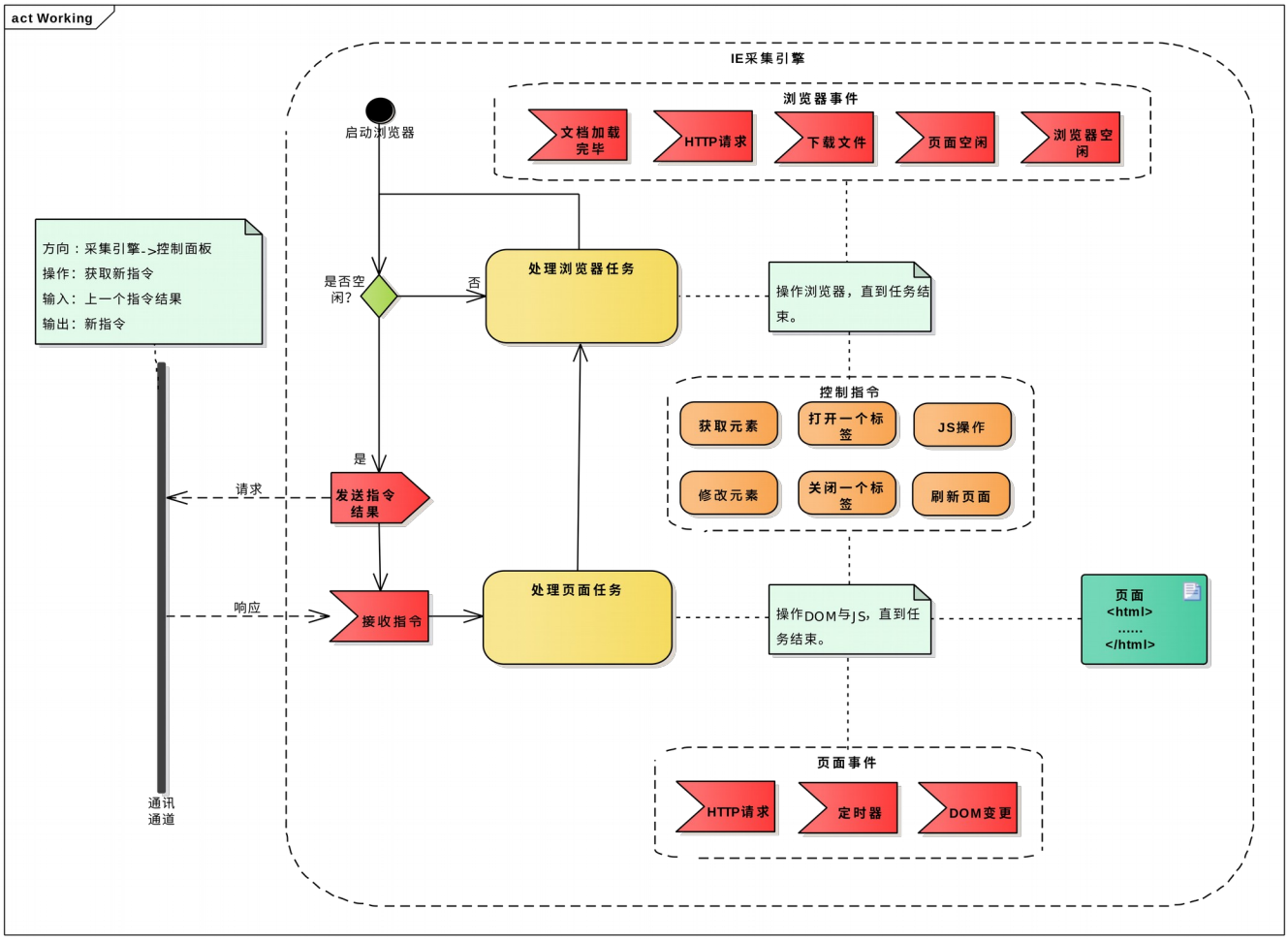
解决：

2. 技术架构

2.1. 驱动引擎

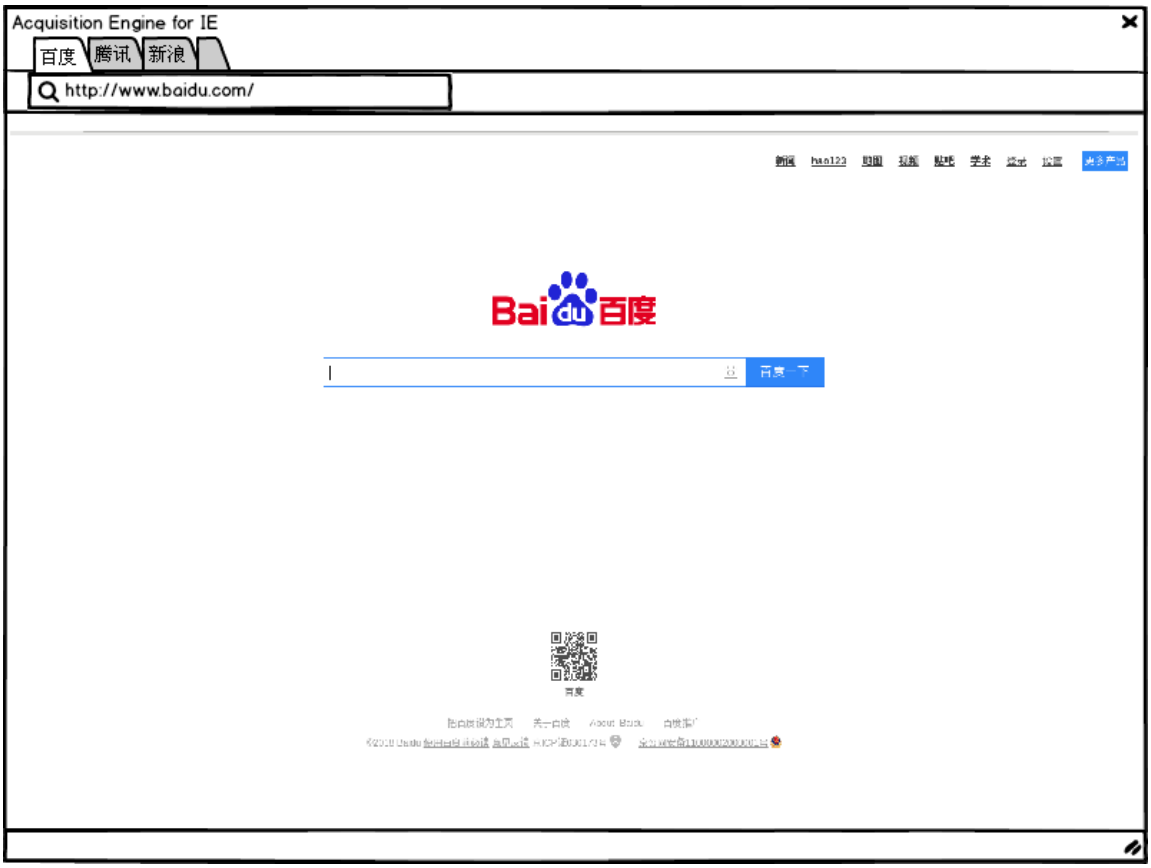


2.2. 采集引擎

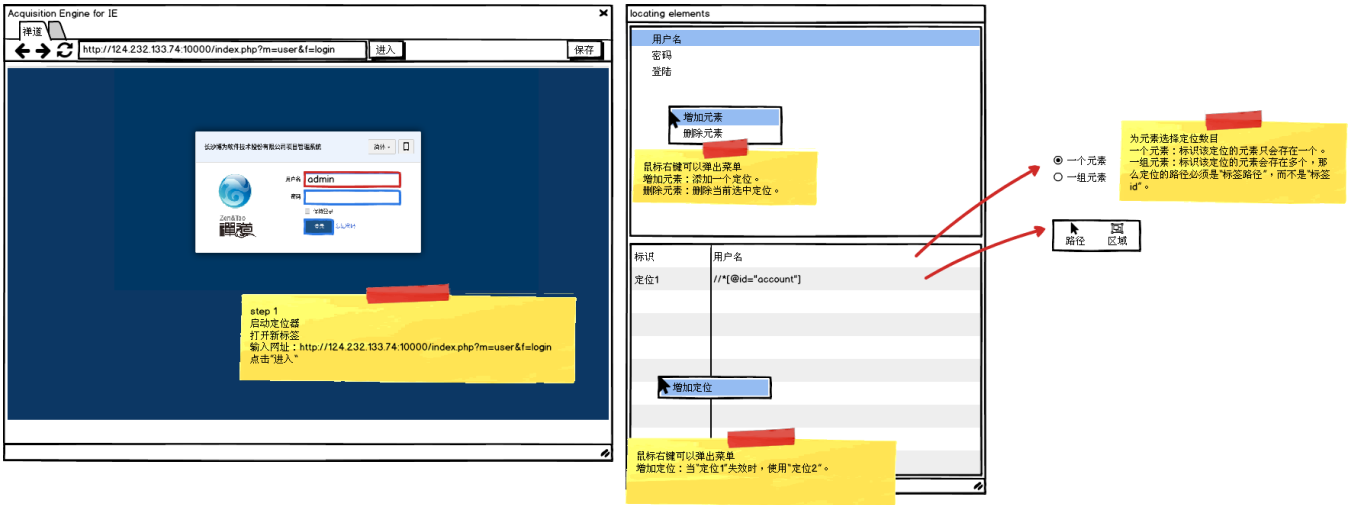

act Working
IE采集引擎
浏览器事件
文档加载完毕
HTTP请求
下载文件
页面空闲
浏览器空闲
启动浏览器
方向：采集引擎->控制面板
操作：获取新指令
输入：上一个指令结果
输出：新指令
是否空闲?
否
处理浏览器任务
操作浏览器，直到任务结束。
控制指令
获取元素
打开一个标签
JS操作
修改元素
关闭一个标签
刷新页面
是
请求
发送指令结果
响应
接收指令
处理页面任务
操作DOM与JS，直到任务结束。
页面
<html>
.....
</html>
页面事件
HTTP请求
定时器
DOM变更
通讯通道

3. 交互原型

3.1. 采集引擎

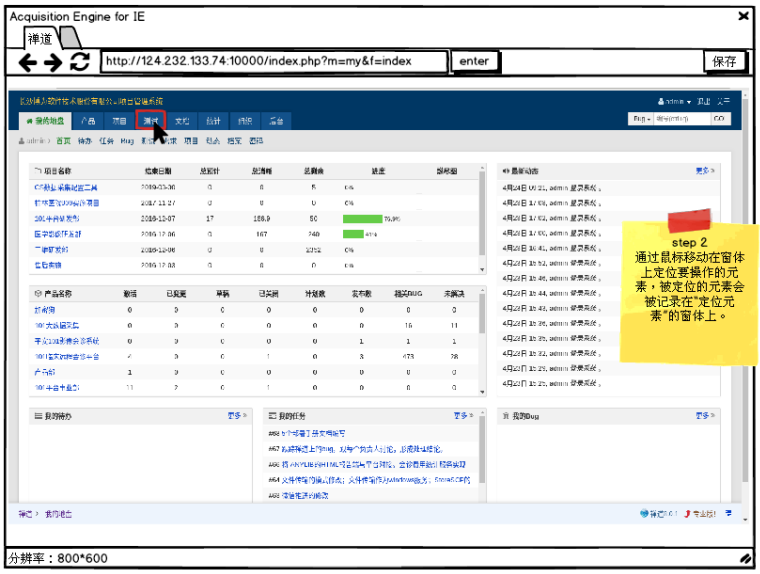


3.2. 定位工具



路径　区域

工具栏：
路径定位
区域定位



step 2
通过鼠标移动在窗体上定位要操作的元素，被定位的元素会被记录在"定位元素"的窗体上。

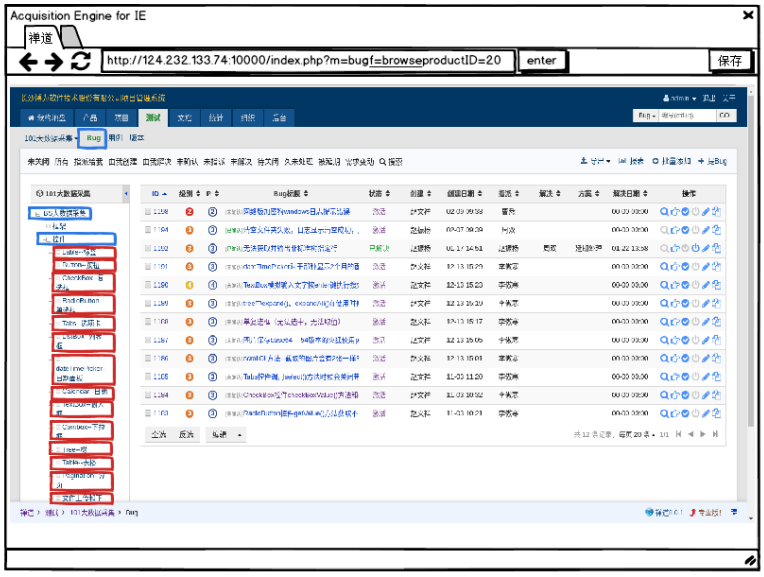
locating elements

- 用户名（不可见）
- 密码（不可见）
- 登陆（不可见）
- 测试

标识	用户名
定位1	//*[@id="menuqa"]

通过鼠标移动在窗体上定位要操作的元素，被定位的元素会被记录在"定位元素"的窗体上。

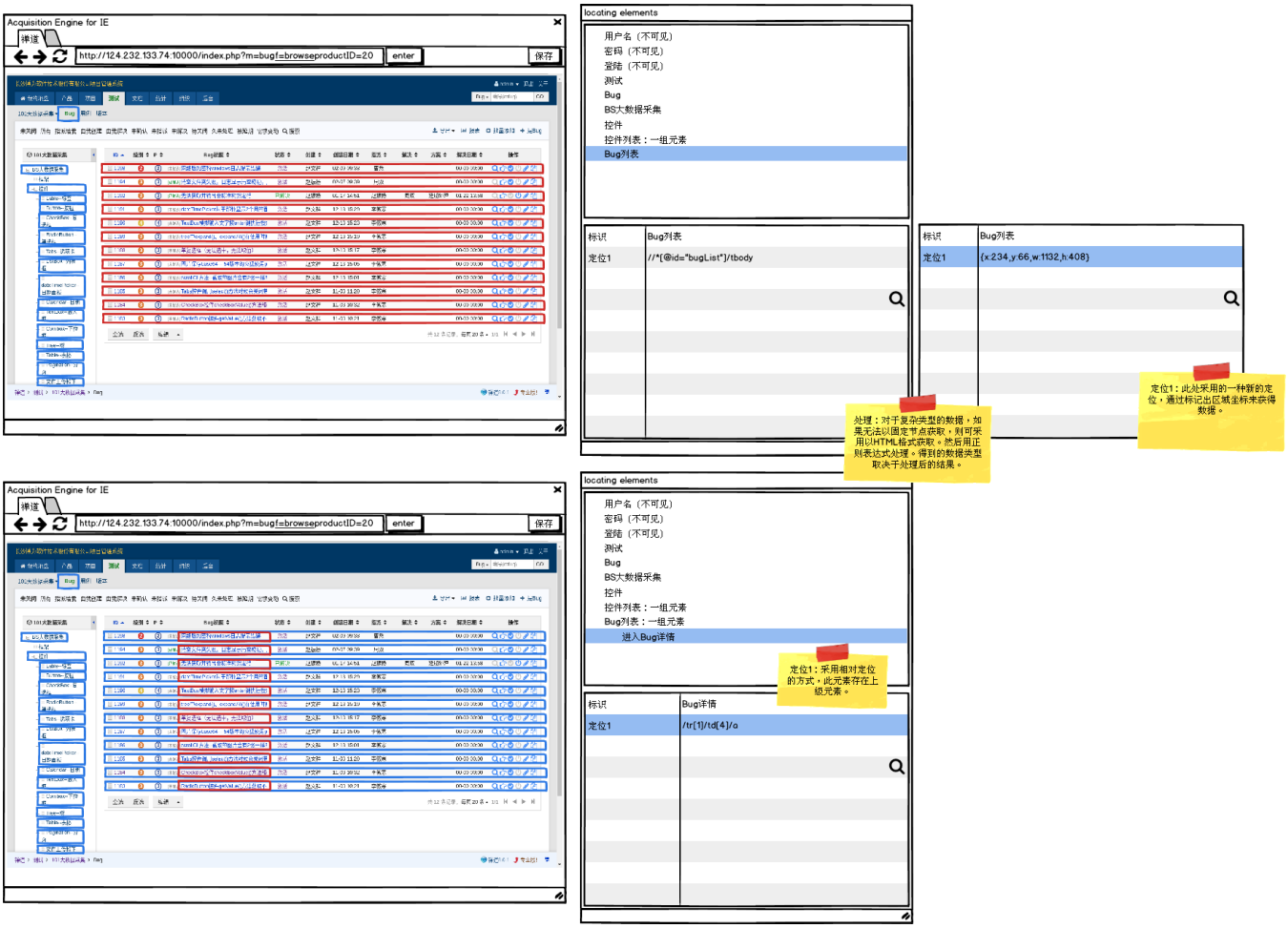
路径　区域



locating elements

- 用户名（不可见）
- 密码（不可见）
- 登陆（不可见）
- 测试
- Bug
- BS大数据采集
- 控件
- 控件列表：一组元素

标识	控件列表：一组元素
定位1	//*[@id="treebox"]/div/div/div[2]/ul/li[1]/ul/li[2]/ul/li

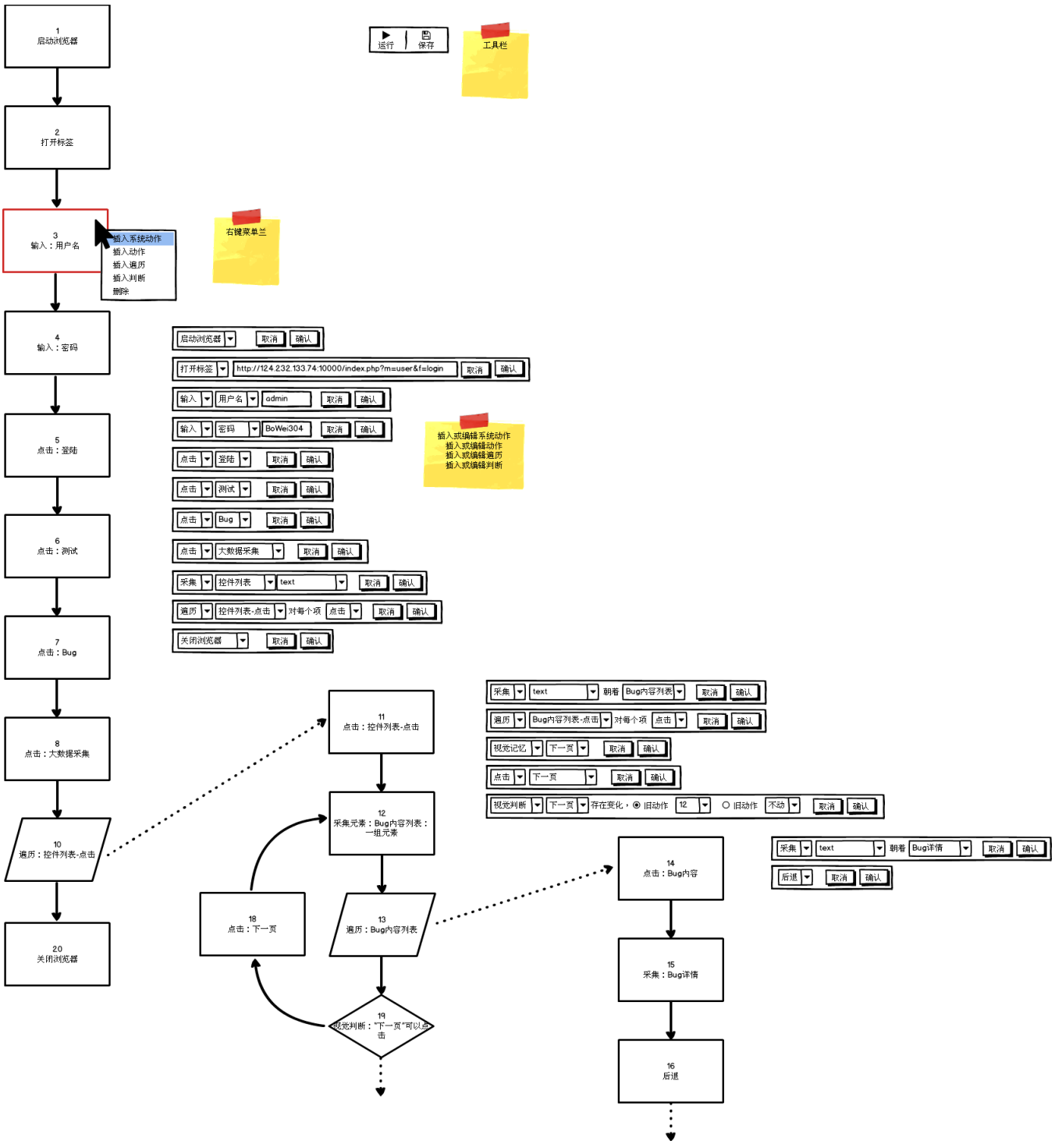


3.2.1.生成定位配置

当点击保存后，生成定位配置到“./driver/locations.js”，当启动程序后默认载入这个配置。

```
module.exports.用户名 = '//*[@id="account"]'
module.exports.密码 = '//*[@id="login-form"]/form/table/tbody/tr[2]/td/input'
module.exports.登陆 = '//*[@id="submit"]'
module.exports.测试 = '//*[@id="menuqa"]'
module.exports.当前选项 = '//*[@id="currentItem"]'
module.exports.数据101 = '/html/body/header/nav[2]/ul/li[1]/div/div/div[1]/ul/li[7]/a'
module.exports.点击采集 = '/html/body/div[1]/div[1]/div[2]/div/div/div[2]/ul/li[1]/div'
module.exports.控件 = '/html/body/div[1]/div[1]/div[2]/div/div/div[2]/ul/li[1]/ul/li[2]/div'
module.exports.控件列表 = '//*[@id="treebox"]/div/div/div[2]/ul/li[1]/ul/li[2]/ul/li'
module.exports.控件列表$项 = './/a'
module.exports.Bug列表 = '//*[@id="bugList"]/tbody/tr'
module.exports.Bug列表$项 = './/td[4]'
module.exports.Bug列表项详情 = '//*[@id="wrap"]/div[1]'
module.exports.Bug列表$下一页 = '//*[@id="wrap"]/div[1]'
module.exports.Bug列表$下一页$$视觉记忆 =
'iVBORw0KGgoAAAANSUhEUgAAABkAAAAeCAYAAADZ7LXbAAAArElEQVRIie3UsQ2DMBAFUIa7xrSGnhnMEDAE8gpuLG/AEHgDqJEr7J/KVaSEkBxK4S/
99r/qrsINqQpSkKeklPiRtmlhjMFxHHwIEYGI0HUd5nnmRXKVUliWhRfJHYYB67ryIkQEIQSmacK
+73xIrpQS1lpepGkaOOd4kLquobVGCOElcBkZxxHbtr0dv4T0fQ/v/enxjxDWY7zlrcQYvxo/hfwqBSlIQf4EeQBsoxMlOiK7zgAAAABJRU5ErkJggg==
='
```

3.3. 驱动工具


1
启动浏览器
2
打开标签
3
输入：用户名
4
输入：密码
5
点击：登陆
6
点击：测试
7
点击：Bug
8
点击：大数据采集
10
遍历：控件列表-点击
20
关闭浏览器
插入系统动作
插入动作
插入遍历
插入判断
删除
右键菜单兰
运行
保存
工具栏
启动浏览器 取消 确认
打开标签 http://124.232.133.74:10000/index.php?m=user&f=login 取消 确认
输入 用户名 admin 取消 确认
输入 密码 BoWei304 取消 确认
点击 登陆 取消 确认
点击 测试 取消 确认
点击 Bug 取消 确认
点击 大数据采集 取消 确认
采集 控件列表 text 取消 确认
遍历 控件列表-点击 对每个项 点击 取消 确认
关闭浏览器 取消 确认
插入或编辑系统动作
插入或编辑动作
插入或编辑遍历
插入或编辑判断
11
点击：控件列表-点击
12
采集元素：Bug内容列表：一组元素
13
遍历：Bug内容列表
18
点击：下一页
19
视觉判断："下一页"可以点击
采集 text 明着 Bug内容列表 取消 确认
遍历 Bug内容列表-点击 对每个项 点击 取消 确认
视觉记忆 下一页 取消 确认
点击 下一页 取消 确认
视觉判断 下一页 存在变化，◉ 旧动作 12 ○ 旧动作 不动 取消 确认
14
点击：Bug内容
15
采集：Bug详情
16
后退
采集 text 明着 Bug详情 取消 确认
后退 取消 确认

4. 工作计划

4.1. 分析阶段(3.28 - 4.13)

难点分析，技术架构，交互原型。

4.2. 验证阶段(4.16 – 5.14)

演示版的实现，附带基本的工具。

刘楚门：负责交互设计，技术架构，与产品文档的输出。

邓广湖：负责定位工具的实现，熟练 WinForms 编程，WebBrowser 控件，具备基本的 HTML，JavaScript 操作。

组件	模块	工作	人员	时间
IE 采集引擎-演示版	采集界面	显示单个标签	刘楚门	2 天
		多个标签切换		
		刷新，前进，后退		
	Http 节点查找	节点路径查找算法	刘楚门	1 周 3 天
	浏览器空闲事件	启动代理服务器，并设置任意进程的代理	刘楚门	1 周 3 天
		Http 请求结束事件		
	指令	接收指令	刘楚门	1 周 1 天
		执行指令： 启动采集，关闭采集 打开标签，关闭标签，切换标签，刷新标签 前一个页面，后一个页面 采集元素，修改元素，点击事件		
		发送执行结果		
IE 驱动引擎-演示版	指令	发送指令	刘楚门	2 天
		接收执行结果		
IE 定位工具-演示版	定位工具界面	显示单个页签	邓广湖	2 天
		多个页签切换		
		刷新，前进，后退，进入		
		定位元素 定位一组元素		
	定位 HTML 元素	单个元素标签	邓广湖	4 天

	输出 HTML 元素界面	元素定位 路径信息 定位一组元素	邓广湖	2 天
	输出元素定位	单个元素定位	邓广湖	1 周
		一组元素定位		
	制定元素信息格式	路径信息	邓广湖 刘楚门	1 周
驱动工具	界面	动作，遍历，判断，运行，保存	刘楚门	剩余时间
	其他	生成脚本	刘楚门	剩余时间
组件同步联调	IE 驱动引擎 驱动引擎 IE 定位工具 驱动工具	同步联调	刘楚门 邓广湖	剩余时间
WebDriver 框架	探索	确定 WebDriver 的潜力与不确定性	田洋	无限制
	思路	网络中断的处理 浏览器崩溃时脚本断点继续运行的处理	田洋	无限制

4.3. 实现阶段

商用版的实现，附带完整的工具。

刘楚门：负责交互设计，技术架构，与产品文档的输出。

邓广湖：负责定位工具的实现，熟练 WinForms 编程，WebBrowser 控件，具备基本的 HTML，JavaScript 操作。

谢洪亮：暂无。

田洋：负责系统测试。

业务	模块	工作	人员	时间
采集引擎	稳定性测试-24 小时	IE8（可用） IE9（可用） IE10（可用） IE11（可用，慢） Firefox chrome（删除，截图 API 不兼容）	田洋	1 周
安装服务	业务不间断运行	浏览器重启	谢洪亮	1 周
定位工具	视觉记忆	截取元素屏幕快照	邓广湖	1 周
	组定位	多层组相对定位		

驱动工具	浏览器指令	启动浏览器 关闭浏览器 打开网页 关闭网页 切换		
	动作指令	采集 点击 输入		
	遍历指令	遍历组		
	判断指令	视觉判断 值判断		
	跳转指令	跳转		
	等待指令	等待元素 等待时间		
	删除	删除一个活动		
	保存	保存设计图		
	运行	运行设计图		
	载入	载入设计图		
	标识	设计图名		
技能提升	WPF	图形界面的开发	邓广湖	1 月
	Wform	flowchart 控件库的使用，用于制作驱动工具		

4.4. 佛山项目-需求未确认

业务	模块	工作	人员	时间
采集	待办信息列表	登录操作 选择侧边待办操作 选择待办环节操作 遍历待办列表操作	田洋	1 周?
	待办信息详细	查看待办操作 办事人信息-15 字段 设立登记-6 字段 刻章、开户-?字段 受理页面-未确认 流程信息-未确认 工作流跟踪-未确认	刘楚门	
交互	配置	采集频率 采集状态查看 采集状态控制	刘楚门	3 天
	日志	推送记录	谢洪亮	1 天
接口	接口文档	待办信息	甲方	甲方时间
	接口测试			
	推送数据		谢洪亮	3 天
流程测试	采集->接口	功能正常	田洋，谢洪亮	剩余时间
	交互->采集			
稳定性测试	运行 24 小时	业务正常	刘楚门	