软件分析测试实验报告

DZ1833036 张霄炜

# 一．工具名称及功能描述

工具名称：整站链接有效性检查工具

功能描述：用于检查整个网站中所有父链接及子链是否是可到达的。

# 二．研究背景

死链是Web网站中处于路径不可达状态的Web链接，也是导致网站性能下降的重要原因．随着互联网的快速发展，网络中的各种信息也在急速增长，维护网络内容的完整性和维持网络的性能也变得十分复杂．网络的一个重要特点是动态性，即网页内容可以任意增加、删除和更新．这个特点使网络成为具有吸引力的媒介，同时也使它的完整性受到了威胁，导致了死链的产生．不仅如此，多数链接是相互引用的，再加上网络的频繁更新和删除，使得死链问题更加严重．

# 三．研究内容

## 3.1研究内容

本次研究主要实现了以下两点：①根据Web链接的调度过程，自动获取链接信息；②解释了死链产生的原因并提出了对应的检测方案。

## 3.2死链产生原因

死链接也称无效链接，即那些不可达到的链接．以下情况会导致死链接的出现：动态链接在数据库不再支持的条件下，变成死链接；某个文件或者网页移动了位置，导致指向它的链接变成死链接；网页内容更新并换成其他的链接，原来的链接变成死链；网站服务器配置错误．

# 四．技术路线

## 4.1死链接检测方法

此次研究中对死链的检测步骤如下：

1）爬取网页内容，并保存完整页面信息；

2）把页面中的链接抽取出来；

3）对这些链接进行预处理；

4）检验链接的可达性．

# 五．详细设计

## 5.1设计流程图：

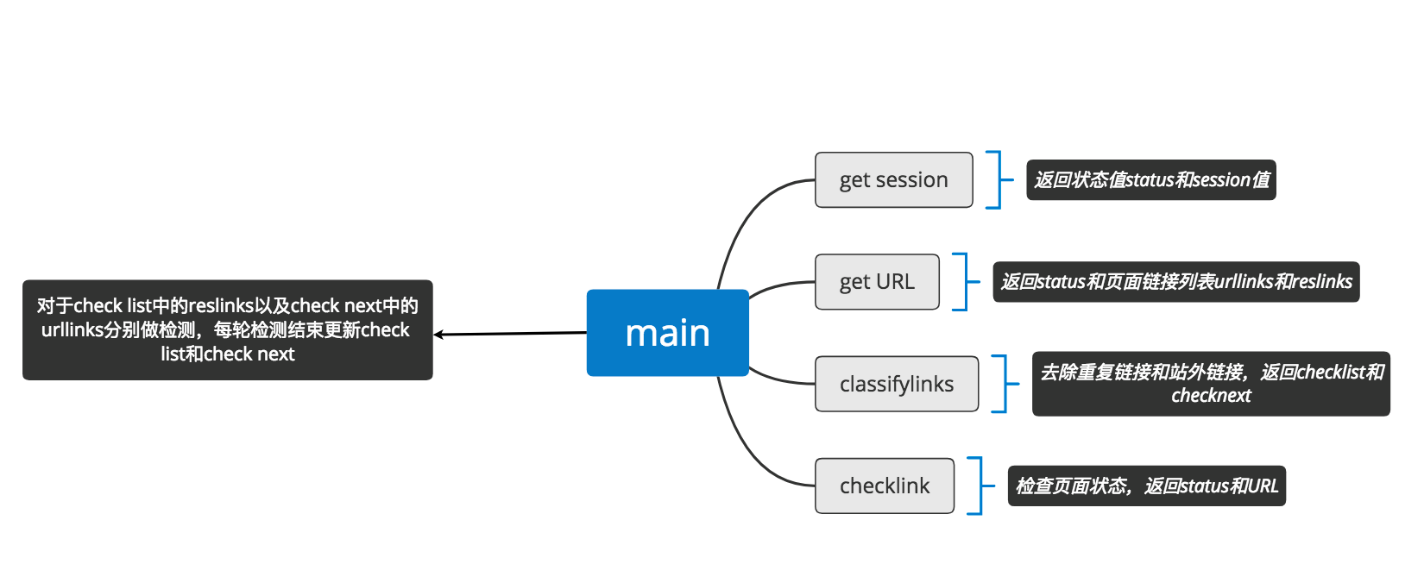


图5-1设计思路图

对于网站的死链接的检查过程中，我首先自动获取了网站链接，实现了连接的自动获取，我以Web页面之间的相互链接为路径，按照广度优先策略，遍历网站中的每一个链接，进行链接初步处理，例如删除链接开头的/’,/”，补全../链接等，并将其存储到本地，以便进行信息的预处理。数据预处理主要包括链接去重和链接筛选，在这一步中去除掉已经check过的链接以及根链接与本站链接不相同的站外链接。在筛选完成后，将链接存储至checklist和check next，然后进行链接有效性的判定，返回链接状态以及url。其中的获取页面链接的算法流程如下：

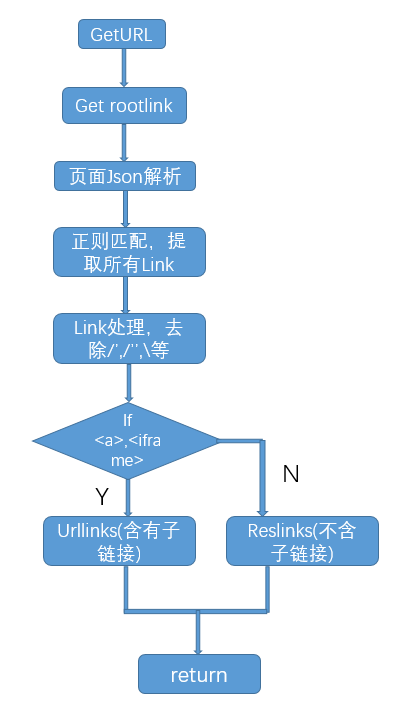


图5-2 GetURL流程图

# 六．实验结果

本次实验选取了一个较为小众的网站进行测试，共检查 33634 个链接，其中有 608 个异常链接。

# 七．总结与反思

在本次链接的检查中，检查了<http://g.yeeyan.org> 网站，其中进行了132层链接的检查，耗时2小时，其实在后续的链接的跳转中，我认为用户可能不需要进行多次跳转，或许跳转五六次之后用户便不再对接下来的页面感兴趣，所以我认为后续实验或许可以调研一下用户一般在页面跳转几次后便不再对页面感兴趣，然后站点检查时只检测这些层的链接就可以。Maybe我需要先看看有没有人做这个的研究。