**软件分析与测试**

Web链接有效性检测

学号：MF1833053

姓名：钱烨

邮箱：[MF1833053@smail.nju.edu.cn](mailto:MF1833053@smail.nju.edu.cn)

**一、问题描述**

在现实生活中，我们常常通过浏览网页来获得或者查找信息。而我们经常访问的网页并不是一个孤立的页面，其中往往包含了许多通向内部和外部的链接。

这些链接通常指向另一个网页或者资源，用户只需要通过点击的操作，就能跳转到该网页或者下载资源。但是有时候我们会发现，点击链接后并没有发生我们期待的事件。而这些链接我们称之为死链接， 也称无效链接，即那些不可达到的链接。一个网站存在死链接不是什么好事，首先一个网站如果存在大量的死链接，必将大大损伤网站的整体形象，导致用户流失。

产生无效链接的情况一般有以下几种：

* **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接。**
* **网站服务器设置错误。**

由上面可以看出，导致无效链接的情况有很多，所以我们要想办法解决这个问题。

**二、解决思路**

为了留住用户，提高用户的使用体验，网站需要定期对网页中的链接进行检查，来排查页面中是否存在无效链接。 但是，由于网页组成复杂，其中的链接更是让人眼花撩乱，如果使用人工去逐个对这些链接进行检查，显然是不现实的。

因此，我想设计一个工具，可以自动地对网页中的所有链接进行检查，并返回其中的死链接，来提供给测试人员有用的信息。

**三、具体实现**

本次实验是用python实现的。

为了最大限度地减少人工投入，这个工具要能够自动获取给定网页中所包含的所有链接；然后逐个对这些链接进行有效性检测。

首先，因为没有现成合适的数据集，所以本实验通过指定网页，然后利用爬虫程序获得网页内容。这些网页内容中就包含了大量的链接，而这些将作为本次实验的数据。

获得数据后，利用正则表达式匹配获得网页中所有链接，得到一个链接列表。然后利用urllib包中提供的工具，逐个向每个链接的服务发出访问请求，然后查看返回信息，确定各个链接的可达性，如果不可达，则却确定该链接为无效链接。

**四、实验结果**

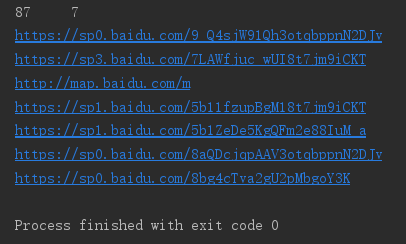
本次实验主要在两个网页上做了测试——CSDN首页和百度首页。实验结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 总连接数 | 有效链接数 | 无效链接数 |
| CSDN | 85 | 73 | 12 |
| 百度 | 94 | 87 | 7 |

最后除了确定无效链接数，测试工具还会返回相应的无效链接，能够让测试人员轻松定位无效链接的位置。

下面是该次实验得到的无效链接：





**五、后续工作与展望**

就当前工作来看，这个测试工具还是略显简陋，但是后续可做的工作还是挺多的。

在问题描述中，介绍了出现无效链接的原因，所以，我觉得可以在后续工作能够直接给出失效链接出现的具体原因。

另外，对于那种存在多个网页的网站，实现能够首页开始，逐步检查整个网站中的所有网页中的链接是否有效，这样也可以提高工具的效率。