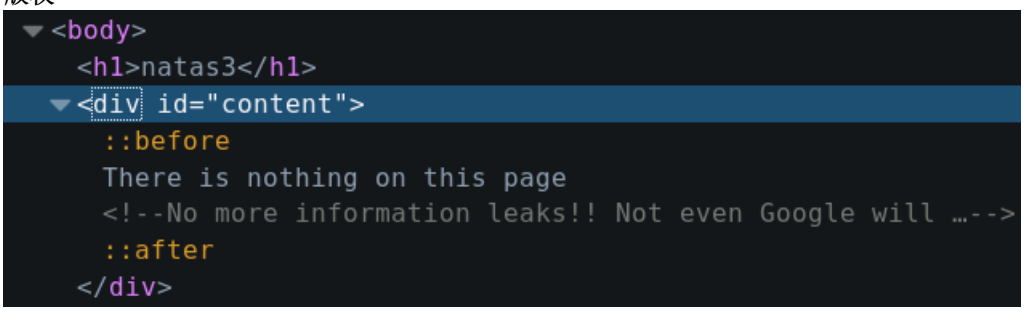
Natas Wargame Level 3 Writeup 与 robots.txt

转载

a_18067 于 2017-05-12 11:54:00 发布 51 收藏
文章标签： 爬虫
原文链接：http://www.cnblogs.com/liqiuhao/p/6844874.html



从HTML的注释代码来看，“google无法搜索到这个页面” -> 说明该网站很可能设置了防爬虫机制：robots.txt

以下是robots.txt的介绍（来自维基百科）：

---

robots.txt（统一小写）是一种存放于网站根目录下的ASCII编码的文本文件，它通常告诉网络搜索引擎的漫游器（又称网络蜘蛛）， 此网站中的哪些内容是不应被搜索引擎的漫游器获取的，哪些是可以被漫游器获取的。因为一些系统中的URL是大小写敏感的，所以robots.txt的文件 名应统一为小写。robots.txt应放置于网站的根目录下。如果想单独定义搜索引擎的漫游器访问子目录时的行为，那么可以将自定的设置合并到根目录下 的robots.txt，或者使用robots元数据（Metadata，又稱元資料）。

robots.txt协议并不是一个规范，而只是约定俗成的，所以并不能保证网站的隐私。注意robots.txt是用字符串比较来确定是否获取URL，所以目录末尾有与没有斜杠“/”表示的是不同的URL。robots.txt允许使用类似"Disallow: *.gif"这样的通配符[1][2]。

其他的影响搜索引擎的行为的方法包括使用robots元数据：

```
<meta name="robots" content="noindex,nofollow" />
```

这个协议也不是一个规范，而只是约定俗成的，有些搜索引擎会遵守这一规范，而其他则不然。通常搜索引擎会识别这个元数据，不索引这个页面，以及这个页面的链出页面。

例子：

允许所有的机器人：

```
User-agent: *
Disallow:
```

另一写法

```
User-agent: *
Allow:/
```

仅允许特定的机器人：（name_spider用真实名字代替）

```
User-agent: name_spider
Allow:
```

拦截所有的机器人：

```
User-agent: *
Disallow: /
```

禁止所有机器人访问特定目录：

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

仅禁止坏爬虫访问特定目录（BadBot用真实的名字代替）：

```
User-agent: BadBot
Disallow: /private/
```

禁止所有机器人访问特定文件类型[2]：

User-agent: * Disallow: /*.php$ Disallow: /*.js$ Disallow: /*.inc$ Disallow: /*.css$

另外还有一些扩展指令。

于是访问根目录下的robots.txt文件：


```
User-agent: *
Disallow: /s3cr3t/
```

得到禁止所有爬虫（spider，crawler）访问的目录s3cr3t/：

Index of /s3cr3t

Name	Last modified	Size	Description
Parent Directory		-	
users.txt	2016-12-20 05:15	40	

Apache/2.4.10 (Debian) Server at natas3.natas.labs.overthewire.org Port 80

得到用户密码：

natas4:Z9tkRkWmpt9Qr7XrR5jWRkgOU901swEZ

总结：网站防止爬虫访问的目录或者文件可能含有敏感文件，这也是一个切入点。

转载于:https://www.cnblogs.com/liqiuhao/p/6844874.html