Robot

1.简介

robots.txt就是一份网站和搜索引擎双方签订的规则协议书。每一个搜索引擎的蜘蛛访问一个站点时，它首先爬行来检查该站点根目录下是否存在robots.txt。如果存在，蜘蛛就会按照该协议书上的规则来确定自己的访问范围；如果没有robots.txt，那么蜘蛛就会沿着链接抓取。

robots.txt必须放置在站点的根目录下，而且文件名必须全部小写。Disallow后面的冒号必须为英文状态的。

User-agent：该项用于描述搜索引擎蜘蛛的名字。

Disallow：该项用于描述不希望被抓取和索引的一个URL，这个URL可以是一条完整的路径。

1. User-agent

有时候我们觉得网站访问量（IP）不多，但是网站流量为什么耗的快？有很多的原因是垃圾（没有）蜘蛛爬行和抓取消耗的。而网站要屏蔽哪个搜索引擎或只让哪个搜索引擎收录的话，首先要知道每个搜索引擎robot的名称。



3.robots.txt文件基本常用写法：

（1）禁止所有搜索引擎访问网站的任何部分。

User-agent: \*

Disallow: /

（2）允许所有的robots访问，无任何限制。

User-agent: \*

Disallow:

或者

User-agent: \*

Allow: /

还可以建立一个空文件robots.txt或者不建立robots.txt。

（3）仅禁止某个搜索引擎的访问（例如：百度baiduspider）

User-agent: BaiduSpider

Disallow:/

（4）允许某个搜索引擎的访问（还是百度）

User-agent: BaiduSpider

Disallow:

User-agent: \*

Disallow: /

这里需要注意，如果你还需要允许谷歌bot，那么也是在“User-agent: \*”前面加上，而不是在“User-agent: \*”后面。

（5）禁止Spider访问特定目录和特定文件（图片、压缩文件）。

User-agent: \*

Disallow: /AAA.net/

Disallow: /admin/

Disallow: .jpg$

Disallow: .rar$

这样写之后，所有搜索引擎都不会访问这2个目录。需要注意的是对每一个目录必须分开说明，而不要写出“Disallow:/AAA.net/ /admin/”。

3.robot特殊参数

（1）Allow

Allow与Disallow是正好相反的功能，Allow行的作用原理完全与Disallow行一样，所以写法是一样的，只需要列出你要允许的目录或页面即可。

Disallow和Allow可以同时使用，例如，需要拦截子目录中的某一个页面之外的其他所有页面，可以这么写：

User-agent: \*

Disallow: /AAA.net/

Allow: /AAA.net/index.html

这样说明了所有蜘蛛只可以抓取/AAA.net/index.html的页面，而/AAA.net/文件夹的其他页面则不能抓取，还需要注意以下错误的写法：

User-agent: \*

Disallow: /AAA.net

Allow: /AAA.net/index.html

原因请看上面Disallow值的定义说明。

（2）使用“\*”号匹配字符序列。

例1.拦截搜索引擎对所有以admin开头的子目录的访问，写法：

User-agent: \*

Disallow: /admin\*/

例2.要拦截对所有包含“?”号的网址的访问，写法：

User-agent: \*

Disallow: /\*?\*

（3）使用“$”匹配网址的结束字符

例1.要拦截以.asp结尾的网址，写法：

User-agent: \*

Disallow:/\*.asp$

例2.如果“:”表示一个会话ID，可排除所包含该ID的网址，确保蜘蛛不会抓取重复的网页。但是，以“?”结尾的网址可能是你要包含的网页版本，写法：

User-agent: \*

Allow: /\*?$

Disallow: /\*?

也就是只抓取.asp?的页面，而.asp?=1，.asp?=2等等都不抓取。

4、robots.txt的好处与坏处（解决方法）。

好处：

（1）有了robots.txt，spider抓取URL页面发生错误时则不会被重定向至404处错误页面，同时有利于搜索引擎对网站页面的收录。

（2）robots.txt可以制止我们不需要的搜索引擎占用服务器的宝贵宽带。

（3）robots.txt可以制止搜索引擎对非公开的爬行与索引，如网站的后台程序、管理程序，还可以制止蜘蛛对一些临时产生的网站页面的爬行和索引。

（4）如果网站内容由动态转换静态，而原有某些动态参数仍可以访问，可以用robots中的特殊参数的写法限制，可以避免搜索引擎对重复的内容惩罚，保证网站排名不受影响。

坏处：

（1）robots.txt轻松给黑客指明了后台的路径。

解决方法：给后台文件夹的内容加密，对默认的目录主文件inde.html改名为其他。

（2）如果robots.txt设置不对，将导致搜索引擎不抓取网站内容或者将数据库中索引的数据全部删除。

User-agent: \*

Disallow: /

这一条就是将禁止所有的搜索引擎索引数据。