链滴

robots.txt 基本介绍

作者：cxJD

原文链接：https://ld246.com/article/1613545713301

来源网站：链滴

在国内，网站管理者似乎对robots.txt并没有引起多大重视，应一些朋友之请求，今天想通过这篇文来简单谈一下robots.txt的写作。

robots.txt基本介绍

robots.txt是一个纯文本文件，在这个文件中网站管理者可以声明该网站中不想被robots访问的部分或者指定搜索引擎只收录指定的内容。

当一个搜索机器人（有的叫搜索蜘蛛）访问一个站点时，它会首先检查该站点根目录下是否存在robots txt，如果存在，搜索机器人就会按照该文件中的内容来确定访问的范围；如果该文件不存在，那么搜机器人就沿着链接抓取。

另外，robots.txt必须放置在一个站点的根目录下，而且文件名必须全部小写。

robots.txt写作语法

首先，我们来看一个robots.txt范例：http://www.seovip.cn/robots.txt

访问以上具体地址，我们可以看到robots.txt的具体内容如下：

Robots.txt file from http://www.seovip.cn

All robots will spider the domain

User-agent: *

Disallow:

以上文本表达的意思是允许所有的搜索机器人访问www.seovip.cn站点下的所有文件。

具体语法分析：其中#后面文字为说明信息；User-agent:后面为搜索机器人的名称，后面如果是*，

泛指所有的搜索机器人；Disallow:后面为不允许访问的文件目录。

下面，我将列举一些robots.txt的具体用法：

允许所有的robot访问

User-agent: *
Disallow:

或者也可以建一个空文件 "/robots.txt" file

禁止所有搜索引擎访问网站的任何部分

User-agent: *
Disallow: /

禁止所有搜索引擎访问网站的几个部分（下例中的01、02、03目录）

User-agent: *
Disallow: /01/
Disallow: /02/
Disallow: /03/

禁止某个搜索引擎的访问（下例中的BadBot）

User-agent: BadBot
Disallow: /

只允许某个搜索引擎的访问（下例中的Crawler）

User-agent: Crawler
Disallow:

User-agent: *
Disallow: /

另外，我觉得有必要进行拓展说明，对robots meta进行一些介绍：

Robots META标签则主要是针对一个个具体的页面。和其他的META标签（如使用的语言、页面的描、关键词等）一样，Robots META标签也是放在页面的 < head > < /head > 中，专门用来告诉搜索擎ROBOTS如何抓取该页的内容。

Robots META标签的写法：

Robots META标签中没有大小写之分，name=”Robots”表示所有的搜索引擎，可以针对某个具体索引擎写为name=”BaiduSpider”。 content部分有四个指令选项：index、noindex、follow、no ollow，指令间以“,”分隔。

INDEX 指令告诉搜索机器人抓取该页面；

FOLLOW 指令表示搜索机器人可以沿着该页面上的链接继续抓取下去；

Robots Meta标签的缺省值是INDEX和FOLLOW，只有inktomi除外，对于它，缺省值是INDEX,NOFLLOW。

这样，一共有四种组合：

< META NAME="ROBOTS" CONTENT="INDEX,FOLLOW" >

< META NAME="ROBOTS" CONTENT="NOINDEX,FOLLOW" >

< META NAME="ROBOTS" CONTENT="INDEX,NOFOLLOW" >

< META NAME="ROBOTS" CONTENT="NOINDEX,NOFOLLOW" >

其中

< META NAME="ROBOTS" CONTENT="INDEX,FOLLOW" >可以写成< META NAME="ROBOTS" CONTENT="ALL" >；

< META NAME="ROBOTS" CONTENT="NOINDEX,NOFOLLOW" >可以写成< META NAME="ROBOTS" CONTENT="NONE" >

目前看来，绝大多数的搜索引擎机器人都遵守robots.txt的规则，而对于Robots META标签，目前支的并不多，但是正在逐渐增加，如著名搜索引擎GOOGLE就完全支持，而且GOOGLE还增加了一个指“archive”，可以限制GOOGLE是否保留网页快照。例如：

< META NAME="googlebot" CONTENT="index,follow,noarchive" >

表示抓取该站点中页面并沿着页面中链接抓取，但是不在GOOLGE上保留该页面的网页快照。

如何使用robots.txt

robots.txt 文件对抓取网络的搜索引擎漫游器（称为漫游器）进行限制。这些漫游器是自动的，在它访问网页前会查看是否存在限制其访问特定网页的 robots.txt 文件。如果你想保护网站上的某些内容被搜索引擎收入的话，robots.txt是一个简单有效的工具。这里简单介绍一下怎么使用它。

如何放置Robots.txt文件

robots.txt自身是一个文本文件。它必须位于域名的根目录中并 被命名为"robots.txt"。位于子目录的 robots.txt 文件无效，因为漫游器只在域名的根目录中查找此文件。例如，http://www.example.com/robots.txt 是有效位置，http://www.example.com/mysite/robots.txt 则不是。

这里举一个robots.txt的例子:

User-agent: *

Disallow: /cgi-bin/

Disallow: /tmp/

Disallow: /~name/

使用 robots.txt 文件拦截或删除整个网站

要从搜索引擎中删除您的网站，并防止所有漫游器在以后抓取您的网站，请将以下 robots.txt 文件放您服务器的根目录：

User-agent: *

Disallow: /

要只从 Google 中删除您的网站，并只是防止 Googlebot 将来抓取您的网站，请将以下 robots.txt 件放入您服务器的根目录：

User-agent: Googlebot

Disallow: /

每个端口都应有自己的 robots.txt 文件。尤其是您通过 http 和 https 托管内容的时候，这些协议都要有各自的 robots.txt 文件。例如，要让 Googlebot 只为所有的 http 网页而不为 https 网页编制引，应使用下面的 robots.txt 文件。

对于 http 协议 (http://yourserver.com/robots.txt):

User-agent: *

Allow: /

对于 https 协议 (https://yourserver.com/robots.txt):

User-agent: *

Disallow: /

允许所有的漫游器访问您的网页

User-agent: *

Disallow:

(另一种方法: 建立一个空的 "/robots.txt" 文件, 或者不使用robot.txt。)

使用 robots.txt 文件拦截或删除网页

您可以使用 robots.txt 文件来阻止 Googlebot 抓取您网站上的网页。 例如，如果您正在手动创建 ro ots.txt 文件以阻止 Googlebot 抓取某一特定目录下（例如，private）的所有网页，可使用以下 robo s.txt 条目：

User-agent: Googlebot

Disallow: /private

要阻止 Googlebot 抓取特定文件类型（例如，.gif）的所有文件，可使用以下 robots.txt 条目：

User-agent: Googlebot

Disallow: /*.gif$

要阻止 Googlebot 抓取所有包含 ? 的网址（具体地说，这种网址以您的域名开头，后接任意字符串然后是问号，而后又是任意字符串），可使用以下条目：

User-agent: Googlebot

Disallow: /*?

尽管我们不抓取被 robots.txt 拦截的网页内容或为其编制索引，但如果我们在网络上的其他网页中发这些内容，我们仍然会抓取其网址并编制索引。因此，网页网址及其他公开的信息，例如指 向该网站链接中的定位文字，有可能会出现在 Google 搜索结果中。不过，您网页上的内容不会被抓取、编制

引和显示。

作为网站管理员工具的一部分，Google提供了robots.txt分析工具。它可以按照 Googlebot 读取 roots.txt 文件的相同方式读取该文件，并且可为 Google user-agents（如 Googlebot）提供结果。们强烈建议您使用它。 在创建一个robots.txt文件之前，有必要考虑一下哪些内容可以被用户搜得到而哪些则不应该被搜得到。 这样的话，通过合理地使用robots.txt, 搜索引擎在把用户带到您网站的时，又能保证隐私信息不被收录。

误区一：我的网站上的所有文件都需要蜘蛛抓取，那我就没必要在添加robots.txt文件了。反正如果文件不存在，所有的搜索蜘蛛将默认能够访问网站上所有没有被口令保护的页面。

每当用户试图访问某个不存在的URL时，服务器都会在日志中记录404错误（无法找到文件）。当搜索蜘蛛来寻找并不存在的robots.txt文件时，服务器也将在日志中记录一条404错误，所以你应做网站中添加一个robots.txt。

误区二：在robots.txt文件中设置所有的文件都可以被搜索蜘蛛抓取，这样可以增加网站的收录。

网站中的程序脚本、样式表等文件即使被蜘蛛收录，也不会增加网站的收录率，还只会浪费服务资源。因此必须在robots.txt文件里设置不要让搜索蜘蛛索引这些文件。

具体哪些文件需要排除， 在robots.txt使用技巧一文中有详细介绍。

误区三：搜索蜘蛛抓取网页太浪费服务器资源，在robots.txt文件设置所有的搜索蜘蛛都不能抓全部的网页。

如果这样的话，会导致整个网站不能被搜索引擎收录。

robots.txt使用技巧

1. 每当用户试图访问某个不存在的URL时，服务器都会在日志中记录404错误（无法找到文件）。每搜索蜘蛛来寻找并不存在的robots.txt文件时，服务器也将在日志中记录一条404错误，所以你应该网站中添加一个robots.txt。

2. 网站管理员必须使蜘蛛程序远离某些服务器上的目录——保证服务器性能。比如：大多数网站务器都有程序储存在“cgi-bin”目录下，因此在robots.txt文件中加入“Disallow: /cgi-bin”是个好意，这样能够避免将所有程序文件被蜘蛛索引，可以节省服务器资源。一般网站中不需要蜘蛛抓取的件有：后台管理文件、程序脚本、附件、数据库文件、编码文件、样式表文件、模板文件、导航图片背景图片等等。

下面是VeryCMS里的robots.txt文件：

User-agent: *

Disallow: /admin/ 后台管理文件

Disallow: /require/ 程序文件

Disallow: /attachment/ 附件

Disallow: /images/ 图片

Disallow: /data/ 数据库文件

Disallow: /template/ 模板文件

Disallow: /css/ 样式表文件

Disallow: /lang/ 编码文件

Disallow: /script/ 脚本文件

3. 如果你的网站是动态网页，并且你为这些动态网页创建了静态副本，以供搜索蜘蛛更容易抓取那么你需要在robots.txt文件里设置避免动态网页被蜘蛛索引，以保证这些网页不会被视为含重复内。

4. robots.txt文件里还可以直接包括在sitemap文件的链接。就像这样：

Sitemap: sitemap.xml

目前对此表示支持的搜索引擎公司有Google, Yahoo, Ask and MSN。而中文搜索引擎公司，显不在这个圈子内。这样做的好处就是，站长不用到每个搜索引擎的站长工具或者相似的站长部分，去交自己的sitemap文件，搜索引擎的蜘蛛自己就会抓取robots.txt文件，读取其中的sitemap路径，接抓取其中相链接的网页。

5. 合理使用robots.txt文件还能避免访问时出错。比如，不能让搜索者直接进入购物车页面。因没有理由使购物车被收录，所以你可以在robots.txt文件里设置来阻止搜索者直接进入购物车页面。

推荐网站：神奇的工作室

推荐网站：野生程序员

推荐公众号：野生程序员基地：yscxyjd