Robots.txt

Michael Borodin
7/23/2014

Contents

Robots Exclusion Standard

The Robot Exclusion Standard, also known as the Robots Exclusion Protocol or Robots.txt Protocol, is a convention to advising cooperating web crawlers and other web robots about accessing all or part of a website which is otherwise publicly viewable. Robots are often used by search engines to categorize and archive web site, or by webmasters to proofread source code. The standard is different from, but can be used in conjunction with, sitemaps, a robot inclusion standard for websites.

About the standard

When a site owner wishes to give instructions to web robots they place a text file called robots.txt in the root of the web site hierarchy (e.g. https://www.example.com/robots.txt). This text file contains the instructions in a specific format (see examples below). Robots that choose to follow the instructions try to fetch this file and read the instructions before fetching any other file from the website site. If this file doesn’t exist, web robots assume that the web owner wishes to provide no specific instructions, and crawl the entire site.

A robots.txt file on a website will function as a request that specified robots ignore specified files or directories when crawling a site. This might be, for example, out of a preference for privacy from search engine results, or the belief that the content of the selected directories might be misleading or irrelevant to the categorizations for the site as a whole, or out of a desire that an application only operate on certain data. Links to pages listed in robots.txt can still appear in search results if they are linked to from a page that is crawled.

A robots.txt file covers one origin. For websites with multiple subdomains, each subdomain must have its own robots.txt file. If example.com had a robots.txt file but a.example.com did not, the rules that would apply for example.com would not apply to a.example.com. In addition, each protocol and port needs its own robots.txt file.

Some major search engines following this standard include ASK, AOL, Baidu, Bing, Google, Yahoo, and Yandex.

Disadvantages

Despite the use of the terms "allow" and "disallow", the protocol is purely advisory. It relies on the cooperation of the web robot, so that marking an area of a site out of bounds with robots.txt does not guarantee exclusion of all web robots. In particular, malicious web robots are unlikely to honor robots.txt; some may even use the robots.txt as a guide and go straight to the disallowed URLs.

How to use the robots.txt file for SEO

There are several best practices that should first be covered:

- As a general rule, the robots.txt file should never be used to handle duplicate content. There are better ways
- Disallow statements within the robots.txt file are hard directives, not hints, and should be thought of as such. Directives here are akin to using a sledgehammer.
- No equity will be passed through URLs blocked by robots.txt. Keep this in mind when dealing with duplicate content (see above)
- Using robots.txt file to disallow URLs will not prevent them from being displayed in Google's search engine (see below for details)
- When Googlebot is specified as a user agent, all preceding rules are ignored and the subsequent rules are followed.

```
User-Agent: *
Disallow: /
```

However, this example of directives applies differently to all user agents, and Googlebot, respectively:

```
User-Agent: *
Disallow: /
User-Agent: Googlebot
Disallow: /cgi-bin/
```

Use care when disallowing content. Use of the following syntax will block the directory /folder-of-stuff/ and everything located within it (including subsequent folders and assets):

```
Disallow: /folder-of-stuff/
```

Limited use of regular expression is supported. This means that you can use wildcards to block all content with a specific extension, for example, such as the following directive which will block PowerPoint's:

```
Disallow: *.ppt$
```

Always remember that robots.txt is a sledgehammer and is not subtle. There are often other tools at your disposal that can do a better job of influencing how search engines crawl, such as the parameter handling tools within Google and Bing Webmaster Tools, the Meta robots tag, and the x-robots-tag response header.

Contents of the robots.txt file

Fig 1

Start each statement with the robots (User agents) you wish to enforce the rule on

User-agent:

*	The user-agent is most often represented with a wildcard (*) which is an asterisk sign that signifies that the blocking instructions are for all robots.
Robot-Name	If you want certain robots to be blocked or allowed on certain pages, you can specify the bot name under the user-agent directive.

Fig 2

Give the robots (user agents a permission) disallow them to crawl (nothing or specific files or directories).

Disallow:

Blank	When disallow has nothing specified it means that the robots can crawl all the pages on a site.
/	Tell to the robots to leave the site alone and not to crawl
/Sub-directory/	Tell the robots not to crawl the specified sub-directory and everything in that sub-directory
/Sub-directory/File-Name/File-Extension	Tell the robots not to crawl a specified file

Fig 3

Specific to the robots the seconds in between page request.

Crawl-Delay:

Blank	Tell the robots that the crawl delay is at default
5	Give the robot 5 seconds in between page request
10	Give the robot 10 seconds in between page request
20	Give the robot 20 seconds in between page request
60	Give the robot 60 seconds in between page request
120	Give the robot 120 seconds in between page request

Fig 4

To provide bots with the location of your Sitemap. To do this, enter a directive in your robots.txt that includes the location of your Sitemap

Sitemap:

http://yoursite.com/sitemap-location.xml

Examples

To allow all bots to access the whole site (the default robots.txt) the following is used:

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

To block the entire server from the bots, this robots.txt is used:

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

To allow a single robot and disallow other robots:

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

To block the site from a single robot:

```
User-agent: XYZbot
 Disallow: /
```

To block some parts of the site:

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

Use this robots.txt to block all content of a specific file type. In this example we are excluding all files that are PowerPoint files. (NOTE: The dollar ($) sign indicates the end of the line):

```
User-agent: *
 Disallow: *.ppt$
```

To block bots from a specific file:

```
User-agent: *
 Disallow: /directory/file.html
```

To crawl certain HTML documents in a directory that is blocked from bots you can use an Allow directive. Some major crawlers support the Allow directive in robots.txt. An example is shown below:

```
User-agent: *
 Disallow: /folder/
 Allow: /folder1/myfile.html
```

To block URLs containing specific query strings that may result in duplicate content, the robots.txt below is used. In this case, any URL containing a question mark (?) is blocked:

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

Sometimes a page will get indexed even if you include in the robots.txt file due to reasons such as being linked externally. In order to completely block that page from being shown in search results, you can include robots noindex Meta tags on those pages individually. You can also include a nofollow tag and instruct the bots not to follow the outbound links by inserting the following codes:
For the page not to be indexed:

```
<meta name=“robots” content=“noindex”>
```

For the page not to be indexed and links not to be followed:

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

NOTE: If you add these pages to the robots.txt and also add the above Meta tag to the page, it will not be crawled but the pages may appear in the URL-only listings of search results, as the bots were blocked specifically from reading the Meta tags within the page.

Another important thing to note is that you must not include any URL that is blocked in your robots.txt file in your XML sitemap. This can happen, especially when you use separate tools to generate the robots.txt file and XML sitemap. In such cases, you might have to manually check to see if these blocked URLs are included in the sitemap. You can test this in your Google Webmaster Tools account if you have your site submitted and verified on the tool and have submitted your sitemap.

Important Rules

- In most cases, Meta robots with parameters "noindex, follow" should be employed as a way to restrict crawling or indexation.
- It is important to note that malicious crawlers are likely to completely ignore robots.txt and as such, this protocol does not make a good security mechanism.
- Only one "Disallow:" line is allowed for each URL.
- Each subdomain on a root domain uses separate robots.txt files.
- Google and Bing accept two specific regular expression characters for pattern exclusion (* and $).
- The filename of robots.txt is case sensitive. Use "robots.txt", not "Robots.TXT."
- Spacing is not an accepted way to separate query parameters. For example, "/category/ /product page" would not be honored by robots.txt.

Appendix A: Specific Robots

Search Engine	Robot Name
Google	googlebot
MSN Search	msnbot
Yahoo	yahoo-slurp
Ask/Teoma	teoma
Cuil	twiceler
CigaBlast	gigabot
Scrub The Web	scrubby
DMOZ Checker	robozilla
Nutch	nutch
Alexa/Wayback	Ia_archiver
Baidu	baiduspider
Naver	naverbot,yeti

Appendix B: Specific Special Bots

Special Robot	Robot Name
Google Image	googlebot-image
Google Mobile	**googlebot-mobile**
Yahoo MM	yahoo-mmcrawler
MSN PicSearch	**psbot**
SingingFish	asterias
Yahoo Blogs	**yahoo-blogs/v3.9**

Links

http://blog.woorank.com/2013/04/robots-txt-a-beginners-guide/

http://moz.com/learn/seo/robotstxt

http://www.mcanerin.com/EN/search-engine/robots-txt.asp

http://searchenginewatch.com/article/2064412/Proper-SEO-and-the-Robots.txt-File

http://www.rimmkaufman.com/blog/robots-txt-best-practices-for-seo/03072012/

http://en.wikipedia.org/wiki/Robots_exclusion_standard#About_the_standard

http://www.highrankings.com/forum/index.php/topic/40716-what-is-the-point-of-using-crawl-delay-in-robotstxt/

http://stackoverflow.com/questions/17377835/robots-txt-what-is-the-proper-format-for-a-crawl-delay-for-multiple-user-agent

https://www.drupal.org/node/14177

http://www.webhostingtalk.com/showthread.php?t=1147278

http://blog.woorank.com/2013/03/all-about-xml-sitemaps/

http://www.robotstxt.org/meta.html