

Slowing down the crawlers

Contributed by Datagod
Thursday, 15 May 2008
Last Updated Thursday, 20 August 2009

My sites routinely get hammered by bots and crawlers. Some are malicious, attempting to find vulnerabilities in unpatched software. Some are link spammers that send fake referrer information containing URL's back to some crappy site selling cheap drugs.

Other crawlers (Google, Yahoo, etc.) are totally legitimate, indexing the pages as they crawl them.

In any case, I want to slow down the page loads for the bots and crawlers to prevent them from overwhelming the webserver. In particular, MyTinyStats has thousands of webpages that are generated via combination of PHP and complex MySQL queries.

I do this with the following code:

```
$BotList = file_get_contents("lookup/BotList.txt");

IF (strstr($BotList,$_SERVER["HTTP_USER_AGENT"]))
{
    sleep(30);
}
```

The BotList.txt file contains a list of 350+ known bots and crawlers that I have identified over the past year. This file gets loaded into a local variable using the "file_get_contents()" function.

I then examine the BotList, searching for the current UserAgent. If found, I slow the page down by sleeping for 30 seconds. If not found, the rest of the page loads.

The file is approximately 23Kb and takes only a few milliseconds to load and parse.

While using this approach, I made an interesting discovery: file_get_contents() accepts a URL as a file parameter.

```
$BotList = file_get_contents("http://mytinystats.com/lookup/BotList.txt");

IF (strstr($BotList,$_SERVER["HTTP_USER_AGENT"]))
{
```

```
sleep(30);
```

```
}
```

Specifying a URL however introduces a 3 second overhead!!

I highly recommend using a relative file name and not a URL, if at all possible.