

Web-Server: Bots & Spider per mod_rewrite aussperren

Problem:

Viele (böartige) Bots beachten die `robots.txt` nicht.

Andere (nicht böartige) Bots sehen Änderungen darin leider erst (viel viel viel) später weil sie diese nicht ständig kontrollieren.

Dadurch wird jedesmal hoher Traffic verursacht, Apache oder gar die Datenbank zu stark belastet.

Lösung:

Wir sperren die Bots dauerhaft per `mod_rewrite` aus.

Voraussetzung: `mod_rewrite` ist bereits installiert und lauffähig im Apache aktiviert.

Basis:

Grundsätzlich ist es egal, wo die Regeln eingebaut werden. Mögliche Orte: `VirtualHost`, `Directory` oder `.htaccess`.

(Wie immer der Hinweis: `.htaccess` ist ein Performance-Fresser!)

Das Grundgerüst bilden folgende Zeilen:

```
RewriteEngine On
# hier dazwischen kommen die u.g. Conditions
#...
# und den Abschluß macht die Rule:
RewriteRule ^.* - [F,L]
```

Bekannt 'böse' Bots (Liste von Server-Wissen.de):

```
RewriteCond %{HTTP_USER_AGENT} ^Alexibot [OR]
RewriteCond %{HTTP_USER_AGENT} ^asterias [OR]
RewriteCond %{HTTP_USER_AGENT} ^BackDoorBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^Black [OR]
RewriteCond %{HTTP_USER_AGENT} ^BlowFish [OR]
RewriteCond %{HTTP_USER_AGENT} ^BotALot [OR]
RewriteCond %{HTTP_USER_AGENT} ^BuiltBotTough [OR]
```

Web-Server: Bots & Spider per mod_rewrite aussperren

```
RewriteCond %{HTTP_USER_AGENT} ^Bullseye [OR]
RewriteCond %{HTTP_USER_AGENT} ^BunnySlippers [OR]
RewriteCond %{HTTP_USER_AGENT} ^Cegbfeieh [OR]
RewriteCond %{HTTP_USER_AGENT} ^CheeseBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^CherryPicker [OR]
RewriteCond %{HTTP_USER_AGENT} ^ChinaClaw [OR]
RewriteCond %{HTTP_USER_AGENT} ^Convera [OR]
RewriteCond %{HTTP_USER_AGENT} ^CopyRightCheck [OR]
RewriteCond %{HTTP_USER_AGENT} ^cosmos [OR]
RewriteCond %{HTTP_USER_AGENT} ^Crescent [OR]
RewriteCond %{HTTP_USER_AGENT} ^Custo [OR]
RewriteCond %{HTTP_USER_AGENT} ^DataFountains [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^DISCo [OR]
RewriteCond %{HTTP_USER_AGENT} ^DittoSpyder [OR]
RewriteCond %{HTTP_USER_AGENT} ^eCatch [OR]
RewriteCond %{HTTP_USER_AGENT} ^Email [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Express WebPictures [OR]
RewriteCond %{HTTP_USER_AGENT} ^Extractor [OR]
RewriteCond %{HTTP_USER_AGENT} ^EyeNetIE [OR]
RewriteCond %{HTTP_USER_AGENT} ^FlashGet [OR]
RewriteCond %{HTTP_USER_AGENT} ^Foobot [OR]
RewriteCond %{HTTP_USER_AGENT} ^GetRight [OR]
RewriteCond %{HTTP_USER_AGENT} ^GetWeb! [OR]
RewriteCond %{HTTP_USER_AGENT} ^Global Confusion [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Go-Ahead-Got-It [OR]
RewriteCond %{HTTP_USER_AGENT} ^Go!Zilla [OR]
RewriteCond %{HTTP_USER_AGENT} ^GrabNet [OR]
RewriteCond %{HTTP_USER_AGENT} ^Grafula [OR]
RewriteCond %{HTTP_USER_AGENT} ^hloader [OR]
RewriteCond %{HTTP_USER_AGENT} ^HMView [OR]
RewriteCond %{HTTP_USER_AGENT} ^httpLib [OR]
RewriteCond %{HTTP_USER_AGENT} ^ia_archiver [OR]
RewriteCond %{HTTP_USER_AGENT} ^IBM_Planetwide [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Image [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Indy Library [OR]
RewriteCond %{HTTP_USER_AGENT} ^InterGET [OR]
RewriteCond %{HTTP_USER_AGENT} ^Internet Ninja [OR]
RewriteCond %{HTTP_USER_AGENT} ^Jakarta [OR]
RewriteCond %{HTTP_USER_AGENT} ^JennyBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^JetCar [OR]
RewriteCond %{HTTP_USER_AGENT} ^Kenjin [OR]
RewriteCond %{HTTP_USER_AGENT} ^Keyword [OR]
RewriteCond %{HTTP_USER_AGENT} ^LexiBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^LibWeb [OR]
RewriteCond %{HTTP_USER_AGENT} ^lwp [OR]
RewriteCond %{HTTP_USER_AGENT} ^Lynx [OR]
RewriteCond %{HTTP_USER_AGENT} ^Mata [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Microsoft.URL [OR]
```

Web-Server: Bots & Spider per mod_rewrite aussperren

```
RewriteCond %{HTTP_USER_AGENT} ^MIDown tool [OR]
RewriteCond %{HTTP_USER_AGENT} ^MIIXpc [OR]
RewriteCond %{HTTP_USER_AGENT} ^Mister [OR]
RewriteCond %{HTTP_USER_AGENT} ^moget [OR]
RewriteCond %{HTTP_USER_AGENT} ^Navroad [OR]
RewriteCond %{HTTP_USER_AGENT} ^NearSite [OR]
RewriteCond %{HTTP_USER_AGENT} ^Net [OR]
RewriteCond %{HTTP_USER_AGENT} ^NICERSPRO [OR]
RewriteCond %{HTTP_USER_AGENT} ^NPBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^Octopus [OR]
RewriteCond %{HTTP_USER_AGENT} ^Openfind [OR]
RewriteCond %{HTTP_USER_AGENT} ^Papa Foto [OR]
RewriteCond %{HTTP_USER_AGENT} ^pavuk [OR]
RewriteCond %{HTTP_USER_AGENT} ^pcBrowser [OR]
RewriteCond %{HTTP_USER_AGENT} ^ProPowerBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^ProWebWalker [OR]
RewriteCond %{HTTP_USER_AGENT} ^QueryN [OR]
RewriteCond %{HTTP_USER_AGENT} ^ReGet [OR]
RewriteCond %{HTTP_USER_AGENT} ^RepoMonkey [OR]
RewriteCond %{HTTP_USER_AGENT} ^RMA [OR]
RewriteCond %{HTTP_USER_AGENT} ^SiteSnagger [OR]
RewriteCond %{HTTP_USER_AGENT} ^SlySearch [OR]
RewriteCond %{HTTP_USER_AGENT} ^Snoopy [OR]
RewriteCond %{HTTP_USER_AGENT} ^SpankBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^spanner [OR]
RewriteCond %{HTTP_USER_AGENT} ^Super [OR]
RewriteCond %{HTTP_USER_AGENT} ^Surfbot [OR]
RewriteCond %{HTTP_USER_AGENT} ^suzuran [OR]
RewriteCond %{HTTP_USER_AGENT} ^tAkeOut [OR]
RewriteCond %{HTTP_USER_AGENT} ^Teleport [OR]
RewriteCond %{HTTP_USER_AGENT} ^Telesoft [OR]
RewriteCond %{HTTP_USER_AGENT} ^The.Intraformant [OR]
RewriteCond %{HTTP_USER_AGENT} ^TheNomad [OR]
RewriteCond %{HTTP_USER_AGENT} ^TightTwatBot [OR]
RewriteCond %{HTTP_USER_AGENT} ^Titan [OR]
RewriteCond %{HTTP_USER_AGENT} ^turingos [OR]
RewriteCond %{HTTP_USER_AGENT} ^TurnitinBot [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^URLy.Warning [OR]
RewriteCond %{HTTP_USER_AGENT} ^VCI [OR]
RewriteCond %{HTTP_USER_AGENT} ^VoidEYE [OR]
RewriteCond %{HTTP_USER_AGENT} ^web [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Wget [OR]
RewriteCond %{HTTP_USER_AGENT} ^Widow [OR]
RewriteCond %{HTTP_USER_AGENT} ^www [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ^Xaldon [OR]
RewriteCond %{HTTP_USER_AGENT} ^Zeus [OR]
```

Web-Server: Bots & Spider per mod_rewrite aussperren

Erweiterungen:

Und hier noch ein paar kritischere Regeln. Evtl. wird hier aber zu viel verboten. Dies muß jeder selber wissen.

```
RewriteCond %{HTTP_USER_AGENT} collect [NC,OR]
RewriteCond %{HTTP_USER_AGENT} crawl [NC,OR]
RewriteCond %{HTTP_USER_AGENT} download [NC,OR]
RewriteCond %{HTTP_USER_AGENT} francis [NC,OR]
RewriteCond %{HTTP_USER_AGENT} grabb [NC,OR]
RewriteCond %{HTTP_USER_AGENT} harvest [NC,OR]
RewriteCond %{HTTP_USER_AGENT} httrack [NC,OR]
RewriteCond %{HTTP_USER_AGENT} larbin [NC,OR]
RewriteCond %{HTTP_USER_AGENT} leech [NC,OR]
RewriteCond %{HTTP_USER_AGENT} libwww [NC,OR]
RewriteCond %{HTTP_USER_AGENT} majestic [NC,OR]
RewriteCond %{HTTP_USER_AGENT} ng-search [NC,OR]
RewriteCond %{HTTP_USER_AGENT} nutch [NC,OR]
RewriteCond %{HTTP_USER_AGENT} offline [NC,OR]
RewriteCond %{HTTP_USER_AGENT} omni [NC,OR]
RewriteCond %{HTTP_USER_AGENT} robot [NC,OR]
RewriteCond %{HTTP_USER_AGENT} suck [NC,OR]
RewriteCond %{HTTP_USER_AGENT} sohu [NC,OR]
```

Folgende Einträge sind gefälschte Browserkennungen die einen normalen User vorgaukeln sollen:

```
RewriteCond %{HTTP_USER_AGENT} ^MSIE 6.0 [OR]
RewriteCond %{HTTP_USER_AGENT} ^Mozilla/4.0 (compatible; MSIE 6.0; Win32) [OR]
RewriteCond %{HTTP_USER_AGENT} MSIE 6.0b [OR]
```

Nun noch ein paar Security-Tips gegen XSS-Angriffe:

```
RewriteCond %{QUERY_STRING} .*'.* [OR]
RewriteCond %{QUERY_STRING} .*%27.* [OR]
RewriteCond %{QUERY_STRING} .*"'.* [OR]
RewriteCond %{QUERY_STRING} .*%22.* [OR]
RewriteCond %{QUERY_STRING} .*`.* [OR]
RewriteCond %{QUERY_STRING} .*%60.* [OR]
RewriteCond %{QUERY_STRING} .*%25.* [OR]
RewriteCond %{QUERY_STRING} .*echr.* [OR]
RewriteCond %{QUERY_STRING} .*esystem.* [OR]
RewriteCond %{QUERY_STRING} .*passthru.* [OR]
```

Web-Server: Bots & Spider per mod_rewrite aussperren

RewriteCond %{QUERY_STRING} .*wget.*

Eindeutige ID: #1285

huschi

2008-02-29 15:02