

# Como configurar Apache para bloquear agentes de usuario indeseables

© Copyright 1999-2003 Joel Barrios Dueñas.

Se otorga el permiso para copiar, distribuir y modificar este documento bajo los términos de la [Licencia de Documentación Libre del GNU](#), versión 1.1 o cualquier otra posterior publicada por [Free Software Foundation](#); este documento no consta de secciones invariantes ni con portadas o contraportadas. [Una copia de la licencia](#) se incluye en la documentación del sitio.

Linux® es una marca registrada de Linus Torvalds, LinuxPPP® es una marca registrada de José Neif Jury Fabre, RedHat® Linux, RPM y GLINT son marcas registradas de RedHat Software, Unix® es marca registrada de X/Open. Apache® es una marca registrada de The Apache Group. Otras marcas registradas son propiedad intelectual de sus autores o propietarios.

La información contenida en este documento se proporciona tal cual es y el autor no asumirá responsabilidad alguna por el mal uso que se haga de esta.

## Introducción.

Los [robots](#) para capturar direcciones de correo, a su vez aprovechados para enviar Spam, y el abuso de ciertos agentes de usuario, como Teleport Pro, continua siendo el dolor de cabeza de todos los administradores de servidores de correo alrededor del mundo.

Los robots especializados en la captura de direcciones de correo electrónico provenientes de nuestros sitios web se ha convertido en un lucrativo negocio para empresas que distribuyen a nivel mundial discos CD-ROM repletos de estas. El principal problema para el administrador consiste en proteger las direcciones de correo electrónico que los visitantes regulares podrían dejar en algún foro de discusión o tablón de mensajes.

Otro problema aún mayor es el abuso de agentes de descarga de sitios web completos a los discos duros locales de los usuarios. Este tipo de actividad no sería tan problemático si no fuese debido al excesivo consumo de ancho de banda y al increíblemente estúpida ocurrencia de los usuarios de que descargarán más rápidamente un sitio web haciéndolo hasta con **diez o veinte hilos simultáneos**. Esto significa problemas para aquellos sitios web que utilizan bases de datos, debido a que este tipo de clientes llegan a abrir hasta cientos de conexiones simultáneas conllevando al bloqueo de la base de datos, y subsecuente *Denial of Service* que nos perjudicará al no permitir que el resto de los usuarios puedan acceder a nuestros sitios web de manera legítima.

Una forma de combatir estos molestos fenómenos consiste en bloquear el acceso de robots y agentes de usuario desde nuestros servidores web.

# Los procedimientos

## Spambots

**Nota:** Para mayores referencias sobre este método, por favor consulte uno de los artículo original (en inglés) que nosotros hemos consultado para la elaboración de este manual, y que se encuentra localizado en [http://evolt.org/article/Using\\_Apache\\_to\\_stop\\_bad\\_robots/18/15126/](http://evolt.org/article/Using_Apache_to_stop_bad_robots/18/15126/).

Determinar que agentes de usuario están accediendo a un sitio web no es complicado, y solo bastará con revisar el fichero de registro de acceso de Apache, regularmente localizado en `/var/log/httpd/access_log`. Desde este podemos examinar que agentes de usuario han sido utilizados para acceder al servidor.

Los agentes de descarga o copia web son fáciles de identificar, pues regularmente llevan nombres descriptivos (Web COpies, WebStreaper, WebReaper), o bien se pueden consultar en [Robotstxt.org](http://Robotstxt.org) para una [lista detallada por tipo](#) y así no confundir un *robot* útil con uno perjudicial. Lo más difícil es determinar que Agentes de Usuario (User Agents) o clientes se están utilizando para realizar la captura de direcciones de correo. Sin embargo puede establecerse una "trampa" utilizando el fichero robots.txt. Esta consiste en añadir la siguiente línea:

```
Disallow: /email-addresses/
```

El directorio `/email-addresses/` no debe existir. A diferencia de los robots útiles y que nos llevarán valioso tráfico, como los de los indexadores de los sitios de búsqueda (no queremos bloquear los robots de Google ni otros buscadores), la mayoría de los robots utilizados para capturar direcciones de correo no respetan las reglas establecidas en robots.txt y suelen buscar y acceder directorios que puedan contener direcciones a como de lugar. Debe de esperarse al menos un par de semanas, o, mejor aún, un mes. Pasado este tiempo, puede revisarse el contenido de `/var/log/httpd/access_log` y revisar aquellas líneas que indicarán que clientes fueron utilizados para acceder a `/email-addresses/`.

Puede utilizarse el siguiente comando para determinar quienes accedieron a `/email-addresses/`:

```
grep /email-addresses access_log | awk '{print $12}' |  
uniq
```

O bien:

```
cat access_log |grep email-address
```

Ejemplo:

```
216.219.236.81 - - [24/Oct/2001:10:45:16 -0600] "GET  
/email-addresses/ HTTP/1.0" 404 294 "-" "EmailWolf"
```

Una vez determinados los culpables, solo hay que editar `/etc/httpd/conf/httpd.conf` y algunas líneas que califiquen a los clientes justo encima de `<Directory "/var/www/html">`. Ejemplo:

```
SetEnvIfNoCase User-Agent "^EmailSiphon" bad_bot  
SetEnvIfNoCase User-Agent "^EmailWolf" bad_bot
```

Esto establece como variable **bad\_bot** a cualquier cliente que acceda a nuestro sitio web y que se identifique como EmailSiphon o EmailWolf. Por favor recuerde que se requiere se encuentre habilitado el módulo de Apache [mod\\_setenvif](#). A continuación añadimos **Deny from env=bad\_bot** a la configuración de nuestro directorio raíz de Apache:

```
<Directory "/var/www/html">
  Options Indexes Includes FollowSymLinks
  AllowOverride None
  Order allow,deny
  Allow from all
  Deny from env=bad_bot
</Directory>
```

Esto establece que se denegará el acceso, devolviendo un error 403 común y corriente, a cualquier cliente que sea clasificado como **bad\_bot**.

En resumen, esta sería un ejemplo de toda la configuración a aplicar:

```
SetEnvIfNoCase User-Agent "^EmailSiphon" bad_bot
SetEnvIfNoCase User-Agent "^EmailWolf" bad_bot

<Directory "/var/www/html">
  Options Indexes Includes FollowSymLinks
  AllowOverride None
  Order allow,deny
  Allow from all
  Deny from env=bad_bot
</Directory>
```

## Agentes de copiado o descarga web

Como es sabido, es común que en sitios web, que hacen utilizar MySQL, en ocasiones se bloquee el acceso a la base de datos, devolviendo el mensaje:

```
Warning: Host 'tu_dominio.com' is blocked because of many connection errors. Unblock with 'mysqladmin flush-hosts'
```

En muchos casos es debido a que no se trata de la única aplicación con que se accede a una misma base de datos y se utiliza *mysql\_pconnect()* (conexión persistente) en lugar de un más saludable y convencional *mysql\_connect()*, algo ya bien conocido por usuarios de PHP-Nuke. Pero cuando usamos el parámetro de conexión correcto y aún así sufrimos de recurrentes bloqueos de bases de datos, esto es culpa total de los clientes de copia web como WebZIP, eCatch, WebReaper, WebStripper, WebCopier y otros *robots* como Scooter-W3-1.0. Veamos entonces por que y cómo solucionar este problema.

Dichos clientes suelen abrir (de manera simultánea) **cientos de páginas en solo unos minutos**, y por lo tanto estableciendo **docenas y hasta cientos de conexiones a MySQL**. La utilización de dichos clientes sobre sitios web que utilizan MySQL puede ser definitivamente abusivo de cualquier modo que se quiera ver.

¿Cómo se puede solucionar esto? La respuesta es bloqueando el acceso a nuestros web, o secciones críticas que utilicen MySQL, a dichos clientes. Esto se puede hacer con Apache en el fichero */etc/httpd/conf/httpd.conf* o bien desde *robots.txt*.

Bloquear desde `httpd.conf` puede hacerse del mismo modo que establecimos en el punto anterior, sobre como bloquear los *spambots* o colectores de direcciones de correo.

```
SetEnvIfNoCase User-Agent "^EmailSiphon" bad_bot
SetEnvIfNoCase User-Agent "^EmailWolf" bad_bot
SetEnvIfNoCase User-Agent "^WebZIP" bad_bot
SetEnvIfNoCase User-Agent "^WebStripper" bad_bot
SetEnvIfNoCase User-Agent "^Teleport Pro" bad_bot
SetEnvIfNoCase User-Agent "^eCatch" bad_bot
SetEnvIfNoCase User-Agent "^WebCopier" bad_bot
SetEnvIfNoCase User-Agent "^Wget" bad_bot

<Directory "/var/www/html">
    Options Indexes Includes FollowSymLinks
    AllowOverride None
    Order allow,deny
    Allow from all
    Deny from env=bad_bot
</Directory>
```

Para *robots.txt*, otro método menos efectivo pero más simple y tolerante, consiste en poner un fichero denominado *robots.txt* en el directorio raíz del sitio web. Dicho fichero es utilizado de manera estándar por los robots (o webbots) para determinar que directorios acceder o no acceder para distintas funciones, como indexar sitios web en los motores de búsqueda. A los clientes de copia web, por lo general, se les añade también soporte para lectura de *robots.txt*, como una manera de evitar el abuso sobre sitios web.

En este fichero se pueden establecer reglas para distintos clientes. Dicho fichero puede definir a un agente en particular de y una regla específica para éste. Ejemplo:

```
User-agent: WebZIP
Disallow: /
```

Lo anterior define que para cualquier cliente identificado como "WebZIP" no le estará permitido indexar (y por lo tanto descargar) el contenido de todo el sitio web. Otro ejemplo:

```
User-agent: WebStripper
Disallow: /phpnuke
```

Lo anterior especifica que cualquier cliente identificado como WebStripper no le estará permitido indexar (y por lo tanto descargar) el contenido de */phpnuke* en el sitio web.

Ahora, poniendo en práctica lo anterior, el siguiente sería el ejemplo de como se puede mantener saludable nuestros sitios con PHP + MySQL:

```
User-agent: WebZIP
Disallow: /

User-agent: WebStripper
Disallow: /

User-agent: Teleport Pro
Disallow: /

User-agent: Wget
Disallow: /

User-agent: eCatch
Disallow: /

User-agent: WebCopier
Disallow: /
```

Hay varias docenas de clientes más que pueden añadirse de este mismo modo. No hay nada de malo en bloquear dichos clientes, siendo que no son visualizadores web, sino herramientas de descarga que no solo no hacen un uso ético de los recursos web sino que abusan de ellos, y que además, en varios casos particulares (no el nuestro en LPT), violan leyes de derechos de autor.

¿Por que bloquear el accesos a todo el sitio web? Simple: dichos clientes hacen que se consuma más ancho de banda del necesario, y además es raro que los usuarios que utilizan estas herramientas realmente vean TODO el material que descargaron. Simplemente es contenido web que almacena en el disco duro y que puede que nunca lleguen siquiera a consultar. Así que de cualquier modo que se quiera ver, es un desperdicio de valioso y costoso ancho de banda permitir a los visitantes utilizar dichos clientes sobre nuestros sitios web.

## El fichero Htaccess

También es posible habilitar filtrado de agentes de usuario editando y añadiendo parámetros en el fichero `.htaccess` del directorio raíz. Ejemplo:

```
SetEnvIfNoCase User-Agent "^Bloodhound" bad_bot
SetEnvIfNoCase User-Agent "^eCatch" bad_bot
SetEnvIfNoCase User-Agent "^GetRight" bad_bot
SetEnvIfNoCase User-Agent "^LeechFTP" bad_bot
SetEnvIfNoCase User-Agent "^Mass Downloader" bad_bot
SetEnvIfNoCase User-Agent "^Prozilla" bad_bot
SetEnvIfNoCase User-Agent "^Offline Explorer" bad_bot
SetEnvIfNoCase User-Agent "^RealDownload" bad_bot
SetEnvIfNoCase User-Agent "^SiteSnagger" bad_bot
SetEnvIfNoCase User-Agent "^Teleport Pro" bad_bot
SetEnvIfNoCase User-Agent "^WebCopier" bad_bot
SetEnvIfNoCase User-Agent "^Web Downloader" bad_bot
SetEnvIfNoCase User-Agent "^webfetcher" bad_bot
SetEnvIfNoCase User-Agent "^WebFountain" bad_bot
SetEnvIfNoCase User-Agent "^Wget" bad_bot
SetEnvIfNoCase User-Agent "^WebMirror" bad_bot
SetEnvIfNoCase User-Agent "^WebReaper" bad_bot
SetEnvIfNoCase User-Agent "^WebStripper" bad_bot
SetEnvIfNoCase User-Agent "^WebZIP" bad_bot
SetEnvIfNoCase User-Agent "^e-collector" bad_bot
SetEnvIfNoCase User-Agent "^EmailSiphon" bad_bot
SetEnvIfNoCase User-Agent "^EmailWolf" bad_bot

deny from env=bad_bot
```

## La alternativa con PHP

En casos en los cuales la configuración de el servidor no permite al administrador del sitio web el poder hacer uso de funciones en *.htaccess* o bien no se tienen derechos o posibilidad alguna para modificar **httpd.conf**, el uso de guiones PHP son la perfecta solución.

Si se tienen instaladas las extensiones de PHP, puede utilizarse un guión que logrará el efecto deseado, bloqueando el accesos al contenido del sitio web a determinados agentes de usuario. El siguiente ejemplo, [UA BLOCK](#), creado por Christopher Lover, funciona excelentemente:

### Código fuente de BLOCK.PHP

Bastará con añadir al inicio del código del documento, el cual debe tener extensión php, lo siguiente:

```
<?PHP include "/ruta/hacia/donde/lo/tenga/block.php"; ?>
```

Cualquier agente de usuario que trate de acceder a un documento que utilice el código anterior, recibirá una página de error como respuesta. Este método es más versátil y permite bloquear documentos o secciones de manera selectiva.