¿Cómo limitar el número de visitas a tu web?

escrito por Andy Garcia | 03/04/2024

Si tienes un sitio web con muchas **páginas web auto-generadas mediante algún algoritmo o IA**, te puede pasar que de vez en cuando los robots o spiders colapsen tu servidor, me ha pasado y creo haberlo solucionado, a continuación te cuento como lo he hecho…

> *He aplicado esta solución en un proyecto web (desarrollo web desde cero con PHP) en el que son los propios usuarios los que generan la mayoría de los contenidos web.*

Cada vez que un usuario hace una consulta, el servidor web guarda el resultado de dicha consulta, para tenerla a mano para la siguiente vez que el mismo u otro usuario haga la misma consulta.

También se guardan decenas o cientos de otros contenidos relacionados o muy relacionados, en algunos casos se trata literalmente de miles de posibles consultas y cada una de ellas genera igualmente decenas o cientos de consultas adicionales, **los contenidos web crecen de forma exponencial**.

> *Al ser consultas de contenidos de texto sin formato, que ocupan poco espacio, el espacio en disco duro no es un problema grave, aunque he tenido que borrar algunos contenidos en alguna ocasión para limitar el tamaño del backup.*

El problema grave viene cuando los robots o spiders intentan indexar tu web y se encuentran con semejante maraña de contenidos.

Si no se imponen unos límites a los robots sencillamente éstos intentarán **indexar todos los contenidos**, entonces el servidor web recibirá muchísimas peticiones, se saturará, se ralentizará o incluso se caerá.

¿Se puede solucionar poniendo indicaciones en el archivo robots.txt?

Puedes intentarlo usando el siguiente código de ejemplo:

```
User-agent: Googlebot
Disallow:
User-agent: *
Disallow: /
Crawl-delay: 10
```

El código anterior colocado en el archivo robots.txt permite que Googlebot acceda a todas las páginas del sitio web, mientras que bloquea el acceso a todas las páginas para otros robots de búsqueda. Además, se establece un

retraso de rastreo de 10 segundos para cualquier robot que respete estas reglas.

> *En teoría esto debería bastar, pero en la práctica muchos robots (diría que la mayoría) no respetan las reglas del archivo robots.txt y todo lo anterior no sirve para absolutamente casi nada.*

¿Cómo limitamos el acceso a las páginas web de forma que funcione?

Mi idea ha sido la siguiente, en cada visita comparo la fecha y hora de la visita anterior con la actual, si coinciden se trata de más de una petición en el mismo segundo, poco probable si se trata de humanos y muy posible en el caso de robots, así que le prohíbo el acceso y a continuación guardo la fecha y hora actual para la siguiente petición.

A continuación el código PHP que limita los accesos a tu web

```
// Control de acceso (máx. una visita por sg.)
$ultTxt = file_get_contents('ult.txt'); // Recupera datos de última visita (guardado en anterior)
if ($ultTxt == "$hoy $ahora") {
header("HTTP/1.1 429 Too Many Requests");
header("Retry-After: 60"); // Tiempo de espera en sg.
exit ("El servidor web está recibiendo demasiadas solicitudes.
Por favor, inténtelo de nuevo un minuto más tarde.");
}
file_put_contents('ult.txt', "$hoy $ahora"); // Guarda datos de visita actual (para la siguiente)
```

El proyecto web del que estoy hablando está en fase beta y no me preocupa que de vez en cuando un usuario reciba un

aviso de el servidor está saturado y tiene que esperar un minuto antes de volver a intentarlo, mientras tanto me estoy ahorrando servir páginas web a docenas de miles de peticiones hechas por robots que tenía de vez en cuando antes de implementar esta medida.

Sin embargo sería más interesante determinar si la visita procede de un robot o un humano para limitar el acceso sólo a los primeros, eso también lo estoy haciendo pero es más complicado y se excede del propósito de este post, quizá para el siguiente, estate atento al blog.