Bisoños Usuarios de Linux de Mallorca y Alrededores | Bergantells Usuaris de Linux de Mallorca i Afegitons

Filtrar contenidos de Internet para un instituto usando Squid

Por **Carlos Yaniz**, ***cyaniz*** (http://www.terra.es/personal6/iescapdepera/)
Creado el 11/04/2003 22:37 y modificado por última vez el 11/04/2003 22:37

En este documento explico mi experiencia instalando un filtro de páginas web para las aulas de informática de un instituto de educación secundaria. Para ello he utilizado un ordenador con Linux, el servidor web Apache[3], el proxy Squid[4] y el filtro SquidGuard[5]. Todos estos programas son libres y gratuitos, se pueden descargar por Internet o instalar desde una distribución como Mandrake[6]. He intentado que la explicación sea lo más detallada posible para que un administrador con pocos conocimientos de Linux pueda seguirla.

EL CONTEXTO

El IES de Capdepera (Mallorca) tiene unos 400 alumnos de ESO, Ciclos Formativos y Garantía Social, que tienen acceso a 2 aulas de informática de 14 ordenadores cada una. Ocasionalmente utilizan también algunos ordenadores de la biblioteca y de las aulas de apoyo.
El centro está cableado gracias al proyecto Xarxipèlag de la Conselleria d'Educació[1]. Todos los ordenadores están en red y salen a Internet por una línea ADSL de 512 Kb.
Conectar los ordenadores de las aulas a Internet ha sido un gran avance, ¡todavía recordamos hace tan sólo unos años, el único ordenador con módem de 56 Kb que tardaba tanto! Sin embargo el uso que se hace de Internet se puede convertir fácilmente en abuso:

- algunos alumnos que se conectan a páginas porno en cuanto el profesor se da la vuelta o se ausenta 5'
- otros que con el pretexto de buscar información para un trabajo, se bajan las últimas canciones del triunfito de moda
- alumnos que no pueden ponerse a trabajar si antes no han comprobado su correo, operación que dura la mitad de la clase
- chat
- ...etc..

Todo esto puede hacer que ciertos profesores se desanimen y vean Internet como un problema y no como una herramienta educativa, mostrándose reacios a utilizar el aula de informática. Está claro que el profesor no puede vigilar la actividad de todos los ordenadores todo el tiempo, y tampoco se trata de ir al aula de informática para acabar haciendo de policía.

Así pues sería deseable controlar el acceso a Internet de manera automática a fin de:

- evitar páginas web inadecuadas mediante un filtro
- registrar las páginas visitadas desde cada ordenador
- establecer un horario para que Internet esté disponible únicamente en las horas en que realmente haga falta

LA SOLUCIÓN

Las funciones requeridas se pueden conseguir mediante el filtro SquidGuard, que usa una base de datos con miles de direcciones web clasificadas en grupos (porno, violencia, publicidad...). Hay múltiples opciones de configuración: bloqueo por grupo o por palabra clave, identificación del cliente por su dirección IP o por nombre de usuario, creación de horarios, actualización de la base de datos por un robot,...etc... SquidGuard funciona asociado al proxy Squid que nos va a permitir registrar las páginas web visitadas en ficheros .log y además hace de caché acelerando el acceso a las

páginas visitadas recientemente.

A continuación, explicaré paso por paso como configurar todo esto en un ordenador (ordenador servidor) partiendo de una distribución Mandrake 8.2 recién instalada. Evidentemente podéis usar vuestra distribución favorita ya que el proceso de configuración es el mismo.

El ordenador que he utilizado tiene un procesador Celeron a 1,8 GHz y 256 MB de RAM. He probado también en un AMD−K6 a 500 MHz con 64 MB de RAM y no noto la diferencia. Supongo que es porque hay pocos accesos simultáneos por segundo.

CONFIGURACIÓN DE APACHE

En el momento que se solicite una página web inadecuada, debe aparecer un mensaje disuasorio del tipo *"Página no accesible..."*. Para ello, es necesario tener funcionando un servidor de páginas web, por ejemplo Apache.

Si hemos instalado la Mandrake con las opciones de servidor, Apache debería estar instalado y funcionando. Para comprobarlo, si la dirección IP del servidor es 172.16.61.98, escribimos `http://172.16.61.98` en el navegador de Internet de un ordenador cliente. Saldrá la página inicial.

Tenemos que asegurarnos también que se puedan ejecutar las CGI ya que por defecto Apache no lo permite. En el fichero de configuración `/etc/httpd/conf/commonhttpd.conf` añadimos el permiso escrito en negrita para el directorio `/var/www/cgi−bin`

<Directory /var/www/cgi−bin>
  AllowOverride All
  Options ExecCGI
  Allow from 127.0.0.1
</Directory>

archivo /etc/httpd/conf/commonhttpd.conf

Desde el terminal, reiniciamos Apache con el comando

#apachectl restart

También hay que comprobar que se pueda acceder a `/var/www/cgi−bin` y que `test−cgi` sea ejecutable. Si no fuera ejecutable, haríamos:

#chmod +x /var/www/cgi−bin/test−cgi

Ya podemos probar desde el cliente con `http://172.16.61.98/cgi−bin/test−cgi`

CONFIGURACIÓN DE SQUID

Squid es el programa que hace de proxy−caché. De nuevo, si hemos instalado la Mandrake con las opciones de servidor, Squid debería estar instalado y funcionando. Para comprobarlo, podemos hacer:

#ps −e | grep squid

Para configurarlo editamos el archivo `/etc/squid/squid.conf` y añadimos las líneas en negrita:

...
acl localhost src 127.0.0.1/255.255.255.255
 # el instituto tiene las IP 172.16.61.*
acl ieshosts src 172.16.61.0/255.255.255.0
...
http_access allow localhost
 # autorizamos a los ordenadores del instituto
http_access allow ieshosts
...
 # queremos un fichero de registro con fecha y hora
emulate_httpd_log on
...

archivo /etc/squid/squid.conf

Reiniciamos Squid

#squid −k reconfigure

Vamos a probar en un ordenador cliente. Debemos hacer que pase por el proxy para acceder a Internet. Por ejemplo, si es Windows, vamos a *"Internet Explorer > Herramientas > Opciones de Internet > Conexiones > Configuración LAN > marcar Usar servidor proxy > Dirección 172.16.61.98 Puerto 3128"*. Los alumnos avispados podrían desactivar la casilla *"Usar servidor proxy"* para saltarse el filtro. En ese caso protegeríamos con políticas de Windows el acceso al panel *"Opciones de Internet"*.

Accedemos a una página web cualquiera y luego vamos al servidor a comprobar el fichero de registro:

#cat /var/log/squid/access.log

Como este fichero aumenta de tamaño muy rápido, conviene hacer que se reinicie cada día. Guardaremos los registros de los 7 últimos días. Para ello, en el archivo `/etc/squid/squid.conf` estableceremos una rotación de 7 con el TAG `logfile_rotate`:

...
logfile_rotate 7
...

archivo /etc/squid/squid.conf

Reiniciamos Squid, y a partir de ahora cada vez que hagamos `squid −k rotate` se producirá una rotación: El registro actual access.log pasa a access.log.0, éste pasa a access.log.1, ..etc..

Para no tener que hacer la rotación manualmente, podemos programar un trabajo que la haga y se autoprograme para el día siguiente. Creamos el script `/etc/squid/rotacion`:

squid −k rotate
at −f /etc/squid/rotacion 23:00 tomorrow

script /etc/squid/rotacion

Hacemos que se ejecute por ejemplo a las 23:00

#at −f /etc/squid/rotacion 23:00

Así pues, access.log será el registro del día actual, access.log.0 será el registro de ayer, access.log.1 será el registro de anteayer, ...etc...

CONFIGURACIÓN DE SQUIDGUARD

SquidGuard es el filtro asociado al proxy Squid. Se instalará con la Mandrake, pero no estará funcionando.

Para configurarlo editamos el archivo `/etc/squid/squidGuard.conf`. Vamos a prohibir todos los grupos de direcciones problemáticas, a excepción de los servidores de correo gratuito, que usan muchos alumnos y profesores.

```
#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
# SquidGuard CONFIGURATION FILE
#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−

# DIRECTORIOS DE CONFIGURACION
dbhome /usr/share/squidGuard−1.2.0/db
logdir /var/log/squidGuard

# GRUPOS DE DIRECCIONES
dest porn {
  domainlist porn/domains
  urllist porn/urls
  expressionlist porn/expressions
}
dest audio−video {
  domainlist audio−video/domains
  urllist audio−video/urls
}
dest hacking {
  domainlist hacking/domains
  urllist hacking/urls
}
dest warez {
  domainlist warez/domains
  urllist warez/urls
}
dest ads {
  domainlist ads/domains
  urllist ads/urls
  # la publicidad es reemplazada por una imagen vacia
  redirect http://127.0.0.1/nulbanner.png
}
dest aggressive {
  domainlist aggressive/domains
  urllist aggressive/urls
}
dest drugs {
  domainlist drugs/domains
  urllist drugs/urls
}
dest gambling {
  domainlist gambling/domains
  urllist gambling/urls
}
# permitimos los servidores gratuitos de correo
#dest mail {
```

```
#  domainlist mail/domains
#}
dest proxy {
  domainlist proxy/domains
  urllist proxy/urls
}
dest violence {
  domainlist violence/domains
  urllist violence/urls
  expressionlist violence/expressions
}

# CONTROL DE ACCESO
acl {
  # por defecto bloqueamos los grupos de direcciones creados
  default {
    pass !porn !audio−video !hacking !warez !ads !aggressive !drugs !gambling !proxy !violence all
    # redireccionamos a una pagina web disuasoria
    redirect http://127.0.0.1/prohibit.html
  }
}
```

archivo /etc/squid/squidGuard.conf

Para conectar Squid con SquidGuard, editamos `/etc/squid/squid.conf` Buscamos el TAG redirect_program y ponemos:

```
...
redirect_program /usr/bin/squidGuard −c /etc/squid/squidGuard.conf
...
```

archivo /etc/squid/squid.conf

Como es fácil cometer errores en el fichero de configuración de SquiGuard, conviene crear un directorio log en `/var/log/squidGuard`. El propietario debe ser el usuario `squid`.

```
#cd /var/log/squidGuard
#mkdir log
#chown squid:squid log
```

Reiniciamos Squid

```
#squid −k reconfigure
```

Comprobamos que aparecen 5 procesos SquidGuard, que es el número por defecto

```
#ps −e | grep squidGuard
```

Comprobamos el log

```
#cat /var/log/squidGuard/squidGuard.log
```

Al final, debe salir el mensaje *"squidGuard ready for requests"*

Comprobamos también que la base de datos de direcciones se ha cargado bien

#cat /var/log/squidGuard/log/squidGuard.log

Si hubiera algún problema, aparecería el mensaje *"Emergency mode..."*. En modo de emergencia, el filtro queda desactivado.

No hay que olvidar crear la página web `/var/www/html/prohibit.html` que es hacia donde se redireccionan las páginas bloqueadas.
También hay que copiar la imagen vacía `nullbanner.png` para la publicidad.

#cp /var/www/cgi−bin/nullbanner.png /var/www/html

Ya podemos probar. Vamos al ordenador cliente, escribimos `http://www.playboy.com` y debe salir la página `prohibit.html`. Por si algo fuera mal, ¡conviene antes cerrar la puerta del aula con llave!

ACTUALIZAR LA BASE DE DATOS

En relación a la base de datos de direcciones hay que señalar que:

- tal y como advierte la documentación, las direcciones han sido buscadas y clasificadas por un programa robot, y por lo tanto, puede haber errores. En el mes que lo llevo probando no me he encontrado con este problema, pero se me ocurre que por ejemplo podría quedar bloqueada alguna web útil dedicada a educación sexual o a prevención de drogas.
- la lista de direcciones no es, evidentemente, exhaustiva. Es más, muchos sitios web inapropiados en español no están fichados. Supongo que se debe a que el robot es angloparlante y ha buscado por palabra clave en inglés.

Actualización de la base de datos completa

Como a diario aparecen nuevos sitios web, conviene bajarse periódicamente toda la base de datos de direcciones actualizada.

Bajamos el archivo `blacklists.tar.gz` de `http://www.squidguar7d.org/blacklist` al directorio `/usr/share/squidGuard−1.2.0`

#cd /usr/share/squidGuard−1.2.0
#tar −xzf blacklists.tar.gz
#mv db db_old
#mv blacklists db
#chown −R squid:squid db

Creamos un script llamado `actualiza` con las siguientes instrucciones:

squidGuard −c /etc/squid/squidGuard.conf −C all
squidGuard −c /etc/squid/squidGuard.conf −u
chown −R squid:squid db
squid −k reconfigure

script actualiza

Ejecutamos este script para actualizar la base de datos.

Comprobamos que no hay ningún error en los logs

#cat /var/log/squidGuard/squidGuard.log
#cat /var/log/squidGuard/log/squidGuard.log

Actualización automática usando el robot

En el directorio `/usr/share/squidGuard-1.2.0/contrib` se encuentra el Robot que busca y clasifica las direcciones. Está escrito en Perl. Todavía no he podido hacerlo funcionar por falta de tiempo y porque no existe documentación. La única ayuda que he encontrado por Internet son unas indicaciones del autor: http://ftp.teledanmark.no/pub/www/proxy/squidGuard/contrib/squidGuardRobot/(2)

Actualización manual

Si se dispone de tiempo, se puede mirar de vez en cuando los registros de páginas visitadas `/var/log/squid/access.log`. Si vemos que los alumnos tienen especial fijación por alguna web inapropiada no fichada, además de echarles la bronca pertinente, podríamos añadir la dirección a la base de datos.

Por ejemplo, en mi instituto se meten a veces en `www.petardas.com` que no está fichado por la base de datos original. Lo podríamos añadir al archivo predefinido `porn/domains`, pero es mejor meterlo en un grupo nuevo, que será más fácil de preservar cuando actualicemos la base de datos completa.

Editamos `/etc/squid/squidGuard.conf` y creamos un grupo de direcciones llamado `nuevas`. Lo añadimos también al control de acceso

...
GRUPOS DE DIRECCIONES
dest nuevas {
** domainlist nuevas/domains**
}

...
CONTROL DE ACCESO
...

  pass **!nuevas** !porn !audio-video ...

archivo /etc/squid/squidGuard.conf

Luego creamos su directorio en la estructura de la base de datos

#cd /usr/share/squidGuard-1.2.0/db
#mkdir nuevas
#cd nuevas

Creamos un archivo domains con el siguiente contenido:

petardas.com
207.44.222.52

archivo /usr/share/squidGuard-1.2.0/db/nuevas/domains

Esto bloqueará todas las páginas del dominio `petardas.com` y también su dirección IP por si algún listillo se la

sabe.

Después ejecutamos el script `actualiza` del apartado anterior y comprobamos que no haya ningún error en los logs.

A partir de ahora, añadiremos todas las direcciones nuevas que encontremos al archivo `/usr/share/squidGuard-1.2.0/db/nuevas/domains`. Si actualizamos la base de datos completa como se explica en el apartado anterior, debemos respetar el directorio `nuevas`.

Bloqueo mediante expresiones

También podemos usar expresiones para prohibir ciertas URLs que contengan alguna palabra clave determinada. Por ejemplo, las webs `www.sexo.com, www.sexo-casero.com, www.lawebsexy.com`, se pueden bloquear identificando las palabras clave *"sexo"* y *"sexy"*. Hacerlo mediante la palabra clave *"sex"* podría resultar demasiado restrictivo ya que muchas URLs de webs útiles de educación sexual, como por ejemplo `www.medusex.com`, contienen esas 3 letras.

Editamos `/etc/squid/squidGuard.conf` y añadimos la línea en negrita:

```
...
# GRUPOS DE DIRECCIONES
dest nuevas {
  domainlist nuevas/domains
  expressionlist nuevas/expressions
}
...
```

archivo /etc/squid/squidGuard.conf

Vamos al directorio `nuevas` que hemos creado en la base de datos y creamos un archivo expressions con el siguiente contenido:

```
(sexo|sexy)
```

archivo /usr/share/squidGuard-1.2.0/db/nuevas/expressions

Finalmente ejecutamos el script `actualiza`.

PONER UN HORARIO

No todos los profesores que van al aula de informática necesitan tener conexión a Internet. Con SquidGuard podemos crear un horario de acceso a Internet. Para ello definimos un horario `clases_aula1` y un grupo de direcciones IP para esa aula. Luego creamos una regla de control de acceso para ese grupo. Si se intenta acceder a Internet fuera del horario establecido saldrá la página `horari.html` con un mensaje del tipo *"En este horario no se puede acceder a Internet..."*.

```
#-------------------------------------------------------------------
# SquidGuard CONFIGURATION FILE
#-------------------------------------------------------------------

# DIRECTORIOS DE CONFIGURACION
dbhome /usr/share/squidGuard-1.2.0/db
logdir /var/log/squidGuard

# HORARIO DE ACCESO A INTERNET DE AULA 1
time clases_aula1 {
```

```
 # dia            hora (HH:MM)   grupo    profesor   internet
 #weekly mondays    08:00−09:00    # 1C     Carlos     no
 weekly mondays     09:00−09:50    # 3A     Peiro
 weekly mondays     12:00−12:50    # 4BC    Carlos
 #weekly mondays    14:00−15:00    # 3CD    Carlos     no
 #weekly tuesdays   09:50−10:40    # 1B     Carlos
 weekly tuesdays    10:40−11:10    # patio  Carlos
 #weekly tuesdays   14:00−14:50    # 1D     Carlos     no
 weekly wednesdays  12:00−12:50    # 4BC    Carlos
 #weekly thursdays  09:00−09:50    # 1A     Rita no
 #weekly thursdays  13:10−14:00    # 3B     Ventura    no
 weekly fridays     09:50−10:40    # 4div   Carlos
}

# AULA 1: IP de la 100 a la 113
src aula1 {
  ip 172.16.61.100−172.16.61.113
}

# GRUPOS DE DIRECCIONES
# ... dejarlo igual ...

# CONTROL DE ACCESO
acl {
  # el aula 1 queda bloqueada fuera del horario definido arriba
  aula1 outside clases_aula1 {
    pass none
    redirect http://127.0.0.1/horari.html
  }

  # por defecto, bloqueamos los grupos de direcciones creados
  default {
    pass !nuevas !porn !audio−video !hacking !warez !ads !aggressive !drugs !gambling !proxy !violence all
    redirect http://127.0.0.1/prohibit.html
  }
}
```

archivo /etc/squid/squidGuard.conf

Hay que tener cuidado al escribir las horas porque se debe usar el formato HH:MM. Si escribimos 8:00 en vez de 08:00, el parser dará un error.

LIMITAR A UN GRUPO DE PÁGINAS

Otra posibilidad de SquidGuard es limitar el acceso a Internet permitiendo únicamente ciertos dominios. Esto nos puede servir si tenemos un grupo de alumnos especialmente difícil pero que tiene que hacer uso puntual de Internet para hacer algo concreto.
Por ejemplo, en mi instituto los alumnos del grupo PAC van al aula 1 y necesitan Internet sólo para leer el correo electrónico en Hotmail. Sin embargo, aprovechan el menor descuido del profesor para navegar por otros sitios. Para limitar el acceso sólo al correo de Hotmail, debemos crear un grupo de direcciones que llamaremos, por ejemplo, `unicas`. Hay que señalar que para leer el correo de Hotmail, no sólo se debe dar acceso a `hotmail.com` sino también a otros dominios y direcciones IP asociados.
En la estructura de la base de datos creamos un directorio llamado `unicas` y un archivo `domains` con el siguiente contenido:

hotmail.com
passport.com
msn.com
msn.es
64.4.36.24

archivo /usr/share/squidGuard−1.2.0/db/unicas/domains

Después en el archivo de configuración de SquidGuard creamos un horario para el PAC, un grupo de direcciones llamado `unicas` y una regla de control de acceso:

```
...
time clases_PAC {
  weekly fridays 09:00−09:50 # PAC Carlos
}
...
dest unicas {
  domainlist unicas/domains
}
...
# CONTROL DE ACCESO
acl {
  # al PAC solo se le permite las direcciones del grupo unicas
   aula1 within clases_PAC {
   pass unicas none
   redirect http://127.0.0.1/unic.html
  }
 ...
}
```

archivo /etc/squid/squidGuard.conf

Algunas observaciones:

- He tenido que colocar la definición del grupo `unicas` (`dest unicas ...`) al final de todas las otras definiciones. De otra manera me daba un error *"Core dumped"* al actualizar.
- Cuando hay varias reglas de acceso, sólo se ejecuta la primera que sea cierta. La regla de control de acceso (`aula1 within clases_PAC...`) debe ir la primera, porque la otra (`aula1 outside clases_aula1...`) también es cierta y no queremos que se ejecute.

CONCLUSIÓN

SquidGuard ofrece muchas posibilidades de configuración. Para otras opciones no explicadas en este artículo, existe una documentación bastante completa en la web de SquidGuard. De todas maneras, probablemente no convenga complicarse demasiado, ya que la gestión de actualizaciones, horarios,... puede llevar mucho tiempo.
El único inconveniente que he encontrado es que en la base de datos hay catalogadas pocas direcciones en español. Estaría bien conseguir que el robot actualizara automáticamente la base de datos buscando por palabra clave en español o en otros idiomas distintos del inglés.

Lista de enlaces de este artículo:

1. http://www.caib.es/
2. http://ftp.teledanmark.no/pub/www/proxy/squidGuard/contrib/squidGuardRobot/
3. http://www.apache.org/

4. http://www.squid−cache.org/
5. http://www.squidguard.org/
6. http://www.mandrake.org/

E−mail del autor: cyaniz@educacio.caib.es
Podrás encontrar este artículo e información adicional en: http://bulmalug.net/body.phtml?nIdNoticia=1729