

El día a día del Administrador de Sistemas: FuzzyOCR

# 1000 OBRAS MAESTRAS

La última moda entre los spammers consiste en ocultar el spam en imágenes. Los administradores reaccionan: una herramienta OCR que extrae el texto y lo envía al filtro de spam. **POR CHARLY KÜHNAST**

Si ejecuta Spamassassin, el plugin FuzzyOCR [1] es una buena opción como herramienta para la evaluación de imágenes. No es complicado de instalar, excepto porque tiene que cumplir con unas cuantas dependencias. Asegúrese de que su versión de Spamassassin esté lo más actualizada posible; se recomienda tener la 3.1.4.

También necesitará las herramientas NetPBM, de Imagemagick *convert*, Giflib, dos módulos de Perl y *gocr* para reconocer los caracteres ópticos. Puede que sea mucho pedir, pero actualmente la mayoría de las distribuciones las llevan incorporadas. El siguiente comando ejecutará los módulos Perl por usted:

```
cpan -i Digest::MD5
String::Approx
```

Se encuentra a un par de pasos para la evaluación de la imagen: necesita *FuzzyOcr.cf* y *FuzzyOcr.pm* en el directorio de Spamassassin, que en mi equipo está en */etc/mail/spamassassin*. FuzzyOCR le proporciona un diccionario de muestra, *FuzzyOcr.words*, que contiene los términos que FuzzyOCR tiene que buscar en las imágenes. Puede modificar la lista para adaptarla a sus necesidades. De nuevo, deberá copiar este fichero a su directorio de Spamassassin.

El siguiente paso es definir la ruta para los ficheros de registro y de diccionario en *FuzzyOcr.cf*. Una vez hecho, FuzzyOCR debería

```
Oct 10 12:04:28 spamfilter2 amavis[2595]: (02595-05) SPAM-TAG,
<rob.lleqyn@par.dcona.co.br> -> <charly@kuehnast.com>,
Yes, score=6.882 tagged above=-99 required=1.5
tests={BAYES_00=-2.599, FORGED_RCVD_HELO=0.135, FUZZY_OCR=6.000,
HTML_50_60=0.134, HTML_MESSAGE=0.001, HTML_TITLE_EMPTY=0.214, INFO_TLD=1.273,
SUBJECT_ENCODED_TWICE=1.723, UNPARSEABLE_RELAY=0.001}
```

Figura 1: FuzzyOCR detecta el texto en los ficheros de imágenes y les asigna puntuaciones si descubre palabras no deseadas.

estar listo para echar a volar, ya que Spamassassin descubrirá el módulo en la ruta de inicio y lo integrará automáticamente.

## Dibuja Esto

Aunque FuzzyOCR viene con unos cuantos ficheros de spam de muestra para pruebas, es mucho más divertido intentarlo con los tuyos propios. La Web es un lugar extraño y maravilloso: el spam puede aparecer en cualquier formato de imagen común, y los tipos MIME a menudo se definen incorrectamente causando más confusión, por ejemplo, los GIFs se hacen pasar por JPEGs. FuzzyOCR reacciona a estas tácticas asignando puntos negativos adicionales.

Los spammers recurren a los GIFs animados, especialmente a los que muestran píxeles basura antes de revelar el mensaje del spammer. Parece que muchos spammers confían en que los motores OCR sólo analicen la primera fase de la animación, pero afortunadamente, FuzzyOCR no se queda ahí.

## Crimen y Castigo

Hora de estudiar a *FuzzyOcr.cf*. Si tiene una versión anterior a la 3.1.4 de Spamassassin, necesitará establecer aquí una entrada para *focr\_pre314 = 1*. Es más importante configurar bien la puntuación que FuzzyOCR asigna cuando encuentra algo sospechoso. Por defecto, el programa es bastante estricto.

Por ejemplo, un mensaje con una imagen adjunta que coincida con dos entradas del diccionario se le asigna una puntuación spam de cuatro y uno respectivamente, y la mitad de los puntos se añaden



den si el tipo MIME está declarado de forma incorrecta. Se añaden dos puntos y medio extra por las imágenes corruptas y cinco si el error no es corregible. Los puntos se añaden para dar un total global como se muestra en la Figura 1.

Las configuraciones muy estrictas incrementan el peligro de los falsos positivos: no olvide que Spamassassin probablemente notificará unos cuantos errores con un mensaje de spam, lo que puede llevar a obtener una puntuación increíblemente alta. Mi consejo es reducir la puntuación a la mitad de los valores en *FuzzyOcr.cf*.

¿Y qué viene ahora? Tan sólo tumbese y espere a ver qué es lo siguiente que se van a inventar los spammers.

## RECURSOS

[1] FuzzyOCR: <http://users.own-hero.net/~decoder/fuzzyocr/>

## SYSADMIN

### Especial NTFS ..... 50

VFAT ha muerto (al menos para Microsoft)... ¿Viva NTFS? Con la nueva versión de Windows, no hay escapatoria: hay que leer y escribir a NTFS sin fallos si queremos mantener nuestras redes híbridas.

### ntfsprogs .....52

Los ntfsprogs son una colección de programas y utilidades para manejar datos en sistemas de ficheros NTFS. Vemos lo que hay y cómo utilizarlo.

### Captive .....56

Captive es el primer driver libre que permite un acceso total a particiones NTFS.

### NTFS en vivo .....59

Cuando todo lo demás falla, no hay más remedio que intentar acceder a un sistema de ficheros desde un Live CD. Demostramos la manera de hacerlo desde Knoppix.