Combatiendo el Spam por Imágenes

www.bitdefender.es

El Problema

Los Spammers están desarrollando métodos cada vez más sofisticados para evitar los filtros antispam. Generalmente, todos estos intentos de envío de oleadas de emails son únicos y diferentes a otros enviados anteriormente. Los spammers analizan las oleadas de spam que consiguen tener éxito, y utilizan los hallazgos para incluirlos como "nuevas características" en los siguientes envíos masivos.

Los nuevos métodos de detección de oleadas de Spam, que estudian y extraen las características principales de este tipo de mensajes y las envían a los usuarios a través de firmas de spam, están en la última fase de desarrollo. Se están realizando investigaciones para encontrar nuevos métodos que permitan predecir los cambios en los mensajes de spam.

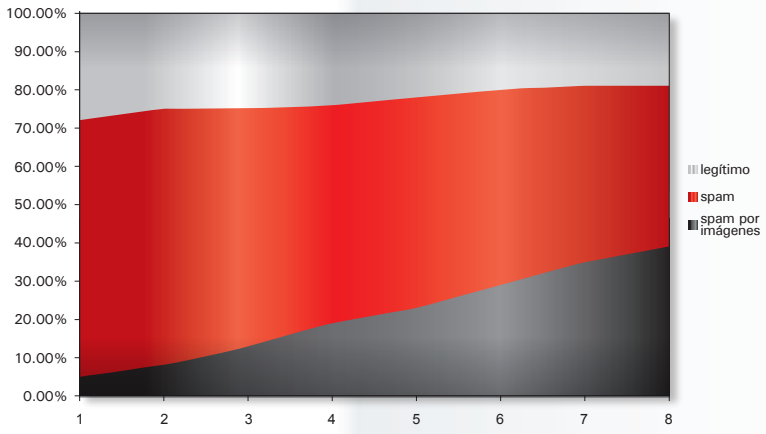
Muchos de los filtros utilizados por BitDefender, se han vuelto más robustos y efectivos, pudiendo detectar todas aquellas pequeñas variaciones encontradas en los flujos de spam. No obstante, el 2006 se ha caracterizado por el aumento de otro tipo de spam más evolucionado: el spam por imágenes. Estos simples emails que aparentemente contienen imágenes similares (aunque únicos, desde el punto de vista computacional) comenzaron a contaminar nuestras bandejas de entrada en grandes cantidades.

En ese momento las técnicas para combatir el spam por imágenes acababan de aparecer, y un eficaz método de detección fueron las firmas de spam basadas en los meta datos de esas imágenes. Mientras tanto, el Laboratorio Antispam de BitDefender estuvo investigando mensajes en circulación que utilizaban técnicas de manipulación de imágenes, y se llegó a la conclusión que era necesaria una nueva tecnología para combatir esta nueva tendencia.

El Punto de Partida

En 2005, el "spam por imágenes" suponía aproximadamente un 10% del Spam. Estas series de mensajes normalmente contenían 5-6 imágenes de spam con pequeñas modificaciones para evitar su detección.

En los últimos meses, los spammers se dieron cuenta que muchas de las soluciones Antispam eran ineficaces contra este nuevo truco, así que empezaron a atacar este segmento. El spam por imágenes provocó un aumento del 30-40% de mensajes de spam en circulación, y cada mensaje contenía una imagen con distorsiones aleatorias respecto a los otros mensajes. En cambio, los ratios de detección disminuyeron notablemente, de un 97% a casi un 65-75%.



Evolución del Spam por Imágenes en los 8 primeros meses del 2006

El spam por imágenes suele contener imágenes de pastillas de medicamentos, hardware, imágenes para adultos, o el clásico mensaje de spam (un poco de texto y un enlace), pero escrito dentro de una imagen con ruido y distorsiones.

Aparentemente, para analizar el contenido de estos mensajes se tendría que ejecutar un módulo de reconocimiento óptico de caracteres (OCR). Pero los filtros OCR tienen un coste computacional muy alto y su exactitud se aleja mucho de los resultados deseados.

bitdefender
secure your every bit

El Enfoque de BitDefender

Para conseguir una detección más fiable, BitDefender ofrece una alternativa al OCR: un filtro que ignora el texto dentro de las imágenes (el mensaje, desde el punto de vista humano), y en cambio, aprende de la experiencia de algunas características comunes de las imágenes.

Esta alternativa se basa en el uso de dos técnicas, la extracción y comparación de histogramas*), que han demostrado ser eficaces a lo largo del tiempo, en aquellas aplicaciones que requieren del procesamiento de imágenes.

Estas técnicas se usan generalmente en la recuperación de imágenes basadas en el contenido (Ej: para extraer las imágenes de delfines en un álbum de fotos de las últimas vacaciones), con un ratio bastante alto de falsos positivos. Al principio, era bastante problemático usar estas técnicas como herramientas antispam, ya que esos falsos positivos se traducían en mensajes que el usuario perdía.

La experimentación con estas técnicas desvelaron una nueva fórmula para utilizar este algoritmo denominado SID (Spam Image Distance).

El algoritmo SID selecciona las imágenes basándose en la similitud de los colores, en lugar de buscar la similitud en las formas. Por ejemplo, desde la perspectiva de SID, aunque todas las imágenes de las páginas impresas parezcan similares (todas contienen el color blanco, blanco roto o cantidades de gris oscuro), una página de la Enciclopedia Británica no se parece a la página de un anuncio de texto, porque las proporciones de los blanco y gris son muy diferentes.

El SID se utiliza para comparar imágenes y mide la "distancia" entre ellas, lo que básicamente significa encontrar las diferencias entre ellas. Las distancias obtenidas a partir de la fórmula SID se utilizan para comparar las imágenes ya incluidas en la base de datos de spam con la nuevas imágenes que pueden ser spam. Si el análisis de la imagen obtiene una puntuación más baja que la indicada por el umbral, entonces la imagen se añadirá a la base de datos de imágenes spam de BitDefender. Por este motivo SID es la mejor técnica cuando se trata de imágenes de spam que son variaciones de otras, o antiguas imágenes de spam.

Aunque esta nueva técnica funciona bien en imágenes "limpias", aún quedaba el problema de las imágenes con ofuscaciones (Ej: con ruido añadido). Afortunadamente las técnicas de ofuscación usadas por los spammers son ya conocidas y el arsenal de medidas preventivas es igualmente amplio. Por ejemplo, los spammers dividirán una imagen en subimágenes y la reconstruirán con una tabla HTML. Este problema puede solventarse uniendo los histogramas de las subimágenes, reconstruyendo así el histograma de la imagen inicial para analizarlo con el algoritmo SID.

Técnicas habituales de 'ruido' o distorsión:

- Añadir píxeles aleatorios en la imagen
- Gifs animados con fotogramas falsos llenos de "ruido"
- Uso de colores similares ante las diferentes partes del texto de la imagen
- Líneas largas al final de la imagen (como si fuera un borde) con partes aleatorias en blanco
- División de la imagen en sub-imágenes, y reconstruirla con una tabla HTML
- Envío de la misma imagen en diferentes tamaños
- Envenenamiento de la imagen - inserción de imágenes legítimas, como logos de empresas, en los mensajes de spam
- Envío de imágenes legítimas distorsionadas, para confundir a los filtros
- Envío de imágenes legítimas con contenido parecido al spam (ej. imágenes de hipotecas provenientes de compañías hipotecarias legítimas)

*) Un histograma puede definirse como una lista de colores y su preponderancia en una imagen, indica qué colores y cuantos píxeles de determinado color existen en esa imagen.

Ratios de Detección

SID muestra un ratio de detección del 98,7% de imagen del spam (unos cuantos millones de muestras extraídos de un spam verdadero) Un 1,23% de estas imágenes están malformadas, lo que significa que sus histogramas no pueden extraerse, pero tampoco pueden visualizarse. Más del 0,07 % presenta falsos positivos como resultados. Si las imágenes que están malformadas se eliminan del cuerpo del mensaje, el ratio de detección llega al 100%

Con estos resultados tan prometedores, el algoritmo SID es un valioso complemento dentro del arsenal de soluciones antispam modernas, y se esperan avances en la reducción de ruido para mejorar todavía más el potencial de esta herramienta tan útil.

Detalles de Contacto:

País: **España**
Dirección: c/ Balmes 191, 2ª planta
08006 Barcelona
Tel: (+34) 93 218 96 15
E-mail: *comercial@bitdefender.es*
Web: *www.bitdefender.es*

bitdefender
secure your every bit