novática
CEPIS UPGRADE Revista de la Asociación de Técnicos de Informática

Novática, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática), organización que edita también la revista **REICIS** (Revista Española de Innovación, Calidad e Ingeniería del Software). ***Novática*** edita asimismo **UP**GRADE, revista digital de **CEPIS** (*Council of European Professional Informatics Societies*), en lengua inglesa, y es miembro fundador de **UP**ENET (**UP**GRADE ***E**uropean **NET**work*).

<http://www.ati.es/novatica/>
<http://www.ati.es/reicis/>
<http://www.upgrade-cepis.org/>

ATI es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*) y es representante de España en **IFIP** (*International Federation for Information Processing*); tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*), así como acuerdos de vinculación o colaboración con **AdaSpain**, **AI2, ASTIC, RITSI** e **Hispalinux**, junto a la que participa en **ProInnova**.

Coordinación Editorial, Redacción Central y Redacción ATI Madrid
Padilla 66, 3º, dcha., 28006 Madrid
Tlfn.914029391; fax.913093685 <novatica@ati.es>
Composición, Edición y Redacción ATI Valencia
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn./fax 963330392 <secreval@ati.es>
Administración y Redacción ATI Cataluña
Via Laietana 46, ppal. 1ª, 08003 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
Redacción ATI Andalucía
Isaac Newton, s/n, Ed. Sadiel,
Isla Cartuja 41092 Sevilla, Tlfn./fax 954460779 <secreand@ati.es>
Redacción ATI Aragón
Lagasca 9, 3-B, 50006 Zaragoza.
Tlfn./fax 976235181 <secreara@ati.es>
Redacción ATI Asturias-Cantabria <gp-astucant@ati.es>
Redacción ATI Castilla-La Mancha <gp-clmancha@ati.es>
Suscripción y Ventas <http://www.ati.es/novatica/interes.html>, ATI Cataluña, ATI Madrid
Publicidad
Padilla 66, 3º, dcha., 28006 Madrid
Tlnf.914029391; fax.913093685 <novatica@ati.es>
Imprenta: Derra S.A., Juan de Austria 66, 08005 Barcelona.
Depósito legal: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC
Portada: "Gaia gateway" / © Concha Arias Pérez
Diseño: Fernando Agresta / © ATI 2006

Nº 185, enero-febrero 2007, año XXXIII

sumario

Monografía del próximo número: "*Informática para deficientes visuales*"

Recuperación de información con adversario en la Web

Ricardo Baeza-Yates[1], Paolo Boldi[2], José María Gómez Hidalgo[3]

[1]Yahoo! Research Barcelona (España) y Santiago (Chile); [2]Università degli Studi di Milano Milan(Italia); [3]Universidad Europea de Madrid (España)

<ricardo@baeza.cl>, <boldi@dsi.unimi.it>, <jmgomez@uem.es>

Resumen: *la Web es la aplicación de Internet por excelencia. Como tal, y del mismo modo que pasa con el correo electrónico, es un objetivo claro para el abuso. El spam ha invadido los motores de búsqueda, las redes sociales, y aun más, la Web no sólo es objeto de abuso por los proveedores de contenidos, sino por sus propios usuarios. La Recuperación de Información con Adversario (Adversarial Information Retrieval, AIR) se centra en la clasificación de los contenidos o de su uso en relación con su forma de abuso, y se enfrenta a un adversario (el abusador), que tiene como objetivo engañar al clasificador. El spam de buscadores y el filtrado de contenidos Web son dos ejemplos de tareas de AIR en la Web. En este trabajo de revisan una serie de problemas de AIR en la Web, junto con algunas soluciones propuestas. Prestamos especial atención a la detección de spam basado en enlaces en motores de búsqueda, y al filtrado de contenidos Web, como representantes de un amplio rango de técnicas propuestas para alcanzar grados de efectividad altos en el control del abuso relacionado con la Web.*

Palabras clave: *análisis de enlaces, filtrado Web, PageRank, Recuperación de Información con Adversario, Spam Web, spam de motores de búsqueda.*

Autores

Ricardo Baeza-Yates es el director de los nuevos laboratorios de investigación de Yahoo! en Barcelona y en Latinoamérica (Santiago, Chile). Previamente ha sido catedrático y director del Centro para la Investigación en la Web del Departamento de Informática de la Universidad de Chile, y Catedrático ICREA (*Institució Catalana de Recerca i Estudis Avançats*) en el departamento de Tecnología en la Universidad Pompeu Fabra en Barcelona. Ricardo es Doctor en Informática por la Universidad de Waterloo (Canada). Es coautor del libro *Modern Information Retrieval*, publicado en 1999 por Addison-Wesley, y también de la segunda edición del *Handbook of Algorithms and Data Structures* (Addison-Weesley, 1991). También fue coeditor del libro *Information Retrieval: Algorithms and data Structures* (Prentice-Hall, 1992). Es el primer científico informático elegido para la Academia de Ciencias de Chile, en 2003.

Paolo Boldi obtuvo su doctorado en informática en la Universidad de Milán, donde es actualmente profesor asociado en el Departamento de Ciencias de la Información. Sus intereses investigadores han tocado muy variados temas de la informática teórica y aplicada, tales como: la teoría de dominios, la teoría no clásica de la computabilidad, la computabilidad distribuida, las redes anónimas, el sentido de la dirección, y los sistemas auto-estables. Más recientemente, sus trabajos se han centrado en problemas relacionados con la World Wide Web, un campo de investigación en el que también ha aportado sistemas software utilizados por muchos otros especialistas en el tema. En particular, ha contribuido a escribir un motor de Recuperación de Información sobre texto altamente eficiente (MG4J), y una herramienta de compresión de grafos (WebGraph) que alcanza las tasas de compresión habituales en las herramientas actuales.

José María Gómez Hidalgo es Doctor en Matemáticas, y ha sido profesor e investigador en la Universidad Complutense de Madrid, y lo es en la Universidad Europea de Madrid desde hace 10 años, donde actualmente dirige el Departamento de Sistemas Informáticos. Sus principales intereses investigadores incluyen el Procesamiento del Lenguaje Natural y el Aprendizaje Automático, con aplicaciones al Acceso a la Información periodística y biomédica, y la Recuperación de Información con Adversario, con aplicaciones en el filtrado de correo basura y en la detección de pornografía en la Web. Ha participado en 10 proyectos de investigación, dirigiendo algunos de ellos. José María es coautor de múltiples artículos científicos centrados en los temas mencionados, que pueden accederse por medio de su página Web <http://www.esi.uem.es/~jmgomez/>. Es miembro del Comité de Programa del CEAS (*Conference on Email and Anti-Spam*) 2007, del Spam Symposium 2007 y de otras conferencias, y ha revisado artículos de JASIST (*Journal of the American Society for Information Science and Technology*), ECIR (*European Conference on Information Retrieval*) y otras. También es revisor de proyectos para la Comisión Europea.

1. Introducción

Del mismo modo que crece la cantidad de información y el uso de la Web, crece también su valor económico y el interés en sacar partido fraudulento de ella. Dado que los motores de búsqueda son el punto de entrada a la Web más desatacados, han sido objeto de complejos ataques denominados *spam*[1] en motores de búsqueda. Existen otras formas de abuso, como la navegación por contenidos Web inapropiados en las escuelas, las bibliotecas o el puesto de trabajo. Estos tipos de abuso y otros tienen en común que pueden ser tratados como tareas de clasificación de documentos en un marco que llamaremos Recuperación de Información con Adversario (*Adversarial Information Retrieval*, AIR) [7].

La tarea de clasificación con adversario más representativa es el filtrado de correo basura o correo *spam* [12]. Cada día se envían cientos de millones de correos basura anunciando sitios pornográficos, Viagra, o solicitudes de datos bancarios de los usuarios, causando pérdidas de tiempo y graves perjuicios económicos. El filtrado de correo basura es una tarea de clasificación de texto en la que los mensajes se clasifican como *spam* o legítimos. En el filtrado de correo basura, el proveedor de contenidos (el emisor de *spam*) abusa del medio (el correo electrónico) enviando al consumidor (el usuario) mensajes no deseados, específicamente confeccionados para evitar la detección por parte del filtro del usuario.

Lo que hace que el problema sea especialmente retador es que los emisores de *spam* están enormemente motivados para evitar los filtros, porque llegar a más y más usuarios es la forma de lograr un mayor beneficio económico. Basta con que sólo uno de cada cien mil usuarios compre una caja de Viagra fraudulento para lograr un beneficio económico importante. La principal característica de las tareas de Recuperación de Información con Adversario es la existencia de un individuo, el adversario, que está constantemente mejorando sus métodos para que el sistema se equivoque y tome decisiones erróneas. De hecho, se trata de una guerra sin fin entre los abusadores y los desarrolladores de sistemas de clasificación, y no cabe esperar que exista un ganador.

En este trabajo se revisan una serie de problemas de Recuperación de Información con Adversario en la Web, concentrándose en dos de ellos porque proporcionan una buena perspectiva de las técnicas usadas en un rango más amplio de problemas. Los problemas que discutimos con más detalles son el *spam* basado en enlaces en buscadores [13], y el filtrado de contenidos Web [15].

2. Problemas de Recuperación de Información con Adversario en la Web

Como el filtrado de correo basura tiene un

papel paradigmático, son muchas las tareas de Clasificación con Adversario las que hacen uso de la palabra *spam*. Pero lo más importante es que lo que tienen en común es la existencia de un abusador o adversario que trata de que el clasificador se equivoque. Aunque el abusador es muchas veces el proveedor de contenidos (el emisor de correo basura que prepara sus correos para engañar a los usuarios, el administrador de sitios Web – *webmaster* – que intenta que su sitio Web tenga una posición alta en las búsquedas más populares, etc.), incluso los propios usuarios pueden tratar de engañar al sistema (los niños que intentan acceder a contenidos inapropiados en una escuela, como la pornografía).

El objetivo de esta sección es presentar una perspectiva de las tareas de Recuperación de Información con Adversario más relevantes en la Web, y referencias a sus soluciones cuando existan. Hemos organizado las tareas que presentamos en tres grandes grupos: *spam* de contenidos, *spam* de enlaces, y *spam* de uso. En la última sección de este trabajo discutimos las formas emergentes de *spam* en la Web.

2.1. *Spam* de contenidos

El *spam* de contenidos tiene que ver con la preparación de los contenidos de una página Web con el objeto de engañar a un motor de búsqueda, y obtener un puesto alto en consultas muy populares. En este tipo de *spam* también incluimos el abuso de navegación, y en consecuencia, el filtrado de contenidos Web.

2.1.1. Spam *de buscadores basado en contenidos*

Los buscadores tienen la misión de devolver páginas Web de acuerdo con las palabras clave introducidas por sus usuarios. Los proveedores de contenidos pueden incrementar el impacto de los suyos insertando palabras clave populares en sus páginas Web [13]. Por ejemplo, el administrador de un sitio Web de comercio electrónico puede incluir en sus páginas las palabras clave más populares según las listas públicas como el Zeitgeist[2] de Google, aunque no tengan que ver con su negocio. El objetivo es que su sitio Web sea devuelto como resultado de una búsqueda usando palabras clave populares, de modo que se atraigan consumidores a su negocio. Esta forma de abuso se suele llamar "*spam* de palabras clave" (*keyword spamming*).

El abusador (el proveedor de contenidos) trata de obtener una posición muy alta para búsquedas con palabras clave populares insertándolas en el título, en la URL, en las etiquetas HTML de hiperenlaces, o haciéndolas muy frecuentes en su página Web, e incluso copiando grandes cantidades de texto de sitios Web populares. Las palabras clave se ocultan con frecuencia insertándolas en el mismo color que el fondo de la página Web.

Ntoulas et al. [20] han investigado las propiedades del texto y del formato de las páginas con *spam* de palabras clave. Por ejemplo, han detectado una correlación clara entre el grado de *spam* de la página, y el número de palabras clave de la página (cuantas más palabras clave, mayor probabilidad de ser *spam*). Otros atributos útiles para la detección de este *spam* de texto son el número de palabras del título de la página, la longitud media de las palabras, la cantidad de texto en etiquetas de hiperenlaces, o la fracción de contenido visible.

2.1.2. *Encubrimiento (*cloaking*)*

El encubrimiento o *cloaking* es una forma de ocultamiento de información en la que los proveedores de información muestran distintas versiones de una página Web a los usuarios y a los robots de indexación de los buscadores [25]. De este modo, se puede recuperar una página en una búsqueda con la que no tiene relación alguna.

El modo más obvio de detectar el encubrimiento es comparar las versiones de la página recuperadas por distintos agentes (por ejemplo, los robots oficiales de Google o Yahoo!, y un robot "disfrazado" de usuario de Firefox). Sin embargo, este método está limitado por la naturaleza siempre cambiante de las páginas dinámicas: para usuarios distintos y en momentos distintos, los servidores ofrecen páginas Web con distintas noticias, anuncios, etc. Wu y Davison [25] proponen comparar copias distintas de la página obtenidas por dos robots de indexación y por dos robots simulando usuarios.

Si las páginas obtenidas por los robots son similares entre si, y las de los usuarios también, pero las primeras son diferentes de las segundas, la página es un buen candidato de encubrimiento. Como descargar cuatro versiones de cada página puede ser muy ineficiente, este método se refina en [26] filtrando en primer lugar aquellas páginas en las que no se encuentra diferencia significativa entre una versión para robot y una para usuario.

2.1.3. *Contenidos sensibles y su abuso*

Es claro que algunos tipos de información de la Web son inapropiados, dependiendo del sitio en el que se acceda a ellos. Por ejemplo, no es adecuado que los trabajadores de una compañía accedan a juegos de casino desde el puesto de trabajo, o en general, no es aceptable que los niños accedan a contenidos pornográficos. Con el objeto de evitar este tipo de abuso, que tienen un impacto económico evidente (en términos de tiempo de trabajo y ancho de banda, por ejemplo), se han desarrollado filtros y monitores y se ha extendido su uso por un número creciente de instituciones. Discutimos con detalle las técnicas usadas por estos sistemas en la **sección 3**.

2.1.4. Spam *de bitácoras*

Las bitácoras, Web *logs* o simplemente *blogs* pueden buenamente ser la más popular de las herramientas de la Web 2.0, y permiten a millones de usuarios tener su voz en la Web, creando redes sociales de extrema utilidad para la personalización de anuncios publicitarios. Como crear, editar y comentar bitácoras ha de ser una tarea tan simple, han sido especialmente objeto de abuso, especialmente con el objeto de canalizar visitas incluso a páginas Web que no son bitácoras, a través de comentarios masivos y de otras tácticas [16].

Las bitácoras son un género específico de la Web, con sus motores de búsqueda especializados como Technorati. Por tanto, se han propuesto técnicas específicas para detectar las distintas formas de *spam* en bitácoras, llamado con frecuencia "*splog*" (de *spam blog*). Por ejemplo, Kolari et al. [17] han propuesto la utilización de técnicas de Aprendizaje Automático aplicado a atributos del texto para la detección de *spam* de bitácoras; este enfoque es muy similar a algunos de los discutidos en nuestras **secciones 3 y 4**. Mishne et al. [19] se centran en los comentarios de bitácoras, y presentan un enfoque para la detección de comentarios *spam* comparando modelos del lenguaje empleado en las entradas de la bitácora, sus comentarios, y las páginas enlazadas desde los comentarios. En contraste con otras técnicas de filtrado de *spam*, este método no requiere aprendizaje o entrenamiento, ni reglas codificadas a mano, ni conocimientos globales de conectividad entre páginas Web.

2.2. *Spam* de enlaces

Probablemente, una de las razones principales del éxito de Google es la capacidad de sus diseñadores para reconocer la naturaleza social de la Web, es decir, ver la Web como un lugar en el que se construyen relaciones humanas. La información social más fiable en la Web está representada por los hiperenlaces, a través de los cuales los proveedores de contenidos demuestran su interés y confianza en el contenido de terceros. Hoy en día, la mayoría de buscadores hacen uso de alguna forma de medida de calidad basada en enlaces para los sitios Web, con un impacto directo en las posiciones de los mismos en sus listas de resultados. La más popular es la usada por Google, el PageRank [21]. En último término, estas métricas no son otra cosa más que variaciones de las medidas de prestigio usadas en la literatura científica, basadas en los grafos de citas.

Dado que obtener una posición alta en un buscador atrae usuarios a los sitios Web, esto tiene un valor económico importante, y en consecuencia, algunos proveedores de contenidos tienen un interés explícito en abusar de los posicionamientos basados en enlaces. El método más obvio para mejorar el posicionamiento de una página es construir una red altamente interconectada de páginas Web (una granja de enlaces – *link farm*) apuntando a ella, con la esperanza de que estos sitios propaguen el posicionamiento a través de ella hacia el sitio objetivo. En la **sección 4** se presenta una descripción detallada de los métodos usados para detectar y evitar este tipo de abuso.

2.3. *Spam* de uso

Otra medida importante usada en los buscadores es las pulsaciones (o *clicks*) de los usuarios. Si un usuario pulsa sobre una página tras una búsqueda, ello indica que se trata de una página de calidad. Aunque las pulsaciones sobre páginas están sesgadas por las posiciones otorgadas por el buscador, y por su interfaz, su distribución real puede aproximarse reduciendo este sesgo. Sin embargo, también es posible abusar de este tipo de medidas. ¿Cómo distinguir entre las pulsaciones efectuadas por un usuario real y las realizadas por un agente software? Por el momento, hay pocos resultados públicos sobre este problema, que incluyen la detección de IPs recurrentes o de patrones de pulsación anómalos.

Se trata de un problema muy complejo que está aún en investigación, en particular porque este problema está relacionado con una forma de *spam* de uso aún más importante: las pulsaciones falsas sobre anuncios. En este caso, cada pulsación implica un coste real, por lo que el *spam* de pulsaciones debe detectarse para cobrar a cada anunciante sólo las pulsaciones efectuadas por personas reales.

3. Caso de estudio: filtrado de contenidos Web

Una forma muy relevante de abuso es el acceso a contenidos Web inapropiados en el puesto de trabajo o en las escuelas. Las técnicas de análisis de contenidos usadas por los filtros Web son buenas representantes de la clasificación con adversario de contenidos, y se discuten a continuación.

3.1. El problema del acceso a contenidos Web inapropiados

La naturaleza autoreguladora de la publicación de contenidos Web, junto con la facilidad para hacer disponible la información en la Web, ha permitido la aparición de sitios Web con contenidos ofensivos, dañinos o incluso ilegales a lo largo y ancho de la Web. Por ello, la utilización de sistemas de filtrado y monitorización se ha convertido en una necesidad en los entornos educativos y en el puesto de trabajo, para proteger a los niños y prevenir el abuso de Internet.

En contraste con las tareas previas, los adversarios en esta tarea son tanto los proveedores de contenidos como sus usuarios. Por un lado, algunos proveedores de contenidos introducen material ilegal y dañino en la Web, como ciertos contenidos violentos (xenofobia y racismo), convenientemente disfrazados de opiniones políticas legítimas. Por el otro, los consumidores realizan un uso inadecuado de los recursos de Internet en el puesto de trabajo (por ejemplo accediendo a juegos de casino) o en las escuelas y bibliotecas (por ejemplo, accediendo a videojuegos o a sitios Web pornográficos). Aunque se requieren tanto iniciativas de carácter educativo como políticas de uso de Internet para prevenir este tipo de abuso, se pueden usar los filtros y monitores para detectar o evitar comportamientos anómalos[3].

3.2. Técnicas para el filtrado y monitorización Web

Existe una gran variedad de soluciones de filtrado disponibles en el mercado, incluyendo productos comerciales como Cyberpatrol o NetNanny, y sistemas de software libre como SquidGuard o DansGuardian. De acuerdo con las evaluaciones en profundidad de estos productos realizadas hasta la fecha (como la realizada en el proyecto europeo NetProtect [6]), la efectividad del filtrado está limitada por el uso de técnicas simples, como el bloqueo de URLs, o la detección de palabras clave. Existe la necesidad de enfoques más sofisticados e inteligentes para incrementar la efectividad de las soluciones de filtrado. En esta sección describimos las técnicas más populares para el filtrado de Web, y nos concentramos en la más prometedora, esto es, en el procesamiento inteligente de la información.

Las técnicas de filtrado de contenidos Web pueden clasificarse en cuatro grandes grupos [18]:

- Los etiquetados propios o de terceros, especialmente usando las escalas PICS (*Platform for Internet Content Selection*) o ICRA (*Internet Content Rating Association*). Los autores o revisores de páginas Web las etiquetan usando distintas etiquetas según sus tipos de contenidos y los niveles de los mismos, que luego se usan por los sistemas de filtrado para permitir o bloquear páginas de acuerdo con la configuración establecida por el administrador. Desgraciadamente, sólo una pequeña fracción de la Web está etiquetada, y los autores pueden etiquetar sus páginas de manera incorrecta bien por descuido o intencionadamente.
- Las listas de URLs (*Uniform Resource Locator*), es decir, mantener listas de sitios Web bloqueados y/o permitidos. Por ejemplo, una página Web se bloquea si su URL contiene una URL bloqueada, o si sus enlaces salientes apuntan a direcciones URL bloqueadas. Estas listas se pueden construir de manera manual o automática, pero es complicado mantenerlas actualizadas, y muchas veces no tienen en cuenta los alias de dominio.
- La detección de palabras clave, en la que se construye de manera manual o automática un conjunto de palabras o de frases clave ("sexo", "fotos gratis") a partir de una serie de páginas a bloquear (por ejemplo, pornográficas). Una página Web se bloquea si el número o frecuencia de las palabras clave que aparecen en ella excede un umbral determinado. Este enfoque es propenso al bloqueo por exceso, es decir, a bloquear paginas Web seguras en las que aparecen las palabras clave seleccionadas (por ejemplo, las páginas de salud sexual en el caso de bloqueo de pornografía).
- El análisis inteligente del contenido, que conlleva una comprensión más profunda del significado del texto y de otros elementos (especialmente las imágenes), usando técnicas de análisis lingüístico, aprendizaje automático, y procesamiento de imágenes. Este enfoque está limitado por el alto coste en la construcción de analizadores lingüísticos y de imágenes, su dependencia del dominio (por ejemplo, detectar desnudos en imágenes es muy distinto de detectar símbolos nazis), y por el retardo que el análisis detallado produce en la carga de las páginas.

Los tres primeros enfoques, ampliamente utilizados en las soluciones de filtrado actuales, se han demostrado bastante poco efectivos, y tienen limitaciones serias [6]. Nosotros opinamos que el análisis inteligente del contenido es factible, siempre que el diseño del sistema tenga en cuenta los temas de retardo, y las técnicas de análisis lingüístico y de imágenes se mantengan tan superficiales como sea posible. Por ejemplo, en el proyecto POESIA [15] se ha diseñado un sistema con dos niveles de filtrado: un filtrado ligero, para aquellas páginas que no son sospechosas o son claramente inapropiadas; y un filtrado pesado, para aquellas páginas en las que los filtros ligeros no son capaces de decidir con claridad. Las técnicas de análisis lingüístico y de imágenes usadas en los filtros ligeros son muy limitadas y eficientes, mientras que los filtros pesados utilizan métodos más avanzados (pero superficiales aún), que permiten tomas de decisión más precisas pero con un mayor retardo.

Como ejemplo de técnica de análisis inteligente del contenido, presentamos a continuación un enfoque de filtrado de Web basado en categorización automática de textos, usado en el filtro de pornografía en español dentro del proyecto POESIA.

3.3. El filtrado de contenidos como Categorización de Texto

La Categorización Automática de Texto consiste en la asignación automática de documentos en categorías predefinidas. Los documentos de texto son normalmente artículos periodísticos, informes científicos, mensajes de correo electrónico, páginas Web, etc. Las categorías son usualmente temáticas, como las clasificaciones bibliotecarias (por ejemplo, los descriptores de la Biblioteca Nacional de Medicina de los EE.UU.), las palabras clave de las bibliotecas digitales, las carpetas personales de correo electrónico, las categorías de los directorios Web (como las de Yahoo!), etc. Se pueden construir sistemas de Categorización Automática de Texto de manera manual (por ejemplo, confeccionando reglas de filtros de correo que archivan los correos en las carpetas personales), pero también de manera automática, entrenando un sistema de clasificación de texto con aprendizaje sobre un conjunto de documentos manualmente etiquetados. Este último enfoque, basado en aprendizaje automático, se ha convertido en el dominante, y las técnicas actuales permiten construir sistemas de clasificación tan precisos como lo son los seres humanos especializados [23].

La detección de pornografía ha sido tratado como un problema de Categorización de Texto en varios trabajos recientes, incluyendo por ejemplo [9] [10] [15] [18]. A partir de estos trabajos se puede modelar la detección de pornografía como un problema de aprendizaje automático sobre dos clases: inducir un clasificador que decida si una página Web es pornográfica o no. En la fase de aprendizaje, y dados dos conjuntos de páginas Web, uno pornográfico (P) y otro seguro (S), que conforman la colección de entrenamiento, se dan los siguientes pasos:

1. Cada página en P y S se procesa para extraer de ella el texto que incluye (dentro de las etiquetas TITLE, H1, P, META, etc.). El texto se divide en palabras, que pueden ser reducidas a su raíz, eliminando las más frecuentes (adverbios, pronombres, etc.), obteniéndose una lista de unidades de representación o términos.
2. Opcionalmente, se puede seleccionar un subconjunto de los términos de acuerdo con métricas de calidad como la Ganancia de Información. Este paso permite reducir la dimensionalidad del problema, acelerando el aprendizaje e incluso aumentado su eficacia. La lista resultante de términos constituye el léxico final.
3. Cada página se representa como un vector de pesos de términos (o de pares atributo-valor). Los pesos pueden ser binarios (un término aparece en una página Web o no), Frecuencia de Término (*Term Frequency* – el número de veces que un término aparece en una página), TF.IDF (el anterior multiplicado por la frecuencia inversa en documentos – *Inverse Document Frequency*), la frecuencia relativa de términos (*Relative Term Frequency*), etc. [23].
4. Finalmente se construye o se induce un clasificador usando un algoritmo de aprendizaje sobre la colección de vectores de entrenamiento y sus clases asociadas. Algunos de los algoritmos usados en este tipo de problemas incluyen los algoritmos probabilísticos Bayes Ingenuo y las redes de inferencia bayesianas [9], variantes de aprendizaje perezoso [10], redes neuronales semi-supervisadas [18], y las *Support Vector Machines* en su versión lineal [15].

Los pasos primero y tercero definen el modelo de representación del texto, que en este caso se denomina con frecuencia el modelo de "bolsa de palabras" Se corresponde con el Modelo del espacio Vectorial usado tradicionalmente en Recuperación de Información. La fase de clasificación consiste en, dada una nueva página cuya clase es desconocida, representarla usando un vector de pesos de términos similar a los usados para los documentos de entrenamiento, y clasificarla de acuerdo con el modelo generado en la fase de aprendizaje. Esta fase debe ser extremadamente eficiente, evitando largos retardos en la entrega de páginas Web cuando han sido clasificadas como seguras.

Hepple et al. (2004) [15] han aplicado esta técnica para construir filtros de pornografía basados en análisis de texto para el inglés y el español, alcanzando niveles de efectividad altos. Por ejemplo, el filtro inglés es capaz de detectar el 95% de las páginas Web pornográficas.

4. Caso de estudio: *Spam* de buscadores basado en enlaces

Como representante de las técnicas de análisis de enlaces en la Web, nos concentramos en esta sección en como los motores de búsqueda sacan partido de la naturaleza interconectada de la información en la Web, y de cómo este tipo de información es objeto de abuso. Describimos a nivel general las soluciones más prometedoras para este tipo de *spam*.

4.1. Ordenación de resultados basada en enlaces

La idea de la ordenación de resultados basada en enlaces es que la relevancia de una página con respecto a una consulta no puede determinarse únicamente sobre la base del contenido de la página (ya sea del texto o de otros tipos), sino también usando la estructura de hiperenlaces de la página y su vecindad, e incluso de toda la Web. La ventaja de este tipo de ordenación sobre las tradicionales, basadas exclusivamente en los contenidos de la página, es que proporciona una medida de la relevancia exógena, que además de resultar más fiable, parece más resistente al abuso.

Las técnicas de ordenación basadas en enlaces pueden clasificarse en dinámicas y estáticas. Las primeras ordenan los documentos con respecto a una consulta concreta (y sólo pueden computarse en el momento de la consulta), mientras que las segundas son independientes de la consulta, y se pueden interpretar como una medida absoluta de la importancia de un documento Web. Aunque en la literatura científica actual sobre este tema se discuten ambos tipos de ordenaciones, el uso masivo de los algoritmos dinámicos en entornos realistas es aún imposible a gran escala, mientras que las técnicas estáticas son enormemente populares, y se cree que la mayoría de los motores de búsqueda actuales han adoptado alguna forma de ordenación basada en análisis estático de hiperenlaces.

El tipo de ordenación basada en análisis estático de enlaces es el llamado grado entrante (*in-degree*): una página se considera importante si y sólo si tiene muchos enlaces entrantes, es decir, si hay muchas páginas (llamadas vecinos entrantes – *in-neighbors*) que tienen enlaces hacia ella. De hecho, el grado entrante es una medida trivial de popularidad: no es difícil abusar de ella, porque cualquiera puede crear muchos vecinos entrantes para una página con el único fin de engañar al algoritmo de ordenación, haciéndole creer que es más importante de lo que realmente es[4].

Una generalización bien conocida de esta idea es el algoritmo PageRank [21]. Una metáfora muy sugerente que describe la idea del PageRank es la siguiente: consideramos un proceso iterativo en el que cada página tiene una cierta cantidad de dinero que al final será proporcional a su importancia. Inicialmente, todas las páginas reciben la misma cantidad de dinero. Entonces, en cada paso, cada página reparte todo su dinero de manera equitativa entre las páginas a las que apunta. Esta idea tiene un límite, porque puede haber grupos de páginas que "absorben" dinero del sistema sin devolverlo nunca. Como se desea evitar la creación de esos oligopolios, se fuerza a cada página a que entregue una parte fija de su dinero al estado (una especie de impuesto), y el dinero recolectado de esta manera se distribuye entre todas las páginas.

Se puede demostrar que este proceso alcanza un comportamiento estable para cualquier tasa de impuesto entre 0 y 1 (pero estrictamente inferior a 1); este parámetro se denomina factor de suavización (*damping factor*), y es una medida de la importancia que damos a los enlaces Web (en particular, cuando todas las páginas tienen el mismo

PageRank). Por razones de tradición (que en parte son aun un misterio), el valor 0,85 funciona particularmente bien, y es el valor usado para el cómputo de PageRank en la mayoría de los casos.

Google fue el primer buscador en usar el PageRank, y muchos creen que la mayor parte de su popularidad proviene del éxito de esta técnica.

4.2. Detección de *spam* de enlaces

Con el creciente uso del análisis de enlaces y de la popularidad de las páginas para mejorar la precisión de los resultados de las consultas en los buscadores, un número creciente de diseñadores Web han empezado a desarrollar técnicas para engañar a los algoritmos de ordenación de resultados que se basan en enlaces, con el fin de mejorar el posicionamiento de sus sitios Web de manera inmerecida: estas técnicas se denominan *spam* de enlaces. Este fenómeno está en pleno crecimiento, y es enormemente relevante desde un punto de vista económico. Amit Singhal, científico jefe de Google Inc., ha estimado que la industria del *spam* de buscadores habría tenido un beneficio potencial 4.500 millones de dólares si hubiese sido capaz de engañar a los buscadores en todas las consultas comercialmente viables [3].

Se puede definir una página de *spam* de enlaces como una página que está conectada de manera que alcance una posición inmerecidamente alta en un buscador. Esta definición es, sin embargo, bastante vaga, porque puede depender de una consulta y de un buscador específico (o de manera más precisa, de las técnicas de ordenación empleadas por el buscador), y además requiere que un humano evalúe lo que significa "inmerecidamente alto", algo que es difícil de definir e incluso de decidir. Por otra parte, en lo que se refiere al PageRank y a otros algoritmos estáticos similares, el *spam* de enlaces no depende de la consulta, y el objetivo del mismo es lograr que una página parezca más importante o valiosa de lo que realmente es, engañando en consecuencia al algoritmo.

Como algunos han observado [3], una página pueden tener una posición inmerecidamente alta incluso si no es una página de *spam*; en este sentido, la definición del *spam* de enlaces depende de la intención del diseñador de la página, y es por ello relativamente subjetivo: podemos llamar *spam* al conjunto de páginas y enlaces que un diseñador no habría agregado a su sitio Web si los motores de búsqueda no existieran [22]. Merece la pena observar, de manera tangencial, que lo que nosotros llamamos "*spam* de enlaces", también se denomina "optimización de buscadores" (*search engine optimization*, SEO) por parte de los diseñadores de sitios Web, una expresión que tiene la intención de inducir al oyente la percepción de que el *spam* es un comportamiento correcto, o al menos, inocente. También pueden encontrarse sitios cooperando de manera encubierta para mejorar sus posiciones [2], por lo que es muy difícil distinguir las optimizaciones éticas de las que no lo son. En especial, ¿hay *spam* cuando los buenos sitios Web tratan de contrarrestar el *spam* usando las mismas técnicas que los abusadores?

Como muchos autores han hecho notar [2] [4], el *spam* de enlaces se suele poner de manifiesto a través de la creación de un conjunto de páginas o dominios con una estructura fuertemente conectada, llamados usualmente "granjas de enlaces" (*link farms*), diseñados para "canalizar" posicionamiento hacia las páginas objetivo. Los enlaces creados con el único fin de hacer *spam* de enlaces se llaman a veces "nepotísticos" (*nepotistic*). Algunos autores han estudiado específicamente el impacto de las granjas de enlaces en los valores del PageRank [2] [8] [27], que acaba dependiendo del factor de suavización y de la estructura de la granja de enlaces propiamente dicha. En [13] se presenta un estudio sobre la estructura óptima de una granja de enlaces, que es aquella que maximiza la ganancia en posicionamiento.

Este último trabajo constituye un esfuerzo en proporcionar una taxonomía del *spam* de enlaces. En lo que a los abusadores concierne, existe un conjunto de dominios en los que tienen todo el poder, y en los que pueden crear nombres de servidores y páginas prácticamente sin coste económico o de esfuerzo. Existe además un segundo conjunto de páginas sobre las que tienen un poder moderado, como los blogs (en los que pueden incorporar comentarios). Finalmente, existe un tercer conjunto formado por una gran parte de la Web, sobre el que no tienen ningún control. Por supuesto, la detección de *spam* se basa en gran medida en la comparación entre las páginas de *spam* y la parte de la Web inaccesible a los abusadores. Algunas técnicas usadas por los abusadores incluyen la creación de tarros de miel (*honey pots*, sitios Web que contienen recursos útiles como manuales, etc., y que también contienen enlaces a las páginas objetivo para potenciar su posicionamiento), la infiltración en directorios Web públicamente accesibles, la inclusión de comentarios con enlaces de *spam* en bitácoras y foros de mensajes (usualmente usando robots o programas automáticos), la compra de dominios cuya posesión ha expirado, etc.

4.3. Detección y penalización

Aunque el *spam* de contenidos puede ser identificado en la mayoría de los casos por inspección directa, el *spam* de enlaces es mucho más escurridizo [4]: las páginas diseñadas para engañar a los algoritmos de posicionamiento pueden parecer completamente inocentes y su contenido puede parecer valioso (con frecuencia, incluso ha sido copiado de páginas Web que no son *spam*). Es absolutamente necesario intentar diseñar algoritmos para la *detección* automática de páginas no fiables. Algunos algoritmos intentan identificar las páginas que son *spam* en si mismas, es decir, que existen con el solo propósito de mejorar el posicionamiento de otras páginas. Otras técnicas se basan en la búsqueda de enlaces nepotísticos, es decir, enlaces usados para canalizar posicionamiento inmerecido.

Una vez que un enlace o una página ha sido marcado como no fiable, debemos encontrar un modo de *penalizarlo* con el fin de que se reduzca su importancia en relación con la que tendría en otro caso. La solución más radical consiste en eliminar directamente la página, desterrándola del índice de manera definitiva e incluyéndola en una lista negra para no ser visitada en el futuro. El buscador puede ser más indulgente, y poner la página en cuarentena: la página no aparecerá en el índice, pero se visitará en el futuro para darla una segunda oportunidad de redención. Una acción aún más indulgente (pero sensata y efectiva) consiste en penalizar la página para que reciba un posicionamiento menor que el que habría recibido normalmente.

Con frecuencia, la detección y la penalización se implementan de manera integrada en la técnica de ordenación de resultados, de modo que esta obtiene los mismos valores que otro algoritmo estándar (como el PageRank) sobre las páginas legítimas, penalizando al mismo tiempo las páginas de *spam*.

Por supuesto, existen soluciones más triviales que algunas veces son efectivas:

- Mantener una lista negra de páginas que abusan de los enlaces entrantes.
- Eliminar los enlaces dentro de un mismo servidor o incluso en el mismo dominio [8].
- Contar el número de nombres de dominio para una misma dirección IP: si el número es muy alto, es muy probable que se trate de un sitio *spam* [11].

Algunos enfoques para detectar el *spam* de enlaces, que se pueden considerar clásicos a fecha de hoy, consisten en tratar de clasificar las páginas como *spam* o legítimas con ayuda de un clasificador estándar que utiliza atributos adecuados de las páginas, de un modo muy similar a como se detecta el correo electrónico basura. Uno de los primeros trabajos en esta línea fue el de Davison [8], que trató de caracterizar los enlaces nepotísticos usando un clasificador C4.5 con atributos puramente sintácticos (como

por ejemplo, si los enlaces son entre páginas del mismo servidor o dominio, el número de palabras en común entre los títulos de las dos páginas, etc.), alcanzando una efectividad muy alta (por encima del 90% en todos los casos). En [1] se siguió un enfoque parecido en un marco más general de categorización de páginas.

Más recientemente, y en la misma línea de trabajo, Baeza-Yates et al. [2] han tratado de combinar un gran número de medidas basadas en enlaces, y usarlas como atributos para la clasificación automática por medio de árboles de decisión, alcanzando una tasa óptima de detección de *spam* en torno al 79% con sólo un 2,5% de falsos positivos. Los atributos incluyen el grado entrante y saliente de la página, la reciprocidad (la fracción de vecinos salientes que también son entrantes), la variabilidad (la razón entre el grado externo de la página y el grado externo de sus vecinos), el PageRank máximo del sitio, la desviación típica, el PageRank de sus vecinos, etc.

4.4. Difundiendo la confianza y la sospecha

Existe la creencia entre la comunidad SEO de que algunos motores de búsqueda, y en especial Google, están haciendo uso de un algoritmo de detección de *spam* de enlaces llamado BadRank [24]. Este algoritmo comienza con un conjunto de páginas marcadas como malas (*bad*), y usa una técnica similar al PageRank, pero invirtiendo el sentido de los enlaces, para difundir la "maldad". La justificación de este algoritmo es que una página es mala si apunta a muchas páginas malas; de modo opuesto, el eslogan del PageRank es que una página es buena si es apuntada por muchas buenas. Aunque todos suponen que esta técnica está en uso en Google, no conocemos ninguna evidencia de ello.

Un algoritmo conceptualmente similar, pero dual, es TrustRank [14]: comienza con un conjunto de nodos fiables, y usa un algoritmo de difusión tipo PageRank para extender el valor de confianza a través del grafo. Los nodos con un valor pequeño de confianza se penalizan como sospechosos.

Una idea relacionada [24] es tratar de localizar las granjas de enlaces consistentes en TKCs (*Tightly Knit Communities*, comunidades densamente entretejidas), siguiendo el mismo camino que en [4], pero obteniendo un algoritmo que se puede aplicar fructíferamente a grandes grafos (del tamaño de la Web). Se considera que una página es "mala" si y sólo si hay demasiados dominios que son simultáneamente vecinos entrantes y salientes de la página. El conjunto de páginas malas se denomina semilla. A continuación, la semilla se expande recursivamente de acuerdo con la idea de que una página que apunta a muchas páginas malas también lo es, como en el BadRank. Finalmente, se computa el PageRank penalizando o eliminando los enlaces de sitios malos. También se puede decidir la calidad de una página usando las dos medidas: fiabilidad y abuso.

4.5. Medidas estadísticas de irregularidad

Una observación clave que subyace a numerosos algoritmos de detección de *spam* es que las páginas de *spam* presentan determinados atributos que no siguen el comportamiento estadísticamente familiar de las páginas legítimas; esta observación se ha tenido en cuenta en los métodos basados en clasificadores expuestos arriba. En algunos trabajos recientes [27] se ha demostrado experimentalmente que existe una correlación entre el abuso y el modo en que los valores del PageRank reaccionan ante cambios en el factor de suavización. Esta observación sugiere que se puede calcular el PageRank con distintos valores de este factor, y penalizar aquellos nodos que muestren un comportamiento sospechoso. Este enfoque ha sido objeto de críticas porque exige múltiples cómputos del PageRank, lo que parece computacionalmente inadmisible. Sin embargo, en [5] se ha presentado una técnica que permite aproximar el PageRank para todos los valores del factor de suavización con un esfuerzo computacional que es sólo un poco mayor que el necesario para calcular un solo vector de PageRank.

Un algoritmo muy general basado en la irregularidad estadística es SpamRank [3].
SpamRank actúa en tres fases:

- En primer lugar, se usa un algoritmo iterativo de Monte Carlo para determinar el conjunto de partidarios (*supporters*) de cada página, definido como aquellas páginas que realizan una contribución mayor al PageRank de la página.
- A continuación, se calcula alguna medida de regularidad para el conjunto de partidarios de una página dada, como la correlación estadística (por ejemplo, la de Pearson) entre los valores de PageRank y una distribución de Zipf ideal. Esta medida se basa en la observación de que los valores del PageRank siguen la distribución de Zipf, y que esta propiedad es en gran medida independiente del subconjunto del grafo Web considerado en cada momento.
- Finalmente, si el conjunto de partidarios es muy irregular, se penalizan de manera proporcional a su irregularidad.

5. Tendencias en Recuperación de Información con Adversario en la Web

La naturaleza evolutiva de la Web conlleva la aparición de nuevas aplicaciones de moda, y en consecuencia, de nuevas forma de abuso. Aquí pretendemos mostrar los retos a los que se enfrentan los operadores de buscadores y los usuarios de la Web en el futuro cercano, con respecto a estas aplicaciones.

A medida que las aplicaciones de Redes Sociales se vuelven más y más populares, atraen en mayor medida la atención de los abusadores. Algunas formas de mercadotecnia pueden ser consideradas legítimas. Por ejemplo, los estudios y las productoras de cine recurren a los sitios de intercambio de videos como YouTube para promocionar sus próximos estrenos. Este tipo de mercadotecnia es legítimo, en tanto que son los usuarios los que difunden su interés como lo hace normalmente para otros contenidos más informales.

Pero los sitios Web de Redes Sociales están plagados de *spam*. Hay formas muy específicas de *spam*. Por ejemplo, como los videos de YouTube se presentan al usuario a través del fotograma justamente intermedio, algunos abusadores suben videos con anuncios en ese fotograma. Por otra parte, cuando se ha informado a la gente de que el avión de Google Maps esta tomando fotos de su vecindad, entonces tratan de lograr que se graben mensajes pseudocomerciales, como ha ocurrido recientemente en Sydney (Australia)[5]. Otros sitios afectados por formas específicas de *spam* son la Wikipedia, el Open Directory Project, Flickr, Orkut, Digg, etc.

Sin embargo, la forma más pura de *spam* en Redes Sociales (llamado recientemente "snam", *Social Network Spam*) es la utilización de la funcionalidad de mensajería FOAF (*Friend of a Friend*, amigo de un amigo) que aparece en este nuevo tipo de redes. Por ejemplo la red Orkut de Google, con más de 200.000 usuarios, permite enviar mensajes de este tipo. Con frecuencia, los mensajes FOAF contienen promociones de conferencias u ofertas laborales, y aunque pequeños en volumen por el momento, exigen una acción diaria por parte de los administradores de las Redes Sociales. De hecho, este tipo de *spam* también afecta a los blogs.

Las formas específicas de *spam* requieren métodos específicos de detección y control. Sin embargo, pensamos que la mayoría de las técnicas usadas actualmente para atajar el *spam* en bitácoras pueden perfectamente ser útiles para afrontar el *spam* en Redes Sociales, y sin duda deben considerarse un punto de partida muy útil.

Referencias

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. HYPERTEXT'03: *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pp. 38–47, New York, NY, USA, ACM Press, 2003.

[2] R. Baeza-Yates, C. Castillo, V. López. Pagerank increase under different collusion topologies. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[3] A. Benczúr, K. Csalogány, T. Sarlós, M. Uher. Spamrank—fully automatic link spam detection work in progress. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[4] K. Bharat, M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA, ACM Press, 1998.

[5] P. Boldi, M. Santini, S. Vigna. Pagerank as a function of the damping factor. *Proceedings of the 14th international conference on World Wide Web*, pages 557–566, New York, NY, USA, ACM Press, 2005.

[6] S. Brunessaux, O. Isidoro, S. Kahl, G. Ferlias, A. Rotta Soares. NetProtect report on currently available COTS filtering tools. Technical report, *NetProtect Deliverable NETPROTECT:WP2:D2.2 to the European Commission*, 2001. Disponible en: <http://www.netprotect.org>.

[7] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma. Adversarial classification. *Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA, August 2004). KDD '04. ACM Press, New York, NY, 99-108.

[8] B. Davison. Recognizing nepotistic links on the web. *Artificial Intelligence for Web Search*, pages 23–28. AAAI Press, 2000.

[9] L. Denoyer, J.N. Vittaut, P. Gallinari, S. Brunessaux. Structured multimedia document classification. DocEng '03*: Proceedings of the 2003 ACM Symposium on Document Engineering*, ACM Press, 153–160.

[10] R. Du, R. Safavi-Naini, W. Susilo. Web filtering using text classification. *Proceedings of the 11th IEEE International Conference on Networks*, 2003, Sydney, IEEE, 325–330.

[11] D. Fetterly, M. Manasse, M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. *Proceedings of the 7th International Workshop on the Web and Databases*, pp. 1–6, New York, NY, USA, ACM Press, 2004.

[12] J.M. Gómez, E. Puertas, M. Maña. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data*, Palais du Grand Large, St-Malo / France, March 13-15, 2002.

[13] Z. Gyöngyi., H. Garcia-Molina. Web spam taxonomy. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[14] Z. Gyöngyi., H. Garcia-Molina, J. Pedersen. Combating web spam with TrustRank. *Proceedings of the 30th International Conference on Very Large Databases*, pp. 576–587, Morgan Kaufmann, 2004.

[15] M. Hepple, N. Ireson, P. Allegrini, S. Marchi, J.M. Gómez. NLP-enhanced Content Filtering within the POESIA Project. *Fourth International conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, May 26-28, 2004.

[16] P. Kolari, A. Java, T. Finin. Characterizing the Splogosphere. *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, WWW Conference 2006*, <http://www.blogpulse.com/www2006-workshop/>.

[17] P. Kolari, A. Java, T. Finin, T. Oates, A. Joshi. Detecting Spam Blogs: A Machine Learning Approach. *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. July 16–20, 2006, Boston, Massachusetts. Published by The AAAI Press, Menlo Park, California.

[18] P. Lee, S. Hui, A. Fong. A structural and content-based analysis for web filtering. *Internet Research* 13, 27–37, 2003.

[19] G. Mishne, D. Carmel, D. Lempel. Blocking Blog Spam with Language Model Disagreement. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web* (AIRWeb 2005).

[20] A. Ntoulas, M. Najork, M. Manasse, D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 83-92.

[21] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank citation ranking: Bringing order to the web. *Technical Report 66, Stanford University*, 1999. Disponible en <http://dbpubs.stanford.edu/pub/1999-66>.

[22] A. Perkins. White paper: *The classification of search engine spam*, Septiembre 2001. Disponible en <http://www.silverdisc.co.uk/articles/spam-classification/>.

[23] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1–47, 2002.

[24] B. Wu, B. Davison. Identifying link farm spam pages. *Proceedings of the 14th International World Wide Web Conference*, Industrial Track, May 2005.

[25] B. Wu, B. Davison. Cloaking and Redirection: A Preliminary Study. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web* (AIRWeb 2005).

[26] B. Wu, B. D. Davison. Detecting semantic cloaking on the web. *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May, 2006). WWW'06. ACM Press, New York, NY, 819-828.

[27] H. Zhang, A. Goel, R. Govindan, K. Mason, B. Van Roy. Making eigenvector-based reputation systems robust to collusion. *Proceedings of the third Workshop on Web Graphs* (WAW), volume 3243 of Lecture Notes in Computer Science, pages 92–104, Rome, Italy, Springer, 2004.

Notas

[1] Spam es una palabra de difícil traducción al castellano. Su origen es un chiste de los humoristas británicos Monty Pyton, y ha sido aplicada tradicionalmente al correo electrónico masivo no deseado, llamado también correo basura. Su uso se ha extendido a otros ámbitos de Internet y de la comunicación.

[2] <http://www.google.com/press/zeitgeist2006.html>.

[3] En nuestra opinión, el filtrado de contenidos Web inapropiados en el puesto de trabajo o en las escuelas no debe verse como censura, sino como prácticas de refuerzo de políticas acordadas, y como una optimización razonable de los recursos de tiempo y ancho de banda.

[4] Las cosas pueden hacerse un poco más difíciles si se cuenta el número de servidores distintos que apuntan a una página: de este modo, los abusadores se ven forzados a dispersar los enlaces de spam por gran variedad de servidores distintos, que pueden ser costosos y difíciles de obtener.

[5] <http://blogs.smh.com.au/mashup/archives/009502.html>.