VNiVERSiDAD
D SALAMANCA

UNIVERSIDAD DE SALAMANCA
Departamento de Informática y Automática

La incidencia del Web Spam en los sistemas de Recuperación de Información

---

Resumen de Tesis Doctoral

D. Armando Carlos Costa Carvalho

Director:

Dr. D. José Luis Alonso Berrocal

Enero 2010

Jose Luis Alonso Berrocal, *Profesor Titular de Universidad del Departamento de Informática y Automática de la Universidad de Salamanca*

HACE CONSTAR: *Que* D. Armando Carlos Costa Carvalho, *Licenciado in Informática por lo IPA - Instituto Politécnico Autónomo en Lisboa (Portugal) ha realizado bajo mi dirección la Memoria que lleva por título* La incidencia del Web Spam en los Sistemas de Recuperación de Información, *con el fin de obtener el grado de Doctor por la Universidad de Salamanca.*

Y para que surta los efectos oportunos firmo en Salamanca, a trinta de enero de dos mil diez.

Contenido

Capítulo 1

Introducción

Este documento es un resumen de la tesis que presento conjuntamente, pero escrita en portugués. Por eso, en primer lugar, agradezco la oportunidad de poder haber escrito toda la obra en mi lengua materna y presentar en castellano tan sólo este resumen.

La organización de este resumen corresponde enteramente a la tesis, tanto en la numeración de los capítulos como en los temas que son abordados.

Para reducir, y poder ser clasificada como resumen, suprimí detalles explicativos, en la mayor parte de los casos, de encuadramiento o de mejor presentación de ideas, por lo que, en esta forma resumida, los temas pueden, a veces, aparecer de forma demasiado abrupta, dadas las reducciones realizadas desde el original.

Por la misma necesidad fueron suprimidas la mayoría de las referencias bibliográficas, presentándose al final una bibliografía tan sólo de las principales obras mencionadas en la tesis.

El trabajo, encuadrado dentro de las tecnologías de Recuperación de Información (RI) en ambiente WEB, se centra en el estudio particular de las dificultades causadas por sofisticadas introducciones de SPAM, deteriorando los resultados de las búsquedas efectuadas por los motores de búsqueda, y se encuentra dividido en 6 capítulos organizados de la siguiente forma:

1. En el primer capítulo se aborda el concepto de la RI general y en

contexto web; se presentan algunos conceptos de búsqueda y de motores de búsqueda.

2. En el segundo capítulo se presenta el Web Spam, a nivel de definición, propiedades y evolución histórica de los motores de búsqueda y de algunos algoritmos de ranking; evolución de motores de búsqueda y la identificación de sus propiedades fundamentales.

3. En el tercer capítulo mencionamos estrategias con el objetivo de mejorar la visibilidad de los sites. Dedicamos algún estudio a los SEO - Search Engines Optimization.

4. En el cuarto capítulo dedicamos especial atención a los principales tipos de Web Spam conocidos y a las formas desarrolladas en investigación para combatirlos.

5. En el quinto capítulo se abordan estudios prácticos de clasificación de sites, en el sentido de colaborar con la validación de modelos matemáticos. Presentamos un caso de estudio relacionado con la forma en cómo los seres humanos pueden verse influenciados por sus vecinos (sociedad física o sociedad digital) en la forma en cómo clasifican o evolucionan su clasificación binaria y las dificultades que puede provocar la clasificación de 'borderline'.

6. En el sexto capítulo se presentan las principales conclusiones del trabajo desarrollado y las posibles líneas futuras de investigación que se abren a partir de este trabajo

1.1. Objetivos

Obtener, a partir de una colección de documentos, **aquellos que satisfagan una necesidad específica del usuario, de tal forma que la mayor parte de los documentos recuperados sean relevantes para esa necesidad**.

Hacemos referencia a que la cuestión de la 'Problemática de la RI' puede ser estudiada desde dos puntos de vista: **el computacional y el humano** y que, desde un elevado nivel de abstracción, puede caracterizarse por:

- ◇ **Existir una colección de documentos** que contienen información realmente de interés sobre variadísimos temas
- ◇ **Existir usuarios** interesados en acceder a esa información
- ◇ **Retornar el sistema, como respuesta, a una lista ordenada** de referencias a documentos relevantes para el usuario / pregunta.

Lo que implica necesidades de precisión en la respuesta que debe obedecer a un concepto subyacente de relevancia en relación a lo solicitado en la pregunta. Se mencionan en el estudio diversos modelos de RI, de entre los cuales se destacan: los Clásicos, los Booleanos, los Vectoriales y los Probabilísticos, y que son construidos para responder de forma más objetiva al criterio de relevancia.

El crecimiento exponencial de la Web transformó conceptos y necesidades: formato de los documentos, tamaño, número de fuentes de información disponibles, evolución de los recursos de hardware, de los recursos de comunicaciones, pero, por encima de todo, el alargamiento al comercio electrónico y a nuevas formas de transacciones conducentes a movimentación de capitales.

Esta expansión de Internet, basada en **Nuevas tecnologías** y **Software de comunicación fáciles**, transformó la producción para la Web en una producción favorecida. De un momento para el otro todos quieren tener (casi) todo en Internet, para aprovecharse de su naturaleza:

- **Desreglamentada**: Sin dueño ni normas de utilización.
- **Descentralizada y abierta**: Disponible siempre y desde cualquier local.
- **No jerarquizada e interactiva** posibilitando el desdoblamiento jerárquico entre emisores y receptores, creando una 'inteligencia colectiva' que permite 'reciprocidad en la comunicación y la división de un contexto', creando una nueva idea de 'una cultura humana de producción'.

Tales características confirman que **los motores de búsqueda son la puerta de entrada para la web actual** - Surgió una nueva necesidad: buscar en tan vasto universo.

En la tesis presentamos algunos modelos y conceptos de motores de búsqueda, que poseen básicamente tres módulos:

- **El colector (crawling)** - formado por robots que son responsables de la recogida de los documentos web;
- **El indexador (indexing)** - que incluye los documentos en el índice de la máquina de origen, poniéndolo disponible para consultas y
- **El módulo de consultas (searching)** - que ofrece la interfaz con el usuario.

El conjunto de resultados es sometido a **algoritmos de ordenación por relevancia**, para atribución de métricas que relacionan los documentos indexados con las informaciones contenidas en los índices.

Existen varios tipos de motores de búsqueda:

- **Globales**: son motores que buscan en la web creando listas organizadas de respuesta.
- **Verticales o temáticos**: devuelven referencias a documentos individuales, 'especializadas' de acuerdo con sus especializaciones temáticas. Ejemplo www.youtube.com
- **Guías locales** de ámbito territorial más reducido: tan solo a nivel local o regional.
- **Guías de búsqueda local** *o buscador local*: este tiene un ámbito nacional y la lista de las empresas y prestadores de servicios próximas a la dirección del internauta a partir de un texto .
- **Diretorios de websites**: son índices de sites.
- **Onto-buscadores**, buscadores basados en Ontologías, que deben responder a algunas preocupaciones.

Además

- Hay un crecimiento rápido que dificulta a los 'crawler' indexar toda la información.

- Muchas páginas se actualizan constantemente.
- Falsos positivos.
- Sites generados dinámicamente.
- Ordenación por coparticipación monetaria (tasa).
- Trucos para manipular los mecanismos de busca y obtener posiciones más favorables en la clasificación final

Principalmente por esta última razón es importante que prestemos atención al capítulo siguiente sobre la temática del Web Spam.

Capítulo 2

Motores de búsqueda y algoritmos de clasificación

En este capítulo se abordan los conceptos más directamente relacionados con la tesis, principalmente a nivel de identificación de los diversos algoritmos usados durante las fases de *crawling, indexing y searching* de los motores de búsqueda y de cómo son discutidos los primeros lugares de las listas de clasificación, principalmente con la introducción de técnicas para eludir los algoritmos.

2.1. Lo que es *Web Spam o Spamdexing*

Cualquiera que sea la definición se concluye que, también por razones sociales, *Spam* hace referencia a **algo indeseable, mismo perturbador, que influencia negativamente el proceso de selección de información tratada en ambiente web, con utilización de los protocolos disponibles**.

Nos centramos, a continuación, en el modo de operar de los spammers sobre los procesos de RI en la web, evidenciando su ligación a los motores de búsqueda, y a las indisociables estrategias de ranking.

Con esta finalidad verificamos que los motores de búsqueda ven el Web Spam como una interferencia para sus operaciones. Los estudios para limitar las restricciones encuentran dificultades en la identificación

automática de spam con base tan solo en algoritmos matemáticos (graph isomorphism) [Bharat *et al.*, 2001; Metaxas & DeStefano, 2005]. En efecto, **necesitamos comprender socialmente la cuestión del Web Spam** y sólo después analizar las cuestiones técnicas, dado que es en el área de los comportamientos sociales donde el spam está siempre más actualizado.

El principal modo de operar de los *spammers* es promover el ataque a los motores de búsqueda a través de texto [Gyongyi & Garcia-Molina, 2005; Henzinger *et al.*, 2002; Metaxas & DeStefano, 2005] y de la manipulaciones de links: **Text spam** y **Link spam**.

2.2. Evolución de los motores de búsqueda y de los modelos de clasificación

Analizamos también lo que habrá sido la primera forma de clasificación: Palabras raras *TF.IDF ranking*, o sea, **cuanto más palabras raras (poco usadas) dos documentos compartiesen, más semejantes eran consideradas** [Henzinger, 2001; Metaxas & DeStefano, 2005]. De este modo una búsqueda Q es tratada como un documento corto, como mínimo de una sóla palabra, y los resultados de una búsqueda de Q son clasificados de acuerdo con su similitud con la consulta (palabras raras coincidentes).

El primer ataque a este 'tf.idf ranking', como es conocido, vino de dentro de los propios motores de búsqueda. Más o menos en 1995, los motores de búsqueda comenzaron a vender palabras clave para anunciantes como una forma de generar recetas.

Para evitar los spammers, los motores de búsqueda intentan mantener secretos sus algoritmos de ranking. No obstante este secreto no duró mucho, pues no consiguió resistir a las nuevas técnicas de reingeniería (reverse engineering) [Marchiori, 1997; Metaxas & DeStefano, 2005; Pringle *et al.*, 1998].

La segunda generación de motores de búsqueda - como el modelo de Metaxas [Metaxas & DeStefano, 2005] - inició una técnica más sofisticada para la clasificación en un esfuerzo por anular los efectos de las 'palabras raras'.

Una de las más exitosas técnicas se basó en el **principio de la votación por 'link'**: Cada web site *s* tiene un valor igual al de su popularidad, que está influenciada por un conjunto de sites *Bs* que apuntan hacia ese site *s*.

La introducción de un nuevo método designado por **PageRank**, en 1998, fue uno de los principales desarrollos para los motores de búsqueda, porque este incorpora mayor sofisticación para proporcionar una solución anti-spam. **En el PageRank, ni todos los links contribuyen igualmente en la reputación de una página**. En vez de eso, las relaciones de alta reputación contribuyen de forma muy superior a links de otros sites menos reputados. De esa forma, las redes de sites desarrolladas por spammers (clusters y otros) no irían a influenciar mucho su propia PageRank.

El concepto inherente al PageRank [Gyongy *et al.*, 2004] es que **una página web es realmente importante si otras páginas importantes apuntan hacia ésta**, donde, el PageRank está basado en un refuerzo mutuo entre las páginas: la importancia de una determinada página *influencia* y *es influenciada* por la importancia de otras páginas.

Surge así el concepto de **popularidad del site** siendo una medida de **cantidad y de calidad de los links apuntados hacia éste**. Un site que recibe muchos links con origen en otros de contenido de calidad y que también tengan alta popularidad, será considerado de alta popularidad y podrá ser considerado relevante para determinadas búsquedas.

En otras palabras, un site es considerado importante cuando otros sites importantes lo recomiendan a través de links. Cada link es considerado un voto y los votos de sites importantes tienen un peso mayor. También podemos decir que los links de sites de contenido relacionado con el suyo tiene un peso mayor que el de sites de contenido no relacionado.

Una de las deficiencias del algoritmo de PageRank es que cualquier link, en cualquier página contenida en el índice, aumenta el PageRank (y mejora el ranking) de la página que recibe el link. Entre otros, existen dos problemas mayores que preocupaban a los investigadores: la compra de In-links y link-farms.

El algoritmo **Hilltop** [Bharat & Mihaila, 2000] responde positivamente a la resolución de estos dos problemas, al intentar detectar hosts

afiliados, pero si un link apunta hacia una página en un hosts afiliado, el valor del link es descontado.

Este mejoramiento permite que los links dejen de tener el mismo peso, pues mientras que el PR determina la 'authority' de una página dentro del concepto general, el algoritmo Hilltop (LocalScore) determina la 'authority' de una página con respecto al objetivo específico de los términos de búsqueda.

El Algoritmo **TrustRank**, presentado por [Gyongy *et al.*, 2004], pretende separar, de forma semiautomática, los websites acreditados de los que contengan spam, según el concepto de 'Atenuación por Confianza' (*trust attenuation*).

Para este algoritmo todo comienza con la creación de semillas (seed) que necesitan, a través de la intervención de especialistas humanos, ser catalogadas como páginas buenas o páginas conteniendo spam. Su contribución para la detección de spam es de tal manera importante que fue incluido en los algoritmos de los principales motores de búsqueda.

Influenciable por la Edad del dominio (historia), por la calidad de los 'Back links' y por la originalidad de los contenidos, el TrustRank considera el alejamiento de las semillas como un 'atenuante' según dos vertientes: Trust Dampening y Trust Splitting.

Después de identificar manualmente la reputación de las semillas, utiliza la estructura de links de la web para descubrir otras páginas que sean susceptibles de ser buenas.

Explicamos el algoritmo y sus etapas de implementación, con el pormenor de que la confianza se va reduciendo en la medida en que nos apartamos de las semillas buenas.

HITS, otra forma socialmente inspirada para determinar el ranking, [Kleinberg, 1999; Metaxas & DeStefano, 2005] divide los sites relacionados a una consulta entre 'hubs' y 'autoridades'. Hubs son sites que contienen apuntadores para muchas autoridades, en la medida en que las autoridades son locales y están apuntados por los hubs. La referencia mutua hace que ambos ganen créditos en la posición clasificadora.

El algoritmo de Kleinberg atribuye además a cada página dos índices, un 'hub index' y un 'authority index'. El peso de cada link dependerá de los índices hub y authority de la página en que se encuentra. El proceso

de cálculo es recursivo y puede envolver billones de páginas. Durante su concepción, el algoritmo se mostró impracticable porque exigía volúmenes enormes de recursos computacionales.

2.2.1. Las diversas generaciones de motores de búsqueda

Los algoritmos de Ranking (PageRank e HITS) marcaron el desarrollo de una nueva generación de motores de búsqueda, que pretendían dar una respuesta ordenada a las solicitudes de los usuarios. Infelizmente, una vez más los spammers han encontrado formas de esquivarlos e incluir respuestas indeseables en relación a la búsqueda pretendida.

Para controlar el PageRank los Spammers implantaron sites con conocimientos sobre asuntos irrelevantes y consiguieron adquirir una alta clasificación en sus sites especializados. Estos utilizaron la técnica de 'bandwagon' para su red de sites (vecindad), creando lo que se puede llamar 'sociedad de admiración mutua' (MAS). Por esta última razón, el modelo HITS se mostró muy permisivo a los spammers, debido al hecho de que su eficacia depende de la precisión del cálculo inicial de vecindad.

[Metaxas & DeStefano, 2005] resumen algunas búsquedas sobre las primeras generaciones de motores de búsqueda y los principales modelos de ataques que sufrieron.

SE	**Ranking**	**Spamming**	**Propaganda**
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation	+ mutual admiration	+ testimonials societies

Tabla: 2.1: Primeras generaciones de los motores de búsqueda

Esta 'cultura humana de producción' mencionada por [Schons, 2007], no jerárquica e interactiva: por su alto grado de interactividad, promueve la remodelación en la estructura del flujo de información. En esa perspectiva, el mismo autor, reflexiona acerca de la contribución entre los internautas y les atribuye el término 'inteligencia colectiva', porque todos pueden contribuir en la concretización de una 'tecnodemocracia' por medio de sus percepciones e inteligencias.

Estudiamos además otras formas de clasificación como fue el caso del modelo de Hoeschl [Hoeschl, 2006] que presenta una **clasificación temporal basada en generaciones** con respecto a los mecanismos de búsqueda en la Web.

Los especialistas en motores y técnicas de RI buscan una nueva generación de motores de búsqueda [Broder, 2002; Metaxas & DeStefano, 2005], que consigan interpretar las necesidades del usuario 'por detrás de la consulta', usando técnicas de interpretación semántica (Ontologías, RDF's, etc) de las consultas realizadas por el usuario.

Se pretende que la web sea capaz de aprender (almacenar, recuperar y procesar informaciones) de forma inteligente, similar a un gran cerebro global [Oliveira & Vidotti, 2004].

Considerándose como parámetro la mente humana, el conocimiento y significado derivan de un proceso de aprendizaje en el que, cuanto mayor es el uso de determinados conceptos, más fuertemente estos se unen. Para la web el análisis es semejante: **Con base en los caminos más recorridos por los internautas, algunas conexiones se vuelven más importantes, mientras que los links poco utilizados se vuelven menos importantes.**

La Web 3.0, como prototipo, tiene como premisa la de tener en cuenta el sentido de cada palabra del usuario, actuando con inteligencia e intuición. Sobre esta nueva fase de la web, que será dotada de la capacidad de 'aprender', 'raciocinar' y 'entender', [Johnson, 2003] apunta que:

> *Por la imposibilidad de la web llegar a ser semejante a la consciencia humana, no podemos probar que sea incapaz de aprender. Todo lo contrario: 'Una red de información adaptable, capaz de reconocer patrones complejos, podrá llegar a ser una de las invenciones más importantes de toda la historia de la humanidad'.*

También la robustez (scalability) de los motores de búsqueda se vuelve cada día más importante, pero son otros los factores preponderantes para la capacidad de los motores de búsqueda, como la relevancia de los resultados devueltos a los usuarios y la posibilidad de implementación de una clasificación independiente.

Scalability	Crawling	Agrega datos volviéndolos buscables, siendo responsable por toda la recuperación, procesamiento y almacenamiento de documentos; evocan lo imaginario, lo intangible y son muy rápidos
	Indexing	Módulo, dentro del concepto referencial de motor de búsqueda, tiene una colección de documentos o datos y construye un índice buscable a partir de éstos. Prácticas comunes son los archivos invertidos, espacios vectoriales, estructuras e híbridos de éstos.
	Searching	El 'searcher' trabaja a nível del indexador, *en la salida*. Acepta las consultas del usuario, las ejecuta a lo largo del índice y devuelve los resultados a la entidad que pregunta.
	Ranking	Infelizmente, los motores de búsqueda no tienen la capacidad de hacer algunas preguntas, como hace un bibliotecario, para centrar la búsqueda. Tampoco pueden invocar experiencias anteriores para clasificar las páginas de la web, de la misma forma en que nosotros, seres humanos, podemos. Entonces los indexadores siguen un conjunto de reglas, conocidas como algoritmos de ranking, cuyo objetivo es responder a la pregunta que originó la búsqueda con una lista de apuntadores para sitios que puedan dar respuesta a esa pregunta Esta lista está ordenada según los algoritmos de ranking, que han evolucionado con el tiempo y con la necesidad de superar las restricciones que van siendo detectadas.
Relevance	FeedBack	Idea polémica en cuanto al carácter de relevancia de los documentos, dado que la relevancia depende del usuario y no del buscador. [Hiemstra & Robertson, 2001] introducen el análisis del feedback del usuario, como siendo un concepto de relevancia, en la medida en que considera todas las acciones producidas por el usuario en documentos anteriormente recuperados para la construcción de una consulta, usando los modelos de lenguaje natural (language models) y el modelo probabilístico de independencia binaria (binary independence model)
Static Ranking	Content Quality	Incluye en los resultados de la ordenación, además del 'dynamic ranking', dependiente del query inicial, la idea de un 'static ranking' o clasificación independiente. Cómo: usando recursos que son independientes de la estructura de links de la web, como son los datos sobre la frecuencia con que los usuarios visitan las páginas en la web. Un ejemplo es el algoritmo de RankNet que clasifica los conjuntos de frases (factores de producción) en un documento.

Tabla: 2.2: Lo que se espera de los motores de búsqueda

Capítulo 3

Visibilidad de los sites

Search Engine Optimization (SEO) es simultáneamente un arte y una ciencia. [1]

La expresión 'optimización de sites para la búsqueda', hace referencia a un conjunto de estrategias con el objetivo de mejorar la posición en los resultados naturales (orgánicos) en los sites de búsqueda/buscadores.

Basado en la máxima de que los SEO's tan sólo pretenden '*colocar las piezas en el lugar correcto*' para obtener buena visibilidad en los motores de búsqueda, normalmente, crean solicitudes derivadas de esos sites para que se queden bien posicionados sin que lo merezcan.

Esta controversia generó un 'brazo de hierro' [Castilho *et al.*, 2006; Sydow *et al.*, 2008] entre los administradores de los motores de búsqueda, que intentan mantener estable la credibilidad que les crearon a los usuarios, contra todos los que usan técnicas designadas por 'black hat', que manipulan el orden natural del ranking. [Becchetti *et al.*, 2008; Svore *et al.*, 2007] hacen referencia también a la existencia de un gran área gris entre actitud 'ética', concretamente los 'white hat' defendida por los detentadores de las SEO's y la opinión de los administradores de los motores de búsqueda que clasifican esta actitud de spam 'antiético'.

[1] http://www.free-ebooks.net/ebook/apr07/MYSEOEGUIDE.pdf

3.1. Técnicas de 'Search Engine Optimization' - SEO

En el análisis de las diversas técnicas utilizadas por los SEO, conviene que admitamos que no todos los objetivos son coincidentes, concretamente en lo referente al público objeto que pretende alcanzar:

- **Un gran público objeto** - optimización de frases comunes. (Informativos Periódicos, Guías Locales, Sites con publicidad basada en CPM, etc).
- **Una mayor preparación para la venta** - optimización de palabras clave altamente específicas.

[Chambers, 2006 y Visser *et al.*, 2007] estudian un modelo marcando la diferencia entre elementos visibles que deben ser incluidos ('Essentials'), que pueden ser incluidos ('Extras'), los que deben ser evitados ('Cautions') y los que no deben ser utilizados ('Dangers').

Del análisis que efectuamos en este capítulo hay que considerar que si queremos usar SEO's de forma correcta, o 'White Hats' hay diversas formas de hacerlo. Destacan algunas:

- Crear referencias para los usuarios y no para las máquinas;
- Hacer los contenidos fácilmente accesibles por los indexadores;
- Y no intentar 'engañar' al sistema.

Desde el punto de vista de los sistemas de indexación y búsqueda, son importantes:

- Títulos cortos, exclusivos y relevantes para cada página del site;
- Encontrar la terminología correcta, objetiva y relevante para el contenido de la página de forma que se supriman formulaciones vagas;
- Aumentar la cantidad de contenido original, lo más exclusivo posible;

- Utilización razonable de los *metatag* sin uso excesivo de palabras clave u otras referencias fuera de contexto;
- Asegurar que todas las páginas son accesibles a través de relaciones regulares, y no tan solo a través de Java, Javascript o Macromedia Flash o incluso por redireccionamiento (meta refresh);
- Permitiendo que los mecanismos de búsqueda puedan indexar páginas del site sin tener que aceptar cookies o IDs de sesión;
- Estructura participativa con otros sites independientes (web ring), con partición de temas de calidad comparables;
- Escribir artículos informativos y útiles ofreciendo gratuitamente la posibilidad de impresión y uso, a cambio de un hiperlink que apunte hacia la fuente.

Como conclusión se puede decir que las prácticas de SEO recomiendan que los creadores de sites centren su acción en aquello que los motores de búsqueda realmente buscan, aliadas a las directrices de code publicadas por el World Wide Web Consortium3 : Contenido relevante y útil para los usuarios, identificado por:

- Contenido actualizado;
- Contenido útil;
- Contenido original;
- Contenido significativo;
- Inbound links.

Capítulo 4

Detección de Web Spam

4.1. Introducción

El Web Spam es reconocidamente uno de los principales desafíos en la industria de los motores de búsqueda [Henzinger *et al.*, 2002]. Se han detectado muchas técnicas diferentes [Collins, 2004; Gyongyi & Garcia-Molina, 2005; Perkins, 2001; Wu, 2007], pero todos reconocen que se trata de una lucha sin fin, pues en la medida en que se desarrollan 'vacunas', nuevas técnicas maliciosas aparecen con el único objetivo de confundir los algoritmos y las técnicas de indexación y de ranking y, por último, a los visitantes del site, sobre la verdadera naturaleza de su contenido [Taveira *et al.*, 2006].

El inmenso trabajo desarrollado en los últimos años en este área permite consenso, entre la mayoría de los investigadores, al clasificar las técnicas de spam en dos áreas específicas: **contenidos** [Gyongyi & Garcia-Molina, 2005; Ntoulas *et al.*, 2006] y **links** [Gyongy *et al.*, 2004;Wu & Davison, 2006]. Hay investigadores, como [Wu, 2007], que consideran una tercera gran división que se designa como 'page-hiding', donde engloba las técnicas de camuflaje y de redireccionamientos.

Chellapilla [Chellapilla & Maykov, 2007] define mejor este concepto al concluir que, en cuanto el 'link spam' y el 'content spam' son métodos que objetivamente pretenden alcanzar el PR, mejorándolo, otros métodos conocidos como de camuflaje (cloaking) y de redireccionamiento, esconden otras técnicas que no siempre son detectables en el momento

del acceso directo.

[Gyongyi & Garcia-Molina, 2005] aprovechando la división de Chellapilla, concluyen que las páginas que contengan Web Spam pueden ser subdivididas en 2 grandes grupos:

- **Boosting del Page Rank** - Los que usan técnicas para impulsar los rankings (link o content spam)
- **Camuflage** - Técnicas camufladas

Por otro lado, las técnicas de boosting pueden ser subdivididas en dos subcategorías: una referente a los contenidos y otra referente a los links.

Para juntarlas a este primer planteamiento hay que tener presente que las técnicas de spam son muy diversas, volviendo muy difícil a los motores de búsqueda detectar o crear métodos válidos de detección para todas las variantes. Por ejemplo los motores de búsqueda pueden usar métodos estadísticos para detectar 'keyword stuffing', pero estas estadísticas ya no son válidas en la detección de 'cloaking'. Es casi una guerra sin que sea visible el opositor.

La parte positiva de esta dificultad es que es bien reconocida la grandeza, y el continuo crecimiento, de la Web - y con esta los sistemas de IR - lo que hace imposible cualquier tipo de clasificación manual. ¡Todas las respuestas tienen que ser automáticas! Como complemento, tiene que haber garantía de pruebas en bases de datos suficientemente representativas y actualizadas, una vez que los algoritmos pueden responder en ambientes de pruebas, normalmente limitados, y no tener igual comportamiento en ambientes grandes y crecientes. Añade además que los algoritmos que funcionan hoy ¡no están garantizados mañana!

En este contexto, Fetterly desarrolló un estudio estadístico de las propiedades de las páginas con spam [Fetterly *et al.*, 2004] donde demostró que las páginas infectadas típicamente difieren de las páginas 'buenas' en varias funcionalidades. Esos desvíos funcionales fueron posteriormente usados por Ntoulas [Ntoulas *et al.*, 2006] para construir un clasificador de detección de spam. Entre otras aplicaciones, ese algoritmo, es aplicado en el estudio de la Spamrank [Benczúr *et al.*, 2005b].

En este capítulo analizamos los principales tipos de Spam identificados y algunas formas de detectarlos.

4.2. Tipos de Web-Spam

4.2.1. Content Spam

Content Spam	Keyword Stuffing	Es el uso de una o más palabras con la única finalidad de aumentar su frecuencia en un página
	K. Stufing en los títulos	Los títulos con más de 24 palabras tienen más probabilidad de ser spam que de ser páginas normales
	Meta tag Stuffing	es la repetición de palabras clave en la zona de Meta tags, pero de palabras no relacionadas con el contenido del site
	Non-markup caracteres	Estudia el tamaño medio de palabras, con respecto al número de caracteres
	Textos âncora	Texto áncora de los links como una referencia descriptiva del contenido de la página apuntada por el link
	Stuffing	Texto colocado en el atributo ALT de una imagen, pueden ser usados para producir resultados más relevantes para los propietarios de las páginas y de los sitios que se pretenden autopromover
	Compresibilidad	Análisis de repetición dentro de las páginas
	Palabras populares	Uso de alfabetos temáticos especializados, con minimización de artículos o conjunciones
	Spam-oriented blogging	Colocar comentarios o links en el sentido de promover websites para que, quién efectua el post, está ligado

Tabla: 4.1: Content Spam

[Ntoulas *et al.*, 2006], presentan un estudio que demuestra que entre un 82-86 % de las páginas, con alteraciones en el contenido, pueden ser detectados por un clasificador automático. Los recursos usados en este análisis y clasificación incluyen, entre otras: el número de palabras dentro del texto de la página, el número de hyperlinks, el número de palabras en el título de las páginas, la redundancia del contenido, etc.

Existe consenso entre la comunidad científica a la hora de considerar que hay una gran relación entre spam a nivel del link y a nivel de contenido. Su interacción lleva a que haya una connivencia recíproca. Como menciona Becchetti [Becchetti *et al.*, 2006] el análisis basado en links y el análisis basado en contenidos ofrecen dos aproximaciones ortogonales que de modo alguno se pueden considerar alternativas, todo lo contrario, deben ser usadas en conjunto.

Infelizmente, ni siempre es posible detectar spam a través de un simple análisis, automático o no, del contenido de las páginas, dado que algunas páginas apenas presentan alteraciones en los links hacia donde apuntan y no en el contenido. Aunque, desde el punto de vista del análisis de contenido, si no fueran analizados los llamados 'out-links' nada podremos concluir en lo que respecta a la posibilidad de estar ante un caso de spam.

Una interesante y diferente forma de abordar el problema es estudiada por [Gibson *et al.*, 2005], aprovechando cada uno de los operandos para un análisis individual:

- Por un lado, el análisis basado en links no consigue abarcar todas las hipótesis de spam, una vez que algunas páginas presentan propiedades, tanto en la forma como en la construcción y disposición gráficas, que están, estadísticamente, muy próximas de páginas libres o exentas de cualquier tipo de spam. En este caso, el análisis de contenidos puede volverse muy útil.

- Por otro lado, el análisis basado tan solo en los contenidos parece ser menos resistente a los cambios en las estrategias de spamming. Por ejemplo, un spammer podría copiar un site completo (creando un conjunto de páginas que pudiesen ser capaces de pasar todas las pruebas para la detección de spam en el contenido), y cambiar un poco los 'out-links' en cada página que apunte hacia la página objetivo que se pretende disimuladamente alcanzar). Esta puede ser una tarea relativamente barata para ejecutar de forma automática, del tipo 'change all', mientras que el proceso de creación, manutención, reorganización de un link-farm, posiblemente envolviendo más de un dominio, es, seguramente, más caro.

4.2.2. Link Spamming

Link Spam	Link Farm	Conjunto de páginas densamente ligadas entre sí, creadas con un único propósito: engañar a los algoritmos basados en clasificación por links
	Permuta de links	Se basa en un compromiso entre dos sites de apuntarse recíprocamente, independientemente de que los contenidos de ambos estén relacionados o no
	Compra de links	Se basa en el comercio de links, desde sites especializados en este servicio (directorios, etc)
	Dominios expirados	Herramientas específicas para acompañar la validez de los registros de DNS, en el sentido de controlar dominios que expiren y no sean revalidados por sus propietarios
	Páginas de entrada	Típicamente, son grandes conjuntos de páginas de baja calidad donde cada página es optimizada hacia una keyword o frase específica
	Throwaway sites	Son páginas que están muy pobladas con links y palabras clave para atraer y redireccionar tráfico
	Link Bombing spam	El texto áncora de un link, en cierta forma, describe la página destino, lo que lleva a que los motores de búsqueda puedan dejar los contenidos por el texto asociado al link
	Affiliate	Consiste en crear un link que remita a los visitantes hacia otro site

Tabla: 4.2: Link Spam

En muchos aspectos, y debido a la creciente influencia de las relaciones, se volvió 'inevitable' [Burdon, 2005] el apetito por el 'link spamming'. Siendo una actividad local, o sea, que se ejerce en un servidor propio al que se accede directamente, tiene como fin impulsar la clasificación de una página / site [Zhou *et al.*, 2008], haciendo la gestión de links entre grupos de páginas.

Las técnicas basadas en la estructura del link, la modifican para atacar los motores de búsqueda que utilizan algoritmos de Ranking basados en el link, como es el caso de PageRank [Brin & Page, 1998] y HITS [Kleinberg, 1999]. La técnica más conocida basada en links incluye 'link farms', 'link exchanges', 'link bombs' y el comment spam en los blogs y

wikis [Wu, 2007].

Esta técnica de spam representa un problema creciente, sobre todo después de volverse pública la importancia que los motores de búsqueda (Google en primer lugar) colocan en los links, con respecto al cálculo interno de la ordenación. La parte positiva de esto es la facilidad de implementación que de ahí surge.

De hecho cualquier persona puede crear varios sites en Internet, con diferentes nombres de dominio, en el que cada uno *linka* hacia todos los demás, o, de otra manera, pueden aprovechar [Zhou *et al.*, 2008] aplicaciones web ya existentes, tales como wikis y weblogs que exhiben hyperlinks presentados de forma anónima o bajo seudónimos.

El objetivo final, como mencionamos en la introducción de este capítulo es el de eludir los algoritmos de Ranking (también en las versiones mejoradas para Google), para forzar a que sea atribuida una clasificación superior a un site y, por consecuencia, a todos los otros que apuntan hacia él.

De entre los efectos perniciosos apuntados, hay que considerar otros efectos, internos, en los propios motores de búsqueda, dado que, además de disminuir la calidad de los resultados de búsqueda, el gran número de páginas con spam (o sea, las páginas creadas expresamente para 'spam'), también aumenta el coste de crawling, la indexación y el almacenamiento en motores de búsqueda [Gan & Suel, 2007].

En virtud de esta casi dedicación al Ranking, no es de extrañar que muchas de las técnicas propuestas [Gyongy *et al.*, 2004; Saito *et al.*, 2007; Wu & Davison, 2005b], a veces también de forma tendenciosa [Jiang *et al.*, 2008], se refieran principalmente a los algoritmos de PageRank.

Este tema ha proporcionado diversos estudios científicos, incluyendo mayor detalle sobre análisis de links y de clasificación con ayuda computacional, basada en métodos de clasificación de detección de spam.

Pero la situación está lejos de ser pacífica, no sólo en la lucha planteada por los investigadores contra los detractores, sino también internamente.

Por ejemplo, se cree que la propagación de spam, conocido por inversión de links [Sobek, 2002], pueda ser usado por algunos motores de búsqueda, mientras [Gyongy *et al.*, 2004] propone la idea de promover

buenas prácticas de confianza en buenos sites, con el fin de desvalorizar el concepto spam.

4.2.3. Camuflage o Page-hiding

El spam basado en 'Page-hiding' esconde de los motores de búsqueda la totalidad o parte de las páginas, en el sentido de obtener mejor ranking.

Tales elementos destinados a ser utilizados como indicadores de los contenidos de la página o imagen, son muchas veces explotados por páginas de spam como un objetivo invisible transformándolos en palabras claves usadas en búsquedas.

Camuflagem ou Page-hiding	Texto escondido	Llenar las páginas con palabras invisible al ojo humano
	Tiny text	Texto escrito con caracteres en formato muy pequeño, imperceptible al ojo humano, tan sólo legible por máquinas
	Links escondidos	Colocar links de forma que parezcan invisibles para el usuario, en el sentido de aumentar la popularidad
	Cloaking	El contenido de las páginas enviadas a los crawlers y a los browsers son significativamente diferentes
	Mirror sites	Hospedaje de múltiples websites todos con el mismo contenido, pero usando diferentes URL's
	Code swapping	Se trata de optimizar una página para obtener una alta clasificación de ranking y después colocar otra página en su lugar, cuando el ranking es máximo
	Redireccionamiento	Transferir al usuario hacia otra página sin intervención humana directa, usando META refresh tags, CGI scripts, Java, JavaScript, Server side redirects u otras

Tabla: 4.3: Camuflaje o Page-hiding

El redireccionamiento deliberado y el camuflaje son dos técnicas muy usadas y conocidas por todos los internautas, que forman parte del grupo de 'page-hiding', en tres áreas: (**i**) Content hiding; (**ii**) Cloaking y (**iii**) Redirection.

4.3. Sinopsis de técnicas anti-spam

Contenidos	NLP	Analizador de páginas estructurado en algunos pormenores, incluyendo frecuencia de ocurrencia de 'stopwords'
	Aproximación Estadística	Proponen que los motores de búsqueda eliminen palabras repetidas si esas palabras aparecen en el texto por encima de un determinado umbral o número de veces
	Machine Learning	Aislaron procedimentos comportamentales, como son el número de palabras en el título y la población de palabras comunes y desarrollaron un clasificador que, basados en estos comportamientos heurísticos, reconozcan páginas con spam
	SLM	Utiliza modelos *n-grama* que buscan consistencia entre las palabras de los contenidos. La falta de consistencia indica la presencia de anomalías y será un indicador de la probabilidad de presencia spam

Tabla: 4.4: Técnicas para combatir el Spam basado en contenidos

Existe una cantidad significativa de publicaciones científicas que básicamente se vuelcan sobre la detección automática y semiautomática de Web-Spam [Attenberg & Suel, 2008; Benczúr *et al.*, 2005b; Castillo *et al.*, 2007a; Davison, 2000; Drost & Scheffer, 2005; Gan & Suel, 2007; Gyongy *et al.*, 2004; Ntoulas *et al.*, 2006; Wu & Davison, 2005; Wu *et al.*, 2006].

Gran parte de esos trabajos se centran en métodos para detectar *link farms* [Becchetti *et al.*, 2006a; Gyöngyi & Garcia-Molina, 2005; Gyongyi *et al.*, 2006; Wu & Davison, 2005].

En cuanto al análisis de spam en los contenidos, se han publicado una menor cantidad de trabajos [Attenberg & Suel, 2008]. Normalmente se trata de técnicas de simples copias de contenidos entre sites, o de contenidos generados automáticamente [Benczúr *et al.*, 2007a; Drost & Scheffer, 2005; Gan & Suel, 2007; Ntoulas *et al.*, 2006].

Estructuras de Links	Textos áncora	Si consideramos los textos áncora como parte integrante del link, entonces las link farms pueden adoptar una característica particular detectable con el uso de la función matemática 'bipartite graph'
	Aproximación Estadística	Examinar la estructura de links en el sentido de detectar estructuras estadísticamente poco habituales que podrán ser indicadores de link spam
	Distribución de hosts	Distribución del número de hosts diferentes mapeando hacia la misma dirección IP
	Machine Learning	Usados sobre todo para detectar link-farms
	Damping Factores	Para detectar conspiraciones entre páginas y servidores
	Vectores (Google)	Google puede estar creando un sistema de vectores personalizados para evitar link-spam
	Host Rank	Se muestra más resistente que el PageRank a link-farms
	IMP	Extensión del HITS que mejora el problema del refuerzo mutuo entre Hub y Authority
	SALSA	Pretende ser más resistente que el HITS al efecto TKC
	Árboles DOM	Identificación del subárbol más relevante para la consulta de otras partes que reciben mayor refuerzo mutuo
	División de HITS	Randomized HITS: para los incoming links; Subspace HITS para los outgoing links
	Clusters	Encontrar clusters de páginas, interligándolas a través de los out-links y calcular el valor del parámetro 'authority'

Tabla: 4.5: Técnicas para combatir spam basado en estructuras de Links

Muchas técnicas, como las referidas por [Ntoulas *et al.*, 2006] y [Gan & Suel, 2007], consiguieron tener algún éxito al identificar un gran conjunto de términos identificadores de spam. No obstante, y como respuesta, los spammers se actualizaron e 'inventaron métodos más inteligentes' [Attenberg & Suel, 2008]. Nuevas técnicas conocidas como de 'costura y tejedura' de las frases han generado, con éxito, páginas que contienen muchas palabras clave y frases, evitando elevadas frecuencias para

palabras individuales incluso para combinaciones de palabras. Al hacer eso, esas páginas pueden muchas veces ser spam, y al mismo tiempo ser capaces de frustrar los filtros existentes.

Otros trabajos analizan un nuevo tipo de síntesis estadística basada en la proximidad de términos (idénticos y relevantes) dentro de todo el contenido de la página.

En esta tesis presentamos además otras técnicas genéricas contra spam basadas en links.

Camuflaje o 'page-hiding'	Najork	Patente que propone la comparación de la firma de la página visitante (construida a través de la anexión de una toolbar a los browsers) con la existente en los motores resultado del crawling
	Penalizaciones	Aplicar penas a los actos de cloaking debidamente comprobados
	Monetarismo	Cuando se verifica que existe cloaking asociado a cuestiones monetarias estamos (98 %) ante el spam
	STRIDER	Analisar redireccionamentos usando uma pré-lista de páginas identificadas como spam
	Resultados del query	Analizados los resultados desarrollados por los motores (con su propia detección ya aplicada) son filtrados para remover el restante spam
	Modelación de lenguaje	Modelos de máxima verosimilitud suavizante para construir modelos probabilísticos exactos

Tabla: 4.6: Técnicas para combatir 'page-hiding'

Capítulo 5

Experiencias

5.1. Introducción

En este capítulo se expone la forma en cómo se desarrollaron la investigación y las experiencias realizadas, comenzando por efectuar una pequeña referencia histórica a los proyectos de investigación sobre WEBSPAM, realizados en Europa, desde 2006.

También con mi participación en el proyecto Web Spam UK2007 (Spam labeling), se realizó un trabajo de campo, con la misma tarea de anotar un largo número de servidores, con exactos rótulos que indicasen la presencia de spam en esas máquinas. Los rótulos definidos para esa clasificación fueron: 'Spam', 'Borderline', 'Don´t Know', o 'Normal' [Castillo *et al.*, 2008].

Concretamente se pretendía analizar si los algoritmos de clasificación de spam reaccionan de forma positiva en relación a la clasificación humana.

La pregunta que se nos colocó fue: '*are there aspects of this page that are mostly to attract and/or redirect trafffic*', y fueron definidas líneas de orientación para utilizar en el momento de la clasificación, realizada dentro de la plataforma desarrollada por el equipo del proyecto.

Con el objetivo de uniformizar lo más posible la actuación de cada uno de los asesores, necesidad fundamental como analizaremos con más detalle más adelante, al equipo responsable por el proyecto, se le definió

un conjunto de reglas de clasificación de los servidores, ilustradas con casos reales.

- Incluye aspectos concebidos para atraer o redireccionar tráfico;
- Casi siempre tienen intención comercial;
- Raramente ofrecen contenidos relevantes para los usuarios.

Los aspectos más indicadores de Web Spam son:

- Incluyen muchas palabras clave y links sin relación;
- Usan muchas palabras claves y muchos signos de puntuación en el URL (Ej.: ../../..);
- Redirecciona al usuario hacia una página que no tiene relación, en cuanto al contenido, con la página inicial;
- Crea muchas copias de la página original con contenido duplicado, aunque a veces presente algo nuevo en relación a la página original.

5.2. La importancia de la unanimidad en la clasificación humana de servidores

Para complementar el trabajo mencionado en la sección anterior, elaboramos un estudio sobre el comportamiento de los asesores en dos momentos distintos: (**1**) En la fase que fue designada por INITIAL y que comprendió no sólo la primera clasificación, sino también eventuales alteraciones efectuadas dentro del plazo establecido, y (**2**) en la fase designada por REVISION en la que se presentan a cada uno de los intervinientes las clasificaciones que él mismo efectuó y todas las demás clasificaciones, de otros intervinientes, para esos mismos sites.

Se mantiene como un dato adquirido que, quien clasifica y tiene que decidir, adopta como principal criterio de decisión, en los casos frontera entre spam y el no spam, la percepción del esfuerzo realizado por los autores de las páginas web para proporcionar buenos contenidos, contra el esfuerzo realizado en la tentativa de puntuación elevada en motores de búsqueda.

Esta decisión de hacer clara la definición sobre si estamos o no ante spam, es una tarea difícil desde el punto de vista matemático y algorítmico, una vez que no está clara la manutención de la misma posición del ser humano cuando, después de revisar posibles clasificaciones, le son planteadas para comparación otras posturas, de otros clasificadores, también humanos, pero con otra perspectiva sobre el límite teórico del spam.

5.2.1. Trabajo relacionado

Este trabajo tiene otros aspectos relacionados, donde se reflejan las preocupaciones sobre '*Web Spam classification*', y principalmente los integrados en las presentaciones efectuadas para los congresos, designados por AirWeb, relacionados con las mejoras en los procedimientos de recuperación de información.

En nuestro caso particular incluimos en el estudio las dificultades provocadas por la introducción de SPAM en los resultados de la selección llevada a cabo por los motores de búsqueda [Castillo *et al.*, 2007a] y de las búsquedas desarrolladas para depurar y mejorar los algoritmos de detección de spam en los sistemas de recuperación de la información.

Tan sólo con el análisis y estudio del comportamiento humano en situación real, es posible mejorar los algoritmos de computación, principalmente desde el punto de vista del ranking de clasificación final, evitando lo más posible los trucos introducidos como inductores de mejoras de clasificación.

Para que sea posible una respuesta con calidad, que permita un acceso lo más directo posible a las páginas que realmente interesan, desde el inicio del WWW, es necesario clasificar las páginas de acuerdo con la posibilidad de respuestas conocidas a la pregunta efectuada a los motores de búsqueda y por estos a los indexadores.

Este trabajo, caracterizado en 'Social Media Research Blog'[1] no siempre es fácil de obtener, principalmente sin estudiarse el comportamiento entre los anotadores.

[1] http://socialmedia.typepad.com/blog/2008/06/the-cost-of-manual-annotation.html

5.2.2. Descripción de las bases de datos

Se utilizó la versión pública de la base de datos del WEBSPAM-UK2006 [Castillo *et al.*, 2007a] y la colección de WEBSPAM-2007 [Bíró *et al.*, 2008].

Al final de los procedimientos del WEBSPAM-UK2006, 6.552 servidores fueron clasificados [Castillo *et al.*, 2007a]; en la colección WEBSPAMUK2007, que incluía 114,529 servidores fueron clasificados con 6,479.

Analizaremos la primera parte (SET 1 - reservado para la comunidad científica), que incluye 4,275 servidores.

En un primer estudio podremos comparar los clasificadores del WEBSPAM 2006 con los de 2007, donde se destaca que el clasificador Normal fue el más utilizado en ambas, sin embargo, en 2007, el valor aumenta más de un 20 %, lo que puede ser un primer indicador de la evolución del control sobre el spam, pero, por otro lado, puede querer representar que la vulgarización de algunos actos o la mejora de algunos trucos hizo imperceptibles páginas con spam (5.19 % vs 22.08 %). Por aquí se abren, desde luego, interesantes perspectivas de análisis.

Por otro lado - y teniendo presente el objetivo final de que la clasificación debe ser lo más clara posible sobre Sí o NO - el clasificador de frontera (BORDERLINE) desciende de 10.82 % a 2.39 %, lo que es excelente desde el punto de vista del objetivo mencionado que implica alcanzar, en el límite, un valor de un 0 % para las incertezas.

	WEBSPAM-UK2006		WEBSPAM-UK2007 SET 1	
Classificación	**Frequencia**	**Perc.**	**Frequencia**	**Perc.**
Normal	4,046	61.75 %	3,776	88.33 %
Spam	1,447	22.08 %	222	5.19 %
Borderline	709	10.82 %	102	2.39 %
Could not be classified	350	5.34 %	175	4.09 %

Tabla: 5.1: Clasificación de los servidores efectuada por los asesores

5.2.3. Distribución de los clasificadores

Desde el punto de vista del interés de los ataques de los spammers un primer 'filtro', o si queremos desde el punto de vista contrario, un primer

'objetivo' está determinado por el subdominio a que el site está ligado.

Interesa por eso saber que, de entre el universo de donde fue extraída nuestra colección (.uk), el subdominio más voluminoso es el co.uk, dada su implantación y usabilidad por la comunidad comercial del Reino Unido, como en la Tabla 5.2.

Subdomain	Quant.	Normal	Spam	Borderline	Unclassified
ac.uk	171	165	–	–	6
bl.uk	2	2	–	–	–
co.uk	3,354	2,916	201	93	144
gov.uk	79	75	-	–	4
ltd.uk	6	6	–	–	–
me.uk	8	7	1	–	–
mod.uk	2	2	-	–	–
net.uk	1	1	–	–	–
nhs.uk	17	15	-	–	2
org.uk	574	527	20	9	18
plc.uk	3	3	–	–	–
sch.uk	58	57	–	–	1
Total	4,275	3,776	222	102	175

Tabla: 5.2: Distribución de los clasificadores por subdominio

Los subdominios más representados (co.uk y org.uk) presentan elevados valores para la clasificación de Normal, como en la Tabla 5.3, destacando que el 'co.uk', de características eminentemente comerciales, presentando valores más bajos para sites NORMAL (86.94 % vs 91.81 %) y más elevados para SPAM (5.99 % vs 3.48 %).

El propio indicador Borderline es más elevado para ese subdominio comercial. Hay también que hacer referencia a la paridad con la tabla 5.1, antes mencionada.

Subdomain	Quant.	Normal	Spam	Borderline	Unclassified
co.uk	3,354	86.94 %	5.99 %	2.77 %	4.29 %
org.uk	574	91.81 %	3.48 %	1.57 %	3.14 %

Tabla: 5.3: Principales subdominios de distribución, en porcentaje

5.2.4. Representatividad de la muestra usada

La tabla 5.4 presenta el universo tratado tras la segunda fase de clasificación (fase de revisión), *versus* el universo total de datos (114,524 servidores).

Subdomínio	Dataset	Percentagem	Classif	Percentagem
ac.uk	5,063	4.42 %	171	4.00 %
bl.uk	18	0.02 %	2	0.05 %
co.uk	89,953	78.54 %	144	78.46 %
gov.uk	1,708	1.49 %	79	1.85 %
ltd.uk	257	0.22 %	6	0.14 %
me.uk	216	0.19 %	8	0.19 %
mod.uk	52	0.05 %	2	0.05 %
net.uk	34	0.03 %	1	0.02 %
nhs.uk	339	0.30 %	17	0.40 %
nic.uk	1	0.00 %	0	0.00 %
nls.uk	1	0.00 %	0	0.00 %
org.uk	15,141	13.22 %	574	13.43 %
parliament.uk	6	0.01 %	0	0.00 %
plc.uk	25	0.02 %	3	0.07 %
police.uk	57	0.05 %	0	0.00 %
sch.uk	1,658	1.45 %	58	1.36 %
Total / Avg	114,528	–	4,275	–

Tabla: 5.4: Representatividad de la muestra

El modelo usado en la distribución de servidores por los participantes, para la clasificación, asegura las relaciones entre el universo individual y el universo del DataSet, por lo que podemos concluir la validad de la muestra.

5.2.5. Experimentos y Resultados

En este estudio analizamos, para una mejor comprensión, el comportamiento humano relacionado con:

- ¿Cuál es el grado de concordancia en spam / no spam?

- ¿Cuán diferentes son las diferentes personas (algunas usan mucho la clasificación spam otras evitan esa clasificación)?
- ¿En qué grado la opinión de las personas muda cuando se les presenta las opiniones de otras, sobre el mismo asunto?
- La clasificación con el rótulo 'borderline' - ¿cómo se comportan los asesores?

5.2.5.1. ¿Cuál es el grado de concordancia en spam / no spam?

En la tabla 5.5 analizamos los diversos grupos, tras la clasificación final.

	Nivel de Concordancia		
	Total	Parcial	Classif.
Nonspam	3,504	272	3,776
Spam	157	65	222
Borderline	102	0	102
Unknown	175	0	175
Total	3,938	337	4,275

Tabla: 5.5: Grado de concordancia

El análisis de esos resultados revela que un 92.80 % (3,504 de 3,776) de los sites clasificados como 'No spam' merecieron la concordancia total de los intervinientes, mientras que en tan solo un 7.20 % (372 de 3,776) se verificó la no concordancia total entre los asesores, lo que, en principio, puede ser un indicador del nivel de los criterios de los asesores bastante equilibrado.

La utilización del clasificador 'Spam' fue mucho más bajo. Un 70.72 % de los humanos concordaron plenamente, habiendo un acuerdo parcial del 29.28 %.

Pensamos que este análisis se enriquecería más si fuese posible incluir otros elementos para el análisis, como son el país de origen, el nivel de estudios, la edad y el sexo del asesor, de un modo particular en los casos en que no se verifica acuerdo total.

5.2.5.2. ¿Cuán diferentes son las diferentes personas (algunas usan mucho la clasificación spam otras evitan esa clasificación)?

Como mencionamos al final del punto anterior, estamos convencidos de que existen muchos factores de personalidad que constriñen la objetividad pedida en el momento de la atribución de una clasificación.

Como podemos analizar en la Tabla 5.6, los primeros 10 asesores, por orden del número total de servidores clasificados, de un modo predominante, clasificaron como NONSPAM. La excepción puntual para el primero en el que el uso de BORDERLINE o UNKNOWN es francamente elevado, lo que puede llevar a concluir que: (1) estos consultores efectuaron una selección inicial para validar, desde el principio, los servidores que creen que son de confianza o (2) que los primeros puedan ser miembros del equipo de concepción y que conocen bien el modelo.

Assessor	Spam	Nonspam	Borderline	Unknown	Total
1	14	292	8	8	322
2	24	257	23	16	320
3	24	239	15	21	299
4	6	215	12	50	283
5	7	249	14	11	281
6	5	243	14	18	280
7	16	236	22	6	280
8	22	226	15	15	278
9	12	235	12	15	274
10	6	221	11	34	272

Tabla: 5.6: Recuento de las Clasificaciones efectuadas por los primeros 10 asesores con el mayor número de sites clasificados

El análisis de los datos demuestra que, como se ha mencionado en la referida tabla 5.7, para la clasificación inicial, se verificó una clasificación masiva como NONSPAM (76.30 %), existiendo también un porcentaje considerable de indecisos (BORDERLINE 6.21 %). Estos últimos se aproximan bastante de los sites clasificados como SPAM (6.54 %).

Phase / Classif.	Borderline	Nonspam	Spam	Unkn.	Total
Initial	619	7,606	652	1,092	9,969
Revision	186	301	142	35	664
(A) %(Revis./Initial)	30.05 %	3.96 %	21.78 %	3.21 %	6.66 %
(B) %(Initial/Total Initial)	6.21 %	**76.30 %**	6.54 %	10.95 %	-
(C) %(Revis./Total Revis.)	28.01 %	45.33 %	21.39 %	5.27 %	-

Tabla: 5.7: Evolución de la clasificación por fases

5.2.5.3. ¿En qué grado la opinión de las personas muda cuando se les presenta las opiniones de otras, sobre el mismo asunto?

Considerando el objetivo: 'Alteración de clasificador', usamos tan solo los datos referentes a los servidores clasificados en la segunda fase (REVISION). De ahí obtuvimos la primera clasificación por asesor y todas las clasificaciones de otros asesores para los mismos servidores.

Este análisis nos permite evaluar el grado de influencia de los 'vecinos' que analizaron el mismo contenido.

De aquí que, de los 9,969 sites accedidos en la fase inicial, 664 fueron alterados (Tabla 5.7), tan sólo un 6.66 %.

De este grupo de 664 servidores, se verificaron alteraciones en un 61.60 % (100.00 % - 38.40 %), motivadas por la visualización de las clasificaciones de los otros asesores (Tabla 5.8).

Podemos entonces sacar una **primera conclusión**, que **en caso de no concordancia**, existe una alta probabilidad (61.60 %) de ser mudada la opinión inicial, lo que nos puede llevar a concluir que el **espíritu de exención absoluto que las medidas de clasificación deben tener, no fueron totalmente alcanzadas**.

De entre los que revieron la clasificación inicial, concluimos que NONSPAM es la clasificación menos alterada (18.7 %), y que existe una convergencia en el sentido de esta clasificación. Se verifica una tendencia para dar más fácilmente como probada esta situación de NONSPAM que lo contrario, o sea, una **segunda conclusión**, de que **a falta de clara evidencia de spam no se usa este clasificador**.

En la tabla 5.8 analizamos las procedencias, o sea, de que forma cada uno de los productos finales reciben información de las otras.

REVISTA (A)		INICIAL (B)		CAMBIARON (B/A)	MANTUVIERON
SPAM	142	**SPAM**	52	36.62 %	7.83 %
		NONSPAM	35	24.65 %	
		BORDERLINE	36	25.35 %	
		UNKNOWN	19	13.38 %	
NONSPAM	301	SPAM	76	25.25 %	18.70 %
		NONSPAM	120	39.87 %	
		BORDERLINE	93	30.90 %	
		UNKNOWN	12	3.99 %	
BORDERLINE	186	SPAM	45	24.19 %	8.58 %
		NONSPAM	74	39.78 %	
		BORDERLINE	57	30.65 %	
		UNKNOWN	10	5.38 %	
UNKNOWN	35	SPAM	3	8.57 %	3.92 %
		NONSPAM	3	8.57 %	
		BORDERLINE	3	8.57 %	
		UNKNOWN	26	74.29 %	
	664	-	664	-	38.40 %

Tabla: 5.8: Fase de Revisión: Cambio de Clasificador proveniencias

De aquí también se puede concluir que de los 142 sites ahora clasificados como SPAM, tan sólo 52 (7.83 %) mantuvieron la clasificación. Cambiaron de clasificador para SPAM 90 sites con otras clasificaciones anteriores (35 de NONSPAM, 36 de BORDERLINE y 19 de UNKNOWN). Es también relevante destacar que ni todas las clasificaciones mantenidas (en este primer caso las que mantuvieron el clasificador de SPAM y que son 52) son todas clasificaciones de unanimidad, dado que la clasificación final resulta de aquella que fuese más preponderante, i.e. por ejemplo si dos asesores se clasifican como SPAM y uno de UNKNOWN el site se marcará como SPAM.

De hecho, idéntica interpretación puede ser aplicada a todos los clasificadores finales aquí estudiados. Calculados esos valores, conforme a la tabla 5.9 verificamos posibles nuevas razones.

Hay que destacar que es posible sacar una **tercera conclusión**, resultante de la línea c), de que la **vecindad ayuda a tomar la decisión**.

	Inicial	Final	Evolución	%	
SPAM	176	142	-34	-19.32 %	a)
NONSPAM	232	301	69	29.74 %	b)
BORDERLINE	189	186	-3	-1.59 %	-
UNKNOWN	67	35	-32	-47.76 %	c)

Tabla: 5.9: Evolución de la clasificación por fases:
a) La disminución de la clasificación como spam puede tener que ver con lo que se conoce como 'votar vencido', en este caso en NONSPAM;
b) El aumento podrá tener que ver con las razones anteriores;
c) La vecindad ayuda a tomar la decisión.

5.2.5.4. La clasificación con el rótulo 'borderline' - ¿cómo se comportan los asesores?

La clasificación del tipo 'borderline' es, sin margen de duda, una clasificación alternativa, que se pretende que sea de utilización muy reducida, mismo residual, de forma que la clasificación sea limitada a SPAM y/o NONSPAM. De ahí que, las conversaciones mantenidas, principalmente en la primera fase, en salas de chat destinadas a esclarecer al grupo de voluntarios, eran en el sentido de ser creadas alertas esclarecedoras, en los casos de difícil clasificación tan solo con SPAM o NONSPAM, que ayudasen en la mejora de las reglas heurísticas a incluir en los procedimientos matemáticos que iban a ser desarrollados.

Por eso verificamos una especial preocupación humana en la utilización del clasificador 'borderline'. En los análisis referidos en las tablas 5.10 y 5.11, la convergencia es absoluta para las clasificaciones de SPAM o NONSPAM. Tan sólo en un 7.57 % de los casos fue utilizado 'Borderline'

	INICIAL	REVISTA	TOTAL
Spam	652	142	794
Nonspam	7,606	301	7,907
Borderline	619	186	805
Unknown	1,092	35	1,127
TOTAL	9,969	664	10,633

Tabla: 5.10: Evolución de las clasificaciones por tipos

	INICIAL	REVISTA	TOTAL
Spam	6.54 %	21.39 %	7.47 %
Nonspam	76.30 %	45.33 %	74.36 %
Borderline	6.21 %	28.01 %	7.57 %
Unknown	10.95 %	5.27 %	10.60 %

Tabla: 5.11: Medias de Clasificación por tipos

En la fase inicial (primera clasificación de los consultores, sin cualquier forma de revisión y participando un mayor número de individuos) el nivel de clasificación como 'Borderline' fue bastante bajo: solamente un 5.47 %. Ante estos análisis, podemos concluir - como mencionamos atrás - que la opción de **'Borderline' es de uso bastante raro**.

Y finalmente podremos evaluar cuantos fueron influenciados con el cambio de clasificador. Como se refleja en la Tabla 5.12 la mayoría (61.59 %) alteraron la clasificación.

CAMBIO	35+36+19+76+93+12+ ...	409	61.59 %
MANTUVIERON	52+120+57+26	255	38.41 %

Tabla: 5.12: Análisis sintético del grado de cambio

5.2.6. Conclusiones

Resalta, como primera evidencia que es muy difícil, desde el punto de vista estrictamente humano, encontrar una definición sobre lo que es SPAM y lo que no es.

La subjetividad de la clasificación reside fundamentalmente en la interpretación humana del concepto, lo que provoca un área gris alrededor de la deseable clara definición de SPAM o NO SPAM. Esta frontera, cuyo límite se pretende que sea cero, está influenciada por la clasificación de los vecinos, como queda indicado por las alteraciones hechas durante la revisión.

Podemos también concluir que la clasificación inicial de NO SPAM y SPAM son menos volátiles y más duraderas. De estos dos, la clasificación más permanente es la de NO SPAM.

Como referencia final, pensamos que también sería importante evaluar el tiempo que cada asesor tarda en efectuar cada tarea de clasificación, así como la profundidad de la decisión, evaluable por el número de Inbounds y Outbounds analizados inmediatamente antes de la toma de decisión.

Respondiendo a la principal preocupación que demostramos al principio del trabajo, podemos concluir que la clasificación como 'Borderline' es reducida y de utilización muy cuidada por el ser humano y de que su aumento, que podremos connotar como de 'indecisos', por un lado, y el valor de un 61.59 % (Tabla 5.12) para los que alteraron su clasificación inicial, son indicadores de que **la vecindad, más allá de influenciar, puede generar alguna confusión**.

Capítulo 6

Conclusiones y trabajo futuro

¿Qué es el Spam? ¿De qué modo dificulta los sistemas de recuperación en la Web?

Estas fueron las principales preocupaciones, desde diversas perspectivas, que nos acompañaron durante todo nuestro trabajo de investigación sobre Web Spam.

Primero identificamos conceptos: motores de búsqueda, algoritmos de clasificación y su evolución temporal. Después, aprovechando lo que mejor conocemos de la comunidad científica, presentamos una sinopsis de técnicas anti-spam. Como trabajo propio presentamos un estudio que añade un factor sociológico a todas las cuestiones estudiadas.

Todo se debe al gran crecimiento de la web.

De hecho, desde su tímido inicio - con la función de compartir datos entre Físicos - la web creció, y continúa creciendo, siendo hoy también un polo central de cultura, de educación y, por encima de todo, de vida comercial. Millones de usuarios ejecutan diariamente transacciones financieras en sus páginas de la web, que pueden ir desde compra de utensilios, de libros, de viajes y hoteles, hasta la gestión de carteras de aplicaciones financieras. Es conocida por todos nosotros, sobre todo desde el punto de vista de las accesibilidades, la forma en como la web modificó el paradigma de la información.

Debido a la increíble cantidad de información disponible en la web, los usuarios - todos nosotros - rápidamente nos habituamos a consultar /

buscar contenidos usando los motores de búsqueda. Para cada consulta, un motor de búsqueda identifica las respectivas páginas en la web y les presenta a los usuarios los links para esas páginas, generalmente en lotes de 10/20 respuestas. De cara a las respuestas, ordenadas por grado de comparación con la pregunta, cada usuario puede optar por cada una de las respuestas proporcionadas.

Este 'click' que determina la primera opción, se efectúa, de una forma general, en el primer grupo de páginas que se le proporciona al utilizador - un 85 % de las veces, las personas apenas consultan los primeros 10 resultados de la respuesta [Metaxas & DeStefano, 2005; Silverstein *et al.*, 1999] -, de ahí la alta ventaja (según el concepto comercial) de aparecer colocado en ese 'top ranking'.

La tentación de colocar, de una forma más o menos clara, las páginas en ese grupo primero de selección, existe prácticamente desde que existe Internet [Ntoulas *et al.*, 2006], por lo menos con divulgación visible hacia fuera de las paredes de las Universidades y de los grandes centros de cálculo y computación.

Esta tentación, a la que llamamos de popularidad del site, asociado al ranking, se fue materializando creando otro gigante.

Fruto de la ganancia por dinero, el mismo creador de la 'Bella' creó el 'Monstruo'.

Este 'Monstruo', al que le hemos llamado Spam y en nuestro caso específico de Web Spam, se ha vuelto más fuerte en los últimos años, en la medida en que más operadores osaron utilizarlo como un medio para aumentar el tráfico, con la esperanza de que puede también venir a aumentar los ingresos. Fruto también de una atención - creciente - de la comunidad académica, el crecimiento se intenta frenar.

> La propia definición de 'Web Spam', que ha sufrido con el tiempo matices sintácticos, apunta siempre para lo que [Ntoulas *et al.*, 2006] se define como: '*la inyección de páginas creadas artificialmente para influenciar los resultados de los motores de búsqueda, para orientar el tráfico de ciertas páginas para la obtención de lucros, o simplemente por placer.*'

Evidenciamos que '*Cualquiera que sea la definición es cierto que Spam se refiere a algo indeseable, perturbador, que influencia negativa-*

mente el proceso de selección de información tratada en ambiente web, con utilización de los protocolos ahí implementados'.

Efectuado un listado de los diversos modos base de ataque del Web Spam, estudiamos algunas de sus variantes, también algunas soluciones encontradas para rebasar esos actos maliciosos y reponer los conceptos de credibilidad y de fiabilidad que, al margen, de estos inconvenientes, representan los sistemas de recuperación de información. En el caso de la Web, los crawlers, los spiders, los motores de búsqueda, etc, son todos elementos estructurales de la idea de sistemas de recuperación de información que son objetivos preferenciales.

La evolución para el mayor aprovechamiento de las automatizaciones de los sistemas, como son las técnicas de 'Machine Learning', o la adopción de algoritmos matemáticos más precisos y complejos (ej. graph isomorphism) [Bharat *et al.*, 2001; Metaxas & DeStefano, 2005], están cada vez más insertadas en los procesos de detección de spam [Hidalgo, 2002; Sahami *et al.*, 1998]. Estas técnicas, sin embargo, se enfrentan con fuertes restricciones relacionadas con la diferente disponibilidad de contenidos para los crawlers y para los browsers.

De cara a los numerosos estudios que pretenden detectar y minimizar el spam, verificamos que existe convergencia en el sentido de agruparse en tres grandes especies: 'Link Spam', 'Content Spam' y 'Cloaking': Las dos primeras ligadas a técnicas de boosting del PR y la última a técnicas de camuflaje. Está claro que esta es una respuesta a la importancia que se le da a esos elementos por los principales algoritmos de Ranking [Brin & Page, 1998; Kleinberg, 1999; Metaxas & DeStefano, 2005].

Verificamos que, por el número significativo de páginas infectadas - por lo menos un 8 % de todas las páginas indexadas son spam [Fetterly *et al.*, 2004; Metaxas & DeStefano, 2005] - , el web spam es uno de los mayores desafíos proporcionados a quien se dedica a la recuperación de información en ambiente web [Henzinger *et al.*, 2002; Metaxas & DeStefano, 2005]. Los estudios para limitar esas restricciones encuentran dificultades en la identificación automática de spam basadas apenas en algoritmos matemáticos [Bharat et al., 2001; Metaxas & DeStefano, 2005]. En efecto necesitamos comprender socialmente la cuestión del web spam y sólo después analizar las cuestiones técnicas, dado que es en el área de los comportamientos sociales donde el spam está siempre más actualizado.

También aquí se aplica, aunque de forma adaptada, la máxima de que personas buenas se relacionan con personas buenas, en este caso [Benczúr *et al.*, 2007b] el spam normalmente apunta para spam y páginas buenas apuntan hacia páginas buenas.

El futuro prevé un crecimiento de complejidad de los ataques, por lo que cualquier especulación predictiva de lo que podrá ocurrir es utópica. Basta que nos inclinemos sobre la velocidad con la que diariamente son detectados nuevos 'malwares', comparado con lo que ocurría hace cinco años, para darnos cuenta de la seriedad de la situación.

No obstante de algunas cosas tenemos la certeza:

- El número y la variedad de los ataques continúa creciendo, de forma cada vez más estructurada en cadenas de crimen organizado, que pretende usurpar información y recursos;
- La fuga de información será cada vez más preocupante, especialmente con la utilización creciente de las tecnologías móviles. Muchos países ya introdujeron leyes estrictas sobre divulgación de información. Estas leyes pretenden penar a todas las empresas que buscan la manera de saltar las barreras de seguridad, una vez que, una violación muy extensa de datos, se haya divulgado, puede afectar a la confianza global en una organización de productos y servicios.

También los Weblogs, o simplemente blogs, han crecido últimamente como una nueva e importante forma de publicar informaciones, participar en discusiones y formar comunidades. La creciente popularidad de los blogs ha dado origen a motores de búsqueda y análisis centrada en la 'blogosfera'.

Un requisito fundamental de esos sistemas es el de identificar los blogs mientras rastrean la Web. Pero eso garantiza que solamente los blogs serán indexados, los motores de búsqueda son también muchas veces sobrecargados por 'splogs' (spam blogs), que acaban por influenciar negativamente las indexaciones, aunque de una generación más reciente, esta es una forma de spamdexing, que podemos fácilmente incluir en el grupo de 'content spam'.

La inseguridad de la web, fruto de su forma concepcional, enflaquecida contra ataques remotos automatizados, fruto también de los crecientes

modelos basados en P2P, continuará siendo la principal forma de distribución de malware específico.

Es también en este sentido de especialización, en esta nueva área, que una parte de la comunidad científica está trabajando. [Kolari *et al.*, 2006b], por ejemplo, usa un modelo identificado como SVM (Support Vector Machines) para identificación de blogs en el que abre puertas para el desarrollo de esta tecnología.

Cada uno de nosotros, normales usuarios de ordenadores, seremos continuamente desafiados en cuestiones de seguridad y control de nuestros propios equipos, para poder defendernos de los ataques, muchas veces por el simple gusto de 'penetrar' sistemas.

No obstante, cuando se tratan convenientemente, los problemas son siempre superables: mejorando nuestras defensas básicas, protecciones actualizadas y un empeño personal por mantenernos informados, pueden proporcionar estabilidad en nuestros sistemas particulares y colectivos. Como complemento, y como buenas noticias, el software de seguridad están mejorando día a día, produciendo alertas para la defensa de posibles nuevas formas de ataque.

Porque la guerra contra los motores de búsqueda continua evolucionando con rapidez, nuevas técnicas spam aparecerán, lo que implica que nuevos abordajes anti-spam también se desarrollarán.

¡La lucha va a continuar!

6.1. Futuro spam

La mayoría de los trabajos publicados discuten métodos para combatir spam de los que ya se conocen atributos y formas de actuación. De hecho raros son los trabajos que abordan la cuestión de futuros ataques de spam. Pero, y porque la recuperación de información (IR) en la web es un campo con cada vez más servicio de búsqueda, es útil prever posibles técnicas de spam que puedan aparecer, en el sentido de poder concretizar alguna anticipación.

Decididamente los algoritmos de ranking serán siempre el primer objetivo. De ahí que los spammers continuarán atacando a los factores a los que esos algoritmos les den relevancia. La puntuación de los mode-

los TF-IDF y de análisis de links deberá continuar siendo utilizada por los motores de búsqueda para elaborar los rankings. Por eso los spammers continuarán invirtiendo en técnicas de manipulación de esos valores, aumentando la complejidad de los métodos. Por ejemplo, en el mismo sentido de obtener mejores valores en el análisis de links, estos pueden desarrollar estructuras de links más complejas y más difíciles de detectar, o, por otro lado, pueden encontrar mejores técnicas de camuflaje, tales como la profundización de Javascripts, para manipular esos valores.

Además de eso, si los motores de búsqueda anuncian públicamente los nuevos componentes que incluirán en los nuevos algoritmos de ranking, serán, seguramente, objetivos de los spammers. Como resultado, nuevas técnicas de spam irán a aparecer.

La construcción de sistemas de búsqueda patrocinadas, que incluye publicidad en la página de respuesta de los motores de búsqueda, han permitido un gran éxito financiero.

Como esos sistemas están agilizados para la consecución de dinero, será probable que ésta sea un área de gran inversión de los spammers. De hecho algunas técnicas de spam dirigidas a esos sistemas fueron ya descubiertas.

Por ejemplo, fraudes provenientes de clicks, que ocurren en los sistemas de 'pay per click' cuando una persona, scripts automatizados, o incluso programas de ordenador, usan los browsers para pinchar en zonas de publicidad que generan dinero, sin tener interés en los productos hacia donde el link apunta.

De hecho y debido al componente financiero que representan, esas búsquedas patrocinadas pueden ser nuevos focos de spam.

Los motores de búsqueda proporcionan cada vez más servicios relacionados con la recuperación de información. Cualesquiera de los nuevos servicios que en un futuro vayan a estar disponibles serán, seguramente, inmediatamente estudiados por los spammers en el sentido de encontrar alguna forma de beneficiarse (de forma ilícita) de esos nuevos servicios.

Son ya referencias de esta certeza, por ejemplo, el spam sobre video o en el Twiter.

No importa la complejidad de las técnicas de spam una vez que todas ellas son proyectadas para manipular los factores determinantes en los

algoritmos de clasificación. Por ejemplo, como los spammers continúan creyendo que los algoritmos basados en estructuras de links continuarán siendo usados, continuarán invirtiendo tiempo y dinero intentando manipular esas estructuras.

Aunque el futuro spam pueda asumir diferentes formas y métodos, podemos combatirlos incidiendo sobre las características que han permanecido relativamente inalteradas o utilizando algunas metodologías anti-spam ya testadas y maduradas.

Con el fin de obtener una mejor clasificación para sus páginas, los spammers tienen la necesidad de reforzar algunos puntos, lo que continuará haciendo con que las páginas manipuladas presenten algunas características especiales, comparadas con páginas normales [Benczúr *et al.*, 2005b; Fetterly *et al.*, 2004]. Por eso, la detección de esas características especiales es un buen indicador sobre la presencia de Spam. Por ejemplo, sites con muchos links incoming y outgoing iguales pueden indicar la presencia de un link farm. Del mismo modo, la inserción de gran cantidad de palabras (eventualmente palabras-clave) [Davison, 2000; Drost & Scheffer, 2005], incluso sueltas, o el empleo de las mismas páginas por links inespecíficos son probablemente fruto de un relleno poco serio. En este sentido, tenemos que buscar más recursos y páginas con características especiales. Por ejemplo, si la mayoría de los outgoing links de un site apuntan hacia otros sites que proporcionan técnicas de 'affiliate' entonces este site usa técnicas 'affiliate'.

Asociado a este planteamiento estadístico de recuento y análisis de links, los avances en las técnicas de 'machine learning' pueden innovar procedimientos principalmente con la posibilidad de plantear de forma diferente de un mismo algoritmo de cara a algunas clases de características que sean detectadas. Por ejemplo, puede construirse un clasificador [Ntoulas *et al.*, 2006] para saber si un click es un click fraudulento, o podemos utilizar un clasificador para saber si un link es un link spam o no, entre otras inmensas posibilidades.

Quedó claro que, cada vez más, hay mezclas de técnicas basadas en links, con otras técnicas basadas en los contenidos [Gibson *et al.*, 2005].

El 'brazo de hierro' entre spammers y motores de búsqueda está ya lejos. Los contra-ataques de los spammers, después de que los motores de búsqueda han neutralizado ataques previos, mantienen actual esta designación. «La carta o as» que entre tanto los investigadores comienzan

a usar en las nuevas técnicas de spam, relacionadas con los ataques preventivos, puede desmotivar a los 'hackers' durante, por lo menos, algún tiempo.

Cuanto mayor fuera la disponibilidad de colaboración, tanto del mundo industrial como del mundo científico, entonces mejores y más sofisticados algoritmos anti-spam podrán ser producidos, por un lado, y por otro, cuanto más evolucionemos socialmente, en el sentido de minimizar las limitaciones encontradas en la necesaria unanimidad en la clasificación humana de los servidores, menos credibilidad daremos a impostores.

Serán factores determinantes para la disminución de la inversión en spam cuando:

- Fuera más caro producir páginas de spam que páginas oficiales con calidad, o
- Fuera más caro producir páginas de spam que los beneficios que se obtendrían.

Por la misma razón de, como dice el pueblo, lo óptimo es enemigo de lo bueno, también la victoria no requiere obligatoriamente la perfección. Por eso tenemos la esperanza de que las búsquedas continuas puedan volver la 'piratería más cara que el original', o sea, de que lo genuino compensa.

!Por esto, saldremos vencedores!

Bibliografia Resumida

Abernethy, Jacob, Chapelle, Olivier, & Castillo, Carlos. 2008a. Web spam Identification Through Content and Hyperlinks. *AIRWeb '08, Beijing, China*, May, 22.

Abernethy, Jacob, Chapelle, Olivier, & Castillo, Carlos. 2008b. *WITCH: A New Approach to Web Spam Detection.* Tech. rept. Yahoo! Research.

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Leonardi, Stefano, & Baeza-Yates, Ricardo A. 2006. Link-based characterization and detection of web spam. *In: In AIRWeb.*

Becchetti, Luca, Castillo, Carlos, Donato, Debora, Baeza-Yates, Ricardo, & Stefano, Leonadi. to appear (In Press). Link Analysis for Web Spam Detection. *ACM Transactions on the Web Journal(TWJ)*, March. Work in Progress.

Benczúr, András A., Csalogány, Károly, Sarlós, Tamás, & Uher, Máté. 2005. Spamrank - fully automatic link spam detection. *In: In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb.*

Bharat, Kirshna, & Mihaila, George A. 2000. Hilltop: A search engine based on expert documents. *In Poster proceedings of WWW*, 72–73.

Castillo, Carlos, Donato, Debora, Gionis, Aristides, Murdock, Vanessa, & Silvestri, Fabrizio. 2007. Know your Neighbors: Web Spam Detection using the Web Topology. *In: Proceedings of SIGIR.* Amsterdam, Netherlands: ACM.

Chellapilla, Kumar, & Maykov, Alexey. 2007. A taxonomy of JavaScript redirection spam. *Pages 81–88 of: AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web.* New York, NY, USA: ACM.

Fetterly, Dennis, Manasse, Mark, & Najork, Marc. 2004. Spam, Damn Spam, and Statistics. *Seventh International Workshop on the Web and Databases (WebDB 2004)*, June 17-18. Paris, France.

Gyongy, Zoltán, Garcia-Molina, Hector, & Pedersen, Jan. 2004. Combating Web Spam with TrustRank. *Proceedings of the 30th VLDB Conference, Toronto, Canada.* Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

Gyongyi, Z., & Garcia-Molina, H. 2005. Web spam taxonomy. *First International Workshop on Adversarial Information Retrieval on the Web.*

Henzinger, M. R. 2001. Hyperlink analysis for the web. *IEEE Internet Computing*, **5(1)**, 45 – 50.

Krishnan, Vijay, & Raj, Rashmi. 2006. Web Spam Detection with Anti-Trust Rank. *Pages 37–40 of: AIRWeb.*

Lempel, R., & Moran, S. 2001. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, **19**(2), 131–160.

Li, Longzhuang, Shang, Yi, & Zhang, Wei. 2002. Improvement of HITS-based algorithms on web documents. *Pages 527–535 of: WWW '02: Proceedings of the 11th international conference on World Wide Web.* New York, NY, USA: ACM Press.

Liu, Yiqun, Cen, Rongwei, Zhang, Min, Ma, Shaoping, & Ru, Liyin. 2008. Identifying Web Spam with User Behavior Analysis. *AIRWeb '08, Beijing, China*, April, 22.

Metaxas, Panagiotis T., & DeStefano, Joseph. 2005. Web Spam, Propaganda and Trust. *AIRWeb2005, May 10, 2005*, 5.

Ntoulas, Alexandros, Najork, Marc, Manasse, Mark, & Fetterly, Dennis. 2006. Detecting Spam Web Pages through Content Analysis. *International World Wide Web Conference Committee (IW3C2)*, May 23-26.

Sánchez, Montserrat Mateos. 2006 (07). *Aplicación de Técnicas de Clustering en la Recuperación de Información Web*. Ph.D. thesis, Universidad de Salamanca - Departamento de Informática y Automática. Director Dr. D. Carlos García-Figuerola Paniagua.

Weideman, Melius. 2007. Use of Ethical SEO Methodologies to Achieve Top Rankings in Top Search Engines. Cape Peninsula University of Technology, Cape Town, South Africa.

Wu, Baoning, & Davison, Brian D. 2005a (May, 10). *Cloaking and Redirection: A Preliminary Study*. First International Workshop on Adversarial Information Retrieval (AIRWeb).

Wu, Baoning, & Davison, Brian D. 2005b. Identifying link farm spam pages. *Pages 820–829 of: WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*. New York, NY, USA: ACM.

Zhou, Bin, Pei, Jian, & Tang, ZhaoHui. 2008. A Spamicity Approach to Web Spam Detection. *Pages 277–288 of: SDM*. SIAM.