Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 39-46 recibido 09-01-09, aceptado 02-03-09

Detección de Web Spam basada en la Recuperación Automática de Enlaces *

Detecting Web Spam using a Recovering Web Links System

Lourdes Araujo
NLP Group at UNED
28040 Madrid, Spain
lurdes@lsi.uned.es

Juan Martinez-Romo
NLP Group at UNED
28040 Madrid, Spain
juaner@lsi.uned.es

Resumen: Actualmente el *Web Spam* es una guerra abierta entre los motores de búsqueda, tratando de garantizar unos resultados relevantes al usuario, y una comunidad, cuyo interés reside en intentar engañar a los primeros en busca de un mejor ranking para sus páginas. En este trabajo presentamos un estudio preliminar sobre distintas medidas que podrían ser útiles para la construcción de un sistema novedoso en la detección de *Web Spam*. Algunas de estas medidas se basan en los resultados de un sistema de recuperación automática de enlaces web rotos. El sistema utiliza distintas fuentes de información de la página analizada y la información extraída de estas fuentes se utiliza para realizar una consulta a un motor de búsqueda usual, como *Google* o *Yahoo!*. Las páginas recuperadas son ordenadas posteriormente en base a su contenido, utilizando técnicas de recuperación de información. Finalmente, el análisis del grado de recuperación de los enlaces es empleado, junto a otras medidas, como un indicador de *Spam*.
Palabras clave: recuperación de información, *World Wide Web*, enlaces rotos, web spam

Abstract: Nowadays, *Web Spam* is a war between search engines, trying to ensure that the results are relevant to the user, and a community that tries to mislead the search engine to attract to the former ones to its pages.
In this work, we present a preliminary study about several features that can be useful for building a novel web spam detection system. Some of these features are obtained from a system for automatic recovery of broken Web links. This system uses several sources of information from the analyzed page to extract useful data that are used later to perform a query to a typical search engine, as Google or Yahoo!. Afterwards, retrieved pages are ordered based on its content, using information retrieval techniques. Finally, the recovery links degree is used, along with other features, as an indicator of Spam.
Keywords: information retrieval, *World Wide Web*, broken links, web spam

1. Introducción

Hoy en día, la creciente popularidad de Internet entre los usuarios como fuente de información, ha convertido a los buscadores en un objetivo de la publicidad. Los buscadores a su vez, basan su modelo de negocio en la publicidad que añaden a los resultados de una consulta. Pero además de esta publicidad relevante a las consultas realizadas, una manera muy económica de conseguir publicidad, consiste en aparecer en los primeros puestos de las respuestas del buscador. En este sentido, estar entre los 30 primeros resultados es muy importante ya que hay estudios(Jansen y Spink, 2003) que reflejan que la probabilidad de que un usuario llegue a mirar más allá de la tercera página de resultados es muy baja. Ante esta manera de aumentar los ingresos por publicidad ha surgido un fenómeno denominado *Web Spam* o *Spamdexing*.

Según (Gyöngyi y Garcia-Molina, 2005) Web Spam podría definirse como cualquier acción destinada a mejorar el ranking en un buscador por encima de lo que se merece. En general en la literatura (Gyöngyi y Garcia-Molina, 2005; Baeza-Yates, Boldi, y Hidalgo, 2007) se distinguen tres tipos de Web Spam: *Link Spam*, *Content Spam* y *Cloacking*.

El *Link Spam* o *Spam de Enlaces* consiste

* Trabajo financiado por el proyecto TIN2007-67581-C02-01

en añadir enlaces superfluos y/o engañosos a una página Web o bien crear páginas superfluas que sólamente contienen enlaces. Uno de los primeros trabajos que trataron este tipo de Spam fue (Davison, 2000), donde se consideraba el nepotismo en los enlaces como una forma de ser más relevante ante los buscadores. La manera más frecuente de encontrar este tipo de Spam es en forma de granjas de enlaces (*Link Farms*) donde un conjunto de páginas son enlazadas entre sí empleando alguna de las topologías estudiadas en (Baeza-Yates, Castillo, y López, 2005), con el objetivo de incrementar la importancia de una de ellas. Estas topologías han sido estudiadas en (Gyöngyi y Garcia-Molina, 2005).

El *Content Spam* o *Spam de Contenido* es la práctica de realizar ingeniería sobre el contenido de una página con el objetivo de resultar relevante para un conjunto de consultas. En (Fetterly, Manasse, y Najork, 2004) se presenta un análisis estadístico sobre diferentes propiedades del contenido para detectar Spam. Entre las técnicas más habituales se encuentran el incluir términos engañosos en las Urls, en el cuerpo (*body*) y en el texto del ancla y cada vez menos habitual como una *Meta Tag*. En (Ntoulas et al., 2006) se realiza una serie de medidas sobre el contenido y luego se construye un árbol de decisión mediante el cual se realiza una clasificación de este tipo de Spam. También existen trabajos (Abernethy, Chapelle, y Castillo, 2008) que combinan información tanto de los enlaces como del contenido para construir un clasificador con SVM y detectar eficientemente distintos tipos de Spam.

Finalmente, el *Cloaking* o *Encubrimiento* consiste en diferenciar a un usuario de un robot de búsqueda para responder con una página distinta en cada caso. En (Gyöngyi y Garcia-Molina, 2005) se presentan las técnicas más utilizadas en este tipo de Spam.

En la literatura existen múltiples trabajos que exploran por separado o de manera conjunta estos tipos de Spam. Sin embargo, estos estudios trabajan habitualmente con una colección etiquetada en la que previamente se ha realizado un crawling y se han precalculado una serie de medidas relevantes.

En este trabajo analizamos la utilidad de los distintos datos extraidos sobre los enlaces de una página para la detección de Spam. En particular, estudiamos la forma de utilizar los resultados extraídos de la aplicación de un mecanismo de recuperación automática de enlaces para la detección de páginas de Spam. Esta técnica, además de aplicar un nuevo indicador de Spam, proporciona un sistema de análisis *online* frente a las tradicionales colecciones.

Nuestro sistema de recuperación de enlaces rotos se basa en técnicas clásicas de recuperación de información para extraer información relevante y realizar consultas a un motor de búsqueda como *Google* o *Yahoo!*. El sistema comprueba los enlaces de la página que se le indica. Si alguno de ellos está roto, hace una propuesta al usuario de una serie de páginas candidatas para sustituir el enlace roto. Las páginas candidatas se obtienen mediante búsquedas en *Internet* compuestas de términos extraídos de distintas fuentes. A las páginas recuperadas con la búsqueda *Web* se les aplica un proceso de ordenación que refina los resultados antes de hacer la recomendación al usuario. La figura 1 presenta un esquema del sistema propuesto.

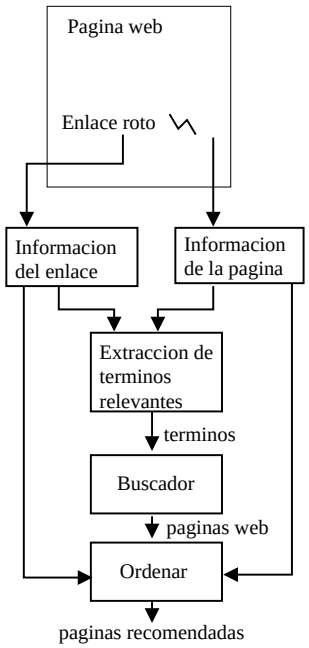

Figura 1: Esquema del funcionamiento del sistema de recomendación para la recuperación de enlaces rotos.

Al analizar los resultados de algunos experimentos, encontramos casos excepcionales en los que el grado de recuperación de enlaces tenía una gran desviación con respecto a la media. Estos casos consistían en páginas con muchos enlaces en los que no se recuperaba ningún enlace o bien se recuperaban las páginas originales de prácticamente todos los

enlaces. En ambos casos se ha comprobado manualmente que se trataba de páginas de Spam. Esto sugiere la utilidad de aplicar estas técnicas a la detección de Spam. El resultado de la recuperación de los enlaces rotos puede tomarse como un indicador de la coherencia entre un enlace y la página a la que enlaza, dato que es útil para la detección de Spam.

Existen algunos trabajos enfocados a la recuperación de enlaces, aunque se basan en información anotada por anticipado en el enlace. El sistema *Webvise* (Grønbæk, Sloth, y Ørbæk, 1999), permite cierto grado de recuperación de enlaces *Web* rotos utilizando información redundante sobre los enlaces, almacenada en bases de datos de servidores de *Internet*. Davis (Davis, 2000) analiza las causas del problema de los enlaces rotos y propone soluciones enfocadas a la recopilación de información sobre la estructura de la red de enlaces. Nakamizo y colaboradores (Nakamizo et al., 2005) han desarrollado un sistema de recuperación de enlaces basado en lo que denominan "enlaces con autoridad" de una página. Shimada y Futakata (Shimada y Futakata, 1998) propusieron la creación de una base de datos de enlaces, *SEDB*, en la que son posibles ciertas operaciones de reparación de los enlaces almacenados.

Nuestro trabajo difiere de los anteriores ya que no presupone la existencia de ninguna información almacenada de antemano sobre los enlaces y es aplicable a cualquier página de *Internet*, lo que le hace útil para analizar el Spam de las páginas web.

El resto del artículo se organiza de la siguiente forma: en la sección 2 se describen las técnicas que utilizamos para la recuperación automática de enlaces web rotos. La sección 3 analiza la relación de distintos datos sobre los enlaces de una página con su identificación como Spam, en particular los resultados de la aplicación de las técnicas de recuperación automática. Finalmente, en la sección 4 se realiza una discusión sobre los resultados y se extraen una serie de conclusiones.

2. Técnicas de recuperación de enlaces

En esta sección analizamos cada una de las fuentes de información consideradas, extrayendo estadísticas de su utilidad para la recuperación de enlaces cuando se aplican por separado o combinadas.

2.1. Uso del Texto del ancla de los enlaces

En muchos casos las palabras que componen el texto del ancla de un enlace son la principal fuente de información para identificar la página apuntada. Para verificar esta teoría, hemos realizado un estudio del número de casos en los que los enlaces rotos se han recuperado buscando en *Google* el texto del ancla entrecomillado.

Para considerar que un enlace se ha recuperado, aplicamos el modelo de espacio vectorial (Manning, Raghavan, y Schütze, 2008), representando cada una de las páginas (original y candidata) a comparar por un vector de términos, y hayamos la distancia dada por el coseno entre ellos. Si este valor es mayor de 0.9, consideramos la página recuperada. Para valores menores que este umbral, como un 0.8, aunque en la mayoría de los casos se trata de la misma página con pequeños cambios como los mencionados, hemos encontrado algún caso en que se trataba de páginas distintas, aunque del mismo sitio *Web*.

De esta forma se ha conseguido recuperar un 41 % de los enlaces entre las diez primeras posiciones (*Google*). Además un 66 % de los enlaces recuperados han logrado encontrarse en la primera posición. Estos datos demuestran que el texto del ancla de un enlace es una gran fuente de información de cara a recuperar un enlace roto.

En este trabajo hemos optado por realizar un reconocimiento de entidades nombradas (nombres de personas, organizaciones o lugares) sobre el texto del ancla, para poder extraer determinados términos cuya importancia sea mayor que la del resto. Para tal fin, existen varias soluciones software como *LingPipe*, *Gate*, *FreeLing*, etc. También existen múltiples recursos en forma de *gazetteers*, pero el amplio dominio sobre el que trabajamos ha impedido conseguir resultados precisos. Estamos en un entorno en el que analizamos páginas aleatorias cuyo único factor común es el idioma (inglés). Además, el hecho de que el texto de las anclas sean conjuntos muy reducidos de palabras y/o números, hace que los sistemas usuales de reconocimiento de entidades proporcionen resultados muy pobres. Por estos motivos, hemos decidido emplear la estrategia opuesta. En lugar de encontrar entidades nombradas, hemos optado por recopilar un conjunto de diccionarios y descartar las palabras comunes y números,

suponiendo que el resto de palabras son entidades nombradas. Aunque hemos encontrado algunos *falsos negativos*, como por ejemplo la compañía *"Apple"*, en el caso de las anclas hemos obtenido mejores resultados con esta técnica.

La tabla 1 muestra los resultados de la recuperación de enlaces en función del contenido de entidades nombradas de las anclas y del número de términos de las mismas. Los resultados demuestran que la presencia de entidades nombradas en el ancla favorece la recuperación del enlace. Además cuando hay entidades nombradas el número de casos recuperados es importante.

	Tipo de ancla			
	Ent. Nomb.		No Ent. Nomb.	
Terms.	E.N.R.	E.R.	E.N.R.	E.R.
1	102	67	145	7
2	52	75	91	49
3	29	29	27	45
4+	57	61	33	47
total	240	232	296	148

Cuadro 1: Análisis de los enlaces no recuperados (E.N.R.) y recuperados (E.R.) en función del tipo de ancla — con (Ent. Nomb.) y sin (No Ent.) entidades nombradas — y del número de términos del ancla. *4+ term.* se refiere a anclas con cuatro o más términos.

2.2. El texto de la página

Los términos más frecuentes encontrados en una página *Web* son una forma de caracterizar el tema principal de dicha página. Esta técnica requiere que el contenido de la página sea suficientemente grande. Un ejemplo claro de utilidad de esta información son los enlaces a páginas personales. Es muy frecuente que el ancla de un enlace a una página personal esté formada por el nombre de la persona a la que corresponde la página. Sin embargo, en muchos casos los nombres, incluido el apellido, no identifican a una persona de forma unívoca.

Hemos aplicado técnicas clásicas de recuperación de información para extraer los términos más representativos de la página. Una vez eliminadas las palabras vacías, generamos un índice de términos ordenado por frecuencias. Los diez primeros términos de este índice se utilizan, uno a uno, para expandir la consulta formada por el texto del ancla. Es decir, se expande con cada uno de ellos y se toman los diez primeros documentos recuperados en cada caso.

En la tabla 2 se puede observar como la expansión mejora globalmente los resultados aumentando el número de enlaces recuperados en las diez primeras posiciones y por tanto reduciendo los enlaces no recuperados. A pesar de esto, el número de enlaces recuperados en primera posición se ve reducido.

Análisis.	1 pos.	1-10 pos.	E.N.R.
No EXP	253	380	536
EXP	213	418	498

Cuadro 2: Análisis del número de documentos recuperados en primera posición (1 pos.), entre las diez primeras posiciones (1-10 pos.) o no recuperados (E.N.R.) en función de utilizar (EXP) o no (No EXP), el método de expansión de la consulta.

Por ello consideramos que lo más adecuado es aplicar ambas formas de recuperación, y ordenar después los resultados para presentar al usuario los más relevantes en primer lugar.

Analizando los casos en los que se consigue recuperar la página correcta con y sin entidades nombradas y en función del número de términos del ancla (tabla 3) vemos que las proporciones obtenidas recuperando sin expandir la consulta se mantienen. Es decir, los mejores resultados se obtienen cuando hay entidades nombradas y cuando hay dos o más términos. Sin embargo, en este caso, es decir con expansión, el número de enlaces recuperados cuando el ancla consta de un único término y no es una entidad nombrada es 25, que ya puede ser una cantidad significativa. Esto sugiere intentar recuperar con expansión también en este caso, siempre que sea posible comprobar la validez de los resultados.

2.3. Ordenación de los enlaces por relevancia

En este momento hemos recuperado un conjunto de enlaces candidatos a sustituir al enlace roto, procedentes de la búsqueda con el ancla y con el ancla expandida con cada uno de los diez primeros términos que representan a la página padre. Ahora queremos ordenarlos por relevancia para presentarlos al usuario. Para calcular esta relevancia hemos considerado dos fuentes de infor-

	Tipo de ancla			
	Ent. Nomb.		No Ent. Nomb.	
Terms.	E.N.R.	E.R.	E.N.R.	E.R.
1	104	65	127	25
2	55	72	70	70
3	30	28	22	50
4+	59	59	31	49
total	248	224	250	194

Cuadro 3: Análisis de los enlaces no recuperados y recuperados en función del tipo de ancla y del número de términos del ancla cuando la expansión es aplicada.

mación. En primer lugar, si existe, la página a la que apuntaba el enlace roto almacenada en la caché del buscador, en nuestro caso de *Google*. Si esta información no existe, entonces utilizamos la página padre que contiene el enlace roto. La idea es que la página enlazada tratará en general sobre una temática relacionada con la página en la que se encuentra el enlace.

De nuevo hemos aplicado el modelo de espacio vectorial (Manning, Raghavan, y Schütze, 2008) para estudiar la similitud entre la página que contenía el enlace roto y las páginas recuperadas. Con esta técnica calculamos la similitud o bien con la caché o bien con la página padre. La figura 2 muestra los resultados correspondientes. En el primer caso, la mayoría de los documentos correctos recuperados se presentan entre los diez primeros documentos, con lo que si se dispone de la caché, podremos hacer recomendaciones muy fiables. En el caso de la similitud con la página padre, el orden de los resultados es peor. Por lo que sólo recurriremos a esta información si no se dispone de la caché.

2.4. Colección de páginas y Resultados de la Recuperación Automática de Enlaces

Si analizamos la utilidad de las distintas fuentes de información utilizadas directamente sobre enlaces rotos, es muy difícil evaluar la calidad de las páginas candidatas a sustituir el enlace. Por ello, en esta fase de análisis trabajamos con enlaces *Web* tomados de forma aleatoria, que no están realmente rotos, y que denominamos *supuestamente rotos*. De esta forma disponemos de la página a la que apuntan y podemos evaluar la recomendación

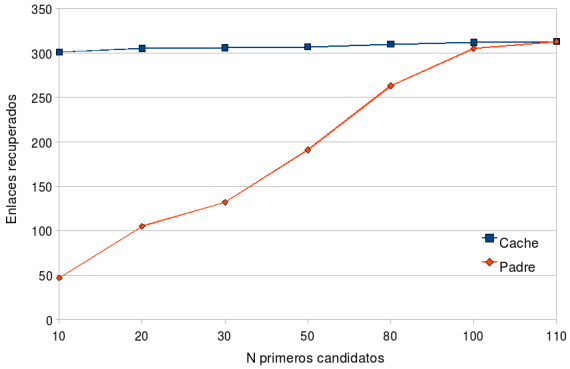

Figura 2: Número de apariciones de páginas correctas en el ranking elaborado, seleccionando los N mejores candidatos según la similitud con la página caché y padre.

que hacemos utilizando cada fuente de información. Para realizar el análisis, tomamos diez enlaces por cada página elegidos aleatoriamente de un conjunto de 100 seleccionadas igualmente de manera aleatoria mediante peticiones sucesivas a *www.randomwebsite.com*, un sitio que proporciona páginas *Web* aleatorias. Este conjunto de páginas además deben cumplir una serie de requisitos en cuanto a su contenido como tener 250 palabras, estar escritas en ingles y tener al menos cinco enlaces activos, ajenos al propio sitio y cuyo texto de anclaje sea mínimamente descriptivo (no sea únicamente un número, una Url, un signo de puntuación o esté vacío).

Los resultados del análisis descrito en las secciones anteriores sugieren criterios para decidir en qué casos hay información suficiente para intentar la recuperación del enlace y qué fuentes de información utilizar. Esta información se ha modelado dando origen a un algoritmo cuyos resultados pasamos a describir.

Hemos aplicado este algoritmo a enlaces que están realmente rotos, pero sólamente de los que se dispone de caché, para poder evaluar los resultados. La tabla 4 muestra los resultados de la posición de los documentos más relevantes en una ordenación por similitud con la página padre. La relevancia se mide por similitud con la caché. Hemos comprobado que en unos casos se trata de la página original, que ha cambiado de Url, y en otros casos de páginas con contenido muy relacionado en una localización diferente. Podemos observar, que aún si no contamos con la caché y ordenamos por similitud con la página padre, el sistema es capaz de presentar

documentos sustitutos relevantes entre las 10 primeras posiciones en un 48 % de los casos y entre las 20 primeras en un 76 %.

N primeros	E.R
1-10	12
10-20	7
20-50	6

Cuadro 4: Número de apariciones de páginas sustitutas (de acuerdo con su similitud con el contenido de la caché) entre los N primeros documentos ordenados por similitud con la página padre.

3. Detección de Web Spam

Nuestro sistema de recuperación de enlaces analiza una Web tanto desde el punto de vista de sus enlaces como desde el punto de vista de su contenido. Aplicando esta metodología a los enlaces de una página (no rotos), puede extraerse información relevante sobre la coherencia de los enlaces y las páginas apuntadas por ellos, que es útil para determinar si una página esta realizando *Spamdexing*. Nuestra propuesta es novedosa para la detección de Spam, ya que habitualmente los sistemas que se encargan de esta tarea realizan un crawling previo, reuniendo una colección de páginas Web junto a una serie de medidas relevantes. Posteriormente y de una manera *offline*, se realiza una clasificación sobre la colección. En los últimos años existe una colección de referencia (Castillo et al., 2006) siendo la primera que incluye las páginas y sus enlaces y que además ha sido etiquetada manualmente por un conjunto de voluntarios. No obstante existen otros trabajos que emplean colecciones propias elaboradas de una forma similar. Este sistema sería novedoso ya que no necesita una colección con información sobre su contenido ni sobre sus enlaces, sino que de una manera *online* extrae de la red información relevante sobre una Web dada para posteriormente ser clasificada según su grado de Spam. Hemos realizado un estudio comparativo aplicado a dos conjuntos de 67 páginas Web, clasificadas previamente como Spam o No Spam, en el que tomando una serie de medidas podemos apreciar ciertas características propias de cada conjunto. Estos dos conjuntos han sido tomados de (Castillo et al., 2006), teniendo en cuenta su clasificación en cuanto a su grado de Spam. Además fue imprescindible que las páginas estuvieran online y que su cuerpo contuviera al menos 100 palabras y un enlace externo.

La primera medida corresponde a la diferencia entre los enlaces recuperados y no recuperados por cada una de las páginas. El análisis en este caso se ha realizado mediante una recuperación de los enlaces activos para poder verificar su recuperación. La intuición en la interpretación de este valor es que una página que hace Spam está enlazando a otras páginas poco conocidas y por tanto, difíciles de recuperar. Por lo tanto, cuanto más negativa es la diferencia entre los enlaces recuperados y no recuperados, mayor es la probabilidad de que la página esté haciendo Spam. En la figura 3 se pueden apreciar las dos distribuciones de estas medidas para cada una de las páginas y por cada uno de los dos conjuntos (Spam y no Spam). También se puede observar como en el caso de las páginas que no hacen Spam, sus valores casi siempre están por encima de los de las páginas de Spam.

Las dos siguientes figuras 4 y 5 corresponden a las páginas de Spam y las de no Spam respectivamente. En ellas se muestra la relación entre las distribuciones de los enlaces de cada página y las páginas que las enlazan. Se puede comprobar como en el primer caso la diferencia es mucho mayor, estando siempre los enlaces entrantes por debajo de los salientes. Estos datos indican que las páginas de Spam contienen muchos enlaces pero en cambio no reciben el mismo número.

En la figura 6 se muestran las distribuciones de la siguiente medida para cada una de las páginas y por cada uno de los dos conjuntos (Spam y no Spam). Esta medida corresponde al valor absoluto de la diferencia entre los enlaces externos y los que son internos. Las páginas de Spam normalmente toman dos estrategias distintas en cuanto a los enlaces, o bien la mayoría son externos con el objetivo de crear granjas de enlaces o por el contrario intentan absorber la mayoría del *PageRank* conteniendo mayoritariamente enlaces al mismo sitio. De esta forma, en la figura 6 se puede comprobar como el equilibrio entre este tipo de enlaces es mayor en el caso de las páginas que no realizan Spam.

Otras dos medidas interesantes (no mostradas en este trabajo) para clasificar una página son el número de las mismas que la enlazan así como el número de enlaces cuyo

texto del ancla es una Url. Para la primera hemos tomado el valor correspondiente aproximado que proporciona el buscador. De esta forma hemos podido comprobar como las páginas de Spam tienen valores muy por debajo, verificando la teoría de que las páginas de prestigio transfieren su confianza a páginas igualmente prestigiosas. Cabe mencionar que existen páginas de Spam con valores elevados, sin embargo corresponden a sitios importantes pero clasificados como Spam por su alto contenido de publicidad. En cuanto al número de enlaces cuyo texto del ancla es una Url, en términos generales las páginas de Spam contienen una mayor cantidad.

Todas estas medidas junto con otras estudiadas en la literatura, tanto en relación al contenido como a la estructura de los enlaces, podrían ser utilizadas para la detección de páginas de Spam.

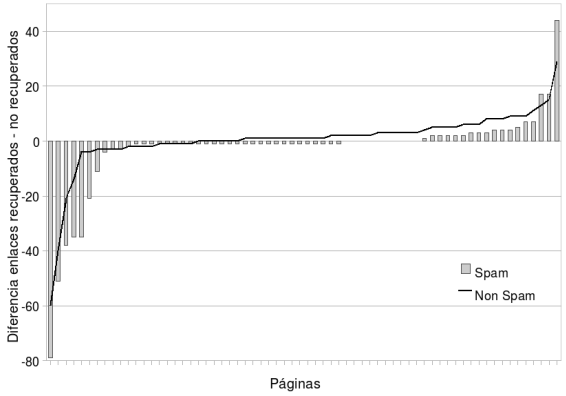

Figura 3: Distribución de la diferencia entre los enlaces recuperados y no recuperados para dos conjuntos de páginas (Spam y no Spam).

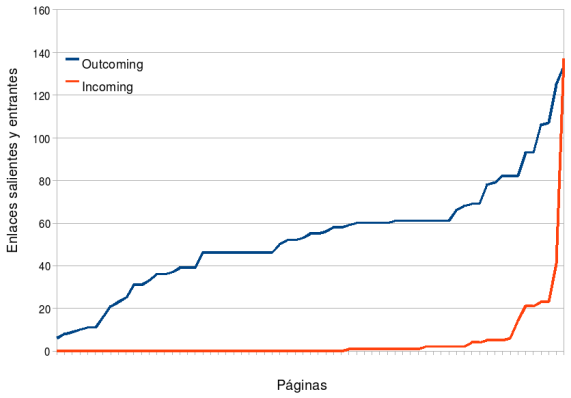

Figura 4: Distribución de los enlaces salientes y entrantes para las páginas de Spam.

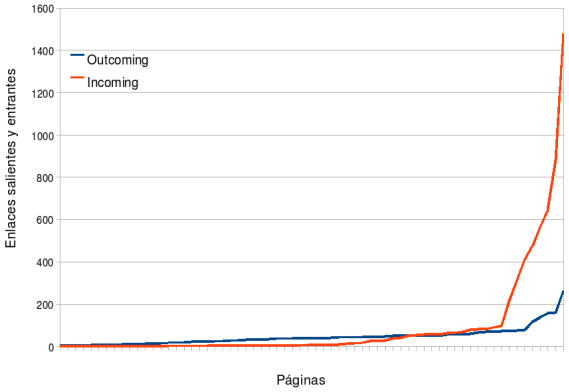

Figura 5: Distribución de los enlaces salientes y entrantes para las páginas de No Spam.

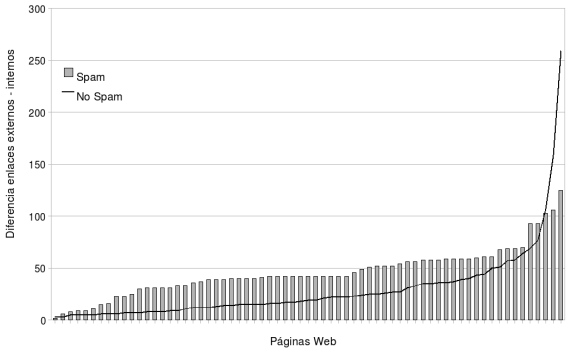

Figura 6: Distribución de la diferencia entre los enlaces externos e internos para dos conjuntos de páginas (Spam y no Spam).

4. *Conclusiones y Futuros trabajos*

En este trabajo presentamos un estudio preliminar sobre una serie de medidas que podrían ser útiles para la detección de Spam en la Web. En particular, analizamos la medida de coherencia entre los enlaces y las páginas apuntadas por ellos. Otras medidas analizadas son las diferencias entre los enlaces entrantes y salientes, entre los enlaces externos e internos o el número de enlaces cuyo texto de anclaje es una Url. Estas medidas son obtenidas a su vez gracias a un sistema de recuperación de enlaces. El sistema resultante resultaría novedoso ya que no necesitaría de una colección con información precalculada sino que funcionaría de una manera *online*.

En cuanto al sistema de recuperación de enlaces, hemos analizado distintas fuentes de información que podemos utilizar para hacer una recuperación automática de enlaces *Web* que han dejado de ser válidos. Los resultados indican que los términos del ancla pueden ser muy útiles, especialmente si hay más

de uno y si contienen alguna entidad nombrada. Hemos estudiado también el efecto de añadir términos procedentes de la página que contiene el enlace, con el fin de reducir la ambigüedad que puede conllevar la cantidad limitada de términos del ancla. Este estudio ha mostrado que los resultados mejoran a los obtenidos utilizando sólo los términos del ancla. Sin embargo, como hay casos en los que la expansión empeora el resultado de la recuperación, hemos decidido combinar ambos métodos, ordenando después los documentos obtenidos por relevancia. El resultado de este análisis ha sido un algoritmo que ha conseguido recuperar una página muy cercana a la desaparecida entre las diez primeras posiciones de los documentos candidatos en un 48 % de los casos, y entre las 20 primeras en un 76 %.

En este momento trabajamos en analizar otras fuentes de información que pueden ser útiles tanto para la recuperación de enlaces como para la detección de Spam, como las propias Urls, las páginas que apuntan a la página analizada o el contenido de sus distintas partes.

Bibliografía

Abernethy, Jacob, Olivier Chapelle, y Carlos Castillo. 2008. Webspam identification through content and hyperlinks. En *Proceedings of the fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Baeza-Yates, Ricardo, Paolo Boldi, y José María Gómez Hidalgo. 2007. Recuperación de información con adversario en la web. *Novática: Revista de la Asociación de Técnicos de Informática*, 185:29–35.

Baeza-Yates, Ricardo A., Carlos Castillo, y Vicente López. 2005. Pagerank increase under different collusion topologies. En *AIRWeb*, páginas 17–24.

Castillo, Carlos, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, y Sebastiano Vigna. 2006. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24.

Davis, H.C. 2000. Hypertext link integrity. *ACM Computing Surveys Electronic Symposium on Hypertext and Hypermedia*, 31(4).

Davison, B. 2000. Recognizing nepotistic links on the web.

Fetterly, Dennis, Mark Manasse, y Marc Najork. 2004. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. En *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, páginas 1–6, New York, NY, USA. ACM.

Grønbæk, Kaj, Lennert Sloth, y Peter Ørbæk. 1999. Webvise: Browser and proxy support for open hypermedia structuring mechanisms on the world wide web. *Computer Networks*, 31(11-16):1331–1345.

Gyöngyi, Zoltán y Hector Garcia-Molina. 2005. Web spam taxonomy. En *Proceedings of the first International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Jansen, Bernard J. y Amanda Spink. 2003. An analysis of web documents retrieved and viewed. En *International Conference on Internet Computing*, páginas 65–69.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Nakamizo, A., T. Iida, A. Morishima, S. Sugimoto, , y H. Kitagawa. 2005. A tool to compute reliable web links and its applications. En *SWOD '05: Proc. International Special Workshop on Databases for Next Generation Researchers*, páginas 146–149. IEEE Computer Society.

Ntoulas, Alexandros, Marc Najork, Mark Manasse, y Dennis Fetterly. 2006. Detecting spam web pages through content analysis. En *WWW '06: Proceedings of the 15th international conference on World Wide Web*, páginas 83–92, New York, NY, USA. ACM.

Shimada, Takehiro y Atsushi Futakata. 1998. Automatic link generation and repair mechanism for document management. En *HICSS '98: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 2*, página 226, Washington, DC, USA. IEEE Computer Society.