Recuperación de información Web: 10 años de cibermetría

Web Information Retrieval: 10 years of cybermetrics

José L. ALONSO BERROCAL, Carlos GARCÍA FIGUEROLA y Ángel F. ZAZO

Grupo de Recuperación Avanzada de la Información – REINA, Departamento de Informática y Automática, Universidad de Salamanca, C/ Francisco Vitoria 6-16, 37008 Salamanca, España.
{berrocal | figue | afzazo}@usal.es

Resumen

El objetivo de esta ponencia es hacer un repaso de la evolución, en los últimos 10 años, en el campo de la recuperación de información web. Con la implantación de las diferentes técnicas cibermétricas la evolución de los estudios de la web ha sido espectacular y es en estos momentos un campo inagotable de estudio.

Palabras clave: Recuperación de información. Cibermetría.

Abstract

The objective of this communication is to make a review of the evolution, in the last 10 years, in the field of the Web information retrieval. With the implantation of the different cybermetrics techniques the evolution from the studies of the Web has been spectacular and is at the moment an inexhaustible field of study.

Keywords: Information retrieval. Cybermetrics.

1. Introducción

Ningún periodo de la historia ha presenciado cambios y desafíos tan profundos en la organización y la difusión de la información como el actual. Los descubrimientos a nivel mundial cada vez mayores en tecnologías de la información y en servicios nos han incitado a dar pasos en un mundo artificial cuyos elementos, fenómenos y seres son totalmente diferentes del mundo en el cual vivimos. El mundo que se conoce como ”Ciberespacio”.

Un espacio en el que la principal criatura viva es la información. Alvin Toffler lo denominó como Infoesfera. Quizás, cuando William Gibson acuñó el término Ciberespacio (Gibson, 1984) en su Neuromancer no podría imaginarse cómo los horizontes insondables tendrían un futuro tan cercano. Hoy, estamos haciendo frente a un Ciberespacio mucho más complejo y multidimensional que el percibido por Gibson. El término Ciberespacio, ampliamente utilizado hoy en día, define las complejas comunicaciones del Web a nivel mundial (Haynes, 1995).

Hoy, no solamente los profesionales de la información, los documentalistas y los bibliotecarios utilizan los amplios potenciales del Ciberespacio, sino también todas las personas de una gran variedad de profesiones y de empresas que utilizan las diferentes capacidades de este espacio de la información. Esta es la razón por la cual hay varias opiniones respecto a la noción de Ciberespacio. Una definición de Ciberespacio la considera como un espacio de posibilidades de computación interactivas, donde están disponibles los ordenadores y su contenido para los usuarios de cualquier ordenador dondequiera que se encuentren (Bauwens, 1996). Esta definición continúa con una interpretación orientada hacia la información, que considera que el Ciberespacio es donde se almacenan y se transmiten cada vez más información y conocimiento, siendo muy importante el lugar donde estamos al comunicarnos con un colega a través de los ordenadores.

Haynes también cree que el Ciberespacio es más amplio que el World Wide Web e incluso que Internet. Hay varios miles de redes de comunicaciones que demuestran la globalidad del ciberespacio (Haynes, 1995). Hojeando estas definiciones, un aspecto que sobresale en la mayoría de ellas, es que inciden más en los medios que en el significado. La explicación de Bauwens sobre el Ciberespacio parece ser más intensa en el aspecto de la información que las otras definiciones. Acentuando la importancia de la información en nuestra discusión de Ciberespacio, quisiera señalar la atención de los profesionales de la información a esta realidad indiscutible de que la información es el corazón de tal atmósfera, aunque en la creación de dicho espacio intervengan muchas clases de tecno-

logías del ordenador, de la telecomunicación y de la información. Así pues, esta ciberinformacion marca una nueva frontera de la investigación de la información.

De forma general, la ciberinformación implica la información comunicada a través de medios electrónicos. Para clarificar más el concepto de ciberinformación, parece necesario indicar los medios de información principales, que constituyen la base del Ciberespacio. Pueden ser detallados como sigue:

1. Redes de información de todas las clases y alcances.
2. Bases de datos y metabases en línea.
3. Herramientas de Internet y medios incluyendo homepage, sedes Web, Email, grupos de discusión y de noticias.
4. Escuelas virtuales, universidades y organizaciones.
5. Sistemas del tablón de anuncios.
6. Conferencias electrónicas, asociaciones, y sociedades.
7. Libros electrónicos, bibliotecas, archivos y servicios información.
8. Sistemas de información multimedia, hipermedia, polymedia y telemedia.

Es evidente que la lista anterior no está completa, pero proporciona algunos componentes importantes del Ciberespacio. Muchos otros términos se pueden agregar a la lista anterior precedidos de adjetivos como ”electrónico” o ”digital” y el prefijo ”Ciber”. Un aspecto importante, que debería ser recalcado, es que muchas actividades, individuales y sociales, que tienen algún tipo de comunicación y de intercambio de información se incorporan cada vez más al ciberuniverso.

2. La cibermetría

El término Cibermetría aparece claramente definido en el año 1998 por Shiri (1998), y añadió una nueva dimensión en la investigación cuantitativa de la información electrónica en el ciberespacio. Este neologismo se utilizó, por Shiri, para destacar los aspectos modernos de la investigación de la información en un ambiente electrónico. El autor se centra en los aspectos cuantitativos de dicha investigación con estudios sobre sedes web, homepages, redes, así como con conceptos tales como análisis de citas electrónicas, estudios básicos sobre revistas electrónicas y recursos electrónicos, etc.

2.1. Antecedentes

El crecimiento rápido y cada vez mayor en la información electrónica junto con los amplios potenciales de las tecnologías y de los medios de información recientemente emergentes, han atraído la atención de los investigadores de la información para reflejar sobre la medida y la métrica cuantitativas de las fuentes de información, de los servicios y de los medios en esta esfera emergente, el cibercosmos. Las investigaciones principales en este área se han emprendido desde 1996 hacia adelante. Arnzen se refiere a cibercitas y a los ejemplos de citas del correo electrónico, website, ftp, Gopher, USENET, o listas de correo (Arnzen, 1996).

Clausen ha realizado otra investigación. Usando métodos de investigación mediante el empleo de encuestas, ha estudiado el uso de los recursos de Internet, usuarios y sus categorías de edad, el número de usuarios y conferencias electrónicas. En Dinamarca, se ha esforzado en cuantificar los hábitos de los usuarios de Internet y también sus actitudes hacia Internet como recurso de la información (Clausen, 1996).

Un estudio cuantitativo orientado a los medios, referente a Internet, es el trabajo de McMurdo. Su trabajo gira principalmente sobre los medios, más que sobre la información. Contando los host de Internet y sus dominios, la distribución de host por dominios, crecimiento del Web, número de hosts de las sedes Web y las relaciones de transformación de las sedes Web están entre los parámetros principales que él ha estudiado a través de su investigación. En el estudio se han utilizado algunas fuentes de información estadística y demográfica sobre internet (McMurdo, 1996).

Una de las investigaciones principales realizadas sobre la métrica del cibermedio es la de Almind e Ingwersen (1997). Procuraron introducir la aplicación de métodos informétricos al World Wide Web (WWW) denominándolo ”Webmetría”. Realizaron un estudio comparando la proporción Danesa de WWW a la de otros países nórdicos. La metodología usada era de análisis bibliométrico. Dentro de este estudio se analizaron cinco aspectos fundamentalmente:

- Un análisis de la posición de Dinamarca respecto al Web.
- Un análisis de la distribución de las páginas Web Danesas en grandes centros de enseñanza en Dinamarca.
- Un análisis de la distribución de los dominios científicos sobre una muestra.

- Un análisis de la distribución de las páginas Web sobre el tipo de documento.
- Un análisis de la distribución de frecuencias seleccionadas para una muestra de páginas Web.

Este estudio también ha explorado el número medio de hiperenlaces por página Web y la densidad de enlaces para los diferentes tipos de dominio. Este estudio webométrico era una investigación de todas las comunicaciones basadas en la red usando la informetría u otras medidas cuantitativas. Sin embargo, debemos considerar que se han centrado principalmente en el análisis cuantitativo del World Wide Web.

También el trabajo de Abraham (1997) —y empleando el término Webmetría— habla de la necesidad de aplicar técnicas de redes neuronales para el mejor conocimiento del Web, representando las conexiones de los nodos mediante número reales, que indicaran la fuerza de la conexión. Indica la necesidad de emplear matrices, aunque en ningún momento hace referencia a la teoría de grafos.

En 1997 se inició una investigación, en la Escuela Real de Bibliotecarios de Dinamarca, dirigida a explorar por estudios cuantitativos ciertos fenómenos y acontecimientos actuales de la información. Uno de los primeros objetivos de esta investigación es el análisis de la creación, uso y del estudio de las homepages danesas y n3órdicas. Este estudio también se ha referido a Internetmetría (Informetría). Parece estar más orientado a la información que las investigaciones anteriores.

Un estudio realizado sobre el factor de impacto del Web (Ingwersen, 1998) informa sobre las investigaciones para ver la viabilidad y la fiabilidad en el cálculo del factor de impacto de las sedes Web llamado factor de impacto del Web. El estudio demuestra que el factor de impacto del Web es calculable y fiable, con la precaución necesaria, para estimar el número de las páginas del Web que señalan a las páginas de una sede determinada.

Dahal (1999) aplicó las leyes bibliométricas al análisis del desarrollo de los sistemas de información en ciencia y tecnología del Nepal, empleando finalmente el término cibermetría para explicar las técnicas empleadas. Parece evidente que la aplicación de la métrica y de las medidas cuantitativas a la información electrónica se está convirtiendo cada vez más un área significativa para la investigación.

La historia está plagada de acontecimientos, pero, para terminar, indicaremos el trabajo de Thelwall (2007) donde aborda el cambio desde la bibliometría hasta la webometría actual e indicando las posibilidades de trabajo sobre la web 2.0.

A partir de estos antecedentes iniciales, el término cibermetría ha convivido desde la segunda mitad de los 90 con una amplia cantidad de términos, utilizados para designar trabajos o campos de estudio de naturaleza similar: Netometrics (Bossy, 1995); Webometry (Abraham, 1997); Internetometrics (Almind e Ingwersen, 1996); webometrics (Almind e Ingwersen, 1997); cybermetrics; o web bibliometry (Chakrabarti et al., 2002).

También ha habido una gran variedad de planteamientos, surgidos desde mediados de los noventa, con nombres como Ecología Web (Chi el al., 1998; Huberman, 2001), Inteligencia Web (Yao, Zhong, Liu u Ohsuga, 2001) y análisis de grafos Web (Broder et al., 2000; Chakrabarti et al., 1999; Kleinberg, 1999).

La razón de ser del término Webometría es que puede ser considerado como un vestigio de la bibliometría e informetría y enfatiza una perspectiva documental de los estudios sobre la Web. Muchos han sido los trabajos de investigación que han trabajado y tratado sobre el tema.

2.2. Definición y campo de estudio

Björneborn e Ingwersen (2004) proponen una diferenciación terminológica distinguiendo entre estudios de la Web y estudios de todos los servicios de Internet. Ellos usan una definición del mundo de la Documentación, según la cual, la Webometría *sería el estudio de aspectos cuantitativos de la construcción y uso de recursos, estructuras y tecnologías de la información en la WWW a partir de planteamientos bibliométricos e informétricos*. Esta definición cubre aspectos cuantitativos tanto de la construcción como del uso de la web, abarcando las cuatro áreas principales de la investigación actual:

- Análisis de contenido de páginas web.
- Análisis de la estructura de enlaces web.
- Análisis del uso web (por ejemplo, explotando las conductas de navegación y búsqueda de los usuarios a través de ficheros de transacciones web).
- Análisis de tecnologías web (incluyendo diseño de buscadores).

Esto incluye formas híbridas como las propuestas por Pirolli et al. (1996), que exploraron técnicas de análisis web para la clasificación automática utilizando teoría de grafos, contenido

textual y similaridad en los metadatos, además de datos de uso.

La Cibermetría es propuesta como término genérico para *el estudio de aspectos cuatitativos de la construcción y uso de recursos, estructuras y tecnologías de información de toda Internet, a partir de planteamientos bibliométricos e informétricos*. La Cibermetría aglutina así estudios estadísticos de grupos de discusión, listas de correo y otras formas de comunicación en la red, incluyendo la Web. Junto a todo tipo de comunicaciones desarrolladas a través de Internet, esta definición de Cibermetría también cubre estudios cuatitativos de tipologías y tráfico.

La amplitud de cobertura de la Cibermetría y la Webometría implica un solapamiento con la proliferación de planteamientos basados en computación para el análisis de contenidos web, estructura de enlaces, uso y tecnologías web.

En la actualidad, el término más común y aceptado es el de Webometría (Webometrics), aunque en España es el término Cibermetría el que posee una mayor implantación. Tiene la ventaja añadida, además, de que aglutina las dos visiones, tanto de la webometría como de la cibermetría.

El principal incentivo de la cibermetría es la amplia variedad de nuevos medios electrónicos por medio de los cuales se comunica una amplísima gama de informaciones. Desde que los servicios de información tradicional y las fuentes, en gran parte, han sido transformadas en nuevos soportes y formatos que reclaman un cambio en el acercamiento a los estudios de la información, la necesidad urgente de reconsiderar nuestros esfuerzos investigadores en esta área parecen evidentes.

Las redes de información como mecanismo importante para la comunicación de la información puede considerarse como una de las áreas principales para ser estudiada. Existen redes funcionando a nivel nacional, internacional o globalmente. El número de cada clase de red, su cobertura temática, el número de usuarios y su dispersión geográfica son elementos para su investigación.

Internet como red de información global nos ha provisto de una amplia gama de servicios informativos y de medios. Las sedes Web, las homepages, el E-mail, grupos de discusión y de noticias son algunas de las herramientas principales de Internet a través de las cuales todas las clases de información pueden ser transmitidas. Estas herramientas han ofrecido el motivo para publicar en los nuevos medios, tales como los libros electrónicos, las revistas, las bibliotecas y los archivos. Junto con el desarrollo de tales recursos, una amplia variedad de herramientas de búsqueda, de recuperación y el empleo de técnicas como el hipertexto, los agentes inteligentes, los knowbots, etc., que permiten a los usuarios que busquen eficientemente la información necesaria. De forma similar, la convergencia de varios medios en una sola plataforma ha originado sistemas de información como multimedia, hipermedia y polimedia.

Ahora, la pregunta es ¿qué se puede medir en este contexto? Si nos referimos a los elementos que se mencionaron al hablar del Ciberespacio, podríamos clarificar esta circunstancia. Se proponen a continuación algunos de estos elementos:

1. El número, el alcance y los temas de las redes de información.
2. Distribución de las redes por países.
3. Volumen de las colecciones de información en las redes por tamaño y tipo.
4. Distribución de los diversos tipos de redes.
5. Evaluación de los tiempos de respuesta de las redes y provisiones de acceso.

Internet, como enorme autopista de la información, ha proporcionado argumentos muy interesantes para el estudio. Por ejemplo para el estudio del E-mail podemos hacer lo siguiente:

1. El número de direcciones de correo.
2. Distribución de las direcciones de correo por países, organismos e instituciones.
3. Uso del correo en los sectores público y privado.
4. El volumen, el tipo y el tamaño de la información enviada a través del correo.
5. Distribución de los usuarios de correo por profesiones y empresas.
6. Proporciones de diversos tipos de documento enviados por correo.

Éstas son las áreas, que se pueden cuantificar usando medidas estadísticas y técnicas informétricas. Uno de los medios de información que más profundamente han influido el mundo de la información en el mundo entero es el World Wide Web, un Web de información hipertexto multimedia que opera como una de las autopistas de Internet. Esta tecnología, siempre en expansión, ha provocado cambios tanto a nivel individual como en las diversas actividades sociales. Hoy en día, todas las organizaciones, instituciones tanto públicas como privadas tie-

nen sus propias sedes y homepages. Podemos encontrar fácilmente todas las clases y formatos de la información en el World Wide Web. Una gran cantidad de productores, de proveedores y de vendedores de la información han puesto sus colecciones en el Web Por lo tanto, la métrica y la medida de estos medios impresionantes, sin ninguna duda, son un área interesante para la investigación. Algunas de estas áreas de estudio son las siguientes:

1. El número de sedes Web y de homepages en el mundo y también su distribución por países.
2. Clasificación de las páginas Web por tipos de documentos.
3. Número de páginas Web por dominios.
4. Clasificación de páginas Web por el idioma de los documentos y por los modos de representación de la información.
5. Estadísticas de uso y usuarios de las páginas Web en un período de tiempo dado.
6. El número de citas recibidas por cada página Web.
7. Ordenar los Web más citados y páginas personales según el tipo de documento.
8. Los tipos de colecciones electrónicas disponibles en cada sede Web.
9. Factor de Impacto del Web y productividad de los autores.
10. Análisis del contenido de las páginas Web.
11. Identificar la variedad de publicaciones electrónicas por el tipo, el idioma y la distribución geográfica.

Estas medidas cuantitativas del Web no pueden solamente mostrar la anchura y la amplitud del WWW sino también pueden mostrar las etapas de desarrollo de los recursos del WWW a través del mundo. Midiendo recursos electrónicos tales como libros electrónicos, revistas, bibliotecas y fuentes de referencia se pueden elaborar otras investigaciones que nos permitan reconocer la transición revolucionaria de lo impreso al mundo electrónico. Para tener una idea del análisis cuantitativo de estos recursos electrónicos, algunos de los principales aspectos que se pueden tratar son:

1. Estadística de bibliotecas digitales.
2. Número de revistas electrónicas por temas e idiomas.
3. Número de revistas publicadas en ambos formatos (electrónico y papel).
4. Número de fuentes de referencia electrónica disponibles.
5. Análisis de citas de revistas electrónicas.
6. Utilización de las revistas electrónicas.
7. Distribución de recursos electrónicos por tipo, país e institución.
8. Productividad científica en el entorno electrónico.
9. Crecimiento de la literatura electrónica y su obsolescencia.

Así los diferentes estudios han profundizado en el trabajo de diferentes medidas o índices y entre estas medidas, se encuentran:

- El tamaño medio de los documentos analizados.
- Los protocolos utilizados por los URLs de los documentos HTML analizados.
- Los tipos de ficheros.
- El recuento de dominios científicos.
- La tipología documental de las páginas Web.
- Los recursos: página Web con datos textuales o audiovisuales.
- El número medio de enlaces por página.
- La densidad media de enlaces.
- El tamaño documental.
- El tamaño informático.
- La densidad hipertextual.
- La densidad multimedia.
- La profundidad.

A estos, pueden añadirse otros elementos como: el número de revistas electrónicas según su temática e idioma, el número de revistas publicadas en formato electrónico e impreso, el número de obras de referencia disponibles electrónicamente, la distribución de recursos electrónicos por tipo, país e institución, así como la productividad científica en el entorno electrónico. Estos últimos elementos apuntarían, sobre todo, a la medición de la comunicación científica en el Web y como puede observarse constituyen sólo adaptaciones al entorno digital, porque se utilizan también en los estudios métricos tradicionales.

La cantidad de trabajos que abordan la disciplina es enorme y algunos que podemos destacar son (Pinto Molina et al., 2003; Aguillo Caño et al., 2004; Alonso Berrocal et al., 2004; Aguillo Caño et al., 2005; Baeza-Yates et al., 2005;

Alonso Berrocal et al., 2006; Baeza-Yates et al., 2006; Ortega Prieto, 2007; Cordón García et al., 2007; Baeza-Yates et al., 2007; Tolosa et al., 2007).

3. Estudios avanzados de la web

Si consideramos la web como una colección de páginas conectadas a través de enlaces, y no consideramos toda la información sobre su contenido, localización y URLs, entonces nos encontramos ante un grafo matemático. Físicos e informáticos han intentado construir modelos de la web a través de sus enlaces y la forma en que se relacionan.

Estos planteamientos pueden ser definidos como topológicos porque ellos tratan la web como un grafo ignorando las relaciones espaciales entre el contenido de las páginas.

La web es la mayor red existente desde este punto de vista, lo que la ha convertido en el terreno de pruebas para muchos esfuerzos actuales de modelización (Albert y Barabási, 2002).

Según (Baeza-Yates et al., 2006) el estudio se puede abordar de la siguiente forma:

1. Vista macroscópica: estructura general.
2. Vista Microscópica: nodos.
3. Vista mesoscópica: regiones.

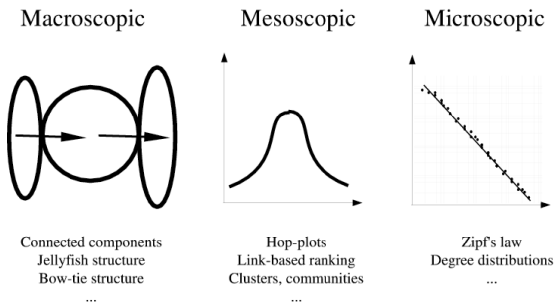

Figura 1. *Obtenido de Baeza-Yates et al. (2006)*

3.1. Vista Macroscópica

La teoría de grafos introduce planteamientos topológicos sobre la estructura completa y la conectividad de la red. La primera aportación establecía que cualquier par de páginas web podían ser conectadas por una cadena corta de enlaces, con sólo 19 enlaces como media (Albert et al., 1999).

Pero se descubrió más tarde que esto no era correcto (Broder et al., 2000), ya que muchos pares de páginas web no estaban conectadas con todas. Esto es debido a que la web es una red dirigida, esto es, los enlaces permiten ir únicamente en una dirección y no volver, así pues podemos llegar a un grupo de páginas con escasamente un par de clics, pero sin embargo debemos realizar una gran cantidad de clics para volver al punto original.

Broder et al. (2000) presentaron el estudio más innovador extrayendo los datos almacenados en Altavista, procesando 200 millones de páginas y 1,5 billones de enlaces. Entorno al 90% de los enlaces formaban un enorme grupo conectado. Este grupo estaba formado por cuatro partes iguales, las cuales formaban el modelo de ”lazo de pajarita” (bow-tie model) debido a su forma.

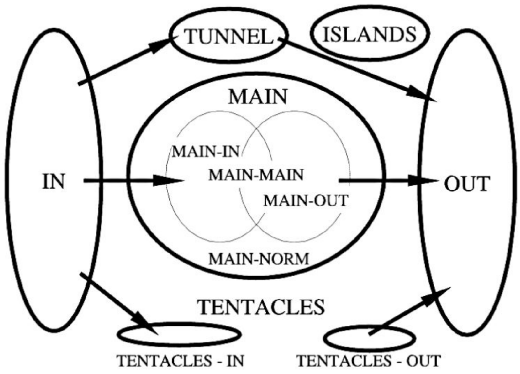

Figura 2. *Modelo Bow-tie (Broder et al., 2000)*

El centro es un componente fuertemente conectado (*strongly connected component*, SCC) en el que todas las páginas pueden trazar una ruta de enlace directo a otras del grupo SCC. La otra parte comprende un grupo de páginas (OUT) a las que se pueden acceder desde el núcleo SCC a través de un enlace directo, pero desde las que no se puede volver al núcleo SCC. Otra es un grupo de páginas (IN) desde donde se puede acceder de forma directa al núcleo SCC, pero no se puede salir. Otro pequeño grupo lo forman los tubos (TUBES) que son nodos que conectan la zona IN y OUT sin mediación del núcleo central SCC. Las restantes (TENDRILS) constituyen nodos que toman una ruta al exterior, fuera del núcleo de páginas estudiadas. Las restantes páginas que no están conectadas de ninguna forma al núcleo del 90% son DISCONNECTED.

A partir de aquí otros estudios han tratado de ver esta disposición general como en el modelo *jellyfish* (medusa) o en el modelo corona, que pueden verse en las siguientes imágenes.

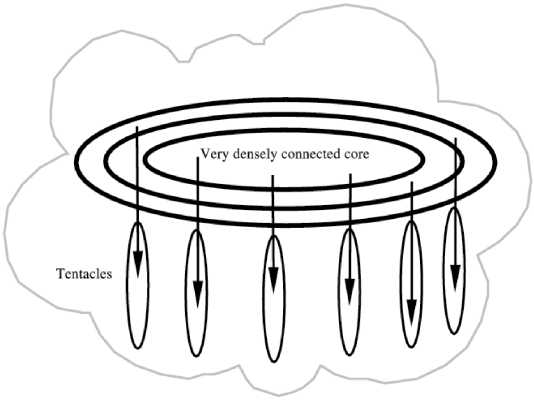

Figura 3. *Modelo medusa (Tauro et al., 2001)*

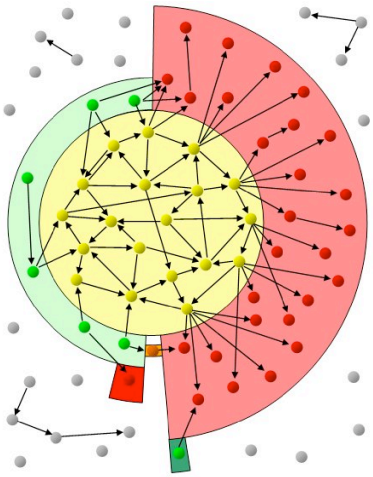

Figura 4. *Modelo corona de los componentes del grafo web (Björneborn, 2004)*

3.2. Vista macroscópica

Para entender cualquier proceso de conexión en la web es fundamental el concepto de ”red de escala libre” (scale-free network) (Barabási et al., 2000) y Barabási y Albert (1999).

En una red de escala libre, el grado de conectividad presenta una característica escalar, donde sólo unos pocos nodos atraen una gran cantidad de conexiones y la restante mayoría sólo recibe apenas unos pocos enlaces (Ball, 2000).

Así pues las distribuciones de enlaces entrantes y salientes en redes de escala libre muestran una tendencia potencial (Power-law). Distintas distribuciones potenciales se han identificado en la web y bajo este prisma están los estudios que tratan de analizar el comportamiento concreto de los nodos del grafo que forma el web. Se analizan las distribuciones del grado de los grafos o de las leyes de Zipf.

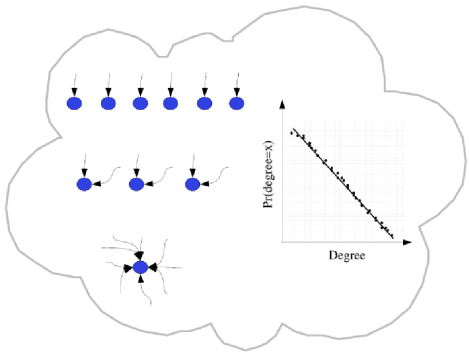

Figura 5. *Estudio del comportamiento de los nodos (Barabási, 2002)*

3.3. Vista mesoscópica

Bajo este criterio se intenta analizar la situación de determinadas regiones de la web. Uno de los primeros trabajos que abordó este criterio mediante las denominadas leyes de exponenciación fue el de (Faloutsos et al., 1999) y en su tercera ley permitía este tipo de estudio. En el trabajo de (Alonso Berrocal et al., 2004) se aplicaban estas técnicas a las universidades españolas.

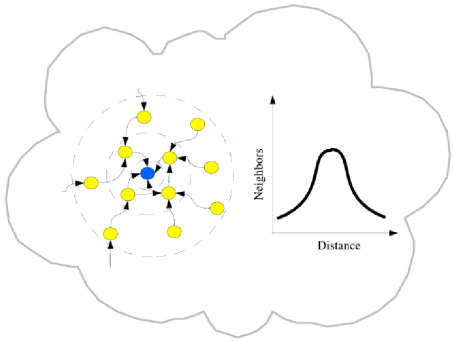

Figura 6. *Estudio del comportamiento de determinadas regiones*

4. Técnicas de posicionamiento

Las técnicas de posicionamiento las podemos entender como el conjunto de procedimientos que permiten colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda. Estas técnicas han tenido y tienen un campo de trabajo y estudio muy activo y en el que se trabaja de forma constante.

Algunas de las técnicas más utilizadas han sido:

4.1. Hits

Este algoritmo desarrollado por Kleinberg (Kleinberg, 1999) depende de la consulta y con-

sidera el conjunto de páginas S que *apuntan a* o *son apuntadas* por la respuesta.

- Las páginas que tienen muchos links que apuntan a ellas en S son llamadas autoridades (authorities)

$$A(p) = \sum_{v \in S | v \to p} H(v)$$

- Las páginas que tienen muchos links de salida son llamadas conectores (hubs)

$$H(p) = \sum_{u \in S | p \to u} A(u)$$

Las mejores páginas authorities vienen de links de entrada desde buenos conectores (hubs) y buenos hubs vienen de enlaces de salida de buenas autoriades (*authorities*).

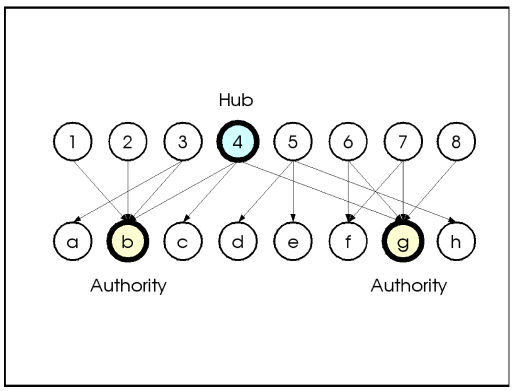

Figura 7: Algoritmo HITS

4.2. PageRank

El PageRank (Page et al., 1998) es la técnica de posicionamiento de mayor exito y, aunque se han descrito diversos problemas en su mecanismo básico de obtención, se han planteado soluciones a los mismos (Sung Jin y Sang Ho, 2002) y constantemente se publican artículos sobre la mejora del mismo. La técnica del PageRank ha demostrado suficientemente sus características como técnica de posicionamiento en los procesos de recuperación de información (Dominich y Skrop, 2005).

- El PageRank simula un usuario que navega aleatoriamente en la Web, quien salta a una página aleatoria con probabilidad q, o que sigue un hyperlink aleatorio (en la página actual) con probabilidad 1 - q.
- Este proceso es modelado como una cadena de Markov, donde la probabilidad estacionaria de estar en cada página puede ser calculada.
- La importancia de una página viene dada por la importancia de las páginas que la enlazan.

$$PR(a) = q + (1-q)\sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

5. Web Spam

Un campo de trabajo de gran actualidad son las investigaciones sobre Web Spam.

No podemos decir con certeza que exista una única definición de Web spamming, referido por muchos autores (Gyongyi y Garcia-Molina, 2005) como Spamdexing, y muchas veces definido como una práctica para conseguir una posición elevada en los resultados de los motores de búsqueda, utilizando técnicas para engañar a los algoritmos de clasificación.

El término ”spam” según (Castillo et al., 2006) ha sido utilizado en los últimos años referido a los mensajes no solicitados (normalmente comerciales).

El Spamdexing es definido por (Gyongyi y Garcia-Molina, 2005) y referido por (Castillo et al., 2006), como “cualquier acción con la intención de conseguir un aumento injustificado de la relevancia o importancia de una página web, considerando su valor real”.

Cualquiera que sea la definición, es cierto que el spam se refiere a algo indeseable, incluso perturbador, con una influencia negativa en el proceso HTTP, que, al basarse en el paradigma solicitud-respuesta, imposibilita el envío directo de las páginas por los spammers hacia los usuarios finales. Para superar esta defensa del protocolo, los spammers utilizan otras técnicas y medios. La más utilizada es a través de mensajes, aparentemente unidireccionales, via e-mail.

Pero si nos centramos en el modo de operar de los spammers sobre los sistemas de recuperación de información en el web, veremos que es diferente del resto. En este caso los principales destinatarios son los motores de búsqueda y su intereés es encontrar la forma de engañar y minar las relaciones de confianza establecidas entre los usuarios de los motores de búsqueda (Gyongyi y Garcia-Molina, 2005).

Estas técnicas de spam orientadas a los motores de búsqueda, pretenden obtener la atención de los usuarios finales, con fines normalmente comerciales. Una de las razones que lo subyacen está en las dificultades que tienen los usuarios finales para distinguir las informaciones confiables de las no confiables debido al éxito

de los motores en las últimas décadas (Metaxas y DeStefano, 2005).

Los usuarios han ido aumentando su confianza en los motores de búsqueda como medio de obtención de información, y los spamers han logrado, con éxito, conducir esa confianza a los resultados de cada consulta.

Para que sea posible continuar con la confianza en los resultados de las consultas, los constructores de motores de búsqueda, deben realizar un gran esfuerzo para proporcionar respuestas sin spam. Realizarán sofisticadas estrategias de ranking para detectarlo y eliminarlo (Becchetti et al., 2008).

De forma general, algunas de las principales formas de realizar spam web son: *Keyword stuffing* (relleno), *Link farms* (granjas), *Spam blogs* (splogs) y *Cloaking*.

5.1. SEO vs. SPAM

La Optimización para Máquinas de Búsqueda (S.E.O. por sus siglas en inglés) tiene que ver con asegurarse que un sitio sea encontrable por los buscadores. Los servicios que ofrecen los spammers incluyen la creación de miles o millones de páginas falsas que tienen como propósito el engañar a las máquinas de búsqueda y a sus usuarios.

En cualquier caso, la relación entre el administrador de un sitio Web que intenta tener un alto posicionamiento y el administrador de la máquina de búsqueda es una relación entre adversarios en un juego de suma cero. Cada ganancia inmerecida de ranking para una página es una pérdida de precisión para la máquina de búsqueda.

Existen dos grandes tipos de técnicas SEO:

- Técnicas SEO legítimas (técnicas de sombrero blanco), cuyo objetivo es hacer aparecer en lo más alto la página en cuestión, cuando un cliente está buscándolos (en contraposición a una página elaborada por personas que odian a su cliente). Es más eficaz, pues pregunta a los sitios web legítimos para vincularse al cliente.
- Spam (técnicas de sombrero negro), para crear lotes artificiales de los sitios web que enlazan a una página que promueve un producto (por ejemplo, Viagra).

El problema es que la línea de separación entre ambas técnicas es muy delgada y por ello es vital la investigación en esta línea.

6. Conclusiones

La cibermetría y su aplicación a la recuperación de información web tiene ya un amplio recorrido y ha aportado valiosas soluciones tanto en la caracterización de espacios web, como en la recuperación de información web. Infinidad de trabajos avalan su aplicación y sigue adaptándose a las nuevas tecnologías para seguir proporcionando adecuada respuesta a los problemas planteados.

Referencias

Abraham, R. (1997). Webometry: measuring the complexity of the world wide web. // World Futures. (50) 785–791.

Aguillo Caño, I. F.; Granadino Goenechea, M. B.; Ronda Laín, C.; Pareja Pérez, V. M.; Arroyo Vázquez, N.; Prieto Valverde, J. A.; Ortega Priego, J. L.; Amieva Vega, A.; Llamas Arigita, G. (2004). Factor de impacto y visibilidad de 4.000 sedes web universitarias españolas. Programa de estudio y análisis para la mejora de la calidad de la enseñanza superior y profesorado universitario. http://www.mec.es/univ/html/informes/estudios_analisis/resultados_2004/ea0020/EA2004-0020.pdf (2008-05-27).

Aguillo Caño, I. F.; Granadino Goenechea, M. B.; Zamora Meca, H.; Pareja Pérez, V. M.; Prieto Valverde, J. A.; Ortega Priego, J. L.; Amieva Vega, A.; Vega, A. A.; Martín Cuesta, B.; Canencia Rabazo, C.; Urdín Caminos, C. (2005). Impacto y visibilidad de las revistas electrónicas universitarias españolas. Programa de estudio y análisis para la mejora de la calidad de la enseñanza superior y profesorado universitario. http://www.mec.es/univ/ proyectos2005/EA2005-0008.pdf (2008-05-27).

Albert, R., Jeong, H., & Barabási, A.-L. (1999). The diameter of the World Wide Web. // Nature. 401 (1999)130–131.

Almind, T. C.; Ingwersen, P. (1996). Informetric analyses on the World Wide Web: A methodological approach to ”interneometrics”. Technical report, Centre for Informetric Studies, Royal School of Library and Information Science. (CIS Report 2).

Almind, T. C.; Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to ”webometrics”. // Journal of Documentation. 53:4 (1997) 404–426.

Alonso Berrocal, J. L.; Figuerola, C. G.; Zazo, Á. F. (2004). Cibermetría: nuevas técnicas de estudio aplicables al Web. Gijón: Trea, 2004.

Alonso Berrocal, J. L.; García-Figuerola Paniagua, C.; Zazo Rodríguez, N. F.; y otros (2006). Análisis de los flujos de información desde las webs universitarias españolas a las webs universitarias europeas en el marco del espacio europeo de educación superior. Programa de estudio y análisis para la mejora de la calidad de la enseñanza superior y profesorado universitario. http:// xww.centrorecursos.com/mec/ayudas/repositorio/20061129125858EA2006-0080.pdf(2008-05-27).

Arnzen, M. A. (1996). Cyber citations: Documenting internet sources prsents some thorny problems. // Internet Worls. 7:9 (1996) 72–74.

Baeza-Yates, R.; Becchetti, L.; Boldi, P.; Castillo, C.; Donato, D., S., L.; Poblete, B. (2006). Link Analysis on the Web. http://www.slideshare.net/ChaToX/link-analysis-123800/.

Baeza-Yates, R.; Castillo, C.; Efthimiadis, E. (2007). Characterization of national web domains. // ACM Transactions on Internet Technology, 7:2 (2007).

Baeza-Yates, R.; Castillo, C.; López, V. (2005). Characteristics of the Web of Spain. // Cybermetrics, 9:1 (2005).

Baeza-Yates, R.; Castillo, C.; López, V. (2006). Características de la web de españa. // El Profesional de la Información, 15:1 (2006).

Ball, P. (2000). The art of networking. // Nature Science Update. (Oct 25, 2006).

Barabási, A.-L. (2002). Linked: the new science of networks. Perseus Publishing, 2002.

Barabási, A. L.; Albert, R. (1999). Emergence of scaling in random networks. // Science. 286:5439 (1999), 509–512.

Barabási, A. L.; Albert, R.; Jeonh, H. (2000). Scale-free characteristics of random networks: the topologu of the world wide web. // Physica. 281:1–4 (2000) 69–77.

Bauwens, M. (1996). Knowledge transfer in cyberspace: A model for future business practices. // FID News Bulletin, 46:1/2 (1996) 46–54.

Becchetti, L.; Castillo, C.; Donato, D.; Baeza-Yates, R.; Leonardi, S. (2008). Link analysis for web spam detection. // ACM Trans. Web, 2:1 (2008) 1–42.

Björneborn, L. (2004). Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach. Ph.d. thesis from the department of information studies, Royal School of Library and Information Science, 2004.

Björneborn, L.; Ingwersen, P. (2004). Toward a basic framework for webometrics. // Journal of the American Society for Information Science and Technology. 55:14 (2004) 1216–1227.

Bossy, M. J. (1995). The last of the litter: 'netometics'. // Solaris. 2 (1995).

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. (2000). Graph structure in the web: Experiments and models. // Proceedings of the Ninth Conference on World Wide Web. Amsterdam, Netherlands: ACM Press, 2000. 309–320

Castillo, C.; Donato, D.; Becchetti, Luca, B.; Paolo, L.; Stefano, Santini, M.; Vigna, S. (2006). A reference collection for web spam. // SIGIR Forum. 40:2 (2006).

Chakrabarti, S.; Joshi, M.; Punera, K.; Pennock, D. (2002). The structure of broad topics on the web. // Proceedings of the WWW2002 Conference.

Clausen, H. (1996). Looking for the information neddle in the internet haystack. // Proccedings of the 20th Online Information Meeting. Oxford: Learned Information, 1996. 115–123.

Cordón García, J. A.; Benito Martín, F.; Pinto Molina, M.; Alonso Berrocal, J. L.; Fernández López, R.; Valentín Centeno, A. (2007). Informe sobre los servicios de Publicaciones de las Universidades andaluzas, chapter Análisis cibermétrico de los servicios de publicaciones. Dirección General de de Universidades de la Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, 2007. 231–279.

Dahal, T. M. (1999). Cybermetrics: The use and implications for scientometrics and bibliometrics; a study for developing science & technology information system in nepal. // IIIrd. National Conference on Science & Technology. Royal Nepal Academy of Science and Technology (RONAST).

Dominich, S.; Skrop, A. (2005). Pagerank and interaction information retrieval. // Journal of the American Society for Information Science and Technology. 56:1 (2005) 63–69.

Faloutsos, M.; Faloutsos, P.; Faloutsos, C. (1999). On power-law relationships of the internet topology. // Proceedings of the ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, August 30 - September 3, 1999, Cambridge, MA, USA. ACM.

Gibson, W. (1984). Neuromancer. New York: Ace Books, 1984.

Gyongyi, Z.; Garcia-Molina, H. (2005). Web spam taxonomy. // First International Workshop on Adversarial Information Retrieval on the Web.

Haynes, C. (1995). How to succeed in Cyberspace. London: Aslib, 1995.

Ingwersen, P. (1998). The calculation of web impact factors. // Journal of Documentation. 54:2 (1998) 236–243.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. // Journal of the ACM. (1999) 668–677.

McMurdo, G. (1996). Net by numbers. // Journal of Information Science. 22:5 (1996) 381–390.

Metaxas, P. T.; DeStefano, J. (2005). Web spam, propaganda and trust. // AIRWeb2005, number 5.

Ortega Prieto, J. L. (2007). Visualización de la Web Universitaria Europea: Análisis cuantitativo de enlaces a través de técnicas cibermétricas. PhD thesis, Universidad Carlos III de Madrid, Departamento de Biblioteconomía y Documentación.

Page, L.; Brin, S.; Motwani, R.; Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Pinto Molina, M.; Alonso Berrocal, J. L.; Cordón García, J. A.; Fernández Marcial, V.; García-Figuerola Paniagua, C.; Javier, G. M.; Gómez Camarero, C.; Zazo Rodríguez, N. F. (2003). Visibilidad de la investigación de las universidades españolas a través de sus páginas web en el ámbito del Espacio Europeo de Enseñanza Superior: análisis, evaluación y mejora de la calidad. Programa de estudio y análisis para la mejora de la calidad de la enseñanza superior y profesorado universitario. http://wwwn.mec.es/univ/html/informes/estudios_analisis/resultados_2003/EA2003-0012/VISIWEB.pdf (2008-05-27).

Pirolli, P. ; Pitkow, J.; Rao, R. (1996). Silk from a sow's ear: extracting usable structures from the web. // CHI 86 Electronic Proceedings.

Shiri, A. A. (1998). Cybermetrics; a new horizon in information research. // 49th FID Conference and Congress, New Delhi, India, 11-17 october.

Sung Jin, K. & Sang Ho, L. (2002). An improved computation of the pagerank algorithm. // ECIR. volume 2291 of Lecture Notes in Computer Science. Springer. 73–85ISBN: 3-540-43343-0.

Tauro, L.; Palmer, C.; Siganos, G.; Faloutsos, M. (2001). A simple conceptual model for the internet topology. In IEEE Global Internet, San Antonio, Texas, USA. IEEE CS Press.

Thelwall, M. (2007). Bibliometrics to webometrics. // Journal of Information Science, 34:4 (2007) 1–18.

Tolosa, G., Bordignon, F., Baeza-Yates, R., & Castillo, C. (2007). Characterization of the argentinian web. // Cybermetrics, 11(1), 3+.