La cibermetría en la recuperación de información en el Web

José L. Berrocal, Carlos G. Figuerola, Ángel F. Zazo y **Emilio Rodríguez**
Grupo de Recuperación de Información
Departamento de Informática y Automática
Facultad de Documentación. Universidad de Salamanca
C/ Francisco Vitoria, 6-16
37008. SALAMANCA - SPAIN
{berrocal–figue–afzazo–aldana}@usal.es

Resumen El crecimiento exponencial del web, así como sus características de datos distribuidos, alta volatilidad, datos sin estructurar, redundantes y muy heterogéneos, han introducido nuevos problemas en los procesos de recuperación de información. Por ello es necesario abrir nuevas vías de investigación que nos permitan obtener unos buenos niveles de precisión. Los trabajos que se basan en la explotación de las características hipertexto del web están alcanzando gran notoriedad. La cibermetría está aportando muchas opciones de trabajo con los enlaces y está ofreciendo en este momento interesantes opciones, y buena parte de las técnicas empleadas en la misma pueden ser útiles en los procesos de recuperación de información en el web.

1. Introducción

El estudio del World Wide Web (WWW) se está convirtiendo en uno de los campos de investigación más interesantes y como dice [23] pocos eventos de la historia de la computación han tenido tanta influencia en la sociedad como la llegada y crecimiento del Web. Precisamente este crecimiento (2000 millones de páginas Web) y su influencia (basada en los contenidos) han creado un sistema de comunicación de información muy potente, pero que al mismo tiempo tiene enormes carencias desde el punto de vista documental. Por ello es necesario abordar su estudio.

Para algunos autores este estudio debe realizarse con las técnicas bibliométricas clásicas y de análisis de citas, sin embargo es necesario realizar otros estudios y abrir nuevas vías de investigación que nos permitan caracterizar adecuadamente el Web, porque no hay que olvidar, por ejemplo, que el tipo de información con el que estamos trabajando tiene unos niveles de permanencia [25] concretos que nos obliga a ajustar nuestras técnicas de estudio.

2. La Cibermetría

2.1. Estudios previos a la cibermetría

Una de las investigaciones principales realizadas sobre la métrica del cibermedio es la de Almind y de Ingwersen [2]. Procuraron introducir la aplicación de métodos informétricos al Web denominándolo "Webmetría". Realizaron un estudio comparando la proporción Danesa del WWW a la de otros países nórdicos. La metodología usada se basaba en el análisis bibliométrico. Este estudio también ha explorado el número medio de hiperenlaces por página Web y la densidad de enlaces para los diferentes tipos de dominio [2]. Este estudio webmétrico, 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.

Algunos trabajos emplean 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 números reales, que indicaran la fuerza de la conexión. Se indica la necesidad de emplear matrices, aunque en ningún momento hace referencia a la teoría de grafos.

En 1997 una investigación, en la Escuela Real de Bibliotecarios de Dinamarca, dirigido a explorar por estudios cuantitativos ciertos fenómenos y acontecimientos actuales de la información tuvo como objetivo el análisis de la creación, uso y del estudio de las *homepages* Danesas/Nórdicas. Este estudio también se ha referido a "Internetmetría" (Informetría) y parece estar más orientado a la información que las investigaciones anteriores.

Una investigación sobre el factor de impacto del Web [22] nos 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 dicho factor 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 [14] 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 en un área significativa para la investigación.

2.2. Definición del término

El incremento en la transición de los materiales impresos a los recursos electrónicos y a recursos de red ha originado a su alrededor nuevas perspectivas para estudiar las fuentes, los servicios y los medios de información. Es decir, si queremos estar enterados de qué información aparece en nuestro entorno, el análisis cuantitativo y el estudio de los fenómenos que operan dentro de este entorno es tan importante como ha sido el estudio cuantitativo de las características de los materiales impresos en el pasado. A través de estos estudios podemos hacer una estimación de qué se conoce como información electrónica y evaluar las características de tal información. En el desarrollo de estos estudios apareció el término "Cibermetría", que fue acuñado por Shiri [32], y lo debemos entender como la medida, el estudio, y el análisis cuantitativo de todas las clases de información y de los medios de información que existen y que funcionan dentro del ciberespacio, empleando las técnicas bibliométricas, cienciométricas e informétricas. 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, se han transformado 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 pueden 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.

Para llevar a cabo estas investigaciones es preciso trabajar con agentes inteligentes, robots del conocimiento, así como con motores de búsqueda del Web, que son herramientas eficaces para extraer apropiadamente la información relevante y que nos va a permitir automatizar todo el proceso, con el fin de poder dar respuesta adecuada al incremento exponencial de la información.

3. Posibilidades de estudio

El estudio del web puede realizarse básicamente desde tres puntos de vista:

- Análisis cuantitativo.
- Medidas topológicas.
- Leyes de exponenciación.

Dentro de cada una de estas vías de investigación, con una gran variedad de cálculos posibles, parte de los estudios se pueden aplicar a recuperación de información. Nuestro grupo de investigación está desde hace algunos años trabajando en estas posibles vías de aplicación a recuperación de información como indicaremos posteriormente.

3.1. Análisis cuantitativo

Autores como [1] constataron la necesidad de aplicar nuevas medidas e interpretaciones en los intentos de medir e interpretar la estructura, tamaño y conectividad del Web, en constante evolución y con una alta volatilidad [25]. Los primeros trabajos cuantitativos relacionados con el Web trataron fundamentalmente de estudiar la evolución del tamaño y la descripción de los primeros motores de búsqueda, tratando de conocer la entidad y cobertura de dichos motores. De esta época son los trabajos de [13, 28]. Más recientemente se han empleado los buscadores en una investigación que trataba de comprobar si el Web podía resultar una buena fuente de datos para la investigación. Otros estudios utilizaron los datos recogidos por algunos motores para realizar un estudio de estas características, como fueron los trabajos de [34], basado en la recogida de datos del motor Inktomi, y de [8], con el motor Open Text. Este último autor realizaba una buena clasificación de los diferentes aspectos que se deberían tratar, además de indicar la importancia de las sedes, e incluía alguna de la terminología que posteriormente otros autores utilizan o redefinen y mejoran. Sin embargo, en estos primeros trabajos no se tenían en cuenta ni la cobertura global del Web ni su naturaleza hipertextual, que evidentemente han modificado los planteamientos en los trabajos posteriores. Uno de los primeros trabajos que tiene en cuenta la naturaleza hipertextual del Web y que permitió la aplicación de técnicas bibliométricas fue el trabajo de [27]. Los trabajos de [21] sobre el estudio de los enlaces establecidos entre revistas convencionales y las revistas electrónicas supuso un nuevo planteamiento en la aplicación de estas técnicas. El concepto de "sitation", de McKiernan, introdujo un nuevo elemento previo al desarrollo de diferentes indicadores cibermétricos. La variedad de cálculos que pueden realizarse en un estudio de tipo cuantitativo es enorme, pero de posible aplicación a recuperación de información. Posteriormente indicaremos algunas de las vías de trabajo.

3.2. Medidas topológicas

Una buena forma de analizar la evolución de los dominios Web, que se centra fundamentalmente en su naturaleza hipertextual, es calcular un conjunto de medidas que tengan en cuenta los enlaces que se producen entre los diferentes documentos que conforman el dominio correspondiente como justifica [15]. Otros estudios posteriores se han centrado en el estudio de la conectividad y la estructura topológica del Web como los de [23] basándose en los trabajos anteriormente mencionados. Incluso [23] relaciona este tipo de trabajos con los análisis de citas y con el factor de impacto. Estas medidas se basan en la consideración del Web como un grafo [15, 24] y la aplicación de diferentes técnicas propias de esta teoría. En la consideración del Web como un grafo, los nodos se representan mediante las páginas HTML y los enlaces se representan mediante los bordes dirigidos. Diferentes estudios [6] sugieren la existencia de varios cientos de millones de nodos en el grafo Web (con un crecimiento importante), y el número de enlaces alcanzaría varios billones [24]. Algunos de los trabajos que han manejado el Web como un grafo han utilizado un volumen de información realmente importante con 200 millones de páginas y 1,5 billones de enlaces mostrando la consistencia de los planteamientos y con la aplicación de algoritmos adecuados para el tratamiento de esta gran cantidad de información [24]. El análisis de la estructura del grafo Web se ha empleado en ocasiones para mejorar la calidad de las búsquedas en el Web como en [9, 23]. También se ha utilizado para clasificación de páginas Web en función de las materias de las páginas a las que apunta una página concreta como en [10], para mostrar la información [7] o en minería del Web [26]. Algunos autores han utilizado el Web como grafo para crear de forma automática hipertextos, partiendo de textos carentes de enlaces [33]. La aplicación de la teoría de grafos al web puede realizarse mediante índices de nodo e índices de grafo, siendo de interés para recuperación de información los índices de nodo, pero sin olvidarnos de los índices de grafo, que pueden ofrecer vías interesantes de estudio como indicaremos posteriormente.

La estructura de enlaces del Web contiene también información sobre las diferentes comunidades Web que se pueden crear y que se reflejan mediante la topología del Web como apunta [19] y también permite aplicar técnicas de similaridad, basadas en los enlaces, para estructurar y visualizar el Web [11].

3.3. Las leyes de exponenciación

Estas leyes, de reciente estudio en el web, tratan de analizar las pautas y mecanismos de crecimiento que gobiernan el web.Una primera aproximación del análisis de dichas leyes puede encontrarse en [16, 29].

4. Recuperación de información

El principal problema con el que deben enfrentarse los sistemas de Recuperación de Información es el de tener que trabajar con información no estructurada (al menos de una forma explícita). De hecho, el fundamento de los diferentes modelos teóricos que se han planteado, y de sus correspondientes implementaciones operativas, consiste en la aplicación de algún formalismo que permita representar adecuadamente cada uno de los documentos almacenados en la base de datos, así como las consultas que puedan generar los usuarios de la misma. La resolución de una consulta requiere la computación de alguna función de similitud que permita establecer el grado de adecuación entre una consulta y cada uno de los documentos [31]. Naturalmente, la efectividad en la resolución de las consultas depende directamente de la bondad del formalismo empleado para representar los documentos.

El uso de términos como elementos básicos de la representación de un documento se ha demostrado eficaz, pero plantea algunos problemas que con la tecnología actual no están bien resueltos. Entre ellos, el de la normalización de dichos términos, es decir, la reducción a una forma común de las distintas variantes (tanto flexivas como derivativas) que puedan aparecer en los documentos [20]. Pero además, en el caso de entornos multilingües (de lo que es un buen ejemplo el Web), tenemos la cuestión de la conversión de términos en una lengua determinada a sus equivalentes correctos (en función del contexto) en otra u otras lenguas. Por estas razones, ha habido diversos intentos de representar documentos atendiendo a otros aspectos.

Un caso notable es el aplicado desde diversas instancias a la literatura científica. Los trabajos científicos se caracterizan, entre otras muchas cosas, por ir acompañados de un aparato bibliográfico más o menos importante: cualquier artículo científico contiene varias citas o referencias, con la intención de indicar al lector fuentes adicionales de conocimiento, o para apoyar las propias tesis en los trabajos o descubrimientos publicados en otros lugares. Así, cuando operamos con colecciones documentales constituidas por artículos científicos es planteable representar dichos documentos a través de las referencias que contienen a otros artículos. Dicho de una forma simple: si dos artículos contienen las mismas citas o referencias, deben ser muy similares en cuanto a contenidos y temas que traten. Así pues, el grado de coincidencia en referencias o citas puede utilizarse para calibrar la semejanza entre dos artículos científicos. De esta forma, dado un artículo como punto de partida, es posible obtener aquéllos dentro de la colección que son parecidos en cuanto a temática o contenido.

Este tipo de planteamientos podría ser extrapolado al Web, considerado éste como una colección de documentos. Las páginas web poseen una característica que las hace especiales (prescindiendo de imágenes, sonido, elementos de captación de datos -formularios- y otras maravillas): las páginas web tienen hipervínculos o enlaces con otras páginas o recursos en la red. A partir de esos enlaces el espacio Web puede ser considerado como un grafo dirigido, en el cual los nodos serían las diferentes páginas existentes y los arcos los hipervínculos que enlazan un nodo con otro [15]. Consiguientemente, y dado que un hipervínculo se activa en un nodo determinado y nos dirige hacia otro nodo concreto, debemos distinguir entre enlaces entrantes y salientes. De esta forma, haciendo abstracción del contenido interno de cada nodo (página web, documento), podríamos definir cada uno de ellos en función de su situación en el grafo, es decir, sobre la base de los enlaces que mantiene hacia otros nodos y los que otros nodos mantienen con él.

Se trataría, entonces, de aplicar los mismos planteamientos indicados para la literatura científica, asumiendo para los enlaces de una página el papel de las referencias en los artículos científicos. Así, podríamos asumir que si dos páginas apuntan o enlazan a los mismos sitios, deben ser más o menos similares en cuanto a sus contenidos. Igualmente, si dos páginas son apuntadas desde los mismos lugares, sus contenidos deben guardar una relación más o menos estrecha. Este enfoque ha sido planteado en varios trabajos, entre los cuales cabe destacar los de [11, 6]. De hecho, estos trabajos, al menos como punto de partida, toman la metodología y los algoritmos

del análisis de citas. Algunas aplicaciones de lo expuesto anteriormente se pueden ver en algunos de los trabajos ya desarrollados por nuestro grupo de investigación [3, 4, 5, 17].

Otra aproximación al problema consiste en la utilización de agentes que rastreen la red [35]. Así, podría plantearse el empleo de agentes, al que se le formularían las necesidades informativas como parte de las especificaciones iniciales; éste exploraría la red, eligiendo los enlaces más prometedores, accediendo a nuevas páginas, recopilando las que pudiesen satisfacer las especificaciones iniciales, y así sucesivamente. Puesto que la propia exploración del Web, aún automática, requiere grandes cantidades de tiempo, un enfoque de este tipo tiene de entrada algunas limitaciones. No es esperable una respuesta inmediata, ni siquiera probablemente con la agilidad suficiente para plantear una dinámica especialmente interactiva con el usuario. Antes bien, y muy en la línea de lo que se entiende por agentes inteligentes, de alguna forma el usuario delega en el agente, después de haberle facilitado algunas instrucciones (por ejemplo, indicándole qué clase de información se desea). Se deja al agente hacer su trabajo de forma autónoma y tomándose su tiempo, en espera de que en un plazo razonable (el propio usuario podría establecer plazos máximos) entregue el resultado de su trabajo, esto es, las páginas web relevantes encontradas. La otra limitación importante de este enfoque es la renuncia implícita a la exhaustividad. Dado el tamaño del Web, parece claro que la exploración completa, o incluso de una parte significativa de él, resulta implanteable; antes al contrario, agentes de este tipo trabajando para usuarios individuales o personales, por ejemplo, explorarían tan sólo una pequeña parte del Web. Se espera, en contrapartida, que los resultados obtenidos alcancen una notable precisión. Esta clase de agentes permitirían obviar el efecto de sobrecarga de información. En este proceso de automatización hay algunos aspectos importantes a tener en cuenta:

4.1. La elección de los puntos de partida

Puesto que un agente de este tipo debe explorar gran cantidad de páginas, es preciso determinar algún punto de partida. Como la distancia entre el nodo por el que se empieza a explorar y cualquiera de los nodos relevantes puede ser muy grande, es crítico localizar previamente nodos o puntos de partida que puedan estar lo más cercanos posible a nodos o páginas relevantes. La distancia a recorrer (el número de nodos por los que hay que pasar) no sólo depende del tamaño del Web, sino que incluso podemos encontrar nodos con vías muertas que se extinguen sin permitir proseguir con la exploración.

Un enfoque utilizado frecuentemente para elegir buenos puntos de partida es comenzar el trabajo del agente con una búsqueda al estilo clásico en las bases de datos de diferentes buscadores convencionales. En estos casos tales búsquedas previas suelen enviarse a servicios metabuscadores, los cuales tratan con los diferentes buscadores, recogen los resultados de cada uno de ellos, los organizan y los devuelven a quien hizo la consulta. En este caso sería el propio agente quien enviaría la consulta a esos metabuscadores, recogiendo las páginas devueltas por éstos. Tales páginas son las candidatas a ser puntos de entrada o de comienzo de exploración. Dichos puntos de entrada pueden manejarse de forma secuencial, empezando la exploración por uno de ellos, hasta una determinado distancia prefijada de antemano, o en paralelo, utilizando varios agentes para ello. En este caso los agentes deben hacer uso de sus capacidades cooperativas, no sólo para compartir criterios de selección de páginas relevantes, sino también para evitar exploraciones de los mismos nodos. La exploración de la red con varios agentes tomando diferentes puntos de entrada ofrece el atractivo de permitir utilizar procesamiento paralelo o varios ordenadores para el proceso, pero incluso sin ello, presenta la ventaja de obviar en alguna medida problemas derivados de las comunicaciones, como cuellos de botella, líneas o servidores lentos, etc., redundando en una mejora en el tiempo de respuesta.

De otro lado, el hecho de disponer de varios puntos de entrada puede implicar la selección de parte de ellos (en un número razonable), así como posiblemente la priorización. Hay diversas estrategias automáticas para abordar esta cuestión, desde tomar simplemente los n primeros, hasta aplicar medidas de similitud entre las especificaciones del usuario y el contenido de las páginas, pasando por el análisis de aspectos como el número de enlaces de cada punto de entrada, o incluso prospecciones de tiempos de respuesta. Del mis-

mo modo, es posible una realimentación por parte del usuario, dejando que sea éste quien seleccione los que estime como mejores puntos de entrada. Naturalmente, estos diversos enfoques son combinables entre sí.

4.2. Activación de enlaces

Dado un punto o página de partida, un agente que pretenda explorar el Web debe extraer los enlaces (direcciones URL) que esa página contenga y guardarlos en una lista. Posteriormente, irá tomando enlaces de esa lista, recuperando las páginas a las que apuntan y así sucesivamente. Si la exploración se ha de llevar a cabo por varios agentes de forma cooperativa, esa lista debería ser compartida en alguna forma, a fin de no duplicar exploraciones de los mismos nodos. El almacenamiento y posterior seguimiento de todos los enlaces en la lista llevaría, teóricamente, a la exploración de todo el Web. Sin embargo, como suele tenerse limitaciones de recursos de almacenamiento, capacidad de proceso o comunicaciones, etc., y especialmente de tiempo, se hace preciso establecer un orden de prioridad para los elementos de la lista.

Este orden atiende a dos premisas fundamentales: en primer lugar, la relevancia de los enlaces (o su presunción) respecto de las necesidades informativas del usuario. En segundo lugar, las posibilidades de acceder a mayores espacios del Web desde unos enlaces que desde otros. Empezando por este último aspecto, se han propuesto diversos sistemas para seleccionar aquellos enlaces más prometedores desde ese punto de vista. Para determinar la importancia de una página, una posibilidad consiste en utilizar los backlinks de la misma, esto es, las páginas que tienen enlaces hacia la página en cuestión [12]. El mecanismo más simple es contar el número de backlinks, pero el problema es disponer de dicha información.

Más sofisticado que el simple recuento de backlinks es el algoritmo conocido como PageRank [30]. La idea básica es que la importancia de un nodo o página es directamente proporcional al número de backlinks que éste tiene, pero no todos los backlinks pesan lo mismo, sino que su valor está en función de la importancia de la página de la que procedan. Y la página de procedencia tiene, a su vez, una importancia que viene determinada por los backlinks que recibe, y así sucesivamente. Según este algoritmo, el cálculo del PageRank ha de hacerse de forma iterativa, asignando de antemano pesos a determinados nodos o páginas, ya sea de forma aleatoria o en función de algún otro criterio, y asumiendo que, de una forma u otra, en algún momento de la computación se llega a esos nodos. Se trata de una visión muy genérica, en la que hay que resolver otros detalles, pero lo que importa resaltar aquí es que se trata de un cálculo costoso en términos de tiempo de proceso. Éste es el mismo problema que encontramos para calcular otro tipo de coeficientes, cuya finalidad es también estimar la importancia de unos determinados nodos frente a otros [15]. Parece que tales índices no son aplicables en una exploración directa del Web, aunque algunos buscadores basados en búsquedas en bases de datos de páginas web previamente recopiladas los utilizan para ordenar los resultados obtenidos en una búsqueda de este tipo.

4.3. Selección de páginas por contenido

Más allá de la mayor o menor importancia de una página (en el sentido de la mayor o menor facilidad de exploración del Web a partir de la misma), lo que realmente nos interesa es disponer de medios para estimar la proximidad de un nodo a las necesidades informativas del usuario. Esto debe permitir, naturalmente, seleccionar páginas para que el agente las entregue al usuario como resultado. Pero también, en conjunción con la estimación de importancia vista antes, para determinar cuáles son los enlaces más prometedores para proseguir la exploración. En esta línea, diversos mecanismos pueden ser utilizados, y muchos de ellos pueden combinarse o compaginarse entre sí, como las técnicas de recuperación de información basadas en el empleo del llamado modelo vectorial, el cual es también representativo de las limitaciones que presenta la aplicación de estos métodos.

5. Nuestra investigación actual

Nuestra investigación actual, una vez que hemos pasado por algunas de las fases previas de conocimiento del medio [3, 4, 5, 17] pasa por calcular determinados índices que nos puedan ayudar en la fase de elección de puntos de partida y en la activación de enlaces. Dentro del análisis cuantitativo, se están valorando la densidad hipertextual, el índice de desarrollo hipertextual, índice de endogamia,

factor de impacto web y visibilidad. Mención aparte merece el análisis de citas ampliamente utilizado en recuperación de información y en algunos casos como parte de un nuevo algoritmo para el desarrollo de nuevos buscadores basados en el análisis de los enlaces [18].

El análisis topológico nos ofrece una gran variedad de índices que pueden ser utilizados, a priori, para nuestro planteamiento de investigación. En principio podríamos pensar que los índices de nodo serían los más adecuados en el proceso de recuperación de información y efectivamente gran parte de los autores han planteado de forma teórica la posibilidad de emplearlos en recuperación de información. Destacaríamos índices como el grado de apertura, ROC, así como los índices de Dice y del coseno. Sin embargo no hay que olvidar algunos de los índices de grafo, que aplicados a una determinada distancia de un nodo origen, nos pueden permitir tomar decisiones sobre el camino más adecuado, a priori. Entre estas medidas podemos destacar la Compactación que indica si los nodos que forman parte del grafo se pueden alcanzar o enlazar fácilmente, sugiriendo un amplio número de referencias cruzadas o enlaces entre los nodos.

6. Conclusión

Las técnicas cibermétricas son una opción interesante para ofrecer soluciones a la recuperación de información en el web. Es necesario perfilar y ajustar los procedimientos, pero después de un largo proceso de desarrollo del agente, de la creación de algoritmos, y del estudio de las diferentes técnicas cibermétricas en la caracterización del web, estamos en disposición de comenzar a aplicarlo a la recuperación de información. Si el análisis del web, mediante algunos de los índices cibermétricos enumerados, nos permite reducir el número de nodos que es preciso recorrer, para poder obtener altos niveles de precisión, estaremos dando un paso muy importante, teniendo en cuenta que estamos trabajando con la información hipertextual de los documentos y por otro lado ofreceríamos al mismo tiempo un mecanismo de recuperación multilingüe.

Referencias

[1] L. A. Adamic. The Small World Web. In *Proceedings of ECDL'99*, 443-452.

[2] T. C. Almind y P. Ingwersen. September 1997. Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation,* 53 (4): 404-426.

[3] J. L. Alonso Berrocal, C. G. Figuerola y A. F. Zazo Rodríguez. 1999. Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información. *Scire*, 5 (2):91-98.

[4] J. L. Alonso Berrocal, C. G. Figuerola y A. F. Zazo Rodríguez. 2001. Cibermetría del web: las leyes de exponenciación. *Revista General de Información y Documentación*, 11(1):209-217.

[5] J. L. Alonso Berrocal. 2001. CIBERMETRÍA: Análisis de los dominios web españoles. *Tesis Doctoral. Usal. Dpto. de Informática. Director Dr. C. G. Figuerola*, Junio.

[6] K. Bharat y A. Broder. A tecnique for measuring the relative size and overlap of public Web search engines. In *Proc. of the Seventh WWW Conference*, (Brisbane, Australia, 1998).

[7] R. A. Botafogo y B. Shneiderman. Identifying aggregates in Hypertext structures. In *Proceedings of Hypertext'91*, (Diciembre de 1991), 63-74.

[8] T. Bray. Measuring the Web. In *Fifth International World Wide Web Conference*, (Paris, France, 6-10 May 1996).

[9] S. Chakrabarti y otros. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th International World Wide Web Conference*, (1998).

[10] S. Chakrabarti y B. I. P. Dom. Enhanced hypertext categorization using hyperlinks. In *Proceedings ACM SIGMOD*, (1998).

[11] S. Chakrabarti y otros. August 1999. Mining the link structure of the World Wide Web. *IEEE Computer,*.

[12] J. Cho, H. García-Molina, y L. Page. 1998. Efficient crawling through url ordering. *Computer Networks and ISDN Systems*, 30: 161-172.

[13] S. J. Clarke y P. Willett. July 1997-August 1997. Estimating the recall performance of Web search engines. *Aslib Proceedings,* 49 (7): 184-189.

[14] T. M. Dahal. Cybermetics: The use and implications for Scientometrics and Bibliometrics; A study for Developing Science & Technology Information System in Nepal. In *IIIrd National Conference on Science & Technology*, (March 8-11, 1999. Royal Nepal Academy of Science and Technology (RONAST)).

[15] D. Ellis, J. Furner-Hines y P. Willett. June 1994. On the creation of hypertext links in full-text documents: measurement of inter-linker consistency. *Journal of Documentation,* 50 (2): 67-98.

[16] M. Faloutsos, P. Faloutsos y C. Faloutsos. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM*, Cambridge, MA, September:251-262.

[17] C. G. Figuerola, J. L. Alonso Berrocal y A. F. Zazo Rodríguez. 2000. El contenido semántico de los enlaces de las páginas web desde el punto de vista de la recuperación de información. IN *I jornada de terminologia i documentació*, Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

[18] G. Flake, S. Lawrence, L. Giles y F. Coetzee. 2002. Self-Organization of the web and identification of communities. In *IEEE Computer*, 35 (3): 66-71

[19] D. Gibson, J. Kleinberg y P. Raghavan. 1998. Inferring web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia.*

[20] R. Gómez Díaz. 1998. La recuperación de la información en español: evaluación del efecto de sus peculiaridades lingüísticas. *Trabajo de Grado. Universidad de Salamanca. Director Dr. C. G. Figuerola.*

[21] S. P. Harter. 1996. The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis. *Public Access Computer Systems Review,* 7 (5).

[22] P. Ingwersen. March 1998. The calculation of web impact factors. *Journal of Documentation,* 54 (2): 236-243.

[23] J. M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM,* 668-677.

[24] J. M. Kleinberg, R. Kumar y P. Raghavan. The web as a graph: measurements, models, and methods. In *Proceedings of the Fifth Annual International Computing and Combinatorics Conference*, (1999).

[25] W. C. Koehler. 1999. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science,* 50 (2): 162-180.

[26] R. Kumar, P. Raghavan, S. Rajagopalan y A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th VLDB Conference*, (Edinburgh, 1999).

[27] R. R. Larson. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. In *Annual meeting of the American Society for Information Science*, (Baltimore, October 19-24, 1996), 71-78.

[28] M. Mauldin y J. Leavitt. Web agent related reserach at the Center for Machine Translation. In *Reunión del ACM Special Interest Gropu on Networked Information Discovery and Retrieval*, (McLean, VA, USA, 4 de Agosto de 1994).

[29] A. Medina, I. Matta y J. Byers. 2000. On the origin of power laws in internet topologies. In *Computer Communications review*, 30 (2).

[30] L. Page y otros. 1998. The pagerank citation ranking: Bringing order to the web. Technical report. URL: citeseer.nj.nec.com/page98pagerank.html

[31] G. Salton. 1987. On the relationships between theoretical retrieval models. *Informetrics,* 87/88 263-270.

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

[33] A. F. Smeaton. 1992. Information retrieval and hypertext: competing technologies or complementary access methods. *Journal of Information Systems,* 2 221-233.

[34] A. Woodruff. An Investigation of Documents from the World Wide Web. In *Fifth International World Wide Web Conference*, (París, May 6-10 1996).

[35] N. R. Wooldridge y M. Jennings. 1995. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10 (2): 115-152.