LA RECUPERACIÓN DE INFORMACIÓN EN EL WEB: RETOS Y ¿SOLUCIONES?

Berrocal, J.L.; Zazo, Ángel F.; Figuerola, Carlos G.; Rodríguez, Emilio
Grupo de Recuperación Automatizada dde la Información - REINA
Dpto. de Informática y Automática. Universidad de Salamanca.
Fac. Documentación. C/ Francisco Vitoria 6-16. 37008 Salamanca
[berrocal|afzazo|figue|aldana]@usal.es
http://reina.usal.es

RESUMEN

Los sistemas de recuperación de información clásicos se han encontrado con problemas a la hora de ser implementados en la información del web. Las particularidades de esta información están obligando a diseñar nuevos mecanismos que permitan unos niveles de precisión mucho más elevados y que posibiliten que el usuario obtenga lo que realmente necesita. Ante los nuevos retos aparecidos, nuestro grupo de investigación REINA está trabajando en las posibles soluciones. Se analizarán algunas de las teorías de mejora de la recuperación de información y se presentará la herramienta Sacarino bot como posible software que facilite esta tarea.

Palabras clave: Recuperación de información, internet.

1. LA RECUPERACIÓN DE INFORMACIÓN EN EL WEB

A finales de la década de los ochenta la interconexión de miles de redes de área local había convertido Internet en el mayor almacén de datos que jamás hubiese existido, pero también en el más caótico. Las posibilidades eran enormes, pero las dificultades resultaban frustrantes: formatos incompatibles, programas distintos, protocolos heterogéneos, etc. Se imponía pues la necesidad de simplificar el acceso a este caudal de información, hacerlo más

sencillo y homogéneo. En las figuras 1, 2 y 3 podemos ver el incremento claramente exponencial de la red.

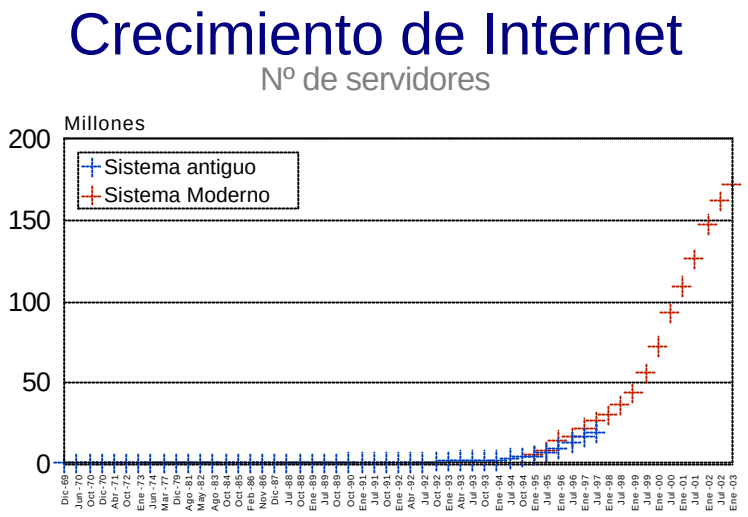


Figura 1 – Fuente: Hobbe's Internet Timeline. Robert H. Zakon
http://www.zakon.org/robert/internet/timeline/

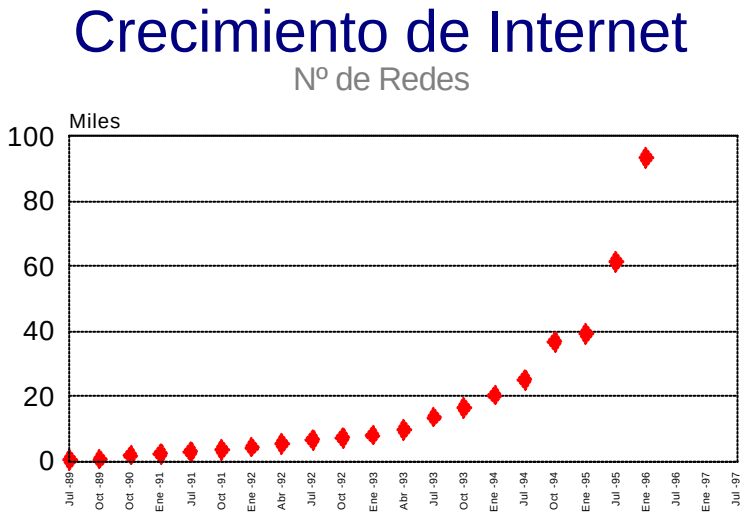


Figura 2 – Fuente: Hobbe's Internet Timeline. Robert H. Zakon
http://www.zakon.org/robert/internet/timeline/

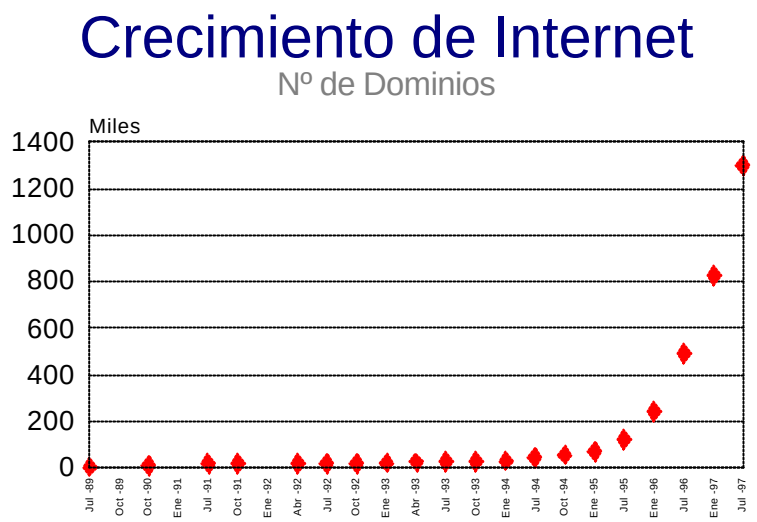


Figura 3 – Fuente: Hobbe's Internet Timeline. Robert H. Zakon
http://www.zakon.org/robert/internet/timeline/

WAIS, desarrollado a partir de 1989 por un grupo de empresas [1] sólo fue una solución parcial: los datos debían indexarse con el nuevo software y distribuirse por medio de un nuevo protocolo, es decir, había que realizar un trabajo de adaptación de lo ya existente al nuevo sistema. El Gopher de la Universidad de Minnesota [2], ampliamente difundido desde 1991, aportó algo más: por medio de un sistema simple de ventanas (o de menús) se accede a todo tipo de archivos de texto, imágenes, bases de datos, etc., sin tener que preocuparse por su localización física en la red, el formato o el protocolo de recuperación: ftp y wais, por ejemplo, son protocolos que el gopher maneja desde el principio, además del suyo propio. Un interface unificado para el acceso a información distribuida: este ha sido el objetivo del gopher, y también el del Web.

El proyecto World-Wide Web del CERN (*Centre Européenne pour la Recherche Nucléaire)* supuso otra vuelta de tuerca en el intento de poner, efectivamente, al alcance de los usuarios el espacio virtual de conocimiento que es Internet.

La experiencia de la proliferación del conocimiento y de la angustia derivada de no poder abarcarlo todo no es nueva, no ha surgido con los ordenadores y la conectividad. Ya en 1945 Vannevar Bush [3] se lamentaba:

> La suma de la experiencia humana se está expandiendo a un ritmo prodigioso y los medios que utilizamos para seguir el hilo a

> través del consiguiente laberinto de ítems momentáneamente importantes son los mismos que usábamos en los días de los barcos de vela. [3].

En su opinión el problema no era tanto una cantidad excesiva de publicaciones como el nulo avance de las tecnologías con que se gestionaba su manejo. Con los rudimentos tecnológicos de su época en mente, Bush fue capaz de idear un sistema llamado memex que permitiría archivar el conocimiento de un modo más eficaz: una especie de escritorio futurista en el que se guardarían, microfilmados, los libros, actas, ficheros, etc. Cada elemento de información se visualizaría en pantalla tecleando su código mnemotécnico correspondiente y, esto es lo más importante, podríamos registrar las conexiones observadas entre elementos distintos. Un usuario del memex que contase con una buena base de datos podría anotar conexiones entre, digamos, un artículo de enciclopedia sobre el escritor angloamericano H. Ph. Lovecraft, una fotografía suya y alguno de sus cuentos. Al leer el artículo, la simple pulsación de un botón le permitiría hojear "El horror de Dunwich" o visualizar la fotografía. Más tarde podría conectar con este conjunto la biografía de Lovecraft escrita por Pierre Bourbonnais.

Bush remarcaba que este tipo de asociación no lineal de ideas era el modo de funcionamiento natural de la mente humana, y confiaba en que dispositivos semejantes al memex lo reproducirían en el futuro más adecuadamente. Es un hecho que los artículos de una enciclopedia, las notas al pie o las referencias bibliográficas contienen conexiones no lineales de aquel tipo, pero los medios tradicionales resultan inadecuados para gestionarlas. Cuando nos encontramos con una referencia bibliográfica que nos interesa, todo lo que podemos hacer es acudir a una biblioteca o una librería. Con el memex, idealmente, pulsaríamos un botón para consultar en nuestra pantalla el libro en cuestión. En el futuro, profetizaba Bush, las enciclopedias serían redes de conexiones que el usuario podría anotar y modificar a su antojo.

Bush era un visionario. En 1945 sus ideas no eran técnicamente realizables. Ni lo eran aún en 1965, cuando otro visionario, Ted Nelson, las ordenó conceptualmente. Fue Nelson quien acuñó el término **`hipertexto'** para referirse a "un cuerpo de material escrito o gráfico interconectado de un modo

complejo que no se puede representar convenientemente sobre el papel; puede contener anotaciones, adiciones y notas de los estudiosos que lo examinan" [4]. La idea es que el lector examina los nodos de una red, y pasa de unos a otros siguiendo las conexiones (links, en inglés). El hecho de que los nodos pueden contener texto, pero también pueden integrar otros medios: imagen, sonido, etc. es lo que se quiere remarcar con otro término complementario: `**hipermedia'**.

Durante las dos décadas siguientes se vivió el auge de los ordenadores, el almacenamiento digital y las redes. El propio Nelson cobró conciencia de lo apropiado de estas nuevas tecnologías para la realización del sueño de una red de elementos de información libremente accesible alrededor del mundo. Sin embargo, se diría que sus ideas sólo han llegado a concretarse recientemente con el World-Wide Web. Ha habido numerosos proyecto de sistemas hipertexto, encontrando una relación exhaustiva en [5].

1.1. El proyecto World-Wide Web.

En 1989 la red mundial de datos, el memex global, ya existía en potencia. Internet, que se originó en el ámbito militar durante la guerra fría [6], se había desarrollado más allá de los propósitos originales como resultado de su uso por parte de la comunidad científica internacional, que necesitaba nuevos sistemas de distribución de la información. Lo único que se requería, como decíamos al principio del artículo, eran vías de acceso sencillas y homogéneas. Este era uno de los objetivos que Tim Berners-Lee se planteó en 1989 cuando presentó a sus superiores del CERN la propuesta original para el proyecto World-Wide Web. Otro era la posibilidad de gestionar conexiones no lineales.

Se pretendía pues que los recursos disponibles en formato electrónico, que residen en ordenadores distintos conectados a la red, fuesen accesibles para cada investigador desde su terminal, de un modo transparente y exento de dificultades, sin necesidad de aprender a utilizar varios programas distintos. Además, debería posibilitarse el salto entre elementos de información conexos. Los recursos existentes deberían integrarse en una red hipertextual distribuida gestionada por ordenadores.

El éxito del WWW, el crecimiento de la telaraña, ha sido espectacular. Durante 1993 se pasó de 50 a 500 nodos. En 1994 se contabilizan ya miles de servidores en el WWW que distribuyen todo tipo de información (de ellos, trece en España; el primero fue el del Departamento de Educación de la Universitat Jaume I, en septiembre de 1993). En las figuras 4 y 5 podemos ver este crecimiento del web.

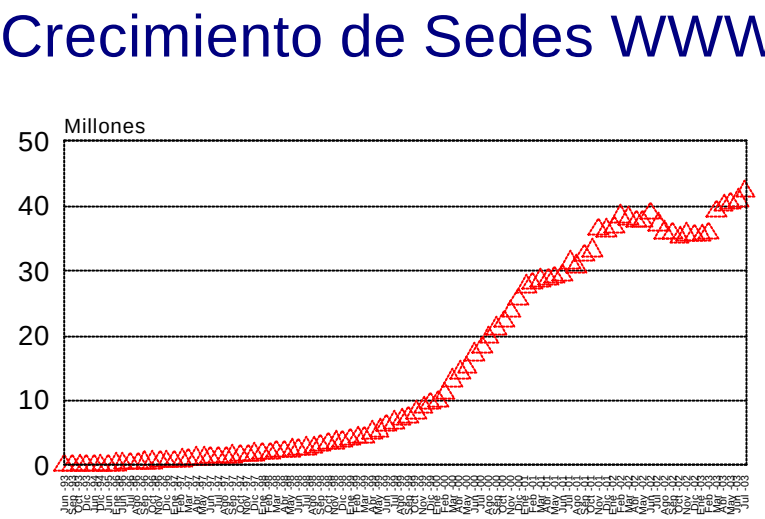


Figura 4 – Fuente: Hobbe's Internet Timeline. Robert H. Zakon
http://www.zakon.org/robert/internet/timeline/

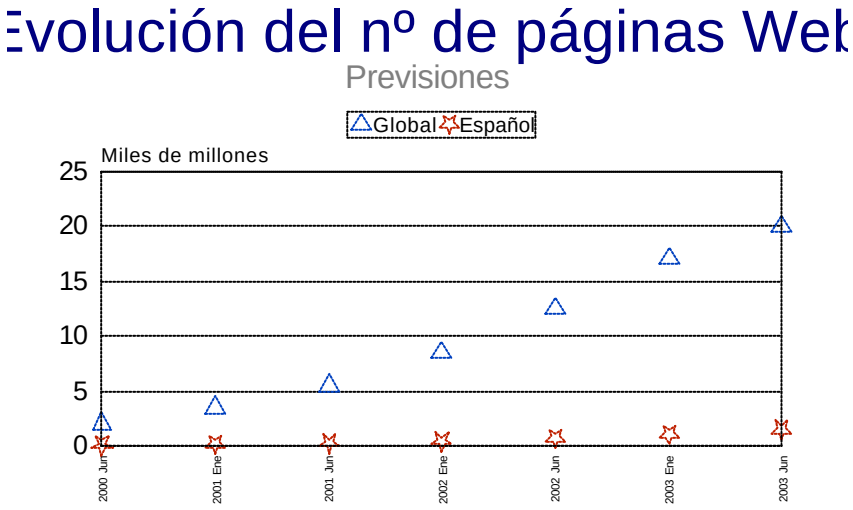


Figura 5 – Previsión de evolución

La importancia del Web como sistema de difusión de la información se ve avalada por el importantísimo crecimiento tanto en número de sedes como en

el número de páginas Web, que poseen un crecimiento claramente exponencial como además Huberman [7] refleja en la recogida de datos sobre los motores Alexa e Infoseek. Trabajos que nos permiten valorar el crecimiento del Web son los de [8], [9], [10], [11], [12].

Este crecimiento exponencial de la información en el web, así como la falta total de estructura (a efectos de la recuperación de información) de la información en él contenida es lo que ha complicado en cierta medida los intentos por crear sistemas eficaces de RI en este entorno.

Otro problema añadido es que las técnicas de RI empleadas en Internet se han estado basando en los SRI tradicionales comenzando a surgir problemas evidentes en dicho proceso. Así mismo surgen nuevos problemas propios del nuevo tipo de información surgida como recoge Baeza-Yates[13]:

1. Datos distribuidos.
2. Alto índice de volatilidad de los datos.
3. Volumen muy grande (crecimiento exponencial).
4. Datos sin estructura y redundantes.
5. Mala calidad de los datos.
6. Datos heterogéneos.

Ante estos problemas están surgiendo dos líneas o corrientes de trabajo, bastante diferentes que intentan abordar la problemática surgida.

En primer lugar, profesionales que abogan por emplear los lenguajes documentales controlados, como una solución adecuada para los problemas, y así están surgiendo un amplio número de iniciativas en este sentido, destacando entre todos ellos los metadatos. Desde nuestro punto de vista, esta es una pobre solución que se va a enfrentar con gran número de problemas, derivados todos ellos de la propia naturaleza de la información con la que estamos trabajando. Para que este planteamiento tuviera alguna posibilidad de funcionar, sería necesario que toda la información generada tuviera la adecuada asignación de los metadatos, posibilidad bastante improbable dada la naturaleza de la información y los flujos de creación de la misma. Este planteamiento sí puede ser adecuado para la información generada en una determinada institución, en la que se podría tener los profesionales adecuados para formalizar adecuadamente los datos. Las consultas que se realizaran a

este profesional, conocedor del mecanismo, darían bueno frutos, pero ¿qué sucedería en consultas realizadas por usuarios que no conociesen los mecanismos empleados? Es un hecho que la asignación de descriptores normalizados, el empleo de tesauros, etc. ofrecen resultados adecuados cuando las consultas las realiza el profesional que ha realizado el trabajo, pero los resultados de los usuarios externos no son tan buenos como en ocasiones se nos pretende hacer creer. Experiencia en este sentido tenemos con el motor de recuperación Karpanta, desarrollado por el Dr. Carlos García-Figuerola Paniagua [14], implementado en http://milano.usal.es/dtt.htm. Del análisis de las consultas realizadas en el sistema se pueden sacar conclusiones muy interesantes en este sentido.

En segundo lugar tendríamos todas aquellas modalidades que tratan de analizar la información en el web de forma automática y autónoma entre las que tendríamos el empleo de los motores de búsqueda, de directorios, etc. y una modalidad más ambiciosa que trata de valorar la carga semántica que los enlaces pueden tener en la estructura de grafo que tiene la información en el web.

Algunos motores de búsqueda, aparte de emplear los sistemas de indización de la información, añaden la posibilidad de mejorar los resultados teniendo en cuenta esa carga semántica de los enlaces.

Estos procesos de rastreo, indización y ranking ofrecen nuevos modelos en la recuperación de la información, aunque evidentemente requieren de mayor trabajo para mejorar los niveles de precisión.

Nuestro grupo de trabajo, como ya hemos apuntado en algún foro, plantea el empleo de las denominadas medidas topológicas y de las leyes de exponenciación, que nos ofrecen otras posibilidades para poder realizar la recuperación de información y trabajan en la línea de tener en cuenta la capacidad semántica de los enlaces.

En el empleo de las medidas topológicas hay que considerar el Web como un grafo, donde los nodos se representan mediante las páginas html y los enlaces se representan mediante los bordes dirigidos. Diferentes estudios [15] sugieren la existencia de varios cientos de millones de nodos en el grafo Web y con un crecimiento importante, y el número de enlaces alcanzaría varios billones [16].

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 [17] mostrando las consistencia de los planteamientos y con la aplicación de algoritmos adecuados para el tratamiento de esta gran cantidad de información [16].

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 [18], [19], [20], [21], [22].

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 [23]; para mostrar la información [24], [25], [20]; en minería del Web [26], [27].

Algúnos autores han utilizado el Web como grafo para crear de forma automática hipertextos, partiendo de textos carentes de enlaces [28], [29].

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 [30] y también permite aplicar técnicas de similaridad, basadas en los enlaces, para estructura y visualizar el Web [31].

Aplicando la terminología de [32] en teoría de grafos al grafo Web, las páginas Web se denominarían nodos y los enlaces se denominarían como arcos o aristas.

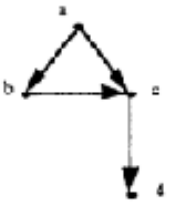
La adecuada representación de un grafo, para su análisis automatizado [33], puede realizarse mediante la creación de las llamadas matrices de adyacencia.

Para un grafo dirigido *G*, con *p* nodos, podemos definir la matriz de adyacencia *A*, formada por *p x p* elementos de la forma a_{ij}. Los valores que pueden tener los elementos de la matriz *A*, pueden ser:

I. Valor 1 si los dos nodos v_i y v_j están conectados mediante un enlace que va de v_i a v_j.
II. Valor 0 si no existe enlace desde v_i a v_j.
III. Valor nulo si i=j.

Un ejemplo de lo comentado con anterioridad se refleja en la siguiente matriz de adyacencia que representa al grafo adjunto, formado por cuatro

nodos y que tiene una distribución de enlaces dirigidos en una dirección muy concreta. La matriz es el reflejo de dicha distribución.



a_{ij}	a	b	c	d
a	-	1	1	0
b	0	-	1	0
c	0	0	-	1
d	0	0	0	-

De forma similar, podemos definir la matriz de distancia *D*, formada por *p x p* elementos de la forma d_{ij}, donde d_{ij} es la longitud (medida en enlaces) del camino más corto entre los nodos v_i y v_j si están conectados, valor cero si no están conectados y valor nulo si i=j.

[34] introduce el concepto de Matriz de distancia convertida, que se puede representar mediante D'_k y se define por *p x p* elementos de la forma d'_{ijk}, donde d'_{ijk} es igual a d_{ij} (si d_{ij} ? 0) o tiene un valor que coincide con el número de nodos (si $d_{ij} = 0$).

Botafogo pretende indicar con ello que la distancia entre dos nodos no conectados no tiene un valor infinito como se puede interpretar del valor 0.

De esta forma podemos analizar diferentes índices que podemos clasificar en índices de nodo, que resumimos en la siguiente tabla:

Tipo de Matriz Valor de la fila i, columna j	ADYACENCIA a_{ij}	DISTANCIA d_{ij}	DISTANCIA CONVERTIDA d'_{ij}
N1: Suma de los valores en la fila i	$? \sum_{j?1}^{n} a_{ij} ? ?_i$ Grado de Apertura de v_i	$? \sum_{j?1}^{n} d_{ij} ? S_i$ Status de v_i	$? \sum_{j?1}^{n} d'_{ij} ? S'_i$ COD de v_i

N2: Suma de los valores en la columna j	$\sum_{i=1}^{p} a_{ij}$ Grado de entrada de v_j	$\sum_{i=1}^{p} d_{ij}$ Contrastatus de v_j	$\sum_{i=1}^{p} d'_{ij}$ CID de v_j
N3: Suma de los valores en la fila h menos la suma de los valores en la columna h	$\sum_{j=1}^{n} a_{hj} - \sum_{i=1}^{p} a_{ih}$	$\sum_{j=1}^{n} d_{hj} - \sum_{i=1}^{p} d_{ih}$ Prestigio (status de red) de v_h	$\sum_{j=1}^{n} d'_{hj} - \sum_{i=1}^{p} d'_{ih}$
N4: Ratio de la suma de valores en todas las filas por los valores en la fila h	$\frac{\sum_{i=1}^{p} \sigma_i}{\sigma_h}$	$\frac{\sum_{i=1}^{p} S_i}{S_h}$	$\frac{\sum_{i=1}^{p} S'_i}{S'_h}$ ROC de v_h
N5:		$\sum_{j=1}^{n} \frac{1}{2^{d_{ij}}}$ Textura de v_i (Bernstein, 1992)	
S_{Dice}	$\frac{2\sum \sigma_{jk} \cdot \sigma_{jl}}{\sum \sigma_{jk}^2 + \sum \sigma_{jl}^2}$		
S_{cos}	$\frac{\sum \sigma_{jk} \cdot \sigma_{jl}}{\sqrt{\sum \sigma_{jk}^2 \cdot \sum \sigma_{jl}^2}}$		

Los índices de nodo son utilizados en recuperación de información clásica y pueden tener su aplicación en la recuperación de información en el web.

Podemos tener índices de grafo, que pueden permitirnos decidir que partes del grafo hay que recorrer y cuales no, reduciendo considerablemente el tiempo de rastreo de la información y que resumimos en la siguiente tabla:

Tipo de Matriz Valor de la fila i, columna j	ADYACENCIA a_{ij}	DISTANCIA d_{ij}	DISTANCIA CONVERTIDA d'_{ij}
G1: Suma de valores en todas las filas (o en todas las columnas)	$\sum_{i=1}^{p} \sigma_i$	$\sum_{i=1}^{p} s_i$	$\sum_{i=1}^{p} s'_i$ Distancia convertida de G
G2: Media de la suma de las filas	$\frac{\sum_{i=1}^{p} \sigma_i}{p}$	$\frac{\sum_{i=1}^{p} s_i}{p}$	$\frac{\sum_{i=1}^{p} s'_i}{p}$

G3:			$\frac{p^3 ? p^2 ? ? \sum_{i?1}^{p} S'_i}{P^3 ? 2P^2 ? P}$ Compactación de G
G4:		$? \sum_{h?1}^{p} \left\| ? \sum_{j?1}^{n} d_{hj} ? ? \sum_{i?1}^{p} d_{ih} \right\|$ Suma de Prestigio absoluto de G	
G5: LAPS=$p^3/4$ si p es par LAPS=$(p^3-p)/4$ si p es impar		$\frac{? \sum_{h?1}^{p} \left\| ? \sum_{j?1}^{n} d_{hj} ? ? \sum_{i?1}^{p} d_{ih} \right\|}{LAPS}$ Stratum de G	
G6: la suma es sobre todos los nodos adyacentes	$? \sum_{?i?j?}^{?½} ? ?$ Índice de conectividad de G (índice de Randic)		
G7:	$\frac{?}{r}$ Índice de Randic normalizado		

1.3. Nuestra solución software

Con el fin de poder dar adecuada respuesta a muchos de los planteamientos aquí expuestos, hemos venido desarrollando desde el año 1997 [35] un robot que nos permita el adecuado trabajo con la información del web. Al tratarse de un software de desarrollo propio nos permite, así mismo, elaborar las modificaciones que consideramos oportunas para conseguir la mejora del mismo.

El robot desarrollado (Sacarino Bot – Sonda AutomátiCA para Recuperación de INformaciÓn en el web) se encuentra programado en este momento en Visual Basic 6 y se está programando totalmente de nuevo para Visual Basic .NET con lo que estamos consiguiendo una considerable reducción de código y aumento de velocidad y con capacidad multiprocesamiento (esta versión será de disponibilidad pública y gratuita cuando esté finalizada).

La versión actual de Sacarino pretende ser un robot altamente configurable, en la que tengamos la posibilidad de recoger toda la información que deseemos y esta información se almacena en un SGBD con el que podemos trabajar a través de SQL.

El robot sigue el Estándar de Exclusión de Robots (en adelante SRE) [36], generado el 30 de Junio de 1994 en la lista sobre robots robots-request@nexor.co.uk [36].

En este momento el SRE se basa en dos implementaciones (cubiertas por Sacarino), complementarias entre sí:

- Protocolo de exclusión de robots.
- Etiqueta META para robots.

Pasaremos a mostrar algunas de las funcionalidades del robot.

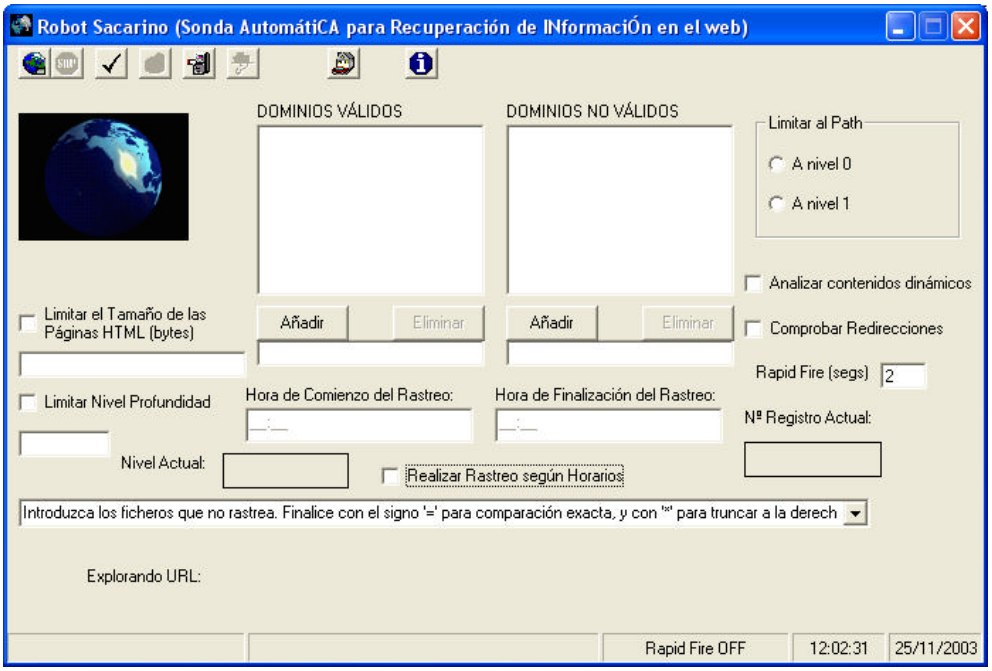


Figura 6- Pantalla principal de Sacarino

En la pantalla principal del programa tenemos la posibilidad de definir varios parámetros que modificarán el comportamiento del programa. Podemos limitar por tamaño de las páginas, fijar el nivel de profundidad de la recogida (que se hace en anchura), podemos fijar el Rapid Fire del robot, realizar o no comprobación de redirecciones, analizar o no contenidos dinámicos. Podemos especificar el ámbito de la recogida en cuando a dominios válidos (o máquinas concretas) o no válidos a efectos de la recogida. Por defecto, si no ponemos

limitaciones a efectos de la recogida, se obtiene solamente la información contenida en la máquina del URL de partida (que podemos centrar en un directorio concreto solamente modificando los límites del path).

Se puede realizar una recogida de datos en horas concretas (por si fuera necesario recoger solamente por ejemplo en horas nocturnas).

En el apartado de opciones en la recogida podemos hacer lo siguiente:

- ?? Podemos realizar el recuento de diferentes etiquetas añadidas por nosotros, realizando el programa diferentes estadísticas y gráficos con el mismo. Ver la figura 7
- ?? Podemos realizar el recuento de diferentes formatos de ficheros, obteniendo estadísticas y gráficos. Ver la figura 8
- ?? Podemos recoger en la base de datos la información específica de diferentes etiquetas, extrayendo de las mismas la información que contienen. Ver la figura 9
- ?? Podemos almacenar todas las páginas recorridas en el path deseado para posteriores tratamientos. Ver la figura 10

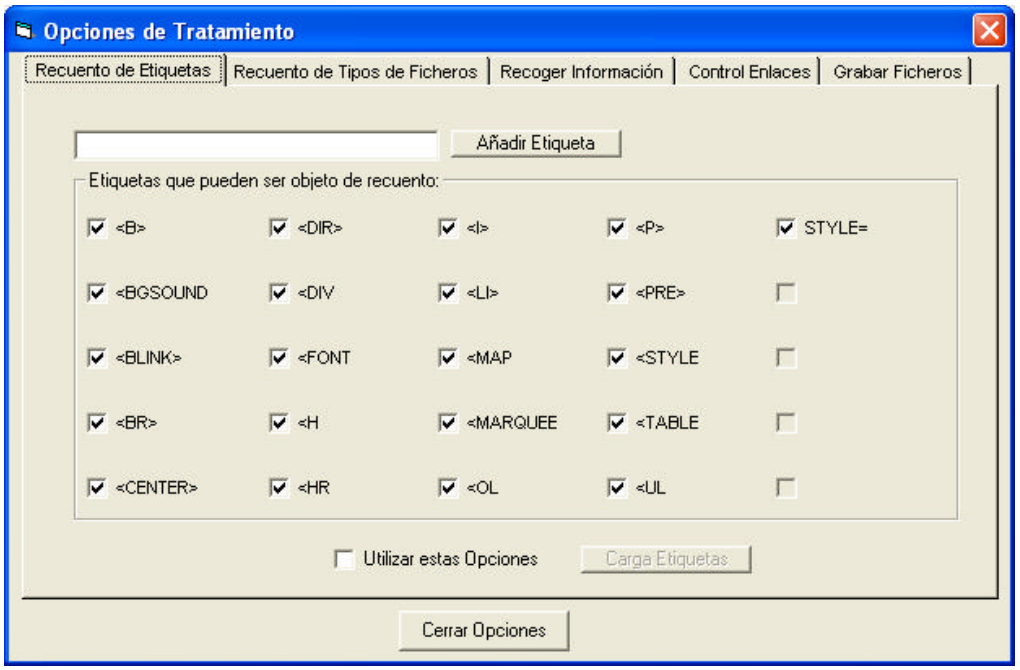


Figura 7- Etiquetas

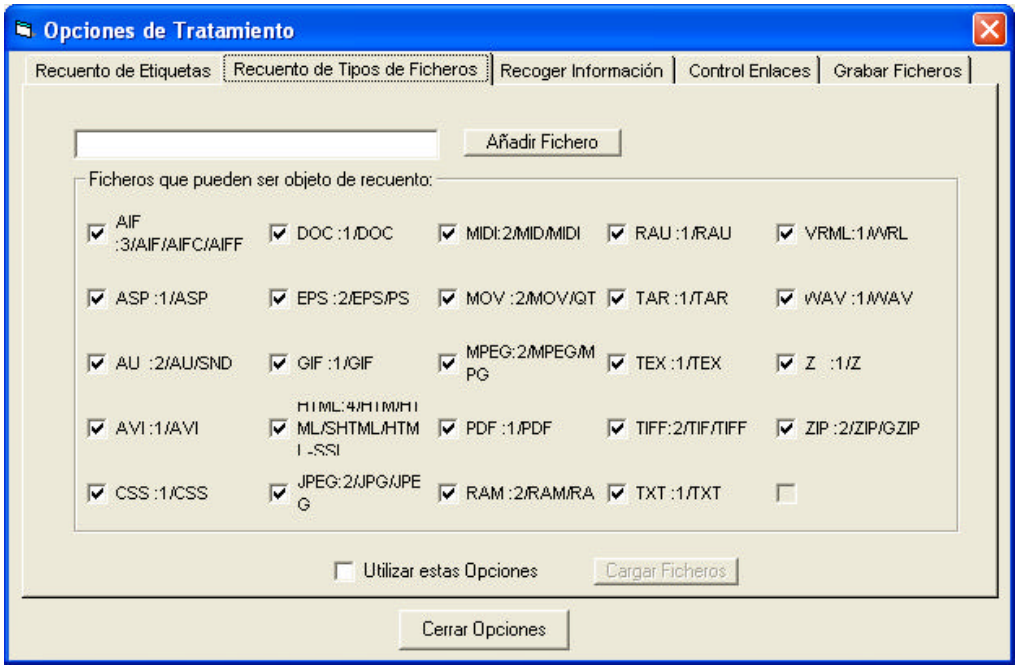


Figura 8- Ficheros

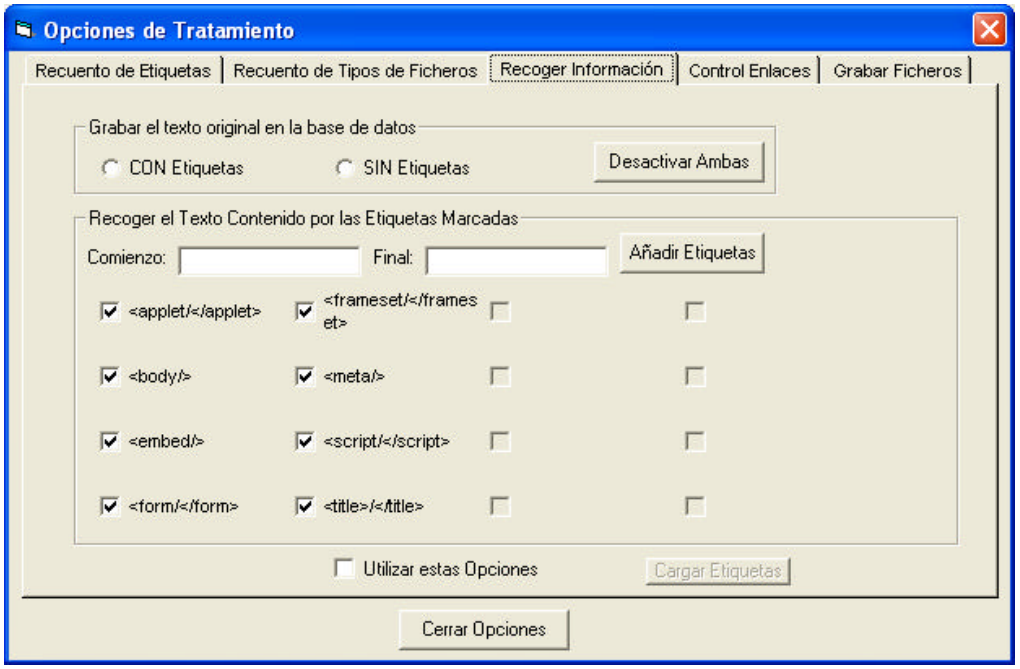


Figura 9- Información a recoger

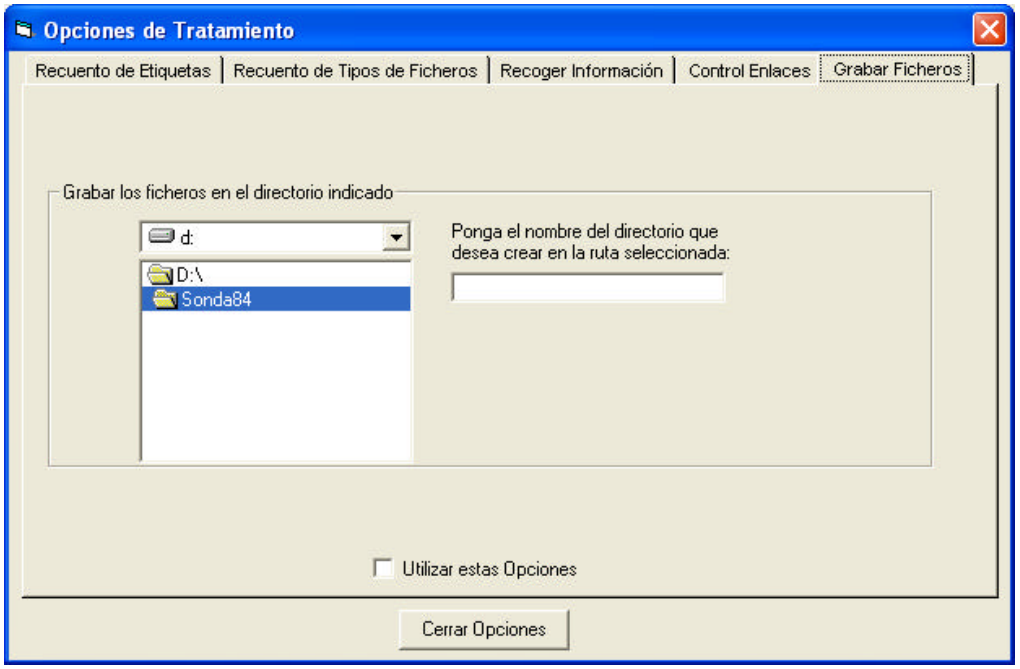


Figura 10- Almacenar los ficheros

Toda la información se recoge en un SGBD, donde además de la información comentada anteriormente, se recoge toda la información relacionada con los enlaces existentes en las páginas analizadas. Con esta información podemos, por lo tanto, obtener información esencial sobre la estructura del grafo con la que se ha trabajado. Podemos generar las matrices de adyacencia (en formato Matlab) del grafo y a partir de ellas podemos realizar todos los cálculos sobre medidas topológicas. Es interesante indicar, que desde el propio programa se puede enlzar con Matlab y ejecutar los scripts elaborados para realizar todos los cálculos.

El programa también nos permite generar un fichero de salida (con las relaciones de los enlaces) para poder procesar el PageRank de la información recogida. Este fichero se puede grabar en formato Matlab para ejecutar los programas de obtención del PageRank.

Como podemos grabar todas las páginas recorridas, estas se pueden tratar con los programas desarrollados por el Dr. Ángel Francisco Zazo Rodríguez que han dado lugar al motor URRAKA (Una Reforma Revisada y Ampliada de KArpanta). Este nuevo motor de recuperación, además de incorporar nuevas opciones en los módulos de indización y búsqueda ya existente en Karpanta, añade un módulo adicional de expansión de consultas.

El módulo de indización se ha visto ampliado por dos aspectos importantes:

- En primer lugar, el procesado de texto se ha enriquecido con la incorporación de un nuevo tipo de lematización, que es una ampliación del S-Stemmer ya utilizado en Karpanta, al que se le añade la normalización de género.
- En segundo lugar, se ha incorporado una subrutina de detección de nombres propios y siglas, importantes en el proceso de recuperación. Esta subrutina además unifica nombres propios (normalización terminológica).

En el módulo de búsqueda se ha permitido aplicar todos los esquemas de pesado conocidos del modelo TF-IDF, sobre todo, por su importancia, aquellos que incorporan información de la diferente longitud de los documentos.

El nuevo módulo de expansión pretende mejorar la recuperación de una consulta inicial ampliando nuevos términos a la consulta original. Varias técnicas se han incorporado en este módulo para permitir seleccionar aquel que mejor convenga en cada momento.

De esta forma hemos conseguido un entorno global para el adecuado tratamiento y recuperación de información en el web. Todas estas iniciativas, que mantiene nuestro grupo de investigación, siguen dando frutos y nos están permitiendo avanzar en diferentes aspectos de la recuperación de información.

CONCLUSIONES

La recuperación de información en el web nos plantea nuevos retos a la hora de poder recuperar la información, pero tenemos ha nuestro alcance soluciones que nos pueden permitir realizar una buena recuperación. Posiblemente muchos de los problemas que surgen se deben al empleo de técnicas pensadas para otro tipo de información y ello requiere nuevas técnicas de estudio que se adapten a la nueva realidad de la información. Las soluciones creemos que existen, aunque pasan por nuevas líneas de investigación que muchos profesionales de la recuperación de información no están dispuestos a afrontar. Nuestro grupo de investigación lleva varios años trabajando sobre estos temas y los resultados en algunos casos empiezan a dar frutos, aunque por el alto coste computacional la rapidez no sea la deseada.

REFERENCIAS

1. KAHLE, B. Wide Area Information Servers Concepts. [en línea]. 1989 [Citado: Septiembre 1999]. Disponible en Internet: ftp://ftp.wais.com/pub/wais-inc-doc/wais-concepts.txt

2. LINDNER, P. Frequently asked questions about Gopher. [en línea]. 1994 [Citado: Septiembre 1999]. Disponible en Internet: ftp://rtf.mit.edu/usenet/news.answers/gopher-faq

3. BUSH, V. As We May Think. *Atlantic Montly*, 1945, Vol. 176, No. 1, p. 101-108.

4. NELSON, T. H. A file Structure for the Complex, The Changing and The Indeterminate. *ACM 20th National Conference*, (1965).

5. BALASUBRAMANIAN, V. Hypermedia Issues and Applications: A State-of-the-Art Review. *Graduate School of Management, Rutgers University, Newark, New Jersey, 1994*.

6. HARDY, H. The History of the Net. [en línea]. Master Thesis, School of Communications, Grand Valley State University [Citado: Septiembre 1999]. Disponible en Internet: http://www.ocean.ic.net/ftp/doc/nethist.html

7. HUBERMAN, B. A. y ADAMIC, L. A. Evolutionary dynamics of the World Wide Web. *Tech. Rep., Xeros Palo Alto Reserach Center*, (February, 1999).

8. GRAY, M. Measuring the growth of the Web. [en línea]. 1995 [Citado: Octubre 1999]. Disponible en Internet: http://www.mit.edu/people/mkgray/growth/

9. GRAY, M. Internet Statistics: Growth and usage of the Web and the Internet. [en línea]. 1996 [Citado: Noviembre 1999]. Disponible en Internet: http://www.mit.edu/people/mkgray/net/

10. BRAY, T. Measuring the Web. *Fifth International World Wide Web Conference*, (Paris, France, 6-10 May 1996).

11. COFFMAN, K. G. y ODLYZKO, A. The size and growth rate of the internet. *First Monday*, 1998, Vol. 3, No. 10.

12. HOBBES ZAKON, R. Hobbes' Internet Timeline v5.2. [en línea]. 2000 [Citado: Febrero 2000]. Disponible en Internet: http://www.zakon.org/robert/internet/timeline/

13. BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval. New York: ACM Press ; Harlow [etc.] : Addison-Wesley, 1999.
14. FIGUEROLA, C.G.; BERROCAL, J.L.; ZAZO, A.F. Diseño de un motor de recuperación de información para uso experimental y educativo. BID. Textos Universitaris de Biblioteconomia i Documentació, 4. [en línea]. 2000 [Citado: Diciembre 2003]. Disponible en Internet: http://www.ub.es/biblio/bid/bid04.htm

15. BHARAT, K. y BRODER, A. A tecnique for measuring the relative size and overlap of public Web serach engines. *Proc. of the Seventh WWW Conference*, (Brisbane, Australia, 1998).

16. KLEINBERG, J. M., KUMAR, R. y RAGHAVAN, P. The Web as a graph: measurements, models, and methods. *Proceedings of the Fifth Annual International Computing and Combinatorics Conference*, (1999).

17. KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S. y TOMKINS, A. Trawling the Web for emerging cyber-communities. *8th. International World Wide Web Conference*, (Toronto, Canada, May 11-14, 1999).

18. BHARAT, K. y HENZINGER, M. R. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information retrieval*, (1998), p. 104-111.

19. BRIN, S. y PAGE, L. The anatomy of a large-scale hypertextual Web search engine. *Proc. 7th. WWW conference*, (Brisbane, Australia, 14-18 April 1998). Url: http://www7.scu.edu.au/

20. CARRIERE, J. y KAZMAN, R. Webquery: searching and visualizing the Web through connectivity. *Sixth international World Wide Web conference*, (Santa Clara, California, USA, April 7-11, 1997).

21. CHAKRABARTI, S., DOM, B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D. y KLEINBERG, J. Automatic resource compilation by analyzing hyperlink structure and associated text. *Proc. 7th International World Wide Web Conference*, (1998).

22. KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, p. 668-677.

23. CHAKRABARTI, S. y DOM, B. I. P. Enhanced hypertext categorization using hyperlinks. *Proceedings ACM SIGMOD*, (1998).

24. BOTAFOGO, R. A. y SHNEIDERMAN, B. Identifying aggregates in Hypertext structures. *Proceedings of Hypertext'91*, (Diciembre de 1991), p. 63-74.

25. PIROLLI, P., PITKOW, J. y RAO, R. Silk from a Sow's ear: extracting usable structures from the Web. *Conference on Human Factors in Computing Systems, CHI'96*, (Vancouver, April 13-18, 1996).

26. MENDELZON, G. M. y MILO, T. Querying the World Wide Web. *Journal of Digital Libraries* , 1997, Vol. 1, No. 1, p. 68-88.

27. KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S. y TOMKINS, A. Trawling the Web for emerging cyber-communities. *8th. International World Wide Web Conference*, (Toronto, Canada, May 11-14, 1999).

28. SMEATON, A. F. y MORRISEY, P. J. Experiments on the Automatic Construction of Hypertext from Text. *The New Review of Hypermedia and Multimedia: Applications and Research*, 1995, Vol. 1. Url: http://www.compapp.dcu.ie/~asmeaton/pubs/Hypermedia_Paper.ps

29. GOLLOGLEY, G. y SMEATON ALAN F. Assisting the Hypertext Authoring Process with Topology Metrics and Information Retrieval. *Working Papers*, (1997).

30. GIBSON, D., KLEINBERG, J. y RAGHAVAN, P. Inferring Web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia*, (1998).

31. CHEN, C. Structuring and Visualising the WWW by Generalised Similarity Analysis. *Proceedings of Hypertext'97*, (Southampton, UK, 1997), p. 177-186.

32. HARARY, F. *Graph Theory.* Reading, MA: Adison Wesley, 1969.

33. ELLIS, D., FURNER-HINES, J. y WILLETT, P. On the creation of hypertext links in full-text documents: measurement of inter-linker consistency. *Journal of Documentation*, June 1994, Vol. 50, No. 2, p. 67-98.

34. BOTAFOGO, R. A., RIVLIN, E. y SHNEIDERMAN, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*, April 1992, Vol. 10, No. 2, p. 142-180.

35. BERROCAL, J.L. *Cibermetría: Análisis de los dominios web españoles*. Salamanca: Ediciones Universidad de Salamanca, 2002.

36. KOSTER, M. A Standard for Robot Exclusion. [en línea]. 1994 [Citado: Diciembre 1998]. Disponible en Internet: http://info.webcrawler.com/mak/projects/robots/norobots.html