Métodos y técnicas para la indización y recuperación de los recursos de la *World Wide Web*

DRA. Mª DOLORES OLVERA LOBO
Facultad de Documentación. Universidad de Granada

Las herramientas de búsqueda de información en la *World Wide Web* desarrollan diferentes métodos y mecanismos para la recopilación e indización de la información que incorporan a sus bases de datos. La diversidad documental, de contenidos y formatos dificultan el proceso. El método adoptado incidirá directamente en la eficacia de la recuperación de los recursos. Se describen algunos de los métodos utilizados para adecuar el funcionamiento de los buscadores a las necesidades de búsqueda de información en Internet. Además, se examina el nuevo papel a desempeñar por los intermediarios de la información en el entorno de la red.

PALABRAS CLAVE: recuperación de información, indicación automática, Internet, World Wide Web, herramientas de búsqueda de la información

METHODS AND TECHNIQUES FOR INDEXING AND RETRIEVING WORLD WIDE WEB RESOURCES

World Wide Web information searching tools use different methods and mechanisms to collect and index the information they keep in their databases. The variety of document types, contents and formats hinders this process. The method used will impinge directly on the effectiveness of the retrieval of these resources. Some of the methods used to adapt search engines to the information-seeking behaviour in the Internet are described. Furthermore, the new role to be played by information intermediaries in the Net environment is examined.

KEYWORDS: Information retrieval, Automatic indexing, Internet, World Wide Web, Information searching tools.

1. INTRODUCCIÓN

La *World Wide Web*, W3, Malla Mundial Multimedia o telaraña mundial de información ha evolucionado hacia lo que podría considerarse un dinámico almacén donde albergar informaciones muy diversas en contenidos, relevancia y utilidad. Por el momento, gran parte de la responsabilidad en la búsqueda y localización de la

información dispersa en la red recae en los motores de búsqueda o buscadores (Lynch, 1997).

Los buscadores de la W3 presentan una estructura constituida por: un *robot* o *araña*, es decir, un programa que cruza la W3 moviéndose de un documento a otro, descendiendo progresivamente a través de los hiperenlaces; un *programa de indización* que indiza la información de los millones de páginas web ubicadas en servidores conectados a la red y enormes *bases de datos* a las que acceden los usuarios a través de la *interfaz* del buscador. Por tanto, los buscadores no sólo deben facilitar la localización de los recursos incluidos en sus bases de datos sino que, además, deben compilarlos.

Salvo en el caso de los directorios temáticos como Yahoo, Olé y otros, la indización automática es el método predominante utilizado por las herramientas de consulta de la W3. No obstante, pese a la ventaja de la rapidez, este sistema, que conlleva necesariamente el uso de robots automáticos, también cuenta con detractores (Desai 1997; Lynch, 1997) los cuáles identifican los siguientes problemas:

- El uso de robots automáticos de búsqueda de manera indiscriminada produce un incremento del tráfico en la red, una sobrecarga de los servidores y otros problemas de uso de las infraestructuras (Koster, 1995)
- La justificación de sistemas de indización mediante robots sería difícil si la red cambiara hacia un uso no gratuito de los recursos.
- El tipo de datos recogidos por los robots no es útil, ya que aun hoy las arañas presentan un funcionamiento demasiado simple.
- Los sistemas de comprensión del lenguaje natural no están lo suficientemente avanzados como para extraer el significado de los recursos.
- La mayor parte de los buscadores sólo reconocen texto, por lo que se hace más difícil generar una identificación automática de las características de recursos tales como las imágenes digitales o los diferentes ficheros multimedia.
- La indización automática tiende a una perspectiva simplista, poco selectiva, que provoca que la localización de recursos en la red y la recuperación de la información (RI) solicitada llegue a ser cada vez menos factible.
- Frente a los indizadores humanos, los programas automatizados tienen dificultades para identificar características de un documento web como el contexto o temática general en la que se engloba, y el género, por ejemplo una comunicación científica, información profesional o informal, al que ese recurso pertenece.
- La W3 carece de reglas para facilitar la indización automática. Los documentos no están estructurados de forma que los programas puedan obtener de

modo fiable la información, conocida como metadatos —autor, título, longitud del texto, materia— que un indizador humano detectaría fácilmente tras una rápida revisión.

- Los editores y/o creadores de estos recursos a veces abusan del carácter indiscriminado de la indización automática. Un servidor web puede falsear el proceso de indización con el fin de atraer la atención de los usuarios, mediante la repetición en el documento de una palabra como *sex*, muy utilizada en las búsquedas, aunque sea otro su contenido.
- El indizador profesional puede describir los componentes de páginas individuales de distinto tipo y puede aclarar de qué forma han de incluirse esas partes en una base de datos de información.
- La información cambia con frecuencia y las arañas únicamente actualizan las bases de datos de los buscadores con cierta periodicidad. Muchas páginas web no son ficheros estáticos sino que recogen información dinámica, perecedera y en constante cambio lo cuál dificulta el que puedan ser analizadas e indizadas por los programas.

Frente a este panorama, quizá en exceso pesimista, se impone el hecho de que no sería realista ni factible pretender hoy día, la indización manual de todo el espacio web. Además, la W3 cuenta con potencialidades aún no totalmente explotadas. Con las formas actualmente adoptadas para representar la información en Internet, se están desaprovechando casi absolutamente las posibilidades del hipertexto. La adopción de estructuras de almacenamiento donde existan dos redes, la de documentos y la de conceptos, podría ayudar a controlar la situación. Estas redes deberían y podrían tener una riqueza semántica de la que carece actualmente la W3. La ampliación de la tipología de relaciones entre los recursos y la aplicación del tesauro a la red de documentos junto con una indización normalizada, homogénea y fácilmente accesible, puede ser una buena, aunque lejana, solución al problema de la recuperación de información en la W3 (Pastor, 1997). Por otro lado, el uso correcto y normalizado de los metadatos podría ayudar a paliar gran parte de los inconvenientes derivados del uso de robots y favorecer una indización de calidad por parte de éstos.

2. MÉTODOS PARA LA INDIZACIÓN Y RECUPERACIÓN DE RECURSOS

Los procesos de indización y recuperación llevados a cabo por las diferentes herramientas de RI y localización de recursos disponibles en la W3, pueden contemplarse desde diferentes ángulos (Ellis, 1998):

a) Representación de los datos

Las herramientas de búsqueda utilizan distintos métodos para indizar los recursos que incorporan a sus bases de datos. La indización puede plantearse en tres niveles: submorfológico, por palabra clave y por conceptos.

La indización en el nivel submorfológico, esto es, sin análisis morfológico, sintáctico o semántico, ofrece un método muy flexible para la recuperación. Así las fuentes de información se indizan como patrones de bits o *bit patterns* de manera que texto, sonido e imágenes en movimiento, pueden indizarse y recuperarse usando la misma forma de representación. Algunas herramientas de consulta comienzan a incorporar sistemas como, por ejemplo, Excalibur Visual RetrievalWare, que ofrecen recuperación de imágenes y de texto.

Sin embargo, la indización por palabra clave o por conceptos es la que se utiliza principalmente para la representación e indización de la información. Estos métodos se desarrollan gracias a la aplicación de técnicas estadísticas de RI ahora incorporadas a una amplia gama de buscadores (Barlow, 1997):

- *Indización por palabra clave*. Mediante este sistema se crean índices inversos de raíces y palabras clave, direcciones, ubicación y frecuencia de apariciones. Este enfoque, esencialmente morfológico y estadístico, basa la RI en la similitud formal de las palabras, y las estadísticas de su presencia en documentos y colecciones de documentos. Es la forma más común de indización de textos en la W3. Algunos buscadores obtienen las palabras clave de determinados campos, las metaetiquetas HTML, pero la mayoría indiza el texto completo de las páginas, incluyendo o no las palabras vacías de significado y eliminando a veces las más frecuentes.
- *Indización por conceptos*. Existen varios procedimientos para construir bases de datos basadas en conceptos, algunas de ellas muy complejas y basadas en sofisticadas teorías lingüísticas y de inteligencia artificial. En otros casos, como Excite, se basan en una aproximación numérica, calculando la frecuencia de aparición de ciertas palabras significativas. A partir de análisis estadísticos el buscador determina qué conceptos aparecen juntos o relacionados en textos que se centran en un tema concreto. Mediante este sistema se pueden recuperar recursos que tratan un tema dado, incluso aunque las palabras incluidas en el documento no coincidan formalmente con las de la pregunta.

Otros sistemas, como Dr-Link, realizan un análisis más profundo e indizan a nivel sintáctico, semántico e incluso pragmático. Sin embargo, el mayor nivel de análisis semántico, posiblemente sea el de los sistemas que ofrecen información evaluada, revisada e indizada por humanos, que se presenta en algunos directorios temáticos, como por ejemplo en Excite e Infoseek.

b) Procesos de equiparación (matching processes)

Los servicios de búsqueda en la W3 han incorporado técnicas de recuperación avanzadas para intentar superar los problemas del sistema clásico de recuperación basado en el método de la lógica booleana, muchas de cuyas prestaciones se consideran demasiado complejas para el usuario medio. Por esta razón, la mayor parte de estas herramientas de consulta han incorporado la posibilidad de plantear preguntas en «lenguaje natural», la ordenación de los resultados según su relevancia, la ponderación de los términos de la consulta dependiendo de los intereses del usuario, la búsqueda mediante ejemplos y la ayuda en la formulación de las preguntas (Croft, 1995). Aunque la aplicación de estas técnicas avanzadas en el entorno de la W3 no es uniforme ni se ciñe a un único modelo preestablecido, todas las herramientas de búsqueda hacen uso de métodos de equiparación parcial o *partial match*. Es decir, cualquiera que sea el modelo formal teórico —probabilístico, de espacio vectorial o de conjuntos difusos— en estos buscadores, lo que los caracteriza es que permiten una comparación perfectamente matizada y no una igualación exacta entre los términos de la búsqueda y los de los documentos (Belkin y Croft, 1987). De esta manera, la equiparación se convierte en un problema matemático consistente en establecer el grado de similitud entre la representación numérica de los términos de la búsqueda planteada por los usuarios y la de los términos incluidos en la base de datos. No obstante, Frakes y Baeza (1992) señalan que la taxonomía anterior —modelo probabilístico, de espacio vectorial o de conjuntos difusos— es inexacta, dado que un sistema puede integrar características de más de una de las categorías expuestas.

Uno de los métodos utilizados para mejorar la recuperación es la búsqueda automática por conceptos o *conceptual retrieval* (Haverkamp y Gauch, 1998), una forma de expansión automática de las búsquedas (*query expansion*) utilizada por herramientas como Excite y Magellan, que supone una alternativa a la coincidencia exacta de los términos pregunta-documento. Otro de los métodos para mejorar los resultados consiste en utilizar un «tesauro» para que el usuario pueda refinar las búsquedas mediante la adición o eliminación de palabras clave de la ecuación de búsqueda. Altavista y Excite, por ejemplo, presentan esta opción donde, en respuesta a una consulta planteada, se muestran términos relacionados con los de la pregunta y se pide al usuario que indique si desea incluirlos o excluirlos para reformular su consulta más acertadamente. Hay que señalar que, a pesar de que los buscadores se refieran a esta prestación como tesauro, realmente no se trata de un lenguaje documental normalizado sino de una serie de términos que el buscador ha identificado como próximos o relacionados con los de la ecuación de búsqueda. La generación automática de tesauros que establezcan relaciones rigurosas entre los conceptos ha propiciado una interesante línea de investigación (Chen, 1998).

Otra posibilidad de extensión de las búsquedas es el truncamiento implícito (*stemming*) o reducción automática de los términos de búsqueda a su raíz, basado

en la premisa de que los términos similares morfológicamente lo son también semánticamente. Sin embargo, si esta prestación no se aplica adecuadamente, puede dar lugar a un elevado ruido documental. Por otro lado, la asignación automática o humana de descriptores, en forma de términos, categorías temáticas o símbolos de clasificación, representa una opción más a la equiparación exacta de palabras. Muchos son los directorios en la W3, como Yahoo, que siguen este método ofreciendo acceso a documentos web a partir de listas alfabéticas precoordinadas de encabezamientos de materias.

Sin embargo, y ante el uso generalizado de técnicas avanzadas de recuperación, en ocasiones se añora la capacidad de búsqueda mediante coincidencia exacta o *exact match* entre el enunciado de búsqueda y palabras o expresiones contenidas en el documento, puesto que, quizá, sea eso precisamente lo que, en muchos casos, pueda satisfacer las necesidades de los usuarios (Hahn, 1998). Belkin (1995) señala, por el contrario, que en determinadas circunstancias y pese a todos los defectos de las búsquedas booleanas, éstas pueden ser tan aconsejables como la búsqueda *best-match,* aunque admite que lo más adecuado sería una combinación de ambas, ya que hay estudios que demuestran que el uso de diferentes tipos de representación de las preguntas incrementa la efectividad de la recuperación.

c) Capacidad de aprendizaje

Los robots que rastrean la red pertenecen a un tipo de programas informáticos denominados agentes, es decir, son aplicaciones que pueden trabajar de forma autónoma y realizar actividades sin la supervisión directa de los humanos, de ahí que se les atribuya un cierto grado de «inteligencia» e «independencia» en el desarrollo de ciertas tareas. Algunos sistemas, sobre todo agentes de búsqueda personalizada, emplean el *feedback* de relevancia para mejorar su funcionamiento a través del tiempo. Partiendo de la relevancia determinada por los usuarios para los documentos recuperados en una primera búsqueda, el sistema pondera las palabras clave. Otras herramientas, como Direct Hit, utilizan la interacción con el usuario como medio para mejorar la relevancia. Este buscador trabaja «observando» y «registrando» el comportamiento de los usuarios en la realización de las búsquedas, de esta forma «aprende» y es capaz de ofrecer, cuando se le solicita, una lista donde las páginas se ordenan según su popularidad para los internautas. Direct Hit comprueba si anteriormente ya se ha hecho esa misma pregunta u otra parecida en el buscador y ordena los resultados según el número de usuarios que han preferido esas referencias, y las han consultado, de entre todos los resultados. Metabusca es otro de los sistemas que también sigue este método.

En los últimos años se vienen adoptando varios paradigmas de aprendizaje automático para la recuperación de información y el análisis textual como, las redes neuronales, el aprendizaje simbólico y los algoritmos genéticos. Una forma de apren-

dizaje automático que no requiere *feedback* de usuario es la representada por el método de «vida artificial» aplicado a la recuperación donde agentes con capacidad de aprendizaje dependen para su supervivencia de la RI que realicen en respuesta a las consultas. Los agentes de búsqueda examinan intranets e Internet procesando información, emplean técnicas de aprendizaje automático y adaptan dinámicamente su reproducción y actividad usando técnicas de vida artificial, en un intento de optimizar su funcionamiento (Haverkamp y Gauch, 1998).

d) Sintaxis de la consulta (input)

Dadas las ventajas que presentan para los internautas tanto los directorios temáticos, con índices navegables y organizados de recursos, como los buscadores, con grandes bases de datos e interesantes prestaciones de búsqueda, la tendencia actual se dirige a incorporar ambas posibilidades dentro de un mismo servicio. De esta forma, el *browsing* y la búsqueda basada en términos van hoy juntos y son la forma predominante de RI en Internet.

La mayor parte de los buscadores permiten formulaciones booleanas donde el usuario cuenta con un gran nivel de control lingüístico. También suelen permitir la búsqueda en lenguaje natural, lo que libera al usuario de tener que ejercer ese estrecho control. La búsqueda mediante ejemplos o *query by example* invita, además, a identificar documentos relevantes sobre los que basarse para mejorar la recuperación como la opción «más como éste» en Excite, aunque no siempre los criterios utilizados se hacen explícitos para los usuarios.

e) Coordinación de las búsquedas

La precoordinación es inherente a muchos de los servicios basados en directorios organizados mediante listados de materias o clasificaciones bibliotecarias, aunque esto último, con menor frecuencia. Muchos buscadores que efectúan las consultas mediante palabras clave ejecutan búsquedas postcoordinadas. Sin embargo, como se ha indicado, la mayor parte de estos sistemas ofrecen una combinación de pre y postcoordinación.

3. TENDENCIAS ACTUALES

El inmenso volumen de información, la aparición de nuevos formatos, el creciente desarrollo de archivos multimedia y las diferentes «normas de etiquetado» para la identificación de objetos, causan problemas a los robots o agentes automáticos encargados de la localización de recursos en la W3. La naturaleza cambiante de Internet, el modo de funcionamiento de los robots, los programas de indización

de documentos, las técnicas de recuperación utilizadas, así como el procedimiento de recopilación de datos seguido para la elaboración de directorios en la W3, siguen dificultando la localización de una ingente cantidad de información valiosa residente en la red.

Para paliar estos problemas, una de las tendencias que más claramente se observan en relación con la búsqueda de información es la aparición de numerosos buscadores temáticos y directorios especializados. Se trata de herramientas de consulta con bases de datos de menor tamaño pero que ofrecen resultados de búsqueda más ajustados a los intereses de los usuarios puesto que recopilan recursos de la W3 de un área de interés determinada. Las colecciones de buscadores como Buscopio o Tematicos ofrecen completos y actualizados listados de los buscadores y directorios existentes.

Por otra parte, la indización a texto completo llevada a cabo por los buscadores generales no permite incluir, entre otros: ficheros con formato de tipo PDF, servidores cuya consulta exige que el usuario se registre e identifique previamente, servicios que no muestran los datos directamente sino que solicitan un perfil o un enunciado de búsqueda específico, o ficheros gopher, ftp, telnet, de correo electrónico, etc.

Para superar esta limitación los buscadores siguen varias estrategias:

a) Incorporan nuevas prestaciones basándose principalmente en las extensiones del nombre de los ficheros y en el texto que extraen de estos recursos, por ejemplo:

- **Altavista** permite realizar búsquedas mediante las etiquetas HTML *image* para búsqueda de imágenes y *applet* para búsqueda de aplicaciones Java.
- **Hotbot** permite restringir la consulta a determinados tipos de archivos por ejemplo: imagen, Schockware, JavaScript, Java, audio, Acrobat, VBScript, ActiveX, video, VRM.
- **Lycos** ofrece estrategias de recuperación especializadas en imágenes y sonidos basándose en el texto de las etiquetas y en el nombre de los ficheros.

b) Han desarrollado secciones especializadas para la búsqueda de diferentes tipos de recursos, así:

- **Altavista** cuenta con un buscador de *medios* con más de 17 millones de imágenes consultables, clips de audio o archivos de películas, (http://image.altavista.com).
- **Lycos** presenta una sección (http://mp3.lycos.com) con un buscador de archivos de música MP3 (MP3 Search).
- **Yahoo**! incluye un buscador de imágenes organizadas en categorías temáticas: arte, entretenimiento, ciencia, etc. (http://ipix.yahoo.com).

c) Sirven de punto de partida proporcionando enlaces hacia servicios especializados en la localización de este tipo de materiales, como:

- **Scour.net**, un buscador y una guía para recursos multimedia en Internet: audio, video, imágenes y animaciones relativas a películas, música, radio, deportes y televisión, noticias y educación (http://www.scour.net).
- **Tile** (http://tile.net) y **Topica** (http://www.topica.com): buscador de listas de correo electrónico o *e-lists* y de grupos de noticias o *newsgroups*.

Aunque el desarrollo de nuevos productos, métodos y estrategias para mejorar la búsqueda automatizada de información en la W3 continúa imparable, se está observando un creciente interés por potenciar servicios de búsqueda gestionado por expertos. Un ejemplo es el servicio gratuito ofrecido por HumanSearch (http://www.humansearch.com). En este caso no son las máquinas, sino especialistas humanos los que interpretan la necesidad de información del usuario, elaboran una ecuación de búsqueda que la represente, realizan la consulta en diferentes buscadores, analizan los resultados y los ordenan según su relevancia a la pregunta planteada. El gran éxito de esta idea debería conducir nuevamente a la reflexión sobre el papel de los intermediarios de la información en el universo de la red Internet.

4. LOS INTERMEDIARIOS DE LA INFORMACIÓN

Internet está creando un inmenso número sin precedentes de usuarios noveles de sistemas complejos de información que están desarrollando nuevas formas de integrar las herramientas en red en su trabajo, estudio y entretenimiento diario (Nahl, 1998). A los usuarios de Internet normalmente se les considera usuarios finales aunque estos usuarios también incluyen a los propios creadores de información y la mayoría de los participantes en la infraestructura de información que dependen de Internet para ofrecer sus servicios (King, 1998). El concepto de usuario final queda difuminado en esta transmisión de información.

El crecimiento de publicaciones electrónicas en Internet, las iniciativas desarrolladas en torno a proyectos de bibliotecas digitales, el incontrolable y dinámico volumen de datos disponibles conducen a que los tradicionales servicios documentales de indización y resumen no sean suficientes para la búsqueda y recuperación de información en estas grandes bases de datos hipermedia. Con frecuencia se compara a Internet con una inmensa biblioteca mundial, la gran biblioteca virtual de la edad digital. Los legos y los profesionales familiarizados con el acceso a la información automatizada pueden, en un principio, albergar la idea —errónea como más tarde comprueban— de que la W3 es una gran biblioteca virtual o una inmensa y casi ilimitada base de datos. Desde el momento en que se establece un primer contacto con la red se puede comprobar fácilmente que esta opinión no se sostiene de ninguna manera. Internet y, concretamente, la colección de recursos multimedia

conocidos como W3, no fue diseñada para soportar la edición y recuperación de información de forma organizada como en las bibliotecas. Ambas nociones —biblioteca y base de datos— implican organización y control (como sinónimo de orden y no de censura) y una cierta normalización. Esto no se produce en el mismo grado en la W3, ya que no se trata de un sistema plenamente estructurado. Esta cuestión es importante porque condiciona la búsqueda y localización de información. Siguiendo el símil, la W3 sería una inmensa biblioteca desorganizada, sin catálogos. Como se ha señalado reiteradamente, la red constituye un depósito caótico para la publicación y distribución de materiales provenientes de todo el mundo, de enorme variedad en cuanto a su contenido (libros, artículos de publicaciones periódicas y aportaciones a congresos, datos científicos originales, páginas personales, menús de restaurantes, publicidad), formato (registros de vídeo y de audio, imágenes, diseño gráfico) y perdurabilidad (lo efímero se mezcla con trabajos de permanente importancia).

El continuo crecimiento de la W3 y su popularización hace que se haya afianzado como una nueva forma de comunicación. Los servicios de búsqueda de la W3 utilizan distintos métodos de indización, organización y acceso a la información. Pero la red, como se ha dicho, no es una biblioteca digital. Precisamente debido a esta circunstancia hay quien sigue viendo su futuro comprometido por tanta desorganización. Una de las grandes dificultades de un sistema descentralizado como Internet, donde cientos de ordenadores almacenan y aportan diferentes datos, documentos e interfaces, es la recuperación de la información.

Quizá por esto llegue a ser necesario, e incluso imprescindible para mantener la vigencia y utilidad de este sistema, algo tan tradicional como los servicios bibliotecarios para organizar, ofrecer acceso y preservar la información en la red (Lynch, 1997). Pero aun en el caso de que esa perspectiva llegara a cumplirse, la red no se parecería a una biblioteca tradicional, porque sus contenidos seguirían estando mucho más dispersos que en una colección tradicional. Además, la tarea del actual gestor de información, acostumbrado a trabajar con una colección mucho más estable, está sufriendo una evolución para adaptarse a las peculiaridades de este nuevo medio, en el que hay muchos agentes implicados, y que no ha terminado de configurarse definitivamente.

La implicación del profesional de la información puede ser, por tanto, determinante para decidir el futuro de la red Internet como defienden las pretensiones más optimistas. Sin embargo, hasta el momento no se vislumbra claramente este futuro alentador. Para conseguirlo los documentalistas, como *facilitadores* e intermediarios de la información, deben formular nuevos planteamientos, nuevas soluciones, ofrecer servicios adecuados a un nuevo tipo de usuarios, ayudándoles a mitigar los problemas que se vienen produciendo y que van en aumento porque, cuanto mayor es la cantidad de información disponible, tanto más se multiplican los problemas de recuperación. Por eso se ofertan y funcionan en la red servicios para filtrar datos y

ofrecer al usuario únicamente aquello que le interesa. Estos servicios ya tienen considerable éxito con relación a las noticias o *news* y, aunque en muchos casos prometen más de lo que ofrecen realmente, quizá constituyan la tendencia que observemos en el futuro, con versiones mejoradas. La solución a la saturación de información puede venir de la mano de servicios no gratuitos y que el usuario opte por pagar para recibir información de calidad filtrada y organizada óptimamente.

BIBLIOGRAFÍA

BARLOW, L. *The Spider's Apprentice: how to use web search engines*. Monash Information Services, 1997. Disponible en:

http://www.monash.com/spidap.html (Consultado 4 julio 97).

BELKIN, N.J.; CROFT, W.B. "Retrieval techniques". *Annual Review of Information Science and Technology*. 22, 109-146, 1987.

BELKIN, N.J. et al. "Combining the evidence of multiple query representations for information retrieval". *Information Processing and Management*. 31(3) 431-448, 1995.

CHEN, H. et al. "Internet browsing and searching: user evaluations of category map and concept space techniques". *Journal of American Society for Information Science*. 49(7) 582-603, 1998.

CROFT, W.B. "What do people want from information retrieval?: the top 10 research issues for companies that use and sell IR systems". *D-Lib Magazine*. nov 1995. Disponible en: http://ciir.cs.umass.edu/info/people/staff/croft.html (Consultado 1 jun. 96).

DESAI, B.C. "Supporting discovery in virtual libraries". *Journal of the American Society for Information Science*. 48(3) 190-204, 1997.

ELLIS, D. Ford, N. "In search of the unknown user: indexing, hypertext and the world wide web". *Journal of Documentation*. 54(1) 28-47, 1998.

FRAKES, W. B.; BAEZA YATES, R. *Information retrieval: data structures and algorithms*. Englewood Cliffs: Prentice Hall, 1992. ISBN 0134638379.

HAHN, T.B. "Text retrieval online: historical perspective on web search engines". *Bulletin of the American Society for Information Science*. 7-10, april/may 1998.

HAVERKAMP, D.S.; GAUCH, S. "Intelligent information agents: review and challenges for distributed information sources". *Journal of the American for Information Science*. 49(4) 304-311, 1998.

KOSTER, M. "Robots in the Web: threat or threat?". *ConneXions*. 9(4), april 1995. Disponible en: http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html (Consultado 7 enero 96).

KING, D.W. "Some economic aspects of the Internet". *Journal of the American Society for Information Science*. 49(11) 990-1002, 1998.

LYNCH, C. "Searching the Internet". *Scientific American*. marzo 1997. Disponible en: http://www.sciam.com/0397issue/0397intro.html (Consultado 14 mar. 98).

NAHL, D. "Learning the Internet and the structure of information behavior". *Journal of the American Society for Information Science*. 49(11) 1017-1023, 1998.

PASTOR SÁNCHEZ, J.A. "Limitaciones del WWW en el ámbito de la información documental". *Information World en Español*. 6(4) 11-13, 1997.