Buscadores de información en la World Wide Web: Características y tendencias

María Dolores Olvera Lobo

RESUMEN

Se analizan las principales características y prestaciones presentes en la mayor parte de buscadores web de Internet. Se indican los criterios utilizados para la presentación de los resultados de las consultas, así como los nuevos métodos para establecer el ranking de resultados. Finalmente, se señalan algunas de las líneas de investigación emprendidas para la mejora de estas herramientas destinadas a la búsqueda y recuperación organizada de información.

ABSTRACT

The main features and services of the majority of Internet web searchers are analyzed. Criteria used for presenting the results of searches, as new methods for ranking results are suggested. Finally, some research policies undertaken for the improvement of these organized information retrieval tools are.

1. Introducción

Internet se ha constituido en una plataforma ideal para la publicación, difusión e intercambio de datos. La red ha creado un entorno totalmente innovador para la búsqueda de información y su gran versatilidad ha generado múltiples usos. Internet y, en especial, la *World Wide Web* (también conocida como WWW, Web, W3, en español malla mundial multimedia o telaraña mundial de información) no se ideó para atender la publicación y recuperación organizada de información. Por el contrario, la red ha evolucionado hacia lo que podría considerarse un caótico (o, en sentido optimista, dinámico) almacén donde albergar informaciones muy diversas en contenidos, relevancia y utilidad.

Los usuarios que buscan información en la red lo hacen con dos fines fundamentales: o bien para explorar el espacio de información con la intención de familiarizarse con él y localizar algo de su interés o bien para buscar y recuperar información relevante de forma más concreta [1]. Estas dos situaciones son las que, con ciertos matices si se quiere, se denominan *browsing* (una estrategia de búsqueda de información exploratoria, no planificada y casual especialmente apropiada para problemas mal definidos y como una alternativa a la estrategia de búsqueda *booleana* compleja) y búsqueda por palabra clave. Por el momento, es la técnica informática la que asume casi toda la responsabilidad de la organización de la información en la red mediante los "motores de búsqueda" o "buscadores" [2]. Por tanto, es la indización automática la que se aplica, de forma predominante, a tan amplia cantidad de recursos heterogéneos y la que incide en la búsqueda y recuperación de información realizada directamente por parte de los usuarios a través de los nuevos sistemas de recuperación de información que se desarrollan en el marco de la W3.

La novedad del tema y las rápidas transformaciones que se suceden provocan una falta de normalización y una patente confusión terminológica en relación con las herramientas de consulta en la W3 de manera que se pueden encontrar términos diferentes para designar el mismo concepto y viceversa: "agentes" (*agents*, *sofbots*), "arañas" (*spiders*), "motores de búsqueda" o "buscadores" (*search engines*), "robots", "índices", "directorios" y las ya en desuso "gusanos" (*webcrawlers*, *worms*), "vagabundos" (*webwanderers*), "rastreadores", etc.

El término *search engine* abarca una amplia variedad de servicios que ofrecen acceso a los recursos de Internet. Todos los buscadores presentan una estructura similar constituida principalmente por la base de datos, la interfaz, el programa de indización y el robot de búsqueda. El robot o araña es el programa que cruza la web moviéndose de un documento a otro, descendiendo progresivamente a través de los hiperenlaces. El programa de indización indiza la información de los millones de páginas ubicadas en servidores conectados a la red, formando así enormes bases de datos a las que acceden los usuarios a través de la interfaz del buscador.

Por otro lado, los *agentes* son programas 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 tales como la recuperación de información, transacciones comerciales, interacción con usuarios de juegos de ordenador o, incluso, los temidos virus informáticos, etc. Se puede decir, pues, que los robots son un tipo de agentes. Por último, los *índices* o *directorios* son un tipo de servicios de búsqueda diferente, aunque complementarios de los buscadores web, creados generalmente con intervención humana, donde las páginas web se presentan organizadas temáticamente y pueden consultarse mediante ojeo o navegación a través del directorio (*browsing*), además de tener un motor interno para las consultas.

Los motores de búsqueda surgieron para facilitar la localización de información. La primera generación de buscadores (programas sencillos, públicos y gratuitos) hace su aparición en los años 1993 y 1994. De entre los primeros en surgir destacaban el ya desaparecido WWWWorm y WebCrawler.[1] Sin embargo, también por esas fechas comenzaron a darse a conocer buscadores más potentes, como Altavista,[2] Excite ,[3] Infoseek ,[4] Lycos[5] y Opentext .[6] Los motores crearon sistemas de búsqueda cada vez más avanzados, integrando tecnologías innovadoras con interfaces sencillas. Las mejoras constantes (búsqueda por conceptos, ordenación de los resultados según su popularidad, etc.) hacen que hoy se hable de una nueva generación de buscadores, aunque sus cambios y su evolución han sido y son constantes.

La creciente cantidad y calidad de las prestaciones de búsqueda contribuyeron desde un primer momento a hacer estas herramientas imprescindibles para los internautas. En 1996 se produce el despegue definitivo de los buscadores web, que comienzan a recabar la atención de publicaciones del campo de la informática y de la documentación, de revistas profesionales de muy diversos ámbitos, de la prensa general, etc. Su éxito y su calidad dan lugar a una variada gama de herramientas de consulta muy bien acogidas por los usuarios, tales como los metabuscadores, los agentes personales de búsqueda, los servicios de búsqueda especializados, etc.

2. Características de los buscadores web

Cuando accede a un buscador (*query-based engine*), el usuario normalmente encuentra una página web que presenta una "plantilla" o formulario en la que introduce la ecuación de búsqueda constituida por palabras clave, operadores (booleanos, de proximidad), indicación de la etiqueta HTML (*hypertext markup language* [lenguaje de marcas hipertextuales]) donde se han de encontrar los términos en el documento y demás datos que se consideren necesarios para delimitar y centrar la búsqueda. Una vez procesada, el programa muestra los resultados de la consulta ordenados según su relevancia probable relativa a la pregunta planteada. Los primeros buscadores se caracterizaron por ser internacionales (es decir, por permitir el acceso a recursos ubicados en servidores dispersos por todo el mundo) y generales (que ofrecen datos relativos a los más diversos contenidos). Sin embargo, también han surgido servicios de búsqueda que poseen robots programados para localizar e indizar informaciones que se ajustan a un patrón específico de contenido y limitan el descubrimiento a los recursos apropiados: son los motores de búsqueda o buscadores especializados. Una opinión muy extendida es que, ante la avalancha constante de información, este tipo de buscadores se impondrá en el futuro y quizá únicamente sobrevivan unos cuantos buscadores generales [3, 4]. Algunos de los más destacados buscadores generales e internacionales son: Altavista, Excite, Infoseek, Hotbot,[7] Northern Light.[8]

1 <*http: //www.webcrawler.com/*>.
2 <*www.altavista.com*> y <*www.altavista.digital.com*>.
3 <*http://www.excite.com*>.
4 <*http://www.infoseek.com*>.
5 <*http://www.lycos.com*>.
6 <*http://opentext.com*>.
7 <*http://www.hotbot.com*>.

Teniendo en cuenta su finalidad y uso, algunas de las cualidades necesarias que deben presentar los buscadores web están en función de su sencillez, adaptabilidad y relevancia [5]. Todas las características que en su día se consideraron imprescindibles en una herramienta "ideal" [6] están hoy presentes en los motores de búsqueda, además de otras que se han ido incorporando. Estas son algunas de las prestaciones que más comúnmente presentan los buscadores [7, 8].

a) Las combinaciones *booleanas* son comunes pero no siempre de forma clara. A veces las relaciones lógicas son automáticas o implícitas, aunque puede ser difícil para el usuario determinar en qué casos se realiza y si se utiliza de forma automática el operador de intersección *and* o el de suma *or*. En algunos casos se permite el uso de paréntesis o anidamiento para formular ecuaciones complejas.

b) Los operadores de proximidad, con múltiples variantes, y las búsquedas de frases son frecuentes pero no siempre están disponibles.

c) Raramente permiten preguntas en lenguaje natural con resultados aceptables.

d) La búsqueda difusa (*fuzzy search*) es una característica estándar. Es más rara la búsqueda por coincidencia exacta a la pregunta.

e) El truncamiento, final o interno, está disponible prácticamente en todos los sistemas y, en muchos casos, automáticamente, lo cual puede inducir a resultados no deseados. Quizá por ello, lo que realmente se aprecia es la posibilidad, no muy común, de buscar cadenas de caracteres.

f) Delimitar por fechas, dominio, lengua o tipo de fichero, la ponderación de los términos de la pregunta y la búsqueda por etiquetas HTML del documento web son características disponibles en distinto grado y con grandes variaciones entre los diferentes buscadores.

g) Es habitual el uso de una lista de palabras vacías aunque suele resultar difícil para los usuarios determinar a priori cuáles se han considerado así.

h) El uso de un vocabulario más o menos delimitado y organizado, de una jerarquía de conceptos o de un "tesauro" para la expansión de la pregunta aún no se ha generalizado. A veces, sin embargo, la expansión de la formulación de búsqueda mediante incorporación automática de sinónimos o términos relacionados se hace automáticamente, sin conocimiento o control del usuario, que no puede desactivarla cuando no la necesita.

i) La discriminación mayúsculas-minúsculas a veces está disponible otras no y, en ocasiones, es difícil determinarlo.

j) Tanto la ordenación de los resultados por relevancia como el proceso de retroalimentación para afinar la recuperación (*relevance feedback* o *iterative search*) pueden estar disponibles, pero basándose en una variedad de criterios desconocidos para los usuarios.

El diseño de la página principal de un buscador (que suele incluir un directorio temático o un enlace directo hacia este) habitualmente presenta una ventana para introducir la ecuación de búsqueda y menús desplegables para aplicar operadores o limitar por etiquetas HTML o tipo de fuentes (noticias, ficheros de sonido, de imágenes, etc.), así como la posibilidad de dos modos de búsqueda: simple y avanzada. Por otra parte, el gran número de documentos que los motores devuelven en respuesta a las consultas hace que una de las prestaciones que el usuario agradece más sea la ordenación de los resultados según su relevancia a la pregunta. Estos servicios listan las referencias en función de cuán pertinentes resultan (al menos, probablemente) respecto a la búsqueda planteada, mostrando en primer lugar los documentos que más se ajustan a la necesidad informativa expresada por el usuario. Sin embargo, estas listas con frecuencia causan cierta sorpresa y confusión al presentar, en muchos casos, resultados que parecen completamente irrelevantes.

3. Presentación de los resultados

Lo cierto es que ningún buscador de la W3 es perfecto. Es más, a menudo producen la sensación contraria. Si lo que el usuario recibe del buscador al final de la interacción con el sistema es precisamente una lista de referencias, ¿por qué se recuperan páginas que poco o nada tienen que ver con la búsqueda planteada previamente?. Esto se debe a que los buscadores utilizan un algoritmo (cuya fórmula exacta es siempre secreta) para ordenar los resultados. Uno de los criterios más utilizados es la frecuencia de aparición de los términos de la pregunta en el documento. Consecuentemente, si la palabra clave es común o tiene otros significados, se pueden recuperar gran cantidad de referencias irrelevantes. Los

8 <http://www.nl.com>.

buscadores suelen considerar, además de la frecuencia, la posición de las palabras clave en el documento para determinar su relevancia. Localización y frecuencia no son, pues, los únicos factores pero sí tienden a ser los factores dominantes. La mayor parte de los buscadores utiliza una combinación de indicadores para determinar la mayor o menor relevancia de los documentos recuperados. El algoritmo utilizado cuenta con varios de los indicadores que se señalan a continuación si bien su combinación y ponderación varía de un sistema a otro:

a) Frecuencia de la palabra o frase de la consulta en el documento. Generalmente se da prioridad en el *ranking* a las páginas que contienen un gran número de veces las palabras clave de la pregunta. Sin embargo, algunas palabras consideradas raras y significativas se ponderan mejor que otras palabras comunes.

b) Longitud del documento. Si tiene poca extensión y contiene repetidamente los términos de búsqueda tiene prioridad sobre otro más extenso que también repite las palabras con frecuencia. Es decir, se prima la cantidad relativa de menciones de las palabras clave respecto del total del documento.

c) Presencia de todas y cada una de las palabras o frases de la consulta en el documento recuperado.

d) Hay buscadores que, además, dan prioridad a los documentos donde aparecen en el mismo orden que en el de la pregunta.

e) Proximidad entre sí de las palabras clave de una ecuación compleja en el documento recuperado.

f) Palabras o frases de la pregunta al principio del texto, mejor si es en el título o en los encabezamientos.

g) Presencia de las palabras de la ecuación de búsqueda en las meta-etiquetas (*keywords, title, author, description,* etc.).

h) Grado de "popularidad" del documento, es decir, si ese recurso es muy citado en otras páginas web.

i) Si forma o no parte de los recursos comentados y evaluados por el personal de ese servicio de búsqueda (en el caso de que se ofrezca esta posibilidad añadida).

j) En el caso de los directorios, las categorías situadas en las ramas superiores del árbol jerárquico, que corresponden a encabezamientos más generales, pueden considerarse más relevantes que las subordinadas.

k) Algunos creadores de páginas web pretenden "engañar" al buscador repitiendo términos en los primeros párrafos o en las metaetiquetas, lo que se conoce como *spamming*. Estos documentos son penalizados y remitidos al final de la lista.

Aunque no es lo más frecuente, varios buscadores permiten la participación del usuario en las decisiones relativas a la ordenación de los resultados. La opción avanzada de Altavista y de Lycos son dos significativos ejemplos de ello.

La adecuada presentación de los resultados es de tal importancia para el usuario que, aunque se recuperen miles de referencias en respuesta a una pregunta, lo que realmente importa es si el algoritmo de ordenación es verdaderamente eficaz puesto que, en el mejor de los casos, el usuario no revisará más allá de las diez o veinte primeras referencias. Con el fin de perfeccionarlos se están desarrollando algunas técnicas que utilizan nuevos métodos para establecer el *ranking* de resultados [9]:

- Feedback *de relevancia.* Algunos nuevos sistemas, como Direct Hit <http://www.directhit.com>, 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 toda la lista de resultados. Metabusca <www.metabusca.com> también sigue este método.

- *Ponderación de los enlaces entre documentos.* Este sistema consiste en considerar los hiperenlaces incluidos en las páginas web para establecer la relevancia de cada documento recuperado y ubicarlos en el *ranking* de resultados. Los dos principales proyectos en este sentido son el sistema Clever, de IBM, y Google <http://google.com>, de la Universidad de Stanford. En el caso de Clever el proceso comienza cuando recoge de un buscador un conjunto de unos pocos cientos o miles de páginas relevantes para una búsqueda concreta. Las páginas que tienen mayor número de enlaces apuntando hacia ellas tienen mejor

puntuación. Esta se recalcula para evitar el ruido y se asigna más peso a las páginas realmente más importantes creando un conjunto depurado y completamente nuevo respecto al inicial. Para determinar la relevancia, Clever tiene también en cuenta como un componente clave tanto el texto incluido en el enlace como el que está próximo a él. Por el contrario, Google es en sí mismo un buscador con un robot que rastrea la red y también pondera la popularidad de los enlaces como parte principal de su mecanismo de ordenación. También tiene en cuenta los términos que aparecen en negrita, en los encabezamientos, con el texto en letra de mayor tamaño, etc. ofreciendo muy buenos resultados en las consultas.

- *Criterios comerciales.* RealNames <www.realnames.com> y GoTo <goto.com> han puesto en marcha tecnologías que permiten comprar el privilegio de aparecer en los primeros lugares del *ranking* de resultados. En el primer caso, los creadores o responsables de páginas web pagan para que cuando se realice una consulta mediante una palabra clave el sistema ofrezca, además de una lista de resultados realizada normalmente, aquellas páginas registradas en RealNames para ese concepto. En el segundo caso, cuanto más se pague, más arriba se estará en el *ranking* de resultados. Ciertamente, estos no parecen criterios muy fiables.

Las opciones de presentación de los resultados en la mayoría de los sistemas son bastante limitadas. El buscador ha de facilitar el proceso de búsqueda y otorgar al usuario la mayor autonomía posible no sólo para plantear las consultas sino también para configurar la manera en que quiere ver los resultados. Es conveniente que se ofrezcan, como mínimo, dos formatos de presentación, uno breve y otro más completo. Óptimamente deben ser tres: uno breve o simple, uno estándar y uno detallado. Algunos buscadores ofrecen más. Los formatos (en sus diferentes versiones) pueden incluir varios de los siguientes elementos:

- Título.
- Puntuación o valor de la relevancia.[9]
- Dirección o URL.
- Resumen.
- Tamaño del archivo en *bytes.*
- Fecha de creación del archivo.
- Fecha de entrada en la base de datos.
- Lengua.
- Categoría temática en la que se ha incluido (si el servicio posee directorio).
- Términos de búsqueda presentes en esa página web.
- Otros (por ejemplo, la opción "más como este" de Excite).

Si bien no se le permite al usuario limitar el número de referencias recuperaradas, este sí puede establecer cuántas referencias desea ver por pantalla. Los resultados suelen aparecer de diez en diez, pero se puede cambiar esta cifra a veinte, veinticinco, cincuenta o cien. Aunque la medida de la exhaustividad de la recuperación en el entorno de la W3 no es concluyente, puede ser útil conocer la cantidad de documentos que, según ese servicio, responden a la consulta, por ello es frecuente que se indique el número total de referencias obtenidas. Este dato, en algunos casos, puede utilizarse para comparar el funcionamiento de distintos servicios de búsqueda, comprobar la presencia de unos temas sobre otros, la necesidad de delimitar más la búsqueda, etc.

Aunque estén ordenadas por relevancia, cuando las búsquedas devuelven gran cantidad de referencias es muy útil que estas aparezcan numeradas ya que, de otro modo, el usuario podría perderse en la larga lista de resultados.

4. Reflexiones finales

Los usuarios de los servicios de búsqueda demandan cada vez más prestaciones, facilidad de uso y posibilidades de interacción. Actualmente se están comenzando a desarrollar los mecanismos para crear herramientas innovadoras que incorporen técnicas avanzadas de análisis de textos que permiten, con buenos resultados, búsquedas en lenguaje natural o por extensión semántica. Otras técnicas resumen documentos y los clasifican en categorías temáticas de forma automática con métodos estadísticos. Asimismo, pueden sugerir sinónimos para las

9 *Para decidir en qué grado responden los documentos recuperados a la consulta formulada, además de ordenar los resultados según su relevancia, el servicio debe mostrar el valor de relevancia asignado a cada referencia recuperada respecto a la consulta planteada. Este es un dato muy importante para el usuario ya que le permite determinar el grado de pertinencia, normalmente expresado en tantos por ciento, del documento a la consulta.*

preguntas de usuario y responder a cuestiones simples. Estas prestaciones están siendo integradas progresivamente en los buscadores de la W3. En el futuro, las interfaces de usuario pueden evolucionar incluso más allá de la presentación bidimensional y tridimensional, potenciando otros sentidos, como el oído, para ayudar a explorar nuevos paisajes en las fronteras de la información [10].

La sobrecarga de información y las variaciones significativas de formatos y estructuras de bases de datos, la riqueza de los medios de información (texto, audio, vídeo) y una abundancia de contenido de información multilingüe han creado severos problemas de interoperabilidad de la información. Por ello hay varios campos de experimentación abiertos para mejorar estas circunstancias, entre ellos, la aproximación mediante la inteligencia artificial [11]. Aplicando técnicas redimensionables desarrolladas en diferentes subáreas de la inteligencia artificial y campos relacionados, tales como segmentación e indización de la imagen, reconocimiento de voz, procesamiento del lenguaje natural, sistemas neuronales, aprendizaje automático, *clustering* y categorización y agentes inteligentes [12]. Sin embargo, en la actualidad, los programas que realizan automáticamente las tareas de análisis e indización de recursos en Internet presentan numerosas dificultades y aún están lejos de alcanzar el grado de calidad de los indizadores humanos.

Referencias

1) 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.

2) Lynch, C. Searching the Internet [en línea]. *Scientific American*. marzo 1997. <http://www.sciam.com/0397issue/0397intro.html> [Consulta: 14 de marzo de 1998].

3) Díez Ferreira, M. Buscadores temáticos. *IWorld*. 2(1)54-60, 1998.

4) "Guía completa para encontrar los recursos de la Web" *PC Magazine: edición española* 11(11)147-172, 1998.

5) Eagan, A., L. Bender. Spiders and Worms and Crawlers, Oh My: searching on the World Wide Web [en línea]. *Untangling the Web: Proceedings of the Conference.* April 26 1996, University of California, Santa Barbara. <http://www.library.ucsb.edu/untangle/eagan.html> [Consulta: 10 de junio de 1996].

6) Jakob, D. Finding Information on the Web [en línea]. *Network Notes*, 15. 10 oct. 1995. <http://www.nlc-bnc.ca/publications/netnotes/notes15.htm> [Consulta: 22 de mayo de 1996].

7) 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.

8) Schwartz, C. Web search engines. *Journal of the American Society for Information Science* 49(11)973-982, 1998.

9) Sullivan, D. (ed.) Counting clicks and looking at links [en línea]. *The search engine report* 21(4), agosto 1998. <http://searchenginewatch.com> [Consulta: 5 de agosto de 1998].

10) Hearst, M. A. Interfaces for searching the Web. *Scientific American* [en línea]. marzo 98. <www.sciam.com/0397issue/0397hearst.html> [Consulta: 14 de mayo de 1998).

11) Chen, H. Introduction. *Journal of the American Information Society for Information Science* 49(7)597-581, 1998.

12) Chen, H. *et al*. A smart itsy bitsy spider for the Web. *Journal of American Society for Information Science* 49(7)605-618, 1998.

Recibido: 1 de noviembre de 1999.

Aprobado: 27 de diciembre de 1999.

María Dolores Olvera Lobo

Universidad de Granada
Facultad de Documentación
Campus de Cartuja
Colegio Máximo
18071 Granada, España
Correo electrónico: <molvera@platon.ugr.es>