RENDIMIENTO DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN EN LA WORLD WIDE WEB: REVISIÓN METODOLÓGICA

M.ª Dolores Olvera Lobo*

Resumen: Este estudio pretende contribuir a establecer una metodología para la evaluación de la recuperación de información de las herramientas de búsqueda en el entorno de la World Wide Web. Se detalla el método diseñado (y aplicado con éxito), para evaluar los resultados de las búsquedas, adaptando las técnicas tradicionales de evaluación a las particularidades de la Web y empleando las medidas de la precisión y exhaustividad, basadas en la relevancia, para los 20 primeros resultados recuperados.

Palabras clave: evaluación de la recuperación de información; World Wide Web; buscadores web.

Abstract: This study is an attempt to establish a methodology for the evaluation of information retrieval with search engines in the World Wide Web. The method, which is explained in detail, adapts traditional techniques for evaluating web peculiarities and makes use of precision and recall scores, based on the relevance of the first 20 results retrieved. This method has been successfully applied to the evaluation of ten different search engines.

Keywords: information retrieval evaluation; World Wide Web; Web search engines.

1 Introducción

Los años 90 asisten a la aparición de un nuevo fenómeno de gran repercusión a todos los niveles y, por supuesto, en las Ciencias de la Información. La World Wide Web (W3), Malla Mundial Multimedia o telaraña mundial de información, creada en 1990, es el acontecimiento que más ha contribuido a popularizar y extender el uso de la red. Sin embargo, Internet, y en especial la W3, no se creó en un principio para atender la publicación y recuperación organizada de información. Su amplio desarrollo y crecimiento posterior dificultan la localización de los documentos pertinentes y ha suscitado la necesidad de contar con herramientas de búsqueda que faciliten esta tarea. Este mundo multimedia de gran riqueza informativa requería de nuevas herramientas de búsqueda, más eficaces y sofisticadas, que permitiesen explotar todas sus posibilidades. La respuesta fueron en primer lugar los índices o directorios y, posteriormente, los denominados motores de búsqueda o buscadores, las «estrellas» de la localización de información en Internet.

Los directorios *(classified lists)* son un tipo de servicios de búsqueda creado generalmente con intervención humana, donde las páginas web se presentan organizadas temáticamente y pueden consultarse mediante ojeo o navegación a través del directorio, o bien, planteando consultas por palabras clave mediante su motor de búsqueda inter-

* Universidad de Granada. Facultad de Documentación. Correo-e: molvera@platon.ugr.es.
Recibido: 1-7-99. Segunda versión: 29-11-99.

no. Ésta es la baza fundamental de los servicios de búsqueda basados en directorios como Yahoo! y LookSmart. Los directorios constituyen una opción a la búsqueda basada en palabras clave. Son una forma organizada de navegar por la información, ya que clasifican los recursos incluidos en su base de datos para facilitar el acceso a los mismos. Un directorio temático es una base de datos de documentos web compilados por el personal empleado en ese servicio de búsqueda, ayudado en muchas ocasiones por los propios creadores de esos documentos *(webmasters)* y por *robots* de localización automática de recursos en la red. Los recursos se organizan en grandes divisiones temáticas que se subdividen en categorías más específicas sucesivamente. El directorio puede —y suele— ser selectivo en la elección de los servidores que incluye en la base de datos ya que, dado que el proceso de indización y organización de los recursos se realiza principalmente de forma manual, es imposible para un servicio de estas características abarcar toda la información disponible en la telaraña. Por tanto, un claro inconveniente es que recogen una proporción relativamente reducida de documentos en relación con todos los existentes en la W3.

Los motores de búsqueda o buscadores *(query-based engine)* son las herramientas de consulta que permiten formular preguntas en la W3 y localizar la información que se necesita. Éstos cuentan con *robots* de búsqueda —programas «inteligentes» que localizan automáticamente los documentos presentes en la red— y potentes programas de indización —que indizan autónomamente cada página formando inmensas bases de datos—. Todos los buscadores de la W3 presentan una estructura similar: un robot o araña, es decir, un programa que cruza la Web 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.

Los directorios y los buscadores surgieron, en un principio, como dos tipos diferentes e independientes de servicios de búsqueda pero los últimos, además de posibilitar la consulta en su base de datos compilada automáticamente, tienden a incorporar también un directorio temático porque son muy apreciados por los usuarios y suponen un valor añadido a las posibilidades del buscador. Por otra parte, los directorios han ido incorporando sofisticadas prestaciones de búsqueda similares a las de los buscadores y han comenzado a ofrecer, además, acceso a éstos, por si el usuario no resuelve la necesidad de información en el propio índice. Dadas las ventajas que presentan para los usuarios tanto los directorios como los buscadores, la tendencia actual se dirige a incorporar ambas posibilidades dentro de un mismo servicio. De esta forma, el *browsing* —a través de los recursos organizados en forma de índice temático— y la búsqueda basada en términos van hoy juntos y son la forma predominante de recuperación de información en Internet (1).

La creciente cantidad y, sobre todo, calidad de las prestaciones de los buscadores, contribuyeron desde un primer momento a hacer estas herramientas imprescindibles para los usuarios. Las continuas mejoras (por ejemplo, la búsqueda por conceptos o utilizar el «índice de popularidad» de los documentos) hacen que hoy se hable de una nueva generación de buscadores, aunque sus cambios y su evolución han sido y son constantes. Su éxito y su calidad dan lugar a una variada gama de herramientas de consulta muy bien acogidas por los usuarios, tales como los servicios especializados, los metabuscadores o los agentes personales de búsqueda, entre otros.

El objetivo fundamental de este trabajo consiste en diseñar y desarrollar una propuesta metodológica válida para la evaluación de la recuperación de información (RI) en la W3. Este objetivo se encamina a la constatación de que las técnicas de evaluación usadas tradicionalmente son perfectamente adaptables y aplicables al entorno de la W3, aunque la tipología documental sea muy diferente, puesto que están, conceptualmente, bien fundamentadas. Debido a que, en Internet, tanto los buscadores como la propia información son muy dinámicos, la finalidad principal de este estudio no es la de determinar cuál es el mejor buscador actualmente presente en la red, puesto que las conclusiones a las que se puede llegar son poco perdurables, sino desarrollar un método que pueda aplicarse a la evaluación de los resultados por éstos ofrecidos, permitiendo así la realización de estudios periódicos que determinen y analicen la evolución de estas herramientas de búsqueda tan apreciadas por los usuarios.

2 Estudios relacionados

La novedad del tema y el interés que despierta entre los usuarios han favorecido la aparición de numerosos estudios comparativos de buscadores, principalmente en revistas de divulgación, aunque éstos no suelen estar realizados por investigadores o profesionales de las Ciencias de la Información. Habitualmente se trata de evaluaciones a muy pequeña escala, que abarcan pocos servicios de búsqueda, plantean pocas consultas o analizan muy pocos de los resultados obtenidos. En su mayoría presentan un carácter meramente descriptivo, y en ellos se llega a conclusiones contradictorias. Los estudios con un enfoque más cuantitativo son mucho más escasos y tampoco suelen indicar detalladamente la metodología empleada en el experimento, lo cual les resta fiabilidad. Cuando sí la indican, el método es, a menudo, deductivamente poco coherente o inductivamente poco riguroso.

Los numerosos SRI que compiten por atraer a nuevos usuarios y su variedad de prestaciones y características propician su evaluación. Por otra parte, al no contar más que con tímidos precedentes en este entorno, las tareas de evaluación suponen un nuevo reto y un incentivo para la investigación en RI. El método de compilación de información seguido, la propia naturaleza de los motores de búsqueda y el carácter dinámico de sus bases de datos, en constante cambio y crecimiento, dificultan su evaluación. Varios son los problemas generados, algunos de los cuales están relacionados con el cálculo de la exhaustividad, el modo de tratar la interacción usuario-sistema y la existencia de gran número de registros duplicados en los resultados de búsqueda (2). Por ello, en los trabajos sobre evaluación de la recuperación, incluso en los pocos casos donde se ha realizado alguna evaluación crítica, la discusión ha sido, hasta hace muy poco, de carácter anecdótico o cualitativo, mostrando experiencias individuales, utilizando un reducido número de preguntas del test y renunciando, generalmente, al uso de las metodologías cuantitativas desarrolladas para evaluar la eficacia de los SRI. Buenos ejemplos de evaluaciones cualitativas son, entre otras, las de Lebedev (3), Leonard (4), Lindop (5) Winship (6) y Zorn (7). El tipo de estudios menos frecuente y, sin duda, más valioso es el que se centra en evaluar la recuperación, es decir, los de carácter cuantitativo. Aunque no todos los autores hacen un análisis con profundidad (7, 8, 9) hay estudios amplios y rigurosos que aportan datos de interés tanto por los resultados obtenidos como por el método aplicado (10, 11, 12, 13, 14).

Marchionini, Barlow y Hill (15) realizaron un estudio comparativo de WAIS y Dialog —como ejemplo de un sistema basado en la lógica booleana— en una base de datos de 200.000 registros bibliográficos y, aunque usaron los clásicos instrumentos de Cranfield para su evaluación, comentaron en sus conclusiones que se constataba la necesidad de nuevas medidas de evaluación orientadas a la naturaleza interactiva de sistemas como WAIS y ya aludieron a la falta de estudios sistemáticos sobre el funcionamiento de la recuperación de los sistemas en red, indicando el reto de analizar sistemas que aplicaran técnicas e interfaces diferentes, así como evaluar esos sistemas desde una perspectiva formativa y comparativa.

El estudio de Ding y Marchionini (11), que evaluaba InfoSeek, Lycos y OpenText, se basó en Cranfield, introduciendo algunas variaciones. Realizaron una detallada comparación de las características de los sistemas y una evaluación de la eficacia de la recuperación. Se utilizaron sólo cinco preguntas, considerando los primeros 20 registros recuperados por consulta en cada motor de búsqueda y asignando sus propios juicios de relevancia, con seis niveles. El funcionamiento se evaluó usando la precisión y *salience* (una nueva medida propuesta por los autores), definida como la suma de las puntuaciones de relevancia de las referencias obtenidas en cada servicio dividida por la suma de la puntuación para todos los servicios. No utilizaron la exhaustividad. Lycos y OpenText fueron considerados superiores a InfoSeek.

Courtois, Baer y Stark (16) evaluaron el funcionamiento de varios buscadores mediante tres preguntas. Para ello procedieron a identificar recursos de información que ellos esperaban que los motores serían capaces de identificar, basándose en la experiencia de los autores, más que en una evaluación cuantitativa de la exhaustividad. Su comparación de CUIW3 Catalog, Harvest, Lycos, OpenText, WebCrawler, W3Worm y Yahoo mostró que sólo Lycos y Opentext Index identificaban toda la lista de recursos esperados. En un estudio posterior, Courtois (8) incluyó información de nuevos buscadores, como Altavista y Excite.

Leighton (17) planteó ocho consultas de diferente dificultad en Infoseek, Lycos, WebCrawler y WWWWorm. Los mejores resultados en cuanto a precisión y tiempo de respuesta fueron los de Lycos e Infoseek. Posteriormente, Leighton y Srivastava mejoraron y ampliaron este estudio, comparando Altavista, Excite, Hotbot, Infoseek y Lycos (12). Usaron para ello diez preguntas planteadas en el servicio de referencia de una biblioteca universitaria y añadieron cinco preguntas adicionales de otros estudiantes. Aplicaron sus propios juicios de relevancia y compararon los resultados utilizando varias medidas de eficacia basadas en la relevancia, pero no usaron la exhaustividad porque consideraban prácticamente imposible determinar el número total de páginas relevantes a una pregunta concreta existente en la W3. Además, efectuaron un riguroso análisis estadístico de todos los datos que se presentaban. En este caso, fueron Altavista, Excite e Infoseek los que ofrecieron resultados más relevantes. En 1999 los autores publican nuevamente este estudio sin introducir revisiones ni modificaciones al método que habían propuesto con lo que demuestra su plena vigencia (13).

Chu y Rosenthal (10) evaluaron Altavista, Excite y Lycos. Usaron diez preguntas de referencia reales —de diferentes niveles de complejidad— obtenidas a partir de consultas planteadas en bibliotecas universitarias y seleccionadas con el fin de analizar las siguientes características de los servicios de búsqueda: tiempo de respuesta, precisión (calculada para los primeros 10 documentos resultantes de cada consulta), opciones de presentación de los resultados, documentación del sistema e interfaz. La eficacia del

sistema fue medida a partir del tiempo de respuesta y la precisión usando las valoraciones de relevancia binaria (sí/no) de los registros recuperados. Tampoco consideraron la exhaustividad, por las mismas razones que Leighton y Srivastava (12). Chu y Rosenthal concluyeron que Altavista es la mejor elección para usuarios que necesitan alta precisión, mientras que factores tales como documentación e interfaz de usuario pueden estar basados en preferencias personales.

Randall (18) examinó 14 buscadores a partir de una serie de consultas que abarcaban desde cotizaciones de bolsa hasta ficheros de sonido de música independiente. Los buscadores fueron evaluados basándose en su facilidad de uso *(usability),* teniendo en cuenta la interfaz, las instrucciones o ayuda en las búsquedas, su adaptabilidad y su eficacia *(effectiveness),* considerando el número y precisión *(accuracy)* de los resultados. Infoseek fue el mejor puntuado, seguido por Lycos, WebCrawler y WWWWorm. Scoville (9) llevó a cabo un estudio de buscadores, contando el número de referencias obtenidas en consultas específicas y determinando la proporción de referencias relevantes entre los diez primeros resultados. Excite, Infoseek y Lycos fueron los mejor puntuados.

Venditto (19), por su parte, analizó Altavista, Infoseek, Lycos, Opentext, WebCrawler y WWWWorm utilizando docenas de términos de búsqueda durante un periodo de dos semanas. La relevancia fue determinada para los primeros 25 resultados de cada consulta. La eficacia de los buscadores se evaluó identificando documentos conocidos en la red sobre un tema particular, a modo de «documentos fuente». A partir de ahí se diseñaban las consultas en lenguaje natural y se comprobaba si los buscadores recuperaban las referencias conocidas. Venditto observó que todos los buscadores funcionaban bien cuando se les planteaban consultas simples pero descubrió que las diferencias surgían una vez que se utilizaban consultas más complejas. InfoSeek fue el que ofreció resultados de búsqueda más relevantes mientras que Altavista destacó por su alto índice de exhaustividad.

Tomaiuolo y Packer (14) realizaron uno de los estudios más extensos en cuanto al número de consultas empleadas, donde consideraron más de 200 temas. Las consultas fueron planteadas en Magellan Point, Lycos, InfoSeek y Altavista, teniendo en cuenta las diez primeras referencias resultantes en cada consulta, para las que se determinaba su relevancia. Altavista fue el que ofreció mejores resultados. Otro estudio de estos autores (20) identificó páginas web relevantes para una serie de consultas (de nuevo, «documentos fuente»). Después ofrecieron las preguntas a una serie de voluntarios con distinta experiencia en la W3, a los que se les pidió que plantearan las búsquedas relativas a esos temas en Altavista y Opentext y que dieran una puntuación tomando como base los documentos fuente encontrados. Este método ofrece más información sobre el funcionamiento de los dos SRI con relación a la exhaustividad que sobre la eficacia de las búsquedas realizadas. La necesidad de estudios centrados en los usuarios ya ha sido manifestada por Dong y Su (21).

Dania y George Meghabghab (22), compararon cinco buscadores, utilizando cinco preguntas y varias medidas comparativas —aunque no la exhaustividad— incluyendo definiciones de precisión así como páginas web relevantes. Desai (23) comparó trece buscadores mediante una sola pregunta (su propio nombre) y, aunque se conocían 24 páginas web relevantes creadas por el autor, éste no calculó la exhaustividad ni la precisión, limitándose únicamente a ofrecer directamente un recuento de las referencias obtenidas. El destacable trabajo de Dong y Su (21) realiza una revisión de los estudios sobre evaluación de buscadores en Internet llevados a cabo hasta esa fecha.

Clarke y Willett (24) consideran que la exhaustividad es una medida también importante, además de la precisión, y que prácticamente ningún estudio de este tipo la tiene en cuenta. Por ello, los autores desarrollan y aplican una metodología que evalúa tanto el funcionamiento de la precisión como de la exhaustividad, además de la *cobertura,* definida como el número total de páginas relevantes localizadas por un buscador dividido por el número total de páginas relevantes encontradas entre todos los sistemas estudiados. Los buscadores analizados fueron Altavista, Excite y Lycos. La metodología, en este caso, se basaba directamente en la aproximación a la colección de documentos para la evaluación que ofreció la base principal para la investigación en RI desde los primeros experimentos del proyecto Cranfield y que —según los autores— es totalmente vigente, ya que se sigue usando en las series actuales de experimentos como la Conferencia de Recuperación de Textos, TREC (Text REtrieval Conference). Se utilizaron 30 preguntas y se analizaron los diez primeros resultados para cada consulta y buscador. Los cálculos de exhaustividad y precisión se realizaron a partir de un *pooling* de los documentos relevantes recuperados entre los tres buscadores. Altavista dio mejores resultados en precisión y cobertura y Excite la mejor exhaustividad.

Por último, Gordon y Pathak (25) publican un trabajo sobre el rendimiento en la recuperación de varios buscadores de la W3 basándose en las medidas de la exhaustividad y precisión para los veinte primeros documentos recuperados, si bien utiliza una técnica de evaluación que extrapola esos resultados a doscientos documentos para cada SRI analizado. Los autores analizan ocho buscadores (Altavista, Excite, InforSeek, Open Text, Hotbot, Lycos y Magellan) utilizando treinta y tres preguntas, y realizando diversos tests. De forma global Altavista y Opentext obtuvieron los mejores resultados frente a Yahoo como el peor situado.

Los trabajos aparecidos sobre los SRI en la W3 cubren, por tanto, una gama bastante extensa de posibilidades. Un rápido repaso a la bibliografía existente permite observar publicaciones y recursos web relacionados con los aspectos comerciales, de gestión y financiación de los servicios de búsqueda. Con frecuencia se publican, en revistas profesionales o bien como páginas web (a cargo de bibliotecas, usuarios individuales o de las empresas de los propios buscadores), guías de motores, tablas comparativas de características, consejos prácticos de búsqueda, etc. Asimismo, se celebran encuentros, seminarios, reuniones científicas y congresos sobre este tema en el ámbito de diferentes disciplinas científicas, como la Informática, Ingeniería, Biblioteconomía y Documentación, etc.

Al principio, los artículos sobre buscadores prácticamente se circunscribían al ámbito de revistas centradas en analizar e informar sobre aspectos diversos de las tecnologías de la información —*Internet World, Online, Database* y otras— o de revistas de informática de amplia difusión (*PC World, PC Magazine, PC Computer, PC Week,* etc.), ya que el interés por la búsqueda de información en Internet no se circunscribe únicamente al área de influencia de la Documentación o disciplinas afines, sino que afecta e interesa a cualquier usuario de la red. Sin embargo, como se ha señalado, los nuevos SRI presentes en la telaraña han abierto una nueva vía para la investigación. Un hecho significativo es que han comenzado a proliferar —sobre todo en los dos últimos años— otros muchos trabajos, publicados en revistas que difunden resultados de investigación sobre diferentes aspectos de estos SRI, como *JASIS* e *Information Processing and Management,* aunque también en *Aslib Proceedings, ARIST, Electronic Li-*

brary, Computers in Library, y en el ámbito español en *El Profesional de la Información* o la *Revista Española de Documentación Científica,* así como en actas de congresos y reuniones científicas. Por último, hay que señalar que la evaluación de los buscadores necesita del desarrollo y establecimiento de metodologías adaptadas a sus exigencias que reflejen, de forma realista y rigurosa, el funcionamiento y utilidad de estas herramientas.

3 Método de evaluación

Cuando accede a un buscador, 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, indicación de la etiqueta HTML donde se han de encontrar los términos en el documento y demás datos que se consideren necesarios para delimitar y centrar la consulta. Una vez procesada, el buscador muestra los resultados ordenados según su relevancia probable relativa a la pregunta planteada. En un principio, la mayor parte de los buscadores tenían un carácter internacional, proporcionando acceso a recursos ubicados en servidores dispersos por todo el mundo, y general, ofreciendo informaciones relativas a los más diversos contenidos. Sin embargo, también han surgido buscadores que poseen robots programados para localizar e indizar informaciones que se ajustan a un patrón temático específico y limitan el descubrimiento y localización a los recursos apropiados. Son los buscadores especializados. Algunos autores opinan que, ante la avalancha constante de información, este tipo de buscadores, más selectivos, se impondrá en el futuro y que quizá únicamente sobrevivan unos cuantos buscadores generales (26, 27). Sea como fuere, el método propuesto permite su aplicación, tanto a los buscadores generales internacionales como a los buscadores especializados, aunque en éste último caso habría que prestar especial atención al tipo de preguntas planteadas para la realización del estudio.

El método consta de cinco etapas principales:

a) Determinación de las necesidades de información de los usuarios.
b) Elaboración del enunciado de búsqueda.
c) Realización de las consultas.
d) Valoración de la relevancia.
e) Análisis de los resultados.

El proceso de evaluación se inicia, pues, con la elaboración de las ecuaciones de búsqueda mediante la sintaxis correspondiente a partir de las necesidades de información planteadas por los usuarios. Tras realizar las consultas en los buscadores de Internet, los asesores externos valoran la relevancia de los ítemes recuperados. Finalmente, se analizan los resultados conforme a las medidas de exhaustividad y precisión.

3.1 Los usuarios y sus necesidades de información

En estudios como los de Clarke y Willet (24), Desai (23) y Leighton y Srivastava (12, 13) los propios investigadores proponían las preguntas para interrogar al sistema, lo que puede conllevar un sesgo y una falta de imparcialidad, al menos, potencial. Aun-

que importantes proyectos como la Conferencia de Recuperación de Textos, TREC, usan la colaboración de asesores externos para esta tarea (28, 29), la tendencia en la evaluación del funcionamiento de los SRI en Internet se orienta a recoger preguntas del servicio de referencia de bibliotecas (10, 12, 13, 14) o de estudiantes (12, 13, 24), es decir, de usuarios reales de información. Algunos servicios de búsqueda en la W3 ofrecen la posibilidad de observar, en tiempo real, las consultas que están realizando otros usuarios. Éste es el caso de Webcrawler o de Metaspy el cual muestra las búsquedas planteadas en el metabuscador Metacrawler. Sin embargo, y aunque puede resultar muy útil, los investigadores no suelen utilizar este método para recopilar las preguntas ya que presenta varios inconvenientes. Por ejemplo, los usuarios no siempre expresan de la forma más adecuada posible sus necesidades de información en la ecuación de búsqueda planteada, con lo que no siempre resultaría fácil saber qué se quiere buscar realmente. Por otra parte, la gran mayoría de las ecuaciones de búsqueda planteadas por los usuarios son demasiado genéricas, es decir, contienen uno o dos términos e incluso contienen errores tipográficos y/o gramaticales y no se aprovechan al máximo las prestaciones de los programas, puesto que no hacen uso de los delimitadores y operadores, todo lo cuál conduce a que los propios perfiles empleados por los usuarios suelen ser difíciles de extrapolar a un estudio riguroso de evaluación.

La selección de las preguntas es un aspecto clave, ya que, en gran medida, de ella depende el éxito o fracaso de la prueba. Las preguntas ofrecen el punto de partida para realizar las consultas, para controlar el proceso de búsqueda y para valorar los resultados ofrecidos por el sistema (30). Las preguntas deberían presentar las siguientes características, algunas ya señaladas por otros autores como deseables en la realización de pruebas de evaluación de la RI de buscadores en la W3 (6, 17): que sean preguntas sobre las que, muy probablemente, haya recursos en la W3; que constituyan una combinación de preguntas «fáciles» —con un alto nivel de respuesta— y «difíciles» —con resultados más restringidos— en relación a la cantidad de recursos que sobre ellas se pudieran encontrar; que unas preguntas sean de temas académicos y/o especializados y otras de temas más comunes y que se trate de preguntas heterogéneas, relacionadas con temas diversos.

Entre los investigadores que se han interesado por los buscadores de páginas web no hay acuerdo en cuanto al número de preguntas a utilizar. En algunos trabajos se han utilizado únicamente una, como en Desai (23) o dos preguntas, lo que es claramente insuficiente. Courtois, Baer y Stark (16) y Zorn (7) formulan tres. Ding y Marchionini (11) y Westera (31) usan cinco. Leighton (17) emplea nueve y posteriormente amplía este número a quince (12, 13), aunque Gordon y Pathak (25) llegan hasta treinta y tres. El trabajo de Tomaiuolo y Packer (14), el más exhaustivo en este punto, usó 200 preguntas.

Para un estudio de estas características, el uso de veinte preguntas aquí se considera suficiente y representativo del funcionamiento de los diferentes SRI en la recuperación de información.

3.2 La sintaxis de búsqueda

El reto principal al realizar una consulta es conseguir que la pregunta recupere los documentos que se consideran realmente relevantes (17). La elaboración de la sintaxis

de búsqueda es un aspecto fundamental. Para realizar las consultas en los SRI, las preguntas son traducidas a las expresiones o enunciados de búsqueda correspondientes. Dicha expresión de búsqueda puede constar de varios elementos: términos, operadores lógicos, uso de paréntesis, truncamiento, formulación de la búsqueda en lenguaje natural, etc. En este sentido, una cuestión de trascendencia en el proceso de RI y que ha generado una línea de investigación ciertamente interesante es la selección y eficacia de los términos de búsqueda utilizados en la interacción con el sistema de recuperación (32).

Una buena decisión es la de realizar las búsquedas en inglés, por ser la lengua de uso mayoritario en Internet, lo que aumenta las posibilidades de encontrar información en las consultas planteadas, sobre todo en los buscadores generales e internacionales, caso al que este estudio se refiere principalmente. Para plantear las consultas se ha de elegir entre la expresión booleana o lo que Leighton y Srivastava (12, 13) denominan «expresión de búsqueda desestructurada», esto es, consultas en lenguaje natural. La naturaleza variopinta de las preguntas demanda sintaxis de búsqueda diferentes —booleana, de frase, de un término, etc.— y se ha de escoger la que en cada caso resulte, probable e intuitivamente, más adecuada sin descuidar que se ha de contribuir a la homogeneidad de los resultados para facilitar su comparación. Por esto, es una buena opción seleccionar la sintaxis y el modo de funcionamiento del motor con formatos más simples.

No hay pues una única manera de plantear la consulta, ya que para elaborar la expresión de búsqueda hay que decidir cuántos y qué términos de la pregunta incluir, además hay que elegir si se formula la pregunta en lenguaje natural o usando la lógica booleana y, en este último caso, el modo de plantearla, además de otras opciones del programa —uso de mayúsculas, truncamiento, etc. —. Esto da lugar a expresiones de búsqueda de distinto tipo (10): unas utilizan términos más generales y otras más específicos; algunas constan de una sola palabra, otras, constituyen frases de búsqueda; unas usan la lógica booleana, otras se plantean como búsquedas de frase y otras como búsquedas en lenguaje natural; las hay que son nombres de persona; en algunos casos se utiliza la mayúscula y el truncamiento, etc.

3.3 Ejecución de las búsquedas

En esta fase se plantean las preguntas en todos los servicios de búsqueda a analizar. Sólo una mínima parte de los estudios publicados sobre evaluación de los SRI de la W3 describen con detalle el método seguido para evaluar la recuperación de información. Los que lo hacen, no siempre indican datos de interés tales como el tiempo empleado en el desarrollo de las consultas (10). No obstante, para conseguir que un examen de estas características sea realmente riguroso es deseable formular la misma pregunta en todos los buscadores sin que transcurra demasiado tiempo entre el uso de los distintos motores. Idealmente la misma consulta debería ser simultánea en todos los servicios evaluados, para evitar la inclusión de páginas insertas con posterioridad a la primera de las búsquedas de una misma serie (12). Lo normal es realizar las búsquedas para una misma pregunta con un intervalo máximo de un buscador a otro de veinte minutos (11) o de media hora (12).

Otra cuestión clave es el análisis de los resultados, que debe hacerse tan rápida-

mente como sea posible tras obtenerlos, porque el retraso aumenta las probabilidades de modificación de las páginas recuperadas. Quien ejecuta el estudio pudiera computar una página como inactiva cuando en el momento de la búsqueda no lo era (12,13). Asimismo, hay que tener en cuenta las horas punta electrónicas, para minimizar el «efecto red» y disponer de mayor ancho de banda, ya que uno de los aspectos a considerar en el estudio de la eficacia del funcionamiento de los SRI es el tiempo de respuesta.

3.4 Los juicios de relevancia

En este paso se determina la relevancia de cada documento recuperado. Se considera ítem relevante todo aquél que versa sobre el tema de la pregunta, es decir, que responde a las necesidades de información tal y como habían sido expresadas por los usuarios. La tarea de juzgar la relevancia puede ser realizada por asesores externos, como se lleva a cabo en TREC (28, 29, 33, 34).

La mayor parte de los buscadores de la W3 —y, por supuesto, los más reputados— ordenan los resultados en función de su relevancia respecto a la pregunta planteada. Los supuestamente mejores resultados aparecen siempre en la parte superior de la lista de referencias. Las medidas utilizadas en muchos de los estudios publicados para el análisis de buscadores de información en la W3 suelen basarse en la precisión de los diez (9, 10, 14, 35), veinte (11, 12, 13, 25, 36) e incluso veinticinco primeros (19) resultados. Considerar los veinte primeros resultados arroja datos suficientes para realizar la evaluación. Además, se ha de acceder a todos ellos para juzgar la relevancia desde el documento web a texto completo, lo que se considera indispensable para tener suficientes elementos de juicio al determinar su adecuación respecto a la pregunta planteada y evaluar de forma rigurosa la relevancia. No obstante, otros estudios, por demás serios y ambiciosos, como el de Tomaiuolo y Packer (14) no realizaron esta comprobación sobre el documento completo.

Para evaluar la relevancia se utiliza una escala constituida por varias categorías: a) enlaces duplicados, inactivos e irrelevantes, todos ellos puntuados con 0; b) enlaces técnicamente relevantes, que reciben un punto; c) enlaces potencialmente útiles, a los que los evaluadores asignan dos puntos; y d) los enlaces probablemente más útiles, que reciben tres puntos (10, 11, 12, 13, 24). Esto propicia y facilita la realización de diferentes tipos de pruebas y análisis posteriores según los diferentes grados de relevancia de los ítemes recuperados: por ejemplo, considerando todos los documentos relevantes (puntuados con 1, 2 o 3), los documentos potencialmente relevantes u óptimos (de 2 o 3 puntos) o únicamente los documentos óptimos (puntuados con 3), e incluso otras pruebas en las que se eliminen los duplicados de los cálculos que se realicen. A continuación se describe con más detalle la escala de relevancia.

Duplicados

Si el enlace en cuestión tiene el mismo URL (Uniform Resource Locator) básico que un enlace anterior de la lista de resultados, se lo considera en la categoría de duplicados, independientemente de sus otras cualidades (inactivo, irrelevante o válido). Esta categoría incluye variantes muy obvias pero otras son más sutiles: si un nombre del directorio en el URL está en mayúsculas en un caso pero no en otro, cuenta como duplicado. Los espejos (*mirror sites* o alias), servidores idénticos que tienen direccio-

nes IP *(Internet Protocol)* o directorios diferentes, incluso cuando dos archivos son el mismo o versiones ligeramente diferentes, no se consideran como duplicados.

Inactivos

Se consideraron enlaces inactivos los que se encuentran entre los casos siguientes:

- Error 404: el servidor ha sido contactado pero no se consigue localizar ese fichero.
- Error 603: el servidor no responde, para los errores 404 y 603 se comprueban los enlaces varias veces —por ejemplo, en un periodo de una semana.
- Mensajes que indican que el acceso a la página está prohibido o que se necesita clave de acceso.
- Mensajes que anuncian que la página deseada ha sido eliminada o trasladada a otro servidor.

Relevantes

Los criterios generales para asignar un valor de 0 a 3 en el juicio de relevancia de los documentos recuperados son:

0. Una página web que no satisface la pregunta ni recoge los términos de la ecuación de búsqueda.
1. Una página técnicamente adecuada pero no útil. En este caso, el documento recoge, en el código HTML (HyperText Markup Language), las diferentes partes de la pregunta pero no en el contexto adecuado. El documento puede contener los términos o componentes de la pregunta, pero bastante alejados entre sí y, aunque la consulta puede plantearse mediante una expresión booleana o bien en lenguaje natural, para que la página sea relevante, habitualmente los elementos de la pregunta han de aparecer próximos. También se asigna 1 punto a páginas que mencionan el tema en el contexto adecuado pero que sólo contienen un mínimo de información realmente útil.
2. Páginas que pueden tener alguna utilidad, aunque no necesariamente, para quien plantee la búsqueda. Obtienen 2 puntos las páginas que no abordan el tema con profundidad o que se centran en algún aspecto específico del mismo. También obtienen 2 puntos las páginas con al menos un enlace a otra página a la que se asignan 3 puntos, aunque la primera no contenga otras informaciones relevantes.
3. Páginas web que, probablemente, serían útiles para quien plantee la consulta. Pueden tratar el tema extensamente y con detalle o contener enlaces a otros documentos que tratan ese tema u ofrecen una bibliografía de páginas web o «webibliografía».

Además de estos criterios generales se habrá de establecer, de manera clara y concreta, la puntuación de 0 a 3 para todas las preguntas formuladas en el desarrollo del estudio en función del tema de cada una.

3.5 Análisis de los resultados

Como se ha indicado, la medida utilizada para analizar los resultados es la «precisión y exhaustividad de los veinte primeros» documentos recuperados. En la W3, la medida real de la exhaustividad no se puede calcular debido a la dificultad de determinar de forma absoluta el número total de páginas relevantes para una pregunta específica (12). En otras palabras, el número total de enlaces cambia muy rápidamente y es prácticamente incognoscible tanto por la naturaleza dinámica de la W3 (10) como por su tamaño. Por ello, son mucho menos frecuentes las evaluaciones que incorporan, además de la precisión, también la exhaustividad en buscadores web.

Algunos estudios (24) sin embargo, han considerado la exhaustividad de las herramientas de búsqueda en la W3 mediante una metodología experimental basada directamente en el «método de la colección de documentos para la evaluación», aplicando técnicas de muestreo donde los valores de relevancia se establecen únicamente para un subconjunto de ítemes en una colección. Otra posibilidad es procesar una pregunta concreta mediante varias búsquedas y métodos de recuperación diferentes o bien mediante un metabuscador, asumiendo que todos los documentos relevantes serán recuperados en estas diferentes búsquedas. Los resultados se combinan entre sí y el conjunto de documentos relevantes a esa pregunta se obtiene analizando la relevancia de cada referencia recuperada. A este método se le denomina *pooling*.

El método aquí propuesto responde a que las medidas de exhaustividad y precisión dependen fundamentalmente de la relevancia de los primeros documentos recuperados en respuesta a una pregunta, es decir, aquellos documentos que presentan una mayor similitud pregunta-documento. Para esos ítemes, los valores de relevancia obtenidos a partir de juicios diferentes son bastante congruentes y los datos resultantes de la evaluación son razonablemente similares a los obtenidos mediante otros métodos (37). Aquí únicamente se evalúa la relevancia de los documentos recuperados (38), que es lo que realmente hace un usuario cuando consulta una base de datos, no de todos los documentos ni de una muestra de los mismos. Los asesores examinan, pues, las primeras veinte referencias de cada lista de resultados para determinar los valores de relevancia correspondientes a cada documento recuperado por los diferentes buscadores en respuesta a cada pregunta.

Los cálculos de precisión y exhaustividad se realizan según el método propuesto por Salton y McGill (39) para aquellos SRI que ordenan los resultados según la relevancia de los documentos a la pregunta. En este caso, los autores proponen recurrir a la evaluación por cortes, lo que Blair (38) denomina "umbral de futilidad", es decir, el punto en el que el usuario cesa de examinar la lista de resultados. En el método que se propone —y siguiendo esta línea— se considera como número total de documentos relevantes los obtenidos entre los veinte primeros. El par de valores exhaustividad-precisión se calcula para cada posición en la lista de resultados, para cada rango, usando el rango como un nivel de recuperación (Tague 30). El método desarrollado por Salton y McGill goza de gran aceptación en la comunidad investigadora y puede aplicarse a la evaluación de los resultados ofrecidos por los buscadores de Internet con los debidos ajustes que lo adapten a las particularidades de la W3 (40).

4 Conclusiones

El método empleado permite analizar la eficacia en el funcionamiento de los buscadores de la W3. Con el fin de alcanzar el objetivo propuesto, a saber, contribuir a establecer una metodología para la evaluación de la recuperación de información de las herramientas de búsqueda en el entorno de la World Wide Web, hay que recordar que los rasgos principales del mismo son:

1) La incorporación de usuarios reales que plantean preguntas reales.
2) El uso de las medidas de exhaustividad y precisión para evaluar la RI.
3) Una adaptación del método de Salton y McGill (39) para su aplicación a los buscadores web en dos sentidos:
   a) Los autores no se definen sobre cómo evaluar la relevancia, ni la escala a utilizar, ni quién ha de evaluarla: los asesores externos analizan la relevancia de los veinte primeros resultados expresada en una escala de cuatro grados;
   b) Se ha utilizado la regresión logarítmica como forma de representación de los valores de exhaustividad-precisión por ser la función que ofrece mejores resultados de ajuste y porque no supone una gran diferencia con respecto al método de representación propuesto por Salton y McGill.

El método, que fue aplicado con éxito en la evaluación de la RI de diez buscadores generales internacionales diferentes (40), produce resultados bastante razonables, que demuestran la viabilidad de adaptar técnicas ya existentes de evaluación de la recuperación de información a los servicios de búsqueda en Internet.

Referencias bibliográficas

1. 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.
2. HARTER, S. P.; HERT, C. A. Evaluation of information retrieval systems. En: Willian, M. E. (ed.), *Annual Review of Information Science and Technology,* 32, 3-94, 1997.
3. LEBEDEV, A. *Best search engines for finding scientific information on the Web.* Sept. 29, 1996. Disponible en: http://www.chem.msu.su/eng/comparison.html (consultado 12 dic. 97).
4. LEONARD, A. *Search engine: where to find anything on the net.* Disponible en: http://www.cnet.com/Content/Reviews/Compare/Search. 1996 (consultado 1 abril 97).
5. LINDOP, L. et al. Catching sites. *PC Magazine,* 6(2), 108-153, 1997.
6. WINSHIP, I. R. *World Wide Web searching tools: an evaluation.* Vine 99, 49-54, 1995. Disponible en:http://www.bubl.bath.ac.uk/BUBL/IWinship.html (Consultado 16 enero 96).
7. ZORN, P. et al. Advanced Web searching: tricks of the trade, *Online,* 20(3), 15-28, 1996.
8. COURTOIS, M. P. Cool tools for web searching: an update, *Online,* 20(3), 29-36, 1996.
9. SCOVILLE, R. Special Report: Find it on the Net!, *PC World,* 125, 14(1), 1996.
10. CHU, H.; ROSENTHAL, M. *Search engines for the World Wide Web: A comparative study and evaluation methodology,* oct. 1996. Disponible en: http://www.asis.org/annual-96/ElectronicProceedings/chu.html (consultado 5 febr. 97).
11. DING, W.; MARCHIONINI, G. A comparative study of web search service performance, *Proceedings of the ASIS Annual Conference,* 33. 136-142, 1996.
12. LEIGHTON, V. H.; SRIVASTAVA, J. *Precision among World Wide Web Search Servi-*

ces (Search Engines): Altavista, Excite, Hotbot, Infoseek, Lycos. act. 10 jul. 97. Disponible en: http://www.winona.msus.edu/is-f/library/webind2/webind2.htm (consultado 12 jul. 97).

13. LEIGHTON, V.H.; SRIVASTAVA, J. First 20 Precision among World Wide Web Search Services (Search Engines), *Journal of the American Society for Information Science,* 50 (10), 870-881, 1999.
14. TOMAIUOLO, N. G.; PACKER, J. G. An analysis of Internet search engines: assessment of over 200 search queries, *Computers in Libraries,* 16(6), 58-62, 1996.
15. MARCHIONINI, G.; BARLOW, D.; HILL, L. Comparing WAIS and boolean capabilities. *Journal of the American Society for Information Science,* 45(8), 561-564, 1994.
16. COURTOIS, M. P.; BAER, W. M.; STARK, M. Cool tools for searching the Web: a performance evaluation, *Online,* 19(6), 14-32, 1995.
17. LEIGHTON, V. H. *Performance of Four World Wide Web (WWW) Index Services: Infoseek, Lycos, Webcrawler and WWWWorm. 1995.* Disponible en: http://www.winona.msus.edu/is-f/library-f/webind.htm (consultado 2 enero 97).
18. RANDALL, N. Hide and go seek: rating Internet search engines, *PC Computing Online.* Ziff-Davis Publishing Company. 1995. Disponible en: http://www4/zdnet.com/pccomp/features/internet/subl.html (consultado 3 mar. 97).
19. VENDITTO, G. Search engine showdown, *Internet World,* 7(5), 79-86, 1996.
20. TOMAIUOLO, N. G.; PACKER, J. G. *Results of 200 subject searches in Altavista, Infoseek, Lycos, Magellan and Point, performed Oct. to Dec. 1995. 20 may 1996.* Disponible en: http://neal.ctstateu.edu:2001/htdocs/websearch.html (consultado 13 oct. 97).
21. DONG, X.; SU, L. T. Search engines on the World Wide Web and information retrieval from the Internet: a review and evaluation, *Online & CD-ROM Review,* 21(2), 67-82, 1997.
22. MEGHABGHAB, D. B.; MEGHABGHAB, G. V. Information retrieval in cyberspace. En: Whitney, G. (ed.), *The digital revolution: proceedings of the American Society for Information Science (ASIS). Mid-year meeting 1996 may 18-22,* San Diego: Information Today, 1996, p. 224-237, ISBN 1573870285.
23. DESAI, B. C. Supporting discovery in virtual libraries, *Journal of the American Society for Information Science,* 48(3), 190-204, 1997.
24. CLARKE, S.; WILLET, P. Estimating the recall performance of web search engines, *Aslib Proceedings,* 49(7), 184-189, 1997.
25. GORDON, M.; PATHAK, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines, *Information Processing and Management,* 35, 141-180, 1999.
26. DIEZ FERREIRA, M. Buscadores temáticos, *IWorld,* 2(1), 54-60, 1998.
27. Guía completa para encontrar los recursos de la Web, *PC Magazine,* edición española, 11(11), 147-172, 1998.
28. HARMAN, D. K. Overview of the Second Text REtrieval Conference (TREC-2), *Information Processing and Management,* 31(3), 271-289, 1995.
29. HARMAN, D. K. The TREC Conferences. En: Sparck Jones, K.; Willett, P. (ed.), *Readings in information retrieval,* San Francisco: Morgan Kaufmann, 1997. p. 247-256, ISBN 1558604545.
30. TAGUE-SUTCLIFFE, J. The pragmatics of information retrieval experimentation revisited, *Information Processing and Management,* 28(4), 467-490, 1992.
31. WESTERA, G. *Robot-drive search engine evaluation: overview, 4 july 1997.* Disponible en: http://www.curtin.edu.au/curtin/library/ staffpages/gwpersonal/senginestudy/ (consultado 13 oct. 97)
32. SPINK, A.; SARACEVIC, T. Interaction in information retrieval: selection and effectiveness of search terms, *Journal of the American Society for Information Science,* 48(8), 741-761, 1997.

33. HARMAN, D. K. A special conference report: the First Text REtrieval Conference (TREC-1), Rockville, Md, USA, 4-6 november, 1992, *Information Processing and Management,* 29(4), 411-414, 1993.
34. HARMAN, D. K. Text Retrieval Conferences (TRECs): providing a test-bed for information retrieval systems, *Bulletin of the American Society for Information Science,* 11-13, abril/mayo 1998.
35. MUNRO, J.; LIDSKY, D. Web search sites, *PC Magazine,* 15(21), 232, 1996. Disponible en: http://www8.zdnet.com/pcmag/iu/srchsite/test.htm (consultado 4 febr. 97).
36. GAUCH, S.; WANG, G. *Information Fusion with ProFusion,* Webnet 96 Conference, San Francisco, CA, October 15-19, 1996. Disponible en: http://www.csbs.utsa.edu:80/info/webnet96/html/155.htm (consultado 22 febr. 97).
37. SALTON, G. The state of retrieval system evaluation, *Information Processing and Management,* 28(4), 441-449, 1992.
38. BLAIR, D. C.; MARON, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system, *Communications of the ACM,* 28(3), 281-299, 1985.
39. SALTON, G.; McGILL, J. *Introduction to modern information retrieval.* Nueva York: McGraw-Hill, 1983. ISBN 0070544840.
40. OLVERA, M.D. *Evaluación de la recuperación de información en Internet: un modelo experimental. Universidad de Granada* (tesis doctoral defendida el 2 de marzo de 1999).