RENDIMIENTO DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN EN LA WEB: EVALUACIÓN DE SERVICIOS DE BÚSQUEDA *(SEARCH ENGINES)*

M.ª Dolores Olvera Lobo*

Resumen: Se han evaluado diez servicios de búsqueda: Altavista, Excite, Hotbot, Infoseek, Lycos, Magellan, OpenText, WebCrawler, WWWWorm, Yahoo. Se formularon 20 preguntas a cada uno de los 10 sistemas evaluados por lo que se realizaron 200 consultas. Además, se examinó la relevancia de los primeros 20 resultados de cada consulta lo que significa que, en total, se revisaron aproximadamente 4.000 referencias, para cada una de las cuales se calcularon los valores de precisión y exhaustividad. Los análisis muestran que Excite, Infoseek y Altavista son los tres servicios que, de forma genérica, muestran mejor rendimiento. Se analizan también los resultados en función del tipo de pregunta (booleanas o de frase) y del tema (ocio o especializada). Se concluye que el método empleado permite analizar el rendimiento de los SRI de la W3 y que los resultados ponen de manifiesto que los buscadores no son sistemas de recuperación de información muy precisos aunque sí muy exhaustivos.

Palabras clave: sistemas de recuperación de información, evaluación de SRI, Internet, World Wide Web, buscadores web.

Abstract: Ten search engines, Altavista, Excite, Hotbot, Infoseek, Lycos, Magellan, OpenText, WebCrawler, WWWWorm, Yahoo, were evaluated, by means of a questionnaire with 20 items (adding up to a total of 200 questions). The 20 first results for each question were analysed in terms of relevance, and values of precision and recall were computed for the resulting 4000 references. The results are also analyzed in terms of the type of question (Boolean or natural language) and topic (specialized vs. general interest). The results showed that Excite, Infoseek and AltaVista performed generally better. The conclusion of this methodological trial was that the method used allows the evaluation of the performance of Information Retrieval Systems in the Web. As for the results, web search engines are not very precise but extremely exhaustive.

Keywords: Information Retrieval Systems, evaluation, Internet, World Wide Web, search engines.

1 Introducción

El trabajo que se presenta, continuación de otro publicado en esta misma revista, se basa en una tesis doctoral defendida en la Universidad de Granada en marzo de 1999 (1). Este dato es de importancia puesto que las consultas planteadas en los dife-

* Universidad de Granada. Facultad de Documentación. Correo-e: molvera@platon.ugr.es
Recibido: 13-12-99.

rentes sistemas de recuperación de información de la World Wide Web (W3) para la realización de este estudio se llevaron a cabo en agosto de 1997. Debido a la constante evolución de estos servicios de búsqueda, perfectamente documentada en páginas web como la que mantiene Danny Sullivan (2), los datos que arrojaron las búsquedas tienen hoy, al igual que en la mayor parte de análisis de este tipo (3), un valor «histórico». Por ello los resultados de este estudio no serían útiles hoy día para ayudar a un usuario a elegir el buscador web a utilizar; sin embargo, esa tampoco fue en ningún momento su finalidad, ya que la investigación realizada se centró en el diseño del método de evaluación. El análisis de los buscadores generales más sobresalientes de la W3 se realizó con el objeto de ilustrar el proceso de evaluación y no para determinar cuál es el mejor buscador de Internet, ya que tanto sus motores como la propia información, son muy dinámicos en la W3 y las conclusiones a las que se pueden llegar son poco perdurables.

La selección de los buscadores a evaluar constituyó el primer paso del trabajo desarrollado. Éstos debían cumplir los siguientes requisitos:

a) que se tratara de buscadores generales de manera que su base de datos incluyese información sobre los más variados temas;
b) que tuviesen un carácter internacional, es decir, que no limitasen la información a ninguna zona geográfica y
c) que fuesen ampliamente utilizados entre los usuarios y estudiosos de Internet para que la muestra fuera lo más representativa posible.

Los buscadores objeto de este estudio fueron seleccionados en base a su popularidad. Se examinaron un total de 23 artículos publicados en revistas especializadas entre enero de 1996 y julio de 1997, donde se describían, comparaban o evaluaban las características y funcionamiento de diferentes servicios de búsqueda. Una vez realizado el análisis se encontró que los más estudiados eran, en este orden, Lycos, Altavista, Infoseek, Excite, OpenText, WebCrawler, Hotbot, Yahoo, Magellan y WWWWorm por lo que éstos son los que aquí se han considerado. Entre ellos, había buscadores «puros» (es decir, sin índice temático de páginas web en esas fechas, como Altavista y Hotbot), otros «híbridos» (como Infoseek y Lycos, que incluían, además de una base de datos compilada por un robot o araña, un directorio) y algunos que incorporan servicios de evaluación de páginas web, como Excite y Magellan. Unos destacan claramente por su directorio, como WebCrawler y Yahoo. Otros han ido en declive hasta desaparecer, como WWWWorm, o especializarse, como Opentext. Es su popularidad durante el período de muestra la que justifica su inclusión, independientemente del servicio que en ellos destaque. En todos los casos, este estudio se centra en el funcionamiento de sus motores de búsqueda.

2 Aplicación del método propuesto

El método de evaluación se organizó en torno a varias etapas. A partir de las necesidades de información planteadas por los usuarios, se elaboraron las ecuaciones de búsqueda mediante la sintaxis que se consideró más adecuada para fines de evaluación y comparación de los SRI de la W3. Luego se realizaron las consultas en los buscadores de la W3, tras lo cual los asesores externos valoraron la relevancia de los pri-

meros 20 ítemes recuperados en respuesta a cada una de las preguntas. Por último, se analizaron los resultados usando las medidas de exhaustividad y precisión.

2.1 Determinación de las necesidades de información de los usuarios

Para llevar a cabo el estudio se contó con la colaboración de diez usuarios de Internet elegidos al azar, relacionados con el ámbito académico —alumnos, profesores y bibliotecarios de la Universidad —. Esta circunstancia no perseguía determinar la naturaleza de las preguntas planteadas, como se comprobará a continuación, sino que vino provocada únicamente por las facilidades ofrecidas —y que muy difícilmente se hubiesen podido encontrar en otro entorno—. A los usuarios se les pidió que indicasen las necesidades de información que les surgieran en un periodo de dos semanas sobre cualquier tema relacionado de una u otra forma con España. Esta última condición respondía a dos motivaciones:

a) En primer lugar, poner en situación adversa a buscadores generales e internacionales, ya que España representa —y aún más en la fecha de realización de las búsquedas— una muy pequeña parte en Internet, y los sistemas de recuperación debían responder a preguntas relativas a temas no estrictamente anglosajones.
b) En segundo lugar, la voluntad de delimitar en cierta manera las búsquedas para obtener resultados mejores y más homogéneos, una de las recomendaciones en las que más se insiste en todos los buscadores consultados, dada la magnitud y variedad de documentos incluidos en sus bases de datos.

Los usuarios tuvieron que cumplimentar un formulario donde debían indicar de forma tan completa y explícita como les fuera posible la necesidad de información que pretendían resolver con su búsqueda en Internet. Realizado un examen a posteriori de dichas necesidades de información, se pudo observar que el conjunto de los temas planteados por los usuarios tenía un carácter heterogéneo y se consideró que eran representativos de diversos tipos de búsquedas que pueden plantear los usuarios en la red. Eran preguntas sobre las que, muy probablemente, había recursos en la W3, condición imprescindible puesto que no se puede evaluar sin resultados. Constituían una combinación de preguntas con un nivel de respuesta potencialmente alto y otras con resultados más restringidos. Por otra parte, unas preguntas eran de temas académicos y/o especializados y otras de temas más comunes. Como se ha indicado, el ámbito temático se limitaba a preguntas relacionadas de cierta forma con España, incluyendo preguntas sobre diferentes aspectos de nuestro país: Arte, Ciencia, Costumbres, Economía, Geografía, Historia, Política, etc.

En este estudio se decidió utilizar un total de 20 preguntas, cantidad que, a la vista del número empleado en la mayoría de los estudios, parece suficiente para realizar la evaluación. Una vez recopiladas, las preguntas resultaron ser las que se muestran en el Apéndice I. Dado que no se trataba, en ningún caso, de evaluar la cobertura de áreas temáticas en la W3 sino de analizar la eficacia en la RI de algunos de los más reputados servicios de búsqueda, se puede afirmar que estas preguntas son válidas para los fines planteados y propician la objetividad en el proceso de evaluación.

2.2 Elaboración del enunciado de búsqueda mediante la sintaxis correspondiente

Cada buscador presenta un motor de búsqueda y unas prestaciones diferentes por lo que las ecuaciones de búsqueda entre ellos también pueden ser diferentes. No obstante, para contribuir a la homogeneidad de los resultados y posibilitar su comparación, se adoptaron una serie de criterios. La primera decisión importante fue la de realizar las búsquedas en inglés por ser la lengua de uso mayoritario en Internet, lo que debía aumentar las posibilidades de encontrar información en las búsquedas planteadas. La naturaleza variopinta de las preguntas demandaba sintaxis de búsqueda diferentes —booleana, de frases, de un término, etc.— y se escogió en cada caso la que resultaba intuitivamente más adecuada. Aunque algunos autores (4, 5, 6, 7, 8) han optado por optimizar la ecuación de búsqueda según el sistema interrogado, en la mayor parte de los estudios realizados (3, 9, 10, 11, 12, 13, 14, 15, entre otros), se seleccionó la sintaxis y el modo de funcionamiento del motor con formatos más simples. Este último es el criterio que aquí se ha seguido. Sólo se optó por la búsqueda avanzada en las raras ocasiones en que la simple no permitía realizar la búsqueda por frase o mediante lógica booleana. Para plantear las consultas se prefirió la expresión de búsqueda estructurada (con operadores lógicos o delimitadores) en lugar de las consultas en lenguaje natural. Siempre que fue posible, en las búsquedas se utilizaron los limitadores + y –, en lugar de los operadores lógicos. En ocasiones hubo que elegir alguna opción específica del menú de búsqueda para plantear la consulta, como *búsqueda de todas las palabras, búsqueda de frase*, etc. En algunos buscadores no se realizó el truncamiento, por diversas razones: o bien no lo permitía, o el número de palabras así buscadas era excesivo o realizaban las llamadas «búsquedas inteligentes» o búsqueda por conceptos que ya incluye esta posibilidad.

Según se observó, las preguntas (*queries*) presentaban las siguientes características:

- Nueve preguntas (preguntas n.º 2, 7, 11, 12, 13, 15, 16, 17 y 20) eran de temas generales, que se han consignado como «ocio» y las once restantes trataban temas más específicos, señaladas como preguntas «especializadas».
- En ocho preguntas (n.º 1, 5, 6, 9, 13, 14, 17 y 19) se usaba la lógica booleana; otras (n.º 2, 3, 4, 7, 8, 10, 11, 12, 18 y 20) se plantearon como búsquedas de frase —es decir, una serie de términos que han de aparecer necesariamente juntos y en ese orden en el documento recuperado— y dos (n.º 15 y n.º 16) como búsquedas en lenguaje natural.
- Algunas de las preguntas, como la n.º 15, constaban de una sola palabra y varias (n.º 9, 10 y 20) eran nombres de persona.
- En algunos casos se utilizó la mayúscula y el truncamiento.

A riesgo de críticas, se escogió también intuitivamente la posibilidad que se consideraba la mejor formulación de cada pregunta. De estas preguntas, seis se formularon con el operador «and» o el signo +, la opción más afín al uso de este operador lógico en algunas interfaces. Las preguntas n.º 1, 17 y 20 requerían un planteamiento de búsqueda más complejo y, además de este operador, se usaron otros, paréntesis o búsquedas por frase. La de búsqueda por frase es otra categoría con bastante representación, pues consta de siete preguntas. En el Apéndice II (Protocolos de la investigación I) se presenta, a modo de ejemplo, la sintaxis para la formulación de dos de

las preguntas en los buscadores. También se detalla si hubo que elegir alguna opción específica como *all the words, search for this phrase*, etc.

2.3 Realización de las consultas en los sistemas

Las páginas web pueden modificar su contenido y ubicación en la red con facilidad. La evaluación de los buscadores se ve dificultada por el sistema de compilación de información seguido, las características de sus motores de búsqueda y el carácter dinámico de sus bases datos, en constante mutación y crecimiento. Las consultas habían de realizarse pues en condiciones especiales para adaptar la metodología respecto a bases de datos tradicionales donde la información no varía. Por tanto, para conseguir que un estudio de estas características fuera realmente riguroso había de transcurrir un intervalo mínimo de tiempo durante la realización de las búsquedas así como en el análisis de las referencias recuperadas.

Las búsquedas para una misma pregunta se realizaron el mismo día con un espacio de tiempo de 3 a 5 minutos entre los buscadores. Otra cuestión clave fue la de examinar las referencias recuperadas, para lo que los colaboradores habían de acceder al documento íntegro. Esto debía hacerse con suficiente rapidez, porque el retraso podía aumentar las probabilidades de modificación, eliminación o cambio de localización de las páginas recuperadas, y hacer menos fiable el análisis. Los resultados de cada pregunta fueron analizados en un periodo de 1 a 7 días. Durante ese periodo se accedió a los enlaces inactivos en varias ocasiones.

En resumen, se formularon 20 preguntas a cada uno de los 10 buscadores evaluados por lo que se realizaron 200 consultas. Además, se examinó la relevancia de los primeros 20 resultados de cada consulta (si bien algún buscador, en alguna ocasión, devolvió un número más reducido de resultados), lo que significa que, en total, se analizaron aproximadamente 4.000 referencias.

2.4 Valoración de la relevancia por asesores externos

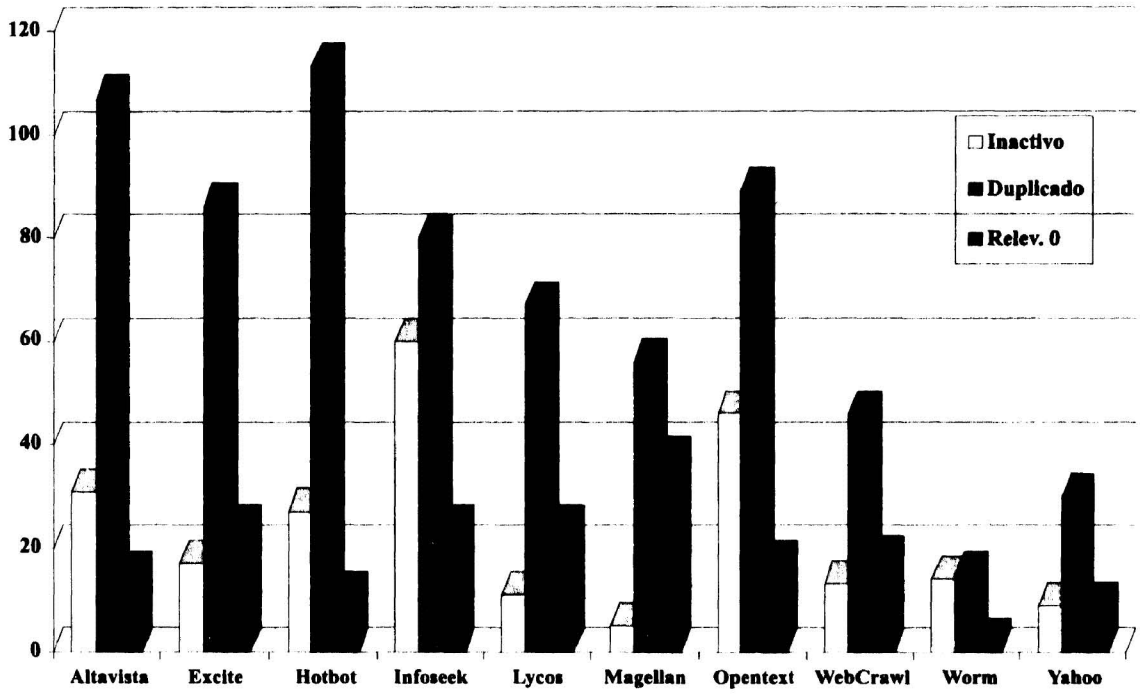
Esta tarea fue realizada por diez asesores externos, cada uno de los cuales evaluó dos de las búsquedas realizadas. Para evitar distorsiones, el mismo colaborador evaluó los resultados de todos los buscadores para una misma pregunta sin conocer de qué servicio procedían. Se consideró ítem relevante todo aquél que versaba sobre el tema de la pregunta, es decir, que respondía a las necesidades de información tal y como habían sido expresadas por los usuarios.

Se accedió a cada uno de los primeros veinte resultados recuperados para juzgar la relevancia desde el documento web a texto completo. Se pidió a los asesores que utilizaran una escala constituida por varias categorías: *a)* enlaces duplicados, inactivos e irrelevantes, todos ellos puntuados con 0, *b)* enlaces técnicamente relevantes, que recibían un punto; *c)* enlaces potencialmente útiles, a los que los evaluadores asignaban dos puntos y *d)* los enlaces probablemente más útiles, que recibían tres puntos (3, 4, 5, 12, 16). En el Anexo III (Protocolos de Investigación II) se muestran de forma detallada los criterios seguidos, en dos de las preguntas planteadas, para valorar la relevancia de cada documento recuperado.

2.5 Análisis de los resultados: exhaustividad-precisión

Para cada uno de los veinte primeros resultados recuperados se determinó su condición de inactivo, duplicado o la puntuación de relevancia, a saber, 0, 1, 2 ó 3. De forma resumida la gráfica 1 muestra los resultados inactivos, duplicados e irrelevantes, es decir, el ruido en la recuperación de información.

Gráfica 1
Inactivos, duplicados y de relevancia 0 (de entre los 20 primeros resultados) para las 20 preguntas



Infoseek, junto a Opentext, es el que mayor número de enlaces inactivos presenta, lo que hace pensar que su base de datos realmente no se actualiza con tanta asiduidad como sería deseable. Hotbot aparece como el buscador con mayor número de duplicados, lo que confirma las conclusiones de Leighton y Srivastava (3, 11). A éste le siguen Altavista y Opentext. Aunque desde el punto de vista de los intermediarios de la información se perciban como ruido documental, desde el punto de vista del usuario los duplicados no siempre son un aspecto negativo en los motores de búsqueda. Por el contrario, la repetición de páginas web de la misma *sede* web e, incluso, de una misma página, puede conferir al usuario una sensación de máxima relevancia. La recuperación de páginas irrelevantes no es muy destacada, pero hay que señalar que es en Magellan donde claramente se produce en mayor grado. Por el contrario, Hotbot, Altavista y Opentext ofrecen el menor número de enlaces no relevantes. En el caso de WebCrawler y, sobre todo, en el de WWWWorm y Yahoo, el bajo índice de inactivos, duplicados y documentos irrelevantes se debe, fundamentalmente, a su escaso índice de respuesta en algunas búsquedas.

En cuanto a los resultados relevantes, se realizaron una serie de pruebas para los diferentes grados de relevancia de los ítemes recuperados según la habían establecido los colaboradores. En las pruebas 4 y 5 se eliminaron los duplicados de los cálculos de exhaustividad y precisión:

Prueba 1: Documentos relevantes (relevancia 1, 2 ó 3)
Prueba 2: Documentos potencialmente relevantes u óptimos (relevancia 2 ó 3)
Prueba 3: Documentos óptimos (relevancia 3)
Prueba 4: Documentos relevantes (relevancia 1, 2 ó 3) sin duplicados
Prueba 5: Documentos relevantes (relevancia 2 ó 3) sin duplicados

Es decir, los resultados considerados relevantes pueden serlo, dependiendo de la prueba, cuando su puntuación de relevancia es 3, cuando es 2 ó 3 o cuando es 1, 2 ó 3. Asimismo, se pueden analizar para los diez, quince o veinte primeros resultados. Se realizaron distintos análisis atendiendo a los diferentes criterios:

a) Resultados por pruebas

Número total y promedio de documentos relevantes en las 20 preguntas, para los 10, 15 y 20 primeros resultados y para los diferentes niveles de relevancia: documentos relevantes, potencialmente relevantes u óptimos y únicamente documentos óptimos, es decir, para las pruebas 1, 2 y 3 respectivamente.

b) Resultados por tipo de pregunta (tema)

Se analizaron los documentos relevantes (10, 15 y 20 primeros resultados) para las pruebas 1 y 2 distinguiendo los resultados según el tema de las preguntas (ocio o especializadas)

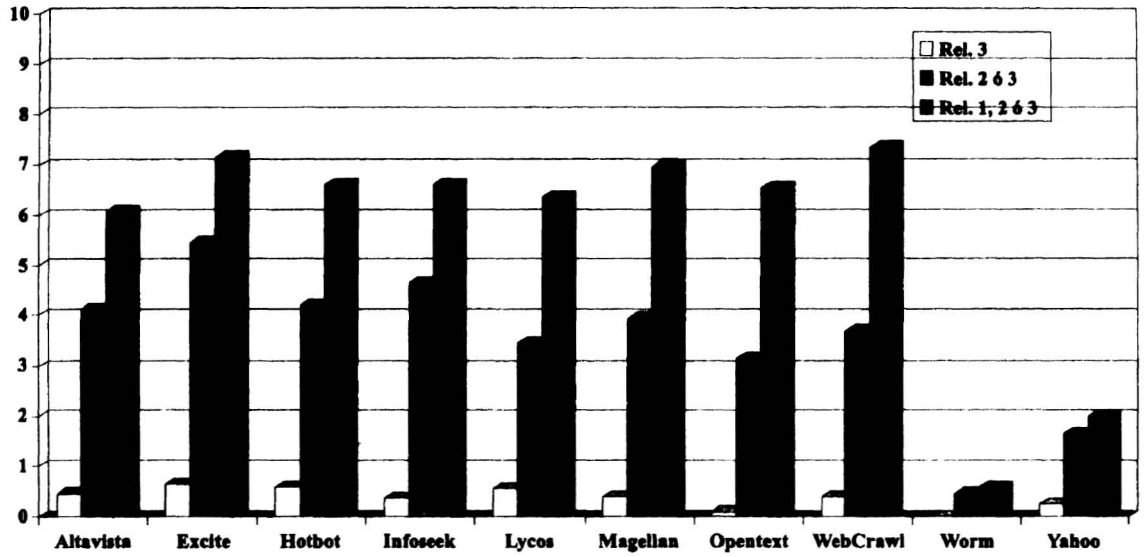
c) Resultados por tipo de búsqueda (sintaxis)

En relación con la sintaxis—booleana o de frase—de la ecuación de búsqueda se analizaron los 10, 15 y 20 primeros resultados tanto para la prueba 1 como para la 2.

Las representaciones gráficas muestran, de forma resumida, estos datos. La gráfica 2 «Promedio de resultados relevantes (de entre los 10 primeros) por pruebas» permite comparar el número medio de documentos relevantes entre los diez primeros resultados para diferentes «grados de exigencia» en cuanto a la relevancia. Se observa que, para la prueba 3, los resultados son claramente insatisfactorios en todos los casos, el bajo promedio de documentos relevantes (inferior a 1) así lo demuestra. Para la prueba 2, la más representativa, Excite siempre muestra mayor número de documentos relevantes y, a distancia considerable, aparecen Infoseek (4,7), Hotbot (4,2) y Altavista (4,1). En la prueba 1, donde también se incluyen los resultados «técnicamente» relevantes, puntuados con 1 pero que, en la práctica, producen ruido documental, se altera la tónica general encontrada hasta el momento y WebCrawler presenta el mejor promedio de documentos relevantes en los diez primeros resultados, seguido de Excite y Magellan, tras los cuales se sitúan Hotbot e Infoseek, con el mismo número de documentos relevantes de media.

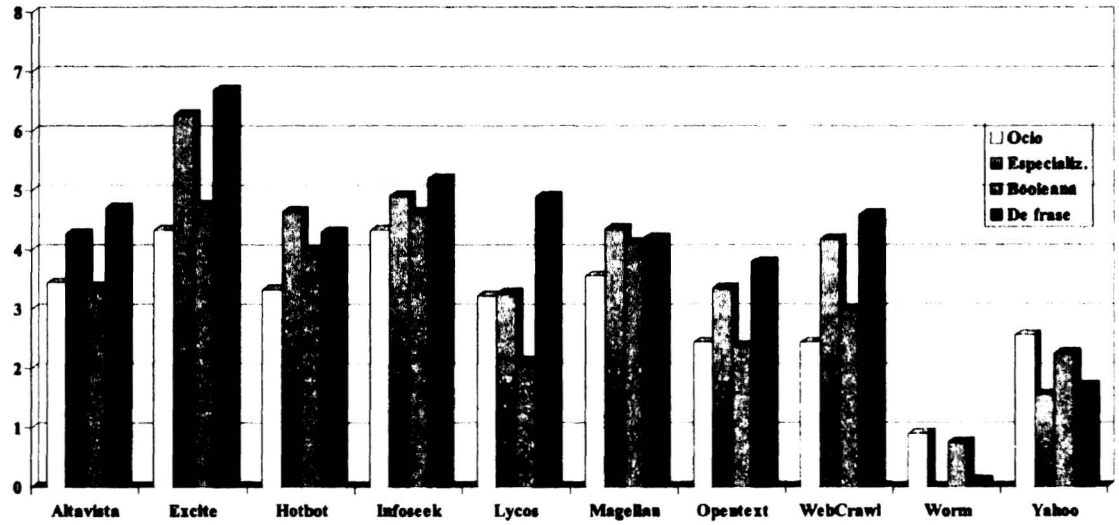
Los resultados pueden analizarse para todas las preguntas en general o bien, a título orientativo, distinguiendo el tema (preguntas especializadas o de temas relacio-

Gráfica 2
Promedio de resultados relevantes (de entre los 10 primeros) por pruebas



nados con el ocio) por una parte, y la sintaxis de la pregunta (preguntas booleanas o de frase) por otra, que se muestra en la gráfica 3 «Promedio de resultados relevantes (rel. 2 ó 3) por tipos de preguntas (de entre los 10 primeros)». Según se observa, los mejores promedios para Excite o Infoseek con relación a cada tipo de pregunta se relacionan con el hecho de que las medidas de exhaustividad y precisión han sido mejores en estos servicios para el total de las preguntas planteadas. Tanto en preguntas sobre «ocio», como para las «especializadas», Excite e Infoseek destacan claramente con un mayor promedio de documentos relevantes, seguidos de Hotbot y Altavista.

Gráfica 3
Promedio de resultados relevantes (rel. 2 o 3) por tipos de preguntas (de entre los 10 primeros)



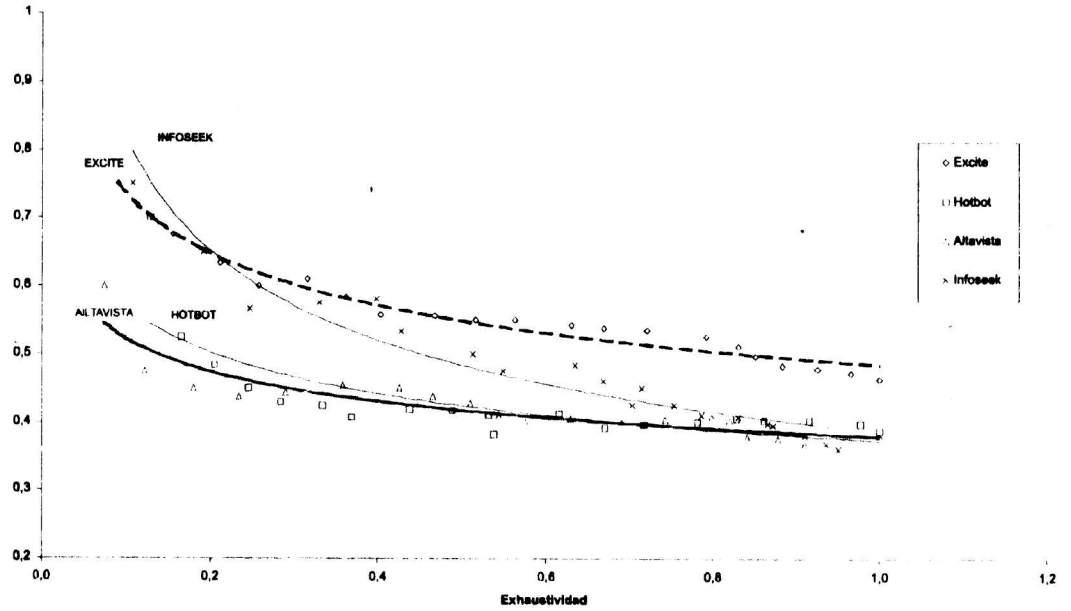
Generalmente, los buscadores ofrecen mejores resultados tanto para búsquedas precisas como exhaustivas de tema especializado con relación a las de «ocio»: posiblemente porque los «temas» especializados suelen estar mejor definidos en la ecuación de búsqueda y favorecen una mejor ordenación por relevancia. En relación con la sintaxis de la ecuación de búsqueda, se distingue entre lo que aquí se han denominado búsquedas booleanas, es decir, aquéllas con uso de operadores lógicos o delimitadores, y las de frase. Las preguntas de frase, en la mayor parte de los casos, ofrecen mayor promedio de relevantes que la búsqueda booleana. Nuevamente destacan Excite e Infoseek. Generalmente, para búsquedas más precisas, la sintaxis de frase es más adecuada, mientras que para preguntas exhaustivas la sintaxis booleana ofrece mejores resultados.

A continuación se analizó la «precisión y exhaustividad de los veinte primeros», que permite realizar la evaluación a partir de los primeros veinte documentos recuperados, presentados en orden de relevancia decreciente respecto a la consulta planteada. Se utilizó la exhaustividad para medir el rendimiento de los SRI atendiendo a dos premisas:

- Únicamente se evalúa la relevancia de los documentos recuperados, que es lo que realmente hace un usuario cuando consulta una base de datos, no de los documentos de la colección (17), ni de una muestra de la misma. Esto implica no realizar juicios de relevancia sobre una pregunta de manera previa a la evaluación. En consecuencia, al carecer de una base de documentos sobre los que calcular la exhaustividad, según lo habitual establecido en los tests de Cranfield, el método empleado ha de ser otro.
- Los cálculos de precisión y exhaustividad se realizan según el método propuesto por Salton y McGill (18), para SRI que ordenan los resultados por relevancia.

Los cálculos realizados fueron los siguientes:

Gráfica 4
Promedio Exhaustividad-Precisión para relevancia 2 o 3 (prueba 2)



a) Valores de exhaustividad y precisión por prueba

Se obtuvieron los valores de exhaustividad y precisión relativos a cada rango de documentos recuperados por cada buscador y pregunta, calculados para los 20 primeros. Los datos se analizaron para la prueba 1, 2 y 3 pero también para la prueba 4, que considera relevantes los puntuados con 1, 2 ó 3 pero donde se han eliminado de los cálculos los duplicados, y la prueba 5, para los documentos de relevancia 2 ó 3 también sin duplicados.

b) Valores de exhaustividad y precisión por tipo de pregunta

Se calcularon los valores medios de exhaustividad y precisión, para los 20 primeros resultados, en relación con el tipo de pregunta —sobre ocio y especializadas— para la prueba 2.

c) Valores de exhaustividad y precisión por tipo de búsqueda

Se calcularon los valores medios de exhaustividad y precisión, para los 20 primeros resultados, por tipo de sintaxis —booleana y de frase— para la prueba 2.

Al calcular el promedio de los valores de exhaustividad-precisión para un conjunto de diferentes preguntas de usuario se obtiene lo que Salton y McGill (18) denominan nivel medio de respuesta (*user-oriented recall-level average*) que refleja el funcionamiento que un usuario estándar puede esperar obtener del sistema. La gráfica 4 muestra las curvas que representan la relación exhaustividad-precisión para la prueba 2, posiblemente la más significativa, ya que tiene un «nivel de exigencia» medio en relación con la relevancia considerada. Estos valores se han representado mediante regresión logarítmica por ser la función que ofrecía los mejores resultados de ajuste. No se han mostrado las representaciones correspondientes a todos los buscadores evaluados debido a que, de esta manera, es más fácil visualizarlos. En las relaciones exhaustividad-precisión, a valores superiores de precisión, menos exhaustividad y viceversa. En resumen, cuando Excite sale de valores de precisión y exhaustividad medios y, de forma más acentuada, para búsquedas más exhaustivas y menos precisas, muestra peores resultados aunque, según la comparación que se puede establecer, es adecuado tanto para búsquedas exhaustivas como precisas. Infoseek es mejor para búsquedas más precisas, mientras que Hotbot y Altavista son menos buenos en ambas con valores de precisión casi siempre inferiores a 0,5, pero se mantienen en una franja de precisión útil lo que se puede interpretar como que tienen peor ordenación por relevancia. En casi todos los casos los buscadores con peores resultados presentan valores de exhaustividad inferiores a 0,5 y, en muchos, los valores de precisión caen entre 0,3 y 0,2. Magellan ofrece valores satisfactorios siempre que la exhaustividad sea baja pero, a medida que ésta aumenta, su precisión disminuye llamativamente presentando valores de precisión en torno al 0,3. Opentext y Lycos tienen un comportamiento similar, aunque el primero presenta mayor precisión. WebCrawler tiene resultados muy irregulares y valores bajos de exhaustividad y precisión. Yahoo y WWWWorm también tienen resultados muy pobres. Este hecho es fácilmente comprensible si se tiene en cuenta que el primero es, principalmente, un directorio con amplias posibilidades para el *browsing* y el otro, ya desaparecido, siempre tuvo una muy pequeña base de datos, pocas prestaciones en las búsquedas y ni siquiera ordenaba los pocos resultados recuperados por relevancia.

3 Conclusiones

La aplicación del método diseñado permite obtener las siguientes conclusiones generales:

Primera. Los cambios y la evolución experimentados en estos servicios de búsqueda podrían provocar que algunas de las conclusiones a las que aquí se ha llegado no respondieran exactamente a la realidad actual. Sin embargo, hay que recordar que los valores reales aquí descritos son menos importantes que la metodología usada para obtenerlos. Concretamente, se demuestra que el método de la exhaustividad y precisión de los 20 primeros, propuesto por Salton y McGill (18), puede aplicarse igualmente a la evaluación del funcionamiento en la recuperación de información, si se presta atención especial a las características particulares que presenta la W3. Este método permite obtener datos sobre la eficacia en la RI que permiten una estimación más fiable de la lograda por otros medios para los SRI de la W3. El método produjo unos 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. Como recordatorio, los rasgos principales del mismo son: *a)* incorporar usuarios reales que plantean preguntas reales; *b)* analizar la relevancia de los veinte primeros resultados expresada en una escala de cuatro grados; *c)* usar las medidas de exhaustividad y precisión para evaluar la RI .

Segunda. Los resultados ofrecidos ponen, una vez más, de manifiesto que los buscadores no son sistemas de recuperación de información muy precisos aunque sí muy exhaustivos. Los resultados confirman la coherencia interna del método utilizado.

Cuarta. En términos generales, se puede establecer el siguiente *ranking* de rendimiento: 1.º Excite, 2.º Infoseek, 3.º Hotbot, 4.º Altavista, 5.º Magellan, 6.º Opentext, 7.º Lycos, 8.º WebCrawler, 9.º Yahoo, 10.º WWWWorm. La evolución experimentada por los servicios analizados en fechas posteriores a la realización de las pruebas de este estudio corroboran los datos obtenidos. Los resultados de WWWWorm no respondieron a la máxima de relación inversa entre exhaustividad y precisión porque este SRI no ordenaba los resultados por relevancia. WWWWorm, que obtiene la peor calificación, desapareció de la red. Opentext, que tampoco destacó entre los mejores, se convirtió en un buscador especializado para negocios. Otros han cambiado de manos en un intento por mejorar su presencia en el mercado: WebCrawler, en octavo lugar (hay que suponer que su algoritmo interno de ordenación por relevancia no es el más adecuado), pertenece hoy a la misma compañía que Excite y Magellan. Lycos ha comprado Hotbot y, en un movimento sin precedentes, se transformó en un directorio temático. Yahoo presenta una tendencia distinta de la de los situados en cabeza porque se trata de un directorio.

4 Líneas de investigación futuras

En torno a este tema, hay diferentes posibilidades que pueden originar útiles e interesantes trabajos de investigación.

- Para analizar los servicios de búsqueda con mayor profundidad deben realizarse estudios donde la evaluación sea mayor, orientada a comparar las expresio-

nes de búsqueda estructuradas (expresiones booleanas) *versus* expresiones de búsqueda en lenguaje natural.

- De forma ideal, se podrían comparar *todos* los principales servicios de búsqueda (incluyendo Northern Light, por ejemplo).
- El método aquí aplicado a servicios de búsqueda generales e internacionales podría aplicarse a servicios de búsqueda con cobertura más limitada en razón de su contenido (especializados) o de su cobertura geográfica (nacionales)
- También se debe investigar en qué condiciones es correcto comparar servicios de búsqueda con servicios de directorio o de evaluación de páginas web como Magellan o Yahoo.
- Naturalmente, los resultados de cualquier estudio de la precisión de cualesquiera SRI de la W3 pierden vigencia con rapidez —recordemos la máxima de que los conocimientos científicos llegan a los manuales introductorios cuando ya no son verdad— y se hace necesario repetir estudios similares con periodicidad.

Apéndice I (Pregunras formuladas)

1. Servicios de traducción inglés-español.
2. La cocina española.
3. La monarquía española.
4. La obra de Velázquez.
5. Arquitectura española en el Barroco.
6. La Alhambra de Granada.
7. La guitarra flamenca.
8. La Guerra Civil española.
9. El Presidente Aznar.
10. Federico García Lorca.
11. Las fiestas de San Fermín.
12. El aceite de oliva.
13. El fino (vino fino).
14. El terrorismo de ETA.
15. Telefónica.
16. El movimiento gay en España.
17. El Real Madrid.
18. El Camino de Santiago.
19. La investigación en España.
20. Julio Iglesias.

Apéndice II (Protocolos de la investigación I)

PREGUNTA 1: Servicios de traducción inglés-español		
TIPO	Búsqueda booleana (con frases de búsqueda y operador «y»)	
RELEVANCIA	3	Una página que ofrezca un amplio listado de empresas y/o traductores que presten este tipo de servicios.
	2	Un servicio de traducción o traductor determinado, o una página que remita a 3.
	1	Una página que contenga todas las palabras pero no en el contexto adecuado o bien que mencione el tema superficialmente.
	0	Página que no contiene todas las palabras clave del enunciado de búsqueda.
PREGUNTA 2: Cocina española		
TIPO	Búsqueda de frase	
RELEVANCIA	3	Una página que ofrezca un estudio general sobre el tema o bien un listado de recetas de cocina española o enlaces a diversos documentos en la red sobre el tema.
	2	Una página con información sobre algún aspecto del tema: libros de cocina, restaurantes, recetas, vinos... o con enlaces hacia éstos o a una página 3
	1	Una página que contenga la frase de búsqueda pero donde el tema se mencione de pasada
	0	Página que no contiene las palabras de búsqueda o están dispersas en el documento

Apéndice III (Protocolos de investigación II)

PREGUNTA 1: Servicios de traducción inglés-español		
SINTAXIS	Altavista	+«translation services» +english +spanish
	Excite	+«translation services» +english +spanish
	Hotbot	***all the words***: «translation services» english spanish
	Infoseek	+«translation services» +english +spanish
	Lycos	+«translation services» +english +spanish
	Magellan	+«translation services» +english +spanish
	Opentext	translation services ***and*** english ***and*** spanish **(búsqueda avanzada)**
	Webcrawler	«translation services» ***and*** english ***and*** spanish
	WWWWworm	***URL references***: «translation services» ***and*** english ***and*** spanish
	Yahoo	+«translation services» +english +spanish
PREGUNTA 2: Cocina española		
SINTAXIS	Altavista	«spanish cooking»
	Excite	«spanish cooking»
	Hotbot	***the exact phrase***: spanish cooking
	Infoseek	«spanish cooking»
	Lycos	***all the words***: «spanish cooking»
	Magellan	«spanish cooking»
	Opentext	***search for this phrase***: spanish cooking
	Webcrawler	«spanish cooking»
	WWWWworm	***URL references***: spanish cooking
	Yahoo	«spanish cooking»

Bibliografía

1. 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).
2. SULLIVAN, D. (ed.). *Search Engine Watch.* Mecklermedia: cop.1996-1999. Disponible en: http://searchenginewatch.com/ (Consultado 2 marzo 96).
3. LEIGHTON, V. H. y SRIVASTAVA, J. First 20 Precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Information Science,* 1999, 50 (10) 870-881.
4. CHU, H. y ROSENTHAL, M. *Search engines for the World Wide Web: A comparative study and evaluation methodology,* octubre de 1996. Disponible en: http://www.asis.org/annual-96/ElectronicProceedings/chu.html (Consultado el 5 de febrero de 1997).
5. DING, W. y MARCHIONINI, G. A comparative study of web search service performance. *Proceedings of the ASIS Annual Conference 33,* 1996, 136-142.
6. GORDON, M. y PATHAK, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management,* 1999, 141-180, 35.
7. LAKE, M. 2nd Annual search engine shoot-out. *PC Computing.* Disponible en: http://www4.zdnet.com/pccomp/features/exc10997/sear/sear.html. (Consultado el 3 de enero de 1998.)
8. TOMAIUOLO, N. G. y PACKER, J. G. An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries,* 1996, 16(6) 58-62.
9. FELDMAN, S. Just the answers, please: choosing a Web search service. *Searcher Magazine,* 29 agosto 1997. Disponible en: http://www.infotoday.com/searcher/may/story3.htm (Consultado 13 oct. 97)
10. LEBEDEV, A. *Best search engines for finding scientific information on the Web.* Septiembre 29 de 1996. Disponible en: http://www.chem.msu.su/eng/comparison.html (Consultado 12 dic. 97).
11. 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 el 2 de enero de 1997).
12. LEIGHTON, V. H. y SRIVASTAVA, J. *Precision among World Wide Web Search Services (Search Engines): Altavista, Excite, Hotbot, Infoseek, Lycos.* actualizado el 10 de julio de 1997. Disponible en: http://www.winona.msus.edu/is-f/library/webind2/webind2.htm (Consultado el 12 de julio de 1997).
13. OVERTON, R. Search engines get faster and faster, but not always better. *PC World,* septiembre de 1996. Disponible en: http://www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html (Consultado el 10 de noviembre de 1996).
14. SCHLICHTING, A. y NILSEN, E. *Signal detection analysis of WWW search engines.* 17 de diciembre de 1997. Disponible en: http://www.microsoft.com/usability/webcong/schlichting/schlichting.htm (Consultado el 13 de octubre de 1997).
15. WESTERA, G. *Robot-drive search engine evaluation: overview.* 4 de julio de 1997. Disponible en: http://www.curtin.edu.au/curtin/library/ staffpages/gwpersonal/senginestudy/ (Consultado el 13 de octubre de 1997)
16. CLARKE, S y WILLET, P. Estimating the recall performance of web search engines. *Aslib Proceedings,* 1997, 49(7) 184-189.
17. BLAIR, D. C. y MARON, M. E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM,* 1985, 28(3) 281-299.
18. SALTON, G. y McGILL, J. *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983. ISBN 0070544840.