Evaluación del rendimiento de los motores de búsqueda en la recuperación de información en la WWW

Autores: Salvador Oliván, J.A.; Vidal Bordés, F.J.

Salvador Oliván, José Antonio
Profesor del Departamento de Ciencias de Documentación e Historia de la Ciencia
Universidad de Zaragoza
Teléfono: 976761000 ext. 3562
E-mail: jaso@posta.unizar.es

Vidal Bordés, Francisco Javier
Biblioteca de la Facultad de Filosofía y Letras
Universidad de Zaragoza
E-mail: fjvidal@posta.unizar.es

Resumen: Este estudio evalúa el rendimiento de nueve buscadores en la World Wide Web en temas específicos de Biblioteconomía y Documentación midiendo características objetivas como enlaces no operativos, duplicidad de páginas en un mismo motor de búsqueda, cobertura única y solapamiento. Para ello se realizan cuatro búsquedas y se examinan las primeras 50 páginas recuperadas de cada tema y en cada buscador. Los resultados globales revelan que Google, Excite y Northern Light consiguen mayor puntuación y, por tanto, son los más eficaces en las características estudiadas.

Palabras clave: evaluación de motores de búsqueda; world wide web; recuperación de información.

Abstract: This study evaluates the retrieval performance of web serch engines nine on library and documentation subjects. We measure the effectiveness with a objective evaluation in terms of the death links, duplicity, unique hits and overlap using four simple queries and examining the web pages top fifty for each query and search engine. The results obtained show that Google, Excite and Northern Light had the highest score and performed better on the studied characteristics.

Keywords: search engines evaluation; world wide web; information retrieval.

Introducción

En la actualidad Internet se puede considerar como la principal fuente de recursos de información[1], convirtiéndose de hecho en el método más utilizado para conseguir información por millones de individuos[2] ya que, después del correo electrónico, la actividad más común realizada por los usuarios es la búsqueda de información[3]. Básicamente existen cinco formas de acceder a información en Internet[4]: a través de una lista de correo o grupo de noticias; ir directamente a una página web de la que se conoce su dirección URL; navegando a través de los enlaces presentes en las diferentes páginas; usando los motores de búsqueda, ya sea expresando las necesidades de información mediante una sentencia de búsqueda o desplazándose por la estructura jerárquica de su directorio; y, por último, explorando la información almacenada en la "web invisible" (bases de datos no accesibles a través de los motores de búsqueda).

Los servicios de búsqueda se han convertido en unas herramientas imprescindibles para buscar en la Web. En las estadísticas publicadas por CyberAtlas según PC Data Online[5] sobre los 50 primeros sitios con mayor número de visitas por usuarios únicos en Diciembre de 2000, destacan varios de los motores de búsqueda como Yahoo, Altavista, Lycos, Excite, Google, etc., lo que implica que constituyen una alternativa muy utilizada para buscar información en la Web, y en realidad, se construyeron para ayudar a los usuarios a recuperar información en la Web de una forma fácil y rápida. Los motores de búsqueda pueden dividirse en cuatro categorías[3]: robots, directorios, metabuscadores y agentes de software. En este artículo sólo se evalúan buscadores pertenecientes a las dos primeras categorías, por lo que a efectos prácticos, de ahora en adelante sólo utilizaremos el término de motores de búsqueda, que es el más utilizado comúnmente para los robots, si bien la distinción entre robots y directorios es cada vez más difusa, ya que los robots licencian directorios web y viceversa[6].

Actualmente existen muchos motores de búsqueda, tanto generales como especializados, diferenciándose no solamente en los recursos web que indizan, sino también en la forma de interrogar en sus bases de datos y en los algoritmos de relevancia que utilizan para presentar los resultados. A pesar del elevado número de motores de búsqueda, sólo unos pocos son los más utilizados, y como quiera que cada uno proporciona un servicio diferente a sus usuarios, se aconseja buscar en más de un motor, por lo que resulta conveniente conocer las diferencias que pueden existir entre ellos y realizar evaluaciones para saber cómo trabajan y con qué precisión devuelven los resultados.

Una de las principales quejas cuando se busca información utilizando los motores de búsqueda es que ante cualquier tema devuelven miles de resultados y que muchas de las páginas devueltas son poco relevantes para el tema de búsqueda[7], y aun cuando son presentados generalmente por orden de relevancia (diferente según el algoritmo utilizado y que no es hecho público por razones comerciales), es necesario ir uno a uno para conocer si satisfacen en realidad las necesidades de información del usuario, lo que puede suponer una gran cantidad de tiempo invertido en función del número de recursos a los que se acceda.

Por otra parte, de sobras es conocido que ningún motor de búsqueda indiza la web completa. En la actualidad no se conoce con exactitud cuál es el tamaño real de la web, si bien uno de los estudios más recientes y completos es el de Lawrence y Giles[8], estimándose un total de páginas web públicas de 800 millones en Febrero de 1999, y ningún motor cubría más de un 16%; esta baja cobertura es debida a la naturaleza dinámica de Internet, donde documentos nuevos aparecen cada día y desaparecen ya existentes[9]. El solapamiento entre los diversos motores de búsqueda es muy pequeño[10], de ahí, que para búsquedas amplias se aconseje buscar en varios motores de búsqueda para conseguir una mayor llamada. Sin embargo, FastSearch[11] declara que de estos 800 millones, la mitad son duplicadas, y que contiene una base de datos de 300 millones de páginas. Estos comentarios vienen a reflejar la escasa coincidencia entre los diferentes estudios y lo que declaran los propios servicios de búsqueda en cuanto al tamaño de la web y a su cobertura. Aunque como bien dice Phil Bradley[12] refriéndose a los motores de búsqueda, lo más grande no significa necesariamente lo mejor y señala que los esfuerzos deberían dirigirse a presentar mejores (precisos) resultados, que en definitiva es lo que le interesa a la persona que busca. Así pues, resulta conveniente conocer lo eficaces que son los motores de búsqueda para buscar y recuperar información.

Las evaluaciones publicadas en la literatura sobre los motores de búsqueda son de dos tipos: aquellas que estudian las características técnicas relacionadas fundamentalmente con las posibilidades de búsqueda (operadores booleanos y de proximidad, búsqueda por campos, etc.) y ofrecen sugerencias y recomendaciones sobre qué motor de búsqueda utilizar, convirtiéndose en guías y tutoriales de cómo utilizarlos (abundantísimos en la propia red), y aquellas otras que estudian la eficacia en la recuperación midiendo diferentes parámetros.

En general, los criterios utilizados por los investigadores para evaluar los motores de búsqueda se pueden resumir en los siguientes[3]: cobertura, enlaces rotos, relevancia, sintaxis de búsqueda, área temática, cambios producidos en las bases de datos a lo largo del tiempo, tiempo de respuesta, características del sistema, opciones de búsqueda, características del interfaz de búsqueda y calidad de los resúmenes.

Los resultados de los estudios indican diferencias que son debidas a los criterios de evaluación y a la metodología utilizada, que en algunos casos no son declaradas, además de los cambios continuos en los motores de búsqueda que afectan a sus interfaces, políticas de indización y algoritmos de relevancia[13].

Objetivos

El objetivo general del presente trabajo consiste en evaluar los resultados proporcionados por los motores de búsqueda ante determinados temas de consulta, centrándonos en el estudio de aquellas características que nos permitan conocer mejor su funcionamiento y determinar qué motores de búsqueda seleccionar para realizar una consulta. En este sentido, los aspectos a evaluar son la operatividad, duplicidad de recursos en un mismo motor de búsqueda, cobertura de páginas únicas y solapamiento entre diferentes motores de búsqueda.

No se estudia la relevancia de los resultados por considerar a ésta como una medida subjetiva, que además puede ser variable de un evaluador a otro, y por tanto sólo nos

basamos en medidas objetivas a fin de no introducir un sesgo subjetivo, ya que cuando se estudia esta característica las conclusiones de los estudios suelen ser bastante contradictorias sobre qué motores proporcionan mayor precisión[14], debido a diferentes sesgos producidos por el planteamiento de la búsqueda, el tema elegido y las personas que evalúan la relevancia. Por lo tanto, el objetivo es estudiar el rendimiento de los motores de búsqueda no utilizando ninguna medida que implique un criterio subjetivo.

Material y Método

Para llevar a cabo el estudio se han seleccionado nueve motores de búsqueda (Altavista, Excite, FastSearch, Google, Hotbot, Infoseek, Lycos, Norther Light y Yahoo. Estos buscadores representan las herramientas de búsqueda generales más populares o utilizadas y más estudiadas en la literatura.

Se eligieron cuatro temas de búsqueda, expresados en lengua inglesa, relacionados y de interés actual en el campo de la Documentación, siendo los siguientes:

Primer tema: Information retrieval

Segundo tema: Knowledge management

Tercer tema: Digital library

Cuarto tema: Distributed information retrieval

Las búsquedas se realizaron durante la primera semana de mayo del 2000, utilizando el interfaz simple de cada uno de los motores de búsqueda. Los términos se encerraron entre dobles comillas para forzar una búsqueda por frase. De esta manera, utilizamos una opción soportada por todos los buscadores, por lo que al estudiar el solapamiento de un determinado intervalo de páginas evitamos el posible sesgo que se pudiera introducir si hubiéramos utilizado opciones de búsqueda sólo permitidas por algunos (operadores booleanos, de proximidad, etc.).

Tras la operación de búsqueda, se han visitado los 50 primeros resultados recuperados de cada motor, por considerar que es un número suficientemente amplio y que en pocas ocasiones se visitan más recursos. Diversos estudios indican que los usuarios no suelen ver más de 10 resultados[15, 16] y aquellos que estudian la precisión de los resultados suelen examinar las 10-20 primeras páginas de cada motor[13, 14, 17, 18, 19, 20].

Se creó una base de datos en Excel con las siguientes variables: dirección URL del recurso, motor de búsqueda en el que aparecía, registros duplicados con la misma URL y con diferente URL, y si el registro era accesible o tenía el enlace roto.

Los aspectos evaluados han sido los siguientes:

- Operatividad o funcionalidad: consiste en analizar si el recurso recuperado es accesible (activo) o, por el contrario, no está operativo. Se han considerado no operativos aquellos a los que por diversos motivos no se pudo acceder (mensajes como “no encontrado”, “no hubo respuesta” o “no se encuentra el servidor”). Aquellas páginas que habían cambiado de dirección pero especificaban la nueva dirección o automáticamente redireccionaban al sitio adecuado eran contabilizadas como operativas. El índice de operatividad (recursos operativos/recursos recuperados visitados) indica en cierto grado la actualización

de las bases de datos de cada motor (con qué frecuencia y meticulosidad examina los enlaces) y sería deseable que fuera próximo a 1. Los enlaces rotos se consideran debidos a un malfuncionamiento del motor de búsqueda[17].

- Duplicidad: consiste en averiguar qué porcentaje de recursos recuperados en cada motor aparecen repetidos. La duplicidad se ha estudiado desde la perspectiva de dos situaciones diferentes: aquellos que contienen la misma dirección URL y aquellos que con diferente URL llevan a la misma página web (diferente dirección IP). Esto se hace así porque aunque la segunda situación puede ser normal (la misma página puede ser accesible a través de múltiples alias, o bien copias de la página pueden estar presentes en otros sitios web), que se de la primera indica que el funcionamiento de un motor de búsqueda no es correcto, ya que para incluir una dirección en su base de datos debería comprobar que ya está registrada.

  Conocer el grado de duplicidad proporciona información de interés para ayudarnos a elegir la herramienta que recupere menor número de recursos duplicados, y que redundará en una mayor eficiencia en el tiempo empleado para visitar los recursos y en una mayor relevancia conseguida, ya que registros recuperados iguales producirían un menor índice de precisión y de relevancia.

- Cobertura: El grado de cobertura viene determinado por el porcentaje de recursos únicos que proporciona cada motor de búsqueda con respecto al total de recuperados por todos los motores. Aquel que consiga mayor grado de unicidad, tendrá mayuor cobertura de información potencialmente relevante.

- Solapamiento: se calcula comparando los resultados de los motores de búsqueda por parejas contabilizando el número de recursos que coinciden en cada par. El grado de solapamiento entre los diferentes motores de búsqueda nos permitirá conocer en qué medida coinciden sus bases de datos y/o el algoritmo de relevancia que utilizan para presentar los resultados

El rendimiento global se mide asignando a cada motor un número de rango del 1 al 9 para las diferentes características; la suma total de los rangos proporciona qué motores de búsqueda obtienen mayor puntuación y tienen mejor rendimiento. No se incluye el solapamiento ya que en cierta manera viene reflejado por el grado de cobertura y unicidad, y motores de búsqueda con un bajo solapamiento en global podrían tener porcentajes de solapamiento alto al analizarlo con cada uno de los buscadores si coincidieran entre ellos las páginas solapadas, con lo que el sesgo sería importante.

El proceso de los datos se realizó con el paquete estadístico SPSS, versión 8.0.

Resultados

Los resultados del estudio se presentan, según los aspectos estudiados, individualmente para cada tema de búsqueda y de forma global en cada motor de búsqueda utilizado.

1. **Operatividad**. La Tabla I muestra el número de enlaces no operativos y el índice de operatividad global para cada motor de búsqueda. Los motores de búsqueda que mayor número de enlaces muertos presentan y, por consiguiente, un menor índice de

operatividad son Lycos e Infoseek, seguidos de Yahoo, FastSearch y Northern Light, con un índice muy similar, mientras que HotBot, Google y Excite son los que tienen un índice de operatividad mayor.

Este índice es un indicador de lo actualizadas que están las bases de datos o índices de los motores de búsqueda.

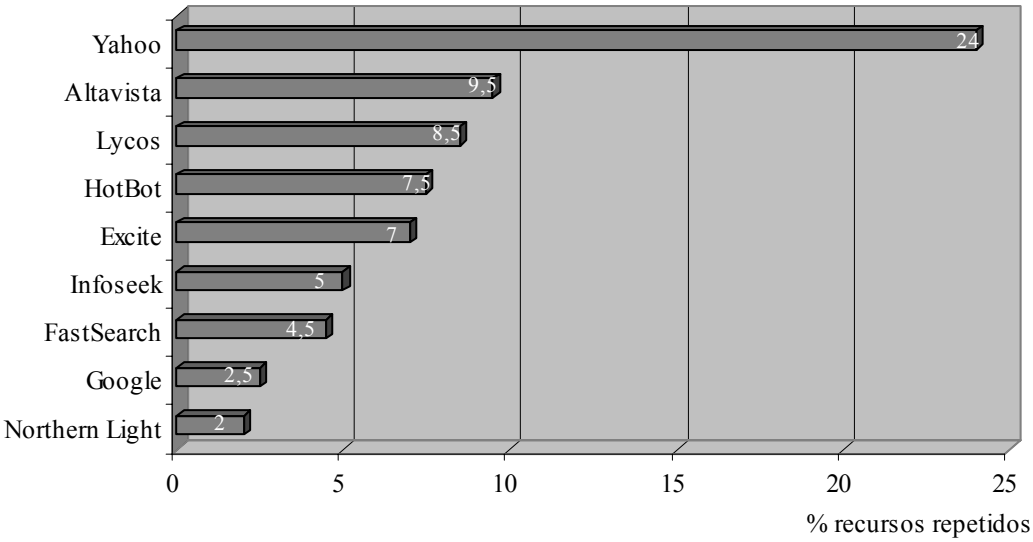
Tabla I. Distribución del número de recursos recuperados no operativos e Indice de Operatividad Global (I.O. Global) por motores de búsqueda

Motor	Tema-1	Tema-2	Tema-3	Tema-4	Total	I.O. Global
Lycos	6	11	9	7	33	0,835
Infoseek	13	8	5	1	27	0,865
Yahoo	8	5	3	0	16	0,920
FastSearch	4	2	3	6	15	0,925
Northern Light	7	2	4	1	14	0,930
Altavista	4	1	1	3	9	0,955
Excite	1	0	1	4	6	0,970
Google	2	1	1	0	4	0,980
HotBot	0	1	1	0	2	0,990

Los peores resultados de Lycos coinciden con los registrados en el estudio realizado por Danny Sullivan (Agosto de 1999)[21] y Leighton (1997)[14], aunque en otros estudios realizados periódicamente por Greg R. Notess (2000)[22] cada vez es un motor diferente el que consigue los peores resultados, siendo Altavista el que encabeza la clasificación en Febrero del 2000. Por lo tanto, no existe una coincidencia general sobre qué motores de búsqueda tienen las bases de datos menos actualizadas, por lo que cabría interpretar esta gran variabilidad al período en el que se realizan los estudios y a su coincidencia con el menor lapso de tiempo en el que se han actualizado las bases de datos. De todas las maneras, sí que se coincide en todos los estudios en que HotBot es uno de los que menor porcentaje de enlaces muertos consigue.

2. **Recursos repetidos**. El Gráfico I muestra el porcentaje de recursos repetidos en cada motor de búsqueda en relación a los 200 documentos recuperados de las cuatro búsquedas.

Gráfico I. Porcentaje total de recursos repetidos recuperados en cada motor de búsqueda



Destaca por encima de todos Yahoo, con un 24% de documentos repetidos (48 documentos), y que puede deberse a que al ser un directorio donde los recursos pueden estar clasificados simultáneamente en varias categorías, la lista de resultados contiene recursos recuperados de diferentes categorías temáticas, de ahí que 46 de ellos contengan la misma dirección URL (Tabla II). Siguen a continuación Altavista (9,5%), Lycos (8,5%), HotBot (7,5%) y Excite (7%), siendo Google y Northern Light (2,5% y 2% respectivamente) los que menor número de registros repetidos recuperan.

En la Tabla II aparece el número de recursos recuperados repetidos en cada motor de búsqueda, diferenciando tanto los que contienen la misma dirección URL (columna *=URL*) como aquellos que con distinta URL conducen a la misma página web (columna *≠URL*).

Tabla II. Recursos recuperados repetidos en cada motor de búsqueda, con la misma URL (=URL) o distinta URL (≠URL).

Motor	**Tema-1**		**Tema-2**		**Tema-3**		**Tema-4**		**Total**	
	=URL	≠URL	=URL	≠URL	=URL	≠URL	=URL	≠URL	=URL	≠URL
Altavista	0	3	2	2	4	4	0	4	6	13
Excite	1	1	0	1	5	2	3	1	9	5
FastSearch	1	0	1	0	1	1	4	1	7	2
Google	0	2	0	0	2	0	1	0	3	2
HotBot	2	1	4	2	4	0	0	2	10	5
Infoseek	3	1	0	1	0	1	1	3	4	6
Lycos	2	1	4	4	4	2	0	0	10	7
Northern Light	0	1	2	0	0	0	0	1	2	2
Yahoo	14	0	12	0	20	1	0	1	46	2

Sería deseable que los motores de búsquedas devolvieran resultados únicos, o al menos, con diferente URL, ya que deberían ser capaces de identificar aquellas páginas con la

misma dirección antes de incluirlas en sus bases de datos, o bien si se trata de directorios, como Yahoo, eliminar aquellas repetidas. En este caso, los buscadores que tienen mayor proporción de páginas repetidas con la misma URL son Yahoo (por lo comentado anteriormente), FastSearch (7 de 9, el 77,7%), HotBot (10 de 15, el 66,6%), Excite (9 de 14, el 64,2%) y Lycos (10 de 17, el 58,8%). Sin embargo, es de destacar que mientras Yahoo tiene el mayor número (excesivamente alto) de páginas repetidas con la misma URL en las tres primeras búsquedas, en la cuarta obtiene 0 resultados, coincidiendo con el tema más específico, lo que puede explicarse que mientras temas más generales pueden clasificarse en varias categorías, cuando el tema es más concreto sólo aparecería en una sola categoría y entonces los resultados en cuanto a páginas repetidas son excelentes.

Por otra parte, están las páginas recuperadas idénticas con diferente URL y que puede deberse a diversos factores, como la existencia de diferentes alias para un mismo nombre de dominio, o la ubicación de los mismos recursos en varios mirrors. En este caso son Altavista (13 de 19, el 68,4%) e Infoseek (6 de 10, el 60%) los que devuelven mayor proporción de páginas idénticas con diferente URL.

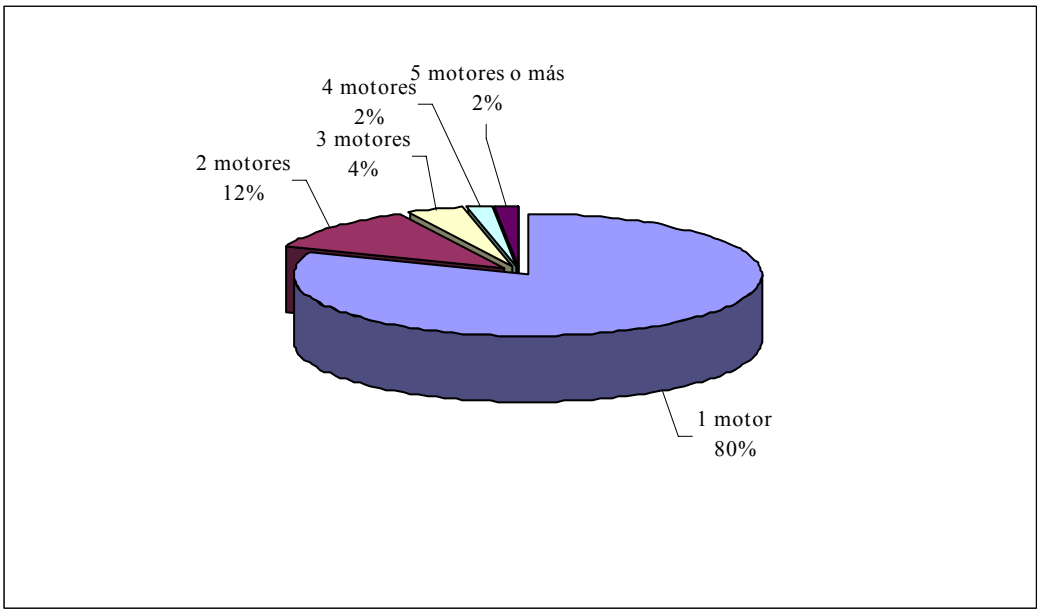
3. **Cobertura**. En cada búsqueda se examinaron 450 páginas recuperadas (50 por motor de búsqueda), lo que hace un total de 1.800 páginas. Para analizar la cobertura, se eliminaron las páginas repetidas en cada motor y aquellas a las que no se puede acceder (enlace roto), quedando un total de 1568 páginas potencialmente útiles o relevantes, de las que 1.183 son páginas únicas, lo que representa un solapamiento del 24,6% para todos los temas de búsqueda.

Tabla III. Cobertura de páginas potencialmente relevantes por los motores de búsqueda.

Páginas recuperadas	Tema-1	Tema-2	Tema-3	Tema-4	Total
1 motor	285	215	174	283	957
2 motores	29	45	20	45	139
3 motores	9	12	18	10	49
4 motores	3	5	10	1	19
5 motores		6	4	1	11
6 motores			3		3
7 motores			3		3
8 motores			2		2
Total páginas únicas	326	283	234	340	1183
Total páginas pot. relevantes	382	391	383	412	1568

De las 1183 páginas únicas, 957 (80%) fueron recuperadas por un solo motor de búsqueda y 139 (12%) por dos motores de búsqueda, y sólo el 8% fueron recuperadas por tres o más buscadores (Gráfico I). En términos generales, coincide con estudios que indican un solapamiento entre los motores de búsqueda de un 16-25%[12].

Gráfico I. Distribución de las páginas únicas recuperadas por los motores de búsqueda



En la Tabla IV se observa cómo se distribuyen las 957 páginas únicas recuperadas por un solo motor. De las 4 búsquedas, Excite consigue encontrar más páginas únicas en 3 de ellas, y Northern Light en 1. En el análisis global, del total de páginas únicas recuperadas por un solo motor de búsqueda, Excite es el buscador que contribuye con mayor porcentaje (14,6%), seguido de cerca por Lycos y Norther Light (14% y 13,8% respectivamente). En el otro extremo, se sitúan HotBot con un 7,8% (casi la mitad que Excite) y Yahoo con un 8,5%.

Tabla IV. Páginas únicas recuperadas en cada motor de búsqueda.

	Tema-1	**Tema-2**	**Tema-3**	**Tema-4**	**Total**
Excite	43	32	26	39	140 (14,6%)
Lycos	39	28	30	37	134 (14,0%)
Northern Light	38	26	31	37	132 (13,8%)
Google	29	29	18	29	105 (11,0%)
FastSearch	35	21	13	30	99 (10,3%)
Infoseek	24	22	19	34	99 (10,3%)
Altavista	30	16	12	33	91 (9,5%)
Yahoo	22	24	13	23	82 (8,5%)
HotBot	25	17	12	21	75 (7,8%)
Total	285	215	174	283	957 (100%)

4. **Solapamiento**. Para estudiar el solapamiento de las bases de datos de los diferentes motores de búsqueda en las cuatro búsquedas, se tienen en cuenta todas las páginas recuperadas, tanto las duplicadas como aquellas con enlace roto, ya que todas constituyen registros de las bases de datos.

Al estudiar el solapamiento de cada motor de búsqueda con los demás, destaca que Altavista y FastSearch tienen el mayo solapamiento con un 24% de páginas que

aparecen en los dos motores de búsqueda. Excite tiene el mayor solapamiento con HotBot (10%), Google tiene un solapamiento del 17% con Altavista, HotBot e Infoseek, HotBot tiene el mayor solapamiento con FastSearch (19,5%) y Yahoo (19%), Infoseek con HotBot (17,5%) y Google (17%), Lycos con Google (5,5%), Northern Light con Altavista (13%) y Google (12,5%), y Yahoo con HotBot (19%) (Tabla V).

El que menor porcentaje de solapamiento consigue en todas las posibles comparaciones por parejas es Lycos, con unos valores que oscilan entre 1,5% y 5,5%; a continuación le siguen Excite y Norther Light. Esto coincide con los tres motores de búsqueda que mayor número de páginas únicas recuperan (Tabla IV), aunque en diferente orden, lo cual puede ser debido a que las páginas repetidas en Lycos y en los demás motores sean diferentes, mientras que el solapamiento de Excite corresponda a páginas que aparecen repetidas en diferentes motores de búsqueda.

Tabla V. Solapamiento entre los diferentes motores de búsqueda (Base=total registros en cada motor).

	Altavista	Excite	FastSearch	Google	HotBot	Infoseek	Lycos	N Light	Yahoo
Altavista	200	17 (8,5%)	48 (24%)	34 (17%)	32 (16 %)	22 (11%)	3 (1,5%)	26 (13%)	12 (6%)
Excite	17 (8,5%)	200	16 (8%)	12 (6%)	20 (10%)	14 (7%)	3 (1,5%)	10 (5%)	14 (7%)
FastSearch	48 (24%)	16 (8%)	200	28 (14%)	39 (19,5%)	21 (10,5%)	4 (2%)	21 (10,5%)	15 (7,5%)
Google	34 (17%)	12 (6%)	28 (14%)	200	34 (17%)	34 (17%)	11 (5,5%)	25 (12,5%)	20 (10%)
HotBot	32 (16%)	20 (10%)	39 (19,5%)	34 (17%)	200	35 (17,5%)	4 (2%)	16 (8%)	38 (19%)
Infoseek	22 (11%)	14 (7%)	21 (10,5%)	34 (17%)	35 (17,5%)	200	5 (2,5%)	9 (4,5%)	15 (7,5%)
Lycos	3 (1,5%)	3 (1,5%)	4 (2%)	11 (5,5%)	4 (2%)	5 (2,5%)	200	4 (2%)	7 (3,5%)
Northern Light	26 (13%)	10 (5%)	21 (10,5%)	25 (12,5%)	16 (8%)	9 (4,5%)	4 (2%)	200	8 (4%)
Yahoo	12 (6%)	14 (7%)	15 (7,5%)	20 (10%)	38 (19%)	15 (7,5%)	7 (3,5%)	8 (4%)	200

Este buen comportamiento de Lycos en cuanto a poco grado de solapamiento contrasta con los resultados de la Tabla I donde aparece como el buscador con mayor número de recursos no operativos. Al objeto de poder averiguar si este solapamiento real difiere del solapamiento existente con sólo las páginas potencialmente relevantes, quitando del numerador y denominador las páginas duplicadas en cada motor de búsqueda y aquellas con enlaces rotos, se muestran los resultados en la Tabla VI.

Como se puede observar, el número de páginas de la Tabla VI es prácticamente idéntico, salvo diferencias de 1 caso en alguna celda, con los datos de la Tabla V, lo que confirma que casi todas las páginas solapadas corresponden a páginas operativas y no duplicadas en el mismo buscador. Lo que sí cambia de manera importante es el

denominador utilizado para calcular el porcentaje de la Tabla VI, y que es la diagonal de donde coincide el motor de la fila con el motor de la columna.

En este caso, la Tabla VI ya no es simétrica y los porcentajes corresponden a cada columna, de manera que la primera columna son los porcentajes de solapamiento de Altavista con los demás motores de búsqueda, la segunda el solapamiento de Excite con respecto a los demás, y así sucesivamente. Aunque cambian ligeramente los porcentajes en relación con los de la Tabla V, el análisis de los resultados llevaría a los mismos comentarios, por lo que la conclusión sería que el solapamiento real de las bases de datos no difiere del solapamiento eficaz (páginas potencialmente relevantes).

Tabla VI. Solapamiento entre los diferentes motores de búsqueda (Base=registros operativos y no duplicados).

	Altavista	Excite	FastSearch	Google	HotBot	Infoseek	Lycos	N Light	Yahoo
Altavista	179	16 (8,8%)	48 (26,8%)	34 (17,4%)	32 (17,5%)	21 (12,6%)	3 (1,9%)	26 (14,1%)	12 (8,4%)
Excite	16 (8,9%)	182	16 (8,9%)	12 (6,2%)	20 (10,9)	14 (8,4%)	3 (1,9%)	10 (5,4%)	14 (9,8%)
FastSearch	48 (26,8%)	16 (8,8%)	179	28 (14,4%)	39 (21,3%)	21 (12,6%)	4 (2,6%)	21 (11,4%)	14 (9,8%)
Google	34 (19%)	12 (6,6%)	28 (15,6%)	195	34 (18,6%)	34 (20,4%)	11 (7,1%)	25 (13,5%)	20 (14%)
HotBot	32 (17,9%)	20 (11%)	39 (21,8%)	34 (17,4%)	183	35 (21%)	4 (2,6%)	16 (8,6%)	38 (26,6%)
Infoseek	21 (11,7%)	14 (7,7%)	21 (11,7%)	34 (17,4%)	35 (19,1%)	167	4 (2,6%)	9 (4,9%)	15 (10,5%)
Lycos	3 (1,7%)	3 (1,6%)	4 (2,2%)	11 (5,6%)	4 (2,2%)	4 (2,4%)	155	4 (2,2%)	7 (4,9%)
Northern Light	26 (14,5%)	10 (5,5%)	21 (11,7%)	25 (12,8%)	16 (8,7%)	9 (5,4%)	4 (2,6%)	185	8 (5,6%)
Yahoo	12 (6,7%)	14 (7,7%)	14 (7,8%)	20 (10,3%)	38 (20,8%)	15 (9%)	7 (4,5%)	8 (4,3%)	143

En resumen, podemos destacar que el solapamiento no es muy alto, lo que sugiere que pueda ser debido a que en realidad los documentos indizados en los motores de búsqueda sean bastante diferentes, o a que los algoritmos de relevancia sean muy diferentes y den un rango (orden) a los documentos tan distinto que coinciden muy pocos dentro de las primeras 50 páginas. Considerando que los algoritmos de relevancia no deberían de cambiar tanto, o dicho de otra manera, que en los primeros 50 resultados debería de haber un solapamiento mayor aún cuando el rango de relevancia no fuera el mismo, nos induce a pensar que hay otro factor que influye en el bajo solapamiento y puede ser debido a que los documentos indizados en cada base de datos de los motores de búsqueda son diferentes.

5. **Rendimiento global**. En la Tabla VII se muestra el rango obtenido por cada motor de búsqueda en las diferentes características estudiadas. Los tres motores de búsqueda que

mejor puntuación total obtienen son Google, Excite y Northern Light, con bastante diferencia con respecto a los demás, y Yahoo el que menor puntuación consigue.

Tabla VII. Rendimiento global de los motores de búsqueda

Motor	Enlaces no operativos	Duplicados	Unicos	Total
Google	8	8	6	22
Excite	7	5	9	21
Northern Light	5	9	7	21
FastSearch	4	7	4,5	15,5
HotBot	9	4	1	14
Infoseek	2	6	4,5	12,5
Lycos	1	3	8	12
Altavista	6	2	3	11
Yahoo	3	1	2	6

También es importante conocer el grado de solapamiento entre los motores de búsqueda que han conseguido mayor rendimiento; así, Google, tiene un solapamiento de sólo un 6% con Excite y un 12,5% con Northern Light, y Excite un 5,5% con Northern Light.

Conclusiones

Los estudios sobre evaluación y rendimiento de los motores de búsqueda son importantes al permitirnos conocer mejor sus características y tomar la decisión de cuál(es) elegir ante determinados temas de búsqueda, aunque las diferentes metodologías utilizadas llevan a resultados discrepantes y conclusiones contradictorias.

En este estudio se han analizado 50 páginas de cada búsqueda por motor, que supera ampliamente las 10 o 20 que utilizan la gran mayoría de los estudios que a nosotros nos parece claramente insuficiente para medir cualquier característica precisamente por la variabilidad existente entre los motores de búsqueda. Nos hemos centrado en la medida de características objetivas: enlaces no operativos, duplicidad en un mismo motor de búsqueda, cobertura de páginas únicas y solapamiento, huyendo de la subjetividad que implica el estudio de la precisión de los resultados basándose en la relevancia y de la falta de definición de ésta, ya que un documento puede ser relevante para la necesidad de información, para la expresión de la necesidad de información (consulta), o para la estrategia de búsqueda, y para cada una de las situaciones las personas encargadas de evaluar la relevancia pueden ser los propios usuarios, expertos en el tema o los que realizan las búsquedas.

Las limitaciones de este estudio vienen determinadas por el ámbito temático de las búsquedas realizadas y por la población a estudio (50 primeras páginas recuperadas), sin realizar una selección aleatoria de las páginas recuperadas en cada motor por las dificultades que esto entraña, por lo que las conclusiones no se pueden extrapolar al comportamiento de estos motores de búsqueda ante cualquier tema; sería necesario realizar otros estudios utilizando la misma metodología pero con diferentes temas de

búsqueda para averiguar si el comportamiento de los motores de búsqueda ofrece los mismos resultados que en este trabajo.

Aún así, las conclusiones de este estudio son las siguientes:

1. HotBot, Google y Excite son los buscadores que menor número de enlaces no operativos presentan, lo que indica que son los que mantienen sus bases de datos más actualizadas.
2. Northern Light y Google son los que recuperan menor número de páginas duplicadas, lo que indica que sus programas de indización funcionan más correctamente que los demás, ya que son capaces de detectar registros ya incluidos en sus bases de datos.
3. Excite, Lycos y Northern Light contribuyen con más páginas únicas y, por consiguiente, tienen menor solapamiento con los demás motores de búsqueda, lo que indica que ante estos temas de búsqueda ofrecen una mayor cobertura y exhaustividad de documentos potencialmente relevantes.
4. Ante cualquier tema de búsqueda, es recomendable utilizar varios buscadores. Google, Excite y Northern Light han sido los más eficientes en la recuperación de información, exigiendo menor esfuerzo y tiempo por parte del usuario para ver páginas potencialmente relevantes y con mayor cobertura de información única.

Debido a la naturaleza dinámica de Internet y a los cambios que se producen en los motores de búsqueda, coincidimos con Olvera[23] en que es necesaria la realización de estudios de evaluación en estas herramientas para conocer y analizar su evolución a través del tiempo.

Bibliografía

1. Maldonado Martínez, A.; Fernández Sánchez, E. (1999). "Comparing Internet search tools". *Online Information 99 Proceedings*, pp. 263-266.

2. Oppenheim, C.; Morris, A.; McKnight C. (2000). "Progress in documentation. The evaluation of WWW search engines". *Journal of Documentation*, vol. 56, nº 2, Marzo, pp. 190-211.

3. Sullivan, Danny. (2000). Search Engine Watch. http://searchenginewatch.com/reports/seindex.html (consultado el 25 de Enero del 2001).

4. Cohen L. (2000). Conducting research on the Internet. http://library.albany.edu/internet/research.html (consultado el 25 de Enero del 2001).

5. Cyberatlas. (2000). Top 50 sites of December 2000. http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,,5931_558551,00.html (consultado el 25 de Enero del 2001).

6. Green, D. (1999). "The evolution of Web searching". *Online Information 99 Proceedings*, pp. 251-258.

7. Lawrence, S.; Giles, C.L. (1999). "Searching the Web: general and scientific information access". *IEEE Communications*, vol. 37, nº 1, pp. 116-122. http://www.neci.nj.nec.com/~lawrence/papers/search-ieee99/ (consultado el 26 de Enero del 2001).

8. Lawrence, S.; Giles, C. (1999). "Accesibility of information on the web". *Nature*, vol. 400, pp. 107-109. (El resumen se puede consultar en http://wwwmetrics.com).

9. Bar-Ilan, J. (1998/9). "Search Engine results over time – A case study on search engine stability". *International Journal of Scientometrics, Informetrics and Bibliometrics*, vol. 2/3, nº 1.
http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html (consultado el 26 de Enero del 2001).

10. Krishna Bharat; Andrei Broder. (1998) "A technique for measuring the relative size and overlap of public web search engines". In *Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia*, pp. 379-388. Elsevier Science. (http://www-sor.inria.fr/mirrors/www7/programme/fullpapers/1937/com1937.htm, consultado el 22 de Enero del 2001).

11. FastSearch (2001). http://www.alltheweb.com/faq.php3 (consultado el 26 de Enero del 2001).

12. Bradley, P. (1999). "The great search engine con-trick". *Online Information 99 Proceedings*, pp. 259-262.

13. Gordon, M.; Pathak, P. (1999). "Finding information on the World Wide Web: the retrieval effectiveness of search engines". *Information Processing and Management*, vol. 35, pp. 141-180.

14. Leighton, H.V.; Srivastava, J. (1997). Precision among World Wide Web search services (search engines): Altavista, Excite, HotBot, Infoseek, Lycos. http://www.winona.msus.edu/library/webind2/webind2.htm (consultado el 24 de Enero del 2001).

15. Jansen, B.; Spink, A.; Sarcevic, T. (2000). "Real life, real users and real needs: A study and analysis of user queries on the web". *Information Processing and Management*, vol. 36, nº 2, pp. 207-227.

16. Silverstein, C.; Henzinger, M.; Marais, H; Moricz, M. (1999). "Analysis of a very large Web search engine query log". *SIGIR Forum*, vol. 33, nº 1, pp. 6-12.

17. Nasios, Y.; Korinthios, G.; Despotopoulos, Y. (1998). "Evaluation of search engines". Proyecto PIPER (número SU 1112 [AD]). 60 páginas. http://piper.ntua.gr/reports/searcheng.pdf

18. Leighton, H.; Srivastava, J. (1999). "First 20 precision among World Wide Web search services (search engines)". *Journal of the American Society for Information Science*, vol. 50, nº 1, pp. 870-881.

19. Chu, H.; Rosenthal, M. (1996). "Search engines for the World Wide Web: a comparative study and evaluation methodology". *ASIS 1996 Annual Conference Proceedings*, October 19-24.

(http://www.asis.org/annual-96/ElectronicProceedings/chu.html, consultado el 21 de Enero del 2001).

20. Wishard, L. (1998). "Precision among Internet search engines: an earth sciences case study". *Issues in Science and Technology Librarianship*, nº 18. http://www.library.ucsb.edu/istl/98-spring/article5.html (consultado el 26 de Enero del 2001).

21. Sullivan, Danny. (1999). Search engine covered study published. http://searchenginewatch.internet.com/sereport/99/08-size.html

22. Notess, Greg R. (2000). http://www.searchengineshowdown.com/stats/sizeest.shtml

23. Olvera Lobo, Mª D. (2000). "Rendimiento de los sistemas de recuperación de información en la world wide web: revisión metodológica." *Revista Española de Documentación Científica*, vol. 23, nº 1, pp. 63-77.