Tesis Doctoral

Evaluación del funcionamiento y recuperación de información textual de los principales motores de búsqueda y metabuscadores genéricos de la World Wide Web

Memoria presentada por

Francisco Javier Vidal Bordes

en la Facultad de Filosofía y Letras de la Universidad de Zaragoza

para optar al título de Doctor.

Director: **Dr. José Antonio Salvador Oliván**

2008

A la memoria de

Antonio Vidal Bordes

EVALUACIÓN DEL FUNCIONAMIENTO Y RECUPERACIÓN DE INFORMACIÓN TEXTUAL DE LOS PRINCIPALES MOTORES DE BÚSQUEDA Y METABUSCADORES GENÉRICOS DE LA WORLD WIDE WEB

Agradecimientos

Quiero agradecer a todas las personas que de un modo u otro han posibilitado la realización del presente trabajo. En primer lugar tengo que dar las gracias al Profesor José Antonio Salvador Oliván, por su confianza en mí, al proponerme la realización de un trabajo tan interesante y por su inestimable colaboración en la resolución de cuantos problemas se me han ido planteando en las diferentes fases de su realización.

También tengo que agradecer el apoyo prestado a personas como José Mª Meléndez Vidal, que inicialmente me facilitó la consulta de los primeros trabajos en formato electrónico, y a Jorge Vidal Serrano por sus propuestas para facilitar la lectura de este trabajo.

Tampoco puedo olvidarme del grupo de amigos que se prestaron de forma voluntaria para la realización de las búsquedas en los diferentes buscadores. En este sentido, quiero agradecer a Mª del Carmen Montolio, Susana Casaña Oliver, Blanca Angulo, Jorge Vidal Serrano, Domingo Bel Gaudó, Luis Fatás Fernández, Ernesto Hernández Mareca, José Romero Sánchez y al profesor Vicente Ramón Palerm, el entusiasmo con el que participaron en las dos sesiones de esta fase.

No puedo olvidarme de todas aquellas personas que han estado cerca de mí, animándome y facilitándome lo necesario para que este trabajo saliera adelante. Así pues, debo dar las gracias a Matilde Cantín Luna, Directora de la Biblioteca María Moliner de la Facultad de Filosofía y Letras de la Universidad de Zaragoza por facilitarme el uso de sus instalaciones para llevar a cabo la parte práctica de la evaluación, a Antonio Cubillo, Operador Informático de la Facultad de Filosofía y Letras, por su cooperación en el almacenamiento de la documentación electrónica en el Servidor de la Facultad, y a Agustín Urdangarín, director del Centro de Documentación Científica de la Universidad de Zaragoza, por su colaboración en la solución a alguno de los problemas informáticos que surgieron a lo largo del trabajo, y al Profesor Angel López Molinero, de la Facultad de Ciencias por sus consejos para la realización del material gráfico. También quiero mencionar, por su constante apoyo y ánimo recibido a lo largo del trabajo, al Profesor Francisco Marco Simón y a su mujer, Cristina y a su hija, Cristina Marco.

Finalmente, he de dar las gracias a mi familia, que desde el principio han mostrado todo su apoyo y comprensión, facilitándome la realización de este trabajo.

"La reflexión utópica es irrenunciable para el pensamiento político y social. Es una prueba insustituible para demostrar qué poco resistentes son de hecho los fines y los prejuicios que guían la acción política. Su principal objetivo es aumentar la precisión de lo que estamos en condiciones de pretender, de lo que podemos esperar razonablemente. También nos ayuda a clarificar qué es lo que podemos exigirnos unos a otros como miembros de comunidades locales y globales, por qué situaciones vale la pena ponerlo todo en juego."

Daniel Innerarity *La sociedad invisible.* Pozuelo de Alarcón (Madrid), Espasa-Calpe, 2004.

I. INTRODUCCIÓN

Introducción y objetivos

Uno de los fenómenos más destacados, no sólo en el campo de la información, y que más ha marcado el final del siglo anterior, ha sido el nacimiento y desarrollo de la red Internet.

Como señala Villaseñor (2000:31) el término Internet se refiere a:

> "todos los instrumentos y recursos que sirven para satisfacer las necesidades informativas de cualquier persona, se hayan creado o no con ese fin y sean utilizados directamente o por un profesional de la información como intermediario".

Por tanto la red Internet y concretamente uno de sus servicios más utilizados, la World Wide Web, constituyen una nueva fuente de información a la que cada vez se accede con mayor frecuencia, pues permite obtener diferentes recursos[1] e información, bien sea comercial, técnica o científica sobre un determinado asunto o tema de interés.

[1]Las ISBD(ER) (International Standard Book Description. Electronic Resources) consideran "recurso electrónico" a todo material que requiere el uso del ordenador y otros componentes periféricos para su manipulación, por lo que en nuestro caso utilizaremos la expresión "recurso de información" para referirnos tanto a los sitios, servicios, como a las páginas web y en general a todo tipo de archivos electrónicos que contienen información de tipo textual y audiovisual accesible a través de Internet.

No vamos a detenernos aquí a enumerar las ventajas que este medio ofrece para la difusión y acceso a la información, pues son de sobra conocidas. Nuestro interés reside, por un lado, en analizar la problemática que presenta el acceso a la documentación especializada en la World Wide Web (en adelante la Web) a través de los servicios de búsqueda de información generales, ¿A qué se debe?¿Qué aspectos influyen negativamente en la recuperación?

En primer lugar hemos de referirnos a las características de su contenido, ya que por ejemplo, el libre intercambio de información y la facilidad de publicación es una de las causas de que la información disponible tanto en Internet como en la Web, crezca a gran ritmo[2]. Este gran cúmulo de información, en constante transformación, contiene recursos electrónicos de todo tipo (textual, imagen, sonido, audiovisual o una mezcla de ellos en cualquiera de sus formatos) de desigual contenido intelectual y por regla general, sin un tratamiento documental básico. Los recursos pueden ser catálogos, bases de datos, enciclopedias, programas de software, además de bibliotecas digitales, etcétera.

En relación con el primero de estos aspectos, el contenido intelectual, conviven documentos con contenidos de escaso interés junto a información de tipo especializado y científico como por ejemplo la existente en sitios Web o en Weblogs de un gran número de investigadores o la que ofrecen los editores de publicaciones, o los responsables de grupos y centros de investigación, de empresas, etcétera. La mayor parte de estos trabajos tiene un gran interés para la investigación. El problema reside en conocer hasta qué punto se recupera esta información de calidad, en saber cómo funcionan los principales buscadores ante búsquedas especializadas y si responden a las necesidades de información de los usuarios.

Respecto al tratamiento documental, la mayoría de recursos carecen de descripción y clasificación, lo que dificulta su recuperación de una forma precisa. Como solución al problema se han ido desarrollando, por parte de la comunidad científica, iniciativas basadas tanto en la utilización de sistemas de metadatos para la descripción y clasificación de recursos, como en el uso del formato MARC y de lenguajes estructurados.

[2] Estudios sobre el número de páginas web existente en Internet estimaban una cifra superior a 1.500 millones en 1999 (Aguillo 1999). Otro estudio de Murray y Moore (2000) publicado por Cyveillance señalaba unos 2.000 millones de páginas calculando para el 2001 una cifra de 4000 millones.

No obstante, estas acciones, aunque están muy desarrolladas, por el momento no han resuelto la totalidad del problema, ya que a su vez generan otros problemas que han de resolver, como los relacionados con su normalización, con quién añade los metadatos a los recursos, etcétera.

La disposición de la información junto a la cantidad de recursos, influyen tanto en el diseño de las herramientas de búsqueda, como en su funcionamiento, cuyo rendimiento no puede ser satisfactorio, o mejor dicho, no puede ser igual de satisfactorio para todo tipo de usuarios, por lo que nuestro estudio trata de tener en cuenta a los usuarios especialistas que utilizan los buscadores genéricos para obtener información sobre uno o varios temas.

Conscientes de los problemas de recuperación que los buscadores generales plantean, surgen nuevas herramientas especializadas, en la recuperación de recursos de un determinado tema o tipología documental[3].

En este sentido, uno de los principales problemas que se plantean al realizar búsquedas es saber qué buscador utilizar[4]. El usuario habitualmente usa un determinado buscador por costumbre, sin preocuparse por la calidad de los resultados ya que entre ellos siempre es posible recuperar algo de interés. Sin embargo, una correcta elección puede dar mejores resultados y ahorrar un tiempo valiosísimo.

Un interesante estudio realizado por Stobart y Kerridge (1996) en la Universidad de Sunderland, en Inglaterra, sobre el uso de los motores de búsqueda en el ámbito académico, en el educativo y en el empresarial, señalan una frecuencia de uso muy alta para fines de investigación. A pesar de ello, los usuarios señalan, por este orden, los siguientes problemas: la lentitud (26%), el exceso de información recuperada (20%), recuperación de recursos anticuados (20%), la poca o nula información sobre el contenido de los enlaces (8%), recuperación de información no requerida (6%), explicación poco clara de los resultados recuperados (6%), poca claridad de las instrucciones de funcionamiento (5%), correspondiendo en 9% restante a otros problemas.

[3]Lycos estima en más de 100 millones el número de buscadores potenciales en la Web (http://insite.lycos.com).

[4] Hípola y Bargas-Quesada (1999b) señalan la existencia de más de 3.000 buscadores en Internet.

Superado el primer problema de tipo técnico mediante la extensión de la banda ancha, los siguientes hacen referencia al exceso de información, a la poca actualidad de los recursos, a la poca o nula información de los registros sobre el contenido y a la poca precisión.

Para el usuario que requiere información especializada relacionada con uno o varios aspectos de un determinado campo científico, es de gran importancia conocer cuál o cuáles recuperan mayor información, y sobre todo, más precisa. Para buscar información de este tipo dispone de una serie de herramientas, en unos casos más especializadas y en otros más genéricas, pero en todo momento debe conocer qué ofrecen cada una de ellas, si son útiles, y en qué medida.

Ante este panorama es lógico que investigadores y profesionales de la recuperación de información se hagan preguntas como: ¿Qué ocurre en este contexto cuando un usuario utiliza alguna de las principales herramientas de búsqueda para recuperar información especializada? ¿Están preparadas estas herramientas para solucionar necesidades de información de este tipo?¿Qué función desarrollan estos sistemas ante este tipo de búsquedas? ¿Cuál o cuáles lo hacen mejor?¿Es aconsejable su uso? ¿Cuáles ofrecen un mejor rendimiento en estas búsquedas? ¿Cuáles ofrecen mejores resultados en relación con los temas de consulta? En definitiva ¿Son útiles los motores de búsqueda y en qué medida, para recuperar información de carácter especializado? o ¿Están principalmente orientados a facilitar información de tipo comercial o general?

La evaluación debe facilitarnos la respuesta a estas cuestiones, ya que debe permitirnos analizar, detectar y valorar las fortalezas y debilidades de estas herramientas. Solo así podrán ser corregidas estas últimas, lo que contribuirá, además, a mejorar su funcionamiento. La evaluación va a permitir conocer mejor sus prestaciones, saber hasta qué punto estos sistemas pueden resolver determinadas necesidades de información y permitirnos, finalmente, seleccionar las mejores.

Varios son los aspectos que hacen necesaria la evaluación. En primer lugar, como ya hemos señalado, el gran número de herramientas de recuperación de recursos web existente, calculándose que el número de buscadores supera los 3.700 (Delgado Martínez, 2001), lo que hace necesario conocer cuáles son los que recuperan recursos de mayor calidad respecto a los términos de búsqueda.

Por otro lado tenemos que la segunda función más importante de Internet es la búsqueda de información, y que como han demostrado diferentes autores, los buscadores

de carácter general son las herramientas que se usan de forma más frecuente para la búsqueda y acceso a recursos de información. Lawrence (2000) cita un estudio de GVU donde se muestra que el 85% de los usuarios utilizan estas herramientas para localizar información. En este sentido, las listas de sitios Web, muestran varios de estos motores y directorios como los servicios más visitados de la red[5].

Por otro lado cada vez van teniendo mayor importancia los servicios de referencia digital que requieren el conocimiento de las mejores herramientas para facilitar búsquedas precisas y resolver problemas de información.

La evaluación que planteamos debe ayudarnos no sólo a resolver estos problemas sino también a conocer el comportamiento y la utilidad de estas herramientas en búsquedas sobre temas especializados.

En el presente trabajo nos planteamos, mediante su evaluación, valorar cómo funcionan los motores de búsqueda y metabuscadores más utilizados ante búsquedas simples y complejas, ante temas especializados, en nuestro caso relacionados con el campo de la Documentación ya que, aunque han sido muchos los trabajos de evaluación sobre los motores de búsqueda genéricos de la Web que se han publicado en los últimos años, han sido pocos los que se han ocupado de valorar su utilidad en recuperación de información científica.

La evaluación ha de realizarse teniendo en cuenta los objetivos que se persiguen. Por tanto, a la vista de los problemas señalados, nos planteamos evaluar los servicios de búsqueda desde la perspectiva de su rendimiento técnico, para lo cual nos centraremos en el análisis de la formación de la base de datos, de los índices, de la consulta a las bases de datos y la recuperación de información de documentos de tipo textual[6]. Nos interesa conocer su comportamiento ante diferentes tipos de búsqueda, la información que aportan sus registros, la actualización de sus bases de datos, duplicados, las características de los recursos que aportan, la tipología documental, el carácter comercial o más especializado la precisión técnica, el solapamiento entre motores y la ordenación. Para ello deberemos

[5]Hípola y Vargas-Quesada (1999) señalan que en 1998 la media de búsquedas diarias se situaban en torno a los 18.300.000.

[6]Cada vez esta más generalizada la separación de los diferentes tipos de recursos en bases de datos distintas ya que se pueden consultar individualmente según tengamos necesidad de recuperar imágenes, documentos sonoros o audiovisuales, noticias de prensa o de otro tipo.

establecer los criterios e indicadores de evaluación, cuyos valores nos permitirán cuantificar determinados aspectos relacionados con su funcionamiento y recuperación.

Planteamos una metodología centrada en el análisis de los aspectos relacionados con los objetivos propuestos, que sea simple pero consistente, de manera que permita, como hemos señalado, no sólo realizar comparaciones entre los sistemas, sino también repetirse cada cierto tiempo, y observar su evolución.

Para la obtención de los datos que nos han de servir de muestra en el análisis, lanzamos diferentes tipos de búsqueda. La primera se basa en la recuperación de documentos que traten sobre un término de búsqueda. La segunda es una sucesión de términos que corresponde a una búsqueda en lenguaje natural. Para analizar su funcionamiento ante búsquedas avanzadas hemos realizado una tercera búsqueda basada en el uso de operadores booleanos, la cuarta incluye operadores de existencia, la quinta, es una búsqueda de frase y la sexta es una búsqueda en el campo título.

Un estudio de evaluación previo al presente trabajo[7], nos permitió observar la especificidad de la mayoría de trabajos de evaluación de estas herramientas, centrándose en la valoración de determinados criterios y fundamentalmente en la precisión, ya que se considera uno de los aspectos más importantes en los sistemas de recuperación de información. Por otro lado advertimos la necesidad de desarrollar una metodología que permitiera valorar estas herramientas atendiendo tanto a las características como al funcionamiento de sus componentes.

Para evitar o al menos minimizar la subjetividad, utilizaremos una metodología basada en valores estadísticos.

El presente estudio se estructura básicamente en tres grandes apartados. En el primero de ellos, de carácter introductorio, analizamos el marco contextual en el que se halla la información, es decir, la red Internet y la World Wide Web, prestando especial atención tanto al origen y la evolución como a las características de la información y servicios que contienen. El siguiente punto de interés lo componen los principales sistemas que se utilizan para recuperarla, es decir los buscadores y sus clases, su funcionamiento, ocupándonos también de los sistemas de recuperación de información tradicionales, ya

[7]SALVADOR OLIVÁN, José Antonio y VIDAL BORDES, Fco. Javier. (2000)

que suponen el origen de las herramientas Web, aunque ambos desempeñan sus funciones ante colecciones documentales sensiblemente diferentes. Esto nos lleva a analizar los factores internos y externos que influyen en la recuperación de recursos web. Al final del apartado introductorio abordamos el tema de la evaluación de los sistemas de recuperación, fundamentalmente de los buscadores de Internet, analizando experiencias previas que recogemos en el estado de la cuestión, y que a su vez nos han servido para seleccionar los indicadores utilizados en nuestra evaluación.

El segundo apartado se ocupa de explicar la metodología adoptada para la evaluación, dedicando el tercer apartado a exponer y analizar los resultados, de los que se extraen las conclusiones que reflejamos en el apartado final.

1. Internet y la World Wide Web

1.1. La red Internet

La red Internet, cuyo término responde a la contracción de Internetworking of computers, es una red de ordenadores conectados siguiendo una arquitectura distribuida, que está formada por la interconexión de diversas redes de ordenadores de todo el mundo que utilizan protocolos de comunicación TCP/IP y soporta diferentes tipos de plataformas.

El Federal Networking Council (FNC) definió en 1995 Internet, como un sistema de información global que está interconectado por un único espacio de direcciones basado en el Internet Protocol (IP) o sus futuras extensiones o adiciones, capaz de soportar comunicaciones que usen el conjunto Transmission Control Protocol/Internet Protocol (TCP/IP), sus futuras extensiones o adiciones y otros protocolos compatibles, y proporciona, utiliza o facilita el acceso público o privado, a servicios de alto nivel basados en la infraestructura de comunicaciones.

Nogales (1999a) recoge una definición muy acertada del ISOC, en la que se hace referencia no sólo a lo que es, sino también a lo que supone. Así, Internet es una:

> "red global de redes que permite a toda clase de ordenadores comunicarse y compartir servicios de forma directa y transparente a través de buena parte del mundo. Puesto que Internet es un potencial enormemente valioso que ofrece tantas posibilidades para tantas personas y organizaciones, también constituye un recurso global y compartido de información y conocimiento, y un medio de colaboración y cooperación entre innumerables comunidades diferentes".

Podemos destacar, por tanto, entre sus funciones básicas, el constituir un sistema de comunicación y de información, lo que permite conectarse a la red, prácticamente desde cualquier ordenador personal, tanto para difundir como para acceder a la información.

1.1.1. Breve introducción histórica

Aunque no es objetivo prioritario de nuestra investigación tratar de la historia de Internet, una visión histórica desde sus orígenes hasta la actualidad puede ayudarnos a comprender mejor tanto el concepto y su situación actual, como la variedad de servicios que la componen.

1.1.1.1. Primeras investigaciones. Las redes y los protocolos

Existe en la actualidad un importante número de estudios dedicados a analizar la red Internet desde distintos puntos de vista. El profesor Ubieto Artur (1995) y Alonso Berrocal y otros (2004), en sus respectivas obras sobre el tema, dedican un interesante capítulo a explicar la gestación del fenómeno Internet.

La red Internet como la conocemos en la actualidad, es fruto tanto de la evolución de las redes y protocolos de comunicación, como de programas informáticos, y su éxito a finales de los 80, coincide con la aparición de la Web y la adquisición de un mayor carácter comercial, frente a su utilidad inicial, más relacionada con el ámbito militar y científico.

Aunque no hay acuerdo unánime sobre la fecha que marca el inicio de Internet, algunos autores la sitúan a inicios de los años 60 con Robert Taylor, director de la oficina de técnicas de proceso de datos de ARPA[8] y del Computer Research Program, quien empezó a investigar la posibilidad de conectar ordenadores de diferentes centros de investigación. Poco después, en 1964 Paul Baran, del laboratorio de investigación americano RAND Corporation, desarrolla los conceptos de redes de comunicación distribuida, formadas por varios nodos interconectados sin que exista uno central, y de conmutación de paquetes, que se desarrollarán más adelante en el seno de ARPA, financiando proyectos de investigación dedicados a la informática, a la tecnología y al tratamiento de la información. Los primeros frutos en la creación de la red darán lugar a ARPANET[9] en 1969, constituyéndose en el elemento aglutinante de las investigaciones desarrolladas hasta el momento y aún de las que tendrán lugar en años posteriores. Para Leiner (1997) los orígenes hay que situarlos en 1962, cuando Joseph Carl Robnett Licklider, del MIT (Massachussets Institute of Technology), planteó la posibilidad de llevar a cabo una red interconectada globalmente que permitiera el acceso desde cualquier lugar a datos y programas desarrollando la idea posteriormente, con su paso a ARPA.

[8]Institución americana dedicada al desarrollo de proyectos de investigación que ha ido cambiando de nombre varias veces consecutivas, entre ARPA y DARPA (Defense Advanced Research Projects Agency), que es como se la conoce en la actualidad.

[9]Inicialmente la constituyeron nodos localizados en la Universidad de California, Los Angeles, (UCLA), el Instituto de Investigación de Stanford, la Universidad de California, en Santa Bárbara, y la Universidad de Utah. La red ARPANET dejó de funcionar en 1991.

Leonard Kleinrock también se ocupaba desde 1961 de desarrollar la conmutación de paquetes, método de telecomunicación que resultó fundamental y que llevado a la práctica en 1967 por Lawrence G. Roberts, de DARPA, permitió conectar dos ordenadores remotos a través de la línea telefónica para el intercambio de datos.

Por otro lado, en el NPL (National Physical Laboratory), Donald Davies y Roger Scantlebury se ocupaban a su vez de la investigación en redes de transmisión.

En 1972 se produce la primera demostración pública de la red a cargo de Robert E. Kahn en la International Coference on Computer Communications, y ya en 1988, la red Internet alcanza una mayor popularidad, con la creación de redes locales y privadas que favorecerán la introducción de contenidos más variados, y de un uso más comercial.

Al margen de estos aspectos puntuales, respecto a la evolución técnica, un hecho importante es la adopción en 1983 como estándar del conjunto de protocolos de comunicación TCP/IP (Transmission Control Protocol/Internet Protocol)[10], más seguro que su antecesor NCP (Network Control Protocol). Esto, unido al desarrollo de las pasarelas entre redes (Gateway), permitió la convivencia de distintas tecnologías y diferentes redes. El trabajo fue fruto de la colaboración entre Robert E. Kahn y Vinton Cerf de la Universidad de Stanford.

Los protocolos OSI (X.25, X.400, X.500)[11], que se venían utilizando en la década de los 80 en las redes académicas y de investigación europeas, fueron sustituidos por servicios basados en protocolos TCP/IP. De aquí que algunos autores sitúen en 1983 la fecha de nacimiento de Internet, ya que es cuando se generaliza el uso de este protocolo, cuyo nombre, además, contiene dicho término. De hecho, en los años 80 se observa un gran interés en el desarrollo de redes en universidades americanas (USENET, BITNET, etcétera), que a lo largo de esta década irán uniéndose entre si, para favorecer el intercambio de información.

[10]Siglas que significan Transmission Control Protocol/Internet Protocol, que en 1978 sustituyen al Protocolo NCP (Network Control Protocol) en el que se aprecian limitaciones.

[11]Sanz considera las conexiones que se realizaban desde el Departamento de Ingeniería Telemática de la Escuela Superior de Ingenieros de Telecomunicaciones de la Universidad Politécnica de Madrid con la red EUnet, y a través de ella con la USENET americana, como pioneras en el uso de Internet desde España. Los servicios que se obtenían se basaban en la mensajería electrónica y en grupos de noticias.

La unión de ARPANET con otras redes, tanto americanas (NSFNet[12] a mediados de los 80 o la National Research and Education Net, NREN)[13], como europeas (EUNet)[14] y de otros países asiáticos, además de la creación y conexión a finales de esta década, de otras redes científicas regionales[15], supuso un paso adelante en el desarrollo de la red, al alcanzar contenidos y fines favorables a la docencia e investigación. ARPANET acabará dividiéndose en dos redes, una de carácter militar (MILNET) y otra para fines científicos y de investigación (NSFNet).

Pese a la desaparición de la red ARPANET en 1988, Internet cuenta ya con una gran popularidad, incrementada en los años noventa, en los que se aprecia un gran interés por conectarse de forma particular y acceder a los diferentes servicios soportados por la red.

1.1.1.2. Los servicios y sus protocolos. Las primeras herramientas de recuperación de información en Internet

La mayoría de servicios de la red utilizan la estructura cliente-servidor, que consiste en la existencia de un ordenador con una serie de programas que le permiten actuar como servidor, al que el usuario se conecta desde el ordenador que contiene el programa cliente, y mediante el cuál puede solicitar bien un documento, una información, un servicio, etcétera. Para que la comunicación pueda producirse, ambos han de estar conectados a la red. De aquí que el primer servicio de la red suponía permitir la conexión entre diferentes servidores.

El correo electrónico se venía utilizando desde 1972 gracias al trabajo de Ray Tomlinson, que se ocupó de desarrollar un programa de gestión del correo. Posteriormen-

[12]La National Science Foundation Network fue creada en 1985 conectando inicialmente a cinco centros estadounidenses, posibilitando poco después la conexión al resto de la comunidad científica.

[13]Red que surgió con objetivos claros de impulsar el uso y desarrollo de las nuevas tecnologías en el ámbito educativo, a todos sus niveles, así como posibilitar la infraestructura necesaria para alcanzar estos objetivos.

[14]En 1982, unía Holanda, Dinamarca, Suecia y Reino Unido. Inicialmente tiene carácter académico pasando después a tener una dedicación más comercial.

[15]En Europa existían entre otras NORDUnet en los países nórdicos, INRIA en Francia, CNUCE en Italia, UCL en Reino Unido y CWI en Holanda. En 1984 surgen JANET en Reino Unido, DFN en Alemania y SUNET en Suecia. Dos años más tarde SURFnet en Holanda, en 1987 SWITCH en Suiza y al año siguiente RedIRIS en España y GARR en Italia. (Sanz, M. A.)

te se utilizó además para poner en contacto a equipos de personas que se inscribían en los llamados Grupos de discusión y Listas de distribución o de correo[16].

A finales de los años 70 se había creado, en el seno de la Universidad de Duke, en Carolina del Norte, la red USENET[17] centrada en favorecer el desarrollo de los grupos de discusión o Newsgroups[18].

Los protocolos utilizados por estos servicios varían; así, el correo electrónico utilizaba inicialmente el protocolo SMTP (Simple Mail Transfer Protocol) adoptando posteriormente el protocolo MIME (Multipurpose Internet Mail Extensions) que se mantiene en la actualidad, y que permite el envío y recepción de ficheros con distintos formatos. Dos nuevos protocolos de correo aparecen en los años 80, el POP (Post Office Protocol) que facilita la gestión del correo en los ordenadores personales y no en servidores específicos como hasta entonces y el IMAP (Internet Message Access Protocol), interfaz gráfica que mantiene los mensajes en el servidor. Las news utilizaron inicialmente el protocolo UUCP (Unix to Unix Copy), y desde 1984 el protocolo NNTP (Network News Transfer Protocol).

Aunque inicialmente estos servicios requerían programas específicos para acceder a ellos, por ejemplo LISTSERV y Majordomo para los grupos de discusión, actualmente son accesibles a través de la Web mediante los navegadores que se usan habitualmente.

Para localizar direcciones de correo se utilizaron tanto el programa Netfind, servidores WHOIS y los directorios X.500, que por problemas de actualización fueron perdiendo interés, por lo que se derivó, hacia las llamadas "páginas blancas", elaboradas con direcciones electrónicas de particulares, con los datos enviados por los propios interesa-

[16] En 1975 Steve Walker crea la primera Lista de correo (*mailing list*)

[17] Contracción de "User's Network" se refería tanto a la red como al servicio que a través de la línea telefónica permitía el intercambio de información especializada mediante un programa de mensajería electrónica. Posteriormente utilizó la red Internet. Los mensajes se depositaban en un servidor a modo de tablón de anuncios, por lo que podía ser consultado libremente.

[18] El motor de búsqueda Google permite actualmente acceder a la base de datos de USENET para realizar búsquedas de documentación relacionada con estos grupos.

dos, y las "páginas amarillas", con direcciones de empresas. Podemos mencionar en este sentido los directorios BigFoot[19], Four11[20] y WhoWhere[21].

Por otro lado, el servicio de transferencia de ficheros (FTP) entre ordenadores se desarrolla en los 90[22] junto con la conexión Telnet a ordenadores remotos, si bien, los orígenes del protocolo FTP se remontan a principios de los 70.

Se denomina File Transfer Protocol (FTP) tanto a la aplicación como al servicio que permite la transferencia de ficheros entre ordenadores, facilitando a los usuarios autorizados (User FTP), realizar un intercambio de ficheros entre distintos ordenadores, y a todo tipo de usuarios (anonymous FTP), la simple transferencia de ficheros al ordenador propio. Los servidores FTP pueden contener documentos en formato ASCII (American Standard Code for Information Interchange), PostScript, SGML, programas de ordenador y ficheros de imágenes y sonido (García Camarero, 2001).

Para facilitar la búsqueda de ficheros en este tipo de servidores, se creó en 1990 Archie[23], accesible al público desde 1992. Su capacidad de búsqueda en un gran número de servidores seleccionados es una de sus principales características. Constaba de una base de datos, de actualización quincenal en la que se registraban los nombres de archivos localizados en sitios FTP anónimos de Internet, otra de descripción del software, y un sencillo interfaz que se debía configurar en cada conexión. Para aligerar su trabajo se creó una red de servidores que contenían copias de la base de datos. Este servicio se utilizaba fundamentalmente mediante una conexión Telnet, un cliente Archie o también a través del correo electrónico[24].

Su utilización mediante programas cliente de carácter gráfico como Wsarchie facilitó en gran medida su mayor difusión. A pesar de todo, la capacidad de recuperación de información seguía siendo limitada.

---

[19]Facilita las búsquedas de personas de EEUU. Puede accederse a través de la Web en la dirección http://Search.bigfoot.com

[20]Actualmente pertenece a Yahoo.

[21]Actualmente lo controla Lycos.

[22] Existe cierta confusión entre las fechas que ofrecen determinados investigadores sobre estos servicios debido a que se trata de sistemas que pueden tener su origen experimental en años anteriores pero que o bien se dan a conocer más tarde o bien es posteriormente cuando adquieren interés o vigencia en la red.

[23]Evolucionó al buscador ArchiePlex, siendo sustituido en la actualidad por el buscador FTPSearch (Ubieto, 2002).

[24] Para más información sobre el funcionamiento de estas herramientas véase la obra de Gilster (1996).

Para realizar una conexión Telnet se requiere una aplicación como NCSA Telnet u otra similar, que permita la conexión de un ordenador a otro más potente o host, convirtiendo al ordenador que inicia la conexión en un Terminal, lo que hace posible la interactuación entre ordenadores, la compartición de aplicaciones y la gestión de un gran número de datos, por lo que este tipo de conexiones se utilizó desde un principio, fundamentalmente para acceder a bases de datos y a catálogos de bibliotecas. En este sentido, hemos de señalar la existencia de herramientas como Hytelnet (Hypertext browser for telnet-accesible sites)[25] y LIBS[26] que facilitaban el acceso, mediante Telnet u otro tipo de conexión, a los catálogos de un gran número de bibliotecas de forma directa o a través del protocolo Z.39.50. Les caracteriza el facilitar la búsqueda de recursos mediante la navegación a través de enlaces.

Importa destacar el hecho de que los lenguajes hipertextuales[27] comienzan a tener aplicación en Internet. El siguiente paso en este sentido se dará con el sistema Gopher.

Según Paul Gilster (1996) el programa Gopher y la tecnología que le acompaña aparecen en 1991 en el seno de la Universidad de Minnesota, en EEUU, como método de organización y recuperación de información relacionada con esta institución, extendiéndose posteriormente como sistema de navegación por la red, ya que permitía la conexión entre distintos servidores de información mediante la activación de enlaces y el acceso a diferentes recursos. Basado en la estructura cliente-servidor, soporta el protocolo y programa de consulta del mismo nombre, con la peculiaridad de permitir el acceso a documentos en formato texto, imagen o sonido, pero de forma aislada, es decir, sin integrarlos en un mismo documento.

La información disponible en este tipo de servidores se caracteriza por aparecer organizada jerárquicamente por temas, con documentos que pueden estar alojados en un

[25]Programa permitía acceder no sólo a sesiones Telnet en determinados servidores Telnet, Gopher, WAIS y World Wide Web, sino también a OPACS que no eran accesibles a través de la Web. El directorio Web-CATS (http://www.lights.com/webcats/) siguió sus pasos entre 1995-2000 y en la actualidad se accede desde el Directorio Libdex (http://www.libdex.com)

[26] (Library Internet Browsing Software). Información sobre este programa puede obtenerse en el trabajo de Stanton y Hooper (1992).

[27]Lenguaje que se estaba estudiando desde 1945 por Vannevar Bush, posteriormente por Ted Nelson que lo aplica en la creación de textos no lineales. Douglas Egelbart lo aplicó a textos en red.

mismo servidor o en servidores distintos, a los que se accede finalmente pulsando sobre el nombre del documento que actúa como hiperenlace.

La búsqueda de información en servidores Gopher se realizaba desde 1992 a través de los programas Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) y Jughead (Jonzy's Universal Gopher Hierarchy Excavation and Display). El primero recuperaba los términos de los títulos y títulos de apartados, pero no del texto (Winship, 1995). Soportaba la lógica booleana, truncamientos y permitía dirigir la búsqueda sobre determinado tipo de recursos, aunque planteaba el problemas, como señala la Poulter, de que la información quedaba frecuentemente desfasada en algunos servidores y que el acceso se hacía cada vez más difícil, a causa de las limitaciones de los servidores. Por otro lado los menús en los que se basaba la indización no requerían una norma estándar, lo que planteaba problemas para la recuperación. Este autor menciona además la existencia de determinados servidores con información especializada que pasaban a formar parte de directorios organizados por materias, entre los que destacaba el denominado "Gopher Jewels". Respecto al buscador Jughead, el funcionamiento era muy similar al anterior, si bien permitía dirigir la búsqueda a un espacio Gopher más limitado, al centrarse en el servidor de una determinada empresa, institución, etcétera.

Los servidores WAIS (Wide Area Information Server o Servicios de Información de Área Extensa) a los que se accedía a través del programa cliente del mismo nombre, conectaban con bases de datos distribuidas, fundamentalmente de tipo textual, indizadas a texto completo y dispuestas a modo de directorio. Esto permitió no sólo consultar estas bases de datos para localizar información sobre determinados términos, sino también facilitó el acceso a catálogos automatizados de bibliotecas, a revistas electrónicas así como a documentos multimedia. El sistema, operativo desde 1991, utilizaba una extensión del protocolo Z39.50[28] y se podía acceder a través de una conexión Gopher o bien a través de la World Wide Web. Tras su consulta se obtenía una lista de artículos o citas (Ruiz de Osma, 1998) ordenados en función de la mayor o menor aparición de los términos de búsqueda, posibilitando la obtención del documento completo, ya fuera de tipo

[28]Conjunto de normas y procedimientos que regulan el comportamiento entre ordenadores con diferentes sistemas informáticos posibilitando que se comuniquen entre sí, facilitando la búsqueda de información en varias bases de datos de forma simultánea, unificando sus opciones de búsqueda. Se trata de una norma americana que tienen sus equivalentes en las normas ISO 10162 e ISO 10163.

texto, imagen o sonido. La búsqueda se realizaba mediante palabras clave y lenguaje natural, y dejaba recuperar todo tipo de documentos de Internet así como información existente en listas de correo y news. Desde el punto de vista técnico también supuso un avance al facilitar la retroalimentación de la relevancia, permitiendo redirigir las búsquedas a partir de los resultados obtenidos. Fue comprado por la compañía American Online (AOL), y en la actualidad el proyecto no ha tenido continuación, pudiéndose consultar algunas de las bases de datos disponibles a través de pasarelas Web.

Para que todo funcione, existen organizaciones dedicadas a la coordinación y desarrollo de diferentes aspectos relacionados con Internet como la ICB (International Cooperation Board), el IETF[29] (Internet Engineering Task Force), el IRTF[30] (Internet Research Task Force), dependiendo estos dos últimos del IAB (Internet Architecture Board) o la Internet Society (ISOC)[31], y el W3C (Word Wide Web Consortium)[32], estos dos últimos de más reciente fundación, destacando W3C por su papel normalizador. Asimismo, la creación de grupos de investigación dentro de estas instituciones, también jugó un papel determinante en la expansión de Internet.

De forma paralela a estos avances, en la década de los noventa, se produce el traspaso de la gestión de estas redes a empresas privadas. Así, a finales de 1994 y principios de 1995 comienzan a aparecer compañías que facilitan la conexión a Internet como CompuServe, America Online y Prodigy, aspecto que incide en la explosión del fenómeno Internet y en el auge de la Web. Simultáneamente, y a medida que la información en este medio va aumentando, van a ir surgiendo los servicios de búsqueda, de los que más adelante nos ocuparemos.

---

[29]Este grupo, entre otros aspectos, se ocupa de elaborar una serie de informes conocidos como RFC (Request for Comment), que son considerados como recomendaciones sobre distintos aspectos de la red. Tras ser discutidos, se envían a IAB para ser estudiados y propuestos como norma.

[30]Se ocupa del desarrollo de la red con vistas al futuro.

[31]De ella forman parte más de 150 organizaciones y contiene más de 6.000 socios, principalmente empresas, agencias gubernamentales, instituciones, fundaciones, etc. Entre otros aspectos, se preocupa del desarrollo de la red tanto en los países avanzados como en los en vías de desarrollo. Más información sobre sus principios y actividades podemos obtener en http://www.isoc.org/isoc/mission/goals/index.shtml

[32]Consorcio creado en 1994, fruto de la colaboración del CERN, DARPA y la Comisión Europea, con representantes del MIT (Massachusetts Institute of Technology) por parte americana, del INRIA (Institut National de Recherche en Informatique et en Automatique) por parte europea y la Universidad Japonesa de Keio desde 1996. Se puede acceder a su servidor web en la dirección: http://www.w3.org. Un interesante resumen de sus actividades puede consultarse en la obra de Vianello Osti, M. (2004:159-161).

En España, los acontecimientos más importantes relacionados con Internet parten de la creación en 1971 de la red RETD (Red Especial de Transmisión de Datos) por la Compañía Telefónica Nacional de España (CTNE), que permitía la transmisión de datos por conmutación de paquetes, pasándose a denominar IBERPAC a partir de 1982.

En 1985 Fundesco (Fundación para el Desarrollo Social de las Comunicaciones) y la Secretaría de Estado de Universidades e Investigación crean el Proyecto IRIS (Interconexión de Recursos Informáticos), para permitir la comunicación entre las distintas universidades y centros de investigación españoles y posibilitar su conexión con redes internacionales. Desde entonces cuenta con el patrocinio del Plan Nacional de I+D.

En 1990 se realiza la primera conexión directa a la red Internet a través de la red troncal ARTIX, creada por RedIRIS, que en esta fase experimental unía Fundesco, el Departamento de Ingeniería Telemática de la Universidad Politécnica de Madrid, el Centro de Información Científica de Andalucía (CICA) y el CIEMAT.

En 1991 se pone en marcha un servicio que conectaba redes de área local IP con acceso a Internet, denominado SIDERAL (Servicio de Interconexión de Redes de Área Local), que facilita enormemente la conexión y el posterior éxito de la red en nuestro país.

A partir de 1994 la gestión de RedIRIS la lleva a cabo el C.S.I.C. (Centro Superior de Investigaciones Científicas).

La conexión internacional se hacía a través de la red europea IXI (International X25 Infraestructure), que posteriormente se denominó EUROPANET, gestionada por COSINE (Cooperation for Open System Interconection Networking in Europe).

RedIRIS, cuya gestión lleva a cabo el C.S.I.C. desde 1994, es socio de DANTE (Delivery of Advanced Network Technology to Europe), proveedor de servicios Internet a redes académicas europeas. Además es socio de TERENA[33] (Trans-European Research and Education Networking Association), participa en RIPE, foro europeo de proveedores de servicios Internet y colabora con el Centro Europeo de Coordinación Internet (RIPE-

[33]Antes RARE (Réseaux Associés pour la Recherche Européenne). TERENA unió en 1994 a RARE, EARN (European Academic and Research Network), el CERN y ECMWF.

NCC) y en la DG XIII de la Unión Europea dedicada a proyectos sobre servicios avanzados de red. Por último, es miembro internacional de Internet2[34].

La red se basa en la existencia de nodos en todas las comunidades autónomas de nuestro país, conectándose al exterior por medio de la Red TEN-34, red IP de alcance europeo que conecta a las redes académicas y de investigación europeas.

Con la llegada de la Web crecen los contenidos comerciales y el desarrollo de redes como INFOVIA, puesta en marcha por Telefónica en 1995 para permitir la conexión de usuarios particulares a Internet. Con la liberalización de las telecomunicaciones en España, Telefónica se vio obligada a modificar su red, dando lugar al establecimiento de INFOVIA PLUS, red IP que entre otras características de interés permite a las empresas establecer intranets virtuales y nuevos servicios basados en el uso de la voz. A raíz de la liberalización surgieron otras redes IP en nuestro país, como es el caso de la red de British Telecom.

Sobre el uso de Internet en España a finales del siglo pasado, Carballar (1998) da las siguientes cifras: de los 200 millones de documentos en Internet, el 70% está escrito en inglés y el resto en otras lenguas. Sólo el 1,8 % está en español. En 1997 existían en España un millón cuatrocientas mil personas conectadas a Internet. Al año siguiente la cifra aumentó a dos millones doscientas cincuenta mil. Pero estas cifras se han ido incrementando año tras año debido a que Internet se encuentra en pleno proceso de crecimiento.

Como consecuencia de todo esto hay que destacar la popularidad alcanzada y que se mantiene en la actualidad de aquí que el número de servidores y ordenadores conectados a la red se duplique cada poco tiempo, y los recursos electrónicos disponibles se

[34]Iniciativa que tiene su origen en Estados Unidos, en 1996, propiciada por un grupo de universidades y que trata de ofrecer una serie de servicios y aplicaciones avanzadas de red para centros de investigación y de enseñanza superior. En Europa existe una iniciativa similar que se conoce con el nombre de TEN-155 (Trans European Networking at 155 Mbps). Más información referente a este proyecto se puede consultar en el documento DANTE, The Next Generation of European Research Networking, rev. Dec.-98. disponible en http://www.dante.net/ten-155.html

multipliquen[35]. Información actualizada sobre este aspecto puede consultarse en la página web del Internet Domain Survey, proveedor de dominios de Internet[36].

Esta visión histórica puede acercarnos a comprender la realidad acerca de qué es y cómo es Internet en la actualidad: una red formada por variado tipo de espacios, servicios, documentos, de carácter científico y divulgativo, de desigual interés para el usuario y en constante crecimiento. Otro tema, al que nos referiremos más adelante, es la problemática que el fenómeno genera respecto a localización y acceso a estos recursos.

1.2. La World Wide Web

No vamos a detenernos a considerar la importancia que ha supuesto para el desarrollo de la red Internet la aparición de la Web, pero sí debemos destacar que el fenómeno ha influido en aspectos económicos, políticos, culturales y sociales teniendo una especial incidencia como vehículo de difusión de la información.

Al igual que ocurrió con los servidores WAIS, tanto los sistemas de información mencionados como las herramientas que permiten la búsqueda de información en cada uno de los espacios de Internet, han ido adoptando un nuevo modo de acceso a través de la Web, que es la predominante en la actualidad. Los sistemas de recuperación de la información de la Web también han tenido que adaptarse para facilitar el acceso a todo tipo de información existente bien utilizando la misma base de datos o creando otras específicas.

Arms (2001) define la Web como colección de información accesible a través de enlaces y alojada en un gran número de ordenadores conectados a Internet, denominados servidores web. De esta definición se desprende que sus contenidos son fundamentalmente informativos, aunque no debemos olvidar que los hay de otro tipo y que los documen-

[35]Existen servidores web que actualizan continuamente ambos aspectos. Las cifras pueden quedar anticuadas en muy poco tiempo por lo que es recomendable la consulta de dichos servidores para obtener visiones tanto históricas como actuales para ver su evolución hasta la fecha. Información útil en este sentido puede consultarse en: http://www.aimc.es, servidor de la Asociación para la Investigación de Medios de Comunicación y http://www.aui.es, donde la Asociación de Usuarios de Internet publica encuestas, estudios estadísticos sobre Internet tanto a nivel nacional como internacional y finalmente en http://www.ojd.es, web de la Oficina de Justificación de la Edición.

[36] http://www.isc.org/ds/WWW-200201/index.html

tos pueden contener texto, imágenes fijas o en movimiento y sonidos, formando todo ello una página web.

Desde el punto de vista técnico, podemos afirmar que constituye la World Wide Web el conjunto de ordenadores, llamados servidores web conectados a la red Internet, a los que se accede desde otras máquinas, mediante un programa cliente denominado navegador. El acceso implica la descarga en el ordenador propio del recurso solicitado. Los servidores pueden llevar a cabo una gran variedad de funciones, como por ejemplo soportar un determinado servicio de Internet ya sea de correo o de transmisión de ficheros, o bien albergar documentos electrónicos, ya sean de texto, imágenes, sonidos, así como programas informáticos u otro tipo de servicios, como el acceso a bases de datos, a catálogos de bibliotecas o a cualquiera de las herramientas de búsqueda.

En la configuración de estos recursos, hemos de destacar la utilización de lenguajes hipertextuales ya que posibilitan la existencia de términos o imágenes que actúan de enlace con otras partes del documento o con otros documentos[37].

Analizamos a continuación los orígenes de la Web así como algunos aspectos técnicos que facilitan su utilización, como son los servidores web, las direcciones URL, deteniéndonos también en alguna de las características de este espacio de información así como en los problemas que le afectan.

1.2.1. Orígenes de la World Wide Web

Simultáneamente a la intensa evolución en la informática que hemos señalado anteriormente, que afectó tanto al desarrollo de redes, del software y del hardware, y sobre todo a su aplicación en la disposición y recuperación de información, también se produjeron interesantes experiencias basadas en los lenguajes hipertextuales. Ello posibilitó que en 1990 Tim Berners-Lee y Robert Cailliau presentaran en el CERN (Centre Européene pour la Recherche Nucléaire)[38] un proyecto titulado World Wide Web: proposal for a hipertext project, que dará lugar a la creación de un navegador básico o visualizador y a la implantación del primer servidor Web. El proyecto se basaba en la utilización del

[37] Las etiquetas HTML que se utilizan siguen el esquema <A HREF="#('texto del enlace o dirección URL')" A NAME="texto del enlace o dirección URL" /A>.
[38] Actualmente recibe el nombre de European Laboratory for Particle Physics

lenguaje HTML (HiperText Markup Language) y el protocolo de transmisión HTTP (HiperText Transfer Protocol).

En 1993 la compañía NCSA (National Center for Supercomputing Applications) y la Universidad de Illinois presentan Mosaic, un navegador de visualización gráfica que soportaba plataformas con diversos sistemas operativos (MS Windows, Macintosh y estaciones de trabajo UNIX con X-Windows) que evolucionará tras la denominación Navigator, a Netscape, actualmente Mozilla. En 1995 aparece la primera versión de Explorer, de la compañía Microsoft. Desde entonces estos programas no han dejado de evolucionar y en sus últimas versiones ofrecen, además de la propia herramienta de navegación, otras utilidades como editores de páginas web, programas de acceso a News, al correo electrónico, al Chat, etcétera, permitiendo no sólo el acceso a servidores Web sino también conexiones Telnet, Gopher, etcétera.

En los últimos años ha proliferado el uso de programas gráficos, de sonido, video, multimedia y de comunicación en tiempo real. Así el popular Chat, término con que se expresa la técnica conocida como IRC (Internet Relay Chat), utilizado desde 1988, tiene su origen en lo que se conocía como Talk, que permitía la comunicación en tiempo real entre dos o más personas a través de mensajes escritos. Internet Phone es otra herramienta de este tipo, que facilita la comunicación mediante voz, siendo cada vez más utilizados otra serie de programas que facilitan el contacto audiovisual entre varias personas, permitiendo servicios como la videoconferencia. Las tecnologías relacionadas con la imagen y el sonido, y su difusión a través de la red, como la tecnología multicast, basada en el envío de imágenes y sonido a las máquinas desde las que se solicita, pueden ser las que obtengan un mayor desarrollo en los próximos años, ya que hasta la actualidad se han visto limitadas por el insuficiente ancho de banda que las soporta. Para su visualización se utilizan programas como NetMeeting de la compañía Microsoft, o bien otros de carácter comercial como CUseeMe, etcétera.

Por otro lado, hemos de señalar que la utilización del lenguaje HTML y la adopción de sistemas gráficos ha permitido ir dejando atrás la utilización de comandos para la consulta de bases de datos, siendo sustituidos por los denominados “botones” y otros elementos de carácter hipertextual.

Finalmente, otros aspectos técnicos que han influido de forma notoria en el desarrollo de la Web, fueron la evolución de los ordenadores personales, las redes LAN (Local Area Network) y el desarrollo del Sistema de Nombres de Dominio (DNS).

1.2.2. Características técnicas y problemas de la información en la Web

La forma de acceder a los recursos que la Web ofrece es variada, pero generalmente se realiza: por medio de la activación de enlaces a partir de una determinada página de inicio, técnica que forma parte del concepto de "navegación" ya que permite ir de una página a otra; por medio de los enlaces facilitados en los mensajes de correo o en información de las News, etcétera; mediante la introducción de una dirección URL (Uniform Resource Locutor) en la ventana prevista para ello que ofrecen los navegadores[39]; utilizando algunos de los portales especialmente diseñados para acceder a los servicios de Internet, por ejemplo las páginas de la mayoría de universidades así como las dependientes de empresas privadas como Terra, etcétera, o a través de búsquedas mediante sistemas de recuperación de recursos, ya sean directorios, agentes de búsqueda, o buscadores.

Al conectarnos a cualquiera de los recursos alojados en los diferentes servidores, en nuestro navegador aparece la dirección URL perteneciente al sitio, página o recurso web. Estas conexiones son posibles gracias a los servidores DNS (Domain Name System).

Los servidores DNS comenzaron a utilizarse en 1984 para facilitar la conexión entre ordenadores. Constan de una base de datos distribuida que contiene nombres de dominios de Internet que son traducidos a su correspondiente dirección IP[40], que es el número asignado a cada máquina conectada a la red. Aunque inicialmente a cada máquina o grupo de máquinas se le aplicaba un nombre de dominio, en la actualidad una máquina puede tener varios dominios y un grupo de máquinas contener una misma dirección, como es el caso de servicios amplios.

Las direcciones de Internet contienen el protocolo (ftp, telnet, gopher, news, http, etcétera), seguido de la dirección del ordenador al que nos queremos conectar, es

[39]Actualmente determinados navegadores como Netscape y otros, han incorporado a esta ventana sus propios motores de búsqueda, por lo que además de una dirección URL admiten la inserción de términos de búsqueda que recuperan los documentos o sitios web que los contienen.

[40]La expresión consta de cuatro números separados entre ellos por un punto, cuyos valores van de 0 a 255 (por ej.: 199.72.1.1). InterNIC fue la primera empresa que se ocupó de la distribución de la numeración tanto en el ámbito americano como en otros países, bien mediante el contacto con entidades responsables o a través de empresas multinacionales.

decir, el dominio del ordenador servidor. Este dominio está representado por una dirección IP a la que corresponde un determinado nombre de dominio.

Los DNS que se utilizan en las conexiones Web constan de varios niveles. De derecha a izquierda podemos observar que el primer nivel, no siempre presente, puede representar al país al que pertenece el carácter de la información del sitio web. El resto de niveles o subdominios contiene información sobre la institución que aloja las páginas web y el nombre de la máquina (Hostname) en que se alojan. Por ejemplo, en Estados Unidos, existen una serie de dominios genéricos de máximo nivel que indican el carácter de la información que contienen, bien sea comercial (.com), militar (.mil), educativo (.edu) o gubernamental (.gov) y otros más recientes relacionados con el ocio (.rec), industria aeroespacial (.aero), empresas (.biz), cooperativas (.coop), museos (.museum), etcétera. Algunos de estos se utilizan también en otros países, además de los dominios geográficos relacionados con su nombre, por ejemplo .uk para Reino Unido, .fr para Francia, .es para España[41]. Los recursos Web pueden ubicarse en diferentes directorios y subdirectorios del servidor, lo que influye en que estas direcciones Web sean más o menos amplias en función del nivel jerárquico de la carpeta en el que dicho recurso se encuentra. Estos niveles suelen separarse mediante el signo slash "/" y se sitúan a la derecha del primer nivel.

Estas direcciones siguen una organizan jerárquica. Una dirección genérica suele acabar con la indicación del dominio de máximo nivel, por ejemplo http://www.unizar.es/index.html pero una dirección más específica puede contener a continuación los nombres de los directorios y subdirectorios en los que se encuentra un documento. Siguiendo el esquema anteriormente señalado, la dirección de la página Web del Defensor Universitario de la Universidad de Zaragoza (http://www.unizar.es/defensor _universitario/), aparece en primer lugar el protocolo (http) seguido por un primer bloque hasta el primer signo de slahs que contiene de derecha a izquierda el nombre del dominio de máximo nivel (.es), al que precede el nombre del dominio de segundo nivel, relacionado con la organización (unizar). Le sigue al bloque el nombre del directorio en el que se encuentran las páginas de esta institución (/defensor_universitario/). Esta expresión

[41]Existe una orden ministerial, la número 6100 de 21 de marzo de 2000 que regula en España el sistema de asignación de nombres de dominio.

completa constituye la URL de la página principal de acceso a la información que ofrece este servicio.

Las direcciones URL contienen pues el protocolo de comunicación o transmisión, representado por el tipo de servidor (http, para documentos hipertextuales; ftp, para ficheros o programas, etcétera); en segundo lugar el nombre del protocolo Internet separado del anterior por dos puntos y dos barras (Ej.: ftp o www, aunque no siempre aparece); el nombre del servidor o del dominio, también conocido como host (Ej.: unizar.es) y el nombre del directorio y de la página (/defensor_universitario/intermediainforma.htm).

Además los navegadores, en servidores que contienen una página de inicio, interpretan que hay que abrir la misma página independientemente de que se exprese este término en su dirección. Así, se accederá al mismo recurso al teclear http://www.unizar.es que http://www.unizar.es/inicio. Dicho aspecto lo hemos tenido en cuenta a la hora de comparar distintos aspectos del rendimiento de los motores de búsqueda, como son el solapamiento de recursos, cuyo cálculo se basa en la comparación de direcciones URL idénticas.

Estas direcciones pueden estar formadas además por caracteres que para su codificación requieran una notación especial. Este problema aparece frecuentemente en los multibuscadores, que añaden o sustituyen signos de las URL de los recursos por otros. Por ejemplo es frecuente la sustitución del signo tilde "~", que se utiliza para señalar directorios personales, por su valor codificado %7E. Su finalidad es evitar conflictos de interpretación por los navegadores. En otras ocasiones añaden cadenas de términos en las que se expresa una acción, por ejemplo "search=" o bien el signo de interrogación "?" junto a los parámetros de búsqueda.

Uno de los principales problemas que plantean los nombres de dominio, del sitio y de la página es su falta de continuidad, es decir que se modifican cuando se cambia un sitio o página web a otro servidor, o cuando se cambia el nombre del servidor o los nombres de los directorios que contienen los sitios o páginas web, o simplemente cuando la página deja de existir. La no actualización automática de los enlaces que apuntan a estos recursos provoca problemas de acceso. Para solucionarlo se han propuesto un URL[42] per-

[42] Abreviatura de *Uniform Resource Locator* según las RFC (Request for Comments) 1738.

sistente, el PURL (Persistent Uniform Resource Locator) o el DOI (Digital Object Identifier) promovido inicialmente por una asociación americana de editores. También hay otras soluciones que están previstas en el software de mantenimiento de los sitios web, aunque unas y otras propuestas no terminan de imponerse.

Cuando el navegador no localiza una dirección, devuelve al usuario un mensaje de error, generalmente con un número. Así, el error 404 indica que la dirección es incorrecta, el recurso no está disponible, ha cambiado de ubicación o ya no existe. El error 403 indica que el recurso no es accesible de forma libre. Otras incidencias pueden ser debidas a que el recurso se encuentre demasiado solicitado en un determinado momento o que, por diversas circunstancias, se supere el tiempo de conexión.

En el ámbito internacional, el organismo que se ocupa de coordinar y gestionar los nombres y números de los diferentes dominios es el ICANN (Internet Corporation for Assigned Names and Numbers)[43]. De la solicitud de dominios de segundo nivel, por ejemplo (.com.es) se ocupa el DYNS (Dynamic Network Services)[44] o Red.es dominios[45] en el caso de España.

El acceso a los recursos generalmente suele ser libre y gratuito, siendo necesario, en algunas ocasiones, una simple inscripción. Otras veces el acceso es restringido y puede exigirse el pago de una cuota o suscripción como es el caso de los sitios Web que, por ejemplo, permiten el acceso a determinadas bases de datos y revistas electrónicas. También hay servidores de acceso mixto que permiten la consulta y utilización libre de una parte de sus recursos, limitando la consulta del resto. Estas limitaciones tienen también sus repercusiones en los motores de búsqueda cuyas arañas no pueden acceder a determinados recursos para indizarlos, por lo que son difíciles de localizar para el usuario, de aquí que formen parte de la denominada Web oculta.

Desde el punto de vista terminológico, los servidores Web contienen lo que se denomina “sedes web”, constituidas, además de por una URL, por un conjunto de páginas web. Pueden formar parte de un único servidor o de varios, aunque existen casos en

[43]http://www.icann.org
[44]http://dyndns.org
[45] Entidad pública empresarial delegada para la gestión del registro de nombres de dominio en Internet bajo el código “.es”.

que un mismo servidor aloja distintos sitios web. Codina (2000) las define como entidades digitales identificadas por una URL que contiene uno o más recursos.

Aguillo (1999) define la página web como fichero o conjunto de ficheros informáticos que constituyen un documento en lenguaje de maquetación hipertextual (Hypertext Markup Language o HTML), es decir hipertextual y multimedia, identificable a través de la red con un URL propio, una dirección en la Web. Una definición posterior de este autor (Aguillo 2000) matiza y aclara la anterior, al definirla como "unidad de visualización que produce un navegador WWW cuando interpreta una dirección URL de un documento HTML o similar con todos los ficheros asociados".

Por tanto, una página web está formada por un fichero electrónico y los que lo acompañan, bien en forma de imagen o sonido, y que además, y esto es más importante, es identificable a través de un URL propio, por lo que, cada documento con dirección propia es una página web.

Cada uno de estos componentes pueden ser documentos del tipo PDF (Portable Document Format), Office, o imágenes de tipo JPEG (Joint Photographic Experts Group), TIFF (Tagged Image File Format), PNG (Portable Network Graphics), o bien imágenes en movimiento en formato MPEG (Moving Picture Experts Group), MOV-Quicktime, AVI (Audio Video Interleave). Para sonidos se utilizan ficheros midi o WAV (Windows Wave), entre otros. Los elementos integrantes también pueden presentarse en formato comprimido, generalmente en formato ZIP. La existencia de esta variedad de documentos, se debe en parte a la fuerte presencia que en los últimos años están teniendo los medios audiovisuales en la Web.

Como hemos señalado, estas páginas se realizan generalmente mediante programas editores que utilizan un lenguaje de marcas denominado HTML (HiperText Markup Language), del que existen diferentes versiones, o mediante el lenguaje XML (Extensible Markup Language) más actual, derivados ambos de SGML (Standard Generalized Markup Language). Estas marcas permiten tanto formatear el documento y dotarlo de imágenes y sonidos, que son ficheros con extensiones propias, como crear enlaces a diferentes partes del documento o a otros documentos. Para ello utilizan los elementos propios como anclas y enlaces. Proporcionan además al documento una estructura que es aprove-

chada por determinados sistemas de búsqueda de información para elaborar sus índices, extrayendo información de determinadas partes como el título, el cuerpo o las etiquetas Meta[46].

El lenguaje HTML se caracteriza por su fácil aplicación, lo que explica, en parte, el gran éxito de la Web en cuanto a producción de información. Sin embargo, como aspectos negativos, hay que señalar que no facilita la descripción de los documentos, pues un gran número de páginas se hallan desprovistas de algo tan básico para su localización como es el título, el autor o la filiación, por no mencionar uno de los aspectos a los que a menudo se refieren los investigadores como posible solución a muchos de los problemas que plantean los motores en la recuperación de información: el mínimo uso de los metadatos. Esta situación ha dado lugar al surgimiento de iniciativas que, aprovechando el lenguaje de marcas, permiten una descripción cuya finalidad es obtener unos resultados más precisos en las búsquedas de información[47].

El problema puede ser solucionado con la adopción de un nuevo modelo de representación de la información, el RDF (Resource Description Framework), que admite metadatos de recursos en XML y de otros lenguajes, así como de recursos que contienen identificadores URI (Uniform Resource Identifiers). No sólo permite realizar una descripción de la página web, de una parte de ella o de un recurso que la integre, sino que también posibilita vincular conceptos, lo que va a permitir que las herramientas de búsqueda preparadas para ello, recuperen información más precisa al facilitar el cruce de datos y términos de diferentes índices[48].

Los creadores de páginas web son conscientes de las limitaciones del lenguaje HTML en cuanto a las posibilidades de ampliación, de mejora en el diseño de las páginas y de estructuración del propio documento y de la información. El lenguaje XML da mayor importancia a las partes constitutivas del documento y a sus datos, lo que permite

---

[46] Para más información sobre la aplicación de este lenguaje en herramientas de búsqueda puede consultarse el trabajo de Hu y otros (2001b).

[47] En el ámbito de la Unión Europea hemos de destacar el proyecto Renardus basado en la utilización de Metadatos que ha dado lugar a la existencia del buscador del mismo nombre que permite recuperar recursos de interés académico, principalmente en lengua inglesa. Utiliza diversos directorios de calidad y organiza los recursos del directorio de acuerdo con la clasificación de Dewey. Más información puede obtenerse en <http://www.renardus.org>

[48] Para más información véase Manola, F. y Millar, E. (2003)

superar los inconvenientes del HTML, añadiendo ventajas como facilitar el desarrollo del comercio electrónico y de la recuperación e intercambio de información. En este sentido, hay que mencionar la función de indización automática de la información contenida en sus etiquetas, que le permite reconocer el tipo de información que contiene, es decir, si se trata de un título, autor, palabra clave, etcétera. Dado que XML está basado en un estándar abierto, no hay problemas de incompatibilidad entre sistemas operativos. Finalmente, otro lenguaje que se va imponiendo tras serle otorgado por el W3C 1999 el estatus de "Recomendación" es XHTML que se caracteriza por sintetizar las ventajas de los dos anteriores.

Al margen de estos aspectos que tratan de desarrollar y mejorar la Web, es fácil comprender el éxito de un medio que pone al alcance del usuario instrumentos de fácil e inmediato acceso, que además permiten de forma sencilla la realización de todo tipo de documentos, simples y complejos, y que además integra otros servicios de Internet como el correo electrónico, consulta de noticias o news, descargas de programas y archivos, etcétera.

La integración de diferentes servicios de Internet con sus respectivos servidores se produce gracias a una serie de pasarelas que utilizan un software específico como es el caso de CGI (Common Gateway Interface), ASP (Active Server Pages), y PHP (Hipertext Preprocessor) que añaden estas mismas terminaciones a sus páginas. La tecnología plug-in que, mediante pequeños programas, facilita la visualización de gran número de elementos multimedia. Hay que señalar también por su importancia los programas que se realizan con Java y VRML (Virtual Reality Modeling Language) especializado en recreaciones virtuales.

Es de destacar la facilidad de uso tanto de las aplicaciones necesarias para los diferentes tipos de conexiones de Internet, como de los programas navegadores, que facilitan el uso del resto de servicios de la red. En este sentido hemos de referirnos a proyectos, como el Proyecto Oxígeno, que se lleva a cabo en el MIT y que trata de desarrollar tecnologías que faciliten la navegación. Este aspecto, junto con el desarrollo de la Web semántica, a la que más adelante nos referiremos, y la tendencia hacia la integración no sólo de diferentes utilidades de Internet, sino de un número menor de herramientas con mayores capacidades y con la posibilidad de concentrar diferentes funciones, son los aspectos que marcan el desarrollo de los últimos años de la Web.

Su expansión es tal que ya se puede conectar a ella a través de otras gamas de aparatos distintas al ordenador personal, como son la televisión, el teléfono móvil o las agendas electrónicas.

Por tanto, es un hecho que la Web está en continua evolución, lo que le confiere un gran dinamismo que influye, como no podía ser de otro modo, en todo su contenido.

El éxito y acogida de la Web ha sido tal que está siendo cada vez más utilizada por servicios de búsqueda y distribuidores de bases de datos[49] para ofrecer sus servicios mediante éste sistema gráfico que facilita la interactividad. Lo mismo ocurre con las bibliotecas, ya que ofrecen no sólo información y diferentes servicios como el acceso a sus catálogos y bases de datos a través de servidores web, sino la consulta de documentos de todo tipo en formato digital. Así, para Rodríguez (2002:178) la Web es "un servicio de transferencia de información en línea que permite distribuir texto, imagen y sonido, posibilita la comunicación, facilita la realización de transacciones y entretenimiento…"

La mayor parte de estudios que se refieren a la Web aluden a su carácter informativo. Pero ¿cómo es la información que contiene? Para contestar a esta pregunta, analizamos a continuación cómo esta dispuesta la información en la Web y las características más destacadas de las páginas web.

Uno de los aspectos fundamentales de la información en la Web, y a la que tal vez en parte debe su éxito, es el carácter distribuido de la información, con recursos de acceso totalmente libre y gratuito aunque, como hemos visto, existen otros de acceso restringido, e incluso de pago. Podemos citar en uno y otro sentido el acceso a bases de datos, a catálogos de bibliotecas y a revistas y libros electrónicos, así como a otro tipo de documentos de carácter científico como son: publicaciones de congresos, tesis doctorales y trabajos de investigación relacionados con todas las materias.

Otro de los aspectos destacables de la información en la Web, al que constantemente se hace referencia, es el gran número de documentos existente y en constante cre-

[49]Muchas son las bases de datos, en su mayor parte de acceso restringido, que se ofrecen a través de Internet. Para el campo de la Documentación hemos de señalar la incorporación de la base de datos LISA del Cambridge Scientific Abstracts, accesible a través de Internet Database Service en la dirección http://www.csa1.co.uk

cimiento[50]. Ello es debido a que en este medio, crear y dar a conocer la información es tan sencillo, que cada vez es mayor el número de personas que publican documentos, no siempre de buena calidad tanto en la forma como en el contenido. Por otro lado, existen grandes intereses comerciales cada vez más patentes, lo que influye en la existencia de un gran cúmulo de páginas de este tipo. No obstante hay que hablar también de la existencia de una parte importante de documentos de alto nivel científico. Un estudio de Jiménez Piano (2000), aporta datos en este sentido, al señalar que el 83% de la información de la Web es de carácter comercial, frente a un 6% de carácter científico. Un trabajo posterior de Lossau, (2004), estima en veintidós billones, el número de páginas de contenido científico existentes en la Web, incluyendo a las existentes en la Web Invisible.

Como vemos, los servidores web pueden contener información y obedecer a fines de lo más variado. Así, podemos observar la convivencia de servidores como los dispuestos por las universidades y centros de investigación, cuyo fin principal puede ser dar a conocer información relacionada con la propia institución, y con la producción científica propia, con un marcado carácter de difusión de información y otros pertenecientes al ámbito empresarial, con un enfoque más comercial. Otros se centran más en ofrecer una serie de servicios, ya sea de localización y acceso a información específica, por ejemplo sobre viajes, economía, etcétera, o a determinado software, material audiovisual, ocio, etcétera. En este contexto, resulta llamativo el contraste entre la existencia de una amplia variedad de recursos, que van desde los grandes y sofisticados servicios web, a los servidores con simples páginas de texto colocadas por usuarios a nivel particular. No obstante, un estudio de Lawrence y Giles (1999) menciona el alto valor científico de los contenidos de la web, lo que unido al relativamente bajo número de servidores con este tipo de información, podría hacer viable la indización de la información de carácter científico.

Por otro lado, cada vez es más frecuente, que tanto los proveedores de las bases de datos en línea como las grandes firmas distribuidoras de publicaciones periódicas a texto completo, utilicen la Web para acceder, mediante suscripción, a sus servicios. Con

[50]Delgado Martínez (2001) estimaba el número de páginas web en 1.200 millones. Otro estudio (Bergman, 2001) señala la existencia de un billón de documentos en la Web visible y 550 billones en 200.000 sitios web de la Web oculta. Estimaciones más recientes para la web visible, recogidas por D. Sullivan (2005), señalan una cifra superior a 11,5 billones de páginas.

ello podemos observar que se está polarizando cada vez más en este medio el acceso a un importante número de recursos de investigación.

Pero lo que nos interesa es analizar el acceso a este tipo de documentación. En este sentido, cada vez son más numerosas las iniciativas y servicios que tratan de ayudarnos a recuperar recursos de este tipo. Por ejemplo para la localización de revistas en línea nos podemos conectar a InfoJump[51] y para buscar artículos es interesante Findarticles[52] o Dialnet[53] para el ámbito hispano, ambas gratuitas. Servicios similares de pago ofrece Northern Light[54]. BUBL[55] ofrece recursos de información al igual que los directorios SOSIG[56] y WWW Virtual Library[57]. De especial interés es el Proyecto CiteSeer.IST para crear una biblioteca digital de literatura científica en el que participan la National Science Foundation y Microsoft Research[58].

Existen otras iniciativas con base territorial, como es el caso de los denominados Portales académicos (Scholars Portal) que para el ámbito americano promueve ARL (American Research Libraries)[59] o el proyecto RENARDUS para Europa[60], por no citar otras de carácter local como las promovidas en Alemania por el Deutsche Forschungsgemeinschaft.

Otro recurso son las bibliotecas virtuales que contienen recursos compilados por especialistas como es el caso de Argus Clearing House[61], Bubl Link[62], INFOMINE[63], Librarian Index to the Internet[64] y eLIB Subject-based Information Gateways/RDN (Resource Discovery Network)[65].

---

[51]Accesible en http://www.infojump.com
[52]http://findarticles.com/PI/index.jhtml
[53] http://www.dialnet.unirioja.es
[54]http://www.Northernlight.com
[55]http://www.bubl.ac.uk
[56]http://www.sosig.ac.uk
[57]http://www.vlib.org
[58]Esta herramienta constituye una excelente ayuda ya que además de facilitar búsquedas bibliográficas y obtener información sobre sus citas, permite acceder a los documentos a texto completo en varios formatos. Más información sobre este proyecto en http://citeseer.ist.psu.edu/cs
[59]http://www.arl.org/access/scholarsportal/
[60] http://www.renardus.org
[61]http://www.argusclearinghouse.net
[62]http://www.bubl.ac.uk/link
[63]http://infomine.ucr.edu/Main.html
[64]http://lii.org
[65]http://www.rdn.ac.uk

Otras características de la información que estos servidores contienen tiene que ver con la actualización, con su carácter efímero, con la facilidad de duplicación, con el carácter interactivo y con la variedad de formatos.

La actualización varía sensiblemente de unos servidores a otros dado que unos actualizan su información diariamente mientras que otros apenas lo hacen.

Su carácter efímero, está relacionado con el carácter instantáneo y con el tipo de información que se edita en este medio ya que posiblemente, en otro soporte no merecería la pena difundirla.

Hemos de señalar que la Web es un medio en el que está permitido disponer la información en varios servidores además de su duplicación, constituyendo los llamados mirrors o espejos[66] dando lugar a recursos que aparecen alojados en distintos sitios web, lo que influye decisivamente en la recuperación de un importante número de duplicados por parte de los motores de búsqueda.

Respecto al tiempo al que una información es accesible, hay estudios que tratan de calcular la permanencia media de las páginas web en los servidores, estimándose entre los 44 días que señala Michael Lesk (1999) y los 70 días que mantiene Brewster Kahle (1998), mientras que un estudio de Koechler (1999) demuestra que la vida media de los sitios web supera por poco los dos años.

Debemos indicar, además, el carácter interactivo de algunos páginas y servicios, que permite actuar sobre ellos para obtener una visión o acción determinada.

Respecto a la variedad en que pueden aparecer los archivos, éstos puede ser SIMP para los documentos de correo, documentos FTP, otros pertenecientes a las news, y finalmente los documentos HTTP más propios del Web. Todos ellos son producto de los diferentes servicios que se prestan. No debemos olvidar, además que servidores del tipo FTP, HTML pueden contener a su vez, documentos electrónicos de una amplia variedad de formatos.

[66]Se denominan así a los servidores que para evitar problemas por exceso de conexiones en el servidor original, contienen información duplicada alojada en otros servidores.

A esta diversidad hay que unir la variada tipología documental de los archivos Web, ya que comprende tanto documentos como weblogs, revistas, enciclopedias y libros electrónicos, etcétera.

Otra característica que no debemos olvidar es la facilidad con que se modifican los documentos electrónicos. Se trata de una información a la que resulta fácil añadir cambios tanto de contenido (dinamismo) como de ubicación (volatilidad), ya que pueden trasladarse a otro lugar del servidor o a otros servidores. Esta movilidad afecta especialmente en un sistema como la Web, basado en el hiperenlace, en el que los documentos se enlazan unos a otros por medio de su URL[67] que debe modificarse al cambiar de ubicación un documento en el servidor o al cambiarlo de nombre, de lo contrario el navegador no lo encuentra y produce un error. Esto influye de forma muy negativa en el acceso a la información ya que los datos del nuevo URL no se actualizan en la dirección de destino de los enlaces que apuntan a dicho recurso. En otros casos se mantiene la información en servidores diferentes, lo que explica la aparición de documentos duplicados con diferente URL. Estos aspectos inciden también en la recuperación de la información en la Web mediante el uso de buscadores, ya que si no se actualizan frecuentemente los índices, puede dar lugar a la recuperación de un importante número de registros a los que o no se puede acceder o el contenido aparece duplicado. En este último caso, cuando se actualiza la página no siempre se hace en la duplicada, dando lugar a páginas similares pero diferentes en función de que los cambios hayan afectado sólo a una de ellas.

En este sentido, Koechler (1999) sostiene en su estudio que en un periodo de seis meses cambia el 97% de los sitios y el 98,3% de las páginas web. De aquí que no resulte difícil intentar acceder a recursos que han cambiado de dirección URL y por tanto no son accesibles desde los enlaces iniciales.

Pero tal vez una de sus características más destacada sea el carácter hipertextual. Esto facilita poder navegar a través de sus enlaces. Esta característica no sólo afecta a páginas web sino también a la disposición en que determinados servicios organizan la información. Este aspecto va a influir considerablemente en la creación tanto de bases de datos, que contienen texto con hiperenlaces que apuntan directamente al recurso, de Direc-

[67]Hasta la fecha constituye el identificador más preciso de los recursos en la Web, si bien se espera que otros identificadores más precisos como el URN (Uniform Resource Name) acaben imponiéndose.

torios y bibliotecas digitales que presentan los recursos ordenados generalmente por materias.

Otras características de la información en la red son su instantaneidad, en cuanto a que cualquier información puede estar disponible, y por tanto ser utilizada de forma rápida; la universalidad, que juega un importante papel en el carácter variado de la información, y el estar sometida a actualización permanente.

También hemos de referirnos a su accesibilidad, pudiendo ser consultada, por regla general, a cualquier hora, en cualquier momento y en cualquier lugar.

Olvera Lobo (1999c) señala además otras características como la dispersión de la información y su carácter multilingüe.

Ante esta situación, podemos hacernos una idea de las dificultades que plantea localizar la información en la red, ya que en primer lugar supone cierto esfuerzo crear programas y bases de datos que permitan un almacenamiento y recuperación eficaz de tanta información, al tiempo que su variedad y creciente número, hace obligado el manejo de diferentes herramientas de búsqueda. Su constante evolución, dificulta además que el usuario especializado pueda llegar, no a dominarlas todas, sino simplemente a conocer alguna de las más adecuadas.

La Web realiza un importante papel como medio de difusión de la información, pero como veremos más adelante, algunas de las características señaladas junto con otros factores, pueden influir negativamente en su recuperación y acceso, especialmente cuando se utilizan para ello los motores de búsqueda generales. Estas herramientas son las encargadas de facilitarnos la recuperación y acceso a los recursos de la Web, de aquí la necesidad de conocer de qué modo lo hacen y si es mejorable.

Como conclusión podemos decir que la Web, por sus características es un sistema complejo, al igual que la información que contiene, que requiere grandes esfuerzos en los sistemas para ofrecer no sólo un correcto funcionamiento sino también para facilitar al usuario la búsqueda y recuperación de recursos de calidad.

2. Los buscadores de información de la WWW en el contexto de los sistemas de recuperación de la información. Procesos, funciones y problemas

2.1. Los Sistemas de Recuperación de la Información y los buscadores Web

La Recuperación de la Información como disciplina, con los Sistemas de Recuperación de la Información[68], constituyen el contexto y punto de partida de las herramientas de recuperación Web.

Abadal y Codina (2005:29) señalan que la Recuperación de Información "es la disciplina que estudia la representación, la organización y el acceso a la información..." Tiene por objetivo solucionar los problemas de información que requieren información cognitiva. Los tres rasgos que la caracterizan son: el uso de ordenadores, el uso de la información textual y el contexto de descubrimiento, relacionado este último con la necesidad del usuario de descubrir qué entidades cumplen una o más condiciones.

Respecto a los principios teóricos, estos autores señalan la existencia de dos corrientes de estudio, una de tipo algorítmico, de orientación informática y centrada en los programas o algoritmos, y otra cognitiva, centrada en aspectos propios o cercanos a las ciencias cognitivas: lenguaje, semántica documental, psicología, etcétera. Esta última sería la más cercana a los profesionales de la documentación.

La información, para su posterior recuperación, suele almacenarse en bases de datos que llevan consigo un importante número de programas informáticos que se ocupan de su gestión y mantenimiento. Este conjunto de programas forman los denominados Sistemas Generales de Bases de Datos. Dichos sistemas pueden estar basados en modelos de datos relacionales, dando lugar a Sistemas de Gestión de Bases de Datos Relacionales, cuyos datos físicos, como indica Chowdury (1999), tienen un alto grado de independencia y disponen de poderosos lenguajes para facilitar y mejorar la recuperación. Se caracterizan por contener la información altamente estructurada y constan además de una serie de programas que no sólo intervienen en la creación de las partes del sistema, sino que participan en su interacción.

[68]Existen diversas obras de gran interés que tratan de la Recuperación de la Información como la de Inwersen (1992) o la de Baeza-Yates y Riviero-Neto (1999).

Por otro lado, como Abadal y Codina indican, están los Sistemas de Gestión Documental (Information Retrieval System) dentro de los cuales se encuentran, por un lado los Sistemas Gestores de Bases de Datos Documentales, a los que caracteriza la formación de diccionarios de datos y la gestión de referencias, y por otro los Sistemas de Indización o motores de búsqueda que se ocupan de la indización y la gestión de bases de datos a texto completo.

Con la llegada de Internet, éstos últimos, han alcanzado un gran desarrollo ya que proporcionan, junto a la referencia, el texto del documento en formato digital.

Abadal y Codina (2005:43) señalan que los Sistemas de Recuperación de la Información (en adelante SRI), "pueden consistir en programas informáticos o conjuntos de rutinas aislados o bien pueden estar integrados en el seno de un sistema de gestión de bases de datos documentales..."

Dreilinger y Howe (1997) definen los SRI como herramientas de software que ayudan a los usuarios a encontrar los documentos contenidos en una base de datos o corpus documental determinado.

Es interesante la clasificación de tipos básicos de SRI que proponen Abadal y Codina (2005:37) en la que tienen en cuenta los componentes intelectual y automático de estos sistemas. En este sentido, señalan que los motores de búsqueda se caracterizan por requerir para su uso, un proceso intelectual mínimo pero unos procesos automáticos intensivos.

Para Furner (1996) la principal función del sistema es asistir o servir de soporte al usuario en sus tareas. Dichas tareas son definidas como cualquier actividad o acción que requiere por parte de las personas, la manipulación de objetos o conceptos, para alcanzar una meta u objetivo, que supone obtener un resultado específico. Estos sistemas deben de proporcionar asistencia al usuario de forma fácil, eficiente y efectiva, y así permitir alcanzar con éxito su objetivo, que no es otro que obtener información de interés. Evidentemente esto implica un correcto funcionamiento de las partes que los componen, aspecto que hemos de tener en cuenta a la hora de evaluarlos.

De su estructura se ha ocupado Purificación Moscoso (2002:521) quien señala que en los sistemas de información documental "la información se estructura en una base de datos que consiste en un conjunto de datos almacenados en soporte informático y organizados de tal forma que puedan recuperarse de determinadas maneras, de acuerdo con las necesidades expresadas en la estrategia de búsqueda." En la organización de la base

de datos juegan un papel fundamental tanto el "diccionario de datos" que contiene toda la información relativa a los campos de la base de datos, como el índice invertido, formado por los términos que aparecen en los diferentes campos y por la información del documento que los contiene.

El segundo componente básico de estos sistemas lo constituyen los programas de gestión y recuperación, en los que radican una serie de funciones fundamentales en la recuperación tales como: soportar la lógica booleana, a través de las expresiones AND, OR y NOT; operadores de proximidad como NEAR y ADJ; de presencia o ausencia, mediante los signos modificadores (+) y (-); truncamientos mediante el signo asterisco (*) o interrogación (?); permitir la búsqueda por campos; redefinir una búsqueda y finalmente, facilitar una serie de técnicas como la ordenación o ranking, la equiparación exacta o Best match, el uso de modelos de recuperación como los probabilísticos o vectoriales así como técnicas de inteligencia artificial aplicadas a la recuperación de información.

Como tercer componente Moscoso señala el software de interfaz "que es el que condiciona y determina la comunicación entre usuarios y el sistema".

Lancaster (1979) y Kent (1971), de una forma más específica señalan las siguientes funciones de los SRI:

- Identificar la información relevante para el usuario.
- Analizar su contenido.
- Representar el contenido de forma que sea accesible mediante las consultas del usuario.
- Analizar las consultas del usuario y en su caso transformarlas para consultar de forma eficiente la base de datos.
- Permitir la búsqueda.
- Recuperar la información relevante.
- De ser necesario, permitir ajustes en la búsqueda para lanzarla de nuevo.

Estas funciones las llevan a cabo una serie de subsistemas que Lancaster denomina subsistema documental, de indización, de vocabulario, de búsqueda, de usuario o interface y subsistema de comparación.

Chowdury (1999:2), por su parte, señala tres subsistemas principales en los SRI: el subsistema documental, el de los usuarios y el de búsqueda/recuperación. Estos subsistemas se ocupan del análisis de documentos y organización de la información, es decir, de la creación de la base de datos; de analizar las preguntas de los usuarios, de la búsque-

da en la base de datos y de la recuperación. Para este autor, su objetivo es recopilar y organizar la información de uno o más temas para facilitarla al usuario, tan pronto como la solicite.

Schwartz (1998) señala que los usuarios de estos sistemas raramente entienden o tienen en cuenta los mecanismos, y más raramente aún, hacen completo uso de las capacidades que facilitan las herramientas de búsqueda. Esto es así a pesar de que una buena recuperación de información depende tanto de la pregunta como del buen manejo de los sistemas.

Desde el punto de vista histórico, debemos señalar que en la RI tienen lugar procedimientos tan tradicionales como el sistema de tarjetas perforadas, pero es a partir de los años 50, con la aparición de los SRI Automatizados[69] (SRI) y más concretamente con la generalización de los procesadores de texto a partir de los 70, cuando alcanzan una mayor importancia, ya que estos nuevos sistemas van a ser utilizados tanto en bibliotecas, para la descripción, clasificación y búsqueda de documentos; en los archivos; en las bases de datos en línea y en los CD-ROM. En esta década sitúa Moya (2002) el punto de partida de la Moderna Recuperación de la Información, que adquiere un gran auge en los años 80 con el desarrollo de técnicas avanzadas como la ordenación por relevancia y la indización a texto completo. Para Salton, la Moderna Recuperación de la Información arranca de 1955 en una primera etapa, marcada por el uso de bases de datos bibliográficas y referenciales, que dura hasta 1975, fecha a partir de la que se ponen en práctica tanto técnicas de recuperación desarrolladas en el periodo anterior, como son los modelos basados en el espacio vectorial[70] y probabilístico, como nuevos lenguajes de búsqueda que facilitarán el acceso directo al recurso buscado.

Dichas técnicas y modelos se irán aplicando a los SRI accesibles a través de Internet adaptándose en unos casos y desarrollándolos en otros.

[69]Cleverdon y Mills (1963) definen a los Sistemas de Recuperación de la Información como una organización completa para obtener, almacenar y facilitar información. Les caracteriza además la existencia de personal capaz de evaluar la información antes de ofrecerla al solicitante así como la existencia de un índice de materias para facilitar la búsqueda.

[70] Utilizado inicialmente por Salton en la evaluación del Sistema Smart.

Posteriormente, con la aparición de la Web, se han ido probando continuamente nuevos sistemas que han dado lugar a la existencia de un numeroso y variado número de herramientas que tratan de hacer posible la RI.

De este modo, los antecedentes más directos de las actuales herramientas de búsqueda web son los buscadores que se venían utilizando en Internet, a los que se les han ido aplicando las técnicas resultantes de la investigación en el campo de los SRI, cuyo desarrollo y mejora, a lo largo de los últimos años, ha sido constante. Así, las herramientas web han heredado de aquellos sistemas sus capacidades de almacenamiento y recuperación de información y han incorporado de la investigación aquellas técnicas que facilitan un mejor desarrollo de las funciones para las que han sido diseñadas. Podemos citar en este sentido técnicas de recuperación como la equiparación (best-match), la Retroalimentación por relevancia (Relevance Feedback), que permite reformular las búsquedas y la Clasificación automática (Data Clustering) con el fin de mejorar la recuperación. A pesar de todo, y tal vez debido a que nos encontramos en el momento inicial de su puesta en marcha, estas herramientas, presentan en la recuperación una serie de problemas, a los que deben hacer frente y solucionar.

Abadal (2001) señala que el acceso y recuperación de la información en la Web "descansa sobre sistemas informáticos, básicamente sistemas de gestión de bases de datos, complementados con sistemas de navegación hipertextuales". Esta frase sintetiza de forma clara cuál es el nivel de desarrollo en el que se encuentran actualmente estos sistemas.

De acuerdo con el modo de navegación por la red, este autor clasifica las herramientas de acceso y recuperación de la información de la siguiente forma: Sistemas de navegación de tipo jerárquico, sistemas de navegación hipertextuales y SRI. De estos sistemas nos interesan tanto los sistemas de clasificación de tipo jerárquico, dado que en ellos la información se estructura jerárquicamente en clases y subclases, como es el caso de los Directorios, como los SRI, mediante los cuales se interroga a la base de datos por medio de un motor de búsqueda en el que se insertan los términos sobre los que se quiere obtener información, bien de forma aislada o mediante el uso de operadores y otras opciones de búsqueda.

Esta variedad de sistemas da lugar a que la búsqueda de información en la Web se realice principalmente por medio de dos técnicas: la selección de hiperenlaces y la

interrogación de las bases de datos[71]. Algunos de estos sistemas, como AltaVista, Yahoo, etcétera, integran los dos modos de búsqueda.

La consulta de las bases de datos de los buscadores se realiza, como en el resto de sistemas, de varias formas: bien insertando el término o términos en la ventana de búsqueda, ya sea de forma aislada o mediante operadores; realizando una búsqueda por frase; utilizando el lenguaje natural o con la ayuda de un tesauro. El motor lanza la consulta sobre el índice y presenta una página de resultados generalmente ordenados por relevancia.

No obstante, hasta presentar los resultados, estos sistemas llevan a cabo con la información que extraen o que se les proporciona, como en el caso de los metabuscadores, una serie de procesos, cuyo objetivo es facilitar, de forma rápida, la necesidad de información planteada por el usuario.

El uso de estas técnicas, unido a otras características de estos sistemas, como son la interactividad, deberían dar lugar a la formación de instrumentos válidos para recuperar información relevante, pero el escaso tratamiento documental de la documentación existente en la Web, y otras características que presenta la información en este medio, dificultan su efectividad. Por tanto debemos pensar en los procedimientos relacionados con la descripción como una de las soluciones para la mejora del funcionamiento de estas herramientas. En este sentido hemos de señalar el importante papel de los sistemas de metadatos y las iniciativas de un gran número de bibliotecas, bases de datos y organismos como OCLC (Online Computer Library Center) que incorporan registros de recursos electrónicos a sus catálogos, así como otras iniciativas en el seno de la Web que constituyen auténticas bases de datos para recuperar recursos web.

2.2. Procesos de los SRI y su repercusión en las herramientas de recuperación Web

De todos es conocido que las bases de datos contienen registros formados por los datos que almacenan. Estos datos son indizados, dando lugar a índices de palabras

[71] Baeza-Yates y Riviero-Neto (1999) añaden además una tercera técnica basada en la simple activación de enlaces, para ir de un documento a otro

clave o de frases. En las bases de datos textuales o referenciales, sobre éstos índices se lanza la búsqueda, con la intención de localizar los documentos o referencias que contienen o se relacionan con los términos solicitados, siendo presentados en una pantalla de resultados. De aquí que las funciones más importantes sean la formación de la base de datos, la indización, la búsqueda y recuperación. Estas funciones se realizan de un modo relacionado, de ahí su carácter de sistema. Su adecuado funcionamiento les ha de permitir alcanzar el fin para el que han sido diseñados: recuperar la información precisa. De aquí que la recuperación de información haya adquirido una gran importancia ya que se ocupa tanto de la representación, de la organización y del acceso a la información. (Abadal y Codina, 2005:29).

A continuación nos referimos, de forma breve, a cada una de las funciones que se desarrollan en estos sistemas, para ponerlos en relación con los SRI de la Web.

2.2.1. Formación de la base de datos

La formación de las bases de datos se lleva a cabo de acuerdo con una serie de principios como puede ser el carácter científico de la información que contienen, así como los propios de las empresas gestoras y creadoras de dichas bases de datos, que deciden dedicarse a compilar información de uno o más temas. La cobertura tiene que ver con su contenido y ha de posibilitar la obtención de un conjunto, lo más completo posible, de referencias, y en su caso documentos, sobre el tema de búsqueda. Como indica Abad (2004:98) puede verse influenciada por otros aspectos como: "el periodo de tiempo sobre el que el sistema puede proporcionar información retrospectivamente (cobertura temporal), la procedencia geográfica (cobertura geográfica), los idiomas en los que están escritos los documentos (cobertura idiomática) y el tipo de documento (libro, informe, carta) ..."

En la evaluación de bases de datos, indicadores para valorar la cobertura son el alcance y continuidad de los contenidos, la exhaustividad de la cobertura, la exclusividad de la cobertura o solapamiento y la puntualidad en la actualización del sistema de información (Abad, 2004).

La formación de la base de datos es uno de los aspectos más importantes en las herramientas Web. Los motores de búsqueda básicamente contienen la información que han ido obteniendo las arañas de los servidores Web, desplazándose, a través de enlaces, a las páginas de interés. Otra parte es fruto de los envíos de información sobre determinadas páginas por parte de sus creadores o responsables. El número de consultas es tan

amplio que hace necesario la existencia de distintas versiones de sus bases de datos, a las que denominan espejos o mirrors, que al alojarse en diferentes dominios pueden presentar variantes en función del país al que pertenece dicho dominio.

En la Web se utilizan diferentes métodos para calcular la cobertura de los buscadores. Bharat y Broder (1998) mencionan el Melee's Indexing Coverage Analysis (MICA), basado en el recuento de páginas web por dominio y, por otro lado, el utilizado en el sitio web especializado en analizar motores de búsqueda conocido como Search Engine Watch[72] que realiza esta función a través del motor Seach Engine EKG y proponen otro método basado en análisis estadístico, llevándolo a cabo mediante el lanzamiento de dos tandas de diez mil consultas cada una. Según estos autores, este método no proporciona valores absolutos sino estimados y a partir de ellos se llega a valorar el número de documentos existentes en la Web.

2.2.2. El análisis documental

Como señalan Peña y otros (2002:212) el análisis documental consiste en:

La selección, almacenamiento y organización de ideas informativamente relevantes para facilitar su localización posterior. Incluye técnicas tradicionales de la documentación como catalogación, indización, clasificación y resumen; y no tradicionales como indización automática, extracto automático y evaluación de concordancia (*matching*, en terminología inglesa).

Estas técnicas se agrupan en lo que se denomina análisis formal y de contenido, formando parte del primero la catalogación o descripción bibliográfica y del segundo la indización, clasificación y resumen.

2.2.2.1. La descripción

La descripción de un recurso o documento trata de representar, mediante un registro, dicho recurso, que al formar parte de una base de datos, permite su recuperación e identificación.

[72] http://www.searchenginewatch.com

En los sistemas de recuperación modernos, la estructuración de la información en campos facilita considerablemente la recuperación. Sin embargo, en el ámbito de los motores de búsqueda de la Web, el análisis formal es prácticamente inexistente, ya que la mayoría de recursos carecen de elementos descriptivos, o bien, la metainformación que pueden contener tanto los documentos como los registros, no es tan descriptiva ni está tan normalizada como la que por ejemplo caracteriza a la catalogación bibliográfica o a la descripción de registros de bases de datos.

No obstante cada vez están adquiriendo mayor importancia sistemas de metadatos que tratan de hacer frente a este problema como Dublin Core, TEI, etcétera, aunque, la mayor utilización del lenguaje XML y de sistemas como RDF, basado en su sintaxis, ha de contribuir a obtener una recuperación más precisa.

2.2.2.2. Indización

El diccionario de Documentación define la indización como:

> "Técnica del Análisis Documental que describe y representa el contenido de las fuentes de información documentales mediante un número limitado de conceptos extraídos del texto de los documentos (palabras clave) o de vocabularios controlados (clasificaciones, listas de materia, tesauros), que van a permitir el control y la recuperación de la información de un conjunto documental dado."

Consiste en asignar al documento una serie de términos o descriptores, que expresan su contenido, con el fin de facilitar su localización y acceso. Esto se puede hacer partiendo de un índice de clasificación ya elaborado (lenguaje precoordinado) o bien, en el caso de la indización basada en el lenguaje libre (lenguajes postcoordinados), en el que los propios términos permiten elaborar el índice. El uso de un lenguaje u otro da lugar, en el primer caso, a la indización elaborada, y por otro, a la indización libre, siendo esta última la más habitual entre los motores de búsqueda de la Web. Este tipo de indización es más amplia al superar la indización basada en el contenido.

Por otro lado, como indica Lancaster (1998), la indización puede ser exhaustiva o específica. Aunque la primera trata de utilizar el mayor número de descriptores de un documento, en la Web, se ve superada por el funcionamiento de los buscadores que indizan a texto completo. La específica, por su parte, trata de expresar de forma más concreta el contenido del recurso. Puede ser realizada por personas o bien de forma automática. En las herramientas de búsqueda se utilizan ambos tipos de indización, correspondiendo la primera a los directorios y la segunda a los buscadores.

Como señala Delgado (2001) la indización automática utiliza varios modelos que han sido recogidos por Ingwersen (1994) a los que denomina: unitérmino, en el sentido ya asignado por Taube a mediados de siglo; unitérmino ponderado; unitérmino en contexto, principalmente utilizado en la realización índices KWIC y KWOK, y finalmente la indización estructurada, que como indica Vianello (2004:240) "La representación resultante puede ser expresada a través de espacios vectoriales o clustering."

Caracteriza a la indización automática la aplicación de determinados algoritmos que tienen como finalidad representar lo más fielmente posible el contenido del documento para facilitar su recuperación.

En la indización automática los términos y frases que se extraen, no siempre son representativos sino que más bien son indicativos de que los documentos contienen dichos términos, lo que unido a la no utilización de sistemas estructurados, da lugar a un exceso de ruido en la recuperación.

Para evitar este problema, y también para facilitar la ordenación por relevancia, a estos términos se les aplican una serie de valores como pueden ser los basados en cálculos de frecuencias que otorgan un peso a un término en función de su frecuencia en el documento o en la base de datos. Los cálculos se basan en la ley de Zipf, posteriormente desarrollada por otros autores como Gerard Salton (1983), Rijsbergen (1979) y otros, dando lugar a conceptos como el de "ponderación de frecuencia inversa de un documento" (FID o IDF en el ámbito anglófono) y "frecuencia absoluta de un término" (FT o TF), que en ocasiones se utilizan de forma combinada. Los cálculos son asignados a los términos que finalmente representan el contenido del documento para favorecer tanto la recuperación de documentos relevantes como su ordenación.

Las técnicas aplicadas a la indización tienen pues como finalidad, permitir seleccionar los documentos más representativos en relación con la ecuación de búsqueda y colocarlos en los primeros puestos de la lista de resultados. Es decir, tratan de proporcionar la mayor precisión en los resultados. Le Loarer (1994) distingue entre indización plana, cuando todos los términos que se extraen tienen igual importancia; ponderada, cuando se da mayor importancia a unos términos que a otros; por roles o facetada y estructurada. Dejando al margen ésta última, a la que acabamos de referirnos, la indización por roles, según Vianello (2004) requiere para su correcto funcionamiento su aplicación a dominios específicos que permitan apoyarse en repertorios terminológicos muy especializados. De aquí que su aplicación a los buscadores generales de la Web ofrezca grandes dificultades.

El resultado de este proceso es la generación de un índice invertido, que se utilizará en la recuperación, aunque como esta autora indica, su elaboración puede realizarse siguiendo diferentes técnicas: técnica del fichero inverso, los ficheros de patrones de bits, árboles de PAT y grafos.

Por otro lado, el gran desarrollo alcanzado en los últimos años en el campo de la lingüística computacional por la semántica léxica se ha hecho notar como señala Vianello (2004:244) en los sistemas de recuperación multilingüe; en la elaboración de resúmenes automáticos; en la aplicación de clasificaciones documentales y en la traducción automática. Fruto de todo ello son los analizadores lingüísticos que permiten analizar documentos electrónicos para extraer información muy útil de cara a su localización. Otras herramientas que son fruto del desarrollo de la investigación en este campo y que ya empiezan a utilizarse en el ámbito de la Web son las ontologías, que a grandes rasgos podemos decir que se dedican a la clasificación no de términos sino de conceptos. Vianello (2004:249) menciona en este sentido las iniciativas denominadas FrameNet[73], proyecto de la Universidad de Berkeley y WordNet[74]. En España estudios de este tipo se están llevando a cabo por Subirats y Petruck, de la Universidad Autónoma de Barcelona.

2.2.2.3. Clasificación

Según el Diccionario de las ciencias documentales, clasificar es:

> "la operación que consiste en agrupar uno o varios elementos en un conjunto o clase en virtud de al menos una propiedad o atributo que tienen en común."

En el ámbito de las Ciencias de la Documentación, la citada obra define la clasificación bibliográfica o documental como:

> "la operación de agrupar los documentos según su tema –clasificación bibliográfica- o su contexto de producción –clasificación archivística."

Para su realización se requiere un sistema o esquema de clasificación anteriormente establecido.

---

[73] http://www.icsi.berkeley.edu/~framenet/
[74] http://www.cogsci.princeton.edu/%7Ewn/

Algunos de los problemas que presenta la indización libre pueden resolverse bien con la indización controlada o con los lenguajes clasificatorios. De aquí que en descripciones bibliográficas se utilicen ambos.

La Clasificación es un tipo de indización controlada, ya que requiere de la existencia de un índice clasificatorio. En el ámbito bibliográfico se han venido utilizando fundamentalmente las clasificaciones decimales como la Universal Decimal Classification (CDU), la Dewey Decimal Classification (DDC) y la Library of Congress Classification (LCC).

A efectos de indización, entre sus características destaca el carácter unívoco de sus elementos, es decir, un elemento clasificatorio significa una cosa, y sólo una.

En las herramientas de recuperación de la Web, la clasificación se utiliza fundamentalmente en el ámbito de los Directorios y de las bibliotecas electrónicas para organizar los recursos de forma temática. El problema es que unos y otros aplican clasificaciones de elaboración propia, alejándose de la idea de universalidad.

Los motores de búsqueda, en muchos casos incorporan directorios temáticos, bien de elaboración propia o por empresas asociadas, cuyo uso se ha ido haciendo más problemático a medida que han ido aumentando los contenidos de la web. No obstante, existen directorios que utilizan mecanismos automáticos para clasificar los recursos o páginas web[75].

2.2.2.4. El resumen

Las definiciones de este término que ofrecen diferentes autores y que recoge el Diccionario de las Ciencias Documentales tienen en común el referirse a él como:

> "la representación concisa, en lenguaje del autor, de las ideas principales del documento original analizado, evitando cualquier apreciación o juicio crítico"

La realización de resúmenes se ha venido efectuando tradicionalmente por personas especializadas, fundamentalmente del campo de la documentación así como por los

[75] Véase como ejemplo el directorio Scorpion accesible en la dirección <http://orc.rsch.oclc.org:6109>

propios autores. En el ámbito de la Web se están utilizando herramientas que realizan este trabajo de forma automática, como es el caso de WebCompass[76].

Los buscadores presentan tras la búsqueda un extracto de dos o tres líneas que contiene una determinada parte del documento. En algunos casos muestran la parte del texto que contiene el término o alguno de los términos buscados. El proceso se realiza automáticamente. Para su elaboración se tiene en cuenta el texto, etiquetas Meta descriptivas, y la descripción que el sistema de búsqueda aporta a los recursos.

Dada la importancia que estos aspectos tienen en la recuperación y teniendo en cuenta el estudio de Stobart y Kerridge (1996) que señala que para los usuarios esta información es nula o insuficiente, nos parece interesante valorar si la información que ofrecen expresa el contenido. Nos basaremos en la valoración de las técnicas que presenta de forma destacada los términos de búsqueda, así como los registros relacionados de un mismo sitio web.

2.2.3. La búsqueda y recuperación de información

La búsqueda de información es la fase del proceso documental a la que se recurre cuando se necesita información sobre un tema, asunto o la obtención de un dato. Es recomendable para su correcta formulación conocer el lenguaje de búsqueda que utiliza la base de datos. Una búsqueda puede ser simple, basada en uno o varios términos; booleana cuando utilizamos los operadores de este tipo; puede utilizar el lenguaje natural o también puede lanzarse una búsqueda por frase. Chang establece cinco tipos de búsquedas: Búsquedas por término simple; por términos múltiples basados en la lógica booleana; búsquedas basadas en el contexto, que utilizan operadores de proximidad o signos como las comillas para la búsqueda por frase; los que utilizan el lenguaje natural, que permite establecer preguntas directas al índice; y finalmente la búsqueda por correspondencia de patrones, que permiten el uso de operadores de truncamiento. Nos centraremos en ellos a la hora de proponer los temas de búsqueda en la evaluación.

La expresión utilizada se denomina "ecuación de búsqueda" y es analizada por el sistema, que realiza una representación interna, que se compara con los términos del

[76] <ftp://ftp.qdeck.com/demo/webcompass/wc200_tr.exe>

índice. Para mejorarla puede existir como herramienta intermedia un tesauro, pero en los sistemas basados en lenguaje natural no siempre se utiliza.

Existen dos puntos de vista desde los que se han venido analizando los resultados de las búsquedas en estos sistemas. Por un lado el centrado en el sistema y por otro el centrado en el usuario.

Las teorías más modernas en torno al proceso de búsqueda, sobre todo desde mediados de los 80, dejan de centrarse en el sistema y en las funciones del tratamiento documental como eje de la recuperación para dar mayor importancia al usuario, que ha de enfrentarse directamente a los sistemas interactivos. En este sentido, Baeza-Yates y Ribeiro-Neto (1999) ya señalan la existencia de tres tipos de sistema en función de la labor que debe llevar a cabo el usuario para recuperar información: los sistemas adhoc basados en colecciones que no varían con gran frecuencia y que requieren para su consulta la utilización de los modelos clásicos (booleano, vectorial y probabilístico); filtering o sistemas con actualizaciones constantes que requieren la utilización de un perfil de usuario, y finalmente el browsing que es el que utilizan los sistemas hipertextuales. Actualmente, y fundamentalmente en el ámbito de la Web, están adquiriendo cierta importancia otros sistemas de recuperación de carácter visual, basados en interfaces gráficos y caracterizados por la agrupación de recursos relacionados.

También se han establecido diferentes modelos en relación con el proceso de búsqueda de información (ISP). Así, Bates (1986) señala tres etapas: Acceso, Búsqueda y Selección. Kuhlthau (1991), desde el punto de vista del usuario propone un modelo más amplio, similar al de Ingwersen (1992), que consta de seis etapas: iniciación, que se refiere a la necesidad de información de la persona; selección, marcado por el momento de identificación y selección del tema de búsqueda, con los aspectos personales que puedan influir; exploración, que se caracteriza por la posible aparición de sentimientos de confusión, incertidumbre o duda ante la resolución del problema; formulación, que corresponde al momento de superación de la fase anterior y se caracteriza por el incremento de la confianza y disminución de la incertidumbre; colección, que marca el momento en que mejor funciona la interacción del usuario con el sistema y finalmente la presentación, que puede estar marcada por un sentimiento de satisfacción si los resultados han sido positivos o de sorpresa si no ha sido así.

Hernández (1999b) expone el modelo de R. E. Berkowitz denominado Big six skills que también contiene seis grandes pasos:

En primer lugar, la definición de la tarea: consiste en definir el problema de la información e identificar la información que se necesita para resolverlo. A continuación ha de elaborarse la estrategia de búsqueda, determinando las posibles fuentes y evaluándolas para seleccionar las más apropiadas. En tercer lugar, se ha de localizar y acceder a las fuentes para encontrar la información y hacer uso de ella. El cuarto paso requiere organizar la información de las fuentes a las que se ha accedido y presentarla. Finalmente queda juzgar la información obtenida (efectividad) y el proceso de resolución del problema (eficiencia). Chowdury (2001) se refiere a estos últimos conceptos señalando que cuando buscamos información en la Web, bien sea en motores de búsqueda, directorios, bases de datos temáticas y bibliotecas virtuales, queremos efectividad, es decir, encontrar la información que nos interesa, y eficiencia, es decir, de una forma rápida y asequible.

El carácter dinámico de las búsquedas en línea debe ir encaminada a mejorar la efectividad. Chowdury (1999) señala además la necesidad de una estrategia en la que se ha de atender a:

1. Los conceptos o facetas a buscar y su orden.
2. Los términos que representan el concepto.
3. Las características del sistema de recuperación.
4. La revisión de la búsqueda.

Aunque los pasos señalados más arriba por Berkowitz nos parecen acertados, hay que pensar que en la Web, más que en otro sistema de recuperación de información, la búsqueda de información es un proceso que tendrá mejores resultados cuanto mayores sean los conocimientos sobre el funcionamiento y forma de consulta de las herramientas así como sobre sus prestaciones. De aquí que Tramullas (2002:606) proponga frente a este planteamiento clásico un nuevo esquema para enriquecer la recuperación de información en este medio. Señala en primer lugar la selección del recurso o recursos a utilizar, al que seguirían la identificación de las herramientas y sus posibilidades técnicas, la formulación y traducción de las expresiones lógicas a expresiones interpretables por el sistema, la ejecución del proceso de consulta, la obtención de resultados y su posterior valoración para finalizar la consulta o bien intentar mejorarla.

Pero el desarrollo de una buena búsqueda requiere además conocer la naturaleza y organización de la base de datos y las necesidades del usuario. En función de esas necesidades puede requerirse una alta exhaustividad o una alta precisión.

Los sistemas tradicionales de búsqueda que vienen utilizándose en centros de documentación o similares, requieren una breve entrevista o resolución de un simple cuestionario con el usuario, y el proceso de búsqueda comienza decidiendo los términos, la base de datos, uso del tesauro para traducir los términos de la búsqueda a los términos apropiados, seleccionar los que posiblemente haya elegido el indizador, coordinar los términos y lanzar la búsqueda. El proceso continúa repitiendo los dos últimos pasos hasta obtener los resultados deseados e identificar los resultados relevantes. El carácter interactivo de los sistemas que utilizan la Web ha cambiado los mecanismos de búsqueda y recuperación de información, como veremos a continuación.

Como hemos señalado, los principales modelos[77] que se utilizan para realizar búsquedas en bases de datos textuales son: el modelo booleano, caracterizado por el uso de los operadores lógicos y que trata de recuperar los documentos que contienen los términos expresados en la ecuación de búsqueda; el modelo booleano extendido, que valora la existencia de un mayor número de términos coincidentes entre los términos de la expresión de búsqueda y los del documento, es decir, se aplican valores a los términos que permitirán colocar en los primeros puestos los que mejor cumplan la condición que se solicita en la ecuación de búsqueda; el modelo vectorial, basado en la representación en forma de vectores tanto de la ecuación de búsqueda como del documento en el que se recupera no sólo la existencia de unos términos sino también permite expresar su mayor o menor importancia, lo que facilita la recuperación de documentos que se aproximen a la ecuación de búsqueda, permitiendo aplicar técnicas de clustering; el probabilístico, que recupera los documentos en base a cálculos de frecuencias y de probabilidades entre los documentos y la expresión de búsqueda, en función de los cuales son ordenados. En último lugar hay que mencionar la equiparación exacta o best match, que compara los términos de la búsqueda con representaciones del documento ordenando los resultados. Este modelo está centrado en la ordenación. Por un lado se mide la importancia relativa de los ítems recuperados y por otro se asigna un valor a los términos de búsqueda. Así, las búsquedas de este tipo reúnen un conjunto de palabras que se lanzan contra la base de datos,

[77] Baeza-Yates y Riviero-Neto (1999:19-71) establecen 15 modelos de recuperación de la información, si bien los más utilizados son los clásicos, es decir el booleano, el probabilístico y el vectorial.

se calcula la similitud entre ambos y se ordenan los valores de forma decreciente de acuerdo con cálculos de frecuencias.

Los modelos booleanos permiten incluir cálculos estadísticos basados en el análisis de términos y en sus relaciones lo que facilita la ordenación de los resultados de la búsqueda. Las medidas más utilizadas tienden a calcular la frecuencia de términos, el peso, la proximidad y posición de las palabras, co-ocurrencias, etcétera.

Los modelos que utilizan cálculos estadísticos o probabilísticos de ocurrencias de términos, no tienen en cuenta aspectos sintácticos, pragmáticos ni semánticos. Estos aspectos son utilizados por otros modelos como los basados en el conocimiento, a través del uso de archivos de sinónimos, tesauros, etcétera. También son utilizados en el modelo conceptual y en las búsquedas basadas en el lenguaje natural. Finalmente se están desarrollando otros modelos que tienen en cuenta el carácter binario de la información de los documentos electrónicos[78].

Los métodos más sencillos de recuperación son los de búsqueda por texto libre, utilizados sobre todo en pequeñas colecciones, como los buscadores de documentos propios que facilitan determinados sistemas operativos, y por otro lado el modelo booleano.

Actualmente se utilizan también en el contexto de la Web, cuyos sistemas, además, han desarrollado otras formas de búsqueda basadas en la agrupación en categorías que utilizan el hipertexto.

Para estos últimos Hildred (1989) plantea la necesidad de favorecer los procesos de exploración y navegación de estos sistemas como es la doble posibilidad de consulta de la Base de datos. De aquí la preocupación de los desarrolladores de buscadores web por integrar unos y otros sistemas de búsqueda para ser útiles a todo tipo de usuarios.

Como hemos señalado anteriormente, el proceso de búsqueda se ha ido modificando al mismo tiempo que lo han hecho las bases de datos, los motores de búsqueda y los lenguajes de interrogación, que en el caso del lenguaje libre, permite una cierta automatización de las búsquedas y prescindir de determinados operadores, del uso de tesau-

[78] Para más información véase el documento: A practical guide to evaluating information retrieval systems, accesible en la dirección: http://web.archive.org/web/19980522211615/http://www.excalibur.be/GB/WPapers/g...[Consulta 15-9-05].

ros, etcétera. Esta simplificación, sin embargo, debería verse compensada con una mayor calidad en los resultados de las búsquedas.

Por esto podemos decir que sobre todo en la Web la recuperación queda limitada prácticamente al modo en que se realiza la indización, que como hemos visto, prácticamente se basa en la indización de los términos que forman parte del documento. Este es uno de los aspectos más importantes y en el que más están incidiendo los desarrolladores de estas herramientas, ya que en virtud de la representación del documento y de los cálculos de similitud que se apliquen en relación con la representación de las consultas, la recuperación tendrá mayor o menor éxito, ya que se ajustará más a lo buscado. En cualquier caso el usuario ha de tener en cuenta que las búsquedas en la Web se producen fundamentalmente sobre el texto del documento y sólo en determinados buscadores, sobre metadatos.

Así, frente al contexto de la búsqueda de información sobre un tema determinado, cuyo resultado puede ser la obtención de un listado de referencias relacionadas con los temas de búsqueda, en la Web, la mayoría de los buscadores ofrecen un listado de documentos que contienen los términos expresados en la ecuación de búsqueda.

El resultado de la búsqueda ha de ser obtener algo que previamente se ha introducido y la mejor forma de obtenerlo es conociendo de qué manera se ha introducido. Esto es interesante porque nos permite distinguir dos fases, una primera que podemos relacionar con el sistema y la segunda en la que actúa más la intervención y posterior valoración por parte del usuario. Esto puede ser determinante para enfocar de forma correcta la evaluación de los SRI, que ha de contemplar ambos aspectos.

Pero las evaluaciones realizadas desde el punto de vista del usuario ajustan sus criterios para medir en qué modo resuelven sus necesidades de información, valorando aspectos como la satisfacción del usuario respecto al tiempo de respuesta, a la interface, a la documentación en forma de ayudas, etcétera. Nosotros, como hemos señalado, nos preocupamos más de analizar los procesos que desarrollan estos sistemas y de valorar hasta qué punto se efectúan correctamente.

2.3. Los buscadores de información Web

Anteriormente nos hemos referido a la variada y amplia colección documental que forma la Web, a su dispersión y por el momento, escaso tratamiento, que al margen de determinadas iniciativas como las basadas en el uso de los metadatos, influyen en la

deficiente recuperación de información en este medio. Estudiaremos a continuación, tras una breve introducción histórica y un análisis de sus principales componentes, cómo los buscadores se enfrentan a este y otros problema y las soluciones que han ido planteando.

2.3.1. Orígenes de los principales motores de búsqueda y metabuscadores

Al tratar de Internet en la parte inicial del capítulo, nos hemos referido a algunas de las primeras herramientas de búsqueda que se utilizaban para recuperar información de la red. Dichas herramientas se consideran los precedentes inmediatos de las utilizadas en la Web.

Resulta interesante observar la evolución de estas herramientas, ya que podemos apreciar cómo a lo largo de este proceso se han ido fraguando gran parte de las características que actualmente mantienen tanto los directorios como los motores de búsqueda más utilizados actualmente, y lo que es más importante, cómo cada una de estas herramientas trata de solucionar los problemas que se les van planteando en relación con la recuperación de información.

En este sentido, hemos de hacer notar, que desde los comienzos de la Web, aprovechando su carácter hipertextual, en los sistemas de búsqueda se adoptaron las técnicas de búsqueda basadas en la activación de enlaces o navegación. Esto dio lugar a la creación en 1994 de directorios de elaboración manual como EINetGalaxy[79] que contenía principalmente recursos Telnet y Gopher, o GENVL (Generate Virtual Library) aunque ambos con un número muy limitado de opciones de búsqueda de recursos y con problemas de actualización.

En 1995 apareció el directorio Yahoo (Yet Another Hierarchical Officious Oracle), que basaba la formación de su base de datos en el envío por parte de los interesados, de información sobre webs temáticas o páginas web particulares, aplicando posteriormente programas rastreadores, aunque las páginas seguían siendo analizadas y clasificadas por especialistas. Este aspecto, unido a un mayor desarrollo de las clasificaciones, supuso un claro avance respecto a los anteriormente señalados.

[79] http://www.galaxi.com

Les siguieron otras iniciativas dirigidas hacia la especialización por materias de estas herramientas como es el caso de SOSIG[80] para ciencias sociales, EEVL[81] para ingeniería y Biz/ed[82] para ciencias económicas y empresariales.

Los directorios, en muchos casos, iban acompañados además por formularios que permitían la búsqueda en sus índices.

Los SRI web, basados fundamentalmente en la utilización de este último modo de búsqueda, tienen como antecedentes, en el ámbito de la Web, a los motores Harvest[83] y ALIWEB[84] (Archie-Like Indexing of the WEB). Basados en el sistema WAIS, son los más claros antecedentes de los actuales motores de búsqueda, aunque se vieron inmediatamente superados por los primeros buscadores, como es el caso de World Wide Web Wanderer, caracterizados por el mayor uso de los programas araña o spiders, que se utilizaban como medio fundamental para la formación de las bases de datos de direcciones URL.

Wanderer, World Wide Web Worm, JumpStation[85] y Repository-Based Software Engineering (RBSE) surgen a finales de 1993. Frente a la ordenación aleatoria de los registros que ofrecían los dos primeros, el tercero ya aplicaba un mecanismo basado en la relevancia. (Sonnenreich, 1997).

Pronto se pensó en la posibilidad de consultar de forma simultánea estas bases de datos, y así, en 1994 se realiza el primer proyecto de creación de un metabuscador, al que se denominó GIOSS (Glossary-of-Servers Server), al que siguió el sistema Discover, que permitía buscar en más de 500 servidores WAIS. Un año más tarde, Brian Pinkerton, de la Universidad de Washington, crea una herramienta de escritorio que dará lugar al

[80]http://sosig.ac.uk
[81]http://www.eevl.ac.uk
[82]http://www.bized.ac.uk
[83]Fue cancelado en agosto de 1998 (Amat 1999b).
[84]Lanzado en 1993, actualmente carece de mantenimiento. Se basaba en la formación de índices mediante la información aportada por los propios responsables del web. Más información sobre ambos buscadores puede consultarse en: KOSTER, M. ALIWEB (Archie-Like Indexing of the Web). *Computer Networks and ISDN Systems*, vol. 27, 1995: pp. 175-182, y sobre Harvest véase Bowman, C. y otros. The Harvest information discovery and acces system. *Computer Networks and ISDN Systems*, vol. 28, 1995: pp. 119-125.
[85]Considerado el primer buscador para autores como Jenkins y otros.

metabuscador WebCrawler[86], caracterizado por incorporar la indización a texto completo.

En el seno de la Universidad Carnegie Mellon, surge en 1994 el buscador Lycos, que desarrollará los sistemas existentes y ampliará la capacidad de sus bases de datos indizando una gran cantidad de páginas, con lo que adquirió cierta importancia. En 1995 se lanzan InfoSeek[87] y Excite, que tiene sus orígenes en el proyecto denominado Architext, caracterizado por recuperar, además, recursos relacionados con los términos de búsqueda.

También en 1995 surgen Northern Light[88] y AltaVista[89]. Sonnenreich (1997) señala que éste último destacaba por su rapidez, por permitir tanto el uso del lenguaje natural como de operadores booleanos, así como búsquedas en las news, recuperación de imágenes con texto y la búsqueda en determinados campos como el de título. HotBot e Inktomi se lanzan en 1996 destacando el primero por su capacidad al indizar más de diez millones de páginas diarias y el segundo por elaborar un directorio de forma automática con la intervención humana para valorar la calidad de sus recursos. Les siguieron AskJeeves en 1997 y Google en 1998. El primero se caracteriza por permitir realizar preguntas directas y el segundo por el hecho de presentar resultados relevantes ordenados mediante técnicas basadas en el análisis de enlaces.

En 1999 aparecen el motor de búsqueda AllTheWeb, que además de una mayor calidad en la recuperación de recursos, aportó la inclusión de recursos que dependían de otros, y el directorio Open Directory que pronto facilitará asistencia a la mayoría de buscadores importantes como AltaVista, AOL Search, Dogphile, HotBot, Lycos, MetaCrawler y Netscape (Sherman, 2000).

En el año 2000, surge Teoma que será adquirido por AskJeeves al año siguiente. Este buscador plantea una mayor selección de recursos, facilitando la intervención de expertos que proponen y valoran recursos específicos.

---

[86]http://metacrawler.cs.washington.edu:8080/home.html
[87]Actualmente Go.
[88]Actualmente de pago.
[89]Fuera de servicio en la actualidad.

De forma simultánea se siguieron desarrollando listados de recursos organizados por temas, que dan lugar a valiosos instrumentos de recuperación como el World Wide Web Virtual Libray[90], primera lista alfabética de materias puesta a disposición del público por el CERN, el BUBL[91] Subject Tree, NISS Information Gateway[92] y Galaxy[93], algunos de los cuales, como indica Winship (1995) llevan asociado el modo de búsqueda por palabras clave.

La tendencia en los últimos años ha seguido siendo la creación de múltiples herramientas de búsqueda. Es de destacar el desarrollo de herramientas especializadas tanto en una materia, como en ámbitos territoriales o lingüísticos. En este sentido debemos mencionar la aparición del metabuscador SurfWax[94] que no sólo permite dirigir las búsquedas a un campo especializado, sino que se constituye como herramienta de búsqueda en diversas fuentes, incluidos algunos contenidos de la Web invisible como pueden ser las bases de datos.

Los buscadores no han permanecido estancados a lo largo de su existencia sino que, como hemos observado, se trata de herramientas en constante evolución, que generalmente han ido haciendo frente a los problemas que se les plantean mediante la incorporación de nuevas posibilidades de búsqueda, de técnicas de indización, de recuperación y de ordenación de resultados. Especial mención merece el desarrollo alcanzado en técnicas de análisis de enlaces, pues además de Google lo utilizan HotBot, Excite o Clever[95]. IBM quiere ir más allá y está desarrollando un programa que permitirá a los robots recorrer la red, evitando los sitios irrelevantes[96].

Otro aspecto a considerar es el que tiene que ver con la trayectoria de estas herramientas, ya que los motores de búsqueda no sólo han venido aumentando sino que también se han dado casos de desaparición o han sido absorbidos por otros, como en el caso de World Wide Web Worm, anteriormente citado o el de OpenText, que se especia-

---

[90] http://vlib.org
[91] http://bubl.ac.uk
[92] http.//www.niss.ac.uk
[93] http://www.galaxy.com. Adquirido en 2001 por First Search
[94] http://www.surfwax.com
[95]http://www.almaden.ibm.com/cs/k53/clever.html
[96]El programa se llama Focused Crawler, caracterizado por efectuar análisis de enlaces. Se puede incorporar a los buscadores añadiéndoles más funcionalidad.

lizó en información económica y financiera bajo la denominación Livelink Pinstripe hasta su desaparición; o como Northern Light que ha pasado a ofrecer sólo su colección especial de pago. Otros casos de absorción más recientes son los de AltaVista y AllTheWeb[97]. HotBot y Excite, se han convertido en metabuscadores.

En la mayoría de casos, los motores de búsqueda tratan de obtener un rendimiento comercial mediante la inclusión de publicidad en sus páginas o utilizan otras prácticas comerciales como asegurar la aparición de páginas web en los primeros lugares del listado de búsquedas o en un lugar destacado junto al resto de recursos recuperados. Buscadores como Lycos ofrecen en sus contratos, además su inclusión en los índices de otros buscadores que utilizan su base de datos, rápida indización, permanencia asegurada en los índices y actualización del contenido. Por otro lado, pueden asegurase también una serie de ingresos mediante técnicas como el pay per click[98].

Este carácter mercantil va a jugar un papel muy importante en el funcionamiento de los motores ya que da mayores posibilidades de recuperación a empresas que pagan por ser fácil y rápidamente localizables sus productos, en detrimento de una recuperación de mayor precisión e interés científico.

Hay que tener en cuenta además la existencia de buscadores como AllTheWeb e Inktomi, que facilitan sus índices, para ser utilizados por otros buscadores como Lycos, HotBot, Terra.com, Overture, Infospace, Excite y Dogpile, en el caso del primero de los buscadores y el segundo a HotBot, About.com, MSN, Espotting.com, LookSmart, Soneraplaza, Goo y Bluewin[99], asegurando la visibilidad en el 70% de las búsquedas.

Otra característica actual es el establecimiento de alianzas, compras y ventas, para compartir bases de datos. El directorio Yahoo controla desde marzo de 2003 Inktomi[100], y desde octubre del mismo año, Lycos y HotBot y Overture (que a su vez había comprado Go, AltaVista y AllTheWeb).

---

[97] AllTheWeb fue absorbido en 2004 por Overture, que a su vez lo adquirió Yahoo!. AltaVista también fue absorbido por la misma compañía en 2003 y desde 2004 utiliza la misma base de datos de Yahoo!.
[98] Técnica comercial que consiste en que las compañías que lo contratan pagan un canon a los buscadores cada vez que un usuario accede a su página web al activar un enlace de la página de resultados de la búsqueda realizada.
[99] Datos que ofrece Lycos en http://insite.lycos.com/inclusion/searchenginessubmit.as [Consultado en mayo de 2004].
[100] Inktomi previamente colaboró con MSN y HotBot.

AskJeeves compró los buscadores Teoma en 2001, pasando a denominarse Ask, y Excite, que es actualmente un metabuscador que utiliza recursos proporcionados por Google, Yahoo, AllTheWeb, Inktomi, AskJeewes y los directorios LookSmart, About, Open Directory, Teoma y Overture.

Además, estas fusiones pueden dar lugar, como en el caso de Yahoo, a la formación de nuevas herramientas de búsqueda más potentes, como es el lanzamiento de su nuevo motor de búsqueda YahooSearch.

Este carácter dinámico tanto en lo técnico, con la continua incorporación de nuevas técnicas de mejora en la búsqueda, como en lo empresarial, con continuos convenios, compras y ventas entre ellos, exigen tanto al usuario como al especialista en RI un conocimiento actualizado de su evolución ya que es primordial para saber, en función de estas fusiones, a cuáles se les está dando una mayor importancia, cuál la puede ir perdiendo, etcétera. Exigen además la necesidad constante de replantearse modificaciones en la evaluación de motores de búsqueda y su realización de forma periódica para poder constatar los cambios que les afectan. Otro tema es la interpretación de el porqué se producen estas fusiones y compras. Desde nuestro punto de vista se trata de hacer frente al continuo desarrollo de determinadas herramientas que, por sus prestaciones, minimizan las existentes, por lo que la única manera de hacerles frente es mediante la ampliación o unión de las más pequeñas. También las mayores exigencias tanto de software como de hardware, hace que sea necesaria una continua inversión que no todas las empresas pueden soportar.

Google, junto a Yahoo y MSN, son los motores en los que se aprecia una mayor expansión que en los últimos tiempos, incorporando el primero de ellos nuevos espacios de búsqueda como Google Scholar[101], para recuperar principalmente información de carácter científico, Google Plans, para búsquedas de carácter geográfico[102], y en la actualidad se encuentra desarrollando un proyecto de digitalización de 15 millones de libros impresos entre 1850 y 1950 existentes en 4 bibliotecas americanas, una británica y otra española.

[101] La versión beta de esta herramienta puede consultarse en http://scholar.google.com
[102] Se accede a la versión beta en la dirección: http://maps.google.com

En la actualidad, como señala Chaín (2004), se están utilizando otras técnicas avanzadas de búsqueda que se basan en Algoritmos Genéticos y en la inteligencia artificial aplicada al procesamiento del lenguaje natural, enfocados a mejorar la interacción hombre-máquina y que posiblemente sea la tecnología que en los próximos años acabe implementándose en los nuevos sistemas.

2.3.2. Definición y clasificación

Chowdhury (1999) define de forma general a los buscadores como un servicio de recuperación de información que consiste en una base de datos, que contiene principalmente recursos disponibles en la Web. En un trabajo publicado en 2001 distingue entre Motores de búsqueda, Directorios, Bibliotecas virtuales y Bases de datos especializadas.

Poulter (1997) por su parte, aporta alguna novedad en su definición al señalar que se trata de servicios de recuperación que contienen una o más bases de datos, con la descripción de recursos disponibles en la Web, programas de búsqueda y una interfaz de usuario.

En relación con la terminología, hemos de señalar que no hay unanimidad respecto a la denominación de éstas herramientas. Nos encontramos con que existen términos y expresiones distintas referidas a estos servicios, si bien debemos señalar que tanto Servicios como Herramientas de búsqueda son los términos más comúnmente aceptados para referirse tanto a Buscadores y Bibliotecas Virtuales como a Bases de datos, utilizándose el primero de estos tres para referirse a Motores de búsqueda, Directorios. Nosotros, siguiendo a Oppenheim (2000) lo utilizaremos también para englobar a los Metabuscadores.

Chu y Rosenthal (1996) los denominan Ayudantes para búsquedas Web (Web search aids), comprendiendo tanto Catálogos, Directorios, Índices, Motores de búsqueda o Bases de datos Web. Para estos autores, los motores de búsqueda deben permitir al menos, que el propio usuario elabore su petición de búsqueda, frente a otras herramientas que permiten buscar información a través de caminos predefinidos o siguiendo estructuras jerárquicas.

Para Oppenheim (2000) todos son motores de búsqueda, diferenciando entre Robots, como AltaVista, Excite[103], Lycos, HotBot y Go; Directorios, como Yahoo[104], SOSIG[105], EEVL[106], Biz/Ed[107] y UK Directory[108]; Metabuscadores como MetaCrawler[109], Dogpile[110], Profusion[111], Ixquick[112], Vivísimo[113] y finalmente, Herramientas de software. Estas últimas son programas que se instalan individualmente en los ordenadores y que posibilitan acciones como guardar los resultados de las búsquedas en el propio disco duro o informarnos tanto de enlaces inactivos como de recursos duplicados. En este sentido, menciona Websleuth, Copernic 98[114], f.Search 1.3.1[115] y Query-N Metasearch.

Hock (1999) se refiere a Finding tools "Herramientas de Búsqueda" como expresión que abarca tanto a Motores de búsqueda (Search engines) como a Directorios (Directories).

En cambio para Ljoland (2000) una cosa son los motores de búsqueda web, dentro de los cuales sitúa a los metabuscadores y otra los directorios.

Green (1999), en función de sus capacidades técnicas, distingue entre buscadores de primera y segunda generación. Sitúa en el primer apartado a los anteriores a abril de 1998, iniciando Direct Hit[116] los de segunda generación. Concretamente éste último, incorporaba una tecnología que le permitía aprender de otras búsquedas, guardar información de los registros activados por los usuarios en búsquedas previas y así aumentar el valor de la relevancia cuando se recuperaran de nuevo.

---

[103]Actualmente es un metabuscador y pertenece a Askjeeves.
[104]En la actualidad cuenta tanto con un buscador al que denomina Yahoo! Search <http://search.yahoo.com>como con un directorio, Yahoo Directory <> aunque habitualmente se utiliza la página de acceso general al portal que integra ambas herramientas así como otro tipo de servicios. http://www.yahoo.com
[105]Especializado en Ciencias Sociales es accesible en http://www.sosig.ac.com
[106]Especializado en matemáticas e ingeniería es accesible en http://www.eevl.ac.uk
[107]Especializado en economía y turismo accesible en http://www.bized.ac.uk
[108]Accesible en http://www.ukdirectory.co.uk
[109]Accesible en http://www.metacrawler.com
[110]Accesible en http://www.dogpile.com
[111]Accesible en http://p16.profusion.com
[112]Accesible en http://www.ixquick.com
[113]Accesible en http://vivisimo.com
[114]Puede descargarse en http://www.copernic.com
[115]Puede descargarse en http://www.karai.com/software/fsearch
[116] Fue clausurado en 2002.

Hay aspectos técnicos que diferencian a unos de otros ya que por ejemplo, mientras los de primera generación forman sus índices a partir de la visita de los "robots" a sitios web, analizando la localización y frecuencia de las palabras, los más modernos basan su ranking[117] de resultados en diferentes aspectos como la cantidad de veces que es visitado un sitio web. Los recursos obtenidos, en este sentido, tratan de responder tanto a los criterios de búsqueda como al hecho de ser muy visitados o utilizados. Otro aspecto que los distingue es que mientras que los primeros no consideran el contexto de los términos de búsqueda sino que hacen búsquedas literales de los términos, los de segunda generación permiten hacer búsquedas tanto utilizando el lenguaje natural como búsquedas conceptuales. Los motores que utilizan estas tecnologías, fundamentalmente Google y Teoma, utilizan además el análisis de enlaces lo que les permite conocer, en el primero de ellos, los recursos que reciben un mayor número de enlaces, es decir, lo que equivaldría a ser los más citados en los cálculos de impacto, y el segundo de ellos trata de ofrecer recursos de calidad al detectar los considerados como "autoridad" en una materia.

Siguiendo con la clasificación, Tramullas (2002) se refiere a los motores de tercera generación que se caracterizan por utilizar, en la recuperación, operaciones basadas tanto en el cálculo vectorial como en el análisis de enlaces.

Aguillo (1998) distingue entre Índices y Buscadores o Motores de búsqueda. Maldonado y Fernández (2000) distinguen entre Motores de búsqueda e Índices temáticos. Abadal (2001) los denomina Localizadores de recursos, entre los que distingue Índices temáticos, Buscadores o Robots y Metabuscadores. Por su contenido, establece diferencias entre Localizadores generales y especializados, subdividiendo estos últimos a su vez según se ocupen de un tema, ámbito geográfico, de determinados tipos de documentos o de un ámbito lingüístico concreto.

Aunque más adelante nos ocuparemos de una forma más específica de unos y otros, conviene señalar que hay dos aspectos fundamentales que distinguen a buscadores temáticos de los automáticos: Por una lado, respecto a la formación de sus bases de datos, los motores la forman mediante la visita indiscriminada y generalmente no selectiva, de servidores Web por parte del robot, frente a los buscadores temáticos cuyos recursos son

[117]Aguillo (2001) señala la existencia de criterios comerciales en el modo en que determinados buscadores realizan sus listados.

seleccionados por personas que se ocupan de su mantenimiento. Por otro lado, desde el punto de vista de la técnica de la búsqueda, la forma preferente de consulta de sus bases de datos, consiste en los buscadores automáticos en el lanzamiento de los términos expresados en el formulario de búsqueda, mientras que en los temáticos es necesario un examen, análisis o selección de términos o materias de las listas que ofrecen. Estas técnicas se conocen como Interrogación y Exploración respectivamente. Este último término, que puede ser utilizado junto a Examen, concuerda bien con la característica de estos directorios de contener más referencias de sitios web susceptibles de exploración que de visualización directa de lo documentos individuales y por eso suelen recoger las direcciones de las páginas principales de un sitio web.

Aunque Yahoo aún lo mantiene, inicialmente los grandes buscadores como AltaVista, Lycos, etcétera, presentaban en las páginas principales ambas formas de búsqueda. La tendencia actual ya no es tanto mantener ambos modos de búsqueda en la misma página del buscador sino separar ambos modos de consulta mediante interfaces distintos, accesibles a través de distinta URL.

Hay casos como Google que busca al mismo tiempo en la base de datos del Directorio, o como Lycos que indica, tras una búsqueda, las categorías del directorio en el que encontrar recursos relacionados con la expresión de búsqueda. Otros utilizan interfaces visuales y combinan ambos tipos de búsqueda, ya que requieren primero la interrogación de la base de datos y la exploración posterior.

En los casos en que acompañan a los motores de búsqueda, podemos decir que se trata de herramientas complementarias, que en determinadas búsquedas y para diferentes usuarios, puede ser un modo más adecuado de localizar información, ya que generalmente se ofrece con un mayor grado de contextualización.

2.3.2.1. Directorios

El término Directorio es una traducción del inglés Directory que junto con la expresión classified lists, denomina a los buscadores caracterizados por clasificar los registros almacenados en su base de datos bien en grupos temáticos, que a su vez pueden subdividirse en otros más específicos, o bien mediante clasificaciones como la CDU, la de Dewey u otras. Esta labor, como hemos dicho, la realizan especialistas que seleccionan dichos recursos de la Web o de los envíos que los particulares realizan, para que sean incluidas sus páginas. Cada directorio puede tener una política de selección determinada

y el propio tamaño de la Web puede hacernos comprender que en cualquier caso ha de ser restrictiva y limitada. El modo de funcionamiento de los buscadores temáticos o Directorios, está basado en los sistemas hipertextuales, muy utilizados en Internet desde sus orígenes, pues como ha señalado Koch (1996) se puede observar desde la forma más temprana de uso del protocolo FTP, Gopher y actualmente en la Web.

Peña y otros (2002:353), estiman que los directorios más importantes pueden contener más de 1000 categorías y un millón de sitios web.

Con el crecimiento de los recursos disponibles, esta forma de acceso a través de enlaces dio lugar al establecimiento de directorios que recogían ordenadamente los recursos, sin que fuera necesario activar varios enlaces para llegar a la información deseada, actuando de forma similar a como lo hacen los bookmarks.

En estas herramientas podemos distinguir dos partes fundamentales que son la base de datos y la estructura jerárquica de materias que facilitan el acceso a los recursos y sirve de interfaz.

El acceso a los recursos se realiza a través de enlaces que están ordenados por temas o categorías y subcategorías, utilizando índices alfabéticos, numéricos o alfanuméricos que se activan pulsando sobre ellos, al tiempo que se va descendiendo en la estructura arbórea hasta llegar al grupo o submateria que contiene los recursos de interés. El problema aquí radica en que no siempre son personas especializadas las que indican en qué categoría incluir los recursos, sino que en determinados directorios, las personas que envían la información sobre el recurso para ser incluido en el buscador indican en qué categoría ha de incluirse el documento.

En el caso de haber indizado sus recursos, además se puede lanzar una búsqueda en su motor interno, que ofrece un listado con los recursos pertinentes a la ecuación de búsqueda localizados en su base de datos. En estos sistemas la indización se realiza de una forma básica al centrarse en el URL, en el título, en el resumen, si lo hubiere, o en otras partes del texto, y no sobre el texto completo como ocurre en la mayoría de los motores de búsqueda.

Los recursos que muestran pueden ir acompañados de indicaciones o valoraciones sobre la importancia del recurso.

La evolución de estas herramientas se ha visto marcada por tratar de ofrecer un servicio más completo, permitiendo así el acceso tanto a información específica, como a

noticias de todo tipo, comercio electrónico, correo, etcétera, lo que hace que a menudo se les considere como auténticos "Portales de acceso a Internet". Los ejemplos más claros con los que contamos son Lycos y Yahoo.

Amat (1999) señala alguna de sus limitaciones y virtudes. Entre las primeras indica que estos sistemas sólo cubren una mínima fracción de los recursos disponibles; que sus estructuras de navegación no constituyen sistemas controlados, extensibles y reconocibles de estructuración del conocimiento como son algunos de los sistemas de clasificación en uso, lo que conlleva una falta de coherencia y fiabilidad tanto para indizadores como para usuarios. Indica además la existencia de deficiencias en su lógica, en sus jerarquías, en su desglose de categorías, en su terminología, en la forma en que se relacionan diferentes clases y en la capacidad de polijerarquía. Aunque estamos de acuerdo en la mayoría de estas apreciaciones, pensamos que no se debe generalizar ya que por ejemplo existen directorios que utilizan clasificaciones reconocidas no sólo de carácter universal como la Clasificación Decimal Universal, la de Dewey o la del Congreso sino también otras especializadas como la National Library of Medicine Classification o el ACM Computing Classification System.

Otros puntos críticos son la inexistencia de mecanismos ágiles que reflejen los cambios de URL o de contenido de los documentos, y también la demora en la inserción de nuevos registros sobre recursos de calidad. Además el número de recursos existente y en continuo crecimiento, dificultan su mantenimiento.

Amat (1999) señala como ventajas que los recursos que contienen han sido previamente seleccionados y que la realización de índices manuales posibilita una contextualización que facilita la recuperación.

Generalmente se admite además, que son recomendables ante búsquedas de información general sobre un tema, y cuando es difícil expresar mediante términos, una determinada búsqueda que tal vez se solucione mejor expandiendo las diferentes categorías y subcategorías de la clasificación.

Los directorios se clasifican, de acuerdo con los recursos que contienen, en: generales, cuando contienen información de diferentes campos temáticos, como es el caso de Yahoo y LookSmart; y especializados, bien de acuerdo con una determinada área geo-

gráfica, ya sea internacional, nacional o local; con un ámbito lingüístico, como es el caso de Olé[118], con recursos en lengua española. Hemos de mencionar además, la creación de directorios especializados que surgen como solución al contenido general que caracteriza a los primeros directorios, que ofrecen recursos de calidad, y que han sido tratados documentalmente para favorecer su recuperación. Es el caso de CORC (Cooperative Online Resource Catalog)[119], BUBL[120], NetFirst[121], SOSIG[122] y CrossROADS[123], alguno de ellos fruto del Proyecto DESIRE[124], promovido por la UE ó INFOMINE[125], de contenido académico.

Grupos profesionales, principalmente del campo de las bibliotecas están elaborando directorios con recursos de especial interés, como es el caso de Librarians' Index to the Internet[126], WWW Virtual Library[127] y otros.

También existen Directorios de directorios como es el caso de Buscopio[128].

Desde el punto de vista de la evaluación de estas herramientas de búsqueda es necesaria más investigación, dado que apenas existen trabajos de investigación que utilicen una metodología que sirva de referencia, a partir de la cuál poder elaborar otros trabajos que la desarrollen.

2.3.2.2. Motores de búsqueda

Cordón García y otros (1999) definen el motor de búsqueda como:

> "herramienta Web que localiza de forma rápida información existente en Internet y que está formado por tres elementos bien diferenciados: una interface (página Web a la que accede el usuario y en la que realiza la búsqueda), un robot (programa que reco-

118 Adquirido por Terra en 2002 y accesible a través de dicho portal. <http://www.terra.es/ole.cfm>

119 Proyecto liderado por OCLC (Online Computer Library Center), en el que intervienen bibliotecas de todo el mundo, accesible en la siguiente dirección Web: http://www.oclc.org/oclc/research/projects/corc/index.htm

120 http://bubl.ac.uk/link

121 http://www.oclc.org/oclc/netfirst.htm

122 http://www.sosig.ac.uk

123 http://www.ukoln.ac.uk/metadata/roads/crossroads

124 http://www.desire.org

125 Elaborado y mantenido por la Universidad de California, accesible en http://infomine.ucr.edu

126 http://lii.org

127 http://vlib.org/

128 http://www.buscopio.es

> rre la Web analizando páginas Web) y una base de datos (índice de palabras, frases y datos asociados con la dirección URL de las páginas Web [...]".

Amat (1999) indica que estos sistemas, frente a los anteriores, proporcionan una mayor cobertura, una mayor exhaustividad en la indización y representación de los documentos y por último un alto grado de especificidad en la indización y actualización, aunque sobre este último aspecto, señala que los ciclos de actualización se alargan indeseablemente.

Según su contenido, podemos distinguir entre motores de carácter general, que indizan recursos de toda la Web, y motores especializados, que no sólo se ocupan de formar sus bases de datos con recursos que tratan sobre un tema determinado, sino que también pueden estar programados para indizar recursos de determinados sitios web.

Desde el punto de vista geográfico, se distingue entre los de carácter nacional e internacional. Así, por ejemplo, para el ámbito francés, se cuenta con Ecila[129], Echo[130], Lokace[131], Francité[132].

Vaquero (1997) los clasifica, según la información que recuperan en: Generales, de Servicios (realizan búsquedas en determinados servicios de la red, subdividiéndose a su vez en Buscadores de Software o de Direcciones), y Temáticos. Según el acceso distingue entre Libres, Privados y Limitados. Utiliza además otra clasificación que contempla el punto de vista de su adquisición para utilizar en servidores Web, distinguiendo entre Inadquiribles, Shareware, y Comerciales.

Los de carácter general más utilizados son Google y Yahoo a los que les siguen otros como MSN, Teoma, WiseNut, AltaVista, etcétera.

Por otro lado, se denomina buscadores híbridos a los que utilizan en la búsqueda bases de datos de otros motores.

---

[129] http://www.ecila.fr
[130] http://francetelecom.com/en/
[131] http://www.lokace.com
[132] http://www.i3d.qc.ca/

Los hay especializados en buscar personas y empresas como FOUR11, Internet Address Finder y Who is who on line, aunque hay que advertir que están especializados en determinados ámbitos geográficos, generalmente relacionados con EEUU, si bien cada vez resulta más fácil acceder a información de un mayor número de países. Otros buscaban en las News y listas de correo, como Deja News[133] y Liszt[134]. Para recursos especializados en revistas y prensa se puede utilizar MediaUK[135], para imágenes y música, Lyrics.com[136], y para imágenes Picons[137]. Finalmente, para software, Tucows[138], Softonic[139], etcétera.

Para búsquedas de información de carácter científico es recomendable utilizar el buscador Scirus[140] de Elsevier o la versión beta de Google académico[141].

Respecto al modo en que realizan las consultas, caracteriza a estas herramientas el que la ecuación de búsqueda se lanza sobre un índice que contiene los términos extraídos de los sitios o páginas web visitadas por el robot. Por tanto, en la mayoría de los casos, de no utilizar la búsqueda por campo, lo que se está buscando son documentos que contengan unos determinados términos, que es distinto a una búsqueda por materias, que requeriría la existencia de tesauros o listados que facilitaran tanto la indización como la propia búsqueda, y que podríamos relacionar con los sistemas de recuperación basados en el uso de metadatos.

Este aspecto plantea a su vez nuevos problemas que están tratando de resolverse, como son los relacionados con los costes, ya que requieren, en la mayoría de casos, la intervención humana en la selección de los metadatos. Iniciativas como Dublin Core proponen que sea el autor o responsable del recurso quien se ocupe de asignar los diferentes metadatos, información que puede ser utilizada en la elaboración de registros de mayor precisión como los que se están elaborando en las bases de datos especializadas. Su utili-

---

133 Permaneció activo hasta el año 2000 en que fue adquirido por eBay y sus archivos por Google. <http://www.dejanews.com>
134 http://www.liszt.com
135 http://www.mediaUK.com
136 http://www.lyrics.com
137 http://www.cs.indiana.edu/picons/searh.html
138 http://www.tucows.com
139 http://www.softonic.com
140 http://www.scirus.com
141 http://scholar.google.com.mx/

zación sigue siendo muy limitada ya que estudios sobre su uso señalan que un 34% de las páginas principales contienen metadatos en sus etiquetas básicas (keywords y description) y tan sólo el 0,3% utiliza el sistema Dublin Core (Lawrence y Giles).

A pesar de todo, y conscientes de la importancia para una recuperación efectiva de la utilización de estos sistemas, los motores de búsqueda han de estar preparados para hacer posible las búsquedas mediante el uso de metadatos. En este sentido, Excite, HotBot, Lycos y WebCrawler aceptaban "etiquetas Meta" en la indización de páginas web, y HotBot las utilizaba en el cálculo de la relevancia (Amat 1999b). No obstante, el uso indebido de metadatos para favorecer la recuperación de determinados recursos ha ido frenando estas iniciativas.

En cualquier caso, Amat (1999b) señala la emergencia de un tercer modelo, basado en la utilización de metadatos, que combina las arquitecturas distribuidas y aprovecha las descripciones normalizadas.

A medida que la Web ha ido creciendo, las necesidades técnicas de los motores de búsqueda han sido mayores tanto en personal como en software y hardware, convirtiéndose en empresas de gran capital, como es el caso de Google y Yahoo. En algunos casos, para su financiación, recurren a la publicidad o al establecimiento de convenios con diferentes compañías de Internet, lo que les ha permitido continuar con su proceso de desarrollo. Autores como Benito Amat (1999) interpretan esta tendencia como un paso de estos servicios hacia la comercialización. Este ha sido el camino que algunos motores, como es el caso de AltaVista, tuvieron que seguir ante los problemas de financiación que suponían las constantes mejoras tanto de software como de hardware. A pesar de todo, fue absorbido por Yahoo. Otra solución consiste en establecer contratos con empresas o particulares para dirigir preferentemente los robots a sus sitios web y ser incluidos de forma rápida en sus bases de datos además de visitarlos de forma periódica y hacer aparecer sus páginas o recursos en puestos altos de los resultados de búsquedas. En este sentido, el metabuscador Excite facilita servicios de pago a las empresas no sólo para que sus productos puedan ser localizados en no más de 72 horas en su propia base de datos, sino también en la de otros buscadores y metabuscadores como Dogpile, MetaCrawler, WebCrawler, Verizon y NBC.

2.3.2.3. Metabuscadores

Ante el distinto funcionamiento y las limitaciones que supone para los motores ofrecer una cobertura total de la documentación existente en la Web, los metabuscadores aportan, entre otros aspectos, una mayor exhaustividad. Para ello utilizan las bases de datos de diferentes motores de búsqueda.

En primer lugar debemos distinguir entre metabuscadores[142] y multibuscadores, ya que a veces ambos términos se utilizan de forma indistinta. Los primeros surgen entre 1995 y 1996 como es el caso de Mamma[143], Dogpile[144] y MetaCrawler[145]. Se caracterizan porque la búsqueda se lanza de forma simultánea sobre distintos buscadores, siendo recomendable utilizar operadores comunes o no usarlos. En los multibuscadores, la búsqueda se hace de forma secuencial, es decir un buscador tras otro y los resultados aparecen separados por motores.

Algunos metabuscadores permiten seleccionar los buscadores sobre los que se lanzará la búsqueda. MetaCrawler reformula la pregunta, de forma que pueda ser procesada por cada buscador y presenta los resultados de una forma unificada. Para la ordenación utiliza su propio algoritmo[146] que valora el ranking asignado por el buscador, el número de buscadores que contiene el recurso, etcétera.

Los metabuscadores ofrecen como ventajas y la realización de búsquedas en múltiples buscadores mediante una única interface y la eliminación de duplicados. Sin embargo, las búsquedas complejas no siempre se realizan como el usuario las plantea sino en función de las capacidades que el metabuscador tenga para lanzar la búsqueda de forma simultánea sobre los motores que utiliza, ya que no todos soportan la misma sintaxis. Además de esta simplificación, se reducen las opciones para mejorar las búsquedas que presentan los buscadores de forma independiente, limitándose también la capacidad de interacción pueden ser completas ya que todos lo motores no soportan una misma sintaxis. A pesar de ello, para autores como Chignell y otros (1999), resuelven el problema

[142]Chowdury (2001) los considera como un subgrupo de los motores de búsqueda.
[143]http://www.mamma.com
[144]http://www.dogpile.com
[145]Lanzaba las búsquedas sobre Galaxy, InfoSeek, Lycos, WebCrawler y Yahoo. Posteriormente añadió OpenText. Actualmente pertenece a InfoSpace. Accesible en http://www.metacrawler.com
[146]Algoritmo denominado “Normalize-Distribute-Sum”

de la baja exhaustividad que se observa en los buscadores, apoyándose en la idea de que ningún motor de búsqueda por si sólo ofrece más del 45% de recursos relevantes existentes en la Web.

No obstante, Chignell y otros (1999) señalan la existencia de una segunda generación de metabuscadores, que tienen en cuenta tanto el tipo de pregunta como la materia, la estrategia de búsqueda, etcétera para seleccionar los buscadores a los que enviar la búsqueda. De aquí nuestro interés por compararlos con los motores de búsqueda para ver hasta qué punto son útiles en búsquedas simples y complejas sobre temas especializados.

En este sentido, Dreilinger y Howe (1997) señalan tres componentes fundamentales:

- Un mecanismo de envío que se ocupa de seleccionar el servidor y el motor o motores a los que se lanza la consulta.
- Agentes del interfaz: se ocupan de interactuar con cada motor de búsqueda y de reformular la consulta de forma que sea correctamente interpretada por cada motor de búsqueda. Finalmente interpretan los resultados que proceden de cada motor.
- Mecanismo de presentación de resultados: se ocupa de integrar los resultados procedentes de los distintos motores, ordenarlos, y en su caso, eliminar duplicados y verificar los enlaces.

Existen diversas clases de metabuscadores, ya que unos lanzan las búsquedas de forma automática sobre los buscadores, mientras que en otros casos como All-in-one[147] es necesaria la intervención del usuario. No obstante, existen casos como ProFusion[148] que integra ambos métodos. Otra diferencia puede ser el que realicen su función en un

[147]http://www.alonesearch.com/multibib/
[148]http://www.profusion.com

servidor propio o utilicen el cliente, como es el caso de Copernic[149], WebSeeker[150] y otros.

Entre los más utilizados, podemos mencionar MetaCrawler[151], Dogpile[152] y Pro-Fusion. Vivísimo[153], e Ixquick[154] pueden considerarse como más actuales y de una tecnología más avanzada. MetaCrawler se caracteriza por la alta especificidad en la expresión de búsqueda, que permite el uso de la lógica booleana, la aplicación de algoritmos de ranking y agrupamiento, o la extracción de términos para especificar la búsqueda. Dogpile por ordenar por relevancia los resultados.

Por otro lado podemos citar SurfWax[155] y Copernic[156] que recuperan incluso en la web oculta. El segundo cuenta con dos versiones, una libre y otra de pago; EZ2Find[157] que incorpora la opción de búsquedas por categorías especializadas; y Fazzle[158] que permite ordenar los resultados por diferentes conceptos y utilizar operadores booleanos, con lo cual podemos observar cómo algunas de las limitaciones que se apuntan para estas herramientas, comienzan a ser superadas.

Hay que señalar además el metabuscador NECI, desarrollado inicialmente en el Instituto de Investigación NEC, y que en la actualidad se denomina Inquirus[159] que trata de mejorar los resultados ofrecidos por otros metabuscadores en cuanto a precisión y eficiencia, mediante el análisis de los documentos, la detección de duplicados, mejora del algoritmo de ranking y la precisión, analizando tanto la información más próxima como facilitando la elaboración de expresiones de búsqueda más precisas y finalmente, destacando la presencia de los términos de búsqueda en los recursos recuperados. (Lawrence y Giles, 1998d).

---

[149]http://www.copernic.com
[150]http://www.bluesquirrel.com/products/webseeker/
[151] http://www.metacrawler.com
[152]http://www.dogpile.com
[153]http://vivisimo.com
[154]http://www.Ixquick.com
[155]http://www.surfwax.com
[156]http://www.copernic.com
[157]http://www.ez2www.com
[158]http://www.fazzle.com
[159] http://www.inquirus.com

Westerra (1997) señala que por sus características, los metabuscadores pueden ser las herramientas del futuro una vez superen aspectos como son: una mejor interpretación de los términos de búsqueda, la posibilidad de poder hacer las búsquedas en los buscadores de acuerdo con la sintaxis que cada uno soporta y la provisión de resultados completos y ordenados.

En la actualidad se están observando los primeros pasos en este sentido, ya que como señalan Chignell y otros (1999), se han venido desarrollando nuevos metabuscadores que discriminan entre diferentes buscadores en función del tipo de búsqueda, del tipo de usuario e incluso de la materia de búsqueda[160]. Otras veces, un programa intermedio analiza las consultas y las transforma de manera que sean "entendibles" para cada uno de los motores (Hu, W. Chen y Yeh, Jyh-Haw 2002).

Otras iniciativas como la propuesta por Gravano y otros (1997) tratan de establecer un protocolo, al que denomina STARTS, cuya utilización por parte de los motores facilitaría una mayor homogeneidad en los resultados. No obstante, como indican Hu, W. Chen y Yeh, Jyh-Haw (2002), este protocolo apenas se está utilizando.

En cualquier caso se trata de herramientas a tener en cuenta ya que como señala Notess, el metabuscador Excite permite aumentar, de forma simple, el resultado de las búsquedas en torno al 50%.

2.3.2.4. Los agentes inteligentes

Los agentes inteligentes de búsqueda como WiseWire o Nano Espacio Custom Search[161], pueden ser otra de las tendencias en recuperación de la información web en los próximos años. Básicamente podemos decir que son programas que se ejecutan directamente en los ordenadores personales y que realizan periódicamente y de forma automática, las funciones para las que han sido programados. Se denominan inteligentes porque "aprenden" a través de su uso, lo que les permite aplicar la información que van acumulando para realizar mejor trabajos repetitivos. Así, basta con que un usuario utilice estos agentes para la recuperación de información una vez, para que posteriormente sea el pro-

[160]Para más información sobre metabuscadores que dirigen las búsquedas hacia motores especializados en los temas de consulta, véase el trabajo de A. Sugiura y O. Etzioni (2000).
[161] Se caracteriza por utilizar algoritmos de inteligencia artificial. Accesible en http:/www.necuse.com

pio programa, a través del perfil de usuario que va generando, quién se ocupe de buscar la información que le interesa y proporcionarla con la frecuencia con la que el usuario la demande[162].

Entre los más conocidos se encuentran Copernic[163], BullsEye[164] y EZSearch[165].

2.3.3. Principales componentes y funcionamiento

Los buscadores que se utilizan habitualmente en la Web son sistemas de recuperación que llevan a cabo miles de búsquedas al mismo tiempo, en bases de datos amplias y variadas, compuestas por diferentes tipos de documentos y con una arquitectura que se caracteriza por su complejidad, potencia y robustez. Cabe mencionar en este sentido los *clusters* o grupo de ordenadores interconectados que facilitan la rápida consulta de los índices, enviando los resultados a un ordenador encargado de ordenar los registros facilitados por los diferentes ordenadores, para finalmente facilitarlos al usuario.

La complejidad de estos procesos hace necesario que nos detengamos a analizar el funcionamiento de estas herramientas, de sus componentes más importantes, ya que conociendo con mayor profundidad estos aspectos, estaremos en mejor disposición para establecer una serie de criterios enfocados a valorar si efectivamente desarrollan su cometido de forma correcta y útil.

En este sentido, podemos preguntarnos ¿Qué se espera de estas herramientas en las búsquedas de documentos? Desde el punto de vista de la Ingeniería del Software, Ges-Chen Hu y otros (2001a) señalan, en primer lugar, la efectividad y eficiencia del sistema en la localización y ordenación de documentos, teniendo en cuenta además, la cobertura, la actualidad, la imparcialidad en el acceso a los documentos, la expresividad y utilidad de los resultados de búsqueda, la facilidad de uso de la interface y finalmente, la adaptabilidad del sistema a las consultas del usuario.

---

[162] Más información sobre las posibilidades de recuperación de información de estos mecanismos puede consultarse en el trabajo de Berrocal, Figuerola, Zazo y Rodríguez (2003).
[163] http://www.copernic.com
[164] http://www.inteliseek.com
[165] http://www.americansys.com

Un punto de vista distinto supone conocer las funciones que llevan a cabo. Así, Gordón, M. y Pathak, P. (1999) señalan que los motores de búsqueda facilitan tres aspectos: reúnen un conjunto de páginas web sobre las cuales el buscador puede recuperar la información; representan dichas páginas de modo que se permita conocer su contenido; y por último, permiten plantear búsquedas que, mediante algoritmos de recuperación, intentan recuperar información relevante. Estas funciones son llevadas a cabo por sus componentes principales, que como todo sistema de recuperación de la información, está compuesto por la base de datos, por el software que se ocupa de su formación y gestión, y que facilita la recuperación, y por la interfaz de búsqueda. De cara a facilitar la consulta para todo tipo de usuarios, estas herramientas ofrecen cada vez interfaces más simples, y suelen ser evaluados de forma específica.

2.3.3.1. El robot o crawler

En relación con el software de formación y gestión de la base de datos, uno de los elementos más importantes son las denominadas "arañas", robots o crawlers. Básicamente se trata de programas informáticos que se conectan a los servidores web, llevando a cabo una serie de instrucciones relacionadas con el modo en que han sido programadas para, por ejemplo, escanear su contenido y ser posteriormente indizado. Su otra función importante, consiste en la elaboración de un listado de direcciones URL extraídas de los recursos que visita, para examinarlas posteriormente. Este listado contiene además, las direcciones URL enviadas a los buscadores por los responsables o creadores de páginas o sitios web, para que sus páginas sean indizadas. También recurren a servidores y páginas web que recogen novedades, los sitios más visitados, grupos de noticias y listas de distribución.

Otros programas (Lynx, Java.net, Comprehensive Perl Archive Network) les permiten llevar a cabo sus funciones independientemente del tipo de servidor, del tipo de recurso de red, o de la aplicación con que se encuentren.

Google, en este sentido, utiliza un servidor de direcciones URL del que parten los robots para visitar las páginas, enviando posteriormente la información extraída a un servidor de almacenamiento, donde se comprime y se le asigna un código identificativo. Esto facilita su recuperación y consulta, bien en línea, o en la versión guardada en memoria.

Una vez que el robot se dirige y visita el recurso, comprueba si ya lo ha visitado anteriormente, y en caso afirmativo, revisa si ha habido cambios, y si ha sido así, actualiza los datos que tiene de dicho recurso.

El orden de las direcciones en el listado de lugares a visitar puede verse influido por las directrices del motor de búsqueda. Entraría aquí en funcionamiento la programación prevista para los recursos de pago, que requieren aparecer de forma rápida y mejor colocados en el ranking de resultados.

El modo de trabajo de estos robots puede hacer posible que en una búsqueda encontremos las páginas enviadas por los creadores, las páginas más consultadas pero, a pesar de todo, hay una parte importante de la Web que les resulta inaccesible, al menos de forma inmediata, bien sea por no haber sido localizadas por el robot o por no haber sido enviadas por sus creadores para su indización. Aún las que se envían sin mediar un contrato económico para ser visitadas por las arañas, pueden tardar de seis a ocho semanas en ser indexadas. Otras limitaciones al trabajo de estos programas pueden venir impuestas por los mantenedores de los servidores, bien a través de los ficheros robot.txt, que controlan y, en su caso, impiden el acceso a determinados recursos a los robots, o su indicación en una etiqueta Meta[166] de la página en HTML. Otros aspectos técnicos que pueden limitar su cometido son: la existencia de marcos en las páginas web, la generación de páginas dinámicas, de mapas de imágenes y de páginas protegidas (Baeza-Yates y otros 1999). Todo ello da lugar a la denominada "Web oculta".

Continuando con el modo de trabajo de estos programas, Baeza-Yates y otros (1999), señalan que los más rápidos superan los diez millones de visitas diarias a páginas Web.

Pero no todos funcionan del mismo modo ya que pueden estar programados para dirigirse a determinados sitios web y una vez allí, extraer sólo la información de una parte determinada o de todo el sitio o página web, es decir, su programación les permite rea-

[166] Véase Koster (1994)

lizar una extracción intensiva, con mayor profundidad en las páginas, o bien extensiva, tratando de recolectar algo de información de la totalidad de páginas del sitio[167].

Los programas más avanzados permiten centrar sus visitas en sitios que contienen determinadas palabras clave o en páginas de una cierta importancia, aunque no todas las herramientas los implementan.

Todo esto se refleja en la indización, que puede tener un componente en unos casos selectivo, en otros jerárquico, dando más importancia a la indización de las páginas de los niveles superiores, mientras que en otros se realiza con mayor detenimiento sobre los documentos situados en niveles inferiores.

Así pues, del trabajo del robot consideramos interesante para la evaluación conocer hasta qué punto desciende en la jerarquía de los servidores y sitios web, es decir, si lo hace con mayor o menor profundidad.

2.3.3.2. El índice

Para facilitar la consulta a la base de datos, hemos de destacar la labor realizada por el *programa de indización*, que realiza una extracción de términos del documento para formar un índice invertido, esto es, un listado de raíces de palabras, de términos, de frases, etcétera, que apuntan a los documentos que los contienen. Como hemos visto anteriormente al hablar de la indización, se caracterizan por la eliminación de palabras vacías y por la aplicación de una serie de valoraciones basadas en frecuencias de términos, calculando el número de veces que aparece un término en un documento o en la base de datos, el lugar en que aparece tanto en la frase como en el texto, etcétera. En determinados casos también pueden hacer intervenir el análisis de hiperenlaces.

Otros aspectos que varían en función del motor tienen que ver con la conversión de los términos a mayúsculas y minúsculas, utilización de las etiquetas HTML en la indización para indizar títulos, direcciones URL, etcétera.

[167] La investigación en este campo se está centrando en la actualidad en la aplicación de técnicas basadas en el uso de la inteligencia artificial. En este sentido puede consultarse el trabajo de Nick, Z. Z. y Themis, P. (2001).

La obtención de resultados en una búsqueda se ve facilitada y condicionada, entre otros aspectos, por el modo en que se ha realizado la indización. A pesar de su importancia, contamos con poca información sobre el modo en que indizan o el ritmo al que lo hacen. En un estudio de Schwartz (1998) se señala que AltaVista indizaba unos seis millones de recursos diarios.

Además, en determinados motores comerciales es frecuente la creación de un segundo índice, que apunta a recursos de pago y que se ofrecen en lugares destacados del listado de recursos recuperados.

Hernández (1999b) denomina analizador o indexador al programa que facilita la extracción de las palabras del contenido, y señala la existencia de diferentes filtros que dirigen el modo en que se ha de realizar dicha indización. El fichero inverso que se genera es gestionado por una base de datos que soporta las consultas del usuario. El texto se extrae generalmente del título, de los títulos de los principales apartados, de las etiquetas Meta, de los hiperenlaces y, según el buscador, de una parte variable del texto. El problema de este tipo de extracción es que no todos los buscadores utilizan los programas que permiten identificar el título, autor, etcétera, del resto de contenido, lo que influye de forma negativa en la precisión de la recuperación. De aquí que sólo una información correctamente etiquetada, puede permitir obtener resultados de mayor precisión. Por ello hemos querido valorar cómo recuperan actualmente los principales buscadores cuando se les indica que busquen determinadas palabras en el título. Es por eso que en la evaluación hemos utilizado una búsqueda por campo, que nos puede dar datos sobre su funcionamiento.

Cada buscador realiza la indización de un modo propio, por ejemplo Google indiza los documentos que va almacenando creando registros que contienen las palabras, información sobre su posición en el texto, sobre el tamaño de la letra y el uso de mayúsculas. Realiza además un análisis de enlaces que se utilizará tanto para alimentar al servidor de direcciones URL como para realizar cálculos de relevancia. El análisis de hiperenlaces ha adquirido un gran desarrollo en los últimos años, especialmente desde la publicación del estudio de Brin y Page (1998) en el que se expone el PageRank model, del que más adelante nos ocuparemos.

El fichero inverso puede actualizarse con una periodicidad variable (de 24 horas a varias semanas), por lo que puede haber cambios no registrados en la base de datos. De aquí que haya casos de información aún no incluida o direcciones que han cambiado o

desaparecido, etcétera. (Olvera 1999c). Debemos tratar de conocer, si no con qué frecuencia los buscadores automáticos actualizan este índice, sí al menos cuál lo hace con mayor asiduidad. Un modo de saberlo es analizando la validez de las direcciones URL que se ofrecen tras una búsqueda, pues en virtud de su mayor o menor índice de conexión, podemos afirmar un valor de actualización.

La incorporación de nuevos términos o datos se lleva a cabo en cada actualización. Google lo hace mensualmente y según Notess (2002), reindiza diariamente los servidores que varían frecuentemente sus contenidos. Su correcto funcionamiento requiere pues, diferentes tareas de mantenimiento. Así, para eliminar páginas, buscadores como WWLib-TNG, desarrollado en la Universidad inglesa de Wolverhampton, utilizan archivos históricos en los que se registran las veces que un determinado enlace no ha podido utilizarse, y a partir de un determinado número de veces, la página se borra de la base de datos. Por supuesto se tiene en cuenta que los problemas no sean debidos al estado de la red.

Este autor demuestra en otros estudios[168] que motores como MSN, HotBot y otros recuperaban documentos colgados en la Web en las últimas 48 horas.

Desde el punto de vista lingüístico, la utilización por algunos motores de mapas conceptuales o la indización de las propias preguntas, son soluciones planteadas para resolver el problema de la polisemia y la homonimia. (Peña y otros. 2002).

Por otro lado, la Web está formada por diferentes tipos de documentos, bien sean textuales, sonoros o audiovisuales sobre los que, en la medida de lo posible, los programas indizadores tratan de extraer información, para facilitar su recuperación. Al mismo tiempo van surgiendo herramientas de búsqueda especializadas en estos tipos de archivos. No obstante, hay que tener en cuenta que no todos los recursos o todo su contenido pueden ser indizados, no sólo por el amplio número existente y la variada tipología, sino por el carácter especial de su contenido, como por ejemplo ocurre con las fórmulas o ciertas representaciones gráficas.

Buscadores como Google, MSN y Yahoo indizan archivos de diferente fuente y tipo documental como son los grupos de noticias (Usenet), documentos PDF, Power-

[168]Véase la página web http://www.notess.com/search/stats/freshness.shtml

Point, etcétera, aspecto que nos parece de especial interés ya que gran parte de la información y documentación científica, utilizan estos formatos electrónicos.

Por otro lado, Lawrence y Giles (1999) han observado la existencia de limitaciones de capacidad en estos índices que superan los cien millones de páginas, lo que explica las variaciones periódicas de su tamaño. Para Sonnenreich (1998) la capacidad de los motores de indización de estas herramientas está limitada a dos mil cien millones de páginas. Hay que mencionar también la limitación que algunos de estos motores mantienen con respecto al tamaño de los documentos ya que como señala Price (2001), en el caso de Google, no se indizan documentos superiores a 110 Kilobytes y 100 Kilobytes en AltaVista. AllTheWeb no presenta límites en este sentido. No obstante, estas cifras hay que aceptarlas con precaución, ya que las limitaciones señaladas parecen ir quedando atrás, al menos en el caso de los grandes sistemas de búsqueda de la Web.

Otro aspecto que afecta a la indización, y en general al funcionamiento de los buscadores, ha sido estudiado por Bar-Ilan (1998/1999), quien se ocupa de analizar las variaciones temporales de estos índices o estabilidad, llegando a la conclusión de que éstos varían frecuentemente y que esta práctica permite a las herramientas de búsqueda añadir nuevos contenidos a los índices sin necesidad de ampliar su capacidad. Los buscadores evaluados fueron AltaVista, Excite, Hotbot, Lycos y Northern Ligh. El buscador con mayor variación fue Excite, y Northern Light el más estable.

A pesar de su importancia en la Web y fundamentalmente en estas herramientas, Olvera (1999b) señala la existencia de diferentes detractores de la indización automática (Desai, 1997; Lynch, 1997), por su carácter simplista, poco selectiva y que ni tiene en cuenta el contexto ni el carácter más o menos científico del recurso. Se refieren además al funcionamiento simple de los robots, a la incipiente utilización de técnicas basadas en el lenguaje natural[169] para la recuperación, a la mayor facilidad en el reconocimiento de

[169]Lancaster (1995) señala que "[...] *lenguaje natural* no significa otra cosa que el lenguaje del discurso común [...]" y que en estos sistemas "[...] la materia de los documentos y de las necesidades de información está representada por un vocabulario ilimitado de palabras y frases utilizadas habitualmente en el campo temático [...]", y por tanto "[...] no tienen controlado el vocabulario[...]". Más adelante añade que "[...] un sistema con Lenguaje Natural puede estar basado en la indización humana, la indización automática o no existir indización. La indización humana puede extraer o no términos o frases del texto, mientras que la indización automática es siempre una indización por extracción [...], si bien, como veremos, hay sistemas que permiten una "indización inteligente" mediante términos adicionales que no aparecen en el texto.

recursos de tipo textual frente a otros como imágenes y documentos multimedia, etcétera, lo que contrasta con las ventajas de la indización profesional.

Lawrence y Gilles (1999) han observado en los motores de búsqueda comerciales una cierta inclinación por indexar información de interés general frente a información de carácter especializado, ya que para localizar este tipo de información se utilizan los buscadores especializados[170] o los elaborados por las bibliotecas, tales como Guías de recursos, Bases de datos, o la integración de recursos web en los catálogos. Buscadores temáticos y directorios especializados son otras soluciones que se proponen (Olvera 1999b) en la recuperación de información especializada.

En general debemos admitir que este tipo de indización se caracteriza por una alta exhaustividad, que influye en una recuperación numerosa de documentos, lo que puede ir en detrimento de la precisión. Este aspecto se amplifica extraordinariamente en un contexto tan amplio en documentación como es la Web. Como hemos visto, para solucionar este problema los desarrolladores de estas herramientas tratan de aplicar diversas fórmulas que por un lado permitan realizar búsquedas más precisas, y por otro ordenar en los primeros lugares de resultados los recursos que mayor relación tengan con los términos expresados en la ecuación de búsqueda. De aquí que dada la importancia de la indización, en la evaluación nos ocupemos de valorar los aspectos que creemos más destacados y que más preocupan a la comunidad científica en la realización de los índices. Para ello utilizamos diferentes tipos de búsqueda, valoraremos además el uso de la metainformación en la indización, la indización de diferentes tipos de documentos, la existencia, actualización y consistencia.

2.3.3.3. La base de datos

Las bases de datos de las herramientas de búsqueda no son idénticas, ya que recogen, en cada caso, la información requerida por cada uno de los sistemas, como se observa en los registros que se facilitan en la recuperación. Dichos registros suelen contener una descripción básica de los recursos, ya sean sitios o páginas web, incluido el URL, el título de la página a la que se refieren, que actúa como elemento de enlace hacia el recur-

[170]Más información sobre estos buscadores puede consultarse en el artículo de King, D. (2000).

so y un pequeño resumen o extracto del documento, generalmente con los términos de búsqueda de forma destacada.

Los motores de búsqueda contienen varias bases de datos, pudiendo facilitar, en virtud de la firma de convenios con otros buscadores, el acceso a otras bases de datos.

Como hemos señalado con anterioridad, sus bases de datos pueden duplicarse dando lugar a mirrors o espejos, aunque como señala Notess (1999), pueden no ser idénticas, y en su caso, contener un mayor número de recursos pertenecientes a un determinado país, idioma, etcétera.

Diversos autores se han ocupado de calcular la parte de la Web indizada por cada motor de forma individual o globalmente. Lawrence y Giles se refieren a que una gran cobertura no implica que se trate de un buen motor, pues hay que analizar cómo son los registros que recupera, si hay muchos duplicados, enlaces no activos, etcétera. De aquí que una parte importante de los estudios se ocupe de valorar la formación de la base de datos. Como hemos visto, la captación de recursos se realiza, principalmente, de forma automática, dependiendo de la programación de los "crawlers" o robots. Pero hay otros métodos como el utilizado por Teoma, que declara formar su base de datos mediante la colaboración de expertos y entusiastas que clasifican y añaden comentarios a los recursos.

Johnson y otros (2001) señalan además que dentro de la cobertura ha de hacerse referencia al tamaño, a la frecuencia de actualización de la base de datos y a la frecuencia por parte del robot en visitar las páginas que cambian regularmente, a la indización de frames, de imágenes, al tiempo de indización y al uso de técnicas como la popularidad, a través del análisis de enlaces. De todos estos aspectos, los que ofrecen mayor dificultad son los que tratan de medir el tiempo de indización, es decir el que transcurre desde que un buscador dispone de un recurso hasta que es indizado, y la frecuencia de visita a sitios que cambian continuamente, aunque estas herramientas van incorporando tecnologías que miden estos aspectos.

Un estudio de Hock (2002) valora el número de páginas HTML que contienen alguno de los buscadores que analizamos. Señala para Google y AllTheWeb dos mil cien millones, mil seiscientos millones para WiseNut, quinientos millones para HotBot y ciento cincuenta millones para Teoma, aunque para Chris Sherman (2002), este último tiene una cobertura de 200 millones de páginas.

En 1999, Google estimaba tener indexados entre setenta y cien millones de páginas. A finales del 2001 dice contener más de tres billones de páginas web, imágenes y mensajes de listas de discusión. En 2005 los cálculos apuntan a 8.1 billones el número de páginas Web.

El motor de búsqueda de MSN contaba con la asistencia de la base de datos Inktomi[171] y posteriormente de Yahoo y LookSmart. Actualmente consta de base de datos propia y contiene unos 110 millones de páginas web.

En 2002, Greg R. Notess hacía las siguientes estimaciones: AllTheWeb contenía más de 2,1 billones de páginas indizadas a texto completo frente a los 2,4 billones de Google. Datos más recientes señalan que Google tiene indexadas más de ocho billones de páginas, Yahoo más de cuatro billones y Teoma más de un billón y medio[172].

Estas cifras coinciden con un estudio de Gulli y Signorini (2005) en el que señalan que Google contiene más de ocho billones de páginas indizadas, MSN cinco, Yahoo alrededor de 4 billones y Ask/Teoma más de dos.

Greg R. Notess valora las diferencias entre las estimaciones que presentan los diferentes buscadores y las que se desprenden de los estudios realizados por especialistas[173], siendo superiores las cifras aportadas por los motores de búsqueda en su afán de convencer al usuario de que su contenido es superior al del resto.

Bharat y Broder (1998) proponen un método para valorar la cobertura y el solapamiento de los motores, pero una estimación de este tipo requiere un gran número de búsquedas, pues en este caso fue necesario lanzar dos tandas de diez mil preguntas y contar con un léxico de cuatrocientas mil entradas. En este sentido, debemos señalar que existen sitios en la Web que proporcionan valoraciones bastante fiables, siendo utilizados a menudo por los investigadores para comparar sus cifras. Es el caso de Search Engine Watch[174].

---

171 Como señala Price (2001) el acceso no es al total de la base de datos de Inktomi, pudiendo ser distinta en cada motor que tiene conciertos con ella.

172 Fuente: Infopeople. Datos ofrecidos en su página web. http://infopeople.org/search/guide.html

173 Véase http://www.searchenginesshowdown.com/stats/sizeest.shtml

174 http://www.searchenginewatch.com

La información que han de ofrecer ha de ser suficientemente amplia y actualizada, de aquí que destaquemos como aspectos a tener en cuenta en la formación de las bases de datos, su cobertura y tamaño, y su actualización.

En nuestro estudio nos centraremos en valorar la cobertura temática que estas bases de datos hacen de temas especializados, para ello nos basaremos en la información facilitada por las herramientas de búsqueda en los listados, aunque somos conscientes de que la información que facilitan no siempre es exacta.

Analizaremos otras características de la información que ofrecen, atendiendo al tipo de documentos que contiene, es decir, si solamente contiene recursos en HTML o si además contiene documentos en PDF, PowerPoint, Word, a la tipología documental y al carácter de la información.

Otro aspecto de gran interés es conocer el solapamiento entre estas herramientas ya que puede ayudarnos a elegir una de ellas o una determinada combinación, cuando se quiera hacer una búsqueda exhaustiva.

2.3.3.4. Los programas de búsqueda y recuperación de la información

Los programas de búsqueda se ponen en funcionamiento al lanzar una búsqueda, tratando de localizar en el índice los términos o la expresión indicada presentando finalmente los recursos obtenidos.

En primer lugar hemos de referirnos a la interfaz, que no sólo proporciona información sobre el funcionamiento, contenido, modo de consulta, posibilidades, opciones, operadores, botones, menús, grupos temáticos, servicios, etcétera, sino que también, y esto es lo fundamental, actúa de intermediario entre el usuario y la base de datos.

Deming (1998) señala que la interfaz de búsqueda debe cumplir los siguientes objetivos:

a) Ayudar al usuario en la búsqueda.

b) Guiar al usuario en la búsqueda.

c) Permitir filtrar los resultados de una búsqueda.

d) Facilitar la navegación.

e) Ser clara.

f) Ofrecer información de interés al usuario.

En virtud de estos objetivos, y en la importancia de su valoración por parte del usuario, la interfaz de estas herramientas suele ser motivo de una evaluación específica.

En este sentido, los buscadores automáticos suelen presentar un formulario con una o varias ventanas para introducir la expresión de búsqueda que se ha de lanzar sobre las bases de datos. Contiene además los mecanismos de interacción que pueden dar mayor precisión a la búsqueda, a los resultados, ayudas, etcétera. Suelen presentar dos opciones de búsqueda: una simple y otra avanzada. Esta última facilita la búsqueda por campo, cierta asistencia en la realización de expresiones de búsqueda complejas (búsqueda booleana, truncamientos, operadores de adyacencia, etcétera), así como el establecimiento de filtros para periodos cronológicos, tipos de documentos, lengua, país, etcétera. Además utilizan un importante número de opciones propias de la Web, como búsquedas en un determinado sitio, dominio, enlace, en el URL, etcétera.

Para el usuario es fundamental conocer el modo en que se puede interrogar a sus bases de datos e interpretar los resultados, pues el éxito de las búsquedas no sólo depende del correcto funcionamiento del buscador, sino también de saberlo consultar de la forma adecuada. Por ello es recomendable conocer las opciones de búsqueda que ofrece cada buscador, ya que suelen variar de unos a otros, siendo en muchos casos necesario acomodarse al sistema[175].

Los buscadores web, las bases de datos en CD-ROM, las accesibles a través de Internet y los OPAC, tienen una serie de características técnicas comunes que, en determinados casos, permiten realizar búsquedas utilizando: palabras clave, lenguaje natural, lógica booleana, operadores de proximidad, búsqueda por frase, diversos tipos de truncamientos, búsqueda por campos, mayúsculas o minúsculas, compartiendo además la posibilidad de realizar las búsquedas de forma simple y avanzada.

Es necesario conocer, de una forma más específica, qué elementos pueden ayudar a hacer una búsqueda más precisa, por lo que por ejemplo, es importante saber deter-

minados aspectos de la indización; cuál es el operador de búsqueda que utiliza cada motor de forma implícita, ya que ello incide muy directamente en los resultados; cuáles son los operadores que soporta, tanto booleanos como de cercanía, si permite truncamientos, si se puede consultar por palabra clave, frase o lenguaje natural, si es sensible al empleo de mayúsculas o minúsculas, así como algunas características de la búsqueda avanzada o por campos. Presentamos a continuación diferentes aspectos que tienen que ver con la recuperación, y especialmente con la ordenación de los resultados de búsqueda.

Los motores de búsqueda utilizan recursos específicos que permiten modificar el funcionamiento automático de las búsquedas. MSN utiliza un "Generador de búsquedas" además de permitir utilizar alguna de las opciones anteriormente señaladas, permite refinar al añadir otro término de búsqueda a los resultados, restringirlos a un sitio, dominio, obtener vínculos, etcétera. Permite ordenar los resultados por fechas o por popularidad. Para calcular la relevancia se sirve de parámetros como la procedencia de la consulta, es decir, calcula este valor en función del lugar desde el que se realiza la búsqueda. Por su parte, los metabuscadores ofrecen, como en el caso de Ixquick, una búsqueda avanzada universal, es decir que analiza la ecuación de búsqueda y, en la medida de lo posible la transforma para que pueda ser entendida por la mayoría de buscadores; una búsqueda global, es decir en todo el mundo y en cualquier idioma; y un refinamiento avanzado, que permite no volver a ver resultados visualizados previamente así como centrar la búsqueda en determinados recursos recuperados y rechazar otros. Además señala con estrellas el número de buscadores que recupera determinado recurso entre los diez primeros. Finalmente, ofrece diferentes opciones personalizables, como la posibilidad de destacar los términos de búsqueda en los resultados, etcétera. Otras características le permiten corregir términos escritos de forma incorrecta y ordenar los resultados por relevancia o por motor de búsqueda. Además ofrece una serie de enlaces relacionados que permiten tanto centrar como filtrar la búsqueda, y un histórico, con las 15 últimas búsquedas lanzadas. Finalmente, como opción a destacar, permite la búsqueda de recursos indicando la fecha de publicación.

---

[175] Véase en este sentido las recomendaciones que hace Greg Notess (2000b) para consultar Google utilizando operadores booleanos.

Ixquick, para enfocar la búsqueda, facilita diversas categorías que recogen los recursos relacionados con aspectos específicos. Se basa en la coincidencia de los términos de búsqueda en los recursos recuperados. Además, se pueden aplicar filtros por dominio, idioma de los recursos y fecha.

Generalmente, se pueden utilizar además de los términos de búsqueda, con los operadores correspondientes, otros elementos como: los signos "más" (+) para forzar la aparición del término al que preceden y "menos" (-) para excluirlo; paréntesis; truncamientos; comillas en las búsquedas por frase y la coma (,) para separar los apellidos del nombre, en el caso de búsquedas de nombres propios.

Como sistemas interactivos que son, la efectividad de la búsqueda también depende de los objetivos del usuario (Arms, 2001) y de su formación. Así, en función de los resultados que obtenga, podrá redirigir o plantear la búsqueda, intentando obtener otros más relevantes.

Es por ello que conocer interfaz de búsqueda y el lenguaje de interrogación del motor es indispensable no sólo para consultar adecuadamente la base de datos, sino también para obtener unos resultados más precisos.

Estos sistemas se basan en el modelo de recuperación booleano extendido y en algunos casos en el vectorial. En el primer caso los operadores pueden ser implícitos, cuando el sistema interpreta que se utilizan los operadores AND o bien OR entre los términos de consulta, y explícitos cuando han de introducirse expresamente por el usuario. La mayoría de motores de búsqueda de la Web utilizan el operador AND de forma implícita.

El modelo vectorial es menos utilizado ya que como señalan Berry y Browne (1999:67), los motores que utilizan técnicas basadas en espacios vectoriales reconocen a los operadores booleanos como palabras vacías en la ventana de búsqueda, permitiendo en algunos casos la búsqueda en campos.

Igualmente es importante valorar en estos sistemas de recuperación la importancia de la representación de la búsqueda, ya que puede basarse en el modo convencional de acumulación de términos, o en otros más evolucionados, como son las búsquedas expresadas mediante el lenguaje natural, que exige la estructura semántica de la frase o la representación vectorial de los términos, a los que se les adjudica un valor según su importancia.

Además la búsqueda en los motores se puede realizar mediante palabras clave, frases, utilizando truncamientos, expresiones del lenguaje natural y otra serie de operadores, además de los booleanos, y de cercanía. Algunos buscadores especializados facilitan la consulta mediante tesauros[176] y la búsqueda por campos.

Pero cada una de estos tipos de búsquedas puede plantear problemas. En este sentido Delgado (2001:46) señala el defectuoso funcionamiento de la búsqueda booleana en buscadores Web al desactivar con su uso la ordenación por relevancia.

Por otro lado, la utilización del lenguaje natural en la consulta puede influir en la recuperación de un mayor número de recursos, debido a los problemas que por ejemplo plantea el uso de términos sinónimos y homónimos. Tal vez por esto y por su incipiente desarrollo, su uso no se ha generalizado y sólo determinados buscadores como AskJeeves lo utilizaban.

Respecto a los resultados de la búsqueda, una vez el motor compara los términos de búsqueda de la ecuación de búsqueda y el índice, presenta un listado de resultados con información sobre cada uno de ellos.

Teoma los ofrece en cuatro apartados: en primer lugar muestra una serie de lugares comerciales; a continuación el apartado “Refine” que contiene recursos que se ofrecen a modo de sugerencia por si son del interés del usuario; en tercer lugar el apartado denominado “Results” que contiene el gran grupo de recursos; y finalmente el apartado “Resources” que son recursos compilados por expertos en la materia.

Coloca en lugares relevantes recursos de empresas que por su carácter comercial quieren aparecer en los primeros lugares de búsqueda. Se distinguen porque se les asigna la etiqueta Sponsored by. Utilizan para ello las bases de datos de Overture, Sprinks y FindWhat.

De aquí que nos hayamos planteado la necesidad de valorar como funcionan los buscadores ante diferentes tipos de búsqueda, cuando se utilizan opciones de búsqueda avanzada como la búsqueda booleana, por frase, por campo, o con en operador de existencia (+).

[176] Véase por ejemplo ERIC (Education Resources Information Centre) en http://www.eric.ed.gov/

Si todo se desarrolla correctamente, constituye la piedra de toque para valorar la utilidad de estas herramientas, en tanto en cuanto sean susceptibles de colocar entre los primeros lugares los recursos más relacionados con la búsqueda y resolver el problema de necesidad de información planteado.

Por tanto, en esta página o páginas, dos aspectos son fundamentales: por un lado la presentación o descripción de los recursos, que se realiza con el fin de ayudar al usuario a decidir si un recurso es de su interés, y por otro la ordenación o ranking.

Respecto al primero de ellos, generalmente suele aparecer el título que puede servir de enlace con el recurso, un resumen descriptivo del contenido, la dirección URL y otra información, que varía en función de cada buscador, como puede ser: el índice de relevancia, el tamaño del archivo, la fecha de creación, la fecha de entrada en la base de datos, lengua o idioma. Junto a estos resultados aparecen, de forma destacada, otra serie de recursos de carácter comercial, cuya frecuencia trataremos de valorar en la evaluación.

Otra característica de los listados es la aparición destacada de recursos dependientes que aparecen colocados de forma más adentrada respecto a los márgenes, que el resto de recursos de los cuales dependen.

El resumen descriptivo presenta, de forma destacada, generalmente en negrita, los términos de búsqueda junto al conjunto de la frase o frases en que aparecen. No hay uniformidad en el contenido entre unos y otros, ya que por ejemplo, Infoseek mostraba los 300 primeros caracteres del cuerpo de la página HTML, Lycos, en función de su brevedad seleccionaba entre las veinte primeras líneas o el 20 % del texto. AltaVista y HotBot utilizaban la etiqueta Meta “description” para extraer el resumen, aunque lo más usual en la actualidad es la extracción de frases en las que aparece el término o términos de búsqueda.

Dada la importancia que esta primera información tiene para el usuario, ya que en función de ella ha de decidir el interés del recurso, en la evaluación trataremos de valorar la utilidad de estos aspectos, analizando la frecuencia de aparición de los términos de búsqueda, la frecuencia de aparición de recursos comerciales y de recursos dependientes.

Respecto a la ordenación, el gran número de recursos que suele aparecer hace que sea un aspecto fundamental, ya que el usuario no suele consultar más allá de las tres primeras páginas de resultados. Esto requiere un buen funcionamiento de los algoritmos que intervienen.

Excite ofrece dos opciones para ordenar los resultados, bien por buscador o por relevancia. Otra opción permite destacar en el contenido de los recursos recuperados, los términos solicitados.

Courtois y Berry (1999) han observado que los resultados se ordenan de acuerdo con el cumplimiento de ciertos criterios como la cantidad de palabras de la expresión de búsqueda localizadas en el documento, dando mayor importancia a los documentos de menor tamaño que las contienen, a la proximidad de los términos y a su ubicación en el documento. Además, aunque la fórmula exacta para la ordenación varía de unos motores a otros, la localización de las palabras, bien en el título o en etiquetas Meta, puede tener más peso que su aparición en el texto. La valoración de un aspecto u otro puede influir en la diferenciación entre buscadores, ya que mientras algunos de ellos excluyen en la indización la metainformación, otros sí la utilizan.

En la evaluación trataremos de precisar qué motores utilizan la metainformación en el ranking así como la frecuencia de aparición de los términos y su peso.

En determinados motores, la frecuencia relativa, es decir, el valor que representa la repetición de una palabra en la base de datos, también incide en la ordenación de los resultados.

Los algoritmos que se aplican suelen ser secretos, si bien, en algunos casos pueden deducirse. Podemos decir que en función de estos algoritmos se realiza el cálculo de la relevancia.

En la mayoría de casos se aplica una ordenación basada en los términos del propio recurso, pero Google utiliza además aspectos externos al documento como son los enlaces. Este buscador ordena los resultados teniendo en cuenta no sólo de la valoración de los aspectos mencionados anteriormente, sino añadiendo además una serie de valores que asigna en función de la existencia de enlaces de otras páginas que apunten hacia ellas, así como la importancia de estas páginas, que a su vez se valora por el número y calidad de enlaces que les señalan. Este algoritmo se denomina PageRank, y aunque en muchos casos ofrece buenos resultados, se ha criticado que favorece e impulsa la recuperación de recursos ya conocidos, en detrimento de los nuevos recursos. (Savoy y Picard, 2001). Así, la colocación de una página en los primeros puestos de resultados requiere cierto tiempo, hasta que el recurso es conocido y a su vez enlazado por otras páginas importantes.

WebQuery es otro algoritmo que de basa en cálculos de enlaces y actúa basándose en un conjunto de páginas que ordena en función de sus enlaces con otras[177].

Finalmente, como señalan Baeza-Yates y otros (1999), Kleinberg diseñó HITS (Hyperlink Induced Topic Search). Este algoritmo es utilizado por el buscador Clever, de IBM, actualmente en proceso experimental y por Teoma. El algoritmo, para facilitar la recuperación de recursos de calidad, distingue entre páginas autoridad, que son las que reciben un gran número de enlaces y páginas concentradoras o páginas eje, que contienen un importante número de enlaces a otras páginas, fundamentalmente de las consideradas "autoridad". A estos valores añade otros cálculos basados en los métodos utilizados normalmente como la frecuencia y la proximidad de términos.

Este buscador utiliza dos tecnologías, por un lado la denominada Subject-Specific Popularity, que detecta comunidades de recursos en la Web en torno a una materia y las coloca en los primeros lugares de las búsquedas. Por otro lado, Dinamic descriptions, basa su técnica en el análisis del contexto de los términos de búsqueda, lo que le permite mostrar resultados que aunque no contengan los términos de búsqueda, sí que tratan sobre el tema objeto de búsqueda. Wisenut aplica el algoritmo context-sensitive ranking que valora tanto los enlaces como el contenido del documento.

El cálculo de la relevancia se presta también a otras valoraciones. En primer lugar se suele valorar la posición de los términos según se encuentren en las etiquetas Meta, en los títulos, bien sea teniendo en cuenta la proximidad entre términos, o por la frecuencia, que actúa de forma inversamente proporcional al valor del término en la base de datos. Se utilizan, además, valores de corrección según el lugar que ocupa el término dentro del texto, siendo mayor su valor si forma parte del título o de la cabecera del documento y cuanto más cerca del inicio del documento se encuentre.

Como señala Notess (1999c) un valor más real es el que se aplica comparando la aparición del término con el número total de palabras que contiene un documento, lo que evita el problema de dar mayor valor a los documentos más grandes. Este autor señala además, como factor a tener en cuenta en la ordenación, la aparición del término de búsqueda en el URL, lo que suele tener en los motores una importancia especial, al igual que

[177]Para más información véase Li, Y. (1998)

ocurre con los términos que actúan como anclas de hiperenlaces hacia otras páginas, que a su vez reciben enlaces de otras. Google utiliza esta técnica valorando además, como acabamos de ver, si las páginas a las que se dirigen los enlaces tienen cierto prestigio. En otros casos, se basa en el número de enlaces que apuntan a un determinado recurso, en las páginas más visitadas, etcétera.

Otras técnicas que se utilizan para mejorar la relevancia se basan en cálculos vectoriales, en las técnicas de agrupamiento o Clustering y en la retroalimentación de la relevancia (Relevance Feedback).

El tema de la ordenación, por su importancia, ha dado lugar a un amplio número de trabajos dedicados al estudio del posicionamiento en los motores de búsqueda[178]. Estos estudios se centran en analizar cómo ordenan los motores los resultados de búsqueda y en base a qué características deciden qué recursos colocar en las primeras posiciones. Los criterios que utilizan están muy relacionados con lo anteriormente señalado.

Por otro lado existe una corriente de estudio centrada en el análisis del ranking. Nosotros proponemos una valoración del ranking en función de la utilización de las etiquetas Meta Key y Description, y de la frecuencia y peso de los términos de búsqueda en las páginas recuperadas, y trataremos de analizar si hay correlación entre estas variables y la ordenación.

Por tanto, a tenor del análisis de las diferentes partes de los buscadores Web, nos interesa analizar la respuesta de estas herramientas ante los distintos tipos de búsquedas, de aquí que cada una de las búsquedas sea diferente (búsqueda por un término, varios términos, con operadores de existencia, booleana, por frase y por campo). Valoraremos así las capacidades de los programas de búsqueda y recuperación que soportan los diferentes buscadores.

Del funcionamiento del robot analizaremos la profundidad de rastreo de los sitios web, fundamentalmente en cuanto a la profundidad de extracción de información de los servidores web, nos ocuparemos de actualización, valorando la existencia de enlaces con error y duplicados. En relación con la base de datos, la estimación de su tamaño en función de la cobertura sobre las búsquedas planteadas, las características de la informa-

[178] Véase por ejemplo el trabajo de L. Codina y M. C. Marcos (2005).

ción recuperada y la recuperación de páginas únicas y solapamiento entre buscadores. Otros aspectos a analizar son la precisión técnica y el ranking.

2.3.3.5. Identificación de los problemas de recuperación de información en la Web

Hemos visto con anterioridad, en el apartado dedicado a las características técnicas y problemas de la información en la Web, alguno de los aspectos que van a incidir de forma negativa en la recuperación de dicha información y a los que los sistemas de recuperación tienen que hacer frente como son el gran número de documentos existentes, su carácter cambiante y efímero, la variedad de formatos en los que aparece, no siempre legibles por todos los sistemas, el gran interés comercial, su escasa descripción, etcétera.

También hemos mencionado al ocuparnos de las partes que constituyen los sistemas de recuperación Web, aquellos que dependen exclusivamente de ellos mismos, como son los problemas relacionados con los índices, con la actualización de las bases de datos, etcétera.

Tramullas y Olvera (2001) mencionan problemas como: la limitada cobertura de los motores, que por ejemplo, en el caso de HotBot no superaba el 32% del total de recursos existentes; la actualización de sus índices, que no se produce de forma automática, siendo más lenta en información menos solicitada, ya que determinados buscadores relegan estos servidores a los últimos lugares del orden de visita, resultando más frecuentemente actualizados los que contienen información que se consulta periódicamente. Los motores no reflejan la variabilidad espacial y temporal de las páginas web, provocando la aparición del error 404, el acceso a las páginas se suele hacer de forma independiente a su contexto informativo, es decir, sin señalar si determinada página pertenece a un documento electrónico mayor, y por último, la programación de los robots, que incide en la limitación para localizar información en los servidores más allá de un determinado nivel de la estructura de la Web.

De los cuatro problemas que Martínez (2002) señala en relación con la recuperación de la información, dos dependen del usuario (la formulación adecuada de la pregunta y la interactividad con la interface de usuario) y otros dos del funcionamiento del motor (inadecuada indización de los documentos y la limitada actualización de los índices del motor).

Hípola y Vargas-Quesada (1999b) han criticado el desfase de estas herramientas y su actualización irregular.

Lawrence y Giles (1998c) señalan como problemas los periodos de inactividad, baja cobertura, interfaz poco consistente, bases de datos anticuadas, pobre ranking de relevancia, baja precisión y dificultades con las técnicas de spam.

A modo de resumen, planteamos a continuación una sistematización de mayor profundidad de los factores externos e internos que intervienen en la deficiente recuperación de la información que presentan estas herramientas.

Entre lo factores externos, podemos señalar como más importantes: la cantidad de información existente en la Web, el limitado acceso a la información existente que estas herramientas proporciona, como en el caso de la Web Invisible, el carácter dinámico de la información en la Web, y la falta de estructuración de los recursos.

Un trabajo de Lawrence y Giles (1999) indicaba respecto al tamaño de la Web, que en julio de ese año estaban a disposición del público 800 millones de páginas alojadas en tres millones de servidores, y que los seis principales motores de búsqueda cubrían un 60% de los recursos, siendo Northern Light el más destacado con el 16%. El gabinete de estudios americano, Cyveillance cifraba entonces en dos billones, el número de documentos accesibles, aumentando la cifra en siete millones diarios. Señalan, además, que Google cuenta con la base de datos más completa estimada en un billón de páginas. Pero estos datos han quedado anticuados en muy poco tiempo pues un estudio más actual de Gulli y Signorini (2005) estima en 11,5 billones, el número de páginas indizadas y que por tanto forman parte de la Web visible.

Tal cantidad de información dificulta y hace prácticamente imposible su total indización ya que exige un gran despliegue de medios, y más si tenemos en cuenta el importante número de consultas diarias que soportan.

Este aspecto incide en la limitación de acceso al total de los recursos existentes en la Web y exige tanto una selección de recursos como una especialización por materias.

Respecto al segundo punto, es decir el que se refiere a la Web Invisible, es un hecho que los motores de búsqueda no pueden controlar toda la información de la Web. Existe además una gran cantidad de recursos de difícil acceso para los buscadores automáticos, que es lo que constituye la Web Invisible.

La Web invisible está formada principalmente por todos aquellos recursos que requieren la identificación del usuario para su acceso, como es el caso de las Intranets; toda la información alojada en servidores que impiden el acceso a los robots de búsqueda; la información alojada en marcos o frames y las páginas dinámicas, como por ejemplo las que se generan al extraer información de una base de datos[179], un diccionario interactivo, etcétera, que generan diferentes direcciones URL y que suelen alojarse temporalmente en el disco duro del usuario, impidiendo, de este modo, su indización por los motores.

Como ya hemos visto al tratar del trabajo de los Crawlers o arañas, determinados servidores Web, que no desean que los motores indicen sus páginas, utilizan tanto un protocolo de exclusión, que consiste en un documento denominado robots.txt, que se ubica en los servidores y contiene instrucciones sobre los ficheros a los que los robots tienen o no acceso, como una etiqueta Meta, con indicaciones dirigidas a los robots[180].

Otra información de interés que escapa a los buscadores es la contenida en servidores que ofrecen noticias de actualidad que requieren una constante actualización.

Finalmente hay que considerar en este apartado a los sitios y páginas que los robots no indizan en un sitio web, normalmente porque supera la capacidad para la que han sido programados, y las páginas o sitios hacia los que no apunta ningún enlace.

Un estudio de Lyman y Varian, citado por Tramullas (2002:602) recoge cifras referidas al año 2000 en las que muestra que la web visible podría contener unos 2,5 billones de páginas[181] y la web oculta quinientos cincuenta millones. También son intere-

---

[179] Por ejemplo las bases de datos de Access, Oracle, SQL y otras. Sin embargo, como señala Salazar (2005:139) "Google ya está llegando a acuerdos particulares con bases de datos para que su buscador pueda indicar el contenido de éstas." En este sentido, ya ofrece acceso a una pequeña parte de la base de datos WorldCat de OCLC (Online Computer Library Center), esto es a tan sólo dos millones de los cincuenta y ocho millones de registros que contiene. También da acceso a la base de datos de la Biblioteca Nacional de Medicina de Estados Unidos (NCBI).
Yahoo también facilita acceso a bases de datos de la Biblioteca del Congreso de EEUU, de la Universidad de California en los Ángeles, la Radio Pública Nacional de EEUU, la Universidad de Michigan y el Proyecto Gutenberg así como a otras bases de datos de pago (Financial Times, The Wall Street Journal, The New England Journal of Medicine, las publicaciones del IEEE, etcétera). Todas ellas requieren para su acceso al documento pagar una cuota de suscripción.

[180] Los comandos y etiquetas específicas destinadas a los robots pueden consultarse en la siguiente página web: <http://www.robotstxt.org/wc/exclusion.html>

[181] Cifra muy superior a la apuntada anteriormente por Lawrence y Giles, que se acerca más a la aportada por Cyveillance

santes los datos aportados por Chang, recogidos por Martínez y Rodríguez (2003) quienes calculan que los motores de búsqueda indizan entre el 5% y el 30% del total de la Web, y que la unión de los once principales motores alcanza el 50%.

La búsqueda y acceso a recursos de la web invisible va siendo posible gracias a la creación de distintos motores como Invisible Web Catalog[182], Webdata[183] o Direct Search[184] que dan acceso a diferentes bases de datos y directorios como CiteLine Profesional[185], que localiza tanto recursos de la Web visible como invisible. En el ámbito español podemos mencionar Internet Invisible[186], que recopila bases de datos de acceso gratuito. Por su parte, los desarrolladores de los motores de búsqueda comerciales tratan de superar este problema preocupándose cada vez más por ofrecer una cobertura más completa de la web. En este sentido hay que mencionar a motores de búsqueda como Google, MSN y Yahoo que posibilitan la recuperación de documentos distintos de HTML como PDF, Postscript, Flash, Shockwawe, programas ejecutables, archivos comprimidos y documentos de Office (Word, Excel, Power Point), etcétera.

En relación con el tercer aspecto apuntado, esto es, el carácter dinámico de la información, es frecuente en este medio el establecimiento de cambios que afectan a las direcciones URL[187] de los recursos, aspecto que se produce cuando el recurso se traslada a otro servidor o cambia de directorio. También puede dejar de estar disponible y desaparecer, por lo que dejará de estar visible y localizable aunque la URL perdure en los sistemas de recuperación o páginas que lo conectaban. Las soluciones aportadas, en unos casos, dependen del servidor en que se alojan las páginas, como es la utilización de programas que reenvían de la dirección URL no válida a la válida, lo que facilita el trabajo de los motores de búsqueda al permitir que los robots puedan acceder de nuevo a dichas páginas. Este aspecto ha de completarse con la aplicación de técnicas de actualización de índices. En este sentido, generalmente los motores declaran en sus páginas informativas, estar en posesión de tecnología que elimina estos enlaces rotos así como las páginas duplicadas.

---

[182] http://www.invisibleweb.com
[183] http://www.webdata.com/webdata.htm
[184] http://gwis2.circ.gwu.edu/~gprice/direct.htm
[185] http://www.citeline.com/proinfo.html
[186] Es un directorio de bases de datos. Se puede acceder en la dirección: http://www.internetinvisible.com

Otro factor externo, al que ya hemos aludido en diferentes apartados de este trabajo, que influye negativamente en la recuperación ofreciendo resultados poco precisos, es la escasa utilización de metadatos.

No hay que olvidar que los motores de búsqueda, aunque parecen herramientas simples, están configuradas con una serie de programas informáticos complejos que en la medida de lo posible tratan de hacer frente a los problemas señalados, y que exige de forma continua la incorporación de nuevas técnicas de búsqueda y recuperación de información como las basadas en la inteligencia artificial, en el uso de algoritmos complejos, en la realización de resúmenes automáticos, en el uso del lenguaje natural, en las búsquedas por conceptos, etcétera. No obstante, al compararlas con otras herramientas de búsqueda, podemos considerarlas incompletas, ya que, por ejemplo, no soportan determinadas posibilidades de búsqueda. Pero esto es debido a que en la Web predomina la información no estructurada lo que por ejemplo dificulta la realización de índices específicos que se puedan utilizar e incluso combinar en las búsquedas. Esto hace que debamos considerarlas complementarias de otras, generalmente más especializadas, a las que aún no han igualado, y mucho menos superado. Sirvan como ejemplo las bases de datos comerciales disponibles a través de la Web (Bases de datos de la Web on Kowledge, etcétera).

Es, también en este caso, en el campo de la estructuración de los documentos donde puede estar la solución a la recuperación efectiva en la Web. Para que los SRI Automatizados funcionen correctamente, ha de existir un lenguaje de codificación que dé consistencia a dichos sistemas como por ejemplo en los OPACS de bibliotecas, que utilizan el formato MARC. Dicho formato permite, al incorporar la etiqueta 856, su utilización para describir y acceder a recursos web. Los inconvenientes que se apuntan a la incorporación de este formato se centran en la dificultad de su uso y en la utilización de códigos que hacen lenta la descripción. Charton (1997) señala al respecto que pueden existir problemas al aplicar a documentos electrónicos las formas de descripción propias de documentos en formato papel, ya que los electrónicos se caracterizan por: su inmediato acceso, su continua evolución y cambio, y su rápida puesta en circulación. Defiende de este modo únicamente la utilización del formato MARC en documentos web importantes

[187] Spinellis (2003) ha demostrado que la vida media de una dirección URL es de cuatro años.

y perdurables. De aquí el interés por buscar un sistema que aligere la codificación, lo que ha dado lugar a sistemas como Dublin Core y otros.

Otro aspecto necesario y en el que se han conseguido avances, tiene que ver con la normalización en la descripción de recursos electrónicos. En este sentido, debemos mencionar el desarrollo de la norma ISBD (ER) (International Standard Bibliographic Description. Electronic Resources).

Finalmente, hay que señalar la existencia de amplios marcos que, en el seno de la Web, tratan de posibilitar la utilización y validez de los diferentes sistemas de metadatos como es el caso de RDF (Resource Description Framework).

Entre los factores internos, recogemos los señalados por Estibill y Abadal (2000) que se refieren al rendimiento "poco satisfactorio" que presentan los Motores. Los problemas que señalan son los siguientes:

> *"-No disponen de criterios de calidad para seleccionar los recursos que han de formar parte de la base de datos: recuperan todo tipo de documentos de forma indiscriminada.*
>
> *-Las descripciones de las fuentes indexadas son muy elementales, incompletas y muchas veces erróneas. Normalmente sólo incluyen el título, el URL, y en algunas ocasiones un breve resumen o las primeras líneas de la página. Como este proceso es a menudo automático, el número de errores y de omisiones es importante. El usuario puede tener muchas dificultades para decidir los que le interesan cuando el buscador le muestra una lista muy extensa de materiales recuperados a partir de los términos de consulta.*
>
> *-No presentan todas las opciones de recuperación de los catálogos o de las bases de datos comerciales. Por ejemplo, los usuarios no pueden especificar en qué campo del registro se ha de ejecutar la consulta: una búsqueda por "Josep Pla" recuperará de manera indiscriminada información sobre este autor, sus propios escritos y documentos publicados por la Fundación Josep Pla.*
>
> *-La unidad documental de la que parten es el fichero y no el recurso, lo cual provoca un grado muy alto de redundancia.*
>
> *-La consulta es lenta como consecuencia del tráfico de la Red y de la sobrecarga de algunos servidores."*

Oppenheim, Morris y McNight (2000) recogen las siguientes críticas:

- Baja relación respuesta-tiempo.
- Recuperación de registros duplicados.
- Dificultad en recuperar recursos relevantes ante el amplio número de recursos irrelevantes.
- Problemas para recuperar recursos que se sabe están en la red.
- Recuperación de recursos anticuados.

Frente a los apuntados por Estibill y Abadal, se trata de problemas que podemos considerar como cuantificables, y por tanto de un mayor interés como indicadores de evaluación. No en vano, muchos de los trabajos de evaluación se ocupan de la valoración de alguno de estos aspectos.

Otro de los problemas que influye en la recuperación tiene que ver con la financiación de estas herramientas. Las prácticas comerciales que realizan estas herramientas para conseguir financiarse y mantenerse, influyen tanto en la formación de la base de datos, en los índices y como no, en la recuperación.

Hemos de referirnos, en este sentido, a la práctica consistente en la extracción por los robots de información de los sitios o páginas web, cuya indización y posibilidad de localización a través del buscador, puede demorarse cierto tiempo (uno o varios meses). Pero este tiempo puede acortarse, en función del contrato establecido entre los interesados.

Esta práctica viene siendo utilizada por varios motores y consiste en el pago de una tarifa, que permite la rápida indización y alojamiento de un determinado número de páginas durante un periodo de tiempo que puede ser seis meses o un año, lo que resulta interesante para empresas que ofrecen información de tipo comercial. Pero además, el pago de una tasa puede facilitar que una determinada casa comercial aparezca en lugares destacados del listado de resultados de búsqueda. Así ocurre con: OpenText, Yahoo, AltaVista, AskJeewes, Inktomi, MSN Search y LookSmart (Sullivan 2001). Otra práctica comercial consiste en incluir un mayor número de términos de búsqueda al indizar determinadas páginas comerciales, para que sea más fácil recuperarlas. No obstante, para evitar la mezcla de estos recursos con otros no patrocinados, se ha ido generalizando cada vez más, la práctica de colocar los recursos de carácter comercial que responden a una búsqueda, en listados aparte, aunque no siempre es así.

Henshaw (2001) distingue tres modalidades de ubicación comercial de los resultados. Por un lado lo que denomina Paid placement, que garantiza una determinada posición del recurso en la primera página de resultados, sin mezclarse con el resto de resultados; Paid inclusion, cuando los recursos se entremezclan con los demás; y Paid submission, cuando simplemente se quiere incluir el recurso en el índice del motor de búsqueda. Dado que se trata de diferentes prácticas comerciales, están sometidas a frecuentes cambios, por lo que actualmente, las dos últimas modalidades apenas se distinguen. En cual-

quier caso, la modalidad Paid inclusión conlleva la obligatoriedad de revisar frecuentemente el sitio o de indizarlo de forma inmediata.

Tampoco en este sentido había unanimidad, ya que por ejemplo LookSmart e Inktomi no los separaban (Hensaw, 2001). Google los incluye al comienzo del listado, dentro de las categorías Premium Sponsorship programs o a un lado (AdWords).

Respecto a la frecuencia con que los robots visitan los servidores o páginas para actualizar la información extraída, también los aspectos financieros juegan su papel.

No obstante, dado que nosotros nos ocupamos en la evaluación de búsquedas sobre temas especializados, en principio, lo comercial no debería interferir en los resultados, aunque, estudiando el carácter de la información recuperada, trataremos de observar hasta qué punto se da la convivencia en la recuperación de resultados de carácter comercial con otros de carácter científico y si predominan unos sobre otros.

Además de estos aspectos, podemos señalar otros problemas como el amplio número de recursos que se presentan como resultados de la búsqueda, dado que además, el verdadero problema es que no todos suelen responder a lo solicitado en la búsqueda. Cierto es que una mayor especificación podría lograrse consultando la base de datos mediante técnicas avanzadas, pero no siempre están disponibles.

De aquí que uno de los principales problemas sea la baja precisión en la recuperación. Lawrence (2000) señala como solución a este problema la posibilidad de indicar el contexto desde el que se realiza la búsqueda. Propone en este sentido la posibilidad de indicar que la búsqueda se lance sobre páginas personales, sobre documentos de investigación o sobre páginas de información general, como es el caso del metabuscador Inquirus 2[188]. Yahoo y Google asisten al usuario mediante ayudas que permiten contextualizar la búsqueda, aunque de una forma muy básica.

Otra solución al problema de la poca relevancia de los resultados puede ser el desarrollo de herramientas de búsqueda especializada, como es el caso de Directorios temáticos y de otros sistemas que permitan la recuperación de recursos descritos mediante el uso de metadatos, como el sistema Dublin Core, PICS (Platform for Internet Content

[188] Multibuscador en fase de desarrollo que se lleva a cabo en los laboratorios NEC

Selection), TEI (Text Encoding Initiative)[189], DOI (Digital Object Identifier), etcétera, o mediante el formato MARC como es el caso de NetFirst, la base de datos de carácter comercial sobre recursos en Internet y el catálogo InterCat, experimento cooperativo para la utilización de herramientas de catalogación de recursos web, ambos promovidos por OCLC o, finalmente, el programa inglés eLib (Electronic Libraries Programme, 1997) del que forma parte el proyecto Acces to Network Resources[190]. Hay que mencionar también la colección descrita mediante la utilización de metadatos, mantenida por la Universidad de California, con el nombre INFOMINE. Además, desde los comienzos de estas herramientas, buscadores como AltaVista, HotBot, y Lycos[191] utilizaban etiquetas Meta como "description" y "keywords" en la formación de sus índices. Actualmente, buscadores de metadatos como MetaBrowser[192], HotMeta[193], y HiSearch[194] además de los que utilizan los metadatos de Dublin Core.

Schwartz (1998) señala las ventajas que ofrecen los servicios que utilizan metainformación: mejora de las representaciones, motores de búsqueda y directorios desarrollados, control de calidad en la selección y resultados de búsqueda de mayor precisión y exhaustividad. Pero estas prestaciones sólo pueden realizarse de forma limitada, ya que exigen la intervención de gran número de personal especializado, con una gran dedicación de tiempo, lo que traducido en coste económico resulta difícilmente planteable para los motores de búsqueda. Por otro lado, las iniciativas existentes, requieren para su supervivencia un tratamiento comercial que les ayude a mantenerse y desarrollarse.

Otras soluciones apuntan a la utilización de vocabularios controlados, pero el gran número y la naturaleza cambiante de los recursos web no lo hacen posible[195]. El

---

189 Propone el etiquetado de todo tipo de textos, especialmente los de valor literario y lingüístico. Existen en la iniciativa diversos comités que se encargan de estudiar y proponer las etiquetas necesarias, de normalizar la descripción física, de la descripción lingüística y literaria de los textos y de la sintaxis de etiquetación. Sus directrices pueden consultarse en http://etext.virginia.edu/TEI.html

190 Basado en la descripción con metadatos y selección de recursos web de calidad relacionados con el arte, la economía, la medicina, que ha dado lugar a la creación de herramientas de búsqueda como ADAM (Art, Design, Architecture &Meida information Gateway), BUBL y OMNI (Organising Medical Networked Information).

191 Westerra (1997) señala que si bien este buscador utiliza este tipo de etiquetas, no las emplea en los cálculos de relevancia de recursos.

192 http://metabrowser.spirit.net.au

193 http://www.dstc.edu.au/Research/Projects/hotmeta/search.html

194 http://www.hisoftware.com/MCBS/index.html

195 http://www.hisoftware.com/MCBS/index.html

lenguaje XML aporta soluciones, como el uso de etiquetas para contener información específica.

Estas iniciativas por un lado, y la constante evolución de los motores de búsqueda, que tendrán que irse adaptando a estos nuevos sistemas descriptivos, por otro, permite pensar que la recuperación de la información en la Web debe de ir mejorando constantemente, sin perjuicio de irse desarrollando otras iniciativas como es el caso de buscadores especializados, catálogos de recursos, agentes inteligentes, bibliotecas digitales, etcétera.

La constante preocupación por la recuperación efectiva ha dado lugar al desarrollo del concepto de web semántica, concepto creado por Tim Berners-Lee. El proyecto en la actualidad trata de desarrollar sistemas de descripción y consulta a través de estándares que permita describir, buscar y encontrar recursos de una forma normalizada.

En la actualidad, los desarrolladores de los motores de búsqueda, conscientes del problema de la falta de precisión en la recuperación, tratan de resolverlo de diferentes formas, bien mediante el desarrollo de algoritmos basados en la popularidad, como en Google, o mediante otras opciones de búsqueda avanzada, en AltaVista, Lycos, HotBot o Yahoo. Otras opciones utilizan el filtrado automático como es la opción “More like this” y otras similares.

La ordenación por relevancia es otro aspecto de gran importancia en la recuperación. Su importancia viene dada por el hecho de que el usuario pocas veces consulta más de dos páginas de resultados, por lo que si una determinada herramienta quiere ser útil para el usuario, ha de ofrecer en los primeros puestos los resultados más relevantes. Para ello los buscadores han de aplicar técnicas de valoración por relevancia.

Arms (2001:211) se refiere a que el problema del deficiente funcionamiento de los buscadores en este aspecto puede ser debido a que los algoritmos de ranking no tienen suficiente información en la que basar sus resultados.

Los metabuscadores también tratan de solucionar algunos de estos problemas mediante técnicas que valoran el contexto en el que se encuentran los términos de búsqueda en el documento, la identificación tanto de páginas que ya no existen como de las que no contienen los términos de búsqueda, localización de páginas duplicadas, mejora del ranking y de la precisión. (Lawrence y Giles, 1998d).

Es por esto que hay que valorar tanto la precisión como el ranking u ordenación de registros, ya que dichas técnicas no siempre se utilizan o lo hacen de forma eficiente.

Para ello, en la evaluación valoraremos aspectos como la precisión técnica y analizaremos tanto la función que la metainformación puede jugar en la ordenación como las frecuencias y peso de los términos de búsqueda.

3. La evaluación de los sistemas de recuperación de la información y las herramientas de búsqueda de la World Wide Web.

3.1. Concepto y fines de la evaluación. El proceso de evaluación

Según Ingwersen (1992) la Recuperación de la Información como disciplina científica se ocupa de diseñar, construir y probar sistemas de recuperación que faciliten el acceso a la misma. De aquí el amplio número de estudios que se dedica a la evaluación de dichos sistemas.

En el presente capítulo vamos a exponer el contexto en el que tienen lugar los trabajos de evaluación de los SRI ya que en él se enmarcan los estudios de evaluación de los buscadores web, analizaremos las experiencias más importantes desarrolladas sobre estas herramientas y finalmente, trataremos de los indicadores seleccionados para llevar a cabo nuestro trabajo de evaluación.

3.1.1. Concepto de evaluación

El Diccionario de la Real Academia[196] dice en sus dos primeras acepciones que evaluar es: "Señalar el valor de una cosa. Estimar, apreciar, calcular el valor de una cosa".

Ello supone efectuar una medición, que bien puede ser global o basada en la valoración de diferentes elementos, lo que requiere el establecimiento de un sistema de medida.

Pero en la evaluación de los buscadores web no existen estándares que puedan ser utilizados para expresar sus puntos fuertes y débiles o para compararlos cuantitativamente. Tampoco hay unos criterios establecidos que sirvan como referencia en la evaluación sino que, como veremos al tratar sobre el estado de la cuestión, se utilizan diferentes parámetros en función de lo que se quiere evaluar. Además se recurre a medidas y métodos utilizados en la evaluación de otros sistemas de recuperación de información. En este

sentido, la utilización del método estadístico es el más recomendable, ya que facilita, mediante la recogida, análisis, y comparación de datos, suficientes elementos objetivos para valorar los aspectos de interés. La evaluación requiere también contrastar los datos obtenidos con los objetivos de estos servicios y juzgar hasta qué punto se cumplen.

Desde el punto de vista de los especialistas en Recuperación de la Información, Lancaster (1992) señala que es:

> "esencialmente un procedimiento de diagnóstico (y eventualmente terapéutico) en el que interesa la identificación del origen de los fallos del sistema".

En un trabajo anterior[197] indica que la evaluación de un sistema ha de hacerse valorando si alcanza sus objetivos, si es eficiente en este aspecto y si se justifica su existencia.

Resulta interesante esta visión de Lancaster por cuanto requiere un detenido análisis de las partes del sistema; saber por qué se dan esos resultados, y ponerlo en conocimiento de los desarrolladores de estas herramientas para buscar soluciones que mejoren los resultados. En este sentido, nuestro objetivo es también ofrecer una valoración imparcial que permita conocer, tanto a los investigadores como a los especialistas en RI, el rendimiento de estas herramientas en búsquedas de información especializada.

Desde el punto de vista práctico, se trata tanto de una herramienta de investigación, de toma de decisiones y de gestión, que permite no sólo conocer el alcance de alguno de los problemas que afectan a los Sistemas de búsqueda, sino también plantear su posible solución o mejora. Va a facilitar la toma de decisiones tanto a los desarrolladores del sistema como a los usuarios, ya que en función de la valoración de sus fortalezas y debilidades, pueden decidir si usarlas o no y seleccionar la más conveniente. Como herramienta de gestión va a servir para plantearse su futuro, su posible desarrollo, la dirección a seguir, etcétera. Finalmente, como herramienta de investigación permite, utilizando el método científico, analizar y valorar un acontecimiento, un hecho, un objeto, en este caso los buscadores de información de la Web. Steiner (1979) ha señalado además,

---

[196] Diccionario de la Lengua Española. 21ª ed. Madrid, Real Academia Española, 1992.
[197] Lancaster (1971).

su valor como herramienta de control respecto a la consecución de los objetivos marcados.

Abad (2002:671) define la evaluación como:

> "un proceso mediante el cual se intenta obtener un juicio de valor o una apreciación de la bondad de un objeto, de una actividad, de un proceso o de sus resultados. Esto es, la puesta en práctica de un procedimiento con el que poner de relieve las cualidades, ventajas y debilidades de aquello que se evalúa".

No obstante nos parece más acertada la expresada en un trabajo posterior (2005:41) al referirse a la evaluación también como:

> "proceso que tiene como objetivo la realización del diagnóstico de una situación determinada cuyo resultado será la emisión de un juicio de valor acerca del funcionamiento, calidad, aceptación o cualquier otra cualidad de un sistema de información."

En este caso, deja abierta la evaluación, pudiéndose centrar en la valoración de la cualidad o cualidades de su funcionamiento que, en cada caso, interese analizar.

Por otro lado, esta autora señala la necesidad de un referente con el que comparar, que será distinto de acuerdo con los fines que persiga la evaluación. Dado que este aspecto no es posible en la evaluación de buscadores web, por las características dinámicas de este medio, se suele recurrir al método comparativo entre distintos sistemas para destacar alguno de ellos.

Desde un punto de vista más específico, para Harter y Hert (1997) la evaluación es un proceso que trata de valorar la efectividad de un servicio o sistema y en qué medida se cumplen sus metas y objetivos.

Como resumen y en opinión de Hernon (1998), evaluar un sistema es el proceso de identificar o recabar datos acerca de actividades y servicios específicos, estableciendo criterios por los que pueda calcularse su bondad o acierto y determinarse la calidad de la actividad o servicio, y el grado en que éstos logran sus metas y objetivos.

3.1.2. Fines y objetivos

Aunque algunas de las definiciones anteriores llevan implícitas información sobre para qué se evalúa, es interesante profundizar algo más en este aspecto. Abad, en una monografía reciente (2005:20) señala que la evaluación suele responder a alguna de las siguientes razones:

1. Medir la consecución de los objetivos previamente establecidos.
2. Disponer de un instrumento para diagnosticar los puntos débiles en el funcionamiento.
3. Facilitar el proceso de la toma de decisiones.
4. Permitir la comparación entre sistemas mediante la construcción de estándares de referencia.
5. Justificar la existencia de los servicios y sistemas de información.

De este modo se consigue valorar la eficacia, la eficiencia o el impacto de un determinado servicio; valorar la correcta realización de los procesos; tratar de responder bien a la calidad de los procesos o a la corrección de la realización de las operaciones documentales; finalmente, las dos siguientes razones facilitan la toma de decisiones, aunque en el ámbito de la Web, es difícil poder establecer estándares de referencia y sirve para justificar el mantenimiento de un servicio o su mejora. Se tiene en cuenta la satisfacción del colectivo al que va destinado y se ha de justificar el cumplimiento de sus objetivos y el mantenimiento de la calidad, etcétera. Desde nuestro punto de vista, al tratar de valorar su funcionamiento, nos parecen de especial interés las razones señaladas en primero, segundo y quinto lugar. Consideramos también importante la tercera razón, ya que ha de suponer una intención de mejora por parte de los desarrolladores de los sistemas que obtengan los peores resultados. De lo anterior, hemos de destacar la idea de que debemos conocer en primer lugar para qué evaluamos.

En nuestro caso, éstas serían las razones para la evaluación:

-Conocer si son útiles en búsquedas de información científica y cuál o cuáles lo son más.

-Conocer si su funcionamiento es correcto.

En este sentido, hemos de apuntar que es nuestra intención valorar tanto la realización de las operaciones documentales que llevan a cabo los buscadores y metabuscadores de la Web, como su correcto funcionamiento y utilidad, especialmente en recuperación especializada, y establecer comparaciones que permitan seleccionar los más útiles.

Abad ha estudiado el tipo de evaluación correspondiente a cada una de las fases de lo que denomina "ciclo vital" de los sistemas de información[198]. Nuestra experiencia se relaciona con la fase de funcionamiento o rutina, que se caracteriza por un funcionamiento fluido del sistema y unos usuarios más o menos fieles. La finalidad de la evaluación en esta fase es valorar si el sistema cumple con las metas y objetivos propuestos, y cómo los cumple. Las investigaciones pueden centrarse en evaluar el comportamiento de los componentes del sistema o en el funcionamiento global. Entre los primeros señala el *input* del sistema y el proceso documental. Respecto a los segundos, se refiere a la eficacia[199] de la recuperación, la eficiencia y la satisfacción. Esta segunda tendencia está más centrada en la valoración del funcionamiento desde el punto de vista del usuario.

Crawford (1996) y Abad (2005) coinciden en que la evaluación ha de recoger información que facilite la toma de decisiones y la justificación y defensa de los recursos empleados. Señalan además que ha de determinar la calidad del servicio, resolver en la medida de lo posible los problemas que se presentan y descubrir las bases para nuevas mejoras. Estos fines coinciden plenamente con los que nos hemos planteado al evaluar los motores web, ya que seguir este planteamiento nos permitirá seleccionar los buscadores que ofrecen una mayor calidad ante búsquedas sobre temas especializados y conocer cuáles son los que menos acusan los problemas detectados en estas herramientas, lo que ha de servir además para facilitar su mejora.

3.1.3. Proceso de evaluación

Como acabamos de ver, la mayoría de autores especializados en evaluar sistemas de recuperación de la información, se refieren a la evaluación como un proceso. Abad señala las siguientes etapas, que trataremos de seguir:

1. Obtener los datos sobre la situación actual del sistema a evaluar.
2. Decidir los criterios según los que se evaluará el sistema y definir los indicadores para la obtención de resultados.
3. Recoger los datos sobre los aspectos a evaluar.

[198] Señala como fases en las que se puede evaluar: Planificación, Viabilidad, Diseño, Implantación y Funcionamiento o Rutina.

4. Comparar los hallazgos obtenidos con una situación de referencia o estándar.

5. Emitir un juicio de valor basado en el análisis de las diferencias y similitudes entre la situación observada y la situación de referencia.

6. Averiguar el origen de las diferencias encontradas.

7. Establecer unas acciones y recomendaciones para la mejora.

La evaluación es un proceso complejo en el que como indican Large y otros (1999) se ha de tener en cuenta tanto el punto de vista mecánico, el humano y la utilidad para un grupo determinado. Los basados en el punto de vista del usuario utilizan medidas que tratan de valorar su satisfacción con los resultados.

Nosotros nos ocupamos de evaluar los procesos de formación de la base de datos, la composición de índices y las capacidades de búsqueda y recuperación, por lo que deberemos utilizar los parámetros que nos permitan valorarlos.

3.2. Tendencias en la evaluación de SRI

Antes de referirnos a los indicadores, debemos tener claro qué tipo de evaluación queremos llevar a cabo. La evaluación de SRI tradicionales se ha abordado principalmente desde dos puntos de vista: el del sistema y el del usuario. Inicialmente se centraron tanto en medir la validez de sus índices y comparar la recuperación basada en el lenguaje natural con la ofrecida mediante vocabularios controlados, (Cleverdon 1966, Lancaster 1968, Aitchison 1969-1970, Keen y Digger 1972, etcétera), como en valorar la efectividad de los sistemas, midiendo la relevancia de la información recuperada, o el comportamiento de un determinado programa o herramienta de búsqueda.

Sin embargo para Salton (1983), una correcta evaluación debe contemplar ambas perspectivas y Rijsbergen (1979) piensa que la evaluación del Sistema de Recuperación de Información debe reflejar la capacidad del sistema para satisfacer al usuario. La corriente actual trata pues de integrar ambas tendencias considerando tan válida una como la otra, ya que por ejemplo, tan importante puede ser conocer la opinión de un usuario

[199] Entendida como "capacidad del sistema de recuperar información relevante para el usuario." Abad (2005:143).

respecto a determinada característica de un sistema de recuperación como la actualización o la cobertura de su base de datos.

Tanto Ingwersen (1992) como Ellis (1992) señala tres grandes bloques respecto a la investigación en torno a la Recuperación de la Información, a saber: el clásico o algorítmico, el orientado a usuarios y el modelo cognitivo. Este último tiene en cuenta aspectos relacionados tanto con el sistema como con el usuario.

Harter y Hert en un importante estudio publicado en 1997 se ocupan de analizar con cierta profundidad diferentes planteamientos, problemas y métodos de la evaluación de SRI, y reducen a dos las tendencias importantes en este campo: la clásica y la orientada al usuario. La primera se ocupa de los algoritmos y estructuras de datos necesarios para optimizar la eficacia de las búsquedas y la segunda se centra en el usuario, en analizar su interacción con el sistema de recuperación, y en el papel de las fuentes de conocimiento implicadas en la Recuperación de la Información. Estos autores consideran que ha de darse mayor importancia a la interacción usuario y sistema.

La corriente tradicional tiene sus más claros antecedentes en las experiencias desarrolladas en el Proyecto Cranfield, puesto en marcha por Cleverdon en 1957, para el que se señalan dos etapas, Cranfield I (1957-1962) y Cranfield II (1963-1966), de las que surgieron fundamentalmente un marco teórico, una metodología y unas herramientas básicas para la evaluación de SRI. Entre éstas últimas, se estableció la necesidad de contar con una colección de documentos fuente; señalar una serie de valores para medir la efectividad de los sistemas, teniendo en cuenta la recuperación de documentos relevantes y la exhaustividad, estableciendo como medidas la exhaustividad y la precisión.

Estos estudios tuvieron una aplicación inmediata en la evaluación de sistemas como SMART, realizado por Salton desde inicios de los 60, la base de datos MEDLARS (Medical Literature Analysis and Retrieval System) llevado a cabo por Lancaster entre 1966 y 1967[200], o el caso de la base de datos STAIRS (Storage and Information Retrieval System) evaluada por Blair y Maron en 1985.

[200] Abad (2002) señala que el trabajo de Lancaster "puso en evidencia que en el contexto real no sólo era necesario conocer la eficacia del sistema, sino que debían explorar las causas que conducían al éxito y al fracaso de la recuperación".

Como hemos señalado, los principales valores en que se basaban estas evaluaciones son la exhaustividad y la relevancia. Su valoración correspondía a expertos ya que podían juzgar de forma más exacta el valor de los documentos o referencias recuperadas.

Las críticas a esta tendencia arrancan con Doyle (1964), quién señaló su carácter subjetivo. Posteriormente Ellis (1984) reparó en las dificultades a la hora de definir la relevancia. Harter y Hert (1997) por su parte, centran sus críticas en la ausencia del usuario en la evaluación, en las dificultades que plantea el criterio de la relevancia, y en que las experiencias son poco reales por lo que los resultados no pueden compararse con otros basados en consultas y usuarios reales. Korfhage (1997) va más allá, al plantear que no está claro que exhaustividad y precisión sean medidas significativas para el usuario.

Estas críticas se han visto superadas en el ámbito de las experiencias TREC (Text REtrieval Conferences), que se celebran anualmente desde 1992 y cuya metodología para la evaluación de sistemas de recuperación, establece la existencia de una amplia colección de documentos y unos procedimientos normalizados, dirigidos a evaluar, fundamentalmente, los algoritmos que utilizan estas herramientas, así como los contenidos de las bases de datos, su comportamiento ante diferentes búsquedas, utilización de tesauros, recuperación de información en idiomas específicos, los mecanismos de recuperación a través de la voz y la incorporación del punto de vista humano.

Pero tampoco las experiencias TREC han estado exentas de críticas, pues como recoge Chaín (2004:184) diferentes autores han mostrado su desacuerdo en basar la evaluación en juicios de relevancia, en el modo en que se seleccionan los temas de búsqueda y en los juicios poco realistas para usuarios reales. Craswell, Bailey y Hawking (1999) se han referido a la diferencia entre los sistemas TREC que recuperan recursos planos frente a los hipertextuales de los sistemas Web, lo que puede influir en los juicios de relevancia al no valorar positivamente páginas que enlazan con otras que sí son relevantes. Tampoco están de acuerdo los investigadores en que la valoración de la relevancia deba hacerse respecto a la materia y no a la calidad del recurso. Blair (2002) ha criticado la forma de

calcular la exhaustividad[201] ya que en su opinión se ofrecen valores superiores a la realidad.

No obstante, esta corriente también ha dejado su impronta, como veremos, en los estudios de evaluación de buscadores Web llevados a cabo por Ding y Marchionini, Chu y Rosenthal, Clarke y Willett, y Leighton y Srivastava.

Respecto a la tendencia centrada en el usuario, para Olvera (1998:26):

> "[...] trata de representar los problemas de información, comportamiento en las búsquedas y componentes humanos de los sistemas de información en situaciones reales. Se nutre principalmente de la psicología cognitiva y emplea métodos de las ciencias sociales [...]".

En este tipo de investigaciones entran en juego aspectos que tienen que ver con el comportamiento mental de los usuarios y de las características de sus búsquedas de información, teniendo en cuenta tanto al individuo como los contextos sociales y organizativos (Olvera 1998). Este enfoque ha sido analizado en diferentes trabajos de autores como Inwersen (1982), Saracevic y Kantor (1988), Spink y otros (1996), Borlund (2000), y por Su (1994) que lo ha aplicado a los sistemas de recuperación de la Web. En un trabajo de 1989 ya utilizó variables como las aptitudes técnicas y características personales.

Esta tendencia ha ido ganando importancia en los últimos años dado que el usuario tienen cada vez un acceso más directo a estos sistemas, ganando protagonismo en este aspecto, de manera que las evaluaciones cada vez se han venido preocupando más de expresar la opinión del usuario y su grado de satisfacción.

Criterios a tener en cuenta son los aportados por Keen, que recogen Martínez y Rodríguez (2004a): la cobertura, entendida como la proporción de los documentos relevantes conocidos que el usuario ha recuperado; la novedad, que contempla los documentos recuperados relevantes desconocidos para el usuario; la exhaustividad relativa y el esfuerzo de exhaustividad.

[201] La metodología que se emplea se basa en la constitución de una colección de recursos relevantes fruto de una primera búsqueda en los diferentes buscadores a evaluar, que se utiliza para comparar la recuperación individual de los diferentes buscadores.

Peña y otros (2002) definen estos últimos elementos como:

> "la proporción de documentos relevantes que ofrece el sistema al usuario respecto de los que esperaba encontrar [Exhaustividad relativa]. Respecto a la precisión de usuario, nos indica la proporción de documentos que el usuario ha encontrado relevantes en una muestra de un tamaño fijado por él mismo. Finalmente, "el esfuerzo de exhaustividad" es la proporción entre el número de documentos esperados en relación con los que han sido examinados para alcanzarlos".

Según estos autores, son valores que se caracterizan por su gran subjetividad.

Para Hearst (1999) además, extraer datos empíricos a partir de experiencias de evaluación con personas, requiere una gran inversión de tiempo y grandes dificultades para extraer conclusiones. Los estudios psicológicos sólo alcanzan pequeñas conclusiones en un contexto limitado, unas veces valorables empíricamente pero otras no valorables de forma exacta. De aquí que, como este autor cita, Nielsen defienda una evaluación más informal, que denomina "evaluación heurística", centrada en unos valores más generales del interfaz de usuario y no siempre expresados con valores estadísticos.

Respecto a las herramientas de recuperación en la Web, y en este caso hay que tener en cuenta que a la práctica totalidad de sistemas de recuperación se accede a través de ella, en la valoración de la satisfacción del usuario, hay que tener en cuenta que se trata de sistemas interactivos, en los que la forma de consultar la base de datos es muy variada y el conocimiento y la práctica del usuario sobre búsquedas en determinados sistemas puede influir de forma decisiva en los resultados de una búsqueda, y como no, en su opinión sobre ellos.

Esta tendencia como señala Olvera (1998:28) también ha sido criticada

> "[…] al excluir la información contextual, menospreciar el método científico y restringirse al estudio de las mentes de usuarios que no son más que vagos constructos abstractos de los propios investigadores, se llega a construir un cuerpo de conocimientos cuya utilidad es más que dudosa."

Finalmente Ellis (1994) ha criticado además la falta de rigor en sus planteamientos metodológicos.

Para servir de elemento aglutinante de las experiencias en evaluación de SRI, y más concretamente de los sistemas interactivos multimedia, se constituyó en 1996, en el seno de la Comisión de las Tecnologías de la Información de la Unión Europea (ESPRIT), un programa a tres años denominado MIRA en el que se planteó la necesidad de acercar las dos posturas tradicionales en evaluación de sistemas.

En este contexto se han venido desarrollando desde mediados de los 80 diferentes experiencias en el campo de la evaluación de SRI, aplicadas tanto a los OPAC (Online Public Acces Catalog), como a los Sistemas online como DIALOG, ORBIT, BRS y bases de datos en soporte CD-ROM, en algunos casos con el fin de facilitar su adquisición.

Con la llegada de los sistemas de recuperación web en los años noventa, se ha reactivado el interés por analizar su influencia y utilidad en este nuevo medio. Antes de referirnos a los diversos trabajos de evaluación realizados en este medio, nos ocuparemos brevemente de analizar cuáles han sido los indicadores utilizados en la evaluación de SRI.

3.3. La evaluación de los SRI. Indicadores

Uno de los aspectos que más ha preocupado a los autores que se han ocupado de evaluar los SRI ha sido establecer una serie de parámetros entorno a los cuales llevar a cabo la evaluación. Chowdury señala un trabajo de Cleverdon de 1978 en el que los seis criterios que propone para evaluar los SRI son, además de la exhaustividad, la precisión, el tiempo de respuesta, el esfuerzo del usuario, la presentación de resultados y la cobertura. Valores similares han propuesto Lancaster (1971) y Vickery (1970) sustituyendo éste último el esfuerzo por la utilidad, mediante la que se valora la calidad de las referencias recuperadas.

Salton y McGill, en un trabajo publicado en 1983 proponen uno de los cuadros más completos de criterios y medidas para evaluar estos sistemas. Así, dentro de los criterios de exhaustividad y precisión proponen valorar la exhaustividad de indización, la especificidad de los términos, el lenguaje de indización, la formulación de la consulta, y la estrategia de búsqueda. Respecto al tiempo de respuesta, son factores a valorar: la organización de los documentos almacenados, tipo de consulta, ubicación del centro de información, frecuencia de utilización y tamaño de colección. El esfuerzo del usuario viene determinado por: la accesibilidad al sistema, la disponibilidad de ayudas, el número de recursos recuperados y la facilidad de interactuar con el sistema. El cuarto criterio se refiere a la forma de presentación de los resultados y establecen como valores, el dispositivo de visualización y los datos del registro. Finalmente, respecto a la cobertura, valoran: los dispositivos de entrada y de almacenamiento, la profundidad de la clasificación por materias, la clase de consultas del usuario, la temática y la forma física de los documentos.

La evaluación de SRI hasta la llegada de la Web se centraba en las bases de datos en línea y en CD-ROM. Entre estos últimos destaca un estudio de Harry y Oppeheim (1993) en el que indican que estos sistemas han de presentar una información completa sobre la fuente, el contenido y la forma del producto, señalando especificaciones técnicas necesarias para su utilización. Plantean que la evaluación, en un gran número de casos, se lleve a cabo por especialistas, que han de realizarla aplicando criterios objetivos, y aunque son partidarios tanto de la participación de usuarios experimentados y no experimentados, apuntan que debido a problemas logísticos, de tiempo y de dinero, un trabajo de este tipo difícilmente puede desarrollarse, por lo que a menudo, resulta más práctico hacerlo el propio especialista.

Más allá de los criterios a valorar, estos autores señalan una serie de principios, que han de estar presentes en el método de evaluación y que nos parece interesante destacar:

- Consistencia, tanto en el procedimiento como en los resultados. Los procedimientos han de ser estándares, de forma que se puedan volver a repetir cada vez que necesitemos evaluar una base de datos. Los resultados también deben presentarse de forma normalizada.

- Eficaz uso del tiempo. La evaluación ha de desarrollarse en un espacio de tiempo corto y la preparación debe llevarse a cabo en un tiempo mínimo.

- Simplicidad. El método de estudio no ha de ser muy complejo. La simplicidad ha de afectar igualmente a la propia prueba, evitando amplios y complejos test.

- Objetividad. Debe basarse en criterios medibles o que puedan ser descritos. Los resultados tienen que ser producto de una síntesis estructurada de resultados empíricos.

- Flexibilidad. Permitirá adaptarlo a diferentes productos.

Estos estudios, en muchos casos van enfocados a dar a conocer mejor estas fuentes, sus características más importantes, la posibilidad de interactuación a través del interfaz y la información que facilitan a todo tipo de usuarios, lo que permite conocer si cubren todos los aspectos importantes de una disciplina de forma consistente. Por otro lado, los interesados en la adquisición de estos productos, necesitaban saber, a través de estu-

dios basados en principios objetivos y con información sobre los aspectos en que estaban interesados (como por ejemplo, su contenido, la facilidad de instalación y la facilidad de uso), saber si sus fuentes son fiables, su coste, así como su idoneidad. Dejando a un lado criterios de evaluación de carácter comercial que afectan a aspectos como su nombre, dirección, etcétera, es importante y útil, en estos sistemas, analizar lo que distintos autores denominan "consideraciones del producto" (Product considerations), que afectan a todo lo relacionado con: la base de datos, las capacidades de búsqueda, el interfaz de usuario, la documentación que les acompaña, las características para la manipulación de datos y la fiabilidad del producto.

Otra tendencia ha sido evaluar determinados aspectos de estos sistemas. Así, el grupo de trabajo SCOUG (Southern California Online User Group) propuso en 1990 los siguientes criterios para evaluar la calidad de las bases de datos: consistencia, alcance, cobertura, cobertura temporal, errores, exactitud, accesibilidad, facilidad de uso, integración, salida de información, documentación, asistencia al usuario, formación y relación calidad-precio (Abadal y Codina, 2005). La evaluación también se puede enfocar en valorar los mecanismos que intervienen en el funcionamiento de estos sistemas (Automatic Query, Expansion, Relevance Feedback, Ranking). Tiene que ver con los sistemas de recuperación interactivos y han sido estudiados en diversos trabajos presentados a las terceras y octavas Conferencias TREC.

En general, para la evaluación de SRI Meadow (1992) distingue dos tipos de medidas: señala en primer lugar las basadas en el proceso de búsqueda, con parámetros como la selección, el contenido, la traducción de una consulta, errores en establecimiento de la consulta, tiempo medio de realización de la búsqueda, dificultad de su realización, número de comandos precisos, coste de la búsqueda, número de documentos recuperados y número de documentos revisados por el usuario. Por otro lado se refiere a las que tienen que ver con la recuperación, que se basan en el análisis de los resultados obtenidos, contemplando medir la precisión, la exhaustividad, el promedio de efectividad E-P y medidas promedio de la satisfacción del usuario. (F.J. Martínez y J.V. Rodríguez, 2004a).

Lancaster y Warner (1993) proponen tres posibles niveles de evaluación de los sistemas de búsqueda, orientados tanto a valorar la efectividad, la relación efectividad-coste y la relación coste-beneficio. En relación con el primer nivel, los criterios a tener en cuenta tratan de valorar la satisfacción del usuario, y son los siguientes:

A) Criterio del coste. En el que se valoran aspectos como el coste por búsqueda, suscripción y documento junto a otros menos tangibles como el esfuerzo para conocer el uso del sistema, esfuerzo para efectuar la búsqueda, para entender los resultados, etcétera.

B) Tiempo. Se valoran aspectos como el tiempo que se tarda en acceder a las referencias, a los documentos, etcétera.

C) Calidad. Se tiene en cuenta la cobertura de la base de datos, la exhaustividad, la precisión, la actualización de los resultados, etcétera.

El segundo nivel está más relacionado con valorar la satisfacción del usuario en relación con la eficiencia interna del sistema y otras consideraciones de costes relacionados con cada uno de los recursos recuperados.

Finalmente, el tercer nivel trata de valorar la relación coste-beneficio de los recursos recuperados, que como estos autores indican, implica grandes dificultades al requerir una valoración económica de la información.

Baeza-Yates y otros (1999), se han ocupado de analizar los sistemas de recuperación de información teniendo en cuenta los diferentes momentos de su creación y desarrollo. De este modo señalan que estos sistemas suelen evaluarse ya antes de su lanzamiento y difusión. Suele realizarse una fase de análisis funcional, basada en el correcto funcionamiento de sus componentes, a la que deberían seguir diversas pruebas para detectar errores. Posteriormente es necesario realizar una evaluación del rendimiento del sistema en la que las medidas más utilizadas son el tiempo de respuesta y el espacio necesario para almacenar datos. Distinguen entre sistemas de recuperación de datos, que deben tener presentes estos aspectos, y así, valorar el funcionamiento de las estructuras de los índices, la interacción con el sistema operativo, los retrasos producidos por los canales de comunicación y los gastos de software. Los sistemas basados en la recuperación de información requieren, además, valorar la ordenación de los resultados por relevancia, es decir, la precisión del conjunto recuperado. En este caso se evalúa el rendimiento de la recuperación, para lo que es necesaria una colección de examen de referencia, que contenga la colección de documentos, un conjunto de preguntas y un grupo de registros relevantes, seleccionado por especialistas y relacionado con cada consulta. Respecto a la medida de evaluación, cuantifica la similitud entre los resultados recuperados por el sistema y los propuestos por los especialistas. Las medidas más utilizadas en este sentido son, una vez más, la exhaustividad y la precisión, aunque también se utilizan otras como La medida E, la Media Armónica, la satisfacción, el fracaso, etcétera.

Al evaluar el rendimiento en la recuperación, estiman que en primer lugar debemos considerar cuál va a ser la función o funciones a evaluar, ya que pueden variar la metodología y las medidas a aplicar. Distinguen entre la evaluación realizada en sistemas de procesamiento por lotes (in batch) y los interactivos. Si se trata de una búsqueda interactiva habrá que valorar el esfuerzo del usuario, el diseño de la interface, las ayudas y la duración de la sesión de búsqueda. En sistemas no interactivos lo más importante es la calidad de los resultados. Otro aspecto a considerar, además de lo referente a los temas de búsqueda y a la interface, es el contexto en el que se va a desarrollar la evaluación, es decir, si se trata de un experimento de laboratorio o basado en consultas reales. Respecto a la metodología, Baeza-Yates y otros (1999) señalan que la evaluación de SRI, inicialmente se diseñaba como una experiencia de laboratorio, analizando principalmente interfaces no interactivas. A partir de los años noventa, cobran mayor importancia los análisis basados en experiencias reales, si bien, las realizadas en laboratorio siguen teniendo importancia debido principalmente a que se pueden repetir y a la representatividad de los datos (scalability).

Peña y otros (2002), atendiendo a aspectos terminológicos como la eficacia y la eficiencia, se refieren a la necesidad de una doble evaluación en los SRI. Señalan que por un lado será necesario valorar el grado de satisfacción de los objetivos propuestos para el sistema, y por otro, tener en cuenta el costo, en tiempo y consumo de recursos, que conlleva alcanzar ese grado de satisfacción[202]. Desde este punto de vista, un sistema de recuperación de información es eficaz si recupera todos los recursos existentes en la base de datos que tienen que ver con la consulta; es decir, le caracteriza una gran exhaustividad en la recuperación, y por otro lado, el que recupera sólo recursos que son relevantes, con lo cuál el ruido es mínimo. En la eficiencia se valora: la velocidad de proceso, tiempo consumido en una búsqueda, la ayuda, la adaptación de interfaces al usuario, la disponibilidad del documento seleccionado y el idioma.

Estos autores señalan además, como alternativa a la evaluación de la eficacia, la realización de un análisis de la bondad del proceso de recuperación.

---

[202]. Peña y otros (2002:316) señalan como valores de la eficiencia de los sistemas de recuperación de información; el tiempo de ejecución, los requisitos de almacenamiento y la cantidad de memoria.

Como vemos, en función del punto de vista desde el que se aborde la evaluación, se utilizan diferentes medidas. En este sentido, Peña y otros (2002) proponen, según se trate de evaluar el sistema en sí: exhaustividad, precisión, fracaso y sus valores; el sistema desde el punto de vista del usuario: exhaustividad relativa, precisión del usuario, esfuerzo de exhaustividad, cobertura y novedad, o los procesos, proponiendo, en este caso, la evaluación de la realimentación.

Martínez y Rodríguez (2003) han sistematizado los diferentes aspectos que se han utilizado para evaluar SRI, valorando su necesidad y utilidad. Hacen un repaso de las medidas tradicionales, es decir tanto de las basadas en la relevancia (precisión, exhaustividad, tasa de fallo y factor de generalidad) como las orientadas al usuario (cobertura, novedad, exhaustividad relativa y esfuerzo de exhaustividad) y otras medidas alternativas, también denominadas de valor simple, que se basan en cálculos probabilísticos como el Modelo de Swet, basado en cálculos de probabilidad de que los registros recuperados correspondan a documentos relevantes, el Modelo de Robertson, similar al anterior en sus objetivos, el Modelo de Cooper, que trata de medir el ahorro de esfuerzo en la consulta de los listados de resultados de una búsqueda, las medidas SMART, etcétera. Como vemos se trata de métodos que utilizan la probabilidad para valorar la exhaustividad y la precisión, basándose en aspectos muy específicos. De la serie de medidas que analizan para medir la efectividad de la recuperación, presentan como valores positivos: el poder ser medidas de forma intuitiva, mediante cálculos simples; (la precisión; la exhaustividad; la cobertura y la usabilidad), junto a otras menos utilizadas o que no inciden en la valoración de la efectividad como son el formato de presentación, el contenido, el tiempo y el coste de la búsqueda. En todos estos valores, estos autores aprecian un cierto grado de subjetividad.

Respecto a la usabilidad, Harter y Hert (1997) la definen como la capacidad del sistema para facilitar una ejecución efectiva, eficiente y satisfactoria de la labor del usuario. Las medidas son: la exactitud, el porcentaje de errores, la recuperación, las percepciones del usuario sobre la facilidad de uso y su satisfacción. Como señala Johnson, posteriores investigaciones han analizado las relaciones de estas medidas con las capacidades cognitivas del usuario.

Jonhson y otros (2001) citan como aspectos a evaluar: la cobertura, la precisión, el tiempo de respuesta, la utilidad, la presentación de los resultados y el esfuerzo del usuario para obtener resultados satisfactorios. Dentro de la utilidad y valor de los recur-

sos consideran: la calidad, la consistencia de los resultados y los recursos inactivos, caducados y duplicados.

Abad (2005) sigue la tendencia que señala que la evaluación de estos sistemas en entornos reales, puede abordarse desde el punto de vista del propio sistema o del usuario. En el primer caso se trata de poner de manifiesto el rendimiento del sistema en la recuperación y establece como indicadores para ello la precisión, la exhaustividad, un análisis de fallos, en los que se tiene en cuenta diversos aspectos que tienen que ver con las necesidades de información, con la búsqueda, con el conocimiento del sistema y con la interacción del usuario con el sistema. Otros aspectos que deben valorarse son los fallos de precisión, de exhaustividad así como el ruido, el silencio, el índice de irrelevancia o de generalidad. El otro punto de vista debe permitir la adaptación del sistema a las necesidades y al entorno de usuarios al que presta su servicio.

Una de las propuestas más completas la ofrecen Abadal y Codina (2005:193) al agrupar los criterios en tres grupos en función de que tengan que ver con la base de datos (el contenido), con el sistema de recuperación de la información (el continente) o con la gestión o administración de la base de datos. En el primer apartado contemplan la necesidad de expresar el grado de exactitud y precisión, valorando aspectos como los errores gramaticales o de omisión, la fiabilidad de datos y los registros duplicados. Respecto al alcance y cobertura, se han de analizar estos aspectos tanto desde el punto de vista temático como geográfico y lingüístico, el grado de inclusión, la estructura, el tamaño y el nivel de crecimiento. Por otro lado ha de valorarse, respecto a la actualización, tanto el grado de ésta como el periodo que tarda en actualizarse. El último aspecto de la base de datos se refiere a la consistencia, que ha de ponderarse tanto en relación con la catalogación como con el análisis de contenido. En cuanto al sistema de recuperación de la información, los criterios a tener en cuenta son: las prestaciones del lenguaje de interrogación, la precisión, la exhaustividad, el tiempo de respuesta, la utilidad, los formatos de visualización y el interfaz. Por último, respecto a la gestión de la base de datos se debe evaluar su documentación, la atención al usuario, el precio y sistema de facturación y el sistema de distribución.

Desde nuestro punto de vista, estos autores son los que con mayor claridad señalan los problemas que puede plantear aplicar una metodología propia de los SRI tradicionales a los buscadores de la Web.

Así, teniendo en cuenta la opinión de distintos teóricos sobre el tema que señalan como requisito imprescindible para la evaluación, conocer los objetivos del sistema a evaluar, y dado que generalmente se trata de permitir recuperar información útil al usuario, que sólo él, en función de sus necesidades de información puede juzgar, nuestra evaluación se enfoca a la valoración, desde un punto de vista técnico, que permita interpretar si es correcto el funcionamiento de estas herramientas, sino también su utilidad en búsquedas especializadas, pudiendo el usuario, posteriormente calibrar de una forma personal, la utilidad de los recursos recuperados. Por lo que nos acercamos a Cooper (1973) para quien el objetivo de un sistema de recuperación es, o debería ser, recuperar documentos útiles y no solamente relevantes.

Respecto a los criterios, han de estar estrechamente relacionados con los objetivos de la evaluación, ya que puede ir enfocada a valorar una determinada parte del sistema, como puede ser la base de datos, o incluso un determinado parámetro, como por ejemplo la ordenación, etcétera. Además hay que tener en cuenta la metodología que se aplica, ya que muchas de estas propuestas van acompañadas de determinados medios, como pueden ser una determinada colección de búsqueda, unas determinadas herramientas que facilitan ese tipo de evaluación, etcétera.

Podemos apreciar que no todos los criterios son aplicables a los buscadores web, ya que, sobre todo en el aspecto de la gestión, es mucha la diferencia entre unos sistemas y otros. Lo mismo podemos decir de la consistencia o del grado de exactitud, ya que los registros no son creados por personas dependientes de la empresa que mantiene el buscador, sino que, de existir, los datos son introducidos por personas ajenas a la compañía que mantiene el buscador. Por eso es necesario realizar una propuesta de criterios más acorde con el medio en el que nos encontramos. En este sentido, deberemos tener en cuenta las recomendaciones sobre los parámetros a utilizar empleados por diferentes autores en sus estudios, muchos de ellos caracterizados por la aplicación de algunos de estos criterios. Será necesario analizar las distintas experiencias realizadas, aspecto del que nos ocuparemos en el siguiente apartado.

Además de valorar su adecuación a estos nuevos sistemas, deberemos seleccionar los criterios que nos permitan valorar nuestros puntos de interés, que como hemos apuntado, tienen que ver con el funcionamiento y con la utilidad de estas herramientas en búsquedas especializadas. Así, nos planteamos evaluar estos servicios porque ante tal variedad y número de buscadores, a menudo surgen dudas sobre cuál o cuáles pueden ser

los más idóneos, atendiendo a su correcto funcionamiento general, a su correcta recuperación mediante diferentes tipos de búsqueda, a su actualización, a su indización de los sitios web o de las propias páginas, a la menor recuperación de duplicados, de recursos no activos, etcétera. Cada uno de estos aspectos requerirá sus elementos de ponderación.

3.4. La evaluación de las herramientas de búsqueda de la World Wide Web. Estado de la cuestión e Indicadores utilizados

3.4.1. Estado de la cuestión

En el siguiente apartado nos ocuparemos de los diferentes trabajos publicados sobre evaluación de los buscadores con el fin de apreciar tanto los enfoques desde los que se han realizado, los criterios utilizados y la metodología y resultados, aspectos que, en la medida de lo posible, y siempre que persigan los mismos objetivos, trataremos de aplicar en nuestra evaluación.

Aunque hay un gran número de trabajos de evaluación sobre diferentes aspectos de los buscadores, no todos intentan aplicar las directrices que se trazan a partir de los estudios teóricos, echándose de menos en este campo, una metodología estándar propia al respecto, siendo necesario recurrir a experiencias de diferentes servicios y sistemas de información científica, para encontrar un método de evaluación que pueda ser aplicado a estas herramientas.

La evaluación de los motores de búsqueda se ha abordado desde diferentes perspectivas: la matemática (Mizzaro), la de ingeniería mecánica e industrial y la informática (Can, Nuray y Sevdik, Gwizdka y Chignell, Nasios y otros, Rousseau, Bharat y Broder, Brin y Page, Courtois y Berry, Figuerola, Zazo y Berrocal, Leighton y Srivastava, Picard y Saboy, Thelwall, Jansen y Pooch); desde el punto de vista de la biblioteconomía, la bibliometría, las ciencias de la información y documentación, (Aguillo, Bar-Ilan, Chu y Rosenthal, Notess, Johnson, Griffiths y Hartley, Ljosland, Martínez, Maldonado y Sánchez, Olvera, Oppenheim y otros, Snyder y Rosenbaum, Vaughan, Westera, Winship), así como los que desarrollan su labor en laboratorios de investigación o forman parte de grupos multidisciplinares (Lawrence y Giles, etcétera). De aquí que podamos afirmar que la evaluación de los buscadores tiene un marcado carácter interdisciplinar.

En uno o en otro sentido, los puntos de interés de los investigadores en recuperación de la información en la Web giran en torno a los motores y su efectividad; aspecto sobre el que existe un amplio número de trabajos que se presentan tanto en reuniones,

jornadas y congresos monográficos, publicándose en revistas especializadas y en la Web. Ejemplo de ello son las conferencias TREC. A partir de la séptima edición, celebrada en 1999 Hawking ha presentado importantes estudios que tratan de comparar los algoritmos utilizados en sistemas TREC con los utilizados en la Web. Otros autores se han ocupado de estudiar desde el punto de vista teórico aspectos relacionados con el tratamiento y desarrollo de nuevas técnicas y algoritmos de recuperación (como el agrupamiento de recursos relacionados, técnicas de filtrado, etcétera) así como con la clasificación de recursos manual y automática, la elaboración automática de índices y resúmenes, la búsqueda de elementos multimedia, el uso de agentes inteligentes y el desarrollo futuro y tendencias posibles, siendo la evaluación de los sistemas de recuperación uno de los temas más tratados.

Es difícil sistematizar el amplio número de estudios desarrollado a lo largo de los últimos años en torno a la evaluación de los buscadores de información de la Web, sin embargo, Gordon y Pathak (1999) señalan dos grandes grupos: por un lado el que contiene las evaluaciones que denomina testimoniales, y por otro, las cuantitativas. Las primeras son las que se publican en la prensa o por empresas relacionadas con la industria informática, que mediante test, comparan: la velocidad de recuperación de información de los motores, la facilidad de uso y el diseño de interfaces. También evalúan aspectos más técnicos como: la posibilidad de utilizar operadores en las ecuaciones de búsqueda, la valoración de los recursos que contiene la base de datos y el tiempo que les cuesta indizar nuevas páginas. En este sentido podemos señalar los trabajos de Morville, Rosenfeld y Janes (1996), Overton (1996), Courtois y otros (1996), Steinberg (1996) Calafia (1997), Slot (1997) y Lake (1997).

En el otro grupo de trabajos se compara la efectividad entre los motores de búsqueda, siguiendo los métodos tradicionales de evaluación que se proponen en recuperación de la información. En este grupo se encuentran los trabajos más importantes sobre el tema, que tratan de medir aspectos como la relevancia, la pertinencia o la interfaz de usuario. (Winship 1995, Ding y Marchionini 1996, Leighton 1996-1999).

Su (2003) en una brillante sistematización de los estudios de evaluación, establece dos etapas. La primera, que comprendería los años 1995 y 1996, caracterizada por una serie de trabajos enfocados a orientar y facilitar la elección a los usuarios, ocupándose de las características de los motores y capacidades de búsqueda, así como por la existencia de otros estudios, que aunque presentan una metodología más o menos elaborada,

no alcanzan resultados homogéneos, ya que varían los criterios de evaluación tanto en número como en tipo.

Critica la forma en que se desarrollan los test, ya que en muchos casos proceden de consultas reales, otras son elaboradas y otras responden a intereses personales o simplemente no se indica. El número de consultas para la evaluación varía entre dos y diez.

En resumen, pone de manifiesto la falta de una metodología de evaluación sistemática y común que contemple a los usuarios.

La segunda etapa, entre 1997 y 2000 se caracteriza por unos estudios con criterios mejor definidos y una metodología basada en criterios estadísticos y mayor sistematización.

En este periodo las medidas más empleadas son por este orden, la precisión, la validez de enlaces, el solapamiento, la exhaustividad, la cobertura, la ordenación por relevancia, teniendo en cuenta tanto el punto de vista del usuario como el funcionamiento del sistema. Otros valores usados son el tiempo de búsqueda y de respuesta, la actualidad, el coste, etcétera.

Martínez (2002) clasifica los estudios de evaluación en: explícitos e implícitos. Los primeros contemplan aspectos externos al motor como pueden ser: el tamaño del índice, audiencia, el mayor o menor uso para realizar búsquedas, la porción de la página indexada y la fidelidad. Los implícitos, junto a medidas basadas en la relevancia, se ocupan de analizar los enlaces fallidos, el solapamiento y el acierto único. Este autor se ha referido, además a la disparidad y dispersión tanto en los estudios de evaluación como en los resultados.

Hawking y otros (2001a), desde la perspectiva de las experiencias TREC, evalúan la calidad de los motores web. Para ello toman las preguntas de los logs de dos de ellos. Analizan veinte buscadores, de los cuáles dos son metabuscadores (MetaCrawler e Inquirus) y un directorio (LookSmart) mediante 54 consultas sobre diferentes temas. Se valoran los veinte primeros resultados. La relevancia, basada en el contenido textual, es juzgada por las personas que realizan la investigación utilizando valores binarios. Evalúan fundamentalmente la efectividad de los algoritmos de recuperación de documentos, valorando, en la lista de recuperación, la posición de los documentos relevantes. Utilizan sólo los enlaces activos, no penalizando la recuperación de recursos a los que no se puede acceder ya que esto les permite establecer comparaciones con resultados obtenidos en sistemas evaluados utilizando la metodología TREC. No obstante se apartan de ella en

otros aspectos como el prescindir del uso de una colección estándar. Los resultados muestran diferencias en la recuperación entre los diferentes motores, y respecto a la precisión de los cinco primeros resultados, observan que no está en relación con la mayor cobertura del índice de determinados buscadores. Los mejores buscadores utilizan algoritmos cuya efectividad se aproxima a la de los sistemas TREC.

Metodológicamente estos autores señalan la dificultad de realizar evaluaciones que puedan ser repetidas ante la inexistencia de una colección estándar de recursos web que lo facilite. Señalan además que se ha de tener en cuenta al plantear la evaluación, las necesidades de información (según se trate de una pregunta directa, búsqueda de un recurso único, de una selección de documentos o de una necesidad de recuperación exhaustiva de recursos sobre un tema) ya que, en función de ellas, la técnica utilizada será diferente. Igualmente señalan la recomendación de repetirla cada cierto tiempo, así como la necesidad de establecer unas medidas que sirvan de media y como punto de comparación.

Desde nuestro punto de vista, se trata de una metodología muy interesante pero plantea problemas por su carácter restrictivo, al centrar su interés fundamentalmente en valorar la efectividad, sin analizar otros aspectos que nos parecen deben tenerse en cuenta como son el solapamiento, valoración de enlaces fallidos, actualización, etcétera.

Coincidimos con estos autores en que es difícil, si no imposible, establecer una metodología única para evaluar los motores de búsqueda, ya que los estudios se realizan desde diferentes puntos de vista, tanto en función de la formación del evaluador, (especialista en ciencias de la computación, especialista en Documentación, bibliotecario, etcétera), como de la orientación, dependiendo de si el objetivo es valorar la satisfacción del usuario o plasmar hasta qué punto es correcto su funcionamiento. Estas necesidades pueden ser a su vez tan específicas que, dependiendo de ellas, puede establecerse una metodología concreta. Es normal que esto haya sido así, y que cada uno haya optado por aplicar una determinada perspectiva y los conocimientos e incluso fórmulas propias del área de investigación a la que se dedica.

En función de tanta variedad podemos afirmar que prácticamente existen tantos métodos de evaluación como trabajos se han realizado.

Sin embargo, una perspectiva de diez años en el funcionamiento y evaluación de estos motores sí puede darnos unas pautas y unos indicadores suficientemente representativos para poder evaluarlos, pudiendo seleccionar los criterios más importantes que sirvan

tanto a usuarios y desarrolladores, y que en definitiva, nos ayuden a cumplir los objetivos de nuestra evaluación.

Exponemos a continuación, ordenados cronológicamente, los trabajos más interesantes relacionados con la evaluación de los buscadores, ya que alguno de ellos es de obligada referencia, tanto por su importancia a nivel teórico como metodológico. En el caso de autores que han realizado diferentes trabajos de evaluación, hemos preferido presentarlos unos a continuación de otros para observar mejor, en su caso, la evolución metodológica utilizada en cada uno de los trabajos.

Uno de los primeros trabajos es el de Winship (1995), quien comparó los motores World Wide Web Worm, WebCrawler, Lycos, Harvest y los directorios Galaxy y Yahoo. Analizó tanto el contenido y la formación de la base de datos, valorando si indizaban las direcciones URL, los títulos, los resúmenes, el texto completo así como el tamaño y la posibilidad de envío de direcciones por parte de usuarios, la interface, las opciones de búsqueda, la recuperación, y sus prestaciones, tendiendo en cuenta además, los mecanismos de control por parte del usuario, el contenido de los registros y el control sobre el modo de ordenación, considerando en este sentido, el número de documentos recuperados y su presentación. Lycos y Harvest obtuvieron los mejores resultados.

Lycos y OpenText fueron los buscadores que obtuvieron mejor puntuación en la evaluación realizada por Courtois, Baer y Stark (1995) que trataba de averiguar cuáles de los motores existentes (CUIW3, Catalog, Harvest, Lycos, OpenText, WebCrawler, W3Worm y Yahoo) recuperaban recursos fundamentales referidos a tres temas de búsqueda. También valoraron positivamente a WebCrawler por su rapidez y flexibilidad para elaborar la consulta.

El mismo año comenzaron a realizarse los estudios de Leighton (1995) basados en la precisión y tiempo de respuesta utilizando los diez primeros registros. No analizan la exhaustividad, dado el tamaño de la Web y la disposición desestructurada de la información. Valoran la relevancia, los recursos únicos y la validez de los enlaces (enlaces activos, duplicados, existencia de copias en mirrors, etcétera) de InfoSeek, Lycos, WebCrawler y World Wide Web Worm mediante ocho búsquedas. Lycos e InfoSeek obtuvieron la mejor valoración en cuanto a tiempo de respuesta y precisión, calculando esta última en función de su adecuación a los términos de búsqueda. En 1997 desarrolló estos estudios junto a Srivastava, analizando los veinte primeros resultados de quince consultas lanzadas sobre AltaVista, HotBot, Excite, InfoSeek y Lycos, siendo AltaVista, Excite e

InfoSeek los mejor valorados, mientras que HotBot respondía mejor a consultas estructuradas, frente a Lycos que destacaba en búsquedas no estructuradas y breves.

Finalmente, en el trabajo de 1999 se analizaron los primeros veinte recursos obtenidos en quince búsquedas lanzadas sobre AltaVista, Excite, HotBot, InfoSeek y Lycos. Valoran los recursos asignándoles diferente categoría según se ajusten a la búsqueda y cumplan las perspectivas esperadas por el usuario. En esta evaluación se centran en valorar la capacidad de los motores de búsqueda para colocar el mayor número de documentos relevantes entre los veinte primeros resultados.

De las quince consultas, diez son preguntas de referencia realizadas por estudiantes, cuatro fueron extraídas de un estudio anterior (Leighton y Srivastava, 1997) y una se ocupa de localizar información sobre una determinada persona. Las preguntas son de tres tipos: siete están realizadas mediante una mera acumulación de palabras; otras siete son estructuradas y la última contiene el nombre de la persona sobre la que se busca información. Establecieron cuatro categorías de relevancia. Se centran en la valoración de este aspecto porque consideran que para los estudiantes universitarios es más importante que la exhaustividad.

Los mejores resultados se dieron en AltaVista, Excite e InfoSeek, destacando Lycos en búsquedas simples y no estructuradas y HotBot en las búsquedas estructuradas. Llegan a la conclusión de que evaluar la relevancia es uno de los mayores problemas a los que se enfrentan los estudios sobre motores de búsquedas, ya que en este estudio demuestran la diferente puntuación que puede adquirir un motor según lo que se entienda por relevancia. Insisten en la necesidad de establecer una metodología correcta para la evaluación objetiva de los motores de búsqueda, de manera que no se les perjudique ni beneficie.

Lycos y OpenText obtuvieron los mejores resultados en el trabajo realizado por Ding y Marchionini (1996) en el que se evaluaron tres buscadores, mediante cinco preguntas, valorándose relevancia y el solapamiento, en función de los veinte primeros resultados. Otros aspectos que se tuvieron en cuenta fueron la base de datos, la calidad del índice, la funcionalidad y la usabilidad. En los resultados no se detectaron grandes diferencias entre unos motores y otros.

En el mismo año, un estudio bastante limitado, de D. B. Meghabghab y G. V. Meghabghab (1996) analiza la efectividad, por medio de la precisión, de Yahoo, Info-

seek, Lycos, Excite y WebCrawler, obteniendo los tres primeros, por este orden, los mejores resultados.

Leonard (1996) analizó la exactitud de resultados, facilidad de uso y posibilidades de la búsqueda avanzada mediante quince búsquedas en diecinueve motores de búsqueda, llegando a la conclusión de que AltaVista era el motor que mejores resultados ofrecía.

Davis (1996) se centró en el tamaño del índice y otros aspectos relacionados con la recuperación de la información, observando los mejores resultados en AltaVista, HotBot e InfoSeek.

Ese año, también Slot (1996) evaluó dieciséis buscadores centrándose en el tiempo de respuesta y la interface, siendo AltaVista y Yahoo los que ofrecían mejores resultados. Un análisis específico de las posibilidades de búsqueda sitúa también a AltaVista como el mejor valorado.

Es importante el trabajo de Stobart y Kerridge (1996) sobre la fidelidad del usuario al sistema de búsqueda, señalando que los aspectos por los que mantienen su fidelidad son, por este orden: la velocidad, el tamaño y la costumbre.

Chu y Rosenthal (1996) compararon la capacidad y el rendimiento de los motores AltaVista, Excite y Lycos. Para valorar el primer aspecto, utilizaron la lógica booleana, truncamientos, búsquedas por campo, palabra o frase. Para analizar el rendimiento se basaron en la cobertura, la precisión, el tiempo de respuesta, el esfuerzo del usuario y la presentación de los documentos. Utilizaron preguntas realizadas por usuarios a un servicio de referencia, observando que AltaVista superaba en estos aspectos a los otros dos motores.

Venditto (1996) evaluó AltaVista, InfoSeek, Lycos, OpenText, WebCrawler y World Wide Web Worm mediante un importante número de temas de búsqueda a lo largo de dos semanas. Se estudió la relevancia de los veinticinco primeros resultados. La efectividad la calculó teniendo en cuenta en la recuperación la existencia de los sitios web más interesantes y conocidos relacionados con cada tema. Los mejores resultados se alcanzaron en búsquedas simples, siendo mayores las diferencias en las búsquedas complejas. InfoSeek dio los resultados mejores en cuanto a relevancia y AltaVista en cuanto a la exhaustividad.

Tomaiuolo y Packer (1996) lanzan doscientas búsquedas sobre Magellan[203], Point, Lycos, InfoSeek y AltaVista midiendo la relevancia sobre los diez primeros, obteniendo AltaVista los mejores resultados. Este trabajo ha sido criticado por no indicar las expresiones de búsqueda ni el concepto de relevancia en que se basa. (Ming, 1998). Posteriormente publicaron otro trabajo sobre AltaVista y OpenText en el que trataron de valorar la exhaustividad.

Schlichting y Nilsen (1996) comparan la calidad de los diez primeros resultados emitidos en varias búsquedas sobre temas especializados. Se comparan AltaVista, Excite, InfoSeek y Lycos. La calidad se basa en criterios de los usuarios según la utilidad y relevancia de los recursos recuperados. Puntúan de 1 a 7 los recursos relevantes. Pero para no perder el contexto de la búsqueda, utilizan el procedimiento denominado Signal Detection Analysis[204]. El método utilizado trata de medir la relevancia, aunque dado que en este caso, la evaluación se realizó cuando las bases de datos de estos buscadores no contenían un número de páginas indizadas suficientes para recuperar al menos diez, se optó por eliminar términos y añadir los nuevos recursos recuperados al listado. Esto unido al lanzamiento de frases sólo en los buscadores que lo permitían (AltaVista e InfoSeek), pone en cuestión la metodología utilizada por falta de consistencia.

El propio autor pone en entredicho la aplicación de este método en los motores, para lo que es necesario una mejora de la tecnología de búsqueda de estas herramientas. De hecho, no se han realizado experiencias posteriores utilizando esta metodología.

Zorn y otros (1996) evaluaron opciones avanzadas de recuperación como la búsqueda booleana, búsqueda por campo, los operadores de proximidad y los truncamientos en AltaVista, InfoSeek, Lycos y OpenText, analizando además la calidad de su documen-

---

[203] Clausurado en 2001 tras ser adquirido en 1996 por Excite.

[204] Se basa en agrupar los recursos en cuatro grandes grupos, el de los relevantes, irrelevantes, enlaces rotos relevantes y enlaces rotos irrelevantes, aspecto éste que se comprueba con la recuperación del mismo recurso por otro motor de búsqueda que mantiene el enlace activo. A continuación se determina la tasa de éxito (*hit rate*) y falsa alarma (*false alarm*). El primero indica la proporción entre recursos buenos recuperados por un motor y los recuperados por los cuatro buscadores. El segundo valor es la proporción de recursos no útiles y el total de recursos inútiles recuperados por los cuatro buscadores. Aplican diferentes fórmulas a los resultados mediante las que calculan dos valores: por un lado, la sensibilidad del motor, en cuanto a que puede distinguir entre enlaces buenos y malos, y por otro el valor que trata de expresar el carácter conservador de los motores, entendido como la perdida de recursos, con el fin de reducir al mínimo el número de falsas alarmas. Se centran en medir tanto la recuperación de información útil y el carácter conservador o liberal de los motores en la elección de recursos que integran su base de datos.

tación, la cobertura, en la que destacan AltaVista y Lycos, y la profundidad de indización, siendo parcial sólo en Lycos.

Clarke y Willett (1997), siguiendo la metodología de las experiencias TREC, compararon la efectividad de AltaVista, Excite y Lycos en treinta búsquedas sobre temas relacionados con la investigación en el campo de la Información, utilizando los diez primeros resultados. Desarrollaron un método en el que compararon la exhaustividad relativa, la precisión y la cobertura[205] en los diez primeros resultados. Como Chu y Rosenthal, establecieron tres valores para medir la relevancia, en función del contenido de las páginas y siguiendo los criterios que se expresan a continuación:

-Si la página tiene relación cercana con lo que se busca, la puntuación es 1.

-Si la página consiste en enlaces, y no en la información que se solicita, se valora en 0,5 si uno de los dos enlaces es útil.

-A los duplicados con el mismo URL y el mismo contenido se les da un valor 0.

-Los duplicados localizados en *mirrors* (espejos) que contienen diferente URL y un mismo contenido, no se consideran duplicados y se valoran como únicos.

-Los no encontrados indican la falta de actualización del índice y por ello se valoran con 0 puntos.

-En los que aparece el mensaje de "no responde" debido a la caída del servidor o no responde en ese momento, se busca la página posteriormente y en caso de no poder conectar se le da el valor 0.

-Las páginas en lengua no inglesa, dada su dificultad en la valoración, se reemplazaban por el siguiente documento.

Los resultados, en cuanto a la precisión, arrojaron los siguientes datos: una precisión que va desde el 0,25 en Lycos al 0,46 en AltaVista. La exhaustividad relativa va de 0,56 en AltaVista a 0,66 en Excite, siendo la diferencia poco significativa. La cobertura dio peores resultados para Lycos que para Excite que a su vez obtuvo peor resultado que AltaVista. Este, como venía siendo habitual, dio los mejores resultados.

[205]Su cálculo se halla dividiendo el número total de páginas relevantes localizadas por un motor entre el número de páginas relevantes recuperados por el resto de motores (Olvera Lobo, 2000).

Otros investigadores que han utilizado la metodología TREC pero de forma adaptada son Hawking, Craswell, Griffiths y otros, quienes en diferentes trabajos publicados entre 1997 y 2001 aplican cambios como puede ser la valoración de la relevancia utilizando valores binarios y la selección de consultas basándose en las que realizan directamente los usuarios. En el último de los trabajos citados se evalúan veinte motores mediante cincuenta y cuatro consultas y se analizan los veinte primeros recursos, eliminándose los no activos. La relevancia es valorada por un equipo de universitarios ayudándose de un programa denominado Relevance Assessment Tool que podemos traducir como Herramienta de Valoración de la Relevancia.

Su (1997) proporciona una nueva metodología basada en usuarios y consultas reales así como la intervención de éstos en las valoraciones. Un año después realiza una experiencia piloto junto a Cheng y Dong (1998) en la que participan once universitarios que realizan diferentes búsquedas en AltaVista, InfoSeek, Lycos y OpenText[206] y valoran la relevancia, las características del sistema, la interacción, la validez de los resultados y el rendimiento, basándose en los criterios de relevancia, eficiencia, satisfacción del usuario, utilidad y aspectos de la conexión. Se basan en los veinte primeros resultados. En 1999 repitieron la experiencia con treinta y seis usuarios. En 2003 publicó nuevos estudios, en los que destaca la sistematización de los criterios de evaluación.

Lebedev publicó entre 1996 y 1997 dos trabajos en los que valora la utilidad de los motores de búsqueda comparándolos con bases de datos especializadas (INSPEC y CAS). Llega a la conclusión de que las bases de datos son más completas al ofrecer un mayor número de resultados relacionados con los términos de búsqueda, y que sólo entre el diez y el veinte por ciento de los resultados ofrecidos por los motores corresponde a publicaciones de carácter científico. Este trabajo se basa en búsquedas con un sólo término y contempla como indicadores: el número de resultados recuperados, estimaciones sobre la capacidad del índice y la tipología de las publicaciones. Estos estudios le llevaron a afirmar que cuanto mayor es la base de datos, más posibilidades hay de no encontrar lo que se busca. Aunque cabría matizar que esta afirmación es defendible teniendo en

[206] Actualmente en desuso.

cuenta que las herramientas a las que se refiere cuentan con importantes limitaciones en lo que a recuperación de información se refiere.

Bar-Ilan (1998) utilizó en sus búsquedas términos específicos. Este trabajo, basado en la comparación de seis motores de búsqueda, recuperó 6.681 registros referentes al matemático húngaro Paul Erdos, sobre los que se calculó la precisión, solapamiento y la exhaustividad estimada, obteniendo como resultados una alta precisión, mínimo solapamiento y baja exhaustividad. Oppenheim y otros (2000) critican el que estos resultados no pueden aplicarse a otros tipos de búsqueda realizadas por el mismo motor.

Bharat y Broder, en un trabajo publicado en 1998 estudiaron la cobertura, centrándose en el cálculo de la base de datos y el solapamiento de AltaVista, Excite, HotBot e InfoSeek. Realizaron las mediciones a mediados y a finales de 1997 mediante más de diez mil consultas cada vez. Los resultados indicaban que el número de páginas indizadas por HotBot giraba en torno a setenta y siete millones, cien millones para AltaVista, treinta y dos millones en Excite y diecisiete millones en InfoSeek. Respecto al solapamiento, se aprecia entre ambos periodos una variación mínima que va desde el 0,9 al 1,4% en al segunda toma de datos. Algo más significativa es la diferencia en cuanto a la cobertura. Un estudio posterior confirmó los datos iniciales.

Megahaghab y otros (1998), compararon la efectividad en Yahoo, WebCrawler, InfoSeek, Lycos y Excite. Parten de cinco temas de búsqueda hasta dar forma a cincuenta tipos de búsqueda según una mayor o menor especificidad. Los resultados dieron una mayor precisión a Yahoo tanto en búsquedas iniciales como filtradas.

Nasios y otros (1998) demostraron que AltaVista y HotBot eran los mejores buscadores, seguidos de InfoSeek y Excite. WebCrawler y Lycos obtuvieron peores resultados. HotBot destaca en la búsqueda por frase. Excite actuó de modo uniforme en todas las modalidades mientras que InfoSeek mostró carencias en búsquedas booleanas. Aún así los resultados no fueron muy distintos unos de otros.

Tunender, H. y Hervin, J. (1998) estudiaron otros aspectos como el tiempo que tardan los buscadores Yahoo, Excite, InfoSeek y AltaVista en indizar una página, tras el envío de la dirección URL a los motores. InfoSeek tardó tan sólo un día en indizar la página del primer nivel, necesitando once días Yahoo y AltaVista. Excite lo hizo en veintitrés días, si bien, a diferencia de otros, indizó en el mismo tiempo varios niveles y Lycos no lo hizo en los cuarenta y seis que duró la experiencia. Al mismo tiempo analizaron hasta qué nivel indizaban, destacando en este sentido, Excite, que indizó seis de nueve

niveles posibles. Le sigue Yahoo con tres y AltaVista e InfoSeek con dos. Observaron además que las etiquetas Meta no fueron indizadas.

Por otro lado, este último trabajo midió también la actividad del robot, valorando cuándo visitaba el sitio web. AltaVista lo hacía a los veintiocho días de iniciar el estudio. InfoSeek a los diecisiete y diecinueve días y posteriormente a los treinta y ocho, a los cuarenta y uno, y a los cuarenta y cuatro, aunque sin añadir datos al índice de páginas de otro nivel inferior. Lycos a los veintisiete días, pero no indizó nada, ya que nada se recuperó. Excite a los diecisiete días, haciéndolo posteriormente casi a diario. También observó que, determinados días, algunos recursos no se recuperaban.

El estudio que Wishard realizó en 1998, como ella afirma, más que una evaluación estadística de la precisión de los resultados de los buscadores, supone una interpretación de las herramientas y de su utilidad en el campo de la Geología. Para llevarlo a cabo seleccionó 37 buscadores entre directorios, motores y metabuscadores sobre los que lanzó tres búsquedas. En sus resultados analiza: la precisión de los diez a quince primeros resultados; la pertinencia, en función de la exactitud de la información; la ordenación; el grado de exhaustividad y los registros únicos. Los mejores resultados respecto a la precisión se observaron en Excite e InfoSeek. No obstante entre sus conclusiones señaló la imposibilidad de recomendar una única herramienta de búsqueda, ya que no hay ninguna que destaque ampliamente sobre las demás

Rousseau, R. (1998/99) comparó en un estudio realizado diariamente, durante veintiuna semanas, el incremento de las bases de datos de AltaVista y Northen Light, para llegar a la conclusión de que AltaVista actualizaba diariamente sus bases de datos aunque observó períodos con mayor acopio de recursos.

Xie, M., Wang, H. y Goh, T.N. en sus trabajos publicados entre 1998 y 1999 evaluaron la calidad de los motores utilizando el modelo SERVQUAL, desarrollado por Parasuraman, Zeithaml y Berry que se basa en la valoración de cinco puntos, éstos son: la calidad de los aspecto tangibles, la fiabilidad, la rapidez, la seguridad y la empatía con el usuario. Para ello se sirve de la valoración del usuario obtenida a partir de diferentes encuestas. Se trata por tanto de una metodología basada, más en la opinión de los usuarios respecto a diferentes aspectos de los motores de búsqueda, que en la valoración de resultados de búsqueda, que es la tendencia más utilizada en evaluación.

Gordon, M. y Pathak, P. (1999) evaluaron el comportamiento de los buscadores AltaVista, Excite, InfoSeek, OpenText, Hotbot, Lycos y Magellan mediante treinta y tres

consultas reales pero estructuradas por especialistas en recuperación de información. Para medir el comportamiento de estas herramientas utilizaron como valores la exhaustividad, la precisión y el solapamiento. En la evaluación se utilizaron los veinte primeros registros. La valoración de la relevancia la realizaron los propios demandantes de información en una escala de cuatro valores. Como resultado se obtuvo una baja efectividad y la existencia de diferencias en cuanto a la precisión entre los motores y no tanto en cuanto a la exhaustividad. Además se observó un solapamiento mínimo entre motores.

La constante aparición de nuevas herramientas de recuperación de información hizo necesaria la realización de trabajos encaminados a analizar el funcionamiento de los buscadores, sus características y prestaciones. En algunos casos (Lawrence y Giles, 1998a), (Kuk, 1999) y otros, dio lugar al desarrollo de trabajos en los que se analizaba un determinado motor para compararse con los resultados ofrecidos por otros. En este contexto se desarrolla el primer trabajo de Lawrence, S y Giles, L. (1998) en el que analizan la exhaustividad y cobertura obteniendo unos resultados limitados.

En un segundo trabajo (1998b) se ocuparon de la cobertura, actualización, solapamiento y enlaces válidos de seis motores (AltaVista, Excite, HotBot, InfoSeek, Lycos y Northern Light) mediante quinientas setenta y cinco preguntas, basándose en los seiscientos primeros resultados.

Pero tal vez el trabajo más consistente fue el de 1999, donde se ocuparon de la accesibilidad a la información en la Web mediante los motores de búsqueda. Estiman en ochocientos millones el número de páginas de acceso público existente en la Web. Basándose en el análisis de dos mil quinientos servidores, calcularon que el 6% contenía información de carácter científico. Por otro lado observaron una baja utilización de metadatos y más aún de sistemas como Dublin Core, utilizado tan sólo por un 0,3% de las principales páginas web, pudiendo influir todo ello de manera negativa en la recuperación de información de alta calidad.

Analizaron las prestaciones de AltaVista, Euroseek, Excite, Google, HotBot, InfoSeek, Lycos, MSN Microsoft, Northern Light, Snap[207] y Yahoo, mediante mil cincuen-

[207] Fue clausurado en a comienzos del 2001.

ta búsquedas. Eliminaron las páginas que no contienen los términos de la búsqueda y en Northern Light, los registros de la Colección Especial.

Calcularon la cobertura, el solapamiento, la actualización de la base de datos y la indización de la Web alcanzada por los motores, advirtiendo que, respecto a un estudio anterior, en el que se calculaba en una tercera parte la indización de la web visible, ahora era sensiblemente menor.

El solapamiento se mantuvo a un bajo nivel. La cobertura estimada, fruto de combinar los resultados de varios buscadores fue del 42%, cifra que puede mejorarse mediante metabuscadores como MetaCrawler. Analizaron también la cobertura de los motores respecto al tamaño estimado de la web, obteniendo los mejores resultados, por este orden: Northern Light, Snap y AltaVista. La menor cobertura se observó en Lycos y Euroseek.

Otros aspectos de los que se ocuparon fueron la indización y la actualización, analizando además de registros no válidos, el tiempo que tardan en indizar páginas nuevas o actualizadas, estableciendo en ciento ochenta y seis la media de días que tarda un motor en indizar estas páginas. Los que más tardaron fueron Snap y Yahoo, y los que menos Northern Light, InfoSeek y AltaVista.

Gwizdka y Chignell (1999) siguieron el planteamiento de Cleverdon (1966), que proponía como criterios de evaluación la cobertura, el tiempo de búsqueda, la exhaustividad, la precisión, la presentación y el esfuerzo del usuario, aunque desestiman el tiempo de respuesta y la exhaustividad, el primero por estar sujeto a variaciones en función del estado de la red y la segunda por la constante variación de la información en la Web, por su gran volumen y por su carácter dinámico, aspectos, que dificultan su medición.

Además, proponen para la precisión valorar la utilidad de los enlaces que contienen los recursos recuperados, utilizando lo que denominan “usefull precision”. Otros valores para la precisión, son la precisión total, que valora la estimación objetiva otorgada a los registros recuperados, la mejor precisión (best precision), que tiene en cuenta sólo los recursos más relevantes, y la precisión objetiva, calculada en función de la existencia en los documentos de los términos de consulta. Calculan además el ranking, el esfuerzo de usuario y la cobertura, valorando, en este sentido, el número total de recursos, la cobertura relativa y el solapamiento.

Compararon AltaVista, HotBot e Infoseek en relación con la recuperación en el ámbito comercial y de organizaciones de seis dominios, de los cuales, cuatro son de dife-

rentes países (Alemania, Austria, Polonia y Reino Unido) y los otros dos responden a los dominios comercial (.com) e institucional (.org). Utilizaron cuatro temas de búsqueda, en diferentes idiomas en función del dominio. Los mejores resultados en cuanto a precisión y cobertura, los alcanzó AltaVista. Otros resultados de interés señalaron un mal funcionamiento del algoritmo de ranking y un bajo solapamiento.

En un trabajo publicado el mismo año M.H. Chignell, J. Gwizdka, y C. Bodner, (1999) trataron de demostrar mediante dos nuevas evaluaciones, los cambios en la recuperación en función de que las búsquedas se lancen en distintos días y horas, teniendo en cuanta además la cobertura geográfica y de diferentes dominios de Internet. En el primer caso utilizaron los motores Excite, Infoseek y HotBot y en el segundo AltaVista, Infoseek y HotBot. Esta segunda evaluación tuvo en cuenta la cobertura geográfica, y temática analizando los recursos recuperados de Alemania, Austria, Polonia y Reino Unido, así como los recursos con dominio relativo a instituciones (.org) y los de carácter comercial (.com). Como en el trabajo anterior, valoraron distintos tipos de precisión. Como conclusión establecen la variación de los resultados en la recuperación dependiendo del contexto.

Maldonado Martínez, A. y Fernández Sánchez, E. (2000) evaluaron las posibilidades de búsqueda que soportan los buscadores Yahoo, Excite, Lycos, InfoSeek, AltaVista, Hotbot, Herdworld, AOL Netfind y Northern Ligth, valorando: si soportan diferentes tipos de búsqueda booleana; si permiten acotar la búsqueda; si contienen un directorio y si permiten la búsqueda por campos, la visualización de índices y el control de vocabulario. Los resultados indicaron que Northern Light es el que más posibilidades abarcaba, tanto en general como en cuanto a aspectos relacionados con la recuperación de información, entre los que se valora: la posibilidad de búsqueda dentro de un conjunto, la ordenación temática de registros, la posibilidad de realizar búsquedas por campos, la visualización de los índices de los campos existentes y contar con herramientas de control de vocabulario. Estas autoras publicaron en 1998 un estudio descriptivo de los principales buscadores desde el punto de vista documental, atendiendo tanto a la recogida y análisis de la información como a la búsqueda y resultados. Analizaron los buscadores AltaVista, Excite, Lycos, WebCrawler, HotBot, InfoSeek, los índices Magellan, Galaxy, LookS-

mart, Yahoo y los de orientación hispana, Ole[208], ¿Dónde?[209], Ozu.com[210], Ozu.es[211], Elcano[212], Biwe[213], Hispavista[214], Trovator[215], Tarantula[216] y Sol[217], de los que, estos tres últimos son motores.

Westerra (2000) evaluó las interfaces de búsqueda de AltaVista, Google y HotBot diferenciando entre capacidades básicas y especiales. AltaVista destacó en ambas, mientras que Google sólo lo hizo en las básicas y HotBot en las especiales.

Uno de los estudios de mayor relieve es el publicado por Johnson, Griffiths y Harley (2001) en el que proponen un marco para la evaluación de motores de búsqueda desde el punto de vista de la satisfacción del usuario, señalando como puntos a analizar: la efectividad, la eficiencia, la utilidad de los resultados y la interacción con el sistema. Para ello se proponen medidas como la precisión y el ranking, una serie de valores basados en el tiempo de respuesta y, respecto a la recuperación, una serie de medidas que tratan de poner de relieve la utilidad del sistema, como son, la valoración de los enlaces y de la recuperación en general. Respecto a la interacción del usuario con el sistema, se valoró la satisfacción con la interface teniendo en cuenta las posibilidades de formulación de la consulta, modificación y visualización. Finalmente, respecto a los resultados, se analizaron aspectos, como la posibilidad de manipulación y la visualización.

Los trabajos de evaluación más recientes tratan de especializarse en determinados aspectos. De este modo, Amat (2002:337) ha analizado la formación de la base de datos, la indización y la recuperación de diecisiete sistemas españoles de recuperación de información distribuida en Internet, valorando su similitud en la formación de los índices con el esquema Dublín Core, las opciones y mecanismos de recuperación de los sistemas analizados y su cobertura relativa, concluyendo que "no se pueden considerar válidas

[208]http://www.ole.es
[209]http://www.donde.uji.es
[210]http://www.ozu.com
[211]http://www.ozu.es
[212]http://www.elcano.es
[213]http://biwe.cesat.es
[214]http://www.hispavista.com
[215]http://trovator.combios.es
[216]http://www.tarantula.com.mx
[217]http://www.sol.es

"islas de información filtrada" ya que "ofrecen escaso acceso a poca información insuficientemente representada".

Griesbaum (2004) ha evaluado la efectividad de tres motores de búsqueda: AltaVista, Google y Lycos, todos ellos en su versión alemana[218]. Como viene siendo habitual, analizan los veinte primeros resultados de un total de cincuenta consultas. En lo metodológico, para valorar estas herramientas, parten de una colección tanto de registros como de consultas, contemplando además los criterios y medidas para valorar la relevancia. La valoración de este aspecto se lleva a cabo por un jurado compuesto por veintinueve personas. Los temas de consulta se extraen del log de dos buscadores, distintos del evaluado, y en su caso, se traducen al alemán. Google obtuvo los mejores resultados, seguido muy de cerca por Lycos.

Un trabajo de Vaughan y Telwall (2004) ha estudiado la cobertura en la recuperación de recursos de carácter comercial de distintos países (Estados Unidos, China, Singapur y Taiwan) de Google, AltaVista y AllTheWeb, constatando una mayor recuperación de recursos de Estados Unidos debido al modo de trabajo de los programas que forman las bases de datos.

Otro aspecto que en la actualidad ha cobrado una gran importancia es el análisis del ranking destacando los trabajos de Courtois y Berry (1999) y de Vaughan y Thelval (2004).

De la importancia que tienen los estudios de ranking, del que más adelante nos ocuparemos, es fiel reflejo el realizado por Vaughan (2004) en el que se estudia el comportamiento de Google, AltaVista y Teoma. Es interesante desde el punto de vista metodológico ya que critica el uso de los valores de precisión y exhaustividad y proponen en su lugar medidas como la correlación entre el ranking valorado por personas y el ranking del motor de búsqueda a lo que denomina "calidad del ranking de resultados" (Quality of result ranking). Frente a la exhaustividad propone valorar la recuperación de páginas situadas en los primeros lugares del ranking. El buscador que obtuvo los mejores resultados fue Google.

[218] http://www.AltaVista.de, http://www.Google.de y http://www.Lycos.de

Del contenido de todos estos trabajos se desprende la existencia de una dispersión de criterios en la evaluación, que puede ser fruto de la necesidad por parte de los investigadores de buscar nuevos indicadores, como es el caso de Stobart y Kerridge (1996) quienes proponían, además de los valores tradicionales, valorar la fidelidad del usuario, o Leonard (1996) y Zorn y otros (1996), que se ocuparon de las posibilidades de las búsquedas avanzadas. Davis (1996) se interesó por el tamaño del índice, Slot (1996) por el tiempo de respuesta y la interface, mientras que Thunender, H. y Erwing, J. (1998) lo hicieron sobre el tiempo que tardan en indizar una página y el nivel jerárquico del sitio Web al que descienden los robots.

Sobre la tendencia a utilizar aspectos aislados en la evaluación de los motores de búsqueda, hay que señalar que se mantiene hasta nuestros días. Basta con citar los trabajos de Stimson (1999) dedicados a comparar la recuperación de nombres de empresas en motores de búsqueda comerciales y no comerciales, obteniendo en ambos buenos resultados, por lo que recomienda no descartar ni unos ni otros en este tipo de búsquedas. Los primeros pueden ofrecer información recopilada de diferentes fuentes, mientras que los motores no comerciales recuperan recursos con información de la web de la empresa y otras páginas de información general. Hay que decir también que los buscadores comerciales facilitan búsquedas de información más específica gracias a sus avanzadas opciones.

Por otro lado hay que señalar una intensa corriente de evaluaciones basadas en el cálculo de la relevancia de los motores de búsqueda que se plasma en los trabajos de Ding, W. y Marchionini, G. (1996), Leighton, H. V. (1996), Leighton, H. V. y Srivastava, J. (1997 y 1999), Venditto, G. (1996) y Tomaiuolo, N. G. y Packer, J. G. (1996).

No sólo por el tratamiento teórico previo, sino también por su metodología destaca el trabajo de Chu, H. y Rosenthal, M. (1996) que proponen una selección de criterios (cobertura, exhaustividad, precisión, tiempo de respuesta y esfuerzo de usuario) donde se compaginan los dos puntos de vista desde los que se han venido haciendo los trabajos de evaluación de Sistemas de Recuperación de la Información. Este trabajo ha influido claramente en la realización de otros experimentos de gran interés como los desarrollados por M. Gordon y P. Pathak (1999), S. Lawrence y L. Giles (1998-1999).

Respecto a las evaluaciones en las que se aplica una metodología que hace especial incidencia en expresar también el punto de vista del usuario, destacamos los trabajos

de Su, L. T. (1997 y 2003) así como los de Johnson, F. C., Griffiths, J. R. y Hartley, R. J. (2001).

Existe otra tendencia, llevada a cabo por autores como G. Notess[219], D. Sullivan[220] o I. Aguillo[221], que en distintos sitios Web, recogen y publican datos, que actualizan continuamente, sobre determinados aspectos del funcionamiento de los buscadores más importantes. Estos sitios constituyen un lugar de referencia obligado para conocer determinados aspectos y características tanto de las herramientas de búsqueda como de la recuperación de información en la Web, aunque en el caso de Aguillo, sus técnicas están más relacionadas con la Cibermetría.

Respecto a la metodología, los distintos trabajos demuestran que no existe un método estándar sino que, la mayoría de trabajos utiliza un procedimiento y criterios distintos así como medidas variables. Aún en el caso de experiencias de tanta importancia como las desarrolladas en las conferencias TREC, la metodología empleada está continuamente sujeta a variaciones y a continuas críticas. Por ello, para elaborar una evaluación consistente, pensamos que es conveniente detenernos en valorar la metodología utilizada en los trabajos más relevantes.

En este sentido hemos de referirnos a uno los trabajos de Oppenheim, C., Morris, A. y Macknight, C. (2000), en el que se sistematizan los estudios más destacados dentro de cuatro grupos:

1. Los que utilizan el método de Cranfield, cuando se conoce la existencia de un pequeño número de páginas referentes al tema de búsqueda, que además son conocidas por el investigador. Se contemplan en este grupo los trabajos de Both Delezar-Tiedman y el Proyecto Erdos llevado a cabo por Bar-Ilan en 1998.

2. Los que usan el método anterior más el cálculo de la exhaustividad relativa. Este parámetro se calcula sumando los recursos relevantes recuperados en una serie de búsquedas, considerando que representa el universo de recursos relevantes en comparación con los registros recuperados por un motor sobre el que se ha lanzado la búsqueda. Este valor ha recibido críticas por parte de Fricke (1998) y otros autores.

[219]http://searchenginesshowdown.com
[220]http://www.searchenginewatch.com

Estudios de este tipo son los de Chu y Rosenthal (1996), Ding y Marchinioni (1996), Gauch y Wang (1996), Tomaiuolo y Packer (1996), Westerra (1996), Leighton y Srivastava (1997), Clarke y Willett (1997), Gonçalves y otros (1998), Megahaghab y otros (1998), caracterizándose por utilizar términos más amplios de búsqueda.

3. Los que utilizan el método Cranfield más el cálculo de la exhaustividad estimada, basada en una estimación estadística del número de probables recursos existentes en la web.

En este grupo tenemos los trabajos de Bharat y Broder desarrollados en los años 1998 y 1999 y los de Lawrence y Giles (1999).

4. Los que eliminan el cálculo de la exhaustividad. Sitúa en este apartado los trabajos de Feldman (1998) que analiza y estudia las características de AltaVista, Excite, Lycos, HotBot, InfoSeek, Northern Light, SavvySearch, Inference Find y AskJeewes. Kimmel (1996), incide en la cobertura y en las características de los registros. Cita también los trabajos de Tunender y Ervin sobre indización y recuperación, en los que se aprecia una insuficiencia e inconsistencia en la indización de motores como Lycos, InfoSeek, AltaVista, Yahoo y Excite.

Se hallan también en este grupo los trabajos que tratan de establecer alternativas a la precisión en la evaluación de los motores, basándose en el cálculo de la amplitud de búsqueda estimada (ESL o Estimated Search Lenght) así como los de Agata y otros (1997).

A pesar de todo, Oppenheim, Morris y Mcknight (2000) se refieren a la poca consistencia de los trabajos de evaluación de los motores de búsqueda en la Web, así como a la imposibilidad de aplicar el método de Cranfield, dada la dificultad de calcular la exhaustividad, lo que impide una correcta comparación de resultados obtenidos por unos investigadores y otros.

Estos autores señalan además, como hecho a tener en cuenta en la evaluación, que los motores de búsqueda, del mismo modo que la Web, están cambiando continuamente, por lo que los resultados de las evaluaciones son válidas por un periodo de tiempo limitado, y que, como gran parte de los investigadores reconocen, los resultados que sus

[221]http://www.cindoc.csic.es/cybermetrics

trabajos ofrecen, son indicativos de las prestaciones de los motores de búsqueda en el momento en que se realizan. Esto indica que este tipo de evaluaciones son efímeras, por lo que han de repetirse cada cierto tiempo y actualizarse de forma constante para comprobar si los resultados han variado, valorar el dinamismo de la Web o si han sido corregidos los posibles problemas detectados. De aquí la importancia que tiene el diseño de un método que facilite una valoración objetiva, basada en criterios científicos, y que de un modo fácil, permita su repetición periódica.

Johnson y otros (2001) recogen en su trabajo alguno de los problemas que plantean las evaluaciones de motores web realizadas siguiendo el método de Cranfield, refiriéndose además a las dificultades en la utilización de valores como la exhaustividad y la precisión, la falta de estandarización en las medidas que afectan a estos valores en unos experimentos y otros, lo que dificulta la comparación entre diversos estudios e incluso las diferencias de concepto de relevancia que se aplican. Finalmente, se refieren a la necesidad de estudiar los motores para que las búsquedas no favorezcan a unos y otros, haciéndolas de un modo estándar. Llegan a la conclusión de que hay que valorar cuáles son los criterios que pueden interesar al usuario en relación con los motores de búsqueda, analizando el impacto que tienen en sus valoraciones aspectos como la velocidad de proceso, la calidad de los resultados, etcétera.

Podemos ver hasta aquí las tendencias de evaluación que se han desarrollado hasta la fecha. En ellas se observa la clara influencia de estudios de evaluación de Sistemas de Recuperación de la Información anteriores y la importancia de la opinión del usuario en la valoración tanto de los resultados como de la interface y opciones de búsqueda. Pero a pesar de todo, no debemos olvidar que se trata de ámbitos de recuperación diferentes por lo que los indicadores de unos y otros se han ido distanciando.

Debemos estudiar y proponer un método que nos permita comparar diversas herramientas para valorar su utilidad, seleccionar las mejores y, si es posible, que permita comparar los resultados con los de otros estudios, al menos para conocer su evolución. Para ello deberemos seleccionar los criterios necesarios que nos permitan alcanzar los objetivos propuestos. En función de ellos seleccionaremos y propondremos los indicadores de evaluación que nos resulten necesarios.

3.4.2. Propuestas de evaluación e indicadores de los motores de búsqueda de la Web

Es un hecho que hay una serie de criterios comúnmente aceptados por la comunidad científica que nos permiten contrastar las capacidades de los SRI, pero es evidente que los buscadores de la Web, por sus características y las de su entorno, requieren una adaptación de aquellos, además de la utilización de otros más relacionados con el propio funcionamiento de estas herramientas.

La mayoría de autores han intentado aplicar técnicas y métodos de evaluación de estos sistemas de recuperación a las herramientas web, pero no debemos olvidar que en el primer caso se trata de sistemas que contienen bases de datos con información estructurada, acomodándose perfectamente a los documentos que indizan, mientras que en la Web, son herramientas que surgen para facilitar la localización de información en un medio donde ésta aparece de forma poco estructurada y, a menudo, con un escaso valor informativo. Por tanto, es ésta una de las diferencias por las que es necesario plantearse si dichas técnicas, métodos e incluso los criterios de valoración que acabamos de ver, son aplicables a la evaluación de las herramientas web. Por otro lado, los avances técnicos han podido influir en que algunos criterios como el "tiempo de conexión" hayan quedado desfasados o, en otros casos, como en el de la exhaustividad, muy utilizada en sistemas más estructurados y especializados como pueden ser las bases de datos en soporte CD-ROM, no se adaptan a estos sistemas.

Nos ocupamos a continuación de las propuestas elaboradas por diferentes autores que se han ocupado de ofrecer, de una forma sistemática, una serie de criterios a tener en cuenta en la evaluación de los motores web.

Una de las primeras propuestas en este sentido es la realizada por David Jakob (1995) que propone los siguientes aspectos a valorar:

1. Facilidad de uso.

2. Rapidez en las búsquedas.

3. Posibilidad de realizar búsquedas básicas y complejas con operadores booleanos.

4. Búsquedas mediante partes de una palabra y truncamientos.

5. Utilización de búsquedas por frase y por términos próximos.

6. Permitir al usuario dar mayor importancia a un término de búsqueda.

7. Control por el usuario de búsquedas con términos en mayúsculas y minúsculas.

8. Utilización de tesauros.

9. Permitir al usuario indicar el máximo número de registros a recuperar.

10. Indización a texto completo, mejor que sobre determinadas partes del recurso (aunque el usuario pueda controlar sobre qué campos limitar la búsqueda).

11. Proporcionar información que incluya el título y el URL del recurso.

12. Ofrecer una fácil interpretación de los resultados con marcadores de la relevancia o un listado ordenado según su importancia.

13. Indicar cuándo fue indexado el recurso.

14. Actualizar la base de datos periódicamente para eliminar recursos no activos y caducados.

15. Permitir al usuario registrar direcciones URL no incluidas en la base de datos.

En este caso, más que de criterios, este autor presenta una serie de características que estas herramientas deberían cumplir. En la actualidad, dada su evolución, la mayoría de aspectos han sido superados.

Podemos apreciar la gran importancia que se da a todo lo relacionado con la búsqueda y recuperación, analizando también algunos aspectos del funcionamiento como la indización y la actualización de la base de datos, aunque olvida elementos como la formación de ésta y la ordenación.

Koch (1996) desde un punto de vista muy general señala siete grandes aspectos que una evaluación de servicios de búsqueda debe analizar, a saber: el tamaño, la cobertura, la actualización, la formación, la indización, la recuperación, y el interfaz de usuario.

Chu y Rosenthal (1996) proponen para las herramientas de la Web los siguientes criterios de evaluación:

1. Composición de los índices, ya que en su elaboración radica el éxito de una búsqueda en un determinado motor. Exige un conocimiento de: la cobertura, la frecuencia de actualización y la parte de la página web sobre la que se realiza la indización.

2. Capacidades de búsqueda, de modo que se analicen: las prestaciones de los distintos motores de búsqueda en relación con operadores booleanos, la búsqueda por frase, el truncamiento, el filtrado (como puede ser la búsqueda por campos). Cuantas más opciones posibiliten, mejor podremos realizar la búsqueda y mejores resultados obtendremos.

3. Ejecución de la recuperación de información, mediante la valoración de la precisión, la exhaustividad y el tiempo de respuesta.

4. Obtención de resultados. Ha de abordarse desde dos puntos de vista: por un lado, desde las opciones que ofrecen los motores de búsqueda, y por el otro, desde el modo en que presentan los resultados, y la información que contienen, es decir, si ofrecen un extracto incompleto, o si los presentan de forma más elaborada, conteniendo un pequeño resumen, junto a la disposición de otros elementos informativos de interés. Esto puede darnos una idea de como trabajan los motores de búsqueda.

5. Esfuerzo por parte del usuario. A juicio de estos autores, la documentación e instrucciones de uso que ofrecen y la interface juegan un papel importante al influir por parte del usuario de forma determinante en la selección de un buscador.

Estos criterios se corresponden con los propuestos por Schwartz y Maldonado, de carácter más específico y que han sido recopilados por García Giménez (2002):

a) Cobertura, que comprende las páginas web a las que tiene acceso, ámbitos geográficos y de contenido de la base de datos y métodos de recogida y análisis de los documentos web.

b) Formularios de búsqueda, analizando el modo en que se plantean para permitir realizar búsquedas más precisas, facilitar la búsqueda a los no iniciados y posibilidad de realizar búsquedas complejas.

c) Búsqueda por campos en títulos, descripción, URL, palabras clave, localización, idioma, tipos de información y tipos de propietarios.

d) Herramientas de búsqueda, contemplando si el tipo de búsqueda es por palabras clave o mediante clasificación, truncado, operadores booleanos, términos compuestos, búsqueda por frase, proximidad, filtrado por lengua, fechas, tipo de archivo, etcétera.

e) Clasificación temática y control del vocabulario. Es decir, valorar el uso de categorías y fórmulas de control del vocabulario.

f) Detección de novedades.

g) Respecto a los resultados, valorar las posibilidades de modificar su presentación y ordenación.

h) Finalmente proponen valorar el nivel de adaptación al usuario y de recepción de sus ideas y comentarios.

Notess (1997), centrándose en las bases de datos, expone los aspectos que permiten comparar un buscador y otro:

a) Tamaño. Una de las medidas que se puede utilizar es el número de direcciones URL contenidas en el índice, aunque este aspecto puede ser engañoso ya que hay sistemas centrados en recoger un gran número de estas direcciones pero que no indizan exhaustivamente el contenido de las páginas a las que apuntan, como Lycos. Tampoco el tamaño en megas de sus bases de datos puede ser utilizado con suficientes garantías, ya que pueden ocupar mucha memoria pero tener poca información. Tal vez más fiable sea la valoración del número de páginas web indizadas, pero teniendo en cuenta que pueden contener un gran número de duplicados. Finalmente hay que ser conscientes de que las páginas web pueden desaparecer, modificarse o reubicarse sin que se refleje en la base de datos de forma inmediata.

b) Disponibilidad y duplicidad de las páginas. No todos los recursos que se recuperan pueden ser consultados ya que han podido ser cambiados de posición, modificando su URL, o eliminados, sin que haya un reflejo instantáneo de estos cambios. Los duplicados pueden obedecer al gran número de recursos que son señalados por otras páginas, hacia las que muy frecuentemente se dirigen los robots, y también a la existencia de una misma información en diferentes servidores, o en diferentes partes de un mismo servidor.

c) Solapamiento de resultados. Notess señala que el solapamiento en los motores generales no es muy elevado, a pesar de que cada uno de ellos forma su base de datos utilizando sus propios recursos y programas.

Para Abad (2002:673), es necesario para poder emitir un juicio de valor, establecer una comparación del producto con una expectativa de resultado articulada del modo más objetivo posible. Para ello es necesario indicar unos criterios o normas. Esta autora señala además que:

> "el establecimiento de criterios supone la definición de los atributos o acontecimientos respecto de los que se va a juzgar o evaluar el objeto en cuestión (en nuestro caso el sistema o la unidad de información) y la determinación de los indicadores o variables que reflejan dicho atributo. Así pues, es el punto determinante para una correcta evaluación."

Ljosland (2000), contemplando medidas cuantitativas y cualitativas, propone como medidas básicas para comparar motores: los ratios de precisión y exhaustividad en un determinado número de registros, la cobertura, ya sea absoluta o relativa[222] de la base de datos y el porcentaje de enlaces no activos. Entre las medidas de carácter cualitativo indica las siguientes: las opciones de búsqueda disponibles, las ayudas, la información contenida en los registros que recupera, la actualización del índice y su rendimiento.

Oppenheim y otros (2000), recogen un variado número de criterios que han sido utilizados por los investigadores en los trabajos de evaluación de motores de búsqueda. Entre los más utilizados figuran los siguientes:

- Tamaño de la Web y cobertura de los motores de búsqueda.
- Actualización de la base de datos y número de enlaces inactivos.
- Relevancia.
- Sintaxis de búsqueda.
- Materiales de mayor interés y formulación de la ecuación de búsqueda.
- Naturaleza cambiante de la red.
- Tiempo de respuesta.
- Diferentes características del sistema.
- Opciones de búsqueda.
- Factores humanos y problemas de interface.
- Calidad de los *abstracts*.

Por su parte, y ante la inconsistencia de las evaluaciones que se venían realizando, señalan los siguientes aspectos a evaluar:

1. Precisión.
2. Exhaustividad relativa (proponiendo el uso del método de Clarke y Willett[223]).
3. Velocidad de respuesta.

[222]Se calcula dividiendo el número total de URLs recuperadas por un motor por el total de URLs recuperadas.
[223]Véase Clarke y Willett (1997).

4. Consistencia de los resultados en un amplio periodo de tiempo.
5. Proporción de enlaces no activos o caducados.
6. Registros duplicados.
7. Calidad de los resultados estimada por los usuarios.
8. Evaluación de la interface por el usuario.
9. Ayuda y variedad, dependiendo de usuarios expertos o inexpertos.
10. Opciones de visualización de registros.
11. Presencia de publicidad.
12. Cobertura (siguiendo el método de Clarke y Willett[224]).
13. Amplitud estimada del motor.
14. Amplitud y legibilidad de los resúmenes.
15. Efectividad del motor (siguiendo el método de Back y Summers[225]).

Criterios que se pueden evaluar y contrastar mediante tres tipos de búsquedas: una por palabras simples, otra por frases y una tercera que utiliza el operador booleano OR o AND.

Defienden además que los resultados sean examinados por usuarios expertos e inexpertos en el uso de estas herramientas. Como vemos, los criterios se acercan más a las características de los motores de búsqueda, pero metodológicamente se trata de un tipo de evaluación más compleja ya que requiere la existencia de una colección de búsqueda y una serie de juicios de valor de expertos a contrastar con las valoraciones. Otro de los elementos sometidos a crítica es la consistencia de los resultados para un largo periodo de tiempo, dado el carácter cambiante de la Web e incluso del funcionamiento de estas herramientas.

Para Martínez y Rodríguez (2003) el uso de los índices de relevancia, exhaustividad y precisión, por sí solos, no permiten llegar a conclusiones definitivas sobre las prestaciones de los motores de búsqueda. Proponen además, el uso de otras medidas que tengan que ver con el contexto de la Web, señalando, en este sentido: la ratio de enlaces fallidos, el grado de solapamiento, el acierto único y la cobertura del motor. En este trabajo analizan además, los criterios utilizados en los diferentes trabajos de evaluación,

[224]Ibid.
[225]Este trabajo no ha podido ser localizado ya que no se había publicado cuando lo citó Oppenheim.

distinguiendo entre estudios implícitos y explícitos. Estos últimos se ocupan de valorar aspectos externos, formales o testimoniales. Se centran en valores como la amigabilidad de la interface, velocidad de respuesta, formatos de presentación, documentación existente y ayuda del sistema. Las evaluaciones implícitas o experimentales utilizan parámetros que someten a determinados test como son las basadas en el análisis de la relevancia. No obstante, dados los problemas que la web plantea para valorar tanto la exhaustividad como la precisión, se han de tener en cuenta otros parámetros como la ratio de enlaces fallidos, el grado de solapamiento, el acierto único y la cobertura del motor.

Debemos destacar la importancia que para los desarrolladores de estos sistemas han ido adquiriendo los aspectos que tienen que ver más directamente con la recuperación de información, y en este sentido, hay que destacar los criterios de relevancia y ordenación. Son conscientes de que gran parte del éxito de estas herramientas radica en un buen funcionamiento de los algoritmos que se utilizan.

Nos ocuparemos a continuación de recoger y comentar algunas de las propuestas sobre distintos aspectos y criterios a aplicar en la evaluación, la mayoría planteadas por los autores a los que nos hemos referido en el estado de la cuestión. Posteriormente nos ocuparemos de analizar de forma individual los criterios más utilizados tales como la relevancia, la ordenación o *ranking* y el solapamiento.

Respecto a la relevancia, como hemos visto al tratar de la evaluación de los SRI, varios son los puntos de vista desde los que se puede abordar la evaluación de estas herramientas, lo que, en función de los objetivos que se persigan, será necesario seleccionar unos criterios de valoración u otros

Un gran número de estudios que se ocupan de la evaluación de los recursos lo hace desde el punto de vista del usuario, que es quien, efectivamente, puede valorar la utilidad de la información recuperada, pero si nos basamos sólo en este punto de vista, dejaremos de conocer otros aspectos relacionados con las causas que determinan esos resultados, con el funcionamiento, etcétera. De aquí la importancia de las evaluaciones basadas en más aspectos llevadas a cabo por especialistas. Como señalan Salvador y Angós (2000:55):

> "el buscador [...] sólo puede valorar si los documentos o referencias bibliográficas coinciden con la demanda de información hecha por el usuario y con la estrategia de búsqueda ejecutada, determinando si son relevantes o no. En el caso del usuario, éste valorará si los documentos satisfacen su necesidad de información, determinando si son relevantes o no."

Por tanto, hay una relevancia basada en la satisfacción de la necesidad de información y otra, medible por el especialista en recuperación de información, que tiene en cuenta la coincidencia de los resultados obtenidos con los solicitados por el usuario y con la estrategia de búsqueda.

Los estudios desarrollados en torno a la relevancia se centran tanto en la valoración de este aspecto en los SRI como en su desarrollo conceptual, tratando de aclarar su significado, los diferentes puntos de vista y su utilidad. Su importancia es fundamental, ya que se utiliza en el cálculo de los dos valores que tradicionalmente se vienen utilizando para expresar la exhaustividad y la precisión en la evaluación de los SRI.

De aquí que el estudio de la relevancia haya sido una constante en la evaluación de estos sistemas, especialmente cuando se quiere reflejar su efectividad. En este sentido, trata de expresar hasta qué punto un sistema cumple su principal objetivo, esto es, recuperar recursos que satisfagan la necesidad de la búsqueda. Su estudio se aborda desde un doble punto de vista, por un lado el que trata de reflejar la capacidad de un sistema para recuperar los recursos que se asemejan a la expresión de búsqueda, y por otro, el que trata de medir la satisfacción del usuario respecto a la información recuperada. De aquí que, diversos autores (Soergel 1976, Reid 2000, etcétera) traten de valorar la relevancia en función de si estos sistemas permiten al usuario conseguir su objetivo o necesidad de información.

Para Abad (2005:145), se trata de un elemento complejo y difícil de definir al ser éste un concepto multidimensional, "ya que su significado depende en gran medida de las percepciones del usuario y de sus necesidades de información", y dinámico "porque la relación entre la información recuperada y el problema informativo que motive la búsqueda en un momento determinado puede variar con el paso del tiempo". Esta autora también distingue entre relevancia orientada al sistema, valorable en la medida en que "los términos expresados en la búsqueda coincidan con los que están presentes en los documentos recuperados o en sus representaciones" y relevancia de usuario, de carácter más subjetivo.

Desde el punto de vista teórico, Saracevic (1988a, 1996) y Mizzaro (1997) han estudiado a fondo este concepto. El primero establece un marco para su interpretación y señala la multitud de factores que pueden influir en su valoración, como: el conocimiento sobre determinada materia, la literatura existente sobre el tema, el propio documento, el objeto de la búsqueda, el sistema de información, el entorno, los valores de quien la juz-

ga, etcétera. Debemos señalar además, que la relevancia en la recuperación también depende de la expresión de búsqueda y de la pericia de la persona que lanza la búsqueda. Por ello señala cuatro grandes marcos desde los que ha de ser valorada: el sistema, la comunicación, aspectos psicológicos y cognitivos.

Mizzaro (1997) se ocupa tanto del concepto de relevancia, sobre el que señala la diversidad de matices y variantes que contiene este término, como de ofrecer un completo estudio histórico de trabajos sobre el tema, insertándolos dentro de tres periodos: el primero anterior a 1958, el segundo entre 1959 y 1976, caracterizándose por relacionar el documento y la consulta, y el tercero entre 1977 y 1997, con un valor más subjetivo y por tanto variable.

Lancaster (1998) da una definición amplia del concepto de relevancia al señalar que todos los usuarios de un sistema de recuperación tienen una exigencia fundamental en común: esperan que el sistema sea capaz de recuperar uno o más documentos que satisfagan su necesidad de información (documentos relevantes). Y añade que es posible expresar cuantitativamente, mediante la razón de exhaustividad, el grado de éxito del sistema en la recuperación de la literatura relevante[226] de la base de datos.

Para este autor es muy difícil establecer la exhaustividad real de los sistemas, por lo que hay que calcular la exhaustividad estimada.

Estos problemas se repiten de una forma más crítica en los SRI de la Web, debido a su carácter cambiante, siendo difícil determinar cuál es el número de documentos relevantes que contiene un sistema. Por eso, también aquí sólo se puede hablar de exhaustividad relativa. El procedimiento consiste en lanzar una búsqueda sobre varios buscadores y reunir los considerados relevantes, para luego valorar si son recuperados por uno u otro buscador y si aparecen entre los diez primeros resultados. Esto permite cono-

[226]Para Lancaster, un documento relevante es el que contribuye a satisfacer las necesidades de información del usuario (documentos pertinentes) y un documento irrelevante es el que no las satisface. En un trabajo de 1979 definía la pertinencia como la relación entre un documento y la expresión de búsqueda juzgada por el usuario, y la relevancia como la misma relación pero juzgada por un usuario externo. La razón de exhaustividad de la recuperación se halla dividiendo el número de documentos relevantes recuperados por el sistema por el número total de documentos relevantes contenidos en el sistema y multiplicándolo por cien. Por tanto, la exhaustividad representa el porcentaje de documentos relevantes obtenidos en una búsqueda de entre el total de documentos relevantes existentes en la base de datos.

cer en qué medida los recursos relevantes recuperados inicialmente por el grupo de buscadores se recuperan por cada uno de los motores. Esta operación es la que realizan autores como Clarke y Willett (1997) para obtener la exhaustividad estimada.

Dado que este valor se basa en la relevancia respecto a la ecuación de búsqueda, es decir valorando básicamente la coincidencia de términos, es necesaria la utilización de otro valor: la razón de precisión[227], para corregir la posible existencia de un gran número de recursos irrelevantes. Este valor indica el porcentaje de documentos relevantes sobre el total de documentos recuperados.

Lo ideal sería que un sistema alcance unos valores de exhaustividad y precisión del 100%, lo que indicaría que el sistema recuperaba todos los documentos relevantes y que, efectivamente todos ellos cumplen esta condición. Pero esto no es posible, ya que como Lancaster demostró, cuando se quiere aumentar un valor, por ejemplo conseguir una mayor exhaustividad, es normal que la precisión disminuya, y a la inversa, pues para obtener una mayor precisión, se realizan búsquedas más concretas, lo que incide en el valor de la exhaustividad. Esta teoría ha sido contestada por Fugmam (1993) aunque su planteamiento ha sido rechazado de nuevo por Lancaster. Actualmente tiende a valorarse más la precisión puesto que supone para el usuario un ahorro de tiempo y esfuerzo (Chowdury, 1999: 2007), aunque hay que decir que esto puede no ser del todo cierto, pues puede haber búsquedas en las que interese la exhaustividad, aún en un medio que soporta tanta cantidad de recursos como es la Web.

De aquí que podamos observar cómo las necesidades específicas del usuario tienen una gran influencia en la valoración de estas variables, y pueden restar objetividad a la evaluación. De aquí la utilización de colecciones cerradas, y la aplicación de criterios concretos a la hora de evaluar estas herramientas.

En cualquier caso, antes de aceptar la utilización de estos criterios en la evaluación de buscadores Web, hemos de tener en cuenta que se trata de un medio distinto de los sistemas tradicionales, ya que no son bases de datos centralizadas; la expresión de búsqueda es libre y la suele realizar directamente el propio usuario; y las opciones de

[227] La razón de precisión se halla dividiendo el número de documentos relevantes recuperados por el sistema por el número total de documentos recuperados por el sistema y multiplicándolo por cien.

búsqueda por campos y otras herramientas para facilitar búsquedas más precisas como tesauros, etcétera, apenas existen. Estas herramientas han sido sustituidas por otras opciones y mecanismos de búsquedas implementados de forma individual por cada buscador.

Leighton y Srivastava (1997) ya señalaron el problema de la excesiva valoración del criterio de la precisión a la hora de estimar cuál es el mejor buscador, indicando además que evaluar la relevancia sin la intervención del evaluador es uno de los mayores problemas de los estudios sobre evaluación de motores.

Gwizdka y Chignell (1999) aportan un nuevo elemento al cálculo de la relevancia ya que mantienen que puede expresarse a través de diferentes valores, que tratan de tener en cuenta tanto aspectos objetivos como subjetivos, la existencia de los términos de búsqueda, y sobre todo, teniendo en cuenta el entorno de los buscadores web, valorando la utilidad de los enlaces de los recursos recuperados.

Otro aspecto que tiene que ver con la utilización de estos valores es el hecho de que no hay acuerdo sobre cómo ha de valorarse la relevancia, si de forma binaria, que es como defienden Large, Tedd y Hartley, o mediante una serie de valores que para Chu y Rosenthal son tres, para Gwizdka cuatro, y cinco para Su, Ding y Marchionini, y otros.

Large, Tedd y Hartley (2001:282) recogen también diferentes críticas al respecto, señalando que por si solos, no son suficientes para evaluar la recuperación de la información, pues desde su punto de vista, accesibilidad y facilidad de uso son los factores que más influyen para elegir una fuente de información. Señalan que, como ya demostrara Cooper, el usuario tiene mayor interés por recuperar suficientes registros pertinentes a la búsqueda que obtener una gran exhaustividad.

Mónica Landoni y Steven Bell (2000) se han referido a la inadecuación de técnicas clásicas de evaluación de la Recuperación de la Información al aplicarse a sistemas interactivos. Llegan incluso a hablar de situación caótica de la evaluación de los motores de búsqueda de la Web.

Para ambos, los estudios centrados en la evaluación de la relevancia, concretamente en los criterios de precisión y exhaustividad son ya clásicos. Por eso introducen en su estudio otros criterios basados en el carácter interactivo y centrado en el usuario de los sistemas Web, proponiendo como aspectos a valorar, la validez (usability/usefullness) y la satisfacción. Dado que según estos autores la satisfacción sólo puede ser valorada mediante la participación del usuario, los criterios que proponen para medir la relevancia y

utilidad son, en relación con el primer aspecto: la precisión, la llamada, el ranking de relevancia y la cobertura. Respecto a la utilidad, proponen como valores, la accesibilidad y el ruido. Recomiendan además tener en cuenta otros aspectos relacionados con la realización de un análisis del motor de búsqueda en el que se contemple la información del servicio, sus características, descripción de la base de datos, modos y facilidades de búsqueda que ofrece, características de los registros recuperados y otros servicios.

En sus conclusiones, señalan la existencia de dos comunidades: por un lado la de los especialistas en Recuperación de la Información, ocupados en el estudio de nuevos criterios y medidas para la evaluación y la Comunidad Web, en la situación caótica definida anteriormente, fundamentalmente por la variedad de criterios que se utilizan en la evaluación. Dada su coincidencia en fines y objetivos, es lógico que ambas colaboren.

Otro problema es que tampoco hay unanimidad respecto a quién debe juzgar la relevancia. Para Green (1995) es el usuario, pero no siempre. Además ¿Cuándo se ha de realizar el juicio de la relevancia? ¿En los resultados de la consulta? ¿Una vez leído el documento? ¿Una vez podamos decidir si es útil para solucionar el problema?

La relevancia valorada por el usuario, como señala Schamber, puede variar en función de la especialidad de la persona que busca, ya que puede ser que para un especialista, un gran número de documentos recuperados no sean particularmente relevantes. De aquí el marcado carácter subjetivo que estas valoraciones llevan consigo. No debemos olvidar que en una recuperación efectiva intervienen diversos factores, tales como la pregunta, el conocimiento de la herramienta de búsqueda, etcétera. Además en la emisión de un juicio de este tipo, intervienen factores de tipo cognitivo, y pueden variar en función del grado de conocimiento de la persona en el momento en que se valora un recurso.

Ahondando en este aspecto, Green (1995) define la relevancia como la propiedad de un documento de servir potencialmente de ayuda a un usuario en la resolución de una necesidad. De aquí que en estos términos sólo pueda ser valorada por la persona que necesita la información. Además, la razón de juzgar relevante o no relevante puede ser variada dependiendo de la autoridad, autoría, etcétera. Si queremos utilizar el concepto para evaluar el funcionamiento de los motores, deberemos utilizar la relevancia formal o del sistema.

Un aspecto más a añadir a la dificultad de su medición es que los resultados pueden ser muy distintos si aplicamos opciones de búsqueda avanzadas, búsquedas en campos determinados, etcétera, o simplemente activando alguna de las opciones de que

disponen. Es decir, el carácter interactivo de estas herramientas facilita la consecución de resultados más precisos. De aquí que la precisión sea indicativa de la exactitud del proceso de búsqueda. Lo mismo podemos decir si tenemos en cuenta la mayor o menor destreza de la persona que realiza la búsqueda. No hay que olvidar, que la recuperación de información es un proceso que puede requerir el dar diferentes pasos hasta obtener unos resultados acordes con lo que se busca. Por eso pensamos que estas medidas pueden contemplarse con otros valores de los motores como son: la ordenación de los resultados tras las búsquedas, el buen funcionamiento del motor, analizar si buscan los términos solicitados, etcétera.

En opinión de Johnson y otros (2001:15), diferentes estudios de usuarios demuestran que la mayoría de ellos se conforman con encontrar simplemente dos recursos relevantes y que difícilmente se va más allá de las tres primeras páginas de resultados. Este aspecto no es del todo válido para usuarios interesados en temas especializados que a menudo necesitan un variado número de recursos de información sobre su tema de interés. Es por ello que la satisfacción del usuario puede ser tratada como una fase complementaria de la evaluación del sistema.

Podemos señalar que lo que más interesa medir en nuestro caso es un valor que si está bien definido, como es la frecuencia de aparición del término o términos en los documentos, puede darnos unos resultados homogéneos para todos los buscadores, ya que lo que se mide es la correcta recuperación de los documentos en los que los términos de búsqueda aparecen en lugares importantes del documento, de forma más o menos destacada y con una relativa frecuencia.

Fairthorne (1963), Bar-Ilan (1998/99) entre otros, señalan que la relevancia se puede medir en función de que aparezcan los términos en los documentos recuperados, valorando otros aspectos del contenido, como la aparición de enlaces que apuntan a páginas relevantes, etcétera. Estos autores coinciden con Lawrence y Giles (1998c) para quienes sólo se puede valorar tras la descarga de una página y la localización en ella de los términos de búsqueda.

Para Large, Tedd y Hartley (2001:286) la mayoría de expertos en recuperación de la información están de acuerdo en considerar que el mejor criterio para valorar la relevancia es que el documento trate sobre la materia que se busca. Añaden además, que la relevancia ha de evaluarse de forma binaria y no mediante una escala en la que se contemplen diversos grados de relevancia.

Courtois y Berry (1999) citan a Mattew Koll quién ha observado que el usuario accede a los resultados que presentan todos los términos de la consulta, y además señala que bajo ninguna circunstancia un motor debe ordenar de forma preferente sobre otros recursos con menos términos de los expresados en la consulta.

Nosotros utilizaremos la relevancia en los términos expresados por Fairthorne (1963), Bar-Ilan (1998/99), Lawrence y Giles (1998c) y otros, que valoran la aparición de los términos en los recursos obtenidos. Bar-Ilan (1998) calcula la precisión hallando el tanto por ciento que se obtiene al dividir el número de documentos que contienen el término de búsqueda por el número de documentos accesibles.

Como ya hemos señalado, otro de los aspectos de interés en la recuperación es la ordenación de los resultados de búsqueda, en la que intervienen determinados algoritmos para calcular el orden en la presentación de los resultados. Este aspecto tiene una gran importancia para el usuario ya que interesa que los resultados más relevantes aparezcan en los primeros lugares, lo que le puede suponer un gran ahorro de tiempo.

Tradicionalmente en los SRI se han venido utilizando algoritmos basados en la valoración de palabras clave en función de la especificidad o generalidad de los términos, en el cálculo de sus frecuencias, de su importancia, teniendo en cuenta además la extensión y la posición en el documento para realizar dicho cálculo. Con la llegada de de los sistemas de recuperación de la WWW, se han incorporado a este cálculo otros elementos como son el análisis de enlaces, basado en diferentes aspectos como su importancia y popularidad y las frecuencias de acceso a las páginas o sitios Web.

No obstante, en el resultado final intervienen otros cálculos más complicados y de carácter secreto. Landoni y Bell (2000) indican, siguiendo a Ding y Marchionini, que el ranking por relevancia se calcula dividiendo el número de documentos relevantes de la mitad superior de la página de resultados por el número de documentos relevantes.

En el ámbito de la Web hemos de destacar el trabajo de Yuwono y Lee (1996) así como el de Li, L., Shang, Y. y Zhang, W. (2000) sobre los modelos utilizados en los motores web para calcular el algoritmo de ranking. Bar-Ilan (2005) ha comparado la ordenación de los resultados de Google, AllTheWeb, AltaVista y HotBot mediante 15 preguntas relacionadas con el campo de la recuperación de información. Las conclusiones apuntan a que los motores de búsqueda utilizan diferentes algoritmos de ranking y que para valorar cuál lo hace mejor es necesario realizar estudios más amplios basados en la opinión del usuario.

Courtois y Berry (1999) también se han ocupado no sólo de estudiar este criterio, sino de aplicarlo a la evaluación. Su trabajo es de un gran interés porque permite conocer cómo los motores ordenan los resultados de la búsqueda, pudiendo valorar el peso de los criterios comerciales, o bien, si realizan una ordenación lógica.

Estos autores analizan cómo los usuarios juzgan sus resultados, recogiendo la afirmación de Koll, al referirse a que los usuarios tienen una idea intuitiva sobre si los motores ordenan correctamente los registros. También se ocupan de medir el ranking de relevancia, mediante el cuál, los motores ordenan sus resultados, estableciendo como criterios, por este orden, la recuperación del mayor número de términos expresado en la ecuación de búsqueda, la proximidad y la localización de las palabras en el texto.

El usuario es una vez más quién tiene la última palabra a la hora de juzgar cuál ordena mejor según sus preferencias, para lo que además, como herramientas interactivas presentan opciones como por ejemplo el uso de los modificadores (+) y (-), que además de indicar la presencia o ausencia de un término en la búsqueda, en determinados buscadores sirve para indicar los términos a los que dar mayor y menor peso en la ordenación, pudiendo un usuario experimentado o la intervención de un especialista ser decisivos para obtener una mayor relevancia y, por tanto, una ordenación condicionada por el peso otorgado a determinados términos. Leighton y Srivastava (1998c), destacan la importancia del análisis que realizan metabuscadores como Inquirus al facilitar el contexto que rodea a los términos de búsqueda, lo que resulta muy útil para el usuario.

La intervención de nuevas variables en el cálculo del algoritmo de ranking como en el caso de Google, que valora la popularidad de una página o sitio web, puede limitar la validez de los estudios basados en la metodología anterior.

Los buscadores web ordenan los registros en relación con su relevancia que, como hemos visto, unos y otros calculan de forma diferente, y esto si que nos parece importante en la evaluación.

Dado que en nuestro trabajo de evaluación no tenemos en cuenta la valoración del usuario sino que nos centramos en el funcionamiento, trataremos de analizar en qué se basa la ordenación, atendiendo a si utilizan o no la metainformación y el peso que juega la frecuencia de aparición de palabras y el peso. Analizaremos además si hay una relación entre estos valores y la ordenación.

Para calcular estos valores hemos utilizado el modulo HTML Analyzer, programa disponible en la red, que de una forma similar a los buscadores de la Web, calcula el

peso de los términos mediante un algoritmo que tiene en cuenta el número total de palabras clave, la forma y lugar en que aparecen así como la distancia del comienzo del texto. Para nosotros lo importante es poder comparar los resultados ofrecidos por los motores que se van a evaluar, teniendo en cuenta los aspectos comúnmente utilizados para ello, y a partir del análisis de resultados, dejar constancia de cuáles son los que mejor y peor funcionan.

Abad (2005:108 y stes.) trata con mayor profundidad este aspecto, dado que es uno de los criterios comúnmente utilizados en la comparación de bases de datos. Los valores que han tratado de expresarse en relación con este aspecto han sido tanto el porcentaje de cobertura relativa, el porcentaje de solapamiento y el de aporte específico, es decir, las páginas únicas, siendo el segundo el más utilizado. Expresa la proporción de referencias comunes que aparecen en los sistemas que se comparan, entre el total de recuperados por los sistemas. En los estudios sobre los buscadores se realiza sobre una muestra que puede ser de los diez, veinte o, como en nuestro caso, de los cincuenta primeros resultados.

Para esta autora, los estudios de solapamiento tienen como objetivos: conocer la cobertura relativa de un sistema de recuperación de información respecto a otro; conocer los contenidos redundantes y los exclusivos. Pero la cobertura relativa de una materia es difícil de calcular en los buscadores web porque requiere conocer la totalidad de recursos que contienen los sistemas sobre una determinada materia. En base a ello se han de comparar las referencias de uno y otro buscador. La cobertura relativa de estos sistemas se halla dividiendo el número de referencias recuperadas por uno de ellos por la suma de los recuperados por la totalidad de los sistemas que se comparan. Así, teniendo dos sistemas de recuperación Ay B, la cobertura relativa de

$$\text{A es igual a: } \frac{\text{número de referencias recuperadas en A}}{\text{núm. de ref. rec. por A + B - ref.comunes}} \times 100$$

El solapamiento global expresa lo que representan las referencias comunes entre ambos sistemas y las existentes en la fuente de comparación. Así el solapamiento global

$$\text{es igual a: } \frac{\text{número de referencias comunes en A + B}}{\text{núm. de ref. rec. por A + B - ref. comunes}} \times 100$$

El solapamiento relativo es un indicador que permite valorar para una determinada herramienta cómo le afecta el solapamiento y se calcula dividiendo el número de referencias comunes de A y B respecto del total de referencias contenidas en A. El Solapamiento relativo es igual a : $\frac{\text{número de referencias comunes de A + B}}{\text{núm. de ref. rec. por A}} \times 100$

Contrariamente al solapamiento el aporte específico trata de valorar los recursos únicos de un sistema de información y se calcula dividiendo el número de referencias no solapadas de un sistema por el número total de referencias recuperadas por los sistemas a comparar. El Aporte específico es igual a:

$$\frac{\text{número de referencias no solapadas recuperadas en A}}{\text{núm. de ref. rec. por A + B - ref. comunes}} \times 100$$

Una vez analizadas las diferentes propuestas elaboradas por los especialistas en recuperación de información, la que más se aproxima a nuestros objetivos es la planteada por Chu y Rosental (1996) que siguiendo los planteamientos de Lancaster y Fallen (1973) para evaluar sistemas de recuperación de información, proponen una serie de criterios que se adaptan a las herramientas Web. Hemos eliminado de dicha propuesta criterios como la evaluación del tiempo de respuesta, por considerarlo superado además de otros de carácter cualitativos de los que Chu y Rosental se ocupan.

Por tanto, además de analizar las capacidades de búsqueda en función de la respuesta a los diferentes tipos de búsqueda, valoraremos la composición de los índices, teniendo en cuenta la profundidad de indización, la existencia de duplicados y en laces inactivos. Finalmente analizaremos la recuperación, no sólo desde el punto de vista más formal, si no también valorando la precisión y la ordenación de registros, así como el solapamiento entre motores.

II. MATERIAL Y MÉTODO

Nos hemos referido con anterioridad a los factores internos y externos que afectan a la recuperación de la información de los sistemas de búsqueda de la Web. Debemos precisar en qué medida se dan, es decir tratar de cuantificarlos y analizar si afectan a todos los buscadores por igual o si hay diferencias entre ellos. Ello nos permitirá informar tanto a los usuarios como a los especialistas en búsqueda de información sobre cómo afectan realmente estos problemas a la recuperación de la información que ofrecen los motores de búsqueda y metabuscadores más utilizados en la Web.

Abad (2005) presenta un breve esquema sobre los tipos de evaluación existentes, que varían en función del criterio adoptado para su clasificación. Nos parece interesante referirnos brevemente al que las clasifica en función del punto de vista de la técnica empleada en la recogida de datos, distinguiendo entre evaluaciones cualitativas, cuando se basan el la observación, entrevistas, estudios de casos, y cuantitativas, cuando utilizan para la toma de datos, cuestionarios, recogida de datos sistemática, etcétera.

Harry y Oppenheim (1993) exponen en un trabajo dedicado a la evaluación de bases de datos electrónicas y más concretamente a los CD-ROM, que la metodología ha de ser consistente tanto en el procedimiento como en la presentación de resultados, no ha de requerir mucho tiempo, ha se ser simple, objetiva y flexible; aspectos que podemos mantener para la evaluación de buscadores web.

Para una mejor evaluación de estos sistemas hay que pensar en que son algo más que cajas negras en expresión de Harter y Hert (1997), sino como los denominan Sparck Jones y Willett (1999:168), cajas de cristal que permiten examinar su trabajo interno.

La Web, desde sus inicios ha despertado gran interés en la comunidad científica y especialmente todo lo que tiene que ver con la información y su acceso. Prueba de ello es el importante número de estudios con que contamos en la actualidad.

En relación con la búsqueda y recuperación de la información, Chowdury (1999) los ha agrupado en los siguientes puntos de interés:

- La efectividad de los motores
- La calidad de la información
- Estudios de usuarios
- Diseño de interfaces
- Metadatos
- Clasificación
- Indización
- Agentes de búsqueda
- CD ROM y búsqueda on line

El primero de estos puntos es el más relacionado con la evaluación del funcionamiento que es el punto de vista en el que nos centraremos para llevar a cabo nuestro trabajo de investigación.

Seguimos la metodología propuesta por Bell (1998), quién señala los siguientes puntos:

1. Valorar el escenario: decidir qué se va a evaluar, por quién y para qué.
2. Analizar los criterios y medidas relacionadas: ir a la lista de criterios seleccionados y medidas y ver cuáles pueden ayudarnos a cumplir los objetivos de nuestra investigación.
3. Analizar los motores de búsqueda: recoger la información sobre las herramientas que vamos a evaluar.
4. Definir el experimento: diseñar el experimento teniendo en cuenta los objetivos y el entorno de los tres pasos previos.
5. Análisis de los resultados: interpretar los resultados objetivamente y con arreglo a las expectativas señaladas en el punto primero.

Estos principios se complementan con los expuestos por Spark Jones, K. y Willet, P. (1997) para la evaluación, que aluden a la consistencia, eficaz uso del tiempo, simplicidad, objetividad y flexibilidad.

Como quiera que con anterioridad ya nos hemos referido al primer punto de los expuestos por Bell, a continuación exponemos los indicadores en los que vamos a basar la evaluación.

1. Los indicadores de evaluación, valores y medidas a aplicar

Para la selección de indicadores hemos tenido en cuenta no sólo los aspectos que interesa valorar y se desprenden del estudio de los problemas que afectan a la información de la Web sino también, el funcionamiento de sus componentes y los utilizados en otros trabajos de evaluación, todo ello orientado a conseguir cumplir los objetivos propuestos.

En la evaluación utilizaremos indicadores cuantitativos que nos permitirán obtener resultados objetivos.

Para ello hemos tratado de establecer unos criterios intrínsecos, que nos permitan, además de valorar el funcionamiento ante cada tipo de búsqueda, analizar la formación, características y mantenimiento de sus bases de datos, los índices y la búsqueda y recuperación de resultados. Los datos extraídos han de permitirnos analizar el contenido y la cobertura de las bases de datos, la actualización de sus índices, la recuperación mediante operadores booleanos y modificadores, búsquedas por campo y por frase, y el funcionamiento del software de recuperación de los buscadores evaluados. Pero sobre todo, ha de permitirnos conocer su utilidad en recuperación especializada. Una vez analizados los distintos aspectos que entran en juego en la recuperación de información en la Web, pensamos que es interesante conocer la cobertura temática de estas herramientas.

Para tener una idea aproximada de la cobertura[228], consideramos interesante valorar el número total de recursos que localiza cada motor en cada búsqueda en particular y ver si su capacidad de recuperación se mantiene en todas las búsquedas. Sabremos cuál permite acceder a más recursos y cuál a menos. Con estos datos podremos hacernos una idea del tamaño de las bases de datos y de cuál herramienta permite acceder a un mayor número de recursos.

En el análisis de los registros partimos de un análisis sobre la extracción del título, que normalmente toman de la metaetiqueta TITLE valorando si coincide o no con la del propio documento, ya que a menor nivel de coincidencia, menos probabilidades

[228] Tradicionalmente la cobertura ha tratado de valorar, como señala Lancaster "con muchas dificultades" hasta qué punto una base de datos ofrece resultados exhaustivos sobre la totalidad de lo publicado sobre un tema. Lancaster (1993:170).

de que el título que presenta el recurso en el listado de recuperación, exprese el contenido del documento, lo que da lugar a una pérdida de información. Por tanto, a mayor coincidencia entre los títulos, mayor exactitud de la información. Nos interesa valorar hasta qué punto la aparición de los términos destacados en la información del registro es útil para valorar el interés que tiene el recurso. También hemos considerado interesante conocer qué buscadores utilizan técnicas visuales que permitan agrupar recursos pertenecientes a un mismo sitio Web, colocando los recursos de forma dependiente de otro anterior. Finalmente hemos analizado la frecuencia de aparición de recursos comerciales en los listados de resultados, ya que nos parece indicado colocar en lugar aparte este tipo de información comercial y no mezclada entre los resultados de forma no destacada.

Respecto al índice, nos interesa valorar cuál es el buscador que más desciende en la jerarquía de los sitios Web para indizar recursos. Se estudiará el número de duplicados que recupera cada motor en cada una de las búsquedas. Consideramos duplicados los recursos cuyo URL se repite. No se considera duplicado el recurso alojado en un espejo o mirror porque no supone un funcionamiento defectuoso del motor ya que indica que un mismo recurso se encuentra alojado en diferentes servidores.

A estos efectos, las páginas con idéntico contenido se consideran diferentes si poseen distinto URL.

La actualización la analizaremos atendiendo al número de enlaces inactivos, es decir, aquellos que al activarlos dan lugar a avisos del tipo 404 (File not found) ya sea por cambio de nombre, de ubicación o por haber sido eliminado del web, avisos de acceso prohibido o mediante identificación, y el error 603 (server is not responding).

En relación con las características de la información analizaremos la actualidad de los recursos mediante un análisis de fechas basado en las que aparecen en los recursos recuperados. Esto nos permitirá establecer cuál es el motor que recupera los recursos con fechas más recientes. Para ello tomaremos como fecha válida, no la del copyright, ni la que se inserta de forma automática cuando se visita una página, ni la que aparece como fecha de edición, como son trabajos presentados a Congresos o los de edición de una publicación periódica o monografía, sino la propia del recurso, siempre y cuando aparezca en la página.

Sobre el carácter predominante de la información en la base de datos, valoraremos si recupera recursos de carácter científico, en los términos en que Van Slype

(1988:1-3) considera este tipo de información (información cognitiva[229]) analizando si la información tiene este interés o si más bien responde a intereses comerciales, publicitarios, de carácter institucional o simplemente divulgativos. Aunque es de suponer que, con los temas de búsqueda propuestos, estos sistemas recuperen un alto porcentaje de información de contenido académico. Para su valoración se tienen en cuenta aspectos como la autoría, filiación o la entidad que aloja la página. Un nuevo punto de análisis que permita distinguir formalmente el tipo de documento que contiene la información, esto es, si se trata de trabajos presentados a Congresos, pre-prints, publicación en revistas electrónicas, recursos de enseñanza o de tipo bibliográfico, etcétera, también nos ayudará a valorar, de forma más correcta el interés de la información recuperada.

Vamos a medir la recuperación de páginas únicas y el solapamiento entre motores, es decir, el grado de coincidencia entre ellos en la recuperación de idénticos recursos ante una misma búsqueda, lo que nos permitirá, ante la necesidad de una búsqueda exhaustiva, decidir cuáles utilizar, o bien, cuando los resultados de una búsqueda en un motor son insuficientes, evitar lanzarla en motores con gran solapamiento.

Calcularemos la precisión técnica, es decir, la pertinencia en la recuperación analizando si los resultados responden al tema solicitado. Nos basaremos en los criterios utilizados por otros autores como Fairthorne (1963) y Bar-Ilan (1998), que consideran recursos precisos aquellos que contienen los términos que se buscan (excluyendo los duplicados). Dado que Bar-Ilan utiliza para valorar este aspecto una búsqueda de un solo término, nosotros nos basaremos en los datos extraídos de la primera búsqueda, que es la mayor similitud. Bar-Ilan (1998) calcula la precisión hallando el tanto por ciento que se obtiene al dividir el número de documentos que contienen el término de búsqueda por el número de documentos accesibles. En el resto de búsquedas valoraremos la recuperación de recursos con todos los términos de búsqueda o de los más importantes.

Finalmente analizaremos cómo ordenan los recursos, valorando la utilización de la información de las metaetiquetas Key y Description, la frecuencia y el peso de los términos de búsqueda que corresponde a los resultados organizados en grupos consecu-

[229] Para Abadal y Codina (2005:33) "es útil para aumentar nuestros conocimientos sobre algún aspecto de la naturaleza…"

tivos de diez en diez hasta los cincuenta que se valoran y, finalmente comprobaremos la existencia o no de relación entre la Frecuencia y el Peso y su ordenación.

Las correlaciones se utilizan para valorar la relación lineal entre variables. Mediante su valoración, se puede establecer la existencia de relaciones de dependencia entre las variables que se comparan. Así pues, para valorar la existencia o no de una relación lineal entre las variables Frecuencia y Peso y la ordenación de los recursos, así como el grado en el que se da, utilizaremos el coeficiente de Correlación de Pearson.

El análisis de todas las variables ha de permitirnos informar al usuario que utiliza estas herramientas para buscar información especializada, sobre cuál le puede ofrecer una información lo más completa y actualizada posible, accesible, precisa, fiable y que en la recuperación se facilite el acceso a los recursos más idóneos, colocándolos en los primeros lugares.

En este sentido, tratamos de valorar la respuesta de estas herramientas al utilizar diferentes tipos de búsqueda.

Un gran número de estudios utilizan 15 o 20 recursos para valorar los elementos que quieren analizar pero nosotros hemos considerado que 50 puede ser una muestra más completa, ya que en búsquedas especializadas, el usuario suele estar interesado en un mayor número de recursos, lo que nos permitirá profundizar más en la recuperación de los buscadores Web.

Como nuestro interés está en analizar el rendimiento de estas herramientas, no excluimos ningún registro, tengan o no los términos de búsqueda, estén más o menos relacionados con el tema de búsqueda, dando detallada cuenta de cada uno de estos casos.

Los recursos a los que no se ha pudo conectar inicialmente, son objeto de un nuevo intento, de manera inmediata. Si pasado un minuto no se cargaban las páginas, se consideraban enlaces erróneos.

2. Selección de Motores de búsqueda y Metabuscadores

El tercer punto de Bell indica la necesidad de analizar los motores de búsqueda a evaluar. Las características más importantes que hemos ido recogiendo y observando a lo largo del presente trabajo aparecen representadas en tabla de la página siguiente.

Por su carácter amplio centramos la investigación en los buscadores automáticos de carácter general y de acceso libre, que son los que se suelen utilizar para todo tipo de búsquedas, así como los metabuscadores más conocidos.

Para la selección de las herramientas de búsqueda nos hemos centrado en las más conocidas y en las que contienen las bases de datos más amplias. Para ello hemos consultado obras especializadas como el Directorio de recursos de interés académico y profesional, coordinado por Ángeles Maldonado Martínez y sitios web especializados en el análisis y evaluación de herramientas de búsqueda (Searchenginewatch[230], Searchengine Showdown[231]).

Motores de búsqueda	URL	Metabuscadores	URL
Google	http://www.google.com>	**Excite**	http://www.excite.com>
YahooSearch	http://www.searchyahoo.com>	**Search.com**	http://www.search.com>
MSN Search	http://search.msn.com>	**IXQuick**	http://www.ixquick.com>
WiseNut	http://www.wisenut.com>	**Profusion**	http://www.profusion.com>
Teoma (Ask)	http://www.teoma.com>	**Vivisimo**	http://www.vivisimo.com>
		SurfWax	http://www.surfwax.com>
		Dogpile	http://www.dogpile.com>

[230] http://searchenginewatch.com/
[231] http://www.searchengineshowdown.com/

Características, modos y opciones de búsqueda de los buscadores evaluados[232].

	Tipos de recursos						Tipología documental							Modos de Búsqueda			Opciones de búsqueda													
	Noticias	Personas en pág.blancas	Empresas en pág.amarillas	Imágenes	Doc. audiovisual		PostScript	RTF	PDF	Word	Excel	Power Point		Directorio	Simple y avanzada		Búsqueda por Frase	Operadores booleanos			Operador de cercanía NEAR	Modificadores (+) y (-)	Truncamientos	Sensibles a mayúsculas y minúsculas	Búsqueda por campo	Filtros	Sinónimos	Búsqueda en metaetiquetas
																		OR	AND	NOT ó AND NOT									
Google	x			x	x		x	x	x	x	x	x		x	x		x	x		x[233]		x			x	x[1]		x
MSN	x	x	x	x	x		x	x	x	x	x	x		x	x		x	x	x	x	x	x			x			x
Teoma(Ask)	x			x			x		x						x		x	x	x			x			x			
WiseNut							x		x						x														
Yahoo	x			x	x		x	x	x	x	x	x		x	x		x	x	x	x		x			x			
Dogpile	x			x	x		x		x	x								x	x	x								
Excite	x	x	x	x	x		x		x	x																			
Ixquick	x			x	x		x		x	x	x	x					x	x	x	x		x	x	x	x			
Profusion							x		x	x	x	x						x	x	x								
Search	x	x		x	x		x		x	x				x	x		x					x	x					
Surfwax	x																x					x						
Vivisimo							x		x								x	x	x	x		x			x			x

[232]Características extraídas principalmente de la información que ofrecen al usuario en las páginas de ayuda. Para la tipología documental nos hemos basado en los datos obtenidos en la evaluación.

[233] En la búsqueda avanzada

Dado que existen en algunos casos distintas versiones de las bases de datos de los buscadores, optamos por lanzar las búsquedas sobre las bases de carácter internacional, expresando el URL genérico.

Desde el lanzamiento de las búsquedas en enero de 2006, los cambios más importantes relativos a estos servicios de búsqueda son que Teoma ha sido adquirido por Ask Jeeves pasando a llamarse Ask.

3. Test de evaluación

Siguiendo la metodología de Bell, nos queda diseñar el experimento para lo que debemos tener en cuenta tanto los objetivos como, obviamente, los aspectos que acabamos de tratar.

3.1. Modos de búsqueda

Para llevar a cabo la evaluación hemos utilizado el modo de búsqueda simple, ya que es la que se suele utilizar de forma generalizada, excepto en determinado tipo de búsqueda, como las búsquedas por campo o la avanzada, en caso de expertos. Sólo ha sido necesario acceder a la búsqueda avanzada para conseguir búsquedas por frase en los buscadores que así lo requerían, con el objetivo de obtener unos resultados que pudieran contrastarse con los recuperados por otros buscadores que no soportan el uso de los signos + y – para forzar su existencia o inexistencia en las páginas recuperadas, como en el caso de WiseNut, o de las comillas en la búsqueda por frase, como en WiseNut y Profusion.

3.2. Temas de búsqueda y Sintaxis

En primer lugar hemos de referirnos a que el idioma seleccionado para las búsquedas es el inglés debido a que al ser el más utilizado en la Web[234], permite observar mejor el rendimiento de estos servicios en situaciones como ésta, en que se demanda un mayor trabajo al sistema.

[234] Véase el estudio de Aguillo, Ortega y Granadino (2006) en el que se demuestra que los contenidos en este idioma superan las dos terceras partes del total.

Dado que pretendemos evaluar la utilidad de los motores de búsqueda ante consultas especializadas, las búsquedas planteadas contienen términos propios de un determinado campo científico, en este caso el de la Documentación. Los términos y frases de las búsquedas han sido extraídos de diferentes trabajos relacionados con esta disciplina y son de uso habitual en ella. Mediante ellos se elaboró un listado del que se extrajeron las búsquedas definitivas.

Los temas de búsqueda han sido seleccionados pensando en los aspectos que se desea evaluar en las bases de datos, es decir, su comportamiento en búsquedas por palabra única, varios términos en lenguaje natural, utilización de operadores de existencia, búsqueda booleana, búsqueda por frase y por campo. Esta variedad obedece al interés por analizar la capacidad de búsqueda de estas herramientas.

En algunos casos, a las expresiones de búsqueda se las ha dotado de cierta dificultad con la intención de hacer que los buscadores y metabuscadores tengan que trabajar de forma algo forzada, a fin de obtener resultados más representativos. Es el caso de las búsquedas con operadores de existencia, búsqueda booleana, por frase y por campo.

En la siguiente tabla presentamos en la primera columna el número de la pregunta, en la segunda los términos de búsqueda y debajo su traducción al español. La tercera columna contiene el tipo de búsqueda y la cuarta la ecuación de búsqueda, que expresa de qué modo se lanza sobre los respectivos sistemas de recuperación.

Núm	Temas de búsqueda	Tipos de búsqueda	Expresión de búsqueda
1	Softbot Robot de búsqueda	Palabra única	**softbot**
2	Best match information retrieval in web search engines Equiparación exacta en recuperación de la información en motores de búsqueda de la Web	Varios términos	**Best match information retrieval in web search engines**
3	Information retrieval systems and the web (De forma que aparezcan obligatoriamente los términos “and” y “the”. Los sistemas de recuperación de la información y la web.	Operadores de existencia	**Information retrieval systems +and +the web**

4	Information retrieval and digital libraries and electronic libraries and virtual libraries La recuperación de la información y las bibliotecas digitales, bibliotecas electrónicas y bibliotecas virtuales.	Búsqueda Booleana	**Information retrieval AND digital libraries AND electronic libraries AND virtual libraries**
5	Natural language processing Procesamiento del lenguaje natural	Frase	**"Natural language processing"**
6	Information retrieval en el campo de título	Búsqueda por campo	**intitle:information retrieval**

En las búsquedas, los términos se indican en minúsculas, pues no se trata de valorar si son sensibles o no al uso de mayúsculas, sino que de este modo se recuperan los términos escritos en ambas grafías.

Para la búsqueda por frase, se encierran los términos entre comillas lo que obliga al motor a buscar los recursos que contengan los términos indicados en el orden establecido.

3.3. Ejecución de las búsquedas

El test se realiza utilizando los ordenadores de la Sala de Formación de Usuarios de la biblioteca María Moliner de Facultad de Filosofía y Letras de la Universidad de Zaragoza. Estos equipos, con un procesador Intel (IR) Pentium (R) 4 CPU 2.80 GHZ, están conectados a Internet a una velocidad de transferencia de 260 Mbps.

Tienen instalado el sistema operativo Windows 2000 y software básico que permite acceder a la mayoría de documentos de Internet (Office 2000 y Acrobat Reader). El navegador utilizado fue Netscape en su versión 7.0.

Intentamos reproducir una situación lo más próxima posible a la realidad. Para ello se convoca a las personas participantes a una hora determinada y tras explicar como se va a realizar la evaluación, se lanzan las búsquedas y se guardan las páginas que contienen los cincuenta primeros resultados para, a continuación, visitar y guardar las páginas de los recursos recuperados.

Para evitar errores, las expresiones de búsqueda se copian en un documento de Word desde donde se trasladan mediante el sistema de copiar y pegar al formulario de cada herramienta de búsqueda.

WiseNut planteó problemas con la segunda búsqueda al no tener capacidad para procesar más de siete términos de búsqueda, tal y como expresó en un mensaje de aviso, por lo que optamos por eliminar el último de los términos de la ecuación de búsqueda. En la tercera búsqueda, en la que se utiliza el limitador (+) hubo que señalar la opción "with all" ya que este buscador indica que no soporta dichos signos y en la tercera seleccionar la opción de búsqueda por frase de la búsqueda avanzada, al igual que Profusion. En la búsqueda por campo, en este buscador fue necesario separar los términos del siguiente modo: in title. A pesar de que tratamos de seleccionar tipos de búsqueda, que aunque son especializadas, deberían ser soportadas por todas las herramientas, Teoma, WiseNut, Dogpile, Profusion y Surfwax no ofrecieron resultados en alguna de ellas.

Dado el número de herramientas a evaluar y teniendo en cuenta que se lanzan seis búsquedas sobre cada una, y que en la misma sesión se ha de acceder a todos los recursos y guardarlos para insertarlos en la base de datos de recogida de información, se estima oportuno realizar dos sesiones con una separación de seis días entre ellas. La primera sesión, en la que se lanzan las tres primeras búsquedas se lleva a cabo el día 20 de enero de 2006, lanzándose las tres restantes el día 26.

Estas sesiones se realizan con la ayuda de 7 personas que colaboran exclusivamente en el lanzamiento de las búsquedas y en el acceso y almacenamiento de los datos. Cada sesión tiene una duración aproximada de tres horas y media.

Cada persona lanzó las búsquedas sobre dos buscadores, guardó las páginas de resultados, accediendo a continuación a cada uno de los recursos recuperados para guardar la información que contienen, el URL e indicando en su caso las dificultades de acceso al recurso. La inestimable colaboración de estas personas permitió que las búsquedas pudieran lanzarse de forma simultánea en los buscadores y acceder de forma inmediata a los recursos, lo que favorece la homogeneidad de los resultados analizados y un trato igualitario para la evaluación de los buscadores.

3.4. Recopilación y análisis de datos

En la sesión de búsqueda y almacenamiento de los datos, en primer lugar se guardan las páginas en formato HTML que contienen los cincuenta primeros registros recuperados. Esto nos permitirá analizar tanto las páginas de resultados como el contenido de los registros. Se crean diferentes carpetas y subcarpetas que actúan como directorios para contener la documentación y bases de datos de los resultados obtenidos por cada buscador en cada una de las búsquedas. Estos directorios contienen además las bases de datos en las que se recoge la información extraída y que va a ser utilizada en la evaluación.

En las tres primeras búsquedas Search no permitió guardar las páginas de búsqueda mediante el comando "Guardar como" utilizado para el resto de buscadores, lo que ha influido en la falta de datos para valorar el número total de recursos recuperados en estas búsquedas. En la siguiente sesión, se utilizó la técnica de copiar y pegar en un documento Word los resultados y la información que les acompaña.

Las bases de datos se diseñan con el programa FileMaker Pro 5, que es el instalado por defecto en los ordenadores utilizados para las sesiones de búsqueda. La base de datos consta de varias presentaciones, para contener datos de cada sesión de análisis. La primera, más básica contiene los campos URL, donde se pega esta información tras copiarla de cada una de las páginas a las que se accede en esta primera sesión, el número de orden en el listado y un campo de observaciones, donde se indican las incidencias sobre el acceso a cada recurso, sobre todo en los casos de problemas de conexión u otro tipo de error, indicándose en este caso el tipo de error. Por tanto, los datos de esta presentación se rellenan en la primera sesión realizada el día que se lanzan las búsquedas.

Además se guardan en los directorios correspondientes los recursos a los que se ha accedido. Para ello se utiliza el comando "Guardar" de cada uno de los programas que hemos necesitado para abrir el recurso (El navegador Explorer, el lector de archivos PDF Acrobar Reader, el procesador de textos Word, PowerPoint, etcétera). Al guardar cada página o recurso, se añade al final del nombre que ofrece por defecto, el número de orden en el listado, lo que nos facilita la relación entre el recurso y la información de la base de datos.

El paso siguiente consistió en extraer los datos a tener en cuenta en la evaluación. En este caso se utiliza el programa File Maker Pro 7 que permite volcar la información de las bases de datos anteriores y además, mediante la utilización de los llama-

dos campos "contenedores", permite almacenar junto al resto de datos, los archivos recuperados que como hemos indicado habíamos guardado en directorios, en cualquiera de sus formatos (HTML, PDF, RTF, Word, etcétera). Esta posibilidad de acceso a cada uno de los recursos recuperados, nos ha permitido, no sólo en un primer momento acceder a los recursos para analizarlos sino resolver, de una forma fácil, cualquier problema que se plateara a lo largo de las sesiones de análisis de los recursos.

En una segunda fase de análisis de los recursos guardados, se valora cada uno de los puntos de interés, recogiéndose los resultados en una nueva presentación de la base de datos con los siguientes apartados, campos y subcampos:

Información general

Contiene información identificativa del motor y de la búsqueda así como de las incidencias observadas al lanzar la búsqueda. Recoge también aspectos de la página de resultados como son la inclusión de recursos de carácter comercial. También contiene otra información sobre las características de los registros como son el título del recurso, la aparición de los términos de búsqueda destacados y la dependencia o no del recurso anterior.

- Nombre del motor
- Nº de búsqueda
- Expresión de búsqueda
- Observaciones de la búsqueda
- Nº total de recursos recuperados en la búsqueda por el motor
- Nº en el listado [Indica el número de orden que ocupa cada recurso en el listado de resultados]
- Muestra recursos publicitarios (s/n) y número total de recursos publicitarios
- Título del listado [Campo que contiene el título que aparece en el listado]
- Descripción [Campo contenedor del recurso recuperado]
- Palabras destacadas y nº de palabras destacadas

- Dependiente de la anterior (s/n) [Se refiere a los recursos que al pertenecer al mismo servidor que el inmediatamente anterior, aparece en el listado con una sangría mayor que el anterior]

Información de la página o recurso

Este apartado recoge datos específicos del recurso y características de su contenido mediante los siguientes campos:

- Recurso [Campo contenedor que permite insertar los documentos relativos a las páginas HTML, PDF, Word, PowePoint, etcétera, recuperadas]
- Título del recurso [Contiene el título propio del recurso o página web y no del título del sitio web.]
- Título de las propiedades de la página HTML [Título que se extrae de la metainformación a la que accedemos mediante el comando Propiedades de página del navegador, que contiene la información de la Metaetiqueta TITLE]
- URL [Contiene la dirección URL del recurso. En algunos metabuscadores ha sido necesario eliminar parte de la dirección URL que es añadida por el propio sistema de búsqueda. Por ejemplo, en http://www.cs.sfu.ca/fas-info/cs/CC/310/havens/notes/midtermExamKey.pdf#search='Soft, eliminamos el texto situado a partir del signo de almohadilla (#).

  Otros signos que incluyen los metabuscadores como (%) para sustituir a otros también se modifican por los originales. Así, la URL<http://encarta.msn.com/ encyclopedia _761582861/ Spider_ 28computer %29.html> se sustituye por: <http://encarta.msn.com/ enciclopedia_761582861/Spider_(28computer)29.html>
- Profundidad del directorio: [Contiene el nivel jerárquico en el que se encuentra el recurso]

- Activo (s/n y tipo de error)

- Formato del recurso:

○ Html/php/script	○ Recurso en Excel
○ Artículo en pdf o ps	○ Recurso en Word
○ Presentación de Power Point en pdf	○ Otro
○ Recurso ppt (Power Point)	○ N/S
○ Artículo en rtf (rich text format)	○ Disponible en varios formatos

- Tipo de recurso:

○ PÁGINA HTML PROPIA DE LA WWW CON INF. DEL SITIO	○ ARTICULO/PÁG DE INF. ESPECIALIZADA
○ PÁGINA HTML EN BLANCO	○ ARTICULO DE REVISTA ELECTRÓNICA
○ PÁGINAHTML EN LENGUAS ORIENTALES O SIMILARES	○ TRABAJO PRESENTADO A CONGRESO O SIM.
○ PÁGINA HTML DE IMÁGENES	○ MONOGRAFIA
○ BASE DATOS BIBLIOGRÁFICA AATC LIBRE	○ CAPITULO DE MONOGRAFIA O RECURSO TEXTUAL MAS AMPLIO
○ BASE DATOS BIBLIOGRÁFICA AATC RESTRINGIDO	○ ARTICULO DE ENCICLOPEDIA
○ BASE DATOS BIBLIOGRAFICA AARB	○ ENTREVISTA
○ BIBLIOTECA DIGITAL	○ DICCIONARIO
○ REPOSITORIO	○ NOTICIAS
○ DIRECTORIO	○ BLOG O PÁGINA PERSONAL
○ BUSCADOR	○ BLOG COMUN ESPECIALIZADO
○ AGENTE DE BUSQUEDA	○ PAGINA REGISTRO
○ ACCESO A BUSCADOR	○ LISTA DE CORREO
○ LISTA DE CORREO	○ DISCURSO
○ REVISTA ELECTRÓNICA	○ EXAMEN
○ E-LIBRO	○ PROYECTO
	○ RESUMEN
○ PRESENTACIÓN	○ CURSO O INF. DE CURSO
○ BIBLIOGRAFIA	○ FAQ
○ LISTA DE RECURSOS WEB	○ NORMAS

La mayoría de estos supuestos no plantea problemas ya que son suficientemente claros por sí mismos. Tal vez en el caso de Página HTML propia de la WWW con información del sitio merezca aclarar que se trata de casos en los que el recurso es la página principal de un sitio Web.

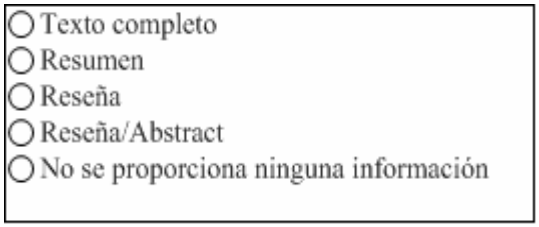
Las abreviaturas de las bases de datos AATC, AARB significan Acceso a Texto Completo y Acceso a Recurso Bibliográfico en función de que la base de datos proporcione el acceso al recurso de un modo u otro.

La diferencia básica entre bibliografía y lista de recursos web es el predominio de referencias de tipo bibliográfico en la primera y en el segundo caso de recursos cuyo acceso se facilita por hiperenlaces.

La localización de libros relacionados con el tema de búsqueda para su venta en librerías a través de Internet, se contempla como bibliografía con interés comercial.

Dentro de proyectos, se incluyen las páginas de grupos de investigación relacionados con el tema de búsqueda.

- Forma de información del recurso:

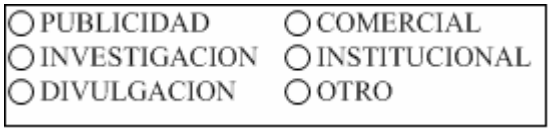


Hay que precisar que la indicación de texto completo tiene sentido sobre todo en los casos en los que hay que especificar si un recurso es simplemente una descripción bibliográfica o bien es el texto completo o un registro que finalmente facilita su acceso, independientemente de los problemas que pueda plantear el acceso al texto completo. Por tanto, si el fin del registro es facilitar el acceso al texto completo, se señala de este modo. Sin embargo si sólo aparece el texto del resumen o de la reseña, se señala en cada uno de estos apartados. Si la recuperación es de la reseña junto al resumen, se señala el apartado Reseña/Abstract.

- Fecha del recurso

Respecto al campo de fecha, no todos los recursos la facilitan y por otro lado la forma en que aparece varía, ya que en unos casos se indica el *copyright*, en otros la fecha que aparece es la relativa al momento en que se consulta el recurso, y por tanto varia a diario, en otros incluyen fechas relativas a la publicación del recurso, como por ejemplo en los artículos de publicaciones periódicas y finalmente otros contienen la fecha en que se ha publicado el recurso en Internet. Para el estudio de la actualidad de los recursos de la base de datos, hemos utilizado todas estas formas a excepción de la primera, ya que no se refiere al recurso sino al acceso. En los casos en los que aparecían varias fechas de estos tipos, optamos por seleccionar la más representativa del recurso, como puede ser la fecha de realización, en detrimento por ejemplo de la del *copyright* que además generalmente abarca un periodo de varios años. Por último, los recursos cuya única fecha es el *copyright* y abarca varios años (ej.: © 2002-2005) la fecha que se utiliza es la primera al ser indicativa del momento en que se realiza o se publica la página.

- Carácter o interés de la información:



Un recurso de investigación se valora atendiendo tanto a la autoría, a la entidad responsable como a su contenido en relación al tema de la búsqueda, o si se trata de un trabajo publicado o presentado a un Congreso. Señalamos en este apartado asimismo los recursos que hacen referencia a información de tipo especializado, por ejemplo la documentación sobre el desarrollo de unas jornadas, que localizándolas, su documentación puede contener documentación de interés. Contemplamos aquí además recursos que si bien por ellos mismos no resuelven la necesidad de información señalada, pueden dirigir a otros recursos que sí lo hacen.

En otras ocasiones nos planteamos si la información que recupera soluciona el problema, y por ejemplo en el caso de acceso a departamentos universitarios donde se informa más bien de aspectos que tienen que ver con el propio departamento, se les aplica el tipo de información señalada como "otro".

Se señala un carácter comercial cuando el objetivo de la página es dar a conocer productos, servicios o programas informáticos para su adquisición o promoción.

En el apartado "otro" se recogen aquellos que no se refieren a ninguno de los apartados anteriores como es el caso de contener información general, noticias cortas, etcétera.

No obstante, hay que advertir que en algunos casos, ha resultado difícil asignar los contenidos a una de las opciones señaladas, pudiéndose clasificar en varias de ellas.

Precisión técnica

Este apartado recoge los datos proporcionados por la aplicación al recurso del Programa HTML analizer que nos proporcionó los valores de cada uno de los recursos.

- Nº de veces que aparecen los términos [Contiene subcampos en los que se contabilizan las frecuencias de aparición de todos los términos o de cada uno de ellos por separado]

- Nº total de palabras en el documento

Ranking

- Frecuencia de aparición del término de búsqueda
- Peso
- Localización de palabras (En título, en texto, en hiperenlace, metaetiquetas Key y Description)

Los mayores problemas para el cálculo del ranking, los hemos tenido con documentos, fundamentalmente en formato PostScript, que no se han podido abrir ni con Acrobat Reader ni con Acrobat Professional. Otros documentos tuvieron que abrirse con el programa de creación de páginas Web, Dreamweaver y guardarlos en un formato legible, generalmente en HTML, desde el que se pudieron analizar.

Tanto para calcular la precisión técnica como el *ranking*, en el análisis de las frecuencias de términos hemos utilizado el programa informático HTML Analyzer, de la casa SEO disponible en Internet[235]. Esta herramienta permite calcular la frecuencia tanto de palabras como frases con un determinado peso, de la misma forma que lo hacen los buscadores de la Web. Para valorar estos aspectos en documentos de World, PDF y RTF se hizo necesaria una conversión al formato HTML. En este proceso, una pequeña cantidad de documentos PDF, debido a sus protecciones, no pudieron ser analizados, para extraer estos valores.

Finalmente, todos los datos de la base de datos son exportados a una hoja de cálculo para su posterior tratamiento con el programa SPSS (Statistical Package for the Social Sciences), de análisis estadístico, mediante el cual se realizan cálculos que permiten hallar medias, medianas, etcétera por grupos de resultados, así como la valoración del coeficiente de correlación de Pearson y otros indicadores de gran interés para la comparación de estas herramientas.

[235] http://www.seodministrator.com/html-analyzer.html

Para la presentación de los resultados ha sido necesaria su tabulación así como su representación gráfica ya que ambos instrumentos permiten el análisis visual y real de los datos, facilitando su comparación.

En el estudio de resultados se nos han planteado algunos problemas como: qué hacer con las páginas en blanco, es decir aquellas que prácticamente no contienen texto, o las escritas en lenguas orientales. En ambos casos se han analizado valorando sólo lo que se conocía, es decir páginas cuya recuperación es correcta porque en la mayoría de casos el término se encontraba en la información del código fuente de la página HTML. Mayor dificultad plantea valorar su contenido, por lo que en estos casos, las características se dejan en blanco. Sin embargo se valoró la frecuencia de aparición en uno u otro buscador al ser tratados como tipos de recursos en el campo correspondiente (Véase campo *Tipo de recurso*).

Las valoraciones de la mayoría de campos no plantean problemas ya que se trata de anotar un dato o característica o seleccionar si cumple o no determinado requisito. También en el tipo de recurso hemos encontrado casos en los que un mismo recurso podía clasificarse en dos o más grupos, optando por el que mejor expresara el contenido y no tanto la forma.

III. RESULTADOS Y ANÁLISIS

1. Datos de la muestra

1.1. Total recursos recuperados por motores de búsqueda y meta-buscadores

Las siguientes tablas recogen el número de recursos que recuperó cada buscador al ejecutar las búsquedas. Estos valores pueden ser indicativos tanto de la capacidad de sus índices como de la cobertura de la base de datos, si bien hemos de tener en cuenta los estudios que demuestran que dichas cifras no siempre son ciertas. En nuestra experiencia también hemos observado que hay buscadores que aunque indican recuperar un determinado número de registros, en realidad, facilitan el acceso a un número sensiblemente inferior. (WiseNut, Ixquick y Surfwax).

Tabla 1.1-1. Motores. Nº de recursos recuperados por búsqueda y motor

	Búsqueda 1 (Término único)	Búsqueda 2 (Varios términos)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Google	26.800	1.520.000	67.700.000	278	2.120.000	1.460.000	**72.827.078**
MSN	4.648	67.555	611.814	61.001	280.661	252.930	**1.278.609**
Teoma (Ask)	3.030	14.800	892.900	Sin resultados	Sin resultados	Sin resultados	**910.730**
WiseNut	459	4.069	189.038	Sin resultados	226.447	128[236]	**420.141**
Yahoo	13.100	245.000	17.500.000	101	1.660.000	1.300.000	**20.718.201**
TOTAL							**96.154.759**

Atendiendo a estos valores, podemos observar que Google es el motor que más recursos recupera en todas las búsquedas a excepción de la búsqueda booleana, en la que sólo recupera 278 recursos. Algo similar le ocurre a Yahoo, que es el segundo buscador en número de recursos. Esto puede ser debido al fuerte filtro que supone la expresión de búsqueda.

[236] Muestra 30

MSN es el tercero en importancia, al recuperar en todas las búsquedas. Destaca frente a los demás en el número de recursos que recupera en la búsqueda booleana, lo cual puede ser indicativo de un funcionamiento anómalo, que no interprete de forma correcta esta búsqueda. Este problema en la recuperación lo hemos podido constatar al valorar la precisión técnica de MSN en esta búsqueda, en la que hemos observado que este buscador es el que recupera un mayor número de páginas que no contienen los términos de búsqueda[237].

Teoma (Ask) supera a WiseNut en todas las búsquedas a excepción de las tres últimas búsquedas, en las que no recupera ningún recurso debido a que su mecanismo de búsqueda no las soporta.

WiseNut no proporciona resultados en la búsqueda booleana, y además, en la búsqueda por campo sólo muestra treinta páginas de las 128 que dice haber recuperado.

Por otro lado, si se quiere valorar la cobertura, resulta más útil tener en cuenta los resultados de la búsqueda simple, en la que no se hace intervenir ningún mecanismo que incida en la selección de unos recursos u otros. En este sentido podemos destacar de nuevo, la recuperación en Google, seguido de Yahoo, aunque en la primera consulta recupera la mitad que Google y en la segunda, una sexta parte. En tercer lugar tenemos a MSN, ofreciendo el resto de buscadores unos resultados inferiores, especialmente en el caso de WiseNut.

Por tanto, aunque como hemos señalado más arriba, estos datos no permiten lanzar afirmaciones concluyentes, pues no todos los recursos contienen todos los términos, ni todos los buscadores facilitan el acceso al número de recursos que dicen recuperar, sí que al menos pueden ser indicativos del tamaño de los índices y de la cobertura de la Web que realizan estos motores en temas especializados.

[237] Véase más adelante el apartado de análisis de la precisión técnica

Tabla 1.1-2 Metabuscadores. Nº de recursos recuperados por búsqueda y metabuscador

	Búsqueda 1 (Término único)	Búsqueda 2 (Varios términos)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Dogpile[238]	Datos no disponibles	43	Datos no disponibles	Sin resultados	Sin resultados	Sin resultados	**43**
Excite	59	73	88	34	76	63	**393**
IXQuick	42 registros únicos de 3.030[239]	41 registros de 14.789[240]	68 registros de 897.497	26 registros de 15.814	52 registros de 1.905.551	25 registros de 972.999.456[241]	**254 de 975.836.137**
Profusion	41	35	Sin resultados	46	39	25	**186**
Search[242]	Datos no disponibles	Datos no disponibles	Datos no disponibles	130	134.040	90.950	**225.120**
SurfWax	Muestra 15 de 459	Muestra 19 de 4.069	Muestra 15 de 459	Sin resultados	Sin resultados	Sin resultados	**49 de un total de 4987**
Vivisimo	Muestra 127 resultados más importantes de un total de 4.660	Muestra 181 de 69.402	Muestra 233 de 898.000	Muestra 187 de 49.982	Muestra 262 de 8.763	Muestra 89 de 40.300	**1.079 de un total de 1.071.107**
						TOTAL	**227124 de un total de 976.907.244**

Los metabuscadores, al no poseer índices propios, proporcionan información de su capacidad de acceso a registros proporcionados por otros buscadores. Aunque no poseemos datos íntegros de todos los metabuscadores, sí que podemos observar la existencia de unos resultados mucho más limitados que los ofrecidos por los motores de búsqueda. Esto parece indicar la existencia de una selección de los recursos recuperados, pues como vemos, Surfwax. Ixquick o Vivisimo sólo ofrecen una parte del total de recursos a los que tienen acceso. Estas herramientas deberían expresar con claridad a qué se debe y, en función de qué aspectos realizan esta selección, ya que en nuestra opinión, dicha selección debería responder a los de mayor relevancia.

Search parece recuperar un gran número de recursos, pero al no poderse guardar las páginas de resultados de las tres primeras búsquedas, no podemos valorarlo globalmente, aunque poseemos datos de las tres últimas búsquedas ya que guardamos las

238 No se recogieron datos del número de recursos que recuperó en las búsquedas 1 y 3.
239 En realidad muestra 32
240 Muestra 39
241 Realmente aparecen 23

páginas de resultados por el procedimiento de copiar y pegar los datos en un documento Word. Le sigue Vivisimo del que poseemos datos de todas las búsquedas. Excite sería el tercer metabuscador en número de recursos recuperados. En las búsquedas simples y complejas Vivisimo es el metabuscador, de los que tenemos datos, que más registros ofrece. Search lo supera en la búsqueda de frase y por campo.

1.2. Total recursos analizados

Tabla 1.2-1 Resultados por búsqueda

Nº de búsqueda	Páginas recuperadas
Búsqueda 1 (Término único)	**538**
Búsqueda 2 (Lenguaje natural)	**536**
Búsqueda 3 (Operadores de existencia)	**536**
Búsqueda 4 (Operadores booleanos)	**351**
Búsqueda 5 (Frase)	**433**
Búsqueda 6 (Campo título)	**378**
Total	**2772**

Si todos los buscadores hubieran recuperado en todas las búsquedas, al seleccionar 50 resultados por cada búsqueda para el análisis, deberíamos contar con un total de 600 resultados en cada una de las búsquedas, pero ya hemos visto que en unos casos no recuperan y en otros no llegan a los cincuenta recursos. Sin embargo, podemos apreciar globalmente, cómo las búsquedas con resultados más numerosos son las tres primeras. El número de páginas desciende en la búsqueda booleana a 351 y en la búsqueda por campo a 378, lo que es debido tanto a los que no recuperaron, como a los buscadores para los que la ejecución de ésta, actuó de filtro en la recuperación. La búsqueda por frase ocupa un lugar intermedio con 433 recursos analizados.

[242] No se pudo obtener información de las búsquedas 1, 2, 3 al no poderse guardar la página de búsqueda desde la que posteriormente se extrajeron los datos.

2. El software de recuperación

2.1. Análisis del funcionamiento de los motores ante los distintos tipos de búsqueda

Analizamos a continuación el funcionamiento de los buscadores ante las búsquedas planteadas para conocer la capacidad de recuperación que tienen estas herramientas ante consultas especializadas. Basándonos en los cincuenta primeros resultados, podemos conocer cuáles de estas herramientas alcanzan en las diferentes búsquedas este número de recursos y cuáles recuperan un menor número o ningún recurso.

2.1.1. Capacidad de búsqueda

Los resultados muestran los problemas que las diferentes búsquedas han supuesto para los buscadores ya que en algunos casos no recuperan documentos y en otros, el número de recursos recuperados no llega a los cincuenta. Ambos aspectos pueden resultar indicativos de las limitaciones de estas herramientas en la recuperación, por un lado al no estar provistas tecnológicamente de mecanismos que interpreten, traduzcan y ejecuten de forma correcta este tipo de búsquedas, y por otro, la existencia de una base de datos reducida en cuanto a recursos especializados.

Tabla 2.1.1-1. Motores. Capacidad de búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total
Google	50	50	50	50	50	50	**300**
MSN	50	50	50	50	50	50	**300**
Teoma (Ask)	50	50	50	Sin resultados			**150**
WiseNut	50	50	50	Sin resultados	50	30	**230**
Yahoo	50	50	50	50	50	50	**300**
Total pág. analizadas	**250**	**250**	**250**	**150**	**200**	**180**	**1280**

Tabla 2.1.1.-2. Metabuscadores. Capacidad de búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total
Dogpile	50	43	50	Sin resultados			**143**
Excite	50	50	50	34	50	50	**284**
Ixquick	32	39	50	21	44	23	**209**
Profusion	41	35	Sin resultados	46	39	25	**186**

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total
Search	50	50	50	50	50	50	**300**
Surfwax	15	19	36	Sin resultados			**70**
Vivisimo	50	50	50	50	50	50	**300**
Total pág. analizadas	**288**	**286**	**286**	**201**	**233**	**198**	**1492**

Podemos observar que el funcionamiento entre motores de búsqueda y metabuscadores en las distintas búsquedas da lugar a resultados muy diferente entre ellos. Así tenemos que la búsqueda por un solo término no plantea problemas en ninguno de los casos, salvo la limitación de recursos recuperados que se observa en los metabuscadores Ixquick con 32 recursos, Profusión con 41 y Surfwax con 15.

La búsqueda que utiliza el lenguaje natural plantea problemas para el buscador WiseNut que no funciona con más de siete términos de búsqueda. Los metabuscadores Dogpile, Ixquick, Profusion y Surfwax ofrecen un número limitado de recursos que en ningún caso, en esta búsqueda llega a los 50.

En la búsqueda que utiliza el operador de existencia (+) para forzar a los buscadores a recuperar recursos que contengan, además de las palabras clave solicitadas, los términos señalados con este signo, aunque se trate de *stop words,* se plantean mayores problemas ya que, entre los metabuscadores, Profusion no ofrece resultados y Surfwax lo hace en un número limitado (36).

En la búsqueda booleana los motores Teoma (Ask) y WiseNut no ofrecen resultados al igual que los metabuscadores Dogpile y Surfwax. En Ixquick también se aprecia un funcionamiento anómalo ya que indica en la página de resultados que recupera 26 registros únicos cuando en realidad son 21. Excite y Profusión ofrecen respectivamente 34 y 46 recursos, lo cuál, como veremos al analizar la precisión técnica, no es indicativo, por desgracia, de un correcto funcionamiento basado en la filtración de sólo los recursos que contienen los términos tal y como aparecen expresados en la ecuación de búsqueda, sino que se trata de una recuperación con unos resultados limitados e insuficientes.

En la búsqueda por frase es Teoma (Ask) el único motor que no ofrece resultados al igual que ocurre con los metabuscadores Dogpile y Surfwax. Ixquick y Profusion, también en esta ocasión facilitan un número limitado de recursos que no llega a cincuenta.

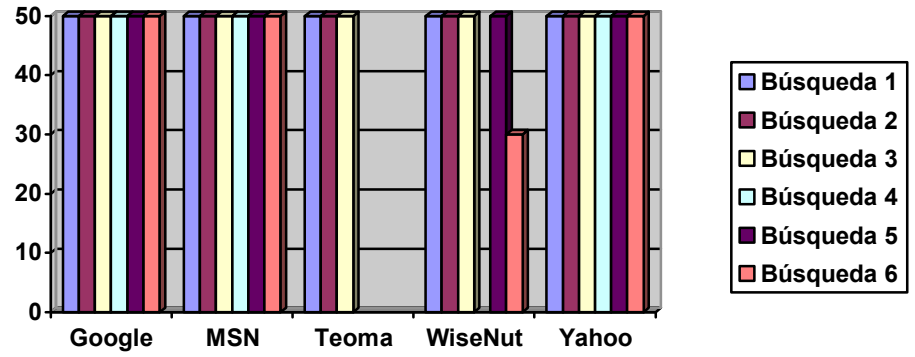
En la búsqueda por campo Teoma (Ask) no recupera resultados y WiseNut sólo ofrece treinta. También los metabuscadores Ixquick y Profusión ofrecen resultados limitados al presentar veintitrés y veinticinco resultados respectivamente. Dogpile y Surfwax tampoco recuperan en esta búsqueda.

Por tanto los buscadores que mejor han funcionado en las seis búsquedas han sido Google, MSN y Yahoo. Respecto a los metabuscadores, han sido Search, Vivisimo y Excite, auque éste último de forma más limitada al no obtener tantos registros como los otros dos en la búsqueda booleana. A continuación podemos situar a Profusión, que no funcionó en la búsqueda con operadores de existencia, y finalmente Dogpile y Surfwax, que no funcionaron en las tres últimas consultas.

A lo largo de las seis búsquedas podemos observar las limitaciones que presenta el buscador Teoma (Ask) en la búsqueda booleana, por frase y por campo, y WiseNut en la búsqueda booleana, que es similar a la que se aprecia en los metabuscadores Dogpile y Surfwax.

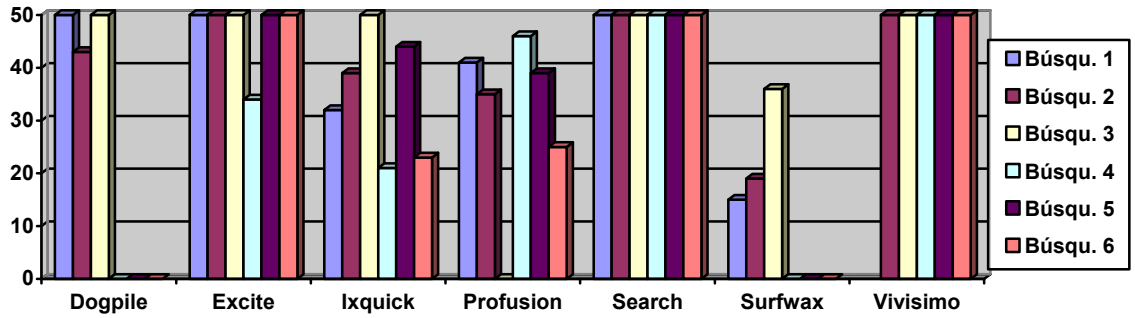
Análisis global

Gráfico 2.1.1-1 Recuperación de los motores en las seis búsquedas



Los buscadores que mejor responden a las búsquedas especializadas son Google, MSN y Yahoo, frente a WiseNut que no funcionó en la búsqueda booleana y recuperó treinta resultados en la búsqueda en el campo de título. Teoma (Ask) no recuperó resultados en ninguna de las tres últimas búsquedas.

Gráfico 2.1.1-2 Recuperación de los metabuscadores en las seis búsquedas



Los metabuscadores que recuperan en todas las búsquedas son Excite, Ixquick, Search y Vivisimo. Dogpile y Surfwax sólo recuperan en las tres primeras. No obstante se aprecian limitaciones en la recuperación de Excite, que en la búsqueda con operadores booleanos, recuperó treinta y cuatro recursos. Ixquick sólo alcanzó los cincuenta resultados en la búsqueda con operadores de existencia y Profusión, que no llegó a cincuenta resultados en ninguna de las búsquedas, falló en la búsqueda con operadores de existencia. El peor comportamiento corresponde al metabuscador Surfwax, que sólo obtuvo resultados, y de forma limitada, en las tres primeras búsquedas.

2.2. Análisis de la presentación de los resultados y de la información de los registros recuperados

Nos hemos referido con anterioridad a la importancia de la presentación de los registros para decidir si interesa o no la consulta de los recursos recuperados. De aquí que a continuación valoremos la información que ofrece el título, para conocer hasta qué punto se utiliza la metainformación en la elaboración de los listados, la frecuencia de aparición de los términos de búsqueda de forma destacada en el registro, la indicación en el listado de recursos relacionados con otros del mismo sitio web, y la frecuencia de aparición de recursos de carácter comercial.

2.2.1. Análisis del uso de metainformación en función de la coincidencia de los títulos de la etiqueta <*title*> y del listado de recuperación

El objeto de este análisis es valorar en la presentación de los resultados el uso que los buscadores dan a la metainformación, en este caso a la etiqueta <*title*> para que aparezca su contenido encabezando el recurso. Para ello comparamos la información de esta etiqueta con el título del recurso, lo que nos permite conocer cuál herramienta ofrece una información más relacionada con el título del recurso y por tanto con su contenido.

Las tablas siguientes muestran, en las diferentes búsquedas, los casos de coincidencias entre unos y otros. Podemos apreciar igualmente cuales son los motores que más uso hacen de la información de esta etiqueta para insertar su contenido en el encabezamiento del registro.

Tabla 2.2.1-1. Motores. Uso de metaetiqueta <title> en la elaboración de los listados

	Búsqueda 1 (Término único)			**Búsqueda 2 (Lenguaje natural)**			**Búsqueda 3 (Operadores de existencia)**		
Coincidencia de Títulos	**Si**	**No**	**n.s.a.[1]**	**Si**	**No**	**n.s.a.**	**Si**	**No**	**n.s.a.**
Google	19 (38%)	16 (32%)	15 (30%)	23 (46%)	19 (38%)	8 (16%)	32 (64%)	18 (36%)	0
MSN	34 (68%)	15 (30%)	1 (2%)	25 (50%)	19 (38%)	6 (12%)	33 (66%)	17 (34%)	0
Teoma (Ask)	30 (60%)	19 (38%)	1 (2%)	23 (46,9%)	21 (42,9%)	5 (10,2%)	35 (70%)	15 (30%)	0
WiseNut	32 (64%)	17 (34%)	1 (2%)	37 (74%)	13 (26%)	0	31 (62%)	19 (38%)	0
Yahoo	31 (62%)	10 (20%)	9 (18%)	36 (72%)	11 (22%)	3 (6%)	19 (38%)	31 (62%)	0

	Búsqueda 4 (Operadores booleanos)			**Búsqueda 5 (Frase)**			**Búsqueda 6 (Campo título)**		
Coincidencia de Títulos	Si	No	n.s.a.	Si	No	n.s.a.	Si	No	n.s.a.
Google	25 (50%)	19 (38%)	6 (12%)	36 (72%)	14 (28%)	0	39 (78%)	10 (20%)	1 (2%)
MSN	37 (74%)	11 (22%)	2 (4%)	34 (68%)	16 (32%)	0	37 (74%)	13 (26%)	0
Teoma (Ask)	Sin resultados								
WiseNut	Sin resultados			32 (64%)	18 (36%)	0	14 (46,7%)	16 (53%)	0
Yahoo	41 (82%)	8 (16%)	0	39 (78%)	7 (14%)	4 (8%)	43 (86%)	6 (12%)	1 (2%)

[1] N.s.a. es la abreviatura de No se pudo analizar que corresponde a los recursos, que al tratarse de documentos en pdf o Word, etc. no contienen etiqueta TITLE.

En la primera búsqueda MSN, seguido de WiseNut, Yahoo y Teoma (Ask) son los motores que recuperan más recursos en los que la metainformación y el título del registro coinciden. En Google, la coincidencia es menor.

En la segunda búsqueda, los mayores porcentajes corresponden a WiseNut (74%) y Yahoo (72%), no superando el resto de buscadores el 50%. En la tercera búsqueda aumenta su utilización en Teoma (Ask) (70%), MSN (66%) y Google (64%). Sin embargo, Yahoo desciende considerablemente en su uso (38%).

En la cuarta búsqueda son de nuevo Yahoo (82%), MSN con un (74%), los motores que más utiliza la metainformación. La quinta búsqueda da un mayor uso de la metainformación a Yahoo (78%), seguido de Google (72%). Finalmente en la sexta búsqueda es Yahoo (86%) seguido de Google (78%) y MSN (74%).

Los metabuscadores aportan los siguientes datos:

Tabla 2.2.1-2. Metabuscadores. Uso de metaetiqueta <title> en la elaboración de los listados

	Búsqueda 1 (Término único)			Búsqueda 2 (Lenguaje natural)			Búsqueda 3 (Operadores de existencia)		
	Si	No	n.s.a.	Si	No	n.s.a.	Si	No	n.s.a.
Dogpile	38 (76%)	7 (14%)	5 (10%)	24 (55,8%)	16 (37,2%)	3 (7%)	24 (49%)	25 (51%)	0
Excite	27 (54%)	15 (30%)	8 (16%)	33 (66%)	13 (26%)	4 (8%)	26 (52%)	24 (48%)	0
Ixquick	20 (62,5%)	7 (21,9%)	5 (15,6%)	24 (61,5%)	11 (28,2%)	4 (10,3%)	21 (42%)	29 (58%)	0
Profusion	26 (63,4%)	13 (31,7%)	2 (4,9%)	13 (37,1%)	20 (57,1%)	2 (5,7%)	Sin resultados		
Search	38 (76%)	7 (14%)	5 (10%)	35 (70%)	9 (18%)	6 (12%)	33 (66%)	14 (28%)	3 (6%)
Surfwax	9 (60%)	6 (40%)	0	6 (31,6%)	13 (68,4%)	0	5 (13,9%)	31 (86,1%)	0
Vivisimo	33 (66%)	13 (26%)	4 (8%)	29 (58%)	18 (36%)	3 (6%)	25 (50%)	25 (50%)	0

	Búsqueda 4 (Operadores booleanos)			Búsqueda 5 (Frase)			Búsqueda 6 (Campo título)		
	Si	No	n.s.a.	Si	No	n.s.a.	Si	No	n.s.a.
Dogpile	Sin resultados								
Excite	20 (58,8%)	14 (51,2%)	0	14 (28%)	34 (68%)	2 (4%)	28 (56%)	19 (38%)	0
Ixquick	12 (57,1%)	8 (38,1%)	1 (4,8%)	24 (54,5%)	20 (45,5%)	0	17 (73%)	6 (26%)	0
Profusion	12 (26,1%)	33 (71,7%)	1 (2,2%)	16 (41%)	23 (59%)	0	11 (44%)	13 (52%)	0
Search	37 (74%)	13 (26%)	0	33 (66%)	16 (32%)	1 (2%)	22 (44%)	18 (36%)	10 (20%)
Surfwax	Sin resultados								
Vivisimo	28 (56%)	15 (30%)	7 (14%)	14 (28%)	35 (70%)	1 (2%)	38 (76%)	12 (24%)	0

En la primera búsqueda los metabuscadores que más utilizan la metainformación para insertarla como título de los listados son Dogpile y Search. También es utilizada por Vivisimo, Excite e Ixquick. En Surfwax, aunque el número pueda parecer escaso (9), el porcentaje es elevado (60%).

En la segunda búsqueda las cantidades y porcentajes descienden con respecto a la anterior, sobre todo en el caso de Dogpile, Vivisimo y Surfwax. Ixquick mantiene cifras similares. En Search y Excite hay una mayor coincidencia entre los títulos.

En la tercera búsqueda se mantiene la tendencia descendente y en la cuarta búsqueda destaca Search con un porcentaje de coincidencia del 74%, aunque en el resto de metabuscadores los porcentajes son próximos al 60%. Las cantidades vuelven a bajar en la quinta búsqueda respecto a la anterior, a excepción del caso de Profusión.

Finalmente, en la búsqueda por campo, hay metabuscadores que alcanzan de nuevo altos porcentajes de coincidencia como Vivisimo 38 (76%), Excite 28 (56%) o Ixquick 17 (73%).

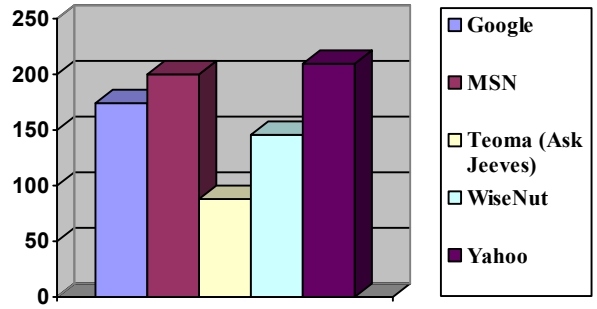
Análisis global

La siguiente tabla recoge los datos correspondientes a las coincidencias entre ambos títulos registrados en cada buscador en las seis búsquedas.

Tabla 2.2.1-3. Utilización de metainformación

Motor	Total
Google	174
MSN	200
Teoma (Ask Jeeves)	88
WiseNut	146
Yahoo	209

Gráfico 2.2.1-1. Motores. Utilización de metaetiqueta <title> en las seis búsquedas



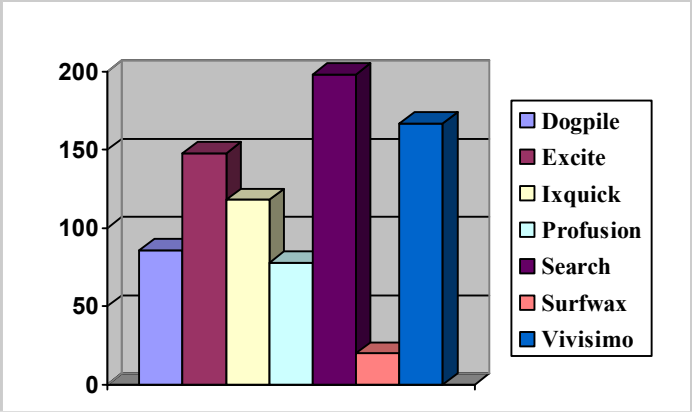
El gráfico que la acompaña facilita la visualización de los resultados, lo que nos permite observar que Yahoo es el buscador en el que más coincidencias hay entre los títulos, presentando por lo tanto una información en los listados más acorde con el contenido de las páginas. MSN también utiliza asiduamente la metainformación, y algo menos Google y WiseNut. Teoma (Ask) es el buscador en que se observa una menor utilización de la información en los listados de recuperación.

Como en el caso anterior, la siguiente tabla recoge los datos correspondientes a las coincidencias entre ambos títulos registrados en cada metabuscador en las seis búsquedas.

Tabla 2.2.1-4. Utilización de metainformación

Metabuscador	Total
Dogpile	86
Excite	148
Ixquick	118
Profusion	78
Search	198
Surfwax	20
Vivisimo	167

Gráfico 2.2.1-2. Metabuscadores. Utilización de metaetiqueta <title> en las seis búsquedas



Podemos observar cómo Search es el metabuscador que más utiliza la información de esta etiqueta, seguido de Vivisimo, Excite e Ixquick. Un menor uso corresponde a Dogpile, Profusión y Surfwax.

2.2.2. Términos de búsqueda destacados

Presentamos a continuación los resultados ofrecidos al contabilizar el número de recursos cuyos registros contienen los términos de búsqueda destacados en los listados de recuperación y en cada una de las búsquedas. Esta información es útil de cara al usuario al facilitarle el contexto en el que aparecen los términos solicitados permitiendo valorar su interés. Por tanto, el buscador que en mayor medida presenta los términos destacados, está ofreciendo una mayor información al usuario.

Tabla 2.2.2-1. Motores. Términos de búsqueda destacados

	Búsqueda 1 (Término único)		Búsqueda 2 (Lenguaje natural)		Búsqueda 3 (Operadores de existencia)	
	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas
Google	47 (94%)	3 (6%)	50 (100%)	0	50 (100%)	0
MSN	43 (86%)	7 (14%)	49 (98%)	1 (2%)	50 (100%)	0
Teoma (Ask)	31 (62%)	19 (38%)	50 (100%)	0	49 (98%)	1 (2%)
WiseNut	46 (92%)	4 (8%)	50 (100%)	0	50 (100%)	0
Yahoo	50 (100%)	0	50 (100%)	0	46 (92%)	4 (8%)

	Búsqueda 4 (Operadores booleanos)		Búsqueda 5 (Frase)		Búsqueda 6 (Campo título)		Total
	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº reg. palabras destacadas
Google	50 (100%)	0	47 (94%)	3 (6%)	50 (100%)	0	**6**
MSN	48 (96%)	2 (4%)	46 (92%)	4 (8%)	48 (96%)	2 (4%)	**16**
Teoma (Ask)	Sin resultados		Sin resultados		Sin resultados		**20**
WiseNut	Sin resultados		50 (100%)	0	30 (100%)	0	**4**
Yahoo	33 (66%)	17 (34%)	50 (100%)	0	50 (100%)	0	**21**

En la primera búsqueda Teoma (Ask) es el motor que presenta un mayor número de registros con el término de búsqueda destacado 19 (38%), MSN presenta 7 (14%) y Google 3 (6%). En la búsqueda por varios términos, con operadores de existencia y en la búsqueda por campo, apenas se utiliza esta técnica, a excepción de los pocos resultados registrados por MSN, Teoma (Ask) y los cuatro de Yahoo. Hay que señalar además que éste buscador no elimina palabras vacías como "in", de aquí que el número registros con palabras destacadas puede corresponder a casos en los que se destaquen simplemente palabras vacías, lo que denota un incorrecto funcionamiento ya que no aportan nada de información al usuario, y el buscador debería presentarlas destacadas sólo si aparecen junto al resto y no de forma aislada como lo vienen haciendo. Esta observación es extensible a todos los buscadores ya que en todos los casos hemos apreciado que términos, como la conjunción inglesa "and", aparece resaltada.

En la cuarta búsqueda destaca Yahoo con 17 (34%) registros con los términos de búsqueda destacados. De nuevo, la explicación es que este buscador destaca palabras

como "and", que en este caso además, se trata de un operador lógico. Generalmente es el único término que destaca.

En la quinta búsqueda Google y MSN utilizan en tres y cuatro casos respectivamente la presentación de registros con los términos de búsqueda destacados. En Yahoo, la página de resultados no destaca todos los términos de la búsqueda aún estando junto a otros que sí destaca, utilizando por tanto esta técnica de una forma poco consistente.

En la sexta búsqueda tan sólo MSN presenta dos registros con los términos destacados.

Estos resultados muestran, por un lado el limitado uso que de esta técnica hacen los motores de búsqueda y por otro, en los casos en los que se utiliza, el deficiente funcionamiento y la poca utilidad para el usuario, ya que a menudo se destacan palabras vacías y no las representativas de la búsqueda.

Tabla 2.2.2-2. Metabuscadores. Términos de búsqueda destacados

	Búsqueda 1 (Término único)		Búsqueda 2 (Lenguaje natural)		Búsqueda 3 (Operadores de existencia)	
	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas
Dogpile	36 (72%)	14 (28%)	30 (69,8%)	12 (27,9%)	47 (94%)	3 (6%)
Excite	33 (66%)	17 (34%)	39 (79,6%)	10 (20,4%)	46 (92%)	4 (8%)
Ixquick	32 (100%)	0	39 (100%)	0	48 (96%)	2 (4%)
Profusion	37 (90,2%)	4 (9,8%)	35 (100%)	0	Sin resultados	
Search	48 (96%)	2 (4%)	49 (100%)	0	49 (98%)	1 (2%)
Surfwax	4 (26,7%)	11 (73,3%)	4 (21,1%)	15 (78,9%)	15 (41,7%)	21 (58,3%)
Vivisimo	45 (90%)	5 (10%)	49 (100%)	0	48 (96%)	2 (4%)

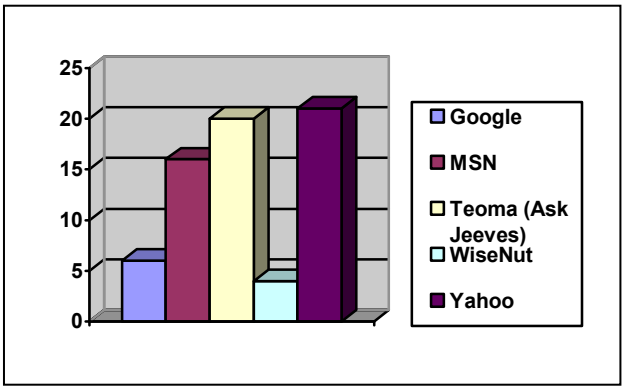
	Búsqueda 4 (Operadores booleanos)		Búsqueda 5 (Frase)		Búsqueda 6 (Campo título)		Total
	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº de reg. que no presentan palabras destacadas	Nº de reg. que presentan palabras destacadas	Nº reg. palabras destacadas
Dogpile	Sin resultados		Sin resultados		Sin resultados		**29**
Excite	32 (94,1%)	2 (5,9%)	46 (92%)	4 (8%)	25 (50%)	25 (50%)	**62**
Ixquick	21 (100%)	0	46 (92%)	4 (8%)	21 (91,3%)	2 (8,7%)	**10**
Profusion	0	46 (100%)	0	39 (100%)	0	25 (100%)	**113**
Search	49 (98%)	1 (2%)	50 (100%)	0	44 (88%)	6 (12%)	**10**
Surfwax	Sin resultados		Sin resultados		Sin resultados		**47**
Vivisimo	42 (84%)	8 (16%)	50 (100%)	0	49 (100%)	0	**15**

Respecto a los metabuscadores, en la primera búsqueda Excite y Dogpile son los que más registros presentan con esta técnica. Surfwax, aún con menos recursos recuperados (15), presenta un alto porcentaje (73,3%) de recursos con los términos destacados. Esta tendencia se advierte en las dos siguientes búsquedas, destacando este mismo metabuscador en la tercera búsqueda con 21 (58,3%) registros que resaltan alguno de los términos de búsqueda. En las tres últimas búsquedas, Profusion muestra un importante uso de esta técnica.

Análisis global

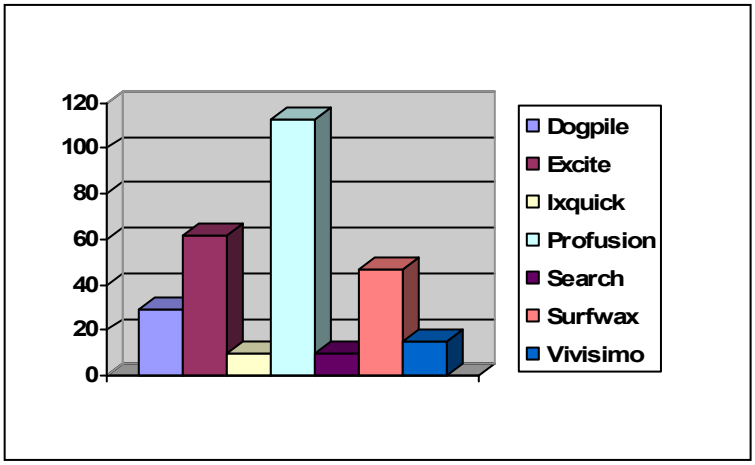
El siguiente gráfico muestra la utilización por parte de los motores de búsqueda, de la técnica de destacar los términos en los registros de los listados de recuperación. Los datos se refieren al total de las seis búsquedas.

Gráfico 2.2.2-1. Utilización de palabras destacadas por los motores en las seis búsquedas



Podemos afirmar que es una técnica poco usada por los buscadores siendo Yahoo, Teoma (Ask) y MSN los motores que con mayor frecuencia destacan, en los registros encontrados los términos de búsqueda, y Google y WiseNut los que apenas la utilizan.

Gráfico 2.2.2-2. Utilización por los metabuscadores de palabras destacadas en las seis búsquedas



Profusión utiliza esta técnica de forma destacada, seguido de Excite y Surfwax. Con menor frecuencia destacan los términos Dogpile, Vivisimo, Ixquick y Search.

Al margen de los datos cuantitativos, hemos podido observar el mal uso de esta técnica tanto por parte de buscadores como de los metabuscadores, ya que a menudo ofrecen de forma destacada términos que no son informativos para el usuario, como es el caso de palabras vacías.

2.2.3. Recursos dependientes

Analizamos a continuación la frecuencia de aparición, en los listados de resultados, de recursos dependientes de otros del mismo sitio web, y que se identifican porque el registro aparece con unos márgenes mayores que el recurso del cuál depende. Este aspecto resulta de gran utilidad para el usuario ya que al existir una relación de dependencia entre los recursos, puede decidir, consultando uno de ellos, si el otro, alojado en el mismo sitio web, es de su interés. Por otro lado también puede ser significativo de la profundidad o extensión de la indización de un sitio web. No obstante hay que advertir que es una técnica que no todas las herramientas de búsqueda utilizan.

Tabla 2.2.3-1. Motores. Recursos dependientes

	Búsqueda 1 (Término único)		Búsqueda 2 (Lenguaje natural)		Búsqueda 3 (Operadores de existencia)	
	Si	No	Si	No	Si	No
Google	4 (8%)	46 (92%)	4 (8%)	46 (92%)	0	50 (100%)
MSN	0	50 (100%)	0	50 (100%)	0	50 (100%)
Teoma (Ask)	4 (8%)	46 (92%)	0	49 (98%)	1 (2%)	49 (98%)
WiseNut	0	50 (100%)	0	50 (100%)	0	50 (100%)
Yahoo	0	50 (100%)	0	50 (100%)	0	50 (100%)

	Búsqueda 4 (Operadores booleanos)		Búsqueda 5 (Frase)		Búsqueda 6 (Campo título)		Total
	Si	No	Si	No	Si	No	
Google	2 (4%)	48 (96%)	1 (2%)	49 (98%)	0	50 (100%)	**11**
MSN	1 (2%)	49 (98%)	0	50 (100%)	0	50 (100%)	**1**
Teoma (Ask)	Sin resultados		Sin resultados		Sin resultados		**5**
WiseNut	Sin resultados		0	50 (100%)	0	30 (100%)	**0**
Yahoo	0	50 (100%)	0	50 (100%)	0	50 (100%)	**0**

Como podemos observar en las búsquedas 1, 2, 4 y 5 Google es el motor que presenta en los listados, de una forma regular, recursos dependientes de otro inmediatamente anterior, lo que puede ser indicativo de una mayor profundidad en la indización de sitios

web. Teoma (Ask) también presentó recursos dependientes en la búsqueda 1 y 3. MSN sólo presentó un resultado en la búsqueda 4, mientras que Yahoo y WiseNut no utilizan esta técnica.

Tabla 2.2.3-2. Metabuscadores. Recursos dependientes

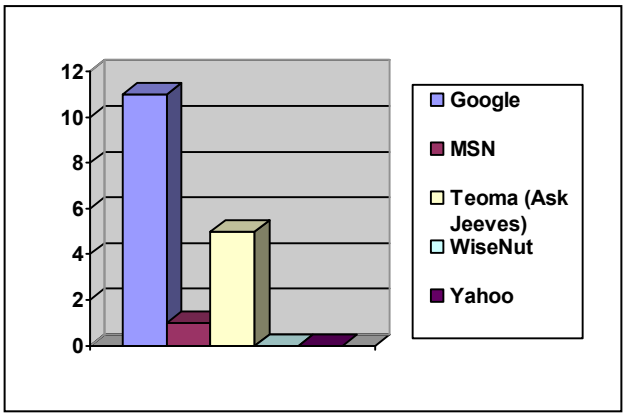
	Búsqueda 1 (Término único)		Búsqueda 2 (Lenguaje natural)		Búsqueda 3 (Operadores de existencia)	
	Si	No	Si	No	Si	No
Dogpile	0	50 (100%)	0	43 (100%)	0	50 (100%)
Excite	3 (6%)	47 (94%)	0	50 (100%)	0	50 (100%)
Ixquick	0	32 (100%)	0	39 (100%)	0	50 (100%)
Profusion	0	41 (100%)	0	35 (100%)	Sin resultados	
Search	2 (4%)	48 (96%)	0	50 (100%)	0	50 (100%)
Surfwax	0	15 (100%)	0	19 (100%)	0	36 (100%)
Vivisimo	0	50 (100%)	0	50 (100%)	0	50 (100%)

	Búsqueda 4 (Operadores booleanos)		Búsqueda 5 (Frase)		Búsqueda 6 (Campo título)		Total
	Si	No	Si	No	Si	No	
Dogpile	Sin resultados		Sin resultados		Sin resultados		**0**
Excite	0	34 (100%)	0	50 (100%)	0	50 (100%)	**3**
Ixquick	0	21 (100%)	0	44 (100%)	0	23 (100%)	**0**
Profusion	0	46 (100%)	0	39 (100%)	0	25 (100%)	**0**
Search	0	50 (100%)	0	50 (100%)	0	50 (100%)	**2**
Surfwax	Sin resultados		Sin resultados		Sin resultados		**0**
Vivisimo	0	50 (100%)	0	50 (100%)	2 (4%)	48 (96%)	**2**

Los metabuscadores utilizan esta técnica de forma muy limitada como podemos observar en la tabla de resultados. Excite, Search y Vivisimo son los únicos que la utilizan. En la búsqueda 1, Excite y Search son los únicos que presentan resultados dependientes y en el resto de búsquedas no se dan más casos, a excepción de dos registros en la sexta búsqueda por parte de Vivisimo, lo que indica que se trata de una técnica no muy implantada en los metabuscadores.

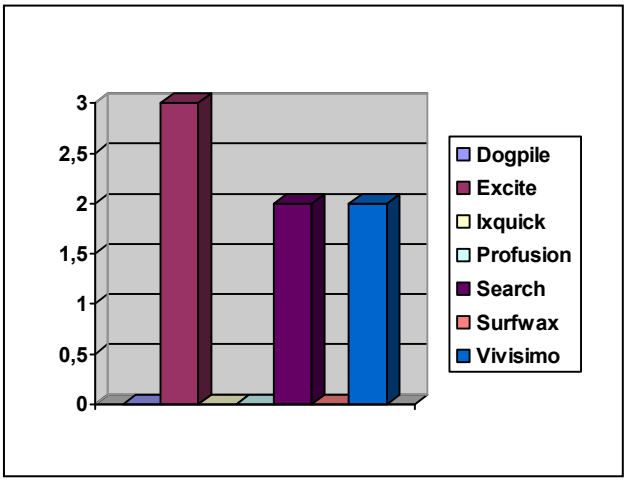
Análisis global

Gráfico 2.2.3-1. Motores. Recursos dependientes



Google seguido de Teoma (Ask) es el motor que utiliza con mayor frecuencia esta técnica, que facilita la consulta de los recursos recuperados por los buscadores, al colocar juntos en los listados, y de forma visible, recursos relacionados de un mismo sitio web. WiseNut y Yahoo no presentan juntos los recursos dependientes.

Gráfico 2.2.3-2. Metabuscadores. Recursos dependientes



Entre los metabuscadores Excite, Search y Vivisimo son los que, de forma esporádica, presentan resultados dependientes, frente a Dogpile, Ixquick, Profusion y Surfwax, que no lo hacen.

El poco uso de esta técnica entre los metabuscadores puede ser indicativo de la selección de recursos que realizan, siendo poco frecuente que, de las bases de datos de las que se sirven, extraigan más de un recurso del mismo sitio web.

2.2.4. Enlaces a páginas de contenido publicitario en los listados

En los listados de resultados suelen aparecer, en diferentes partes de la página, una serie de enlaces a páginas cuyos propietarios pagan por aparecer de forma destacada. Recogemos a continuación una valoración que nos permita conocer hasta qué punto los motores se sirven de esta técnica comercial para favorecer su financiación. Contabilizamos los enlaces que aparecen en el total de páginas[2] que contienen los resultados de las búsquedas analizadas.

Tabla 2.2.4-1. Motores. Contenido de enlaces a páginas de publicidad

Nº de recursos publicitarios	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Google	0	1	1	1	1	3	**7**
MSN	0	0	45	0	0	4	**49**
Teoma (Ask)	0	15	3	Sin resultados	Sin resultados	Sin resultados	**18**
WiseNut	0	1	0	Sin resultados	0	0	**1**
Yahoo	0	0	0	0	0	0	**0**

El buscador que inserta un mayor número de registros de carácter comercial en el listado de resultados es MSN, que proporciona en cada una de las páginas de recursos una serie de enlaces, tanto al comienzo como al final de la página, pertenecientes a empresas que pagan por hacer que aparezcan sus recursos en lugares destacados.

Google ofrece resultados comerciales pero en menor medida que el motor anterior. Yahoo no inserta los resultados de carácter comercial junto al resto de resultados.

Tabla 2.2.4-2. Metabuscadores. Contenido de enlaces a páginas de publicidad

Nº de recursos publicitarios	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Dogpile	0	0	0	Sin resultados	Sin resultados	Sin resultados	**0**
Excite	0	0	0	0	0	0	**0**
Ixquick	0	0	2	0	2	0	**4**
Profusion	0	2	Sin resultados	0	1	0	**3**

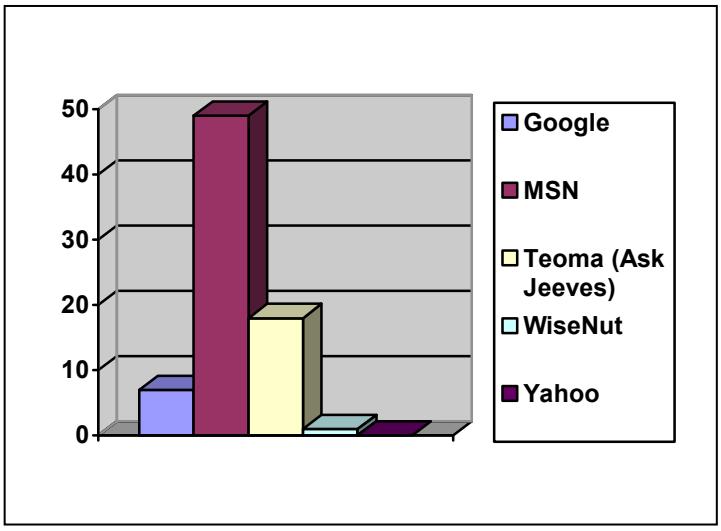
[2] Normalmente se trata de las cinco páginas en las que aparecen los cincuenta resultados utilizados en la evaluación.

Nº de recursos publicitarios	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Search[3]	No se pudo analizar este aspecto	No se pudo analizar este aspecto	No se pudo analizar este aspecto	3	4	5	**12**
SurfWax	1	2	1	Sin resultados	Sin resultados	Sin resultados	**4**
Vivisimo	1	2	2	0	0	0	**5**

Search es el metabuscador que más recursos comerciales ofrece, como podemos observar en sus tres últimas búsquedas, seguido por Vivisimo, que los contiene en las tres primeras.

Análisis global

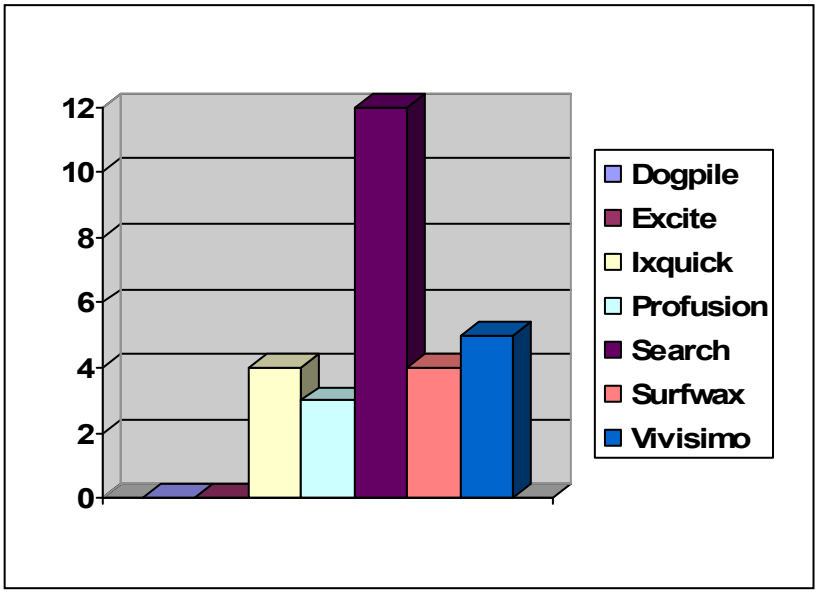
Gráfico 2.2.4-1. Motores. Enlaces a recursos publicitarios



MSN seguido de Teoma (Ask) son los motores de búsqueda que más enlaces comerciales proporcionan, eso sí, siempre de una forma destacada del resto de recursos. Google, WiseNut y fundamentalmente Yahoo, son los que menos recursos comerciales ofrecen en las páginas de resultados.

[3] No se pudo obtener información de las búsquedas 1, 2, 3 al no poderse guardar la página de búsqueda, desde la que posteriormente se extrajeron los datos.

Gráfico 2.2.4-2. Metabuscadores. Enlaces a recursos publicitarios



Entre los metabuscadores es Search el que más recursos de carácter comercial ofrece. Ixquick, Profusion, Surfwax y Vivisimo apenas insertan enlaces comerciales, y Dogpile y Excite, no ofrecen recursos de este tipo.

El aspecto analizado puede ayudarnos a comprender la filosofía de estas herramientas, en cuanto a que han de utilizar técnicas comerciales para obtener recursos económicos que faciliten su mantenimiento y desarrollo. La presentación de recursos seleccionados en función de estas técnicas comerciales resulta de gran interés y pensamos que hay que tenerlo en cuenta en las evaluaciones de estas herramientas ya que, bien sea presentando los resultados de este tipo en lugares destacados del listado de resultados, incorporándolos rápidamente a la base de datos y actualizándolos de forma más frecuente que el resto, son aspectos que influyen en la recuperación y no pueden pasar desapercibidos. Lo que resulta interesante es la existencia de técnicas que, igual que se utilizan para los recursos comerciales, podrían usarse para destacar recursos especializados o de investigación.

3. Los componentes de los buscadores y características de la información recuperada

3.1. Aspectos relacionados con el robot o crawler

Para conocer la profundidad con la que los buscadores rastrean los sitios web al localizar e indizar sus páginas, hemos elaborado las siguientes tablas en las que se señala el nivel en el que se encuentran las páginas recuperadas dentro de la jerarquía de directorios del sitio web.

3.1.1. Profundidad de indización del sitio web

Tabla 3.1.1-1. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por término único

	Búsqueda 1 (Término único)									
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**	**9º**	**13º**
Google	0	8 (16%)	15 (30%)	11 (22%)	8 (16%)	4 (8%)	3 (6%)	0	0	1 (2%)
MSN	1 (2,1%)	10 (20,8%)	21 (43,8%)	8 (16,7%)	4 (8,3%)	1 (2,1%)	2 (4,2)	1 (2,1%)	0	0
Teoma (Ask)	7 (14%)	13 (26%)	10 (20%)	9 (18%)	8 (16%)	0	1 (2%)	2 (4%)	0	0
WiseNut	3 (6%)	12 (24%)	18 (36%)	10 (20%)	1 (2%)	4 (8%)	2 (4%)	0	0	0
Yahoo	1 (2%)	9 (18%)	7 (14%)	11 (22%)	9 (18%)	5 (10%)	3 (6%)	4 (8%)	1 (2%)	0

Tabla 3.1.1-2. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por varios términos

	Búsqueda 2 (Lenguaje natural)						
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**
Google	1 (2%)	2 (4%)	19 (38%)	25 (50%)	1 (2%)	2 (4%)	0
MSN	4 (8%)	10 (20%)	15 (30%)	15 (30%)	5 (10%)	1 (2%)	0
Teoma (Ask)	0	9 (19,1%)	14 (29,8%)	11 (23,4%)	9 (19,1%)	2 (4,3%)	2 (4,3%)
WiseNut	4 (8%)	19 (38%)	14 (28%)	9 (18%)	4 (8%)	0	0
Yahoo	0	7 (14,6%)	16 (33,3%)	20 (41,7%)	4 (8,3%)	1 (2,1%)	0

Tabla 3.1.1-3. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda con operadores de existencia

	Búsqueda 3 (Operadores de existencia)							
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**
Google	18 (37,5%)	6 (12,5%)	11 (22,9%)	7 (14,6%)	4 (8,3%)	1 (2,1%)	0	1 (2,1%)
MSN	17 (34,7%)	6 (12,2%)	14 (28,6%)	8 (16,3%)	4 (8,2%)	0	0	0
Teoma (Ask)	9 (18,4%)	5 (10,2%)	11 (22,4%)	14 (28,6%)	6 (12,2%)	3 (6,1%)	1 (2%)	0
WiseNut	0	7 (14,9%)	15 (31,9%)	16 (34%)	8 (17%)	0	1 (2,1%)	0
Yahoo	7 (15,2%)	10 (21,7%)	12 (26,1%)	9 (19,6%)	4 (8,7%)	3 (6,5%)	1 (2,2%)	0

Tabla 3.1.1-4. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda con Operadores booleanos

	Búsqueda 4 (Operadores booleanos)							
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**11º**
Google	0	6 (12,2%)	15 (30,6%)	13 (26,5%)	9 (18,4%)	4 (8,2%)	2 (4,1%)	0
MSN	3 (6%)	11 (22%)	16 (32%)	9 (18%)	8 (16%)	2 (4%)	1 (2%)	0
Teoma (Ask)	Sin resultados							
WiseNut	Sin resultados							
Yahoo	0	6 (12%)	13 (26%)	18 (36%)	5 (10%)	4 (8%)	3 (6%)	1 (2%)

Tabla 3.1.1-5. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por frase

	Búsqueda 5 (Frase)										
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**	**9º**	**10º**	**11º**
Google	6 (12%)	14 (28%)	13 (26%)	9 (18%)	7 (14%)	0	0	0	0	0	1 (2%)
MSN	9 (18%)	11 (22%)	13 (26%)	10 (20%)	5 (10%)	0	0	1 (2%)	1 (2%)	0	0
Teoma (Ask)	Sin resultados										
WiseNut	8 (18,4%)	12 (24,5%)	12 (24,5%)	9 (18,4%)	6 (12,2%)	1 (2%)	0	0	0	0	0
Yahoo	4 (8%)	5 (10%)	15 (30%)	17 (34%)	6 (12%)	1 (2%)	0	1 (2%)	1 (2%)	0	0

Tabla 3.1.1-6. Motores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por campo

	Búsqueda 6 (Campo título)							
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**11º**
Google	7 (14,6%)	8 (16,7%)	14 (29,2%)	12 (25%)	6 (12,5%)	1 (2,1%)	0	0
MSN	10 (20,4%)	5 (10,2%)	22 (44.9%)	9 (18,4%)	2 (4,1%)	1 (2%)	0	0
Teoma (Ask)	Sin resultados							
WiseNut	0	5 (17,9%)	11 (39,3%)	6 (21,4%)	2 (7,1%)	1 (3,6%)	2 (7,1%)	1 (3,6%)
Yahoo	9 (18%)	7 (14%)	13 (26%)	14 (28%)	5 (10%)	2 (4%)	0	0

En la búsqueda por un término, el máximo nivel de profundidad lo registra Google con un recurso en un directorio de decimotercer nivel. Sin embargo, no presenta recursos de un nivel octavo que a excepción de WiseNut, el resto sí lo hacen. El máximo número de recursos recuperados por Google 15(30%), pertenece al tercer nivel igual que ocurre con MSN y WiseNut, sin embargo en Yahoo, el máximo número con 11(22%), pertenece al cuarto nivel y, frente al resto, destaca la frecuencia de recursos de octavo nivel 4(8%) y uno del nivel noveno. Teoma (Ask) es el motor que más recursos de primer nivel recupera 7(14%) y 13(26%), correspondiéndole el rastreo más superficial.

Sin embargo, en la segunda búsqueda, sólo Teoma alcanza con dos recursos, el séptimo nivel. En Google, MSN y Yahoo, los niveles en los que se aprecia mayor concentración son el tercero y el cuarto, aunque WiseNut recupera más del segundo y tercer nivel, lo que unido a los cuatro recursos de primer nivel, indica que este motor realiza un rastreo más superficial que el resto.

La búsqueda con operadores de existencia, ofrece los resultados más generalistas, pues las frecuencias más altas corresponden a los niveles más bajos, como en el caso de Google y MSN. Yahoo y WiseNut concentran la mayoría de resultados en los niveles segundo, tercero y cuarto. De nuevo Google alcanza el nivel más alto, con un recurso de octavo nivel, mientras que Teoma (Ask), WiseNut y Yahoo recuperan un recurso de séptimo nivel.

En la cuarta búsqueda, teniendo en cuenta que Teoma y WiseNut no ofrecen resultados, Yahoo recupera un recurso de nivel undécimo y tres del séptimo nivel, que es el segundo en importancia. MSN se muestra en esta búsqueda como el motor más generalista, ya que no sólo es el único motor, de los tres que ofrecen resultados que recupera re-

cursos de primer nivel 3(6%), sino que los de segundo y tercer nivel son superiores a los otros dos buscadores.

En la quinta búsqueda, en la que Teoma (Ask) no ofrece resultados, destaca la recuperación por parte de Google de un recurso de nivel undécimo, ofreciendo en resto un comportamiento similar, a excepción de Yahoo que recupera menos registros del primer y segundo nivel, destacando la frecuencia de recursos de tercero y cuarto nivel.

En la sexta búsqueda, corresponde a WiseNut la recuperación de un recurso de nivel undécimo. Google supera a MSN y WiseNut en recursos de quinto nivel, y Yahoo supera a todos en recursos de sexto y cuarto nivel. MSN y Yahoo son los motores que más recurso de primer nivel recuperan.

En general, podemos observar que Google, es el motor que recupera registros de mayor nivel en la mayoría de las búsquedas. Yahoo recupera más recursos de los niveles intermedios. WiseNut ofrece resultados más variables en las diferentes búsquedas y MSN recupera más recursos de los primeros niveles, por lo que es el motor que ofrece acceso a sitios web o recursos más generalistas.

Las siguientes tablas recogen los resultados aportados por los metabuscadores.

Tabla 3.1.1-7. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por término único

	Búsqueda 1 (Término único)								
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**	**9º**
Dogpile	4 (8%)	14 (%)	11 (22%)	13 (26%)	3 (6%)	1 (2%)	2 (4%)	2 (4%)	0
Excite	3 (6%)	6 (12%)	10 (20%)	17 (34%)	6 (12%)	2 (4%)	3 (6%)	3 (6%)	0
Ixquick	5 (15,6%)	6 (18,8%)	7 (21,9%)	4 (1,6%)	5 (15,6%)	2 (6,3%)	1 (3,1%)	2 (6,3%)	0
Profusion	3 (7,3%)	9 (22%)	6 (14,6%)	7 (17,1%)	5 (12,2%)	2 (4,9%)	5 (12,2%)	4 (9,8%)	0
Search	4 (8%)	12 (24%)	15 (30%)	7 (14%)	2 (4%)	2 (4%)	3 (6%)	4 (8%)	1 (2%)
Surfwax	2 (13,3%)	5 (33,3%)	5 (33,3%)	1 (6,7%)	0	1 (6,7%)	1 (6,7%)	0	0
Vivisimo	4 (8%)	12 (24%)	15 (30%)	11 (22%)	4 (8%)	1 (2%)	1 (2%)	2 (4%)	0

Tabla 3.1.1-8. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por varios términos

	Búsqueda 2 (Lenguaje natural)						
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**9º**
Dogpile	8 (18,6%)	3 (7%)	15 (34,9%)	7 (16,3%)	10 (23,3%)	0	0
Excite	2 (4,1%)	4 (8,2%)	18 (36,7%)	12 (24,5%)	12 (24,5%)	1 (2%)	0
Ixquick	0	7 (18,9%)	16 (43,2%)	8 (21,6%)	4 (10,2%)	2 (5,4%)	0
Profusion	1 (2,9%)	4 (11,4%)	13 (37,1%)	10 (28,6%)	5 (14,3%)	2 (5,7%)	0
Search	1 (2,1%)	9 (19,1%)	15 (31,9%)	14 (29,8%)	4 (8,5%)	4 (8,5%)	0
Surfwax	2 (11,1%)	10 (55,6%)	3 (16,7%)	3 (16,7%)	0	0	0
Vivisimo	0	14 (28%)	13 (26%)	18 (36%)	3 (6%)	1 (2%)	1 (2%)

Tabla 3.1.1-9. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda con operadores de existencia

	Búsqueda 3 (Operadores de existencia)						
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**
Dogpile	17 (41,5%)	3 (7,3%)	8 (19,5%)	10 (24,4%)	3 (7,3%)	0	0
Excite	9 (19,6%)	2 (4,3%)	8 (17,4%)	14 (30,4%)	6 (13%)	7 (15,2%)	0
Ixquick	8 (18,2%)	10 (22,7%)	12 (27,3%)	8 (18,2%)	6 (13,6%)	0	0
Profusion	Sin resultados						
Search	5 (10,6%)	6 (12,8%)	12 (25,5%)	11 (23,4%)	12 (25,5%)	1 (2,1%)	0
Surfwax	9 (26,5%)	8 (23,5%)	3 (8,8%)	9 (26,5%)	3 (8,8%)	1 (2,9%)	1 (2,9%)
Vivisimo	14 (29,2%)	4 (8,3%)	11 (22,9%)	11 (22,9%)	7 (14,6%)	1 (2,1%)	0

Tabla 3.1.1-10. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda booleana

	Búsqueda 4 (Operadores booleanos)							
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**11º**
Dogpile	Sin resultados							
Excite	1 (3%)	0	14 (42,4%)	11 (33,3%)	5 (15,2%)	2 (6,1%)	0	0
Ixquick	1 (5,3%)	0	3 (15,8%)	11 (57,9%)	3 (15,8%)	1 (5,3%)	0	0
Profusion	1 (2,3%)	4 (9,3%)	10 (23,3%)	19 (44,2%)	7 (16,3%)	1 (2,3%)	1 (2,3%)	0
Search	0	8 (16,3%)	17 (34%)	16 (32,7%)	4 (8,2%)	4 (8,2%)	0	0
Surfwax	Sin resultados							
Vivisimo	5 (10,6%)	11 (23,4%)	13 (27,7%)	12 (25,5%)	4 (8,5%)	0	1 (2,1%)	1 (2,1%)

Tabla 3.1.1-11. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por frase

	Búsqueda 5 (Frase)									
Nivel al que se encuentra el recurso	1º	2º	3º	4º	5º	6º	7º	8º	9º	11º
Dogpile	Sin resultados									
Excite	9 (18,4%)	8 (16,3%)	13 (26,5%)	7 (14,3%)	5 (10,2	0	3 (6,1%)	1 (2%)	2 (4,1%)	1 (2%)
Ixquick	9 (21,4%)	11 (26,2%)	13 (31%)	4 (9,5%)	2 (4,8%)	0	1 (2,4%)	1 (2,4%)	1 (2,4%)	0
Profusion	4 (10,3%)	8 (20,5%)	11 (28,2%)	9 (23,1%)	4 (10,3%)	0	1 (2,6%)	1 (2,6%)	1 (2,6%)	0
Search	5 (10,2%)	13 (26,5%)	15 (30,6%)	7 (14,3%)	6 (12,2%)	0	0	1 (2%)	1 (2%)	1 (2%)
Surfwax	Sin resultados									
Vivisimo	19 (38,8%)	9 (18,4%)	13 (26,5%)	6 (12,2%)	0	1 (2%)	0	0	1 (2%)	0

Tabla 3.1.1-12. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en la búsqueda por campo

	Búsqueda 6 (Campo título)							
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**
Dogpile	Sin resultados							
Excite	6 (12%)	7 (14%)	29 (40%)	10 (20%)	5 (10%)	1 (2%)	1 (2%)	0
Ixquick	6 (26,1%)	2 (8,7%)	10 (43,5%)	2 (8,7%)	2 (8,7%)	1 (4,3%)	0	0
Profusion	0	7 (28%)	7 (28%)	5 (20%)	3 (12%)	2 (8%)	1 (4%)	0
Search	6 (12,2%)	6 (12,2%)	14 (28,4%)	8 (16,3%)	5 (10,2%)	5 (10,2%)	2 (4,1%)	3 (6,1%)
Surfwax	Sin resultados							
Vivisimo	10 (20,8%)	6 (12,5%)	15 (31,3%)	6 (12,5%)	7 (14,6%)	3 (6,3%)	0	1 (2,1%)

Los metabuscadores, al obtener sus datos de los motores, concentran los recursos en los mismos niveles que éstos.

En la primera búsqueda destaca por la profundidad Search, que ofrece un recurso con un recurso del nivel noveno, cuatro del octavo y tres del séptimo. Profusion también ofrece resultados de los niveles octavo y séptimo. Dogpile y Vivisimo recuperan un mayor número de recursos de los primeros niveles, por lo que utilizan fuentes más genéricas para la selección de sus resultados.

En la segunda búsqueda, los resultados más superficiales corresponden a Dogpile, con ocho recursos de primer nivel y ninguno en los niveles máximos. Surfwax recupera

recursos de los niveles intermedios pero ninguno de los superiores, y Excite tiene un buen comportamiento con altas frecuencias a partir del tercer nivel. Search es el metabuscador que más recursos de sexto nivel recupera 4(8,5%). Vivisimo tiene altas frecuencias en los niveles más bajos, si bien, no recuperó ningún recurso de primer nivel. Es el único metabuscador que recupera un recurso de noveno nivel.

En la tercera búsqueda es Surfwax el que denota una mayor profundidad al recuperar un recurso de séptimo nivel. Search tiene un buen comportamiento en la recuperación de resultados a partir del tercer nivel, mientras que Dogpile y Vivisimo son los metabuscadores que más recursos de primer nivel ofrecen en esta búsqueda.

En la cuarta, de los que recuperan, es Vivisimo el metabuscador que más recursos de los primeros niveles ofrece, aunque a su vez es el único que ofrece un resultado del nivel undécimo.

En la búsqueda por frase, los metabuscadores que ofrecen resultados de mayor profundidad son Excite y Search. Vivisimo destaca por recuperar el mayor número de recursos de primer nivel.

En la última búsqueda, Search ofrece el mayor número de los recursos de octavo nivel 3(6,1%), seguido de Vivisimo 1(2,1%), que además es el metabuscador que mas recursos de los primeros niveles ofrece.

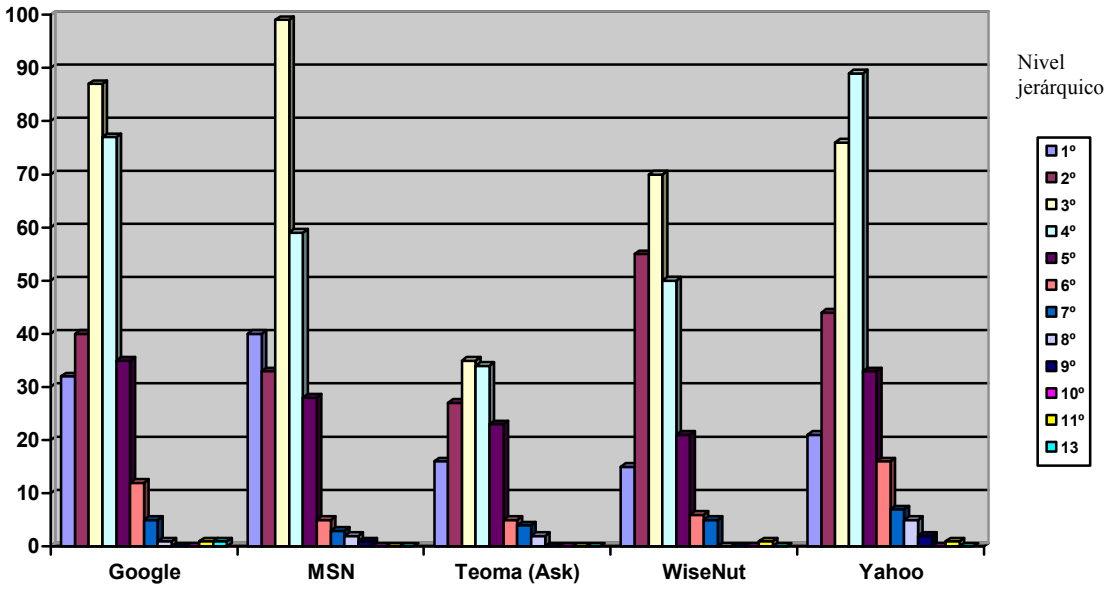
En los metabuscadores no se aprecia una tendencia que permita hablar de regularidad en los resultados, ya que como hemos visto en Surfwax, no ofrece recursos de los máximos niveles en la segunda búsqueda y sí lo hace en la tercera. Search es la excepción ya que ofrece buenos resultados en la primera, segunda, quinta y sexta búsquedas. Vivisimo se caracteriza principalmente por recuperar mayor número de recursos genéricos.

Análisis global

Tabla 3.1.1-13. Buscadores. Recursos recuperados en los distintos niveles del directorio en las seis búsquedas

	Total de las seis búsquedas											
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**	**9º**	**10º**	**11º**	**13º**
Google	32	40	87	77	35	12	5	1	0	0	1	1
MSN	40	33	99	59	28	5	3	2	1	0	0	0
Teoma (Ask)	16	27	35	34	23	5	4	2	0	0	0	0
WiseNut	15	55	70	50	21	6	5	0	0	0	1	0
Yahoo	21	44	76	89	33	16	7	5	2	0	1	0

3.1.1-1. Motores. Recursos recuperados en los distintos niveles del directorio en las seis búsquedas



En la tabla podemos apreciar de forma conjunta, lo que ya hemos señalado al valorar los resultados individuales de cada pregunta, como es el hecho de que Google es el motor que recupera más registros de los niveles superiores, lo que indica que es el buscador que recupera registros de mayor profundidad dentro de los sitios web. Le sigue Yahoo, que sin alcanzar el nivel de Google, también recupera recursos correspondientes a los mayores niveles. No obstante, y esto se aprecia claramente en el gráfico, podemos observar que el funcionamiento de ambos es distinto pues Yahoo, recupera más recursos de cuarto nivel y menos del nivel genérico mientras que en Google, el mayor número de recursos corresponde al tercer nivel y el nivel genérico es superior. Esto permite afirmar que aunque Google y Yahoo realizan una indización de cierta profundidad respecto a los demás, existen pequeñas diferencias entre ellos, ya que el primero recupera más recursos de los niveles máximos pero también del más genérico mientras que Yahoo recupera menos del nivel genérico.

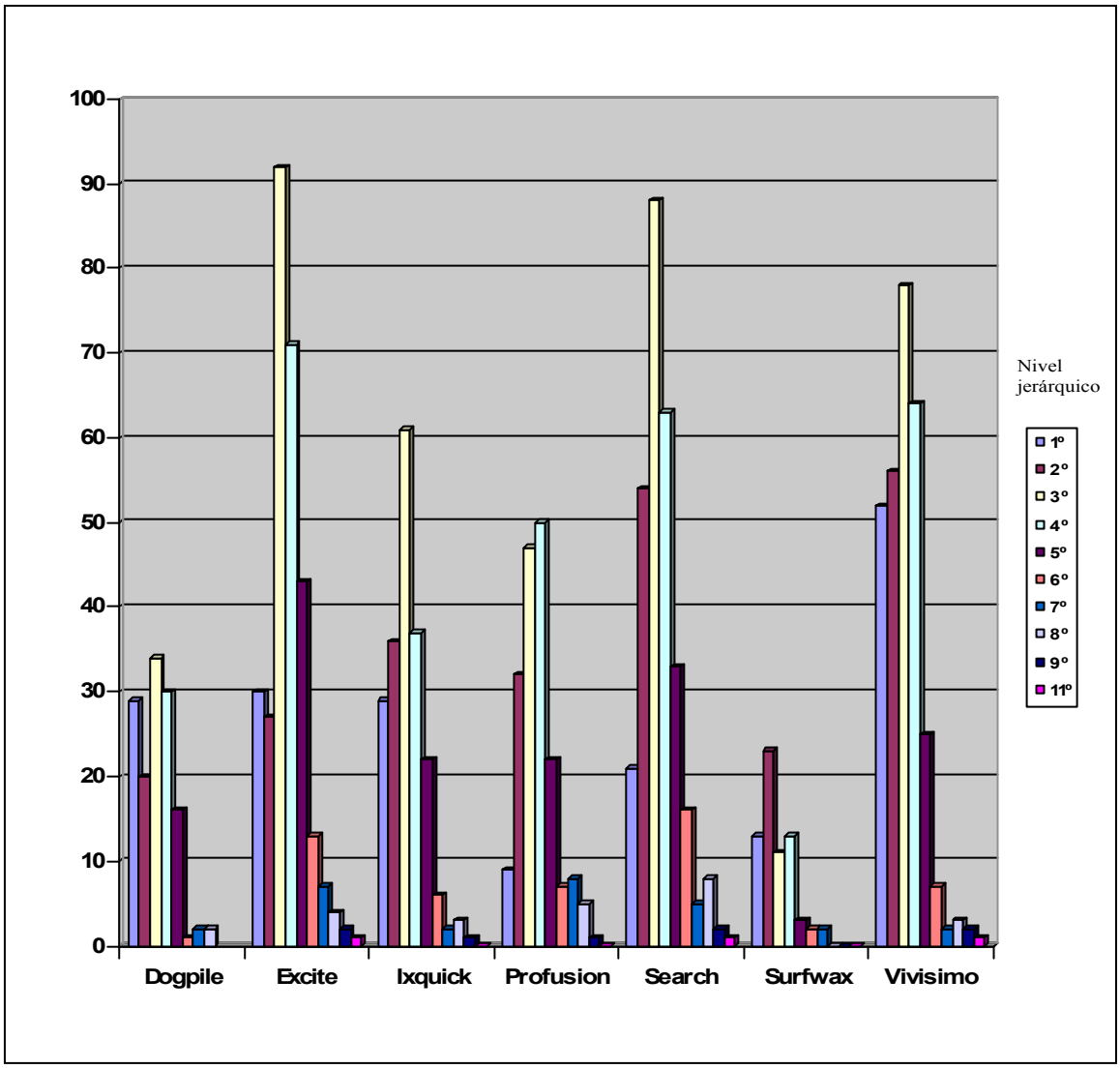
Por otro lado, MSN tienen una tendencia a recuperar más registros de carácter genérico que los anteriores.

WiseNut tiene un comportamiento más irregular, pues además de recuperar un recurso de undécimo nivel, se caracteriza por ser el motor que más registros de segundo nivel recupera.

Tabla 3.1.1-14. Metabuscadores. Recursos recuperados en los distintos niveles del directorio en las seis búsquedas

	Total de las seis búsquedas									
Nivel al que se encuentra el recurso	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**7º**	**8º**	**9º**	**11º**
Dogpile	29	20	34	30	16	1	2	2		
Excite	30	27	92	71	43	13	7	4	2	1
Ixquick	29	36	61	37	22	6	2	3	1	0
Profusion	9	32	47	50	22	7	8	5	1	0
Search	21	54	88	63	33	16	5	8	2	1
Surfwax	13	23	11	13	3	2	2	0	0	0
Vivisimo	52	56	78	64	25	7	2	3	2	1

Gráfico 3.1.1-2 Metabuscadores. Recursos recuperados en los distintos niveles del directorio en las seis búsquedas



La presente gráfica permite observar cómo Vivisimo proporciona mayor número de recursos de primer y segundo nivel, facilitando el acceso a recursos de carácter más genérico. Excite y Search tienen una mejor recuperación de recursos de tercer a sexto nivel. Los recursos de mayor profundidad los facilitan Excite, Search y también Vivisimo.

3.2. Aspectos relacionados con el índice de los buscadores

3.2.1. Duplicados

La existencia de duplicados es uno de los criterios más utilizados en la evaluación de los buscadores de la Web, ya que es un importante elemento para valorar el correcto funcionamiento de estas herramientas. Un elevado número de duplicados en los resultados, es indicativo de un deficiente funcionamiento de los programas que se ocupan de la indización de las páginas web ya que han de comprobar si ya se encuentran indizadas previamente.

Las siguientes tablas recogen los resultados relativos a cada una de las seis búsquedas y la suma total de duplicados recuperados por cada motor de búsqueda. Para la valoración de los duplicados nos hemos basado en las URL de los recursos, señalando como duplicados las páginas que mantienen una misma URL.

Tabla 3.2.1-1 Motores. Nº de duplicados por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total
Google	1	8	0	0	0	0	**9**
MSN	0	0	0	1	0	0	**1**
Teoma (Ask)	3	1	1	Sin resultados			**5**
WiseNut	1	0	0	Sin resultados	0	0	**1**
Yahoo	1	0	0	0	0	0	**1**

Los buscadores que recuperan menor número de registros duplicados, y por tanto que mejor funcionamiento demuestran son MSN y Yahoo con un único resultado en las seis búsquedas. El mayor número de duplicados corresponde a Google con un total de nueve recursos en las seis búsquedas, destacando la segunda búsqueda, en la que se le contabilizan ocho duplicados.

Teoma (Ask) recupera duplicados en todas las búsquedas en las que ofrece resultados y WiseNut, que no funcionó en la búsqueda con operadores booleanos, sólo ofrece un duplicado en la búsqueda por término único.

Estos resultados muestran por un lado, unos buenos resultados ya que el número de duplicados no es excesivo, y por otro, una mejora en los resultados de Yahoo respecto a

trabajos anteriores[1] y no tanto en Google, aunque en este caso, el alto número de duplicados corresponden a una sola búsqueda, por lo que creemos que se debe más a un fallo puntual en la recuperación que a una característica propia de este buscador, que en el resto de búsquedas apenas recupera duplicados. En cualquier caso, debemos hacer constar en este sentido, un mal funcionamiento de este motor de búsqueda.

Tabla 3.2.1-2. Metabuscadores. Nº de duplicados por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total
Dogpile	0	0	5		Sin resultados		**5**
Excite	0	0	3	1	0	1	**5**
Ixquick	0	0	0	0	1	0	**1**
Profusion	9	7	Sin resultados	1	4	0	**21**
Search	3	0	1	5	5	1	**15**
Surfwax	0	0	0		Sin resultados		**0**
Vivisimo	2	1	0	0	0	0	**3**

En cuanto a los metabuscadores, la herramienta que mayor número de recursos duplicados recupera en las seis búsquedas es Profusion, con un total de 21 recursos duplicados. Le sigue Search con 15, y Dogpile y Excite con 5. El menor número de duplicados corresponde a Vivisimo con tan sólo tres, ya que aunque Surfwax no parece recuperar duplicados, en las tres últimas búsquedas no funcionó.

El mayor número de duplicados en una búsqueda lo alcanzó Profusion, al registrar 9 casos en la primera búsqueda. Además; en la segunda obtuvo 7, y en la tercera no funcionó.

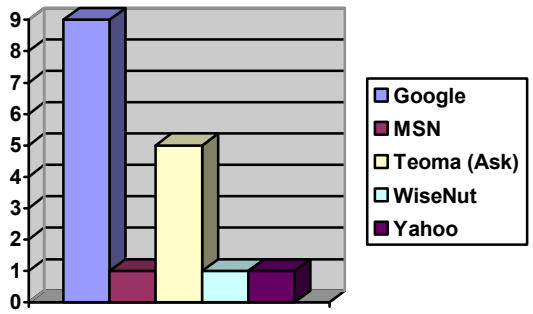
En Search también es frecuente la aparición de duplicados entre sus resultados.

El aumento de duplicados en los metabuscadores denota un mal funcionamiento de los programas que los soportan, ya que conscientes de ofrecer altos índices de duplicados, implementan este tipo de programas que les permiten eliminarlos, aunque como vemos no siempre lo consiguen, especialmente en el caso de Profusión y Search. Sin embargo Dogpile y Excite muestran resultados más relacionados con los motores de búsqueda.

[1] Véase Salvador Oliván, J.A. y Vidal Bordes F.J. (2000)

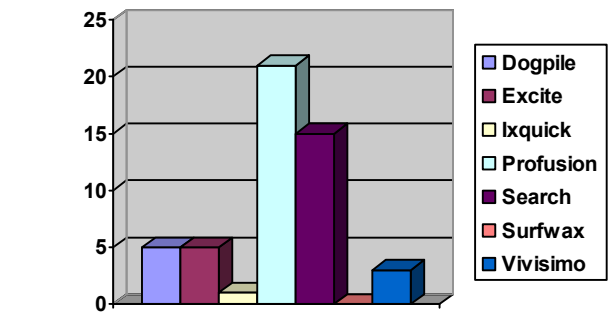
Análisis global

Gráfico 3.2.1-1. Motores. Total Duplicados



Como muestra la gráfica anterior, hemos de destacar el buen comportamiento, en cuanto a la menor recuperación de recursos duplicados, de los buscadores MSN y Yahoo frente a Google. WiseNut, a falta de los resultados de la búsqueda en la que no recuperó, también ofreció un buen comportamiento en este sentido.

Gráfico 3.2.1-2. Metabuscadores. Total Duplicados



Respecto a los metabuscadores que recuperan en todas las búsquedas, el mejor comportamiento se observa en Ixquick, seguido de Vivisimo y Excite. Por otro lado, Search y en mayor medida Profusion ofrecen, en el total de las seis búsquedas, un importante número de duplicados.

En comparación con los motores, las cifras de duplicados son sensiblemente superiores en los metabuscadores, por lo que deberían revisar los mecanismos de detección de

duplicados, especialmente en los casos que ofrecen un número más elevado (Search y Profusion).

3.2.2. Enlaces inactivos

Tabla 3.2.2-1. Motores. Nº de recursos inactivos por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total inactivos
Google	0	1 (2%)	2 (4%)	5 (10%)	1 (2%)	2 (4%)	**11**
MSN	6 (12%)	3 (6%)	1 (2%)	0	1 (2%)	1 (2%)	**12**
Teoma (Ask)	1 (2%)	3 (6%)	1 (2%)		Sin resultados		**5**
WiseNut	3 (6%)	3 (6%)	4 (8%)	Sin resultados	1 (2%)	3 (10%)	**14**
Yahoo	3 (6%)	4 (8%)	7 (14%)	0	0	0	**14**
Total por Búsqueda	**13**	**14**	**15**	**5**	**3**	**6**	**56**

Los que presentan mayor número de recursos no operativos tanto en la primera como en la segunda búsqueda son MSN, Yahoo y WiseNut. En la segunda, hay que añadir Teoma (Ask) con 3 (6%). En la tercera destaca el incremento de Yahoo con 7 (14%) frente a MSN y Teoma con 1 (2%). En la cuarta Google recupera cinco inactivos (10%), mientras que en Yahoo y MSN todos los enlaces fueron activos.

En la quinta búsqueda, Google, MSN y WiseNut recuperaron un recurso inactivo y en Yahoo ninguno. Finalmente en la sexta búsqueda, WiseNut fue el motor con más recursos inactivos con 3(10%), seguido de Google con 2(4%) y MSN con 1(2%). En Yahoo funcionaron, de nuevo, todos los enlaces recuperados.

A pesar de los resultados de estas últimas búsquedas, los motores que acusan mayor número de recursos inactivos son Yahoo y WiseNut con 14, seguidos por MSN con 12 y Google con 11. Como vemos las diferencias son poco significativas entre los motores, y a efectos de la recuperación, aunque las cantidades no podemos considerarlas problemáticas, dado el gran número de recursos que se analizan, si que son indicativos de la falta de actualización de los índices de los buscadores.

Estos datos guardan relación con los ofrecidos en otros estudios (Bar-Ilan, 1998), pues señalan un porcentaje del 2% de recursos a los que no se puede acceder.

Tabla 3.2.2-2. Metabuscadores. Nº de recursos inactivos por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Operadores booleanos)	Búsqueda 5 (Frase)	Búsqueda 6 (Campo título)	Total inactivos
Dogpile	0	6 (14%)	10 (20%)	Sin resultados			**16**
Excite	1 (2%)	1 (2%)	6 (12%)	1 (2,9%)	1 (2%)	1 (2%)	**11**
Ixquick	1 (3,1%)	2 (5,1%)	6 (12%)	3 (14,3%)	2 (4,5%)	0	**14**
Profusion	1 (2,4%)	0	Sin resultados	4 (8,7%)	0	0	**5**
Search	1 (2%)	3 (6%)	3 (6%)	3 (6%)	2 (4%)	2 (4%)	**14**
Surfwax	0	1 (5,3%)	2 (5,6%)	Sin resultados			**3**
Vivisimo	1 (2%)	2 (4%)	4 (8%)	3 (6%)	2 (4,1%)	2 (4,1%)	**14**
Total por búsqueda	**5**	**15**	**31**	**14**	**7**	**5**	**77**

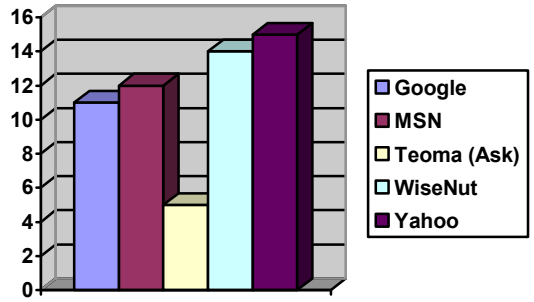
Los metabuscadores con mayor índice de operatividad en la primera búsqueda son Dogpile y Surfwax, aunque el resto de metabuscadores sólo recupera un enlace inactivo. En la segunda, Profusion mejora el resultado anterior, mientras que Dogpile empeora considerablemente. En la búsqueda con operadores de existencia, son Surfwax con 2 (5,6%), y Search con 3 (6%) los que menor número de inactivos recuperan. El peor comportamiento corresponde a Dogpile con 10 (20%), aunque los inactivos que presentan Excite e Ixquick también son elevados, 6 (12%). En la cuarta búsqueda no se dan resultados significativos correspondiendo a Excite el menor número de recursos no activos 1 (2,9%) y a Profusion el mayor 4 (8,7%). En porcentaje, el máximo corresponde a Ixquick (14,3%).

La quinta búsqueda no ofrece resultados significativos, si bien destaca Profusión, al no recuperar registros con enlaces inactivos, aspecto que se repite en la última búsqueda junto a Ixquick.

Por tanto, podemos afirmar que, en este caso, los peores resultados corresponden a Dogpile que, sin haber funcionado en las tres últimas búsquedas, presenta el mayor número de recursos inactivos (16). Ixquick, Search y Vivisimo coinciden en el total de recursos inactivos (14). El mejor comportamiento, con sólo 5 inactivos corresponde a Profusion, si bien hay que tener en cuenta que no recuperó en la tercera búsqueda.

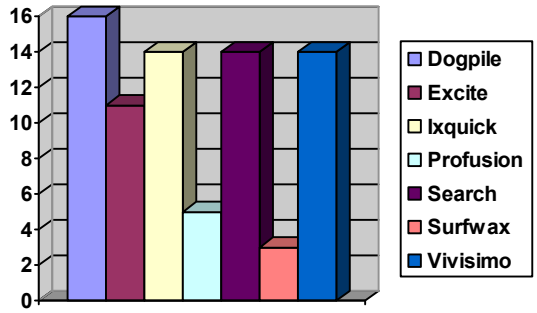
Análisis global

Gráfico 3.2.2-1. Motores. Total enlaces inactivos



Como podemos observar es poca la diferencia entre los buscadores en la recuperación de recursos inactivos. A excepción de Teoma (Ask), cuyos datos no se pueden tener en cuenta al no contabilizar tres de las búsquedas, Google es el motor que recupera menos enlaces inactivos en las seis búsquedas y por lo tanto podemos decir que actualiza de forma más frecuente los índices, al contrario que Yahoo, que recupera un mayor número de inactivos.

Gráfico 3.2.2-2. Metabuscadores. Total enlaces inactivos



Dejando a un lado los metabuscadores que no recuperaron en todas las búsquedas (Dogpile, Profusion y Surfwax), podemos observar que las diferencias también aquí son pequeñas, aunque Excite es el metabuscador que menos registros inactivos proporciona, por lo que sus fuentes están un poco más actualizas que el resto.

En este caso, las diferencias entre motores de búsqueda y metabuscadores apenas son apreciables, existiendo una gran relación entre unos y otros.

3.3. Aspectos relacionados con la base de datos

En el análisis de las bases de datos, nos ocupamos de analizar la actualización de la información recuperada, para lo que tenemos en cuenta la fecha de creación del recurso o página; el carácter de la información, es decir si tiene fines comerciales, de divulgación o si se trata de una información de carácter científico. También se analiza el tipo de fichero que soporta la información, bien sea una página web, un documento PDF, Word o PowerPoint, etcétera, y finalmente nos ocupamos del tipo de información que ofrecen, mediante una amplia tipología documental.

3.3.1. Análisis de las características de la información recuperada

3.3.1.1. Actualización de la información proporcionada

Las siguientes tablas recogen los resultados relativos a los registros recuperados por los buscadores en las diferentes búsquedas, en las que figura una fecha Las cifras se refieren al número de recursos con fechas de creación o copyright más reciente.

Tabla 3.3.1-1. Motores. Nº de recursos más recientes por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Google	6	6	13	6	9	13	**53**
MSN	13	9	14	10	11	14	**71**
Teoma (Ask)	11	7	5	Sin resultados			**23**
WiseNut	11	10	14	Sin resultados	11	1	**47**
Yahoo	7	9	12	5	7	9	**49**

Teniendo en cuenta las fechas reflejadas en los recursos, podemos observar que en la primera búsqueda, MSN con trece, es el buscador que ofrece un mayor número de recursos actuales. Le sigue Teoma y WiseNut. El último lugar corresponde a Google. En la segunda búsqueda es WiseNut el motor que ofrece los recursos más actualizados seguidos de MSN y Yahoo con nueve, quedando en los últimos lugares Google y Teoma.

En la tercera búsqueda MSN y WiseNut ocupan los primeros lugares, con muy poca diferencia sobre Google y Yahoo. Teoma, es en esta ocasión, el motor con menos recursos actualizados.

En la búsqueda booleana, MSN con diez páginas es, de los tres motores que obtuvieron resultados, el que recupera el mayor número de páginas actuales, seguido por Google y Yahoo con seis y cinco páginas, respectivamente.

En la quinta búsqueda no hay grandes diferencias entre los buscadores, aunque el mayor número de páginas actuales corresponde a MSN y WiseNut.

En la búsqueda por campo, MSN y Google obtienen el mayor número de páginas actuales, y es WiseNut el que menor número obtiene.

El hecho de que en Google, en todas las búsquedas sean inferiores a las de MSN pueden ser reflejo del algoritmo que utiliza en la ordenación, y que como hemos visto, basa su cálculo en la valoración de los sitios web que recogen un determinado enlace. En este sentido, hasta que el recurso no es suficientemente conocido y añadido como enlace es un determinado número de páginas y sitios Web, no es ofrecido en los primeros puestos por este buscador. Ello influye, como vemos en el carácter algo más anticuado de los recursos, aspecto que, en cuanto a su mayor perdurabilidad, también puede ser interpretado como una garantía de su importancia.

Si observamos la columna de totales, los resultados de Google, WiseNut y Yahoo son similares, destacando MSN, que es el motor que recupera un mayor número de recursos de actualidad.

Tabla 3.3.1-2. Metabuscadores. Nº de recursos más recientes por búsqueda

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Dogpile	8	17	10	Sin resultados			**35**
Excite	5	15	13	7	16	5	**61**
Ixquick	6	10	14	5	13	4	**52**
Profusion	8	9	Sin resultados	10	10	3	**40**
Search	13	13	9	11	8	7	**61**
SurfWax	4	11	18	Sin resultados			**33**
Vivisimo	8	11	9	8	6	8	**50**

En cuanto a los metabuscadores, Search, en la primera búsqueda recupera el mayor número de páginas actualizadas (13) seguido de Dogpile, Profusion y Vivisimo con 8.

Ixquick, Excite y Surfwax, por este orden son, por este orden, los que menor número de recursos actualizados recuperan.

En la segunda búsqueda, los más actualizados serían Dogpile y Excite con diecisiete y quince páginas, seguidos de Search con trece. Profusion e Ixquick recuperan el menor número de páginas actualizadas.

En la tercera búsqueda es Surfwax el metabuscador con más recursos actuales (18), seguido de Ixquick y Excite con catorce y trece páginas respectivamente. Dogpile, Search y Vivisimo son los que proporcionan un menor número.

En la búsqueda booleana, corresponde a Search y Profusion la recuperación de un mayor número de recursos posteriores al 2000. Ixquick es el que menos recursos actuales recuperó en esta búsqueda.

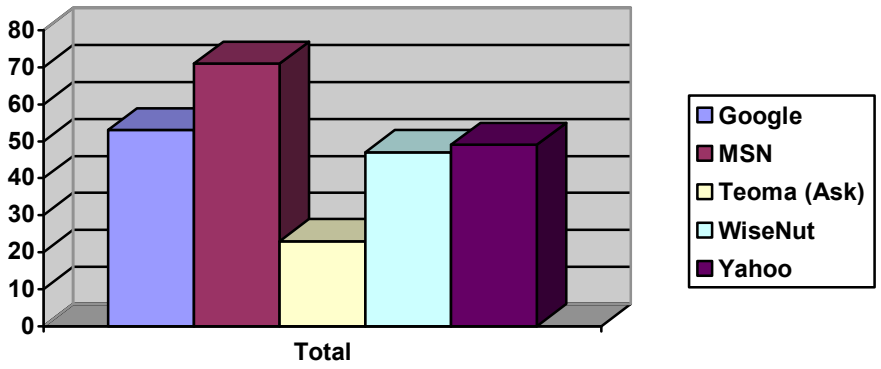
En la búsqueda por frase es Excite el más actualizado. Le siguen Ixquick y Profusion. Vivisimo y Search obtienen el menor número de recursos actualizados.

Finalmente, en la sexta búsqueda son Vivisimo y Search los que contienen más recursos actuales, frente a Profusión que obtiene el menor número.

En base a lo anteriormente expuesto podemos afirmar que entre los metabuscadores no hay ninguno que destaque de forma regular, aunque teniendo en cuenta las cifras totales, Excite y Search recuperan el mayor número de recursos actuales.

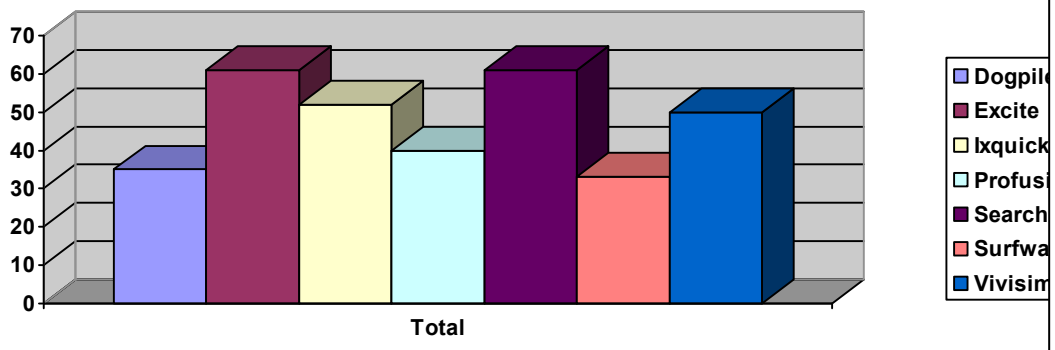
Análisis global

Gráfico 3.3-1 Motores. Total registros actuales



Así pues, teniendo en cuenta la fecha que aparece en los registros, MSN es el buscador que recupera registros más actualizados, seguido por Google, Yahoo y WiseNut, con resultados muy similares.

Gráfico 3.3-2 Metabuscadores. Total registros actuales



Entre los metabuscadores, son Excite y Search los que facilitan mayor numero de recursos actualizados, seguidos por Ixquick y Vivisimo.

Una vez más, los resultados están muy relacionados con los ofrecidos por los motores de búsqueda.

3.3.1.2. Carácter de la información

Las tablas que aparecen a continuación muestran los resultados de analizar los recursos recuperados por cada buscador, clasificándolos teniendo en cuenta la utilidad de su contenido y el objetivo que persiguen.

Tabla 3.3.1-3. Motores. Carácter de la información

	Publicidad	Investigación	Divulgación	Institucional	Comercial	Otro	Total
Google	3	252	1	4	11	16	**287**
MSN	1	232	2	1	35	17	**288**
Teoma (Ask)	3	113	4	0	17	3	**140**
WiseNut	3	147	5	1	35	29	**220**
Yahoo	3	247	0	0	17	19	**286**
						Total	**1221**

De los 1280 recursos recuperados por los motores de búsqueda, se analizan 1221 una vez descontadas las páginas erróneas, páginas en blanco o imposibles de analizar, en este sentido, generalmente por aparecer en lenguas orientales.

El primer dato a destacar es el predominio de los relacionados con la investigación. Google con 252, es el motor que más resultados de este tipo recuperó, seguido muy de cerca por Yahoo con 247 y a mayor distancia MSN con 232. Peor fue la recuperación de este tipo de registros llevada a cabo por WiseNut y Teoma, que se ven afectados además por los problemas de recuperación que presentan en la cuarta, quinta y sexta búsqueda.

Los recursos de otro tipo apenas son reseñables, aunque se puede observar que los motores que recuperan mayor número de recursos comerciales son MSN y WiseNut.

Tabla 3.3.1-4. Metabuscadores. Carácter de la información

	Publicidad	Investigación	Divulgación	Institucional	Comercial	Otro	Total
Dogpile	1	94	1	0	19	10	**125**
Excite	0	189	3	0	41	39	**272**
Ixquick	2	149	4	2	24	16	**197**
Profusion	0	152	3	0	6	19	**180**
Search	0	163	2	2	21	20	**208**
Surfwax	0	29	2	1	13	22	**67**
Vivisimo	4	136	3	2	23	31	**199**
						Total	**1248**

Excite es el metabuscador que más recursos de investigación recupera, seguido de Search, Ixquick y Vivisimo, mientras que Surfwax, teniendo en cuenta que sólo recuperó en tres búsquedas, es el que menos. recursos de este tipo recupera, ofreciendo unos resultados muy bajos, al igual que ocurre con Dogpile.

También es significativa la recuperación de información de tipo comercial, en este tipo de búsquedas especializadas tanto en Excite como en Vivisimo, mientras que en Profusión apenas se dan recursos de este tipo.

Análisis global

Los siguientes gráficos ilustran de una forma clara los datos de las tablas precedentes al tiempo que nos permiten comparar la recuperación de motores y metabuscadores. La línea blanca representa el total de recursos que se analizan en las seis búsquedas. Gracias a ella podemos observar el lugar que ocupa la documentación recuperada de un determinado carácter, respecto del total de los recursos de cada motor. En este sentido, vemos que la cifra de recursos de investigación recuperada por Google se aproxima bastante al total, mientras que en MSN la distancia es mayor, y aún lo es más en WiseNut.

Por otro lado, la visualización de los dos gráficos muestra claramente que el tipo de información predominante en ambos tipos de herramientas es el relacionado con el campo de la investigación. Le sigue a cierta distancia la información comercial que, como hemos visto anteriormente, en Google y Yahoo es menor que en el resto, siendo MSN y WiseNut los motores que mayor número de recursos de este tipo recuperan.

Gráfico. 3.3-3. Motores. Carácter de la información recuperada por las herramientas de búsqueda

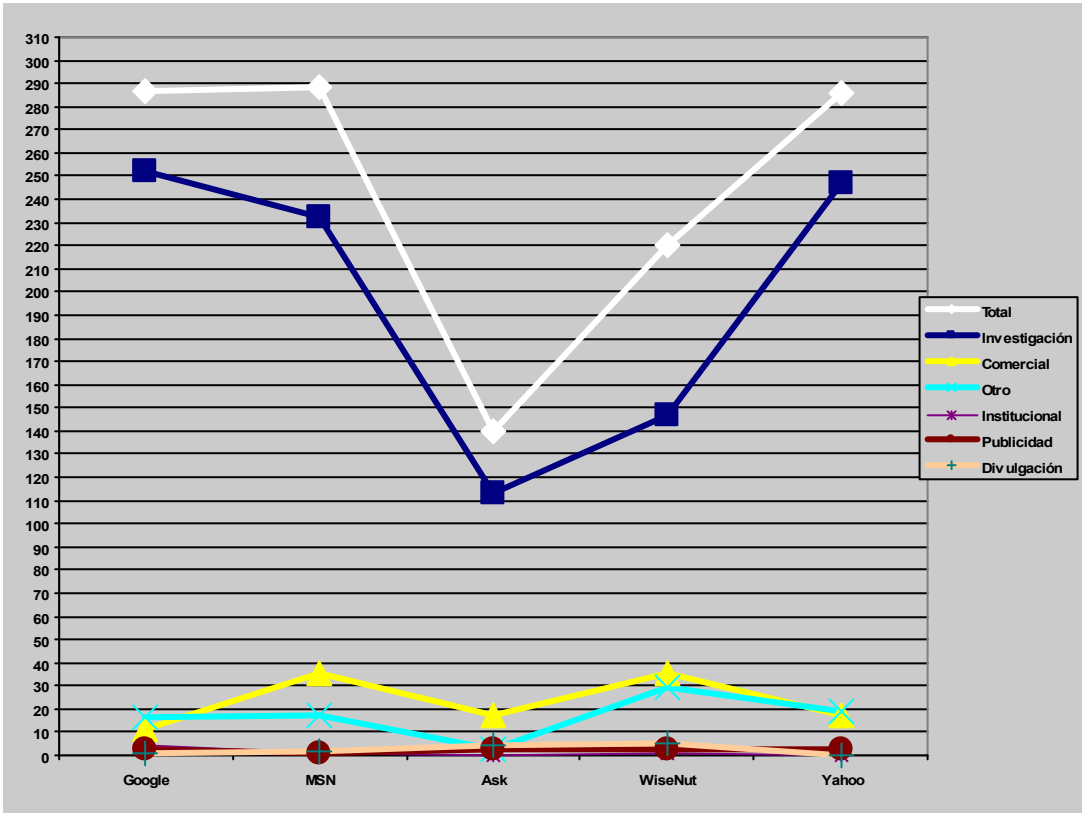
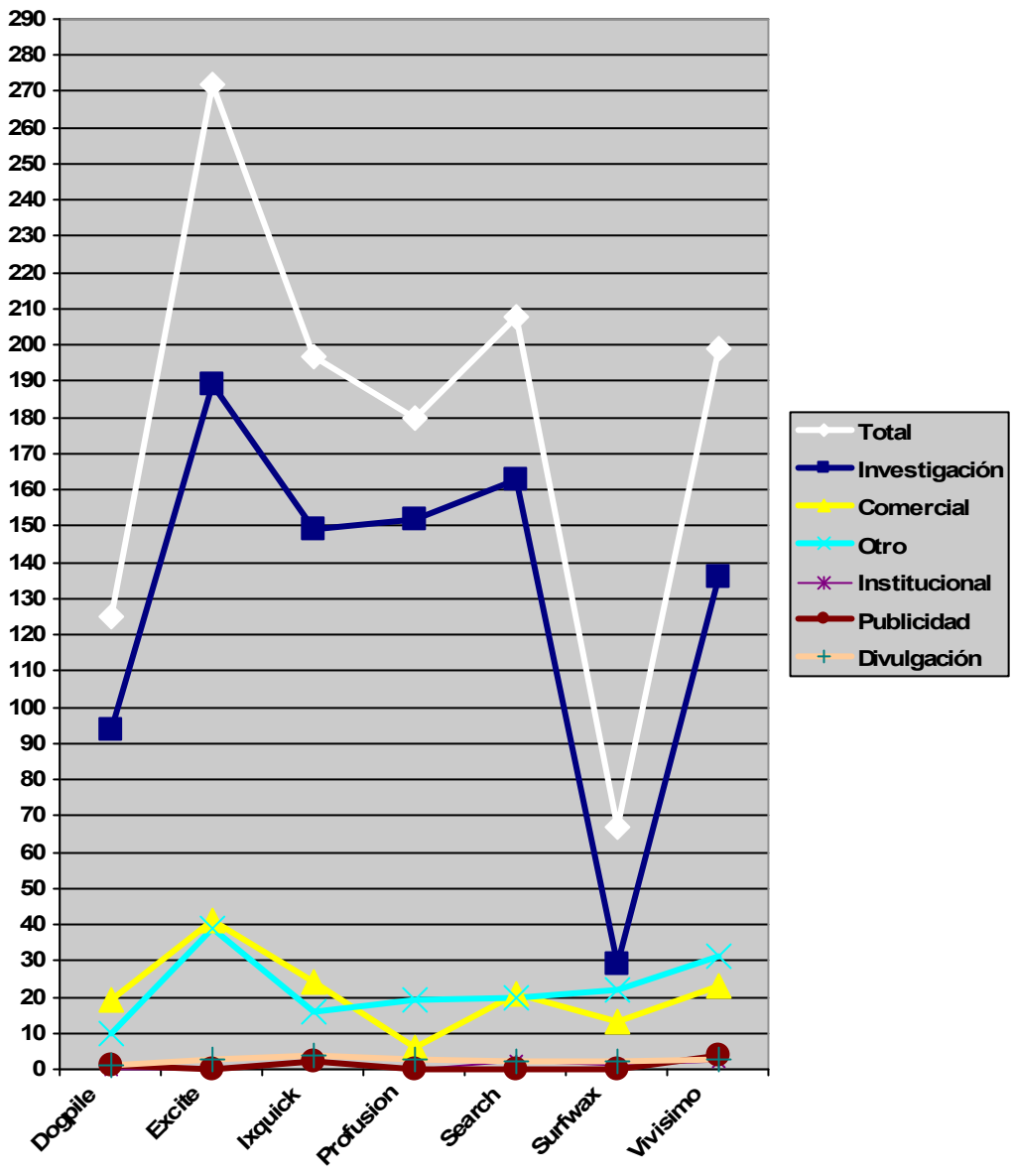


Gráfico 3.3-4. Metabuscadores. Carácter de la información recuperada por las herramientas de búsqueda



Excite es el metabuscador que más recursos de investigación recupera, seguido por este orden de Search, Profusión, Ixquick y Vivisimo, siendo Surfwax el que peores resultados ofrece. Por otro lado, son también Excite, Vivisimo y Search los que más recursos de carácter comercial recuperan. Profusión recupera menos recursos de este tipo.

En cuanto a la comparación entre motores y metabuscadores, destacan los primeros, a excepción de Teoma y WiseNut, por ofrecer un mayor número de recursos de investigación.

3.3.1.3. Tipo de fichero

Analizamos a continuación la tipología documental de los recursos recuperados ya que constituye un elemento a tener en cuenta para valorar el contenido de la base de datos, ya que artículos y otros documentos de investigación utilizan generalmente, además del propio HTML, otros formatos como PDF, PowerPoint, etcétera.

Tabla 3.3.1-5. Motores. Formatos de ficheros

	HTML/PHP/ SCRIPT	PDF/PS	Power Point en PDF	Power Point	RTF	Excel	Word	Otro	Varios formatos	Total
Google	260	21	1	3	0	0	3	0	0	**288**
MSN	282	6	0	0	0	0	1	0	0	**289**
Teoma (Ask)	139	6	0	0	0	0	0	0	0	**145**
WiseNut	217	2	0	0	0	0	0	0	0	**219**
Yahoo	267	13	2	2	1	0	1	2	0	**288**

El tipo de fichero más abundante en la Web es la página HTML y por ello también, es el tipo de fichero más común en la recuperación de búsquedas especializadas. En este sentido, es MSN el buscador que más páginas de este tipo recupera. En segundo lugar está Yahoo, con una cifra similar a Google. Sin embargo, estos dos buscadores, especialmente Google, recuperan un mayor número de recursos en formato PDF, que con frecuencia, se suele utilizar en la elaboración de artículos de carácter científico o en documentos de cierta importancia para evitar su modificación. Esto, unido a la recuperación de algunos documentos en formato PowerPoint o Word, hace de Google el buscador con mayor versatilidad en la recuperación, al facilitar una mayor variedad en la tipología documental, lo que indica la existencia de un software que facilita la indización de diferentes tipos de recursos.

Tabla 3.3.1-6. Metabuscadores. Formatos de ficheros

	HTML/PHP/SCRIPT	PDF/PS	Power Point en PDF	Power Point	RTF	Excel	Word	Otro	Varios formatos	Total
Dogpile	117	4	1	0	0	0	1	2	0	**125**
Excite	251	20	0	0	0	0	1	1	0	**273**
Ixquick	185	8	0	1	0	1	1	0	0	**196**
Profusion	175	2	1	2	0	1	1	0	0	**182**
Search	263	18	1	0	0	0	1	2	0	**285**
Surfwax	67	0	0	0	0	0	0	0	0	**67**
Vivisimo	274	13	0	0	0	0	0	1	1	**289**

Entre los metabuscadores, es Vivisimo con 274 el que mayor número de páginas en lenguaje HTML recupera. Le siguen Search con 263 y Excite con 251. Sin embargo es este último el que más recursos en formato PDF recupera, seguido de Search y Vivisimo, con un comportamiento en este sentido similar al de los mejores motores de búsqueda.

Análisis global

Google es el buscador que recupera mayor número de páginas en formatos propios de información especializada como son PowerPoint, que se utiliza en presentaciones relacionadas con investigación sobre los temas de búsqueda, y documentos en PDF, que contienen artículos de investigación. Entre los metabuscadores destacan en la recuperación de este tipo de documentos Excite y Search.

3.3.1.4. Tipología documental

Exponemos a continuación los resultados desprendidos del análisis sobre la tipología documental de los resultados aportados por los buscadores en cada una de las búsquedas. Este aspecto puede ayudarnos a conocer qué herramientas recuperan en este tipo de búsquedas recursos de carácter más especializado, pudiendo comparar unos buscadores y otros.

Adjuntamos a las tablas gráficos para visualizar dichos resultados, así como diferenciar los de unos motores y otros.

3.3.1.4.1. Análisis individualizado de las búsquedas

3.3.1.4.1.1. Búsqueda de un término

Tabla 3.3.1-7. Motores. Búsqueda 1. Tipología documental

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Página html común	9 (18%)	10 (20,8%)	16 (32,7%)	14 (30,4%)	8 (17%)
Página html en blanco		2 (4,2%)	5 (10,2%)		
Página html en lenguas orientales		3 (6,3%)			2 (4,3%)
Imagen					
Base de datos a texto completo libre					
Base datos acceso restringido	3 (6%)				
Base datos acceso a registros bibliogr.	1 (2%)		2 (4,1%)		1 (2,1%)
Biblioteca Digital	5 (10%)		9 (18,4%)		3 (6,4%)
Repositorio					
Directorio		1 (2,1%)		2 (4,3%)	1 (2,1%)
Buscador					3 (6,4%)
Agente de búsqueda			3 (6,1%)		
Normas					
Lista de correo					
Revista electrónica					
E-libro					
Presentación	6 (12%)	1 (2,1%)			4 (8,5%)
Bibliografía		3 (6,3%)	3 (6,1%)	5 (10,9%)	3 (6,4%)
Lista de recursos web	1 (2%)	1 (2,1%)	5 (10,2%)	9 (19,6%)	3 (6,4%)
Artículo/Inf. especializada	14 (28%)	2 (4,2%)	2 (4,1%)	5 (10,9%)	7 (2,9%)
Artículo de rev. Electrónica					
Congreso/Trabajo congreso		2 (4,2%)		1 (2,2%)	
Monografía					
Capítulo de mon.	2 (4%)	6 (12,5%)		1 (2,2%)	4 (8,5%)
Art. de Enciclopedia	1 (2%)		2 (4,1%)	1 (2,2%)	
Entrevista	3 (6%)	1 (2,1%)		2 (4,3%)	3 (6,4%)
Diccionario	5 (10%)	3 (6,3%)		1 (2,2%)	
Noticias		2 (4,2%)		1 (2,2%)	
Blog o pág. personal		8 (16,7%)	2 (4,1%)		3 (6,4%)
Blog común especializado				1 (2,2%)	
Página registro					
Lista de correo				1 (2,2%)	
Discurso					
Proyecto					

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Curso o inf. de curso					
Resumen					
Repositorio					
FAQ					
Normas					
Examen		3 (6,3%)			1 (2,1%)
Registro					1 (2,1%)
Banco de datos					
Repositorio					

Una vez eliminados los registros no válidos, respecto a la tipología documental, podemos observar que existe un tipo común, que recuperan prácticamente la totalidad de buscadores y metabuscadores, y que corresponde a las Páginas HTML de carácter general, seguidas por otras con características específicas en la presentación de la información, por ejemplo listas de recursos web, artículos especializados y bibliografías. Así, podemos destacar el número de artículos que Google recupera (14) frente a Yahoo (7) o WiseNut (5). MSN y Teoma sólo recuperaron dos. Sin embargo MSN destaca en la recuperación de capítulos de monografías (6) frente a Yahoo con cuatro, Google con dos y WiseNut con uno.

Son menos los motores que dan acceso a bibliotecas digitales (Google, Teoma y Yahoo), Directorios (MSN, WiseNut y Yahoo), Blogs o páginas personales (MSN, Teoma y Yahoo), artículos de enciclopedias (Google, Teoma y WiseNut) y diccionarios terminológicos (Google, MSN y WiseNut). Sólo recuperan documentación presentada a congresos o relativa a ellos, MSN y WiseNut. Google y WiseNut son los únicos que recuperan documentos en PowerPoint en esta búsqueda. MSN se caracteriza por el alto número de accesos a Blogs y a capítulos de monografías, y Teoma por facilitar más recursos de bibliotecas digitales que el resto.

Por otro lado hay que mencionar la recuperación de páginas en blanco, llevada a cabo por MSN y Teoma, que denotan un mal funcionamiento de estos motores, ya que deberían estar configurados para evitar ofrecer este tipo de recursos sin contenido.

MSN y WiseNut facilitan el acceso a un mayor número de noticias con información actualizada.

Gráfico 3.3-5. Motores. Búsqueda 1. Tipología documental

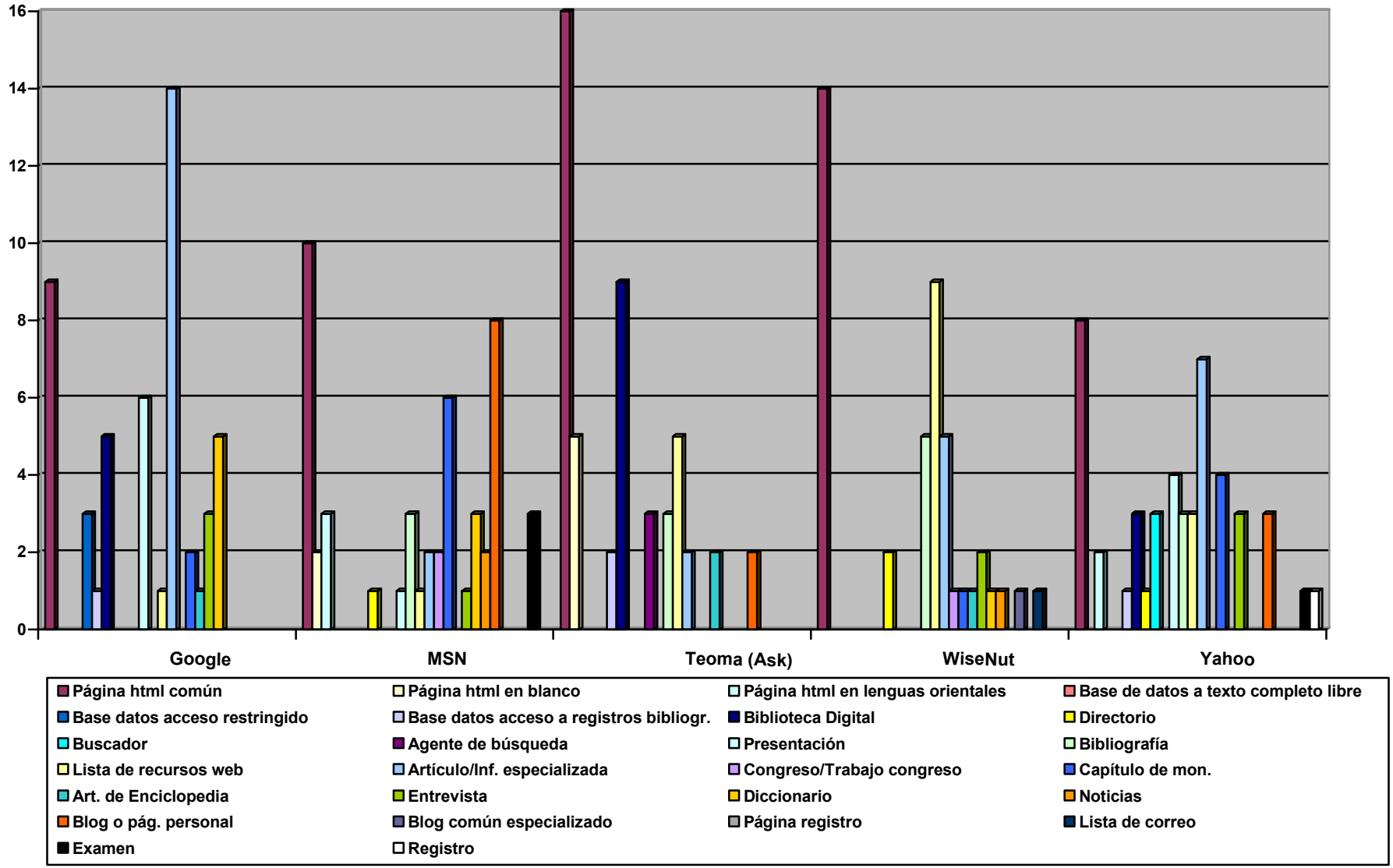


Tabla 3.3.1-8. Metabuscadores. Búsqueda 1. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común	12 (24%)	7 (14,3%)	6 (19,4%)	8 (20%)	14 (28,6%)	4 (26,7%)	12 (24,5%)
Página html en blanco	2 (4%)	1 (2%)	1 (3,2%)	1 (2,5%)			2 (4,1%)
Página html en lenguas orientales							
Imagen							
Base de datos a texto completo libre							
Base datos acceso restringido							
Base datos acceso a registros bibliogr.				2 (5%)	1 (2%)		1 (2%)
Biblioteca Digital	3 (6%)	6 (12,2%)	4 (12,9 %)	5 (12,5%)	3 (6,1%)	1 (6,7%)	6 (12,2%)
Repositorio							
Directorio	1 (2%)	1 (2%)	3 (9,7%)	1 (2,5%)	2 (4,1%)	2 (13,3%)	
Buscador							
Agente de búsqueda	1 (2%)						1 (2%)
Normas							
Lista de correo							
Revista electrónica							
E-libro							
Presentación		1 (2%)		1 (2,5%)	1 (2%)		
Bibliografía	3 (6%)	3 (6,1%)	1 (3,2%)	5 (12,5%)	3 (6,1%)	1 (6,7%)	2 (4,1%)
Lista de recursos web	6 (12%)	1 (2%)	2 (6,5 %)	1 (2,5%)	3 (6,1%)	2 (13,3%)	5 (10,2%)
Artículo/Inf. especializada	5 (10%)	14 (28,6%)	4 (12,9 %)	3 (7,5%)	3 (6,1%)		4 (8,2%)
Artículo de rev. Electrónica							
Congreso/Trabajo congreso	1 (2%)						2 (4,1%)
Monografía							
Capítulo de mon.	3 (6%)	3 (6,1%)	2 (6,5 %)	1 (2,5%)	2 (4,1%)		2 (4,1%)
Art. de Enciclopedia	1 (2%)	2 (4,1%)	2 (6,5 %)	2 (2,5%)	2 (4,1%)	2 (13,3%)	1 (2%)
Entrevista	1 (2%)	2 (4,1%)		1 (2,5%)	1 (2%)		1 (2%)
Diccionario	2 (4%)	1 (2%)		1 (2,5%)	1 (2%)		2 (4,1%)
Noticias			1 (3,2%)			1 (6,7%)	
Blog o pág. personal	7 (14%)	3 (6,1%)	2 (6,5 %)	7 (17,5%)	9 (18,4%)	1 (6,7%)	7 (14,3%)
Blog común especializado			1 (3,2%)	1 (2,5%)	1 (2%)	1 (6,7%)	
Página registro							
Lista de correo		1 (2%)					
Discurso		1 (2%)					
Proyecto	1 (2%)	1 (2%)	1 (3,2%)		1 (2%)		

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Curso o inf. de curso							
Resumen							
Repositorio							
FAQ							
Normas							
Examen	1 (2%)	1 (2%)	1 (3,2%)		1 (2%)		
Registro							1 (2%)
Banco de datos							
Repositorio							

La tipología documental recuperada por todos los metabuscadores coincide con la de los motores de búsqueda. Así, todos ellos recuperan páginas HTML, páginas de acceso a bibliotecas digitales, a bibliografía y a listas de recursos web. Dogpile, Search y Vivisimo son los metabuscadores que mayor número de páginas web de carácter general ofrecen.

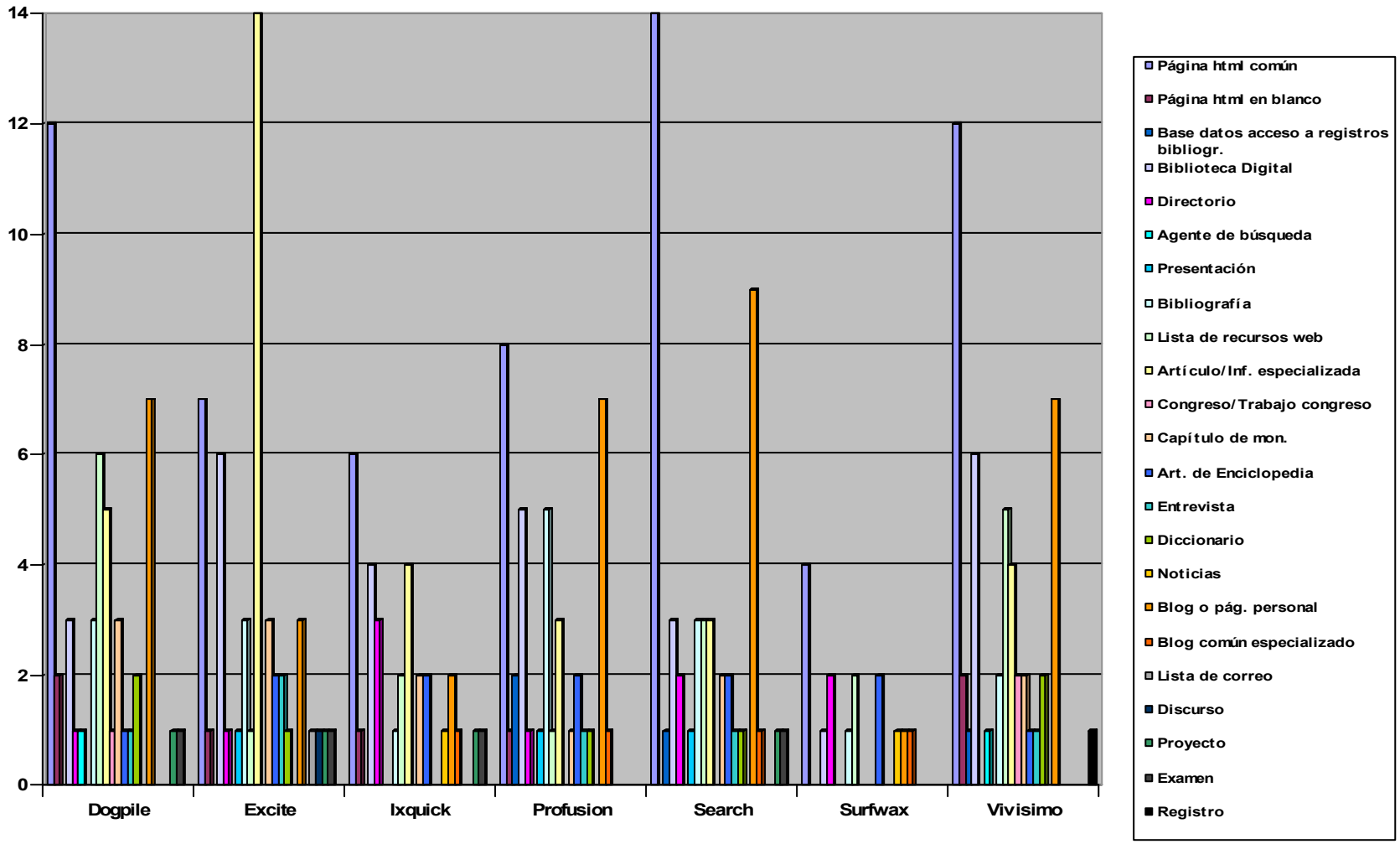
Pero a diferencia de los motores de búsqueda, todos los metabuscadores recuperan artículos de enciclopedia y facilitan el acceso a Blogs. Search con nueve, y Dogpile y Profusion con siete, son los que más recursos de este tipo ofrecen.

En la recuperación de artículos destaca Excite con catorce, seguido a cierta distancia por Dogpile (5), Ixquick (4) y Vivisimo (4). Trabajos presentados a congresos son facilitados por Vivisimo (2) y Dogpile (1). En la recuperación de capítulos de monografías y de artículos de enciclopedias están todos ellos muy igualados, si bien Ixquick y Surfwax no proporcionan acceso a entradas de diccionario, aunque sí lo hacen a noticias.

Los metabuscadores proporcionan acceso a la totalidad de proyectos de investigación recuperados.

Llama la atención la mayor recuperación de recursos pertenecientes a bibliotecas digitales, directorios, blogs, proyectos, que la observada en los motores.

Gráfico 3.3-6. Metabuscadores. Búsqueda 1. Tipología documental



3.3.1.4.1.2. Búsqueda utilizando el lenguaje natural

Tabla 3.3.1-9. Motores. Búsqueda 2. Tipología documental

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Página html común	3 (6,1%)	6 (12,8%)	2 (4,3%)	13 (26,5%)	3 (6,5%)
Página html en blanco	2 (4,1%)				
Página html en lenguas orientales					
Imagen					
Base de datos a texto completo libre					
Base datos acceso restringido					
Base datos acceso a registros bibliogr.					
Biblioteca Digital				2 (4,1%)	
Repositorio					
Directorio	1 (2%)	1 (2,1%)			
Buscador	2 (4,1%)	5 (10,6%)	2 (4,3%)		4 (8,7%)
Agente de búsqueda					
Normas					
Lista de correo					
Revista electrónica					
E-libro					
Presentación		1 (2,1%)	1 (2,1%)		
Bibliografía		1 (2,1%)	5 (10,6%)	1 (2%)	1 (2,2%)
Lista de recursos web	7 (14,3%)	3 (6,4%)	8 (17%)	1 (2%)	3 (6,5%)
Artículo/Inf. especializada	23 (46,9%)	17 (36,2%)	19 (40,4%)	5 (10,2%)	18 (39,1%)
Artículo de rev. Electrónica		1 (2,1%)			1 (2,2%)
Congreso/Trabajo congreso	1 (2%)	2 (4,3%)	3 (6,4%)	2 (4,1%)	4 (8,7%)
Monografía		1 (2,1%)			
Capítulo de mon.	3 (6,1%)	3 (6,4%)	3 (6,4%)	4 (8,2%)	3 (6,5%)
Art. de Enciclopedia	1 (%)	1 (2,1%)	1 (2,1%)		2 (4,3%)
Entrevista					
Diccionario	3 (6,1%)	1 (2,1%)	1 (2,1%)		
Noticias	2 (4,1%)	1 (2,1%)	1 (2,1%)	14 (28,6%)	3 (6,5%)
Blog o pág. personal					
Blog común especializado		1 (2,1%)		5 (10,2%)	1 (2,2%)
Página registro					
Lista de correo					
Discurso					
Proyecto					
Curso o inf. de curso	1 (2%)	1 (2,1%)	1 (2,1%)	1 (2%)	3 (6,5%)

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Resumen					
Repositorio					
FAQ		1 (2,1%)		1 (2%)	
Normas					
Examen					
Registro					
Banco de datos					
Repositorio					

El primer dato a destacar en esta búsqueda es que la tipología documental es más reducida que en la anterior. Los documentos más recuperados son páginas HTML, si bien, en esta ocasión en menor cantidad. El buscador que más recursos de este tipo recupera es WiseNut.

El segundo tipo documental de mayor frecuencia corresponde a artículos, destacando Google con veintitrés, frente a WiseNut con cinco. Sin embargo éste último destaca en el apartado de "noticias" con catorce recursos de este tipo y cinco accesos a blogs especializados. MSN experimenta un importante incremento en la recuperación de artículos respecto a la búsqueda anterior. Otros tipos documentales con un importante número de recursos recuperados, son las listas de recursos web, en las que destacan Google y Teoma; los capítulos de monografías y la documentación relativa a congresos.

Con respecto a la anterior búsqueda, sólo WiseNut facilita el acceso a recursos de bibliotecas digitales, y descienden también los recursos con información recogida en enciclopedias y diccionarios.

Ente Google, MSN y Yahoo, el primero recupera menos recursos HTML, superando a los demás en la recuperación de artículos. MSN recupera más páginas HTML que los otros dos, y proporciona más páginas con listados de recursos web. Yahoo tiene un comportamiento similar al señalado en Google, aunque generalmente, con cifras más bajas y lo supera en el número de recursos que facilitan el acceso a un buscador, y en recursos que facilitan información sobre congresos. Google es en esta ocasión el único motor que recupera páginas en blanco.

Gráfico 3.3-7. Motores. Búsqueda 2. Tipología documental

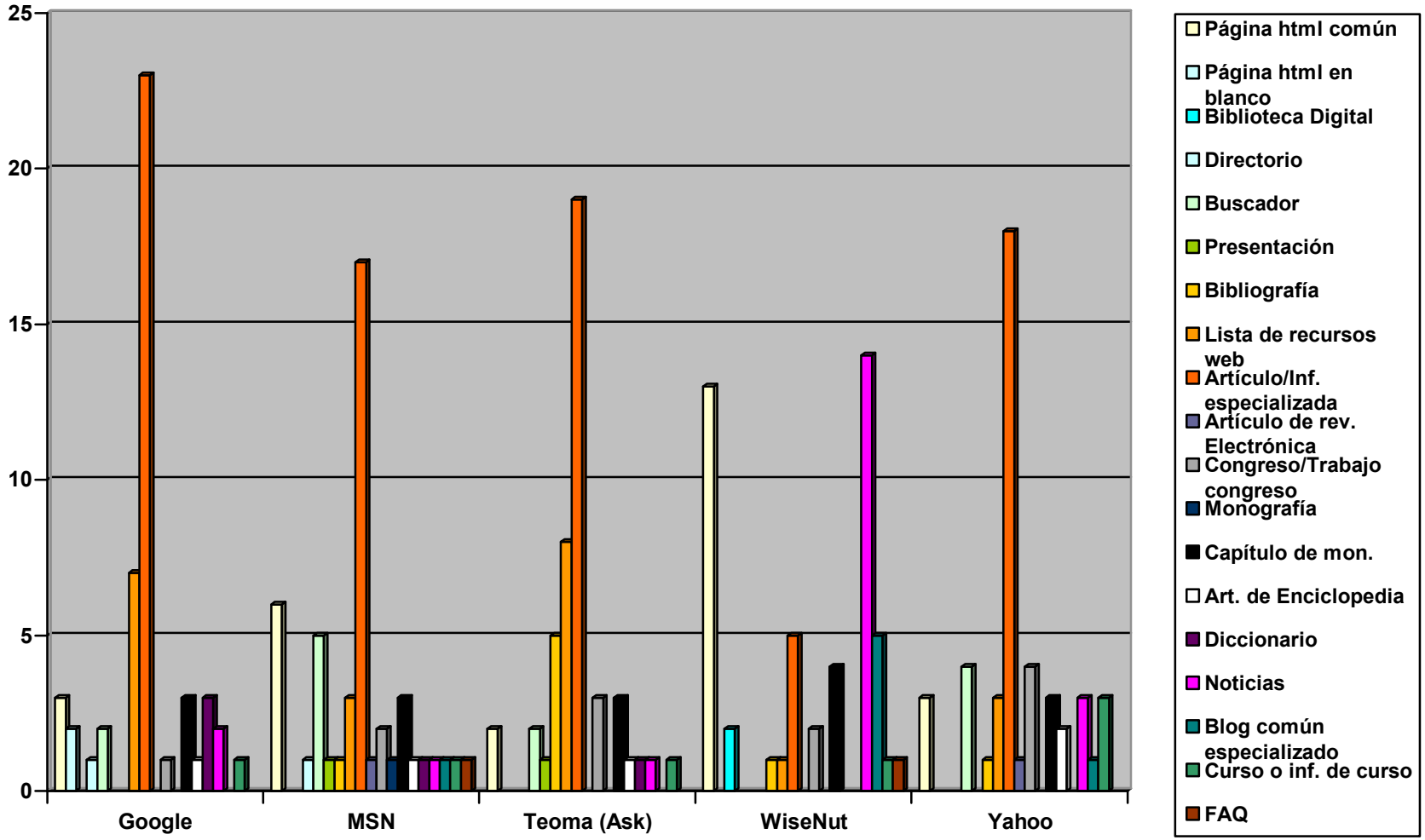


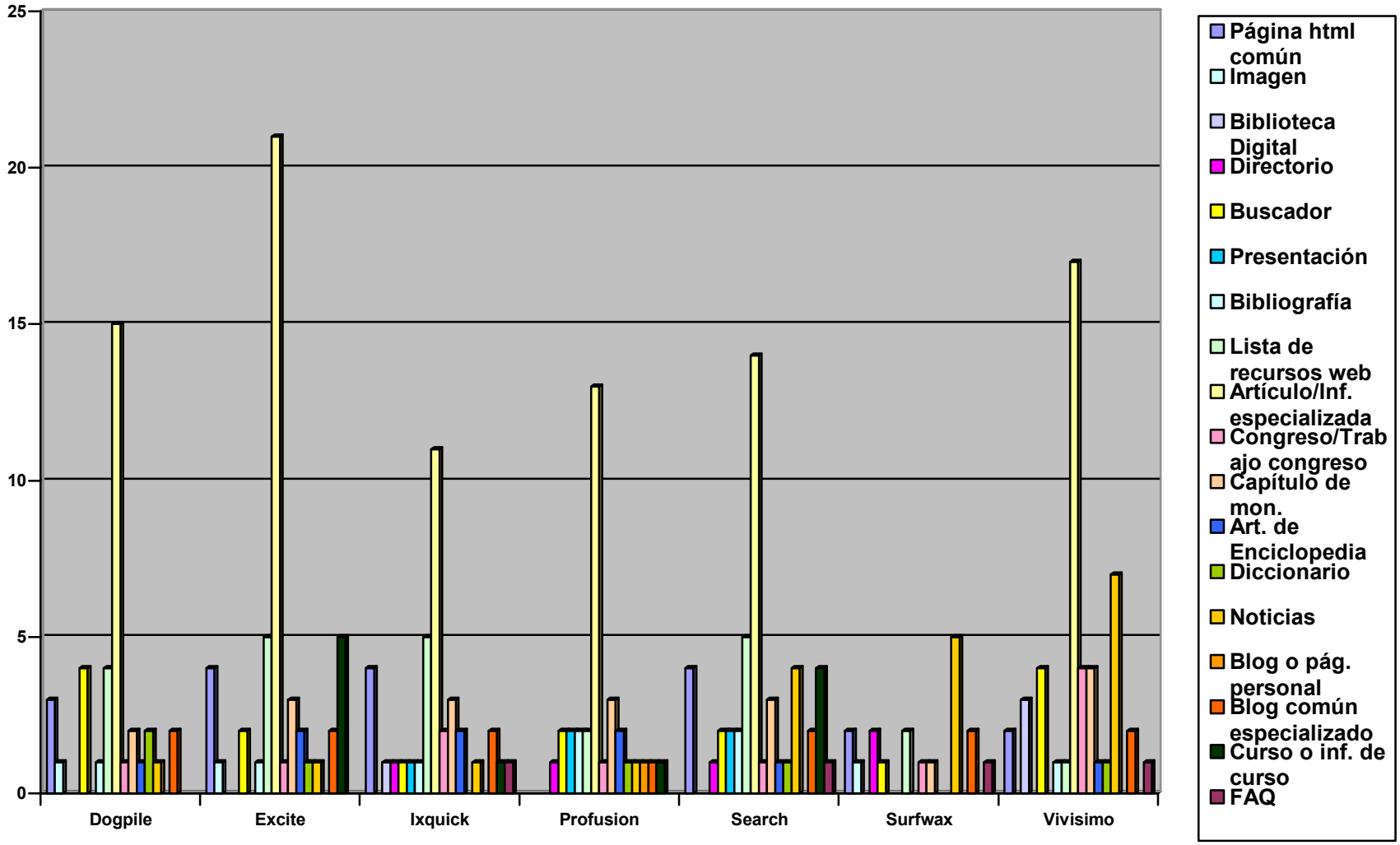
Tabla 3.3.1-10. Metabuscadores. Búsqueda 2. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común	3 (8,1%)	4 (8,2%)	4 (10,8%)		4 (8,5%)	2 (11,1%)	2 (4,2%)
Página html en blanco							
Página html en lenguas orientales							
Imagen	1 (2,7%)	1 (2%)				1 (5,6%)	
Base de datos a texto completo libre							
Base datos acceso restringido							
Base datos acceso a registros bibliogr.							
Biblioteca Digital			1 (2,7%)				3 (6,3%)
Repositorio							
Directorio			1 (2,7%)	1 (2,9%)	1 (2,1%)	2 (11,1%)	
Buscador	4 (10,8%)	2 (4,1%)	1 (2,7%)	2 (5,7%)	2 (4,3%)	1 (5,6%)	4 (8,3%)
Agente de búsqueda							
Normas							
Lista de correo							
Revista electrónica							
E-libro							
Presentación			1 (2,7%)	2 (5,7%)	2 (4,3%)		
Bibliografía	1 (2,7%)	1 (2%)	1 (2,7%)	2 (5,7%)	2 (4,3%)		1 (2,1%)
Lista de recursos web	4 (10,8%)	5 (10,2%)	5 (13,5%)	2 (5,7%)	5 (10,6%)	2 (11,1%)	1 (2,1%)
Artículo/Inf. especializada	15 (40,5%)	21 (42,9%)	11 (29,7%)	13 (37,1%)	14 (29,8%)		17 (35,4%)
Artículo de rev. Electrónica							
Congreso/Trabajo congreso	1 (2,7%)	1 (2%)	2 (5,4%)	1 (2,9%)	1 (2,1%)	1 (5,6%)	4 (8,3%)
Monografía							
Capítulo de mon.	2 (5,4%)	3 (6,1%)	3 (8,1%)	3 (8,6%)	3 (6,4%)	1 (5,6%)	4 (8,3%)
Art. de Enciclopedia	1 (2,7%)	2 (4,1%)	2 (5,4%)	2 (5,7%)	1 (2,1%)		1 (2,1%)
Entrevista							
Diccionario	2 (5,4%)	1 (2%)		1 (2,9%)	1 (2,1%)		1 (2,1%)
Noticias	1 (2,7%)	1 (2%)	1 (2,7%)	1 (2,9%)	4 (8,5%)	5 (27,8%)	7 (14,6%)
Blog o pág. personal				1 (2,9%)			
Blog común especializado	2 (5,4%)	2 (4,1%)	2 (5,4%)	1 (2,9%)	2 (4,3%)	2 (11,1%)	2 (4,2%)
Página registro							
Lista de correo							
Discurso							
Proyecto							
Curso o inf. de curso		5 (10,2%)	1 (2,7%)	1 (2,9%)	4 (8,5%)		

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Resumen							
Repositorio							
FAQ			1 (2,7%)		1 (2,1%)	1 (5,6%)	1 (2,1%)
Normas							
Examen							
Registro							
Banco de datos							
Repositorio							

Los metabuscadores no se alejan mucho de lo observado en los motores de búsqueda. Profusión es el único que no recupera páginas en HTML aunque da acceso a trece artículos, repartiéndose el resto de registros entre la variada tipología, sin que destaque ninguno sobre los demás. Todos, excepto Surfwax, recuperan un importante número de artículos, aspecto en el que sobresalen Excite y Vivisimo. Otros tipos documentales frecuentes son las listas de recursos web, información relacionada con congresos, listados proporcionados por buscadores y artículos de monografías. Disminuye en esta consulta la recuperación de páginas relativas a bibliotecas digitales. Vivisimo destaca en esta búsqueda en la recuperación de noticias y en recursos con información sobre congresos.

Gráfico 3.3-8. Metabuscadores. Búsqueda 2. Tipología documental



3.3.1.4.1.3. Búsqueda con operadores de existencia

Tabla 3.3.1-11. Motores. Búsqueda 3. Tipología documental

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Página html común	12 (25%)	19 (38,8%)	5 (10,2%)	7 (15,2%)	7 (16,3%)
Página html en blanco					
Página html en lenguas orientales					
Imagen					
Base de datos a texto completo libre	1 (2,1%)				3 (7%)
Base datos acceso restringido	1 (2,1%)			1 (2,2%)	
Base datos acceso a registros bibliogr.		1 (2%)			1 (2,3%)
Biblioteca Digital					
Repositorio					
Directorio				2 (4,3%)	1 (2,3%)
Buscador	3 (6,3%)	1 (2%)	1 (2%)	3 (6,5%)	
Agente de búsqueda					
Normas					
Lista de correo					
Revista electrónica	2 (4,2%)	1 (2%)		1 (2,2%)	1 (2,3%)
E-libro	1 (2,1%)				
Presentación					
Bibliografía	2 (4,2%)	3 (6,1%)	3 (6,1%)	4 (8,7%)	4 (9,3%)
Lista de recursos web	9 (18,8%)	7 (14,3%)	6 (12,2%)	4 (8,7%)	
Artículo/Inf. especializada	7 (14,6%)	6 (12,2%)	21 (42,9%)	9 (19,6%)	15 (34,9%)
Artículo de rev. Electrónica					
Congreso/Trabajo congreso	1 (2,1%)	6 (12,2%)	5 (10,2%)	2 (4,3%)	1 (2,3%)
Monografía					
Capítulo de mon.	1 (2,1%)		3 (6,1%)	2 (4,3%)	1 (2,3%)
Art. de Enciclopedia	1 (2,1%)	1 (2%)			1 (2,3%)
Entrevista					
Diccionario					
Noticias	3 (6,3%)	1 (2%)	1 (2%)	3 (6,5%)	1 (2,3%)
Blog o pág. personal			1 (2%)	1 (2,2%)	
Blog común especializado					
Página registro					

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Lista de correo				1 (2,2%)	
Discurso					
Proyecto			1 (2%)		
Curso o inf. de curso	2 (4,2%)	2 (4,1%)	1 (2%)	3 (6,5%)	5 (11,6%)
Resumen					
Repositorio				3 (6,5%)	
FAQ					
Normas	1 (2,1%)				
Examen					
Registro	1 (2,1%)	1 (2%)			1 (2,3%)
Banco de datos					
Repositorio					

Como en la búsqueda anterior es común en todas las herramientas la recuperación de páginas HTML. Profusión en esta búsqueda no ofreció resultados.

En esta búsqueda, el mayor número de recursos corresponde a páginas HTML y a artículos. Yahoo, a diferencia de Google y MSN, facilita menos páginas en HTML y más artículos, mientras que MSN es el buscador que más páginas en HTML recupera. Teoma recupera veintiún artículos, seguido de Yahoo con quince, Google con siete, y MSN seis. Éste último destaca, junto a Teoma, en la recuperación de recursos relativos a congresos. Llama la atención el descenso en artículos recuperados por Google y por MSN. Por otro lado hay que destacar la recuperación en esta búsqueda de recursos en forma de bases de datos.

Gráfico 3.3-9. Motores. Búsqueda 3. Tipología documental

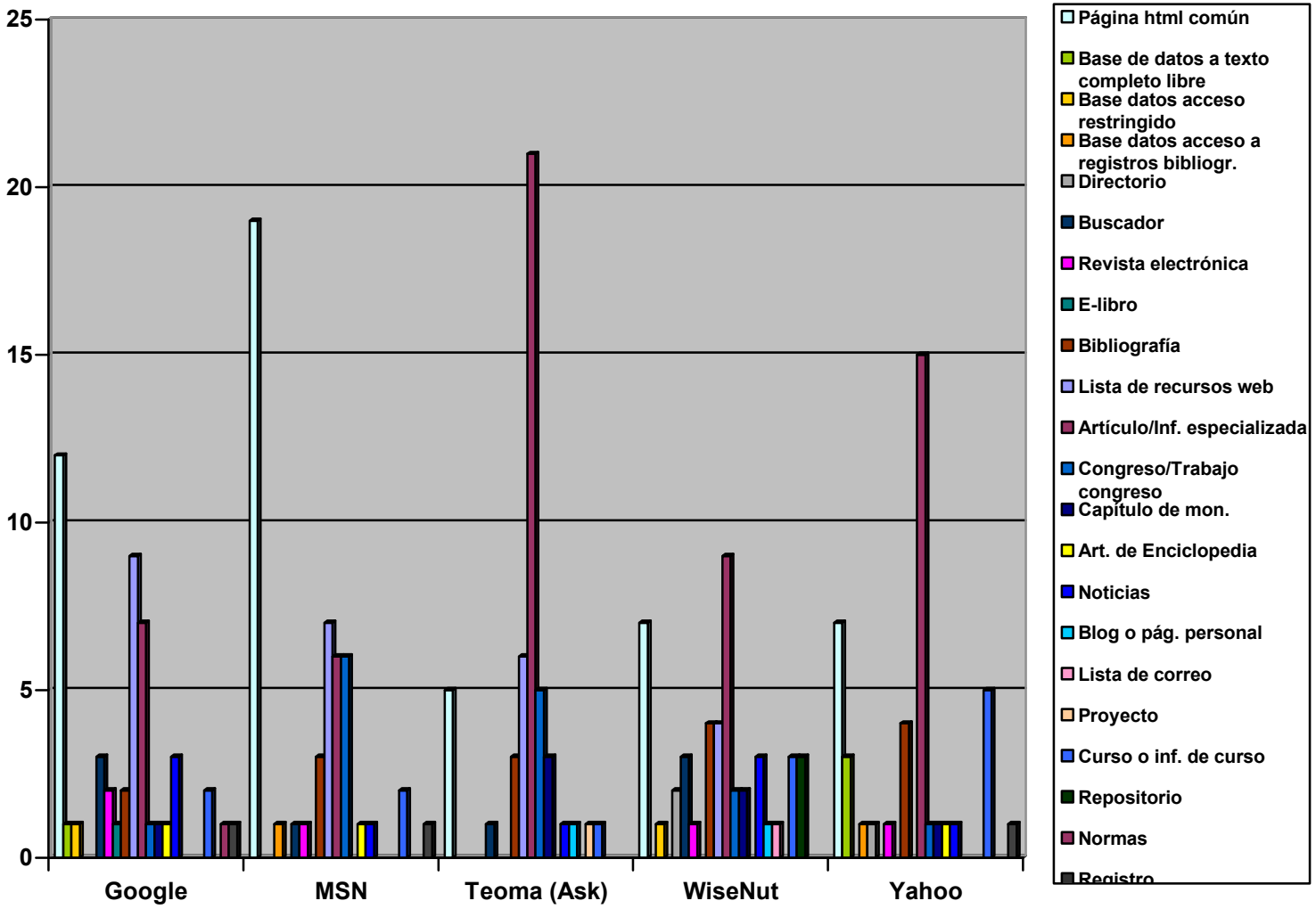


Tabla 3.3.1-12. Metabuscadores. Búsqueda 3. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común	14 (35%)	9 (20,5%)	12 (26,7%)		8 (17%)	7 (20,6%)	15 (31,9%)
Página html en blanco			1 (2,2%)		1 (2,1%)	1 (2,9%)	
Página html en lenguas orientales							
Imagen							
Base de datos a texto completo libre	1 (2,5%)	1 (2,3%)	1 (2,2%)		1 (2,1%)		
Base datos acceso restringido			1 (2,2%)		1 (2,1%)	1 (2,9%)	
Base datos acceso a registros bibliogr.							
Biblioteca Digital							
Repositorio							
Directorio		7 (15,9%)	1 (2,2%)			4 (11,8%)	
Buscador	1 (2,5%)		3 (6,7%)			1 (2,9%)	
Agente de búsqueda							
Normas							
Lista de correo							
Revista electrónica			1 (2,2%)		1 (2,1%)	1 (2,9%)	2 (4,3%)
E-libro							
Presentación							
Bibliografía	3 (7,5%)	5 (11,4%)	5 (11,1%)		4 (8,5%)	2 (5,9%)	1 (2,1%)
Lista de recursos web	8 (20%)	6 (13,6%)	3 (6,7%)		8 (17%)		7 (14,9%)
Artículo/Inf. especializada	6 (15%)	12 (27,3%)	8 (17,8%)		9 (19,1%)	7 (20,6%)	10 (21,3%)
Artículo de rev. Electrónica						1 (2,9%)	
Congreso/Trabajo congreso	1 (2,5%)		1 (2,2%)		1 (2,1%)	1 (2,9%)	2 (4,3%)
Monografía							
Capítulo de mon.		1 (2,3%)	1 (2,2%)		2 (4,3%)		2 (4,3%)
Art. de Enciclopedia	1 (2,5%)		1 (2,2%)		4 (4,3%)		1 (2,1%)
Entrevista							
Diccionario							
Noticias	2 (5%)	1 (2,3%)	2 (4,4%)		3 (6,4%)	7 (20,6%)	2 (4,3%)
Blog o pág. personal			1 (2,2%)		1 (2,1%)		1 (2,1%)
Blog común especializado							
Página registro							
Lista de correo							
Discurso							
Proyecto	1 (2,5%)	1 (2,3%)					1 (2,1%)
Curso o inf. de curso	1 (2,5%)		1 (2,2%)		4 (8,5%)	1 (2,9%)	3 (6,4%)

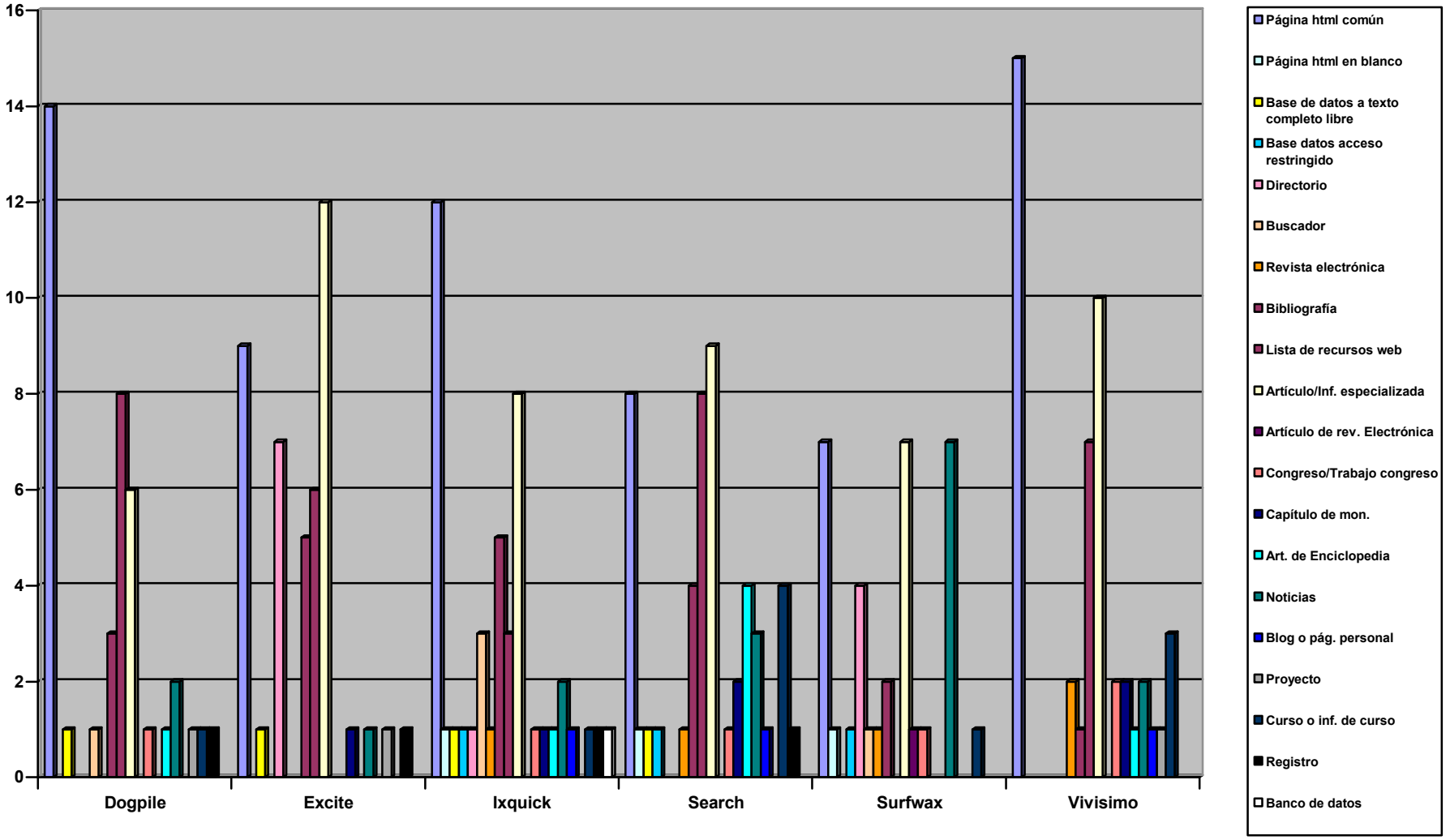
	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Resumen							
Repositorio							
FAQ							
Normas							
Examen							
Registro	1 (2,5%)	1 (2,3%)	1 (2,2%)		1 (2,1%)		
Banco de datos			1 (2,2%)				
Repositorio							

Entre los metabuscadores hay cierto predominio de páginas en HTML con información común, aunque tiende a igualarse con los artículos. Vivisimo y Dogpile recuperan más páginas web con información común. Surfwax aumenta en la recuperación de artículos respecto a la búsqueda anterior, si bien no alcanza las cifras que ofrecen para este tipo de documentos metabuscadores como Excite con doce 12, Vivisimo con diez y Search con nueve. También la recuperación de recursos en forma de bases de datos aparece entre los metabuscadores. En esta ocasión, el mayor número de recursos relacionados con noticias corresponde a Surfwax.

Los listados con recursos web son el tercer grupo de recuperación, en el que todos los metabuscadores están bastante igualados a excepción de Ixquick.

Excite destaca en esta búsqueda en la recuperación de recursos en forma de directorio y Surfwax en noticias. Ixquick, Search y Surfwax recuperaron páginas en blanco. Search ofrece un mayor número de artículos de enciclopedias y de páginas con información sobre cursos. El acceso a páginas con bibliografía es común en todos los metabuscadores destacando, Excite e Ixquick, seguidos por Search.

Gráfico 3.3-10. Metabuscadores. Búsqueda 3. Tipología documental



3.3.1.4.1.4. Búsqueda booleana

Tabla 3.3.1-13. Motores. Búsqueda 4. Tipología documental

	Google	MSN	Teoma	WiseNut	Yahoo
Página html común	1 (2,2%)				
Página html en blanco					
Página html en lenguas orientales					
Imagen					
Base de datos a texto completo libre					
Base datos acceso restringido					
Base datos acceso a registros bibliogr.					
Biblioteca Digital		1 (2%)			1 (2%)
Repositorio					
Directorio					
Buscador	1 (2,2%)	2 (4%)			
Agente de búsqueda					
Normas					
Lista de correo					
Revista electrónica		4 (8%)			
E-libro					
Presentación					
Bibliografía	4 (8,9%)	3 (6%)			6 (12%)
Lista de recursos web	11 (24,4%)	21 (42%)			13 (26%)
Artículo/Inf. especializada	20 (44,4%)	11 (22%)			16 (32%)
Artículo de rev. Electrónica					
Congreso/Trabajo congreso	2 (4,4%)	1 (2%)			4 (8%)
Monografía					
Capítulo de mon.	2 (4,4%)				1 (2%)
Art. de Enciclopedia					
Entrevista					1 (2%)
Diccionario					
Noticias	2 (4,4%)	1 (2%)			2 (4%)
Blog o pág. personal		2 (4%)			
Blog común especializado		1 (2%)			
Página registro					
Lista de correo					
Discurso					
Proyecto	1 (2,2%)	1 (2%)			3 (6%)
Curso o inf. de curso	1 (2,2%)				3 (6%)

	Google	MSN	Teoma	WiseNut	Yahoo
Resumen					
Repositorio					
FAQ		1 (2%)			
Normas		1 (2%)			
Examen					
Registro					
Banco de datos					
Repositorio					

La tipología documental es en esta búsqueda la mas reducida, correspondiendo el mayor número de recursos a artículos. Google con veinte, es el buscador que recupera mayor número de artículos, seguido de Yahoo con dieciséis y MSN con once. Yahoo destaca en la recuperación de recursos relativos a bibliografía, congresos, proyectos de investigación e información sobre cursos. MSN proporciona acceso a revistas electrónicas, blogs tanto personales como especializados, acceso a FAQ y Normas pero en lo que más destaca es en facilitar listados de recursos web.

Gráfico 3.3-11. Motores. Búsqueda 4. Tipología documental

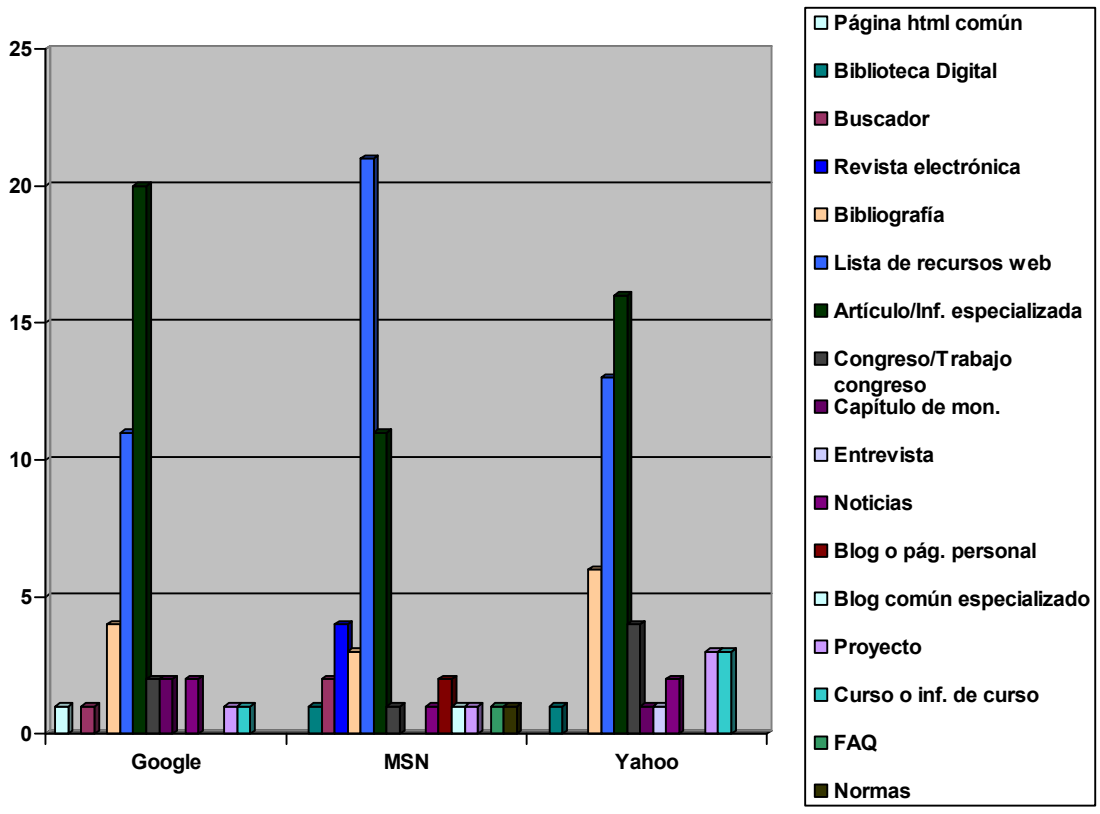


Tabla 3.3.1-14. Metabuscadores. Búsqueda 4. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común		1 (3%)					
Página html en blanco							
Página html en lenguas orientales							
Imagen							
Base de datos a texto completo libre							
Base datos acceso restringido							
Base datos acceso a registros bibliogr.		1 (3%)					
Biblioteca Digital		2 (6,1%)	2 (10,5%)	1 (2,4%)	1 (2,1%)		2 (4,3%)
Repositorio							
Directorio		3 (9,1%)	1 (5,3%)	2 (4,8%)			
Buscador		3 (9,1%)		1 (2,4%)			
Agente de búsqueda							
Normas							
Lista de correo							
Revista electrónica		2 (6,1%)	1 (5,3%)	2 (4,8%)	2 (4,3%)		
E-libro							
Presentación							
Bibliografía				2 (4,8%)	6 (12,8%)		1 (2,1%)
Lista de recursos web		7 (21,2%)	4 (21,1%)	12 (28,6%)	16 (34%)		23 (48,9%)
Artículo/Inf. especializada		10 (30,3%)	8 (42,1%)	13 (31%)	12 (25,5%)		11 (23,4%)
Artículo de rev. Electrónica				1 (2,4%)	1 (2,1%)		
Congreso/Trabajo congreso		1 (3%)		1 (2,4%)	4 (8,5%)		3 (6,4%)
Monografía							
Capítulo de mon.			1 (5,3%)		1 (2,1%)		
Art. de Enciclopedia			1 (5,3%)	1 (2,4%)			
Entrevista			1 (5,3%)	1 (2,4%)			
Diccionario							2 (4,3%)
Noticias				1 (2,4%)			2 (4,3%)
Blog o pág. personal							
Blog común especializado							1 (2,1%)
Página registro							

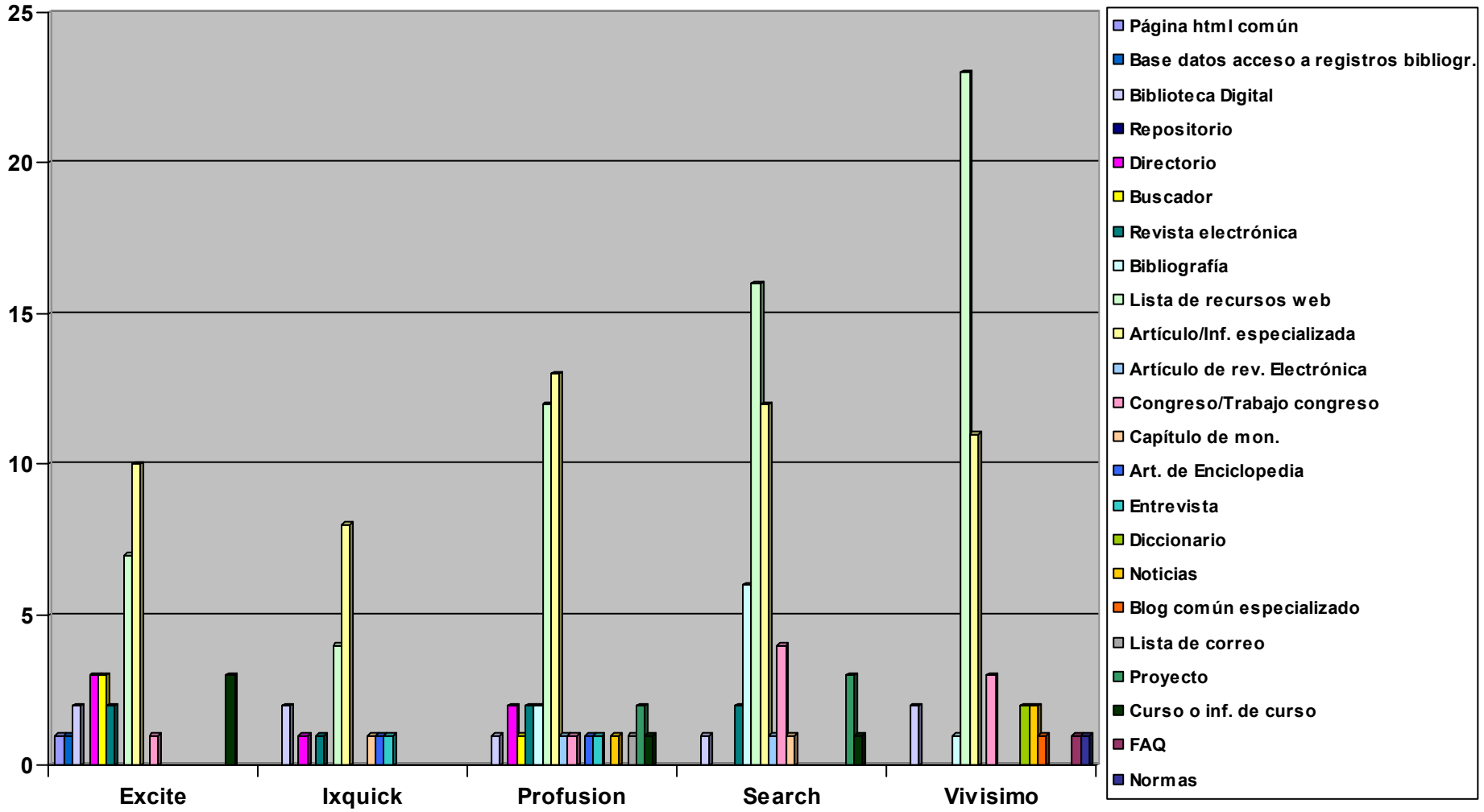
	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Lista de correo				1 (2,4%)			
Discurso							
Proyecto				2 (4,8%)	3 (6,4%)		
Curso o inf. de curso		3 (9,1%)		1 (2,4%)	1 (2,1%)		
Resumen							
Repositorio							
FAQ							1 (2,1%)
Normas							1 (2,1%)
Examen							
Registro							
Banco de datos							
Repositorio							

En esta búsqueda el número de páginas HTML desciende considerablemente respecto al resto de consultas. Son las listas de recursos web y los artículos especializados los tipos de documentos más recuperados, por lo que ante este tipo de búsquedas, la especialidad de los recursos aumenta.

El mayor número de artículos lo recupera Profusión, seguido de Search y Vivisimo. Este último recupera el mayor número de listados de recursos web, con 23, seguido por Search y Profusión, con muy poca diferencia sobre Excite e Ixquick.

Search y Excite recuperan un mayor número de recursos en forma de bibliografías, y respecto a información relacionada con congresos, son Search y Vivisimo los que más páginas de este tipo recuperan. Search aventaja a los demás también en la recuperación de recursos de carácter bibliográfico y proyectos. Vivisimo supera a los demás en entradas de diccionario, FAQ, normas y acceso a noticias. Excite por su parte, recupera más recursos relacionados con directorios y buscadores e información sobre cursos.

Gráfico 3.3-12. Metabuscadores. Búsqueda 4. Tipología documental



3.3.1.4.1.5. Búsqueda de frase

Tabla 3.3.1-15. Motores. Búsqueda 5. Tipología documental

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Página html común	2 (4,1%)	4 (8%)		4 (8,2%)	3 (6%)
Página html en blanco		1 (2%)			
Página html en lenguas orientales				1 (2%)	
Imagen					
Base de datos a texto completo libre					
Base datos acceso restringido					
Base datos acceso a registros bibliogr.					
Biblioteca Digital					
Repositorio					
Directorio	2 (4,1%)	1 (2%)			2 (4%)
Buscador	1 (2%)				
Agente de búsqueda					
Normas					
Lista de correo					
Revista electrónica					
E-libro	1 (2%)				
Presentación					
Bibliografía	4 (8,2%)	4 (8%)		2 (4,1%)	3 (6%)
Lista de recursos web	3 (6,1%)	4 (8%)		6 (12,2%)	7 (14%)
Artículo/Inf. especializada	1 (2%)	5 (10%)		5 (10,2%)	7 (14%)
Artículo de rev. Electrónica				1 (2%)	
Congreso/Trabajo congreso	1 (2%)	4 (8%)		4 (8,2%)	
Monografía					
Capítulo de mon.	1 (2%)	2 (4%)		1 (2%)	
Art. de Enciclopedia	1 (2%)	1 (2%)		1 (2%)	1 (2%)
Entrevista					
Diccionario	1 (2%)			1 (2%)	1 (2%)
Noticias	2 (4,1%)	1 (2%)		1 (2%)	2 (4%)
Blog o pág. personal		2 (4%)			
Blog común especializado					
Página registro					
Lista de correo					

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Discurso					
Proyecto	23 (46%)	14 (28%)		18 (36,7%)	15 (30%)
Curso o inf. de curso	5 (10,2%)	4 (8%)		3 (6,1%)	6 (12%)
Resumen					
Repositorio					
FAQ	1 (2%)	2 (4%)			3 (6%)
Normas		1 (2%)		1 (2%)	
Examen					
Registro					
Banco de datos					
Repositorio					
Total					

En esta búsqueda, el tipo documental más frecuente son los proyectos de investigación. Google recupera un total de veintitrés recursos de este tipo, WiseNut dieciocho, Yahoo quince y MSN catorce. En relación con los artículos, llama la atención el descenso de este tipo documental respecto al resto de búsquedas; Yahoo, con siete, es el buscador que más recursos de este tipo recupera, seguido por MSN y WiseNut, ambos con cinco.

Los siguientes tipos en importancia son, de nuevo, las listas de recursos web y la información sobre cursos, ambos con Yahoo a la cabeza. Google recupera un libro electrónico y MSN y WiseNut recuperan un mayor número de recursos relacionados con congresos. MSN facilita el acceso a capítulos de monografías.

MSN es el único buscador que facilita páginas en blanco en esta búsqueda.

Gráfico 3.3-13 Motores. Búsqueda 5. Tipología documental

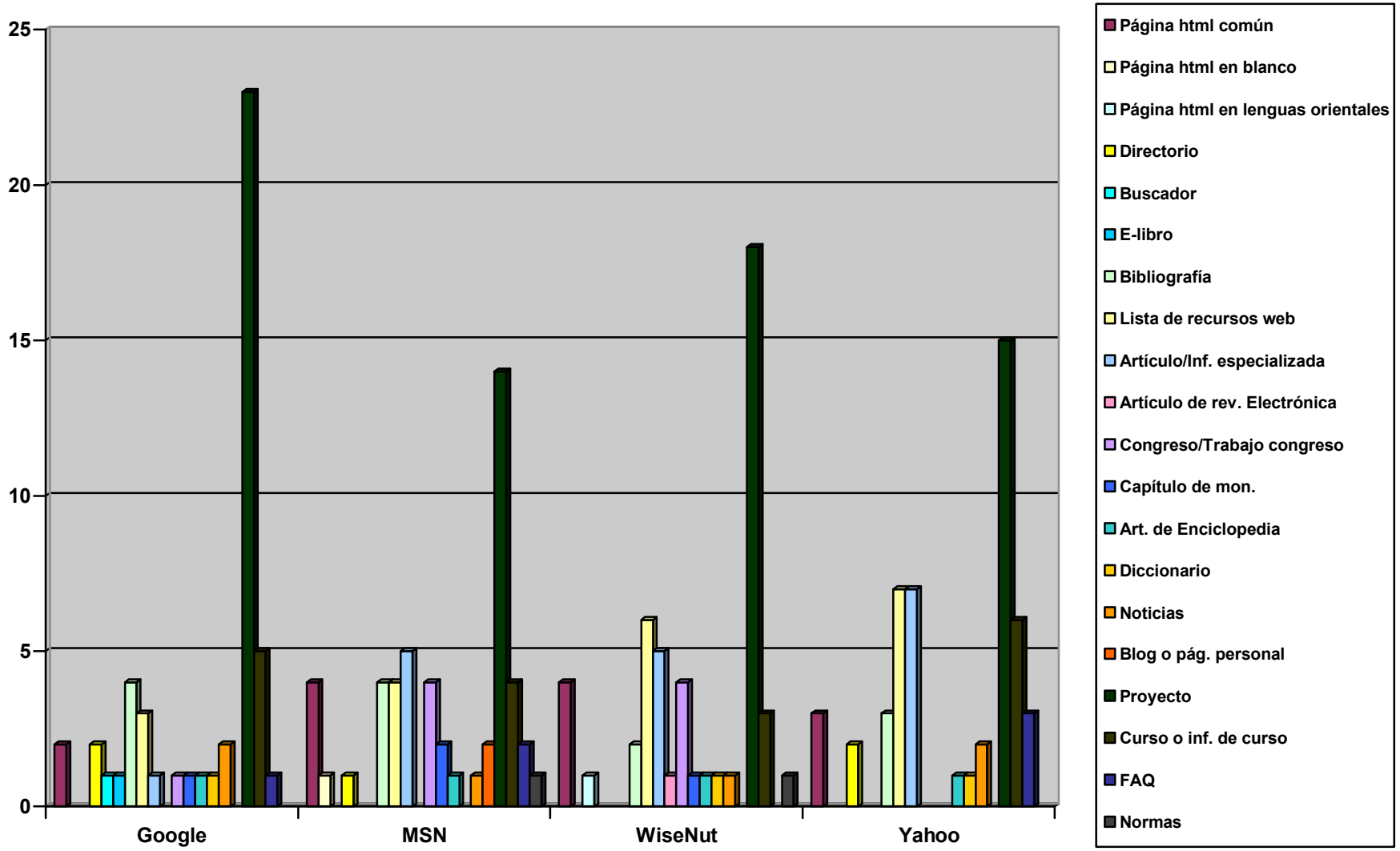


Tabla 3.3.1-16. Metabuscadores. Búsqueda 5. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común		10 (20,4%)	7 (16,7%)	2 (5,1%)	2 (4,1%)		3 (6,1%)
Página html en blanco					1 (2%)		1 (2%)
Página html en lenguas orientales							
Imagen							
Base de datos a texto completo libre			1 (2,4%)				
Base datos acceso restringido							
Base datos acceso a registros bibliogr.							
Biblioteca Digital							
Repositorio							
Directorio		3 (6,1%)	3 (7,1%)	1 (2,6%)	1 (2%)		
Buscador		3 (6,1%)					
Agente de búsqueda							
Normas							
Lista de correo							
Revista electrónica							
E-libro							1 (2%)
Presentación							
Bibliografía		4 (8,2%)	3 (7,1%)	2 (5,1%)	3 (6,1%)		3 (6,1%)
Lista de recursos web		4 (8,2%)	3 (7,1%)	2 (5,1%)	4 (8,2%)		4 (8,2%)
Artículo/Inf. especializada		5 (10,2%)	7 (16,7%)	2 (5,1%)	1 (2%)		18 (36,7%)
Artículo de rev. Electrónica			1 (2,4%)	1 (2,6%)			
Congreso/Trabajo congreso			1 (2,4%)	1 (2,6%)	6 (12,2%)		1 (2%)
Monografía							
Capítulo de mon.				1 (2,6%)	1 (2%)		
Art. de Enciclopedia		1 (2%)	1 (2,4%)	2 (5,1%)	1 (2%)		1 (2%)
Entrevista							
Diccionario		1 (2%)	1 (2,4%)	1 (2,6%)			
Noticias		2 (4,1%)	2 (4,8%)	3 (7,7%)			2 (4,1%)
Blog o pág. personal				1 (2,6%)			5 (10,2%)
Blog común especializado				4 (10,3%)			
Página registro							
Lista de correo							
Discurso							
Proyecto			11 (26,2%)	14 (35,9%)	22 (44,9%)		5 (10,2%)

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Curso o inf. de curso					1 (2%)		4 (8,2%)
Resumen							
Repositorio							
FAQ		3 (6,1%)	1 (2,4%)	2 (5,1%)	3 (6,1%)		1 (2%)
Normas							
Examen					1 (2%)		
Registro							
Banco de datos							
Repositorio							

Los metabuscadores reflejan una situación similar a la de los buscadores, y en relación con el acceso páginas con información sobre proyectos de investigación, Search recupera veintidós, Profusion catorce e Ixquick once. Excite no recupera recursos de este tipo.

El número de páginas HTML es superior a los facilitados por los motores de búsqueda.

El metabuscador Vivisimo destaca en la recuperación de artículos especializados, al recuperar dieciocho, seguido a gran distancia por Ixquick con siete. También hay que destacar que es el único metabuscador que facilita el acceso a un libro electrónico.

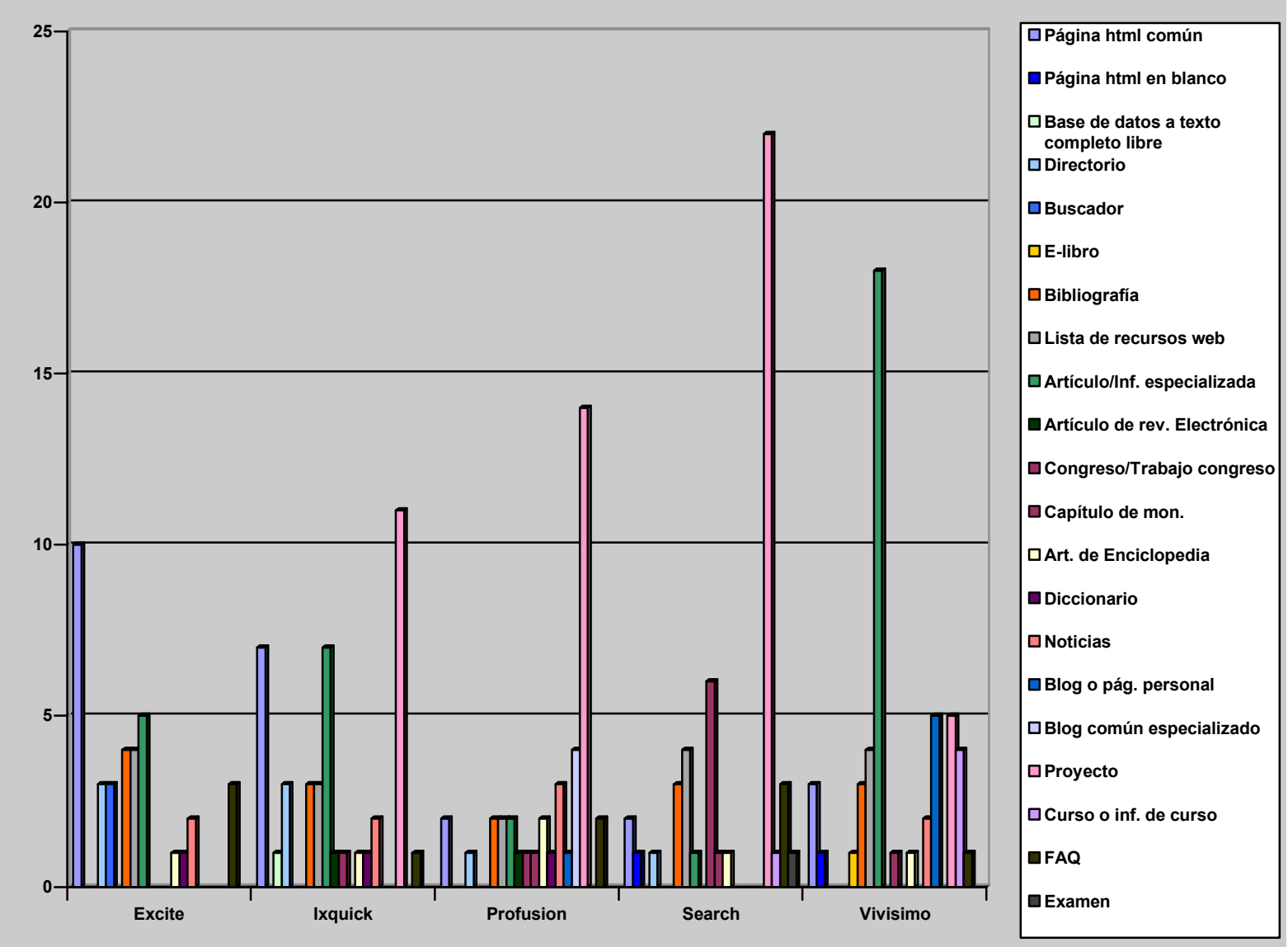
Los proyectos, ocupan en esta búsqueda un lugar destacado, como podemos apreciar en Search que recupera veintidós, seguido por Profusion e Ixquick.

Otro indicador de interés es la información relativa a Congresos, destacando el metabuscador Search con seis.

Excite facilita más páginas en HTML y listados proporcionados por otros buscadores.

Search y Vivisimo facilitan en esta búsqueda el acceso a una página en blanco.

Gráfico 3.3-14. Metabuscadores. Búsqueda 5. Tipología documental



3.3.1.4.1.6. Búsqueda por campo

Tabla 3.3.1-17. Motores. Búsqueda 6. Tipología documental

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Página html común	3 (6,3%)	5 (10,2%)		1 (3,7%)	5 (10%)
Página html en blanco					
Página html en lenguas orientales					
Imagen					
Base de datos a texto completo libre					
Base datos acceso restringido	1 (2,1%)				
Base datos acceso a registros bibliogr.					
Biblioteca Digital		1 (2%)			
Repositorio					
Directorio	2 (4,2%)	2 (4,1%)			2 (4%)
Buscador		1 (2%)		1 (3,7%)	
Agente de búsqueda					
Normas		1 (2%)			
Lista de correo					
Revista electrónica	1 (2,1%)				2 (4%)
E-libro	2 (4,2%)	2 (4,1%)			2 (4%)
Presentación					
Bibliografía	4 (8,3%)	2 (4,1%)		4 (14,8%)	4 (8%)
Lista de recursos web	8 (16,7%)	8 (16,3%)		2 (7,4%)	14 (28%)
Artículo/Inf. especializada	7 (14,6%)	14 (28,6%)		11 (40,7%)	3 (6%)
Artículo de rev. Electrónica					1 (2%)
Congreso/Trabajo congreso	6 (12,5%)	2 (4,1%)		1 (3,7%)	
Monografía					
Capítulo de mon.	3 (6,3%)	2 (4,1%)		1 (3,7%)	2 (4%)
Art. de Enciclopedia	2 (4,2%)	1 (2%)			2 (4%)
Entrevista					
Diccionario					1 (2%)
Noticias	1 (2,1%)	1 (2%)		4 (14,8%)	1 (2%)
Blog o pág. personal				2 (7,4%)	
Blog común especializado		1 (2%)			
Página registro					
Lista de correo		1 (2%)			

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Discurso					
Proyecto	4 (8,3%)	3 (6,1%)			5 (10%)
Curso o inf. de curso	4 (8,3%)	2 (4,1%)			5 (10%)
Resumen					
Repositorio					1 (2%)
FAQ					
Normas					
Examen					
Registro					
Banco de datos					
Repositorio					
Total					

En esta búsqueda, la recuperación de los motores se centra en facilitar artículos y ofrecer listados con recursos web. En este sentido, los buscadores Yahoo, Google y MSN ofrecen, por este orden, el mayor número de este tipo de listados. El mayor número de artículos de información especializada lo recupera MSN con catorce, seguido de WiseNut con once y Google con siete. MSN se diferencia poco de los demás buscadores aunque es el único que facilita el acceso a bibliotecas digitales, normas, blogs especializados y listas de correo.

Yahoo es el único que recupera artículos de revistas electrónicas, facilita entradas de diccionario y el acceso a repositorios. Su recuperación se centra más en ofrecer listados con recursos, páginas en HTML, información sobre proyectos y cursos, y acceso a recursos bibliográficos.

Google, al margen de lo antes señalado, se caracteriza en esta búsqueda por la recuperación de páginas con información sobre Congresos y bases de datos de acceso restringido, y junto a Yahoo, por recuperar revistas electrónicas.

La recuperación de WiseNut se caracteriza además de por el alto número de artículos, por ofrecer pocas páginas en HTML y un mayor acceso a noticias y blogs personales.

Gráfico 3.3-15. Motores. Búsqueda 6. Tipología documental

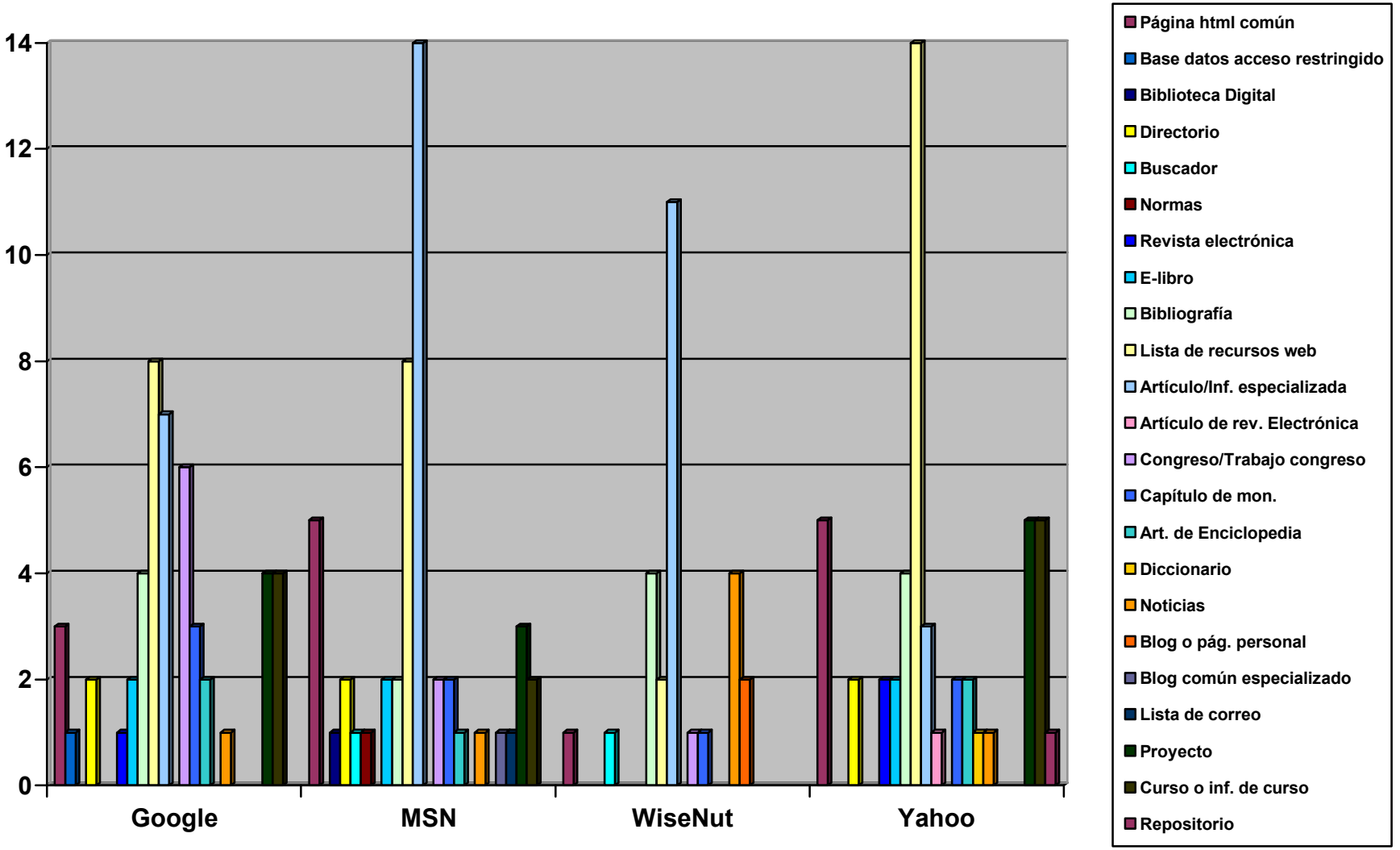


Tabla 3.3.1-18. Metabuscadores. Búsqueda 6. Tipología documental

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Página html común		4 (8%)	3 (13%)		2 (4,2%)		2 (4,2%)
Página html en blanco							
Página html en lenguas orientales							
Imagen							
Base de datos a texto completo libre							
Base datos acceso restringido			1 (4,3%)				
Base datos acceso a registros bibliogr.							
Biblioteca Digital				2 (8%)			
Repositorio				1 (4%)			
Directorio		2 (4%)					
Buscador		2 (4%)	1 (4,3%)	2 (8%)			
Agente de búsqueda							
Normas							
Lista de correo							
Revista electrónica					1 (2,1%)		
E-libro		1 (2%)	3 (13%)	1 (4%)	1 (2,1%)		
Presentación							
Bibliografía		5 (10%)	3 (13%)	3 (12%)	6 (12,5%)		3 (6,3%)
Lista de recursos web		5 (10%)	4 (17,4%)		6 (12,5%)		7 (14,6%)
Artículo/Inf. especializada		12 (24%)	3 (13%)	3 (12%)	17 (35,4%)		12 (25%)
Artículo de rev. Electrónica							
Congreso/Trabajo congreso		2 (4%)	1 (4,3%)	4 (16%)	1 (2,1%)		8 (16,7%)
Monografía							
Capítulo de mon.							3 (6,3%)
Art. de Enciclopedia		1 (2%)			4 (8,3%)		
Entrevista							
Diccionario							
Noticias		6 (12%)	2 (8,7%)		3 (6,3%)		3 (6,3%)
Blog o pág. personal		2 (4%)			2 (4,2%)		
Blog común especializado		4 (8%)					
Página registro							
Lista de correo		1 (2%)		2 (8%)			
Discurso							
Proyecto		2 (4%)	1 (4,3%)	1 (4%)	3 (6,3%)		3 (6,3%)

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Curso o inf. de curso		1 (2%)	1 (4,3%)	6 (24%)	2 (4,2%)		6 (12,5%)
Resumen							1 (2,1%)
Repositorio							
FAQ							
Normas							
Examen							
Registro							
Banco de datos							
Repositorio							

En relación con los metabuscadores, destaca en la recuperación de artículos por parte de Search, con diecisiete, seguido de Vivisimo y Excite, ambos con doce, superando en algunos casos las frecuencias ofrecidas por los motores de búsqueda.

Search se distingue por la recuperación de bibliografía sobre el tema y entradas de enciclopedias.

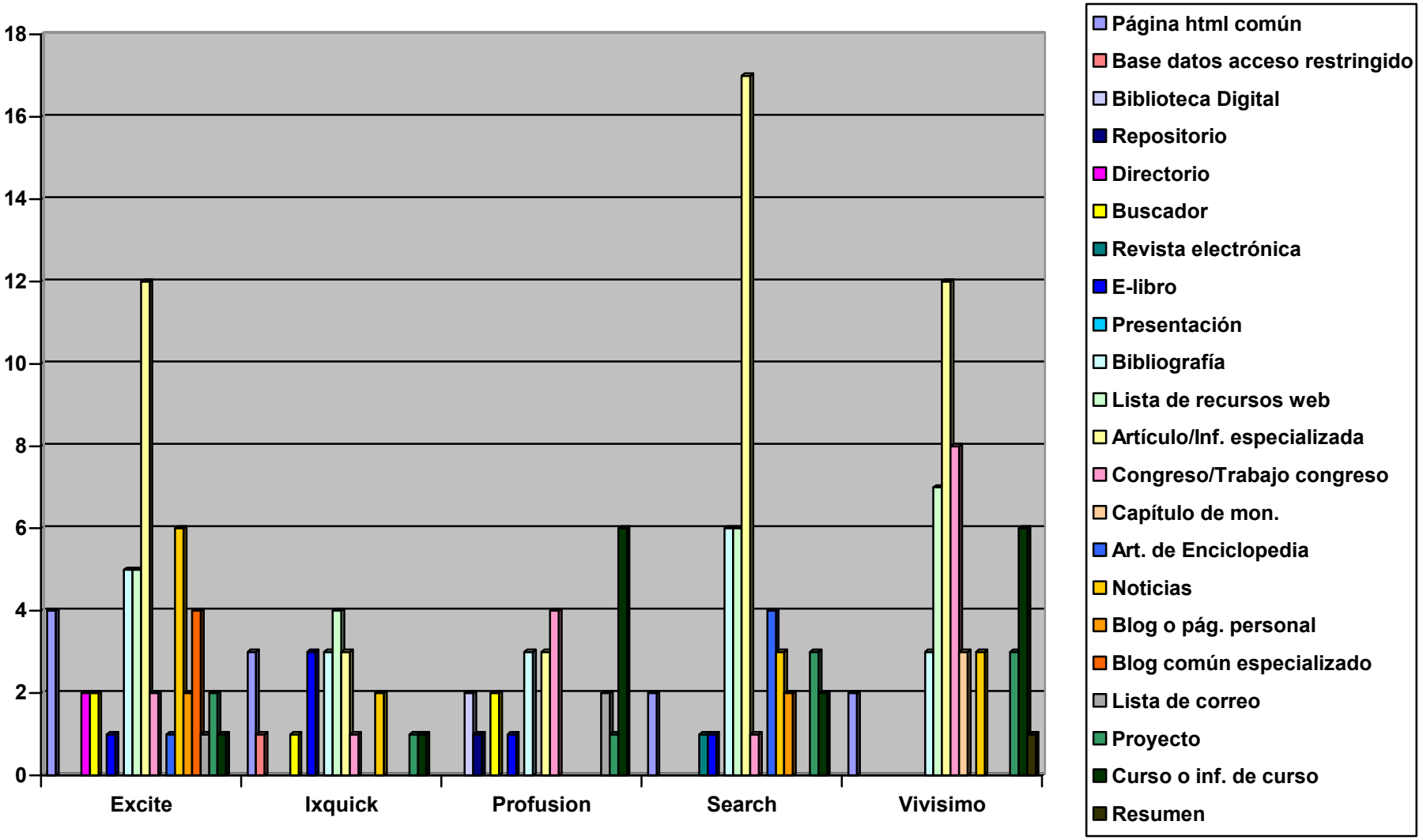
Vivisimo destaca en la recuperación de recursos relacionados con congresos, ocho y más ligeramente en los listados de recursos web. Por otro lado, es el único que recupera capítulos de monografías.

Excite tiene un comportamiento caracterizado por recuperar recursos de variada tipología documental, destacando en parte, por la recuperación de páginas en HTML, recursos bibliográficos, y acceso a noticias. Es el único metabuscador que recupera recursos ofrecidos por directorios de la web, y acceso a blogs especializados.

Profusión mantiene las características señaladas para Excite, si bien con menores frecuencias y destacando en recursos relacionados con bibliotecas digitales y acceso a repositorios. Con Vivisimo, tiene en común una buena recuperación de información sobre congresos y cursos.

Finalmente Ixquick destaca por ser el metabuscador que más libros electrónicos recupera.

Gráfico 3.3-16. Metabuscadores. Búsqueda 6. Tipología documental



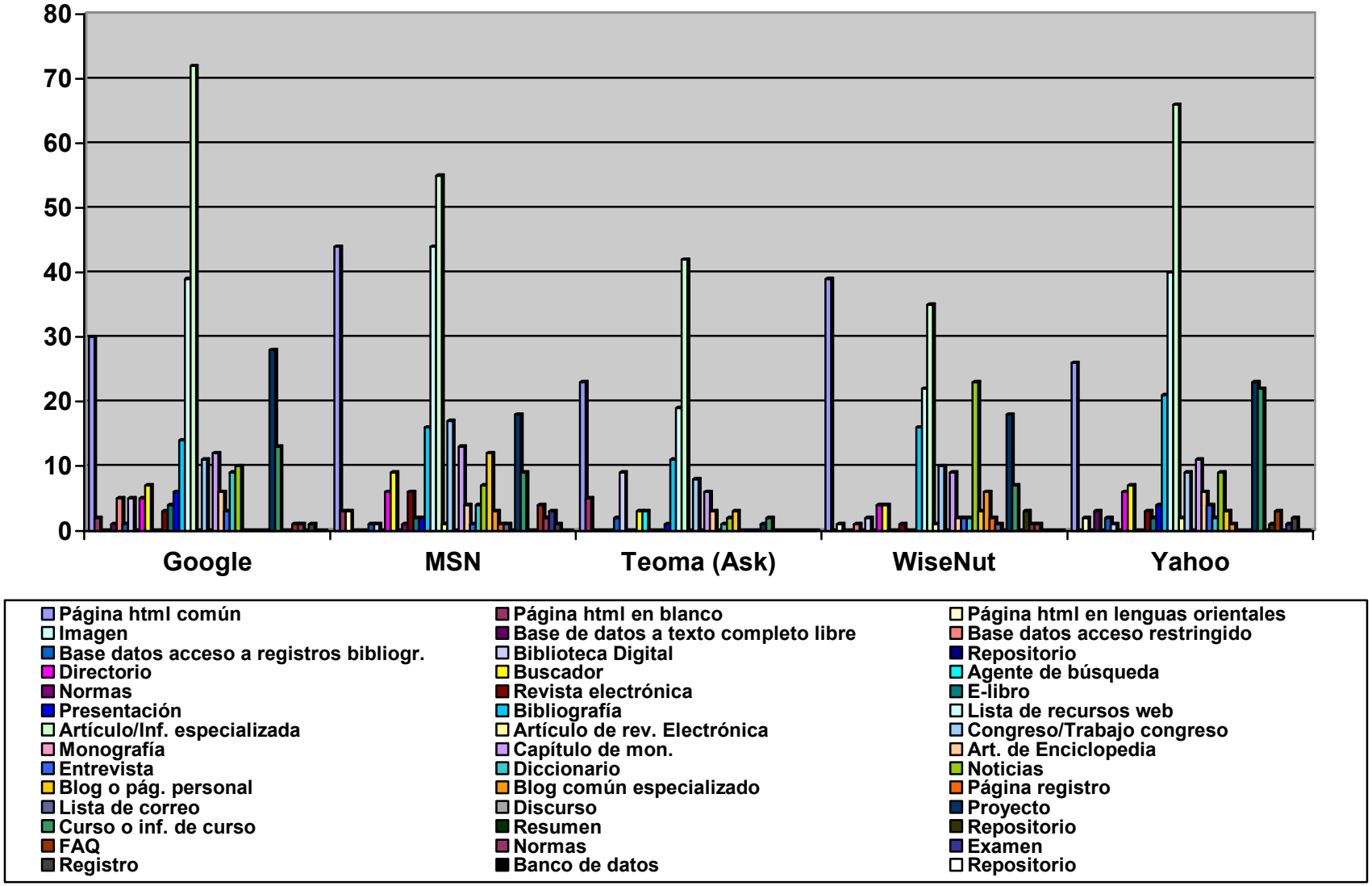
Análisis global de las seis búsquedas

Tabla 0-1 Motores. Tipología documental de las seis búsquedas

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Total
Página html común	30	44	23	39	26	162
Página html en blanco	2	3	5	0	0	10
Página html en lenguas orientales	0	3	0	1	2	6
Imagen	0	0	0	0	0	0
Base de datos a texto completo libre	1	0	0	0	3	4
Base datos acceso restringido	5	0	0	1	0	6
Base datos acceso a registros bibliogr.	1	1	2	0	2	6
Biblioteca Digital	5	1	9	2	1	18
Repositorio	0	0	0	0	0	0
Directorio	5	6	0	4	6	21
Buscador	7	9	3	4	7	30
Agente de búsqueda	0	0	3	0	0	3
Normas	0	1	0	0	0	1
Revista electrónica	3	6	0	1	3	13
E-libro	4	2	0	0	2	8
Presentación	6	2	1	0	4	13
Bibliografía	14	16	11	16	21	78
Lista de recursos web	39	44	19	22	40	164
Artículo/Inf. especializada	72	55	42	35	66	270
Artículo de rev. Electrónica	0	1	0	1	2	4
Congreso/Trabajo congreso	11	17	8	10	9	55
Monografía	0	0	0	0	0	0
Capítulo de mon.	12	13	6	9	11	51
Art. de Enciclopedia	6	4	3	2	6	21
Entrevista	3	1	0	2	4	10
Diccionario	9	4	1	2	2	18
Noticias	10	7	2	23	9	51
Blog o pág. personal	0	12	3	3	3	21
Blog común especializado	0	3	0	6	1	10
Página registro	0	1	0	2	0	3
Lista de correo	0	1	0	1	0	2
Discurso	0	0	0	0	0	0

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Total
Proyecto	28	18	1	18	23	88
Curso o inf. de curso	13	9	2	7	22	53
Resumen	0	0	0	0	0	0
Repositorio	0	0	0	3	1	4
FAQ	1	4	0	1	3	9
Normas	1	2	0	1	0	4
Examen	0	3	0	0	1	4
Registro	1	1	0	0	2	4
Banco de datos	0	0	0	0	0	0
Repositorio	0	0	0	0	0	0
Total	289	294	144	216	282	1225

Gráfico 3.3-17. Motores. Tipología documental de las seis búsquedas



Google es el buscador que recupera un mayor número de artículos de información especializada, seguido por las listas de recursos web, por las páginas HTML, por proyectos, cursos o información de cursos, documentación sobre congresos, noticias y finalmente acceso a diccionarios. Supera a Yahoo en recuperación de presentaciones y proyectos.

De forma similar a Google, MSN recupera, aunque en menor número, artículos y páginas HTML comunes, pero recupera un mayor número de recursos relacionados con revistas electrónicas, congresos y sobre todo facilita mayor acceso a blogs particulares. Sin embargo es inferior en la recuperación sobre cursos.

Yahoo sigue a Google en la obtención de artículos, listados de recursos web y proyectos de investigación, pero le supera en la recuperación de libros electrónicos e información de cursos.

Teoma destaca por ser el buscador que más accesos ofrece a bibliotecas digitales, pero también es el buscador que recupera mayor número de páginas en blanco, lo que supone un funcionamiento defectuoso.

WiseNut destaca en la recuperación de noticias, mostrándose aceptable en el resto de tipología documental.

Por tanto, teniendo en cuenta la tipología que puede ser más interesante en recuperación de recursos especializados, esto es artículos, bibliotecas digitales, revistas y libros electrónicos, información sobre congresos, artículos de enciclopedias y entradas de diccionario, blogs especializados y proyectos de investigación, el mejor comportamiento corresponde a Google, seguido de Yahoo y de MSN.

Tabla 0-2 Metabuscadores. Tipología documental de las seis búsquedas

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo	Total
Página html común	29	35	32	10	30	17	34	187
Página html en blanco	2	1	2	1	2	1	3	12
Página html en lenguas orientales	0	0	0	0	0	0	0	0
Imagen	1	1	0	0	0	1	0	3
Base de datos a texto completo libre	1	1	1	0	1	0	0	4
Base datos acceso restringido	0	0	2	0	1	1	0	4

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo	Total
Base datos acceso a registros bibliogr.	0	1	0	2	1	0	1	5
Biblioteca Digital	3	8	7	8	4	1	11	42
Repositorio	0	0	0	1	0	0	0	1
Directorio	1	16	11	5	4	8	0	45
Buscador	5	10	5	5	2	2	4	33
Agente de búsqueda	1	0	0	0	0	0	1	2
Normas	0	0	0	0	0	0	0	0
Revista electrónica	0	2	2	2	4	1	0	11
E-libro	0	1	3	1	1	0	1	7
Presentación	0	1	1	3	3	0	0	8
Bibliografía	7	18	13	14	24	3	11	90
Lista de recursos web	18	28	21	17	42	4	47	177
Artículo/Inf. especializada	26	74	41	34	56	7	72	310
Artículo de rev. Electrónica	0	0	1	2	1	1	0	5
Congreso/Trabajo congreso	3	4	5	7	13	2	20	54
Monografía	0	0	0	0	0	0	0	0
Capítulo de mon.	5	7	7	5	9	1	11	45
Art. de Enciclopedia	3	6	7	7	12	2	4	41
Entrevista	1	2	1	2	1	0	1	8
Diccionario	4	3	1	3	2	0	5	18
Noticias	3	10	8	5	10	13	0	49
Blog o pág. personal	7	5	3	9	12	1	13	50
Blog común especializado	2	6	3	6	3	3	3	26
Página registro	0	0	0	0	0	0	0	0
Lista de correo	0	2	0	3	0	0	0	5
Discurso	0	1	0	0	0	0	0	1
Proyecto	2	4	13	17	29	0	9	74
Curso o inf. de curso	1	9	3	8	12	1	13	47
Resumen	0	0	0	0	0	0	1	1
Repositorio	0	0	0	0	0	0	0	0
FAQ	0	3	2	2	4	1	3	15
Normas	0	0	0	0	0	0	1	1
Examen	1	1	1	0	2	0	0	5
Registro	0	0	0	0	0	0	0	0
Banco de datos	0	0	0	0	0	0	0	0
Repositorio	0	0	0	0	0	0	0	0
Total	126	260	196	179	285	71	269	1386

También entre los metabuscadores la tipología documental más común en este tipo de búsquedas corresponde a artículos con información especializada, a páginas comunes en HTML y a listas de recursos Web.

Excite destaca por ofrecer un mayor número de accesos a artículos especializados, así como a directorios y buscadores. Esto unido a la importancia de los recursos bibliográficos, hace que sea un metabuscador a tener en cuenta para recuperar obras de referencia.

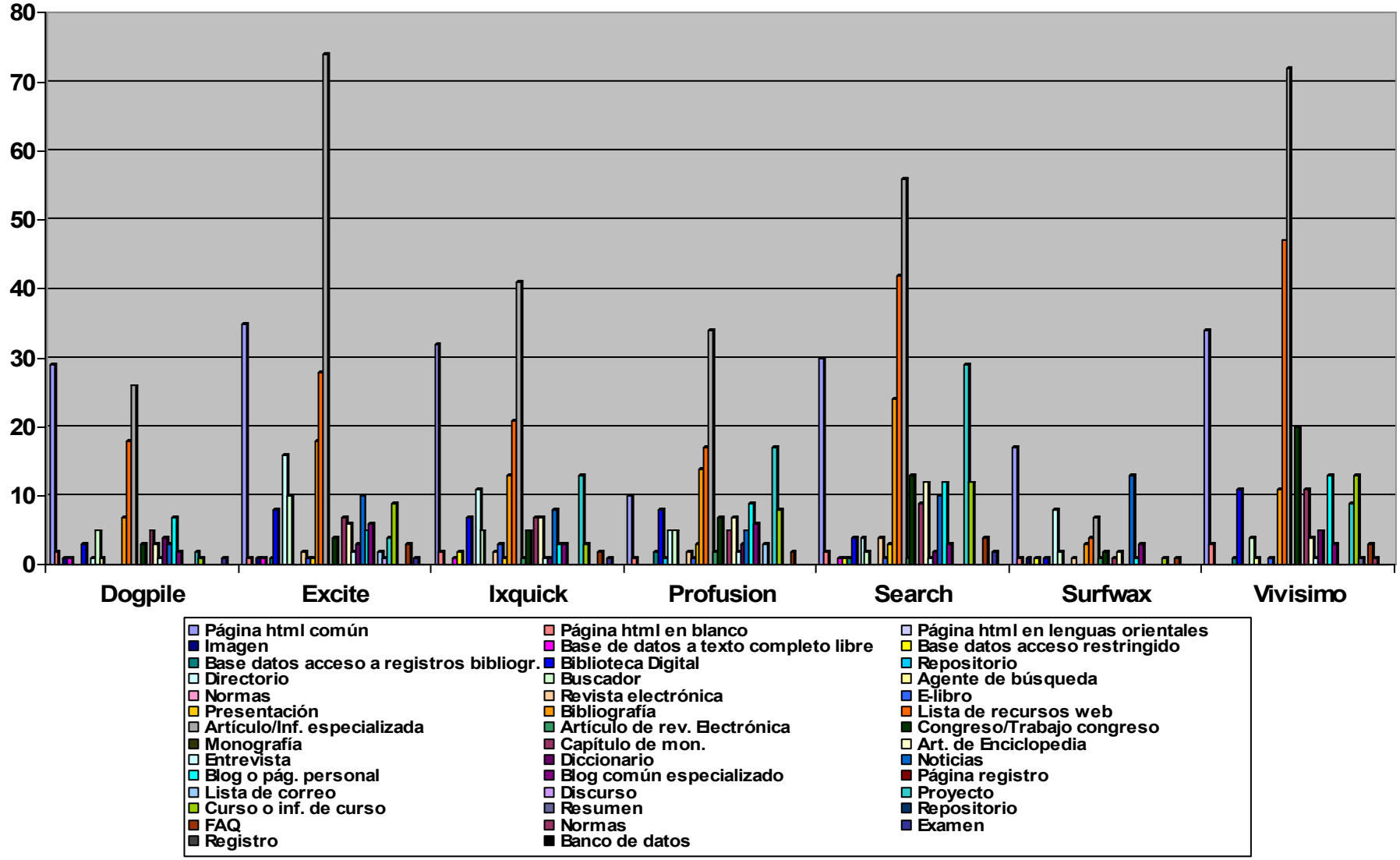
Vivisimo es el segundo metabuscador que más artículos recupera, destacando además por ser el que mayor número de accesos a bibliotecas digitales proporciona. Otros tipos documentales en los que sobresale son las listas de recursos, la información de congresos, acceso a capítulos de monografías, blogs y la información sobre cursos.

Search es el tercer metabuscador en recuperación de artículos. En el resto de tipología documental se puede comparar a Excite, pero le supera en accesos a revistas electrónicas, presentaciones del tipo PowerPoint, bibliografías, información de congresos, capítulos de monografías, artículos de enciclopedia, listas de recursos web, pero sobre todo en la recuperación de proyectos.

Ixquick supera al resto, en libros electrónicos. Profusion se asemeja a los mejores metabuscadores en cuanto a la recuperación de proyectos de investigación y acceso a bibliotecas digitales.

En resumen, Excite es el metabuscador que mejor combina la recuperación de recursos de interés directo, como puede ser el acceso a artículos de información especializada con el acceso a otros recursos de referencia. Search es el metabuscador que ofrece mayor cantidad de recursos de todo tipo, sin centrarse tanto en la recuperación de artículos especializados. Surfwax es el metabuscador que peores cifras alcanza en la recuperación de tipología documental propia de la información especializada, mientras que Vivisimo es un metabuscador que, ofreciendo un importante acceso a artículos especializados y a bibliotecas digitales, facilita la recuperación de otros tipos documentales de tipo más informativo, como puede ser la información de cursos, listas de recursos e información sobre Congresos.

Gráfico 3.3-18. Metabuscadores. Tipología documental de las seis búsquedas



4. Cobertura y solapamiento de los buscadores

4.1. Análisis de páginas únicas y solapamiento

Para valorar la cobertura entre buscadores se examina el total de páginas recuperadas por cada buscador en cada una de las búsquedas (generalmente cincuenta resultados, a excepción de los motores que en determinadas búsquedas no llegan a recuperar cincuenta resultados, o simplemente no funcionan ante una determinada búsqueda). En función de los listados de las URL y del motor que los recupera y se extraen los datos relativos a los recursos que recuperan bien sea un sólo motor, dos, tres y así hasta el total de buscadores evaluados.

Tabla 4.1-1. Motores de búsqueda y metabuscadores (Base= total excluidos duplicados y errores)

	Búsqueda 1 (Término único)		Búsqueda 2 (Lenguaje natural)		Búsqueda 3 (Operadores de existencia)		Búsqueda 4 (Búsqueda booleana)		Búsqueda 5 (Búsqueda de frase)		Búsqueda 6 (Búsqueda por campo)		Total en las seis búsquedas	
Páginas recuperadas	Frec.	%	Frec.	%	Frec.	%	Frec.	%	Frec.	%	Frec.	%	Frec.	%
1 buscador	113	55,1	161	66,3	212	74,1	160	72,4	137	65,9	167	71,1	950	47,5
2 buscadores	26	12,6	26	10,7	28	9,8	33	14,9	26	12,5	45	19,1	184	9,2
3 buscadores	19	9,2	14	5,8	16	5,6	20	9,0	15	7,2	8	3,4	92	4,6
4 buscadores	12	5,8	12	4,9	7	2,4	6	2,7	8	3,8	3	1,3	48	2,4
5 buscadores	15	7,3	16	6,6	12	4,2	0	0	6	2,9	4	1,7	53	2,6
6 buscadores	11	5,3	6	2,5	6	2,1	1	0,5	4	1,9	3	1,3	31	1,5
7 buscadores	2	0,9	3	1,2	4	1,4	1	0,5	5	2,4	5	2,1	20	1
8 buscadores	6	2,9	3	1,2	1	0,3	0	0	6	2,9	0	0	16	0,7
9 buscadores	0	0	0	0	0	0	0	0	1	0,5	0	0	1	0,05
10 buscadores	0	0	1	0,4	0	0	0	0	0	0	0	0	1	0,05
11 buscadores	0	0	1	0,4	0	0	0	0	0	0	0	0	1	0,05
12 buscadores	1	0,4	0	0	0	0	0	0	0	0	0	0	1	0,05
Total páginas únicas	205	100	243	100	286	100	221	100	208	100	235	100	1998	100
Total pág. potencialmente relevantes	485		485		476		323		412		366		2547	

Una vez suprimidas las páginas repetidas en cada motor y aquellas a las que no se pudo acceder (enlaces rotos), contamos con un total de 2.547 páginas útiles de las que 1998 (78,4%) son páginas únicas, es decir aparecen una sola vez. A partir de estas cifras, podemos calcular el porcentaje de solapamiento que corresponde a los recursos cuya

URL aparece más de una vez. Así pues, el solapamiento estimado para todos los temas de búsqueda es el 21,6%.

Por otro lado tenemos que casi el 47,5% de los recursos es recuperado por un buscador. En otras experiencias similares, como la llevada a cabo por Bar-Ilan (1998), el porcentaje de recursos recuperados por un buscador fue del 75%, siendo en nuestro caso mayor el solapamiento en función de la utilización de metabuscadores en la evaluación, ya que la utilización de las bases de datos de los motores da como resultado la recuperación de recursos compartidos por ambos tipos de buscadores.

Gráfico 4.1-1. Solapamiento

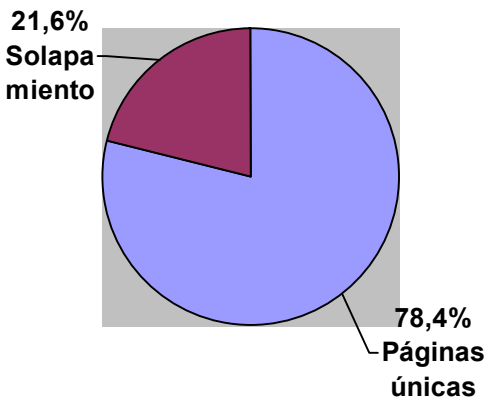


Tabla 4.1.-2. Total páginas únicas

Páginas potencialmente relevantes	2.547
Páginas únicas	1.998

No obstante este solapamiento es inferior al observado en trabajos anteriores, en los que se obtuvo un porcentaje del 24,6%. (Salvador y Vidal, 2000).

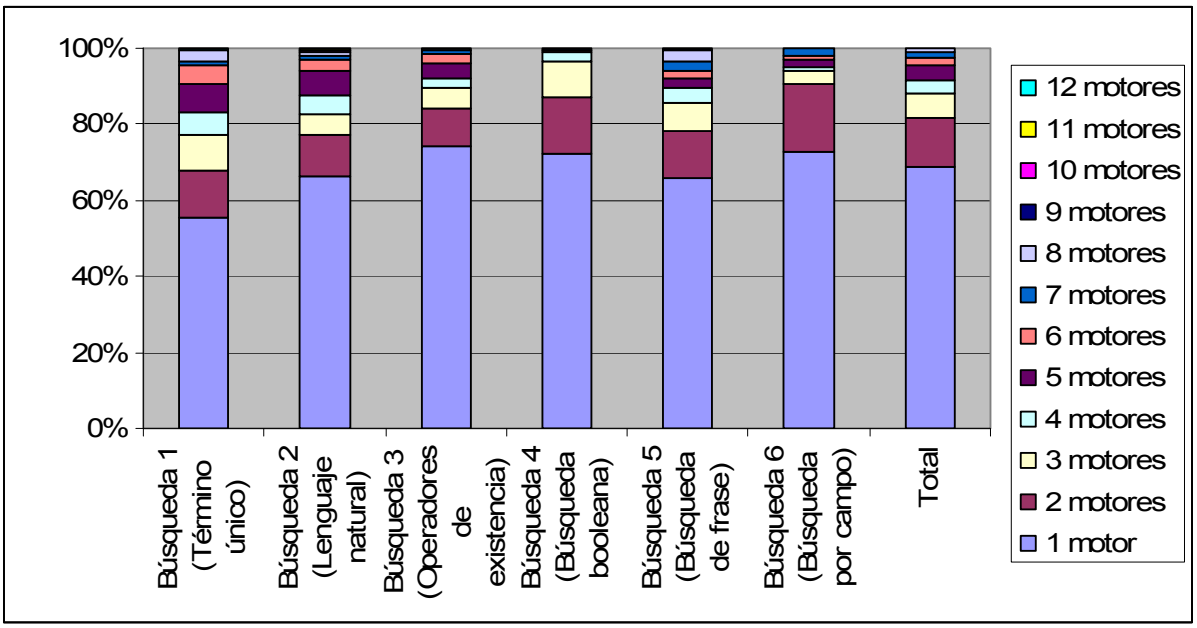
De las 1998 páginas únicas, 950 (47,5%) fueron recuperadas por un único motor de búsqueda, 184 (9,2%) páginas fueron recuperadas por dos motores y 92 (4,6%) por tres motores.

Las búsquedas con menor solapamiento son, por este orden la booleana y la búsqueda por campo. El menor solapamiento se observa en la búsqueda con operadores de existencia.

En la búsquedas por término único, sólo uno de los resultados fue recuperado por los doce buscadores. En la búsqueda que utiliza lenguaje natural, once de ellos recupera-

ron el mismo recurso. Otro fue recuperado por diez. En la búsqueda por frase, fueron nueve los motores que recuperaron un mismo recurso.

Gráfico 4.1-2. Registros únicos en las diferentes búsquedas



El gráfico 4.1-2 recoge los resultados analizados por porcentajes, y permite apreciar que el mayor índice de recursos recuperados por un solo motor corresponde a las búsquedas en las que se utilizan operadores de existencia, seguida de la búsqueda bolean y la búsqueda por campo.

Por otro lado, vemos que el porcentaje de recurso compartidos por dos buscadores es muy similar en todas las búsquedas, aunque es algo mayor en la búsqueda booleana y por campo. Finalmente, el porcentaje de recursos recuperados por tres motores es mayor en las búsquedas booleana y en la que se utilizó únicamente un término.

Análisis global

El solapamiento observado entre los buscadores evaluados (21,6%) resulta elevado, en comparación con otros trabajos de evaluación llevados a cabo con anterioridad, pero dado que en ellos no se utilizaron metabuscadores, este valor da idea de la relación entre las bases de datos de estas herramientas. Así mismo se observa que los motores recuperan un mayor número de recursos únicos en las búsquedas por campo y con operadores de existencia.

4.1.1. Páginas únicas por motor de búsqueda

La distribución de páginas únicas por buscador y búsqueda aparece recogida en las siguientes tablas. Esto nos permite conocer cuáles son los buscadores que recuperan un mayor número de páginas únicas, la influencia en este sentido del tipo de búsqueda y si los datos son constantes.

Tabla 4.1.1-1. Motores. Páginas únicas (Base=Excluidos duplicados y errores)

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Google	23 (20,4%)	22 (13,7%)	28 (13,2%)	20 (12,5%)	18 (3,1%)	15 (8,4%)	126 (13%)
MSN	17 (14,7%)	23 (14,3%)	35 (16,5%)	31 (19,4%)	14 (10,2%)	26 (14,5%)	146 (15,1%)
Teoma (Ask)	16 (13,8%)	24 (14,9%)	31 (14,6%)	Sin resultados	Sin resultados	Sin resultados	71 (7,3%)
WiseNut	25 (21,6%)	31 (19,5%)	30 (14,2%)	Sin resultados	13 (9,5%)	15 (8,4%)	114 (11,8%)
Yahoo	13 (11,2%)	26 (16,1%)	29 (13,7%)	26 (16,3%)	22 (16,1%)	24 (13,4%)	140 (14,5%)
Total	**94 (100%)**	**126 (100%)**	**153 (100%)**	**77 (100%)**	**67 (100%)**	**80 (100%)**	**597 (100%)**

Tabla 4.1.1-2. Metabuscadores. Páginas únicas (Base=Excluidos duplicados y errores)

	Búsqueda 1 (Término único)	Búsqueda 2 (Lenguaje natural)	Búsqueda 3 (Operadores de existencia)	Búsqueda 4 (Búsqueda booleana)	Búsqueda 5 (Búsqueda de frase)	Búsqueda 6 (Búsqueda por campo)	Total
Dogpile	3 (2,6%)	8 (5,0%)	2 (0,9%)	Sin resultados	Sin resultados	Sin resultados	13 (1,3%)
Excite	11 (9,5%)	10 (6,2%)	11 (5,2%)	20 (12,5%)	14 (10,2%)	27 (15,1%)	93 (9,6%)
Ixquick	2 (1,7%)	3 (1,9%)	11 (5,2%)	2 (1,3%)	8 (5,8%)	4 (2,2%)	30 (3,1%)
Profusion	1 (0,9%)	3 (1,9%)	Sin resultados	15 (9,4%)	8 (5,8%)	22 (12,3%)	49 (5%)
Search	0	1 (0,6%)	6 (2,8%)	3 (1,9%)	6 (4,4%)	19 (10,6%)	35 (3,6%)
Surfwax	0	7 (4,3%)	18 (8,5%)	Sin resultados	Sin resultados	Sin resultados	25 (2,5%)
Vivisimo	2 (1,7%)	3 (1,9%)	11 (5,2%)	43 (26,9%)	34 (24,8%)	27 (15,1%)	120 (12,4%)
Total	**19 (100%)**	**35 (100%)**	**59 (100%)**	**83 (100%)**	**70 (100%)**	**99 (100%)**	**365 (100%)**

Tabla 4.1.1-3. Total de páginas únicas por búsqueda

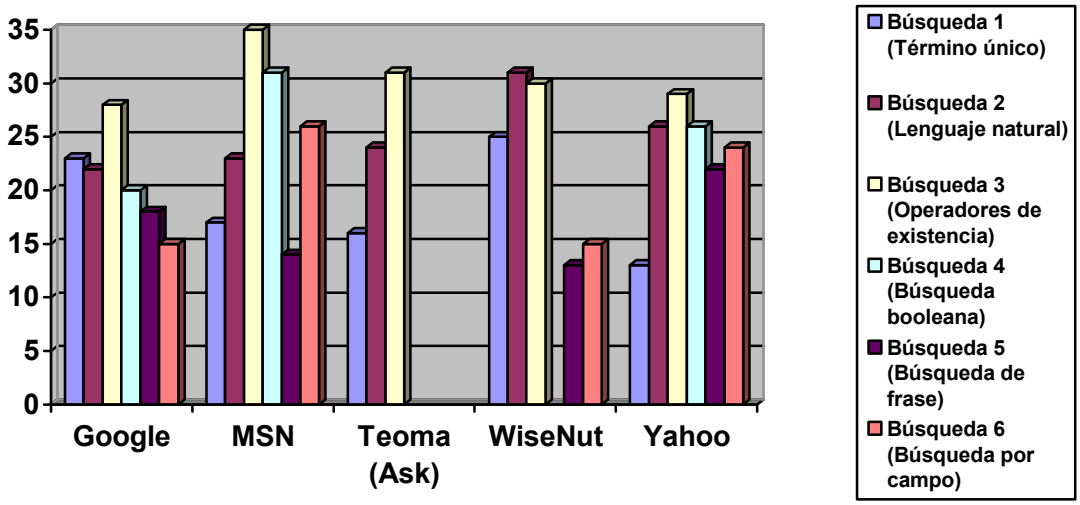
Total	**113 (100%)**	**161 (100%)**	**212 (100%)**	**160 (100%)**	**137 (100%)**	**179 (100%)**	**962 (100%)**

Esta última tabla, que suma los resultados de las dos tablas anteriores, muestra que la búsqueda que ofrece un mayor número de páginas únicas (212), recuperadas por un sólo buscador es la que utiliza los operadores de existencia, seguida por la búsqueda por campo con 179. En tercer lugar está la búsqueda de varios términos con 161 páginas únicas, seguida de la búsqueda booleana con 160. Las menores cifras de recursos únicos, es

decir donde hay más solapamiento corresponden a la búsqueda por frase (137) y a la búsqueda por término único (113).

Por otro lado, si observamos en las tablas anteriores las cifras totales por buscador, podemos observar que el motor de búsqueda que más páginas únicas recupera es MSN con 146, seguido por Yahoo con 140. En tercer lugar aparece Google con 126, y en cuarto y quinto puesto están WiseNut y Teoma.

Gráfico 4.1-3. Motores. Registros únicos por búsqueda



Atendiendo al gráfico que muestra los resultados alcanzados por cada buscador en las seis búsquedas podemos observar que en la búsqueda por un término, a la que corresponde la barra azul, WiseNut recupera el mayor número de recursos únicos (25), seguido por Google con 23. Los motores con menor número de recursos únicos en ésta búsqueda son por este orden Yahoo, Teoma y MSN, siendo estos a los que corresponde un mayor solapamiento.

En la segunda búsqueda, representada por la barra color granate, sigue siendo WiseNut el motor con mayor número de páginas únicas (31), seguido por Yahoo con 26, correspondiendo el mayor solapamiento a Google, MSN y Teoma.

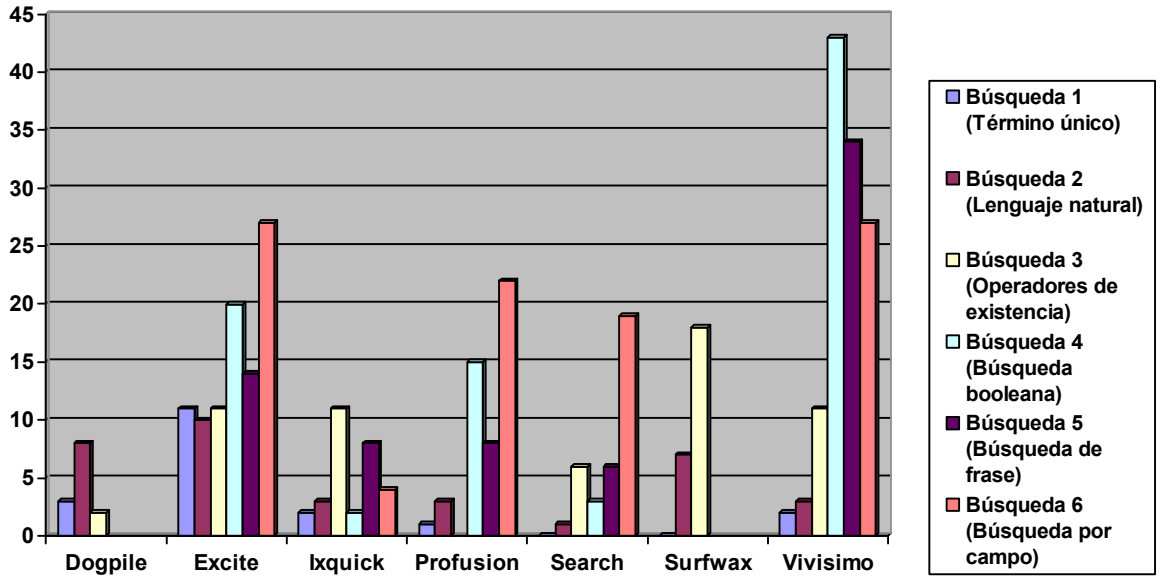
En la búsqueda con operadores de existencia, es MSN el motor que más recursos únicos recupera (35), seguido por Teoma y WiseNut. El mayor solapamiento lo registran Google y Yahoo.

En la cuarta búsqueda, MSN con 31 recursos únicos es de nuevo el buscador con más recursos únicos; le sigue Yahoo con 26 y Google con 20 que pasa a ser en este caso el que registra mayor solapamiento.

En la búsqueda por frase, es Yahoo con 22 el motor con más recursos únicos, seguido por Google con 18 y MSN con 14. WiseNut recuperó trece recursos únicos, siendo estos últimos a los que corresponde un mayor solapamiento.

La búsqueda por campo, sitúa de nuevo a MSN con 26 recursos únicos en primer lugar, seguido de Yahoo con 24 y WiseNut y Google con 15 son los que registran un mayor solapamiento.

Gráfico 4.1-4. Metabuscadores. Registros únicos por búsqueda

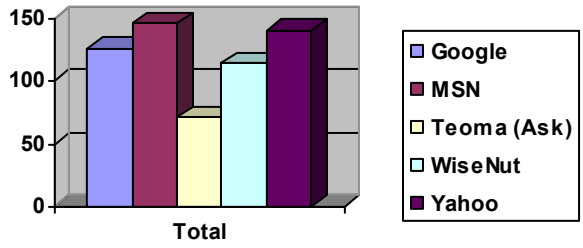


En los metabuscadores, podemos observar tanto en Vivisimo como en Excite, Profusion y Search, un mayor número de recursos únicos en las búsquedas avanzadas. Los datos más destacados corresponden a la búsqueda booleana, en la que Vivisimo recupera 43 recursos únicos, a la búsqueda por frase, con 34 y a la búsqueda por campo con 27. Los metabuscadores con mayor solapamiento en estas búsquedas son Ixquick y Search.

Análisis global

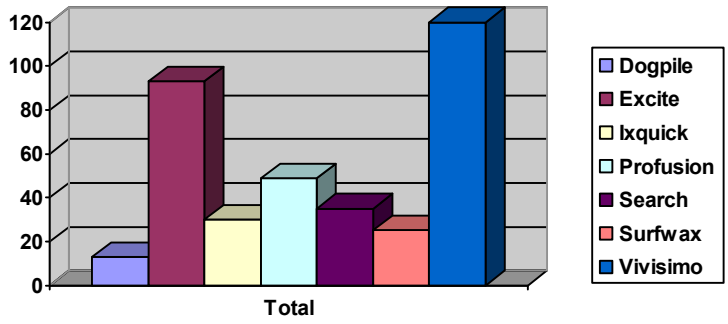
Las búsquedas con mayor número de páginas únicas son la búsqueda con operadores de existencia y la búsqueda por campo, existiendo un mayor solapamiento en las búsquedas por un término y en la búsqueda por frase.

Gráfico 4.1-5. Buscadores. Total páginas únicas



Como podemos apreciar en el gráfico MSN y Yahoo son los motores que más páginas únicas recuperan, correspondiendo a Google el mayor solapamiento.

Gráfico 4.1-6. Metabuscadores. Total páginas únicas



Entre los metabuscadores, a pesar del descenso de páginas únicas respecto a las recuperadas por los buscadores, Vivisimo y Excite son los que más se aproximan a las cifras presentadas por aquellos. Constituyen por tanto estos metabuscadores un buen complemento a utilizar en combinación con los motores en los que se registra menor solapamiento en búsquedas en las que se requiere un alto grado de exhaustividad.

4.1.2. Solapamiento entre buscadores. Análisis por búsquedas

La presente tabla así como las tablas siguientes recogen el número de páginas en que coinciden los distintos buscadores en cada una de las búsquedas. Hemos colocado en

las primeras columnas a los motores seguidos de los metabuscadores para valorar primero el solapamiento entre los motores y a continuación con los metabuscadores.

El porcentaje se calcula sobre el total de URL pertenecientes a los recursos recuperados por dos motores, una vez eliminadas las páginas inactivas y enlaces duplicados.

4.1.2.1. Búsqueda de un término

Tabla 4.1.2-1. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Google	-	4 (1,9%)	4 (1,9%)	4 (1,9%)	15 (7,2%)	12 (5,8%)	17 (8,2%)	8 (3,8%)	9 (4,3%)	13 (6,3%)	3 (1,4%)	6 (2,9%)
MSN	4 (1,9%)	-	3 (1,4%)	7 (3,4%)	4 (1,9%)	18 (8,7%)	3 (1,4%)	6 (2,9%)	10 (4,8%)	13 (6,3%)	4 (1,9%)	18 (8,7%)
Teoma (Ask)	4 (1,9%)	3 (1,4%)	-	2 (1%)	8 (3,8%)	10 (4,8%)	13 (6,3%)	8 (3,8%)	10 (4,8%)	8 (3,8%)	2 (1%)	20 (9,6%)
WiseNut	4 (1,9%)	7 (3,4%)	2 (1%)	-	2 (1%)	9 (4,3%)	3 (1,4%)	11 (5,3%)	10 (4,8%)	16 (7,7%)	14 (6,7%)	8 (3,8%)
Yahoo	15 (7,2%)	4 (1,9%)	8 (3,8%)	2 (1%)	-	14 (6,7%)	22 (10,6%)	10 (4,8%)	13 (6,3%)	18 (8,7%)	1 (0,5%)	9 (4,3%)
Dogpile	12 (5,8%)	18 (8,7%)	10 (4,8%)	9 (4,3%)	14 (6,7%)	-	21 (10,1%)	14 (6,7%)	18 (8,7%)	28 (13,5%)	4 (1,9%)	30 (14,4%)
Excite	17 (8,2%)	3 (1,4%)	13 (6,3%)	3 (1,4%)	22 (10,6%)	21 (10,1%)	-	17 (8,2%)	15 (7,2%)	20 (9,6%)	3 (1,4%)	13 (6,3%)
Ixquick	8 (3,8%)	6 (2,9%)	8 (3,8%)	11 (5,3%)	10 (4,8%)	14 (6,7%)	17 (8,2%)	-	14 (6,7%)	22 (10,6%)	11 (5,3%)	9 (4,3%)
Profusion	9 (4,3%)	10 (4,8%)	10 (4,8%)	10 (4,8%)	13 (6,3%)	18 (8,7%)	15 (7,2%)	14 (6,7%)	-	25 (12%)	10 (4,8%)	16 (7,7%)
Search	13 (6,3%)	13 (6,3%)	8 (3,8%)	16 (7,7%)	18 (8,7%)	28 (13,5%)	20 (9,6%)	22 (10,6%)	25 (12%)	-	14 (6,7%)	17 (8,2%)
Surfwax	3 (1,4%)	4 (1,9%)	2 (1%)	14 (6,7%)	1 (0,5%)	4 (1,9%)	3 (1,4%)	11 (5,3%)	10 (4,8%)	14 (6,7%)	-	3 (1,4%)
Vivisimo	6 (2,9%)	18 (8,7%)	20 (9,6%)	8 (3,8%)	9 (4,3%)	30 (14,4%)	13 (6,3%)	9 (4,3%)	16 (7,7%)	17 (8,2%)	3 (1,4%)	-

Entre los motores, en la búsqueda por un término, el mayor solapamiento de Google se da con Yahoo, con el que coincide en la recuperación de quince páginas (7,2%), mientras que con el resto de buscadores coincide sólo en cuatro resultados. MSN tiene su mayor coincidencia con WiseNut, con el que comparte 7 recursos (3,4%). Teoma coincide con Yahoo en 8 resultados (3,8%) y tan sólo en 2 con WiseNut. WiseNut muestra su mayor solapamiento con MSN y en menor medida, como acabamos de ver con Teoma y Yahoo.

En resumen y atendiendo a los buscadores, podemos apreciar que el menor solapamiento se da entre WiseNut, Teoma y Yahoo (1%), seguido del registrado entre MSN y Teoma (3%). En tercer lugar hemos de señalar a MSN, Google y Teoma con un solapa-

miento entre ellos del 4%, correspondiendo el mayor solapamiento a Yahoo y Google (7,2%).

Con los metabuscadores, el solapamiento de los motores de búsqueda es muy variable ya que por ejemplo Yahoo tiene un solapamiento bajo con Surfwax (0,5%), pero con Excite asciende al 10,6% y con Search al 8,7%. De nuevo vuelve a ser bajo entre WiseNut y MSN con Excite (1,4%), sin embargo entre Google y Excite es del 8,2%, y entre MSN y Teoma con Vivisimo, oscila entre el 8,7% y el 9,6% respectivamente.

Por otro lado, podemos observar que los indicadores del solapamiento entre metabuscadores son más elevados, ya que por ejemplo entre Vivisimo y Dogpile hay un solapamiento del 14,4%. También se puede apreciar que el metabuscador con más solapamiento con el resto es Search, con cifras entre el 6,7% y el 13,5%. El menor solapamiento corresponde a Surfwax.

4.1.2.2. Búsqueda utilizando el lenguaje natural

Tabla 4.1.2-2. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Google	-	8 (3,3%)	4 (1,6%)	2 (0,8%)	6 (2,5%)	7 (2,9%)	10 (4,1%)	8 (3,3%)	6 (2,5%)	13 (5,3%)	1 (0,4%)	6 (2,5%)
MSN	8 (3,3%)	-	4 (1,6%)	1 (0,4%)	9 (3,7%)	5 (2,1%)	7 (2,9%)	4 (1,6%)	10 (4,1%)	15 (6,2%)	1 (0,4%)	16 (6,6%)
Teoma (Ask)	4 (1,6%)	4 (1,6%)	-	1 (0,4%)	6 (2,5%)	11 (4,5%)	12 (4,9%)	12 (4,9%)	14 (5,8%)	5 (2,1%)	1 (0,4%)	16 (6,6%)
WiseNut	2 (0,8%)	1 (0,4%)	1 (0,4%)	-	2 (0,8%)	1 (0,4%)	1 (0,4%)	10 (4,1%)	1 (0,4%)	13 (5,3%)	10 (4,1%)	16 (6,6%)
Yahoo	6 (2,5%)	9 (3,7%)	6 (2,5%)	2 (0,8%)	-	10 (4,1%)	12 (4,9%)	9 (3,7%)	9 (3,7%)	13 (5,3%)	1 (0,4%)	9 (3,7%)
Dogpile	7 (2,9%)	5 (2,1%)	11 (4,5%)	1 (0,4%)	10 (4,1%)	-	27 (11,1%)	17 (7%)	13 (5,3%)	14 (5,8%)	1 (0,4%)	11 (4,5%)
Excite	10 (4,1%)	7 (2,9%)	12 (4,9%)	1 (0,4%)	12 (4,9%)	27 (11,1%)	-	21 (8,6%)	18 (7,4%)	22 (9,1%)	1 (0,4%)	14 (5,8%)
Ixquick	8 (3,3%)	4 (1,6%)	12 (4,9%)	10 (4,1%)	9 (3,7%)	17 (7%)	21 (8,6%)	-	15 (6,2%)	23 (9,5%)	9 (3,7%)	20 (8,2%)
Profusion	6 (2,5%)	10 (4,1%)	14 (5,8%)	1 (0,4%)	9 (3,7%)	13 (5,3%)	18 (7,4%)	15 (6,2%)	-	12 (4,9%)	1 (0,4%)	14 (5,8%)
Search	13 (5,3%)	15 (6,2%)	5 (2,1%)	13 (5,3%)	13 (5,3%)	14 (5,8%)	22 (9,1%)	23 (9,5%)	12 (4,9%)	-	10 (4,1%)	28 (11,5%)
Surfwax	1 (0,4%)	1 (0,4%)	1 (0,4%)	10 (4,1%)	1 (0,4%)	1 (0,4%)	1 (0,4%)	9 (3,7%)	1 (0,4%)	10 (4,1%)	-	10 (4,1%)
Vivisimo	6 (2,5%)	16 (6,6%)	16 (6,6%)	16 (6,6%)	9 (3,7%)	11 (4,5%)	14 (5,8%)	20 (8,2%)	14 (5,8%)	28 (11,5%)	10 (4,1%)	-

La búsqueda por varios términos disminuye los porcentajes de solapamiento respecto de la búsqueda anterior. No obstante hemos de señalar que Yahoo mantiene los porcentajes más altos de solapamiento con el resto de buscadores, a excepción de WiseNut (0,8%), al que corresponden los índices menores de solapamiento en esta búsqueda. MSN tiene mayor solapamiento con Yahoo (3,7%) y Google (3,3%), siendo entre estos tres

buscadores entre los que se aprecia mayor solapamiento en esta búsqueda, aunque hay que advertir que no es muy elevado pues como hemos visto no supera el 3,7%.

El solapamiento con los metabuscadores es alto en el caso de MSN, Teoma y WiseNut con Vivisimo, ya que el solapamiento con ellos es del 6,6%. Google tiene con Search el mayor solapamiento (5,3%) y con Surfwax el menor (0,4%). Algo similar le ocurre a MSN, teniendo en cuenta que el mayor solapamiento, como acabamos de ver es con Vivisimo (6,6%). Teoma muestra un elevado solapamiento con la mayoría de metabuscadores, a excepción, una vez más, de Surfwax. WiseNut es el caso más desigual, ya que mantiene con Dogpile, Excite y Profusión porcentajes muy bajos de solapamiento (0,4%), mientras que con Surfwax, Search y Vivisimo los porcentajes van del 4,1% al 6,6%.

El solapamiento entre metabuscadores destaca en el caso de Search y Vivisimo (11,5%) y de Dogpile y Excite (11%). El menor solapamiento se da entre Surfwax, Dogpile y Profusion (0,4).

Vivisimo, que suele tener mucho solapamiento tanto con buscadores como con metabuscadores, con Google y Yahoo tiene bajo solapamiento, lo que puede ser una buena opción para utilizar de forma combinada en búsquedas de este tipo.

4.1.2.3. Búsqueda con operadores de existencia

Tabla 4.1.2-3. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Google	-	7 (2,4%)	6 (2,1%)	3 (1%)	5 (1,7%)	12 (4,2%)	8 (2,8%)	8 (2,8%)	(Sin resultados)	15 (5,2%)	0	8 (2,8%)
MSN	7 (2,4%)	-	2 (0,7%)	3 (1%)	2 (0,7%)	6 (2,1%)	3 (1%)	4 (1,4%)		11 (3,8%)	1 (0,3%)	9 (3,1%)
Teoma (Ask)	6 (2,1%)	2 (0,7%)	-	1 (0,3%)	3 (1%)	11 (3,8%)	7 (2,4%)	5 (1,7%)		6 (2,1%)	1 (0,3%)	9 (3,1%)
WiseNut	3 (1%)	3 (1%)	1 (0,3%)	-	3 (1%)	3 (1%)	3 (1%)	5 (1,7%)		12 (4,2%)	5 (1,7%)	9 (3,1%)
Yahoo	5 (1,7%)	2 (0,7%)	3 (1%)	3 (1%)	-	6 (2,1%)	4 (1,4%)	5 (1,7%)		10 (3,5%)	0	5 (1,7%)
Dogpile	12 (4,2%)	6 (2,1%)	11 (3,8%)	3 (1%)	6 (2,1%)	-	22 (7,7%)	17 (5,9%)		20 (7%)	6 (2,1%)	15 (5,2%)
Excite	8 (2,8%)	3 (1%)	7 (2,4%)	3 (1%)	4 (1,4%)	22 (7,7%)	-	19 (6,6%)		16 (5,6%)	5 (1,7%)	14 (4,9%)
Ixquick	8 (2,8%)	4 (1,4%)	5 (1,7%)	5 (1,7%)	5 (1,7%)	17 (5,9%)	19 (6,6%)	-		14 (4,9%)	13 (4,5%)	17 (5,9%)
Profusion	(Sin resultados)											

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Search	15 (5,2%)	11 (3,8%)	6 (2,1%)	12 (4,2%)	10 (3,5%)	20 (7%)	16 (5,6%)	14 (4,9%)		-	4 (1,4%)	20 (7%)
Surfwax	0	1 (0,3%)	1 (0,3%)	5 (1,7%)	0	6 (2,1%)	5 (1,7%)	13 (4,5%)		4 (1,4%)	-	12 (12,2%)
Vivisimo	8 (2,8%)	9 (3,1%)	9 (3,1%)	9 (3,1%)	5 (1,7%)	15 (5,2%)	14 (4,9%)	17 (5,9%)		20 (7%)	12 (4,2%)	-

En esta búsqueda el solapamiento desciende, correspondiendo los mayores índices a Google, con valores entre el 1% con WiseNut y el 2,4% con MSN. El resto de valores es muy similar a los proporcionados en la búsqueda anterior.

Los metabuscadores con mayor solapamiento con los motores de búsqueda son Search, Vivisimo y Dogpile. Menor solapamiento se aprecia, por este orden, en Surfwax, Excite e Ixquick, lo que teniendo en cuenta otros aspectos del funcionamiento de estos metabuscadores, puede suponer una alternativa de búsqueda, en casos determinados como puede ser el interés en una mayor exhaustividad. En este caso es interesante conocer las combinaciones más oportunas.

Así, corresponden a Google los valores más altos de solapamiento, ya que con Search se alcanza el 5,2% y con Dogpile el 4,2%. Con el resto de metabuscadores se mantiene un porcentaje del 2,8%, a excepción de Surfwax, con el que Google no registra solapamiento. MSN registra el mayor solapamiento con Search (3,8%) y con Vivisimo (3,1%), y la menor coincidencia de resultados corresponde a Surfwax (0,3%). Teoma tiene el mayor solapamiento con Dogpile (3,8%) y Vivisimo (3,1%). WiseNut, al igual que en las búsquedas anteriores, tiene un elevado solapamiento con Search (4,2%), y con Vivisimo (3,1%). Finalmente Yahoo, sigue manteniendo el mayor solapamiento con Search (3,5%), y como le ocurre a Google, no hay solapamiento en esta búsqueda con Surfwax.

Entre los propios metabuscadores, Dogpile mantiene una alto porcentaje con Excite (7,7%) y con Search (7%). Excite, al margen de este aspecto, se solapa de forma similar con los demás metabuscadores, salvo con Surfwax. La misma característica corresponde a Ixquick, cuyos índices son similares, pero en este caso, incluso con Surfwax.

Como hemos señalado, Search tiene un importante solapamiento con Dogpile (7%) y con Vivisimo. Surfwax con los que más se solapa es con Ixquick (4,5%) y con Vivisi-

mo (4,2%). El solapamiento de Vivisimo oscila ligeramente entre el 4,2% de Surfwax y el 7% de Search.

4.1.2.4. Búsqueda booleana

Tabla 4.1.2-4. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask) (Sin resulta-dos)	WiseNut (Sin resulta-dos)	Yahoo	Dogpile (Sin resulta-dos)	Excite	Ixquick	Profu-sion	Search	Surfwax (Sin resulta-dos)	Vivisimo
Google	-	1 (0,5%)			18 (8,1%)		1 (0,5%)	1 (0,5%)	6 (2,7%)	18 (8,1%)		0
MSN	1 (0,5%)	-			1 (0,5%)		5 (2,3%)	6 (2,7%)	8 (3,6%)	16 (7,2%)		1 (0,5%)
Teoma (Ask) (Sin resultados)												
WiseNut (Sin resultados)												
Yahoo	18 (8,1%)	1 (0,5%)			-		2 (0,9%)	1 (0,5%)	6 (2,7%)	17 (7,7%)		0
Dogpile (Sin resultados)												
Excite	1 (0,5%)	5 (2,3%)			2 (0,9%)		-	7 (3,2%)	8 (3,6%)	4 (1,8%)		2 (0,9%)
Ixquick	1 (0,5%)	6 (2,7%)			1 (0,5%)		7 (3,2%)	-	13 (5,9%)	4 (1,8%)		3 (1,4%)
Profusion	6 (2,7%)	8 (3,6%)			6 (2,7%)		8 (3,6%)	13 (5,9%)	-	13 (5,9%)		2 (0,9%)
Search	18 (8,1%)	16 (7,2%)			17 (7,7%)		4 (1,8%)	4 (1,8%)	13 (5,9%)	-		1 (0,5%)
Surfwax (Sin resultados)												
Vivisimo	0	1 (0,5%)			0		2 (0,9%)	3 (1,4%)	2 (0,9%)	1 (0,5%)		-

De la presente tabla, pocos datos podemos extraer dado el limitado número de resultados en esta búsqueda. No obstante llama la atención el alto solapamiento de Google y Yahoo (8,1%), frente al 0,5% que mantiene con MSN. El otro dato destacado está en el solapamiento entre Google, MSN y Yahoo con Search, con porcentajes que oscilar entre el 7,2 y el 8,1%.

Entre los metabuscadores no se observan altos índices de solapamiento en esta búsqueda, destacando en todo caso en que mantiene Profusion con Ixquick y con Search (5,9%).

El menor solapamiento se da en esta ocasión entre Search y Vivisimo, (0,5%) con Search y entre Profusion y Excite (0,9%).

4.1.2.5. Búsqueda de frase

Tabla 4.1.2-5. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask) (Sin resultados)	WiseNut	Yahoo	Dogpile (Sin resultados)	Excite	Ixquick	Profu-sion	Search	Surfwax (Sin resultados)	Vivisimo
Google	-	18 (8,7%)		17 (8,2%)	16 (7,7%)		19 (9,1%)	17 (8,2%)	18 (8,7%)	21 (10,1%)		3 (1,4%)
MSN	18 (8,7%)	-		22 (10,6%)	11 (5,3%)		16 (7,7%)	15 (7,2%)	17 (8,2%)	19 (9,1%)		5 (2,4%)
Teoma (Ask) (Sin resultados)												
WiseNut	17 (8,2%)	22 (10,6%)		-	13 (6,3%)		11 (5,3%)	14 (6,7%)	15 (7,2%)	22 (10,6%)		7 (3,4%)
Yahoo	16 (7,7%)	11 (5,3%)		13 (6,3%)	-		22 (10,6%)	17 (8,2%)	15 (7,2%)	20 (9,6%)		5 (2,4%)
Dogpile (Sin resultados)												
Excite	19 (9,1%)	16 (7,7%)		11 (5,3%)	22 (10,6%)		-	25 (12%)	20 (9,6%)	22 (10,6%)		4 (1,9%)
Ixquick	17 (8,2%)	15 (7,2%)		14 (6,7%)	17 (8,2%)		25 (12%)	-	23 (11,1%)	19 (9,1%)		8 (3,8%)
Profusion	18 (8,7%)	17 (8,2%)		15 (7,2%)	15 (7,2%)		20 (9,6%)	23 (11,1%)	-	22 (%) (10,6%)		4 (1,9%)
Search	21 (10,1%)	19 (9,1%)		22 (10,6%)	20 (9,6%)		22 (10,6%)	19 (9,1%)	22 (10,6%)	-		6 (2,9%)
Surfwax (Sin resultados)												
Vivisimo	3 (1,4%)	5 (2,4%)		7 (3,4%)	5 (2,4%)		4 (1,9%)	8 (3,8%)	4 (1,9%)	6 (2,9%)		-

En la búsqueda por frase llama la atención el aumento de las cifras de solapamiento, lo que puede ser debido a la especificidad de la búsqueda. En general, los motores se solapan de forma semejante entre si. Destaca el solapamiento de MSN con WiseNut (10,6%) y con Google (8,7%). El menor solapamiento se da entre este buscador y Yahoo (5,3%).

Respecto al solapamiento de motores con los metabuscadores, podemos observar unos índices más o menos similares, que sólo destacan en el caso de Yahoo con Excite (10,6%) o de Google con Search (10,1%). En el otro extremo destaca Vivisimo, siendo, en esta búsqueda, el metabuscador con menor solapamiento entre las herramientas analizadas.

Entre los metabuscadores, se mantienen los altos porcentajes y sólo disminuyen en el caso de Vivisimo.

4.1.2.6. Búsqueda por campo

Tabla 4.1.2-6. Solapamiento (Base=registros operativos y no duplicados)

	Google	MSN	Teoma (Ask) (Sin resulta-dos)	WiseNut	Yahoo	Dogpile (Sin resulta-dos)	Excite	Ixquick	Profu-sion	Search	Surfwax (Sin resulta-dos)	Vivisimo
Google	-	15 (6,1%)		0	23 (9,3%)		13 (5,3%)	14 (5,7%)	0	12 (4,9%)		17 (6,9%)
MSN	15 (6,1%)	-		0	12 (4,9%)		7 (2,8%)	10 (4%)	2 (0,8%)	12 (4,9%)		13 (5,3%)
Teoma (Ask) (Sin resultados)												
WiseNut	0	0		-	0		4 (1,6%)	0	0	13 (5,3%)		0
Yahoo	23 (9,3%)	12 (4,9%)		0	-		9 (3,6%)	8 (3,2%)	0	10 (4%)		11 (4,5%)
Dogpile (Sin resultados)												
Excite	13 (5,3%)	7 (2,8%)		4 (1,6%)	9 (3,6%)		-	14 (5,7%)	0	14 (5,7%)		11 (4,5%)
Ixquick	14 (5,7%)	10 (4%)		0	8 (3,2%)		14 (5,7%)	-	1 (0,4%)	9 (3,6%)		12 (4,9%)
Profusion	0	2 (0,8%)		0	0		0	1 (0,4%)	-	0		0
Surfwax (Sin resultados)												
Search	12 (4,9%)	12 (4,9%)		13 (5,3%)	10 (4%)		14 (5,7%)	9 (3,6%)	0	-		11 (4,5%)
Vivisimo	17 (6,9%)	13 (5,3%)		0	11 (4,5%)		11 (4,5%)	12 (4,9%)	0	11 (4,5%)		-

En la búsqueda por campo se observa un alto solapamiento de Google y Yahoo (9,3%), siendo representativo el que mantiene con MSN (6%) y éste con Yahoo (4,9%).

Con WiseNut no hay solapamiento, en ningún caso.

Respecto al solapamiento con los metabuscadores, lo primero que llama la atención es que WiseNut, que apenas tiene solapamiento, alcanza un índice del 5,3% con Search.

Google se solapa con Vivisimo (6,9%), con Ixquick (5,7%), con Excite (5,3%) y con Search (4,9%). MSN se solapa con Vivisimo, Search e Ixquick. El solapamiento de Yahoo con el resto de metabuscadores es similar, a excepción de Profusión, con el que no hay solapamiento.

Entre los metabuscadores, Excite se solapa con Ixquick, Profusion y Vivisimo, pero no hay solapamiento con Profusion, que en esta búsqueda apenas tiene solapamiento tan-

to con buscadores como con metabuscadores. Vivisimo es, en esta ocasión el metabuscador con mayor solapamiento con el resto.

Análisis global

Tabla 4.1.2-7. Solapamiento (Base=registros operativos y no duplicados en las seis búsquedas)

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Google		53 (11,7%)	14 (6,3%)	26 (7,8%)	83 (17,2%)	31 (7,5%)	68 (10,7%)	56 (8,9%)	39 (8,5%)	92 (11,6%)	4 (2,4%)	40 (7,2%)
MSN	53 (10,4%)	-	9 (4,1%)	33 (9,9%)	39 (8%)	29 (7%)	41 (6,4%)	45 (7,1%)	47 (10,3%)	86 (10,8%)	6 (3,7%)	62 (11,1%)
Teoma (Ask)	14 (2,7%)	9 (2%)	-	4 (1,2%)	17 (3,5%)	32 (7,7%)	32 (5%)	25 (3,9%)	24 (5,2%)	13 (1,6%)	4 (2,4%)	45 (8,1%)
WiseNut	26 (5,1%)	33 (7,3%)	4 (1,8%)	-	20 (4,1%)	13 (3,1%)	22 (3,4%)	40 (6,3%)	26 (5,7%)	76 (9,6%)	29 (17,9%)	42 (7,5%)
Yahoo	83 (16,4%)	39 (8,6%)	17 (7,7%)	20 (6%)	-	30 (7,2%)	71 (11,2%)	50 (7,9%)	43 (9,4%)	88 (11,1%)	2 (1,2%)	39 (7%)
Dogpile	31 (6,1%)	29 (6,4%)	32 (14,6%)	13 (3,9%)	30 (6,2%)	-	70 (11%)	48 (7,8%)	31 (6,7%)	62 (7,8%)	11 (6,7%)	56 (10%)
Excite	68 (13,4%)	41 (9,1%)	32 (14,6%)	22 (6,6%)	71 (14,7%)	70 (16,9%)	-	103 (16,4%)	61 (13,3%)	98 (12,4%)	9 (5,5%)	58 (10,4%)
Ixquick	56 (11%)	45 (10%)	25 (11,4%)	40 (12%)	50 (10,3%)	48 (11,6%)	103 (16,2%)	-	66 (14,4%)	91 (11,5%)	33 (20,3%)	69 (12,4%)
Profusion	39 (7,7%)	47 (10,4%)	24 (10,9%)	26 (7,8%)	43 (8,9%)	31 (7,5%)	61 (9,6%)	66 (10,5%)	-	72 (9,1%)	11 (6,7%)	36 (6,4%)
Search	92 (18,2%)	86 (19,1%)	13 (5,9%)	76 (22,9%)	88 (18,2%)	62 (15%)	98 (15,4%)	91 (14,5%)	72 (15,7%)	-	28 (17,2%)	83 (14,9%)
Surfwax	4 (0,7%)	6 (1,3%)	4 (1,8%)	29 (8,7	2 (0,4%)	11 (2,6%)	9 (1,4%)	33 (5,2%)	11 (2,4%)	28 (3,5%)	-	25 (4,5%)
Vivisimo	40 (7,9%)	62 (13,7%)	45 (20,5%)	42 (12,6	39 (8%)	56 (13,5%)	58 (9,1%)	69 (11%)	36 (7,8%)	83 (10,5%)	25 (15,4%)	-
	505	**450**	**219**	**331**	**482**	**413**	**633**	**626**	**456**	**789**	**162**	**555**

La presente tabla muestra el solapamiento entre motores de búsqueda y metabuscadores teniendo en cuenta los cincuenta primeros recursos recuperados en las seis búsquedas.

El primer problema al que nos enfrentamos en el análisis de los datos es que hay búsquedas en las que los buscadores no recuperaron recursos, lo que da lugar a que en estos casos, las cifras de solapamiento sean bajas. Esto ocurre sobre todo con Teoma, Surfwax y Dogpile que no recuperaron en tres de las búsquedas. También, aunque en menor medida se aprecia en WiseNut y Profusión que no recuperaron en una de las búsquedas.

Por tanto, nos centraremos en valorar el solapamiento entre buscadores y metabuscadores que recuperaron recursos en todas las búsquedas, lo que nos permitirá conocer el mejor modo de combinarlos para obtener búsquedas más completas. Las cantidades que

aparecen en la tabla indican el número total de recursos con solapamiento recogido en las seis búsquedas entre las herramientas de búsqueda correspondientes. Los porcentajes se han calculado sobre el total de recursos solapados que corresponden a cada herramienta de búsqueda, que es la cifra que aparece al final de las columnas.

Teniendo en cuenta lo anterior, podemos apreciar que las herramientas de búsqueda en las que se da un mayor solapamiento son los metabuscadores. En este sentido, Search es la herramienta de búsqueda con mayor solapamiento (789 recursos), seguido de Excite con 633 e Ixquick con 626, correspondiendo el menor solapamiento de los metabuscadores que recuperan en todas las búsquedas a Vivisimo.

Entre los buscadores es MSN el que menos solapamiento presenta seguido de Yahoo, siendo Google el que más resultados solapados ofrece.

MSN tiene mayor solapamiento con Google que con Yahoo y con los metabuscadores Search y Vivisimo, con los que coincide en la recuperación de 86 resultados, un 19,1% y 62 (13,7%) páginas respectivamente. Corresponde a Excite la recuperación del menor número de recursos comunes, esto es 41 (9,1%).

Google tiene el mayor solapamiento con Yahoo, con el que coincide en 83 páginas, (16,4%) y con el metabuscador Search, con el que coincide en 92 páginas (18,2%). El menor solapamiento lo obtiene con MSN y con el metabuscador Vivisimo.

De forma recíproca, Yahoo tiene un alto solapamiento con Google, 83 (16.4%) descendiendo considerablemente el que mantiene con MSN 39 (8%). Con Search y Excite también mantiene un alto solapamiento, al coincidir en 88 recursos con el primer caso, (18,2%) y en 71 (14,7%) con el segundo. El menor solapamiento de este motor se da con Vivisimo.

Teoma registra el mayor solapamiento con Google y Yahoo así como con el metabuscador Vivisimo, con el que llama la atención el hecho de que coincidan en la recuperación de 45 recursos (8,1%) en tan sólo tres de las consultas en las que Teoma aportó resultados, cifras que en proporción, resultan elevadas.

El mayor solapamiento de WiseNut se da con MSN y Google y con el metabuscador Search. Dogpile es con el que menor coincidencia de resultados existe.

En cuanto a los metabuscadores, Excite tiene un solapamiento en torno al 11% con Yahoo y con Google, disminuyendo al 6,4% con MSN. Con los metabuscadores el solapamiento es mayor ya que alcanza un 16,2% con Ixquick y un 15,4% con Search, descendiendo a un 9,1% con Vivisimo.

Ixquick tiene un solapamiento similar con los tres buscadores que recuperaron en las seis búsquedas, esto es Google, Yahoo y MSN con porcentajes del 8,9%, 7,9% y 7,1% respectivamente. Como ya hemos señalado, hay un alto solapamiento con el metabuscador Excite y también con Search (14,5%). Con Vivisimo el porcentaje disminuye al 11%.

El solapamiento de Search es similar con los tres buscadores, oscilando entre el 10,8% con MSN y el 11,6% de Google, aspecto que se repite en los metabuscadores con los que los porcentajes van del 10,5% con Vivisimo al 12,4 con Excite.

Finalmente Vivisimo alcanza su mayor solapamiento con MSN (11,1%), siendo muy similar al obtenido con Google y Yahoo (7,2% y 7% respectivamente). Search es el metabuscador con el que más solapamiento existe (14,9%), seguido de Ixquick (12,4%) y Excite (10,4%).

De todas estas cifras podemos deducir que una buena combinación para obtener resultados distintos y posiblemente más completos sea la combinación de Google con MSN o de Yahoo con MSN, puesto que son los que ofrecen menos solapamiento entre ellos.

Respecto a los metabuscadores, el que mejor se complementa con el resto de buscadores, y que por tanto es aconsejable en búsquedas exhaustivas es Vivisimo, si bien, hay que tener en cuenta su solapamiento con MSN. En este mismo sentido, también podemos decir que resulta menos aconsejable el uso de Excite y Search.

Por otro lado, también debemos de tener en cuenta, al valorar el solapamiento entre buscadores, que se deben considerar además otros aspectos como por ejemplo el carácter de la información, ya que un bajo solapamiento por si sólo no puede ser del todo definitivo en la elección de dos o más herramientas complementarias puesto que puede ocurrir que alguna de ellas recupere preferentemente recursos comerciales o de otro tipo, y no de investigación.

5. Análisis de la precisión técnica

Para el análisis de la precisión técnica en la primera búsqueda, nos basamos en las frecuencias con las que aparece el término en los documentos recuperados, que es la metodología utilizada por los investigadores que se han ocupado de estudiar este aspecto. En el resto de las búsquedas, dado que prácticamente no se recuperan recursos conteniendo todos los términos, optamos por analizar las frecuencias con las que aparecen de forma individual determinados términos o frases de búsqueda.

5.1. Búsqueda de un término

Para el cálculo de la precisión técnica en esta búsqueda utilizaremos la fórmula propuesta por Bar-Ilan (1998) que calcula la precisión hallando el tanto por ciento que se obtiene al dividir el número de documentos que contienen el término de búsqueda por el número de documentos accesibles. Para ello se basa en los resultados ofrecidos en la búsqueda de un término. De aquí que este cálculo sólo se pueda aplicar a la primera búsqueda.

La siguiente tabla recoge los datos aportados en dicha búsqueda por los buscadores evaluados.

Tabla 5.1-1. Búsqueda 1. Recursos analizados

Recursos que contienen el término de búsqueda:	206 (38,2%)
Recursos que no contienen el término de búsqueda en el texto:	316 (58,7%)
Recursos a los que no se pudo acceder por dar error:	16 (2,9%)
Total recursos analizados	**538 (100%)**

En base a estos datos, la frecuencia técnica de las herramientas de búsqueda analizadas es del 38,2%, muy inferior a la obtenida por Bar-Ilan (77,2%), para quién esta alta precisión, como expresa en sus conclusiones fue motivo de sorpresa, por lo que plantea la necesidad de más investigación en este aspecto, y que trate de responder por qué los usuarios se quejan de la poca precisión frente a los resultados por él obtenidos.

La diferencia de valores obtenidos entre el trabajo de Bar-Ilan y el nuestro puede ser debida a los términos de búsqueda, ya que este autor utiliza en su trabajo un término muy específico, concretamente el apellido “Erdos” para recuperar registros relacionados

con este matemático. Este tipo de búsquedas con un término tan concreto requiere de los buscadores una alta precisión ya que es difícil recuperar, como ocurre con otros términos, palabras derivadas o lingüísticamente relacionadas que utilicen la misma raíz. En nuestro caso al tratarse del término "Softbot", los buscadores recuperaron un gran número de recursos con el término en plural, que aunque la valoración de la precisión técnica no han sido dados por válidos, no es tan específico como el anterior. Por tanto, podemos observar la existencia de variaciones en los resultados, en función de los términos, por lo que una vez más hemos de tener en cuenta que los datos son indicativos, y que para establecer conclusiones categóricas, es necesario seguir investigando y utilizar para las comparaciones resultados obtenidos utilizando los mismos términos. No obstante nuestros datos sirven para precisar la opinión de Bar-Ilan en cuanto a su extrañeza respecto a los resultados sobre la precisión en su búsqueda.

En cualquier caso, y además, teniendo en cuenta que se trata de una muestra reducida, los datos son suficientemente reveladores de la poca precisión técnica que caracteriza a estas herramientas de búsqueda, utilizando términos relativamente específicos.

Los datos que arrojan de forma individual los buscadores son los siguientes:

Tabla 5.1-2. Motores. Frecuencia de aparición del término "softbot"

	Google	MSN	Teoma (Ask)	WiseNut	Yahoo
Recursos que contienen el término de búsqueda	24 (48%)	13 (26%)	11 (22%)	11 (22%)	25 (50%)
Recursos que no lo contienen	26 (52%)	31 (62%)	36 (72%)	36 (72%)	22 (44%)
Recursos a los que no se pudo acceder	0	6 (12%)	3 (6%)	3 (6%)	3 (6%)
Total recursos analizados	50 (100%)	50 (100%)	50 (100%)	50 (100%)	50 (100%)

Desde este punto de vista, se aprecia una mejora en los resultados al alcanzar, en los mejores casos, una precisión técnica en torno al 50%, como es el caso de Yahoo y Google. No obstante hay que señalar que estos valores siguen estando por debajo de lo que correspondería a herramientas de búsqueda en las que se requiere, al realizar búsquedas especializadas, resultados de una mayor precisión.

Por otro lado, los peores resultados, con altos porcentajes de recursos que no contienen el término de búsqueda, corresponden a MSN, seguido de Teoma y WiseNut. Esto puede ser debido a que recuperan un importante número de páginas que contienen el término en plural, lo que, además de un problema de recuperación, supone una notable falta de precisión.

Tabla 5.1-3. Metabuscadores. Frecuencia de aparición del término "softbot"

	Dogpile	Excite	Ixquick	Profusion	Search	Surfwax	Vivisimo
Recursos que contienen el término de búsqueda	19 (38%)	32 (64%)	10 (31,2%)	21 (51,2%)	20 (40%)	3 (20%)	17 (34%)
Recursos que no lo contienen	30 (60%)	17 (34%)	21 (65,6%)	19 (46,3%)	29 (58%)	12 (80%)	32 (64%)
Recursos a los que no se pudo acceder	1 (2%)	1 (2%)	1 (3,1%)	1 (2,4%)	1 (2%)	0	1 (2%)
Total recursos analizados	50 (100%)	50 (100%)	32 (100%)	41 (100%)	50 (100%)	15 (100%)	50 (100%)

Excite es el metabuscador con mayor precisión técnica ya que presenta un porcentaje superior al de los motores, por lo que constituye una herramienta de búsqueda a tener en cuenta en búsquedas que requieran una alta precisión técnica.

Sólo Profusión supera tímidamente el 50%, correspondiendo los peores resultados a Surfwax con el 20%.

5.2. Búsqueda utilizando el lenguaje natural

En las siguientes búsquedas hemos analizado tanto el número de documentos que no contienen los términos de búsqueda como la frecuencia de aparición de los términos en cada recurso, para lo que, dado que los resultados apenas ofrecen algún recurso con todos los términos de búsqueda solicitados, hemos descompuesto los términos de las búsquedas por palabras y frases.

Los términos y frases seleccionados en los que hemos basado el análisis son los siguientes:

Búsqueda completa:

1. best-match information retrieval in web search engines

Términos y frases:

2. best
3. match
4. best-match
5. information retrieval
6. web search
7. web search engines
8. search engines

Ofrecemos en primer lugar un análisis individual de los resultados relacionados con cada uno de los buscadores, para presentar a continuación, un análisis comparativo entre ellos.

5.2.1. Análisis individualizado de los motores de búsqueda

Las siguientes tablas muestran el comportamiento de los buscadores en cuanto a la frecuencia de aparición de los términos de búsqueda en los recursos recuperados. Dado que como hemos indicado, prácticamente ninguna de estas herramientas recuperó páginas conteniendo todos los términos solicitados en las búsquedas de más de un término, teniendo en cuenta la metodología que diferentes autores utilizan para valorar la precisión técnica, nos pareció interesante valorar en qué medida, los recursos contenían alguno de los términos o frases de búsqueda.

La primera tabla que se muestra a continuación del nombre de cada buscador recoge el número de recursos analizados en cada búsqueda. A continuación, las demás tablas muestran, en la primera columna el número de veces que aparece un término o frase en un recurso. Así, cuando no aparecen los términos en las páginas, el resultado es cero. La segunda columna recoge el número de recursos en los que aparece. La tercera indica el porcentaje que ese número supone entre los recursos recuperados.

Google

Tabla 5.2.1-1. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.1-2. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Como podemos apreciar, en esta segunda búsqueda Google no recupera páginas con todos los términos.

Tabla 5.2.1-3. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	10	20%
1	16	32%
2	7	14%
3	5	10%
4	4	8%
5	3	6%
9	4	8%
10	1	2%
Total	50	100%

El porcentaje de páginas que no contiene este término (20%), es elevado y sensiblemente superior al que ofrecen MSN (8%), WiseNut (12%) y Yahoo (14%) lo que indica que este buscador valora en menor medida que el resto no tiene en cuenta en la recuperación la aparición de los términos de búsqueda en los documentos que recupera. Por otro lado, llama la atención el alto número de documentos en los que el término sólo aparece una vez (16 documentos).

Tabla 5.2.1-4. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	24%
1	14	28%
2	8	16%
3	6	12%
4	2	4%
12	1	2%
15	2	4%
18	4	8%
39	1	2%
Total	50	100%

En relación con el término "match", los resultados de la presente tabla muestran dos grupos, el primero de ellos formado por documentos con frecuencias bajas, no superando cuatro apariciones por documento (el 60% de los recuperados) y un segundo grupo (el 14% de los documentos) con frecuencias de aparición entre 12 y 18 veces por

documento. Finalmente, en un documento aparece en 39 ocasiones. Igualmente se puede apreciar gran similitud respecto a la tabla anterior.

Tabla 5.2.1-5. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	45	90%
1	3	6%
2	2	4%
Total	50	100%

El término compuesto tan sólo aparece en 5 de los documentos recuperados, con frecuencias mínimas de aparición.

El comportamiento es similar al de Excite (Tabla 5.1-58), pero el metabuscador recupera un recurso con más frecuencia de aparición del término (9 veces) frente a los dos documentos que recupera Google en los que aparece 2 veces.

Tabla 5.2.1-6. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	35	70%
3	2	4%
4	2	4%
6	4	8%
15	1	2%
16	1	2%
27	1	2%
29	1	2%
30	3	6%
Total	50	100%

En primer lugar hemos de destacar el alto porcentaje de recursos que no contienen los términos de búsqueda (70%).

Por otro lado llama la atención el distinto comportamiento respecto de los términos analizados anteriormente, pues en este caso, disminuyen los documentos en los que aparecen con poca frecuencia y aumentan los documentos en los que la frecuencia de

aparición es alta (en un documento aparecen en 15, 16, 27 y 29 ocasiones, y en tres documentos, 30 veces.

Las frecuencias son similares a Excite, pero el metabuscador recupera un recurso con mayor frecuencia de aparición de los términos (35 veces).

Tabla 5.2.1-7. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	56%
1	5	10%
2	4	8%
3	3	6%
5	1	2%
6	4	8%
13	1	2%
18	1	2%
24	2	4%
52	1	2%
Total	50	100%

Como en el caso anterior, Google vuelve a recuperar un amplio número de recursos con baja frecuencia de aparición de los términos, pero aparece compensado con otros en los que las frecuencias son más elevadas. El comportamiento es similar en esta ocasión a MSN (Tabla 5.1-19).

En la frecuencia de estos términos, Google supera a Excite (Tabla 5.1-64) al presentar un documento con una frecuencia de 52, frente a 24 que es la máxima de Excite.

Tabla 5.2.1-8. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	40	81,6%
1	1	2%
2	4	8,2%
3	1	2%
9	2	4,1%
32	1	2%
Total	49	100%

La frecuencia de aparición de los tres términos ofrece datos similares a los observados anteriormente, aunque el número de recursos en los que no aparecen los términos es muy superior.

Tabla 5.2.1-9. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	22	44,9%
1	1	2%
2	1	2%
3	2	4,1%
4	2	4,1%
5	1	2%
7	3	6,1%
8	3	6,1%
13	1	2%
17	2	4,1%
18	3	6,1%
25	1	2%
33	1	2%
35	1	2%
36	1	2%
38	1	2%
40	2	4,1%
57	1	2%
Total	49	100%

La recuperación de recursos con estos dos términos es más frecuente que los anteriores aunque un 44,9% no los contiene. Por otro lado se observa un aumento del número de recursos con altas frecuencias de aparición de los términos.

Por tanto podemos decir que Google ofrece unos resultados muy pobres en relación con la aparición de los términos de búsqueda en los documentos que recupera, siendo muy expresivos en este sentido los resultados obtenidos en cuanto al término compuesto "*best-match*" que es el más específico de los que componen la búsqueda.

MSN

Tabla 5.2.1-10. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.1-11. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

MSN como Google tampoco recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.1-12. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	4	8%
1	17	34%
2	7	14%
3	4	8%
4	4	8%
6	2	4%
7	1	2%
8	2	4%
9	1	2%
10	2	4%
11	1	2%
12	2	4%
14	1	2%
24	1	2%
25	1	2%
Total	50	100%

El porcentaje de recursos que no contienen este término de búsqueda (8%), es sensiblemente inferior al que ofrece Google (20%) y determinados metabuscadores como Dogpile (27,9%), Excite (12%), etcétera. También se observa MSN recupera un mayor

número de documentos en los que el término aparece con una mayor variedad de frecuencias.

Tabla 5.2.1-13. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	5	10%
1	19	38%
2	11	22%
3	5	10%
4	1	2%
5	1	2%
7	1	2%
12	1	2%
14	2	4%
15	1	2%
16	1	2%
27	1	2%
39	1	2%
Total	50	100%

Como en el caso anterior, MSN sigue ofreciendo el menor número de documentos que no contienen el término de búsqueda (5 recursos) y mayor variedad de frecuencias del término en diferentes documentos.

Tabla 5.2.1-14. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	49	98%
2	1	2%
Total	50	100%

Los buscadores apenas recuperan recursos con el término compuesto, y MSN, de acuerdo con esta tendencia sólo recupera un documento en el que el término aparece dos veces, teniendo peor comportamiento que Google, que recuperó cinco documentos.

Tabla 5.2.1-15. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	27	54%
1	2	4%
3	1	2%
4	1	2%
5	3	6%
6	1	2%
8	1	2%
11	1	2%
16	1	2%
17	2	4%
18	1	2%
19	1	2%
23	2	4%
24	1	2%
27	1	2%
28	1	2%
40	1	2%
47	1	2%
87	1	2%
Total	50	100%

MSN sigue manteniendo menores porcentajes de páginas que sin estos términos de búsqueda, con lo que se favorece la precisión técnica. Destaca la recuperación de dos documentos que superan las 40 ocurrencias y otro con 87.

Tabla 5.2.1-16. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	32	64%
1	5	10%
2	4	8%
3	3	6%
5	1	2%
6	2	4%
8	1	2%
17	1	2%
52	1	2%
Total	50	100%

MSN ofrece en esta ocasión frecuencias similares a las de Google, superándole éste al recuperar recursos sin los términos, frente a 32 de MSN.

Tabla 5.2.1-17. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	40	80%
1	6	12%
2	1	2%
3	2	4%
32	1	2%
Total	50	100%

Como ocurre con Google, también MSN recupera un alto número de recursos sin los tres términos.

En general la recuperación de estos términos es similar en Google, con una recuperación de recursos con frecuencias bajas y en el otro extremo, la recuperación de algún recurso en el que los términos aparecen con una frecuencia elevada (32 veces).

Tabla 5.2.1-18. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	42%
1	1	2%
3	2	4%
4	1	2%
5	1	2%
7	1	2%
8	2	4%
10	2	4%
11	1	2%
12	3	6%
13	1	2%
15	1	2%
18	2	4%
21	2	4%
27	1	2%
33	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
35	2	4%
36	1	2%
42	1	2%
49	1	2%
57	1	2%
73	1	2%
Total	50	100%

También en estos términos el comportamiento es similar a Google, si bien MSN ofrece un mayor número de páginas con diferentes frecuencias de aparición de los términos. Destaca además por la recuperación de un recurso en el que estos términos aparecen en 73 ocasiones.

En resumen, la precisión técnica en MSN es en esta búsqueda, aunque es muy similar en algunos términos a Google, superándolo en cuanto a porcentajes de páginas que contienen los términos de búsqueda, y en la variedad de frecuencias que muestran los documentos recuperados. Al margen de la recuperación con el término compuesto, best-match que no es muy frecuente en los recursos recuperados, como se ha demostrado en el comportamiento con el resto de términos, este buscador da importancia a la existencia de los términos de búsqueda en los documentos que recupera.

Teoma (Ask)

Tabla 5.2.1-19. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.1-20. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Como en los buscadores anteriores, Teoma tampoco recupera páginas con todos los términos de la búsqueda

Tabla 5.2.1-21. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	24%
1	13	26%
2	7	14%
3	6	12%
4	8	16%
5	1	2%
7	1	2%
9	1	2%
10	1	2%
Total	50	100%

Teoma es el motor que menos recursos con este término recupera, y se caracteriza por una recuperación en dos grupos. En el primero de ellos, con muy bajas frecuencias, que correspondería al 90% de los recursos recuperados, con frecuencias no superan los cuatro casos, y el segundo grupo (8%) sólo aparece en una ocasión.

Tabla 5.2.1-22. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	15	30%
1	15	30%
2	10	20%
3	3	6%
4	2	4%
5	1	2%
10	1	2%
12	1	2%
15	1	2%
39	1	2%
Total	50	100%

Teoma sigue mostrando el comportamiento en dos grupos, si bien, en esta ocasión, si que recupera un recurso con alta frecuencia de aparición del término (39 veces), que también aparecía en los buscadores anteriores. El alto número de recursos sin el término de búsqueda influye en su baja precisión técnica.

Tabla 5.2.1-23. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	76%
1	10	20%
3	1	2%
9	1	2%
Total	50	100%

Teoma destaca en esta ocasión por la recuperación de un mayor número de recursos (10) en los que el término compuesto aparece una vez, pero también hay que destacar la recuperación de un recurso en el que aparece en 9 ocasiones.

Tabla 5.2.1-24. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	27	54%
1	7	14%
4	1	4%
5	2	4%
6	1	2%
12	1	2%
13	2	4%
15	2	4%
16	2	4%
19	1	2%
22	1	2%
24	1	2%
35	1	2%
Total	50	100%

Dentro de las características señaladas, aunque Teoma sigue mostrando un alto número de recursos en los que los términos de búsqueda no aparecen, en esta ocasión supera las frecuencias mostradas por Google, aunque sin alcanzar los datos de MSN.

Tabla 5.2.1-25. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	22	44%
1	10	20%
2	2	4%
3	2	4%
4	3	6%
5	1	2%
6	1	2%
7	2	4%
8	1	2%
9	1	2%
11	1	2%
14	2	4%
21	1	2%
52	1	2%
Total	50	100%

El comportamiento es similar al observado en los términos anteriores, destacando respecto a los anteriores por el número de páginas (10) en las que los términos de nuevo la recuperación de páginas en las que aparecen tan sólo en una ocasión.

Tabla 5.2.1-26. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	56%
1	8	16%
2	6	12%
4	2	4%
5	1	2%
7	2	4%
12	1	2%
14	1	2%
32	1	2%
Total	50	100%

También aquí mantiene las características observadas anteriormente en cuanto a la compensación entre el alto número de resultados con bajas frecuencias y los recursos

con frecuencias superiores. Recupera en esta ocasión un mayor número de recursos con estos términos que Google y MSN.

Tabla 5.2.1-27. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	16	32%
1	2	4%
2	3	6%
3	3	6%
5	1	2%
6	2	4%
7	1	2%
10	1	2%
11	3	6%
16	1	2%
17	1	2%
18	1	2%
19	1	2%
23	1	2%
26	1	2%
30	1	2%
33	1	2%
36	1	2%
38	1	2%
39	1	2%
42	1	2%
46	1	2%
57	1	2%
65	1	2%
77	1	2%
81	1	2%
114	1	2%
Total	50	100%

Teoma (Ask) recupera un importante número de recursos con frecuencias altas, destacando por la recuperación de un recurso en que los términos aparecen en 114 ocasiones. Por otro lado, se mantiene como el resto en cuanto a recursos con menores frecuencias.

En definitiva, Teoma (Ask) tiene para esta búsqueda un comportamiento similar a Google, al que supera, al igual que al resto, en el caso del término "*best-match"* y en la recuperación de un menor número de recursos que no contienen los términos de búsqueda analizados.

WiseNut

Tabla 5.2.1-28. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.1-29. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines "

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

WiseNut tampoco recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.1-30. Frecuencia y nº de recursos en los que aparece el término "best "

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12%
1	12	24%
2	11	22%
3	6	12%
4	2	4%
5	1	2%
6	3	6%
7	1	2%
9	2	4%
11	1	2%
12	1	2%
14	1	2%
16	1	2%
35	1	2%
38	1	2%
Total	50	100%

WiseNut y MSN son los buscadores en los que menor es el número de páginas recuperadas que no contiene el término de búsqueda. En general, este último recupera un gran número de recursos en los que el término aparece una vez (17) frente a 6 recursos en WiseNut. Éste a su vez recupera más recursos en los que el término aparece 2 y 3 veces, aunque WiseNut recupera menor número de recursos de altas frecuencias de aparición del término que MSN. Sin embargo recupera dos documentos de 35 y 38 casos, que no recuperó aquél.

Tabla 5.2.1-31. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	18%
1	21	42%
2	15	30%
3	2	4%
6	1	2%
8	1	2%
19	1	2%
Total	50	100%

WiseNut tiene en esta ocasión un comportamiento más irregular que el observado respecto al término anterior, destacando el alto número de recursos en los que el término aparece en el texto sólo en una o dos ocasiones.

Tabla 5.2.1-32. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	49	98%
1	1	2%
Total	50	100%

WiseNut sólo recupera un recurso en el que el término compuesto aparece una vez, siendo similar a MSN, si bien en el documento recuperado por éste, el término aparecía dos veces.

Tabla 5.2.1-33. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	47	94%
1	1	2%
3	1	2%
19	1	2%
Total	50	100%

WiseNut sólo recupera tres registros con estos términos, ofreciendo en este sentido el peor resultado en comparación, no sólo con los anteriores, sino con el resto de motores de búsqueda.

Tabla 5.2.1-34. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	41	82%
1	2	4%
2	1	2%
3	1	2%
4	1	2%
6	1	2%
12	1	2%
13	1	2%
18	1	2%
Total	50	100%

Las frecuencias de recuperación de páginas que no contienen los términos de búsqueda son también en esta ocasión las más elevadas, lo que confirma la baja precisión técnica de este motor.

Tabla 5.2.1-35. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	47	94%
2	2	4%
11	1	2%
Total	50	100%

Como en los términos anteriores, corresponden a este buscador los mayores porcentajes de recursos sin los términos de búsqueda.

Tabla 5.2.1-36. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	35	70%
1	5	10%
2	3	6%
11	1	2%
13	1	2%
25	1	2%
34	1	2%
36	1	2%
42	1	2%
70	1	2%
Total	50	100%

WiseNut también aquí destaca por el alto porcentaje de recursos que no contienen estos términos (70%), lo que unido a la mayor recuperación de recursos con bajas frecuencias, da lugar a una baja precisión técnica.

En resumen, debemos mencionar en primer lugar el desigual comportamiento de este buscador en la recuperación de documentos con los diferentes términos de búsqueda analizados. Las diferencias observadas por ejemplo entre la recuperación de los dos primeros términos, tal vez puedan ser debidas a la distinta frecuencia con que pueden aparecer estos términos en los documentos ya que el segundo término (*match*) es más específico. No obstante, la recuperación de un único recurso con el término "*best-match*" nos permite afirmar que existe un deficiente funcionamiento en este motor en relación con términos compuestos.

Yahoo

Tabla 5.2.1-37. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.1-38. Frecuencia y nº de recursos en los que aparece el término "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Yahoo no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.1-39. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	7	14%
1	14	28%
2	4	8%
3	4	8%
4	8	16%
5	2	4%
6	2	4%
8	1	2%
9	3	6%
11	1	2%
12	2	4%
13	1	2%
40	1	2%
Total	50	100%

Yahoo presenta frente a Google una mayor variedad de documentos con frecuencias distintas y sobre todo por la recuperación de documentos con frecuencias más altas, ya que Google no recupera recursos en los que el término aparece más de diez veces.

Tabla 5.2.1-40. Frecuencia y nº de recursos en los que aparece el término "match"

Recursos en los que los términos aparecen n veces	Nº de recursos	Porcentaje
0	12	24%
1	14	28%
2	7	14%
3	6	12%
4	6	12%
6	1	2%
7	3	6%
12	1	2%
Total	50	100%

Yahoo presenta mayor variación respecto al comportamiento en la recuperación del término anterior, disminuyendo en esta ocasión los documentos con mayor frecuencia de aparición de los términos. Tanto es así que es superado por Google.

Tabla 5.2.1-41. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	45	90%
1	1	2%
2	1	2%
3	1	2%
8	1	2%
9	1	2%
Total	50	100%

Yahoo muestra un comportamiento sin grandes contrastes en la recuperación de éste término, permaneciendo constante el número de documentos (1) en los que el término aparece en una, dos, tres, ocho y hasta nueve ocasiones, superando los resultados del resto de buscadores.

Tabla 5.2.1-42. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	11	74%
1	1	2%
2	1	2%
3	1	2%
4	1	2%
5	2	4%
8	1	2%
9	1	2%
16	1	2%
17	1	2%
19	1	2%
28	1	2%
90	1	2%
Total	50	100%

Yahoo no tiene un comportamiento tan destacado en la recuperación de recursos con estos términos ya que cuenta con un importante número de recursos que no los contienen, (74%). Es el buscador que recupera el recurso en el que los términos aparecen con mayor frecuencia (90 ocasiones).

Tabla 5.2.1-43. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	56%
1	4	8%
3	5	10%
4	1	2%
5	1	2%
6	1	2%
7	1	2%
8	1	2%
10	1	2%
14	2	4%
17	1	2%
18	1	2%
19	1	2%
24	1	2%
52	1	2%
Total	50	100%

De forma similar a Google y MSN, Yahoo recupera varios recursos en los que la frecuencia de aparición de los términos es baja, pero en esta ocasión, recupera, respecto al resto de buscadores, documentos en los que las frecuencias son más variadas, y en general más elevadas que el resto.

Tabla 5.2.1-44. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	33	66%
1	3	6%
2	1	2%
3	4	8%
4	2	4%
5	1	2%
7	2	4%
9	2	4%
12	1	2%
32	1	2%
Total	50	100%

Yahoo ofrece en esta ocasión unas frecuencias similares a MSN aunque con mayor número de recursos que contienen los términos de búsqueda con frecuencias superiores a las de éste último.

Tabla 5.2.1-45. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	13	26%
1	3	6%
5	3	6%
6	2	4%
7	1	2%
8	2	4%
10	1	2%
11	1	2%
14	2	4%
15	1	2%
16	1	2%
17	2	4%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
18	3	6%
20	2	4%
21	1	2%
26	1	2%
32	1	2%
35	1	2%
36	1	2%
40	1	2%
42	1	2%
45	1	2%
46	1	2%
49	1	2%
51	1	2%
57	1	2%
77	1	2%
Total	50	100%

Respecto a estos términos, llama la atención el importante descenso del porcentaje de recursos que no contienen estos términos de búsqueda (26%) frente al 66% observado en los términos anteriores. Por otro lado, y como se refleja en la amplitud de la tabla, se recuperaron documentos con una mayor variedad de frecuencias de aparición de los términos. En tercer lugar, la recuperación se caracteriza por aportar pocos documentos con frecuencias de repetición bajas.

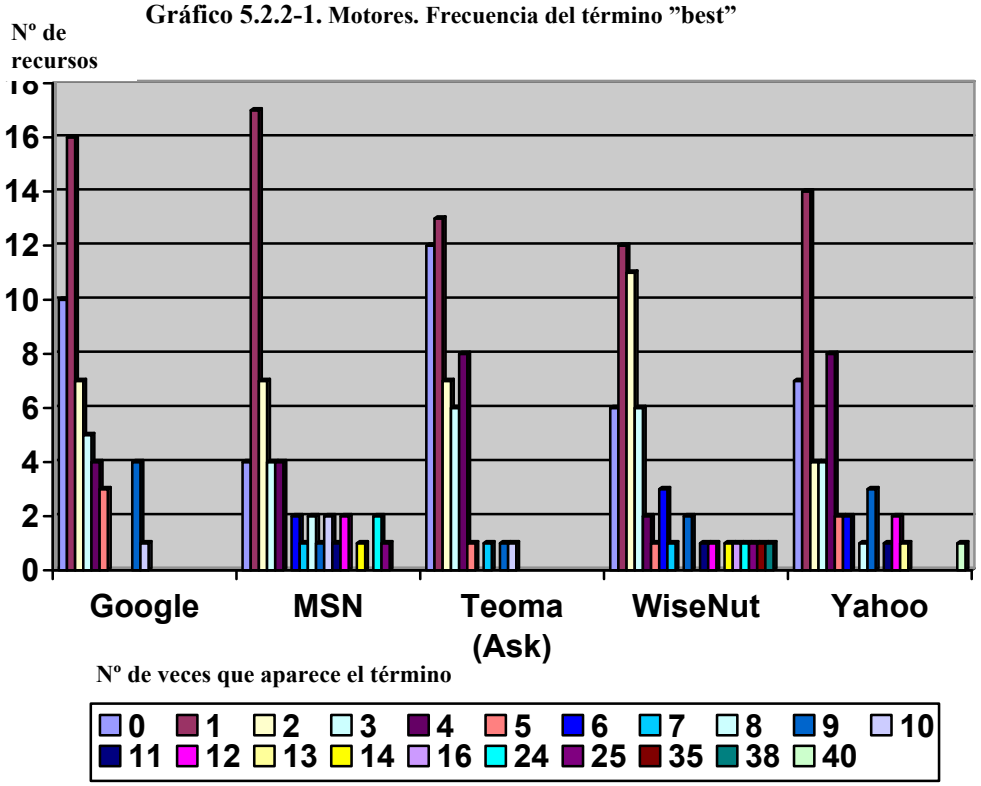
En consecuencia, Yahoo, a pesar de que tampoco recupera recursos con todos los términos de búsqueda, tiene un comportamiento que supera al resto de los buscadores en las frecuencias de aparición de los términos. También debemos destacar a Teoma (Ask), que en gran medida centra la recuperación en la existencia de los términos de búsqueda, seguidos en este aspecto por Google y MSN.

5.2.2. Análisis comparativo de los motores de búsqueda

Aunque ya hemos comparado de una forma básica los resultados de unos y otros motores al analizar la precisión técnica de forma individual, las siguientes gráficas van a facilitarnos un análisis con mayor profundidad.

En el título del gráfico aparecen el término o términos a los que se refiere. La leyenda que acompaña a los gráficos expresa la frecuencia con que aparecen los términos en los documentos. En el eje de ordenadas se expresa el número de documentos que contiene cada término o expresión analizada.

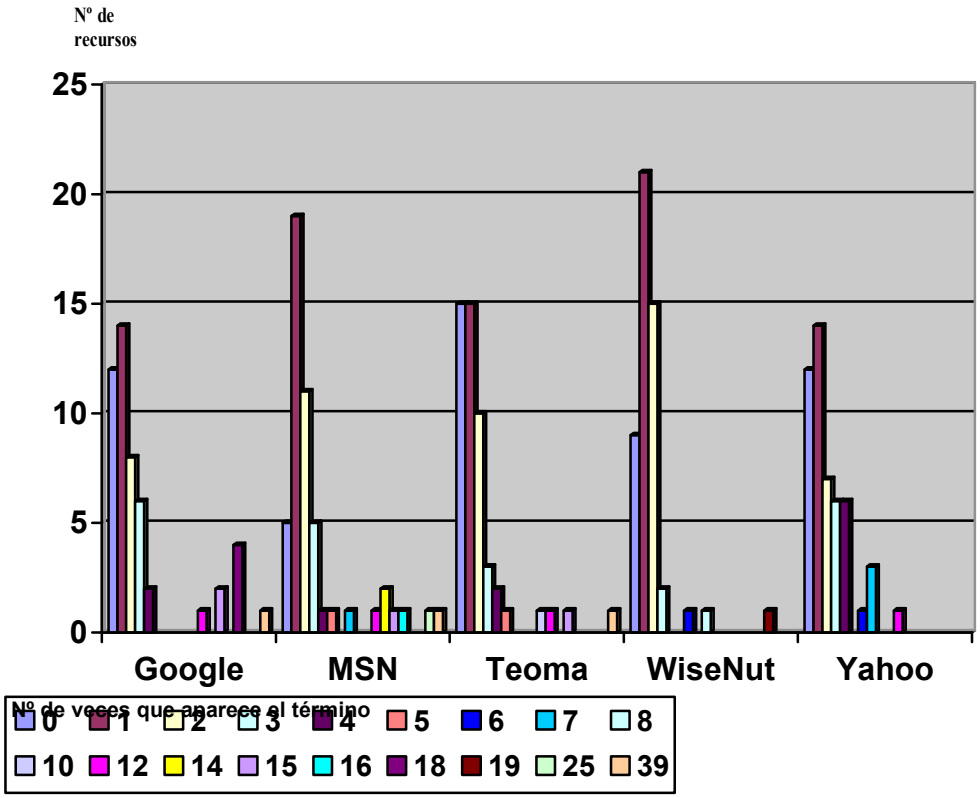
Gráfico 5.2.2-1. Motores. Frecuencia del término "best"



La presente gráfica muestra las diferencias en las frecuencias de aparición del término *best* entre los distintos motores. En Google podemos ver en primer lugar que la cantidad de casos, que aparecen representados por las columnas, es menor que en el resto de buscadores. Por otro lado podemos observar que este buscador tiende a recuperar documentos en que los términos se utilizan con poca frecuencia, pues la más elevada es la columna que representa a la aparición del término una sola vez, seguida por la columna relativa a la aparición de dos veces (en siete documentos), y así sucesivamente. Las mayores frecuencias corresponden a cuatro documentos en los que el término aparece nueve veces y en otro diez. En MSN, aparecen documentos con frecuencias similares a las de Google, pero con mayor variedad, sobre todo en lo que respecta a documentos con frecuencias de aparición del término de búsqueda elevadas. En Teoma podemos observar dos tendencias, una con documentos de altas frecuencias y otra parte con do-

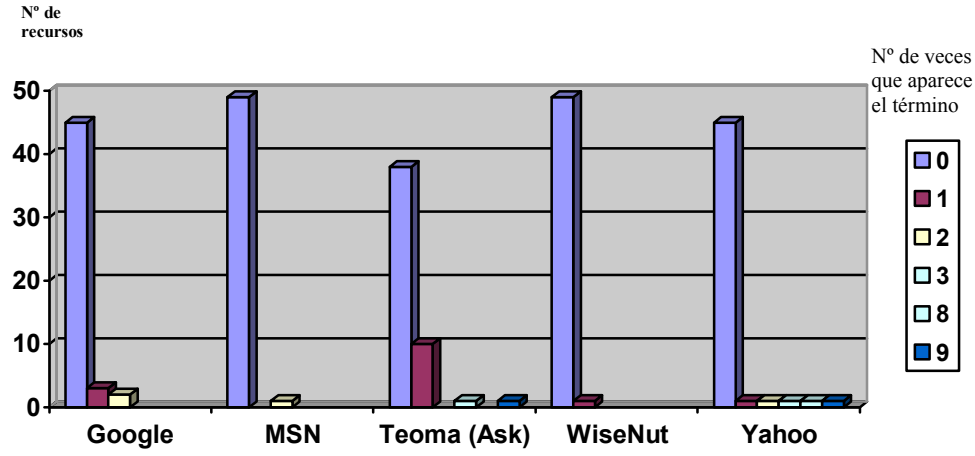
cumentos de frecuencias con valores más bajos. WiseNut se caracteriza por la recuperación de muchos documentos en los que el término aparece en pocas ocasiones y una variedad de documentos en los que aparece muy frecuentemente. Yahoo tiene un comportamiento más variado, destacando por la recuperación de un documento en el que el término se utiliza en cuarenta ocasiones.

Gráfico 5.2.2-2. Motores. Frecuencia del término "match"



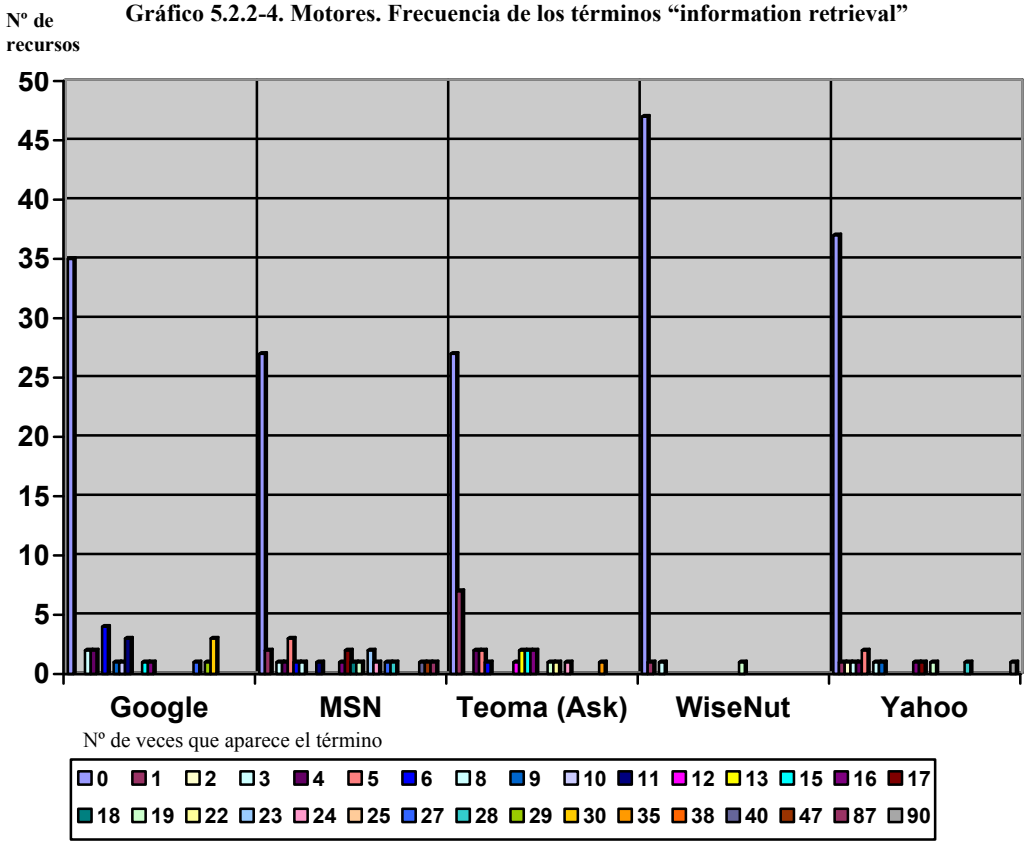
En la recuperación de este término hay cierta relación con lo señalado para el anterior, sobre todo en MSN y Yahoo si bien se observa un crecimiento en la recuperación de documentos en los que el término aparece solo en una ocasión o dos, como podemos apreciar en WiseNut. Google y Teoma recuperan en esta ocasión más recursos con mayores frecuencias de aparición del término.

Gráfico 5.2.2-3. Motores. Frecuencia del término "best-match"



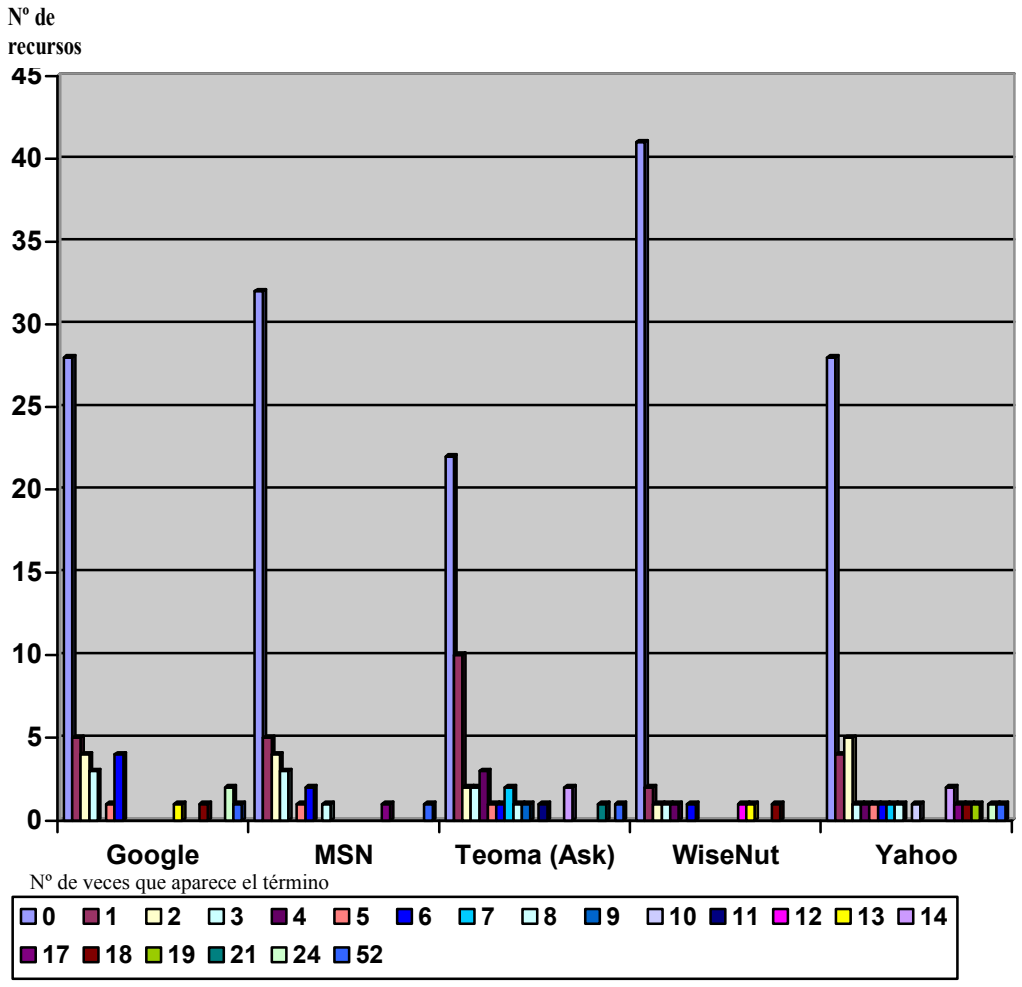
En este caso llama la atención el alto número de recursos que no contienen el término compuesto. Teoma es el buscador que recupera más páginas con el término seguido de Yahoo y Google, aunque todos ellos con una recuperación más que discreta, más aún si tenemos en cuenta que se trata de los términos más específicos de la búsqueda, lo que indica una baja precisión en la recuperación realizada por estas herramientas.

Gráfico 5.2.2-4. Motores. Frecuencia de los términos "information retrieval"



En la gráfica podemos observar que los valores de las frecuencias de aparición de estos términos en los documentos recuperados son variables alternando documentos de bajas frecuencias con otros de altas. Por otro lado el comportamiento de los motores es desigual, destacando el caso de WiseNut por los pocos recursos que contienen el término. Google casi alcanza el valor de cinco repeticiones en un mismo documento en varios de ellos mientras que en MSN, Teoma y Yahoo apenas aparecen tres veces en algunos de los documentos. Sin Embargo MSN y Yahoo recuperan otros documentos con altas frecuencias de aparición. Por otro lado Teoma recupera el mayor número de páginas en las que los términos aparecen sólo una vez.

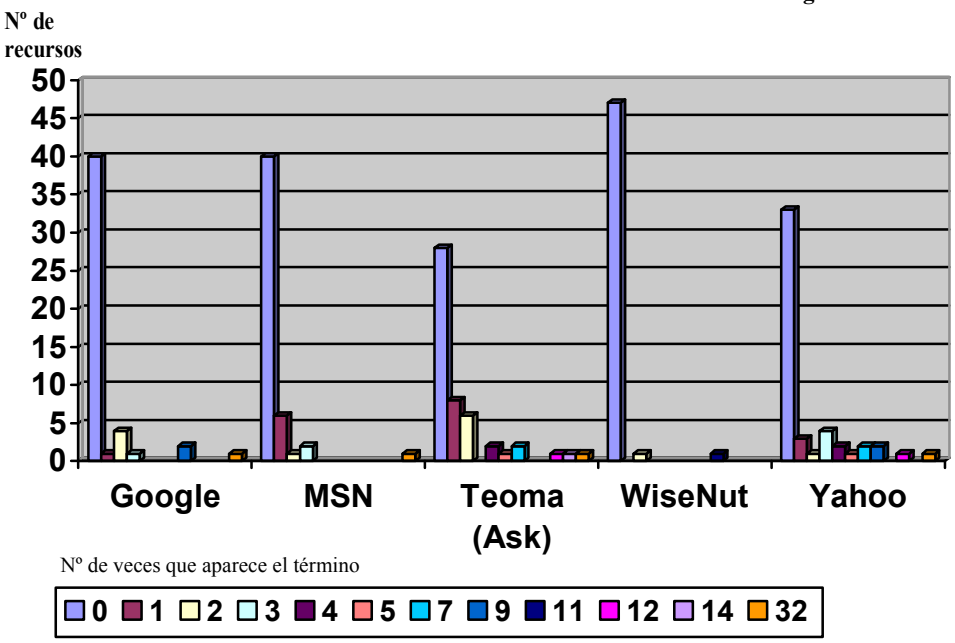
Gráfico 5.2.2-5. Motores. Frecuencia de los términos "web search"



La recuperación de documentos con estos términos es similar a la anterior aunque se aprecia un leve aumento de las frecuencias. También podemos hablar de la existencia de una cierta similitud entre los buscadores, aunque Teoma destaca en la recuperación

de más páginas en las que los términos aparecen con poca frecuencia, y por ser el motor que recupera un menor número de recursos que no contienen estos términos. Además, junto a Yahoo, es el motor que más variedad de frecuencias obtiene. Corresponden a WiseNut los peores resultados.

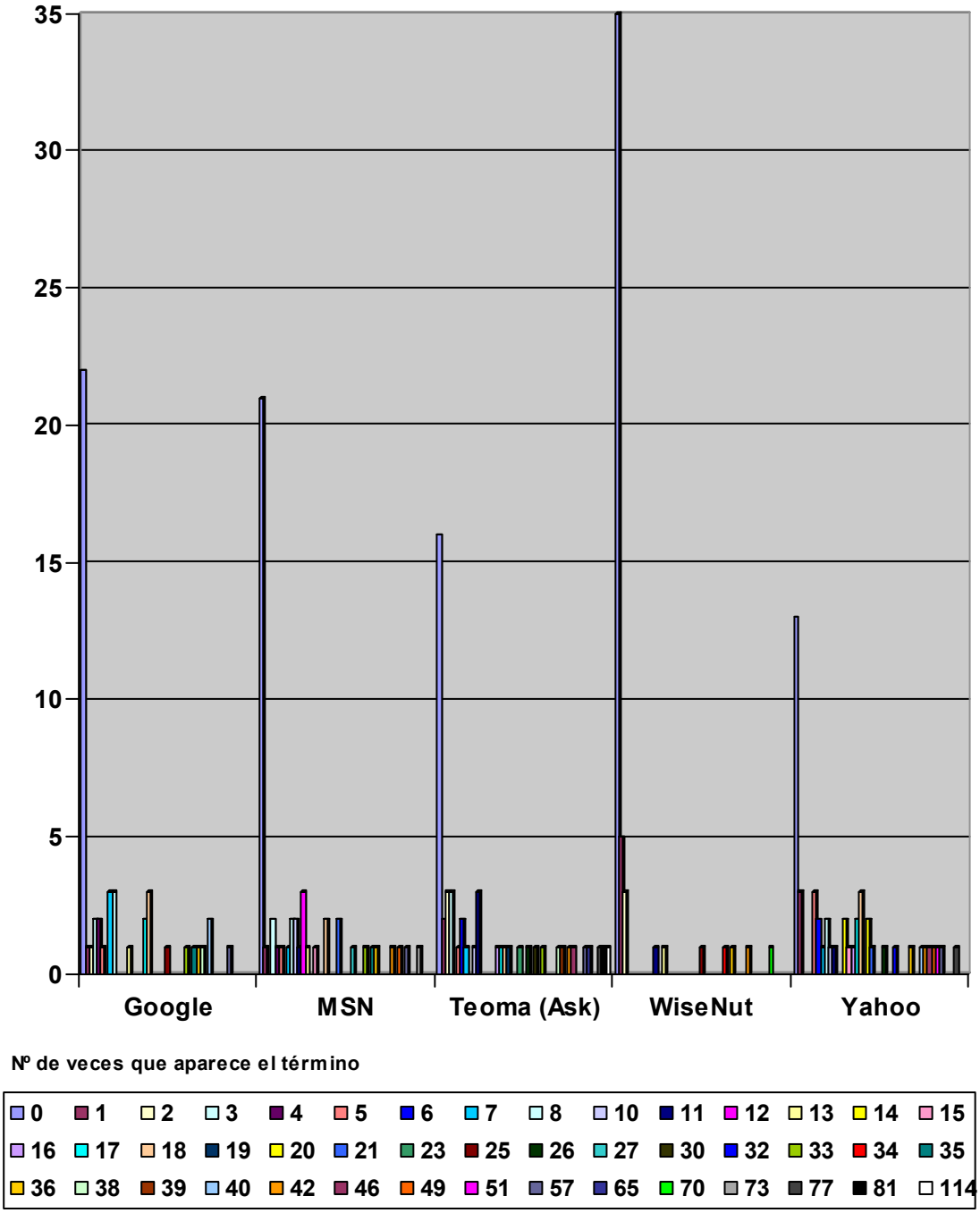
Gráfico 5.2.2-6. Motores. Frecuencia de los términos "web search engines"



De nuevo lo más destacado es la escasez de páginas que contienen los términos. Teoma y Yahoo son los motores con mayor número de páginas con los términos y mayor variedad de frecuencias. Todos los buscadores a excepción de WiseNut recuperan un recurso con la frecuencia más alta (32). MSN recupera un escaso número de páginas con los términos, y en su mayor parte sólo aparece la frase una vez.

Estos términos son más frecuentes en los documentos recuperados pues sólo en WiseNut destaca el número de recursos que no los contienen. Corresponde a Yahoo una mayor precisión técnica ya que ofrece el menor número de páginas que no contienen los términos de búsqueda al mismo tiempo que facilita un importante número de recursos con frecuencias de aparición variadas. En la misma línea, pero con mayor número de páginas sin los temas de búsqueda, podemos situar a MSN y Teoma, destacando éste último por recuperar el recurso con mayor frecuencia. La recuperación de Google tiene menos en cuenta la aparición de los términos.

Gráfico -5.2.2-7. Motores.Frecuencia de los términos "search engines"



En la presente gráfica podemos observar que a excepción de WiseNut, la existencia de estos términos en los recursos recuperados es más frecuente que los inmediatamente anteriores. La mayor precisión técnica corresponde a Yahoo y Teoma que son los que menor número de recursos recuperan sin los términos de búsqueda así como por la recuperación de un mayor número de documentos con las frecuencias más elevadas.

En definitiva, y dado que no se recuperaron recursos con todos los términos, si tenemos en cuenta las expresiones más específicas de la búsqueda como es el caso de "best-match", "information retrieval" y "web search engines", es muy poca la diferencia, a excepción del bajo comportamiento observado en WiseNut, que hay entre los motores de búsqueda, pues aunque en relación con los últimos términos podríamos destacar a Yahoo, por la recuperación del término "best-match" destaca Teoma (Ask).

5.2.3. Análisis individualizado por metabuscadores

Dogpile

Tabla 5.2.3-1. Nº de recursos analizados

Nº Recursos	43

Tabla 5.2.3-2. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web seacrh engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	43	100%

El metabuscador Dogpile no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.3-3. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	27,9%
1	6	14%
2	9	20,9%
3	5	11,6%
4	5	11,6%
5	1	2,3%
8	1	2,3%
9	1	2,3%
10	1	2,3%
12	1	2,3%
13	1	2,3%
Total	43	100%

El término no aparece en un 27,9% de los resultados que ofrece el buscador, lo que nos parece elevado ya que es superior al de la mayoría de motores de búsqueda, y en principio, son éstos los que sirven de fuente a los metabuscadores. En seis páginas aparece 1 vez (14%). En nueve documentos aparece dos veces (20,9%). Hay cinco recursos en los que aparece tres veces, y otros cinco en los que aparece cuatro veces. La mayor frecuencia corresponde a un documento en el que aparece 13 veces.

Tabla 5.2.3-4. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	17	39,5%
1	13	30,2%
2	4	9,3%
3	5	11,6%
4	1	2,3%
5	1	2,3%
6	1	2,3%
39	1	2,3%
Total	43	100%

El término no aparece en un 39,5% de los recursos. En 13 recursos aparece una vez (30,2%) y recupera un recurso en el que aparece en 39 ocasiones, manteniendo, salvo en este último aspecto, una recuperación similar a la del término anterior.

Tabla 5.2.3-5. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	41	95,3%
1	1	2,3%
9	1	2,3%
Total	43	100%

Es de destacar que 41 (95,3%) no contienen estos términos que, en un documento aparece sólo una vez y en otro nueve veces, y aún así es de los metabuscadores que recupera el término compuesto con mayor frecuencia.

Tabla 5.2.3-6. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	36	83,7%
4	1	2,3%
5	2	4,7%
16	1	2,3%
30	1	2,3%
35	2	4,7%
Total	43	100%

Estos términos tampoco son muy frecuentes en Dogpile, ya que en 36 (83,7%) no aparece y la máxima frecuencia se da en dos documentos, en los que aparece 35 veces. También en un documento aparecen en 30 ocasiones.

Tabla 5.2.3-7. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	24	55,8%
1	3	7%
2	1	2,3%
3	3	7%
5	2	4,7%
6	3	7%
8	1	2,3%
14	4	9,3%
15	1	2,3%
24	1	2,3%
Total	43	100%

En Dogpile el porcentaje de recursos recuperados sin estos términos es del 55,8% Los resultados se pueden agrupar en tres grupos correspondiendo al primero (30,3%) las menores frecuencias de aparición no superando 8 veces. En segundo lugar recursos en los que los términos aparecen 14 y 15 veces y finalmente un recurso en el que se repiten 24 veces.

Tabla 5.2.3-8. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	31	72,1%
1	3	7%
2	2	4,7%
4	2	4,7%
5	1	2,3%
6	1	2,3%
7	1	2,3%
9	1	2,3%
12	1	2,3%
Total	43	100%

La frecuencia de estos términos no es muy elevada pues las más elevadas corresponden a un documento en el que aparecen 12 veces, seguido de otro con 9 apariciones. La mayor frecuencia y porcentaje sigue perteneciendo a los documentos en los que no aparecen, esto es un total de 31 recursos (72,1%).

Tabla 5.2.3-9. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	22	51,2%
3	1	2,3%
4	1	2,3%
6	2	4,7%
8	1	2,3%
9	1	2,3%
10	1	2,3%
16	1	2,3%
17	2	4,7%
18	2	4,7%
33	1	2,3%
35	1	2,3%
36	2	4,7%
40	1	2,3%
42	1	2,3%
46	1	2,3%
49	1	2,3%
81	1	2,3%
Total	43	100%

Estos términos no aparecen en aproximadamente la mitad de los recursos recuperados (51,2%). No obstante las frecuencias son bastante altas ya que en un documento aparecen más de 81 veces, descendiendo la frecuencia a otros con 49, 46, 42, 40, etcétera.

Estos datos indican una baja precisión técnica y por tanto un mal funcionamiento ya que deberían ofrecer mejores resultados que los ofrecidos por los motores de búsqueda. Dogpile tiene un comportamiento similar al de los motores de búsqueda a los que, por su condición de metabuscador, debería superar ya que al seleccionar recursos de diferentes buscadores, el resultado debería ser cualitativamente superior al ofrecido por estos.

Excite

Tabla 5.2.3-10. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.3-11. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Este metabuscador no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.3-12. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12%
1	13	26%
2	10	20%
3	5	10%
4	7	14%
5	2	4%
6	1	2%
8	1	2%
9	1	2%
10	2	4%
12	1	2%
13	1	2%
Total	50	100%

En este metabuscador, en comparación con el anterior, descienden los porcentajes de no aparición de los términos, lo que denota un mejor funcionamiento.

Por tanto, la precisión técnica mejora, siendo similar a la mostrada por MSN. Aparecen recursos, tal vez los mismos que recupera el metabuscador Dogpile, con alta frecuencia de aparición como es el caso del recurso en que el término aparece trece veces.

Tabla 5.2.3-13. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	24%
1	16	32%
2	8	16%
3	5	10%
4	2	4%
5	2	4%
6	1	2%
7	1	2%
15	2	4%
39	1	2%
Total	50	100%

En este término ocurre lo mismo que en el anterior, aunque como también ocurre en Dogpile, se recupera un recurso con mayor frecuencia (39 veces) y dos recursos en los que la frecuencia es 15.

Tabla 5.2.3-14. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	47	94%
1	1	2%
2	1	2%
9	1	2%
Total	50	100%

Los resultados que ofrece son muy similares a los de Dogpile, aunque cualitativamente resulta destacable en esta ocasión, dada la escasez de recursos que contienen el término compuesto, recuperar un recurso más en el que el término aparece en dos ocasiones.

Tabla 5.2.3-15. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	36	72%
1	3	6%
3	1	2%
4	1	2%
5	2	4%
10	1	2%
15	1	2%
16	1	2%
23	1	2%
30	1	2%
31	1	2%
35	1	2%
Total	50	100%

Los porcentajes de recursos que no contienen estos términos, aunque similares da los de Google, siguen siendo elevados (72%) para un metabuscador.

No obstante los resultados superan a los que presenta Dogpile, al aparecer los términos en un mayor número de documentos.

Tabla 5.2.3-16. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	4	8%
2	6	12%
3	3	6%
4	1	2%
5	2	4%
6	3	6%
8	1	2%
11	1	2%
14	4	8%
15	1	2%
24	1	2%
Total	50	100%

La recuperación de estos términos es muy similar a la observada en Dogpile, si bien, las cifras de documentos en los que aparecen los términos con poca frecuencia son superiores en Excite, como ocurre en los seis documentos en los que aparecen dos veces.

Tabla 5.2.3-17. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	36	72%
1	4	8%
2	3	6%
4	2	4%
5	1	2%
7	2	4%
9	1	2%
12	1	2%
Total	50	100%

En este metabuscador el 72% de los recursos recuperados no contienen estos términos y en el 28% restante no aparecen en ningún documento más de 12 veces, por lo que las frecuencias no son nada elevadas. También aquí hay gran similitud con Dogpile.

Tabla 5.2.3-18. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	17	34%
1	4	8%
2	2	4%
3	1	2%
4	1	2%
6	2	4%
8	1	2%
9	1	2%
10	2	4%
13	1	2%
14	1	2%
16	1	2%
17	2	4%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
18	2	4%
20	1	2%
33	1	2%
35	1	2%
36	2	4%
38	1	2%
39	1	2%
40	1	2%
42	1	2%
46	1	2%
49	1	2%
81	1	2%
Total	50	100%

Son los términos que se recuperan con mayor frecuencia tanto por los motores como por los metabuscadores.

No obstante, podemos observar un descenso en el porcentaje de páginas que no los contienen. En este caso, en Excite se da una recuperación similar a la observada en Yahoo, con pocos recursos en los que los términos aparecen con bajas frecuencias y una mayor variedad en la que las frecuencias aumentan.

Excite muestra un comportamiento que supera ligeramente a Dogpile, dado que recupera un mayor número de recursos con los términos de búsqueda y con frecuencias mayores tal como acabamos de ver en los resultados de la última tabla.

Ixquick

Tabla 5.2.3-19. Nº de recursos analizados

Nº Recursos	39

Tabla 5.2.3-20. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	39	100%

Ixquick no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.3-21. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	15,4%
1	7	17,9%
2	11	28,2%
3	3	7,7%
4	7	17,9%
5	1	2,6%
9	1	2,6%
10	1	2,6%
12	1	2,6%
38	1	2,6%
Total	39	100%

Este metabuscador realiza, en relación a este término, una búsqueda similar a la de Excite si bien recupera más recursos en los que los términos aparecen en dos ocasiones. Además, recupera un documento en el que el término aparece con más frecuencia (38 veces), pudiéndose comparar en esta ocasión con Yahoo y WiseNut, que son lo únicos que recuperan un documento con esta misma frecuencia.

Tabla 5.2.3-22. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	7	17,9%
1	15	38,5%
2	7	17,9%
3	4	10,3%
4	3	7,7%
5	1	2,6%
6	2	5,1%
Total	39	100%

Frente a lo observado en el término anterior, en esta ocasión no recupera el recurso de mayor frecuencia, que sí aparecía en los anteriores metabuscadores.

Tabla 5.2.3-23. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	35	89,7%
1	3	7,7%
2	1	2,6%
Total	39	100%

Como ocurre con el resto de buscadores, el número de documentos recuperados y las frecuencias de aparición del término compuesto son extremadamente bajas.

Tabla 5.2.3-24. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	71,8%
1	2	5,1%
3	1	2,6%
4	1	2,6%
5	2	5,1%
6	1	2,6%
16	1	2,6%
19	1	2,6%
30	1	2,6%
35	1	2,6%
Total	39	100%

La recuperación en esta ocasión resulta equilibrada en cuanto a que no es elevado el número de recursos con bajas frecuencias que recupera, ni tampoco lo es el de las altas frecuencias, con dos documentos en los que los términos se repiten respectivamente 30 y 35 veces.

Tabla 5.2.3-25. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	15	38,5%
1	2	5,1%
2	6	15,4%
3	3	7,7%
4	1	2,6%
5	2	5,1%
6	3	7,7%
8	1	2,6%
11	1	2,6%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
13	2	5,1%
14	2	5,1%
24	1	2,6%
Total	39	100%

Ixquick tiene para estos términos mejor comportamiento que en los términos anteriores, ya que el porcentaje de recursos que no contienen los términos desciende al 38,5%, lo que influye en una recuperación de más recursos con frecuencias variadas.

Tabla 5.2.3-26. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	25	64,1%
1	2	5,1%
2	4	10,3%
4	2	5,1%
5	1	2,6%
7	2	5,1%
9	1	2,6%
10	1	2,6%
12	1	2,6%
Total	39	100%

Ixquick, en esta ocasión eleva el porcentaje de documentos sin los términos de búsqueda lo que se traduce en un descenso de la precisión técnica respecto a los metabuscadores anteriores.

Tabla 5.2.3-27. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	15	38,5%
2	3	7,7%
4	1	2,6%
6	1	2,6%
8	1	2,6%
13	2	5,1%
16	1	2,6%
17	2	5,1%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
18	1	2,6%
26	2	5,1%
35	1	2,6%
36	2	5,1%
38	1	2,6%
39	1	2,6%
40	1	2,6%
42	1	2,6%
46	1	2,6%
49	1	2,6%
81	1	2,6%
Total	39	100%

Ixquick se mantiene la tendencia observada para estos términos en los metabuscadores anteriores, si bien este metabuscador recupera menos páginas con valores de frecuencias intermedios, como podemos observar al compararlo con Excite.

Ixquick a pesar de que en esta búsqueda recuperó sólo 39 páginas web, tiene un comportamiento que, en determinados aspectos, supera a los dos anteriores, fundamentalmente en cuanto a la disminución del número de páginas que no contienen los términos de búsqueda.

Profusion

Tabla 5.2.3-28. Nº de recursos analizados

Nº Recursos	35

Tabla 5.2.3-29. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	35	100%

Este metabuscador no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.3-30. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	8,6%
1	7	20%
2	9	25,7%
3	3	8,6%
4	7	20%
5	1	2,9%
10	2	5,7%
13	1	2,9%
24	1	2,9%
78	1	2,9%
Total	35	100%

Profusión es, junto a Vivisimo, el metabuscador que menor número de recursos recupera que no contienen este término de búsqueda.

En este metabuscador aparece un documento con una alta frecuencia, aunque también recupera un elevado número de recursos con bajas frecuencias.

Tabla 5.2.3-31. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	8,6%
1	12	34,3%
2	4	11,4%
3	7	20%
4	4	11,4%
5	1	2,9%
15	3	8,6%
39	1	2,9%
Total	35	100%

Mantiene para este término características similares a las apreciadas para el término anterior.

Tabla 5.2.3-32. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	31	88,6%
1	2	5,7%
2	2	5,7%
Total	35	100%

Las frecuencias de aparición del término compuesto son menores que en Excite y Dogpile, ya que no se da ningún caso de nueve apariciones aunque recupera, como Ixquick, un recurso más que aquellos con el término compuesto.

Tabla 5.2.3-33. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	60%
1	3	8,6%
3	1	2,9%
4	1	2,9%
15	2	5,7%
16	2	5,7%
17	1	2,9%
19	1	2,9%
23	1	2,9%
35	2	5,7%
Total	35	100%

Profusión tiene un comportamiento irregular ya que teniendo en cuenta que el 60% de los recursos no contienen estos términos de búsqueda, en el 14,4% de recursos que le siguen los términos no aparecen más de cuatro veces. Finalmente, en el 23,6% restante los términos se repiten entre 15 y 23 veces y en dos documentos aparecen en 35 ocasiones.

En la recuperación de recursos con estos términos supera Dogpile e Ixquick siendo similar a Excite. Respecto a los motores de búsqueda, el comportamiento es similar a MSN.

Tabla 5.2.3-34. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	34,3%
1	3	8,6%
2	4	11,4%
3	5	14,3%
4	1	2,9%
5	1	2,9%
6	2	5,7%
8	1	2,9%
11	1	2,9%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
13	2	5,7%
14	1	2,9%
24	2	5,7%
Total	35	100%

Profusion, teniendo en cuenta que los resultados para esta búsqueda fueron limitados, presenta un porcentaje del 34,4% de páginas que no contienen los términos, recuperando mayor número de recursos con bajas frecuencias de aparición de los términos.

Tabla 5.2.3-35. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	20	57,1%
1	2	5,7%
2	6	17,1%
5	1	2,9%
7	1	2,9%
9	2	5,7%
10	2	5,7%
12	1	2,9%
Total	35	100%

Respecto a estos términos, el porcentaje de recursos que no los contienen, es del 57,1%, destacando por recuperar seis recursos en los que los términos aparecen en dos ocasiones. Por lo demás, el comportamiento es similar al del resto de metabuscadores.

Tabla 5.2.3-36. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	25,7%
2	3	8,6%
6	1	2,9%
10	3	8,6%
13	3	8,6%
16	1	2,9%
18	3	8,6%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
23	1	2,9%
33	1	2,9%
35	1	2,9%
36	1	2,9%
38	1	2,9%
39	1	2,9%
40	2	5,7%
42	1	2,9%
49	2	5,7%
81	1	2,9%
Total	35	100%

Los porcentajes de páginas que no contienen los términos descienden en esta ocasión al 25,7%, observándose menos páginas con bajas frecuencias que en el resto de metabuscadores. La recuperación en esta ocasión, se centra en la recuperación de los recursos con mayores frecuencias.

Profusion, como Ixquick tampoco recupera un gran número de recursos en esta búsqueda ofreciendo sólo 35 páginas web. No obstante da cierta importancia a la existencia de los términos de búsqueda, ya que es uno de los metabuscadores con menores porcentajes de recursos sin los términos de búsqueda, En general podemos apreciar que fundamentalmente presenta recursos con bajas frecuencias, aunque también de forma regular recupera recursos en los que los términos aparecen mayor número de veces.

Search

Tabla 5.2.3-37. Nº de recursos analizados

Nº Recursos	49

Tabla 5.2.3-38. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	45	91,8%
3	3	6,1%
4	1	2%
Total	49	100%

Es el único metabuscador en el que aparecen recursos con todos los términos ya que aparece en un documento en cuatro ocasiones y en otros tres, tres veces.

Tabla 5.2.3-39. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	7	14,3%
1	13	26,5%
2	12	24,5%
3	6	12,2%
4	4	8,2%
5	2	4,1%
6	2	4,1%
9	1	2%
12	2	4,1%
Total	49	100%

Aunque hay que destacar el bajo porcentaje de recursos que no contienen este término, Search se caracteriza por ofrecer mayor número de recursos en los que las frecuencias de aparición del término son bajas, ya que en el 51% de las páginas recuperadas el término aparece una vez o dos. Además no muestra documentos con frecuencias de aparición superior a 12 veces, lo que indica una baja precisión técnica en cuanto a este término.

Tabla 5.2.3-40. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	11	22,4%
1	17	34,7%
2	6	12,2%
3	7	14,3%
4	2	4,1%
6	2	4,1%
7	1	2%
15	1	2%
16	1	2%
39	1	2%
Total	49	100%

La recuperación de recursos con este término es prácticamente una copia del anterior.

Tabla 5.2.3-41. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	46	93,9%
1	1	2%
2	1	2%
9	1	2%
Total	49	100%

La recuperación de recursos con este término es similar a la mostrada por Excite.

Tabla 5.2.3-42. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	26	53,1%
1	2	4,1%
3	2	4,1%
4	1	2%
5	4	8,2%
6	1	2%
10	1	2%
15	1	2%
16	1	2%
17	2	4,1%
19	1	2%
23	2	4,1%
24	1	2%
27	1	2%
28	1	2%
30	1	2%
35	1	2%
Total	49	100%

En esta ocasión los términos parecen en un mayor número de documentos que en Excite. La cifra más representativa en este sentido, la ofrece el porcentaje de recursos que no contienen estos términos que en Search es el 53,1% frente a Excite que es el

72%. Por otro lado recupera más documentos con altas frecuencias de aparición de los términos, por lo que en relación con estos términos le corresponde una mayor precisión técnica que al resto.

Tabla 5.2.3-43. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	27	55,1%
1	6	12,2%
2	5	10,2%
3	3	6,1%
5	2	4,1%
6	1	2%
13	2	4,1%
14	1	2%
18	1	2%
24	1	2%
Total	49	100%

En Search el porcentaje de recursos que no contienen estos términos de búsqueda es del 55,1%, centrándose la recuperación en recursos cuya frecuencia de aparición de los términos es baja, pues en el 32,6% de las páginas, no supera las cinco veces.

Tabla 5.2.3-44. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	40	81,6%
1	2	4,1%
2	3	6,1%
4	1	2%
7	1	2%
9	1	2%
10	1	2%
Total	49	100%

El porcentaje de recursos que no contiene los términos es del 81,6% descendiendo por tanto no sólo las páginas que contienen los términos, sino también las frecuencias de aparición, que en este caso no superan las 10 veces en un documento.

Tabla 5.2.3-45. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	25	51%
1	2	4,1%
2	3	6,1%
8	2	4,1%
10	2	4,1%
12	1	2%
13	2	4,1%
17	2	4,1%
18	1	2%
25	1	2%
33	1	2%
35	2	4,1%
36	1	2%
40	1	2%
42	1	2%
46	1	2%
49	1	2%
Total	49	100%

En esta ocasión, hay un mayor número de páginas que no contienen los términos (51%) y no aparece el recurso recuperado por otros metabuscadores n el que aparecen los términos con mayor frecuencia (81).

Search tiene un comportamiento anómalo dado que a pesar de ser el único buscador que recupera recursos con todos los términos de búsqueda, se caracteriza por la recuperación de altos porcentajes sin los términos de búsqueda o con bajas frecuencias. En cualquier caso, es una herramienta de búsqueda a tener en cuenta ya que ofrece recursos con todos los términos y en el caso del término específico "*best-match*" tiene un comportamiento similar a Excite.

Surfwax

Tabla 5.2.3-46. Nº de recursos analizados

Nº Recursos	19

Tabla 5.2.3-47. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	100%

Este metabuscador no recupera páginas con todos los términos de la búsqueda

Tabla 5.2.3-48. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	7	36,8%
1	2	10,5%
2	3	15,8%
3	3	15,8%
4	1	5,3%
6	1	5,3%
12	1	5,3%
19	1	5,3%
Total	19	100%

Surfwax, aún recuperando un número reducido de recursos en esta búsqueda (19), presenta el mayor porcentaje de recursos sin este término.

Tabla 5.2.3-49. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	47,4%
1	6	31,6%
2	3	15,8%
6	1	5,3%
Total	19	100%

En los resultados relacionados con este término podemos observar como en el 47,4% de los recursos, cifra que coincide con el número de recursos que no contienen el término, la frecuencia de aparición del término no es superior a dos ocasiones y tan sólo en un documento aparece seis veces, lo que indica una baja precisión técnica.

Tabla 5.2.3-50. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	100%

Este metabuscador no recupera páginas con estos términos.

Tabla 5.2.3-51. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	17	89,5%
3	1	5,3%
19	1	5,3%
Total	19	100%

Las cifras en relación con estos términos denotan una deficiente recuperación, si se comparan con las cifras que ofrecen otros metabuscadores ya que el porcentaje de recursos recuperados que no contienen los términos es del 89,5%, apareciendo los términos tan sólo en dos documentos, 3 y 19 veces respectivamente.

Tabla 5.2.3-52. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	16	84,2%
1	1	5,3%
3	1	5,3%
13	1	5,3%
Total	19	100%

También en esta ocasión, la poca frecuencia de aparición de estos términos denota una deficiente recuperación, si se compara con las cifras que ofrecen otros metabuscadores.

Tabla 5.2.3-53. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	17	89,5%
2	2	10,5%
Total	19	100%

Los resultados en este metabuscador siguen siendo muy inferiores a los del resto de metabuscadores.

Tabla 5.2.3-54. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	16	84,2%
2	1	5,3%
36	1	5,3%
42	1	5,3%
Total	19	100%

En comparación con el resto de metabuscadores, aquí también podemos observar que la recuperación es inferior a ellos, pues estos términos aparecen con bajas frecuencias, ya que la máxima corresponde a un documento en el que aparecen 42 veces, cuando en Excite, Ixquick, Profusion, hay documentos en los que esta cifra se eleva a 81 casos, prácticamente el doble.

El reducido número de recursos dificulta la extracción de datos que señalen una determinada tendencia en el modo de recuperar recursos con los términos de búsqueda, no obstante, atendiendo a los porcentajes, es fácil observar que éste metabuscador recupera un alto índice de recursos sin los términos de búsqueda.

Vivisimo

Tabla 5.2.3-55. Nº de recursos analizados

Nº Recursos	50

Tabla 5.2.3-56. Frecuencia y nº de recursos en los que aparecen los términos "best-match information retrieval in web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Este metabuscador no recupera páginas con todos los términos de la búsqueda.

Tabla 5.2.3-57. Frecuencia y nº de recursos en los que aparece el término "best"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	6%
1	14	28%
2	8	16%
3	9	18%
4	4	8%
5	2	4%
6	2	4%
9	1	2%
10	2	4%
12	3	6%
38	2	4%
Total	50	100%

Destaca en Vivisimo la recuperación de un alto número de documentos en los que el término aparece con poca frecuencia, pues en el 60% de los recursos no aparece más de tres veces.

Tabla 5.2.3-58. Frecuencia y nº de recursos en los que aparece el término "match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	6%
1	17	34%
2	11	22%
3	9	18%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
4	4	8%
5	1	2%
6	1	2%
7	1	2%
12	1	2%
16	1	2%
39	1	2%
Total	50	100%

Tiene un comportamiento muy relacionado con lo observado para el término anterior.

Tabla 5.2.3-59. Frecuencia y nº de recursos en los que aparece el término "best-match"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100

Vivisimo no recupera recursos con el término compuesto.

Tabla 5.2.3-60. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	31	62%
1	2	4%
2	1	2%
3	1	2%
4	2	4%
5	4	8%
16	1	2%
17	2	4%
19	1	2%
23	2	4%
24	1	2%
27	1	2%
28	1	2%
Total	50	100%

También en Vivisimo es importante el porcentaje de recursos que no contienen los términos (62%), agrupándose en dos apartados los resultados. En el primero de ellos, que representa el 20%, las frecuencias de aparición no superan las cinco veces y en el segundo, que comprende el 18%, varían entre dieciséis y veintiocho veces, valores que no son tan elevados como los de otros metabuscadores.

Por tanto, aunque su comportamiento pueda asimilarse al de Excite, en los documentos de frecuencias altas es claramente inferior su recuperación.

Tabla 5.2.3-61. Frecuencia y nº de recursos en los que aparecen los términos "web search"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	8	16%
2	3	6%
3	3	6%
4	1	2%
5	2	4%
6	1	2%
8	1	2%
9	1	2%
11	1	2%
13	1	2%
14	2	4%
18	1	2%
21	1	2%
52	1	2%
Total	50	100%

Respecto a los términos anteriores, el porcentaje de recursos sin estos términos desciende al 46%, elevándose al 16% las páginas en las que sólo aparecen en una ocasión. Destaca la recuperación de un recurso en el que los términos aparecen en 52 ocasiones.

Tabla 5.2.3-62. Frecuencia y nº de recursos en los que aparecen los términos "web search engines"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	34	68%
1	4	8%
2	4	8%
3	1	2%
4	1	2%
5	1	2%
7	2	4%
12	1	2%
14	1	2%
32	1	2%
Total	50	100%

Vivisimo muestra unos resultados muy relacionados con lo observado para los términos anteriores, aunque recupera más recursos con los términos de búsqueda que Search, Excite y Dogpile.

Tabla 5.2.3-63. Frecuencia y nº de recursos en los que aparecen los términos "search engines"

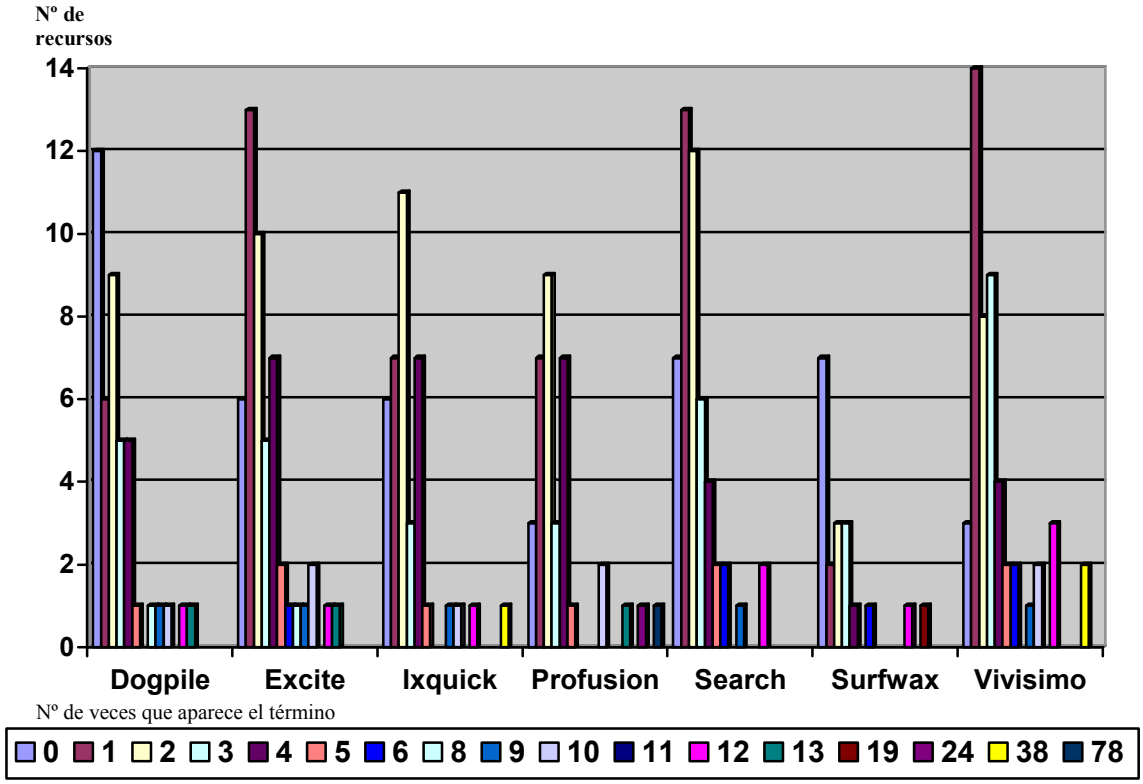
Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	1	2%
2	3	6%
5	1	2%
6	1	2%
8	1	2%
10	2	4%
12	1	2%
15	1	2%
16	1	2%
18	1	2%
21	1	2%
25	1	2%
33	1	2%
35	1	2%
36	2	4%
39	1	2%
42	1	2%
46	1	2%
49	1	2%
57	1	2%
65	1	2%
77	1	2%
81	1	2%
Total	50	100%

Vivisimo muestra aquí un comportamiento similar al de Excite, aunque en esta ocasión recupera un mayor número de páginas sin los términos de búsqueda.

Por tanto, Vivisimo se caracteriza por presentar un comportamiento desigual, ya que en unos casos mantiene altos porcentajes de recursos con los términos de búsqueda y en otros no. Su precisión técnica se resiente también respecto a otros metabuscadores en la recuperación del término compuesto "*best-match*" al no recuperar páginas que lo contengan, aunque los términos "web search engines" aparecen con mayor frecuencia en los documentos recuperados por este metabuscador que en Dogpile, Excite y Search.. Respecto a los documentos con altas frecuencias, también tiene un comportamiento irregular, aunque mejora cuando se trata de varios términos.

5.2.4. Análisis comparativo de los metabuscadores

Gráfico 5.2.4-1. Metabuscadores. Frecuencia del término "best"

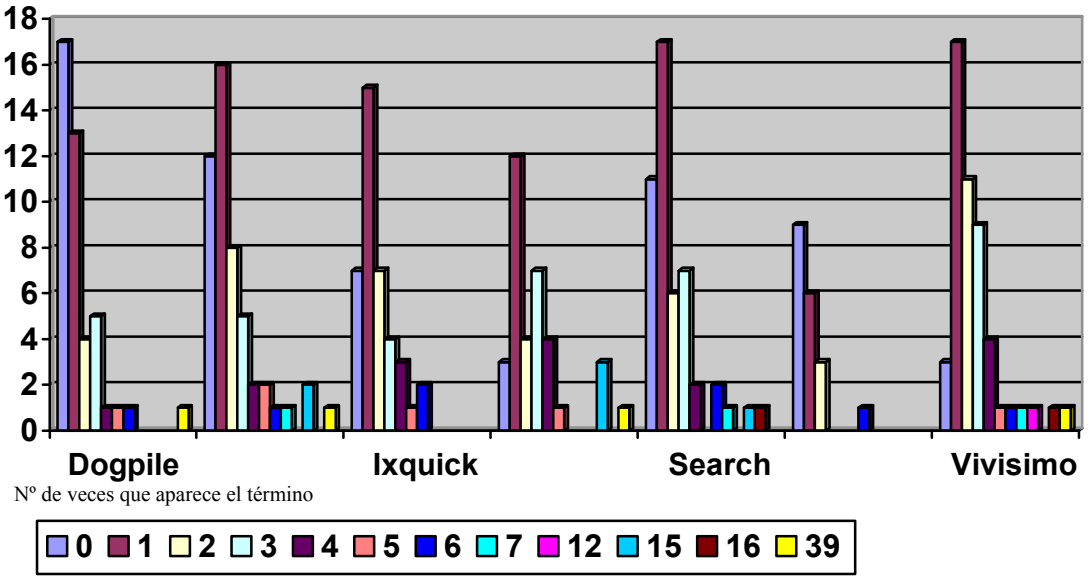


Vivisimo es el metabuscador que recupera mayor número de recursos con este término, correspondiéndole también la recuperación de mayor número de documentos

con bajas frecuencias, lo que incide en un mayor número de páginas con frecuencias medias e incluso en la recuperación de una página en la que el término aparece con gran frecuencia (38). No obstante hay que advertir que se trata de un término genérico, por lo que resulta aventurado extraer en base a ello grandes conclusiones.

Excite también recupera un gran número de recursos con frecuencias de aparición bajas, pero a diferencia de Vivisimo no recupera documentos con altas frecuencias, pareciéndose más a la recuperación que hacen Dogpile y Search, aunque el primero de éstos, recupera un gran número de recursos que no contienen el término. En Search predominan los resultados con bajas frecuencias al igual que en el resto. Profusión recupera recursos tanto con altas como con bajas frecuencias, destacando por ser el metabuscador que recupera el recurso en el que el término aparece con mayor frecuencia (78).

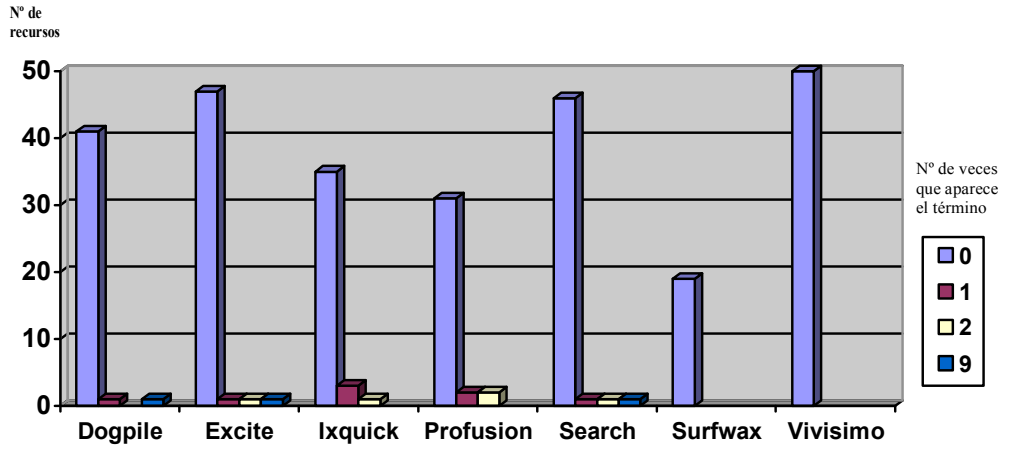
Gráfico 5.2.4-2. Metabuscadores. Frecuencia del término "match"



Vivisimo de nuevo recupera el mayor número de recursos con el término, y combina la recuperación de documentos con bajas frecuencias con otros de frecuencias altas, y todo ello a pesar de que hay una menor recuperación de recursos con estos términos, posiblemente porque no es tan común como el anterior. Por tanto sigue predominando la recuperación de documentos con bajas frecuencias, aunque Dogpile, Excite, Profusión, y Vivisimo recuperan un registro de alta frecuencia de aparición del término (39).

La recuperación en Profusión es equilibrada en cuanto a que recupera recursos tanto con altas como con bajas frecuencias, aunque la precisión es inferior a la registrada por Vivisimo.

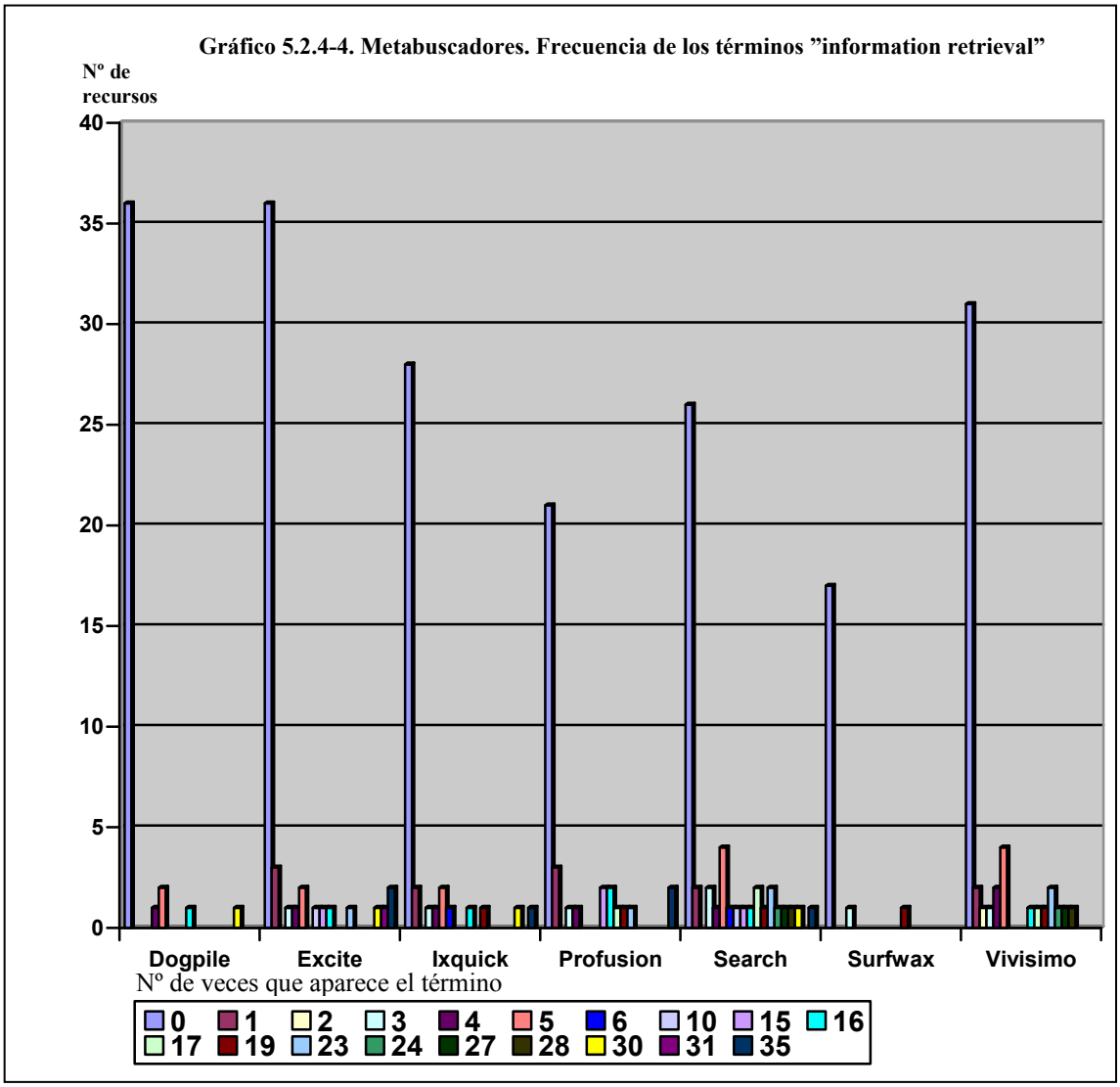
Gráfico 5.2.4-3. Metabuscadores. Frecuencia del término "best-match"



Surfwax mantiene los menores niveles de precisión técnica respecto al resto de metabuscadores.

Las frecuencias de aparición del término compuesto, al igual que ocurría en los motores de búsqueda, también son mínimas en los metabuscadores. Hay cierta similitud entre Excite, Dogpile y Search, correspondiendo por su parte a Ixquick una mayor semejanza con Profusion, en cuanto a la recuperación de registros con bajas frecuencias. Vivisimo y Surfwax no recuperan recursos con el término compuesto.

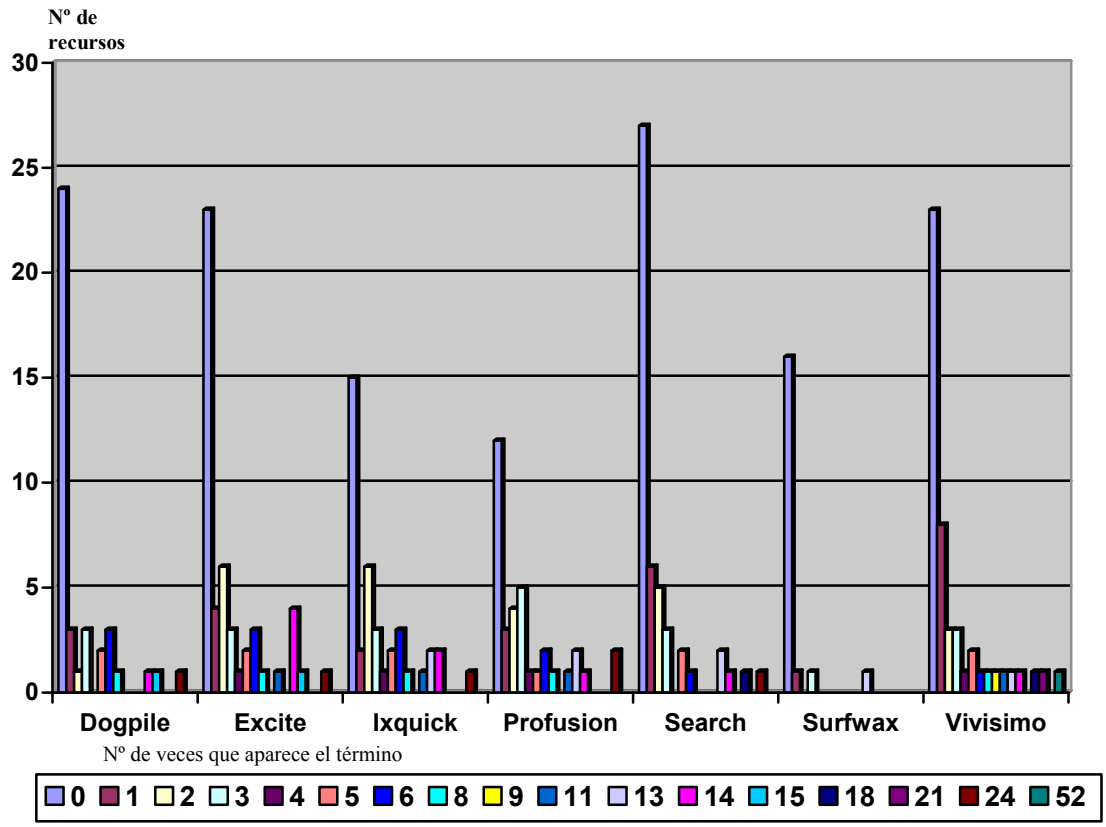
Gráfico 5.2.4-4. Metabuscadores. Frecuencia de los términos "information retrieval"



La recuperación de los términos de esta expresión adquiere unos valores muy bajos tanto en el número de documentos que las contienen como en su frecuencia dentro de ellos. Search, Excite y Vivisimo son los metabuscadores en los que se observan las mayores frecuencias, destacando en este caso el primero de ellos al recuperar mayor número de recursos con estos términos.

Profusión, Dogpile y sobre todo Surfwax son los que ofrecen recursos con menores frecuencias.

Gráfico 5.2.4-5. Metabuscadores. Frecuencia de los términos "web search"



Dado que el uso de estos términos es más frecuente que los inmediatamente anteriores, las frecuencias de aparición, en esta ocasión, son más altas.

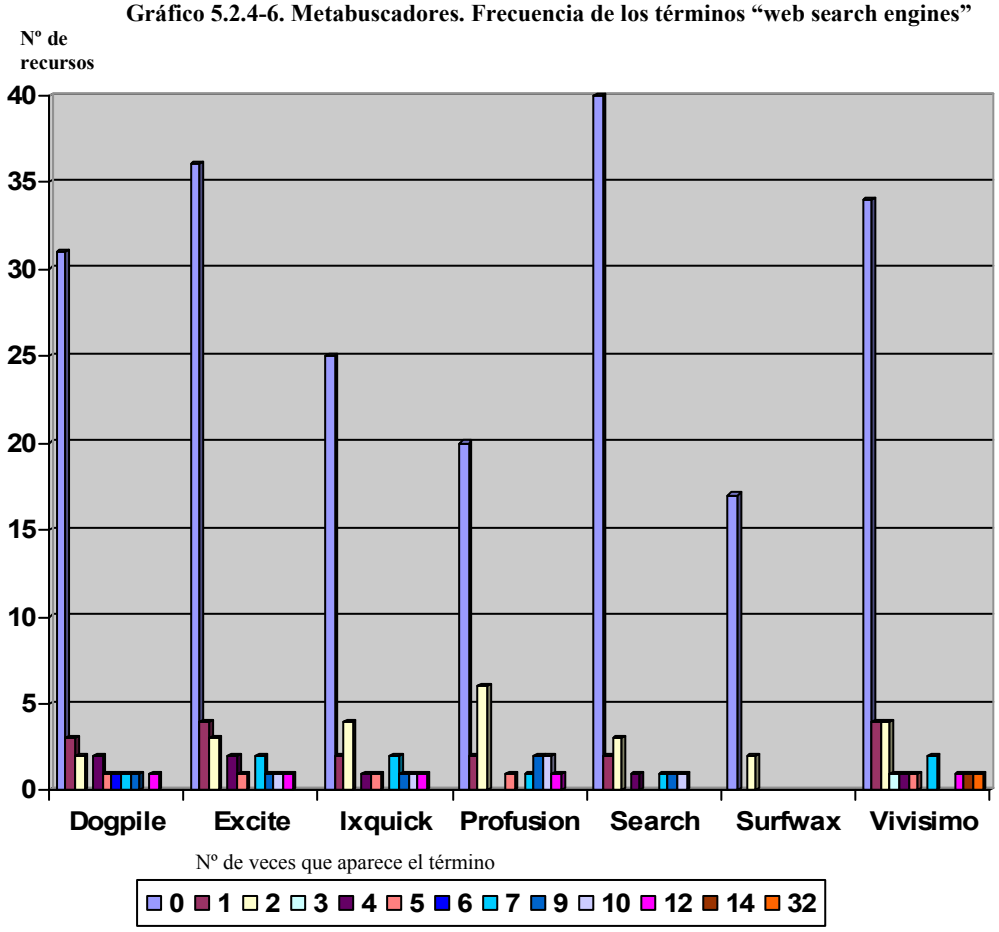
A grandes rasgos, el comportamiento en la recuperación de documentos con estos términos de la mayoría de metabuscadores es muy similar, aunque podemos apreciar pequeñas diferencias. En este caso, entre Excite e Ixquick, se observa en este último, que la aparición de páginas con estos términos es más frecuente. Por su parte, Dogpile recupera un mayor número de recursos con frecuencias intermedias. Profusion no dista mucho de lo señalado para el resto, si bien hay que tener en cuenta que las cantidades son menores al ser menor el número de registros recuperados en esta búsqueda.

Search destaca tanto por el bajo número de documentos que no contienen estos términos como por obtener más recursos con bajas frecuencias.

Surfwax, como hemos venido observando, mantiene un bajo comportamiento también en la recuperación de estos términos.

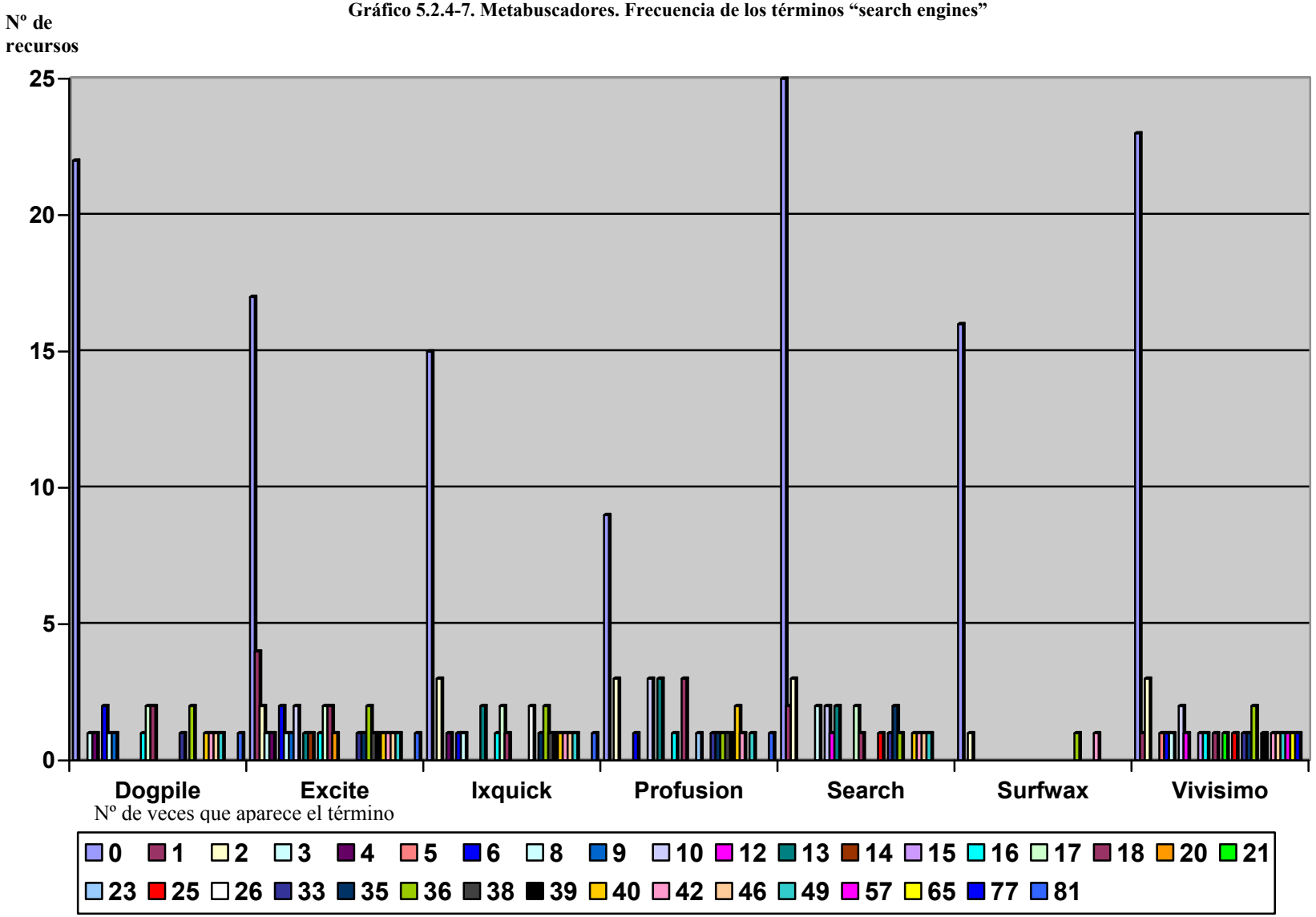
Vivisimo cuenta con un importante número de páginas que no contienen los términos o aparecen con baja frecuencia.

Gráfico 5.2.4-6. Metabuscadores. Frecuencia de los términos "web search engines"



Respecto a estos términos, los metabuscadores también guardan similitud, aunque en el caso de Search llama la atención el alto número de páginas que no contiene estos términos. Vivisimo proporciona de nuevo el acceso a un registro del máximo nivel de frecuencia (32). En definitiva no se aprecian grandes diferencias entre Dogpile, Excite y Vivisimo.

Gráfico 5.2.4-7. Metabuscadores. Frecuencia de los términos "search engines"



Search sigue mostrando el mayor nivel de recursos sin los términos de búsqueda, aunque los valores son similares a los del resto. En Excite y Vivisimo se aprecian las mayores frecuencias de recuperación de estos términos, siendo menores en Dogpile, Ixquick, Profusión y sobre todo Surfwax.

En definitiva, a la vista de los datos analizados podemos decir que no se observa la existencia de un metabuscador que destaque frente a los demás, aunque la mayor precisión técnica corresponde a Excite y Vivisimo, en el caso de éste último, por facilitar el acceso a recursos con alta frecuencia de aparición de los términos de búsqueda, seguidos por Profusion, que se caracteriza por recuperar un menor número de recursos sin los términos.

Análisis global

Teniendo en cuenta que la primera búsqueda sirve para calcular la precisión técnica de todos los buscadores evaluados, y que como hemos visto, el índice de precisión no es muy elevado, el resto de búsquedas, y concretamente el análisis de frecuencias de los términos de búsqueda, nos permiten apreciar que sólo el metabuscador Search recupera recursos con todos los términos de búsqueda. En relación con el funcionamiento en la recuperación de estas herramientas, lo primero que llama la atención es el hecho de que

Por otro lado, también hemos podido observar que las frecuencias de aparición de los términos en los recursos recuperados por los metabuscadores son inferiores a las de los motores, por lo que son más recomendables éstos últimos cuando lo que se requiere son recursos de mayor precisión técnica. No obstante también hay que tener en cuenta, que el número de recursos que no contienen los términos de búsqueda es menor en los metabuscadores que en los motores, por lo que podemos afirmar que los metabuscadores conceden importancia a que los recursos que recuperan de los motores contengan, en la medida de lo posible, los términos de búsqueda.

5.3. Búsqueda con operadores de existencia

El análisis se realiza en función de las frecuencias de los siguientes términos o frases:

Búsqueda completa:

1. information retrieval systems +and +the web

Términos y frases:

2. information retrieval systems
3. information retrieval
4. web
5. information

5.3.1. Análisis individualizado por motores de búsqueda

Google

Tabla 5.3.1-1. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.1-2. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Google no recuperó entre los 50 primeros recursos ningún documento con todos los términos.

Tabla 5.3.1-3. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	39	78%
1	7	14%
2	2	4%
4	1	2%
6	1	2%
Total	50	100%

Un 78% de los recursos recuperados no contienen estos términos, lo que constituye un alto porcentaje. En siete documentos aparece una vez y en dos documentos aparece dos veces. En uno aparece cuatro veces, y en otro seis.

Estos datos no resultan nada positivos de cara a la precisión técnica ya que son claramente limitados.

La recuperación es similar a la que ofrece Excite, si bien el metabuscador supera a Google al recuperar dos recursos en el que estos términos de búsqueda aparecen en 11 ocasiones.

Tabla 5.3.1-4. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	26	52%
1	2	4%
2	4	8%
4	2	4%
6	3	6%
9	2	4%
11	1	2%
12	1	2%
13	1	2%
19	1	2%
22	1	2%
23	1	2%
29	1	2%
30	2	4%
32	1	2%
36	1	2%
Total	50	100%

El porcentaje de páginas que contiene estos términos aumenta en esta ocasión, aunque el hecho que un 52% no los tenga, nos sigue pareciendo un porcentaje elevado. Los porcentajes de frecuencias altas y bajas son diferentes siendo mayores los de frecuencias bajas. La recuperación es similar a Excite.

Tabla 5.3.1-5. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	13	26%
1	5	10%
2	1	2%
3	4	8%
4	6	12%
5	2	4%
6	2	4%
7	3	6%
8	2	4%
11	1	2%
12	1	2%
13	1	2%
14	1	2%
18	2	4%
20	1	2%
25	1	2%
28	1	2%
39	1	2%
82	1	2%
127	1	2%
Total	50	100%

Como podemos observar se trata de un término muy genérico, de aquí que las frecuencias sean superiores tanto dentro de los documentos como en el número de documentos que los contienen.

Una vez más la recuperación es similar a la de Excite, pero en esta ocasión tampoco recupera documentos con frecuencias tan altas como las recuperadas por aquél (196) frente a 127 en Google.

Tabla 5.3.1-6. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	4	8%
1	2	4%
3	3	6%
4	2	4%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
5	2	4%
6	3	6%
7	4	8%
9	2	4%
11	2	4%
12	1	2%
13	2	4%
15	1	2%
17	2	4%
18	1	2%
20	2	4%
21	1	2%
25	1	2%
28	1	2%
29	1	2%
31	1	2%
34	1	2%
35	2	4%
43	1	2%
47	3	6%
63	1	2%
67	2	4%
104	1	2%
117	1	2%
Total	50	100%

Para este término, sólo un 8% de recursos no lo contienen, lo que influye en una buena precisión técnica al no concentrarse los datos en frecuencias bajas, sino que hay un mayor reparto de frecuencias entre los documentos recuperados.

Supera en esta ocasión a Excite en el número de páginas que contienen el término de búsqueda. Sin embargo, en el metabuscador, las frecuencias de aparición de los términos siguen siendo superiores.

Los resultados de esta búsqueda constatan lo observado en la anterior en cuanto a la baja precisión técnica de este buscador para lo que también aquí podemos tener especialmente en cuenta los resultados obtenidos en las frecuencias de los términos "*information retrieval systems*"(Tabla 5.1-114).

MSN

Tabla 5.3.1-7. Nº de recursos analizados

Nº Recursos	48

Tabla 5.3.1-8. Nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	48	100%

MSN tampoco recuperó ningún documento con todos los términos.

Tabla 5.3.1-9. Nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	43	89,6%
1	5	10,4%
Total	48	100%

La frecuencia de aparición de los términos en los documentos es inferior a los recuperados por Google, pues un 89,6% de los recursos recuperados no los contiene. Sólo recupera 5 recursos en los que aparecen una vez.

Tabla 5.3.1-10. Nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	29	60,4%
1	4	8,3%
2	3	6,3%
3	1	2,1%
4	1	2,1%
5	1	2,1%
6	1	2,1%
7	1	2,1%
9	1	2,1%
11	1	2,1%
12	2	4,2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
29	1	2,1%
30	1	2,1%
32	1	2,1%
Total	48	100%

El porcentaje de recursos que no contienen los términos es superior a Google lo que unido a una recuperación de mayor número de recursos de bajas frecuencias y menor de los de altas frecuencias, tiene como consecuencia unos peores resultados en cuanto a la precisión técnica.

Tabla 5.3.1-11. Nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	13	27,1%
1	2	4,2%
2	3	6,3%
3	3	6,3%
4	1	2,1%
5	7	14,6%
6	3	6,3%
8	1	2,1%
10	1	2,1%
11	1	2,1%
12	2	4,2%
14	1	2,1%
18	2	4,2%
20	1	2,1%
28	1	2,1%
31	1	2,1%
32	1	2,1%
41	1	2,1%
43	1	2,1%
61	1	2,1%
81	1	2,1%
Total	48	100%

Recuperación similar a Google aunque este último recupera un recurso en el que el término aparece 128 veces y en MSN la mayor frecuencia es de 81.

Tabla 5.3.1-12. Nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	1	2,1%
1	5	10,4%
2	4	8,3%
3	5	10,4%
4	4	8,3%
5	2	4,2%
6	3	6,3%
8	1	2,1%
9	1	2,1%
10	3	6,3%
12	1	2,1%
13	3	6,3%
14	1	2,1%
15	1	2,1%
17	1	2,1%
18	1	2,1%
19	1	2,1%
20	1	2,1%
25	1	2,1%
29	2	4,2%
47	2	4,2%
49	1	2,1%
100	1	2,1%
104	1	2,1%
120	1	2,1%
Total	48	100%

Recuperación similar a Google aunque en MSN sólo un 2,1% no contiene este término, lo que supone un gran contraste si lo comparamos con los términos anteriores.

El análisis pormenorizado de los términos que componen esta búsqueda permite observar las carencias en la precisión de estos buscadores, ya que llama la atención el bajo porcentaje de recursos que contienen los términos "*information retrieval systems*" (Tabla 5.1-120) frente a la frecuencia de documentos con el término "*information*". En general, la recuperación de este buscador es similar a la observada en la búsqueda anterior.

Teoma (Ask)

Tabla 5.3.1-13. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.1-14. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Teoma (Ask) no recuperó ningún documento con todos los términos.

Tabla 5.3.1-15. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	41	82%
1	5	10%
2	3	6%
11	1	2%
Total	50	100%

La recuperación de recursos con estos términos es baja, aunque destaca frente a los anteriores buscadores por facilitar un recurso en el que los términos aparecen 11 veces.

Tabla 5.3.1-16. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	4	8%
2	2	4%
3	2	4%
4	2	4%
5	2	4%
6	4	8%
7	1	2%
11	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
13	1	2%
15	1	2%
16	1	2%
23	1	2%
26	2	4%
29	1	2%
31	1	2%
43	1	2%
Total	50	100%

La frecuencia de documentos con estos términos es mayor que la de los anteriores lo que incide en la variedad de frecuencias con que aparecen en los documentos. En este sentido, supera a la precisión técnica aportada por Google en función de los porcentajes y las frecuencias de los términos, como muestra la recuperación de un recurso en el que los términos aparecen en 43 ocasiones.

Tabla 5.3.1-17. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	5	10%
1	3	6%
2	3	6%
3	1	2%
4	1	2%
5	5	10%
7	4	8%
8	2	4%
12	1	2%
13	2	4%
14	1	2%
15	1	2%
20	1	2%
22	1	2%
28	1	2%
30	1	2%
31	1	2%
37	1	2%
41	1	2%
44	1	2%
64	1	2%
77	1	2%
81	1	2%
87	1	2%
98	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
99	1	2%
106	1	2%
119	1	2%
124	1	2%
125	1	2%
127	1	2%
137	1	2%
177	1	2%
Total	50	100%

Los porcentajes de páginas que no contienen este término son los más bajos, observándose un gran contraste con Yahoo.

Esto influye en la variedad de páginas recuperadas con distintas frecuencias, superando a las ofrecidas por Google.

Tabla 5.3.1-18. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	6%
1	2	4%
2	1	2%
3	2	4%
5	2	4%
6	7	14%
7	4	8%
10	1	2%
12	1	2%
14	1	2%
15	2	4%
16	1	2%
22	1	2%
25	2	4%
28	1	2%
29	1	2%
33	1	2%
35	1	2%
40	1	2%
41	1	2%
43	1	2%
47	1	2%
49	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
50	1	2%
51	1	2%
52	1	2%
53	1	2%
60	1	2%
70	1	2%
75	1	2%
77	1	2%
104	1	2%
112	1	2%
113	1	2%
Total	50	100%

En la recuperación de este término, de nuevo las páginas con bajas frecuencias son inferiores a Google, aunque la recuperación de documentos en los que aparece el término seis veces lo diferencia del anterior, siendo similares en el resto de resultados.

El comportamiento de Teoma se caracteriza por unas frecuencias de aparición de los términos analizados superiores a las de Google y Yahoo, lo que nos hace pensar que se trata de un buscador que pondera con preferencia los recursos que contienen los términos de búsqueda, todo ello teniendo en cuenta, como ocurre en el resto de buscadores, el mal funcionamiento en la recuperación con el operador de existencia (+).

WiseNut

Tabla 5.3.1-19. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.1-20. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

WiseNut no recuperó ningún documento con todos los términos.

Tabla 5.3.1-21. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	76%
1	5	10%
2	2	4%
3	1	2%
6	2	4%
11	1	2%
12	1	2%
Total	50	100%

En la recuperación de estos términos muestra cierta similitud con Teoma, aunque lo supera al recuperar dos documentos con frecuencias de seis apariciones y otro de tres, que indican un mejor comportamiento de WiseNut en la recuperación de estos términos.

Supera a Google en la recuperación de recursos con los tres términos ya que recupera dos recursos con doce y once apariciones, frente al máximo de Google que son seis apariciones.

Tabla 5.3.1-22. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	42%
1	6	12%
2	3	6%
3	3	6%
4	3	6%
6	1	2%
7	1	2%
9	2	4%
10	1	2%
11	3	6%
12	1	2%
15	1	2%
17	1	2%
23	1	2%
27	1	2%
36	1	2%
Total	50	100%

Para estos términos WiseNut tiene un comportamiento similar a Google y MSN, aunque la recuperación de recursos de bajas frecuencias es superior a los anteriores.

Tabla 5.3.1-23. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12%
1	6	12%
2	3	6%
3	3	6%
4	1	2%
5	3	6%
6	2	4%
7	2	4%
8	2	4%
9	2	4%
10	2	4%
14	1	2%
15	1	2%
16	1	2%
17	2	4%
18	1	2%
19	4	8%
20	1	2%
23	1	2%
37	1	2%
39	1	2%
44	1	2%
46	1	2%
66	1	2%
77	1	2%
Total	50	100%

Recursos con el término "web" son, como en el caso de Teoma, ampliamente recuperados por este motor.

Tabla 5.3.1-24. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12%
1	2	4%
3	4	8%
4	2	4%
5	3	6%
6	2	4%
7	3	6%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
8	2	4%
9	2	4%
12	4	8%
13	3	6%
14	1	2%
18	1	2%
19	1	2%
20	2	4%
25	1	2%
30	1	2%
31	1	2%
33	1	2%
46	2	4%
50	2	4%
59	1	2%
63	1	2%
66	1	2%
75	1	2%
Total	50	100%

Aunque WiseNut muestra altas frecuencias de aparición del término, son inferiores a las de Google, MSN y Teoma.

WiseNut tiene un comportamiento que mejora respecto a los anteriores en la recuperación de los términos "*information retrieval systems*" además de presentar bajos porcentajes de documentos que no contienen el resto de términos analizados por lo que ofrece un amplio número de recursos con diferentes frecuencias, sin superar los resultados de Teoma.

Yahoo

Tabla 5.3.1-25. **Nº de recursos analizados**

Nº Recursos	50

Tabla 5.3.1-26. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Yahoo no recuperó ningún documento con todos los términos.

Tabla 5.3.1-27. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	38%
1	8	16%
2	5	10%
3	7	14%
4	5	10%
5	1	2%
6	1	2%
7	1	2%
12	1	2%
13	1	2%
14	1	2%
Total	50	100%

En Yahoo disminuyen los recursos que no contienen los términos de búsqueda. Ello mejora la precisión técnica al recuperar más documentos con altas frecuencias de aparición de los términos, superando a Google y WiseNut, y de forma más clara a MSN y Teoma.

Tabla 5.3.1-28. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	15	30%
1	3	6%
2	3	6%
3	6	12%
4	2	4%
7	5	10%
8	2	4%
10	2	4%
11	1	2%
13	1	2%
16	2	4%
19	1	2%
20	1	2%
22	1	2%
25	1	2%
32	1	2%
36	1	2%
57	1	2%
97	1	2%
Total	50	100%

También aquí supera en ambos índices a Google, WiseNut, MSN y Teoma.

Tabla 5.3.1-29. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	31	62%
1	10	20%
2	1	2%
3	1	2%
4	2	4%
6	2	4%
7	1	2%
16	1	2%
39	1	2%
Total	50	100%

Es el buscador que más páginas recupera sin este término, convirtiéndolo en el buscador que peor precisión técnica ofrece para este término.

Tabla 5.3.1-30. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	18%
1	4	8%
2	1	2%
3	5	10%
4	1	2%
6	2	4%
7	4	8%
9	2	4%
10	1	2%
13	1	2%
14	4	8%
15	1	2%
16	1	2%
18	1	2%
20	1	2%
24	1	2%
25	1	2%
28	1	2%
33	1	2%
35	2	4%
47	1	2%
63	1	2%
90	2	4%
170	1	2%
174	1	2%
Total	50	100%

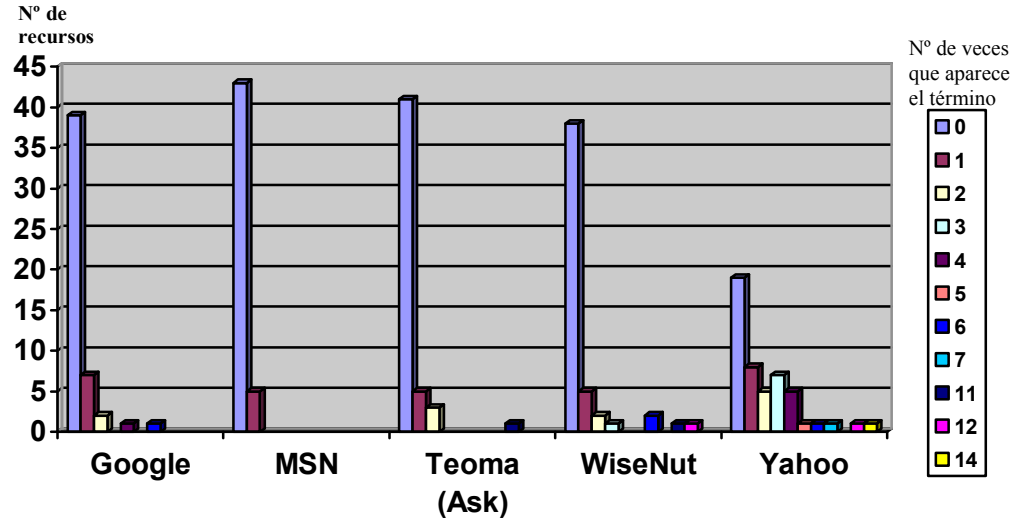
El comportamiento de Yahoo en la recuperación de documentos con este término es superior al mostrado por Google, ya que a pesar de la mayor frecuencia de páginas sin el término de búsqueda, las frecuencias superan las 170 apariciones, frente a las 104 y 117 veces que se dan en los documentos recuperados por Google. Algo similar podemos decir respecto al resto de buscadores. Sólo Teoma, al recuperar un documento con una frecuencia de aparición del término en 177 ocasiones supera, en este caso, a Yahoo.

En el comportamiento de Yahoo en esta búsqueda podemos observar dos aspectos. Por un lado la recuperación de los términos de búsqueda más específicos es mejor que la observada en el resto de motores de búsqueda. Sin embargo, para los términos más generales como "*web*" e "*information*", el número de recursos que no los contienen es superior a los demás.

5.3.2. Análisis comparativo de los motores de búsqueda

Lo primero que llama la atención es el mal funcionamiento de los motores en la búsqueda con operadores de existencia ya que ninguno de los buscadores evaluados recupera la frase con todos los términos de búsqueda, destacando el alto porcentaje de recursos que no los contienen, lo que una vez más indica una baja precisión en la recuperación.

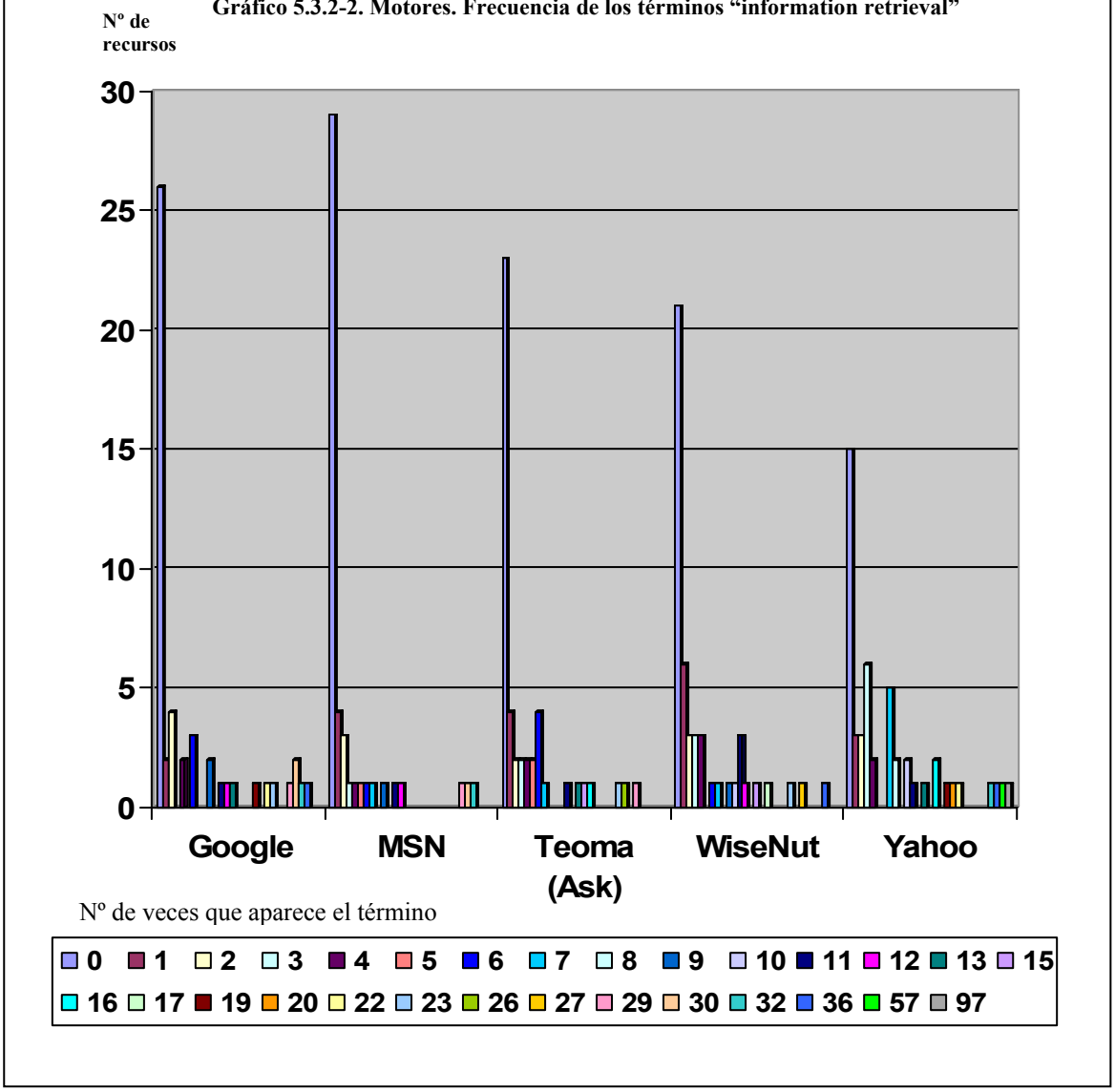
Gráfico 5.3.2-1. Motores. Frecuencia de los términos "information retrieval systems"



Como podemos observar en el gráfico anterior, la mayoría de motores destacan por el alto número de recursos que no contienen los términos de búsqueda

Yahoo es el buscador con mejores resultados en la recuperación de recursos con estos tres términos, destacando frente al resto, por recuperar las páginas con frecuencias superiores. WiseNut es el segundo buscador en recuperación de páginas con estos términos, mientras que Google y Teoma tienen un comportamiento semejante, correspondiendo en esta ocasión los peores resultados a MSN.

Gráfico 5.3.2-2. Motores. Frecuencia de los términos "information retrieval"

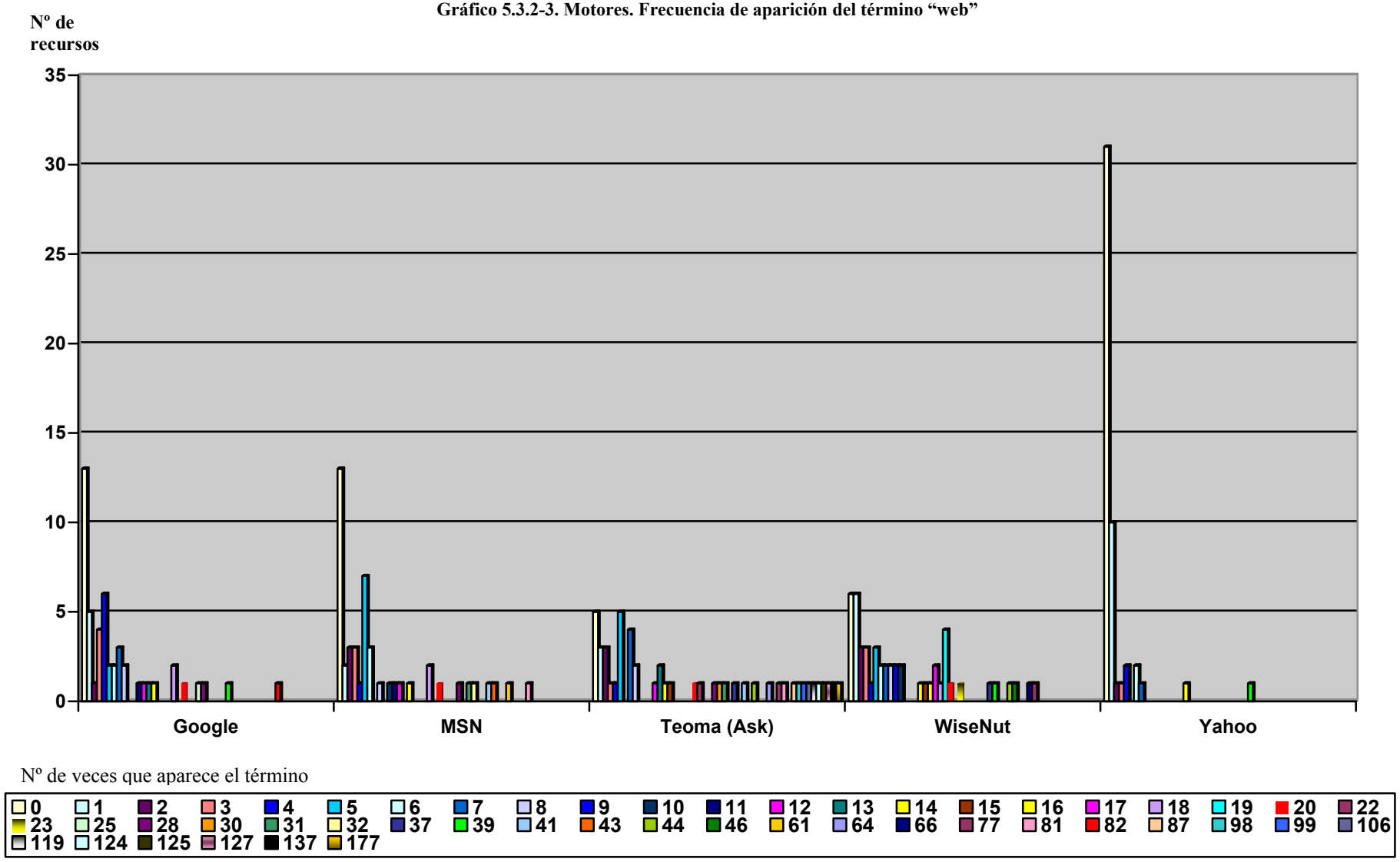


El perfil que ofrecen estos resultados es muy similar al anterior, aunque al tratarse sólo de dos términos, la aparición es más frecuente. También aquí le corresponde a Yahoo el mejor comportamiento, mejorando en esta ocasión al recuperar un mayor número de recursos con altas frecuencias de aparición de estos dos términos. Otra característica es el aumento de recursos que contienen los términos con diferentes frecuencias, sobre todo en el caso de MSN, cuyos resultados también suponen una mejora en comparación con los

términos analizados anteriormente, aunque sigue siendo el buscador con mayor índice de páginas en las que no aparecen estos términos.

Google mejora la recuperación de recursos con frecuencias más altas, teniendo un comportamiento similar a Teoma y WiseNut.

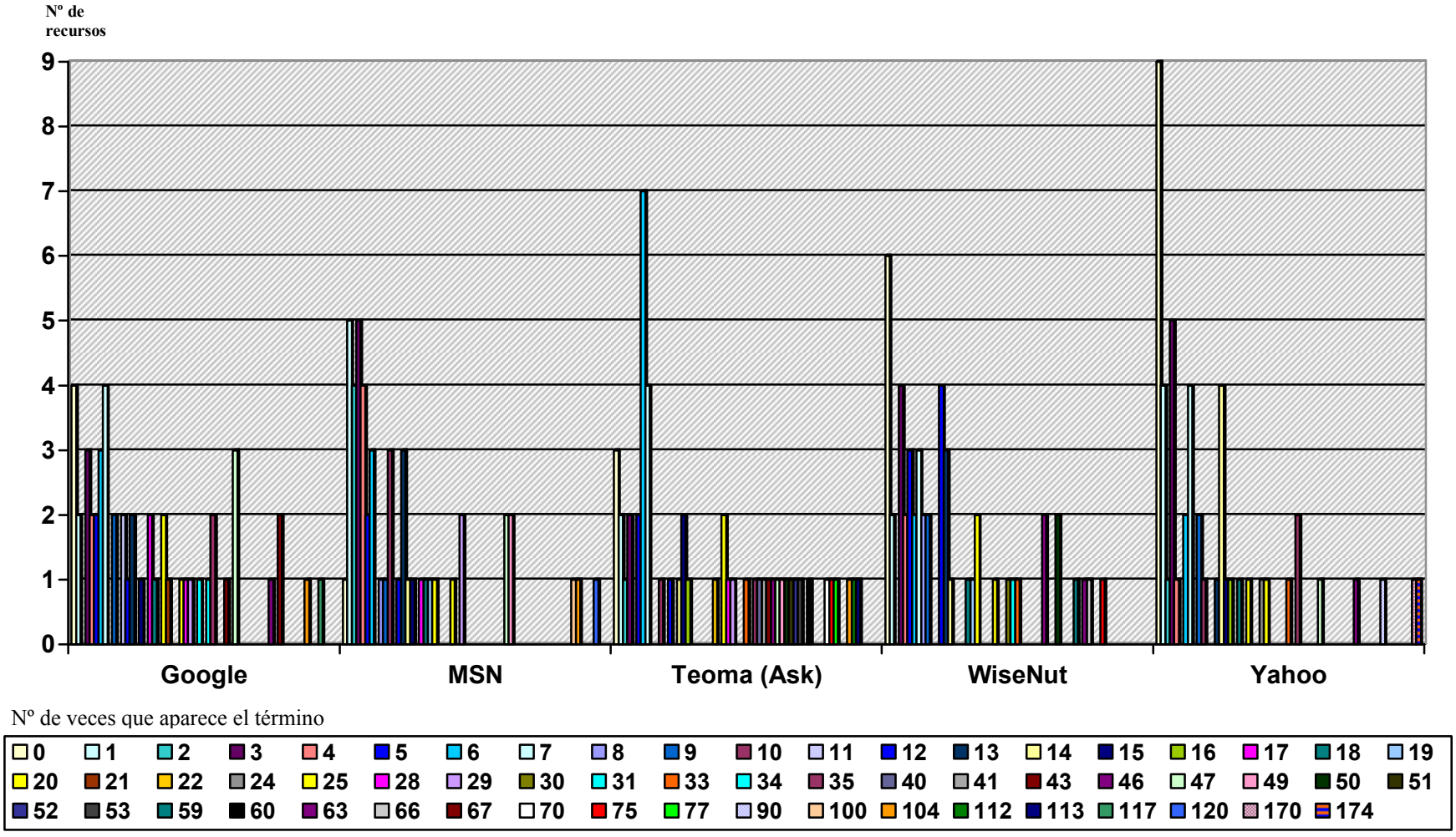
Gráfico 5.3.2-3. Motores. Frecuencia de aparición del término "web"



El funcionamiento de los motores ante la recuperación de frases con este término varía, ya que podemos observar que el número de páginas que no lo contienen desciende considerablemente, a excepción de Yahoo.

Google y MSN tienen un comportamiento similar, siendo en Google algo superiores los documentos con bajas y medias frecuencias. La recuperación en Teoma y WiseNut también es semejante, aunque Teoma recupera un mayor número de páginas con frecuencias tanto bajas como elevadas.

Gráfico 5.3.2-4. Motores. Frecuencia del término ”information”



En la recuperación de este término podemos constatar lo observado anteriormente, aunque aparece con mayor frecuencia y en un mayor número de documentos. Así, llaman la atención los bajos índices de páginas que no contienen el término de búsqueda, aunque Yahoo se caracteriza de nuevo por recuperar el mayor número de recursos sin el término de búsqueda, sin embargo también es el buscador que recupera los documentos en los que aparece con mayor frecuencia.

En definitiva, desechando la valoración de los términos genéricos, y teniendo en cuenta la expresión más específica "information retrieval systems" los buscadores en los que aparece con mayor frecuencia son Yahoo, seguido de Google y WiseNut, correspondiendo en este sentido, la peor recuperación a Teoma y MSN, lo que además coincide con los datos aportados por el resto de términos.

5.3.3. Análisis individualizado por metabuscadores

Dogpile

Tabla 5.3.3-1. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.3-2. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100

Dogpile no recuperó ningún documento con todos los términos.

Tabla 5.3.3-3. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	41	82%
1	5	10%
2	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
3	1	2%
4	1	2%
11	1	2%
Total	50	100%

Como ocurre con los motores de búsqueda, un elevado porcentaje de recursos (82%) no contiene estos tres términos y las frecuencias de aparición son bajas.

Tabla 5.3.3-4. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	32	64%
1	2	4%
2	1	2%
4	1	2%
6	1	2%
7	3	6%
9	2	4%
13	1	2%
14	1	2%
22	1	2%
23	2	4%
30	1	2%
31	1	2%
32	1	2%
Total	50	100%

Estos términos aparecen con mayor frecuencia, si bien, un 64% de los recursos los contienen. Las frecuencias de aparición en los recursos es más variada y mayor que en los término anteriores.

Tabla 5.3.3-5. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	7	14%
4	3	6%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
5	3	6%
6	1	2%
7	1	2%
10	1	2%
12	1	2%
20	3	6%
22	2	4%
28	1	2%
41	1	2%
77	1	2%
127	1	2%
177	1	2%
Total	50	100%

A pesar de ser un término muy utilizado en la Web, el 46% de los recursos recuperados por Dogpile no lo contienen. No obstante, se recuperan dos recursos en los que aparecen 127 y 177 veces, siendo sólo superado por Search, que recupera un recurso en el que aparece en 196 ocasiones.

Tabla 5.3.3-6. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	38%
1	4	8%
2	2	4%
3	1	2%
5	2	4%
6	3	6%
7	3	6%
10	1	2%
11	1	2%
12	1	2%
13	1	2%
25	1	2%
28	1	2%
31	1	2%
33	1	2%
34	1	2%
43	1	2%
47	2	4%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
49	1	2%
52	1	2%
75	1	2%
78	1	2%
Total	50	100%

Es, junto con el anterior, el término más frecuentemente recuperado, pues sólo un 38% de recursos no lo contiene. Es un término que se repite a menudo en cada documento recuperado, si bien con frecuencias algo inferiores al término anterior, ya que el recurso que más ocasiones lo contiene es uno en el que aparece en 78 ocasiones. No obstante, llama la atención frente a los motores de búsqueda, el alto porcentaje de recursos que no contienen este término (38%).

Dogpile, en esta búsqueda, a pesar de no tener un comportamiento muy distante del observado para los motores de búsqueda, llama la atención, frente al resto de metabuscadores, por el alto número de recursos que no contienen los términos de búsqueda.

Excite

Tabla 5.3.3-7. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.3-8. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Excite no recuperó ningún documento con todos los términos.

Tabla 5.3.3-9. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	76%
1	5	10%
2	2	4%
3	1	2%
4	1	2%
7	1	2%
11	2	4%
Total	50	100%

La recuperación de estos términos es similar a la observada en Dogpile, con alto porcentaje de recursos que no contienen los términos (76%) y bajas frecuencias aún en los documentos en los que aparecen más veces.

Tabla 5.3.3-10. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	26	52%
1	2	4%
2	1	2%
3	3	6%
4	1	2%
5	1	2%
6	2	4%
7	1	2%
8	1	2%
9	2	4%
13	1	2%
14	1	2%
16	1	2%
17	1	2%
23	3	6%
30	1	2%
31	1	2%
41	1	2%
Total	50	100%

El comportamiento con estos términos también es similar al de Dogpile, si bien podemos anotar una cierta mejoría al descender el porcentaje de recursos que no contienen

estos términos (52%) y recuperar algún recurso más, en los que la frecuencia de aparición de los términos es elevado, ya que recupera un documento en el que aparecen en 41 ocasiones.

Tabla 5.3.3-11. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	16	32%
1	6	12%
2	1	2%
3	4	8%
4	3	6%
5	3	6%
6	1	2%
10	1	2%
20	4	8%
22	1	2%
28	2	4%
29	1	2%
66	1	2%
73	1	2%
77	2	4%
110	1	2%
127	1	2%
196	1	2%
Total	50	100%

Excite, también demuestra una mejora respecto a Dogpile en la recuperación de recursos con este término, ya que la cifra de recursos que no lo contienen desciende hasta un 32%. Respecto a las frecuencias, el documento que más veces contiene este término es uno en el que aparece en 196 ocasiones. En general, podemos apreciar, respecto al anterior metabuscador, una mejor precisión técnica.

Tabla 5.3.3-12. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	18%
1	4	8%
2	4	8%
3	1	2%
4	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
5	1	2%
6	3	6%
7	4	8%
9	1	2%
10	1	2%
11	2	4%
13	1	2%
14	2	4%
20	1	2%
25	2	4%
28	1	2%
31	1	2%
34	1	2%
43	1	2%
47	1	2%
49	1	2%
56	1	2%
75	2	4%
76	1	2%
78	1	2%
117	1	2%
158	1	2%
Total	50	100%

En primer lugar, llama la atención, frente a los motores de búsqueda, el alto porcentaje (18%) de recursos que no contienen el término de búsqueda, aunque no alcanza las frecuencias máximas vistas en Yahoo (con documentos en los que aparecen 170 y 174 veces).

Supera a Dogpile en cuanto a la recuperación de documentos con frecuencias máximas.

Excite presenta una ligera mejora respecto a Dogpile dado que recupera menos documentos sin los términos de búsqueda, lo que posibilita una mayor variedad en la recuperación y en las frecuencias.

Ixquick

Tabla 5.3.3-13. Nº de recursos en los que aparece el término

Nº Recursos	50

Tabla 5.3.3-14. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

	Nº de recursos	Porcentaje
0	50	100%

Ixquick no recuperó ningún documento con todos los términos.

Tabla 5.3.3-15. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	33	66%
1	7	14%
2	1	2%
3	2	4%
4	1	2%
5	1	2%
6	1	2%
10	1	2%
11	1	2%
12	1	2%
13	1	2%
Total	50	100%

En la recuperación de estos términos demuestra una mejoría respecto a los metabuscadores anteriores no sólo por el número de recursos que no los contienen, (un 66% en este caso) sino también por la recuperación de documentos con mayores frecuencias.

Tabla 5.3.3-16. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	24	48%
1	3	6%
2	3	6%
4	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
5	2	4%
6	1	2%
7	2	4%
10	2	4%
11	1	2%
12	1	2%
13	1	2%
14	1	2%
16	1	2%
17	1	2%
23	1	2%
29	1	2%
30	1	2%
31	1	2%
32	1	2%
97	1	2%
Total	50	100%

También en la recuperación de estos términos supera a Excite, ya que la mayor frecuencia corresponde a un documento en el que los términos aparecen 97 veces frente a 41, que es la mayor frecuencia que ofrece un documento recuperado por Excite.

Tabla 5.3.3-17. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	3	6%
2	3	6%
3	1	2%
4	5	10%
5	3	6%
6	1	2%
9	1	2%
10	1	2%
11	1	2%
14	1	2%
17	1	2%
20	1	2%
28	1	2%
41	1	2%
66	1	2%
77	1	2%
110	1	2%
Total	50	100%

Este término no es tan frecuente en los documentos que recupera Ixquick ya que un 46% no lo contienen, aproximándose en este sentido a Dogpile. En frecuencias de apari-

ción, es algo inferior a Excite ya que recupera un documento con 110 apariciones frente a 196 que es la frecuencia del documento de Excite.

Tabla 5.3.3-18. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	12	24%
1	3	6%
2	5	10%
3	4	8%
4	1	2%
5	2	4%
6	1	2%
7	3	6%
8	1	2%
10	1	2%
11	1	2%
13	1	2%
17	2	4%
28	1	2%
29	1	2%
31	1	2%
32	2	4%
47	2	4%
49	1	2%
56	1	2%
75	1	2%
104	1	2%
158	1	2%
170	1	2%
Total	50	100%

Ixquick muestra para este término una recuperación similar a Excite, caracterizada por tanto por la alta frecuencia de aparición del término en los recursos recuperados, y por las altas frecuencias de repetición del término en determinados recursos.

Ixquick continúa la mejora, en cuanto a la precisión técnica apreciada en Excite, basada una vez más en una recuperación de documentos con los términos de búsqueda.

Search

Tabla 5.3.3-19. Nº de recursos analizados

Nº Recursos	49

Tabla 5.3.3-20. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	49	100%

Search no recuperó ningún documento con todos los términos.

Tabla 5.3.3-21. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	57,1%
1	10	20,4%
2	2	4,1%
3	2	4,1%
4	2	4,1%
5	1	2%
7	1	2%
8	1	2%
11	1	2%
13	1	2%
Total	49	100%

En comparación con Excite e Ixquick, Search recupera un menor porcentaje de recursos que no contienen los términos (57,1%) superándolos también en la recuperación de más documentos con bajas frecuencias, ya que por ejemplo, en el 20,4% de los resultados sólo aparecen una vez. En la recuperación de documentos con altas frecuencias de aparición de los términos supera a Dogpile y Excite.

Tabla 5.3.3-22. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	16	32,7%
1	2	4,1%
2	4	8,2%
4	2	4,1%
5	1	2%
6	1	2%
7	3	6,1%
8	2	4,1%
9	1	2%
11	1	2%
12	1	2%
13	2	4,1%
14	1	2%
16	1	2%
17	1	2%
22	1	2%
23	2	4,1%
30	1	2%
32	2	4,1%
36	1	2%
41	1	2%
43	1	2%
97	1	2%
Total	49	100%

Search mantiene para estos términos una buena precisión técnica fruto del descenso del número de recursos que no contiene los términos, lo que le permite tener mejores resultados en el resto de frecuencias, especialmente en las más altas.

Tabla 5.3.3-23. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	10	20,4%
1	7	14,3%
2	2	4,1%
3	1	2%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
4	4	8,2%
5	4	8,2%
6	1	2%
8	2	4,1%
10	2	4,1%
11	1	2%
12	1	2%
14	1	2%
15	1	2%
16	2	4,1%
18	1	2%
19	1	2%
20	1	2%
28	1	2%
39	1	2%
41	1	2%
73	1	2%
77	1	2%
127	1	2%
196	1	100%
Total	49	

De nuevo muestra unos bajos índices en la recuperación de páginas sin los términos de búsqueda, lo que influye una vez más en una mejora de los resultados de frecuencias. La precisión técnica resultante, aún con una preponderancia de páginas con poca frecuencia de aparición de los términos queda compensada por el número de páginas con frecuencias elevadas.

Tabla 5.3.3-24. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	4	8,2%
1	6	12,2%
2	1	2%
3	2	4,1%
4	1	2%
5	2	4,1%
6	3	6,1%
7	3	6,1%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
8	1	2%
9	1	2%
11	1	2%
13	1	2%
15	1	2%
17	1	2%
18	1	2%
20	2	4,1%
25	1	2%
28	1	2%
31	1	2%
33	1	2%
34	1	2%
43	1	2%
46	1	2%
47	3	6,1%
50	1	2%
63	1	2%
70	1	2%
75	1	2%
76	1	2%
78	1	2%
117	1	2%
170	1	2%
Total	49	100%

Es el metabuscador en el que el término aparece de forma más frecuente, ya que no aparece en el 8,2% de los documentos recuperados.

De aquí que las frecuencias sean variadas en el resto de documentos caracterizándose por una precisión técnica equilibrada, pues aunque el número de recursos en los que aparece una vez es alto (8,2%), el resto de recursos mantiene frecuencias variadas, resultando alguna de ellas de las más elevadas.

Search ofrece en esta búsqueda los mejores resultados en comparación con el resto de metabuscadores.

Surfwax

Tabla 5.3.3-25. Nº de recursos analizados

Nº Recursos	36

Tabla 5.3.3-26. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	36	100%

Surfwax no recuperó ningún documento con todos los términos.

Tabla 5.3.3-27. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	32	88,9%
1	1	2,8%
6	1	2,8%
11	1	2,8%
12	1	2,8%
Total	36	100%

Surfwax ofrece un comportamiento inferior al resto de metabuscadores en cuanto a la precisión técnica, ya que en el 88,9% de los resultados no aparecen los términos, mientras que en el resto figuran una, seis, once y doce veces respectivamente.

Tabla 5.3.3-28. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	77,8%
1	1	2,8%
2	1	2,8%
5	1	2,8%
7	1	2,8%
10	1	2,8%
12	1	2,8%
18	1	2,8%
23	1	2,8%
Total	36	100%

Surfwax muestra también en estos términos, un comportamiento inferior al del resto de metabuscadores, con alto porcentaje de recursos que no contienen los términos (77,8%), manteniendo un equilibrio entre las frecuencias bajas y altas, recuperando sólo un recurso para las distintas frecuencias.

Tabla 5.3.3-29. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	8	22,2%
1	6	16,7%
2	5	13,9%
3	3	8,3%
4	2	5,6%
5	2	5,6%
6	2	5,6%
9	1	2,8%
14	1	2,8%
15	2	5,6%
20	1	2,8%
28	1	2,8%
41	1	2,8%
77	1	2,8%
Total	36	100%

Muestra una mejor recuperación de este término que Dogpile, ya que sólo un 22% no lo contiene. Se caracteriza por la recuperación de recursos con frecuencias bajas pues el 58,5% no supera una aparición del término en más de 9 ocasiones. En Dogpile, sin embargo, las frecuencias de aparición de los documentos son menores para este término que para el resto.

Respecto al resto de metabuscadores, Surfwax no recupera los recursos con altas frecuencias de aparición de los términos.

Tabla 5.3.3-30. Frecuencia y nº de recursos en los que aparece el término "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	13	36,1%
1	5	13,9%
2	4	11,1%
3	1	2,8%
4	2	5,6%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
6	1	2,8%
7	1	2,8%
9	2	5,6%
12	1	2,8%
13	1	2,8%
16	1	2,8%
17	1	2,8%
33	1	2,8%
70	1	2,8%
75	1	2,8%
Total	36	100%

También aquí supera al resto, excepto a Dogpile en cuanto al número de páginas que no contiene el término de búsqueda. Además los resultados se agrupan en los apartados de bajas frecuencias, y como ocurre con el término anterior, no recupera recursos con altas frecuencias de aparición del término.

Surfwax se caracteriza por el alto número de recursos que no contienen los términos de búsqueda, concretamente los más específicos, y recupera un mayor número de recursos con los términos más comunes como "*web*" e "*information*", lo que denota una baja precisión técnica en los recursos recuperados.

Vivisimo

Tabla 5.3.3-31. Nº de recursos analizados

Nº Recursos	50

Tabla 5.3.3-32. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems +and +the web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	50	100%

Vivisimo no recuperó ningún documento con todos los términos.

Tabla 5.3.3-33. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval systems"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	76%
1	9	18%
2	1	2%
11	1	2%
12	1	2%
Total	50	100%

Vivisimo muestra un alto porcentaje de páginas que no contienen los términos de búsqueda (76%), centrando sus resultados en páginas con bajas frecuencias, aunque también recupera dos recursos en los que los términos aparecen once y doce veces.

La recuperación de este metabuscador en comparación con Dogpile y Excite, recupera un recurso en el que la frecuencia de los términos es superior. Sin embargo, Ixquick y Search lo superan en cuanto a porcentaje de recursos que contienen los términos y por la recuperación de documentos con frecuencias superiores de aparición.

Tabla 5.3.3-34. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	25	50%
1	5	10%
2	2	4%
4	2	4%
5	1	2%
6	1	2%
7	3	6%
11	1	2%
12	2	4%
13	1	2%
23	2	4%
30	1	2%
31	1	2%
32	1	2%
36	1	2%
43	1	2%
Total	50	100%

Vivisimo presenta los resultados en tres grupos, el primero de ellos con bajas frecuencias, en el que el 28% de los resultados no supera las siete veces en cuanto a la aparición de los términos en un documento. Un 8% de las páginas no supera las trece veces, y finalmente, en el 14% restante, los términos aparecen entre veintitrés y cuarenta y tres veces.

La precisión técnica es superior a la de Dogpile tanto en cuanto al número de recursos que contienen los términos de búsqueda como en la recuperación de un recurso con mayor frecuencia de aparición de los términos de búsqueda.

Tabla 5.3.3-35. Frecuencia y nº de recursos en los que aparece el término "web"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	14	28%
1	8	16%
3	1	2%
4	4	8%
5	2	4%
6	2	4%
8	2	4%
10	1	2%
12	1	2%
13	1	2%
14	1	2%
15	1	2%
18	1	2%
19	1	2%
20	2	4%
27	1	2%
28	1	2%
39	1	2%
41	1	2%
61	2	4%
77	1	2%
127	1	2%
Total	50	100%

En la recuperación de este término podemos observar un aceptable comportamiento de este metabuscador pues aunque es alto el porcentaje de recursos con bajas frecuencias de aparición del término, las páginas con frecuencias medias y altas son importantes, aunque, como ocurre en el resto de los casos, no muy elevadas.

Como en los términos anteriores, la recuperación es similar a la de Excite y Surfwax, aunque éstos recuperan un recurso con mayor frecuencia de aparición del término de búsqueda.

Tabla 5.3.3-36. Frecuencia y nº de recursos en los que aparece el términos "information"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	10	20%
1	8	16%
2	5	10%
3	1	2%
4	1	2%
5	1	2%
6	3	6%
7	2	4%
8	1	2%
9	1	2%
10	1	2%
13	1	2%
15	1	2%
18	1	2%
19	1	2%
20	1	2%
28	1	2%
43	1	2%
46	1	2%
47	2	4%
49	1	2%
63	1	2%
70	1	2%
75	1	2%
78	1	2%
120	1	2%
Total	50	100%

La recuperación de este término es un claro reflejo de lo señalado en los anteriores.

También aquí supera a Dogpile y Surfwax en las frecuencias que analizamos.

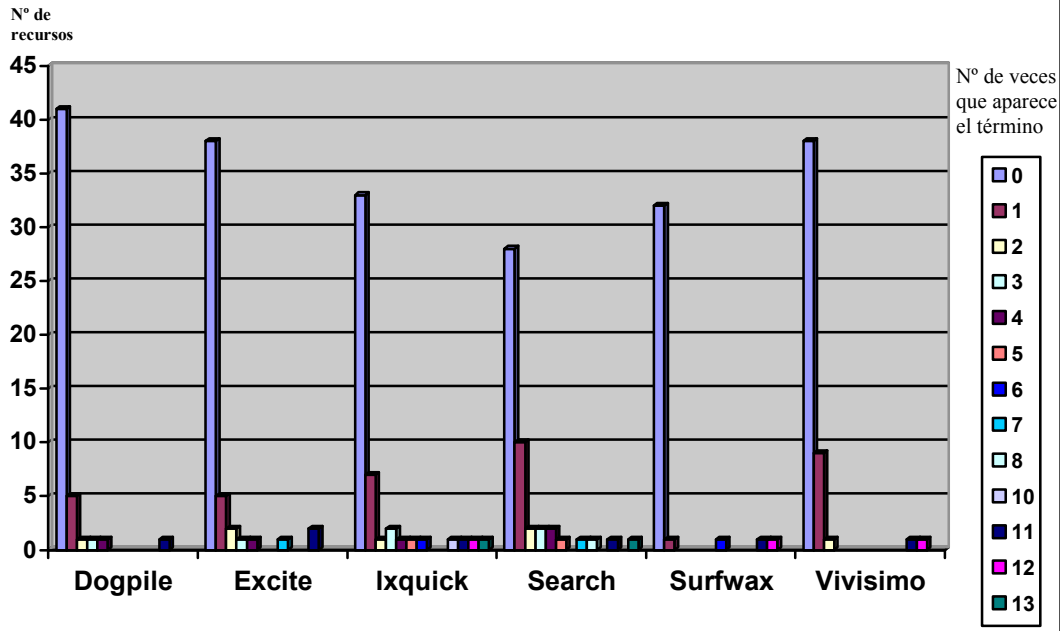
Excite lo supera tanto en los documentos recuperados que contienen el término como en la frecuencia del documento en el que aparece en más ocasiones (120 de Vivisimo frente a 158 de Excite).

Con Search guarda cierta relación en cuanto al porcentaje de recuperación de páginas que contienen el término.

Vivisimo no destaca frente al resto por ningún aspecto en particular, simplemente podemos decir que la mejor recuperación la realiza para los términos más genéricos de la búsqueda.

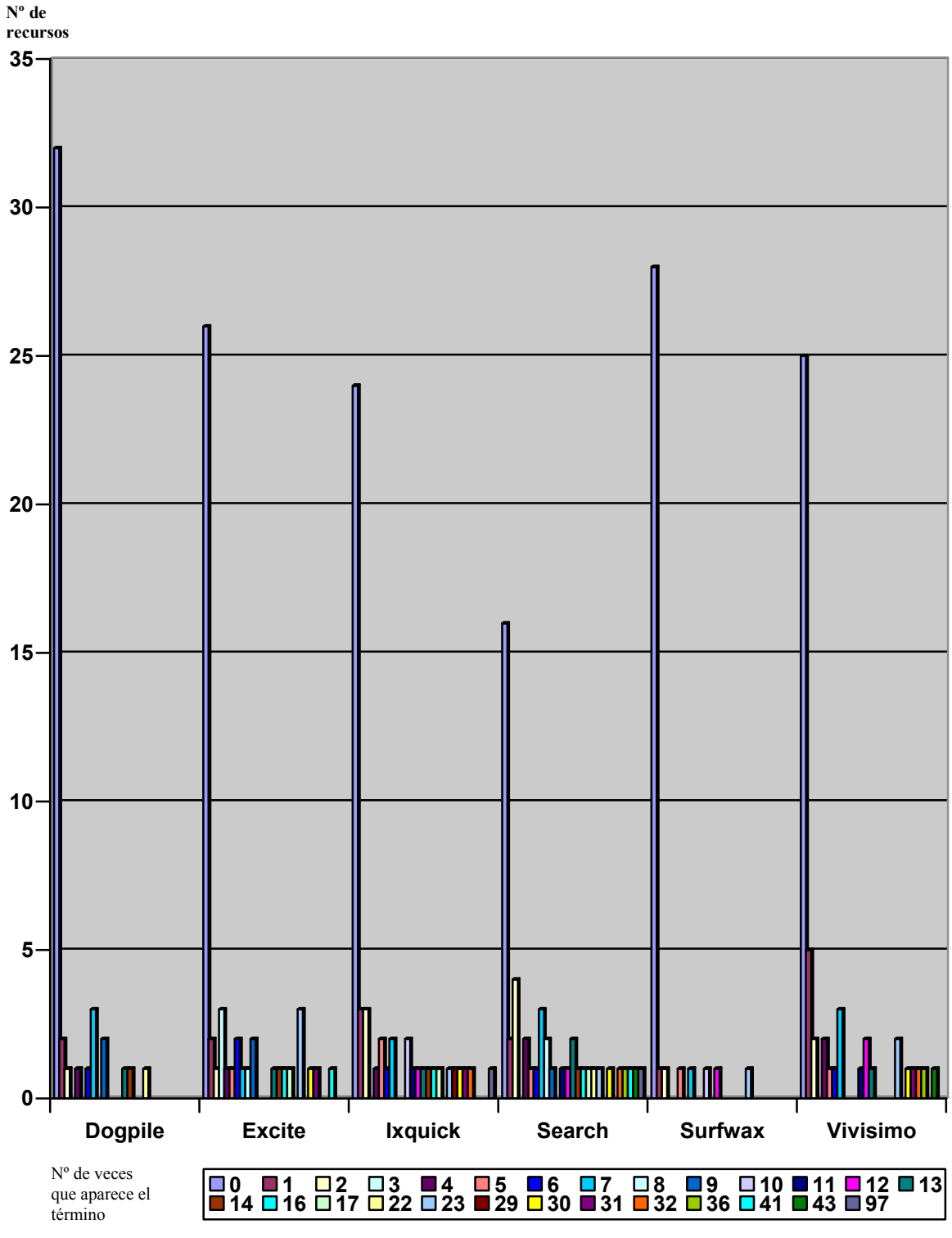
5.3.4. Análisis comparativo de los metabuscadores

Gráfico 5.3.4-1. Metabuscadores. Frecuencia de los términos "information retrieval systems"



Al igual que los motores, los metabuscadores recuperan un gran número de páginas sin los términos de búsqueda. Corresponde a Search la recuperación tanto de los recursos con mayores frecuencias de aparición de los términos como el mayor número de páginas que contienen los términos. Se caracteriza además por recuperar recursos tanto con bajas como con altas frecuencias. Ixquick mantiene en esta búsqueda un funcionamiento similar a Search, aunque recupera un gran número de recursos sin los tres términos de búsqueda.

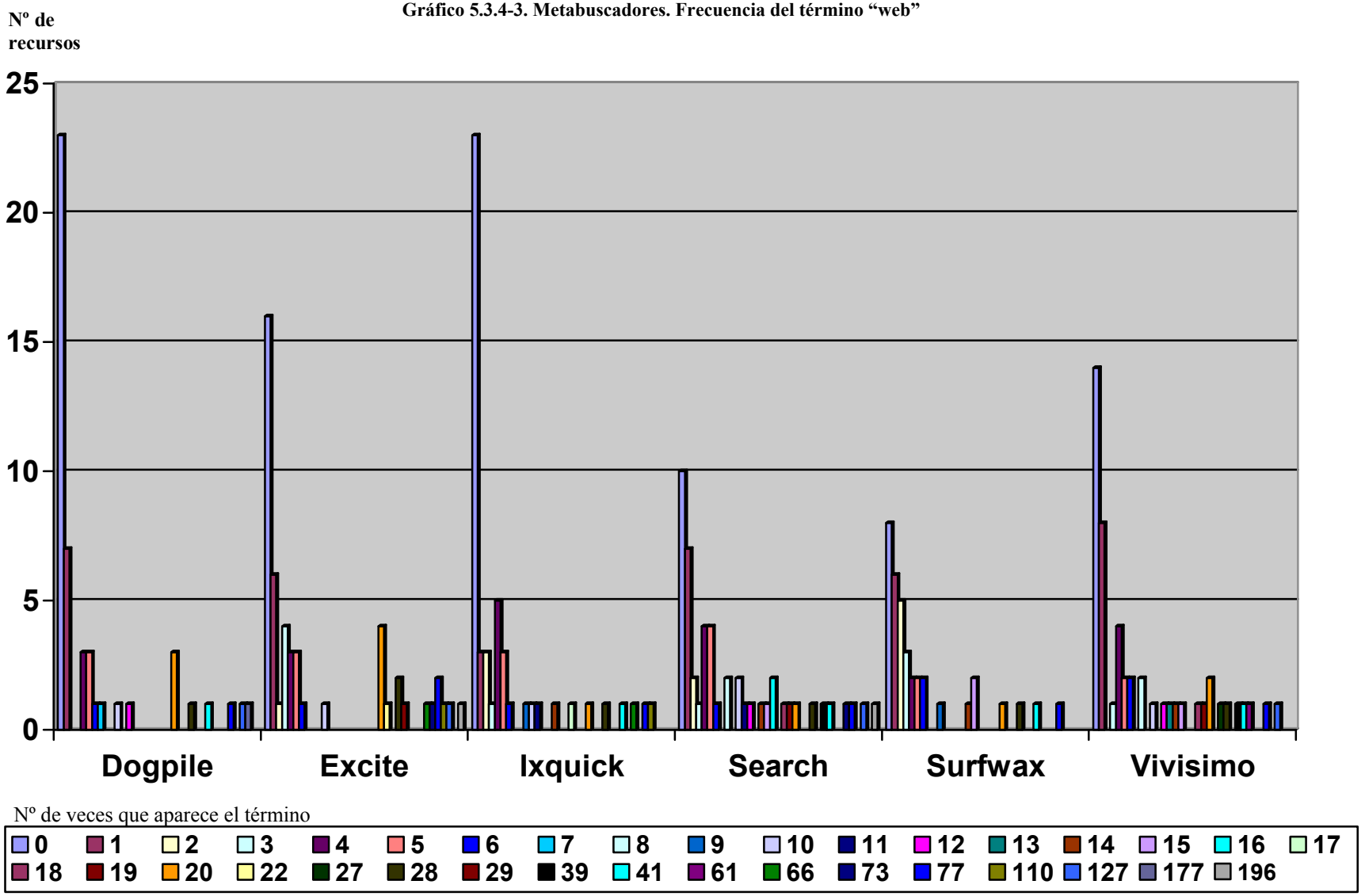
Gráfico 5.3.4-2. Metabuscadores.Frecuencia de los términos "information retrieval"



El perfil de la gráfica anterior es muy similar a la que ofrecen los motores de búsqueda, de aquí que podamos hablar de cierta semejanza en la recuperación de estos términos entre unos sistemas y otros.

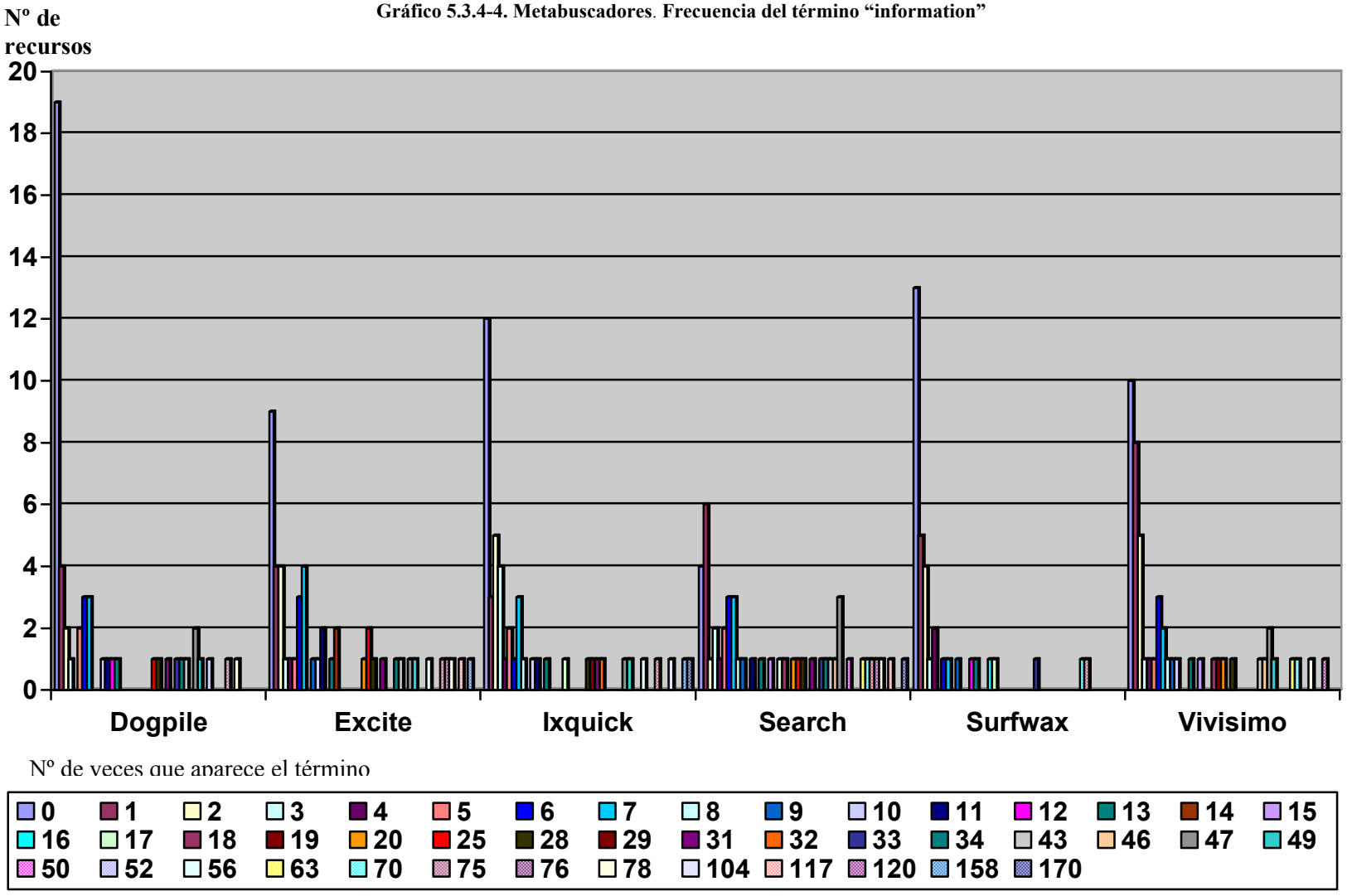
A nivel individual, es de nuevo Search el metabuscador que mejor comportamiento tiene en la recuperación de estos términos. Vivisimo muestra también una buena recuperación de recursos con altas frecuencias, aunque como podemos apreciar, también es elevado el número de recursos con frecuencias bajas. En un grupo intermedio podemos colocar a Ixquick y Excite, correspondiendo el peor funcionamiento a Dogpile y sobre todo a Surfwax.

Gráfico 5.3.4-3. Metabuscadores. Frecuencia del término "web"



Los metabuscadores que recuperan un mayor número de recursos con este término de búsqueda son Search y Vivisimo, aunque en éste último, como en el caso anterior, el número de páginas que no lo contienen es superior. Por otro lado, se caracterizan por la variedad de páginas que contienen el término con distintas frecuencias, así como por la recuperación de los recursos con mayores frecuencias. En Excite, Surfwax e Ixquick el comportamiento es similar, aunque en este último es mayor el número de páginas que no contienen el término. Corresponde una menor precisión a Dogpile, fundamentalmente por el alto número de páginas que no contienen el término.

Gráfico 5.3.4-4. Metabuscadores. Frecuencia del término "information"



Como ocurre con los motores de búsqueda, el funcionamiento en la recuperación del término "information" es similar al observado anteriormente. Las mayores diferencias las ofrece Surfwax, que en este caso recupera un mayor número de páginas sin el término, y a Excite que mejora al recuperar un mayor número de recursos.

Por tanto, si tenemos en cuenta la expresión más específica para valorar la precisión técnica, esto es "Information retrieval systems", Search e Ixquick son las herramientas que mejor funcionaron, seguidos, por este orden, de Excite, Dogpile y Vivisimo, correspondiendo a Surfwax la peor recuperación.

5.4. Búsqueda booleana

Para analizar la búsqueda booleana nos centramos en las frecuencias de aparición de los siguientes términos o frases:

Búsqueda completa:

1. information retrieval AND digital libraries AND electronic libraries AND virtual libraries

Términos y frases:

2. information retrieval
3. digital libraries
4. electronic libraries
5. virtual libraries

Tampoco en esta búsqueda los motores recuperan recursos con todos los términos solicitados. Teoma y WiseNut no funcionaron en este tipo de búsqueda. El análisis de las frecuencias de aparición de cada uno de los temas de búsqueda arrojó los resultados que se recogen en las siguientes tablas.

5.4.1. Análisis individualizado por motores de búsqueda

Google

Tabla 5.4.1-1. **Nº de recursos analizados**

Nº Recursos	48

Tabla 5.4.1-2. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	28	58,3%
1	12	25%
2	2	4,2%
3	2	4,2%
5	2	4,2%
7	1	2,1%
12	1	2,1%
Total	48	100%

Google con el 58,3%, presenta un porcentaje algo menor que Yahoo (60,4%) de recursos que no contienen estos términos de búsqueda. El porcentaje aún es mayor en MSN (69,4%). En el 25% de los recursos recuperados, estos términos aparecen sólo una vez, y sólo en un documento aparece en doce ocasiones. Las mayores frecuencias corresponden a los recursos con bajas frecuencias.

Tabla 5.4.1-3. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	25	52,1%
1	5	10,4%
3	1	2,1%
4	2	4,2%
6	2	4,2%
9	1	2,1%
12	1	2,1%
13	1	2,1%
16	2	4,2%
18	1	2,1%
27	2	4,2%
28	1	2,1%
36	1	2,1%
43	1	2,1%
44	1	2,1%
45	1	2,1%
Total	48	100%

Google supera en esta ocasión a Yahoo en la recuperación de recursos con estos términos, pero está por debajo de MSN.

También en la recuperación de recursos de frecuencias elevadas es superado por MSN, ya que este buscador recupera dos recursos con 90 y 72 apariciones de los términos, mientras que en Google la frecuencia de aparición superior es de 45.

Respecto a los metabuscadores, los términos aparecen en Google con mayor frecuencia que en Excite, aunque no recupera recursos con frecuencias tan elevadas como éste.

Tabla 5.4.1-4. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	32	66,7%
1	12	25%
2	1	2,1%
3	1	2,1%
4	1	2,1%
5	1	2,1%
Total	48	100%

El comportamiento de Google en la recuperación de estos términos es similar a la observada en el caso de "information retrieval". Aparecen con mayor frecuencia en los documentos recuperados por Google que en recuperados por los otros dos buscadores, ya que en MSN un 93,9% no contiene estos términos y en Yahoo un 85,4%.

Tabla 5.4.1-5. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	79,2%
1	6	12,5%
2	1	2,1%
3	1	2,1%
4	1	2,1%
5	1	2,1%
Total	48	100%

Son los términos que aparecen con menor frecuencia en los documentos ya que en Google no aparecen en un 79,2% de resultados, en MSN en un 93,9% y en Yahoo en un 81,3%.

Google recupera, en comparación con Excite, un mayor número de documentos que contienen estos términos, si bien, la frecuencia de aparición, como ocurría con "*digital libraries*", no es tan numerosa.

La precisión técnica en este tipo de búsquedas también es deficiente por los bajos porcentajes de aparición de los términos de búsqueda, más teniendo en cuenta que se trata de términos de uso frecuente. Otro aspecto a considerar es que centra su recuperación en recursos con bajas frecuencias de aparición de los términos.

MSN

Tabla 5.4.1-6. Nº de recursos analizados

Nº Recursos	49

Tabla 5.4.1-7. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	34	69,4%
1	9	18,4%
2	2	4,1%
4	1	2%
6	1	2%
9	1	2%
13	1	2%
Total	49	100%

MSN es el motor con mayor número de páginas que no contienen los términos de búsqueda, (69,4%). En un 18,4% de documentos sólo aparece una vez. De forma similar a Google, con el que se aprecia cierta similitud, recupera un documento en el que los términos se citan en trece ocasiones.

Tabla 5.4.1-8. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	42,9%
1	4	8,2%
2	2	4,1%
3	1	2%
4	1	2%
8	2	4,1%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
10	1	2%
11	2	4,1%
12	1	2%
16	1	2%
17	1	2%
18	1	2%
20	1	2%
23	1	2%
26	1	2%
27	1	2%
32	1	2%
36	2	4,1%
38	1	2%
42	1	2%
72	1	2%
90	1	2%
Total	49	100%

MSN recupera más recursos que Google con estos términos de búsqueda. Presenta además dos recursos con unas frecuencias de 72 y 90 que Google no recupera.

Tabla 5.4.1-9. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	46	93,9%
2	3	6,1%
Total	49	100%

MSN presenta tan sólo tres recursos en los que aparecen estos términos, frente a Google que recuperaba dieciséis.

Tabla 5.4.1-10. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	46	93,9%
3	2	4,1%
22	1	2%
Total	49	100%

También estos términos aparecen en menor medida en los recursos recuperados por MSN que en los recuperados por Google. Si embargo, MSN recupera un documento en el que estos términos aparecen en 22 ocasiones.

MSN es el buscador que peor funcionamiento presenta en la recuperación de los términos que componen la búsqueda booleana a excepción de los términos "Digital libraries", ya que la recuperación del resto de términos se caracteriza por las bajas frecuencias.

Yahoo

Tabla 5.4.1-11. Nº de recursos analizados

Nº Recursos	48

Tabla 5.4.1-12. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	29	60,4%
1	11	22,9%
2	1	2,1%
3	1	2,1%
5	1	2,1%
6	1	2,1%
12	1	2,1%
13	1	2,1%
15	1	2,1%
47	1	2,1%
Total	48	100%

Yahoo, frente a Google y MSN, recupera un mayor número de recursos, con mayor frecuencia de aparición de los términos.

Tabla 5.4.1-13. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	29	60,4%
1	3	6,3%
3	1	2,1%
4	1	2,1%
6	3	6,3%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
9	1	2,1%
12	1	2,1%
16	1	2,1%
18	1	2,1%
24	1	2,1%
25	1	2,1%
32	1	2,1%
36	1	2,1%
43	1	2,1%
44	1	2,1%
45	1	2,1%
Total	48	100%

Yahoo obtiene el mayor número de páginas sin estos términos, lo que influye en una menor precisión técnica de sus resultados.

Tabla 5.4.1-14. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	41	85,4%
1	5	10,4%
2	1	2,1%
5	1	2,1%
Total	48	100%

También en esta ocasión obtiene malos resultados aunque no tanto como los de MSN.

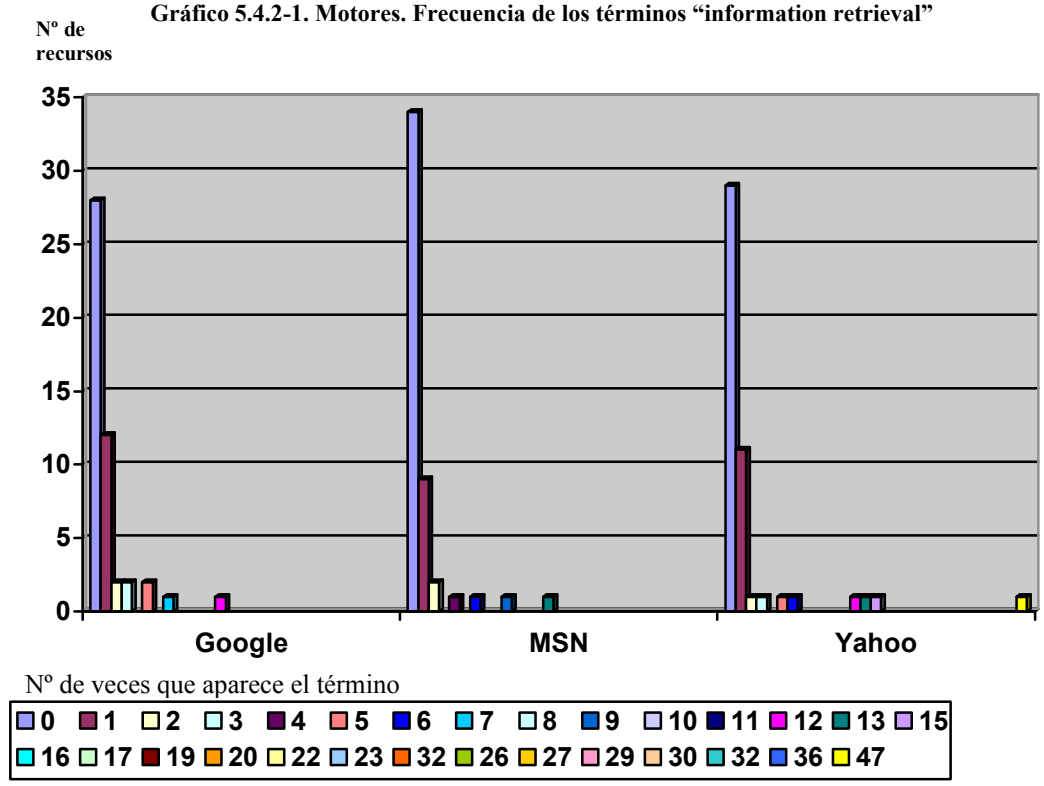
Tabla 5.4.1-15. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	39	81,3%
1	6	12,5%
2	1	2,1%
3	2	4,2%
Total	48	100%

Sin embargo, respecto a estos términos, se da la misma circunstancia que en el caso anterior. En relación con Google, no presenta grandes diferencias ya que, aunque recupera un mayor número de recursos que no contienen los términos, obtiene más documentos con bajas frecuencias.

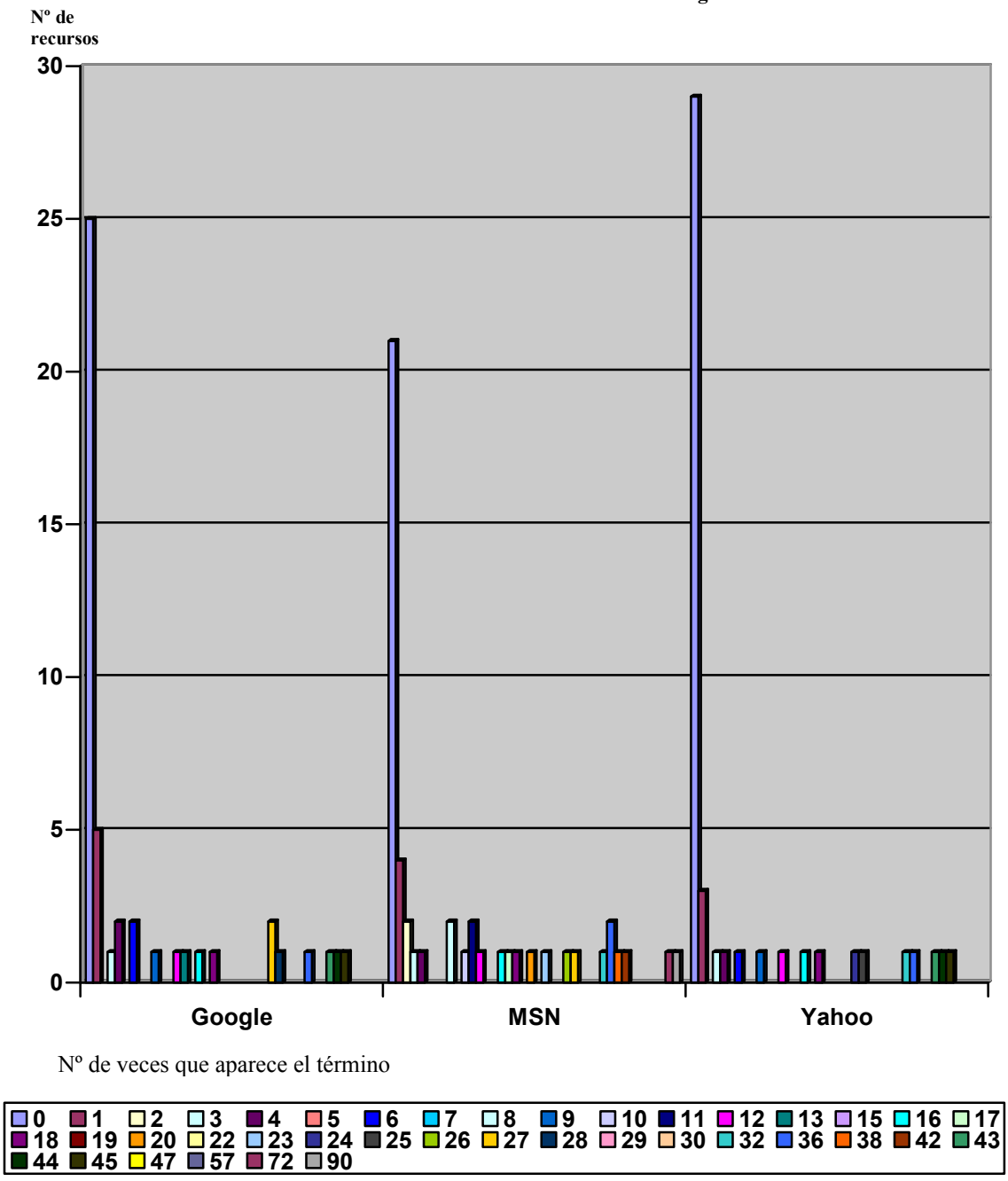
5.4.2. Análisis comparativo de los motores de búsqueda

Gráfico 5.4.2-1. Motores. Frecuencia de los términos "information retrieval"



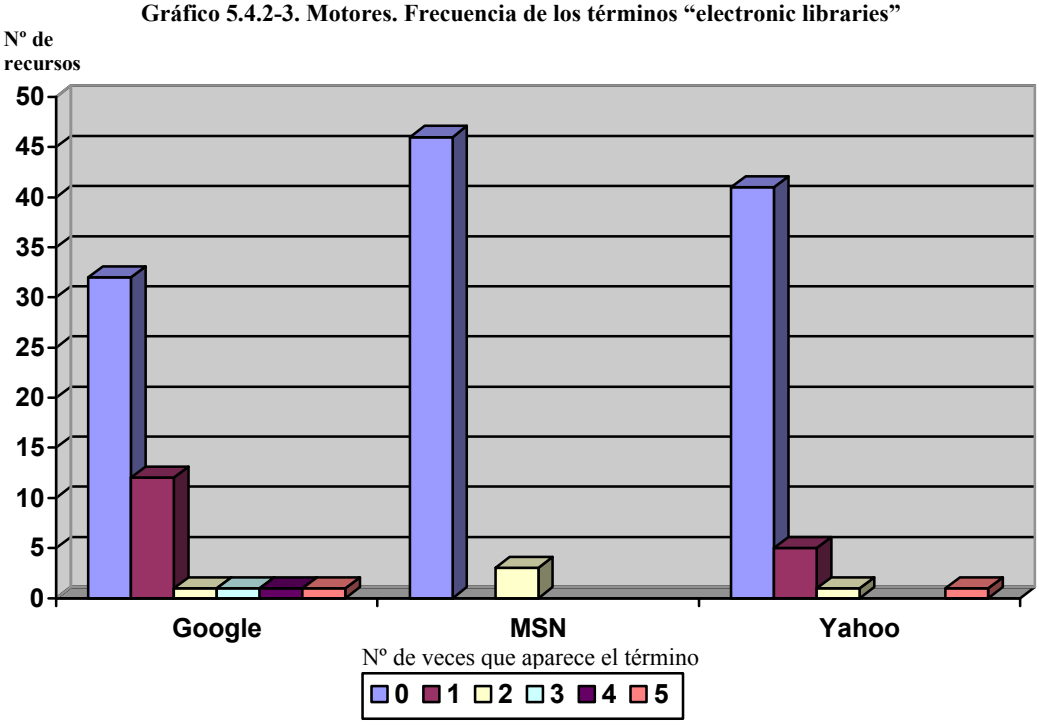
En el análisis de la recuperación por términos, podemos observar, en los tres motores que ofrecieron resultados, el elevado número de recursos que no contienen estos términos, lo que tratándose de una búsqueda booleana indica, también en estos casos, un defectuoso funcionamiento ante este tipo de búsquedas. La recuperación difiere en los tres motores ya que aunque entre Google y Yahoo se aprecia cierta semejanza, éste último recupera un recurso de alta frecuencia de aparición de los términos.

Gráfico 5.4.2-2. Motores. Frecuencia de los términos "digital libraries"



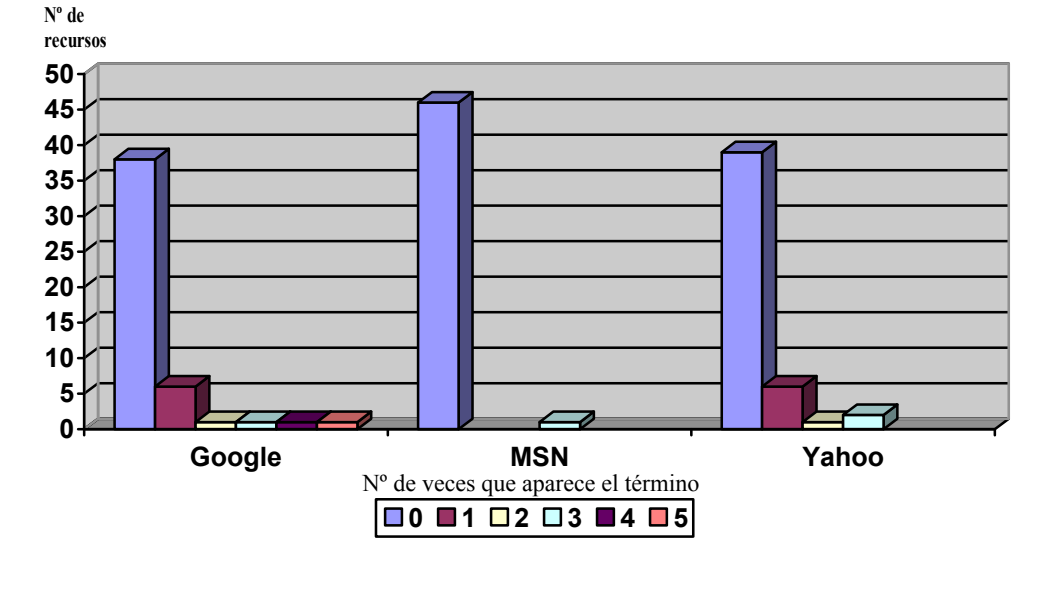
Respecto a la recuperación de recursos con estos términos, se aprecian diferencias respecto a los anteriores ya que en esta ocasión es Yahoo el motor que menos recursos recupera y MSN el que ofrece más páginas con estos términos, además de las mayores frecuencias en un mismo documento. Google mantiene un comportamiento similar al de los términos anteriores, si bien aquí recupera un mayor número de recursos en los que los términos aparecen con mayor frecuencia.

Gráfico 5.4.2-3. Motores. Frecuencia de los términos "electronic libraries"



De nuevo destaca el mal funcionamiento en la recuperación de recursos con estos términos. A pesar de todo debemos mencionar el comportamiento Google, que supera a los otros dos motores al recuperar el mayor número de páginas con los términos de búsqueda.

Gráfico 5.4.2-4. Motores. Frecuencia de los términos "virtual libraries"



La recuperación de recursos con estos términos es prácticamente un calco de la anterior, que confirma la tendencia observada al principio sobre el mal funcionamiento de los motores en este tipo de búsquedas.

5.4.3. Análisis individualizado por metabuscadores

Dogpile y Surfwax no recuperan recursos en esta búsqueda. Otros metabuscadores como Excite e Ixquick no alcanzaron los cincuenta resultados, lo que no dificulta su análisis.

Excite

Tabla 5.4.3-1. Nº de recursos analizados

Nº Recursos	33

Tabla 5.4.3-2. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	63,6%
1	6	18,2%
2	2	6,1%
3	2	6,1%
18	1	3%
20	1	3%
Total	33	100%

En relación con estos términos debemos destacar el alto porcentaje (63,6%) de recursos que no contienen esta expresión, por lo que no se observa mejora respecto a los motores.

Claramente los resultados se reparten en dos grupos, uno de bajas frecuencias y otro de frecuencias altas.

Tabla 5.4.3-3. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	21	63,6%
1	2	6,1%
2	1	3%
18	1	3%
20	1	3%
21	1	3%
22	1	3%
23	1	3%
26	1	3%
29	1	3%
36	1	3%
67	1	3%
Total	33	100%

Estos términos aparecen en los documentos recuperados un número mayor de veces que los términos anteriores, aunque el porcentaje de los que no los contiene sigue siendo elevado (63,6%). Respecto al resto de metabuscadores, Ixquick, Profusion y Search obtienen unas frecuencias mejores que las de Excite, pues el porcentaje oscila entre el 41,3% y el 46%. Vivisimo sigue mostrando el mayor porcentaje de recursos que no contienen los términos.

Las frecuencias también aumentan respecto a los términos anteriores, y el documento en el que más número de veces aparecen es uno en el que se contabilizan 67 apariciones.

Tabla 5.4.3-4. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	33	100%

Este metabuscador, frente al resto, no recupera recursos con estos términos en la búsqueda booleana.

Tabla 5.4.3-5. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	30	90,9%
1	1	3%
3	1	3%
16	1	3%
Total	33	100%

Estos términos apenas aparecen en los recursos recuperados, pues como podemos observar, el 90,9% no los contiene. El porcentaje es similar al registrado por otros metabuscadores, excepto Vivisimo que no recupera recursos con estos términos. Como podemos observar sólo aparecen en tres recursos y sólo en uno de ellos con una frecuencia de dieciséis veces, siendo mínima en el resto.

Excite, como la mayoría de metabuscadores no recupera gran número de recursos con los términos solicitados en la búsqueda booleana, por lo que podemos señalar que, junto a los demás casos, no son las herramientas más adecuadas para este tipo de búsquedas.

Ixquick

Tabla 5.4.3-6. Nº de recursos analizados

Nº Recursos	21

Tabla 5.4.3-7. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	13	61,9%
1	2	9,5%
2	4	19%
16	2	9,5%
Total	21	100%

El porcentaje de no existencia de estos términos (61,9%) en los recursos recuperados es similar a Excite (63,6%). Respecto a las frecuencias, hay una pequeña diferencia ya que, aunque Excite presenta dos documentos con una repetición de los términos de

dieciocho y veinte veces, Ixquick recupera dos cuya frecuencia de aparición de términos es dieciséis.

Tabla 5.4.3-8. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	9	42,9%
1	1	4,8%
2	1	4,8%
7	1	4,8%
11	1	4,8%
15	1	4,8%
18	2	9,5%
20	2	9,5%
29	1	4,8%
36	1	4,8%
76	1	4,8%
Total	21	100%

En comparación con Excite, estos términos aparecen en este buscador con mayor frecuencia, aunque hay que tener en cuenta el menor número de páginas recuperadas por aquél. Ixquick recupera más recursos con bajas frecuencias y respecto a los de mayor frecuencia, éste recupera un recurso en el que el término se repite en 76 ocasiones, frente a las 67 ocasiones del recurso recuperado por Excite.

Tabla 5.4.3-9. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	18	85,7%
1	2	9,5%
2	1	4,8%
Total	21	100%

Ixquick junto con Vivisimo, sólo recupera tres recursos en los que aparecen estos términos, frente a ninguno en Excite. El resto de metabuscadores, aunque de forma limitada, tienen mejor comportamiento.

Tabla 5.4.3-10. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	90,5%
4	1	4,8%
6	1	4,8%
Total	21	100%

La recuperación de estos términos es similar a la presentada por Excite, si bien, Ixquick se ve superado por Excite ya que frente al documento de éste que presenta una frecuencia de dieciséis, el de Ixquick es de seis.

Del comportamiento de este buscador podemos decir que además de recuperar un limitado número de recursos para esta búsqueda, los que ofrece, apenas tienen los términos de búsqueda solicitados.

Profusion

Tabla 5.4.3-11. Nº de recursos analizados

Nº Recursos	46

Tabla 5.4.3-12. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	31	67,4%
1	6	13%
2	3	6,5%
4	2	4,3%
7	1	2,2%
12	1	2,2%
16	2	4,3%
Total	46	100%

Profusión presenta unos resultados en cuanto a estos términos, peores que el resto, dado el mayor número de recursos que no contienen los términos de búsqueda. En los recursos con frecuencias elevadas tiene un comportamiento inferior al de Excite.

Tabla 5.4.3-13. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	19	41,3%
1	3	6,5%
2	1	2,2%
3	2	4,3%
4	1	2,2%
5	1	2,2%
6	1	2,2%
7	1	2,2%
8	1	2,2%
9	2	4,3%
10	1	2,2%
11	1	2,2%
18	2	4,3%
20	2	4,3%
24	1	2,2%
29	1	2,2%
32	1	2,2%
36	2	4,3%
43	1	2,2%
45	1	2,2%
67	1	2,2%
Total	46	100%

Profusión supera en precisión técnica, en relación con estos términos, a los dos metabuscadores anteriores, caracterizándose por la recuperación de un mayor número de páginas en las que las frecuencias varían, así como de recursos en los que los valores de éstas son mayores.

Tabla 5.4.3-14. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	40	87%
1	1	2,2%
2	3	6,5%
3	1	2,2%
5	1	2,2%
Total	46	100%

Profusion, junto con Search, recuperan más recursos con estos términos que Excite, Ixquick y Vivisimo.

Tabla 5.4.3-15. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	40	87%
1	4	8,7%
3	1	2,2%
16	1	2,2%
Total	46	100%

El comportamiento de Profusión es similar al de Excite, pero le supera en los recursos de baja frecuencia de aparición de los términos.

Los porcentajes de recursos sin los términos de búsqueda son similares a los de los metabuscadores anteriores, aunque dado el mayor número de recursos recuperados, las frecuencias son más variadas y elevadas.

Search

Tabla 5.4.3-16. Nº de recursos analizados

Nº Recursos	50

Tabla 5.4.3-17. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	33	66%
1	8	16%
2	1	2%
3	1	2%
4	1	2%
6	1	2%
8	1	2%
12	2	4%
13	1	2%
47	1	2%
Total	50	100%

El porcentaje de recursos que no contienen estos términos es similar a Excite, si bien, en relación con los recursos de mayor frecuencia, Search recupera un recurso con una frecuencia de cuarenta y siete frente a veinte, que es la máxima en Excite.

Tabla 5.4.3-18. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	23	46%
1	1	2%
3	1	2%
4	3	6%
6	1	2%
7	1	2%
8	1	2%
9	2	4%
10	1	2%
11	1	2%
16	1	2%
18	1	2%
23	1	2%
24	2	4%
27	1	2%
32	1	2%
35	1	2%
36	1	2%
38	1	2%
43	1	2%
44	1	2%
45	2	4%
76	1	2%
Total	50	100%

En relación con estos términos, Search recupera un mayor porcentaje que Excite y aunque en este aspecto no supera a Profusión, sí lo hace en cuanto a los recursos con frecuencias medias y altas.

Tabla 5.4.3-19. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	38	76%
1	4	8%
2	3	6%
3	1	2%
4	1	2%
5	3	6%
Total	50	100%

Supera en la recuperación de estos términos a Excite, Ixquick y Vivisimo, y se asimila mucho a los resultados presentados por buscadores como Google o metabuscadores como Profusion.

Tabla 5.4.3-20. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	44	88%
1	4	8%
3	1	2%
6	1	2%
Total	50	100%

La recuperación de recursos con estos términos es similar a la que presentan Profusión y Excite, si bien estos metabuscadores recuperan un recurso en el que los términos aparecen 16 veces frente al máximo de Search en el que aparecen en seis ocasiones.

Search es el metabuscador que ofrece mejores resultados que el resto dado que los recursos que recupera obtienen frecuencias más variadas que el resto.

Vivisimo

Tabla 5.4.3-21. Nº de recursos analizados

Nº Recursos	48

Tabla 5.4.3-22. Frecuencia y nº de recursos en los que aparecen los términos "information retrieval"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	43	89,6%
1	1	2,1%
2	1	2,1%
4	1	2,1%
6	1	2,1%
12	1	2,1%
Total	48	100%

Destaca este metabuscador por el alto número de recursos que no contienen estos términos de búsqueda (89,6%) característica que, en general, se mantiene en todos los términos que afectan a esta búsqueda.

Tabla 5.4.3-23. Frecuencia y nº de recursos en los que aparecen los términos "digital libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	43	89,6%
1	2	4,2%
4	2	4,2%
10	1	2,1%
Total	48	100%

Sigue mostrando respecto al resto, una recuperación más limitada en cuanto a la aparición de los términos de búsqueda, así como en las frecuencias.

Tabla 5.4.3-24. Frecuencia y nº de recursos en los que aparecen los términos "electronic libraries"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	46	95,8%
1	1	2,1%
2	1	2,1%
Total	48	100%

La recuperación de recursos con estos términos es mínima, ofreciendo respecto al resto de buscadores, los peores resultados.

Tabla 5.4.3-25. Frecuencia y nº de recursos en los que aparecen los términos "virtual libraries"

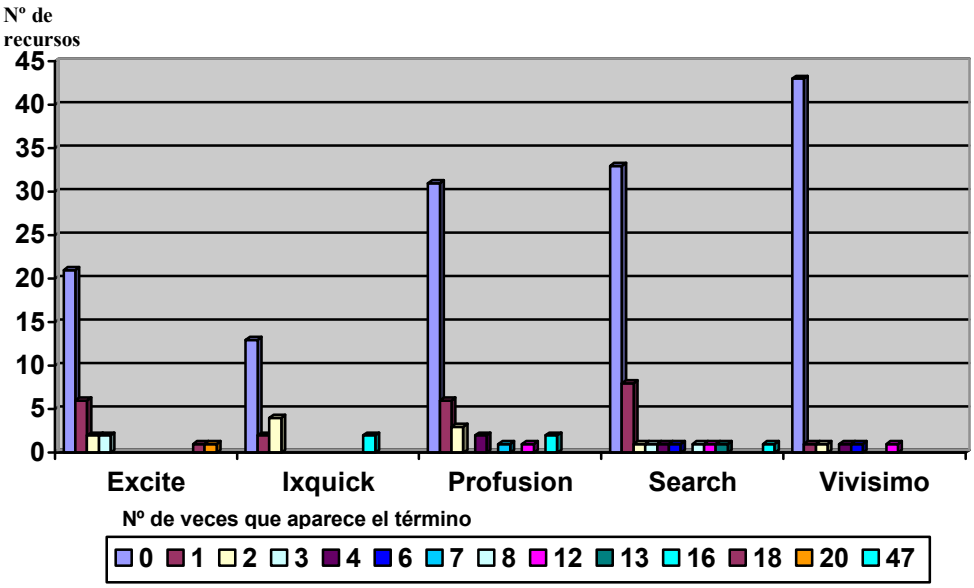
Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	48	100%

No recupera recursos con estos términos.

Los resultados de Vivisimo confirman que los metabuscadores no son las herramientas adecuadas para este tipo de búsquedas ya que apenas obtienen resultados con los términos solicitados.

5.4.4. Análisis comparativo de los metabuscadores

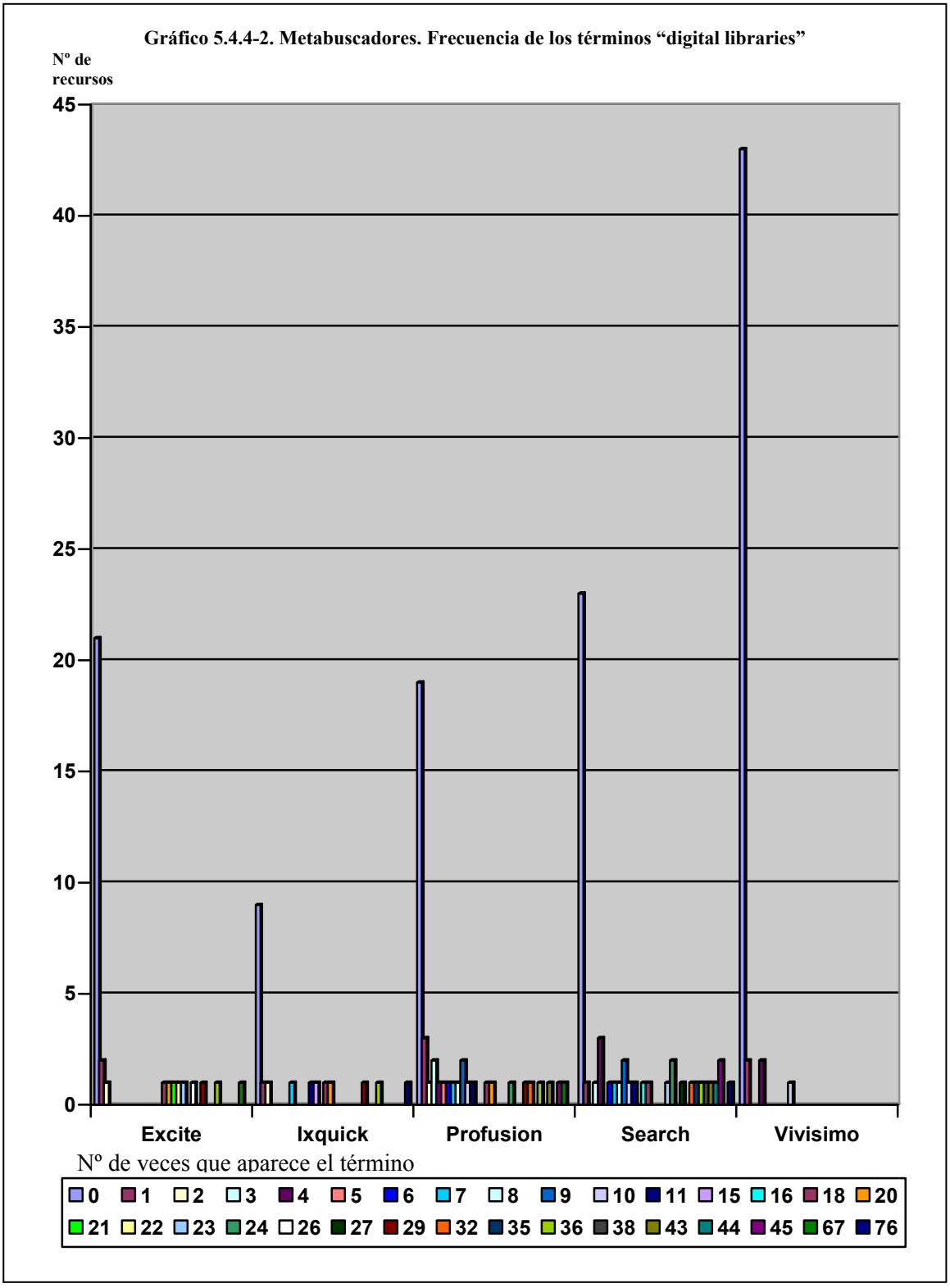
Gráfico 5.4.4-1. Metabuscadores. Frecuencia de los términos "information retrieval"



El comportamiento de los metabuscadores guarda relación con lo observado en los motores de búsqueda, aunque como podemos apreciar, existen claras diferencias entre ellos. Antes de comparar la recuperación realizada por los metabuscadores evaluados, debemos señalar que Dogpile y Surfwax no ofrecieron resultados, e Ixquick recuperó tan sólo veintiún recursos. Sin embargo, este metabuscador y Profusión recuperan los recursos con mayores frecuencias de repetición de los términos de búsqueda.

Vivisimo recupera el mayor número de registros sin los términos de búsqueda frente a Excite, Profusion y Search, que son los que más páginas con estos términos ofrecen.

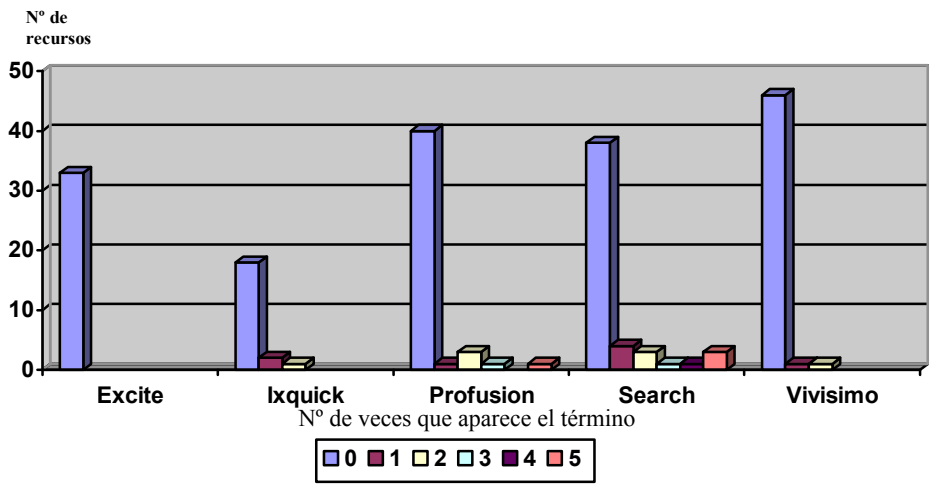
Gráfico 5.4.4-2. Metabuscadores. Frecuencia de los términos "digital libraries"



El perfil que presentan los metabuscadores en esta gráfica también es muy similar al de los motores, destacando en ambos casos el importante número de recursos que no contienen los términos.

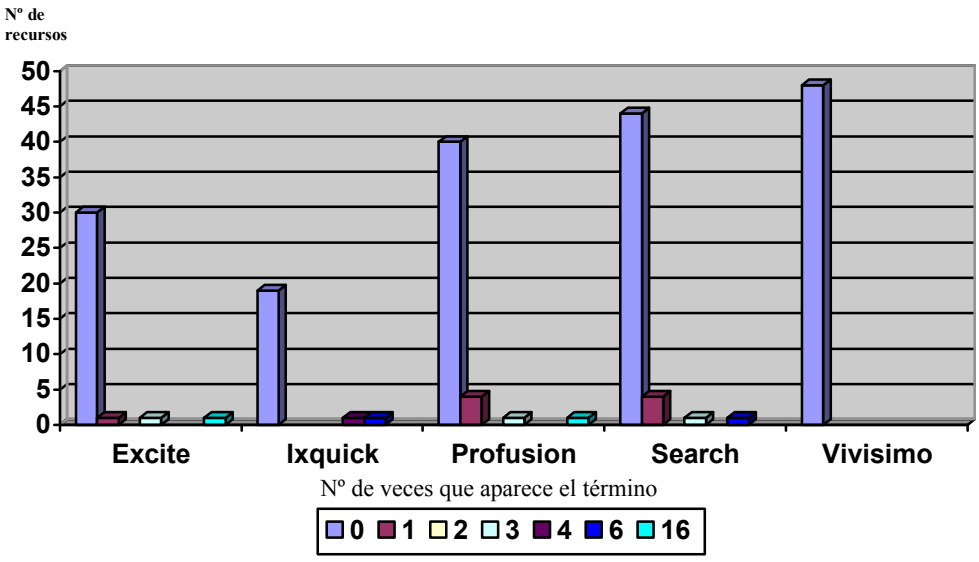
Search y Profusión, por este orden, son los metabuscadores que recuperan un mayor número de páginas con los términos de búsqueda, ofreciendo un variado número de frecuencias. Vivisimo destaca por el alto número de recursos recuperados que no los contienen.

Gráfico 5.4.4-3. Metabuscadores. Frecuencia de los términos "electronic libraries"



Search seguido de Profusion ofrecen mejores resultados que el resto de metabuscadores, que se caracterizan por ofrecer muy pocos recursos y con bajas frecuencias de estos términos. En este sentido debemos destacar el caso de Excite que no recuperó ningún registro con estos términos.

Gráfico 5.4.4-4. Metabuscadores. Frecuencia de los términos "virtual libraries"



Los datos que nos ofrecen los metabuscadores en esta ocasión son reflejo del mal funcionamiento de estas herramientas ya que o recuperan pocos recursos con esto términos o no recuperan, como en el caso de Vivisimo. A éste último le ocurre en esta ocasión lo mismo que a Excite en el caso anterior, lo que indica que estos defectos en la recuperación afectan tanto a unos como a otros.

Esta búsqueda refleja de forma clara la relación entre motores y metabuscadores, que como hemos visto, éstos últimos suelen ofrecer resultados que en ningún caso mejoran los observados en los motores.

5.5. Búsqueda de frase

En este tipo de búsqueda, se trata de valorar si los buscadores funcionan correctamente cuando se requiere que los términos de búsqueda expresados entre comillas, aparezcan todos ellos y en el orden que se solicita. De aquí que en este caso centremos el análisis en las frecuencias de aparición de la frase : "Natural language processing"

5.5.1. Análisis individualizado por motores de búsqueda

Google

Tabla 5.5.1-1. Nº de recursos analizados

Nº Recursos	47

Tabla 5.5.1-2. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	5	10,6%
1	2	4,3%
2	5	10,6%
3	8	17%
4	8	17%
5	4	8,5%
6	2	4,3%
7	1	2,1%
8	1	2,1%
11	2	4,3%
13	2	4,3%
15	1	2,1%
18	1	2,1%
20	1	2,1%
22	1	2,1%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
26	1	2,1%
28	1	2,1%
42	1	2,1%
Total	47	100%

En la búsqueda por frase, los porcentajes de recursos que no contienen los términos de búsqueda disminuyen sensiblemente respecto al resto de búsquedas. En este caso el 10,6% de los recursos recuperados no las contienen, lo que a pesar de todo, nos parece un porcentaje alto, dado que el tipo de consulta requiere que aparezcan.

La frecuencia de aparición de la frase en los documentos recuperados es muy variable ya que recupera tanto dos documentos en los que aparece una sola vez, como otro en el que se repite en 42 ocasiones.

En cualquier caso, y a la vista de los resultados de otras búsquedas, podemos decir que Google muestra un buen comportamiento en la búsqueda por frase.

MSN

Tabla 5.5.1-3. Nº de recursos analizados

Nº Recursos	50

Tabla 5.5.1-4. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	5	10%
1	1	2%
2	5	10%
3	12	24%
4	4	8%
5	7	14%
6	1	2%
7	4	8%
8	3	6%
11	1	2%
13	2	4%
15	1	2%
18	1	2%
28	1	2%
30	1	2%
42	1	2%
Total	50	100%

En MSN, el porcentaje de recursos que no contienen la frase se mantiene, aunque disminuye respecto a Google. A diferencia de éste, ofrece un mayor número de recursos en los que los términos se repiten tres veces, pasando de los ocho de Google a los 12 de este buscador. Sin embargo descienden a la mitad los que aparecen cuatro veces, aunque vuelve a subir a siete el número de recursos en los que aparecen cinco veces. También son superiores las frecuencias de recursos en los que aparecen siete y ocho veces. Por tanto, aunque las diferencias no son muy significativas, sí que nos indican una diferencia en los resultados que ofrecen unos y otros, caracterizando a MSN unas frecuencias medias ligeramente superiores a las vistas en Google. Respecto a los recursos de frecuencias altas, ambos buscadores muestran una recuperación similar, por tanto, en esta búsqueda podemos hablar de similitud entre ambos.

WiseNut

Tabla 5.5.1-5. Nº de recursos analizados

Nº Recursos	49

Tabla 5.5.1-6. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12,2%
1	2	4,1%
2	4	8,2%
3	8	16,3%
4	7	14,3%
5	7	14,3%
6	2	4,1%
7	4	8,2%
8	1	2%
10	1	2%
11	1	2%
13	1	2%
18	1	2%
20	1	2%
25	1	2%
28	1	2%
42	1	2%
Total	49	100%

El porcentaje de recursos que no contienen los términos aumenta en este buscador ligeramente respecto a los anteriores (12,2%) aunque en general, la recuperación tiene relación con la mostrada por ellos.

Yahoo

Tabla 5.5.1-7. Nº de recursos analizados

Nº Recursos	49

Tabla 5.5.1-8. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	2	4,1%
1	2	4,1%
2	6	12,2%
3	6	12,2%
4	4	8,2%
5	3	6,1%
6	2	4,1%
7	5	10,2%
8	1	2%
9	1	2%
11	2	4,1%
13	3	6,1%
15	1	2%
18	1	2%
20	1	2%
21	1	2%
23	1	2%
26	1	2%
28	1	2%
30	2	4,1%
35	1	2%
42	1	2%
63	1	2%
Total	49	100%

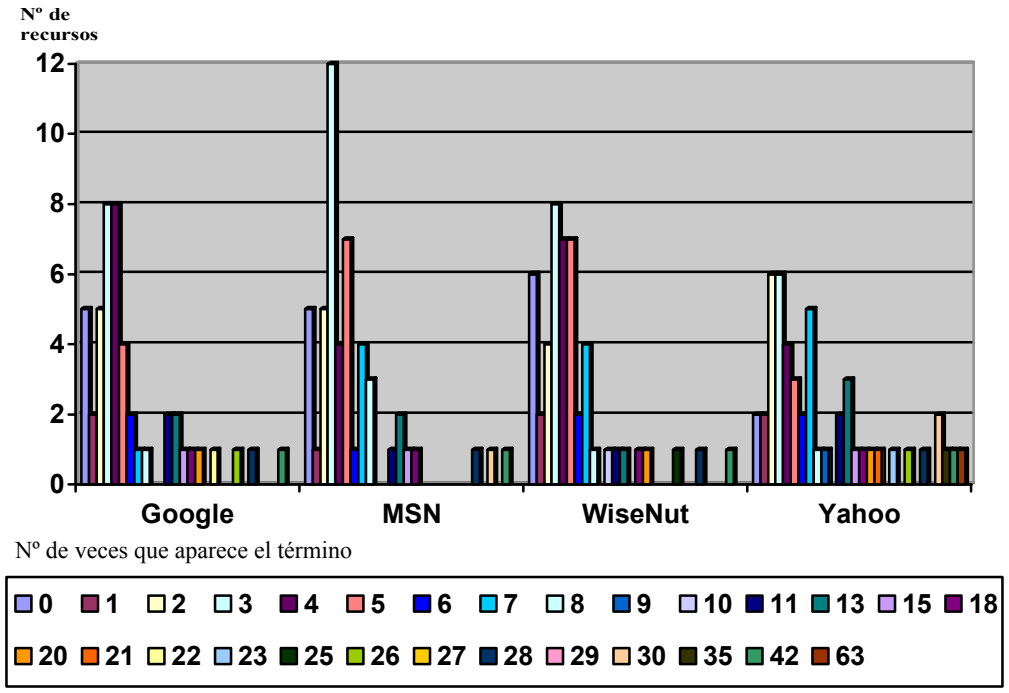
Es el buscador que mejor funciona en esta búsqueda ya que el porcentaje de recursos que no contiene los términos desciende hasta el 4,1%, lo que supone un sensible des-

censo respecto al resto. Esto se traduce en una mejora en la recuperación de documentos en los que los términos de la frase aparecen con frecuencias mayores y muy variadas.

5.5.2. Análisis comparativo de los motores de búsqueda

En la búsqueda por frase, los buscadores, a excepción de Teoma, Surfwax y Dogpile, que no recuperaron recursos en esta búsqueda, funcionan de forma más correcta que en la búsqueda booleana.

Gráfico 5.5.2-1. Motores. Frecuencia de los términos "Natural language processing"



A la vista de los datos anteriores, podemos afirmar que Yahoo es el motor de búsqueda que muestra un mejor funcionamiento en la búsqueda por frase ya que es el que recupera un mayor número de registros con todos los términos y además es el que recupera el registro en el que la frase aparece el mayor número de veces (63). Google y WiseNut tienen un comportamiento similar, en el que predominan documentos con bajas frecuencias, aspecto que se acentúa en el caso de MSN que recupera más recursos en los que la frase aparece un número más limitado de veces que el resto, como es el caso de los doce documentos en los que aparece la frase en tres ocasiones.

5.5.3. Análisis individualizado por metabuscadores

Excite

Tabla 5.5.3-1. Nº de recursos analizados

Nº Recursos	50

Tabla 5.5.3-2. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	6	12%
1	1	2%
2	6	12%
3	12	24%
4	5	10%
5	1	2%
6	1	2%
7	4	8%
8	2	4%
10	2	4%
11	1	2%
13	3	6%
15	1	2%
18	1	2%
20	1	2%
25	1	2%
28	1	2%
42	1	2%
Total	50	100%

Excite presenta en la búsqueda por frase un comportamiento similar al de la mayoría de los buscadores, tanto en lo que se refiere al escaso número de recursos que no contienen la frase, como las frecuencias de ella dentro de los documentos. Así, sólo recupera una página en la que la frase aparece sólo una vez pero aumenta el número de páginas en las que los términos aparecen dos, tres y cuatro veces. La recuperación de páginas en las que aparecen cinco y seis veces desciende para aumentar de nuevo el número de páginas en las que aparece de siete a trece veces. Finalmente se mantienen las frecuencias mayores como en los motores de búsqueda.

Ixquick

Tabla 5.5.3-3. Nº de recursos analizados

Nº Recursos	43

Tabla 5.5.3-4. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	10	23,3%
1	1	2,3%
2	3	7%
3	10	23,3%
4	3	7%
5	3	7%
7	1	2,3%
8	1	2,3%
11	2	4,7%
13	3	7%
18	1	2,3%
19	1	2,3%
25	1	2,3%
28	1	2,3%
30	1	2,3%
42	1	2,3%
Total	43	100%

Ixquick no recupera bien en esta búsqueda ya que el porcentaje de páginas que no contiene los temas de búsqueda duplica los resultados que ofrecen los demás.

La recuperación se basa en facilitar páginas con bajas frecuencias, manteniéndose, como en el buscador anterior, las de frecuencias altas.

Profusion

Tabla 5.5.3-5. Nº de recursos analizados

Nº Recursos	39

Tabla 5.5.3-6. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	10	25,6%
2	3	7,7%
3	6	15,4%

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
4	3	7,7%
5	4	10,3%
7	2	5,1%
8	1	2,6%
11	1	2,6%
13	3	7,7%
15	1	2,6%
18	1	2,6%
26	1	2,6%
28	2	5,1%
42	1	2,6%
Total	39	100%

Profusión tiene un comportamiento similar al de Ixquick, aunque se puede apreciar una ligera mejoría al recuperar más recursos con frecuencias de aparición de la frase medias y altas.

Search

Tabla 5.5.3-7. Nº de recursos analizados

Nº Recursos	47

Tabla 5.5.3-8. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	3	6,4%
1	3	6,4%
2	4	8,5%
3	8	17%
4	10	21,3%
5	4	8,5%
6	1	2,1%
7	2	4,3%
8	1	2,1%
11	1	2,1%
13	2	4,3%
15	1	2,1%
18	2	4,3%
25	1	2,1%
26	1	2,1%
28	2	4,3%
42	1	2,1%
Total	47	100%

Search tiene el mejor comportamiento respecto al resto de metabuscadores ya que el porcentaje de recursos que no contienen los términos desciende al 6,4% y las páginas en las que aparecen cuatro y cinco veces, son superiores a las de Excite, al igual que ocurre con las páginas de frecuencias altas, que también lo superan.

Vivisimo

Tabla 5.5.3-9. Nº de recursos analizados

Nº Recursos	43

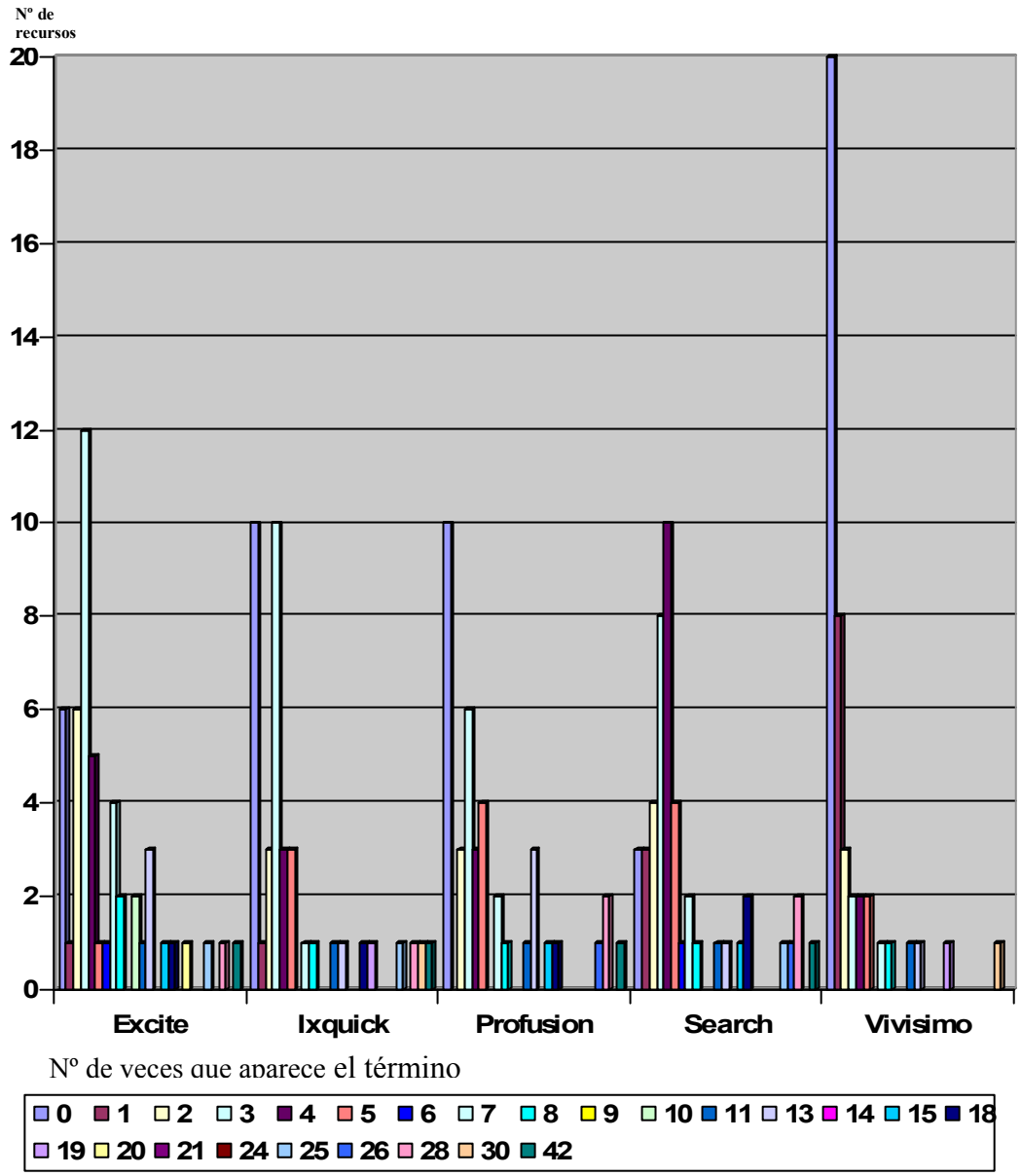
Tabla 5.5.3-10. Frecuencia y nº de recursos en los que aparece la frase "Natural language processing"

Nº de veces que aparecen los términos	Nº de recursos	Porcentaje
0	20	46,5%
1	8	18,6%
2	3	7%
3	2	4,7%
4	2	4,7%
5	2	4,7%
7	1	2,3%
8	1	2,3%
11	1	2,3%
13	1	2,3%
19	1	2,3%
30	1	2,3%
Total	43	100%

Es el metabuscador que peor comportamiento tiene en esta búsqueda ya que el porcentaje de recursos que no contienen los términos asciende al 46,5%. Además las páginas con bajas frecuencias de aparición de la frase son las que tienen mayores valores, no recuperando apenas páginas de altas frecuencias.

5.5.4. Análisis comparativo de los Metabuscadores

Gráfico 5.5.4-1. Metabuscadores. Frecuencia de los términos "Natural language processing"



En los metabuscadores destaca, al igual que ocurría con los motores, la recuperación de recursos con bajas frecuencias de aparición, aunque se combinan con la recuperación de otros recursos en los que el uso de la frase es más abundante. En este sentido podemos decir que todos los metabuscadores realizan una recuperación aceptable, a excepción de Vivisimo, que destaca por el alto número de recursos que no contienen la frase de búsqueda, lo que denota problemas de funcionamiento.

Search y Excite son los metabuscadores con mayor precisión técnica ya que ofrecen, además de no muy alto número de páginas con bajas frecuencias, un importante número de páginas con frecuencias medias, manteniéndose las frecuencias altas. Habría que destacar además en Search el bajo número de recursos que no contienen la frase de búsqueda.

Al igual que en la búsqueda anterior, esta búsqueda es un reflejo de la relación entre motores y metabuscadores en la recuperación.

Por tanto, en cuanto a esta búsqueda, a modo de resumen, podemos resaltar el comportamiento de Yahoo y de Search, correspondiendo a Vivisimo los peores resultados.

5.6. Búsqueda por campo

Las siguientes tablas recogen los resultados que arrojan los buscadores al lanzar sobre ellos la búsqueda en el campo de título (*intitle:*) con la expresión: *information retrieval.* Se analizan las coincidencias entre los términos de búsqueda, bien se encuentren formando parte del título principal del recurso, es decir, el que aparece más destacado, o bien si la coincidencia se da con títulos de apartados o con otro tipo de títulos.

Tabla 5.6-1. Motores. Recursos que contienen los términos de búsqueda en el título

	Título principal	Título secundario	Otros	Total
Google	43 (89,6%)	2 (4,2%)	3 (6,3%)	48
MSN	46 (93,9%)	2 (4,1%)	1(2%)	49
Teoma (Ask)	Sin resultados			
WiseNut	0	0	30 (100%)	30
Yahoo	44 (88%)	1 (2%)	5 (10%)	50

Como podemos apreciar, en esta búsqueda también se aprecia una mayor precisión ya que a excepción de WiseNut, el resto de motores tienen en el título los términos solicitados.

Tabla 5.6-2. Metabuscadores. Recursos que contienen los términos de búsqueda en el título

	Título principal	Título secundario	Otros	Total
Dogpile	Sin resultados			
Excite	17 (34,7%)	1 (2%)	31 (63,3%)	49

	Título principal	Título secundario	Otros	Total
Ixquick	20 (87%)	1 (4,3%)	2 (8,7%)	23
Profusion	14 (56%)	0	11 (44%)	25
Search	22 (46,8%)	1 (2,1%)	24 (51,1%)	47
Surfwax	Sin resultados			
Vivisimo	43 (89,6%)	1 (2,1%)	4 (8,3%)	48

Entre los metabuscadores destaca el comportamiento de Vivisimo, que aunque en las anteriores búsquedas mostraba un mal funcionamiento, en la búsqueda por campo es el metabuscador que mayor porcentaje de recursos con los términos en el título recuperó (89,6%) lo que indica, en este sentido, una correcta recuperación. A continuación destacan los porcentajes que ofrece Ixquick (87%). Search, Excite y Profusion recuperan recursos en un porcentaje claramente inferior a Vivisimo, por lo que podemos afirmar que existe dificultad por parte de algunos metabuscadores, especialmente estos últimos, en traducir y lanzar este tipo de búsqueda más precisa, de forma que sea correctamente interpretada por los buscadores, lo que les impide obtener los resultados apropiados. En definitiva podemos observar que los motores de búsqueda realizan una recuperación más precisa en este tipo de búsquedas por lo que resultan más recomendables en búsquedas por campo.

Análisis global de las seis búsquedas

En la valoración que ofrecemos a continuación de los buscadores, hemos tenido en cuenta tanto el número de recursos que no contienen los términos de búsqueda como la recuperación que estas herramientas hacen de recursos con bajas y altas frecuencias de aparición de los términos. En base a estas observaciones valoramos de uno a tres la recuperación que realiza cada herramienta de búsqueda del término o términos de consulta, obteniendo finalmente la puntuación correspondiente a cada una de las búsquedas. De todo lo observado anteriormente podemos extraer las siguientes conclusiones:

El comportamiento de las herramientas de búsqueda denota en la primera búsqueda una baja precisión que sólo mejora en la búsqueda por frase y por campo.

En relación con la búsqueda por lenguaje natural y teniendo en cuenta el porcentaje de recursos en los que aparecen los términos, Teoma y Yahoo son los buscadores en los que aparecen en un mayor número de documentos. Les siguen MSN y Google, siendo WiseNut el que menor número de recursos con los términos recupera.

Valorando la aparición de términos de consulta específicos como es el caso del término compuesto best-match, así como las altas frecuencias de aparición de los términos en los recursos recuperados, Teoma y Yahoo son los que ofrecen mejores resultados.

En la búsqueda con operadores de existencia, atendiendo a la aparición de los términos en el mayor número de páginas recuperadas, así como a las mayores frecuencias de aparición, la mejor recuperación es la ofrecida por Yahoo y WiseNut, a pesar de que Yahoo no muestra un buen comportamiento en la recuperación de los términos *web* e *information*, lo que tal vez sea debido a que se trata de términos bastante frecuentes en la Web. Google no alcanza tan buena posición debido a los altos índices de recursos que no contienen los términos de búsqueda.

Respecto a la búsqueda booleana, el comportamiento de Google, tanto en la recuperación de documentos con los términos de búsqueda, como en las frecuencias de aparición de los términos en los recursos recuperados, es superior a Yahoo y a MSN. No obstante, esta búsqueda es ilustrativa de la deficiente forma en que recuperan estas herramientas, ya que los recursos recuperados tanto por motores como por metabuscadores, obtienen los más bajos porcentajes de frecuencias registrados en las seis búsquedas.

Por último, en la búsqueda por frase, Yahoo es el motor que recupera más recursos que contienen la frase de búsqueda, manteniendo un buen comportamiento en las frecuencias de aparición de los términos en los documentos recuperados. Le siguen Google y WiseNut, quedando MSN a más distancia al recuperar un gran número de recursos en los que la frase no aparece en más de tres ocasiones. Entre los metabuscadores destaca la recuperación que hace Search, cuyos datos mejoran los presentados por Google, MSN o WiseNut.

Respecto a los metabuscadores, en la búsqueda en lenguaje libre, Profusión y Vivisimo, son los que recuperan un mayor número de páginas que contienen los términos de búsqueda, y ofrecen además, un mayor número de documentos con distintas frecuencias.

Excite también tiene un buen comportamiento, pero se caracteriza por recuperar fundamentalmente, recursos con baja frecuencia de aparición de los términos.

Search registra un aceptable comportamiento en cuanto a que recupera recursos con frecuencias altas de los términos, pero muestra problemas, al recuperar un excesivo número de recursos sin los términos de búsqueda.

Dogpile se caracteriza por recuperar gran número de recursos sin los términos de búsqueda, correspondiendo a Surfwax los peores resultados.

En la búsqueda con operadores de existencia destaca Search, siendo muy similar la realizada por el resto de metabuscadores, principalmente Excite, Ixquick y Vivisimo. Dogpile y Surfwax muestran, en este caso, los peores registros.

La búsqueda booleana es en la que los metabuscadores tienen un peor comportamiento, ya que apenas recuperan registros con los términos de búsqueda. A pesar de todo, el mejor comportamiento corresponde a Profusion y Search, seguidos de Excite e Ixquick, correspondiendo la última posición a Vivisimo.

En la búsqueda por frase es Search el que recupera más páginas que la contienen, y además, registra un buen comportamiento tanto en la recuperación de recursos en los que la frase aparece en pocas ocasiones, como de otros, en los que aparece con mayor frecuencia. Excite tiene un comportamiento similar, pero recupera más recursos en los que la frase se repite en pocas ocasiones. Lo mismo ocurre con Ixquick y Profusión, frente a Vivisimo, que se caracteriza, de nuevo, por el alto número de recursos que no contienen la frase de búsqueda.

En la búsqueda por campo se da una mayor precisión en la recuperación, fundamentalmente en el caso de MSN, que se acerca al 100%, seguidos por Google y Yahoo. Entre los metabuscadores, sólo Ixquick y Vivisimo se aproximan a los datos obtenidos por los motores.

6. Análisis de la ordenación de resultados o ranking

La ordenación de los resultados en los listados de búsqueda es uno de los aspectos más importantes tanto desde el punto de vista del usuario como de los desarrolladores de herramientas de búsqueda, ya que para los primeros, es importante que figuren en los primeros lugares, de forma preferente, los resultados más relevantes, es decir, los que tienen mayor relación con lo que se busca. Por su parte, los desarrolladores de estos sistemas, buscan fórmulas para que, en la medida de lo posible, esto sea así. Nosotros para comprobarlo hemos centrado el análisis del ranking en valorar el uso de la metainformación, la aparición del término de búsqueda y el peso que dicho término tiene en el documento, así como en valorar si existe o no, relación entre estos valores para la ordenación. Para ello nos hemos basado en los datos aportados en la primera búsqueda, que al estar compuesta por un solo término, nos permite valorar estos aspectos con una mayor claridad.

6.1. Utilización de la metainformación

El programa informático utilizado para la valoración de diferentes aspectos relacionados con la evaluación, *HTML Analyzer*, nos facilitó, el cálculo de la existencia de los términos de búsqueda en las etiquetas KEY, que contienen palabras clave relativas al contenido del documento, y DESCRIPTION que recogen un pequeño resumen. Estas etiquetas están presentes en las páginas HTML, de forma visible al visualizarlos en formato fuente.

Para la elaboración de las siguientes tablas hemos agrupado los resultados recuperados por cada buscador en grupos de diez páginas, manteniendo el orden correlativo en el que aparecen en los listados, lo que nos permite analizar si el término de búsqueda aparece en estas etiquetas en los primeros y si existe una clara influencia en su ordenación.

Tabla 6.1-1. Motores. Frecuencia del término de búsqueda en Metaetiquetas.

	Búsqueda 1									
	Etiqueta Key					Etiqueta Description				
	1-10	11-20	21-30	31-40	41-50	1-10	11-20	21-30	31-40	41-50
Google	0	2	0	0	0	0	4	1	2	0
MSN	0	0	1	0	0	0	0	0	0	0
Teoma (Ask)	2	7	7	8	6	2	7	4	3	1
WiseNut	3	1	1	0	1	3	0	1	0	0
Yahoo	2	1	1	1	1	2	1	0	0	0

En la recuperación de los motores de búsqueda, llama la atención el pequeño porcentaje de páginas que contienen el término de búsqueda en las etiquetas META. Teoma constituye la excepción, aunque, al acumular las máximas frecuencias a partir de los diez primeros resultados, podemos afirmar que este aspecto, en este caso, no parece ser determinante para la ordenación. Sí que se aprecia una mejor relación en WiseNut y Yahoo, en los que las mayores frecuencias se hayan en los primeros resultados, apareciendo después en orden descendente.

MSN es el buscador que menos recursos recupera con el término de búsqueda en la metainformación. En Google, podemos observar una mayor frecuencia de recursos que contienen el término en la metaetiqueta DESCRIPTION.

Tabla 6.1-2. Metabuscadores. Frecuencia del término de búsqueda en Metaetiquetas

	Búsqueda 1									
	Etiqueta Key					**Etiqueta Description**				
	1-10	**11-20**	**21-30**	**31-40**	**41-50**	**1-10**	**11-20**	**21-30**	**31-40**	**41-50**
Dogpile	1	4	2	1	0	0	4	2	0	0
Excite	1	3	3	1	1	1	5	1	1	0
Ixquick	3	3	1	0	s.r*.	2	3	1	0	s.r*.
Profusion	3	3	3	2	0	2	4	2	1	0
Search	3	2	1	1	0	2	2	1	3	0
Surfwax	5	0	s.r*.	s.r*.	s.r*.	4	0	s.r*.	s.r*.	s.r*.
Vivisimo	1	2	3	3	3	1	2	2	7	0

*s.r.: sin resultados

En la presentación de los resultados de los metabuscadores, la información de las etiquetas META sí que parece tener un cierto peso no sólo para la ordenación, sino también en la selección de los recursos facilitados por los motores, ya que la aparición del término de búsqueda es más frecuente que en éstos.

En este sentido, podemos afirmar que prácticamente todos los metabuscadores parecen tener en cuenta la metainformación en la presentación de los resultados. Vivisimo y Profusion son los que más recursos recuperan con el término de búsqueda en la metainformación, y llama la atención Surfwax, que parece otorgar gran peso a este aspecto para la ordenación de sus resultados, ya que cinco de los diez primeros recursos contienen el término en la metaetiqueta KEY y cuatro en la etiqueta DESCRIPTION.

Por tanto, ¿Se puede apreciar una relación entre estos aspectos?

Aunque somos conscientes que los resultados son muy limitados para poder establecer unas conclusiones categóricas respecto a la influencia de estas etiquetas en la ordenación de los resultados, sí que nos parecen válidos para conocer el diferente funcionamiento de estas herramientas, y explicar así las diferencias entre unos y otros. Es necesario continuar investigando para conocer mejor los mecanismos que influyen en la ordenación ya que puede servir de ayuda para la elaboración de la forma adecuada para obtener un buen posicionamiento de las páginas de contenido académico.

6.2. Frecuencia y peso del término de búsqueda en las páginas recuperadas

Se analizan a continuación la frecuencia de aparición del término de búsqueda y su peso, así como su aparición en el texto o en hiperenlaces de las páginas HTML. Para valorar su función en la ordenación de los resultados de cada buscador, como en el caso anterior, hemos utilizado el programa HTML Analyzer. Las páginas recuperadas se agrupan, manteniendo el orden de recuperación, en grupos de diez, y se calcula la media, la desviación típica y la mediana de los valores de las variables, lo que nos permite obtener datos de mayor precisión. Así, por ejemplo, ante la variabilidad observada en el valor de la Media de la Frecuencia y el Peso, para el análisis nos basaremos tanto en estos valores como en el de la Mediana, que en este caso, es más representativa.

Motores de búsqueda

Google

Tabla 6.2-1. Google. Frecuencia y peso del término de búsqueda.

		Búsqueda 1			
		Frecuencia de aparición del término	Peso	En el texto	En hiperenlace
Resultados 1 a 10	Media	13,8	30,6	13,7	0,6
	Desv. típ.	13,7	30,4	13,7	0,9
	Mediana	12	17,3	11	0
	Mínimo	1	2,7	1	0
	Máximo	36	78,6	36	2
	N	9	9	9	9
Resultados 11 a 20	Media	10,2	22,7	9,3	0,6
	Desv. típ.	12,4	28,7	12,9	1,0
	Mediana	3,5	12,5	2,5	0
	Mínimo	0	0	0	0
	Máximo	36	84,8	36	3
	N	10	10	10	10
Resultados 21 a 30	Media	2,7	2995,6	2,4	0,3
	Desv. típ.	1,9	6928,4	1,9	1
	Mediana	2	6,5	2	0
	Mínimo	1	0,1	0	0
	Máximo	7	20616	6	3
	N	9	9	9	9
Resultados 31 a 40	Media	6,5	8695,3	5,4	0
	Desv. típ.	6,8	27462,2	7,3	0
	Mediana	5	7,4	1,5	0
	Mínimo	0	0	0	0
	Máximo	21	86854	21	0
	N	10	10	10	10
Resultados 41 a 50	Media	6	628,2	5,5	0
	Desv. típ.	6,3	1731,2	6,6	0
	Mediana	4	5,7	3	0
	Mínimo	1	0,5	0	0
	Máximo	20	4912	20	0
	N	8	8	8	8
Total	Media	7,9	2596,5	7,3	0,3
	Desv. típ.	9,7	13100,0	10,0	0,8
	Mediana	4	7,5	3	0
	Mínimo	0	0	0	0
	Máximo	36	86854	36	3
	N	46	46	46	46

A la vista de los resultados de la tabla anterior podemos afirmar que Google parece tener en cuenta, en los treinta primeros resultados, la frecuencia de aparición de los términos y el peso, pero a partir de ellos, los valores de la Mediana superan a los de las

series anteriores. Sobre todo destaca la frecuencia y el peso de los diez primeros resultados. Además es interesante apreciar, en relación con el algoritmo que Google aplica a la ordenación, que al tener en cuenta la valoración de la aparición de los términos en los hiperenlaces, los dos primeros grupos de páginas (resultados 1 a la 20), tienen como valor 0,6 correspondiendo el valor 0,3 al tercer grupo (del 21 al 30) y 0 en los dos últimos. Este buscador ordena en los diez primeros lugares las páginas en las que los valores de la Media y la Mediana de la frecuencia son más altos. Los valores descienden paulatinamente, ascendiendo de nuevo a partir del treinta. Llaman la atención en Google los altos valores de la Media del Peso, sobre todo para los recursos del veintiuno al treinta.

También es interesante apreciar cómo corresponden a los primeros veinte resultados los valores más elevados para las frecuencias del término, lo que indica una intervención en la ordenación tanto del análisis de frecuencias como de los cálculos que efectúan sus algoritmos específicos, entre los que se aprecia que existe relación, si bien, esta relación se mantiene de forma clara en los treinta primeros resultados, a partir de los cuales, los valores vuelven a aumentar.

Este comportamiento de Google parece reflejar, que los algoritmos de ordenación valoran, siguiendo una determinada fórmula, los treinta primeros resultados, haciendo intervenir, a partir de ellos, otros parámetros para su ordenación. No obstante, es necesario seguir investigando estos aspectos para poder confirmar dicha observación.

MSN

Tabla 6.2-2. MSN. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
		Frecuencia de aparición del término	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	5,7	13,6	5,7	0,7
1 a 10	Desv. típ.	9	25,9	9	1,2
	Mediana	1	3,8	1	0
	Mínimo	0	0	0	0
	Máximo	25	84,8	25	3
	N	10	10	10	10

Resultados	Media	2,6	3	2,6	0,3
11 a 20	Desv. típ.	3,5	5,1	3,5	0,7
	Mediana	1	1,2	1	0
	Mínimo	0	0	0	0
	Máximo	9	15,9	9	2
	N	9	9	9	9
Resultados	Media	1,3	3,1	1,2	0
21 a 30	Desv. típ.	1,2	4,4	1,3	0
	Mediana	1	0,7	1	0
	Mínimo	0	0	0	0
	Máximo	4	13,1	4	0
	N	9	9	9	9
Resultados	Media	1,3	1,7	1,2	0,3
31 a 40	Desv. típ.	0,9	2,4	1	0,7
	Mediana	1	0,6	1	0
	Mínimo	0	0	0	0
	Máximo	3	6,6	3	2
	N	9	9	9	9
Resultados	Media	1,3	1,5	1,3	0,2
41 a 50	Desv. típ.	0,9	2,0	0,9	0,7
	Mediana	1	0,5	1	0
	Mínimo	0	0	0	0
	Máximo	3	4,9	3	2
	N	9	9	9	9
Total	Media	2,5	4,8	2,5	0,3
	Desv. típ.	4,7	12,9	4,7	0,8
	Mediana	1	0,6	1	0
	Mínimo	0	0	0	0
	Máximo	25	84,76	25	3
	N	46	46	46	46

Atendiendo tanto a los valores de la Media y de la Mediana de la frecuencia de aparición del término y sobre todo del Peso, podemos observar que en MSN sí que tienen importancia en la ordenación, destacando en los diez primeros recursos los valores de la Media que son 5,7 para la frecuencia, y 13,6 para el Peso; los valores para los 10 recursos siguientes descienden a la mitad, 2,6 y más aún el Peso, que adquiere el valor 3. También hay una relación en la aparición de los términos en los hiperenlaces, pero en menor medida que en lo observado en Google.

Finalmente, si comparamos los datos de esta tabla con los de Google, podemos observar que los valores de las medias de las variables que ofrece Google son superiores a los de MSN.

Teoma (Ask).

Tabla 6.2-3. Teoma (Ask). Frecuencia y peso del término de búsqueda.

Búsqueda 1					
		Frecuencia de aparición del término	Peso	En el texto	En hiperenlace
Resultados 1 a 10	Media	11	12,5	9,1	0,4
	Desv. típ.	23,6	17,7	22,3	1
	Mediana	1	5,7	1	0
	Mínimo	0	0	0	0
	Máximo	73	55,8	72	3
	N	9	10	10	10
Resultados 11 a 20	Media	5	14,5	2,9	0,8
	Desv. típ.	2,8	11,2	2,6	0,9
	Mediana	4,5	14,7	2	0,5
	Mínimo	1	1,5	0	0
	Máximo	9	40,3	7	2
	N	10	10	10	10
Resultados 21 a 30	Media	4,3	2069,9	2,6	0,5
	Desv. típ.	4,4	6516,4	3,0	1
	Mediana	2	7,5	1,5	0
	Mínimo	1	0,5	0	0
	Máximo	13	20616	9	3
	N	10	10	10	10
Resultados 31 a 40	Media	1,2	0,6	0,1	0
	Desv. típ.	0,6	0,4	0,3	0
	Mediana	1	0,5	0	0
	Mínimo	1	0,5	0	0
	Máximo	3	1,8	1	0
	N	10	10	10	10
Resultados 41 a 50	Media	3,6	7,6	2,7	0,3
	Desv. típ.	4,8	13,5	4,8	0,7
	Mediana	1	0,5	0	0
	Mínimo	0	0	0	0
	Máximo	14	40,3	14	2
	N	10	10	10	10
Total	Media	4,9	4,2	3,5	0,4
	Desv. típ.	10,6	2914,3	10,4	0,8
	Mediana	1	1,7	1	0
	Mínimo	0	0	0	0
	Máximo	73	20616	72	3
	N	49	50	50	50

En Teoma (Ask), atendiendo a los valores de la Mediana, no se observa una clara relación entre el ranking y estos valores, ya que del once al veinte, las cifras son superiores a las de los diez primeros.

La Media de las frecuencias tiene una ordenación descendente en los cuarenta primeros resultados, subiendo en el último grupo (del 41 al 50). Sin embargo, esta relación no se aprecia en la valoración del peso, ya que por ejemplo, los resultados del once al veinte, obtienen mayor valor que los situados entre los diez primeros. Estos resultados hacen aconsejable, como ocurre con otros buscadores, la consulta de más de veinte resultados, ya que los valores tanto de la Frecuencia como del Peso, o se mantienen, o incluso son superiores en los últimos recursos analizados.

También se puede apreciar, al igual que en otros casos, que la frecuencia de aparición de los términos en las páginas que se analizan, ofrece resultados variables, ya que aparecen del 31 al 40, resultados con valores menores que los apreciados para el grupo del 41 al 50. Por tanto, si que hay una relación hasta los treinta primeros resultados, alterándose a partir de ellos.

Las medias de las frecuencias son superiores a las de MSN a excepción de la que representa a los resultados del 31 al 40.

WiseNut

Tabla 6.2-4. WiseNut. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
		Frecuencia de aparición del término	Peso	En el texto	En hiperenlace
Resultados	Media	3	6,5	2,1	0,6
1 a 10	Desv. típ.	2,2	7,3	1,8	0,8
	Mediana	2,5	4,7	1	0
	Mínimo	1	0,6	0	0
	Máximo	7	24	5	2
	N	10	10	10	10
Resultados	Media	1,3	2,4	1,2	0,2
11 a 20	Desv. típ.	1,1	3,5	0,8	0,4
	Mediana	1	1,6	1	0
	Mínimo	0	0	0	0
	Máximo	4	11,8	3	1
	N	10	10	10	10
Resultados	Media	4,9	11939,9	4,6	0,8
21 a 30	Desv. típ.	6,8	28904,8	7,0	1,1
	Mediana	1	2,9	1	0
	Mínimo	0	0	0	0
	Máximo	21	86854	21	3
	N	9	9	9	9

Resultados	Media	1,3	3,0	1,1	0,3
31 a 40	Desv. típ.	1,5	5,1	1,2	0,5
	Mediana	1	1,6	1	0
	Mínimo	0	0	0	0
	Máximo	5	16,9	4	1
	N	9	10	10	10
Resultados	Media	1	2,2	0,7	0,1
41 a 50	Desv. típ.	0,5	3,0	0,7	0,3
	Mediana	1	0,6	1	0
	Mínimo	0	0	0	0
	Máximo	2	8,8	2	1
	N	8	10	10	10
Total	Media	2,3	2195,9	1,9	0,4
	Desv. típ.	3,5	12690,8	3,3	0,7
	Mediana	1	2,2	1	0
	Mínimo	0	0	0	0
	Máximo	21	86854	21	3
	N	46	49	49	49

En WiseNut no se aprecia una relación entre los valores de la Frecuencia de aparición del término y Peso para la ordenación, ya que los valores de los diferentes grupos son irregulares, pues la relación sólo se observa en los veinte primeros resultados, y a partir de estos, los valores de la Mediana de la frecuencia se mantienen mientras que los del peso fluctúan de forma irregular. Este buscador presenta los valores más bajos que el resto de buscadores.

Yahoo

Tabla 6.2-5. Yahoo. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
		Frecuencia de aparición del término	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	20	29,6	19	0,7
1 a 10	Desv. típ.	22,8	20,5	23,0	0,9
	Mediana	12	20,2	11	0
	Mínimo	1	7	1	0
	Máximo	73	65,3	72	2
	N	9	9	9	9
Resultados	Media	13,4	10770,0	12,9	1,4
11 a 20	Desv. típ.	13,0	27505,0	13,4	2,8
	Mediana	8	30,3	7,5	0
	Mínimo	1	1,8	0	0
	Máximo	36	86854	36	9
	N	10	10	10	10

Resultados	Media	15,8	13,2	5,8	0,9
21 a 30	Desv. típ.	24,7	13,1	8,3	1,3
	Mediana	5	10,5	3	0
	Mínimo	1	0	0	0
	Máximo	73	40,5	28	3
	N	8	10	10	10
Resultados	Media	7,6	713,4	5,8	0,3
31 a 40	Desv. típ.	6,8	2097,3	7,3	0,7
	Mediana	5,5	9,8	3	0
	Mínimo	2	0	0	0
	Máximo	23	6306	23	2
	N	8	9	9	9
Resultados	Media	3,3	503,2	2,9	0
41 a 50	Desv. típ.	2,7	1549,2	2,9	0
	Mediana	2,5	9,3	2	0
	Mínimo	1	0,2	0	0
	Máximo	10	4912	10	0
	N	10	10	10	10
Total	Media	11,9	2490,6	9,1	0,7
	Desv. típ.	16,5	12828,5	13,5	1,5
	Mediana	4	12,9	3	0
	Mínimo	1	0	0	0
	Máximo	73	86854	72	9
	N	45	48	48	48

Yahoo sí que tiene en cuenta la Frecuencia y el Peso de los términos a la hora de ordenar sus resultados ya que generalmente aparecen valores elevados, si los comparamos con los que ofrecen otros buscadores y metabuscadores.

Atendiendo a los valores de la Mediana de la Frecuencia, podemos observar una clara relación que sólo se interrumpe mínimamente en los resultados comprendidos entre el treinta y uno y el cuarenta, con una Media de este valor, algo superior a la de los diez que le preceden.

Respecto al peso, no existen una relación tan clara ya que por ejemplo, es superior el valor de la Mediana de los resultados del segundo grupo (del once al veinte) que el del primero (uno al diez).

Finalmente hay que señalar que corresponden a Yahoo los valores más altos de la Media y de la Frecuencia de aparición del término en los documentos.

Metabuscadores

Dogpile

Tabla 6.2-6. Dogpile. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
		Frecuencia de aparición del término	Peso	En el texto	En hiperenlace
Resultados	Media	16,6	23,4	16,3	0,5
1 a 10	Desv. típ.	24,3	27,9	24,2	0,8
	Mediana	2,5	8,8	2,5	0
	Mínimo	1	0,2	0	0
	Máximo	73	65,3	72	2
	N	10	10	10	10
Resultados	Media	13,6	16	6	1
11 a 20	Desv. típ.	22,3	16,2	7,9	1,2
	Mediana	3,5	14,2	2,5	0,5
	Mínimo	0	0	0	0
	Máximo	73	44	25	3
	N	10	10	10	10
Resultados	Media	5,7	7,2	5	0,7
21 a 30	Desv. típ.	8,0	7,6	8,2	0,9
	Mediana	1	7	1	0
	Mínimo	0	0	0	0
	Máximo	25	20,2	25	2
	N	9	9	9	9
Resultados	Media	3,9	11,1	3,8	0,4
31 a 40	Desv. típ.	6,2	26,3	6,3	0,7
	Mediana	1	1,6	1	0
	Mínimo	0	0	0	0
	Máximo	20	84,8	20	2
	N	10	10	10	10
Resultados	Media	1,8	2,4	1,8	0,2
41 a 50	Desv. típ.	2,4	2,5	2,4	0,4
	Mediana	1	2	1	0
	Mínimo	0	0	0	0
	Máximo	8	7,2	8	1
	N	9	8	9	9
Total	Media	8,5	12,5	6,7	0,6
	Desv. típ.	16,2	20,1	13,1	0,9
	Mediana	1	2,7	1	0
	Mínimo	0	0	0	0
	Máximo	73	84,8	72	3
	N	48	47	48	48

Dogpile no tiene en cuenta, de forma determinante en la ordenación, la Frecuencia de aparición del término en los recursos ni el Peso, pues como podemos observar en los resultados del once al veinte, los valores son superiores a los de los colocados en las diez primeras posiciones.

Llama la atención los altos valores de las Medias de la Frecuencia y del Peso, que son superiores a la mayoría de los motores.

Excite

Tabla 6.2-7. Excite. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Excite		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperen-lace**
Resultados	Media	20	27,3	19,5	0,9
1 a 10	Desv. típ.	23,2	25,7	23,1	1,1
	Mediana	11,5	21,4	9,5	0,5
	Mínimo	0	0	0	0
	Máximo	73	65,3	72	3
	N	10	10	10	10
Resultados	Media	11,9	11,9	3,1	0,3
11 a 20	Desv. típ.	23,4	17	4,1	0,5
	Mediana	3	1,9	1	0
	Mínimo	1	0	0	0
	Máximo	73	44	13	1
	N	9	10	10	10
Resultados	Media	6,1	19,1	5,4	0,4
21 a 30	Desv. típ.	6,1	25,7	6	0,9
	Mediana	5	9,1	5	0
	Mínimo	0	0	0	0
	Máximo	20	84,8	20	2
	N	9	9	9	9
Resultados	Media	6,8	10207,8	6,4	0
31 a 40	Desv. típ.	6,2	28787,9	6,3	0
	Mediana	4	18,3	4	0
	Mínimo	1	0,7	1	0
	Máximo	21	86854	21	0
	N	9	9	9	9
Resultados	Media	9	14,5	8,9	1,1
41 a 50	Desv. típ.	8,3	23,4	8,4	2,2
	Mediana	8	7,4	8	0
	Mínimo	0	0	0	0
	Máximo	27	78,7	27	7
	N	10	10	10	10
Total	Media	10,9	1928,7	8,8	0,6
	Desv. típ.	15,9	12538,6	12,9	1,2
	Mediana	5	8,1	5	0
	Mínimo	0	0	0	0
	Máximo	73	86854	72	7
	N	47	48	48	48

En Excite podemos observar que los valores de la Frecuencia y el Peso del término de búsqueda son superiores a los de Dogpile. Tampoco en este metabuscador decrecen de forma homogénea a medida que descendemos en el listado sino que, como veni-

mos observando en la mayoría de los buscadores, en algunos casos adquieren valores superiores a los que les preceden, como en los resultados del veintiuno al treinta y del cuarenta y uno al cincuenta.

Por otro lado, los valores de las variables son, a excepción de Google, superiores a los de los motores de búsqueda, lo que hace recomendable, también en esta ocasión, la consulta de un número de resultados más allá de los veinte primeros, ya que, como hemos señalado, en grupos de resultados posteriores, se registran valores en la Frecuencia de aparición del término y del Peso, superiores a los de los primeros grupos.

Ixquick

Tabla 6.2-8. Ixquick. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Ixquick		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	12,6	14,4	12	0,3
1 a 10	Desv. típ.	23,8	24,5	23,8	0,7
	Mediana	2,5	2,2	1	0
	Mínimo	1	0,6	0	0
	Máximo	73	65,3	72	2
	N	10	10	10	10
Resultados	Media	6,3	14,5	5,2	0,4
11 a 20	Desv. típ.	10,7	20,2	11,0	0,5
	Mediana	2,5	7,8	1	0
	Mínimo	1	0,1	0	0
	Máximo	36	65,3	36	1
	N	10	10	10	10
Resultados	Media	16,3	11,6	5,6	0,25
21 a 30	Desv. típ.	26,4	15,1	8,2	0,5
	Mediana	4	2,5	2,5	0
	Mínimo	1	0	0	0
	Máximo	73	40,5	25	1
	N	7	8	8	8
Resultados	Media	2	6,3	1,5	1
31 a 40	Desv. típ.	0	1,9	0,7	1,4
	Mediana	2	6,3	1,5	1
	Mínimo	2	4,9	1	0
	Máximo	2	7,6	2	2
	N	2	2	2	2
Total	Media	10,6	13,1	7,3	0,4
	Desv. típ.	19,7	19,3	15,5	0,6
	Mediana	2	2,7	1	0
	Mínimo	1	0	0	0
	Máximo	73	65,3	72	2
	N	29	30	30	30

En Ixquick no interviene, de forma determinante en la ordenación, ni la Frecuencia de aparición del término de búsqueda, ni el Peso. Los únicos valores sobre los que

parece haber alguna relación son el Peso y la valoración de la aparición de los términos de búsqueda en el texto, tal como podemos ver en los valores medios.

En comparación a los metabuscadores anteriores, los valores de las variables son más bajos.

Profusion

Tabla 6.2-9. Profusion. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Profusion		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	12,7	17,7	11,7	1
1 a 10	Desv. típ.	21,8	17,5	21,7	1,2
	Mediana	7	14,8	4,5	0,5
	Mínimo	1	0,21	0	0
	Máximo	73	55,8	72	3
	N	10	10	10	10
Resultados	Media	13,9	20,2	12,8	1,4
11 a 20	Desv. típ.	22	17,2	22,1	1,2
	Mediana	4,5	14,2	3,5	1,5
	Mínimo	0	0	0	0
	Máximo	73	55,8	72	3
	N	10	10	10	10
Resultados	Media	8,8	2076	7,1	0,3
21 a 30	Desv. típ.	11,2	6514,3	11,0	0,9
	Mediana	4	7,5	3	0
	Mínimo	0	0	0	0
	Máximo	36	20616	36	3
	N	9	10	10	10
Resultados	Media	8,1	2079,9	7,4	1,5
31 a 40	Desv. típ.	10,3	6513	10,4	2,8
	Mediana	4,5	13,2	3	0
	Mínimo	1	0,6	1	0
	Máximo	32	20616	32	9
	N	10	10	10	10
Resultados	Media	9	15,9	9	2
41 a 50	Desv. típ.	.	.	.	.
	Mediana	9	15,9	9	2
	Mínimo	9	15,9	9	2
	Máximo	9	15,9	9	2
	N	1	1	1	1
Total	Media	10,9	1023,3	9,7	1,1
	Desv. típ.	16,7	44,9	16,6	1,7
	Mediana	4,5	12,9	3	0
	Mínimo	0	0	0	0
	Máximo	73	20616	72	9
	N	40	41	41	41

En Profusion se aprecia una relación similar a la observada para Google, manteniendo un orden jerárquico en los valores la Mediana de la Frecuencia y el Peso para los veinte primeros resultados, volviendo a aumentar los valores en los resultados siguientes, lo que nos puede estar indicando que también aquí, el algoritmo de ranking se aplica para un determinado número de recursos, que pueden ser los veinte o treinta primeros, y que a partir de éstos, los cálculos se aplican a otro grupo que se coloca a continuación y cuyos valores suelen ser superiores a los registrados en el grupo anterior. La confirmación de esta observación requiere la realización de más estudios de evaluación que nos ayuden a extraer conclusiones más determinantes en este sentido.

Search

Tabla 6.2-10. Search. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Search		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	13,5	24,2	12,9	0,7
1 a 10	Desv. típ.	14,1	25,4	14,4	0,8
	Mediana	8,5	13,6	8	0,5
	Mínimo	1	1,5	0	0
	Máximo	36	65,3	36	2
	N	10	10	10	10
Resultados	Media	13,4	20,6	12,7	0,6
11 a 20	Desv. típ.	22,7	28,3	22,7	0,8
	Mediana	2,5	8,8	1	0
	Mínimo	1	0,2	1	0
	Máximo	73	84,8	72	2
	N	10	10	10	10
Resultados	Media	2,9	7	2,4	0,8
21 a 30	Desv. típ.	2,7	7,1	2,7	1,1
	Mediana	3	2,4	2	0
	Mínimo	0	0	0	0
	Máximo	9	16,4	9	3
	N	9	9	9	9
Resultados	Media	6,4	2069,7	5,7	1,3
31 a 40	Desv. típ.	9,6	6516,5	9,7	2,9
	Mediana	1,5	2,3	1	0
	Mínimo	1	0,3	0	0
	Máximo	32	20616	32	9
	N	10	10	10	10
Resultados	Media	12,6	8,3	3,9	0,6
41 a 50	Desv. típ.	24,6	13,2	3,6	0,9
	Mediana	4	1,2	2	0
	Mínimo	1	0	0	0
	Máximo	73	40,5	9	2
	N	8	9	9	9
Total	Media	9,8	443,4	7,7	0,8
	Desv. típ.	16,4	2973,7	13,4	1,5
	Mediana	3	6,6	1,5	0
	Mínimo	0	0	0	0
	Máximo	73	20616	72	9
	N	47	48	48	48

En Search se observa también entre los diez primeros resultados los mayores valores de la Mediana tanto de la Frecuencia como del Peso, descendiendo su valor en los resultados siguientes. Este descenso se produce de forma brusca en los valores de la Frecuencia, mientras que los del Peso, van descendiendo paulatinamente.

Surfwax

Tabla 6.2-11. Surfwax. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Surfwax		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	2,8	6,0	1,6	0,4
1 a 10	Desv. típ.	2,0	7,5	1,6	0,7
	Mediana	2,5	2,2	1	0
	Mínimo	1	0,6	0	0
	Máximo	7	24	5	2
	N	10	10	10	10
Resultados	Media	1	1,7	1	0,2
11 a 20	Desv. típ.	0	1,1	0	0,4
	Mediana	1	2,3	1	0
	Mínimo	1	0,3	1	0
	Máximo	1	2,7	1	1
	N	5	5	5	5
Total	Media	2,2	4,5	1,4	0,3
	Desv. típ.	1,9	6,4	1,3	0,6
	Mediana	1	2,3	1	0
	Mínimo	1	0,3	0	0
	Máximo	7	24	5	2
	N	15	15	15	15

Dado que Surfwax, en esta búsqueda, sólo ofrece quince resultados, no se puede apreciar si hay una relación en la ordenación teniendo en cuenta los aspectos que se valoran, ya que en el segundo grupo de ordenación sólo se valoraron cinco resultados, frente a los diez del primer grupo. No obstante, si que parecen influir estos aspectos en la ordenación de los resultados.

Vivisimo

Tabla 6.2-12. Vivisimo. Frecuencia y peso del término de búsqueda.

Búsqueda 1					
Vivisimo		**Frecuencia de aparición del término**	**Peso**	**En el texto**	**En hiperenlace**
Resultados	Media	4,8	7,8	4,5	1,2
1 a 10	Desv. típ.	7,8	8,4	7,6	1,2
	Mediana	1	4,7	1	1
	Mínimo	0	0	0	0
	Máximo	25	22,6	25	3
	N	10	10	10	10
Resultados	Media	13,2	20,7	11,1	0,2
11 a 20	Desv. típ.	23,3	28,4	22,3	0,6
	Mediana	5	11,8	1,5	0
	Mínimo	0	0	0	0
	Máximo	73	84,8	72	2
	N	9	10	10	10
Resultados	Media	3,3	8,0	2,6	0,5
21 a 30	Desv. típ.	2,9	12,7	2,3	0,8
	Mediana	2,5	2,0	2	0
	Mínimo	1	0,2	0	0
	Máximo	9	40,3	7	2
	N	10	10	10	10
Resultados	Media	4,7	10,4	2,9	0,3
31 a 40	Desv. típ.	4,3	10,1	2,5	0,5
	Mediana	4	8,6	2	0
	Mínimo	0	0	0	0
	Máximo	13	31,2	8	1
	N	9	9	9	9
Resultados	Media	2,3	4,8	1,8	0,2
41 a 50	Desv. típ.	3,0	6,6	2,8	0,4
	Mediana	1	0,5	1	0
	Mínimo	1	0,2	0	0
	Máximo	10	18,3	9	1
	N	9	9	9	9
Total	Media	5,6	10,4	4,7	0,5
	Desv. típ.	11,3	16,0	11,0	0,9
	Mediana	2	2,6	1,5	0
	Mínimo	0	0	0	0
	Máximo	73	84,8	72	3
	N	47	48	48	48

En Vivisimo no hay una relación entre los valores que se analizan y la ordenación de los resultados de la búsqueda. Da la impresión que utiliza otros indicadores como la aparición de los términos en los hiperenlaces, ya que todos los grupos de páginas aparecen con valor, frente a otros buscadores que en gran medida aparecen con un 0 en este indicador.

Llaman la atención los bajos valores tanto de la Mediana de la Frecuencia como del Peso en los diez primeros resultados, aumentando del once al veinte y descendiendo el valor en el resto. Esto demuestra que en este metabuscador intervienen otras variables con mayor repercusión en la ordenación.

En resumen, hemos observado como Google tiene en cuenta, junto a otros valores, la Frecuencia y el Peso en los treinta primeros resultados, y que ya que los que les siguen ofrecen valores superiores a éstos, es aconsejable la consulta más allá de los veinte primeros.

MSN basa más la ordenación en estos valores, pues se mantienen de forma descendente a lo largo de los resultados analizados. Yahoo valora de forma especial las frecuencias de los términos y selecciona en los primeros lugares recursos en los que el término de búsqueda aparece con mayor frecuencia, lo que nos permite destacar el funcionamiento de este motor de búsqueda, al ser el que más en cuenta tiene para la ordenación los valores de frecuencia y peso de los términos.

Sin embargo Teoma no tiene en cuenta el Peso de los términos de búsqueda en la ordenación y WiseNut presenta los valores de la Frecuencia y el Peso más bajos que el resto de buscadores.

Respecto a los metabuscadores, Excite coloca en los primeros puestos documentos en los que la Frecuencia de aparición de los términos es alta, pero también lo es en los del final, siendo recomendable, como en el caso de Google, la consulta de recursos posteriores a los veinte primeros.

Surfwax es el metabuscador que recupera recursos con bajos índices en cuanto a las frecuencias de aparición de los términos de búsqueda y Peso, y Vivisimo tiene un comportamiento irregular en la utilización de estos valores para la ordenación de los recursos, lo cuál puede ser indicativo de un incorrecto funcionamiento o en la intervención de otras variables para la elaboración del ranking.

Por tanto, no se aprecia una uniformidad en la valoración de estos aspectos a la hora de organizar los resultados, o por lo menos no más allá de los veinte primeros, haciendo necesaria la consulta de un mayor número de recursos en los que valores de la Frecuencia y el Peso de los términos de búsqueda vuelven a alcanzar mayores valores.

6.3. Correlación entre la frecuencia de aparición del término de búsqueda y el peso con la ordenación de los resultados de búsqueda

La siguiente tabla muestra los resultados correspondientes al cálculo del Coeficiente de correlación de Pearson, que nos va a permitir valorar la existencia o no de relación entre la posición en los listados que ocupan los recursos recuperados, por un lado de la Frecuencia de aparición del término de búsqueda, y por otro del Peso del término de búsqueda. Como en los casos anteriores, nos basamos en los datos de la búsqueda por un término.

Tabla 6.3-1. Buscadores. Correlación de la frecuencia y el peso con la ordenación

Búsqueda 1			
Buscador	**Estadístico**	**Frecuencia de aparición del término**	**Peso**
Google	Correlación de Pearson	-0,237	-0,094
	Sig. (bilateral)	0,113	0,534
MSN	Correlación de Pearson	-0,307	-0,271
	Sig. (bilateral)	0,038	0,068*
Teoma (Ask Jeeves)	Correlación de Pearson	-0,206	-0,016
	Sig. (bilateral)	0,156	0,913
WiseNut	Correlación de Pearson	-0,106	0,040
	Sig. (bilateral)	0,485	0,785
Yahoo	Correlación de Pearson	-0,340	-0,091
	Sig. (bilateral)	0,022	0,539

A partir de estos datos, podemos observar que la relación de los valores que corresponden tanto a la Frecuencia como al Peso, generalmente son negativos, lo que indica una relación lineal inversa, que se traduce en el hecho de que los motores de búsqueda tienden a colocar en los primeros lugares de la ordenación los recursos en los que estos valores son mayores. Por el contrario, los valores positivos indican que estos motores colocan en primer lugar recursos en los que estos valores son menores, lo que indica un funcionamiento defectuoso de la ordenación.

Por otro lado, la mayoría de los valores de este índice no se acercan al valor -1, que indicaría una relación lineal fuerte, siendo más bien próximos a 0, lo que nos indica que la relación entre la Frecuencia de aparición del término en los recursos y su ordenación no es muy fuerte.

Un análisis más detallado de estos valores muestra que la mayor relación se da en Yahoo (-0,340), seguido de MSN (-0,307) y Google (-0,237) siendo Teoma y WiseNut los motores en los que el valor de la frecuencia tiene menos repercusión en la ordenación.

Respecto al Peso, tampoco se puede apreciar una fuerte relación en la ordenación. Sin embargo, corresponde a MSN el mayor valor de este coeficiente (-0,271), lo que unido a los valores señalados en la variable anterior nos permiten afirmar que MSN es el buscador que más tiene en cuenta ambos valores para ordenar sus resultados. También aquí, Teoma y WiseNut registran los valores más bajos.

En Yahoo hay una mayor relación con la Frecuencia de aparición del término, mientras que en Google no hay una clara relación entre estas variables y la ordenación.

Tabla 6.3-2. Metabuscadores. Correlación de la frecuencia y el peso con la ordenación

Búsqueda 1			
Metabuscador	**Estadístico**	**Frecuencia de aparición del término**	**Peso**
Dogpile	Correlación de Pearson	-0,347	-0,343
	Sig. (bilateral)	0,016	0,018
Excite	Correlación de Pearson	-0,229	0,072
	Sig. (bilateral)	0,122	0,625
Ixquick	Correlación de Pearson	-0,004	-0,086
	Sig. (bilateral)	0,985	0,650
Profusion	Correlación de Pearson	-0,195	0,115
	Sig. (bilateral)	0,229	0,476
Search	Correlación de Pearson	-0,094	0,099
	Sig. (bilateral)	0,532	0,503
Surfwax	Correlación de Pearson	-0,533	-0,420
	Sig. (bilateral)	0,041	0,119
Vivisimo	Correlación de Pearson	-0,176	-0,128
	Sig. (bilateral)	0,235	0,385

En los metabuscadores, los valores de Correlación entre la Frecuencia de aparición del término y la Ordenación no difieren mucho de los registrados por los motores de búsqueda. Los valores más significativos los podemos encontrar en Surfwax (-0,533) y Dogpile (-0,347), lo que indica una mayor correlación en estos metabuscadores. Sin embargo, Ixquick registra un valor del Índice de correlación muy próximo a 0, lo que indica la inexistencia de relación entre la Frecuencia y la Ordenación.

En lo que respecta al Peso, llama la atención Excite cuyo Índice de correlación es positivo, lo que indica que este metabuscador no tiene en cuenta el Peso en la ordenación de resultados, lo que puede influir en una peor recuperación al colocar en los primeros lugares recursos de menor Peso. El máximo valor del Índice de correlación corresponde a Surfwax (-0,420), lo que indica que este metabuscador tiene en cuenta para la ordenación el valor del Peso del término de búsqueda.

En resumen, podemos afirmar que las herramientas de búsqueda ordenan los resultados siguiendo procedimientos y cálculos distintos en los que se tienen en cuenta, en unos casos la metainformación como en Teoma, WiseNut y Yahoo; o las frecuencias de aparición de los términos de búsqueda y el Peso para un determinado número de recursos que pueden ser los diez, veinte o treinta primeros recursos como hemos observado en Google y otros buscadores, o de una forma más regular, en Yahoo y MSN.

Por otro lado, respecto a la correlación entre la Frecuencia y el Peso y la ordenación de los recursos, Yahoo, MSN y Google, por este orden, son los que mantienen unos valores que indican una mayor relación entre las variables.

Tampoco en los metabuscadores hay unanimidad en el funcionamiento ya que a pesar de que coinciden en la no utilización de la metainformación para la ordenación, unos se basan en la Frecuencia de aparición del término y el Peso, aunque en el caso de Excite, se observa sólo en los diez primeros recursos, mientras que Profusión tiene un comportamiento similar a Google, y Vivisimo ofrece unos valores desiguales a lo largo de los resultados analizados. Finalmente Surfwax no los tiene en cuenta.

Sobre la relación que juegan la Frecuencia de aparición y el Peso del término de búsqueda en la ordenación, los resultados también son dispares, destacando Surfwax como el metabuscador en el que mayor relación existe entre las variables.

El análisis del ranking en función de estos parámetros nos ha permitido constatar, además de valorar que aspectos tienen más en cuenta unos buscadores y otros, la importancia de utilizar y revisar más allá de los veinte primeros recursos tanto para la evaluación de estas herramientas como de cara al usuario, ya que en estos recursos aparecen valores de Frecuencia y Peso que denotan la existencia de cierto interés para su consulta.

V. CONCLUSIONES

Conscientes de las limitaciones que este tipo de evaluaciones basadas en pequeñas muestras llevan consigo y que los resultados que ofrecen estas herramientas de búsqueda son muy variables, y pueden depender de multitud de factores, unas veces propios de los sistemas y otras externos a ellas, podemos extraer, en relación con los diferentes aspectos analizados, las siguientes conclusiones:

1. Capacidad de búsqueda

Los buscadores que recuperaron en todas las búsquedas, alcanzando al menos la cifra de cincuenta recursos, han sido Google, MSN y Yahoo, siendo más limitado el funcionamiento de Teoma, que no recuperó en la búsqueda booleana, ni en las búsquedas por frase y por campo. WiseNut tampoco funcionó en la búsqueda booleana y sólo ofreció treinta resultados en la búsqueda por campo.

Respecto a los metabuscadores, hemos de mencionar a Vivisimo, Search y Excite entre los que ofrecen una recuperación más completa en todas las consultas, aunque éste último sólo recuperó treinta y cuatro recursos en la búsqueda booleana. Profusion no recuperó en la búsqueda con operadores de existencia. Dogpile y Surfwax, que no funcionaron en las búsquedas booleana, por frase y por campo, demostraron que poseen unos mecanismos de recuperación más limitados que el resto. Éste último además, ofrece en las demás consultas un reducido número de resultados.

El comportamiento de estas herramientas es diferente en cada tipo de búsqueda, lo que dificulta la posibilidad de establecer unos principios claros e inamovibles sobre su funcionamiento. Las búsquedas que menos problemas plantean tanto a buscadores como a metabuscadores son la búsqueda por un término y la que utiliza el lenguaje natural, si bien WiseNut planteó problemas al no soportar más de siete términos de búsqueda. La búsqueda con el limitador (+) plantea problemas tanto para WiseNut, que requiere utilizar la búsqueda con todos los términos, como para Profusion. La búsqueda boo-

leana también plantea problemas para WiseNut por el número de términos utilizado.

Por tanto, Teoma por un lado, y los metabuscadores Dogpile y Surfwax, han de mejorar sus opciones de búsqueda de forma que permitan recuperar en búsquedas de tipo booleano, por frase y por campo, siendo más aconsejable el uso de Excite, Ixquick, Search y Vivisimo en este tipo de búsquedas.

2. Presentación de los resultados de las búsquedas

- *Uso de la Metainformación en el título del recurso*

En el análisis del uso que estas herramientas hacen de la etiqueta TITLE para utilizarla en la descripción que los buscadores muestran en la página de resultados como elemento inicial y destacado de cada recurso, hemos observado que Yahoo es el motor de búsqueda que más la utiliza, seguido de MSN y Google, y en menor medida la adoptan WiseNut y Teoma.

Entre los metabuscadores, es Search el que hace uso de ella en mayor medida, seguido de Vivisimo, Excite e Ixquick. Dogpile, Profusion y Surfwax la utilizan con menos frecuencia.

Aunque cada vez es menos frecuente en los listados la aparición de recursos con el título <unknow> (desconocido), si que hemos apreciado la utilización de otros términos o frases que nada tienen que ver con el contenido del documento por lo que se ha de recomendar a los creadores de páginas web una mayor atención y evitar la existencia de páginas sin nombre, de forma que reflejen, en la medida de lo posible, su contenido, ya que es fundamental para decidir el posible interés de un recurso. Respecto a los desarrolladores de las herramientas de búsqueda, deberían dar más importancia en la indización a la información alojada en estas etiquetas ya que, si se ha hecho de forma correcta, expresan de forma clara las características y el contenido del documento.

Por tanto, consideramos que la posibilidad de extracción de metainformación de los recursos web es algo que se ha de ir imponiendo, dado su especial interés, tanto para recuperar con mayor precisión documentos o recur-

sos, como para utilizarlos en la descripción. De aquí la importancia de la valoración del uso que hacen estas herramientas de la metainformación.

- *Términos de búsqueda destacados*

  En relación con la utilización de términos de búsqueda destacados en los registros de los listados de recuperación, que de forma rápida permite al usuario valorar el recurso en función del contexto en el que aparecen dichos términos, Yahoo, Teoma y MSN son los motores que con mayor frecuencia los destacan.

  Entre los metabuscadores es más frecuente el uso de esta técnica, sobre todo en el caso de Profusion. Excite Surfwax y Dogpile la utilizan de forma más limitada, y en muy pocos casos destacan los términos Vivisimo, Search e Ixquick.

  Hemos podido constatar el poco uso, por parte de los buscadores, de esta técnica, y que además se utiliza sin tener en cuenta las palabras importantes de la búsqueda, ya que a menudo destacan también palabras vacías. Por tanto, los desarrolladores de estas herramientas deberían procurar que se destaquen sólo los términos representativos y no cualquiera de ellos, ya que generalmente, para el usuario, no cumplen una función orientativa.

- *Recursos dependientes o relacionados con otros de un nivel jerárquico superior*

  Respecto a la aparición de forma destacada en los listados de los recursos dependientes de otros de mayor rango, Google es el buscador que distingue de forma sistemática los recursos pertenecientes a un mismo sitio web. Consideramos que esta técnica es interesante para el usuario puesto que le ayuda a decidir la necesidad de consultar un recurso que tiene relación con otro, ya que es muy posible que si uno es de su interés, posiblemente el otro también, y viceversa. WiseNut y Yahoo son los motores que no presentan recursos basados en esta técnica.

Entre los metabuscadores Excite, Search y Vivisimo son los que, de forma esporádica, destacan los resultados dependientes. El resto tampoco utiliza esta técnica.

- *Recursos publicitarios*

Finalmente, respecto a la aparición en la página de resultados, de recursos de carácter publicitario, su presencia no es elevada. El buscador que con mayor frecuencia los presenta es MSN, y entre los metabuscadores, Ixquick. Yahoo no ofrece ningún recurso de este tipo. Es de destacar que las herramientas que los ofrecen, lo hacen presentándolos al margen del resto de los recursos recuperados, lo que facilita que el usuario pueda centrarse en acceder a los recursos que no tienen carácter comercial, y acceder a ellos cuando las necesidades de información lo requieran.

En resumen, la calidad de los registros debe mejorar si lo que se pretende es reflejar el contenido del recurso al que se refieren y servir de ayuda para su elección, por lo que han de facilitar unos títulos significativos y acordes con el contenido, destacar los términos representativos de la búsqueda en el contexto en que aparecen y finalmente, agrupar, y en la medida de los posible, señalar visualmente los recursos relacionados o dependientes de un determinado sitio Web.

3. Componentes de los buscadores

3.1 Robot ó Crawler

- *Profundidad de la indización*

Hemos observado en cuanto a la profundidad de indización de los robots o *crawlers* que Google seguido de Yahoo, son los motores que indizan recursos de mayor profundidad en los sitios web. Sin embargo entre ellos se diferencian en que en Google son más frecuentes los recursos de primer nivel, que corresponden a las páginas de inicio de los sitios web, mientras que Yahoo recupera más páginas correspondientes a directorios de niveles más bajos, lo que demuestra mayor profundidad en el trabajo de los crawlers o arañas en este buscador. MSN registra peores resultados ya que ofrece mayor

número de páginas genéricas, tanto de primero como de tercer nivel, y WiseNut, a pesar de sus problemas en la recuperación, ofrece fundamentalmente recursos de segundo, tercer y cuarto nivel, aunque, como Yahoo, también ofrece recursos de mayor nivel. Por su parte, Teoma, en comparación con el resto de los motores, contabiliza valores inferiores en todos los niveles.

En los metabuscadores hay que señalar a Search que, en comparación con los demás, se caracteriza por recuperar menos recursos del nivel más superficial y más de los que requieren un rastreo en mayor profundidad. Excite tiene un comportamiento similar ya que, si bien recupera menos recursos de segundo nivel, supera a aquél en recursos de tercero, cuarto y quinto nivel, lo que hace aconsejable su uso, si se requiere una herramienta que revise en profundidad los recursos informativos que ofrecen los sitios web. Vivisimo muestra un comportamiento similar a los anteriores en cuanto a los recursos de mayor profundidad, pero mantiene los más altos índices de recuperación de recursos de menor profundidad.

3.2 Índices

- *Duplicados*

La recuperación de este tipo de recursos, tradicionalmente se viene considerando como uno de los aspectos a tener en cuenta para valorar el correcto funcionamiento de las herramientas de búsqueda, ya que es un elemento de juicio fundamental, que indica la capacidad que estas herramientas tienen para distinguir y eliminar de sus bases de datos los recursos repetidos. Esto supone no sólo una mayor credibilidad respecto a la cobertura que ofrecen las herramientas de búsqueda, sino que para el usuario es más cómodo y rápido el no acceder a recursos previamente recuperados.

En este sentido, son MSN y Yahoo los buscadores que recuperan un menor número de recursos duplicados frente a Google, que es el que más duplicados recuperó, si bien la mayoría se registraron en una misma búsqueda, lo que, en cualquier caso, es indicativo de la existencia de problemas de funcionamiento en este buscador.

En el caso de los metabuscadores, la detección y eliminación de duplicados es también indicativo de un mayor desarrollo de su tecnología puesto que, aunque en algunos casos indican textualmente en sus características, que son capaces de eliminar de sus listados los recursos duplicados, la evaluación demuestra que este aspecto no se cumple.

Ixquick es el metabuscador que menor número de duplicados presenta seguido de Vivisimo, Excite y Dogpile, todos ellos con bajas frecuencias, siendo Profusion, junto a Search los metabuscadores que más registros duplicados recuperan.

Aunque como hemos visto en los resultados el nivel de duplicados no es excesivamente alto, las herramientas de búsqueda han de tratar de eliminarlos totalmente y, en la medida de lo posible, así como aquellos recursos que sin tener la misma URL, su contenido aparece en más de una ocasión.

- *Recursos inactivos*

También en este caso, la recuperación de recursos inactivos, aparte de ser una pérdida de tiempo para el usuario que tiene que esperar hasta que el servidor mande un mensaje sobre la incidencia relacionada con el recurso, es un indicador que se tiene en cuenta para valorar la actualización de los índices, ya que se considera que una vez advertido que un recurso no es accesible, ha de eliminarse de la base de datos de forma inmediata e indizarse de nuevo cuando esté disponible. La mayor existencia de este tipo de enlaces en unas herramientas que en otras nos indica que la revisión de URL se realiza con menor periodicidad.

Aunque apenas hay diferencia entre ellos, Google es el que menos recursos inactivos recupera, seguido de MSN y Yahoo. WiseNut ofrece resultados similares pero hay que tener en cuenta que en la búsqueda booleana no ofreció resultados.

El metabuscador con mayor número de recursos inactivos es Dogpile, seguido de Ixquick, Search y Vivisimo. Los mejores resultados corresponden a Profusion.

3.3 La base de datos

En relación con el tercero de los componentes, la base de datos, nos hemos centrado en valorar las características de la información analizando qué buscador ofrece páginas de mayor actualidad, el carácter de la información recuperada, tipo de archivo y tipología documental más frecuente.

- *Actualidad de los recursos*

Tras analizar los recursos recuperados que contienen la fecha de realización o, en su defecto, la del copyright, podemos afirmar que las herramientas de búsqueda que proporcionan recursos de mayor actualidad son MSN y los metabuscadores Search y Excite.

- *Carácter de la información recuperada*

Éste es sin duda uno de los aspectos más positivos observados en la evaluación del funcionamiento de los buscadores de la Web, ya que las cifras de los recursos recuperados de interés para la investigación son los más frecuentes en todos ellos, destacando en primer lugar Google, seguido por Yahoo y MSN. Lo mismo ocurre entre los metabuscadores, entre los que destaca Excite, seguido por Search, Profusion, Ixquick y Vivisimo. Sin embargo, Surfwax apenas recupera recursos de este tipo.

En función de estos datos se puede afirmar que, fundamentalmente, las herramientas señaladas son útiles para recuperar información especializada, aunque entre los resultados también aparecen recursos de carácter comercial y publicitario, lo cual nos parece aceptable, ya que por ejemplo, aunque las búsquedas sean sobre temas especializados, en un determinado caso, al usuario puede interesarle el recurso que le ofrece una editorial u otro tipo de empresa, institución, etcétera, en relación con el tema de búsqueda. Lo deseable sería que el usuario pudiera decidir sobre la recuperación o no de este tipo de recursos, por lo que es necesario un mayor desarrollo en las opciones de búsqueda que presentan estas herramientas, es decir se ha de incidir en la posibilidad de establecer filtros que permitan reconducir las búsquedas

en función de los intereses del usuario. Para que esto funcione se requiere una descripción y clasificación normalizadas.

- *Tipo de archivo*

  El tipo de archivo que más recuperan las herramientas de búsqueda son las páginas en HTML, lo cuál es lógico en el ámbito de la Web. Los motores que más recursos en PDF y presentaciones de PowerPoint recuperan son Google y Yahoo. Entre los metabuscadores, destacan Excite y Search, y Vivisimo sólo en la búsqueda booleana.

- *Tipología documental*

  Hemos podido constatar en esta evaluación, que la tipología documental de los recursos recuperados por las herramientas de búsqueda varía en función del tipo de búsqueda. Así, en la búsqueda booleana se destaca el alto número de artículos recuperados por la mayoría de buscadores. En el resto de las búsquedas también aparecen de forma frecuente, lo cual indica una correcta recuperación. En este sentido destacan Google y Yahoo, seguidos por MSN y Teoma, quedando WiseNut en último lugar.

  Otro tipo documental que denota especialización del contenido es la información sobre proyectos de investigación en la que destaca Google, aunque con escasa diferencia respecto al resto de buscadores.

  MSN recupera, con mayor frecuencia que el resto, información sobre congresos cuyo tema específico está relacionado con los términos de búsqueda, y páginas que dan acceso a revistas electrónicas. Teoma es el motor que más páginas sobre bibliotecas digitales proporciona, facilitando el acceso a recursos electrónicos especializados.

  En la parte negativa debemos señalar que Teoma es el motor que más páginas en blanco recupera, seguido por Google y MSN. La recuperación de este tipo de páginas es indicativo del mal funcionamiento de estas herramientas, ya que se deberían poder detectar en estos casos la falta de contenido para evitar su recuperación.

La tipología documental que recuperan los metabuscadores es un fiel reflejo de lo observado en los buscadores. Corresponde la mayor recuperación de artículos a Excite seguido de Vivisimo, que destaca además en páginas sobre congresos y bibliotecas digitales. El tercer lugar lo ocupa Search que se caracteriza por facilitar el acceso a revistas electrónicas, presentaciones en PowerPoint, información sobre congresos, capítulos de monografías, artículos de enciclopedias y sobre todo por la información sobre proyectos de investigación, por lo que, en este sentido, puede resultar recomendable la utilización de este metabuscador.

4. Solapamiento

Respecto al solapamiento registrado por motores de búsqueda y metabuscadores en las seis búsquedas, el índice obtenido resulta elevado (21,6%), si bien es inferior al observado en trabajos anteriores.

El análisis de registros únicos recuperados por las herramientas de búsqueda demuestra, que los motores que recuperan un mayor número de recursos únicos, y que por tanto tienen menor solapamiento, son MSN y Yahoo, frente a Google que recupera un menor número de recursos únicos, y al que le corresponde el mayor solapamiento entre estos tres.

Estos datos coinciden con los aportados en el análisis del solapamiento entre buscadores que reúne los resultados de las seis búsquedas, así como en lo observado en el estudio por búsqueda individual. En ellos se demuestra que MSN, seguido de Yahoo y Google son los motores de búsqueda en los que se aprecia un menor solapamiento. Entre éstos, Google y Yahoo son los que más recursos iguales recuperan, siendo, en este sentido MSN, una herramienta de búsqueda recomendable para utilizar en combinación con cualquiera de los otros dos motores. También son complementarios Yahoo y WiseNut.

Search, por su alto solapamiento es el metabuscador menos indicado para utilizar en combinación con los motores de búsqueda y Vivisimo varía el solapamiento en función del tipo de búsqueda.

El alto solapamiento entre metabuscadores hace que no sea recomendable utilizarlos entre sí.

Así pues, el mayor solapamiento corresponde a Search seguido de Excite e Ixquick. El menor solapamiento lo registra Vivisimo.

El análisis del solapamiento por búsquedas también nos ha permitido observar que puede variar en función del tipo de búsqueda, aunque este aspecto se podría precisar mediante la realización de más estudios de evaluación.

5. Precisión técnica

En el análisis de la Precisión técnica, hemos de tener en cuenta que aunque no es el único indicador que se utiliza para valorar la precisión, sí que permite hacernos una idea aproximada de la calidad de los recursos que estas herramientas recuperan, y en este sentido podemos afirmar que los resultados analizados muestran, en todos los sistemas, una baja precisión técnica.

Al analizar la frecuencia de aparición de los términos de búsqueda en los recursos recuperados, llama la atención el alto porcentaje de páginas que no contienen los términos de búsqueda. Así, en la primera búsqueda, en la que se utiliza un único término, las herramientas en las que se observa una mayor precisión, esto es Yahoo, Google y el metabuscador Excite, sólo el 50% de recursos contiene el término de búsqueda.

En el resto de las búsquedas, hay que destacar igualmente que, a pesar del alto número de recursos que suelen recuperar los buscadores, tan sólo hemos observado un caso, recuperado por Search en la búsqueda que utiliza el lenguaje natural, en el que el recurso contiene todos los términos de búsqueda planteados. Aunque sabemos que la aparición de los términos en estas herramientas no es determinante para la recuperación, una presencia tan limitada de los términos de búsqueda es indicativa de una baja precisión.

De aquí que los datos obtenidos no permitan destacar un buscador, que de forma regular, a lo largo de las búsquedas analizadas, ofrezca buenos resultados, observándose cierta variedad en función del tipo de búsqueda. Teniendo en cuenta estos aspectos, podemos decir que Teoma seguido de Ya-

hoo, ofrecen los mejores resultados en la búsqueda por lenguaje natural, al ser los que ofrecen un menor porcentaje de recursos sin los términos de búsqueda, pero en la búsqueda con operadores de existencia, son Yahoo y WiseNut. En la búsqueda booleana es Google el que destaca sobre los demás y en la búsqueda por frase los mejores resultados los obtiene Yahoo, mientras que en la búsqueda por campo, MSN, Google y Yahoo obtuvieron resultados bastante aceptables, al obtener la mayoría de recursos recuperados los términos de búsqueda en el título.

Los peores resultados en la búsqueda en lenguaje natural y por campo corresponden a WiseNut y en la búsqueda por frase a MSN.

En los metabuscadores se observa la misma tendencia, ya que Profusión ofrece buenos resultados en las búsquedas en lenguaje libre y booleana, mientras que es Search el que destaca en las búsquedas con operadores de existencia y por frase. Vivisimo, que ocupa una posición intermedia respecto al resto de metabuscadores, tiene un mal comportamiento en la búsqueda booleana, al recuperar muy pocos recursos con los términos de búsqueda. Excite, por su parte, recupera pocos recursos con los términos de búsqueda en la mayoría de ellas pero es mejor que el resto en la búsqueda por frase.

Finalmente, en la búsqueda por frase, sólo Vivisimo ofrece unos resultados aceptables, ya que en el resto de los metabuscadores, sólo la mitad de los recursos contienen los términos de búsqueda en el título.

En definitiva, en cuanto a la precisión técnica alcanzada por los metabuscadores, sólo cabe mencionar la obtenida por Search y Profusión, correspondiendo a Excite una posición intermedia y a Dogpile, Surfwax y Vivisimo, los peores resultados.

Por tanto, en este sentido hemos de señalar que es necesaria una mejora de los mecanismos de búsqueda de estas herramientas, de forma que permitan, cuando así se requiera, forzar la aparición de los términos en las páginas recuperadas, al igual que las búsquedas por campo, ya que esto posibilitará una mejora en la recuperación dentro del gran número de documentos textuales y multimedia existentes en Internet.

6. Ordenación

Por último, para analizar la ordenación que muestran los buscadores de los recursos localizados, nos basamos en los resultados de la búsqueda por un término. En este sentido valoramos la aparición del término de búsqueda en las metaetiquetas KEY y DESCRIPTION, así como en las frecuencias y peso del término correspondiente en los documentos recuperados.

Por otro lado hemos de advertir que somos conscientes de la limitación que supone analizar la ordenación de resultados basándonos sólo en aspectos como el uso de la metainformación, frecuencia y peso de los términos dado que en los algoritmos que utilizan los buscadores intervienen más factores de los que aquí se reflejan, pero pensamos que esta experiencia puede ayudarnos a comprender, al menos en parte, el funcionamiento de estas herramientas, y conocer hasta qué punto la recuperación refleja el uso de estas variables que consideramos de interés en la recuperación.

- *Utilización de la metainformación*

Como sabemos la metainformación juega un desigual papel entre las herramientas de búsqueda a la hora de ponderar los recursos para ordenar los resultados. Teoma es el buscador que más parece tener en cuenta la metainformación en la ordenación de resultados, ya que es el que más recursos con el término de búsqueda en las etiquetas mencionadas recuperó. Yahoo y WiseNut también utilizan la información de estas etiquetas para la ordenación de los resultados.

Entre los metabuscadores, Vivisimo y Profusión son los que más resultados recuperan con el término en las metaetiquetas, y en Surfwax es en el que mejor se aprecia la relación entre la información allí recogida y la ordenación de los recursos recuperados.

- *Frecuencia y Peso del término de búsqueda*

Respecto al cálculo de frecuencias de aparición de los términos de búsqueda y su peso, teniendo en cuenta su ordenación en grupos de diez en diez, hasta los cincuenta analizados, hemos podido observar una desigual utilización de

estos valores, ya que si bien Yahoo parece tenerlos en cuenta de una forma regular, el resto lo utilizan para ordenar un número determinado de recursos, que bien pueden ser los diez, los veinte o los treinta primeros, apareciendo, de forma anómala, a partir de valores superiores. Este funcionamiento puede ser indicativo de que los cálculos se aplican a un número limitado de recursos, y una vez valorados estos, se vuelven a valorar otros grupos, que pueden adquirir un valor que supera al grupo anterior, cuando una correcta ordenación exigiría una constante decreciente.

Por otro lado, esta forma anómala de ordenar los recursos nos indica que no es suficiente con visualizar los veinte primeros resultados, sino que a partir de éstos, los valores de frecuencia y peso de resultados posteriores pueden ser superiores y por tanto del mismo interés que los aparecidos en lugares anteriores en los listados. Un ejemplo, en este sentido, lo tenemos en Google. En este motor se aprecia una relación entre las frecuencias de aparición de los términos y la ordenación en los veinte primeros resultados, pero en los siguientes esta relación se rompe pasando a depender de otro tipo de cálculos en los que intervienen los algoritmos diseñados de forma específica para la ordenación. Además registra los valores más elevados de las medias de frecuencia de aparición de términos, del peso, de la aparición de los términos en el texto y en hiperenlaces.

Los metabuscadores Dogpile, Ixquick y Vivisimo no utilizan de forma determinante la frecuencia y peso para la ordenación. Excite y Search lo hacen en los diez primeros resultados y con Profusión, tienen, en cuanto a la valoración de las frecuencias de aparición del término, y a la importancia asignada al peso del término en los recursos recuperados, un comportamiento similar al de Google, al colocar los recursos en los que el peso del término es mayor, a partir de los treinta primeros resultados.

- *Coeficiente de correlación de Pearson*

Para conocer si hay alguna una relación entre la ordenación de los recursos en los listados y la frecuencia de aparición del término de búsqueda o su peso, y en cuál de ellos es mayor, el cálculo del Coeficiente de correlación de

Pearson no nos ha proporcionado ningún valor altamente significativo, aunque sí hemos podido observar, que MSN, seguido de Yahoo, son los motores de búsqueda en el que mayor relación hay entre la ordenación, la frecuencia de los términos y el peso.

Entre los metabuscadores la mayor relación se aprecia en Surfwax, lo que llaman la atención frente a los niveles alcanzados en otros aspectos de la evaluación.

Estas conclusiones deben utilizarse con la cautela que unos datos tan limitados exigen, no obstante permiten advertir sobre la necesidad de que estas herramientas sigan mejorando para tratar de resolver los problemas señalados e intenten ajustar más sus resultados tanto a las necesidades de los usuarios como a las expresadas mediante los diferentes tipos de búsqueda. Sería necesaria la posibilidad de expresar determinados aspectos que ayuden a centrar la búsqueda, bien permitiendo seleccionar qué tipo de recurso se busca, o bien mediante la inclusión de filtros tras la recuperación. Tal vez el ofrecimiento al usuario de estas mejoras y su funcionamiento con los recursos correctamente descritos, sirva para que estos aspectos se realicen de forma más generalizada.

Los resultados alcanzados deben servir como punto de partida para seguir investigando y evaluando, y en la medida de lo posible como guía para las personas cuya tarea profesional está relacionada con la recuperación de información bien sea desde el punto de vista de la referencia como de la investigación.

BIBLIOGRAFÍA

ABAD GARCÍA, M. F. 2002. Evaluación de las operaciones de análisis y difusión de la información. En: LÓPEZ YEPES, J., Coord. *Manual de Ciencias de la Documentación,* Madrid, Pirámide, 2002: pp.671-690.

ABAD GARCÍA, M.F. 2005. Evaluación de la calidad de los sistemas de información, Madrid, Síntesis, 2005

ABADAL FALGUERAS, E. 2001. *Sistemas y Servicios de información Digital*, Gijón, Ediciones Trea, 2001.

ABADAL FALGUERAS, E. y CODINA BONILLA, L. 2005. *Bases de Datos documentales: características, funciones y método*. Madrid, Síntesis, 2005.

ACKERMANN, E. y HARTMAN, K. 2003. Searching and Researching on the Internet and the World Wide Web. 3ª ed. Wilsonville, (Oregon), Franklin Beedle and Associates, 2003

AGATA, T. y otros. 1997. A measure for evaluating search engines on the World Wide Web: Retrieval test with ESL. *Library and Information,* 37, 1997: pp. 1-11.

AGUILLO, Isidro. 1998. Hacia un concepto documental de sede web. *El profesional de la información,* 7, (1-2) 1998: pp. 22-41.

AGUILLO, Isidro. 1999. Searching the Web [en línea]. *Cybermetrics*, 1999. <http://www.cindoc.csic.es/cybermetrics/links08.html>[Consulta: agosto de 2001].

AGUILLO, Isidro. 2000. Indicadores hacia una evaluación objetiva (cuantitativa) de sedes web. En: *VII Jornadas Españolas de Documentación* (Bilbao,19-21 octubre, 2000) *La Gestión del conocimiento: retos y soluciones de los profesionales de la información.* Bilbao, Universidad del País Vasco, DL 2000: pp. 233-248.

AGUILLO, I., ORTEGA, J.L. y GRANADINO, B. 2006. Contenidos del buscador Google. Distribución por países, dominios e idiomas. *El profesional de la Información.* 15, (5), septiembre-obtubre 2006: pp. 384-389.

AITCHISON, T.M. 1969-1970. *Comparative evaluation of Index Languages.* London, Institution of Electrical Engineers, 1969-1970. 2 vol.

ALONSO BERROCAL, J.L., FIGUEROLA, C.G. Y ZAZO, A.F. 2004. Cibermetría: nuevas técnicas de estudio aplicables al Web. Gijón, Ediciones Trea, D.L. 2004.

ARMS, W. Y. 2001. *Digital libraries,* Cambridge, The MIT Press, 2001.

BAEZA-YATES, R. y otros. 1999. *Modern information retrieval.* London, Addison Wesley, 1999.

BAR-ILAN, J. 1998.On the Overlap, the precision and estimated recall of search engines. A case study of the query "Erdos". *Scientometrics*, 42, (2) 1998: pp. 207-228.

BAR-ILAN, J. 1998/99. Search Engine Results over Time. A Case Study on Search Engine Stability [en línea]. *Cybermetrics*, Issue 1, Paper 1, 1998/99. <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>[Consulta: septiembre de 2001].

BAR-ILAN, J. 2005. Comparing rankings of search results on the Web. *Information Processing and Management*. 41, 2005: pp. 1511-1519.

BARO I QUERALT, J. 1997. Cerca i recuperació d'informació al World Wide Web: una aproximació a les eines disponibles. En: *6es Jornades Catalanes de Documentació* (Barcelona, 23, 24 y 25 d' octubre de 1997) *Cap a la societat digital: un món en continua transformació.* Barcelona, SOCADI, 1997: pp. 469-479.

BATES, J. M. 1986. Subject acces in online catalogs: a design model. *Journal of the American Society for Information Science*, 37, (5) 1986: pp.: 357-376.

BENITO AMAT, Carlos. 1998. Sistemas de recuperación de información distribuida en Internet. Una revisión de su evolución, sus características y sus perspectivas. Primera Parte. *Revista Española de Documentación Científica*, 21, (4) 1998 pp.:463-474.

BENITO AMAT, Carlos. 1999. Sistemas de recuperación de información distribuida en Internet. Una revisión de su evolución, sus características y sus perspectivas. Segunda Parte. *Revista Española de Documentación Científica*, 22, (1) 1999: pp.: 92-98.

BENITO AMAT, Carlos. 1999b. Sistemas de recuperación de información distribuida en Internet. Una revisión de su evolución, sus características y sus perspectivas. Tercera Parte. *Revista Española de Documentación Científica*. 22, (2) 1999: pp.:268-273.

BERGMAN, M.K. 2001. The deep web: surfacing hidden value. *The Journal of Electronic Publishing* [en línea].<http://www.press.umich.edu/jep/07-01/bergman.html>[Consulta: marzo de 2004].

BERRY, W. M. y BROWNE, M. 1999. *Understanding search engines. Mathematical Modeling and Text Retrieval*, Philadelphia, Siam, cop. 1999

BHARAT, K. y BRODER, A. 1998. A Technique for measuring the relative size and overlap of public Web search engines [en línea]. <http://decweb.ethz.ch/WWW7/1937/com1937.htm>[Consulta: marzo de 2002].

BLAIR, D.C. 1990. *Language and representation in information retrieval*, Amsterdam, Elsevier Sciece Publishers, 1990.

BLAIR, D.C. 2002. Some thoughts on the reported results of TREC. *Information Processing and Management*. 38 (3), 2002: pp. 445-451.

BORGMAN, C. 1989. All users of information retrieval systems are not created equal: an exploration into individual differences. *Information Processing and Management*. 25(3), 1989: pp. 237-252.

BORLUND, P. e INGWERSEN, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53 (3), 1997: pp. 225-250.

BORLUND, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56 (1), 2000: pp. 71-90

BORLUND, P. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8 (3), 2003. [en línea]. <http://informationr.net/ir/8-3/paper152.html>[Consulta: septiembre de 2005].

BRADLEY, P. 1999. The great search engine con-trick. *Online Information 99 Proceeding*, 1999: pp. 259-262.

BRADLEY, P. 2000. *The advanced internet searcher's handbook*, London, Library Association Publishing, 1999, reprinted 2000.

BRIN, S. y PAGE, L. 1998. The anatomy of a large-scale Hypertextual Web Search Engine. Trabajo presentado al Seventh International World Wide Web Conference. Brisbone, Australia, abril de 1998. [en línea] <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm> [Consulta: junio de 2004]

BUCKLAND, M. 1991. *Information and Information Systems*, Westport, Connecticut, Praeger, 1991.

BURDEN, P. 1996. The UK Web Library-WWLib. [en línea]. <http://www.scit.wlv.ac.uk/wwlib/> [Consulta: abril de 2004]

CALAFIA, 1997. *Search engine watch* [en línea]. <http://searchenginewatch.com>[Consulta: mayo de 2001].

CARBALLAR, 1998. Las fuentes de información: estudios teórico-prácticos. En: *VI Jornadas Españolas de Documentación* (Valencia, 29-31 octubre, 1998) *Los Sistemas de Información al sevicio de la Sociedad.* Valencia, FESABID, 1998: pp. .

CARIDAD SEBASTIÁN, Mercedes, 1999. Planes de la Unión Europea para alcanzar el próximo milenio en política del conocimiento. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999, pp. 37-57.

CHAÍN NAVARRO, Celia, 2004. *Técnicas y métodos de recuperación de información*. Murcia, Diego Marín, 2004.

CHARTRON, G. 1997. Repérage de l'information sur Internet: Nouveaux outils, approches bibliohéconomiques et micro-structures [en línea]. <http://www.ccr.jussieu.fr/urfist/cdi99.htm>[Consulta: agosto de 2001].

CHIGNELL, M.H., GWIZDKA, J., BODNER, C. 1999. Discriminating meta-search: a framework for evaluation. *Information Processing and Management*, 35, (3) 1999: pp. 337-362.

CHOWDHURY, G.G. 1999, *Introduction to modern information retrieval*, London, Library Association Publishing, 1999.

CHOWDHURY, G.G. 1999b. The Internet and Information retrieval research: a brief review. *Journal of Documentation,* vol. 55, 2, March 1999: 209-225.

CHOWDHURY, G.G y CHOWDHURY, S. 2001, *Information Sources and Searching on the World Wide Web*, London, Library Association Publishing, 2001.

CHU, H. y ROSENTHAL, M. 1996. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology [en línea]. *Asis 1996 Annual Conference Proceedings*, 1996. <http://www.asis.org/annual-96/ElectronicProceedings/chu.html> [Consulta: febrero de 2000].

CHU, H. 1997. Internet Search Tools: What Can They Offer to Users. *Proceedings of the Eighteenth National Online Meeting.* Medford, NJ, Information Today, Inc, 1997: pp. 73-80.

CLARKE, S.J. y WILLETT, P. 1997. Estimating the recall performance of Web search engines. *ASLIB Proceedings*, 49, (7) 1997: pp. 184-189.

CLEVERDON, C.W. y MILLS, J. 1963. The testing of index language devices. *Aslib Proceedings,* 15, (4) 1963: pp. 106-130.

CLEVERDON, C.W. et al. 1966 *Factors determining the performance of Index Languages.* Cranfield, College of Aeronautics, 1966. 3 vol.

CODINA, Lluis, 1997. Cómo funcionan los servicios de búsqueda en Internet: un informe especial para navegantes y creadores de información (Parte I). *Information World en Español*, 6, (5) 1997: pp. 22-27.

CODINA, Lluis, 1997. Cómo funcionan los servicios de búsqueda en Internet: un informe especial para navegantes y creadores de información (Parte II). *Information World en Español*, 6 (6) 1997: pp.19-27.

CODINA, Lluis, 2000. Evaluación de recursos digitales en línea: conceptos, indicadores y métodos. *Revista Española de Documentación científica,* 23, (1), 2000:pp. 9-14.

COOPER, W.S. 1973. On selecting a mesure of retrieval effectiveness (Parte I). *Journal of the American Society for Information Science,* 24, 1973: pp. 87-100.

CORDON GARCÍA, J.A. 1999. Sobre la información, su necesidad y los modos de acceder a ella. En: DE TORRES RAMÍREZ, Isabel, Coord. *Las fuentes de información: estudios teórico-prácticos*, Madrid, Síntesis, 1999.

COURTOIS, M. P., BAER, W. M. y STARK, M. 1995. Cool tools for searching the Web: A performance evaluation. *Online*, 19,(6) 1995: pp.15-32.

COURTOIS, M. P. 1996. Cool tools for web searching: an update. *Online*, 20 (3) 1996: pp. 29-36.

COURTOIS, M. P. y BERRY, Michael W. 1999. Results Ranking in Web Search Engines. *Online*, 23, (3) 1999: pp. 39-46.

CRASWELL, N., BAILEY, P. y HAWKING, D. 1999. Is it fair to evaluate Web systems using TREC ad hoc methods? *ACM SIGIR '99 Workshop on evaluation of Web Document Retrieval.* 1999.

CRAWFORD, J. 1996. *Evaluation of library and information services*. London: ASLIB, 1996.

CRESTANI, F. y LEE, P.L. 2000. Searching the web by constrained spreading activation. *Information Processing and Management,* 36, 2000: pp. 585-605.

DAVIS, E. 1996. A Comparison of Seven Search Engines [en línea]. Kent, University, 1996. <http://www.iwaynet.net/~lsci/Search/paper.htm>[Consulta: junio de 2001].

DELGADO DOMÍNGUEZ, A. 2001. Herramientas de búsqueda para la WWW [en línea].Trabajo presentado al Congreso Internacional Virtual de Educación, CIVE 2001. <http;//dmi.uib.es/people/adelaida/CIVE/adecive.htm> [Consulta: julio de 2004]

DEMPSEY, L. Meta Detectors [en línea]. <http://www.ariadne.ac.uk/issue3/metadata>[Consulta: marzo de 2000].

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

DIEKEMA, A. et al. 2000. TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. 2000.

DING, W. y MARCHIONINI, G. 1996. A comparative study of web search service preformance. *ASIS 1996 Annual Conference Proceedings,* Baltimore, MD, 1996: pp. 136-142.

DONG, X. and SU, L. 1997. Search engines on the World Wide Web and Information retrieval from the Internet: a review and evaluation. *Online & CDROM review*, 21, (2) 1997: pp. 67-82.

DOYLE, L.B. 1963. Is relevance an adequate criterion for retrieval system evaluation? *Proceedings of the American Documentation Institute,* Part 2, Washington, D.C. 1963: pp.199-200.

DREILINGER, D. y HOWE, A. E. 1997. Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems,* 15, 3, 1997: pp. 195-222.

EÍTO BRUN, R. 2003. Topics maps y la indización de recursos electrónicos en la web. *El profesional de la Información*, 12, (2) 2003: pp. 141-148.

ELLIS, D. 1992. The physical and cognitive paradigms in information retrieval research. *Journal of Documentation,* 45 (3) 1992:171-212.

ELLIS, D. 1994. Paradigms in information retrieval research En: KENT, A. ed. *Encyclopedia of Library and Information Science*, Vol. 54. New York, Marcel Dekker, 1994: pp. 275-291.

ESTIBILL, Asumpció, y ABADAL, Ernest, 2000. Acceso a los recursos web gratuitos desde las bibliotecas. *El profesional de la Información*, 9, (11) 2000: pp. 4-20.

FALK, H. 1997. World Wide Web search and retrieval. *The Electronic Library*, 15, (1) February 1997: pp. 49-55.

FEDERACION INTERNACIONAL DE ASOCIACIONES DE BIBLIOTECARIOS. 1997. *ISBD(ER):International standard bibliographic description for electronic resources: revised from the ISBD(CF): International standard bibliographic description for computer files.* Munchen, Saur, 1997

FELDMAN, S. 1997. Just the answers, please. Choosing a web search service [en línea]. *Information Today*, Inc. <http://www.infotoday.com/searcher/may/story3.htm> [Consulta: abril de 2000].

FELDMAN, S. 1998. Web search services in 1998: trends and challenges. [en línea]. *Searcher,* 6 (6) June 1998: pp. 29-39. <http://www.infotoday.com/searcher/jun98/story2.htm> [Consulta: marzo de 2002].

FELDMAN, S. 2000. Meaning-based search tools: Find what I mean, not what I say [en línea]. *Online,* May 2000. <http://www.findarticles.com/cf_0/m1388/3_24/61640528/print.jhtml> [Consulta: septiembre de 2001].

FIGUEROLA, C.G., ALONSO BERROCAL, J.L., ZAZO RODRÍGUEZ, A.F. 1998. Nuevos puntos de vista en la Recuperación de la Información en el Web. En: *VI Jornadas Españolas de Documentación* (Valencia,29-31 octubre, 1998) *Los Sistemas de Información al sevicio de la Sociedad.* Valencia, FESABID, 1998: pp. 273-280.

FIGUEROLA, C.G., ALONSO BERROCAL, J.L., ZAZO RODRÍGUEZ, A.F. 2000. Diseño de un motor de recuperación de la información para uso experimental y educativo [en línea]. *BiD Biblioteconomia i documentació*, 4, juny, 2000. <http://www.ub.es/biblio/bid/04figue2.htm> [Consulta: julio de 2004].

FRICKE, M. 1998. Measuring recall. *Journal of Information Science*, 24 (6), 1998: pp 409-417.

FUENTES I PUJOL, E. GONZÁLEZ QUESADA, A. y JIMÉNEZ LÓPEZ, A. 2000. Documentación e información electrónica. En: MOREIRO, J.A. (coord.). Manual de documentación informativa. Madrid, Cátedra, 2000: pp. 345-422.

FURNER, J. 1996. The evaluation of hypermedia IR systems: a statement of the problems. [en línea]. *Proceedings of the Second MIRA Workshop*. Padua, 14-15 november. Mark Dunlop University of Padua. 1996 <http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/htir.html [Consulta: octubre de 2002]

GARCÍA CAMARERO, E. y GARCÍA MELERO, L. A. 2001. *La biblioteca digital.* Madrid, Arco Libros, 2001.

GARCÍA FIGUEROLA, C., ZAZO, A. F. y ALONSO BERROCAL, J. L. 2002. La interacción con el usuario en los sistemas de recuperación de información: realimentación por relevancia. *Scire*, 8, (1) enero-junio, 2002: pp. 87-94.

GARCÍA JIMÉNEZ, A. 2002. Organización y gestión del conocimiento en la comunicación. Gijón, Ediciones Trea, 2002.

GARCÍA MARCO, J. y TRAMULLAS SANZ, J. 1996, *World Wide Web: fundamentos, navegación y lenguajes de la red mundial de información*, Madrid: RA-MA, 1996.

GARRIDO PICAZO, P. Y TRAMULLAS SANZ, J. 2004. Un experimento de creación de biblioteca digital con Greenstone. *El profesional de la Información,* 13, (2) marzo-abril, 2004: pp. 84-92.

GILSTER, P. 1996. *Finding it on the Internet*. New York: John Wiley and Son, 1996.

GLOSSBRENNER, A. y GLOSSBRENNER, E. 2001, *Search engines for the World Wide Web*, 3rd ed. Berkeley, Peachpit Press, 2001.

GORDON, M. y PATHAK, P. 1999. Finding information on the World Wide Web: the retrieval effectivness of search engines. *Information Processing and Management*, 35, (2) 1999: pp. 141-180.

GRAVANO, L. y otros 1997. STARTS: Stanford protocol proposal for internet retrieval and search. En: *Proceedings of the ACM SIGMOD International Conference on Management of Data, 1997*

GREEN, D. 2000. The evolution of web searching. *Online Information Review,* 24 (2) 2000: pp. 124-137.

GRIESBAUM, J. 2004. Evaluation of three German search engines: AltaVista.de, Google.de and Lycos.de. *Information Research, 9, 4, paper 189.* [en línea]. <http://InformationR.net/ir/9-4/paper189.html. >.[Consulta: septiembre de 2005]

GRIFONI, G. 1997. Come orientarse tra i motori di ricerca. *Biblioteche oggi,* giugno 1997: pp. 10-16.

GUDIVADA, V. N. et al. 1997. Information Retrieval on the World Wide Web [en línea]. *IEEE Internet Computing*, septiembre-octubre 1997. <http://computer.org/internet>.[Consulta: septiembre de 2001].

GULLI, A. y SIGNORINI, A.2005. The indexable Web is more than 11.5 billion pages. [en línea]. Poster Proceedings of the 14th. International Conference on World Wide Web. <http://www.cs.uiowa.edu/~asignori/web-size/size-indexable.pdf>.[Consulta: septiembre de 2006].

GWIZDKA, J. y CHIGNELL, M. 1999. Towards Information Retrieval measures for evaluation Web search engines. 1999 [en línea]. <http.//www.imedia.ic.utoronto.ca/~jacekg/pubs>.[Consulta: septiembre de 2001].

HARMAN, D. 1995. Overview of the second text retrieval conference (TREC-2). *Information Processing and Management*, 31, (3) 1995: pp. 271-289.

HARMAN, D. 1996. The fourth text retrieval conference (TREC-4). *NIST Special Publication*, Gaithersburg MD, National Institute of Standards and Technology, 1996: pp. 500-236.

HARRY, V. y OPPENHEIM, C. 1993. Evaluations of electronic databases, Part I: Criteria for testing CDROM products. *Online & CDROM review*, vol. 17, (4) 1993: pp. 211-222.

HARRY, V. y OPPENHEIM, C. 1993. Evaluations of electronic databases, Part II: Testing CDROM products. *Online & CDROM review*, 17, (6) 1993: pp. 339-368.

HARTER, S. P. y HERT, C. A. 1997. Evaluation of information retrieval systems: approaches, issues and methods. En: WILLIAMS, Martha E. (ed.) *Annual Review of Information Science and Technology*, Medfor NJ, ASSIS, 32, 1997: pp. 3-94.

HAWKING, D. y otros. 1999. Results and challeges in Web search evaluation. *Computer Networks,* 31, 1999: pp. 1321-1330.

HAWKING, D. y otros. 2001a. Measuring Search Engine Quality. *Information Retrieval*, 4, (1), 2001: pp. 33-59.

HAWKING, D. , CRASWELL, N. y GRIFFITHS, K. 2001b. Which Search Engine is best at finding Online Services? [en línea]. *Proceedings of the Tenth International Worl Wide Web Conference WWW10, Hong Kong, May 1-5, 2001.* <http://citeseer.ist.psu.edu/hawking01which.html>. [Consulta: febrero de 2005].

HEARST, Marti A. 1999. User interfaces and visualization. En BAEZA-YATES, R. Modern Information retrieval (ed.). Harlow, England, Addison-Wesley, 1999: pp. 257-323.

HENSHAW, R. 2001. What Next for Internet Journals? Implications of the Trend Towards Paid Placement in Search Engines. [en línea]. *First Monday,* 6, (9) September 2001. <http://firstmonday.org/issues/issue6_9/henshaw/index.html>.[Consulta: diciembre de 2003]

HERNÁNDEZ PÉREZ, A. 1999a. Las infraestructuras de la Sociedad de la Información: las redes de telecomunicación. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid: Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 113-142.

HERNÁNDEZ PÉREZ, A. 1999b. La búsqueda y recuperación de la información en Internet. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid: Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 213-240.

HERNON, P., ALTMAN, E. 1998. Assesing Service Quality: satisying the Expectations of Library Costumers. Amer Library Assn Editions, 1998

HERTHER, N. 1986. A planning model for optical product evaluation. *Online*, 10, (5) 1986: pp. 128-130

HÍPOLA, P. y VARGAS-QUESADA, B. 1999. Agentes inteligentes: definición y tipología. Los agentes de información. *El profesional de la información,* 8, (4) abril 1999: pp.: 13-21

HÍPOLA, P., VARGAS-QUESADA, B. y MONTES, A. 1999b. Descripción y evaluación de agentes multibuscadores. *El profesional de la información,* 8, (4) abril 1999: pp.: 13-21

HOCK, R. 1998. How to do field searching in Web search engines: a field trip. *Online*, 22, (3) May 1998: pp. 18-22.

HOCK, R. 1999. *The extreme searcher's guide to web search engines: a habdbook for serious searcher*, CyberAge Books, Medford (New Jersey), 1999.

HOCK, R. 2001. Revisiting Web Search Engines: features and commands [en línea]. *Online*, 25, (5) 2001. <http://www.onlineinc.com/onlinemag/OL2001/oltocsept01.html>[Consulta: octubre de 2001].

HOCK, R. 2001. A new era of search engines: not just Web pages anymore. *Online*, 26, (5) Sept-oct. 2002: pp. 20-27.

HOWE, A. E. y DREILINGER, D. 1997. SavvySearch: a meta-search engine that learns which search engines to query. *AI Magazine,* 18, 2, 1997.

HU, Wen-Chen et al. 2001a. An overview of World Wide Web Search Technologies. Proceedings of 5th World Multi-Conference on Systemics, Cybernetics and Informatics SCI 2001, Orlando, Florida, july 22-25, 2001.

HU, Wen-Chen et al. 2001b. An XML World Wide Web search engine using approximate structural matching. En Proceedings of 5th World Multi-Conference on Systemics, Cybernetics and Informatics SCI 2001, Orlando, Florida, july 22-25, 2001.

HU, Wen-Chen y YEH, Jyh-Haw. 2002. World Wide Web Search Technologies. En: *Achitectural Issues of Web-Enabled Electronic Business.* London, Idea Group Publishing, 2002.

INGWERSEN, P. 1992. *Information Retrieval Interaction,* London, Taylor Graham, 1992.

INGWERSEN, P. Y WILLET, P. 1995. An introduction to algorithmic and cognitive approaches for in formation retrieval. *Libri*, 45, (3-4), 1995: pp. 160-177.

JANSEN, Bernard J. et al. 1998. Real Life Information Retrieval: a study of user queries on the Web. *SIGIR Forum,* 32, (1), 1998: pp. 5-17.

JANSEN, Bernard J. y POOCH, U. 2001. A review of Web searching studies and framework for future research. *Journal of the American Society for Information Science*, 52, (3) 2001: pp. 235-246.

JANSEN, Bernard J. y SPINK, A. 2003. An analysis of Web documents retrieved and viewed. *The 4th International Conference on Internet Computing. Las Vegas, Nevada, 23-26 june 2003:* pp. 65-69.

JENKINS, C. et al. 1998. Searching the World Wide Web: an evaluation of available tools and methodologies. *Information and Software Technology*, 39, 1998: pp. 985-994.

JIMÉNEZ PIANO, M. 2001. Evaluación de sedes Web. *Revista Española de Documentación Científica*, 24, (4) 2001: pp.:405-432.

JOHNSON, F.C., GRIFFITHS, J.R. Y HARTLEY, R.J. 2001. DEVISE a framework for the evaluation of Internet search engines. [en línea]. Manchester, Manchester Metropolitan University, Centre for Research in Library and Information Management, 2001 .[Consulta: septiembre de 2003].

KEILY, L. 1997. Improving resource discovery on the Internet: the user perspective. *Proceedings of the 21st International Online Information Meeting*, 1997: pp. 205-212.

KIMMEL, S. 1996. Robot generated databases on the world wide web. *Database*, 19, (1) 1996: pp. 41-49.

KING, D. 2000. Specialized search engines: alternatives to the Big Guys. *Online*, 24, (3) 2000: pp. 67-74

KOCH, T. 1996. Internet search services [en línea]. <http://www.lub.lu.se/tk/demos/DO9603-meng.html>[Consulta: abril de 2002].

KOCH, T. 1998. Searching the Web-Systematic overview over indexes [en línea]. <http://www.lub.lu.se/tk/websearch_systemat.html>[Consulta: abril de 2002].

KORFHAGE, R. R. 1997. *Information Retrieval and Storage*. New York, Wiley Computer Publisher, 1997.

KOSTER, M. 1994. A Standard for Robot Exclusion [en línea]. <http://info.webcrawler.com/mark/projects/robots/robots.html>[Consulta: enero de 2002].

KOSTER, M. 1995 Robots in the Web: threat or treat? [en línea].*ConneXions*, 9 (4) april 1995. <http://info.webcrawler.com/mark/projects/robots/threat-or-treat.htm> [Consulta: enero de 2002].

KOSTER, M. 1998. The Web Robots. FAQ [en línea]. <http://info.Webcrawler.com/mak/projects/robots/faq.html>[Consulta: febrero de 2001].

KOSTER, M. HTML Author's Guide to the Robots META tag <http://info.webcrawler.com/mark/projects/robots/meta-user.html> [Consulta: abril de 2005].

KUHLTHAU, C. C. 1991. Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42, (5) 1991: pp.: 361-371.

KUK, G. 1999. Social Science Information Gateway for Psychology: a utility test of SOSIG. *Social Science Computer Review,* 17, 1999: pp. 451-454

LAKE. 1997. 2nd Annual search engine shoot-out [en línea]. *PC Computing*. <http://www4.zdnet.com/pccomp/features/exc10997/sear/sear.html>[Consulta: febrero de 2002].

LANCASTER, F.W. 1971. The cost-effectiveness analysis of information retrieval and dissemination systems. *Journal of the American Society for Information Science*, 22, (1) 1971: pp.: 12-27

LANCASTER, F.W. 1995. *El control de vocabulario en la recuperación de información*, Valencia, Universidad, 1995.

LANCASTER, F.W. 1998. *Indexing and abstracting in Theory and Practice*, 2nd. ed. Londres, Library Association Publishing, 1998.

LANCASTER, F.W. y FAYEN, E. G. 1973. I*nformation retrieval On-Line*, Los Angeles, Melville Publishing Co. 1973.

LANCASTER, F.W. y WARMER, A. 1993. *Information Retrieval Today*. Allington, Ressources Press, 1993.

LANDONI, M. y BELL, S. 2000. Information retrieval tecniques for evaluating search engines: a critical overview. *Aslib Proceedings*, 52, (3), March 2000, pp. 124-129.

LARGE, A., TEDD, L. A. y HARTLEY, R. J. 2001. *Information seeking in the online age: principles and practice.* München, Saur, 2001.

LAWRENCE, S. y GILES, L. 1998. How big is the Web? How much of the Web do the search engines index? How up to date are the search engines? [en línea]. <http://www.neci.nec.com/homepages/lawrence/websize.html>[Consulta: enero de 2000].

LAWRENCE, S. y GILES, L. 1998b. Searching the World Wide Web. *Science*, 280, 1998: pp. 98-100.

LAWRENCE, S. y GILES, L. 1998c. Context and page analysis for improved web search. *IEEE Internet Computer.* July-August, 1998: pp. 38-46.

LAWRENCE, S. y GILES, L. 1998d. Inquirus, the NECI meta search engine. Trabajo presentado al Seventh International World Wide Web Conference. Brisbone, Australia, abril de 1998. [en línea] <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm> [Consulta: noviembre de 2005]

LAWRENCE, S. y GILES, L. 1999. Accessibility and Distribution of Information on the Web. *Nature,* 400, 1999: pp. 107-109.

LAWRENCE, S. 2000. Context in Web Search [en línea]. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,* 23, (3) September 2000. <http://www.research.microsoft.com/research/db/debull/A00sept/issue.htm [Consulta: septiembre de 2001].

LEBEDEV, A. 1997. Best search engines for finding scientific information in the Web [en línea]. Moscow, State University, 1997.
<http://www.chem.msu.su/eng/comparison.html>[Consulta: junio de 1999]

LEIGHTON, H.V. 1995. Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, WebCrawler and WWWWorm. [en línea]. Winona, Winona State University
<http://www.winona.msus.edu/library/webind.htm> [Consulta: junio de 2002].

LEIGHTON, H.V. y SRIVASTAVA, J. 1997. Precision among World Wide Web search services (search engines): AltaVista, Excite, HotBot, Infoseek, Lycos [en línea].
<http://www.winona.msus.edu/is-f/library-f/webind2.htm>[Consulta: diciembre de 2000].

LEIGHTON, H.V. y SRIVASTAVA, J. 1999. First 20 Precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Information Science*, 50, (10), 1999: pp. 870-881.

LEINER, B. M. et al. 1997. Una breve historia de Internet (Primera Parte) [en línea]. *Novática*, 130, nov-dic. 1997.
<http://www.ati.es/DOCS/internet/histint/index.html>[Consulta: mayo de 2001].

LEINER, B. M. et al. 1998. Una breve historia de Internet (Segunda Parte) [en línea]. *Novática*, 131, ene-feb. 1998.
<http://www.ati.es/DOCS/internet/histint/index.html>[Consulta: mayo de 2001].

LEONARD, Andrew J. 1996. Where to find anything on the net [en línea].
<http://www.cnet.com/Content/Reviews/Search/>[Consulta: octubre de 2000].

LI, Y. 1998. Toward a qualitative search engine. IEEE Internet Computing, 2,(4) 1998 : pp. 24-29.

LI, L., SHANG, y Y ZHANG, W. 2001. Relevance evaluation of search engines' query results. [en línea]. *Proceedings of the Tenth International Worl Wide Web Conference WWW10, Hong Kong, May 1-5, 2001.<http://www10.org/cdrom/posters/frame.html>* [Consulta: marzo de 2005].

LIDDY, E. 2001. How a seach engine works [en línea]. *Searcher,* May 2001 <http://www.infotoday.com/searcher/may01/liddy.htm> [Consulta: marzo de 2002].

LIDDY, E. y Myaeng, S. 2003. DR-LINK: A system updated for TREC2. En: Harman, Donna K. (Editor). The Second Text Retrieval Conference (TREC-2). NIST Special Publication 500-215, Gaithersburg MD, National Institute of Standards and Technology, 2003: pp. 85-99.
http://trec.nist.gov/pubs.html [Consulta: noviembre de 2006].

LJOSLAND, M. 2000. Evaluation of Web search engines and the search for better ranking algorithms [en línea]. Paper presented at the SIGIR99 Workshop on Evaluation of Web retrieval, August 19, 1999.
<http://www.aitel.hist.no/~mildrid/dring/paper/SIGIR.html> [Consulta: noviembre de 2003].

LOSSAU, N. 2004. Search engine technology and Digital Libraries: Libraries need to discover the Academic Internet [en línea]. *D-Lib Magazine,* 10, (6) 2004.
<http://www.dlib.org/dlib/june04/
lossau/06lossau.html> [Consulta: marzo de 2004].

LUCAS, W. y TOPI, H. 2004. Training for web search: will it get you in shape? *Journal of the American Society for Information Science and Technology*, 55, (13) 2004: pp. 1183-1198.

MALDONADO MARTÍNEZ, A. y FERNÁNDEZ SÁNCHEZ, E. 1998. Evaluación de los principales "buscadores" desde un punto de vista documental: recogida, análisis y recuperación de recursos de información. En: *VI Jornadas Españolas de Documentación* (Valencia,29-31 octubre, 1998) *Los Sistemas de Información al sevicio de la Sociedad.* Valencia, FESABID, 1998: pp. 528-551.

MALDONADO MARTÍNEZ, A. y FERNÁNDEZ SÁNCHEZ, E. 1999. Comparing Internet Search Tools. *Online Information 99 Proceedings*, 1999: pp. 263-266.

MALDONADO MARTÍNEZ, A. y FERNÁNDEZ SÁNCHEZ, E. 2000. Análisis comparativo de buscadores en Internet. *El profesional de la información*, 9, (3) 2000: pp. 40-46.

MALDONADO MARTÍNEZ, A. (Coord.) 2001. La información especializada en Internet: directorio de recursos de interés académico y profesional. Madrid, CSIC, 2001.

MALDONADO MARTÍNEZ, A y RODRÍGUEZ YUNTA, L. (Coord.) 2006. La información especializada en Internet: directorio de recursos de interés académico y profesional. Madrid, CSIC, 2006.

MANOLA, F. y MILLER, E. 2003.RDF Primer. *W3C Working Draft.* Januari, 2003.

MARCOS MORA, P. 1998. Motores de recuperación de información: un análisis comparativo. (Parte I). *El profesional de la información*, 7, (1-2) 1998: pp. 18-22.

MARCOS MORA, P. 1998. Motores de recuperación de información: un análisis comparativo. (Parte II). *El profesional de la información*, 7, (3), marzo, 1998: pp. 13-20.

MARTÍNEZ MÉNDEZ, J. 2001. Aproximación general a la evaluación de la recuperación de información por medio de los motores de búsqueda. *6 Encuentros IBERSID*, octubre. Zaragoza, 2001.

MARTÍNEZ MÉNDEZ, J. 2002. Propuesta y desarrollo de un modelo para la evaluación de la recuperación de la información en Internet. [en línea]. Tesis doctoral. 2002, Universidad de Murcia. <http://www.cervantesvirtual.com/servlet/SirveObras/67937253111315940722202/0100> [Consulta: febrero de 2004].

MARTÍNEZ MÉNDEZ, J. y RODRÍGUEZ MUÑOZ, J. V. 2003. Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web [en línea]. *Information Research: an International Electronic Journal*, 8, (2), january, paper nº 148, 2003. <http://InformationR.net/ir/8-2/paper148.html> [Consulta: noviembre de 2003].

MARTÍNEZ MÉNDEZ, J. y RODRÍGUEZ MUÑOZ, J. V. 2004a. Reflexiones sobre la evaluación de los Sistemas de recuperación de información: necesidad, utilidad y viabilidad. *Analesde Documentación*, 7, 2004: pp. 153-170.

MARTÍNEZ MÉNDEZ, J. y RODRÍGUEZ MUÑOZ, J. V. 2004b. Aspectos de la evaluación de los Sistemas de recuperación de información: necesidades y utilidad. *Anales*, 8, 2004. Disponible en Internet <http.//www.um.es/fjmm/anales2004.pdf>[Consulta: junio de 2004].

MARTOS, A. 2001. *Herramientas de búsqueda en Internet*, Madrid, Prentice Hall, 2001.

MEADOW, Charles T. 1992. *Text information retrieval systems.* San Diego, (California), etc, Academic Press, 1992.

MEGHABGHAB, D.B. y MEGHABGHAB, G.V. (1996). Information retrieval in cyverspace. *Proceedings of American Society for Information Science ASIS Mid-Year Meeting, 18-22 mayo,* 1996: 224-237

MÉNDEZ RODRÍGUEZ, E. M.1999a. Política del tándem Clinton-Gore en materia de información: el liderázgo de los Estados Unidos En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 3-36.

MÉNDEZ RODRÍGUEZ, E. M. 1999b. Globalización de la Información. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 59-87.

MEYRIAT, J. 1981. Document, documentation, documentologie. *Revue de bibliologie, schema et schematisation*, 19, 1981: pp. 2-5.

MILLER, D.C. 1987. Evaluating CD-ROMs: To buy or what to buy. *Database*, 10, (3) 1987: pp. 36-42

MILSTEAD, J.; FELDMAN, S. 1999. Metadata: Cataloging by Any Other Name [en línea]. *Online*, January 1999. <http://www. onlineinc.com/onlinemag/OL1999/milstead1.html>[Consulta: febrero de 2000].

MING, H. 2000. Comparison for Three Search Engines [en línea]. Toronto, University, 2000. <http://gypsy.rose.utoronto.ca/people/ming/report.html>[Consulta: enero de 2002].

MIZZARO, S. 1997. Relevance: the whole history. *Journal of the American Society for Information Science*, 48, (9) 1997: pp. 810-832.

MORVILLE, P. ROSENFELD, L. y JANES, J. 1996. *The internet searcher's handbook*, New York, Neal-Schuman Publishers, Inc. 1996.

MOSCOSO, P. 2002. Sistemas de información documental: concepto, modelo, estructura y organización. En: LÓPEZ YEPES, J., Coord. *Manual de Ciencias de la Documentación,* Madrid, Pirámide, 2002: pp.519-536.

MOYA ANEGÓN, F. 2002. Sistemas avanzados de recuperación de la información. En: LÓPEZ YEPES, J., Coord. *Manual de Ciencias de la Documentación,* Madrid, Pirámide, 2002: pp. 553-599.

MURRAY, Brian H. and MOORE, A. 2000. *Sizing the Internet: a white paper.* [en línea]. Civeillance, 2000. <http://www.cyveillance.com/web/corporate/white_papers.htm>[Consulta: mayo de 2002].

NASIOS, Y., KORINTHIOS, G. y DESPOTOPOULOS, Y. 1998. Evaluation of search engines [en línea]: Report undertaken by the National Technical University of Athens on behalf of the European Commission and Project PIPER, july 1998. <http://www.piper.ntua.gr/reports/searching/doc.0000.htm>[Consulta: septiembre de 2001].

NEC Research Institute. 1998.

NICK, Z. Z. Y THEMIS, P. 2001.Web searching using a genetic algorithm. *IEEE Internet Computing,* 5, 2, 2001: pp. 18-26.

NOGALES FLORES, J.T. 1999a. Los usos básicos de Internet. Servicios y aplicaciones En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 143-173.

NOGALES FLORES, J.T. 1999b. La revolución de la Worl Wide Web. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 175-212.

NOTESS, Greg R. 1997. Measuring the size of Internet databases *Database*, 20, (5) 1997: pp. 69-72.

NOTESS, Greg R. 1998. Tips on Evaluating Web Databases. *Database*, 21 (4) 1998.

NOTESS, Greg R. 1999. AltaVista's Internetional Mirrors [en línea]. *EContent,* August 1999. <http://www.ec.mag.net/EC1999/net8.html>[Consulta: septiembre de 2001].

NOTESS, Greg R. 1999b. A multiplicity of Databases on Search Engines [en línea]. *EContent,* October 1999. <http://www.ec.mag.net/EC1999/net10.html> [Consulta: septiembre de 2001].

NOTESS, Greg R. 1999c. On the Net [en línea]. *Online,* may 1999. <http://www.onlinemag.net/OL1999/net5.html> [Consulta: junio de 2002].

NOTESS, Greg R. 2000. Search Engine Statistics: Unique Hits Report [en línea]. <http://www.notess.com/search/stats/unique.shtml> [Consulta: marzo de 2001].

NOTESS, Greg R. 2000b. Search Engine Showdown Analysis: Boolean Searching on Google. [en línea]. <http://searchenginesowdown.com/features/google/googleboolean.html> [Consulta: julio de 2007].

NOTESS, Greg R. 2002. Internet Search Engine Update [en línea]. *Online,* nov-dec 2002. <http://www.infotoday.com/online/nov02/SearchEngine.html> [Consulta: junio de 2004].

OLVERA LOBO, Mª Dolores. 1998, *Evaluación de la recuperación de la información en Internet: un modelo experimental*. Tesis doctoral. Marzo 1998. Universidad de Granada.

OLVERA LOBO, Mª Dolores. 1999. Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias. *El profesional de la información*, vol. 8, nº 11, nov. 1999: pp. 4-14.

OLVERA LOBO, Mª Dolores. 1999b. Métodos y técnicas para la indización y la recuperación de los recursos de la World Wide Web. *Boletín de la Asociación Andaluza de Bibliotecarios,* 57, diciembre 1999.

OLVERA LOBO, Mª D. 2000. Rendimiento de los sitemas de recuperación de información en la World Wide Web: revisión metodológica. *Revista Española de Documentación Científica,* 23, (1) 2000: pp. 63-77.

OLVERA LOBO, Mª D. 2000b. Rendimiento de los sitemas de recuperación de información en la Web: evaluación de los servicios de búsqueda. *Revista Española de Documentación Científica,* 23, (3) 2000: pp. 302-316.

OPPENHEIM, C., MORRIS, A. y McNIGHT, C. 2000. The evaluation of WWW search engines. *Journal of Documentation,* 56, (2) March 2000: pp. 190-211.

ORGANIZACIÓN INTERNACIONAL DE NORMALIZACIÓN (ISO). 1988. *Recueil de Normes ISO 1: Documentation et Information*, Troisième éd. Geneve, ISO, 1988.

OVERTON, R. 1996. Search engines get faster and faster, but not always better [en línea]. *PC World*, septiembre 1996. <http://www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html>[Consulta: julio de 2000].

OXNARD, L. y EVANS, A. (2003). Methodologies for the Automatic Location of Academic and Educational Texts on the Internet. University of Leeds, School of Geography, 2003

PAO, M.L. y WORTHEN, D.B.1989. Retrieval Effectiveness by Semantic and Citation Searching. *Journal of the American Society for Information Science,* 40, 1989: pp. 226-235

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

PEÑA, R., BAEZA-YATES, R. y RODRÍGUEZ, J. V. (2002) Gestión digital de la información: de bits a bibliotecas digitales y la web. Madrid, RA-MA, 2002.

PINTO MOLINA, M. 1999. Tratamiento de los contenidos en la Sociedad de la Información. En: CARIDAD SEBASTIÁN, Mercedes, Coord. *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos*, Madrid, Editorial Centro de Estudios Ramón Areces; Universidad Carlos III de Madrid, D.L. 1999: pp. 267-288.

PINTO MOLINA, M. y otros. 2002. *Indización y resumen de documentos digitales y multimedia: técnicas y procedimientos*. Gijón, Ediciones Trea, 2001.

POLLOCK, A. y HOCKLEY, A. 1997. What's Wrong with Internet Searching [en línea]. *D-Lib Magazine*, marzo 1997. <http://www.dlib.org/dlib/march97/bt/03pollock.html>[Consulta: julio 2000].

POULTER, A. 1997. The design of World Wide Web search engines: a critical review. *Program*, 31, (2) April 1997: pp. 131-145.

PRICE, G. 2001. Web Search Engine FAQS: Questions, Answers and Issues [en línea]. *Searcher*, 9, (9) Oct 2001. <http://www.infotoday.com/searcher/oct01/searcher.htm>. [Consulta: febrero 2002].

RIJSBERGEN, C. J. 1979. *Information Retrieval.* 2nd. ed. London: Butterworth, 1979.

RODRÍGUEZ BRAVO, B. 2002. *El documento, entre la tradición y la renovación*. Gijón: TREA, 2002.

ROUSSEAU, R. 1998/99. Daily time series of common single word searches in AltaVista and NorthernLight [en línea]. *Cybermetrics*, 1, (2) 1998/99. <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>[Consulta: septiembre de 2001].

RUIZ DE OSMA DELATAS. 1998. Las redes como fuente de información. En: TORRES RAMÍREZ, Isabel de, *Las fuentes de información: estudios teórico-prácticos,* Madrid, Síntesis, 1998, pp.: 401-415.

RYEN, W., RUTHVEN, I. y JOSE, M. J. 2001. Web document summarisation: a task-oriented evaluation. [en línea]. International Workshop on Digital Libraries. Proceedings of the 12th International Database and Expert Systems Applications Conference (DEXA 2001). Munich, 3-7 September 2001. <http://www.dcs.gla.ac.uk/%7Ewhiter/dlib.pdf> [Consulta: febrero de 2005].

SALAZAR, Idota. 2005. *Las profundidades de Internet: Accede a la información que los buscadores no encuentran y descubre el futuro inteligente de la Red.* Gijon, Trea, 2005.

SALTON, G. y McGILL, J. 1983, *Introduction to modern information retrieval*, New York, McGraw-Hill, 1983.

SALVADOR OLIVAN, José.Antonio y VIDAL BORDES, Fco. Javier. 2000. Evaluación del rendimiento de los motores de búsqueda en la recuperación de información en la WWW. *Documentación de las Ciencias de la Información,* 23, 2000: pp. 93-108

SANZ, M.A. 1998. Fundamentos históricos de la Internet en Europa y España [en línea]. *Boletín de RedIRIS*, 45, octubre 1998. <http://www.rediris.es/rediris7boletin/45/enfoque2.html>[Consulta: mayo de 2001].

SARACEVIC, T. (et al.) 1988a. A study of information seeking and retrieving. I: background and methodology. *Journal of the American Society for Information Science.* 39 (3), 1988: pp. 161-176.

SARACEVIC, T. y KANTOR, P. 1988b. A study of information seeking and retrieving. II: user, questins, and effectiveness. *Journal of the American Society for Information Science.* 39 (3), 1988: pp. 177-196.

SARACEVIC, T. y KANTOR, P. 1988c. A study of information seeking and retrieving. III: searchers, searches and overlap. *Journal of the American Society for Information Science.* 39 (3), 1988: pp. 197-216.

SARACEVIC, T. 1996. Releance reconsidered ’96. En: *Ingwersen P. y Pors, N. O. eds. Proceedings of CoLIS 2, Second Internacional Conference on Conceptions of Library and Information Science.* Copenhagen, 1996: pp. 201-218.

SAVOY, J. y PICARD, J. 2001. Retrieval effectiveness on the web. *Information Processing and Management,* 37, 2001: pp. 543-569.

SCHLICHTING, C. y NILSEN, E. 1996. Signal detection analysis of WWW search engines [en línea]. <http://www.microsoft.com/Usability/webconf/schlichting/schlichting.htm> [Consulta: mayo de 2002].

SCOVILLE, R. 1996. Special report: Find it on the net! [en línea]. *PC World*, 14, (1) 1996: p. 125. <http://www.pcworld.com/software/internet_www/articles/jan96/jan9635.html> [Consulta: octubre de 2001].

SHERMAN, C. 1999. The future Search Web of Search [en línea] *Online*, 23 (3), May 1999. <http://www.findarticles.com/cf_0/m1388/3_23/54474833/print.jhtml> [Consulta: mayo de 2002].

SHERMAN, C. 2000. The future revisited: what's new with Web Search [en línea] *Online*, 23 (5), May 2000. <http://www.findarticles.com/cf_0/m1388/3_24/6160525/print.jhtm [Consulta: diciembre de 2005].

SHERMAN, C. 2002. Teoma vs Google, round two. [en línea] <http://siliconvalley.internet.com/news/print.php/3531_1002061>

SLOT, M. 1997. The matrix of internet catalogs and search engines [en línea]. <http://www.ambrosiasw.com/~fprefect/matrix/>[Consulta: diciembre de 1999].

SNOW, B. 2000. The Internet's hidden content and how to find it [en línea]. *Online*, 24 (3) 2000: pp.61-66. <http://www.infotoday.com/online/OLtocs/OLtocmay00.html> [Consulta: noviembre de 2003].

SNYDER, H, y ROSENBAUM, H. 1999. Can search engines be used as tools for Web-link analysis? a critical view. *Journal of Documentation*, 55, (4) September 1999: pp. 375-384.

SONNENREICH, W. 1997. A History of Search Engines. [en línea]. <http://www.wiley.com/legacy/compbooks/sonnenreich/history.html> [Consulta: abril de 2002].

SONNENREICH, W. 1998. *Web developer.com Guide to search engines,* New York, Wiley Computer Publishing, 1998.

SPARCK JONES, K. y WILLET, P. (eds) 1997. *Readings in information retrieval*. San Francisco, Morgan Kaufmann, 1997

SPINELLIS, D. 2003. The decay and failures of Web references. *Communications of th ACM,* 46, (1) 2003: pp. 71-77.

SPINK y otros. 1996. Multiple search sessions model of end-user behavior : an exploratory study. *Journal of the American Society for Information Science,* 47 (8), 1996: pp. 603-609.

SPINK y otros. 1997. Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science,* 48 (5), 1997: pp. 382-394.

SPINK y otros. 2001. T. Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology,* 52, (3), 2001: pp. 226-234

STANTON, D. y HOOPER, T. 1992. The LIBS Internet Acces Software: an Overview and Evaluation. [en línea]. *The Public-Access Computer Systems Review* 3, (4) 1992: pp. 4-14. <http:/epress.lib.uh.edu/pr/v3/n4/Stanton.3n4>[Consulta: noviembre de 2006].

STEINBERG, S. G. 1996. Seek and ye shall find (maybe). *Wired*, 4 (05) 1996: 109 ff

STEINER, G. A. 1979. *Planificación de la alta dirección* Pamplona, Universidad de Navarra, 1979.

STINSON, L. 1999. Searching tricky company names. (Dialog, LEXIS-NEXIS, Altavista, Excite, Infoseek, Northern Light evaluated)(Evaluation). *Searcher*, Sept, 1999.

STOBART, S. y KERRIDGE, S. 1996. *WWW search engine study* [en línea]. Sunderland: University, 1996. <http://osiris.sunderland.ac.uk/sst/se/>[Consulta: diciembre de 2001].

SU, L.T. 1992. Evaluation measures for interactive information retrieval. *Information Processing and Management,* 28, (4) 1998: pp. 503-516.

SU, L.T. 1994. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45, 1994: pp. 207-217.

SU, L.T. 1998. Value of search results as a whole as the single measure of information retrieval performance. *Information Processing and Management,* 34, (5) 1998: pp. 557-579.

SUGIURA, A. y ETZIONI, O. 2000. Query routing for Web search engines: Architecture and experiments. En: *Proceedings of the 9th. International World Wide Conference.* Amsterdam, may 2000.

SULLIVAN , D. 1998. How Yahoo Works [en línea]. <http://searchenginewatch.com> [Consulta: octubre de 2002].

SULLIVAN , D. 2000. Media Metrix Search Engine Ratings [en línea]. SearchEngineWatch.com: 2000 <http://searchenginewatch.com/reports/mediametrix.html> [Consulta: julio de 2001].

SULLIVAN , D. 2001. Buying your way in to search engines [en línea]. SearchEngineWatch.com: 2001 <http://searchenginewatch.com/webmasters/paid.html >.[Consulta: septiembre de 2001].

SULLIVAN , D. 2005. New study sizes up the Web [en línea] .http://www.clickz.com/experts/search /article.php/3512376 [Consulta: octubre de 2005].

TAGUE-SUTCLIFFE, J. M. 1992. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28,(4) 1992: pp. 467-490.

TAYLOR, ARLENE, G. y CLEMSON, P. Acces to Networked documents. Catalogs? Search Engines? Both? [en línea]. <http://www.oclc.org/oclc/man/colloq/taylor.htm>[Consulta: octubre de 2001].

THUNENDER, H. y ERWING, J. 1998. How to succeed in promoting your Web site: the impact of search engine registration on retrieval of a World Wide Web site. *Information technology and Libraries,* September, 1998: pp. 173-179.

TORRES, A. 1998. ¿Hay que quemar Internet?". En: *RAMONET, I. Internet, el mundo que llega,* Madrid, Alianza, 1998: pp 165-173.

TOMAIUOLO, N. G. y PACKER, J. G. 1996. An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries,* 16, (6) 1996: pp. 58-62.

TRAMULLAS, Jesús y OLVERA, Mª Dolores. 2001. *Recuperación de la información en Internet*, Madrid, Ra-Ma, 2001.

UBIETO ARTUR, Antonio Paulo. 1995. *Documentación automatizada: manual de uso de la red Internet*. Zaragoza, Anubar, D. L. 1995.

UBIETO ARTUR, Antonio Paulo. 2002. Internet. En: LÓPEZ YEPES, J., Coord. *Manual de Ciencias de la Documentación,* Madrid, Pirámide, 2002: pp. 489-518.

VAN SLYPE, G. *Lenguajes de Indización : concepción, construcción y utilización en los sistemas. Documentales*. Madrid : Pirámide, 1991

VAQUERO PULIDO, J. R. 1997. Recuperación de la información en Internet: motores y otros agentes de búsqueda. *Scire* , 3, (2) julio-diciembre, 1997: pp. 85-100.

VAUGHAN, L. 2004. New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, 40, (4), may 2004: pp. 677-691.

VAUGHAN, L. y THELWALL, M. 2004. Search engine coverage bias: evidence and possible causes. *Information Processing and Management*, 40, (4), may 2004: pp. 693-707.

VENDITTO, G. 1996. Search engine showdown: IW labs test seven Internet search tools. *Internet World*, 7 (5) 1996: pp. 79-86.

VIANELLO OSTI, M. 2004 *El Hipertexto entre la utopía y la aplicación: identidad, problemática y tendencias de la Web*. Gijón: Ediciones Trea, 2004.

VIDAL BORDES, F. J. y SALVADOR OLIVÁN, J. A. 2000. La implementación de metadatos y Dublin Core en sedes y páginas web de bibliotecas y centros de documentación de universidades y centros de investigación de la red IRIS. En: *VII Jornadas Españolas de Documentación* (Bilbao,19-21 octubre, 2000) *La Gestión del conocimiento: retos y soluciones de los profesionales de la información.* Bilbao, Universidad del País Vasco, DL 2000: pp. 197-209.

VILLASEÑOR RODRÍGUEZ, I. 1999. Los instrumentos para la recuperación de la información: Las fuentes. En: *Las fuentes de información. Estudios teórico-prácticos*, Madrid, Síntesis, 1999: p. 29-42.

WEISE, E. 2001a. Search sites brush up on people skills [en línea]. *USA Today.* <http://www.usatoday.com/life/cyber/tech/review/crg841.htm>[Consulta: septiembre de 2001].

WEISE, E. 2001b. Succesful Net search stars with need [en línea]. *USA Today.* <http://www.usatoday.com/life/cyber/tech/review/crg842.htm>[Consulta: septiembre de 2001].

WESTERA, G. 1996. Robot-driven search engine evaluation overview [en línea]. <http://www.cortin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/index.htm> [Consulta: octubre de 2001].

WESTERA, G. 1997. On the Edge of the Abyss: Locating Information of the Vortex of the World Wide Web. Seventh Asian Pacific Specials, Health and Law Librarian's Conference. Perth, Western Australia, 12-16 October, 1997

WESTERA, G. 2000. Comparison of Search Engine User Interface Capabilities [en línea]. Curtin, University of Technology, 2000. <http://lisweb.curtin.edu.au/staff/gwpersonal/compare.html>[Consulta: marzo de 2000].

WILSON, T. D. 1984. The cognitive approach to information seeking behaviour and information use. *Journal of Documentation,* 55(3), 1984: pp. 249-270.

WINSHIP, I. R. 1995. World Wide Web searching tools-an evaluation [en línea]. *VINE*, 99, 1995: pp. 49-54. <http://bubl.bath.ac.uk/BUBL/IWinship.html> [Consulta: septiembre de 1999].

WISHARD, L. 1998. Precision Among Internet Search Engines: An Earth Sciences Case Study [en línea]. Pennsyilvania, State University, 1998. <http://www.library.ucsb.edu/istl/98-spring/article5.html> [Consulta: agosto de 2001].

XIE, M., WANG, H. Y GOH, T.N. 1998. Quality dimensions of Internet search engines. *Journal of Information Science*, 24 (5), 1998: pp. 365-372.

XIE, M., WANG, H. Y GOH, T.N. 1999. Service quality of Internet search engines. *Journal of Information Science*, 25 (6), 1999: pp. 499-507.

XU, J. L. 1999. Internet search engines: Real world IR issues and challenges. Paper presented at the Conference on Information and Knowledge Management. Kansas City, MO. 1999.

YUWONO, B y LEE, Dik L. Search and Ranking Algorithms for Locating Resources on the World Wide Web. [en línea]. <http://www.searchenginewatch.com/webmasters/rank.html> [Consulta: febrero de 2005].

ZORN, P. et al. 1996. Advanced Web searching: tricks of the trade. [en línea]. *Online*, 20 (3), 1996: pp.15-28. <http://www.onlineinc.com/onlinemag/MayOL/zorn5.html> [Consulta: octubre de 2001].