Recuperación de la información en Internet: motores y otros agentes de búsqueda

José Raúl Vaquero Pulido
Universidad de Salamanca

0.1. Resumen

Ante la gran cantidad de documentos existentes en Internet y la imposibilidad de realizar una búsqueda total por hiperenlaces, se diseñaron en Internet unas herramientas de búsqueda que facilitan la recuperación de documentos. El presente artículo se estructura en dos partes: por una parte intenta dar a conocer los motores de búsqueda, su evolución en la reciente historia de Internet, sus elementos componentes, la forma que tienen de trabajar y su clasificación, intentando dar así solución al caos terminológico existente actualmente. Por otra parte se analizan las ventajas e inconvenientes en el proceso de recuperación de información, planteando finalmente una serie de soluciones que el autor considera necesarias para ganar en efectividad. Finalmente, se ha intentado ofrecer una bibliografía lo más amplia posible para mantenerse actualizado en este cambiante mundo de la Red y los motores de búsqueda. (Autor)

Palabras clave : Motores de búsqueda. Recuperación de la información. Internet.

0.2. Abstract

Due to the enormous number of documents that Internet contains and the practical impossibility of achieving search exhaustivity throughout hyperlinks, it was necessary to design some search tools to facilitate retrieval. This article is structured in two parts. Firstly, it gives an introduction to the concept of search engines, their evolution in the recent history of Internet, their components, their procedures and a prospective classification, trying to solve the current terminological chaos. Second, their comparative advantages and disadvantages are analyzed, and some proposals for further effectivity are given. Finally, a comprehensive bibliography for further readings is suggested. (Author amended).

Keywords: Search engines. Information retrieva. Internet.

1. Introducción

Aunque muchas son las transformaciones que ha sufrido Internet desde que a finales de los años 60 el departamento de defensa de los Estados Unidos pusiera en marcha el proyecto ARPANET, los servicios básicos como Mail, Telnet y Ftp siguen funcionando con la misma filosofía, aunque algunos como Gopher estén cada vez más cayendo en el olvido dejando paso al World Wide Web, verdadera revolución que ha acercado Internet al gran público, alejándolo del carácter restringido que durante muchos años tuviera a investigadores, educadores y científicos.

Hoy en día, un usuario individual no sólo tiene la oportunidad de conectarse sin problemas a este medio excitante, sino que además puede introducir de forma sencilla mucha información (que puede ser de calidad o no), originando, hoy más que nunca, una verdadera explosión informativa.

Pero la existencia de tal cantidad de información, lo poco organizada que está, el carácter cambiante de la propia Red y que no toda la información tiene la misma calidad o valor, provoca que el proceso de búsqueda de información se convierta en algo difícil, tedioso e incluso peligroso para nuestro monedero si la búsqueda se alarga excesivamente en el tiempo.

Es por ello por lo que han aparecido en Internet unas herramientas que localizan la información y los recursos deseados. Estas herramientas son los motores de búsquedas o también llamados buscadores.

2. Definición

Podemos definir un motor de búsqueda como "herramienta Web que localiza de forma rápida información existente en Internet".

3. Historia y evolución

Buscar y encontrar algo en Internet ha sido siempre un problema, por lo que desde muy temprano se diseñaron herramientas que permitieran de una forma fácil organizar y encontrar información en Internet.

Hasta 1990 los usuarios de Internet debían conocer comandos Unix para navegar y buscar en la Red. La herramienta principal para localizar ficheros era el comando *grep* de Unix. Archie, que es una elaboración de grep, fue creado en 1989 por un grupo de investigadores del Centro de Computación de la Universidad McGill, en Estados Unidos, con el fin de encontrar ficheros Ftp más fácilmente.

En 1991, investigadores del Centro de Computación de la Universidad de Minnesota inventaron el Gopher, con un interface de usuario más amigable y un

formato de menú de ficheros. Tal fue la popularidad de Gopher que en poco tiempo aparecieron millares de sitios Gopher, para los cuales no existía un índice central o un mecanismo de búsqueda. Es por ello, por lo que en 1992 se creó Veronica, permitiendo a los usuarios buscar en el Gopherespacio.

Tim Berners-Lee creó el World Wide Web en 1991 para ayudar a los científicos del CERN a publicar y compartir sus investigaciones. Pero la aparición de la Web supuso una revolución ya que su carácter gráfico e hipertextual la acercó al gran público, aumentando la cantidad de información existente, por lo que se planteó de nuevo la creación de otras herramientas que permitieran buscar y encontrar en la Web.

Un primer intento de proporcionar un acceso directo a los recursos de la Web fue *The Mother of All the Bulletin Boards*, que aunque no disfrutó de excesiva popularidad, puede considerarse como punto de partida de los sistemas de indización distribuida. Fue desarrollado por Oliver McBryan en la Universidad de Colorado y tenía como objetivo generar un índice global de recursos Web, basado en la colaboración de los administradores de servidores, que notificarían a los gestores de The MABB la presencia de nuevas aportaciones en la Red, intentando los gestores acoplar cada nuevo registro en una ordenación por materias.

En 1993 la primera generación de motores -WWW Worm y WebCrawlerhace su aparición, aunque a principios de 1994 son sustituidos por motores de búsqueda más poderosos como Alta Vista, Excite, Infoseek, Lycos y Open Text. Pero incluso estos motores que usan algoritmos de búsqueda por palabras a menudo fallan para encontrar información relevante, es por ello por lo que actualmente está surgiendo una nueva generación de motores que incorpora algoritmos inteligentes - Inktomi's Slurp- basados en indización por conceptos, es decir, tratan de determinar lo que el usuario piensa, no lo que dice.

4. Elementos componentes

Tres son los elementos que componen un motor de búsqueda :

4.1. Interface

Es la página Web a la que accede el usuario. En ella se establece el tipo de búsqueda y podemos distinguir dos tipos :

- Formularios: Se presenta una página con formularios en los que se introducen las palabras claves de búsqueda relacionadas con el tema que nos interesa, junto con la lógica a emplear.
- Con Directorios (1): Además de contar con formularios, estructuran la información jerárquicamente por materias. Para encontrar la información se desciende de los temas más generales hasta los más específicos.

4.2. Base de datos textual

Índice de palabras, frases y datos asociados con la dirección de páginas Web (URL), programas, ficheros, etc. La información se da de alta y baja, pero siempre dejando una referencia para que después, quien busque un tema relacionado, pueda encontrar la dirección y un pequeño resumen de lo que contiene.

4.3. Robot (2)

Programa de ordenador que está diseñado para recorrer de forma automática la estructura hipertexto de la Web con el fin de llevar a cabo una o varias de las siguientes funciones :

- Análisis estadístico : Miden el crecimiento de la Web, número de servidores conectados, etc.
- Mantenimiento de la estructura hipertextual de la World Wide Web : Verificando la corrección de los enlaces entre documentos y eliminando o guardando información de los denominados "enlaces muertos" (dead links). Es decir, páginas Web que ya han desaparecido.
- Duplicación de directorios Ftp (Mirrors): Incrementando su utilidad a un número mayor de usuarios.
- Creación automática de bases de datos textuales a partir de los documentos html distribuidos por los distintos servidores (3). Debido a que cada robot está programado para buscar en la Red de diferente forma, la información almacenada en cada base de datos puede ser diferente.

Los términos *robots*, *spiders*, *wanderers*, *knowbots*, *webcrawlers* y *web scanners* suelen considerarse como sinónimos, aunque es posible encontrar matizaciones, como la recogida por Michael Berns quien distingue entre :

- *Spiders* (Arañas) : Robots que crean bases de datos tomando como referencia los distintos nexos que configuran la estructura hipertexto de la Red.
- *Knowbots*: Robots programado para localizar referencias hipertexto dirigidas hacia un documento, servidor, etc., en particular. Este tipo de robots permiten evaluar el impacto de las distintas aportaciones que engrosan las distintas áreas de conocimiento presentes en la Red.
- *Wanderers* (Vagabundos) : Robots encargados de la medición del crecimiento de la Web.

Otra matización interesante es la aportada por Martijn Koster, quien entiende como sinónimos robots, spiders y webcrawlers, añadiendo los conceptos de:

- Worms (Gusanos) : Robots encargados de tareas de duplicació (*en mirrors*)

- WebAnts : Conjunto de robots físicamente alejados que cooperan para la consecución de distintos objetivos.

5. Funcionamiento

Los motores realizan un funcionamiento de trabajo doble, bien diferenciado:

- El primero hace referencia al robot y su forma de actualizar la base de datos. Periódicamente, y tomando como partida una URL inicial, el robot recupera un fichero en formato html que transfiere al sistema local, en donde procede a su indización, guardando en su base de datos los términos indizados, así como la localización URL completa del documento. Una vez indizado el documento, el robot identifica los enlaces hipertextuales que contiene el mismo y de forma recursiva procede a recuperar los documentos referenciados en esos enlaces, procediendo a su indización, obtención de nuevos enlaces, etc.
- En el segundo, el usuario acude a ellos como si de una hoja Web más se tratara, donde puede seleccionar de un directorio el tema de interés, o rellenar los campos de búsqueda con las palabras claves relacionadas con el tema de interés, seleccionar el tipo de búsqueda (lógica booleana, truncamientos, etc.) y lanzar la pregunta. El motor revisa su base de datos en busca de términos que coincidan con los de la pregunta y en breves instantes devuelve una lista con las direcciones de las hojas Web que más se ajusten a la pregunta. Dependiendo de la complejidad del motor y del algoritmo de búsqueda empleado, al principio de la lista tendremos las direcciones que mejor se adaptan al perfil de búsqueda. Después de comprobar las direcciones interesantes, sólo habrá que pulsar sobre ellas para conectarse y ver si se ha acertado.

6. Clasificación

Aunque en el fondo todos los motores tienen el mismo objetivo, hacer rápida y sencilla la recuperación de información en Internet, podemos establecer unas clasificaciones de los motores según varios aspectos :

6.1. Según la información que buscan

1. Generales: Buscan información general en la Red (ej. : Alta Vista, Excite, Infoseek, Lycos, Webcrawler, etc.). Son los más comunes.
2. De servicios: Buscan sólo informaciones que se encuentran en algún servicio de Internet (Archie, Finger, Gopher, Newsgroup, Ftp, Telnet, etc.). Dentro de estos destacar :

a) Software : Buscan software para recuperarlo, generalmente shareware o freeware (ej . : Shareware.com).

b) Dirección : Buscan direcciones de personas o instituciones (ej. : Four 11 Directory).

3. Temáticos: Especializados en un tema determinado (ej. : los motores que buscan información en español, sobre arte, etc.). Hacia estos es hacia donde parecen dirigirse actualmente los nuevos motores.

6.2. Según el acceso

1. Libres : Cualquiera puede consultarlos, sin límite de resultados ofrecidos.
2. Privados : Implementados por una persona, empresa o compañía para su uso propio, no estando disponible sus informaciones al público en general.
3. Limitados : Motores a los que hay que pagar una subscripción para un uso completo, pero que permiten un número limitado de resultados en la versión libre.

6.3. Según la forma de adquirir el motor

1. Inadquiribles : Aquellos motores que sólo pueden ser consultados pero de ninguna manera pueden ser adquiridos por los usuarios para su uso.
2. Shareware : Se pueden adquirir de forma gratuita para un uso personal (Excite)
3. Comerciales : Se pueden adquirir para su uso después de haberlo comprado. Es muy común encontrarse con versiones reducida del Motor que sirven de prueba (ej.: CompasSearch Web Server).

Debido a la cada vez mayor cantidad de motores, existen en la Red otras herramientas de gran utilidad para buscar y recuperar información, son los que hemos dado en llamar Recopiladores de motores de búsqueda, que no son motores propiamente dicho, sino páginas Web en las que se introducen las palabras de búsqueda, se eligen los operadores de delimiten la búsqueda y se envía a uno o varios motores de búsqueda, los cuales devolverán los resultados en un breve período de tiempo. En general, el problema que presentan es que no pueden utilizar las herramientas de búsqueda de los motores, por lo que los resultados pueden ser menos ajustados que si se utilizaran estos. Dos tipos de recopiladores:

1. De búsqueda conjunta: Lanzan la pregunta a varios motores al mismo tiempo. (Ej. : Savvy Search y MetaCrawler).

2. De búsqueda individual: Combinan varios formatos de búsqueda dentro de una misma página, pero sólo permiten la búsqueda de uno en uno

7. Lista de los principales motores

Se identifica en la lista que sigue más abajo, el tipo de motor según las siguientes abreviaturas :

1. G : General.
2. R : Recopilador.
3. S : Servicio.
   a) S-NEW : News..
   b) S-MAI : Mail.
   c) S-SOFT : Software.
   d) S-GOP : Gopher.
   e) S-TEL : Telnet.
   f) S-FING : Finger.
4. T : Temático.
   a) T-ESP : Búsqueda en castellano.

NOMBRE	URL	TIPO
Achoo	http ://www.achoo.com/achoo/search.htm	T-ESP
Aliweb	http ://web.nexor.co.uk/public/aliweb/aliweb.html	G
All-In-One	http ://www.albany.net/allinone	R
Alta Vista Web Search	http ://www.altavista.digital.com	G
Apollo Advertising	http ://apollo.co.uk	T
Archie Server	http ://www.fwi.uva.nl/fun/archie.html	S-SOFT
ArchiePlex at CUI	http ://cuiwww.unige.ch/archieplexform.html	S-SOFT
ArchiePlex at NASA	http ://www.lerc.nasa.gov/archieplex/doc/form.html	S-SOFT
Art History Information	http ://www.ahip.getty.edu	T
AstroNet	http ://www.stsci.edu/astroweb/astronomy.html	T
Auyantepui	http ://www.auyantepui.com	T-ESP
B.I.W .E.	http ://biwe.cesat.es	T-ESP
BlackWidow	http ://140.190.65.12/~khooghee/index.html	G
Cadê	http ://www.cade.com.br	T
CD Search	http ://blueridge.infomkt.ibm.com/knudsen/cdsearch .html	T
Cibercentro	http ://www.cibercentro.com/busqueda	T-ESP

CICA Windows Software	http ://www.nova.edu/Inter-Links/ software/ windows.html	S-SOFT
CMC Information	http :// www.december.com/cmc/info/Index.html	T
Computer Vendor Guide	http ://www.ronin.com/SBA	T
CS-HKUST WWW	http ://dbcl13.cs.ust.hk :8001/IndexServer	G
CUSI	http ://pubweb.nexor.co.uk/public/cusi/cusi.html	R
DejaNews	http ://www.dejanews.com	S-NEW
Digisearch	http ://www.digiway.com/digisearch	R
Directorio Global Net	http ://www.golbalnt.com	T-ESP
Directorio Ipl	http ://www.ipl.com.gt/bases/ipl/diresp.htm	T-ESP
DNA	http ://www.iwcc.com/directorio/directorio.html	T-ESP
Dónde ?	http ://donde.uji.es	T-ESP
El Índice	http ://www.globalcom.es/indice	T-ESP
El País Vasco	http ://www.PaisVasco.com/info.htm	T-ESP
Elcano	http ://www.elcano.com	T-ESP
Encis Net	http ://www.encis.es/cast	T-ESP
Encuéntrelo	http ://www.encuentrelo.com	T-ESP
ESP-Email Search	http ://www.esp.co.uk	S-MAI
Excite NetSearch	http ://www.excite.com	G
Explora México	http ://www.explore-mex.com	T-ESP
Fantástico	http ://www.fantastico.com	T-ESP
FOUR11 Directory	http ://www.four11.com	S-MAI
FTP search	http ://ftpsearch.unit.no	S-SOFT
Galaxy	http ://galaxy.einet.net	G
Games Domain Search	http ://www.gamesdomain.co.uk	T
Gauchonet	http ://www.gauchonet.com	T-ESP
GetURL	http ://Snark.apana.org.au/James/GetURL	G
Harvest Broker	http ://town.hall.org/Harvest/brokers/www-home-pages/query.htm	G
HENSA Software	http ://unix.hensa.ac.uk	S-SOFT
HotBot	http ://www.hotbot.com	G
Hytelnet	http ://galaxy.einet.net/hytelnet/START.TXT.html	S-TEL
Indica	http ://www.m3w3.com.mx/INDICA	T-ESP
InfoMacHyperArchive	http ://hyperarchive.lcs.mit.edu/HyperArchive.html	S-SOFT
InfoSeek Guide	http ://www.infoseek.com	G
Inspector de Telépolis	http ://www.telepolis.com	T-ESP

Internet Address Finder	http ://www.iaf.net	S-MAI
Internet en Castellano	http ://buscar.interaccess.cl	T-ESP
Internet Sleuth	http ://www.isleuth.com	R
Jughead	gopher ://logic.uc.wlu.edu :3002/7	S-GOP
Jumbo	http ://www.jumbo.com/home_Page.html	S-SOFT
Katipo	http ://www.vuw.ac.nz/~newbery/Katipo.html	G
La brújula	http ://www.ba.net/robot	T-ESP
LeoNet	http ://www.fut.es/~jbarrat/leonet.html	T-ESP
List Servers	http ://www.nova.edu/Inter-Links/listserv.html	S-MAI
LookUP ! Directory	http ://www.lookup.com	S-MAI
Lycos	http ://www.lycos.com	G
Magellan	http ://www.mckinley.com	G
MetaCrawler	http ://www.cs.washington.edu	R
Mex Master	http ://www.mexmaster.com	T-ESP
Mex Search	http ://www.yellow.com.mx	T-ESP
México Web Guide	http ://mexico.web.com.mx	T-ESP
MOMspider	http ://www.ics.uci.edu/WebSoft/MOMspider	G
Mundivia	http ://www.mundivia.es	T-ESP
Nexor ArchiePlex	http ://pubweb.nexor.co.uk/public/archie/servers.hm	S-SOFT
OKRA net.citizen	http :/okra.ucr.edu/okra	S-MAI
Olé	http ://www.ole.es	T-ESP
OO Bibliography	http ://cuiwww.unige.ch/cgi-bin/bibrefs	T
Open Text Index	http ://www.opentext.com	G
Otis Index	http ://www.interlog.com/~gordo/otis-index.html	R
Ozú	http ://www.ozu.com	T-ESP
RBSE Spider	http ://rbse.jsc.nasa.gov/eichmann/urlsearch.html	G
Rick Harris' Abstracts	http ://daneel.rdt.monash.edu.au/tr/siteslist.html	T
Savvy Search	http ://www.guaraldi.cs.colostate.edu :2000/form	R
SBEL	http ://rtn.net.mx/sbet	T-ESP
Search	http ://www.search.com	R
Seric	http ://empresas.seric.es	T
Shareware.com	http ://www.shareware.com	S-SOFT
Shase	http ://www.shareware.com	S-SOFT
Showbiz	http ://www.showbizdata.com	G
Sift	http ://sift.stanford.edu	S

Snoopie	http ://www.snoopie.com	S-SOFT
Stardot	http ://www.stardot.com/services.html	T
Tarantula	http ://www.tarantula.com.mx	T-ESP
The Peregrinator	http ://www.maths.usyd.edu.au :8000/jimr/pe/	T
Tribal Voice	http ://www.tribal.com/search.htm	G
Trovator	http ://trovator.combios.es	T-ESP
Universid. Buenos Aires	http ://www.uba.ar/Architext/AT	T-ESP
Veronica	gopher ://veronica.scs.unr.edu/11/veronica	S-GOP
Vieiros	http ://www.cesatel.es/Itemaga/vieiros	T-ESP
Vilaweb	http ://vilaweb.com	T-ESP
Wais FAQ Search	http ://www.cs.ruu.nl/cgi-bin/faqwais	T
Wanderer	http ://www.mit.edu/people/mkgray/net	G
WebCrawler	http ://www.webcrawler.com	G
What's New Too	http ://newtoo.manifest.com	G
Who's Who on Internet	http ://web.city.ac.uk/citylive/pages.html	T
WhoWhere	http : //www.whowhere.com/org.hzml	S-FING
Worl File Project	http ://filepile.com	S-SOFT
World Email Directory	http ://worldemail.com/wede4.shtml	S-MAI
WWW Virtual Library	http ://www.edoc.com/vjournal	T
WWWW	http ://www.cs.colorado.edu/home/mcbuyan/ wwww.html	G
Yahoo	http ://www.yahoo.com	G
Yahooligans	http ://www.yahooligans.com	G
Your Personal Network	http ://www.ypn.com	G

8. Ventajas e inconvenientes

Una vez que hemos analizado los aspectos teóricos y técnicos de los motores es necesario ver cuales son sus ventajas e inconvenientes en el proceso de recuperación de información en Internet.

8.1. Ventajas

Sus ventajas han de ser siempre entendidas con un prisma amplio, ya que hoy por hoy no son ventajas completas. En resumen, los motores vendrían a solucionar algunos de los problemas que supone el buscar información en Internet, ya que :

- Localizan información de interés entre la gran cantidad de documentos existentes.
- Organizan la información. Bien en directorios a la hora de realizar la búsqueda, en bases de datos o motores especializados por materias, y lógicamente a la hora de mostrar la información.
- Actualizan sus bases de datos periódicamente.

8.2. Inconvenientes

Podríamos destacar los siguientes inconvenientes:

8.2.1. El hecho de quela efectividad de un motor depende de muy variados factores:

a) *Usuarios*:: Aunque la influencia de los usuarios es de carácter subjetivo, no medible, su importancia es enorme en el proceso de recuperación de la información. Resumiendo, tres serían los factores de los usuarios que afectarían a la efectividad de un motor, y por consiguiente, a la efectividad de los resultados:

1. Conocimientos que el usuario tiene sobre la materia.
2. Conocimientos que el usuario tiene del interface y de las lógicas matemáticas, truncamientos, etc.
3. Modo de expresar su necesidad de información.

b) *Robots*: Del algoritmo (programa) que utilizan, que sea o no capaz de recuperar todos los documentos, dejando escapar las menos excepciones posibles.

c) *Indizacion* : Por ser este un punto medible y objetivo es sobre el que más se ha investigado. La efectividad de un motor desde este punto de vista dependerá de varios factores:

1. El campo de indización: Las técnicas de indización varían de un motor a otro, pero en general se indiza por la dirección de la página, el título, los encabezamientos por los que comienza cada sección, las primeras líneas del texto o el texto completo. Lógicamente, un motor será más efectivo si indiza por el texto completo que si indiza sólo por el título.
2. Lógica de indización: Determina el modo en que se extraen los términos de un documento. Fundamentalmente existen dos métodos :
3. Estadístico: Calcula la frecuencia de aparición de los términos significativos, confrontando en un primer paso las palabras del texto con un antidiccionario de términos vacíos o “stop-word list” para elimi nar las palabras con bajo contenido informativo. Su inconveniente es

que no tiene en cuenta las palabras sinónimas ni homónimas.

4. Asignación : La indización se realiza asignando palabras claves procedentes de un tesauro y se ejecuta en dos etapas : búsqueda de las entradas del tesauro correspondiente a los conceptos presentes en el documento, y traducción de dichos conceptos por los descriptores del tesauro.
5. Autor de la indización : Actualmente se ve que la indización humana es superior a la automática, ya que hoy por hoy sólo el ser humano es capaz de extraer la información presente en un documento, aunque no esté expresada por unas palabras determinadas.

8.2.2. La utilización de estas herramientas sobrecargan el tráfico de la Red**.**

8.2.3. Resulta realmente muy costoso (tiempo / recursos) mantener actualizada una base de datos, ya que el robot debe periódicamente comprobar si la información guardada en ella sigue vigente en el mismo URL.

8.3. Soluciones

Aunque todavía queda mucho terreno por recorrer, se está tendiendo hacia las siguientes soluciones:

- Diseñar interfaces fáciles y amigables al usuario.
- Permitir al usuario realizar sus preguntas en lenguaje natural, es decir, sin utilizar operadores booleanos u otras estructuras lógicas.
- Utilizar lenguajes de programación de última generación que permitan al robot tomar decisiones semejantes al pensamiento humano.
- Indizar la máxima información de los documentos (obtenida del URL, título, texto completo, etc.).
- Utilizar técnicas automáticas que imiten la inteligencia humana para indizar documentos.
- Potenciar la indización distribuida : De forma general consiste en que el administrador de cada servidor se registre en el servidor global encargado de llevar la indización distribuida y cree un fichero índice en la que guarde la información de los ficheros html que desea poner en la Web. Posteriormente, el servidor global recuperará mediante un robot los ficheros índices de los servidores registrados (4). Para realizar una búsqueda ya sólo se acudirá al motor de este servidor global.

Las ventajas que suponen este último sistema son grandes y entre ellas estaría que:

- Se reduce el consumo de tiempo y recursos que de la Red hacen actualmente los robots, ya que únicamente se efectuará una conexión a los servi-

dores registrados para recuperar un fichero único, evitando el tener que recorrer todos los directorios del servidor.

- Se respeta el deseo de los servidores que no desean poner su información, o desean excluir parte de ella, a los usuario de la Web.
- Aumenta la calidad de la indización, y con ella la eficacia de la búsqueda, ya que se sustituye la indización automática (con todos sus problemas) por la indización humana.

Pero para conseguir estas ventajas se debe :

- Normalizar el tipo de fichero-índice. El Internet Anonymous Ftp Archives Working Group de la IETF ha propuesto el formato IAFA, que consiste en documentos de texto plano en el que se consignan pares atributo-valor.
- Normalizar el nombre y ubicación de este fichero, para que el robot del servidor global pueda recuperarlo regularmente.
- Duplicar las bases de datos e interfaces de búsqueda para facilitar la oportunidad de acceso a todos los usuarios de Internet.

9. Conclusiones

Debido a que la "recuperación perfecta" consistente en leer todos los documentos almacenados, reteniendo los relevantes y olvidando todos los demás, es una tarea imposible de realizar en Internet, es necesario utilizar herramientas que faciliten y mejoren el acceso a la información.

Los motores de búsqueda son, hoy por hoy, esas herramientas útiles para todos aquellos que deseen recuperar información o recursos en Internet de una forma rápida, salvando los problemas de volumen y actualización de la información existente en la Red.

Como inconveniente señalar que debido a la gran cantidad de información existente, las búsquedas deben ser muy precisas ya que si no los resultados darán muchos documentos no relevantes. Para realizar estas búsquedas precisas se debe conocer muy bien la lógica booleana y las técnicas de truncamiento, conocimientos muchas veces sólo conocidos por especialistas en documentación.

10. Notas

(1) Una hoja Web que sólo muestre la información por temas, no es un motor, sino un Directorio. Un motor se caracteriza por que cuenta con formularios donde se introducen los términos / operadores de búsqueda, los cuales son contrastados con la BD.

(2) El robot no tiene por que aparecer siempre, ya que en algunos motores son los usuarios o Webmaster los que introducen sus hojas en la base de datos.

(3) Esta función es la que más nos interesa ya que es la que hace relación con los motores de búsqueda.

(4) Actualmente esta experiencia se está llevando a cabo con Aliweb : http ://web.nexor.co.uk/public/aliweb/doc/search.html

11. Referencias

11.1. Libros

Carballar, José A. (1995). Internet : El mundo en sus manos. Ra-ma, 1995.

Cheong, Fah-Chun (1996). Internet Agents Spiders, Wanderers, Brokers and Bots. Indianapolis : New Riders Publishing, 1996.

Gilster, Paul (1995). El navegante de Internet. Madrid : Anaya Multimedia, 1995.

Watson, Mark. Programming intelligent agents for the Internet.

Williams, Joseph (1996). Bots and other Internet beasties. Sams, 1996.

11.2. Artículos de revistas

Daroca, Rafael (1996). Perdidos por la Red : Los Buscadores // JUMPing. (mayo 1996) 60-63.

Egea Alonso, Álvaro (1996). En busca de la Información. ¿Quién sabe dónde ? Netmanía. 1 : 2 (abril 1996) 60-66.

Fletcher, Gordon ; Greenhill, Anita (1995). Academic referencing of Internet-based resources // Aslib Proceedings. 47 : 11/12 (1995) 245-252.

Pareja, Enrique (1997). Como buscar en Internet // Servicom magazine. 12 (1997) 56-61.

Richardson, Eric C. (1996). Addan Engine. Internet World. (mayo 1996) 88-92.

Rodrige, Real (1995). Les réseaux informatiques // Documentation et bibliothèques. (enero-marzo 1995) 5-11.

Rubio, Luciano (1996). Cruzamos el Atlántico : Motores de búsqueda en la Red (II) // PC Actual. (diciembre 1996) 348-351.

Rubio, Luciano (1996b). Los índices de Internet : Motores de búsqueda en la Red (I) // PC Actual. (noviembre 1996) 350-353.

Salvador, Emilio (1996). Internet : Guía de navegantes // ABC informática. 1 (mayo 1996) 34-35.

Seltzer, Richard (1996). Heads Up // Internet World. (abril 1996) 68-69.

Venditto, Gus (1996). Search engine showdown // Internet World. (mayo 1996) 79-86.

11.3. Páginas Web

Alta Vista vs. Lycos (marzo 1996), Ariadne. URL : http : //ukoln.bath.ac.uk/ariadne/issuez/engines.

Berns, Michael (junio 1996). RoboSearch, University of Toronto.URL : http ://www.oise.on.ca/~mberns/RoboSearch.html.

Cómo funcionan los motores de búsqueda (enero 1997), Universidad de Alcalá, Programas de Gerencia. URL : http ://www.alcala.es/internet/buscar/funionam.htm

Cooper, Ian (diciembre 1996). Indexing the World, University of Kent. URL : http ://stork.ukc.ac.uk/computer_science/HTML/Pubs/HC10-94/toc.html.

Eagan, Ann (junio 1996). Spiders and worms and crawlers, Oh My : Searching on the World Wide Web. URL : http ://www.library.ucsb.edu/untangle/eagan.html.

Felt, Elizabeth ; Scales, Jane (mayo 1996). Web robots. URL : http ://www.wsulibs.wsu.edu/general/robots.htm.

Fischer, Keith D. (febrero 1995). The WWW robot and search engine FAQ. URL : http ://science.smsu.edu/robot/faq

Greg, R. (mayo 1996). Internet search engines & finding aids capabilities & features. URL : http ://www.imt.net/~notess/compeng.html.

Hacia la automatización de una tarea rutinaria : empleo de agentes de indización (diciembre 1996), Facultad de Biblioteconomía y Documentación de la Universidad de Granada. URL : http ://www.ugr.es/~felix/g9/Know2.htm.

Hypermail 1.02 (junio 1996). Robots mailing list archive by thread. URL : http ://info.webcrawler.com/mailing-lists/robots/index.html.

Kollar, Charles P. ; Leavitt, John R. R. ; Mauldin, Michael (junio 1996). Robot exclusion standard revisited. URL : http ://www.kollar.com/robots.html.

Koster, Martijn (enero 1996). WWW robot frequently asked questions. URL : http ://info.webcrawler.com/mak/projects/robots/faq.html.

Koster, Martijn (mayo 1996). A standard for robot exclusion. URL : http ://info.webcrawler.com/mak/projects/robots/norobots.html.

Recuperando información desde el cliente Web : Robots personales (diciembre 1996), Facultad de Biblioteconomía y Documentación de la Universidad de Granada. URL : http ://www.ugr.es/~felix/g9/Know3.htm.

Rodríguez, Gerard (febrero 1997) . AlephWeb : Estudio técnico de los buscadores Web. Un nuevo esquema. URL: http ://www.aleph.ac.upc.es/isoccat/prensa/alephweb.html.

Scoville, Richard (enero 1996). Special report : Find it on the net !.URL: http ://www.pcworld.com/reprints/lycos.htm.

Soluciones alternativas al empleo de Robots : indización distribuida (Diciembre 1996), Facultad de Biblioteconomía y Documentación de la Universidad de Granada. URL : http ://www.ugr.es/~felix/g9/Know4.htm.

Sullivan, Danny (junio 1996). Guide to the major players. URL : http ://maxonline.com/webmasters/major.htm.

Sullivan, Danny (junio 1996). How search engine (say they) work. URL : http ://maxonline.com/webmasters/work.htm.

Sullivan, Danny (junio 1996). Search engine comparison chart. URL : http ://maxonline.com/webmasters/char.htm.

Terminología : ¿Robot, Arañas, Vagabundos... ? (diciembre 1996), Facultad de Biblioteconomía y Documentación de la Universidad de Granada. URL : http ://www.ugr.es/~felix/g9/Know1.htm.

Vicent, Antonio (octubre 1996). ¿Cómo buscar ?, Web for Schools. URL : http ://wfs.vub.ac.be/schools/timeline/search/Buscar/Indice.htm

Web search tool features (abril 1996), University of Northumbia, URL: http ://www.unn.ac.uk/features.htm.

Winship, Ian R. (junio 1996). World Wide Web searching tools - an evaluation. URL: http ://www.bubl.bath.ac.uk/BUBL/Iwinship.html.