Del multibuscador al metabuscador: Los agentes trazadores en Internet

Isidro F. Aguillo
CINDOC-CSIC
Joaquin Costa, 22
28002 Madrid
isidro@cindoc.csic.es
91-5635482

Resumen:

La localización y recuperación de información en el World Wide Web es uno de los principales retos que afrontan los documentalistas. Las herramientas de primera generación (motores de búsqueda e índices) no han resuelto adecuadamente los problemas de ruido documental y de escasa exhaustividad de los resultados. Los multibuscadores, dado el bajo nivel de solapamiento entre los grandes motores de búsqueda, pueden ayudar a solucionar estos problemas. Las herramientas de segunda generación, basadas en programas cliente, han expandido el concepto de forma que son capaces de explorar la naturaleza hipertextual de la web. Se acuña el término metabuscadores para describir a los programas capaces de identificar a través de la red, siguiendo la madeja hipertextual, una serie de documentos pertinentes a una estrategia de búsqueda a partir de unas sedes originales o "semilla". Se discuten características y prestaciones prácticas de los principales productos disponibles.

Palabras clave: World Wide Web, recuperación de información, herramientas de segunda generación, agentes autónomos, multibuscadores, metabuscadores

Abstract:

The information retrieval in the World Wide Web is one of the main challenges that confront the documentalists. The tools of first generation (search engines and indices) have not solved suitably the problems of noise and poor coverage of the results. The multisearchers, due to the low overlap between the great search engines, can help to solve these problems. The tools of second generation (client-side) have expanded the concept so that they are able to explore the hipertextual nature of the Web. The term "metasearchers" is coined to describe the programs able to identify through the network a series of pertinent documents according certain strategy and following links from previously defined websites or "seed sites". The characteristics and practical benefits of main products available are discussed.

Keywords:World Wide Web; information recovery; second generation tools; autonomous agents, multisearchers, metasearchers.

Introducción.

Aunque parece que el crecimiento explosivo de la Internet física (número de ordenadores conectados a la red) se ha reducido hasta situarse "solo" en alrededor de un 31% annual en 1998 [1], el ciberespacio, la Internet de los contenidos, y en especial aquellos accesibles desde la World Wide Web están creciendo hasta volúmenes inusitados.

Las últimas estimaciones hablan de más de 400 millones de páginas Web [2], cifra a la que se llegaría teniendo en cuenta el bajo solapamiento descrito entre los grandes motores de búsqueda y el tamaño real de estos: Altavista superaría los 140 millones de páginas, Northern Light rondaría los 120 millones, mientras que Hotbot también superaría la marca de los 100 millones con unos 110 aproximadamente. Infoseek y Google acercándose a los 60 y Lycos y Excite por encima de los 30 completarían el escenario.

El ya señalado bajo solapamiento entre motores podría implicar que aquellas herramientas que explotan conjuntamente los resultados de varios buscadores pueden estar especialmente indicadas para la recuperación de información. Dichas herramientas, extraordinariamente heterogéneas en cuanto a características, potencialidad y flexibilidad,

reciben el nombre genérico de multibuscadores y agrupan a algunas de las sedes más populares de Internet. En la tabla que figura a continuación señalamos aquellas más potentes según nuestro criterio:

ByteSearch	*www.bytesearch.com*
Chubba	*www.chubba.com*
Cyber411	*ww.cyber411.com*
Debriefing	*www.debriefing.com*
Dogpile	*www.dogpile.com*
Highway 61	*www.highway61.com*
HotOIL	*www.dstc.edu.au/cgi-bin/RDU/hotOIL/hotOIL*
Husky Search	*huskysearch.cs.washington.edu/huskysearch*
Inference Find	*www.inference.com/ifind*
Insane Search	*www.cosmix.com/motherload/insane*
IntelliScope	*wizard.inso.com*
Mamma	*www.mamma.com*
MetaCrawler	*www.go2net.com*
MetaFind	*www.metafind.com*
MetaGopher	*www.metagopher.com*
Ms. DaChanni	*www.mochanni.com/index.en.html*
OnPoint	*www.cs.uchicago.edu/~cooper/onpoint*
Profusion	*profusion.ittc.ukans.edu*
Verio Metasearch	*search.verio.net*

Estas herramientas "server-side" (primera generación) ofrecen resultados ya filtrados, eliminando duplicados e incluso direcciones ya no válidas, utilizando criterios propios o extraídos de la estrategia de búsqueda para clasificar las respuestas. No obstante, dado su gran éxito, se enfrentan a ciertas limitaciones de potencia, lo que significa que para dar servicio a un gran número de usuarios las muestras que ofrecen son limitadas y, aunque precisas, poco exhaustivas. La posibilidad de configurar el tiempo de trabajo que ofrecen algunos de ellos puede aumentar el número de resultados hasta cierto punto.

Esta estrategia basada en la búsqueda en paralelo se ha aplicado con éxito a la llamada **segunda generación de herramientas** [3]. Se trata de programas independientes que el usuario final instala en su propio ordenador ("client-side"), lo que le permite tener un mayor control sobre su funcionamiento y prestaciones. Estos multibuscadores avanzados presentan nuevas y sofisticadas opciones que permiten sobre todo automatizar las tareas de búsqueda contra varias bases de datos simultáneamente. Sin embargo, resulta aún más importante el hecho de que sean capaces de generar muestras completas que solo dependen de los recursos informáticos de quien interroga y de su paciencia para compilar un listado exhaustivo. Además ofrecen la posibilidad de generar "informes" o directamente exportar los registros a bases de datos externas, eliminando respuestas duplicadas y con la opción de "anotar" los recursos con descriptores o un pequeño resumen.

Su número se ha multiplicado considerablemente en los últimos meses y ya existe una amplia variedad con diferentes opciones:

Copernic 98 2.51	Agents Technologies (*www.agents-tech.com*)
CrowCall 1.30	(*www.alphalink.com.au/~pbrooks/CrowCall*)
EZSearch 3.0	American Systems (*www.americansys.com*)
FullFind Pro 3.0	JJ Software (*www.jjsoftware.com/fullfind.html*)
Killer Search 1.12	ADR (*www.killersearch.com*)
Hurricane WebSearch 1.21	Gate Comm (*gatecomm.com/websearch*)
Lazo 2.0	VaultBase (*www.vaultbase.com*)
Mata Hari 1.11	The Web Tools (*thewebtools.com*)
MetaQuest 1.21	RareSpecies (*members.tripod.com/~MetaQuest*)
QueryN MetaSearch 2.2	FreeFlow Software (*www.queryn.com*)
Quest 98 2.1	Inforian Inc. (*www.inforian.com*)
Search Stream 1.5	RobSoft Int. (*www.speed.inter.net*)
SearchWolf 2.03	Trellian (*www.msw.com.au/searchwolf*)
WebSeeker'98 3.4	Blue Squirrel (*www.bluesquirrel.com*)
WebFerret Pro 2.7	Ferretsoft (*www.ferretsoft.com*)
WebStorm 2.5	SharpeWare (*www.sharpeware.com*)
ZurfRider 1.0	Zurf Inc.(*www.zurf.com*)

Sin embargo, todas estas herramientas de recuperación de información se basan en la consulta lineal, sin tener en cuenta que el World Wide Web es, en realidad, un cuerpo documental hipertextual con un alto grado de niveles (profundidad hipertextual). Por esta y otras razones (heterogeneidad formal, carencia de estructuras homologables, inexistencia de controles de forma y calidad) no debe considerarse a la Web como una gigantesca (la mayor) base de datos, por lo que una nueva perspectiva puede ser además de innovadora, muy útil para trabajos de localización que exijan un alto grado de exhaustividad.

Hemos acuñado el término de **metabuscadores** para referirnos a una serie de programas, de características similares a los llamados agentes autónomos (algunos autores los incluyen plenamente dentro de estos), que son capaces, de forma automática, de recorrer la telaraña hipertextual hasta un nivel prefijado por el usuario, examinando las páginas visitadas y contrastando su contenido con la estrategia para valorar su relevancia. Utilizan unas sedes predefinidas llamadas "semilla" desde la que inician la navegación a través de los enlaces que emanan de ellas, de forma que estas nuevas sedes, si son pertinentes, se constituyen a su vez en semilla para el siguiente salto hipertextual.

En muchos casos, los programas utilizan como semilla a uno o varios motores de búsqueda, de forma que la estrategia sirve conjuntamente para la obtención de página que sirvan como punto de partida como para evaluar (filtrar) los resultados a medida que se van volcando. Ello implica que en algunos casos nos encontramos con programas mixtos multibuscadores-volcadores (*downloaders*)-metabuscadores, una asociación que resulta especialmente potente.

Metodología.

Con el fin de proceder a evaluar las capacidades y potencia de estos programas se ha realizado un análisis comparativo de los siete programas (seis metabuscadores y un multi-metabuscador) que hemos podido evaluar porque, en su momento, se encuentraban disponibles bajo la formula “ shareware” que permite su utilización gratuita durante un periodo limitado de tiempo. Aunque algunos de ellos trabajan bajo Windows 3.1, se recomienda utilizarlos en entornos W9x o NT y con equipos holgados de potencia y memoria RAM.

- **Agentware Desktop** (versión 2.1) de la empresa Autonomy Systems ya no se comercializa como producto independiente y por tanto ya no se puede descargar para evaluación desde la sede central de la empresa, aunque se puede obtener de algunas sedes recopilatorias.
- **DigOut4U** (versión 1.4) es un producto francés, que ARISEM (www.arisem.com) distribuye ahora como producto exclusivamente comercial
- **Cybot** (versión 2.4.2) de Virtual Gallery (www.theArtMachine.com/Cybot.htm)
- **MacroBot** (versión 3.03 Pro) producido por Information Projects Group (www.ipgroup.com/macrobot)
- **WebBandit** (versión 3.60) de JW Software Gems (www.jwsg.com)
- **SearchPad** (versiones 1.5 y AI 1.2) de Satyam Spark Solutions (www.searchpad.com)
- **WebWolf** (versión 2.03) de la empresa Trellian (www.trellian.net)

Para comprobar las prestaciones de estos programas se ha construido una estrategia de búsqueda compleja:

"fifth framework program"

que según los casos se ha completado con un amplio número de descriptores adicionales:

"european commission"; "v fp"; "dg xii"; "R&D"; "research and development"; "key actions"

y se ha ejecutado con una profundidad máxima de 10 niveles y/o hasta 24 horas de exploración y se han comparado los resultados. Se ha utilizado, cuando fue posible, como páginas “ semilla”

algunos de los principales motores de búsqueda, aunque no se ha hecho ningún esfuerzo por homologarlos dada la diferente forma de tratarlos de cada programa.

Resultados.

Agentware Desktop

Este programa no se ha podido probar con la estrategia descrita puesto que ha dejado de distribuirse independientemente y ya no hay posibilidad de utilizarlo para evaluación. No obstante se ha valorado en el pasado y como pionero de este grupo de programas merece la pena detenerse en algunas de sus características más relevantes.

Es un programa muy visual, que utiliza un gráfico de un perro para ilustrar los diferentes procesos. La mascota es "entrenada" introduciendo los términos de la estrategia de búsqueda en una pizarra. Al soltarla en el globo terráqueo de la web, inicia la exploración a partir de diferentes sede dejando una "huella" de diferente color según la labor que realice. Finalmente las sedes se ordenan de mayor a menor pertinencia, adjudicándoles un hueso de diferente tamaño.

A pesar de la fuerte carga gráfica del programa el funcionamiento del agente es bastante poco intrusivo y no requiere tantos recursos informáticos como otros, aunque en principio solo se puede ejecutar los agentes de uno en uno. No obstante, la última versión tenía algunas facilidades de recursos compartidos, con agentes que interaccionaban entre sí, aunque está opción no estaba plenamente desarrollada.

El programa generaba una "biblioteca" de recursos que podría llegar a tener un gran tamaño, aunque en un formato propietario.

La característica más relevante de este programa era su capacidad de aprendizaje, pues se podía interrumpir la búsqueda para "re-entrenar" al can, instruyéndole cuando una sede era o no pertinente. Aunque es difícil evaluar la verdadera trascendencia de este mecanismo, si que permitía "podar" ciertas ramas de la exploración ahorrando tiempo y recursos. Esta opción es única en la serie de programas analizados.

Señalar, por último, la capacidad del programa para explorar otras partes del ciberespacio, además del web, lo que puede ser de interés en proyectos concretos.

DigOut4U

Este programa de origen francés incluye un gran número de motores de búsqueda francófonos, así como la posibilidad de configurar aquellos otros que se deseen como paginas semilla. Dado nuestro propósito de comparar resultados, seleccionamos los ocho buscadores señalados en la metodología que se ofrecían en su doble variante de búsqueda simple o avanzada, optando por esta última. Hay que tener en cuenta que este programa multilingüe traduce automáticamente la estrategia al francés, por lo que se desactivó esta "alternativa" para que los motores solo efectuaran la búsqueda original.

Esta herramienta viene preconfigurada para lanzar 5 agentes simultáneamente, opción que no se modificó.

El programa visita y vuelca las sedes para realizar su evaluación. Ello le permite dar al usuario la posibilidad de incluir unos extractos de los contenidos en la exportación de los resultados que se realiza en formato html.

El programa filtra a priori los "hosts" de las paginas semilla, pero dado que muchos buscadores construyen búsquedas complementarias sobre otros webs comerciales (*amazon, barnesandnoble*, etc…) se recomienda identificar estas direcciones indeseadas e incorporarlas al directorio de "hosts" prohibidas antes de lanzar la estrategia.

Tras 24 horas de trabajo (primera semana de febrero 1999) el programa identificó 32.717 sedes, de las que pudo analizar 22.780. Encontró que eran relevantes 9903 (%) de 661 hosts diferentes. El fichero generado ocupa más de 500 Megas, pero se puede consultar con cierta facilidad y los resultados ofrecidos eran pertinentes.

Cybot

El funcionamiento es similar a los anteriores, puesto que además de indicar la estrategia se indican las sedes "semilla" antes de lanzar la búsqueda. Hay que señalar, sin embargo, que el control del filtrado se realiza a priori mediante un sistema de pesos que el usuario puede definir. Este proceso exige cierta práctica porque la valoración final de una sede no solo depende de la suma primaria de pesos, sino que va añadiendo los valores de las sedes "hijas" y "nietas" a medida que se profundiza el análisis. Por eso conviene hacer una primera prueba con distintos valores antes de lanzar definitivamente la búsqueda.

Durante el periodo de prueba el programa identificó 32.314 posibles respuestas, de las que analizó 5366 para llegar a un resultado de 1792 sedes válidas. Estas cifras no son estrictamente comparables con las obtenidas por los otros programas puesto que las configuraciones de filtrado son diferentes, pero la inspección de los resultados revela que son muy pertinentes en la mayoría de los casos.

Un importante aspecto de CyBot es que utiliza Microsoft Access 7 para almacenar la información, de forma que a través de este programa se pueden analizar los datos y generar informes sobre ellos.

Macrobot

En teoría nos encontramos con uno de los programas mas potentes del grupo, puesto que dispone de una potente opción de edición de macros. Aunque se pueden descargar distintos "scripts" de su sede, el control de lo que puede hacerse no es ni fácil ni demasiado potente. Por ello, se ha trabajado en formato automático, ya preconfigurado y que resulta comparable al resto de los programas de la serie.

Admite la posibilidad de introducir varias sedes semilla, aunque en el caso de los buscadores es preferible utilizar la URL larga, es decir con la sintaxis ya especificada. Dispone de una serie de direcciones prohibidas para evitar el "ruido documental", que el usuario puede configurar o incrementar.

Como en algún otro programa de este grupo se puede configurar para recopilar direcciones de correo electrónico con fines de "buzoneo" para marketing.

La base de datos interna está en formato Access lo cual es un importante valor añadido. Sin embargo, la prueba no pudo ser completada ya que a partir de 250 registros el sistema comenzó a colapsarse. No obstante, de los resultados obtenidos se puede indicar que es un producto muy potente, con un alto grado de pertinencia y una notable exhaustividad en el seguimiento de los enlaces. El hecho de que parece utilizar un único agente puede explicar algunas de las limitaciones observadas.

WebBandit

Diseñado originalmente con otros fines, puesto que es capaz de generar registros no solo con datos generales de cada web, sino que recupera explícitamente direcciones postales y correos electrónicos, este potente programa presenta importantes características. Destacaremos, sobre todo, su capacidad para exportar tanto en formato html (con 3-4 líneas de los contenidos de cada web) y Access (con un gran número de campo, incluidos los ya indicados para uso en marketing).

Sin embargo, marra considerablemente a la hora de evaluar los registros dando un bajísimo índice de pertinencia. En el periodo señalado fue capaz de recuperar 2991 registros,

pero la mera inspección de los mismos demuestra una gran heterogeneidad que hace que muchos de ellos no sean válidos.

Recomendable únicamente para proyectos muy específicos, ya que requiere estrategias no ambivalentes o demasiado amplias.

SearchPad

Este programa se presenta en dos versiones ligeramente diferentes, que afectan sobre todo al módulo de evaluación. Al contrario que en otros programas la valoración de pertinencia se hace a posteriori, de forma que tienen que ser volcados los registros antes de clasificarlos. Este segundo módulo es el que distingue ambas versiones, ya que AI se refiere a inteligencia artificial.

Se ha evaluado el primero de ellos, que básicamente es multibuscador similar a otros existentes en el mercado. Una vez obtenidos los resultados, estos se pueden filtrar de acuerdo a criterios de pertinencia definidos por el usuario. Estrictamente no debiera considerársele miembro de esta categoría, pero una posible mayor integración de los módulos en el futuro aconsejan seguir su evolución.

WebWolf

De todos los programas probados este resulta el más sencillo, y aunque sus prestaciones generales son similares a las del resto, apenas ofrece mecanismos adicionales de automatización, filtrado o exportación. Esta carencia de opciones lo hace útil fundamentalmente para estrategias no muy complejas con términos raros (poco frecuentes).

La estrategia única es lanzada contra una semilla definida por defecto, aunque se pude forzar otra elección diferente. La semilla parece ser uno de los grandes motores, por lo que también en este programa existe la posibilidad de excluir la visita de ciertas sedes según dominios prohibidos, lo que indudablemente repercute en la velocidad de la búsqueda y en la pertinencia final de los resultados.

Es posible definir una Biblioteca ("Library") de sedes prioritarias, que también se genera automáticamente con las diferentes búsquedas. Siempre se pueden editar esto registros e incluso restringir la navegación a sólo a las páginas de dichas sedes o a los enlaces emitidos desde las mismas.

En el periodo señalado, trabajando con 10 agentes simultáneamente, el programa identificó 2431 enlaces de los que analizó 2341 y seleccionó 1538 páginas. La gran mayoría de las respuestas son pertinentes a la pregunta realizada, pero el número de sedes distintas es muy bajo. Esta baja diversidad puede ser debida a la imposibilidad de filtrar con los términos adicionales.

Los informes se producen como página html donde es posible navegar con una barra alfabética. Además del nombre y url de la página seleccionada se indica la fuente o página padre desde donde se ha alcanzado.

Es interesante la posibilidad de tener acceso a los enlaces no visitados todavía y a las "páginas" que son ficheros para descarga, que obviamente no puede evaluar pero que están oportunamente separados en un documento aparte. Esta posibilidad de acceso a la "internet invisible" es destacada por los propios fabricantes como uno de los puntos fuertes del programa y nos consta que es inédita para el software que hemos analizado.

Señalar por último la existencia de un buscador que permite localizar textos en los informes y genera un nuevo informe "filtrado", aunque solo de los términos que aparecen en los títulos y URL.

Conclusiones.

Los programas probados han requerido para completar la prueba un volumen elevado de recursos informáticos, habiendo colapsado con frecuencia el Pentium II a 350 Mhz y 64 Mb de RAM utilizado a tal fin. Algunos de ello se han mostrado inestables o incluso incapaces para manejar a posteriori los resultados obtenidos. Ello tiene que ser tenido en cuenta si lo que se pretende es realizar una amplia indagación que pueda generar grandes muestras.

Por contrapartida, todos ellos han sido capaces de generar, dentro de sus diferentes posibilidades, enormes cantidades de resultados. Muchas de las respuestas obtenidas poseen un elevado nivel de pertinencia, incluso sin utilizar a fondo ni los mecanismos de filtrado ni de los de aprendizaje que ofrecen algunos de ellos.

En general, los mejores metabuscadores están cualificados para una muy amplia gama de trabajos documentales, aunque sería deseable la incorporación de características ausentes o poco representadas en el grupo analizado. El bajo coste de la mayoría de ellos ofrecidos bajo el modelo "shareware" les hace especialmente atractivos para tareas documentales de muy diverso orden. Además, algunos permiten exportar directamente los registros a bases de datos, aunque dicha opción ni es universal ni está adecuadamente implementada lo que limita considerablemente su uso.

Entre los aspectos negativos, señalaremos que es particularmente importante la ausencia de mecanismos adecuados de asignación de índices o de generación de resúmenes que podría significar una verdadera revolución en la indización de recursos Internet.

Bibliografía.

1. Lottor, Mark. "Network Wizards Internet Domain Survey. January 1999". http://www.nw.com/zone/WWW/top.html (visitado el 15 de febrero de 1999)
2. Aguillo, I. F. "Searching the Web". http://www.cindoc.csic.es/cybermetrics/links08.html (visitado el 15 de febrero de 1999).
3. Aguillo, I.F. "Herramientas de segunda generación". Anuario SOCADI 1998. Barcelona: Sociedad Catalana de Documentalistas.