Motores de búsqueda de información científica y académica

Lluís Codina

Citación recomendada: Lluís Codina. *Motores de búsqueda de información científica y académica* [en linea]. "Hipertext.net", núm. 5, 2007. <http://www.hipertext.net> [Consulta: 12 may. 2007]. undefined.

1. Introducción

Existe mas de una contradicción cuando se unen en la misma frase las palabras "web" y "ciencia". Por un lado, los contenidos de la Web, no sin algo de razón, siempre han despertado recelos en sectores académicos y profesionales:

1. ¿Quién controla la información que se publica en la Web?
2. ¿Hasta qué punto es fiable la información que encontramos a través de los motores de búsqueda?
3. ¿Se aplican a la Web los controles editoriales propios de las publicaciones impresas que tanto han significado para el progreso de la ciencia?

Ciertamente, no faltan casos de fraudes o de manipulaciones más o menos conocidos en el mundo de la Web, tales como las falsas páginas de La Casa Blanca, la manipulación de los resultados de los motores de búsqueda que hacen los practicantes poco escrupulosos del posicionamiento web (el caso más conocido es el llamado "Gooble bombing" que ha sido erradicado por Google solo muy recientemente). Lo anterior ha generado cosas como la reciente prohibición, por parte de una universidad norteamericana, de que sus estudiantes citen la Wikipedia como fuente para sus trabajos académicos.

A todo ello hay que sumar la dificultad para obtener resultados académicos o científicos cuando se utilizan términos que tienen la misma forma (pero distinto significado) que otros términos propios del comercio o de la cultura popular. Por ejemplo, a alguien muy interesado en la fisiología del sueño le resultará muy difícil encontrar información sobre la fase del sueño denominada *Rapid Eye Movement* y que se conoce internacionalmente como REM, ya que si entra esa expresión en Google solamente encontrará resultados vinculados con el grupo musical REM. La palabra clave "Dolly" proporciona otro buen ejemplo: si alguien está interesado en clonación y quiere informarse sobre el famoso experimento de clonación de la oveja Dolly, es probable que en un motor de búsqueda como Google solamente encuentre información sobre la cantante Dolly Parton.

Lo misma dificultad se puede experimentar si tenemos una necesidad de información cuya palabra clave coincide con palabras presentes en temas discutidos en fórums de Internet. Por ejemplo, si alguien

interesado en encontrar información sobre tarjetas gráficas utiliza palabras clave como ATI o NVIDIA, lo último que encontrará serán análisis técnicos o artículos científicos; en cambio obtendrá toneladas de los típicamente caóticos mensajes en foros de discusión y e interminables listas de precio en sitios como e-Bay.

Sin embargo, pese a todas las dudas, la Web no solamente ha llegado para quedarse, sino para tener también un impacto positivo y real en la difusión del conocimiento académico y científico. Durante años, más o menos desde los noventa hasta nuestros días, una de las soluciones que buscó el mundo académico a esta contradicción consistió en desarrollar y promover directorios, portales y servicios de evaluación, como INTUTE (www.intute.ac.uk).

Con el enorme crecimiento que Internet ha experimentado dese el entonces. el problema que plantean los servicios de información creados y mantenidos "a mano" es que apenas pueden abarcar una parte ínfima de los contenidos reales de la Web. De manera, que la contradicción seguía sin resolverse.

2. Motores académicos

Históricamente, la importante editorial Elsevier fue la primera en detectar que existía una nueva necesidad de información académica en la Web y que, por tanto, se necesitaba una nueva clase de sistemas de información para la Web. En concreto, Elsevier concibió un sistema capaz de indizar páginas web de manera automática, es decir, tal como lo hacen los motores convencionales, pero que fuera capaz de filtrar la información de manera que pudiera ser admisible y fiable para los estrictos criterios del mundo académico.

Ese producto se llamó ***Scirus*** (www.scirus.com) y, al parecer su éxito despertó suficientes recelos en Google para que esta empresa intentara una operación parecida, y así tuvimos unos pocos años después ***Google Scholar*** (scholar.google.com) .

Por imitación (y para suerte del mundo académico) Microsoft no quiso ser menos y, desde inicios del 2007 contamos con un nuevo contendiente en este apasionante campo: ***Live Search Academic*** (academic.live.com).

La característica principal de los tres sistemas es que solamente indizan sitios web vinculados con el mundo académico. Qué se entiende por "mundo académico" cambia en cada caso. La perspectiva que combina, a la vez, rigor y máxima amplitud corresponde sin duda a Scirus. La perspectiva que se ciñe con el máximo rigor, pero en esta caso a costa de la amplitud, corresponde a Live Search Academic y, en alguna posición intermedia, se encuentra Google Scholar.

Con el fin de poder presentar una comparativa entre los tres motores, proponemos la siguiente tipología de documentos académicos:

1. ***Tipo 1:*** Páginas web y documentos de todo tipo (word, ppt, etc.) publicados en sitios de instituciones académicas o científicas (p.e., sitios del tipo .edu).
2. ***Tipo 2:*** Artículos de publicaciones científicas tipo *peer review* , ya se trate de publicaciones open acces,o de publicaciones de pago.
3. ***Tipo 3:*** Trabajos académicos tales como tesis doctorales o tesis de licenciatura.
4. ***Tipo 4:*** Documentos depositados en repositorios científicos (e-prints) ya sean pre-pirnts, post-prints, materiales didácicos, etc.
5. ***Tipo 5:*** Patentes
6. ***Tipo 6:*** Libros (monografías)

Los seis tipos de documentos anteriores se solapan entre ellos. Por ejemplo, algunos repositorios incluyen tesis doctorales (aunque no todos); algunos repositorios han sido creados por asociaciones científicas o por agencias gubernamentales, pero otros creados y mantenidos por universidades y se accede a ellos a través de su sitio web, etc. Pese a todo, la distribución anterior nos será útil aquí para situar en contexto a los motores de búsqueda académicos.

A partir de la clasificación anterior, podemos establecer una tabla como la siguiente para presentar una comparativa de los tres sistemas anteriores en relación la clase de documentos que incluyen (o sea, en relación a sus "inputs"):

Sistema	*Tipo 1*	*Tipo 2*	*Tipo 3*	*Tipo 4*	*Tipo 5*	*Tipo 6*
Scirus	x	x	x	x	x	
Live Search Academic		x				
Google Scholar	x	x	x	x		x

Como se puede observar, de los seis tipos posibles, Scirus y Google Scholar tienen 5 de ellos (aunque no coincidentes): Scirus no tiene libros y, por su parte, Google no tiene patentes. Live tiene solamente uno, mientras que el Tipo 2 (revistas científicas) es, como parece lógico si se mira bien, el único común a los tres motores. En lo que sigue presentaremos con un poco más de detalle cada uno de los tres motores.

2.1. Scirus

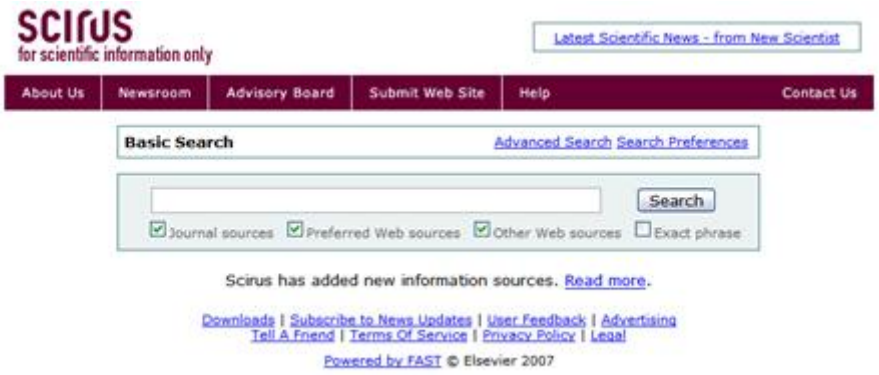

Ilustración 1: La austera pero potente y eficaz pantalla principal de Scirus

2.1.1. Contexto

El motor de búsqueda ***Scirus*** es, como ya se ha apuntado antes, una creación de la importante editorial de revistas científicas holandesa Elsevier (www.elsevier.com) que es parte, a su vez, del gigante editorial anglo-holandés Reed-Elsevier (www.reed-elsevier.com), editor de libros y revistas y productor de bases de datos como Lexis-Nexis.

La cuestión es que Elsevier parece haber comprendido muy bien importantísimo papel que la Web está jugando en la distribución de información académica y dispone de otras dos grandes bases de datos (en este caso y a diferencia de los motores que analizaremos aquí, dirigidas a su utilización en el contexto de bibliotecas universitarias): *Science Direct* (www.sciencedirect.com) y *Scopus* (www.scopus).

Scirus fue fundado en el año 2001 y, poco a poco ha ido ampliando su campo de acción incorporando sucesivamente nuevas fuentes hasta convertirse en un auténtico gigante y en el más completo sistema de los tres (Google Scholar y Windows Live). En un análisis realizado a finales del año 2006 (Jacsó, 2006) se constató que contenía más de 300 millones de documentos (empezó con 50 millones en 2001, de manera que ha multiplicado su contenido por seis desde entonces). Otros dos análisis previos (Giustini y Barksy, 2005; Doldi y Bratengeyer, 2005) confirmaron en su momento que Scirus era, con mucha diferencia más completo que Google Scholar (no existía Live en 2005) por lo que hacía a repositorios científicos del tipo American Physical Society o PubMed.

⟁2.1.2. Inputs

Los inputs de Scirus, es decir, el origen de los documentos que incluye en sus índices son los siguientes (nos guiamos por la propia categorizació de Scirus):

1. *Artículos de revistas:* principalmente, publicaciones académicas de la propia editorial Elsevier (unos 2.000 títulos) más un amplio grupo de publicaciones de tipo *open access* . Son los documentos que Scirus agrupa bajo la denominación ***Journal Sources*** en su página de resultados y la opción del mismo nombre que se puede marcar o desmarcar en su formulario de búsqueda.
2. *Repositorios institucionales o académicos: este apartado incluye repositorios como el de la NASA sobre astronomía o el de la biblioteca de la Cornell University sobre ciencias (física, informática, biología y matemáticas), hasta un total (en teoría) de 18 repositorios, entre los que debemos destacar, además de los mencionados, el de tesis doctorales de la red internacional NDLTD y el de patentes de Lexis-Nexis que incluye patentes de Estados Unidos, Japón y Europa. Decimos "en teoría" porque las pruebas demuestran que en realidad utiliza más repositorios, por ejemplo, hemos podido comprobar que utiliza también E-LIS, un repositorio sobre Bibliteconomía-Documentación que no aparece en la lista "oficial" de fuentes de Scirus. Esta clase de documentos está señala por Scirus bajo la denominación* ***Preferred Web Sources*** .
3. *Páginas y documentos publicados en sitios web:* en este caso se trata excusivamente de servidores de universidades, de instituciones académicas o de departamentos o institutos de I+D de algunas empresas . Desde el punto de vista del dominio, se trata mayoritariamente de sitios del tipo .edu, ac.uk, .gov, etc. Este grupo se identifica en Scirus como ***Other Web Sources.***

⟁2.2. Google Scholar

Ilustración 2: La súper austera interfaz de Google Scholar

⟁2.2.1. Contexto

A estas alturas es difícil presentar a Google. Ha sido la empresa que ha revolucionado de tal manera la búsqueda en la Web que incluso ha acabado afectando a los hábitos de navegación. Por ejemplo, la mayoría de los internautas ya no utiliza los **Preferidos** del navegador: prefiere entrar el nombre de la web en la más famosa caja de búsqueda de la historia. Muchos tampoco entran ya una URL completa si ésta en medianamente complicada. Prefirieren entrar una parte del nombre de la web sabiendo que Google les llevará a ella, probablemente en el primer resultado. Ha empujado a los directorios generalistas, como Yahoo o Dmoz, prácticamente a la clandestinidad y ha barrido a los centenares de directorios nacionales e internacionales que existían antes del 2000. La influencia de Google se ha dejado sentir también en el primer modelo de negocio que ha sido capaz de generar beneficios en la Web: su sistema de anuncios AdWord y AdSense, imitado también por sus competidores.

Por último, prácticamente han creado (u obligado a desarrollar, según se mire) una rama de la matemática: el análisis de enlaces. Lo cierto es que son muchas cosas las que Google ha aportado a la Web. La cuestión es que, en su búsqueda incesante de nuevas actividades (siempre pensado en reforzar su modelo de negocio, no lo olvidemos), desde hace dos años Google se decidió a entrar en el mercado de

los motores académicos y lanzó Google Scholar (Google Académico) con algunas ideas (relativamente) nuevas. La más importante, sin duda, la de llevar a la Web el ***análisis de citaciones*** (por eso decimos que era una idea relativamente nueva).

2.2.2. Inputs

De acuerdo con la documentación oficial (y como es fácil comprobar con un simple test) los *inputs* de Google Scholar consisten en lo siguiente:

1. *Artículos de revistas:* en este caso se trata de artículos de las editoriales académicas que han aceptado formar parte del programa de Google Scholar. En una línea secretista que comienza a ser demasiado característica de Google, no existe una documentación pública (al menos este analista no la ha encontrado) que detalle qué editoriales son en concreto. Mediante pruebas sucesivas es fácil ver que hay una amplia representación de ellas, pero naturalmente, esto no substituye la buena práctica que consistiría en ir publicando periódicamente qué editoriales están en el programa de Google Scholar.
2. *Libros:* al igual que en el caso anterior, se trata de editoriales que han aceptado formar parte de los contenidos de Google Scholar, en este caso, editoriales de libros. Tampoco disponemos de forma pública de una lista de tales editoriales. En todo caso, lo anterior es solamente una de las variedades de esta entrada. La segunda consiste en acuerdos con bibliotecas para obras cuyo derecho de autor haya caducado por haber transcurrido más de los *X* años que cada legislación (la europea, la norteamericana, etc.) establece después de la muerte del autor para que la obra pueda pasar a dominio público. En general, cabe señalar que, en el caso que alguno de los resultados de Scholar sea un libro, el sistema nos remitirá a Google Books para su examen. No obstante, entendemos que debemos incluimos aquí esta categoría documental porque está integrada en las búsquedas de Scholar.
3. *Sitios Web* : Al igual que Scirus, incluye documentos y páginas de sitios web vinculados con el mundo académico. La documentación oficial de Scholar no explica cómo seleccionan estos sitios. Es posible deducir, no obstante, que debe utilizar un sistema similar al de Scirus, a saber, indizar sitio del tipo .edu, etc., sin perjuicio que tengan una lista de URL (sitios) de partida para analizar y a partir de los cuales encuentren otros, etc. En esta categoría, Google Scholar incluye también repositorios de e-prints como los mencionados a propósito de Scirus.

El principal problema de Google Scholar es que no facilita ninguna información precisa sobre sus fuentes concretas. No tenemos una lista ni de editoriales ni de repositorios, ni tampoco una estimación sobre el número de sitios que indizan o sobre el número de documentos que contiene. En su lado positivo, podemos señalar que ha construido su propio índice de impacto, basado en citaciones que se aplica a todos los resultados. De forma que respresenta algo así como la alternativa económica al índice ISI (con muchas menos prestaciones, al menos por el momento).

2.3. Live Search Academic

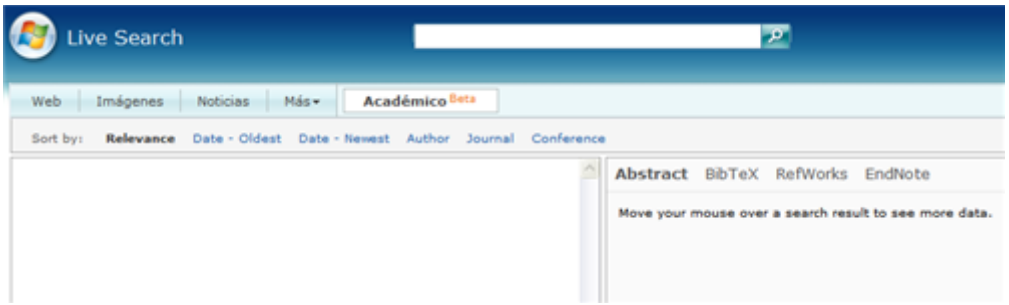

Ilustración 3: Live Search presenta la única interfaz en el mundo de la búsqueda en la Web que no intenta imitar la de Google

2.3.1. Contexto

Microsoft (la compañía propietaria de Live Search) tiene una curiosa historia con la Web: casi siempre llega tarde, pero acaba dominando todo o parte del sector. Les sucedió con los navegadores, con el correo electrónico y les ha sucedido con las búsquedas en la Web. Les ha vuelto a suceder con las búsquedas para objetivos académicos, es decir, en este caso se cumple solamente la primer parte: han llegado tarde. Lo que no sabemos es si acabarán dominado una buena parte del sector, como consiguieron hacer en el caso de los navegadores.

En todo caso, Microsoft es la única empresa del mercado informático que dispone de capacidad tecnológica y financiera suficiente para plantear un desafío creíble al lider actual de las búsquedas generalistas en la web (Google), por un lado, y al líder de las búsquedas académicas por otro (Scirus). Solamente una incomprensible lista de fracasos anteriores de Microsoft en este campo hace difícil pensar en su liderazgo a medio plazo, pese a los medios de que dispone.

2.3.2. Inputs

En el caso de Live Academic, la lista de inputs es simple: artículos de revistas académicas procedentes de diversas editoriales y sociedades científicas ¿Cuáles son estas estas revistas participantes? Por suerte, Live Academics es algo más transparente que Google en este aspecto y proporciona una la lista de lo que denominan " *participating publishers* ". En esta lista aparecen publicaciones como: ACM, Blackwell, Elsevier, Nature, Springer-Verlag y así hasta poco más de cincuenta " *publishers* ". Lo que sucede es que uno solo de estos "publishers" edita hasta 2000 títuos distintos. Lo que no indica aquí Live Academic es cuàntos títulos de estas editoriales incluye, es decir, si incluye todas sus publicaciones o solamente una parte. Las pruebas muestran que, al menos por el momento solo incluye una parte, y no muy amplia, de los títulos de estas editoriales. La lista también también demuestra que su lista no incluye editoriales fuera del ámbito anglosajón. Ciertamente, una búsqueda usando palabras clave en castellano arroja algún resultado, pero siempre corresponde al hecho de que alguna editorial no española, como Elsevier haya publicado alguna vez, casi por casualidad, algún documento en castellano. Nada que ver con el hecho de incluir, por ejemplo, las publicaciones del CSIC o de cualquier otro editor español (en lengua castellana o en cualquier otra lengua)

Si Microsoft piensa tomar en serio su nuevo motor de búsqueda no hay duda que deberá ampliar su lista de "publishers" a varias bandas: editoriales de otros países, pero también mayor número de títulos de cada editorial.

3. Conclusiones

Hay evidencias de que la difusión y, si se nos permite, la promoción del conocimiento, actividad característica de la Documentación, está entrando en una nueva era. Hasta hace poco, la Web había demostrado de sobras su formidable capacidad para actuar como un agente de primer orden en la difusión de la comunicación y de la cultura. Faltaba el elemento de la ciencia y de la información académica.

Contrasta este giro de los motores de búsqueda hacia el mundo académico con su "desentendimiento" del proyecto de la Web Semántica que lleva a cabo el WWW Consortium con un amplio apoyo de instituciones científicas de todo el mundo. No deja de ser curioso que, en esta nueva etapa que está abriendo los motores ninguno de los tres actores (Google, Elsevier, Microsoft) haya considerado incluir alguno de los aspectos de la Web semántica, tales como el uso de ontologías. Tal vez se trate de iniciativas ambas demasiado tempranas como para que puedan pensar en unirse. Probablemente, será necesario que antes maduren cada una de ellas por separado antes de que puedan pensar siquiera en unir esfuerzos. Aún así, es un pena la mutua ignorancia en la que parecen vivir la Web semántica por un lado y los motores de búsqueda por otro.

En todo caso, estas novedades en la búsqueda auguran una nueva etapa en la forma en la cual se gestionará y se difundirán los conocimientos científicos. De momento, las evidencias son muy prometedoras. Corresponde a los documentalistas-bibliotecarios seguir jugando, pero ahora de acuerdo al

nuevo esquema de la Web, el imprescindible papel promotor del conocimiento que nos ha sido siempre tan característico.

4. Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia (España) como parte del proyecto HUM2004-03162/FILO.

5. Bibliografía

Codina, Lluís. (2006). "Motores de búsqueda para usos académicos: ¿Cambio de Paradigma?". *ThinkEPI*, Enero 2006. [Acceso: http://www.thinkepi.net/repositorio/motores-de-busqueda-para-usos-academicos-¿cambio-de-paradigma/]

Doldi, L. M.; Bratengeyer, E. (2005). "The web as a free source for scientific information: a comparison with fe-based databases". Online information review, v. 29, n. 4, p. 400-411

Giustini, D.; Barsky, E. (2005). " *A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations" JCHLA/JABSC, 26, 2005, P. 85-89* . [Acceso: http://pubs.nrc-cnrc.gc.ca/jchla/jchla26/c05-030.pdf]

Grupo Digidoc. *Web semántica y Sistemas de Información Documental* . Acceso: http://www.semanticaweb.net/

Jacsó, Peter. *Péter's Digital Reference Shelf.* December, 2006. [Acceso: http://www.gale.com/reference/peter/]

Rovira, Cristòfol; Marcos, Mari-Carmen; Codina, Lluís (2007). "Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código". *El Profesional de la Información* , v. 16, n. 1, enero-febrero 2007. [Acceso en: http://eprints.rclis.org/archive/00008668/]