



CRÍTICA DE LIBROS / BOOK REVIEWS

Academic search engines: a quantitative outlook

José Luis Ortega

Elsevier Science, Amsterdam, 2014, 198 págs. ISBN 978-1-84334-791-0 (print)

El autor es José Luis Ortega, investigador del CSIC y miembro del Laboratorio de Cibermetría, que cuenta con una importante trayectoria investigadora con publicaciones nacionales e internacionales en el ámbito de la cibermetría, la visualización de información y el análisis de redes. La obra está escrita en un inglés claro y sencillo y su título refleja de forma precisa su contenido: los motores de búsqueda académicos.

El libro hace una revisión de los motores de búsqueda académicos más importantes, ofreciendo para cada uno de ellos una visión global de cuál es el tamaño de su índice, qué funcionalidades tienen, cuales son las ventajas y las posibles limitaciones de cada servicio e incluso una revisión crítica de la fiabilidad para los servicios más importantes que ofrecen cada uno. El autor realiza también un ejercicio de ingeniería inversa para intentar medir de forma cuantitativa, la cobertura y la importancia de las distintas fuentes de información en cada motor de búsqueda.

Los principales motores de búsqueda académicos analizados son: CiteSeerX, Scirus, AMiner, Microsoft Academic Search y Google Scholar, a los que se dedica un capítulo a cada uno. A estos motores hay que añadir también: BASE, Q-Sensei Scholar y WorldWideScience, que son tratados con menor profundidad en un solo capítulo (otros motores de búsqueda académicos) por las características particulares en las que quiere fijarse el autor en cada uno de ellos. BASE porque está especialmente orientado a repositorios institucionales, Q-Sensei por su herramienta de filtrado y WorldWideScience por su modelo de motor de búsqueda federado alimentado por bases de datos bibliográficas. A estos seis capítulos de motores de búsqueda académicos hay que sumar un capítulo introductorio al principio y otros dos, uno de análisis comparativo de los motores y otro de conclusiones al final.

El primer capítulo está dedicado a definir el concepto de qué son los motores de búsqueda académicos y las características que lo diferencian

respecto a otros motores: el autor explica que básicamente son servicios basados en web y que a diferencia de otros servicios deben prestar especial atención a la tipología de documentos (artículos, comunicaciones a congresos, etc.) y a la transparencia y origen de las fuentes de donde extraen sus datos porque para los usuarios de este servicio, los científicos, consumidores pero también creadores de esta información, la fuente de donde salen estos datos (una revista con factor de impacto frente a otra de bajo impacto) o la tipología de los mismos puede ser igual o más importante incluso que la relevancia con la que el motor de búsqueda puede juzgar su contenido. Este capítulo dedica también una interesante y condensada historia de la evolución de los motores de búsqueda académicos, desde 1997 hasta nuestros días. Y finaliza con un epígrafe dedicado a las perspectivas futuras centrándose en las nuevas funcionalidades que empiezan a aparecer en este tipo de motores, con especial énfasis en los indicadores bibliométricos referidos a científicos e investigadores así como a las posibilidades de implementación de funcionalidades de redes sociales o Web 2.0 y las posibilidades que presentan para poder trabajar con otras métricas (altmetrics).

Los capítulos del 2 al 7 están dedicados al análisis más o menos pormenorizado de los motores de búsqueda académicos, de acceso libre y basados en web, más importantes: CiteSeerX, Scirus, AMiner, Microsoft Academic Search Engines y Google Scholar y de forma más breve de BASE, QSensei Scholar y WorldWideScience. Aunque el análisis de cada motor es diferenciado, la estructura de estos capítulos intenta responder a cuestiones similares:

- Una breve presentación de la historia de cada motor de búsqueda, qué organización lo soporta y en términos generales qué y cómo construye sus índices (cómo es el proceso de recolección -crawling- y extracción de datos y qué servicios proporciona.

- Una estimación crítica de la cobertura, del tamaño, de su base de datos.
- Una descripción de sus registros, de la información que recoge el motor de búsqueda de cada documento y/o una exhaustiva explicación de las funcionalidades más importantes de cada motor.
- Y una más que interesante revisión crítica de las principales virtudes pero, sobre todo, de los principales errores y/o carencias detectados por el autor respecto a cada uno de los motores. Para ello, JL Ortega realiza distintos ejercicios de recolección de datos para detectar documentos recogidos en el índice pero borrados (se pueden buscar pero no encontrar porque han sido borrados) o para descubrir registros incluídos con errores respecto a la fuente de información o a la fecha de publicación, como hace en el caso de CiteSeerX.

En el penúltimo capítulo el autor realiza un análisis comparativo de los distintos motores de búsqueda centrado en cuatro aspectos fundamentales:

- El funcionamiento técnico de cada motor con unas más que interesantes tablas comparativas sobre los medios de obtención y acceso a los datos de cada motor de búsqueda y en el que se hace una valoración de las ventajas e inconvenientes de cada uno, refiriéndose a si los datos se obtienen de fuentes secundarias o si se hace un trabajo específico de descubrimiento y recolección (crawling) de datos o si en el acceso a esos datos prima la búsqueda o se facilita la navegación a modo de directorio y aportando ejemplos claros sobre algunos de ellos.
- La estructura de cada motor de búsqueda, refiriéndose a las entidades que recoge cada motor para cada objeto, en algunos casos primando los autores, las revistas u otras fuentes, las organizaciones a las que pertenecen los autores o si ponen a disposición de los usuarios cierta clasificación temática.
- La cobertura, o el tamaño de la base de datos de cada motor, a lo que el autor otorga una gran importancia puesto que a mayor cobertura, mayor tamaño de la base de datos y por tanto, mientras más documentos puedan ser recuperados mayor es la probabilidad de satisfacer las necesidades de los usuarios.
- Y las funcionalidades del interfaz de búsqueda, referido a si los motores ponen a disposición de los usuarios interfaces de búsqueda avanzada, uso de operadores booleanos, posibilidades de ordenación de resultados, límites a los registros que se pueden exportar o formatos de exportación de resultados (APA, MLA, etc.).

El último capítulo recoge algunas conclusiones respecto a lo difícil que resulta realizar una definición integradora de los motores de búsqueda académicos, más allá de que todos ellos contienen información científica y están basados en web, puesto que sus funcionalidades y aplicaciones difieren considerablemente. El autor considera que todos estos motores tienen aún mucho margen de mejora y muchos retos técnicos que superar para ser capaces de extraer citas, autores y organizaciones de los documentos que descubren y recolectan. Y, sobre todo, mucho margen de mejora respecto a la transparencia de las fuentes de información que utilizan, a su frecuencia de actualización y a la tipología de documentos con las que trabajan, lo que hace que aún no puedan ser consideradas como instrumentos del todo fiables para la evaluación de la investigación de autores u organizaciones. El libro finaliza con una concisa, actual y muy pertinente bibliografía y un útil índice de los términos más importantes utilizados en el libro.

No es fácil escribir un libro sobre motores de búsqueda, se trata de un sector tan dinámico que cualquier intento puede dejar la letra impresa del libro con datos desactualizados en cualquier momento. Le pasó al autor, quien ya en el prefacio hace saber al lector que uno de los motores analizados, Scirus, fue cerrado durante el proceso de edición final del libro. O cuando en su crítica a la tardanza en la actualización de Google Scholar Metrics se refiere a que la fecha del último índice era de abril de 2012 y actualmente esa fecha es de junio de 2014. ¿Significa eso que el libro ya está desactualizado y sus contenidos son poco novedosos o aportan poco a la literatura científica? En absoluto, se trata de un libro que recomendaría, sin ninguna duda, a los investigadores y a los estudiantes interesados en la comunicación científica en general. Y en mi opinión sus contenidos seguirán siendo válidos en varios años puesto que incluso aunque veamos una profunda transformación en su estructura, tamaño y funcionamiento el libro permitirá conocer la evolución de estos motores de búsqueda.

El libro está escrito de forma clara y sencilla y el autor aporta numerosos ejemplos para explicar y demostrar algunos de los problemas que presentan estos motores. Y el uso que hace de tablas e ilustraciones es también muy adecuado. Pero quizás el mayor y mejor valor de la obra estriba en el espíritu crítico, en el mejor sentido del término, con el que el autor aborda los problemas e inconvenientes de cada motor. Su uso de la ingeniería inversa le permiten poner de manifiesto, por ejemplo, el problema de CiteSeerX de obtener referencias de bibliografías y presentarla en la página de resultados (figura 2.4) o nos permite saber que Scirus sólo clasificaba un tercio de todos sus documentos, bien porque no tenían categorías para dichos documentos o porque el sistema no la asignaba por cualquier otra razón. Otro tanto ocurre con AMiner al que critica que siendo una plataforma centrada en el autor el sistema indexe solo documentos. O

la crítica a Microsoft Academic Search por su lentitud a la hora de actualizar los índices. O a las inconsistencias de Google Academic en el número de resultados que ofrece. No son las únicas críticas, el autor no se limita a describir el funcionamiento de los distintos motores sino que entra a hacer una valoración. Y en esas valoraciones sobre los servicios, positivas y negativas, el autor introduce multitud de ideas sobre cómo deberían evolucionar estos servicios.

Lo que menos me ha gustado: los apartados referidos al tamaño y cobertura de los motores de búsqueda, especialmente el del capítulo 8, representan un intento de estimación del tamaño de los índices y bases de datos de estos motores. No es fácil encontrar en la literatura científica trabajos que permitan establecer una comparación y el autor tampoco especifica claramente la metodología por lo que es difícil valorar si las estimaciones se acercan a la realidad, máxime en un entorno tan variable en donde de un día para otro pueden verse indexados miles de nuevos documentos por acuerdos, por ejemplo, con empresas editoriales. Y, por

último, si como decía Umberto Eco, el lector tiene la responsabilidad de actualizar los textos del autor, se echa de menos algún apartado que hiciese referencia a servicios como Mendeley, ResearchGate o Academia.edu. Es verdad que se puede discutir si son o no son motores de búsqueda académicos, pero el ecosistema del descubrimiento y el acceso a la información científica es lo suficientemente complejo para que el tema hubiese merecido algo más que una pequeña mención en la página 134, máxime cuando alguno de estos servicios empieza a convertirse en una fuente importante para las llamadas altmetrics.

En definitiva, un libro muy recomendable si está interesado en conocer cómo funcionan y qué pueden ofrecer los motores de búsqueda académicos, que cumple con los objetivos que se marcaba el autor, escrito de forma clara, cuyos contenidos se adecúan a su título y que trata el tema con rigor y profundidad.

Tony Hernández-Pérez

Universidad Carlos III de Madrid