RU&SC Revista de Universidad y Sociedad del Conocimiento Vol. 3 - N.º 1 / Abril de 2006

www.uoc.edu/rusc ISSN 1698-580X

Indicadores web para medir la presencia de las universidades en la Red

Isidro F. Aguillo
Begoña Granadino

Resumen

La cibermetría es una disciplina emergente que utiliza métodos cuantitativos para describir los procesos de comunicación en Internet, los contenidos en la Web, sus interrelaciones y el consumo de esa información por parte de los usuarios, la estructura y la utilización de las herramientas de búsqueda, Internet invisible o las particularidades de los servicios basados en el correo electrónico.

La presencia de las instituciones académicas, y muy especialmente de las universidades, en la Web puede generar información muy útil para la evaluación de sus actividades académicas y de investigación, incluyendo no sólo las que generan producción formal, por medio de artículos y de publicaciones, sino también las que transmiten conocimiento de manera más informal.

Se distinguen tres grandes grupos de indicadores web para el análisis cibermétrico: medidas descriptivas, que miden el número de objetos encontrados en cada una de las sedes web (páginas, ficheros media o ricos, densidad de enlaces); medidas de visibilidad e impacto, que cuentan el número y el origen de los enlaces externos recibidos, como el famoso algoritmo PageRank de Google, y medidas de popularidad, donde se tiene en cuenta el número y las características de las visitas que reciben las páginas web.

Datos empíricos obtenidos para dominios web universitarios muestran que la cibermetría es una interesante herramienta para describir la presencia en Internet de instituciones académicas, pero también evidencia la llamada brecha digital, que puede conducir a un indeseable colonialismo cultural y científico.

Palabras clave

universidades, cibermetría, indicadores web, comunicación científica, Open Access

Abstract

Cybermetrics is an emerging discipline that uses quantitative methods to describe communication processes on the Internet, web contents, their interrelations and consumption of this information by users, the structure and use of search tools, invisible Internet, and the special features of services based on electronic mail.

The presence of academic institutions, and especially that of universities, on the web generates highly useful information for evaluating their academic and research activities, including not only formal activities, through articles and other publications, but also those that transmit knowledge through more informal means.

There are three major groups of web indicators for cybermetric analysis: descriptive measures, which measure the number of objects found in each of the websites (pages, media or rich files, mean number of links); measures of visibility and impact, which count the number and source of external links, such as Google' famous PageRank; and popularity measures, which calculate the number and characteristics of the different visits to web pages.

Empirical data obtained for university web domains show that cybermetrics is an interesting tool to describe the presence of academic institutions on the Internet but that it also shows the so-called digital gap, which could lead to undesirable cultural and scientific colonialism.

Keywords

universities, cybermetrics, web indicators, scientific communication, Open Access

www.uoc.edu/rusc ISSN 1698-580X

INTRODUCCIÓN

En los últimos años hemos asistido a un notable interés en la evaluación de la actividad científica que, poco a poco, se ha ido generalizando a todos los ámbitos de la estructura académica-investigadora. La necesidad de controlar el gasto público, de racionalizar el esfuerzo investigador y de premiar a los investigadores y a los centros de excelencia son objetivos que sólo pueden cubrirse con un conocimiento preciso de la producción y de la productividad de profesores y científicos.

El proceso de evaluación se ha afrontado desde dos vías complementarias: una primera basada en la opinión de expertos, generalmente pares reunidos en comités, donde el consenso diluye los efectos de la subjetividad, y una segunda, basada en técnicas cuantitativas, generalmente bibliométricas, que, además de una cierta objetividad, permite su aplicación a amplios colectivos dada su mayor viabilidad técnica.

La bibliometría ha demostrado ser válida para la medición de los resultados formales de la actividad investigadora, generalmente artículos publicados en revistas de prestigio y monografías especializadas. El análisis de citas ha proporcionado una herramienta eficaz para la evaluación de esa producción, especialmente útil para identificar la élite del sistema. Sin embargo, la dependencia directa de estas técnicas de las bases de datos bibliográficas de citas producidas por el ISI ha dado lugar a ciertos problemas derivados de los sesgos de éstas. Citaremos entre los más destacados la cobertura diferencial geográfica, temática y lingüística de las fuentes, que determina que exista un mayor peso de las revistas publicadas en los países desarrollados, en inglés y del área de las ciencias puras sobre las sociales, humanas o tecnológicas.

Para contrarrestar dicha situación se requiere en la actualidad que otros muchos aspectos de la actividad sean tenidos en cuenta, de forma que, en el caso de un docente, la comunicación informal, la dirección formativa o la divulgación sumen en la evaluación personal. Desafortunadamente sólo a costa de un gran esfuerzo y en casos muy controlados es posible extender esta rigurosa y exhaustiva colecta a grandes instituciones.

Una posible alternativa pasaría por incrementar sustancialmente la presencia de estas actividades en un medio público que las aglutinara y que fuera objeto de análisis unitario. Dicho medio ya existe y se ha convertido en el principal canal de comunicación científica, aunque todavía existen reticencias sobre su uso y debería potenciarse aun más la publicación académica mediante éste. Se trata del soporte electrónico, y más concretamente de la Web, cuya ubicuidad, accesibilidad, asequibilidad, sencillez y potencia abren considerablemente las posibilidades de la comunicación universitaria.

La cibermetría es una disciplina emergente que, a partir de las técnicas y el modelo bibliométrico, pretende extender la aplicación de los métodos cuantitativos a la descripción de los procesos de comunicación científica en Internet, a la determinación del volumen y de la tipología de los contenidos académicos en la Web, y a tratar de desentrañar las interrelaciones sociales y el consumo de información por parte de los usuarios. Otros aspectos también susceptibles de estudio cibermétrico son la descripción de las herramientas de búsqueda en la Web, la llamada Internet invisible o las particularidades de los servicios basados en el correo electrónico y en los foros personales.

La herramienta fundamental son los llamados indicadores, que pueden utilizarse de forma combinada con los equivalentes bibliométricos y que, al igual que éstos, se utilizan para describir distintos aspectos de los procesos de comunicación académica y científica. En este trabajo se presentan indicadores web diseñados para medir la presencia de universidades o centros de enseñanza superior y que se han utilizado en la elaboración de distintos trabajos comparativos.

www.uoc.edu/rusc ISSN 1698-580X

METODOLOGÍA

Unidades

El primer problema al que nos enfrentamos es la identificación de la unidad de trabajo. En la Web las unidades lógicas están subordinadas al sistema físico de almacenamiento, que se ve reflejado más o menos en la nomenclatura de las direcciones de Internet. La URL suele definir unívocamente una página, pero sus componentes fuertemente jerarquizados también pueden referirse a una serie o a un conjunto de ellas, formando una sede, o en el caso de reflejar una gran institución, agrupando varias sedes en un dominio institucional.

Así, la mayoría de las páginas web de la Universidad Complutense se agrupan bajo el dominio «ucm.es», aunque eventualmente algunos grupos, proyectos o congresos tengan nombres independientes bajo dominios internacionales (org, com o info). En Yahoo Search existe un delimitador que nos permite conocer el número de subdominios de la forma «xxx.ucm.es» que se engloban dentro de uno dado (fig. 1).

Otro problema que debe tenerse en cuenta son los *alias* o dominios alternativos, que generalmente sólo afectan a la sede principal. Así, la Universitat Oberta de Catalunya puede localizarse tanto con el dominio «uoc.edu» como bajo «uoc.es». Afortunadamente estos casos son pocos y van solucionándose en los últimos años, aunque fueron especialmente molestos en el caso de las universidades del Reino Unido.

Herramientas

La medida de los componentes de una sede web exige la utilización de un programa que visite el servidor correspondiente e indexe sus contenidos. Este programa, llamado robot o rastreador *(crawler)*, es el componente principal de los motores de búsqueda. Aunque pueden diseñarse o reutilizarse robots personales, su uso es complejo, difícil de interpretar y conflictivo a la hora de analizar sedes ajenas. Por ello, a pesar de las limitaciones y los sesgos de los motores de búsqueda, prefieren

FIGURA 1. Delimitador de subdominios en Yahoo. Estrategia de búsqueda: site:ucm.es feature:index

utilizarse éstos, que no sólo son de manejo más simple, sino que ofrecen una cobertura más universal, prácticamente global, del Webespacio.

Frente a lo que habitualmente se cree, el número de motores con bases de datos propias, independientes, que ofrezcan una cobertura alta de los contenidos web, es en realidad muy reducido. Si además excluimos aquellos que no permiten el filtrado mediante delimitadores, encontramos que sólo cinco son útiles para fines cibermétricos:

– Google (www.google.com)
– Yahoo Search (search.yahoo.com)
– MSN Search (search.msn.com)
– Ask (www.ask.com)
– Exalead (www.exalead.com)

Hace ya tiempo que se sabe que ninguno de los motores cubre de forma exhaustiva la totalidad de la Web, y que las causas de esta cobertura incompleta son difíciles de solucionar. De hecho, los estudios muestran que el solapamiento entre las diferentes bases de datos no es elevado y que por ello resulta recomendable el uso en combinación de varios motores para el cálculo de los indicadores.

Sesgos y limitaciones

Ninguno de los motores ofrece más allá de los primeros mil resultados de una búsqueda, por lo que hay que recurrir al número de resultados que indica el motor como valor de referencia. Esta cifra suele ser representativa del total real, pero suele ofrecerse redondeada o

aproximada, lo que en la práctica supone una tasa de error no inferior al 3%.

Los motores están sujetos a determinantes comerciales, especialmente la garantía del servicio. De esta forma, cuando el servicio se satura, los resultados que se ofrecen son aproximaciones más groseras, que infravaloran el valor real. Otras veces el servicio se ofrece desde *data centres* (servidores de la base de datos en otros lugares) alternativos, cuyos contenidos pueden ser ligeramente distintos.

La recolecta de datos también plantea problemas, de forma que cada uno de los robots tiene su propia idiosincrasia: los hay que no exploran en profundidad ciertas sedes, los que no actualizan la base de datos con la frecuencia adecuada y los que, encontrando problemas de navegación, provocan sesgos, fundamentalmente geográficos. Parece demostrada la cobertura diferencial negativa de servidores asiáticos y africanos por parte de ciertos motores.

Indicadores de contenido

Los principales indicadores son los que describen el volumen de contenidos publicados en la Web. Pueden medirse el número y el tamaño de los objetos informáticos encontrados en cada una de las sedes, pero el segundo dato resulta poco útil porque depende de factores ligados al formato y no al contenido.

El número de páginas html o asimiladas (páginas dinámicas, ficheros ricos, ficheros de texto) puede calcularse con el delimitador «site:», que es útil en todos los buscadores citados excepto en Ask, donde requiere añadir *inurl:dominio* (fig. 2).

En el ámbito académico, la utilización de ciertos formatos documentales para la comunicación científica sirve para derivar indicadores más ajustados de los contenidos. Los llamados ficheros ricos (doc, pdf, ps, ppt) pueden recuperarse directamente de algunos motores de búsqueda y están ligados a actividades de publicación

FIGURA 2. Sintaxis combinada de Ask.
Estrategia de búsqueda: site:uoc.edu inurl:uoc.edu

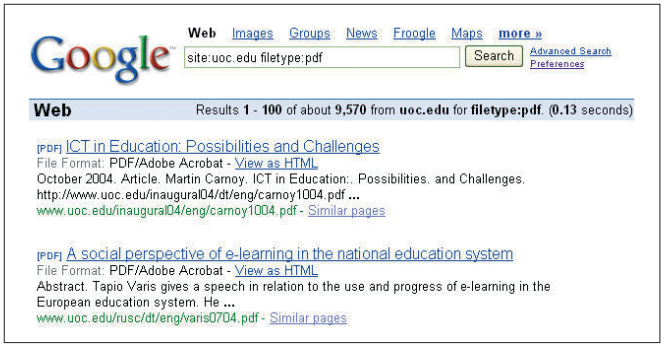

FIGURA 3. Obtención de ficheros ricos en Google.
Estrategia de búsqueda: site:uoc.edu filetype:pdf

(el ps o Postscript es el formato estándar para físicos, matemáticos o ingenieros) o comunicación (el Powerpoint o ppt es el más popular para presentaciones en congresos o transparencias para el aula). En Google el delimitador utilizado es *filetype:* (fig. 3).

Indicadores de visibilidad y de impacto

El carácter hipertextual de la Web ha llevado a muchos autores a homologar la cita bibliográfica con los enlaces web. Aunque las motivaciones para establecer un enlace, incluso en el mundo académico, son más ricas y variadas que las que justifican una cita, las técnicas de análisis de citas pueden aplicarse a la descripción del escenario global.

La medida de visibilidad viene dada por el número de enlaces externos (de terceras sedes) recibidos por un dominio. Desafortunadamente Google no calcula enlaces por dominio, por lo que podrá utilizarse tanto Yahoo como MSN Search, que comparten la misma estrategia básica (fig. 4).

FIGURA 4. Visibilidad o enlaces externos recibidos.
Estrategia de búsqueda: linkdomain:ub.es -site:ub.es

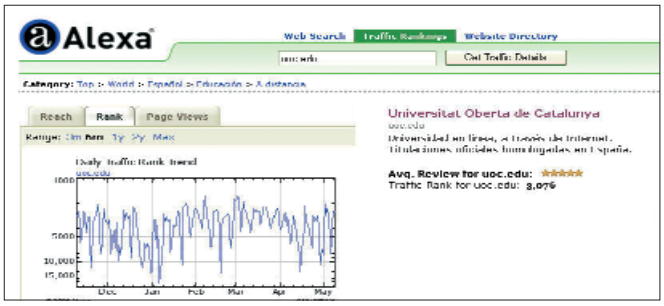

FIGURA 5. Popularidad relativa según Alexa.
Estrategia de búsqueda: uoc.edu

En el pasado se utilizaba el llamado factor de impacto Web (WebIF), que se obtiene como cociente del número de enlaces entre el número de páginas de una sede o de un dominio. Por distintas razones, este índice da lugar a numerosos artefactos matemáticos, por lo que ha dejado de utilizarse.

Una alternativa, bastante difícil de calcular, es el índice que se construye de acuerdo al peso relativo de las sedes de origen de los enlaces: es el famoso PageRank de Google, que puede obtenerse de la barra de navegación de este motor, pero que no permite una segregación eficaz de valores (rango de números enteros entre 0 y 10).

Indicadores de popularidad

El consumo de información puede medirse contando el número y describiendo las características de los visitantes y las visitas que recibe una sede. Esto es notablemente difícil de realizar porque sólo pueden obtenerse estos valores cuando se tiene acceso a todos y cada uno de los ficheros log de cada uno de los servidores.

Alternativamente, resulta más simple obtener los valores relativos proporcionados por el buscador Alexa (www.alexa.com), que intercepta visitas en todo el mundo y establece a partir de ahí un *ranking* de popularidad. El valor proporcionado, posición en el ámbito mundial, puede utilizarse en estudios comparativos regionalizando el análisis, ya que hay ciertos sesgos geográficos (fig. 5).

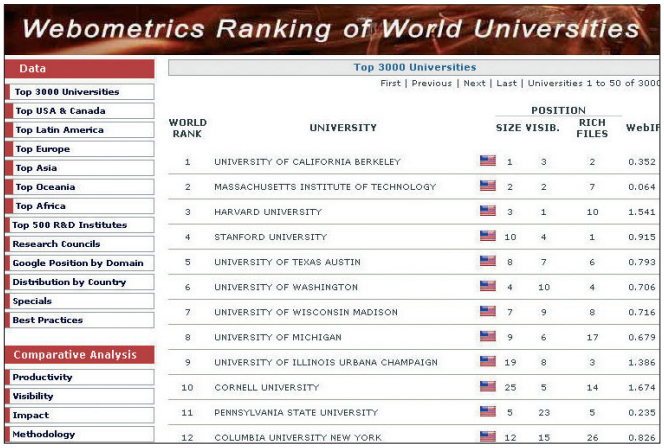

FIGURA 6. Ranking mundial de universidades.

RESULTADOS Y APLICACIONES

Los indicadores descritos se han utilizado para construir el *Webometrics ranking of world universities* (www.webometrics.info), donde, a partir de datos obtenidos de un total de más de diez mil universidades de todo el mundo, se ha procedido a seleccionar las tres mil primeras de acuerdo a un indicador combinado llamado *Webometrics rank* (WR) (fig. 6).

El WR es un indicador que combina la visibilidad y el tamaño de una forma similar al WebIF, pero dando más peso al primer elemento en una proporción 4:3 frente al 1:1 del WebIF. Además, reconoce la importancia de los ficheros ricos como vehículo documental de la actividad académica e investigadora. La formula de cálculo es:

WR: 2*S + 4*V + R

donde todos los valores son rangos calculados de valores normalizados obtenidos de los motores de búsqueda: *S* corresponde a la mediana entre los valores de tamaño de Google, Yahoo, MSN Search y Ask; *V* es la visibilidad mediante la combinación de enlaces en Yahoo y MSN Search, y *R* son los ficheros ricos obtenidos con Google.

El análisis de los resultados muestra, como cabría esperar, que las grandes universidades estadounidenses aparecen en las primeras posiciones. Sin embargo, muchas otras instituciones de este país se muestran abrumadoramente entre las primeras clasificadas y relegan a países (Francia, Italia, Japón) con una fuerte tradición académica e investigadora a posiciones más retrasadas.

Esta «brecha digital», no ligada a condicionantes económicos sino de política científica, gestión de la investigación, y actitudes y comportamientos personales, es especialmente preocupante. Aunque el idioma puede jugar un papel relevante, hay que señalar que, al contrario que en Europa, los profesores e investigadores de Estados Unidos publican libremente sus actividades en la Web y participan más en las iniciativas de Open Access. A medio y largo plazo, la ausencia de contenidos específicamente propios y la incapacidad de nuestros investigadores de comunicar globalmente sus resultados por medio de la Web pueden dar lugar a un colonialismo cultural y científico sobrevenido, y dificultar los procesos de innovación, con el impacto industrial y económico que ello conlleva.

BIBLIOGRAFÍA

AGUILLO, Isidro F. (2002). «Measuring informal scientific publication in the Web». En: *EASST 2002 Conference. International Conference of the European Association for the Study of Science and Technology*. Universidad de Cork (Reino Unido).

AGUILLO, Isidro F. (2005). «Indicadores de contenidos para la web académica iberoamericana». BiD: Textos Universitaris de Biblioteconomia i Documentació [artículo en línea]. N.° 15.
<http://www2.ub.edu/bid/consulta_articulos.php?fichero=15aguil2.htm>

AGUILLO, Isidro F.; GRANADINO, Begoña; LLAMAS, Germán (2005). «Posicionamiento en el Web del sector académico iberoamericano». *Interciencia*. Vol. 30, n.° 12, pág. 1-5.

AGUILLO, Isidro F. [*et al.*] (2005). «Medida de la actividad y comunicación científica mediante indicadores cibermétricos». En: *I Jornadas de Indicadores para la Evaluación de la Ciencia y la Tecnología*. Madrid.
<http://www.cindoc.csic.es/info/fesabid-prog.html>

AGUILLO, Isidro F. [*et al.*] (2005a). «What the Internet says about science». *Scientist*. Vol. 19, n.° 14, pág. 10.

BAR-ILAN, Judit (2005). «Expectations versus reality — Search engine features needed for Web research at mid 2005». *Cybermetrics*. Vol. 9, n.° 1.
<http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html>

BJORNEBORN, Lennart; INGWERSEN, Peter (2004). «Towards a basic framework for webometrics». *Journal of the American Society for Information Science and Technology*. N.° 555, pág. 1.216-1.227.

HARNAD, Stevan; BRODY, Tim (2004). «Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals». *D-Lib Magazine*. Vol. 10, n.° 6.
<http://www.dlib.org/dlib/june04/harnad/06harnad.html>

SWAN, Alma (2005). *Open Access self-archiving: An introduction*. Truro: Key Perspectives.
<http://eprints.ecs.soton.ac.uk/11006/01/jiscsum.pdf>

THELWALL, Mike (2003). «Web use and peer interconnectivity metrics for academic Web sites». *Journal of Information Sciences*. Vol. 29, n.° 1, pág. 11-20.

WILKINSON, Davil [*et al.*] (2003). «Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication». *Journal of Information Science*. Vol. 29, n.° 1, pág. 59-66.

Para citar este documento, puedes utilizar la siguiente referencia:

AGUILLO, Isidro F.; GRANADINO, Begoña (2006). «Indicadores web para medir la presencia de las universidades en la Red». En: ROCA, Genís (coord.). *La presencia de las universidades en la Red* [monográfico en línea]. *Revista de Universidad y Sociedad del Conocimiento (RUSC)*. Vol. 3, n.° 1. UOC. [Fecha de consulta: dd/mm/aa].
<http://www.uoc.edu/rusc/3/1/dt/esp/aguillo_granadino.pdf>
ISSN 1698-580X

www.uoc.edu/rusc ISSN 1698-580X

Isidro F. Aguillo

Laboratorio de Internet (Cindoc-CSIC)

isidro@cindoc.csic.es

Trabaja en el Laboratorio de Internet en el Centro de Información y Documentación Científica (Cindoc) del Consejo Superior de Investigaciones Científicas (CSIC). Realiza tareas relacionadas con el desarrollo de indicadores de la sociedad de la información, análisis documental de recursos web, cibermetría y procesos de comunicación científica por la Red. Dirige o participa en varios proyectos de I+D de la Unión Europea y del Plan Nacional de Investigación Científica. Ha sido miembro de la Oficina Española de Ciencia y Tecnología (SOST) en Bruselas y *Metcalfe visitor professor* en la Universidad de Nueva Gales del Sur (Sídney, Australia).

Licenciado en Biología por la Universidad Complutense de Madrid y máster en Información y documentación por la Universidad Carlos III de Madrid. Edita la revista electrónica *Cybermetrics* desde 1997, es miembro del Comité Asesor del *Profesional de la Información* y de comités científicos de diversos congresos nacionales e internacionales, y participa como evaluador y revisor de proyectos de investigación europeos.

Begoña Granadino Goenechea

Laboratorio de Internet (Cindoc-CSIC)

bgranadino@cindoc.csic.es

Doctora en Ciencias Biológicas (1986), y científica titular del CSIC (2000). Desde 2003 desarrolla su labor de investigación en el Cindoc, en el ámbito de la Cienciometría y la Cibermetría.

Sus trabajos tienen por finalidad contribuir al análisis de la producción científica y tecnológica, fundamentalmente en ciencias de la vida, ciencias ambientales y biotecnología, así como al desarrollo de técnicas cuantitativas para la descripción y la evaluación de los contenidos en Internet en el área de la actividad académica y de la investigación científica-técnica.

Edición: FUOC. Producción: Ediciones Doyma, S.L.
www.doyma.es