F.1. La necesaria evolución de la cibermetría

Isidro F. Aguillo

12 mayo 2011

Aguillo, Isidro F. "La necesaria evolución de la cibermetría".
Anuario ThinkEPI, 2012, v. 6, pp. 119-122.

Resumen: Se explica la génesis, relaciones disciplinares y principales aplicaciones de la cibermetría y la webometría, para luego pasar a discutir la situación actual respecto a los métodos utilizados hasta la fecha para la extracción de información e indicadores web de fuentes de acceso público. Se realiza un análisis comparado de tres técnicas para medir el prestigio, impacto o visibilidad de instituciones académicas, que incluyen: 1) la encuesta, limitada por su subjetividad y pequeños tamaños poblacionales, 2) el análisis de citas, también lastrado por poblaciones reducidas pero que ofrece más fiabilidad por tratarse de opiniones de pares, y 3) el análisis de enlaces. Se considera el análisis de enlaces la piedra angular de la webometría y se valora los grandes números involucrados, defendiéndose que ello permite identificar patrones a pesar de ser el método más "ruidoso" debido a la riqueza y diversidad de las motivaciones para vincular. Se informa de la desafortunada desaparición de la principal fuente de información sobre enlaces, el servicio Yahoo Site Explorer, lo que obligará a un replanteamiento de las técnicas y la búsqueda de fuentes alternativas de este valioso tipo de datos. Se estudian distintas alternativas como el uso de menciones, tales como títulos, nombres de instituciones o personas, y la llamada mención o cita URL, discutiendo sus ventajas e inconvenientes. Se señala que esta nueva aproximación es especialmente útil para las herramientas de la Web 2.0.

Palabras clave: Cibermetría, Webometría, Análisis de enlaces, Buscadores de búsqueda, Menciones, Menciones de urls.

Title: The necessary evolution of cybermetrics

Abstract: The origins, disciplinary relationships and main applications of cybermetrics and webometrics are explained, in order to proceed to a discussion of the current situation regarding the methods used for extracting web information and indicators from publicly available sources. We performed a comparative analysis of three techniques to measure the prestige, impact or visibility of academic institutions, including: 1) the survey, limited by its subjectivity and small sample sizes, 2) citation analysis, also hampered by small populations but which offers more reliability because it relies on peer opinions, and 3) link analysis. Link analysis is considered the cornerstone of webometrics and the large numbers involved is to be appreciated. It can identify patterns despite being the method that is most "noisy" because of the richness and diversity of motivations for linking. The unfortunate demise of Yahoo Site Explorer, the main source of information on links, is reported and its absence will force a rethinking of the techniques and a search for alternative sources of this valuable type of data. Some alternatives are studied, including the use of mentions such as titles, names of institutions or individuals and so-called reference or quote URLs, and their advantages and disadvantages are discussed. It is noted that this new approach is especially useful for web 2.0 tools.

Keywords: Cybermetrics, Webometrics, Links analysis, Search engines, Mentions, Url mentions.

1. Introducción y definiciones

La cibermetría, que incluye a la webometría (término que remarca las raíces comunes con la bibliometría y que, aunque malsonante para algunos, es preferible al de webmetría, puesto que éste define internacionalmente a otra área diferente), es una especialidad emergente dentro del grupo de las llamadas ciencias cuantitativas que ahora globalmente se definen como informetría.

La cibermetría se dedica al análisis de la presencia en la web de los procesos de creación (métodos, herramientas, estructuras, modelos, series de datos) y comunicación de conocimiento académico y científico, tanto formal (revistas electrónicas) como informal (todo lo que genera esa actividad que no acaba en forma de artículo, capítulo de libro o monografía), lo que proporciona un nuevo punto de vista en el estudio y evaluación de universidades y organizaciones dedicadas a la investigación, pero también de grupos y

científicos y profesionales individuales. Aunque había algunos precedentes, la disciplina nace a mediados de los años 90 gracias a los trabajos de un grupo de investigadores que incluye –entre otros– a **Ingwersen** (1997, 1998), **Rousseau** (1997), **Aguillo** (1998), **Bar-Ilan** (1999), **Smith** (1999) y **Thelwall** (2002, 2009). Existe una revista científica, con evaluación por pares, que desde 1997 viene recogiendo trabajos sobre este tema. Se trata de *Cybermetrics* que se publica exclusivamente en formato electrónico y en inglés.
http://cybermetrics.cindoc.csic.es

2. Métodos

Una de las causas del éxito de la cibermetría radica en la sencillez de sus métodos. De forma sucinta, podemos resumir diciendo que los datos se recogen de la llamada web pública, es decir sólo se consideran aquellos contenidos que son abiertamente accesibles en las webs. Entre esos datos podemos incluir, por ejemplo, páginas web, enlaces (hipervínculos), ficheros ricos o multimedia, o entradas en redes sociales y otros servicios de la web 2.0.

"La cibermetría es una especialidad emergente dentro del grupo de las ciencias cuantitativas que se definen globalmente como informetría"

La recogida de datos se puede hacer directamente, mediante robots especialmente diseñados para dicha tarea, o indirectamente extrayendo la información de las bases de datos de los buscadores comerciales (*Google*, *Yahoo!* o *Bing*). Mientras que la programación de los robots puede ser dificultosa y requerir para su funcionamiento importantes recursos humanos y de cómputo, los buscadores disponen de sus propios robots, que son más potentes y ofrecen una cobertura mucho más amplia. Aunque la opacidad del funcionamiento de los buscadores (funcionan con algoritmos que son secretos comerciales) y su comportamiento irregular o impredecible han sido objeto de crítica, su papel en los procesos de comunicación es fundamental. Efectivamente, no se trata de meros intermediarios sino que hoy son el principal mecanismo de visibilidad de los contenidos web. Como se ha comentado alguna vez, lo que no está en *Google* no existe.

Hasta la fecha la herramienta más poderosa de la cibermetría era el análisis de enlaces que, dada la naturaleza hipertextual de la Web, suponía una forma práctica de descubrir patrones entre sedes web, interconexiones entre instituciones o relaciones entre temas. Para entender la importancia del análisis de enlaces hay que poner en contexto este método con los otros habituales en informetría. Aunque la definición de calidad es compleja y suscita mucho debate, la informetría ha utilizado una aproximación transaccional: se estima que una actividad o resultado es de calidad, tiene impacto o alcanza gran visibilidad cuando su popularidad medida en términos de indicadores cuantitativos indica la existencia de un consenso al respecto en una comunidad. En términos prácticos hay tres grandes métodos:

"Hasta la fecha, la herramienta más poderosa de la cibermetría era el análisis de enlaces"

Encuesta

Se solicita a un grupo de pares que valoren, por ejemplo, las publicaciones de una institución, una revista o un científico. El número de opiniones recogidas es muy pequeño, pero provienen de un grupo de expertos de reconocido prestigio. Este método es práctico para microanálisis, pero está sujeto a sesgos fruto de incompatibilidades no reveladas, y es inválido para evaluar universos muy amplios donde difícilmente se encontraría una persona con el conocimiento global requerido (por ej., producción editorial mundial, ranking de universidades…).

Análisis de citas

Al igual que en el caso anterior, se recaba la opinión de pares, pero mediante un método indirecto, contando las citas recibidas. Las citas bibliográficas entre trabajos científicos se han utilizado como indicador de relevancia o impacto y han resultado especialmente prácticas a nivel meso (evaluación de revistas e instituciones) o macro (políticas nacionales o regionales), pero dado los bajos números involucrados (unas pocas docenas) y la baja capacidad de segregación de algunos de sus indicadores (por ejemplo los valores enteros del índice h) resultan inapropiadas para evaluaciones individuales generalizadas. Pero su limitación más importante es que trabajan sobre un universo cerrado, el de los trabajos formalmente publicados en revistas científicas, lo que en términos prácticos sólo representa una pequeña parte de la actividad de científicos o profesores (especialmente cierto en muchas disciplinas) y de su impacto académico, económico o socio-cultural.

Análisis de enlaces

El conteo de enlaces se realiza sobre el webespacio, por definición un universo mucho más abierto y menos estructurado, aunque también sorprendentemente muy auto-organizado. Las motivaciones para enlazar son mucho más diversas y aunque se incluyen auténticas citas ("*sitations*"), también hay razones espurias detrás de ciertas ligas. Sin embargo, las cifras involucradas son enormes, a menudo del orden de millones, y la ley de los grandes números nos informa de la capacidad discriminante de las mismas y de las posibilidades estadísticas de extraer patrones significativos a pesar del enorme ruido existente. Este referéndum virtual incluye a "terceras partes", actores relevantes para cualquier sistema científico que no son académicos, pero que forman parte de una comunidad diferenciada. En este sentido hay que diferenciar enlaces (visibilidad hipertextual) de visitas (popularidad), ya que sólo los editores web pueden enlazar, mientras que cualquier internauta puede realizar una visita.

El análisis de enlaces es una de las claves del éxito de *Google,* ya que su algoritmo *PageRank* organiza las páginas web según un indicador ponderado de los enlaces que reciben. Esto es también relevante para el desarrollo de la cibermetría, pues *Google* utiliza como unidad la página y sólo ofrece información de enlaces recibidos página a página. Dado el crecimiento explosivo de la Web, se hacían inviables los estudios de enlaces, inter-enlaces y co-enlaces con dicho buscador.

"El papel del análisis de citas debe ser ahora asumido por el análisis de menciones, una técnica prometedora"

Hasta 2011 esto no era un problema práctico, pues varios buscadores ofrecían la posibilidad de recolectar los enlaces a dominios o subdominios completos. En los años 90 el favorito era *Altavista*, y ya en el siglo XXI se utilizaban *Yahoo!* (que había comprado *Altavista*), *Bing* (la última encarnación de los buscadores de *Microsoft*) o *Exalead* (un pequeño buscador francés con bastantes sesgos) que ofrecían estos servicios.

Esa información era valiosa también para el posicionamiento de páginas web (*search engine optimization*, SEO) en los resultados de los buscadores, y posiblemente haya sido el abuso de ciertos SEO lo que llevó a *Bing* a suspender esta opción. En 2010, *Bing* llegó a un acuerdo con *Yahoo!* por el cual la base de datos de *Microsoft* sustituiría a la propia de *Yahoo!*, perdiendo así sus operadores específicos. Durante cierto tiempo *Yahoo!* mantuvo *Site Explorer*, pero este servicio cerró a finales de 2011.

3. Nuevos métodos

La pérdida de los operadores de *Yahoo!* ha obligado a una evolución tanto teórica como metodológica de la cibermetría. El papel del análisis de citas debe ser ahora asumido por el análisis de menciones, una prometedora técnica que ya había sido señalada por varios autores (**Aguillo**, 2009; **Thelwall**, 2009), aunque **Blaise Cronin** ya la describía como "invocación" hace más de una década (**Cronin** *et al.* 1998).

Sin abandonar los buscadores de búsqueda, el método consiste ahora en extraer no enlaces sino términos o frases y evaluar su presencia de forma cuantitativa.

"En el marco del proyecto europeo *OpenAire* se están elaborando indicadores web para medir el impacto de los trabajos depositados en repositorios"

Así se puede preguntar por un autor, el nombre de una institución, el título de un trabajo, un código o directamente un url. Esto plantea nuevos problemas, aunque algunos como la homonimia es vieja conocida de los bibliómetras. El nombre "José Gómez" es difícilmente útil en este contexto, incluso filtrando por dominio institucional; "Universidad Complutense" no da un resultado exhaustivo pues a menudo la encontraremos como "Complutense University". Los acentos y otros caracteres no convencionales pueden generar también problemas, pues se puede comprobar que los buscadores dan respuestas distintas para todas las variantes.

En el marco del proyecto europeo *OpenAire* (*Open access infrastructure for research in Europe*) se están elaborando indicadores web para medir el impacto de los trabajos depositados en repositorios. Presentamos a continuación algunos resultados preliminares de los métodos.
http://www.openaire.eu

a) Títulos. En la mayoría de los trabajos científicos suelen tener una gran longitud, lo que reduce las probabilidades de generar ruido. El título ha de ir siempre entre comillas (operador de adyacencia estricta). Cuando el número de términos es bajo, se puede añadir el apellido del primer

http://www.openaire.eu

autor. Es interesante destacar que se puede buscar tanto en el cuerpo de la página web como en el título (<TITLE>) con operadores específicos. Si hay dos versiones del título (original y traducido), se pueden combinar utilizando el operador OR, aunque hay que tener en cuenta las limitaciones de los buscadores, especialmente *Google* cuando se utilizan operadores booleanos.

b) URLs. En el caso de los repositorios se da la circunstancia de que hasta tres tipos de direcciones pueden referirse al mismo documento: la de la página del registro, la del fichero con el documento a texto completo (pdf o similar) y la del redireccionador o *handle*.

c) DOIs. Se va imponiendo poco a poco puesto que sólo se utiliza cuando el trabajo ha sido formalmente publicado en una revista. Una precaución al utilizar estas técnicas es la de excluir siempre las automenciones, utilizando expresiones del tipo "–site:urlrepositorio".

El análisis de menciones se puede generalizar a otras fuentes de información tales como noticias, blogs, wikis, redes sociales, etc. Las precauciones descritas son igualmente aplicables. Estas técnicas alternativas a las bibliométricas y cibermétricas clásicas se engloban en la llamada altmetría (ver *altmetrics.org*).

Posteriormente a esta nota, **Thelwall** y **Sud** (2011) publicaron un artículo en la línea de lo aquí presentado, que incluye soporte empírico para alguno de los indicadores.

En resumen, la cibermetría es una disciplina científica que evoluciona, que lleva a cabo estrategias viables frente a los problemas y cuyo impacto en nuestra actividad no ha hecho más que empezar.

4. Referencias bibliográficas

Aguillo, Isidro F. "STM information on the Web and the development of new internet R&D databases and indicators". En: D. Raitt, (ed.). *Procs of Online information 98*. Oxford: Learned Information, 1998, pp. 239-243.

Aguillo, Isidro F. "Measuring the institution's footprint in the web". *Library hi tech*, 2009, v. 27, n. 4, pp. 540-556. *http://dx.doi.org/10.1108/073788309*

Almind, Tomas C.; **Ingwersen, Peter**. "Informetric analyses on the world wide web. Methodological approaches to 'webometrics'". *Journal of documentation*, 1997, v. 53, n. 4, pp. 404-426. *http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/jdocAlmind1997.pdf* *http://dx.doi.org/10.1108/EUM0000000007205*

Bar-Ilan, Judith. "Search engine results over time - a case study on search engine stability". *Cybermetrics*, 1999, v. 2, n. 1, paper 1. *http://cybermetrics.cindoc.csic.es/articles/v2i1p1.pdf*

Cronin, Blaise; **Snyder, Herbert W.**; **Rosenbaum, Howard**; **Martinson, Anna**; **Callahan, Ewa**. "Invoked on the Web". *Jasis*, 1998, v. 49, n. 14, pp. 1319-1328. *http://dx.doi.org/10.1002/(SICI)1097-4571*

Ingwersen, Peter. "The calculation of web impact factors". *Journal of documentation*, 1998, v. 54, n. 2, pp. 236-243. *http://dx.doi.org/10.1108/EUM0000000007167*

Rousseau, Ronald. "Sitations: an exploratory study". *Cybermetrics*, 1997, v. 1, n. 1, paper 1. *http://cybermetrics.cindoc.csic.es/articles/v1i1p1.pdf*

Smith, Alastair G. "A tale of two web spaces; comparing sites using web impact factors". *Journal of documentation*, 1999, v. 55, n. 5, pp. 577-592.

Thelwall, Mike. "An initial exploration of the link relationship between UK university web sites". *Aslib procs*, 2002, v. 54, n. 2, pp. 118-126. *http://cybermetrics.wlv.ac.uk/paperdata/2002_An_initial_exploration_of_the_link_relationship.pdf*

Thelwall, Mike. *Introduction to webometrics: quantitative web research for the social sciences*. New York, NY: Morgan & Claypool, 2009. *http://dx.doi.org/10.2200/S00176ED1V01Y200903ICR004*

Thelwall, Mike; **Sud, Pardeep**. "A comparison of methods for collecting web citation data for academic organizations". *Jasist*, 2011. v. 62, n. 8, pp. 1488-1497. *http://dx.doi.org/10.1002/asi.21571*