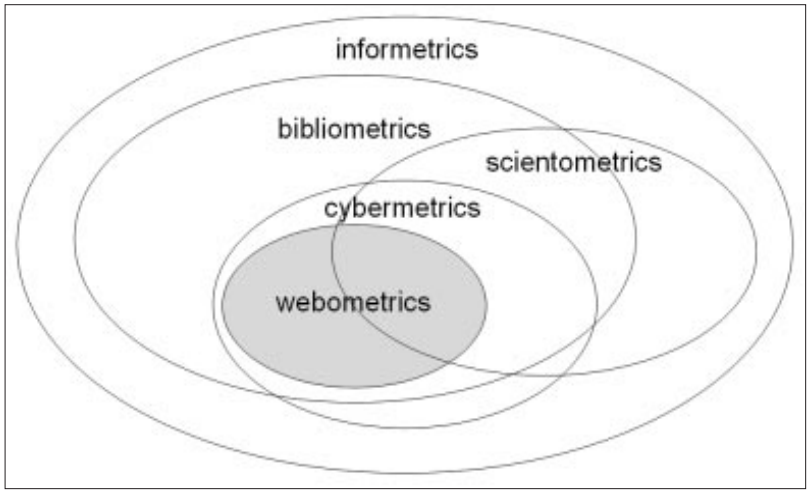
Análisis exploratorio de los enlaces de la Universidad Complutense de Madrid

JUAN-ANTONIO MARTÍNEZ-COMECHE[1]
UNIVERSIDAD COMPLUTENSE DE MADRID, ESPAÑA

INTRODUCCIÓN

Este estudio parte de la consideración de la Webmetría como un área de investigación integrada por completo dentro del campo de la Cibermetría, que a su vez se origina a raíz de los principios y métodos de la Bibliometría. Este enfoque ha sido desarrollado más en profundidad por Björneborn e Ingwersen (2004), y resumido en la siguiente gráfica:



Fuente: Björneborn; Ingwersen (2004)

1 Agradezco la participación de la profesora Michela Montesi en los comienzos de este estudio.

Empezando por el término más general, Informetría (informetrics), esta puede definirse como el estudio de los aspectos cuantitativos de la información en cualquiera de sus formas, no solamente registros o bibliografías, y en cualquier grupo social, no solamente entre los científicos (Tague-Sutcliffe, 1992).

El siguiente término, más restringido, es Bibliometría, que puede definirse a su vez como el estudio de los aspectos cuantitativos de la producción, diseminación y uso de la información registrada (Tague-Sutcliffe, 1992).

La Cienciometría (*scientometrics)* posee una relación parcial con las anteriores en cuanto afronta el estudio de los aspectos cuantitativos de la ciencia como disciplina, pero también acoge el estudio de la ciencia como actividad económica o política (Tague-Sutcliffe, 1992).

Desde su aparición, la Web ha sido una vía más de comunicación académica tanto formal como informal, que introducía novedades importantes con respecto a hábitos anteriores (el acceso a un volumen creciente de información digitalizada, o el uso de herramientas de búsqueda de información inexistentes hasta entonces). Diversas han sido las denominaciones empleadas para el estudio de esta nueva vía de comunicación, inicialmente de carácter académico, pero entre ellas (netometrics, internetometrics) destacan dos: Webmetría (webometrics) y Cibermetría (cybermetrics). Actualmente se establece una diferencia entre ellas (Björneborn, 2004):

- La Cibermetría, más general, se ocupa del estudio de los aspectos cuantitativos de la construcción y uso de fuentes de información, estructuras y tecnologías en Internet en su totalidad, partiendo de una aproximación bibliométrica e informétrica.
- La Webmetría, más restringida, se ocupa del estudio de los aspectos cuantitativos de la construcción y uso de fuentes de información, estructuras y tecnologías en la Web, partiendo de una aproximación bibliométrica e informétrica.

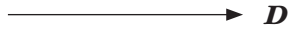
Conforme a esta definición, atañe especialmente a la Webmetría:

- El análisis del contenido de las páginas web
- El análisis de la estructura de enlaces de la Web

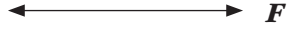
La Web puede ser analizada como un grafo dirigido, en cuanto que consta de nodos (entendiendo por nodo desde una página web hasta dominios de primer nivel correspondientes a países o sectores) y de conexiones entre ellos realizadas mediante enlaces desde un nodo saliente hasta otro entrante. La terminología más frecuente en relación a los enlaces web incluye las siguientes denominaciones:

- Enlace entrante (inlink) en ***A***: un nodo web (página, dominio, etc.) ***A*** recibe un enlace procedente de otro nodo web ***B***:

- Enlace saliente (outlink) de ***B***: un nodo web ***B*** envía un enlace hacia otro nodo web A.
- Auto-enlace (self-link): Un enlace enviado desde un cierto nodo web al mismo nodo
- Co-enlaces (co-link): dos nodos web ***C*** y ***D*** son enlazados simultáneamente desde un nodo ***B*** (nodos ***C*** y ***D*** coenlazados)



- Enlace recíproco: un nodo ***E*** enlaza hacia ***F***, y al tiempo ***E*** recibe un enlace desde el nodo ***F***:



La Web académica ha sido objeto de numerosos estudios desde los inicios de la Webmetría, quizá porque la Web nació precisamente con una finalidad académica. Sin embargo, no abundan los trabajos cuyo objeto de estudio sean los dominios académicos en español, y más en concreto, los procedentes de España. En consecuencia, este estudio preliminar aborda el análisis webmétrico de la Universidad Complutense de Madrid, poniendo el énfasis en la naturaleza y características generales de su red de enlaces.

Nos hemos centrado en tres aspectos básicos:

- ❖ Conocimiento de los principales dominios de primer nivel (TLD) con los que enlaza la UCM
- ❖ Conocimiento de las principales regiones del mundo con las que enlaza la UCM
- ❖ Clasificación temática de los principales dominios enlazados

Metodología

Para conocer los enlaces que establece el dominio de la Universidad Complutense de Madrid (http://*.ucm.es) debe emplearse un programa crawler o rastreador de la Web. En este estudio se ha empleado el programa crawler de código abierto Nutch. Nutch se programó para que efectuase un rastreo de los enlaces salientes (outlinks) de la UCM. De esta manera, se obtuvo un corpus inicial de aproximadamente 1,000,000 de páginas de la UCM que enlazaban con páginas ajenas a la propia Universidad Complutense. Ese millón de páginas de la UCM establecían enlaces con un total de 9899 dominios externos. Estos 10,000 dominios externos enlazados desde la UCM presentan una distribución muy conocida en la Web, de manera que la inmensa mayoría de estos dominios reciben un único enlace desde la UCM, mientras que, si vamos aumentando el número de enlaces recibidos, va disminuyendo el número de dominios que reciban ese número de enlaces cada vez más elevado. En nuestro estudio decidimos analizar únicamente los dominios que recibían 10 enlaces o más. De esta manera, aunque tenemos en cuenta solamente 741 dominios de los 10,000 iniciales, acogemos en nuestro estudio más del 90% del número total de enlaces salientes desde la UCM. A su vez, nos interesaba restringir el corpus a aquellos dominios que mantuviesen una relación lo más estrecha posible con la Universidad Complutense, lo que suponía centrarnos en los dominios que enlazasen simultáneamente con la UCM. De estos 741 dominios nos quedamos finalmente con 355 dominios que no solamente recibían más de 10 enlaces desde la UCM, sino que simultáneamente enlazaban con la Universidad Complutense.

Una vez obtenida la muestra final de 355 dominios, el estudio webmétrico afrontado contempla los dos aspectos clásicos de este tipo de estudio:

- El análisis del contenido de los dominios
- El análisis de las características de los enlaces

En cuanto a las características de los enlaces, se efectuó inicialmente un análisis de los dominios de primer nivel (.es, .com, etc) que presentaban esos 355 dominios. Sin embargo, es sabido que en la Web los dominios de primer nivel (Top Level Domain o TLD) no siempre nos informan del lugar en donde se han desarrollado las páginas web.

Con objeto de conocer mejor las zonas de influencia prioritarias de la UCM, se optó por realizar un análisis manual de los 355 dominios involucrados, tratando de averiguar en qué país se habían desarrollado las páginas del dominio correspondiente. En aquellas ocasiones en las que técnicamente no ha sido posible averiguar dicha localización geográfica, se ha impuesto un localización por defecto (el país donde está registrada la matriz de una compañía, y en última instancia el propio TLD).

En cuanto al análisis de contenido de los dominios, dicho estudio exigía la adopción de una clasificación temática de las páginas. En relación a este aspecto, son muy diversas las clasificaciones empleadas en estudios anteriores. Entre ellas podemos citar las categorías principales del Open Directory Program (dmoz.org) o la tipología empleada en la Dublin Core Initiative.

Se ha preferido partir de una clasificación muy genérica de sitios web que contemplase pocas categorías, pero al mismo tiempo especializada en el campo de la web académica, con el objeto de precisar mejor las relaciones mantenidas desde la UCM con el exterior. Se ha partido, pues, de la clasificación desarrollada en su momento por Vaughan (Vaughan, 2007), que tiene la ventaja añadida de incorporar categorías muy próximas a las de los Dominios de Primer Nivel, lo que permitirá efectuar posteriormente una comparación de los resultados obtenidos con ambos enfoques.

A las 5 categorías empleadas por Vaughan (sitios educativos, organizativos, gubernamentales, comerciales y personales), hemos tenido que efectuar algunas adaptaciones. Se ha añadido una categoría más (sitios Web 2.0), que abarca desde blogs hasta redes sociales. Esta nueva categoría viene impuesta por el indudable crecimiento de este tipo de dominios en la Web. Por otra parte, dado el número apenas existente de webs de carácter personal en la muestra considerada, se ha modificado su denominación por el de "Recursos documentales", más numeroso, que abarca cualquier tipo de repositorio de documentos digitales.

Resultados

Los Dominios de Nivel Superior (.es, .com, etc.) correspondientes a los 355 dominios considerados se pueden resumir en la siguiente gráfica:

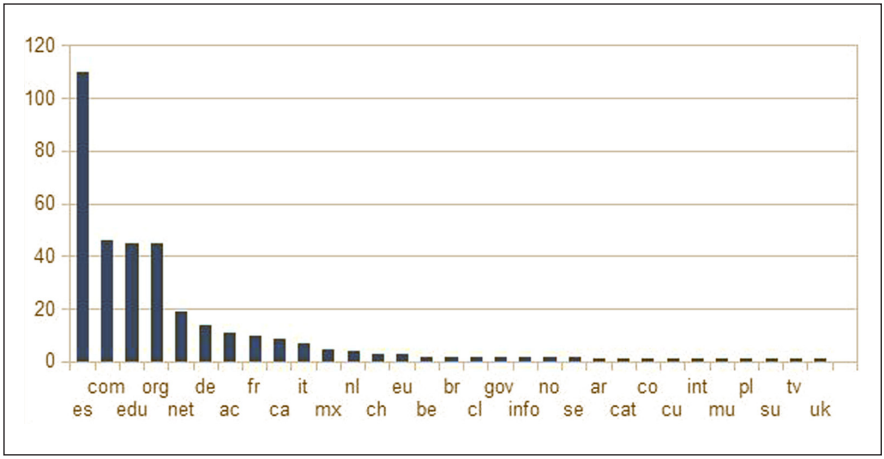


Fig. 1. Dominios de Primer Nivel co-enlazados con la UCM

Como era de prever, el dominio más frecuente corresponde a España (un 31% del total), pero en segundo lugar figuran los dominios de carácter comercial (.com), educativo (.edu) y organizativo (.org) con frecuencias prácticamente iguales, que en conjunto

suman un 38'3% del total. A partir de ahí obtenemos una larga cola donde se mezclan los dominios genéricos (.net, .ac, .info) con los dominios de carácter geográfico (.de, .fr, .it, .mx).

En relación a la distribución de estos enlaces por regiones, los datos pueden resumirse en la siguiente gráfica:

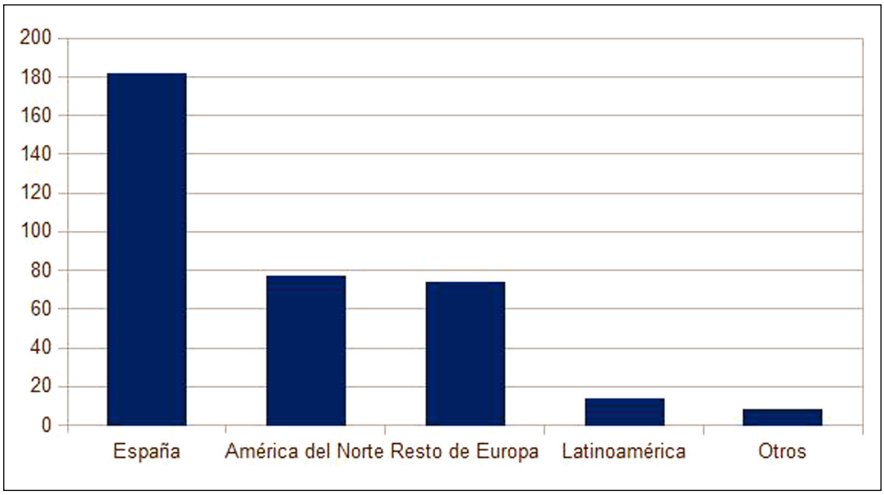


Fig. 2. Regiones co-enlazadas con la Universidad Complutense de Madrid

Destaca sobremanera el número de enlaces de la Universidad Complutense con dominios españoles (un 51% del total), dato que avala, al menos parcialmente, la falta de internacionalización de las webs universitarias españolas que se ha detectado en estudios precedentes (Thelwall y Aguillo, 2003). En efecto, la UCM adolece de una relativamente escasa conexión con Europa, hecho corroborado por la presencia de un porcentaje prácticamente igual de co-enlaces europeos y norteamericanos (21'7% de co-enlaces con América del Norte y un 20'8% de co-enlaces con Europa).

La clasificación de los 355 dominios enlazados con la UCM, conforme a las seis categorías consideradas (sitios web educativos, organizativos, gubernamentales, comerciales, "Web 2.0" y "Recursos documentales") se puede resumir en la gráfica siguiente:

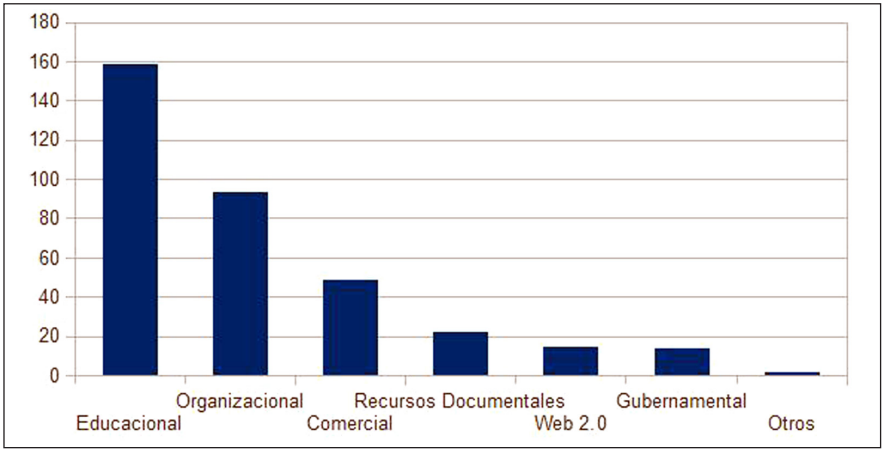


Fig. 3. Clasificación de los dominios co-enlazados con la UCM

En primer lugar, destacar que esta clasificación ratifica los resultados obtenidos mediante el análisis de los Dominios de Primer Nivel, pero al mismo tiempo los complementa, ofreciéndonos una información más detallada sobre la naturaleza de los enlaces de la UCM. En efecto, con ambos procedimientos obtenemos que los dominios de carácter comercial, educativo y organizativo predominan claramente. Pero mientras el análisis de TLD's mostraba unos porcentajes iguales entre ellos, con el análisis clasificatorio somos capaces de discernir qué tipología de sitios web predominan en los enlaces que mantiene la UCM. Como es lógico, entre estas tres categorías principales predominan los sitios web de carácter educacional e investigador (un 44'8% del total), seguidos de los sitios de naturaleza organizacional (26'5%) y de los sitios con una finalidad comercial (13'8%). Esta gradación es lógica en una institución universitaria como la UCM, cuya prioridad es la formación y la investigación, al tiempo que muestra una relación cada vez más intensa con el mundo empresarial.

En segundo lugar, conviene destacar la presencia de categorías no consideradas en estudios precedentes. Los enlaces entre la UCM y los sitios web correspondientes a repositorios y fuentes documentales de todo tipo, por una parte, y la aparición de enlaces con

sitios web característicos de la Web 2.0 por otra (blogs fundamentalmente, pero también otros medios sociales), demuestran que el desarrollo vertiginoso de la Web tiene repercusiones inmediatas en el estudio de cualquier tipo de enlaces en Internet, incluyendo las universitarias.

Conclusiones

En relación a la obtención de estadísticas relativas a los Dominios de Primer Nivel o TLD, se concluye que ofrece una información incompleta, dada la propia naturaleza de los TLD. Una clasificación manual que tenga en cuenta dichas categorías, además de otras nuevas, permite discernir con mayor precisión la importancia relativa de dominios tipo .edu, .org y .com

En relación a las clasificaciones temáticas de la Web, se concluye que es necesario efectuar correcciones y ampliaciones en las clasificaciones previamente existentes. Ello es debido al carácter altamente dinámico y cambiante de la propia Web, que provoca la aparición de categorías novedosas y cambios profundos en las previamente consideradas.

En cuanto a la procedencia geográfica de los dominios más estrechamente relacionados con la UCM, destaca la vinculación prioritaria con otros dominios españoles (un 51% del total), dato que avala la falta de internacionalización detectada en estudios precedentes (Thelwall; Aguillo, 2003). En efecto, la UCM adolece de una relativamente escasa conexión con Europa, prácticamente la misma que con dominios norteamericanos.

Bibliografía

Björneborn, L.; Ingwersen, P. (2004). Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 55(14): 1216-122.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. Information Processing & Management, 28(1): 1–3.

Thelwall, M.; Aguillo, I. (2003). La salud de las web universitarias españolas. *Revista Española de Documentación Científica*, vol. 26(3), 291-305.

Vaughan, L.; Kipp, M.; Gao, Y. (2007). Why are websites co-linked? The case of canadian universities. *Scientometrics*, vol. 72(1), 81-92.