

J.10. Fuentes de enlaces web para análisis cibernéticos

Enrique Orduña-Malea

16 febrero 2012

Orduña-Malea, Enrique. "Fuentes de enlaces web para análisis cibernéticos (2012)".
Anuario ThinkEPI, 2012, v. 6, pp. 276-280.



Resumen: En la edición de 2012 del *Ranking web de universidades del mundo* se observa un cambio importante en la metodología, obligado por la fuente de obtención de enlaces externos a las sedes web de las universidades. Tras el cierre de *Yahoo site explorer* ahora la nueva herramienta utilizada es *MajesticSEO*. Se analizan los principales servicios y productos que ofrecen datos de enlaces, así como su idoneidad para ser utilizados en análisis cibernéticos. Se concluye que existe necesidad de herramientas creadas explícitamente para realizar análisis cibernéticos, dada la evolución de los motores de búsqueda y las limitaciones de otros productos, orientados a la optimización en buscadores y no al análisis informétrico.

Palabras clave: Cibermetría, Análisis de enlaces, *Ranking web de universidades del mundo*, *Exalead*, *Open site explorer*, *MajesticSEO*.

Title: Weblink sources for cybermetric analysis

Abstract: In the 2012 edition of the *Ranking web of world universities* there has been a major change in methodology, forced by changes in the source for obtaining external links to university websites. Following the closure of *Yahoo site explorer*, the new tool currently in use is *MajesticSEO*. The main services and products that offer data links and their suitability for use in cybermetric analysis are discussed. We conclude that there is a need for tools created specifically for cybermetric analysis, given the evolution of search engines and the limitations of other products aimed mainly at search engine optimization rather than at informetric analysis.

Keywords: Cybermetrics, Link analysis, *Ranking web of world universities*, *Exalead*, *Open site explorer*, *MajesticSEO*.

1. Introducción

En enero de 2012 se ha actualizado, como viene siendo habitual desde 2004, la edición del *Ranking web de universidades del mundo*¹ modificándose la fuente utilizada para obtener los datos de enlaces externos a las páginas web del catálogo de universidades (20.300 instituciones que lo convierten en el catálogo más completo de universidades del mundo). Por primera vez, se anuncia que la única fuente de datos de enlaces utilizada es *MajesticSEO*².

El motivo del cambio, como ya anunciaba su propio autor en un artículo previo (Aguillo, 2012), se debe tanto a la inhabilitación de los comandos de consulta de enlaces de *Yahoo search*³ ("linkdomain") como a la desaparición del servicio *Yahoo site explorer*⁴, como consecuencia de los acuerdos comerciales llevados a cabo entre *Microsoft* y *Yahoo!*⁵.

Aguillo partía de la necesidad de plantear nuevos rumbos y horizontes metodológicos dentro de la disciplina de la cibermetría, dado el

gran impacto que suponía la desaparición de los comandos de enlaces de *Yahoo!*, y que se pueden resumir en la necesidad de búsqueda de:

- nuevos indicadores: que puedan simular, predecir o complementar los análisis de enlaces;
- nuevas fuentes: que permitan, aunque a menor escala, la extracción de datos de enlaces.

Aguillo y otros autores han centrado su interés en el primer punto, analizando variantes de indicadores de invocación o mención (Thelwall; Sud, 2011). Sin embargo estas iniciativas no están exentas de la dependencia de los buscadores, y su aplicación no parece ser inmediata, como demuestra la elección de *MajesticSEO* en la nueva edición del *Ranking web*.

Los propósitos de este artículo son los de identificar los principales productos que todavía ofrecen datos de enlaces web, así como evaluar de forma general sus principales prestaciones y limitaciones, con el propósito de ofrecer una visión general de estas herramientas que ayude a conocer la mayor o menor validez de las mismas para fines cuantitativos.

2. Fuentes de enlaces web

Entre los distintos productos que ofrecen la posibilidad de obtener datos de enlaces se pueden distinguir:

- Motores de búsqueda con comandos de enlaces: *Exalead*⁶.
- Productos o servicios (orientados al SEO y a datos de tráfico web): *Open site explorer*⁷ y *MajesticSEO*, entre otros.
- Herramientas para webmasters ofrecidas por los grandes buscadores: *Google webmaster tools*⁸ y *Bing webmaster tools*⁹.

Para cada uno de ellos (descontando el último caso, sólo aplicable para los *webmasters* y, por tanto, con acceso a la administración de los sitios web), se deberían considerar al menos los siguientes aspectos en el caso de que se deseen utilizar como herramientas de análisis cibernético:

“La cibermetría necesita herramientas propias para sus análisis, pues los motores generalistas y los de nueva generación se están alejando de las prestaciones que necesita”

a) ¿Permite el enlazado selectivo?

El indicador más básico es el número total de enlaces que un url recibe. Si se desean realizar análisis cuantitativos, la herramienta debería permitir filtrar este número global, distinguiendo por una parte los enlaces internos y externos y, a partir de éstos últimos, cuantificar la cantidad de enlaces desde un sitio determinado, lo que se conoce como enlazado selectivo.

b) ¿Permite medidas de enlaces agregadas por dominio/subdominio?

Es fundamental conocer si se permiten los cálculos de enlaces a niveles desagregados:

- En la fuente (*source*): todos los enlaces que se reciben agrupados por dominio/subdominio (*domain-level inlink*). Es decir, todos los enlaces procedentes de un mismo dominio/subdominio se cuentan como 1 único enlace, por lo que la métrica en realidad se basa en cuantificar el número de subdominios/dominios desde donde un sitio web es enlazado.

- En el destino (*target*): todos los enlaces que recibe un determinado subdominio de un dominio general (por ejemplo, los enlaces que recibe “ec3.ugr.es”, como subdominio de “ugr.es”).

c) Diferencia de cobertura entre productos

La cobertura del producto es fundamental. Por mucho que una herramienta permita todas las opciones comentadas anteriormente, si la cantidad de enlaces que maneja en su base de datos es poco representativa, sesgada o con injustificadas fluctuaciones en el tiempo, esta fuente nunca deberá ser utilizada a menos que se trate de un análisis de la propia herramienta, y no de un universo de *websites* y enlaces determinado con fines cuantitativos.

3. Motores de búsqueda

Yahoo! ya no habilita ningún comando de enlaces (figura 1), mientras que *Bing* lo inhabilitó en 2007, debido precisamente al incremento de consultas automáticas que recibía el buscador¹⁰.

Google tampoco permite el análisis de enlaces. Si bien es cierto que incluye el comando “link:”, éste no permite la obtención de datos agregados. Es decir, la consulta “link: http://biblioteca.uv.es/valenciano/informacion/historia.php” proporciona el número total de enlaces a dicha página web, pero no es posible conocer todos los enlaces que recibe el valor agregado “biblioteca.uv.es”, ni el general de la universidad (“uv.es”), ni mucho menos cuantificar los enlaces en función de la procedencia.

Actualmente (febrero de 2012), el único gran motor de búsqueda que permite la obtención de enlaces a nivel agregado es *Exalead*. Además, facilita el enlazado selectivo al permitir la combinación de los comandos “linkdomain” y “site”, tal y como hacía *Yahoo search* antaño.

Dejando de lado los buscadores generalistas (y, por tanto, con mayor cobertura), se puede destacar la aparición del buscador *Topsy*¹², orientado a la extracción de información en los sitios de redes sociales y que, entre sus distintas prestaciones, destaca el *Social analytics*¹³, que proporciona datos relativos a enlaces contenidos en *tweets* significativos, y permite realizar comparativas entre dominios (figura 2).

Nuevos buscadores (*Blekko*¹⁴ o *Yebo!*¹⁵) u otros más clásicos (*WolframAlpha*¹⁶) indican que

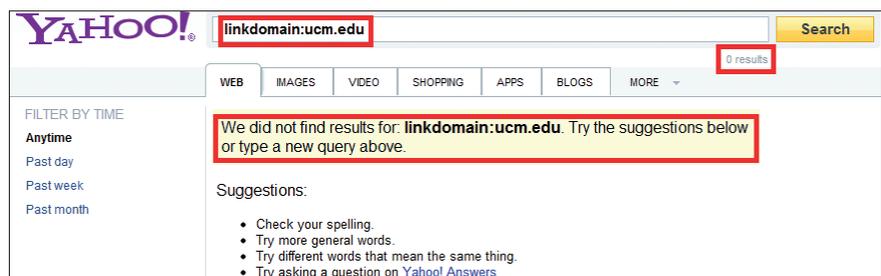


Figura 1. *Yahoo Search* inhabilita el comando “linkdomain”, <http://search.yahoo.com>

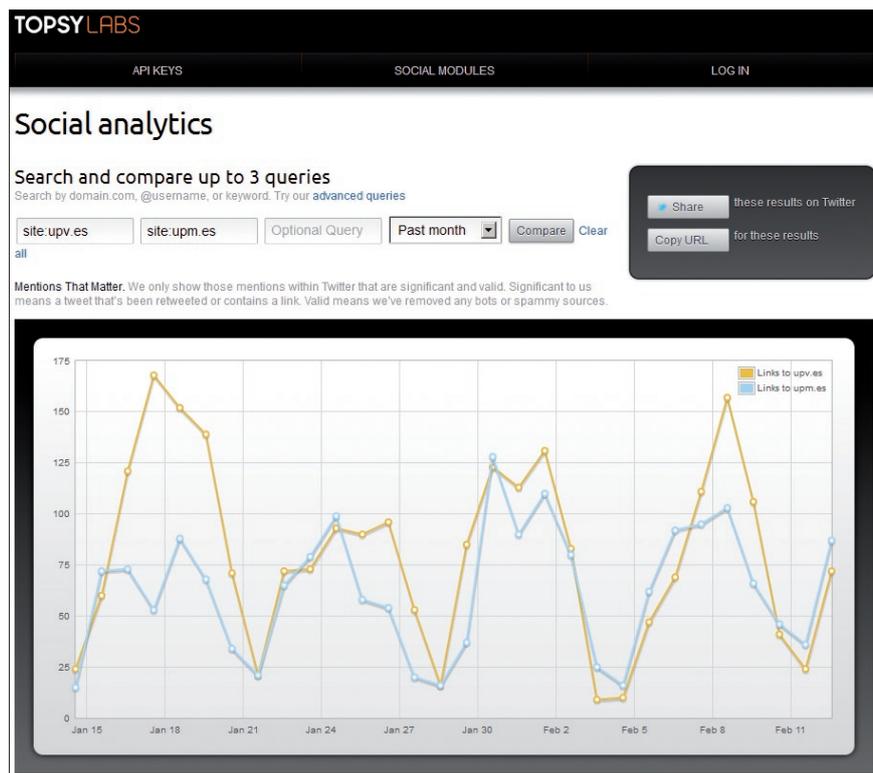


Figura 2. Análisis de enlaces contenidos en Twitter a través de Topsy <http://analytics.topsy.com>

los buscadores de nueva generación no parecen estar interesados en mostrar cuánta información poseen, sino sólo la información que el usuario desea, lo que es lógico por su parte pues no son herramientas creadas con los propósitos de realizar análisis informétricos.

Deberían elaborarse productos y buscadores creados explícitamente para la cuantificación de información.

La recuperación y medición de información contenida en etiquetas de metadatos es una posible vía. Experimentos como *Sindice*¹⁷ abren una esperanza a que se puedan realizar en el futuro análisis métricos dentro de la web semántica, pero todavía queda mucho camino para ello, como muestran los escasos avances en estudios webométricos a partir de la información contenida en metadatos (Longqing; Qingfeng, 2011).

4. Plataformas web

Esta categoría hace referencia a productos, más o menos comerciales, orientados principalmente a los webmasters y a los profesionales del SEO (*Search engine optimization*; optimización en buscadores). Entre los indicadores y métricas que ofrecen suele estar presente el número total de enlaces (llamados en estos contextos *backlinks*). Entre estos servicios, destacan principalmente *OpenSiteExplorer* y *MajesticSEO*.

a) *Open site explorer*

Desarrollado por el equipo de Seomoz¹⁸ a partir del antiguo índice *Linkscope*, que proporciona una gran variedad de métricas asociadas a un dominio web, especialmente el filtrado entre enlaces internos y externos, así como la separación entre enlaces a páginas, subdominios y dominios. Igualmente se ofrecen datos de enlaces externos agregados a nivel de dominio (*linking root domains*) que hacen de este servicio un producto de referencia a corto plazo.

Sin embargo, de cara a estudios cibernéticos, presenta 3 importantes carencias:

- No permite el enlazado selectivo (figura 3): no se puede cuantificar "automáticamente" el número de enlaces a un sitio web que provengan de un determinado dominio (por ejemplo, el número de enlaces que "upv.es" recibe de "Wiki-

pedia"), ni tampoco la exclusión de determinados sitios (por ejemplo, la exclusión de los enlaces provenientes de "ub.cat" si deseamos conocer los enlaces externos reales hacia "ub.edu").

- En el caso de detectar redirecciones (por ejemplo, "ub.cat" redirige a "ub.edu"), sólo proporciona información de la segunda.

- La cobertura es muy pequeña en comparación con la que proporcionaba *Yahoo! site explorer*.

b) *MajesticSEO*

Destacan las herramientas de comparación de rendimiento a lo largo del tiempo, los indicadores brutos de enlaces externos y de agregados a nivel de dominio (*referring domains*), y la gran cobertura de datos que maneja. Sin embargo, es imposible calcular datos de enlazado selectivo, ni excluir determinados dominios en la consulta.

c) *Otros productos*

Existen infinidad de servicios, algunos orientados a prácticas de SEO (*Backlink watch!*¹⁹) y otros a proporcionar otro tipo de datos, pero que incluyen adicionalmente indicadores de enlaces. Por ejemplo *Alexa*²⁰ (orientado a datos de tráfico) todavía proporciona información del número de enlaces externos, agregados a nivel de dominio, concepto llamado "sites linking in", y que algunos autores proponen como alternativa al cálculo

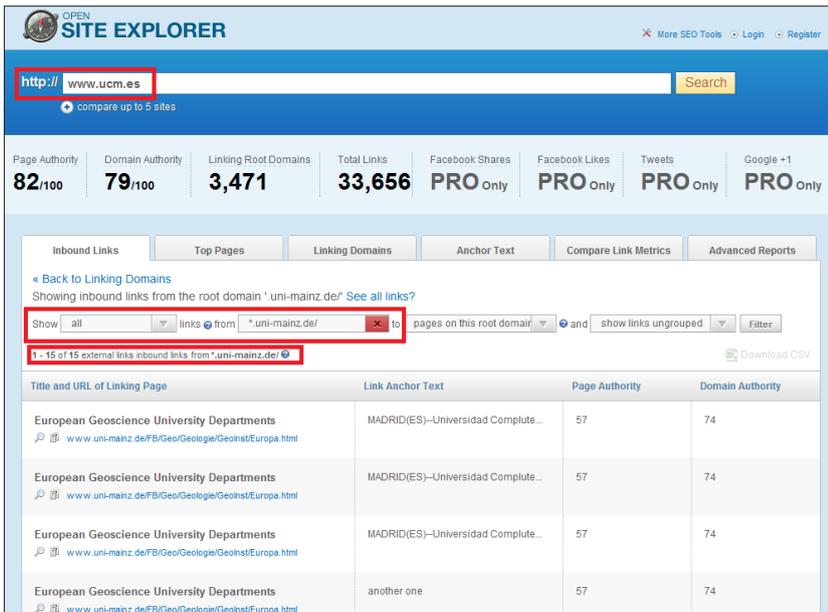


Figura 3. Enlaces selectivos desde Open site explorer, <http://www.opensite-explorer.org>

de enlaces (Vaughan, 2012), Ahrefs²¹ (orientado casi exclusivamente al análisis de backlinks), y SEMRush²² (orientado al SEO y SEM), proporciona un informe de enlaces donde identifica la fuente y el destino del enlace (figura 4).

5. Cobertura

Como punto final, y con el objetivo de mostrar las diferencias de cobertura entre cada uno de los productos comentados, se presenta una tabla comparativa de enlaces totales externos, para las 10 primeras universidades mundiales según el Ranking web (enero 2012).

Se muestran claramente las diferencias de cada producto:

- La poca cobertura de Exalead. Se han analizado universidades estadounidenses, pues la cobertura en otros países (como España) es todavía peor (el caso de Pittsburgh es suficientemente explicativo). El hecho de que la exclusión de datos externos (comando “-site:”) sea manual (en el resto viene por defecto) podría influir en los pocos resultados obtenidos y en la nula correlación con el resto de fuentes.

- Las diferencias en cobertura entre OSE y MajesticSEO no son tan elevadas como a primera vista pueden parecer en esta pequeña muestra de ejemplo, dado que siempre hay que normalizar los datos y tomarlos a nivel macro. En bibliometría 10 citas en ciencias

sociales son muchas (y pasar de 10 a 20 es un mundo), pero en cibermetría se manejan miles de enlaces (Aguillo, 2012), por lo que tener 1.000 o 1.250 supone prácticamente lo mismo en ciertos contextos.

En todo caso, sí se observan algunos casos puntuales: Harvard destaca en OSE, y Cornell en MajesticSEO. Sería necesario tomar una muestra lo suficientemente representativa (y global) para conocer con mayor precisión este dato, sobre todo para dominios con valores extremos (muy pocos o muchos enlaces), donde MajesticSEO parece menos fiable.

No obstante, el Ranking web ha elegido este último producto, lo que supone un punto de debate interesante en la actualidad, sobre todo a la hora de analizar los posibles cambios en las posiciones que el uso de esta nueva fuente haya podido generar.

6. Conclusiones

Los productos que ofrecen actualmente datos de enlaces levantan suficientes dudas metodológicas como para que no sean totalmente válidos en análisis cibernétricos, aunque proporcionan suficiente información para conocer el “rendimiento comparado” de una muestra de sitios web.

La cibermetría precisa por tanto herramientas propias para realizar sus análisis, pues los motores generalistas (los que mayor cobertura ofrecen) y los de nueva generación (buscadores semánticos)

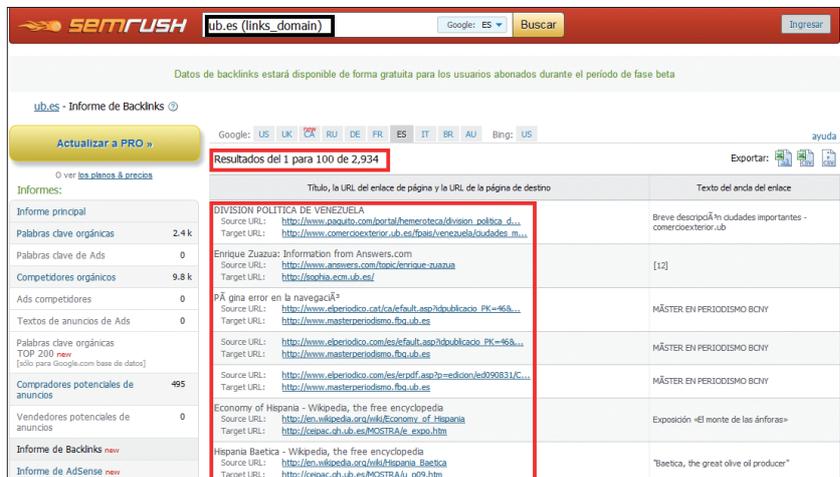


Figura 4. Enlaces externos (source y target) desde SEMRush, <http://es.semrush.com>

Universidad	URL	Exalead	OSE	Majestic
Harvard	harvard.edu	550.176	23.271.358	10.373.807
MIT	mit.edu	449.060	8.610.712	8.246.457
Stanford	stanford.edu	311.406	9.574.644	8.373.193
Michigan	umich.edu	172.259	3.974.570	4.352.221
California-Berkeley	berkeley.edu	287.432	6.113.208	6.480.190
Cornell	cornell.edu	201.813	5.424.199	11.534.443
Michigan	msu.edu	172.357	5.794.637	5.686.311
Wisconsin-Madison	wisc.edu	154.467	3.204.919	4.110.326
Pittsburgh	pitt.edu	88.738	7.339.572	10.136.334
Carnegie Mellon	cmu.edu	155.030	2.113.742	1.479.717

Tabla 1. Enlaces externos en diferentes fuentes para las 10 primeras universidades en el *Ranking web de universidades del mundo* (enero 2012)

se están alejando de las prestaciones que esta disciplina necesita, mientras que otros productos y servicios existentes sólo proporcionan datos de enlaces externos brutos, sin posibilidad de seleccionar la procedencia (filtrado) de los mismos, y con problemas de cobertura que deben ser analizados en profundidad.

7. Agradecimientos

Muchas gracias a **Isidro F. Aguillo** y **Álvaro Cabezas-Clavijo** por sus comentarios que ayudaron a mejorar este texto.

8. Notas

1. Ranking web de universidades del mundo
<http://www.webometrics.info>
2. MajesticSEO
<http://www.majesticseo.com>
3. Yahoo search
<http://search.yahoo.com>
4. Yahoo site explorer
<http://siteexplorer.search.yahoo.com/index.php>
5. "It's official: Yahoo-Microsoft Announce 10-Year search/ad pact". *The Washington post*, 2009.
<http://www.washingtonpost.com/wp-dyn/content/article/2009/07/29/AR2009072901108.html>
6. Exalead
<http://www.exalead.com/search>
7. Open site explorer
<http://www.opensiteexplorer.org>
8. Google webmaster tools
<https://www.google.com/webmasters/tools>
9. Bing webmaster tools
<http://www.bing.com/toolbox/webmaster>
10. "We are flattered, but...". *Bing community*, 2007.
http://www.bing.com/community/site_blogs/blsearch/archive/2007/03/28/we-are-flattered-but.aspx

11. Gigablast
<http://www.gigablast.com>
12. Topsy
<http://topsy.com>
13. Social analytics
<http://analytics.topsy.com>
14. Blekko
<http://blekko.com>
15. Yebol
<http://www.yebol.cn>
16. WolframAlpha
<http://www.wolframalpha.com>
17. Sindice
<http://sindice.com>
18. Seomoz
<http://www.seomoz.org>
19. Backlink watch!
<http://www.backlinkwatch.com>
20. Alexa
<http://www.alexa.com>
21. Ahrefs
<http://ahrefs.com>
22. SEMRush
<http://www.semrush.com>

9. Referencias bibliográficas

- Aguillo, Isidro F.** "La necesaria evolución de la cibernética". *Anuario ThinkEPI*, 2012, v. 6, pp. 119-122.
- Longqing, Shi; Qingfeng, Zhao.** "Data sources on bibliometrics". En: *7th intl conf on computational intelligence and security*, 2011, pp. 1312-1315.
- Thelwall, Mike; Sud, Pardeep.** "A comparison of methods for collecting web citation data for academic organizations". *JIST*, 2011, v. 62, n. 8, pp. 1488-1497.
- Vaughan, Liven.** "An alternative data source for web hyperlink analysis: 'Sites linking in' at *Alexa Internet*". *Collnet journal of scientometrics and information management (en prensa)*.