

PROBLEMAS TÉCNICOS, METODOLÓGICOS Y DOCUMENTALES EN LA ELABORACIÓN DE RANKINGS BASADOS EN INDICADORES WEB

Aguillo, Isidro F.

Laboratorio de Cibermetría
Centro de Ciencias Humanas y Sociales (CCHS - CSIC)
isidro.aguillo@cchs.csic.es

Resumen.

Las técnicas cibernéticas se han extendido considerablemente en los últimos años, pero aún faltan estudios en detalle sobre los problemas que plantean dichas técnicas y las posibles soluciones a los mismos. Esta revisión metodológica se centra en la aplicación de la Cibermetría a la obtención de indicadores Web, especialmente aquellos que son útiles para la elaboración de Rankings de carácter institucional. Se estudian las unidades de análisis, las posibilidades y limitaciones de los motores de búsqueda como fuente de información y se describen los métodos de elaboración de Rankings.

Palabras clave: Internet, Web, Cibermetría, Indicadores Web, Motores de búsqueda, Rankings

Introducción

La cibermetría se ha consolidado en los últimos años como una de las técnicas cuantitativas más interesantes para la descripción y evaluación de la actividad científica (Aguillo et al., 2006), incluyendo la posibilidad todavía no suficientemente explorada de analizar relaciones más amplias del sistema ciencia, economía y sociedad.

Los indicadores Web se han demostrado útiles en contextos de comunicación científica pero el análisis de enlaces requiere todavía de más estudios teóricos antes de entender las y clasificar las motivaciones que llevan al establecimiento de un enlace (Thelwall, 2004). Estas son obviamente más ricas y variadas que la mera cita bibliográfica y por tanto ofrecen una posibilidad inédita en el estudio de las relaciones entre sedes web cualquiera que sea su contenido.

En el presente trabajo se analizan las principales tareas, modelos, técnicas y algunos resultados empíricos de la elaboración de rankings, fundamentalmente universitarios (Aguillo, Ortega y Fernández, 2008), a partir de indicadores Web.

Selección de las unidades de análisis.

El viejo aforismo de “juntar peras con peras y manzanas con manzanas” es aquí plenamente válido. Existen varios niveles de unidades documental en la Web, pero en muchos casos son de difícil aplicación general y solo en ocasiones vamos a encontrar una amplia identificación entre unidades documentales e informáticas. Un único documento puede estar representado no solo por un fichero html, sino que necesita ficheros independientes para imágenes, iconos u otros gráficos. Incluso un fichero pdf puede representar sólo un capítulo o una sección de un documento completo.

A nivel superior, el propio sistema de nombres de dominio hace que nos encontremos con una identidad entre contenidos albergados en un mismo dominio o subdominio y una unidad institucional representada por el mismo. Es el concepto de sede Web (Aguillo, 1998), que permite equiparar unidades documentales (web) con instituciones en sentido amplio.

Este procedimiento institucional es el que se recomienda en este análisis, de forma que solo aquellas organizaciones con un dominio o subdominio propio serán consideradas unidades de análisis.

Aunque excluyamos a aquellas organizaciones sin presencia Web, esto plantea algunos problemas ya que existen algunas cuya URL unitaria esta en un directorio (xxx.es/yy). No se trata solo de instituciones pequeñas, con pocos recursos, en países en vías de desarrollo, sino que también afecta a organizaciones de mayor tamaño, por razones de tipo organizativo o político. Muchas bibliotecas, centros de documentación e incluso institutos de investigación no tienen dominio propio. Numerosos hospitales, especialmente en España, aparecen con dominios de las comunidades autónomas que los

financian y en algún caso con una estructura de nombres compleja y confusa.

Tabla 1. Direcciones de los Hospitales dependientes de la Comunidad Autónoma de Madrid

URLs (http://www.madrid.org/cs/Satellite?...)
...pagename=HospitalPuertaHierroMaja/Page/HPHM_home
...pagename=HospitalHenares/Page/HHEN_home
...language=es&pagename=HospitalInfantaCristina/Page/HSUR_home
...language=es&pagename=HospitalInfantaElena%2FPage%2FHVAL_home
...language=es&pagename=HospitalInfantaLeonor%2FPage%2FHVLL_home
...language=es&pagename=HospitalInfantaSofia/Page/HNOR_home
...pagename=HospitalSureste/Page/HSES_home
...pagename=HospitalTajo/Page/HTAJ_home
...language=es&pagename=HospitalFundacionHospitalAlcorcon%2FPage%2FHALC_home
...pagename=Hospital12Octubre/Page/H12O_home
...pagename=HospitalCarlosIII/Page/HCAR_home&c=Page&site=HospitalCarlosIII
...language=es&pagename=HospitalClinicoSanCarlos/Page/HCLN_home
...language=es&pagename=HospitalCruzRojaSanJoseSantaAdela/Page/HCRU_home
...pagename=HospitalRodriguezLafora/Page/HLAF_home&c=Page&site=HospitalRodriguezLafora
...pagename=HospitalElEscorial/Page/HESC_home&c=Page&site=HospitalElEscorial
...language=es&pagename=HospitalFuenlabrada/Page/HFLA_home
...pagename=HospitalGetafe/Page/HGET_home
...pagename=HospitalGregorioMaranon/Page/HGMA_home
...language=es&pagename=HospitalGuadarrama/Page/HGUA_home
...pagename=HospitalFuenfria/Page/HFUE_home&c=Page&site=HospitalFuenfria
...language=es&pagename=HospitalLaPaz/Page/HPAZ_home
...language=es&pagename=HospitalLaPrincesa/Page/HPRI_home
...pagename=HospitalMostoles/Page/HMOS_home
...pagename=HospitalNinoJesus/Page/HNIJ_home
...pagename=HospitalPrincipeAsturias/Page/HPPE_home&c=Page&site=HospitalPrincipeAsturias
...pagename=HospitalRamonCajal/Page/HRYC_home
...pagename=HospitalSantaCristina/Page/HCRI_home&c=Page&site=HospitalSantaCristina
...pagename=HospitalSeveroOchoa/Page/HSEV_home&c=Page&site=HospitalSeveroOchoa
...pagename=HospitalVirgenPoveda/Page/HVPO_home&c=Page&site=HospitalVirgenPoveda
...pagename=HospitalVirgenTorre/Page/HVTO_home&c=Page&site=HospitalVirgenTorre
...pagename=HospitalJoseGermain/Page/HGER_home&c=Page&site=HospitalJoseGermain

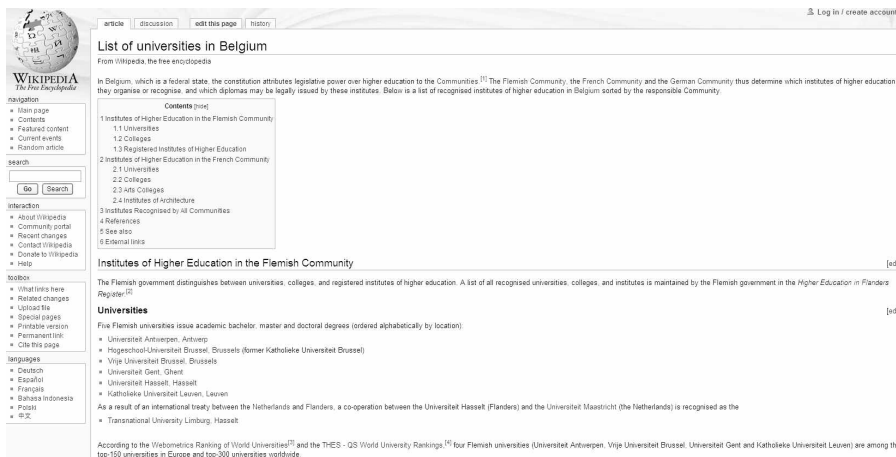
Las fuentes para acceder a listados de direcciones de instituciones son muy variadas. En el caso de Universidades, Hospitales o Bibliotecas existen directorios tanto mundiales como nacionales que pueden ser útiles. Sin embargo, ni la cobertura de los primeros es exhaustiva ni todos los países disponen de índices fiables. Además el mantenimiento de muchos de estos directorios deja bastante que desear y es frecuente encontrar un gran número de enlaces rotos. La Tabla 2 incluye un listado de algunos de los más importantes directorios de Universidades.

Tabla 2. Principales directorios Web de Universidades del Mundo

Colleges and Universities	www.mit.edu/people/cdemello/univ-full.html
Web US Higher Education	www.utexas.edu/world/univ/
Universities Worldwide	univ.cc
Online University Directory	www.braintrack.com
All Universities around the World	www.bulter.nl/universities/
General Education Online	www.findaschool.org
Index of American Universities	www.clas.ufl.edu/au/
Canadian Universities	www.uwaterloo.ca/canu/index.php
International Colleges and Universities	www.4icu.org
List of Universities of the World	www.unesco.org/iau/onlinedatabases/list.html
University Directory	www.university-directory.eu
Universities.ac	www.universities.ac

Hay que señalar que en la actualidad una fuente importante de información al respecto es la Wikipedia (www.wikipedia.org), que no solo incluye entradas con amplios directorios nacionales, sino que proporciona información muy actualizada para casos concretos (Figura 1). Las entradas a organizaciones cuyo nombre ha cambiado, que se han fusionado con otras o que incluso han desaparecido son muy útiles para resolver situaciones complicadas.

FIGURA 1. Volcado de pantalla de una entrada de la Wikipedia que muestra la lista de Universidades de Bélgica (Enero 2009).



La delimitación de la tipología institucional es a menudo difícil. En el caso de las Universidades nos encontramos con instituciones de educación superior que ofertan titulación de tercer ciclo como Escuelas de Negocios, Conservatorios de Música o Escuelas de Arte, Danza, Teatro, Cine o Televisión cuya inclusión dependerá de los objetivos buscados. Otra situación a considerar es la realización o no de actividades de investigación. En general solo las universidades de ciclo completo, incluidos los estudios de doctorado, entrarían en esa categoría. Según Van Raan (2008), el número de dichas universidades que son muy productivas apenas supone unos pocos cientos, pero como demuestran las bases de datos de citas (WoS, Scopus, Google Scholar), la cifra total de organizaciones que publican superan los varios miles.

La clasificación de instituciones educativas ha sido objeto de estudio detallado (The Carnegie Classification of Institutions of Higher Education, www.carnegiefoundation.org/classifications/) y puede ser una guía para delimitar los criterios de inclusión. La Tabla 3 ofrece un resumen integrado de varias propuestas aunque no pretende ser exhaustiva.

Tabla 3. Clasificación de Instituciones de Educación Superior (varias fuentes)

Universidades generalistas	Se incluyen las orientadas a investigación con programas de doctorado así como las multidisciplinares. Se suele tratar de grandes instituciones públicas o privadas con gran tradición.
Universidades Politécnicas	Escuelas superiores y medias de ingeniería, arquitectura o informática
“Liberal Arts Colleges”	Muy frecuentes en EEUU, las Escuelas de Artes Liberales ofrecen una formación menos convencional
Escuelas Universitarias	Ofrecen ciclos cortos, también llamadas “Two-Year Colleges” o “Junior Colleges”
Escuelas profesionales	Generalmente de carácter especializado, de muy diversa orientación disciplinar y que en ciertos países se solapan con categorías anteriores
	Seminarios teológicos y otras instituciones ofreciendo títulos en religión
	Escuelas y centros de medicina
	Otras escuelas relacionadas con medicina (Enfermería,..)
	Escuelas agrícolas, veterinarias o forestales
	Escuelas de deportes y educación física
	Escuelas de ingeniería, tecnología o informática
	Escuelas de negocios y gestión y administración de empresas
	Escuelas de arte, música, diseño, danza, teatro, etc..
	Escuelas de Derecho y para-derecho
	Escuelas de Pedagogía
(Educación de adultos)	

Un problema diferente lo suponen las organizaciones no oficiales o incluso fraudulentas. La legislación de ciertos países permite la existencia de instituciones privadas con nombres confusos que otorgan títulos de validez limitada o nula. En muchos casos se trata de verdaderas estafas (“*diploma mills*”) o de universidades “internacionales”, cuyos títulos no siempre tienen validez, especialmente en los países desarrollados. Muchas de estas instituciones ofrecen cursos a distancia u “online”, por lo que a menudo esto supone un problema adicional al resultar difícil discernirlas de proveedores de educación no presencial perfectamente legítimos.

En Latinoamérica es relativamente común encontrarse dominios de organizaciones educativas que ofrecen todos los ciclos docentes, incluidos el universitario. Se trata normalmente de organizaciones religiosas que reúnen en único campus todas las edades y cuya formación de tercer grado suele tener un fuerte componente tecnológico-profesional. El problema es que solo tienen un único dominio y la sección universitaria representa apenas un directorio.

La UNESCO (<http://www.unesco.org/iau/onlinedatabases/list.html>) ha creado una lista “oficial” de universidades que recoge unas 17000 instituciones cuya validez ha sido refrendada por los países miembros. Sin embargo, no existe un criterio uniforme en todos los casos y mientras que algunos países han sido muy restrictivos, otros han suministrado una lista más flexible. Puesto que no se proporcionan URLs es difícil saber el grado de solapamiento con otros directorios, pero la cifra podría representar el 90% del total del sector a nivel mundial.

Problemas de Multidominio.

Del mismo modo que la firma institucional de un artículo científico no está normalizada y podemos encontrar una misma universidad u hospital bajo incluso varias docenas de variantes de nombres diferentes, la asignación de dominios a instituciones también plantea problemas.

Algunas instituciones permiten el uso de dominios externos para cierto de actividades. Por ejemplo, los proyectos europeos que involucran a varias instituciones tienen dominio org, info o net y desde hace poco tiempo eu, aunque las páginas estén hospedadas en un servidor concreto de una universidad con dominio propio.

En otros casos el servidor institucional hospeda páginas de terceros sin reconocer dicha situación en el dominio que coincide con el de la organización. Congresos o seminarios internacionales, *mirrors* de directorios o bases de datos, repositorios temáticos, documentación de software, portales de revistas, sedes de sociedades científicas y otras situaciones enriquecen los contenidos pero a base de inflarlos con material ajeno.

Una situación más preocupante es la existencia de varios dominios principales o subprincipales. A veces una universidad mantiene dos dominios equivalentes por cuestiones de comodidad de acceso (*nombre.edu* y *nombre.org*, pero también *nombre.edu.pais*), pero en otras lo que ha ocurrido es un cambio de dominio que no se ha generalizado a todos los servidores. Esa convivencia de dominios castiga severamente la visibilidad de una organización en la web. En los casos extremos el cambio solo afecta al servidor principal y se mantiene unos o varios dominios adicionales que afectan a un porcentaje significativo del resto de servidores.

TABLA 4. Problemas de múltiples dominios en las Universidades catalanas (número de objetos de acuerdo diferentes buscadores, Noviembre 2008)

Universidad /Dominio		GOOGLE	YAHOO	EXALEA D	LIVE	SCHOLAR
Univ. de Barcelona	<i>ub.es</i>	222.000	486.990	48.749	52.600	6.570
	<i>ub.edu</i>	241.000	150.036	14.642	57.000	233
	<i>ub.cat</i>	6.800	2.010	5	661	26
Univ. Autónoma de Barcelona	<i>uab.es</i>	581.000	467.098	28.252	66.400	6.510
	<i>uab.cat</i>	1.000.000	221.558	1.272	13.800	852

Univ. Politécnica de Cataluña	<i>upc.es</i>	582.000	469.432	37.549	64.200	5.730
	<i>upc.edu</i>	1.590.000	342.572	11.008	60.700	3.490
	<i>upc.cat</i>	25	16		14	
Univ. Rovira & Virgili	<i>urv.es</i>	277.000	105.866	6.699	30.400	1.040
	<i>urv.net</i>	537.000	34.566	1.546	23.200	56
	<i>urv.cat</i>	215.000	20.356	237	4.670	132
Univ. Pompeu Fabra	<i>upf.es</i>	51.700	97.454	14.537	86.200	905
	<i>upf.edu</i>	175.000	588.165	21.853	97.200	938
	<i>upf.cat</i>	28	3	64	2	14
Univ. de Gerona	<i>udg.edu</i>	370.000	185.008	5.759	34.400	237
	<i>udg.es</i>	42.000	103.978	14.099	31.000	2.010
	<i>udg.cat</i>	135	278	92	225	
Univ. de Lérida	<i>udl.es</i>	99.500	121.814	10.530	30.900	920
	<i>udl.cat</i>	66.100	20.683	1.132	5.030	149
Univ. Ramón Llul	<i>url.es</i>	2.140	1.693	276	6.410	22
	<i>url.edu</i>	59.900	57.308	1.094	24.500	160
	<i>url.cat</i>	553	340		3	
Univ. Intern. de Cataluña	<i>unica.es</i>	18.700	2.817	13	130	
	<i>unica.edu</i>	43.000	32.738	94	229	
Univ. de Vic	<i>uvic.es</i>	741	363	3.910	315	100
	<i>uvic.cat</i>	56.900	12.687	1.602	427	72
Univ. Oberta Catalunya	<i>uoc.es</i>	9.410	19.625	3.516	27.400	277
	<i>uoc.edu</i>	195.000	108.336	9.590	78.100	592
	<i>uoc.cat</i>		3	1	1	

Hay casos especiales que merecen verse en detalle. Las Universidades de Uruguay o Zagreb (Croacia) no tienen un dominio central común, de forma que las principales facultades tienen dominios distintos. La Universidad de Helsinki comparte dominio con el ayuntamiento de la ciudad y los subdominios pueden corresponder indistintamente facultades o departamentos o bien a información turística. Un último caso está representado por varios campus de universidades francesas. El campus (Jussieu en París) o un grupo de universidades (las tres de Estrasburgo) puede tener un dominio común, compartido con centros de investigación independientes (unidades del CNRS, por ejemplo), mientras que cada universidad tiene un dominio diferente que suele ser tener muchos menos contenidos.

Los hospitales universitarios plantean algunos problemas, pues aunque muchos centros están

ligados a facultades de medicina y por tanto comparten el dominio universitario, esto no siempre ocurre así. Hay hospitales con dominio diferente (la mayor parte de los holandeses, que son claramente parte de las respectivas universidades), pero hay situaciones donde la escuela de medicina también tiene dominio distinto (por ejemplo la de la Johns Hopkins).

Muchos de los hospitales no tiene sede propia y son los consorcios los que los reúnen bajo un paraguas común. El poderoso sector sanitario privado estadounidense da entrada Web por corporaciones, no por hospitales individuales, lo que dificulta la comparación de contenidos.

Solo la mitad de los repositorios institucionales tienen dominio o subdominio propio, ya que suelen compartir el dominio de la biblioteca que los hospeda. En algún caso los registros tienen dirección diferente al fichero del artículo que puede estar depositado en un servidor distinto.

Selección de las herramientas.

El análisis cuantitativo exige de herramientas que permitan la recolección automática de las principales variables que describen los contenidos de una sede o un dominio web completo. Se trata de los llamados robots, agentes o “crawlers”, unos programas diseñados para explorar las páginas web siguiendo los enlaces de los árboles hipertextuales para recopilar todos sus contenidos. Aunque se pueden utilizar robots personales diseñados especialmente para tareas cibernéticas, se trata de programas de difícil y compleja personalización, útiles para un número limitado de sedes (Thelwall, 2001).

La alternativa que permite abordar escenarios globales es el uso de las bases de datos de los principales motores de búsqueda. (Aguillo et al., 2006). Aunque no exentos de limitaciones y problemas se pueden diseñar estrategias para disminuir al máximo sesgos e irregularidades en los resultados obtenidos.

El número de motores con grandes bases de datos independientes es limitado y más aún aquellos que permiten la recuperación de datos de carácter cibernético de forma controlada. La Tabla 5 muestra los más importantes y la sintaxis actual

Tabla 5. Sintaxis para la extracción de datos cibernéticos de los principales motores de búsqueda (Noviembre 2008)

	GOOGLE	YAHOO	LIVE	EXALEAD	ASK	GIGABLAST
TLD	site:xx	NO	site:xx	site:xx	site:xx	site:xx
dominio	site:aa.xx	NO 1	site:aa.xx	site:aa.xx	site:aa.xx	site:aa.xx
directorio	site:aa.xx/bb	(inurl:aa.xx/bb)	site:aa.xx/bb	NO	site:aa.xx/bb	NO
palabra url	inurl:xx	inurl:xx	NO	inurl:xx url:xx	inurl:xx	inurl:xx
enlace	link:aa.xx/b.htm	NO 1	NO	link:www.aa.xx	(NO)	(NO)
enlace dominio	NO	(linkdomain:aa.xx)	NO	link:aaa.xx	NO	NO
tipo fichero	filetype:yy	originurlextension:yy	filetype:yy	filetype:yy	filetype:yy	filetype:yy
idioma	Avanzada	Avanzada	Avanzada	Avanzada	Avanzada	NO
país	Avanzada	Avanzada	(Avanzada)	Avanzada	Avanzada	NO

En la mayoría de los casos se pueden obtener resultados fiables de los APIs que ofrecen los diferentes motores aunque hay que tener en cuenta que trabajan sobre bases de datos menos actualizadas y generalmente menores que los interfaces comerciales. Esto justifica el uso de Yahoo Search sobre sus *mirrors* actuales Altavista y Alltheweb.

El caso de Google es ligeramente diferente. Este buscador depende de una serie de Data Centers con contenidos ligeramente diferentes y que responden a las peticiones de forma impredecible. Así dos búsquedas consecutivas a la misma dirección de Google pueden producir resultados incluso bastante distintos pues han sido solucionadas desde diferentes centros. La solución en este caso es identificar una IP concreta (a través del caché por ejemplo) y realizar las peticiones directamente a dicha dirección. La Tabla 6 proporciona algunas direcciones IP de Data Centers de Google:

Tabla 6. Direcciones IP de Data Centers de Google operativos a finales de 2008

http://64.233.161.99/	http://66.249.89.104/
http://64.233.161.104/	http://66.249.91.99/
http://64.233.161.147/	http://66.249.91.104/
http://64.233.167.99/	http://66.249.93.99/
http://64.233.167.104/	http://66.249.93.104/
http://64.233.167.147/	http://72.14.203.99/
http://64.233.169.99/	http://72.14.203.104/
http://64.233.169.104/	http://72.14.205.99/
http://64.233.179.99/	http://72.14.205.104/
http://64.233.179.104/	http://72.14.207.99/
http://64.233.183.99/	http://72.14.207.104/
http://64.233.183.104/	http://72.14.221.99/
http://64.233.187.99/	http://72.14.221.104/
http://64.233.187.104/	http://72.14.235.99/
http://64.233.189.104/	http://72.14.235.104/
http://66.102.1.104/	http://216.239.59.99/
http://66.102.9.99/	http://216.239.59.103/
http://66.102.9.104/	http://216.239.59.104/
http://66.102.9.147/	http://216.239.59.147/
http://66.249.89.99/	

Indicadores Web.

Existe una amplia bibliografía (Codina, 2000,2004; Jiménez Piano, 2001) sobre distintos aspectos de la Web que pueden medirse tanto de forma cualitativa como cuantitativa. Muchos de ellos están centrados en el diseño y usabilidad de las páginas, mientras que otros calibran el seguimiento o cumplimiento de estándares. Sólo alguna de las variables tiene posibilidad de generar indicadores cuantitativos, que son las que describimos a continuación.

Tamaño. Desde un punto de vista informático el tamaño de los ficheros puede estar correlacionado con el volumen de contenidos de los mismos, pero dicha relación que es válida para ficheros textuales, no lo es en absoluto para páginas dinámicas y muy especialmente para aquellas ricas en gráficos de calidad (jpg, png), o con ficheros de audio o video.

Desde un punto de vista documental una página Web puede formar una unidad más adecuada. Hay que tener en cuenta sin embargo que existen varios miles de formatos en la Web y que aunque la gran mayoría son asimilables a los formatos HTML, en otros casos hay ficheros muy complejos y/o grandes.

Profundizando en la aproximación documental existe un grupo de ficheros denominados ficheros en formatos ricos o simplemente ficheros ricos que suelen representar documentos completos unitarios. Los ficheros ricos presentan varias ventajas ya que un único fichero puede contener e integrar un gran volumen de información no solo textual, permite su organización de forma práctica y elegante mediante maquetaciones guiadas por lenguajes de descripción de páginas y se han convertido en estándares tanto dentro como fuera de la Red.

Aunque son varios los formatos ricos, los más importantes en cuanto a número (con mucha diferencia con respecto por ejemplo a los formatos abiertos) son: Adobe Acrobat (pdf), los ofimáticos del Microsoft Office (Word:doc, rtf; Powerpoint:ppt; Excel:xls) y los procedentes de editores de texto especializados (Latex) como el PostScript (ps, eps).

La distribución por idioma debe descartarse dadas las limitaciones de los actuales sistemas de asignación automática.

Artículos. La base de datos Google Scholar (scholar.google.com), desaparecida Live Academic y no incluyendo la más tradicional Scirus, se ha convertido en la principal fuente de información académica de la Red. Aunque todavía se encuentra en versión beta, la inclusión de citas incrementa considerablemente su valor con fines descriptivos y evaluativos. La falta de control bibliográfico no es un inconveniente importante para la obtención de indicadores cibernéticos y aunque no es posible filtrar tipologías ni formatos todavía, las cifras obtenidas pueden ser representativas del volumen de información científica publicada en cierto dominio. La posibilidad de extraer citas de forma automática desarrollada por Harzing (www.harzing.com/pop.htm) es una interesante opción para análisis más detallados.

Enlaces. La estructura hipertextual del Web es muy útil para su descripción. La densidad (media de enlaces por página) puede ser una medida inadecuada dada la existencia de grandes directorios y el comportamiento “power law” de las distribuciones Web. En todo caso parece conveniente distinguir los enlaces internos, con fines de navegación, de los externos que pueden tener unas motivaciones más diversas. La opción obvia es clasificar los enlaces por dominios, tanto de alto nivel como institucionales.

Una medida interesante, aunque no necesariamente ligada a localidad, es el porcentaje de enlaces rotos, es decir no operativos. En general su número está ligado al mantenimiento efectivo de las páginas pero puede depender mucho de la dinámica del área concreta.

El indicador más interesante desde un punto de vista cibernético es la visibilidad, el número de enlaces externos recibido por una página o sede Web (backlinks). El método más eficiente de obtenerlo es a través de ciertos comandos en motores de búsqueda. Un sistema más sofisticado es el PageRank, el algoritmo de Google que tiene en cuenta no solo el número de enlaces sino la importancia relativa de las páginas que los originaron. El PR publicado en diferentes fuentes es inservible dada su escasa capacidad discriminadora y aunque se puede computar el algoritmo esta es una opción normalmente inviable por la necesidad de considerar porciones significativas del webespacio.

El factor de impacto Web, medido como relación entre enlaces recibidos y páginas web de la sede receptora, se ha demostrado sujeto a artefactos matemáticos derivados de la distribución de ambas variables. No es una opción válida para describir sedes de pequeño y mediano tamaño.

Una alternativa es considerar universos cerrados, donde solo se contabilizan los enlaces cruzados entre los miembros de la población y no los de terceros. Es lo que se denomina factor G y en el caso de universidades mide únicamente los enlaces que provienen de otras universidades. Es una medida difícil de realizar por el gran número de peticiones que requiere. Sin embargo abre la puerta a otros análisis como el estudio de co-enlaces.

Visitas. El número y características de las visitas que recibe un servidor Web solo están disponibles para el webmaster del mismo y aunque algunas veces se publican en abierto, la falta de estándares hace difícil la comparación precisa entre los datos disponibles.

Una manera indirecta de acceder a indicador de popularidad (medida en número de visitas en contraste a la visibilidad que considera número de enlaces) es utilizar el Traffic Rank de Alexa (www.alexa.com), un ranking (es decir una medida relativa) de dominios ordenados por número decreciente de visitas interceptadas a través del sistema Alexa, un spyware no dañino que tiene una amplia base de instalaciones, fundamentalmente mediante la barra Alexa.

El dato de Alexa presenta fuertes sesgos regionales y una gran variación tanto anual como semanal. Los valores se calculan para periodos de tres meses para reducir dicha variación.

Modelos de Rankings.

El objetivo de un Ranking es reducir una serie de variables a un único ordinal que represente al conjunto de las mismas. La combinación de las variables exige la utilización de pesos distintivos para cada una de ellas, que se pueden estimar mediante métodos empíricos o a través de un modelo previamente establecido. Es habitual que varias o muchas de las variables involucradas estén fuertemente correlacionadas, por lo que un escenario complejo no necesariamente ha de dar lugar a una mejor clasificación.

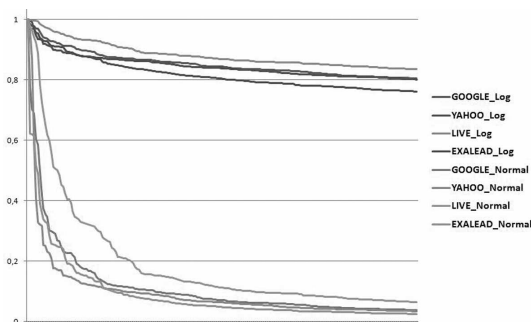
Entre los Rankings de Universidades podemos encontrar que el modelo condiciona la elección de las variables

Tabla 7. Variables principales de los Rankings de Universidades más populares

<----- Orientado a estudiantes			Orientado a investigación ----->		
US News & WR McLeans	THES	Webometrics	Shanghai (ARWU)	Taiwan (HEEACT)	Leiden
Costes	Producción científica				
Opiniones		Visibilidad Web	Impacto Premios	Impacto	
Infraestructuras Servicios	Prestigio	Presencia Web	Excelencia		

Como se ha señalado el número de indicadores cuantitativos disponibles para la descripción de la web es realmente limitado, aunque pueden complementarse con otros ligados a los contenidos y su tipología. Esta es una ventaja evidente de los Rankings pues se pueden combinar posiciones de variables muy distintas con magnitudes heterogéneas entre sí. Obviamente es necesario realizar primero una normalización de los datos para que no influyan los distintos tamaños poblacionales. Puesto que en la mayoría de los casos las distribuciones siguen una ley de potencia ("power law"), una transformación adecuada es la log-normalización que como se demuestra en la Figura 2 es más eficaz que la porcentual (o tanto por uno).

FIGURA 2. Distribución de resultados tras normalización de los datos de acuerdo a dos métodos diferentes.



NORMAL: $B=A_i/\text{MAX}(A_1:A_n)$ LOGNORMAL: $C=\log(a_i+1)/\log(\text{Max}A_1:A_n+1)$

La combinación de variables con sus pesos se puede realizar sobre los valores normalizados o sobre los ordinales. Este segundo caso permite conservar las relaciones entre variables pero puede alterar significativamente las posiciones.

Tabla 8. Posiciones de destacadas Universidades de acuerdo a distintos indicadores individuales (posición=ordinal, frecuencia) y su combinación en un Ranking (ord=suma de ordinales; abs=suma de frecuencias)

NOMBRE	TAMAÑO		VISIBILIDAD		FICH RICOS		SCHOLAR		ORD	ABS
Massachusetts Institute of Technology	1	1,000	2	1,000	1	1,000	8	0,836	1	2
Harvard University	2	0,980	3	0,996	19	0,883	1	1,000	2	1
Stanford University	11	0,924	1	1,000	6	0,953	12	0,819	3	4
University of California Berkeley	3	0,972	4	0,993	2	0,974	25	0,777	4	3
Pennsylvania State University	4	0,970	8	0,946	7	0,934	6	0,855	5	5
University of Michigan	15	0,900	6	0,963	20	0,884	21	0,784	6	8
Cornell University	12	0,923	5	0,983	8	0,940	44	0,752	7	6
University of Minnesota	7	0,935	17	0,927	4	0,979	22	0,781	8	7
University of Wisconsin Madison	8	0,935	11	0,938	9	0,931	36	0,756	9	9
University of Texas Austin	17	0,898	7	0,951	11	0,936	42	0,752	10	10
University of Illinois Urbana Champaign	19	0,891	10	0,940	10	0,936	34	0,758	11	12
University of Pennsylvania	30	0,878	9	0,941	34	0,837	20	0,786	12	15
University of Washington	18	0,893	12	0,935	5	0,970	61	0,721	13	13
Carnegie Mellon University	6	0,942	24	0,912	3	0,996	46	0,748	14	11
Columbia University New York	33	0,877	13	0,932	23	0,890	33	0,759	15	16
Purdue University	9	0,934	31	0,900	12	0,946	66	0,715	16	17
University of California Los Angeles	32	0,877	20	0,922	25	0,864	67	0,715	17	19
University of Florida	16	0,898	29	0,901	18	0,891	86	0,697	18	22
University of Chicago	69	0,859	15	0,930	83	0,820	3	0,913	19	14
University of Maryland	68	0,860	30	0,901	16	0,896	26	0,769	20	20

En el caso de los Rankings Web producidos por el Laboratorio de Cibermetría del CSIC, el modelo por el que se ha optado está basado en el Factor de Impacto, en el que actividad científica (trabajos publicados) e impacto bibliométrico (citas recibidas) tienen el mismo peso. El Webometrics Rank (WR) otorga un 50% de peso al volumen de información publicada en la Web y el otro 50% a los enlaces recibidos por dichas páginas. Es decir se mantiene una tasa 1:1 entre tamaño y visibilidad. Un segundo nivel tiene en cuenta los contenidos específicos de una sede web universitaria y el modelo refuerza en sus pesos la contribución de formatos documentales (ficheros ricos) y más específicamente de artículos científicos y materiales afines. Esto obliga a una redistribución de pesos del 50% correspondiente a la actividad. La Tabla 10 muestra un escenario más sofisticado con estimaciones informadas y posibles futuros desarrollos.

Tabla 9. Modelo actual y variables adiciones viables para la elaboración de Rankings Web.

Actividad			Visibilidad		
Páginas Web	Total	10-20%	Enlaces recibidos	Total	25-50%
	Subdominios	0%		Factor G	0-25%
Ficheros ricos	Total	10-15%	PageRank		0%
	Combinados			Popularidad	
Scholar	Total	15-25%	Visitas	0%	
	Recientes		Visitantes		

Conclusiones.

La construcción de Rankings Web exige el conocimiento y aplicación rigurosa de las técnicas de análisis documental, identificando correctamente los recursos Web y describiendo de forma cuantitativa sus contenidos. El método más viable hoy en día es la utilización de motores de búsqueda como fuente de información. Esto plantea algunos problemas que deben ser resueltos con cierto grado de flexibilidad asumiendo tasas de error que son aceptables dados los grandes volúmenes de datos involucrados.

Sin embargo las decisiones sugeridas respecto a variables y pesos de las mismas están sujetas a discusión y abiertas a experimentación y modificación en la medida que puedan reflejar mejor la situación real o acomodarse a modelos distintos diseñados a priori. Esta es una interesante vía para la investigación futura, que puede enriquecer considerablemente tanto la disciplina ciberométrica como los resultados y aplicaciones de la metodología descrita.

Bibliografía

- Aguillo, I.F. (1998). Hacia un concepto documental de sede web. *El Profesional de la Información*, 7(1-2):45-46.
- Aguillo I.F.; Granadino B.; Ortega J.L.; Prieto JA (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, 57(10):1296-1302.
- Aguillo, I. F.; Granadino, B. (2006). Indicadores web para medir la presencia de las universidades en la Red. *Revista de Universidad y Sociedad del Conocimiento*, 3(1). <http://www.uoc.edu/rusc/3/1/dt/esp/aguillo_granadino.pdf>
- Aguillo, I.F., Ortega, J.L., Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher Education in Europe*, 33(2-3): 233-244.
- Codina, L. (2000). Evaluación de recursos digitales en línea: conceptos, indicadores y métodos. *Revista Española de Documentación Científica*, 23 (1):9-44.
- Codina, L. (2004). Evaluación de calidad en sitios web: proyectos de estudios sectoriales y realización de auditorías. En *Actas IX Jornadas Catalanas de Documentación*. Barcelona, p. 59-72.

Jiménez Piano, M. (2001). Evaluación de sedes web. *Revista Española de Documentación Científica*, 24 (4):405-429.

Liu NC, Cheng Y, Liu L (2005). Academic ranking of world universities using scientometrics - A comment to the "Fatal Attraction". *Scientometrics*, 64(1):101-109.

Thelwall, M. (2001). A web crawler design for data mining, *Journal of Information Science* 27(5), 319-325.

Thelwall, M. (2004). *Link Analysis: An Information Science Approach*. San Diego: Academic Press. 282 pags. ISBN-10: 0120885530

Van Raan AFJ (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62 (1):133-143.

Van Raan AFJ (2008). Bibliometric statistical properties of the 100 largest European research universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology*, 59(3):461-475.