VALORACION DEL IMPACTO DE LA INFORMACION EN INTERNET: ALTAVISTA, EL "CITATION INDEX" DE LA RED

Josep Manuel Rodríguez i Gairín*

Resumen: En el presente artículo se recoge una metodología para la valoración del impacto de la información en Internet, usando las capacidades de indización y recuperación del buscador Altavista. Se aprovecha el contexto para describir la función de los metaelementos del HTML como mecanismo de estructuración y ordenación de la información. Se discuten las limitaciones y fiabilidad del método y se exponen algunos datos que muestran la producción de páginas WWW a nivel de institución y a nivel nacional, así como su comparación con otros países europeos. Se hace especial hincapié en la posibilidad de medir el impacto de estas páginas en función de las veces que son 'enlazadas' desde páginas externas de manera similar a como funciona el 'Citation Index' del Institute for Scientific Information.

Palabras clave: metaelementos, WWW, impacto de la información, análisis de producción de páginas WWW, Altavista, Science Citation Index.

Abstract: This article resumes a methodology for the Internet information impact assessment by using the Altavista indexation and recuperation capabilities. The context is used to describe the HTLM metaelements function as an arrangement and ordering information system. The method limitations and reliability are discussed and some data that show the production of WWW pages are exposed at institutional and national levels and compared to other European countries. The article emphasizes the possibility to measure the impact of these pages depending on the amount of times that they are linked from external pages, similarly to the way in which the Institute for Scientific Information Citation Index works.

Keywords: metaelements, WWW, information impact, WWW pages analysis production, Altavista, Science Citation Index.

1 Introducción

La información existente en Internet es cada día más abundante y difícilmente estructurable. Los buscadores como Altavista, facilitan enormemente la tarea de localizar información en la red; sin embargo, su uso sin conocer algunas de sus características de funcionamiento, provoca en el usuario una sensación de que el resultado de la búsqueda no se ajusta al esperado. Cuando realizamos consultas en Altavista hemos de considerar el tipo de información que éste incluye, sus mecanismos de indización y las posibilidades de recuperación que ofrece. Pero desde el punto de vista del bibliotecario-documentalista, la posibilidad de catalogar correctamente esos documentos facilitará los mecanis-

* Biblioteques de la Universitat Politècnica de Catalunya. Unitat de Suport en Informació Electrònica. Correo electrónico: josep-m@bupc.upc.es. Página personal: http://escher.upc.es/USR/josep-m/inici.htm.
Recibido: 26-2-97.

mos de estructuración de esos robots y como conclusión final permitirá al usuario su correcta localización.

2 Metaelementos para facilitar la recuperación

Sin duda, la mejor manera de recuperar correctamente la información pasa por una pertinente descripción y catalogación de los documentos. En todos los cursos de edición de HTML se hace gran hincapié en las formas y estilo del contenido del documento, pero pocos hablan de la potencialidad de la información adicional que se puede almacenar en la cabecera del mismo.

Todo documento HTML consta de dos partes básicas. La cabecera, delimitada entre las marcas <head> y </head> y el cuerpo delimitado entre las marcas <body> y </body>. La información almacenada en la cabecera no es visible directamente en el visualizador, sin embargo puede contener datos muy valiosos que son empleados por robots como Altavista. Esta información se almacena en el interior de marcas tipo META cuya estructura es :

<META name="Author" content="Josep Manuel Rodríguez">

El atributo 'name' expresa el tipo de metainformación acumulada (etiqueta), mientras que el atributo 'content' indica el contenido de la misma.

Varios grupos están trabajando en las definiciones de posibles tipos de metainformación. Así por ejemplo tenemos como nombres de metainformación el título del documento (title), descripción (description), identificador (identifier), fecha (date), palabras clave (keywords). A su vez cada atributo, 'name', puede venir definido aplicando distintos esquemas según el contenido de la información. Un ejemplo, un identificador de un documento (identifier) puede seguir el esquema ISBN (identifier.ISBN): en este caso la metainformación se escribiría:

<META NAME="ISBN.Identifier" CONTENT="1234-1234">

Podemos encontrar amplia información de metaelementos y esquemas en (OCLC/ NCSA METAData Workshop Report) (http://vancouver-webpages.com/VWbot/VW-dublin-core.html)

```
< !-- Algunos ejemplos de cabecera de un documento HTML -->

<head>

<META NAME="Title" CONTENT="Valorando el impacto de la información en
Internet: Altavista, el Citation Index de la Red.">
<META NAME="Author" CONTENT="Josep Manuel Rodríguez Gairín">
<META NAME="Description" CONTENT="Una descripción de la página ...">
<META NAME="KeyWords" CONTENT="Metaelementos, WWW, Impacto de la información,
Análisis de producción de páginas WWW, Altavista, Science Citation Index,
Institute for Scientific Information">
<META NAME="Date.HTTP" CONTENT="Wed, 26 Jan 1997 12:50:57 GMT">
<META NAME="Publisher" CONTENT="Revista española de Documentación Científica">
<META NAME="MARC.Language" CONTENT="lengua en formato MARC">
<META NAME="Identifier" CONTENT="un número identificativo">
```

```
<META NAME="ISSN.Identifier" CONTENT="el ISSN de la revista si hubiera lugar">
<META NAME="INSPEC.Subject" CONTENT="aquí pondría una clasificación de INSPEC si
fuera el caso.">
<META NAME="MESH.Subject" CONTENT="aquí pondría una clasificación de MEDLINE si
fuera el caso.">
<META NAME="VW96.ObjectType" CONTENT="Artículo">
<META NAME="Form" CONTENT="HTML3.2">
</head>
```

3 Uso de los metaelementos por parte de los buscadores

Muchos motores de búsqueda emplean la información almacenada en los metaelementos para crear sus índices y permitir al usuario acotar por campos. En la actualidad, diversas bases de datos que utilizan tecnología WAIS están implementándolo. De esta manera, el ruido producido por la indización del texto completo puede ser acotado fácilmente.

En el caso de Altavista, por el momento, sólo utiliza la metainformación almacenada en 'KeyWords' para generar sus índices, y la almacenada en 'Description' para elaborar el breve resumen que presenta en la hoja de resultados.

Incluir en nuestras páginas palabras claves en varios idiomas y que sigan algún índice o tesauro facilitará la labor de recuperación del usuario final. En documentos médicos, el grupo de trabajo antes descrito incluye como recomendación de metaelemento la clasificación de la National Library of Medicine.

```
<META NAME="MESH.Subject" CONTENT=" aquí pondría una
          clasificación de MEDLINE">
```

4 Búsqueda de información en Altavista. Uso de la delimitación por campos

Al realizar búsquedas en la opción avanzada de Altavista, normalmente nos limitamos a utilizar los operadores AND, OR, NEAR, NOT. Sin embargo, peculiaridades de ese robot nos permiten ajustar las búsquedas por campos. No se trata de los campos descritos anteriormente como metainformación sino de las siguientes acotaciones:

anchor:xx	Busca documentos con el texto 'xxx' en el interior de un hiperenlace.
host:xxx	Busca páginas con 'xxx' en el nombre del servidor de Web.
image:xxx	Busca páginas que contengan 'xxx' en la marca de una imagen.
link:xxx	Busca páginas que contengan 'xxx' en los enlaces de la página.
text:xxx	Busca páginas que contengan 'xxx' en cualquier parte del texto visible de la página.
title:"xxx"	Busca páginas que contengan 'xxx' en el título.
url:xxx.yy	Busca páginas que contengan 'xxx' e 'yy' a lo largo de la URL de la página. Equivalente a url:"xxx yy".
domain:xx	Busca páginas que contengan 'xx' en el dominio de la institución.

El uso correcto y razonado de estas limitaciones puede ayudar a perfilar notablemente las búsquedas. Por ejemplo, podemos acotar a documentos almacenados en servidores españoles añadiendo a nuestra expresión de búsqueda: '*AND domain:es*' o buscar posibles imágenes de mapas de España con la expresión 'image :spain'.

5 Altavista: El Citation Index de Internet

Sin duda, el uso de los campos anteriormente descritos recordará a aquellos documentalistas que hemos usado el 'Science Citation Index', una de sus peculiaridades más características e importantes: buscar referencias citadas o, en este caso, documentos enlazados. Las posibilidades son muy amplias:

Buscando documentos que citen una página en concreto:	**link:escher.upc.es/USR/josep-m/publica/altavis.htm**
Buscando documentos que citen alguna de mis páginas:	**link:escher.upc.es/USR/josep-m**
Eliminando las autocitas:	**not url:escher.upc.es.josep-m**
Buscando los documentos incluidos de mi servidor	**host:escher.upc.es**
Comparando con el total de documentos de mi institución	**host:upc.es**

El número de accesos a nuestras páginas (medibles con la inclusión de contadores o a través de programas estadísticos) es un parámetro importante pero debe ser complementado con el número de páginas externas que 'las citan'. Este sin duda sería el primer paso para valorar el impacto de nuestra información en la red.

Uno de los datos más valorados a nivel de instituciones como las universidades es el impacto de sus publicaciones recogido en el JCR (Journal Citation Report) que publica anualmente el Institute for Scientific Information. A nivel de productividad en Internet, Altavista puede darnos unos datos que reflejen, no sólo el volumen de páginas que nuestra institución tiene en la red, sino también el impacto que éstas han tenido (medido por el volumen de páginas externas que las enlazan). Una ecuación como la siguiente podría ser el primer punto de partida.

1. host :upc and domain:es
2. (link :upc.es/ or link :upc.es :) and not (host :upc and domain:es)
3. 2 / 1

La descripción del uso de los operadores y la estructura será comentada posteriormente en el análisis del impacto de las páginas WWW del país.

El resultado de dividir las páginas que nos citan (descontados los autoenlaces) por el número total de páginas de que dispone nuestra institución puede ser un buen índice del impacto que nuestra institución tiene en Internet.

Este parámetro por sí mismo puede tener significación si lo comparamos con otro valor, como podrían ser los mismos datos aplicados a todas las páginas del dominio ES. En la siguiente tabla se recogen los valores obtenidos el 26 de enero de 1997.

	Páginas	Enlaces	Enlaces sin autoenlaces	Relación enlaces/páginas
.ES	205,250	131,951	67,851	0,33
UPC.ES	5,217	6,067	3,318	0,63
	2,54%	4,59%	4,89%	

El valor de relación enlaces/páginas=1 correspondería a una cantidad idéntica de enlaces respecto de páginas (por ejemplo, que cada página hubiera sido citada externa-

mente por otra). Los datos recogidos indican que, a nivel español, de 205.250 páginas indizadas, sólo 67.851 páginas externas las citan, mientras que en el caso de la Universitat Politècnica de Catalunya de 5.217 páginas producidas (2.54 % del total nacional) hay 3.318 páginas externas a la UPC que las citan (o sea, aproximadamente una de cada dos es citada externamente).

Un estudio similar a nivel europeo también puede posicionar la producción española en comparación con otros países. La tabla siguiente recoge algunos de estos datos. En este caso conviene ser prudente y tener muy presente la metodología para la recogida de datos.

La ecuación empleada es la siguiente:

páginas de un país : domain:es

Se descartó emplear 'host :.es' ya que recogía datos de países externos como (http :server.es.udm.uk/). Este sistema no contempla las páginas cuyo dominio sea .com.

enlaces en general : link :.es/ or link :.es :

En este caso se es consciente que se pierden aquellos posibles enlaces citados en la página que pudieran incluir el nombre del servidor simplemente. Sin embargo, se prefirió efectuar de esta manera por considerar que lo valorable es la cita de páginas concretas no del servidor.

Enlaces externos (sin autocitas) : (link :.es/ or link :.es) and not (domain:es)

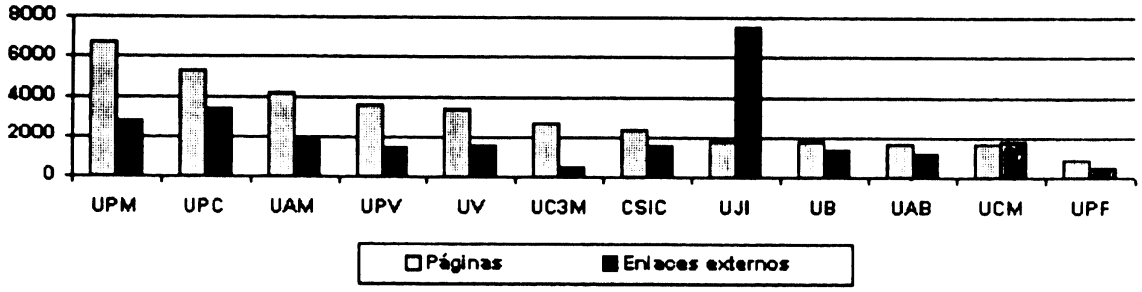
Se efectuaron diversas pruebas y la valoración de los resultados nos lleva a afirmar que la ecuación no se ajusta totalmente a la realidad, pero que las diversas combinaciones no provocan diferencias significativas en el cociente final que nos interesa.

Los datos obtenidos fueron :

	Páginas	Enlaces	Enlaces externos	E. externos /páginas
Francia	452,265	400,81	198,425	0,43
Reino Unido	1.134,989	1.150,25	600,986	0,52
Alemania	594,647	513,50	338,698	0,56
Italia	389,062	382,00	239,619	0,61
Portugal	58,082	77,15	41,361	0,71
España	205,250	131,86	67,851	0,33

Si estos datos son en algún momento valorables, nos indican que el impacto de la producción española a nivel mundial está bastante por debajo del resto de países europeos comparados. El presente estudio simplemente pretende mostrar la metodología; evidentemente, una afirmación como la anterior requeriría una valoración de todos los países europeos. Por ello encarezco al lector lo considere como un ejemplo de la metodología. Quiero pensar que el idioma pudiera ser un factor decisivo ya que no conviene olvidar que Estados Unidos continúa siendo con mucho el principal productor de páginas.

Un estudio igualmente interesante hace referencia a la producción de las universidades españolas y su impacto. La gráfica adjunta recoge los datos obtenidos :



	UPM	UPC	UAM	UPV	UV	UC3M	CSIC	UB	UJI	UAB	UCM	UPF
Páginas	6.708	5.774	4.109	3.577	3.323	2.634	2.371	1.859	1.859	1.720	1.706	869
Enlaces	2.742	3.695	1.809	1.528	1.610	530	1.620	1.455	7.483	1.180	1.861	484
Cociente	0,40	0,63	0,44	0,42	0,48	0,20	0,68	0,78	4,02	0,68	1,09	0,55

Sin entrar en valoraciones específicas, la Universitat Jaume I y la Complutense de Madrid, son las únicas cuyo cociente supera 1, es decir, se las enlaza más veces que páginas propias poseen. Es de destacar la UJI, sin duda por la proyección de sus famosos mapas de recursos.

No se han considerado centros cuyo nivel de páginas fuera inferior a 1000; en esos casos es posible que en algunos de ellos el cociente sea mayor de 1, pero este dato pudiera estar mediatizado por todos aquellos índices o listados de universidades que sin duda llegan al millar.

6 Consideraciones a este modelo de valoración

Sobre el papel, estos datos son muy atractivos aunque las pruebas efectuadas plantean toda una serie de detalles a considerar.

— Altavista, a pesar de ser uno de los robots más potentes de la red, no recoge todas las páginas.

  • No recoge las páginas accesibles bajo suscripción y bloqueadas por password.
  • No recoge las páginas aisladas, desvinculadas de la estructura general del servidor, a menos que sean citadas desde una página externa.
  • No recoge las páginas ni servidores que contengan el protocolo de exclusión de robots (http://info.webcrawler.com/mak/projects/robots/robots.html)

— La valoración del impacto de nuestra información en la red no debe limitarse a las páginas WEB. Mucha información valiosa se mueve a partir de mensajes de correo electrónico, listas de distribución o grupos de 'news'. Si bien es posible buscar información de este estilo en Altavista, no es posible buscar por los cam-

pos indicados y por tanto no se puede aplicar esta metodología a este tipo de información. En el caso de búsquedas en grupos de noticias, Altavista permite delimitar por el campo FROM: de esta manera podemos saber las noticias que una institución o país ha introducido en la red pero no su impacto.

— Los dominios '.com' no salen recogidos dentro del país. Este dato si bien en la actualidad aún es irrelevante sobre todo en España, en otros países europeos sí que se observan dominios '.com' que posiblemente corresponden a empresas de aquel país y no estadounidenses.
— La fiabilidad de las búsquedas en Altavista sería bastante discutible. Por ejemplo, la colocación de espacios entre los operadores booleanos y la marca de campo provocaba resultados diferentes fuera de toda lógica, si bien la diferencia no era altamente significativa (3.500, 3.554, 3.612)
— El número de veces que una página es enlazada por páginas externas no siempre es proporcional a la calidad de la información suministrada. La posibilidad de tener documentos en la red y los criterios de calidad de los mismos aún están en sus primeras fases.

7 Conclusiones

La explosión de información existente en la red Internet obliga a plantearse distintas metodologías que analicen la calidad de la misma. Sin duda, cuantificar la presencia de las instituciones en la red es un parámetro importante, pero aún lo es más valorar la repercusión que ésta tiene en el resto de la comunidad Internet. El buscador Altavista, a través de los criterios descritos, nos permite una primera aproximación.

Sin duda esta metodología debería aplicarse a otros motores como Excite o Webcrawler; sin embargo resulta difícil encontrar información sobre el funcionamiento avanzado de los mismos y, a través de las ayudas en línea, no hemos conseguido determinar posibilidades de limitación por este tipo de campos.

De la misma manera, y al igual que ocurre con el Journal Citation Report, este parámetro debe considerarse como uno más a incluir en la valoración del impacto de la información en la red. Su peso específico dependerá de los criterios aplicados y la necesidad en cada caso concreto.

8 Bibliografía

Koster, Martijn <m.koster@webcrawler.com> "A Standard for Robot Exclusion". 1994. <http://info.webcrawler.com/mak/projects/robots/robots.html> (visitado 23 enero 1997).

"The Dublin Core METAdata Elements" <http://vancouver-webpages.com/VWbot/VW-dublincore.html>

Stuart Weibel, Jean Godby, Eric Miller "OCLC/NCSA Metadata Workshop Report" . 1995 < http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html> (visitado 26 enero 1997)

"The META tag: Controlling how your Web page is indexed by AltaVista." , "Advanced query help" < http://www.altavista.digital.com/cgi-bin/query?pg=ah>

Información sobre el Institute for Scientific Information y el Journal Citation Reports puede encontrarse en <http://www.isinet.com/>