Análisis de la persistencia y del estado de páginas web en los resultados de Google

José Luis Ortega, José Antonio Prieto, Natalia Arroyo, Víctor M. Pareja, Isidro F. Aguillo

Laboratorio de Internet
Centro de Información y Documentación Científica (CSIC)
Joaquín Costa, 22
28002 Madrid
Tfno: 915635482 Fax: 91-5642644
{jortega, joseaprieto, narroyo, vmpareja, isidro}@cindoc.csic.es

Resumen

El propósito de este trabajo es estudiar la persistencia y el estado del buscador Google a través de los documentos recuperados en tres consultas. Los resultados de estas consultas han sido monitorizados durante 15 semanas contrastando en cada momento los resultados con los de la primera consulta. Se ha detectado que estas consultas adoptan en el tiempo una distribución logarítmica que se ajusta a la fórmula de inactividad radioactiva de los isótopos. Además, se ha encontrado un porcentaje elevado de páginas no operativas (>14%) y un decrecimiento de páginas basadas en programación. Finalmente se concluye que el ajuste de las distribuciones nos han permitido calcular la vida media y predecir la persistencia de las consultas en Google, además de poder detectar un núcleo en los resultados de páginas más estables.

Palabras Clave: Cibermetría, World Wide Web, Buscadores, Google, Persistencia Web

Keywors: Webometrics, World Wide Web, Search Engines, Google, Web Decay, Linkrot

Introducción

Desde los 15 años de existencia de la World Wide Web, diferentes aspectos del crecimiento de la red han sido estudiados. Pennock et. al. (2002) descubrieron que el número de enlaces que recibía una sede web se incrementaba a medida que más enlaces recibía, siguiendo un ritmo exponencial. Según Internet Systems Consortium (2004) los dominios web están creciendo desde 1994 a un ritmo similar. Sin embargo el Proyecto

de Caracterización Web de la OCLC (O'Neill et al., 2003), realizado entre 1998 y 2000, nos advierte que aunque la WWW en general sigue creciendo, el ritmo de crecimiento de sedes web se ralentiza hasta llegar, en el periodo 2001-2002, a un decrecimiento del 1%. Pero a pesar de todo, existe un cierto vacío en la bibliografía respecto al decaimiento o desaparición de páginas web en Internet. Podemos decir que Harter y Kim (1996) fueron los primeros en documentar la efímera naturaleza de la web, ellos detectaron que un tercio de las citas electrónicas en revistas electrónicas no estaban disponibles. Lawrence et al. (2001) también estudiaron la problemática de las citas electrónicas llegando a resultados similares. Koehler (1999, 2002, 2004), uno de los autores mas preocupados con este tema, monitorizó 360 páginas y 343 sedes web durante diversos años, encontrando que en 2001 las páginas operativas se habían reducido a un 34,4% y en 2003 a un 33,8%. Nelson y Allen (2002) testaron los contenidos de diversas bibliotecas digitales a lo largo de un año encontrando tan solo un 3% de objetos no disponibles (*linkrot*), aunque ellos mismos nos advierten que este medio es más estable que la web en general y que sus resultados deben tomarse con precaución. Recientemente, Bar-Ilan (2004a) consultó "informetrics" a los más importantes buscadores durante 5 años, con la intención de estudiar la evolución de dicha disciplina en la web a lo largo del tiempo, descubriendo un grado de desaparición del 40%.

Por otro lado diversos trabajos han tratado de evaluar el estado de los buscadores estudiando la estabilidad de sus índices, su cobertura y sesgos comerciales o geográficos. A este respecto Vaughan y Thelwall (2004) encontraron un sesgo en la cobertura de los buscadores de páginas estadounidenses con respecto a páginas asiáticas, concluyendo que este desequilibrio se debe más a la ventaja temporal de EEUU, que a una realidad intencionada. También Bar-Ilan et al. (2004b) encontró que el ranking de resultados en los buscadores no variaba mucho a lo largo del tiempo, encontrando los índices de Google los más dinámicos debido a la actualización asíncrona de sus índices. Por último, existen diversos sedes web que periódicamente nos reportan informes sobre el estado de los buscadores desde distintos puntos de vista (Search Engine Watch, 2004; Search Engine Showdown, 2004).

Objetivo

El propósito de este trabajo es estudiar la estabilidad de los índices del buscador Google (google.com), a través de los documentos recuperados en tres repetidas consultas realizadas a lo largo de 15 semanas (del 29 de enero de 2004 al 6 de mayo de 2004). Con ello se pretende conocer la permanencia y estabilidad de las páginas recuperadas, su desaparición y sus motivos. Además se pretende evaluar el comportamiento de los índices de los buscadores desde el punto de vista de la estabilidad y permanencia de las páginas web.

Metodología

Tres consultas ("*alhambra de granada*", "*mezquita de cordoba*" y "*catedral de jaen*") han sido realizadas al buscador Google, durante 15 semanas (del 29 de enero de 2004 al 6 de mayo de 2004). Se han elegido tres consultas iguales en su formulación y que no expresaran una realidad muy cambiante, como podían ser los temas de actualidad.

Inicialmente, se opto por contar también con los resultados de otros dos buscadores, AltaVista y Alltheweb, pero a mitad del experimento dichos buscadores dejaron de actualizar sus bases de datos y poco después perdieron dichas bases de datos en favor del nuevo buscador Yahoo! Search (Pandia, 2004).

De estos resultados se ha realizado un seguimiento a las páginas obtenidas en la primera consulta, intentando ver su permanencia en los índices del buscador durante las siguientes 14 consultas. Para ello se ha utilizado el programa Web Data Extractor 4.0 (2003) el cual nos permite descargar en un fichero de texto los resultados de las consultas. De los resultados obtenidos sólo se pudieron extraer un promedio de 500 resultados por consulta, salvo en el caso de "*catedral de jaen*" que se obtuvieron sólo 230 resultados. Esto es debido a que dicho programa implementaba por defecto sólo 500 resultados por cada buscador.

En este caso, debido a que el interés está en la permanencia de las direcciones web, sólo se recogieron las URLs sin ningún otro tipo de información, como título, palabras claves, etc. Las consultas se realizaron sin acentos y en minúsculas para evitar la sensibilidad del buscador ante estas leves variaciones en la escritura, además se utilizaron las comillas para buscar sólo la cadena de texto.

Con la intención de estudiar el estado en que se encuentran estas URLs a lo largo del tiempo, se han comprobado las URLs extraídas anteriormente. Para esta tarea se ha contado con el programa Xenu's Link Sleuth (2004) que permite conocer las URLs que están activas o no. Debido al posible estado de la red y al trafico existente en cada momento, las direcciones fueron comprobadas en dos momentos.

Resultados

Permanencia

En un primer lugar se pude destacar que el número de documentos que permanecen en cada consulta, originarios de la primera, desciende describiendo una curva logarítmica inversa muy ajustada en las dos primeras consultas ($R^2 = 0,97$). Hemos comprobado que este comportamiento se asemeja al descenso de la actividad radioactiva de un isótopo a lo largo del tiempo (U.S. Department of Defense, 1996), para ello se debe calcular la vida media de las consultas. La vida media de una consulta se podría definir como el tiempo transcurrido desde la observación original hasta el momento en que sólo recuperamos la mitad de resultados de la consulta original (Spinellis, 2003). La siguiente tabla I muestra la vida media de las tres consultas.

Consultas	*Vida Media*
alhambra de granada	0,0479
mezquita de cordoba	0,0579
catedral de jaen	0,227

Tabla I. Vida Media en años de las consultas observadas

A continuación se calcula la persistencia de los resultados con la siguiente fórmula:

$$R_t = R_0 e^{(-\lambda t)}$$

Donde

$$\lambda = \frac{-0,693t}{T_{1/2}}$$

Los términos de la fórmula serían:

R_t = Persistencia de resultados después del intervalo t

R_0 = Resultados de la primera consulta

e = base del logaritmo natural (2,718...)

t = tiempo transcurrido en años

$T_{1/2}$ = Vida Media de la consulta

En los gráficos 1 y 2 se puede apreciar en escala logarítmica los valores observados y los calculados con la fórmula matemática.

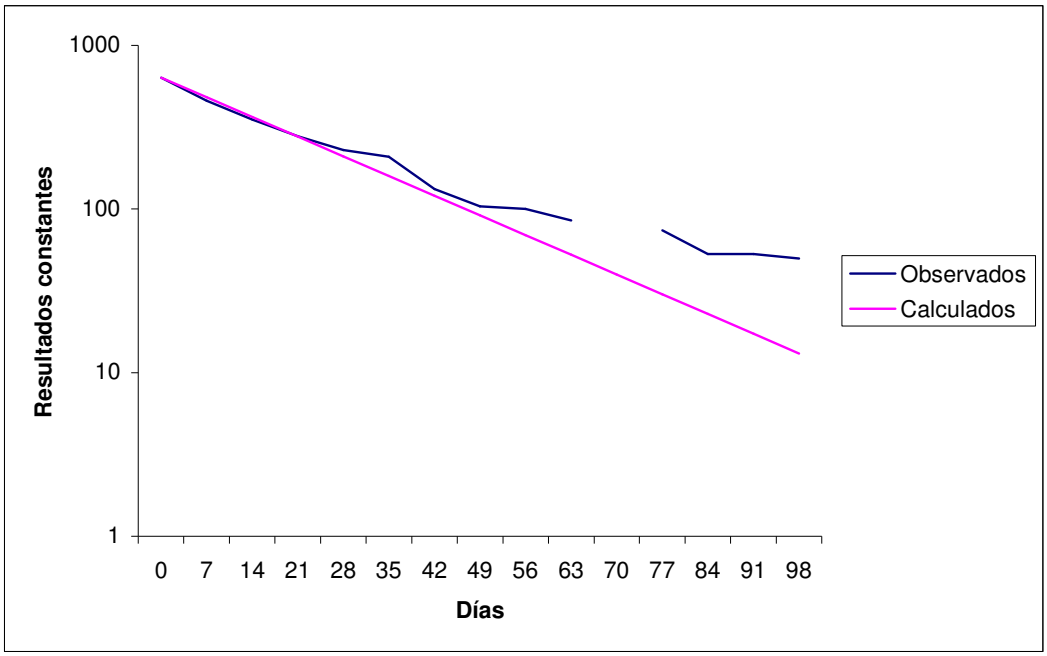

Gráfico 1. Consulta "*alhambra de granada*" a escala logarítmica

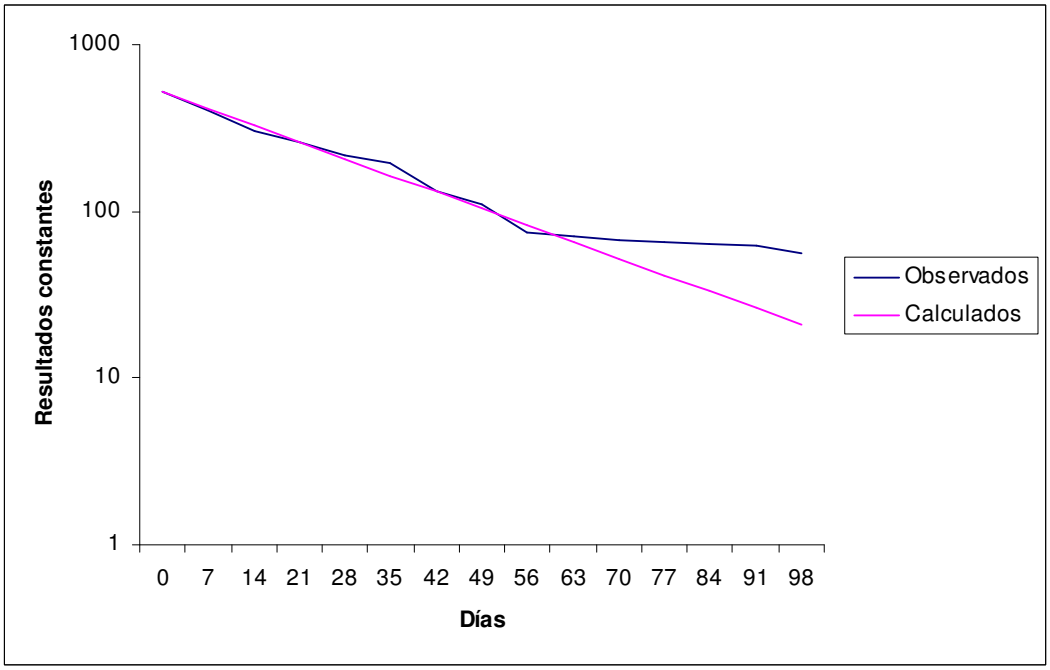

Gráfico 2. Consulta "*mezquita de cordoba*" a escala logarítmica

Como se puede ver, en las dos consultas existe un momento en que la distribución deja de ser logarítmica (y no calculada por la fórmula) y toma una tendencia lineal. Creemos que puede deberse a un núcleo de páginas más estables que permanecen siempre fijas en toda consulta, cuya persistencia es más prolongada a lo largo del tiempo.

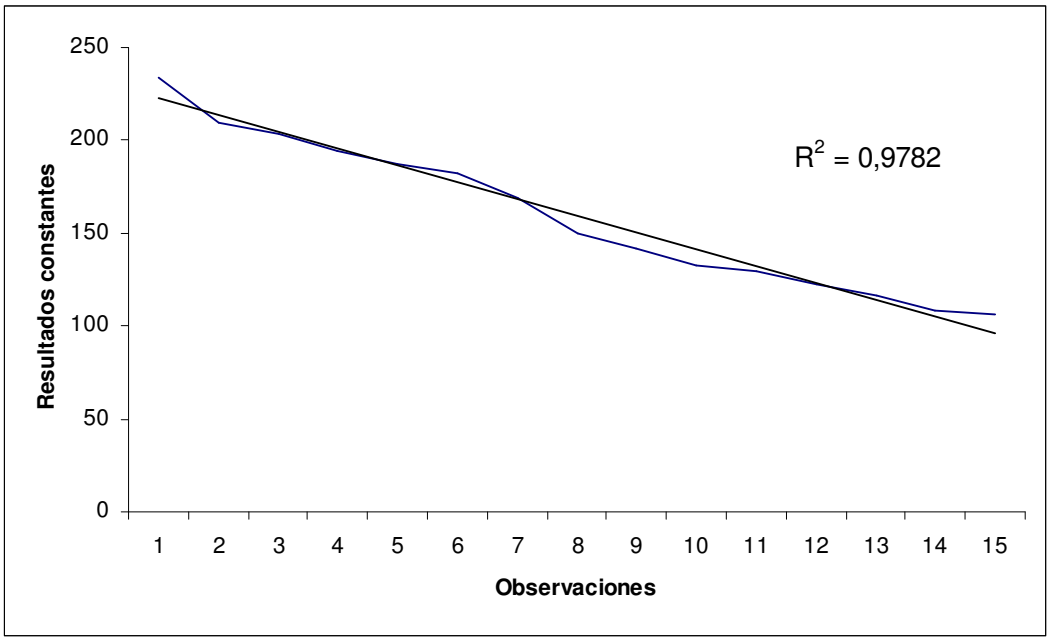

Gráfico 3. Consulta "*catedral de jaen*"

Quizás, en la tercera consulta *"catedral de jaen"*(Gráfico 3), que adopta una tendencia lineal, ese núcleo sea más visible ya que es la consulta que menos páginas recupera. La amplitud de la vida media ($T_{1/2}$ = 0,227 años) nos confirma una mayor estabilidad de los resultados de esta consulta.

Por otro lado, hemos encontrado que de este grupo de páginas el porcentaje de URLs constituidas por documentos HTML o dominios completos aumenta frente a las URLs formadas por programación, tanto cliente (Script) como servidor (PHP, ASP). En la Tabla II se puede ver el porcentaje en que descienden las páginas basadas en programación a lo largo de las 15 observaciones.

Consultas	*Programación*
alhambra de granada	1,2%
mezquita de cordoba	3,67%

catedral de jaen	0,96%

Tabla II. Porcentaje medio de desaparición de páginas basadas en programación.

En ella se puede suponer que gran parte del decaimiento está provocado por la inestabilidad de la programación dinámica y que la estabilidad a la que tiende las distribuciones anteriores la constituye dominios y páginas no basadas en programación. Quizás por esto, el bajo porcentaje de lenguaje de programación en el caso de la tercera consulta repercuta en la estabilidad de sus resultados.

Estado

Consultas	*Forbidden Request*	*Not Found*	*OK*
alhambra de granada	1,99%	12,24%	83,7%
mezquita de cordoba	9,08%	18%	71,26%
catedral de jaen	6,78%	15,77%	72,78%

Tabla III. Porcentaje medio del estado de las consultas.

Por otro lado, el estudio del estado de las páginas indizadas en Google ha permitido detectar un porcentaje, a nuestro entender, muy elevado de páginas no operativas, ya sea Not Found (404) o Forbidden Request (403). En las tres consultas el porcentaje conjunto oscila entre el 14,2% de "alhambra de granada" y el 27% de "mezquita de cordoba". En las tres consultas este porcentaje permanece estable a lo largo de la observación, aunque en el caso de "mezquita de cordoba" tiende a aumentar a lo largo de las 15 semanas.

Discusión y Conclusiones

Por un lado se ha estudiado la permanencia de los índices y por otro el estado en que se encuentran dichas páginas recuperadas en el buscador. No deben confundirse ya que como se ha podido ver, los índices están recogiendo páginas no operativas con un porcentaje elevado.

Creemos que Google no trabaja bien en este aspecto, ya que sus índices están diseñados para el posicionamiento pero no para detectar páginas que han desaparecido. Mas aún, pensamos que el número de páginas no operativas es mayor ya que actualmente se puede encontrar muchos servidores que desarrollan páginas *soft 404*, esto es, mensajes

de error que devuelven un 200 (OK) en vez de 404 (Not Found). Estos servidores, al recibir la petición de una página que ya no está operativa, devuelven una pagina sustitutiva que lleva escrito el mensaje de error o redirecciona a la página principal. Bar-Yossef et al. (2004) estiman que un 25% de los mensajes 200 (OK) en verdad son *soft 404*. Pese a todo creemos que estos resultados son preocupantes ya que se puede pensar que más del 14% (306 millones) de las páginas indizadas por Google no están operativas. Aún así, generalizar los resultados de tres consultas con los 4.000 millones de páginas indizadas por Google puede ser algo temerario, por este motivo animamos a nuevos trabajos que permitan contrastar nuestros resultados.

Por otro lado, se ha podido comprobar que las páginas basadas en programación son más inestables y tiende a desaparecer a lo largo del tiempo de los índices del buscador. Con ello se puede decir que Google tiene muy en cuenta este hecho y que prefiere mantener a lo largo del tiempo formatos más estables y que provoquen escasa variabilidad en sus índices.

Finalmente, se ha podido ver que la permanencia en los índices de las páginas web con respecto a la consulta primera desciende de forma logarítmica, en los dos primeros casos, y lineal en el tercero. El ajuste encontrado en los dos primeros casos en muy elevado lo que ha permitido calcular con una fórmula matemática el comportamiento de los índices de Google respecto a estas consultas. Este hecho permite medir el grado de permanencia de las consultas a través de la Vida Media, y con ella predecir la estabilidad de estas a lo largo del tiempo.

También ha permitido detectar cierto grupo de páginas que no se ajustan a este modelo logarítmico y que, como en el caso de la tercera consulta, tienden hacia un ajuste lineal. Por ello suponemos que en toda consulta existe un núcleo de páginas que permanecen más estables en la Web. Por este motivo y para futuros trabajos, el experimento se ha alargado, no solo con estas consultas sino con otro tipo de muestras con la intención de estudiar si este fenómeno también se reproduce en otros entornos web (dominios, sedes, etc.).

Bibliografía citada

Bar-Ilan, J. ; Peritz, B.C. (2004a) "Evolution, Continuity, and Disappearance of Documents on a Specific Topic on the Web: A Longitudinal Study of 'Informetrics". *Journal of the American Society for Information Science and Technology*, 55 (11): 980-990

Bar-Ilan, J.; Levene, M.; Mat-Hassan, M. (2004b) "Dynamics of Search Engine Rankings – A Case Study". En*: WWW2004, May 17-22, 2004, New York, USA*

Bar-Yossef, Z.; Kumar, R.; Broder, A. Z.; Tomkins, A. (2004) "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay". *WWW2004, May 17-22, 2004, New York, USA*.

Harter, S.; Kim, H. (1996). "Electronic journals and scholarly communication: a citation and reference study". *Information Research* 2 (1) paper 9. http://informationr.net/ir/2-1/paper9a.html [Consulta: 02/09/2004]

Internet Systems Cosortium, Inc. (2004). Redwood, CA http://www.isc.org/index.pl?/ops/ds/ [Consulta: 02/09/2004]

Koehler, W. (1999) "An Analysis of Web page and Web site constancy and permanence". *Journal of the American Society for Information Science*, 50, (2), 162-180

Koehler, W. (2002) "Web page change and persistence – a four-year longitudinal study". *Journal of the American Society for Information Science and Technology*, 53 (2), 162-171

Koehler, W. (2004) "A longitudinal study of Web pages continued: a report after six years". *Information Research*, 9 (2) paper 174. http://informationr.net/ir/9-2/paper174.html [Consulta: 02/09/2004]

Lawrence, S.; Coetzee, F.; Glover, E.; Pennock, D.; Flake, G.; Nielsen, F.; Krovetz, B.; Kruger, A.; Giles, L. (2003) Persistence of Web References in Scientific Research. IEEE Computer, 34(2): 26-31 http://www.neci.nec.com/~lawrence/papers/persistence-computer01/persistence-computer01.pdf [Consulta: 20/09/04]

Nelson, M.; Allen, B. (2002) "Object persistence and availability in digital libraries". *D-Lib Magazine* 8 (1). http://www.dlib.org/dlib/january02/nelson/01nelson.html [Consulta: 02/09/2004]

O'Neill, E. T.; Lavoie, B.F.; Bennet, R. (2003). "Trends in the Evolution of the Public Web 1998-2002". *D-Lib Magazine* 9 (4). http://www.dlib.org/dlib/april03/lavoie/04lavoie.html [Consulta: 02/09/2004]

Pandia. (2004) "The death of AltaVista and AlltheWeb". Pandia Search Central: Oslo http://www.pandia.com/sw-2004/08-yahoo.html [Consulta: 02/09/2004]

Pennock, D.; Flake, G.W.; Lawrence, S.; Glover, E. J.; Giles, C. L. (2002) "Winners don't take all: Characterizing the competition for links on the web". *Proc. Natl. Acad. Sci. USA*, Vol. 99, Issue 8, 5207-5211, April 16, 2002 http://www.pnas.org/cgi/reprint/99/8/5207 [Consulta: 02/09/2004]

Vaughan, L.; Thelwall, M. (2004). "Search engine coverage bias: evidence and possible causes". *Information Processing & Management*, 40(4), 693-707 http://www.scit.wlv.ac.uk/~cm1993/papers/search_engine_bias_preprint.pdf [Consulta: 02/09/2004]

Search Engine Showdown: The Users' Guide to Web Searching. (2004) Notes.com: Bozeman, MT, U.S. http://www.searchengineshowdown.com/ [Consulta: 02/09/2004]

Search Engine Watch: Tips About Internet Search Engines & Search Engine Submission. (2004) Danny Sullivan: [U.S.] http://searchenginewatch.com/ [Consulta: 02/09/2004]

Spinellis, D. (2003) The Decay and Failures of Web References. Communications of the ACM, 46(1) http://www.eltrun.aueb.gr/papers/urlcite.pdf [Consulta: 20/09/2004]

U.S. Department of Defense (1996) FM 8-9 NATO Handbook on the Medical Aspects of NBC Defensive Operations AMedP-6(B). Washington http://www.fas.org/nuke/guide/usa/doctrine/dod/fm8-9/toc.htm [Consulta: 20/09/04]

Web Data Extractor [on-line] Ver. 4.0 [s. l.]: WebExtractor System, c2002-2003. Software. http://www.webextractor.com/ [Consulta: 02/09/2004]

Xenu's Link Sleuth [on-line] Ver. 1.2f [s. l.]: Tilman Hausherr, c1997-2004. Software. http://home.snafu.de/tilman/xenulink.html [Consulta: 02/09/2004]