B.12. Datos numéricos en los listados de resultados de los buscadores

Por **Cristòfol Rovira**

Rovira, Cristòfol. "Datos numéricos en los listados de resultados de los buscadores". En: *Anuario ThinkEPI*, 2008, pp. 76-78.

Resumen: *Análisis de los datos numéricos que proporcionan los listados de resultados de los buscadores usando el coeficiente de correlación. Se llega a la conclusión de que son suficientemente fiables como para ser usados en investigaciones que comparen o evalúen sedes web.*

Palabras clave: *Buscadores, Motores de búsqueda, Inconsistencias en los buscadores, Listado de resultados en buscadores, Fiabilidad de los buscadores, Datos numéricos de los buscadores.*

Title: Numeric data in search engine results pages

Abstract: *The numeric data provided by search engine results pages is analyzed using correlation coefficients. The conclusion is that numeric data is reliable for studies that compare or evaluate web sites.*

Keywords: *Search engines, Reliability of search engines, Inconsistency of search engines, Search results page, Search engines numeric date.*

LOS PROFESIONALES DE LA INFORMACIÓN que usamos los buscadores como herramienta profesional sabemos que debemos usar con mucha precaución los datos cuantitativos que aparecen en el listados de resultados sobre el número de páginas encontradas.

Por otro lado, hay algunos detalles sospechosos, por ejemplo, en *Google* y *Yahoo*, las cifras altas son siempre redondas. En ambos buscadores se indica que son aproximadas. En cambio, en *Live* no aparecen nunca cifras redondas aunque siempre son menores que en *Google* o *Yahoo*.[1]

Los propios buscadores no dan información clara sobre estas disfunciones, y los especialistas especulan, sin que haya consenso, sobre sus causas y su magnitud; como por ejemplo el *post* de *Telendro*, de dudosa rigurosidad, donde se explica cómo calcula *Google* el número de resultados[2].

Datos valiosos

A pesar de estos problemas, los datos numéricos que ofrecen los buscadores son muy valiosos para la cibermetría, en investigación sobre posicionamiento (*SEO*) o en cualquier otro tipo de estudio evaluativo sobre la Red. A menudo en esos ámbitos se usan las cifras de forma relativa para comparar y evaluar diversas sedes web y no como cifras absolutas que se puedan usar en diversos contextos. Aplicando este principio podremos decir que una web, en un momento dado, tiene más visibilidad que otra porque *Google* y/o *Yahoo* dan más resultados en una búsqueda del tipo "link:www.loquesea.com", aunque consideremos que la cifra que ofrece el listado de resultados podría no ser exacta. No parece razonable pensar que el error en esa cifra vaya a afectar de manera distinta a una u otra sede si tomamos la precaución de hacer las diversas búsquedas en las mismas condi-

Los datos numéricos que ofrecen los buscadores son muy valiosas para la cibermetría, en investigación sobre posicionamiento (*SEO*) o en cualquier otro tipo de estudio evaluativo sobre la Red

ciones como, por ejemplo, ejecutándolas casi al mismo tiempo y usando el mismo *datacenter* de *Google*.

En el Grupo de investigación *DigiDoc* (*Universidad Pompeu Fabra*) hemos realizado en los últimos años diversas investigaciones sobre posicionamiento usando, entre otras fuentes, los datos cuantitativos de los buscadores. Estas investigaciones nos han permitido obtener algunas cifras sobre la fiabilidad de los números que dan los buscadores aplicando el coeficiente de correlación.

Coeficiente de correlación

El coeficiente de correlación puede tomar valores desde -1 a 1 y da una idea de hasta qué punto dos variables aumentan (o disminuyen) de manera coordinada, de forma que los valores de coeficiente de correlación cercanos al 1 indican que siempre que una variable aumenta también lo hace la otra. En otras palabras, el coeficiente de correlación da la "probabilidad" de que cuando una variable aumente también lo haga la otra.

Los tres principales buscadores (*Google*, *Yahoo*, *Live*) no suelen coincidir en los números de resultados al hacer la misma búsqueda porque no han indizado exactamente las mismas páginas. No obstante, cabría esperar que los resultados fueran parecidos desde un punto de vista estadístico y que los coeficientes de correlación fueran altos: o sea, que si

Los datos numéricos que proporcionan los listados de resultados de los buscadores son suficientemente fiables como para ser usados con seguridad en análisis que comparen o evalúen sedes web

una web "A" da un mayor resultado que otra "B" en alguna búsqueda (número de enlaces recibidos, número de citas de su *URL*, etc.) también "A" será mayor que "B" en los otros buscadores. Los siguientes datos obtenidos de un grupo de 160 sedes web seleccionadas temáticamente confirman esta idea inicial:

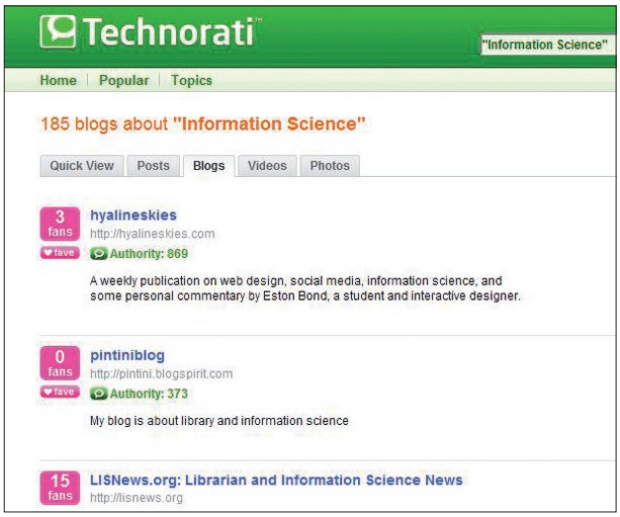
– 0,83 coeficiente de correlación entre citas de las *URL*.

– 0,77 coeficiente de correlación entre citas de las *URL*.

– 0,96 coeficiente de correlación entre citas de las *URL*.

Estos resultados son muy importantes porque refuerzan la fiabilidad de los tres buscadores.

También se obtienen coeficientes de correlación altos con los resultados sobre enlaces recibidos entre buscadores y otros grandes servicios como *Del.icio.us* o *Technorati*, en especial cuando estamos hablando de *Google Blogs*:



– 0,93 coeficiente de correlación entre total de enlaces recibidos según *Google Blogs* y *Technorati*.

– 0,72 coeficiente de correlación entre total de enlaces recibidos según *Google Blogs* y *Del.icio.us*.

No obstante, hay algunos resultados sorprendentes. ¿Cabría esperar un coeficiente de correlación alto entre citas de la *URL* y enlaces recibidos en un mismo buscador? Dicho de otra manera ¿las sedes web cuya *URL* aparece más veces citada en la Red son también las que reciben más enlaces? Parece que no es así, y una de las razones podría ser que las *URL* largas no salgan escritas en los enlaces:

– 0,19 coeficiente de correlación entre citas de las *URL* en *Google* y enlaces recibidos según *Google*.

– 0,07 coeficiente de correlación entre citas de las *URL* en *Yahoo* y enlaces recibidos según *Yahoo*.

Conclusión

En definitiva y resumiendo, de acuerdo con los resultados obtenidos aplicando el coeficiente de correlación, nuestra opinión es que los datos numéricos que proporcionan los listados de resultados de los buscadores son suficientemente fiables como para ser usados con seguridad en análisis que comparen o evalúen sedes web.

Notas

1. *Google Inconsistencies*. searchengineshowdown. 11 de octubre, 2003.
http://searchengineshowdown.com/features/google/inconsistent.shtml#count
Algunos errores detectados en el pasado han sido solucionados, como por ejemplo que "kabul OR kaboul OR kaboel", dé menos resultados que "kabul OR kaboul": **Josu Gómez**. Google y booleanos. *IweTel*. 19 de octubre, 2006.
http://listserv.rediris.es/cgi-bin/wa?A2=ind0610c&L=iwetel&D=1&T=0&O=D&P=6246

O que en la búsqueda por "the" se obtengan hasta 8 billones de resultados, casi tantos como toda la Web: **Jean Véronis**. Web: Google's counts faked. En: *Technologies du langage*. 26 de enero, 2006.
http://aixtal.blogspot.com/2005/01/web-googles-counts-faked.html

2. *Telendro.* Número de resultados para cierta búsqueda. 20 de septiembre, 2006
http://telendro.com.es/2006/09/20/numero-de-resultados-para-cierta-busqueda/