F.2. Analítica de búsqueda: cómo y qué buscan los usuarios

Por **Jorge Serrano-Cobos**

8 noviembre 2010

Serrano-Cobos, Jorge. "Analítica de búsqueda: cómo y qué buscan los usuarios". *Anuario ThinkEPI*, 2011, v. 5, pp. 173-176.

Resumen: *Se comentan ejemplos de frases y términos de búsqueda utilizados por los usuarios del buscador Google, que pueden consultarse gratuitamente en su servicio Google Adwords. El análisis de qué y cómo busca el público puede ofrecer interesantes pautas para decidir los encabezamientos de materias de las bibliotecas, así como para la SEO (optimización de la posición de nuestra web en los resultados de los buscadores).*

Palabras clave: *Google Adwords, Términos de búsqueda, Hábitos, Mejora de la búsqueda.*

Title: *Search analytics: how and what users are looking for*

Abstract: *Some examples of search terms and phrases used by Google users, which are available free of charge on the Google Adwords service, are discussed. The analysis of why and how the public searches for information can offer interesting guidelines for deciding particular subject headings in libraries, and for search engine optimization (SEO) to improve the position of a website in the search engine results pages.*

Keywords*: Google Adwords, Search terms, Habits, Search improvement.*

LAS QUEJAS SOBRE las carencias de los sistemas de búsqueda de los opacs son algo casi tradicional en nuestro entorno profesional[1] (si bien es verdad que además de trabajar en mejorar los algoritmos de recuperación y la presentación de los mismos a los usuarios, podrían mejorarse también los contenidos a recuperar[2]).

Cómo hacerlo es otra cuestión. Las iniciativas se han multiplicado en los últimos años[3] y podemos encontrar, por ejemplo, que se usan:

– *tags* para implantar la recuperación por ojeo o *browsing* en *LibraryThing*[4] o *GoodReads*[5];

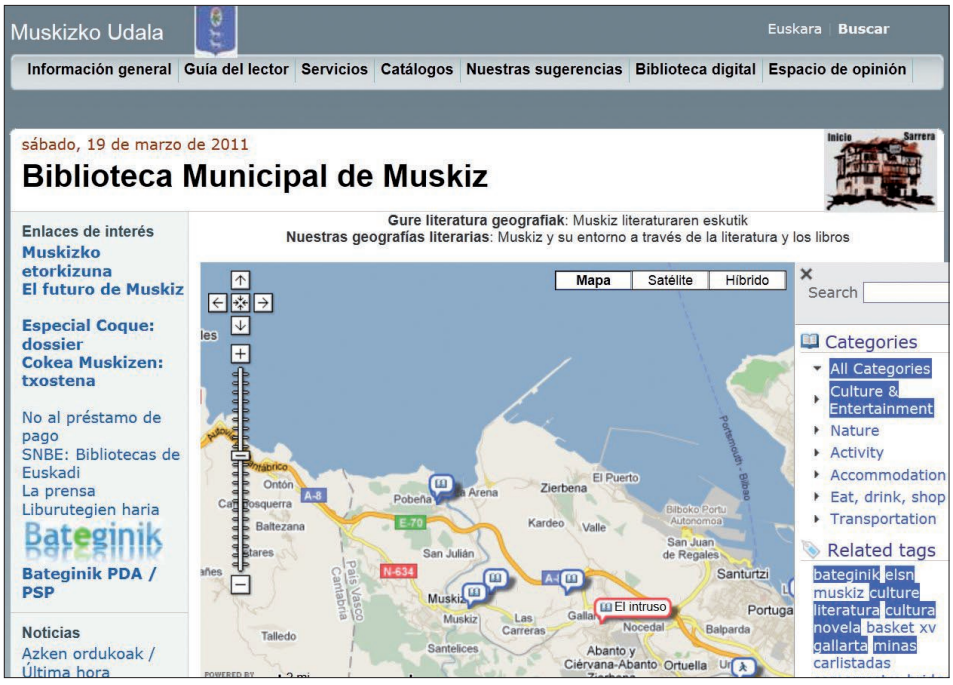
– *rich results*[6] en los resultados de la búsqueda de libros en *Google Books*, infiriendo qué busca el usuario en concreto[7] mediante minería de uso o *web user mining*, con una tendencia a mostrar la última edición (probablemente para promover la compra online);

– *linked data* (dentro del movimiento de la web semántica) en rdf para enriquecer los contenidos de los libros, por ejemplo permitiendo recuperar libros de poemas por tipo de métrica[8] o por caracteres de ficción[9] e incluso libros que hablan en algún momento de un país o una ciudad[10], lo que se puede hacer mediante *mashups* con mapas[11].

Pero a la hora de enriquecer los resultados de una búsqueda es importante entender cómo y para qué buscan nuestros usuarios. Así, sabemos de los 3 tipos de intencionalidad en la búsqueda más conocidos[12]: el 75% de las búsquedas en la Web son informacionales, el 13% de navegación, y un 12% de transacción, aproximadamente[13]. Y también sabemos que en general la mayoría de los usuarios reformula sus búsquedas infructuosas mediante cambios en su contenido[14], aunque harían falta estudios más actuales, una vez popularizados los últimos cambios en la presentación/facetación de resultados de los grandes buscadores de internet.

"A la hora de enriquecer los resultados de una búsqueda, es importante entender cómo y para qué buscan nuestros usuarios"

Otra cosa es que los usuarios y no usuarios de las bibliotecas actúen igual. En cuanto a los no usuarios que buscan en español y en España, mediante técnicas de analítica de búsqueda (*search analytics*[15]) podremos destacar algunos detalles curiosos de algunas de sus cadenas de



http://www.muskiz-liburutegia.org

búsqueda, en este caso simplemente consultando *Google Keywords* (*Adwords*) *Tool*[16]:

– Las búsquedas con errores gramaticales son muy comunes, tanto que en ocasiones se busca más por la suma de los posibles errores que por la palabra clave correcta. Pero *Google* hoy día casi elimina ese problema de las búsquedas en internet al corregir esos errores, y ciertos sistemas bibliotecarios integran software del tipo "quiso decir".

– En otras ocasiones el usuario conoce la enorme variedad de contenidos que se puede encontrar y con su lenguaje natural intenta contextualizar y desambiguar el resultado que busca. Por ejemplo, en búsquedas como "el caballero de la armadura oxidada libros", "cien años de soledad libro", o "don quijote de la mancha libro". Es decir, el usuario faceta o filtra su búsqueda por formato, pero usando su lenguaje.

– En general se usan más los verbos en infinitivo ("comprar libros" más que "compra libros" o "compro libros") pero hay que tener en cuenta que el español es un idioma que usa la forma activa y, al parecer, más aún en España.

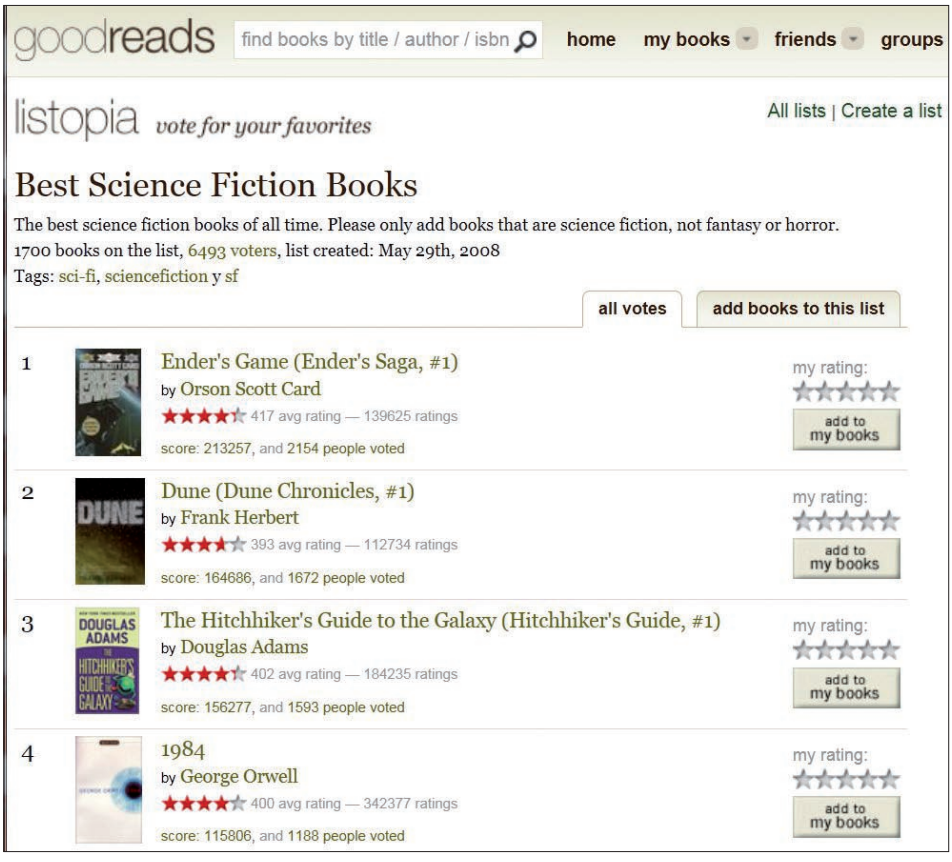
"Es interesante analizar búsquedas como 'autora harry potter' o 'romeo y julieta autor'"

– Dependiendo de lo que se busca, se utiliza más el plural que el singular, o viceversa. Por ejemplo, a la hora de recuperar información general o listas de elementos, se busca más en plural (12.100 veces al mes de media "lecturas para niños", frente a 8.100 veces "lectura para niños"). Sin embargo en el caso de bibliotecas, para ahí comenzar la búsqueda de los ítems que interesan, se usa más el singular, habitualmente acompañado de una localización para desambiguar (2.740.000 veces "biblioteca" frente a 450.000 de "bibliotecas").

– Los sinónimos también deben ser tenidos en cuenta en nuestra búsqueda de la excelencia catalogadora: hemos de preguntarnos, por ejemplo, si los usuarios buscan lo mismo en el caso de "aprendizaje lectura" (2.900 búsquedas de media) que en "enseñanza lectura" (1.600).

http://www.librarything.com

– Deberíamos ser capaces de jugar con las cartas que da el desconocimiento de los usuarios de lo que buscan, que intentan dar un rodeo usando los datos que sí conocen. Así, podemos encontrar a usuarios que buscan "hogar del libro" en lugar de "casa del libro", pero más interesantes –por el problema que pueden acarrear en la recuperación en un motor de búsqueda del catálogo– son las búsquedas como "autora harry potter" (1.900 búsquedas), "romeo y julieta autor" (1.600) o "autor de la eneida" (590). Si el usuario en estos casos lo que busca es una lista de libros de ese autor, o información biográfica del mismo, el opac puede que como mucho les aporte el título buscado; o ni eso si el sistema quiere encontrar todos los términos de la búsqueda.



http://www.goodreads.com

– Por último, es interesante contrastar la polisemia existente entre la intención dada al usar nuestros encabezamientos de materia con el que tiene quien realiza las búsquedas. Por ejemplo, cabe preguntarse cuántas de las 12.100 veces que se busca la expresión "escritores mexicanos" o de las 2.400 al mes que se busca "escritores hispanoamericanos" se hacen con la intención de encontrar uno o más libros que estudien a los escritores mexicanos o hispanoamericanos, o con la de encontrarse con listas de autores con sus obras asociadas.

Al parecer, según *Google* y su uso extenso e intensivo del *crowdsourcing*[17], lo más probable es lo segundo[18].

"Deberíamos ser capaces de jugar con las cartas que da el desconocimiento de los usuarios de lo que buscan"

Referencias bibliográficas

1. **Schneider, Karen G.** "How opacs suck, part 2: the checklist of shame". *ALA TechSource*, 2006.
http://www.alatechsource.org/blog/2006/04/how-opacs-suck-part-2-the-checklist-of-shame.html

2. **Castillo-Vidal, Jesús**. "Descenso del número de visitas a las webs de bibliotecas y opacs". *IweTel*, 29 sept. 2010.
http://www.mail-archive.com/iwetel@listserv.rediris.es/msg04422.html

3. **Serrano-Cobos, Jorge**; **Sellés, Alicia**. "Catálogos online y portales bibliotecarios: ¿un reto para la integración?". *Mi biblioteca*, 2009, n. 19, pp. 70-75.

4. LibraryThing
http://www.librarything.com/work/8653840

5. Goodreads
http://www.goodreads.com/user/new?remember=true

6. **Catacchio, Chad**. "Google Books to get 'rich results' starting today". *The next web* (*TNW*), 1 Nov. 2010.
http://thenextweb.com/google/2010/11/01/google-books-to-get-rich-results-starting-today

7. **Madrigal, Alexis**. "Inside the Google Books algorithm". *The Atlantic*, 1 Nov. 2010.
http://www.theatlantic.com/technology/archive/10/11/inside-the-google-books-algorithm/65422/#

8. Freebase. Poetic verse form
http://www.freebase.com/view/book/poetic_verse_form

9. Freebase. Book character
http://www.freebase.com/view/book/book_character

10. Open Library
http://openlibrary.org/subjects/place:new_york

11. Biblioteca Municipal de Muskiz.
http://www.muskiz-liburutegia.org/mapalit.html

12. **Broder, Andrei**. "A taxonomy of web search". *IBM research*.
http://www.sigir.org/forum/F2002/broder.pdf

13. **Jansen, Jim**. "Classifying the user content of web queries using k-means clustering". *Web search*, 1 Nov. 2010.
http://jimjansen.blogspot.com/2010/11/classifying-user-intent-of-web-queries.html

14. **Young Rieh, Soo**; **Hong (Iris) Xie**. "Patterns and sequences of multiple query reformulations in Web searching: a preliminary study (2001)". En: *Proc. of the 64th Asist annual meeting*, Washington DC, 2001, pp. 246-255.
http://rieh.people.si.umich.edu/papers/rieh_asist2001.pdf

15. **Serrano-Cobos, Jorge**. *Search analytics* [presentación].
http://www.slideshare.net/jorgeserranocobos/search-analytics-2219355

16. Google Adwords
https://adwords.google.com/select/KeywordToolExternal

17. Google "define:crowdsourcing"
http://bit.ly/aZHVXU

18. Google "escritores hispanoamericanos"
http://bit.ly/aTbEW3

IweTel, foro de información y debate de la biblioteconomía y la documentación

Fundada por Tomàs Baiget en 1993, *IweTel* es la lista pionera en español de los profesionales de las bibliotecas, documentación, bases de datos y sistemas de información en general.

Al principio se alojó en *Sarenet* y en 1998 pasó a *RedIRIS*. Posteriormente se han ido creado otras listas más especializadas como *Arxiforum* (archivos), *Bib-Med* (información bio-médica), *Bescolar* (bibliotecas escolares), *Incyt* (indicadores científicos), etc., pero *IweTel*, con más de 5.000 miembros, es la lista de referencia, el medio de comunicación básico y central para los profesionales de la información.

En la lista se cumple la conocida regla del 80/20 (el 80% de los mensajes los genera el 20% de los inscritos), o su reciente reformulación a 90, 9, 1%: el 90% de los inscritos son pasivos, casi nunca envían nada, el 9% (unos 360) participa alguna vez, y existe un 1% (50 personas) que genera la mayoría de mensajes.

Con el aumento de inscritos y el número de mensajes (algunas semanas se distribuyen más de 100) fue necesario hacer la lista moderada, y en ello estamos los 4 firmantes, intentando aplicar nuestro sentido común para decidir cuáles se aprueban y cuáles no, y evitando los mensajes repetidos. Rechazamos alrededor de un 15-20%, lo cual a veces provoca quejas de sus autores, y para dirimir las dudas se creó un Consejo Asesor formado por veteranos de la lista, a quienes los moderadores pedimos consejo.

La lista cumple los dos objetivos básicos típicos: tablón de anuncios (conferencias, cursos, publicaciones, noticias) y foro de debates. Además se usa como sistema abierto de evaluación por pares (*open peer review*) de las notas que los miembros del think tank *ThinkEPI* envían periódicamente a la lista para su pública crítica y discusión. Esas notas y los principales mensajes que generan se publican cada año reeditados en el *Anuario ThinkEPI* de la editorial *EPI*.

Con los cambios tecnológicos habidos a lo largo de estos años y, más recientemente, con las nuevas plataformas web 2.0, se ha planteado muchas veces si las listas de correo se han hecho "obsoletas". La verdad es que pensamos que una lista sigue siendo el medio ideal de comunicación de una comunidad profesional: rápida, limpia, discreta y eficaz, lejos de la faramalla de las redes sociales, también muy interesantes y útiles pero para otras cosas.

Más información e inscripciones:
http://www.rediris.es/list/info/iwetel.html

Javier Leiva-Aguilera (*Catorze.com*), **Paco López-Hernández** (*Universidad Carlos III de Madrid*), **Isabel Olea** (*Universidad de León*) y **Tomàs Baiget** (*El profesional de la información*).