**Menú Principal**

- Inicio
- Editorial
- Contenido
- Keith Raniere
- Reconocimiento
- Contacto
- Búsqueda
- Cultura y Arte

**Edición 72**


## Google, el motor más popular del mundo

# El alma de un buscador de páginas Web

*Doctora Elisa Schaeffer*
*Profesor-Investigador*
*Facultad de Ingeniería Mecánica y Eléctrica / UANL*
*Candidata al Sistema Nacional de Investigadores*

Google sabe todo. Si necesita un restaurante de comida italiana, el nombre de un libro que quie-re comprar, la hora de la función de cine, el hombre moderno le pregunta a Google, el motor de búsqueda más popular del mundo. Prácticamente cada navegador de páginas web incorpora una barra de búsquedas para consultar con Google sobre cualquier tema. Lo que interesa a cualquier proveedor o comerciante es cómo lograr que la página Web de su negocio salga primero en la lista de resultados que muestra el buscador.

Para asegurar un buen lugar en esta lista de resultados, uno necesita saber cómo funcionan Google y sus competidores: cómo coleccionan los datos de los cuales eligen los resultados, en cada consulta, los millones de usuarios; cómo seleccionan las páginas para ofrecer al usuario y cómo ordenan esas páginas seleccionadas. En este artículo nos enfocamos más en Google, pero varios de estos aspectos se aplican también a otros buscadores de la Web.

**NAVEGADORES AUTOMÁTICOS**

La recolección de datos se basa en “navegadores automáticos” que recorren las páginas web. Esos navegadores automáticos se llaman crawler en inglés, un sustantivo que proviene del verbo crawl, lo que se podría traducir como “arrastrarse”. Un crawler comienza su recorrido desde una cierta página asignada y sigue las ligas de esta

página para llegar a más páginas. Una liga, también conocida como un enlace, es simplemente un texto anotado con la dirección de otra página para ofrecer al usuario la posibilidad de ver otra página aparte de la que actualmente está leyendo.

Las páginas visitadas, con toda su información, están guardadas en las computadoras de Google, y las ligas entre ellas están analizadas. Múltiples crawlers recorren la WWW simultáneamente, recopilando así una “imagen” de la estructura y el contenido de ella en un momento dado.

El procesamiento hecho a las páginas obtenidas en esta forma es complejo, y sus detalles cambian de un motor de búsquedas a otro. Sin embargo, el proceso general tiene dos componentes principales e importantes de co-nocer: las palabras que aparecen en una página están procesadas según su frecuencia y posición en el texto. Las pa-labras utilizadas en títulos se consideran más importantes que las palabras incorporadas en el texto y una palabra que se usa mucho es considerada la más representativa del documento que una que se usa unas pocas veces. En este proceso se toma en cuenta la frecuencia global de las palabras en todas las páginas procesadas: si alguna palabra (como por ejemplo las preposiciones y otras pa-labras pequeñas) se usa mucho en todos los documentos, su presencia no se considera relevante para determinar el tema de la página. El procesamiento también cuenta con un diccionario de sinónimos para identificar conceptos en vez de limitarse a nivel de palabras.

**IMPORTANCIA DE LAS PALABRAS**

La selección entre todas las páginas guardadas en las computadoras está basada en la consulta hecha por el usuario; o sea, las palabras clave que utilizó el usuario al lanzar la búsqueda. Es importante reconocer que la búsqueda está realizada en las copias almacenadas en las computadoras de Google, en vez de buscar de forma dinámica en la Web para cada consulta. Las listas de pa-labras clave están comparadas contra las listas de conceptos de las páginas almacenadas para identificar cuáles páginas manejan los mismos conceptos que la consulta.

Google también detecta si una palabra parece estar mal deletreada; si la palabra ocurre en los documentos almacenados con poca frecuencia, pero una variación cercana de la palabra ocurre mucho más, Google pregunta al usuario si realmente quería buscar por la variación en vez de la palabra que definió.

De nuevo, el motor de búsqueda incorpora el uso de sinónimos y ponderación de la importancia de las pa-labras de la consulta según su frecuencia en la colección de páginas. El orden de las palabras clave está interpretado como un orden de prioridad, y típicamente el usuario puede “forzar” la inclusión o exclusión de cierta palabra.

**Para más información:**
**Una guía de uso de Google (en inglés) está disponible en http://www.googleguide.com/**

**Un artículo científico The Anatomy of a Large-Scale Hypertextual Web Search Engine sobre la arquitectura de Google, por los fundadores de la empresa, Sergey Brin y Lawrence Page, está disponible en línea, igual que varios otros artículos sobre el tema.**

**Para buscar publicaciones académicas, es mejor hacerlo en http://scholar.google.com/ que, en vez de juzgar la importancia por ligas, ordena resultados considerando referencias bibliográficas entre textos científicos, como una medida de importancia.**

**IMPORTANCIA DE LAS PÁGINAS**

Para ordenar las páginas encontradas en el "almacén" según su importancia, Google hace matemáticas. Una página se considera "importante", si muchas otras páginas importantes tienen ligas a esa página. Ésta es una definición recursiva de importancia: la importancia está definida en términos de la importancia misma. Matemáticamente se maneja esa definición por crear una matriz que contiene la información de la existencia de ligas en una página A que apuntan a una página B para todos los posibles pares A y B. Esta matriz se convierte a otra matriz que contiene la probabilidad de que un "buscador aleatorio" vaya a una página B si actualmente está viendo una página A.

Esta matriz de probabilidades tiene una estructura especial algebraica que permite interpretar una propiedad matemática suya, específicamente su eigenvector principal, en forma de valores de importancia de las páginas. Es una matriz enorme, por lo cual no existe una computadora con suficiente memoria para calcular ese eigenvector, y aun si existiera, el tiempo de computación no sería nada razonable. Afortunadamente existen métodos aproximados para rápidamente estimar el eigenvector y así obtener valores numéricos de importancia "estructural" para las páginas Web. La lista de documentos seleccionados por las palabras clave de la consulta del usuario está entonces ordenada según estos valores numéricos de importancia, comenzando con la página más importante seleccionada.

**CONSEJOS PARA LOS USUARIOS**

Tomando en cuenta la función de un motor de búsqueda, es posible ofrecer algunos consejos para asegurar que su página reciba la importancia que merece. Es esencial escribir texto claro y bien estructurado sobre los productos, servicios y otros temas relevantes de la empresa o institución en cuestión.

Las palabras utilizadas en los títulos (las directivas <H1>, <H2> etcétera del lenguaje HTML) pesan mucho en el análisis

de la página por el motor de búsqueda. Colocar texto únicamente en figuras e imágenes perjudica la visibilidad de la página en Google, porque ese texto pasará desapercibido por el análisis de contenido. Otro consejo, particularmente apto para optimizar la visibilidad en Google, se refiere a tratar de conseguir ligas que apunten a su página –la importancia de su página proviene de la importancia y la cantidad de las ligas que entran en esa página. Es importante colaborar con los clientes y proveedores, de tal manera que pongan ligas a los sitios de sus colaboradores. También la inclusión en diferentes tipos de directorios en línea puede ayudar a aumentar la importancia relativa de su página.

Para asegurarse de que un crawler visite su página, usted puede solicitarlo directamente a Google por la página http://www.google.com/addurl.html y ellos programarán una visita al sitio. Los crawlers visitan diferentes páginas con diferente frecuencia, buscando mantener las páginas más importantes mejor actualizadas en el almacén. Si su página tiene ligas entrantes desde alguna página que Google ya conoce, tarde o temprano un crawler encontrará su página por las ligas. También las visitas de los usuarios de la Web pueden informar a Google de la existencia de una página –muchos navegadores informan al buscador desde cuál página llegó el usuario a ver la página de Google. Los consejos para una persona que realiza búsquedas en la Web también se basan en función de los motores de búsqueda, como: poner las palabras clave más importantes primero en la consulta, pues esto ayuda a encontrar las páginas más relevantes; utilizar palabras más específicas, lo que puede ayudar a limitar los resultados, si son demasiados. Palabras más generales suelen am-pliar la gama de los resultados si originalmente fueron escasos. Es recomendable evitar palabras que tienen doble significado o que son altamente comunes.

La búsqueda de información, productos y servicios en línea es una parte importante de la vida diaria de una población creciente. Encontrar y ser encontrado tienen un impacto socio-económico más allá de lo que podíamos haber esperado en los primeros años de la WWW.

[Volver]