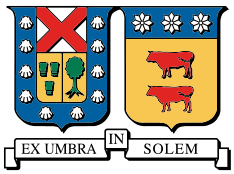

Redes de Computadores

Proyecto: ¿Cómo Funciona Google?

Felipe Fernández Pino
201121011-5

René Fredes Verdugo
201121031-K

Felipe Gil Silva
201121019-0

28 de septiembre de 2015

1. Resumen

El proyecto abarca el funcionamiento general del buscador web Google. Primero se estudia el proceso de rastreo e indexación permanente que realiza Google sobre todos los sitios web encontrados a partir de los ya identificados. Luego se mencionan los principales algoritmos de búsqueda que Google utiliza para mejorar los resultados y personalizarlos de acuerdo a los intereses de cada usuario. Finalmente se realizan unas demostraciones sobre las características de personalización de Google y las restricciones de búsqueda que imponen algunos sitios de importancia.

2. Introducción

En el presente informe se detalla sobre la exposición realizada, llamada: ¿Cómo funciona el buscador Google?. Para cualquier navegante de la red resulta casi imperativo encontrarse en algún punto con este buscador (o algún otro, pero en notable menor medida), dado que suele ser la primera opción para cualquier *plug-in* de búsqueda y además su gran cantidad de aplicaciones ya implementadas en los diferentes dispositivos computacionales existentes en la actualidad.

Este proyecto busca responder a varias preguntas que cualquier usuario puede hacerse muy rápidamente luego de realizar una búsqueda cualquiera, como por ejemplo: ¿De dónde Google obtiene la información buscada?, ¿cómo decide qué resultados entregar?, ¿qué tipo de resultados son discriminados?, etc.

El lector observará que para responder a las preguntas propuestas se llegará a los algoritmos de rastreo e indexación y algoritmos de búsqueda, destacándose las palabras claves: `Googlebot` y `PageRank` respectivamente.

3. Rastreo e Indexación

Seguramente, lo que cada usuario de Google piensa cuando realiza una pregunta es que el motor de búsqueda buscará, valga la redundancia, por toda web el contenido especificado, pero lo que en realidad busca en una especie de índice que se ha creado conforme los años y que ahora mismo sigue en crecimiento.

Este índice es rastreado mediante softwares llamados "Rastreadores Web", también conocidos como *Spiders*, del inglés Arañas. Esta última acepción contribuye con una buena analogía del funcionamiento del programa, que consiste en rastrear a alguna página, para luego rastrear a todos los enlaces a los que ésta lleve, creando así toda una "telaraña" de información sobre las páginas existentes en la web.

Uno de los *spiders* más conocidos es "*Googlebot*". El proceso de rastreo que éste realiza comienza con un listado de sitios web de rastreos anteriores y *sitemaps*. Googlebot recopila tanta información como puede de cada página a la que ingresa, y luego la almacena en un índice dependiendo del tipo de información: el título y los enlaces se almacenan en un índice especializado en búsquedas amplias; mientras que el resto del contenido de la página se almacena en otro índice dedicado a las búsquedas más complejas y específicas. Este comportamiento se ilustra en la Figura 1.

Los *webmasters* (o dueños de sitios web) pueden restringir el acceso a los rastreadores para evitar que algunas páginas de su sitio aparezcan en los resultados de las búsquedas. Esto se logra agregando un archivo de texto `robots.txt` en la raíz del sitio web y especificar en este archivo cuales páginas se deben indexar y cuales no. En la sección de demostración se observan ejemplos de su configuración en algunos

sitios web más reconocidos.

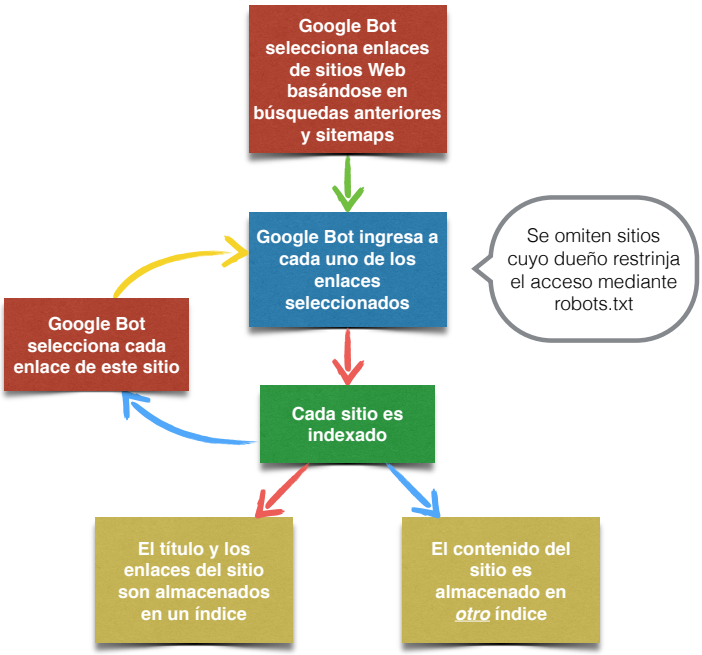

Figura 1: Rastreo e indexación de Google Bot.

Si se piensa con detenimiento, la idea de Google es bastante parecida a la conducta de algún investigador que recurre a un libro, esta persona sabe a qué libro consultar, pero no a qué capítulo ir, por lo tanto su primera idea será buscar en el índice, buscará el capítulo e inciso en cuestión para luego saber a qué página prestar atención.

Los índices son almacenados en los centros de datos (ó *Data Centers*), estos consisten en servidores de gran capacidad de almacenamiento. Google tiene 36 centros de datos activos a lo largo del mundo, 19 en Estados Unidos y 17 más en el resto del mundo, 1 de ellos en América Latina, específicamente en Quilicura, Chile. Entre ellos se tienen aproximadamente 2.000.000 de servidores y más de 10 Exabytes (10^{10} GB) de información. Cada centro de datos puede tener un tamaño de hasta 500.000 metros cuadrados y llegar a costar 600 millones de dolares (El ubicado en chile costo 150 millones de dolares).

Estos tienen la siguiente arquitectura: existen racks de entre 40 y 80 equipos, cada rack tiene un switch de Ethernet, los racks están organizados en clusters conformados por 30 o más racks. El cluster se encarga del balance de carga, es decir, cuando alguien trata de conectarse a Google. Los servidores DNS traducen la dirección `www.google.com` a varias direcciones IP distintas permitiendo que se distribuya la carga a varios clusters, de forma que el hardware envía la consulta al servidor que se encuentre menos ocupado.

Los servidores dentro del centro de datos se dividen distintas tareas, entre ellos se encuentran los servidores *Proxy Squid*, que aceptan la petición y devuelven el resultado desde el cache. Cuando la información no se encuentra en el cache esta es enviada al servidor web, este tiene la función de coordinar los envíos a los servidores índices, los cuales almacenan un conjunto de trozos del índice. Los servidores

índices son llenados con la información recopilada por los servidores de recolección. Por último se tienen los servidores de documentos que como su nombre lo dice almacena documentos y los servidores de anuncios que gestionan la publicidad.

4. Algoritmos de Búsqueda

El objetivo del buscador de Google es conseguir que el usuario obtenga la respuesta que esta buscando lo más rápido posible, creando una conexión casi perfecta entre el usuario y el conocimiento que este busca. Para ello, Google no entrega las millones de paginas que mencionen lo que se esta buscando, si no que filtra principalmente según la actualidad del contenido buscado, la región y el PageRank.

Para realizar estas tarea las base de datos debe actualizarse constantemente, con la finalidad que se entreguen datos de contingencia a la búsqueda. Además se rastrea la IP del usuario para entregar resultados de importancia regional. Luego de filtrar los resultados se les da un valor numérico a las paginas web, esto lo hace mediante el PageRank, con la finalidad de mostrar primero las paginas con el PageRank más alto.

4.1. PageRank

Se le llama PageRank a la gran conglomeración de diferentes protocolos y algoritmos que le dan una connotación numérica a cada sitio que sea indexado. El algoritmo original es una función matemática definida de la siguiente manera:

$$PR(A) = (1-d) + d\sum_{i=1}^{n} \frac{PR(i)}{C(i)} \tag{1}$$

Con:

- PR(A): PageRank asociado a un sitio A.
- d: Factor de Amortiguamiento.
- PR(i): PageRank asociado a un sitio i-ésima que enlaza a la página A.
- C(i): Número total de enlaces salientes de la página i-ésima.

Se puede concluir rápidamente a partir de la Ecuación (1) que se trata de un algoritmo recursivo, y relativamente simple. Se observa que el PageRank de un sitio A depende del de los sitios que enlazan a A, ponderado por el número de sitios enlazados de cada una de aquellas páginas y que es un valor numérico acotado entre 1 y 10. Además existe el factor de amortiguamiento "d" que representa la probabilidad de que se llegue al sitio A a partir de escribir la `url` o acceder a un híperenlace que guíe a la página. Por lo tanto, este protocolo se enfoca en la cantidad de enlaces que citan a una página, y qué página las cite, puesto que una página con PageRank alto le ayudará a una página a que su PageRank sea más alto. Se podría decir que PageRank es una especie de concurso de popularidad donde cada voto es un híperenlace y, bueno, la diferencia clara es que cada voto tiene un valor numérico diferenciado.

5. Demostraciones

5.1. Restricciones de indexación

A modo de ejemplo se ilustran en las Figuras 2 y 3 los archivos `robots.txt` de los sitios web `www.utfsm.cl` y `www.google.com`.

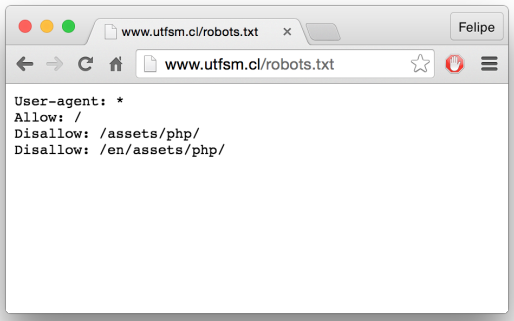

Figura 2: Archivo `robots.txt` del sitio web `www.utfsm.cl`.

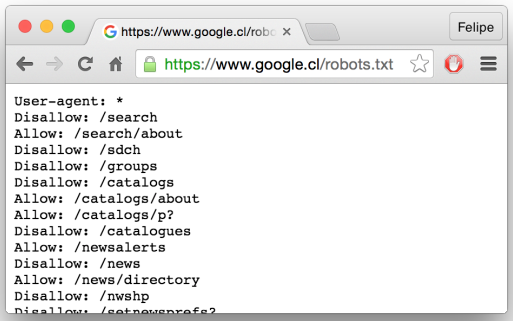

Figura 3: Archivo `robots.txt` del sitio web `www.google.cl`.

Se puede observar a partir de la Figura 2 que se configurar restricciones para todos los rastreadores al utilizar la linea `User-agent: *`. En este caso se permite la indexación de todo el sitio (`Allow: /`) pero se agregan dos subdirectorios como excepción: `/assets/php/` y `/en/assets/php/`

Por otro lado se observa en la Figura 3 que el mismo sitio de Google tiene configurado un archivo `robots.txt` para restringir su propio rastreador. Esto tiene lógica al darse cuenta que no es conveniente entregar como resultado de una búsqueda un enlace a otra búsqueda; esto se evita al agregar la linea `Disallow: /search`, restringiendo la indexación del subdirectorio que almacena las búsquedas.

5.2. Personalización de los resultados de búsqueda

Para demostrar los resultados realizó la búsqueda de una estación de gas y de un hotel utilizando una IP de Viña del Mar, Chile (Figuras 7 y 9) y otra IP de Ámsterdam, Holanda (Figuras 8 y 10) (Anexo 1).

En las figuras se aprecia que al realizar la búsqueda del mismo texto, se obtienen distintos resultados según la ubicación del usuario que realiza la búsqueda (particularmente es la localización de la IP que realiza la búsqueda). En las figuras 7 y 9 se obtienen resultados de estaciones de bencina y hoteles en las proximidades del plan de Viña del Mar, en cambio en las figuras 8 y 10 se tienen resultados de las proximidades de Ámsterdam, Holanda (Anexo 1).

En caso de que el usuario posea una cuenta Google, el algoritmo utilizará información recopilada creando un historial de búsqueda para que en futuras búsquedas arrojar el resultado óptimo según cada usuario.

En las siguientes figuras 4 5 se demuestra la capacidad de almacenamiento de Google.

Figura 4: Historial de búsqueda de Google toda la historia.

Figura 5: Historial de búsqueda de Google en pasado mes.

5.3. Comprobación del rastreador

Como medio de seguridad para los dueños de sitios web se puede verificar si alguno de los accesos al sitio corresponde o no a un rastreador. Aquí se analiza como comprobar que el acceso lo realizo Googlebot.

En un sistema con Windows se puede utilizar el comando `nslookup x.x.x.x` y en sistemas Unix se puede utilizar el comando `host x.x.x.x`; donde `x.x.x.x` corresponde a la dirección IP con la que se desea realizar el requerimiento. En la Figura 6 se observa que el resultado del requerimiento utilizando la IP `66.249.66.1` es el dominio googlebot.com, verificándose así que el acceso lo realizó el rastreador Googlebot.

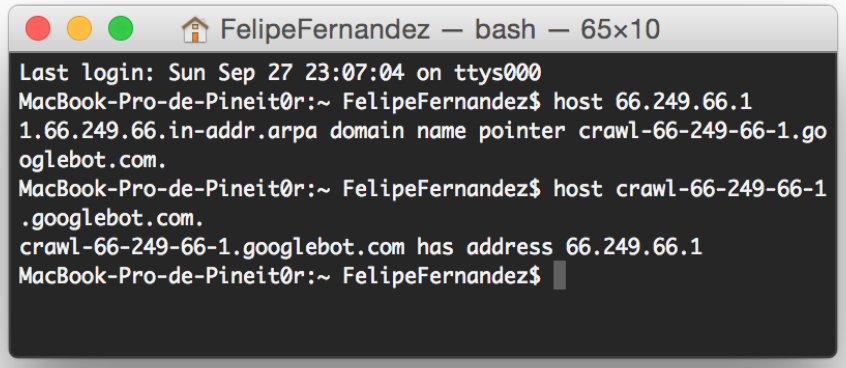

Figura 6: Requerimiento a DNS inverso.

6. Conclusiones

A lo largo de este informe se respondieron cada una de las preguntas propuestas en la etapa introductoria. Se observa que al realizar una búsqueda Google no irá revisando página por página y enviará los resultados de búsqueda, sino que buscará en su propio índice de sitios web, que están jerarquizados mediante el protocolo PageRank. A los datos que entregue les dará particular importancia a la región en donde se realiza la búsqueda, además de la información reciente que se encuentre sobre ella.

El rastreo de páginas web continua actualizándose día a día dando especial cuidado a sitios de contingencia, noticias y avisos publicitarios relevantes.

En un comienzo, el conjunto de algoritmos llamado PageRank contemplaba solamente la ecuación (1), donde se evalúa la importancia de cada sitio indexado, destacando numéricamente el número de enlaces que dirigen al sitio en cuestión, y quienes lo citan.

Mediante la creación de un archivo `robots.txt` en la raíz de un sitio web se puede especificar las páginas de este sitio que se desean indexar y las que se quiere evitar su indexación.

Se puede corroborar la autenticidad de un rastreador mediante un requerimiento a DNS inverso con la IP del rastreador.

Toda la información indexada es guardada en Data Centers, que tienen gran capacidad de almacenamiento, estos distribuyen las tareas a los diversos servidores que lo conforman.

7. Referencias

1. `http://www.google.com/patents/US6285999`
2. `http://infolab.stanford.edu/~backrub/google.html`
3. `http://www.google.com/intl/es/insidesearch/howsearchworks/index.html`

8. Anexo 1

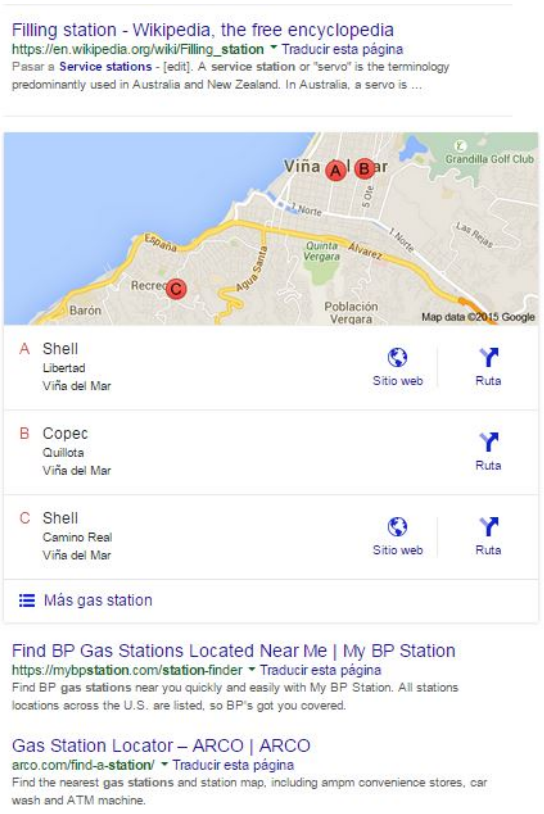

Figura 7: Búsqueda utilizando IP perteneciente a Viña del Mar.

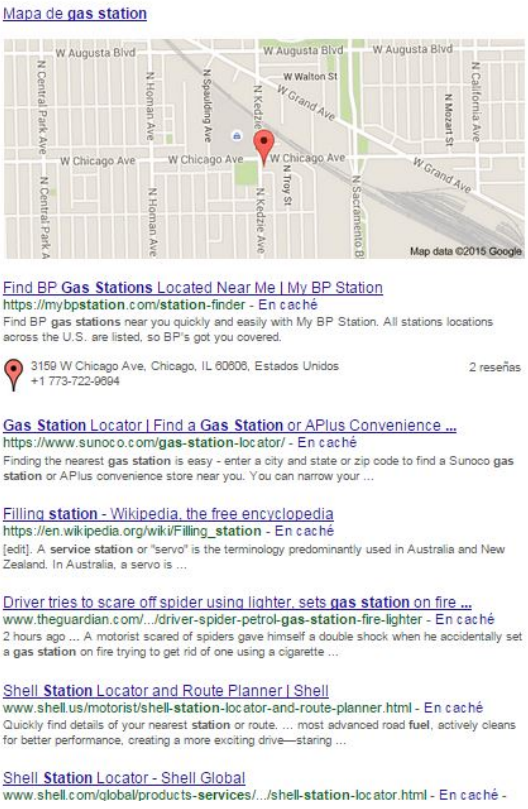

Figura 8: Búsqueda utilizando un proxy con IP de Ámsterdam.

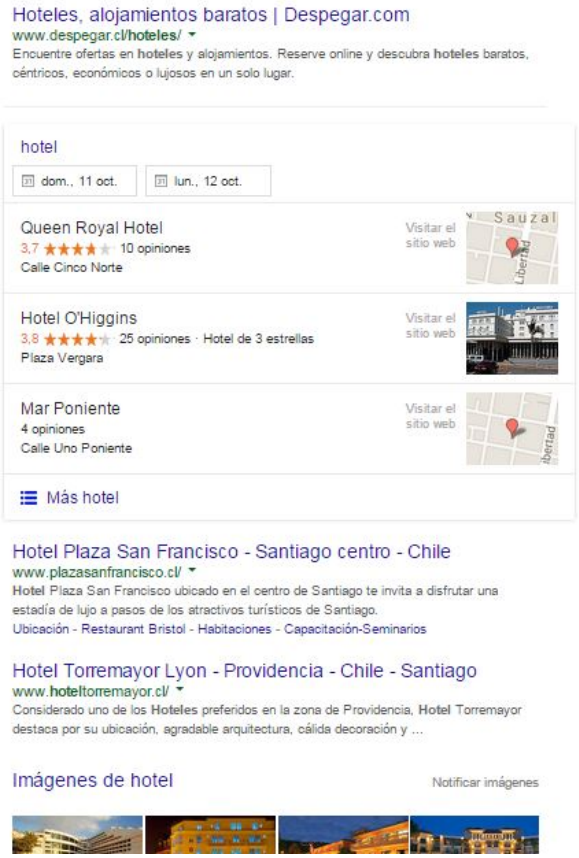

Figura 9: Búsqueda utilizando IP perteneciente a Viña del Mar.

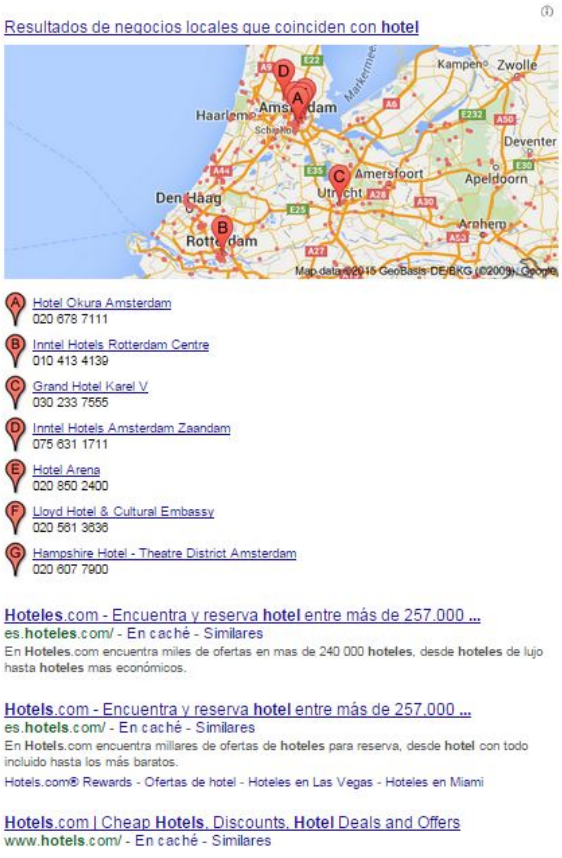

Figura 10: Búsqueda utilizando un proxy con IP de Ámsterdam.