Buscando en la Web

Gonzalo Navarro
Centro de Investigación de la Web
Departamento de Ciencias de la Computación
Universidad de Chile
gnavarro@dcc.uchile.cl

Se dice que los más jóvenes no tienen idea de cómo era buscar información antes que que existiera la Web. Eso es sólo parcialmente cierto. Los menos jóvenes tampoco recordamos gran cosa. Nos resulta un ejercicio de imaginación muy difícil recordar cómo vivíamos cuando, ante cualquier consulta, desde cultural hasta de entretenimiento, no podíamos escribir un par de palabras en nuestro buscador favorito y encontrar inmediatamente montañas de información, usualmente muy relevante.

Gonzalo Navarro es Profesor Titular del DCC, donde obtuvo su doctorado en 1998. Actualmente dirige el Centro de Investigación de la Web (CIW). Sus principales áreas de interés son: algoritmos y estructuras de datos, búsqueda en texto, búsqueda por similitud, y compresión. Ha escrito un libro de búsqueda en texto, y unos 200 artículos en libros, revistas, y congresos internacionales. Ha presidido el Comité de Programa de 6 congresos internacionales y creado el primer congreso en búsqueda por similitud (SISAP).

Para operar este milagro no basta con Internet. Ni siquiera basta con la Web. El ingrediente imprescindible que se necesita son los *buscadores* o *máquinas de búsqueda*. Estos buscadores, cuyos representantes más conocidos hoy en día son probablemente *Google*, *Yahoo!* y *Microsoft MSN*, son los que conocen en qué páginas de la Web aparecen qué palabras (y saben bastante más). Sin un buscador, deberíamos conocer las direcciones Web de todos los sitios de bibliotecas, o de turismo, o de cualquier tema que nos pudiera interesar, y los que no conociéramos sería como si no existieran. En un sentido muy real, los buscadores *conectan* la Web, pues existen grandes porciones de la Web a las que no se puede llegar navegando desde otra parte, a menos que se use un buscador. No es entonces sorprendente que casi un tercio del tiempo que los usuarios pasan en Internet lo dediquen a hacer búsquedas.

Esto nos da una primera idea del gigantesco desafío tecnológico y científico que supone desarrollar un buscador. Debemos resolver cuestiones básicas como ¿qué páginas debería conocer un buscador? ¿qué debería almacenar de esas páginas? ¿qué tipo de preguntas debería aceptar? ¿qué debería responder a esas preguntas? ¿cómo debería mostrar la información? Y ésas son sólo las preguntas más elementales.

Para ordenar la discusión comencemos mostrando la arquitectura típica de una máquina de búsqueda, en la figura 1. Los cuadrados de bordes duros indican procesos, y los de bordes suaves información almacenada. Las flechas representan flujo de información.

En el *crawling* se recolectan páginas de la Web, ya sea nuevas o actualizadas. El proceso de *parsing* es el que extrae los enlaces que parten de las páginas leídas y realimenta el crawling con

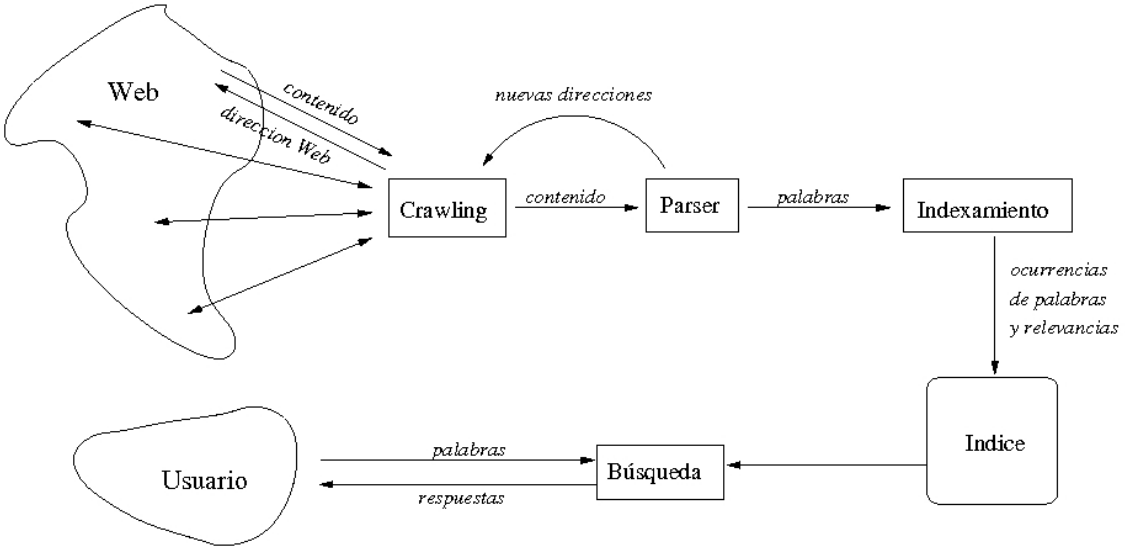

Figura 1. Arquitectura típica de una máquina de búsqueda Web.

nuevas direcciones para visitar, mientras que alimenta al indexador con las páginas depuradas (es decir, sin información irrelevante para el indexamiento). El *indexamiento* almacena en el *índice* la información sobre qué palabras aparecen en qué páginas, junto con una estimación de la importancia de tales ocurrencias. La *búsqueda* usa el índice para responder una consulta, y luego presenta la información al usuario para que éste navegue por ella.

1. Crawling: ¿qué páginas debería conocer un buscador?

Se llama *crawling* al procedimiento de visitar páginas para ir actualizando lo que el buscador sabe de ellas. Un *crawler* es un programa que corre en la máquina del buscador y que va solicitando a distintos computadores de Internet que le transfieran el contenido de las páginas Web que él les indica. Para estos computadores un crawler es prácticamente lo mismo que un humano que visitara sus páginas: debe enviarle el contenido de la página solicitada.

¿Qué páginas debería conocer un buscador? ¡Es tendador responder que todas! Pero lamentablemente esto no es posible. La Web cambia demasiado seguido: un porcentaje alto de las páginas cambia de un mes a otro, y aparece un porcentaje importante de páginas nuevas. Internet no es lo suficientemente rápida: se necesitan meses para transmitir todas las páginas de la Web al buscador. Es simplemente imposible mantener una foto actualizada de la Web ¡Ni siquiera se puede explorarla al ritmo al que va creciendo! La foto que almacena un buscador es siempre incompleta y sólo parcialmente actualizada. No importa cuántas máquinas usemos para el buscador. Los mayores buscadores hoy en día ni se acercan a cubrir la mitad de la Web.

Querer mantener una foto de la Web al día puede compararse con querer estar al tanto de todo lo que ocurre en todas partes del mundo, hasta los menores detalles locales, mediante la continua lectura del diario. Van ocurriendo más novedades de las que es posible ir leyendo. Podemos

pasarnos todo el tiempo leyendo detalles insignificantes y perdiéndonos los hechos más importantes, o podemos tener una política más inteligente de seleccionar las noticias más relevantes, y postergar (tal vez para siempre) la lectura de las menos relevantes. Esto es aún peor si consideramos la llamada *Web dinámica*, formada por páginas que se generan automáticamente, a pedido (por ejemplo al hacer una consulta al sitio de una línea aérea), y que son potencialmente infinitas.

Un tema fundamental en un buscador es justamente el de decidir qué páginas debe conocer, y con cuánta frecuencia actualizar el conocimiento que tiene sobre cada página. Un crawler comienza con un conjunto pequeño de páginas conocidas, dentro de las cuales encuentra enlaces a otras páginas, que agrega a la lista de las que debe visitar. Rápidamente esta lista crece y es necesario determinar en qué orden visitarlas. Este orden se llama "política de crawling". Algunas variables relevantes para determinar esta política son la importancia de las páginas (debería actualizar más frecuentemente una página que es más importante, lo que puede medirse como cantidad de veces que la página se visita, o cantidad de páginas que la apuntan, o frecuencia con que se buscan las palabras que contiene, etc.), y la frecuencia de cambio de las páginas (debería revisitarse más frecuentemente una página que cambia más seguido), entre otras.

2. Indexamiento: ¿qué debería almacenarse de las páginas?

El *indexamiento* es el proceso de construir un *índice* de las páginas visitadas por el crawler. Este índice almacena la información de manera que sea rápido determinar qué páginas son relevantes a una consulta.

¿No basta con almacenar las páginas tal cual, para poder buscar en ellas después? No. Dados los volúmenes de datos involucrados (los mayores buscadores indexan hoy en día miles de millones de páginas, que ocupan varios terabytes), es imposible recorrer una a una todas las páginas almacenadas en un buscador para encontrar cuáles contienen las palabras que le interesan al usuario. ¡Esto demoraría horas o días para una sóla consulta!

El buscador construye lo que se llama un *índice invertido*, que tiene una lista de todas las palabras distintas que ha visto, y para cada palabra almacena la lista de las páginas donde ésta aparece mencionada. Con un índice invertido, las consultas se pueden resolver mediante la búsqueda de las palabras en el índice y el procesamiento de sus listas de páginas correspondientes (intersectándolas, por ejemplo). La figura 2 ilustra un índice invertido.

Los buscadores grandes deben procesar hasta mil consultas por segundo. Si bien este trabajo puede repartirse entre varios computadores, la exigencia sigue siendo alta. El mayor costo para responder una consulta es el de leer de disco las listas de páginas apuntadas por el índice invertido. Es posible usar técnicas de compresión de datos para reducir el espacio en que se representan estas listas. Con esto se logra ganar espacio y velocidad simultáneamente. Pueden hacerse también otras cosas, como precalcular las respuestas a las consultas más populares.

3. Búsqueda: ¿qué preguntas debería responder, y cómo?

Hemos estado considerando que el usuario escribe algunas palabras de interés y el buscador le da la lista de las páginas donde aparecen estas palabras. La realidad es bastante más complicada. Tomemos el caso más elemental, de una consulta por una única palabra. Normalmente hay millones de páginas que contienen esa palabra, y está claro que el usuario no tiene la menor posibilidad

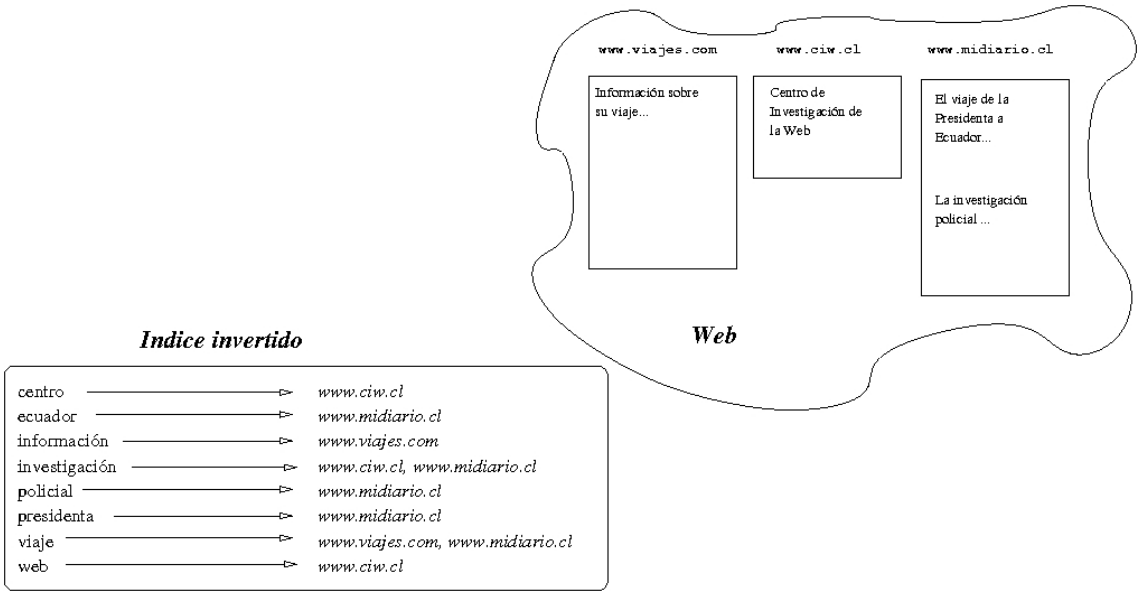

Figura 2. Ejemplo de un índice invertido para tres páginas Web.

de examinarlas todas para ver cuáles satisfacen su necesidad de información. De alguna manera el buscador debe *ordenar* las respuestas por su supuesta *relevancia* a la consulta.

Existen muchas formas de calcular esta relevancia, que dan lugar a mejores o peores heurísticas. Por ejemplo, uno puede considerar que una página donde la palabra aparece varias veces es más relevante que otra donde aparece una vez. Pero si la palabra aparece más veces en una página que es mucho más larga que otra, entonces tal vez la palabra no sea tan importante en esa página. También uno puede considerar cuán importante es la página en sí (por ejemplo si es muy visitada, o muy apuntada por otras). Los buscadores utilizan fórmulas matemáticas para calcular relevancia que tienen en cuenta estos aspectos.

Existen técnicas más sofisticadas, por ejemplo llevar información de cómo se comportaron otros usuarios cuando hicieron esta misma consulta (por ejemplo, el buscador puede saber que la gran mayoría de los usuarios que buscaron "mp3" terminaron yendo a ciertos sitios específicos). Esto se llama *minería de consultas* y es extremadamente útil para dar buenas respuestas a consultas que no dicen mucho. También puede usarse información posicional, por ejemplo si la palabra aparece en el título de la página o de los enlaces que la apuntan, puede ser más relevante que si aparece cerca del final.

La situación se complica cuando la consulta tiene varias palabras, donde algunas pueden ser más importantes que otras. Normalmente las ocurrencias de palabras que aparecen en muchos documentos, como los artículos y preposiciones, son poco importantes porque no sirven para discriminar. Para peor, sus listas de ocurrencias en los índices invertidos son muy largas, ocupando espacio inútil. Por ello muchos buscadores las omiten de sus índices (intente buscar "and" en su buscador favorito). La forma de combinar el peso de las distintas palabras da lugar también a mejores o peores heurísticas. Por ejemplo los buscadores en la Web normalmente muestran sólo

páginas donde aparecen todos los términos, como una forma de eliminar respuestas irrelevantes. Asimismo, los mejores dan preferencia páginas donde las palabras aparecen cercanas entre sí.

La verdad es que en la Web hay mucha, mucha más información de la que se puede obtener mediante la búsqueda de documentos que contengan ciertas palabras. Esta limitación se debe a que no es fácil implementar búsquedas más sofisticadas a gran escala. Conseguir responder consultas más complejas a escala de la Web es un tema actual de investigación. Algunos ejemplos son:

1. Buscar por contenido en fotos, audio o video. Imagínese mostrar una foto de su promoción y poder encontrar otras fotos de las mismas personas en la Web, incluso sin recordar sus nombres. O tararear una parte de una melodía (incluso con errores) y encontrar el mp3 para poder bajarlo. Existen técnicas para hacer esto, pero no a gran escala. Los buscadores ofrecen búsqueda de fotos, pero basada en palabras que se encuentran asociadas a las fotos durante el crawling.
2. Hacer preguntas complejas que se pueden inferir de la Web. Por ejemplo preguntas como ¿cuál es la farmacia más cercana que venda un antigripal a un precio inferior a $ 3.000? y ¿qué universidades dictan una carrera de Diseño Gráfico de 5 años en la Región Metropolitana? Responder este tipo de preguntas requiere normalmente de cierta cooperación de quien escribe las páginas.
3. Hacer consultas con componente temporal, como ¿qué ocurrió con el seguimiento en los medios de comunicación a las consecuencias de la guerra del Golfo en los meses siguientes a su finalización? Esto requiere llevar una cuenta histórica de los contenidos de la Web a lo largo del tiempo.

4. Interacción con el Usuario: ¿cómo presentar la información?

Ya vimos que las respuestas que se muestran al usuario son sólo una mínima parte de las que califican. Los buscadores normalmente presentan una lista de las primeras páginas según el orden que han hecho en base a la consulta. En esta lista se indica la dirección de la página (para que el usuario pueda visitarla con un click) y usualmente el *contexto* del texto donde las palabras aparecen. Esto ayuda al usuario a saber rápidamente si las palabras aparecen en la forma que las esperaba.

Poder mostrar un contexto requiere que el buscador no almacene sólo el índice invertido, sino también el contenido completo de las páginas que indexa, de modo de poder mostrar un pasaje donde aparecen las palabras de la consulta. Si bien el espacio es barato, esto es un requerimiento bastante exigente, pues el buscador debería tener suficiente almacenamiento para duplicar toda la Web en sus discos. Para reducir el espacio, el buscador puede evitar almacenar las imágenes, por ejemplo. La compresión de datos es también útil para aliviar este problema.

Los buscadores suelen ser lo suficientemente buenos como para que, un gran porcentaje de las veces, lo que busca el usuario esté entre las primeras respuestas que ofrece. De todos modos, es posible pedirles que entreguen el siguiente conjunto de respuestas, y el siguiente, hasta hallar lo que uno busca. La experiencia normal es que, si la respuesta no está en las primeras páginas, es raro que esté más adelante. Es mejor en esos casos reformular la consulta, por ejemplo haciéndola más específica (si se hallaron demasiadas páginas irrelevantes) o más general (si se hallaron demasiadas pocas respuestas). Por ejemplo, en la figura 2, si buscáramos "viaje" encontraríamos tanto la página de la agencia de viajes como la noticia sobre el viaje presidencial. Refinando la consulta

a "viaje Presidenta" tendríamos mejor precisión. Esta iteración es frecuente en las sesiones con los buscadores, y con el tiempo el usuario aprende a formular consultas más exitosas.

Existen formas mucho más sofisticadas de presentar la información, pero nuevamente es difícil aplicarlas a sistemas masivos como la Web. Asimismo suele ocurrir que las interfaces demasiado "inteligentes" resultan ser demasiado complejas para la mayoría de la gente. Incluso los lenguajes de consulta más complejos, donde se puede indicar que las palabras A y B deben aparecer, pero no C, normalmente están disponibles en los buscadores Web, pero se usan muy raramente. La regla en este caso es que la simplicidad es lo mejor.

Referencias

- `www.searchenginewatch.com` es un sitio dedicado a las estadísticas sobre las principales máquinas de búsqueda en la Web.
- `http://www.press.umich.edu/jep/07-01/bergman.html` y `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/` son dos sitios dedicados a estudiar el crecimiento de la Web, y en general de la cantidad de información disponible en el mundo.
- `www.ciw.cl`
- `www.todocl.cl`