

Capítulo 4

Buscando en la Web

Gonzalo Navarro

Se dice que los más jóvenes no tienen idea de cómo era buscar información antes que existiera la Web. Eso es sólo parte de la verdad. 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, en general muy relevante.

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 son probablemente Google [1], Yahoo! [2] y Microsoft MSN [3], 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.

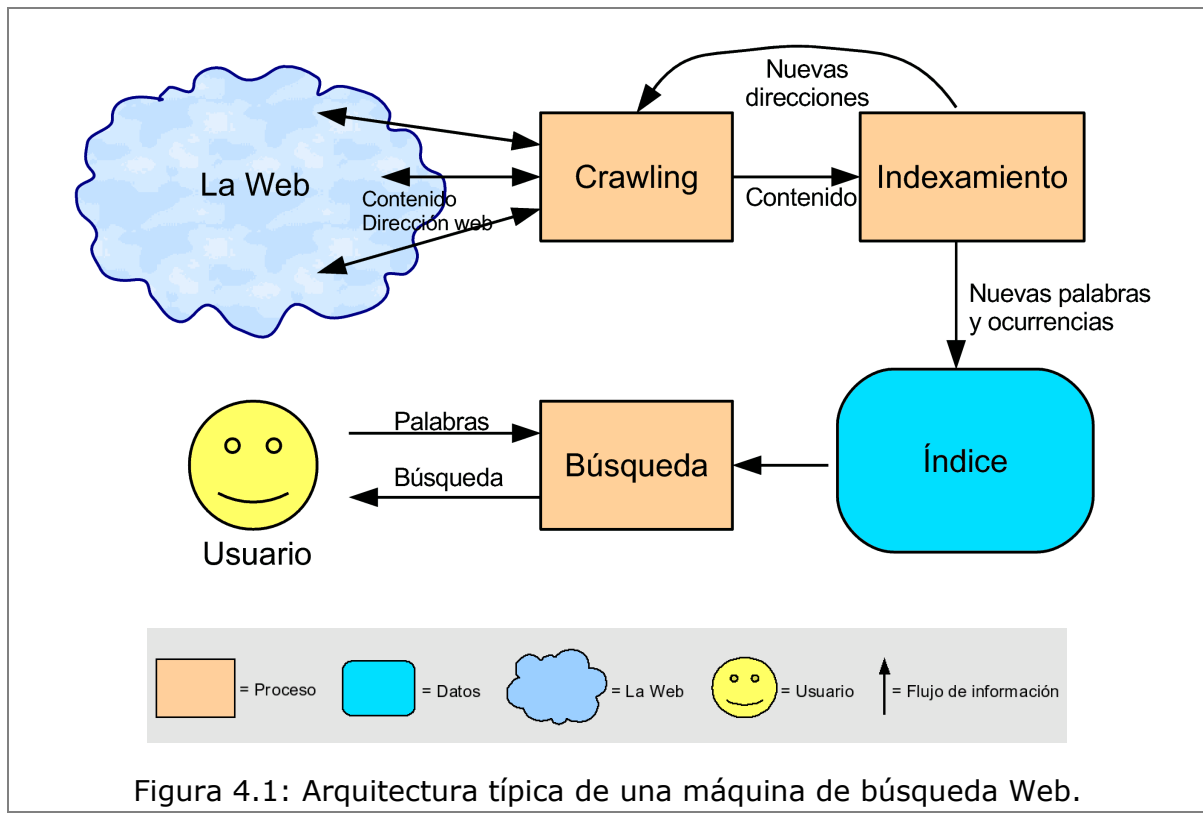


Figura 4.1: Arquitectura típica de una máquina de búsqueda Web.

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 esas son sólo las preguntas más elementales.

Para ordenar la discusión comencemos mostrando la arquitectura típica de una máquina de búsqueda, en la figura 4.1. En ésta, la Web y los usuarios son el mundo exterior al buscador. Todo lo que está a la derecha es parte del buscador.

En el *crawling* se recolectan páginas de la Web, ya sea nuevas o actualizadas. El proceso de *indexamiento* es el que extrae los enlaces que parten de las páginas leídas y realimenta el crawling con nuevas direcciones para visitar, mientras que almacena en el *índice* la información para 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 [4].

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 solicita a distintos computadores de Internet que le transfieran el contenido de las páginas Web que él les indica. Para estos computadores es casi lo mismo que un crawler o un ser humano visite sus páginas: debe enviarle el contenido de la página solicitada.

¿Qué páginas debería conocer un buscador? ¡Es tentador 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 es posible explorarla al ritmo al que va creciendo! La foto que almacena un buscador es siempre incompleta y sólo parcialmente actualizada. No importa cuántos computadores usemos para el buscador. Los mayores buscadores hoy ni se acercan a cubrir el total de la Web. ¡Es incluso difícil saber cuál es el tamaño real de la Web! 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. Y esto considerado que se refieren sólo a la Web pública (de acceso gratuito).

Algunos números pueden dar una idea de las magnitudes involucradas. En 2005 se estimaba que la Web contenía 11.500 millones de páginas, de las cuales los mayores buscadores cubrían a lo sumo el 70%. Algunos estudios calculan que la Web dinámica, por otro lado, puede llegar a los 500 mil millones de páginas.

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 leer el diario continuamente. 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.

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 (el crawler debería visitar más frecuentemente una página que cambia más seguido), entre otras.

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 hoy indexan más de 3 mil 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 sola 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 buscar las palabras en el índice y procesar sus listas de páginas correspondientes (intersectándolas, por ejemplo). La figura 4.2 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 si-

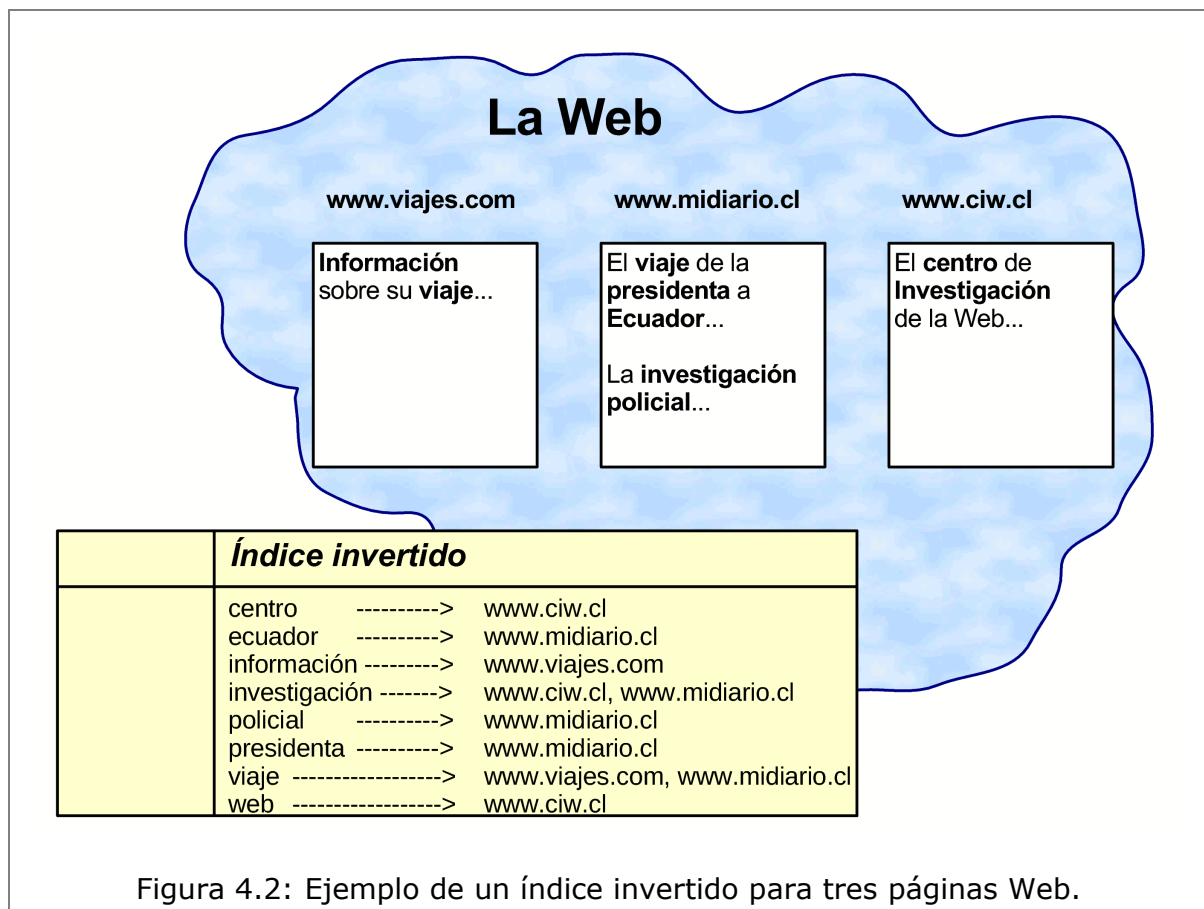


Figura 4.2: Ejemplo de un índice invertido para tres páginas Web.

multáneamente. Pueden hacerse también otras cosas, como precalcular las respuestas a las consultas más populares.

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 millo-

nes de páginas que contienen esa palabra, y está claro que el usuario no tiene la menor posibilidad 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 buscada 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 cuan importante es la página en sí (por ejemplo si es muy visitada, o muy apuntada por otras). Los buscadores utilizan fórmulas matemáticas para calcular la 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 a páginas donde las palabras aparecen cercanas entre sí.

Pero la verdad es que en la Web hay mucha, mucha más información de la que se puede obtener mediante buscar 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 una persona se encarga de asociar a cada foto 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 en el Líbano en los meses siguientes a su finalización? Esto requiere llevar una cuenta histórica de los contenidos de la Web a lo largo del tiempo.

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 esperaba (por ejemplo *investigación* puede referirse a científica o policial).

Poder mostrar un contexto requiere que el buscador no almacene sólo el índice invertido, sino también el contenido completo de las páginas que indexa. 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! Por ejemplo, para reducir el espacio, el buscador puede evitar almacenar las imágenes. La compresión de datos es también útil para aliviar este problema.

Los buscadores suelen ser lo suficientemente buenos como para que, en un gran porcentaje de las veces, lo que busque el usuario esté entre las primeras respuestas que ofrece. De todos modos es posible pedirle que

entregue 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. En esos casos es mejor reformular la consulta, por ejemplo haciéndola más específica (si se encontraron demasiadas páginas irrelevantes) o más general (si se encontraron muy pocas respuestas). Por ejemplo, en la figura 4.2, si buscáramos `investigación` encontraríamos tanto la página del Centro de Investigación de la Web como la noticia policial. Refinando la consulta a `investigación policial` 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 sofisticados, 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.

Para saber más

- ◆ El sitio www.searchenginewatch.com está dedicado a las estadísticas sobre las principales máquinas de búsqueda en la Web.
- ◆ Los sitios <http://www.press.umich.edu/jep/07-01/bergman.html> y <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/> están dedicados a estudiar el crecimiento de la Web, y en general de la cantidad de información disponible en el mundo.
- ◆ El sitio www.todo.cl es el buscador chileno Todo.cl.

Referencias

1. Google. <http://www.google.com>
2. Yahoo! <http://www.yahoo.com>
3. Microsoft MSN. <http://www.msn.com>
4. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley-Longman, 1999. Capítulo 13.