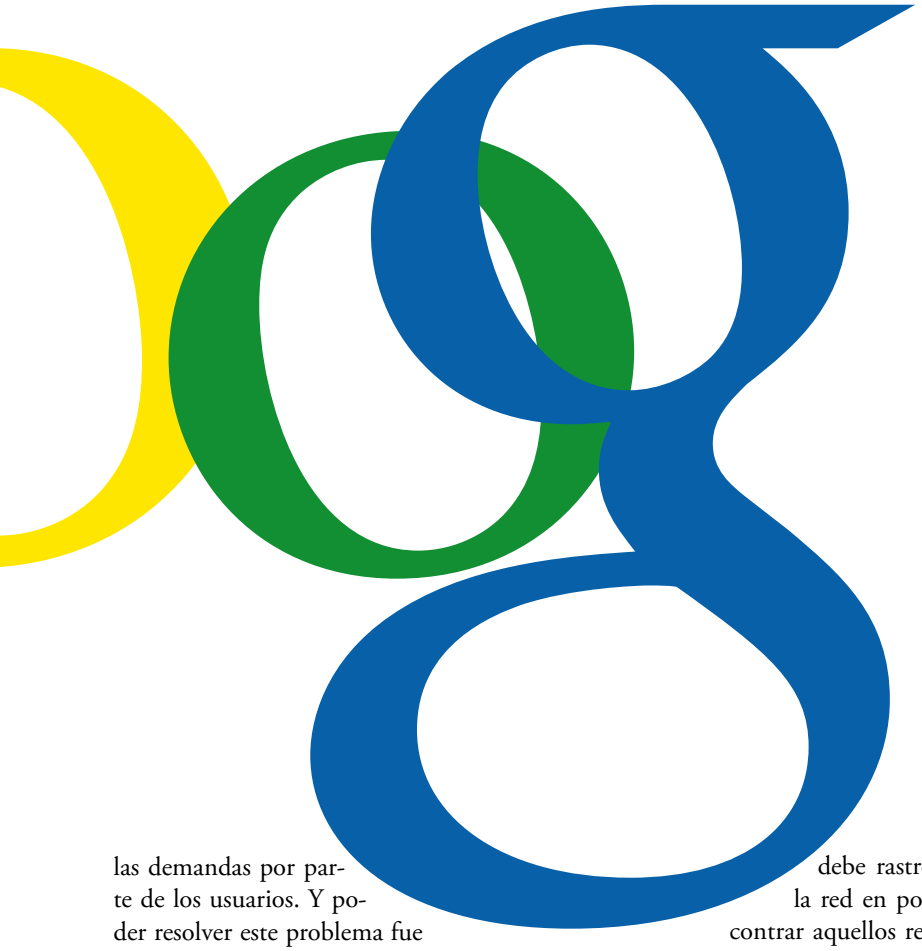
Buscadores de Internet

Sé lo que quiero y lo quiero ya

¿Cómo hace Google para obtener de manera tan rápida y eficiente los resultados de una búsqueda en Internet? ¿Son iguales todas las búsquedas? Mucho depende de aquello que se desea encontrar, el lugar o contexto para la búsqueda, las herramientas que se tienen a disposición para la tarea y los objetivos fijados para la misma.

Fernando Asteasuain
fasteasuain@dc.uba.ar

Todo parte de una búsqueda, de la necesidad de encontrar algo. En algunas ocasiones es algo bien específico y concreto. En otras simplemente se desea conocer más sobre algún tópico en particular. Cada búsqueda es distinta, con diferentes características, urgencias y necesidades. No es lo mismo buscar las llaves de casa cuando apremia el tiempo para no llegar tarde al trabajo, que buscar información sobre posibles destinos turísticos para las próximas vacaciones. En cada caso el contexto cambia, así como también cambian las estrategias usadas para lograr con éxito lo emprendido. Y como es de esperar, la situación es similar en el mundo virtual que representa Internet. Las búsquedas en Internet empezaron a tener relevancia a medida que Internet crecía, no solo en contenidos, sino también porque crecían



las demandas por parte de los usuarios. Y poder resolver este problema fue clave para que Internet pudiera despegar y convertirse en esta gigantesca fuente de información que constituye en la actualidad. De nada sirve un lugar que lo tiene todo, si no podemos encontrar aquello que necesitamos.

A medida que Internet evolucionaba, empezaron a cobrar relevancia un tipo especial de sitios web: los buscadores. Son sitios dedicados principalmente a buscar información por toda la Red. Varias interrogantes surgen en este sentido: ¿qué hay detrás de una búsqueda en la web?, ¿qué significa buscar en la web?, ¿cuáles son los resultados esperados? Como se mencionó anteriormente, hay distintos tipos de búsquedas: buscar restaurantes naturistas que estén abiertos los domingos a la noche, ver cuántos goles hizo Messi el último domingo, averiguar cuándo se estrenó *Volver al Futuro*, chequear los horarios y funciones de los cines o teatros, consultar cómo tramitar el nuevo pasaporte en la modalidad *express*, o indagar sobre la mejor receta hogareña para hacer un pollo al curry. Ante cada consulta, el buscador debe rastrear toda la red en pos de encontrar aquellos resultados relevantes a la búsqueda y ofrecerlos al usuario. A todo esto, se suma un nuevo desafío: ¿qué resultados mostrar primero al usuario?. Este último planteo no es para nada menor, ya que en gran medida el éxito de las búsquedas en Internet depende de este factor. Una posible opción para el buscador es mostrar en primer lugar los resultados que encontró primero: recorrer todos los sitios e ir armando una lista de resultados con los sitios que sean relevantes a la búsqueda. En el primer lugar de la lista figuraría el primer sitio que apareció, luego el segundo, y así sucesivamente. Pero una rápida mirada crítica sobre esta manera de proceder alerta sobre un posible problema. ¿Qué ocurre si el sitio más relevante para la búsqueda se encuentra en los últimos lugares de la lista de resultados? En una lista de pequeña longitud esto no sería un problema, ya que recorrerla toda no implica en principio invertir mucho tiempo. Pero en cambio, si se trata de una lista con millones de resultados, la situación cambia drásticamente. Rara vez los usuarios buscan más allá de los seis ó siete primeros sitios en esa lista de resultados, por lo que los buscadores para tener éxito no solo deben encontrar los sitios relevantes, sino también deben ubicarlos en los primeros lugares al mostrar los resultados. Dado esto, parece sencillo solucionar el problema: los buscadores deben poner primero en la lista los sitios más relevantes, y relegar a los últimos lugares aquellos que menos relación tienen con la búsqueda. Pero si bien es sencilla la enunciación de la solución, su puesta en práctica no es tan trivial: ¿cómo determina un buscador cuándo un sitio es más relevante que otro? En otras palabras, el buscador debe intuir lo que el usuario quiere encontrar, y decidir en base a esa intuición. ¡Y debe hacerlo rápido! No solo ordenar por relevancia los resultados, sino toda la búsqueda. En síntesis, un buscador debe ser *eficiente para encontrar de manera rápida los resultados y, a su vez, inteligente para ordenarlos*. Y entre todos los buscadores, hubo uno que se destacó sobre el resto: Google. Las razones son sencillas, fue el que mejor resolvió estos dos puntos fundamentales.

Búsquedas rápidas

Este punto fue abordado desde muchos lugares. Primero, hay distintas maneras o algoritmos para realizar búsquedas, siendo

algunas más rápidas que otras. Cuando una persona busca una ficha para ubicar en un rompecabezas, bien podría buscar una por una todas las fichas hasta encontrar la deseada. Este enfoque se conoce computacionalmente como "fuerza bruta". Otro enfoque un poco más eficiente podría ser concentrarse únicamente en aquellas cuya forma/color encaja mejor, y así poder encontrar más rápido la pieza buscada. Google trabajó, y trabaja con esmero para poder *desarrollar los algoritmos más eficientes posibles*. Segundo, *mucho poder de cómputo*. Es decir, atacar el problema con artillería pesada: computadoras especializadas, construidas especialmente para trabajar lo más rápido que se pueda. Tercero, *acceso rápido a la información*. Una manera de ver esto es que Google tiene muchas búsquedas "pre-cocidas", listas para ser enviadas al usuario. Por ejemplo, cuando un usuario busca "Messi", Google, antes de realizar la búsqueda desde cero, se fija si no resolvió "hace poco" una consulta parecida. La clave está en dejar cada búsqueda nueva guardada un tiempo, para intentar ahorrarnos tiempo la próxima vez que se busque. Esta técnica se conoce como *caché en memoria* de consultas. Cuarto y último, un concepto fundamental, el cual sostiene a todos los puntos anteriores*: índices*. La idea de usar índices para las búsquedas no es para nada novedosa. Basta con pensar en cualquier libro. Generalmente al principio, cada libro cuenta con un índice, el cual dice, por ejemplo, en qué página comienza cada capítulo. De esta forma, si una persona está interesada en leer el capítulo cuatro de una novela, se dirige al índice para acceder de manera directa a la página inicial del capítulo. De no contar con el índice, la persona hubiera tenido que recorrer el libro hasta encontrar el comienzo del capítulo. En determinadas situaciones es sumamente útil contar con más de un índice. Por ejemplo, es de poca utilidad el índice de comienzo de cada capítulo cuando se desea buscar una ilustración en particular dentro del libro. En este caso, no queda otra opción más que buscar por todo el libro hasta encontrarla. Una posible manera de alivianar esta tarea sería contar también con un índice de figuras, donde establezca la página de cada figura. Otro tipo valioso de índices suele ser el que especifica para un determinado término, los lugares dentro del libro donde está mencionado el mismo. Por ejemplo, poder buscar en un libro de cine, todas las páginas donde esté mencionado "Alfred Hitchcock".

Google incorpora para sus procesos de búsquedas un *avanzado manejo de índices* para poder acceder de manera directa a los resultados deseados, y lograr así agilizar y reducir notablemente el tiempo empleado para lograr sus objetivos. De manera periódica, Google modifica su índice, para incorporar páginas nuevas y actualizar los contenidos de Internet. El encargado de este procedimiento es un programa conocido como *GoogleBot*.

Ordenar por relevancia

Una vez resuelta la búsqueda, el paso siguiente consiste en ordenar la lista con los resultados de manera tal que aquellos más relevantes se encuentren ubicados en las primeras posiciones. El problema entonces es determinar cómo asignar a cada resultado su relevancia y confiabilidad para la consulta dada. Una vez determinado este factor, solo resta ordenar la lista de mayor a menor. Y es este punto uno de los factores decisivos para el éxito de Google. El 9 de enero de 1999, Larry Page y Sergey Brin, creadores de Google, dieron a conocer su ahora famoso algoritmo denominado *PageRank,* cuyo objetivo es asignar un valor numérico a cada resultado, estableciendo qué tan relevante es para la consulta. Siendo ambos de familias con tradición académica, decidieron imitar un sistema conocido en el mundo de la publicación de trabajos de investigación: los trabajos más relevantes son aquellos que mayor impacto tienen, es decir, los que son citados o referenciados más veces. En particular, se trata del modelo Science Citation Index (SCI) creado por Eugene Garfield durante la década del '50.

Aplicar este modelo del mundo científico al mundo de Internet fue casi directo: las páginas más importantes son aquellas que son más "citadas". ¿Cuándo una página cita a otra? Para Google, una página A "cita" a una página B cuando la primera tiene un enlace o vínculo a la segunda. Adicionalmente, se considera un segundo factor: cuanto más citada sea una página, mayor valor tienen las citas que ésta haga. Es decir, la cita de una página muy citada vale más que la cita de una página escasamente citada. De esta manera, Google decide qué resultados mostrar primero. *PageRank* es actualizado periódicamente: durante el 2011 tuvieron lugar dos actualizaciones, una en enero y la más reciente, en junio.

Mejoras

Google fue añadiendo mejoras a su algoritmo de búsquedas, para evitar que los resultados puedan manipularse. Una manera de lograr esto es, por ejemplo, crear muchos enlaces o citas a la página que se desea ubicar en los primeros lugares. Para esto se puede escribir un programa que llene de manera automática blogs, páginas de visitas, etcétera, con la sola finalidad de poner enlaces a la página en cuestión. Esto se conoce como "IndexSpamming", es decir, el viejo y conocido *spam*, pero ahora utilizado para llenar de enlaces artificiales la web para poder posicionar mejor a una página. La última versión del algoritmo de búsqueda lanzada por la empresa, denominada *Google Panda*, busca afanosamente luchar contra el *spam* de índices, y también incluye otras

mejoras. Por ejemplo, busca priorizar contenido original, restándole importancia a aquellas páginas que solo sean duplicados o copias de otra. También incorpora nociones lingüísticas, destacando páginas con buena ortografía, y con frases y oraciones bien construidas. Y penaliza a aquellas páginas o sitios con exceso de publicidad. Finalmente, también las redes sociales impactarán en las búsquedas: es decir, se tendrán en cuenta resultados producidos desde redes como *Twitter, Google Plus, o Youtube*.

Google instantáneo

Una de las últimas funcionalidades con que Google ha sorprendido a sus usuarios se denomina *Google Instant*. Bajo esta modalidad Google muestra los resultados a medida que el usuario va escribiendo en la barra de búsqueda. Para esto, Google calcula a través de un algoritmo cuáles son los resultados más esperados a partir de lo que el usuario está ingresando. Así, parece ante los ojos del usuario que los resultados aparecen de manera instantánea, casi por arte de magia. El principio detrás de *Google Instant* es que los seres humanos en general escriben lento, pero leen con rapidez. Según los estudios de Google pulsar una tecla puede llevar 300 milisegundos, mientras que en solo 30 milisegundos el ojo humano es capaz de mover su atención a distintos lugares de la página. Esto indica que el usuario puede entonces analizar los resultados mientras escribe una consulta. *Google Instant* reduce entre dos y cinco segundos el tiempo destinado a cada consulta, ya que en algunos casos no es necesario terminar de escribirla o, incluso, apretar la tecla *Enter* para iniciar la búsqueda.

Quizás quiso decir

Otra de las virtudes de Google es, también, intuir cuándo el usuario se equivoca en una búsqueda, principalmente por errores sintácticos. Google logra esto "aprendiendo" de los errores de los usuarios. En general, cuando un usuario busca una palabra con errores no hace clic en ningún resultado porque los resultados no son relevantes. Luego, al darse cuenta del error, busca nuevamente, ahora con la palabra corregida. En este caso, Google "aprende" que las palabras están relacionadas, y que la segunda versión sea probablemente una corrección de la primera. Así, al próximo usuario que cometa el mismo error, le sugerirá la palabra correcta. Con las miles y miles de consultas por segundo que se hacen a nivel global, Google es un rápido aprendiz. También puede aprender por ejemplo, si al buscar "orasión" un usuario hizo clic en alguna página que menciona la palabra "oración". La cantidad de resultados también es un punto a tener en cuenta a la hora de sugerir cambios. Por ejemplo, la búsqueda "orasión" da unos 170.000 resultados, que a priori no es un mal número. Sin embargo, los resultados encontrados para una palabra relacionada y similar como "oración" son muchos más, alrededor de 20.700.000 para ser precisos. Una vez que Google tiene la suficiente confianza en las correcciones, hace la búsqueda directamente con la versión corregida de la palabra. En estos casos le avisa al usuario el cambio realizado. Para el caso anterior, Google buscará directamente "oración" y, al mostrar los resultados indicará: Mostrando resultados para oración. Haga clic para buscar "orasión". Es importante notar que Google no tiene un corrector ortográfico, sino que se guía por cómo una palabra aparece escrita mayor cantidad de veces en la web. Esto quiere decir que si, de repente y por alguna razón, muchos usuarios empiezan a escribir "oración" con la letra s, Google empezará a sugerir esta versión de la palabra. Un claro ejemplo de esto aconteció unos días después del descenso de River, ocurrido el día 26 de junio de 2011. Al hacer la búsqueda "chau river", Google sugería la búsqueda "chau Riber", ya que en numerosos sitios figuraba esta última frase, como una broma hacia los hinchas de River, y por lo tanto, tenía muchos más resultados que la búsqueda original (para más información, ver el artículo de Mariano Blejman en *Página 12* del día 29 de junio de 2011).

Entonces, ¿por qué buscar archivos en mi computadora tarda tanto en comparación, si es tanto menor el espacio de búsqueda?

Buscar archivos en una computadora es un proceso notoriamente diferente. Primero que nada, las computadoras de Google están especializadas en búsquedas, tanto en hardware como en software. En la nuestra no hay resultados "pre-fabricados", ni discos optimizados, y todas las búsquedas son casi un procedimiento manual, inspeccionando archivo por archivo.

La palabra clave para mejorar los tiempos de búsqueda no es otra que *índices*. Tener nuestros archivos "indexados" es un paso importante en este sentido. Justamente una de las opciones que tenemos al instalar sistemas operativos como *Windows* es pedir que se utilicen índices para mantener nuestros archivos. Pero sin dudas, la ayuda definitiva viene de aplicaciones que podemos instalar, las cuales aprovechan al máximo el concepto de índices. Dos de ellas son *Windows Desktop Search* y *Google Desktop*.

En pocas palabras

Resumiendo, se puede afirmar que Google combina diversos enfoques para constituir un excelente buscador: un gigantesco poder de cómputo, algoritmos eficientes, superlativa utilización de índices y una eficaz manera de ordenar los resultados más relevantes, junto a novedosas formas de mejorar cada vez más los resultados esperados de una búsqueda. El desafío es simple: solo se trata de un programa, frente a un usuario, intentando adivinar qué es en realidad lo que quiere buscar, y obteniendo los resultados en escasos milisegundos.

Agradecimientos: José Castaño, Esteban Feuerstein, Cecilia Ruz, Diego Gavinowich y Alexis Tcatch.