BUSCADORES ON-LINE: de la recuperación de la información a la generación del conocimiento

Autores: Borja Sanz, Javier Nieves, Carlos Laorden, Igor Santos, Pablo G. Bringas
DeustoTech Computing - S3Lab - Universidad de Deusto, Bilbao

En 2012, el sitio web más popular a nivel mundial según *Alexa* fue *Google.com*[1]. Esto es el buscador *on-line* del gigante de Mountain View. De los múltiples servicios que se ofrecen en Internet, un "simple" buscador copa este ranking, por delante de redes sociales online (*Facebook* en el número 2), servicios de compartición de video (*YouTube* en el número 3) o enciclopedias de conocimiento compartido (*Wikipedia* en el número 6).

Este hecho no hace más que remarcar la importancia que ha adquirido la gestión de la información con el paso de los años, información que a su vez ha experimentado un crecimiento exponencial en su volumen como consecuencia de la evolución que ha vivido la tecnología. Pero, si bien pudiera parecer que la cantidad de información disponible juegue a favor de los usuarios, no hace sino complicar el acceso al contenido deseado, introduciendo mucha información que está fuera de lugar. Y es ahí donde nace la necesidad de un sistema para navegar entre la ingente cantidad de información con la que cuenta la red y que, además, sea capaz de brindar al usuario aquella información que realmente le resulte de utilidad, el buscador de contenidos online.

Evolución de los buscadores online

Si bien el primero de estos buscadores fue el ya desaparecido "*Wandex*", desarrollado por **Mattew Gray** en el MIT (1993), que en realidad estaba formado por un simple índice de contenidos, "*WebCrawler*" sería, ya en 1994, el primer motor de búsqueda tal y como los conocemos hoy en día. WebCrawler permitía al usuario realizar búsquedas por palabras y acabó por marcar un estándar para sus sucesores.

A pesar de ello, la experiencia de usuario de cara a buscar contenido en la librería de conocimiento que se ha convertido la Web, no era ni mucho menos satisfactoria. El hecho de que los buscadores devolvieran sin mucho criterio miles de páginas a partir de una consulta, hacía que acabaran funcionando mejor técnicas empleadas en la antigüedad como el boca a boca o el conocimiento experto. Es decir, se conocían los sitios web interesantes a través de amigos y de sitios recomendados por otras personas, las cuales habían dedicado horas de esfuerzo en realizar las búsquedas.

Pero, todo esto cambió cuando el análisis de los *links* llegó al mundo de la Recuperación de Información en 1998. Los motores de búsqueda más importantes comenzaron a aplicar este análisis, una técnica que explotaba la información adicional inherente en la estructura de enlazado entre los sitios de la Web, para mejorar los resultados ante una consulta dada.

Entre todas las técnicas utilizadas para realizar la clasificación, quizás

[1] *Ranking elaborado por Alexa, una subsidiaria de Amazon que crea un listado de los sitios web más importantes en base al tráfico generado cada 3 meses.*

las más usadas sean las denominadas "clasificación de objetos basados en enlaces" (*link-based object ranking*, LBR). El objetivo de este enfoque es el de extraer información del grafo generado por los enlaces para realizar la clasificación. *PageRank*, es un buen ejemplo de este tipo de clasificación. Este algoritmo, semilla del algoritmo que hoy utiliza Google en su buscador, valora a cada uno de los nodos en función tanto del número de nodos que le apuntan como de la importancia de cada uno de ellos. Otro buen ejemplo es HITS[2]. En este caso, divide las páginas web en dos tipos distintos. Por un lado, las autoridades, "*authorities*", aquellas que son enlazadas a una gran cantidad de concentradores, "*hubs*", que, a su vez, enlazan a una gran cantidad de autoridades. Así, por ejemplo, un periódico como "*El País*"[3] sería una autoridad, mientras que "*Menéame*"[4] sería un concentrador.

Pero en la actualidad, los buscadores online no se mantienen estancados en sus métodos de búsqueda. Las técnicas más peculiares llegan a los diferentes sitios Web. Concretamente, algunos lugares ya permiten seleccionar aquella música que se va a escuchar basándose en cómo se siente la persona[5]. Muchos de estos avances están siendo posibles gracias a la comunidad científica. Por ejemplo, la detección de sentimientos también se ha aplicado a los videos. En concreto, investigadores japoneses han profundizado en el tema, consiguiendo desarrollar un sistema capaz de identificar los videos que el usuario quiere visualizar en base a los colores y los sentimientos[6]. Pero las tendencias actuales para los buscadores no se quedan aquí. Desde el país del sol naciente, otros investigadores han ampliado las características de las búsquedas habituales al añadir una vertiente social[7]. Estos investigadores se centraban en la búsqueda de lugares de interés dentro de tu viaje basándose en las experiencias previas de otros usuarios. Aunque todo esto pudiera parecer ciencia ficción, cada vez están proliferando más este tipo de desarrollos en la red.

Teniendo en mente todas estas técnicas, y conociendo la evolución sufrida en el mundo de los motores de búsqueda, se puede afirmar que, como dogma general, desde *Yahoo*! a *Microsoft*, pasando por el propio *Google*, todos basan la potencia de sus buscadores en: i) la capacidad de indexar el contenido generado en Internet de la manera más rápida y efectiva posible; y ii) una vez adquirida toda esa información, ser capaces de ofrecer al usuario el contenido deseado en base a una consulta determinada, independientemente de la naturaleza de la misma consulta.

Web crawlers, arañas o buscadores jerárquicos: los encargados de indexar la web

Las arañas son los agentes de los motores de búsqueda con el objetivo marcado de recuperar información sobre los diferentes sitios web. Son pequeñas piezas de software a las que se proporciona un conjunto de directorios raíz a visitar para, empezando por ellos, seguir los enlaces a otras páginas y descubrir así nuevo contenido. Pero el volumen de información que se maneja en Internet, la corta fecha de caducidad de dicho contenido y la celeridad con que es actualizado, no hace sino complicar en gran medida esta tarea.

Así, estos sistemas se enfrentan a problemas tales como: ¿Qué contenido indexar? ¿Cada cuánto tiempo volver a visitar un sitio ya indexado? ¿Qué aspectos éticos deberían tenerse en cuenta? ¿Cómo coordinar el trabajo de los diferentes *web crawlers*? El motor de búsqueda capaz de dar respuesta a todas estas preguntas y de realizar el diseño más acorde a estas áreas de actuación, es el que consigue una ventaja competitiva respecto al resto.

Conociendo lo que quieres: el dossier digital del usuario

Una vez adquirida toda la información disponible en Internet, es importante saber qué contenido presentarle al usuario. Desde el comienzo de los motores de búsqueda se han ido evolucionando los sistemas de recuperación de documentos, mejorando los sistemas de indexado y dotándolos de una mayor comprensión de los criterios de búsqueda introducidos, aplicando, por ejemplo, diferentes técnicas de *Procesado de Lenguaje Natural* (PLN). Pero el siguiente paso diferenciador al que apuntan todos ellos no es otro que el de conocer mejor a la persona que está pidiendo la información, conocer mejor al usuario.

En este sentido, el impacto que han tenido tanto las redes sociales *online* como los *smartphones* (teléfonos inteligentes), ha hecho cambiar completamente el paradigma no solo de las tecnologías de la información sino también de los buscadores. Toda la información que el usuario pone a disposición de estos servicios facilita la creación de perfiles digitales muy precisos entre los que se incluyen preferencias, gustos, deseos o sentimientos. Estos perfiles o dosieres digitales se emplean posteriormente para ofrecer búsquedas más satisfactorias (ejemplo de ello es la reciente integración de las búsquedas de Google con su propia red

[2] http://dl.acm.org/citation.cfm?id=324140
[3] http://elpais.com/
[4] http://www.meneame.net/
[5] http://musicovery.com/
[6] http://link.springer.com/chapter/10.1007/978-3-642-32597-7_15
[7] http://link.springer.com/chapter/10.1007/978-3-642-32597-7_13

social *on-line Google+* o la integración de *Bing* con *Facebook*). Pero no sólo conocen lo que el usuario publica en el perfil de su red social on-line preferida. El futuro de la tecnología se acerca cada vez más a los hogares inteligentes, los coches inteligentes, el "todo" inteligente. ¿Qué significa esto? Se podría decir que se traduce en más información para los ya gigantes de la información. Conociendo hábitos alimentarios (p.ej., mediante neveras inteligentes), rutas y medios de transporte (p.ej., pago de transporte con el móvil mediante NFC* o coches conectados a internet) o problemas físicos (p.ej., el expediente médico digital), éstos serán capaces de personalizar los resultados de sus búsquedas hasta límites insospechados.

Y, ¿esta efectividad perseguida por los buscadores qué consigue? Consigue convertirlos en una herramienta por defecto para la búsqueda no sólo de información, sino de recursos y conocimiento. Esto revaloriza sus servicios y facilita la aplicación de sistemas de monetización como los programas de publicidad patrocinada.

Pero el conocimiento que los buscadores quieren tener sobre el usuario no se remite a información sobre sus gustos o preferencias. Una nueva corriente, propiciada en parte por las posibilidades que ofrecen los mencionados *smartphones*, va en la dirección de identificar las necesidades del usuario en base al contexto en el que se encuentre, ya no virtualmente, sino en el mundo real.

Sabiendo dónde estás: dime qué te rodea y te diré lo que quieres

La posibilidad de tener conocimiento de la localización real del usuario permite que los buscadores personalicen aún más los resultados a devolver ante una consulta. Claro ejemplo son los cada vez más populares asistentes de voz virtuales que podemos encontrar para las diferentes plataformas móviles. *Apple* cuenta con *Siri*[8], *Google* con *Google Now*[9], *Samsung* con *S-Voice*[10] y, como alternativa española, *Sherpa*[11].

Algunos medios especializados definen estos nuevos asistentes como el final de las búsquedas tal cual las conocemos hoy en día. El por qué a esta afirmación se encuentra en la capacidad que tienen estas nuevas herramientas de entender preguntas en su contexto geográfico, lo que, sumado a toda la información previamente adquirida sobre el usuario, permite devolver al usuario no una serie de resultados, sino una respuesta precisa y personalizada a su consulta.

Aquí tiene lo que pedía: Ok Glass, búscame un restaurante en la zona

Desde que Internet empezó a crecer, se vio la necesidad de herramientas que ayudasen a localizar la información relevante. Sin embargo, la explosión en el volumen de información de todo tipo (e.g., textos, fotos, vídeos) que se ha producido en los últimos años, unido al aumento en el nivel de exigencia de los usuarios, hacen que los buscadores se esfuercen sobremanera para adaptarse a este nuevo entorno. Un ejemplo claro de ello son las *Google Glass*[12], que integra toda la potencia del buscador de *Google*, con toda tu información alojada en su plataforma disponible en un complemento para las gafas, controlado por la voz.

Lo que nació como una pequeña red de ordenadores conectados se ha convertido en la mayor fuente de datos de la historia de la humanidad, haciendo realidad el sueño de la *biblioteca de Alejandría*. Sin embargo, es muy fácil perderse entre tantos datos e información. Por eso, la labor de los buscadores hoy en día es procesar toda esa información y conseguir un nuevo avance en la historia de la humanidad: pasar de la era de la información a la era del conocimiento.

[8] *http://www.apple.com/ios/siri/*
[9] *http://www.google.es/landing/now/*
[10] *http://www.samsung.com/es/galaxys3/svoice.html*
[11] *https://play.google.com/store/apps/details?id=com.sherpa.asistentesherpa&hl=en*
[12] *http://www.google.com/glass/start/*

** Near-Field Communication o Comunicación de Campo Cercano, es un estándar creado para la comunicación sin cables de corto alcance para la realización de pagos a través de nuestros dispositivos móviles mayoritariamente.*