Recuperación de información en Internet: Google, un buscador singular

Alicia García de León

Adriana Garrido Díaz

RESUMEN

Se parte del análisis de los problemas que plantea la recuperación de información en Internet desde la perspectiva de los profesionales de la información. Se examinan las diferentes herramientas existentes para tal fin: directorios, buscadores y metabuscadores y su diferencia con los portales, la problemática de la web invisible y del procesamiento del lenguaje natural para su recuperación. Se analiza la propuesta que realiza el buscador Google, basada en el ordenamiento de los resultados de una búsqueda según la cantidad de enlaces que tiene cada página en el conjunto de la red y las mejoras que pretende incorporar el proyecto Clever. Se propone que los profesionales de la información, desde espacios multidisciplinarios, optimicen la creación de documentos y trabajen en la normalización de los mismos.

ABSTRACT

This paper is based in the analysis of the problems existing regarding the information retrieval in Internet from the point of view of the information professionals. The authors study the different tools: directories, search engines and meta search interfaces, as well as the difference between these ones and the portals, the invisible web problematic and the processing of natural language for information retrieval. After this they describe the singular proposal that Google carries out, which is based on the order that is given to the results of a search according to the amount of links each page has in the within the Network and the improvements that the Project Clever tries to introduce. The proposal is the following, to try that information professionals could optimize the elaboration of documents and work on its standardization.

Recuperación de información en la red

El escenario

En las últimas décadas, y sobre todo a partir de los avances tecnológicos, hemos asistido al incremento exponencial de la información, aunque la capacidad humana para su lectura se mantiene básicamente constante, por lo cual nos enfrentamos a la ineludible limitación de leer únicamente lo que se ajusta a nuestras estrictas necesidades. Esta problemática ha dirigido grandes esfuerzos teóricos para la generación de modelos conceptuales de recuperación pensando en sistemas que filtren el exceso de información para devolvernos solo lo pertinente.

> Muchas veces no existe, al nivel de usuario final, una clara comprensión del fenómeno recuperación en la web.

La información que se encuentra en Internet presenta ciertas características que hace su recuperación aún más compleja. El volumen abrumador, la capacidad para el cambio constante, la casi inexistencia de normativas para la publicación y edición, la falta de

estructura, así como la ausencia de arbitraje sobre calidad de contenido son algunas de las dificultades que hacen de la información en Internet un recurso tan huidizo como aplastante en su generosidad. La búsqueda en la red puede volverse entonces una tarea titánica que lleva a los usuarios a desistir en sus intentos porque en ese laberinto parece imposible encontrar la información que se busca y, por lo tanto, la mayoría de las veces, los resultados de búsqueda no le hacen justicia a los contenidos de la red. La consecuencia directa de este fenómeno es que muchas veces no existe, al nivel de usuario final, una clara comprensión del fenómeno recuperación en la web.

Herramientas de búsqueda: los medios de recuperación de información en Internet

Desde los inicios de Internet se han creado herramientas de búsqueda con el fin de suministrar a los usuarios acceso a la información que demandan. Aún no desarrollada la web, ya existía Archie, un localizador de archivos disponibles en el protocolo FTP creado en 1990; más tarde, en 1993, se creó Verónica, un buscador que permitía la recuperación dentro de los recursos presentada en Gopher.

Actualmente existen en la red diferentes herramientas de búsqueda; básicamente se pueden distinguir los directorios, los buscadores o motores de búsqueda y los metabuscadores. Aunque son muchas las diferencias que las separan entre sí, comparten la limitación para indizar el contenido de la red por completo y han debido asumir la imposibilidad, hasta el presente, de hacerlo en su totalidad. El volumen abrumador es una de las características de la información en Internet que solo está controlada en forma parcial, por un lado, debido al retraso que llevan y, por el otro, por un problema aún más preocupante conocido como "web invisible" o infranet.

Los motores de búsqueda solo pueden indizar cierto tipo de páginas y quedan fuera las bases de datos, sean estas bibliográficas, a texto completo o las obras de referencia; las revistas electrónicas y también los archivos de documentos; sin olvidar aquellas páginas que desde los metadatos autoexcluyen su indización. A pesar de toda la gama de herramientas que se nos ofrecen en la red**,** las búsquedas nunca serán sobre el total de lo que allí se aloja, por lo menos con la tecnología que se cuenta en el presente.

Volviendo sobre el tema del atraso en la indización, se debe destacar que los directorios llevan meses de retraso porque deben adecuar cada web a la clasificación temática utilizada. Los motores de búsqueda llevan ventaja en este aspecto, pero de todas formas el atraso en la incorporación del material nuevo existe porque este crece cotidianamente. Además del retraso, en la mayoría de los casos estas herramientas indizan aquellas páginas sugeridas por los usuarios; por lo tanto, si el autor no lo hace disminuirá aún más su posibilidad de indización y, por ende, de visibilidad en la red.

Antes de continuar, es preciso determinar la diferencia —cada vez más difícil de establecer a simple vista— entre portales y herramientas de búsqueda.

Actualmente existen en la red diferentes herramientas de búsqueda; básicamente se pueden distinguir los directorios, los buscadores o motores de búsqueda y los metabuscadores.

El usuario medio utiliza Internet para conseguir información y también para realizar operaciones comerciales sencillas y en las que encuentre alguna ventaja sobre su trámite en el mundo real. En un principio, la red se veía como un nuevo medio de comunicación que las facilitaba y abarataba, pero con el tiempo y los avances tecnológicos en lo que tiene que ver con la cuestionada seguridad de Internet, se vio que podía ser un canal satisfactorio para el comercio.

Casi todas las empresas que se precien de estar al día tienen su página web; en un primer momento ofrecía información, pero en la actualidad la mayoría agregó servicios y productos que comercializan por intermedio de la red. Esta transformación, entre otras, introdujo un cambio de enfoque del concepto internauta, el cual pasó de ser un simple usuario a convertirse en un potencial cliente.

Los portales se dieron cuenta que si se quería captar al usuario se le debía ofrecer todo lo que buscaba desde un único sitio con los enlaces adecuados para cubrir un amplio espectro de posibles necesidades. Por tanto, se podría definir portal como un punto de

entrada a Internet con contenidos organizados y servicios y productos, tales como: búsquedas, páginas blancas y amarillas, correo electrónico gratuito, espacio para la generación de páginas web personales, compras en línea, etc., de forma tal que les permitan al usuario hacer cuanto necesite o, al menos, que pueda encontrar allí todo cuanto utiliza a diario sin tener que salir de ese sitio web.

El objetivo es cautivar al usuario para que se acostumbre a navegar a través de las opciones que le ofrece el portal elegido. Esto tiene una explicación económica ya que dichos portales venden sus espacios publicitarios de acuerdo con las visitas obtenidas. La consecuencia directa de esta transformación es que muchas herramientas de búsquedas perfilaron sus páginas como portales porque era una forma rápida de obtener dividendos. A pesar de todo esto, no creemos que sea una opción deseable para aquellos que desean únicamente realizar una búsqueda enfrentarse a una página que si no distrae, por lo menos entorpece el objetivo primario.[1]

Lo dicho hasta aquí, permite apreciar que la definición y los objetivos de un portal difieren del objetivo que persigue un buscador, el cual intenta acercar al usuario a una búsqueda lo más pertinente posible o el que pretende un directorio, ya que intenta auxiliar al usuario en aquellos casos que sus búsquedas son por grandes temas.

No se desconoce la necesidad de financiación de estas herramientas, pero lo más rápido no siempre es lo mejor y si para obtener esos recursos se desvirtúa el contenido, se corre el riesgo de confundir a sus fieles usuarios al ver desdibujado el objetivo principal de la página.

Directorios

Un directorio es una herramienta que proporciona una organización temática jerárquica con la cual se permite recorrer los recursos de información de Internet; en definitiva serían catálogos con una ordenación temática. Bajo cada categoría o subcategoría, se listan los enlaces de los documentos que corresponden a un tema. No admiten la polijerarquía, por lo tanto, cada página web se clasifica en un único epígrafe trasladando de esa forma el problema de la "ubicación física" de los documentos a la red. Cuando se clasifica un documento para establecer su lugar en una colección, aunque el sistema utilizado acepte la relación temática, deberá tener una ubicación única. Sin embargo, cuando hablamos de la red no se entiende un criterio tan pobre sino como respuesta simplificadora a una tarea bien compleja. Para colaborar aún más con esa simplificación se trabaja solo con las páginas de inicio. La clasificación es efectuada por el autor o los autores del documento, o por los administradores del directorio, por lo tanto, se trata de una clasificación manual y es aquí donde se entiende la necesidad de simplificar la forma de trabajo que se realiza.

A pesar de esta debilidad, pueden ser muy útiles en ocasiones. Lo importante es reconocer sus características para saber cuándo puede servirle al usuario. Los directorios son adecuados cuando no se tiene una idea precisa de lo que se busca o cuando interesa buscar por grandes temas: para conocer el estado de una disciplina o el tratamiento que un tema recibe en Internet.

Algunos directorios poseen también mecanismos propios de búsqueda por palabra clave, otros "disparan" la búsqueda que no pueden resolver por su condición de directorios, a un buscador que la resuelve sin que el usuario cambie de interfaz. Si, por ejemplo, se quiere hacer una búsqueda en Yahoo por el término "pelícano", este directorio envía la búsqueda a Google, y en el cabezal de la página de respuesta coloca "Yahoo Powered by Google".

La mayoría de los directorios presentan breves anotaciones e información general sobre sus enlaces, otros incluso, añaden un indicador de pertinencia del documento con respecto al tema buscado.

Los directorios son adecuados cuando no se tiene una idea precisa de lo que se busca o cuando interesa buscar por grandes temas: para conocer el estado de una disciplina o

1 *Escapa al alcance de este trabajo todo lo que tiene que ver con la ergonomía de las páginas, pero hay suficiente bibliografía sobre el tema que orienta en lo que tiene que ver con la necesidad de una clara relación entre lo que se ofrece y los objetivos que se establecen.*

el tratamiento que un tema recibe en Internet.

Algunos ejemplos de directorios:

- Yahoo <http://www.yahoo.com>.
- Looksmart <http://www.looksmart.com/>.
- Ozú <http://www.ozu.es/>.

Buscadores

Con la creación y adopción masiva de la web en 1993 y las posibilidades que la tecnología puso a nuestra disposición, se crearon herramientas llamadas buscadores (conocidas también como *robots*, motores de búsqueda, *search engines, crawlers*, orugas, *spiders* o *worms*, muchos de ellos, términos sin siquiera traducción al español pero que se utilizan en nuestra lengua con mucha frecuencia, para hacer referencia a la misma herramienta).

Los buscadores son programas computacionales que recorren Internet examinando la información de acceso público en la red para su indización y almacenamiento; con este material se generan bases de datos en constante actualización, que permiten su interrogación por palabra clave para la recuperación de la información.

Como señalábamos más arriba, en los buscadores la indización es realizada por los robots que son programas que rastrean la red sin pausa para identificar los recursos de dominio público —ya se hizo mención anteriormente a la posibilidad que tienen las páginas de autoexcluirse de los buscadores— e indizarlos. Esta información pasa a formar parte de la base de datos del buscador y una vez interrogados presentan una lista de enlaces (*links*), muchas veces anotada, que llegan a incluir la extensión del documento y grado de pertinencia. La estructura de los buscadores varía de acuerdo con el tipo de indización, el tamaño de su índice, la frecuencia de actualización, las opciones de búsqueda, el tiempo de demora para su respuesta, la presentación de los resultados y las facilidades de uso. Es por este motivo que una misma búsqueda podrá arrojar resultados diferentes en los distintos buscadores. Si bien todos comparten los rasgos necesarios para entrar en esta categoría de herramientas, las disimilitudes entre ellos pueden ser abismales en algunos casos.

Generalmente ofertan la información recogida en sus propios sitios web, donde pueden ser interrogados por los usuarios, aunque también es posible llegar a sus servicios a través de aquellos portales que los adquieren. Las capacidades de recolección de información, así como de interrogación y suministro de respuestas han variado con el tiempo y varían de acuerdo con las potencialidades de los diferentes buscadores. Inicialmente recogían pocos datos de las páginas que visitaban, como título y primeras palabras del texto, más tarde comenzaron a "leer" los metadatos y considerarlos para la recuperación de la información y su descripción, actualmente algunos llegan a indizar el texto completo. También existe una diferencia entre ellos en cuanto al material recogido, pues mientras unos operan solo sobre recursos web otros empiezan a incluir otro tipo de material.

En los buscadores la indización es realizada por los robots que son programas que rastrean la red sin pausa para identificar los recursos de dominio público e indizarlos.

Si bien existe un amplio abanico a la hora de indizar no ocurre lo mismo en cuanto al criterio para declarar la pertinencia de una referencia, ya que para la mayoría de los motores de búsqueda este es similar. Tomada del título y las primeras palabras, de los metadatos o del texto completo, básicamente los buscadores indizan por el número de ocurrencias de un término y devuelven las referencias a los documentos por un sistema de *ranking* (originalmente se presentaban sin ningún tipo de orden).

Algunos ejemplos de buscadores:

- Google <http://www.google.com>.
- AltaVista <http://www.altavista.com>.
- Northernlight <http://www.northernlight.com>.
- Hotboot <http://www.hotbot.com>.
- Kartoo <http://www.kartoo.com/>.

Las búsquedas en los buscadores

Tradicionalmente los buscadores reciben demandas de los usuarios y suministran, en consecuencia, una lista de enlaces a documentos posiblemente pertinentes, pues poseen el o los términos solicitados.

Las búsquedas pueden ser optimizadas por:

- Uso de operadores boléanos.
- Especificación del alcance geográfico, fechas, idioma del documento, tipo de recurso, tipo de dominio.
- Formulación de su pregunta en forma de frase completa.
- Diferenciación entre letras mayúsculas y minúsculas.
- Búsqueda de un término presente en un URL.
- Sugerencia de expresiones de búsqueda complementarias o presumiblemente vinculadas.
- Truncamiento de palabras clave, posibilitando que se devuelva todo lo que contiene ese principio.
- Incorporación de clasificaciones temáticas (directorios) a sus páginas.
- Suministro de páginas con textos de ayuda y de búsqueda asistida.
- Posibilidad de disparar la misma búsqueda a través de su propia interfase a otros buscadores.
- Gráfico de las relaciones entre los términos existentes en la base, sugiriendo nuevas operaciones de búsqueda.

Criterios para el establecimiento de pertinencia

Luego de presentada la búsqueda en la base de datos, esta ofrece un conjunto de referencias que coinciden con lo solicitado, pero se supone que unas sean más acorde con la búsqueda que otras. En consecuencia, la mayoría de los buscadores recurren a normas heurísticas para establecer el criterio de ordenación del material solicitado. Para efectuar esa ordenación los programas utilizan mecanismos automáticos que colocan primero los sitios que consideran más pertinentes de acuerdo con los criterios que se detallan a continuación:

- *Por frecuencia de ocurrencia de un término:* Se considera que aquellas páginas que reiteran el término de búsqueda una mayor cantidad de veces serán más pertinentes. Este criterio ha generado el *spamming*, fenómeno por medio del cual los constructores de páginas web se aseguran que su página se recupere incluyendo muchas veces el mismo término o incluso introduciendo muchas veces términos que se saben muy buscados —como *sex*— y que no tienen que ver con el contenido de su página, pero que le aseguran la recuperación y de esa forma publicitan su web. Son páginas que se construyen pensando en su difusión y en la venta de sus espacios publicitarios, no en su contenido y para eso se adecuan a las prácticas de indización que se saben utilizan los buscadores. El problema de este fenómeno es que introduce mucho ruido en la recuperación y distorsiona los contenidos. Por esa razón, en muchos buscadores, los sitios que en su texto o dentro de los metadatos incluyen un mismo término en forma reiterada para que obtengan uno de los primeros lugares dentro de los listados, son penalizados por no éticos y no indizados.
- *Frecuencia y lugar de un término en el documento:* Algunas veces la frecuencia de un término se combina con el lugar que ocupa dentro del documento. Por ejemplo, a un término que aparece en el alto del documento se le asigna más valor que a la misma ocurrencia en otra parte del documento. A pesar de los defensores de este criterio, se sabe que no siempre hace justicia al contenido ya que muchas veces se utilizan sinónimos o, en el peor de los casos, el término que designa el tema no se menciona y provoca un silencio que no se ajusta a la realidad. Existen otros documentos, que por su naturaleza (tratados, leyes, convenios, acuerdos, etc.) incluyen información clave para la indización al pie y no pueden ser alterados porque esto forma parte de su estructura. En estos casos la aplicación del criterio que nos ocupa en este apartado es altamente perjudicial.
- *Importancia y popularidad*: Se le asigna mayor valor a un sitio al que enlazan sitios referenciales o un sitio que es muy enlazado por el resto de los sitios y tendría el mismo sustento de fondo que el análisis de citas. Este tipo de clasificación se basa en un relevamiento de los enlaces hipertextuales, si bien en una primera instancia, como es lógico, buscan la ocurrencia del término dentro de los documentos. En este último caso la presentación de la información resultante de la búsqueda es bien diferente a la organización por la frecuencia de ocurrencia. No basta que el término esté presente en el texto, se incorpora una categorización sobre su impacto en la red.

En realidad el tercer criterio detallado implica un cambio radical en la consideración del elemento que establecería la pertinencia mientras que los dos primeros obedecen a una gradación en el mismo concepto. La primera pregunta que deberíamos hacernos es si un único criterio es suficiente para determinar la pertinencia de la respuesta en una búsqueda. Responder esta pregunta es algo bien complejo, pero lo que podemos afirmar, sin temor a equivocarnos, es que la adopción de un criterio único genera prácticas poco lícitas, como el mencionado *spamming*, y proyecta fácilmente las debilidades del sistema.

Ninguno de estos criterios es perfecto y por eso muchas veces los resultados de las búsquedas dejan al usuario un poco perplejo entre el ruido y la constante interrogante que se plantea sobre todo lo que la red alberga y no supo recuperar. Esa es una constante desconfianza que marca la relación que se establece con la red y a debilitar esa desconfianza están orientados todos los esfuerzos de los buscadores para mejorar las herramientas de que disponen.

Si consideramos los dos primeros criterios, nos enfrentamos a que no siempre las páginas pertinentes por sus contenidos, incluyen el término demandado por el usuario. Una de las principales causas de este problema es que el lenguaje humano está lleno de sinónimos, expresiones cotidianas o regionales, y polisemias. Muchos equipos lingüísticos han desarrollado especies de tablas de equivalencias que no son visualizadas por los usuarios pero convierten los términos o los singularizan; sin embargo, todavía estamos ante soluciones parciales y muy lejos de ser las deseadas. Varias herramientas de búsqueda han procurado enriquecerse incorporando estos recursos y no se puede negar que lo han hecho. De todas formas la mayoría de los buscadores dejan al usuario el problema del lenguaje y ya desde sus hojas de ayuda plantean que si una búsqueda recupera demasiados documentos se deben incorporar palabras clave para acotarla, así como si su respuesta es el silencio se debe replantear con sinónimos, plurales/singulares, etc.

El usuario de Internet, no documentalista, desconoce en su mayoría la problemática de la recuperación de información y no está acostumbrado a pensar en esos términos. Se le proponen herramientas como los buscadores que ofrece grandes masas de información, pero con pocas interfaces amistosas que lo orienten adecuadamente.

Los metabuscadores son de estructuras que permiten "disparar" una búsqueda hacia varios buscadores en forma simultánea.

En la actualidad, a pesar de los esfuerzos que han realizado estas herramientas para mejorar la interfaz de usuario, trazar perfiles y acotar las búsquedas mediante las posibilidades de optimización, los buscadores siguen dejando al usuario la tarea más difícil: intuir las intenciones de los autores y sobre esa base seleccionar adecuadamente las palabras clave para realizar su búsqueda. Si al incorporar términos para la recuperación en una base de datos de una biblioteca con lenguaje natural se arrastra al usuario al desconcierto, en Internet el desconcierto se transforma en desconfianza porque la masa documental es abrumadora e intangible. Pedirle al usuario final que se enfrente a este problema no es algo menor y la consecuencia directa es que este se limita a visitar lo conocido y a lo que en una primera experiencia funcionó satisfactoriamente. Es de esta situación que se sirven los portales para ofrecer sus servicios, sus caminos seguros.

Metabuscadores o agentes multibuscadores

Los metabuscadores son de estructuras que permiten "disparar" una búsqueda hacia varios buscadores en forma simultánea. Se les llama *simultaneous inified search interfaces* (SUSIs).

Los metabuscadores no disponen de una base de datos propia, sino que utilizan la información almacenada en las bases de datos de otros buscadores y directorios. En la gran mayoría de los metabuscadores, los usuarios pueden seleccionar los buscadores en los que quieren que se efectúe la búsqueda.

Algunos ejemplos de metabuscadores:

- Dogpile <http://www.dogpile.com>.
- Digisearch <http://digiway.com/digisearch>.
- Meta Crawler <http://www.metacrawler.com/>.
- Mamma <http://www.mamma.com/>.

Google

El buscador Google <http://www.google.com/> es un proyecto de investigación de la Universidad de Stanford, iniciado a finales de 1997. El nombre Google deriva de *googol*, término creado por el famoso matemático Edward Kasner para denominar

al número 1 seguido por 100 ceros. Los creadores de Google, adoptaron este término para simbolizar su objetivo de organizar la enorme cantidad de información disponible en la red.

Algunos motores de búsqueda, por ejemplo, Google y el proyecto Clever, se basan en que la ordenación de los resultados de una búsqueda, se efectúe de acuerdo con el número de enlaces que tiene cada página en el conjunto de la red. De esta manera son las páginas más referenciadas las que se presentan en primer término. Google clasifica las referencias a las páginas web que entrega como respuestas, en función de su popularidad. Esta idea presupone que una página muy citada en la web es más sólida y pertinente que una que posee pocas citas. En consecuencia, no es la sola ocurrencia de un término lo que determina la aparición de una referencia.

Los enlaces son un recurso que caracteriza la estructura de la red y tomarlos en cuenta enriquece enormemente las búsquedas; el seguimiento de los enlaces se puede efectuar en forma manual por técnicos, pero en este caso lo exhaustivo es opuesto a lo preciso. Analizando los sitios manualmente podríamos llegar a ser muy precisos pero solo sobre un número muy restringido de documentos. Google usa tecnología PageRank™, cuya patente está en curso, para incorporar el "factor enlace" a la recuperación. Gracias a esa tecnología, Google suma como valor a la página los enlaces que apuntan a ella en el conjunto de la red y considera que cada vez que un sitio web es presentado como enlace en la red se está realizando un voto implícito a su favor. PageRank™ es el indicador general de importancia de Google y no depende de una consulta efectuada por el usuario, se trata de la característica de la herramienta.

Metodológicamente este mecanismo tiene estrecha vinculación con los conceptos que sustenta el análisis de citas utilizado para la valoración de los artículos científicos. El análisis de citas estudia la regularidad con que los artículos remiten unos a otros y juzga el valor de una publicación por el número de veces que es citada. Google presupone y asume que los enlaces hipertextuales se comportan como las citas bibliográficas.

Considerar el número de citas de un documento como un sinónimo de su valor, es un criterio muy cuestionado y largamente discutido. La existencia de una cita —en el caso de Internet, un enlace— no significa necesariamente un juicio positivo de valor. Sin embargo, esta premisa se sustenta en parte en virtud del "efecto Mateo", señalado por Merton en 1968, y cuyos adeptos defienden fervorosamente. Este nombre alude a un pasaje del Evangelio según San Mateo, la parábola de los talentos: "...al que tiene se le dará más, y tendrá abundancia; pero al que no tenga se le quitará hasta lo poco que posea". Este efecto hace referencia a la forma en que influyen factores, tales como, el reconocimiento que tengan los autores, el equipo de investigación o la universidad responsable en el número de citas que recibe un documento.

Algunos motores de búsqueda, por ejemplo, Google y el proyecto Clever, se basan en que la ordenación de los resultados de una búsqueda, se efectúe de acuerdo con el número de enlaces que tiene cada página en el conjunto de la red.

Es muy fácil que un autor, equipo o universidad exitoso, conocido, asociado a un concepto clave o a una investigación famosa siga siendo citado y que cuando elabore un nuevo documento, se considere a *priori* de gran valor y se remita a él. Mucho más difícil es que un autor, equipo o universidad se "abra paso" desde cero. El valor de su producción por mayor que sea está más oculto y para salir a la luz tendrá dificultades. Las realizaciones pasadas ganan, al menos en una primera instancia, la partida.

El reconocimiento tiende a superponerse y la falta de reconocimiento a reforzarse negativamente. Por tanto, el número de citas o enlaces hipertextuales es un dato atendible, importante, pero no define la relevancia, y menos aún la pertinencia, en consecuencia, no puede sustentar por sí solo la recuperación.

Peter Ingwersen ha elaborado el concepto de factor de impacto web (*web impact factor*), llamado WIF, un indicador que nace del cociente entre el número de citas externas que recibe un sitio web y el tamaño de sí mismo expresado en su número páginas. El WIF puede obtenerse a través de buscadores como Altavista ltavista.com. De hecho, el WIF nunca sustenta por sí mismo una evaluación de calidad,

aparece como un dato más, es una medida relativa, que tiene que ver con la popularidad.

Todos los sitios quieren notoriedad y el WIF puede indicar éxito para los creadores de un sitio web, pero está muy claro que no necesariamente aporta a la pertinencia o relevancia de los contenidos con respecto a una búsqueda.

Otras propiedades de Google

- Presenta un diseño simple, claro y austero con fondo blanco, logotipo de la empresa y cuadro de diálogo muy sencillo. Tiene ayudas disponibles. No busca ser un portal, no presenta ni publicidad, ni correo gratuito, ni *chats*, no hay horóscopos, noticias recientes. No presenta *banners* ni despliega nuevas ventanas. Google es simple y llanamente un buscador.
- Permite al usuario seleccionar el idioma de diálogo de la interface.
- Almacena las páginas web que visita en su memoria *caché* con el fin de recuperarlas para los usuarios como una copia de seguridad, en caso de que el servidor de la página falle temporalmente (el muy conocido *error 404,* por ejemplo). Si el servidor no se encuentra operativo o la página ya no existe, el usuario puede apelar a ese *caché.*
- Permite ir directamente al primer resultado de la búsqueda, al activar la expresión *I'm feeling lucky* (Me siento afortunado), una vez la búsqueda planteada y ejecutada.
- Tiene indizadas en su motor de búsqueda 1 326 920 000 de páginas, un número que le convierte en el buscador con más páginas disponibles en toda la red.
- Ha incorporado en versión beta, un traductor automático del texto completo de las páginas enlazadas y presentadas en los resultados de búsqueda <http://www.google.com/intl/es/machine_translation.html>. En esta primera instancia solo opera sobre páginas escritas en inglés. Ejemplo de enlace presentado como resultado y la oferta de traducción automática:

  Google Search Technology -[¡Novedad! Traduce esta página]
  ... The heart of our software is PageRank™ (TM), a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And <www.google.com/technology/ > -7k - En *caché* -Páginas similares
- Presenta un listado de preguntas frecuentemente formuladas (FAQs).
- Crea automáticamente versiones texto de documentos PDF (*Portable Document Format*) cuando explora la web. Si un documento está originalmente disponible en formato PDF , Google permite, cuando lo presenta como resultado de búsqueda, acceder a él en versión texto. Ejemplo de presentación de resultado:

  [PDF]
  <www.enssib.fr/bbf/bbf-2001-1/10-rostaing.pdf>
  Le Web et ses outils ... média est deve nu BBF 68 2001 Paris, t. 46, n° 1 LE WEB ET SES OUTILS D ... re nvois ve rs Hervé Rostaing est maître de ...
  Versión texto - Páginas similares
- Posee una página especial para establecer y sugerir comunicación con los usuarios. <http://www.google.com/intl/es/contact.html>, diferenciando el tipo de contactos.

El aporte del proyecto Clever

Pensando en incorporar otras variables, sin desconocer el valor de la ocurrencia de un término y la frecuencia de las citas de un sitio, nació el proyecto Clever.

Aún en su etapa experimental (no operativo) este buscador, desarrollado inicialmente por la Universidad de Cornell, está hoy a cargo del Almaden Research Center de la empresa IBM en California. Es en esencia un sistema que selecciona las páginas web considerando la cantidad de enlaces existentes entre ellas. Pero no solo mide el número de citas, considera dos tipos de páginas, que desempeñan distintas funciones y aportan diferente significación a los enlaces:

Páginas que son autoridades (*authorities*): son citadas por muchos otros sitios relacionados con el tema, así la página de la Organización de las Naciones Unidas para la educación, la Ciencia y la Cultura (UNESCO) es una autoridad en los temas de Educación.

Páginas llamadas centrales (*hubs*): tienen muchos enlaces hacia las autoridades. Un buen *hub* presenta muchas autoridades, y una buena autoridad será reafirmada al ser enlazada por muchos *hubs.*

Clever busca medir el impacto de una página pero busca un impacto calificado. Para Clever no basta un gran número de enlaces: esto no es sinónimo de calidad y mucho menos de pertinencia, Clever recalcula el puntaje que significa el número de citas, priorizando aquellos enlaces que vienen de las páginas de autoridades, es decir, con más peso y los vincula con las páginas centrales o *hubs*. Es en este tipo de recursos y en su perfeccionamiento que debe trabajarse para lograr una mejor recuperación de la información existente en la red.

Conclusiones

Todas las herramientas de Internet para la recuperación de documentos necesitan continuar mejorando para vencer las debilidades que presentan hasta el día de hoy. Al mismo tiempo es necesario que el usuario conozca las diferencias entre unas y otras para poder optar por la adecuada para cada ocasión.

Para esta mejora que señalamos no es necesario únicamente que la técnica nos brinde más recursos, sino incorporar otras variables importantes para reducir el ruido y el silencio que caracterizan un gran porcentaje de resultados de búsquedas, así como la indiscutible ampliación de interfaces amistosas que orienten al usuario menos entrenado.

La optimización de la recuperación de documentos en Internet es un gran desafío. Así como dar más valor a un término porque está en el alto de un documento no necesariamente le hace justicia al documento y a su contenido, la existencia de muchos enlaces a un sitio no garantiza su pertinencia y calidad. Basta pensar en los clubes de intercambio de enlaces. La traspolación del análisis de citas a la web deja algunas dudas al descubierto. Más allá de todas las críticas que ha recibido el análisis de citas para el ámbito que fue concebido, debemos plantearnos si es adecuado aplicarlo a la estructura de Internet. Cuando se habla del fenómeno de citación se está haciendo principalmente un reconocimiento al aporte de otros autores en el documento que cita; mientras que al hacer referencia al hipertexto estamos refiriéndonos a la estructura de la información en la web. Esto implica una revolución en la escritura y asociarlo con un rasgo hipertextual de la escritura líneal parece apresurado. No dudamos que pueda ser tan pertinente como el análisis de citas cuando se trabaja con páginas del ámbito científico; pero Internet es más que eso y pueden ser muchos los mecanismos que lleven a generar enlaces. Muchas variables deberían entrar en este análisis como, por ejemplo, que cualquier persona puede generar una página web y es difícil evaluar su comportamiento como colectivo.

La traspolación del análisis de citas a la web deja algunas dudas al descubierto. Más allá de todas las críticas que ha recibido el análisis de citas para el ámbito que fue concebido, debemos plantearnos si es adecuado aplicarlo a la estructura de Internet.

Sin embargo, Google contempla e incorpora una variable importante, no es igual un sitio muy referenciado que uno ignoto. Inferir que un enlace es un voto de aprobación es por lo menos un riesgo y una variante importante en relación con los criterios que utilizan la mayoría de los buscadores. No es esta la única variable importante a tener en cuenta, existen otras que incorporan valor: ¿quién hace el enlace? ¿qué autoridad tiene? ¿desde "dónde" se efectúa el enlace y luego, qué significa? Y a estas preguntas intenta dar respuesta el Proyecto Clever.

Sin embargo, hay una variable que hasta ahora no se incorpora y es una categorización de páginas, que nos posibilite desde la búsqueda determinar el tipo de recursos que necesitamos. Por ejemplo, si me interesa el tema análisis de citas desde un punto de vista teórico, solo quiero recuperar artículos científicos, no todas aquellas páginas sobre cursos en los cuales se trata el análisis de citas, entre otras. Un usuario de Internet experimentado podrá decir que es posible acceder solo a los artículos o por lo menos reducir el ruido, pero lo que necesitamos es que la herramienta de búsqueda lo haga por nosotros facilitando la tarea. Si el usuario pudiera escoger esto de la misma manera que el idioma o la fecha, entre otros, el ruido se reduciría y mucho. No podemos olvidarnos que Internet es un medio de comunicación que se utiliza como canal para documentos bien diferentes y las palabras claves o la popularidad no nos ayudan a delimitarlo por completo, solo son dos variables importantes pero no válidas por sí solas. De la misma manera que hay herramientas adecuadas para el tipo de búsqueda que necesitamos hacer, también debería incluirse esa categorización de páginas para que el usuario pudiera determinar qué es lo que busca.

Otro factor clave en la buena recuperación, es la elaboración de páginas web pensadas para ser encontradas. Las páginas web son estructuras de información y deben crearse contemplando el tratamiento que recibirán, todos los esfuerzos de normalización, estructuración clara y ergonómica serán un aporte.

Muchos problemas de la recuperación de información son tributarios de la falta de conciencia de que, quien

hace páginas web hace documentos, quién crea sitios web crea estructuras de información. Igualmente el trabajo de incorporación de metadatos puede pautar y, de hecho pauta, en muchas oportunidades, la diferencia en términos de recuperación y descripción.

Otra parte de la respuesta está en trabajar apostando a la normalización de la información electrónica, la elaboración de pautas orientadas a la calidad. Es un imperativo trabajar para la normalización y la calidad, apostando siempre a la asistencia y la formación de usuarios. Son los profesionales de la información, trabajando en marcos interdisciplinarios los que deben contribuir a esta tarea, las redes son un espacio privilegiado para su acción, siempre exigentes y desafiantes.

Bibliografía

Abilock, Debbie. Choose the Best Search Engine for Your Purpose [en línea]. <http://nuevaschool.org/~debbie/library/research/adviceengi ne.html>. [Consulta: 7 de agosto del 2001]

Almind, Tomas C. y Peter Ingwersen. Informetric analysis on the World Wide Web: A methodological approach to "webometrics". *Journal of Documentation* (Copenhage) 53(4):404-426, 1997.

Brandt, Scott D. Do you have an ear for searching? *Computers in Libraries* (Pasadena) 19(1):42-44, 1999.

Brin, Sergey y Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine [en línea]. Septiembre 1998. <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.h tm>. [Consulta: 7 de agosto del 2001]

Centre de Recherche Informatique de Móntreal. Ergonomie du web [en línea]. 20 septiembre 2001. <http://www.crim.ca/~arbastie/>. [Consulta: 26 de septiembre del 2001]

Davison, Brian. Web Structure Analysis [en línea]. 10 septiembre 1998. <http://www.cs.rutgers.edu/~davison/web-structure/>. [Consulta: 7 de agosto del 2001]

Eugene Garfield, Ph.D. - Home Page [en línea]. 20 abril 2001. <http://www.garfield.library.upenn.edu/>. [Consulta: 7 de agosto del 2001]

García Gómez, Juan Carlos y Tomás Saorín Pérez. Los portales de Internet [en línea]. Mayo 2001. <http://www.um.es/gtiweb/portales/>. [Consulta: 7 de agosto del 2001]

Google. Todo acerca de Google. [en línea]. <http://www.google.com/intl/es/about.html>. [Consulta: 7 de agosto del 2001]

Ingwersen Peter. The calculation of WEB impact factor. *Journal of Documentation* (Copenhage) 54(2):236-243, 1998.

Lowley, S, C. Oppenheim, A. Morris y C. McKnight. Progress *En* Documentation: The Evaluation Of WWW Search Engines. *Journal of Documentation* (Copenhage) 56(2):190-211, 2000.

Rostaing, Hervé. Le Web et ses outils d' orientation: Comment mieux appréhender l'information disponible sur Internet par l'analyse des citations*. Bulletin des Bibliothèques de France* (Paris) 46(1):68-77, 2001.

Sullivan, Danny. Search Engine Features For Searchers [en línea]. 6 setiembre 2001. <http://www.searchenginewatch.com/facts/ataglance.html>. [Consulta: 26 de septiembre del 2001]

Smith, Alastair G. A tale of two web spaces: comparing sites using web impact factors. *Journal of Documentation* (Copenhage) 55(5): 577-592, 1999.

Thelwall, Mike. Web impact factors and search engine coverage. *Journal of Documentation* (Copenhage) 56(2):185-189, 2000.

Tunender, H y J Ervin. How to Succeed in Promoting Your Web Site. *Information Technology and Libraries* (Washington) 17(3):173-179, 1998.

Recibido: 18 de septiembre del 2001.

Aprobado:15 de octubre del 2001.

Alicia García de León

Red Académica Uruguaya
Servicio Centrak de Informática.
Universidad de la República.
Montevideo, Uruguay.
Correo electrónico: <aliciag@seciu,edu.uy>.