BY NC ND

# Xarxa Punt TIC

## MÓDULO 1 NIVEL BÁSICO
## Búsqueda y recuperación de información en Internet
### Unidad didáctica 3: El Proceso de búsqueda de la información

plan avanza2
GOBIERNO DE ESPAÑA
PlanE
MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES Y PARA LA SOCIEDAD DE LA INFORMACIÓN
Generalitat de Catalunya

## ➔ El proceso de búsqueda de la información:

### *Tipología de búsquedas*

Parte de las horas que pasamos delante del ordenador buscando información se pueden ahorrar aprendiendo cómo obtenerla, y otra gran parte examinando y cribando datos completamente irrelevantes para nosotros.

Las herramientas más utilizadas en Internet se basan en el empleo de patrones de búsqueda. Casi nadie accede a bases de datos utilizando complicadas sesiones Telnet, normalmente se hace mediante sistemas de consulta más potentes y cómodos. La experiencia y habilidad para usarlos correctamente es mucho más importante si deseamos encontrar datos que nos sean útiles.

La crisis de la sobreinformación es bastante reciente y se debe a la incapacidad de una sola persona de retener ni siquiera una mínima parte de la información que le es útil. El problema se agrava cuando la información útil es a su vez una parte ínfima de toda la que recibe e inapreciable entre toda la que tiene a su disposición.

Con tanta información disponible lo que debemos saber es cuál nos interesa, dónde encontrarla, cómo obtenerla y la mejor manera de administrarla. Con la ayuda de los navegadores de WWW se facilitan las tareas de localización y obtención, pero saber qué es lo que hay, lo que nos interesa y cómo utilizarlo son cuestiones más complejas.

Hay varios métodos de afrontar el problema; lejos del “divagar” por la Red, que consiste en ir de enlace en enlace sin rumbo fijo, se puede intentar “navegar “ utilizando guías, listas de recursos clasificados, recetarios, etc. con resultados muy diversos. Esto es probablemente lo que llevamos haciendo desde hace tiempo.

Una vez localizadas páginas que contienen referencias a aquellos ítems de mayor interés, es habitual que el usuario haga una lista de los mismos; de hecho, con la mayoría de navegadores de Web , uno puede crearse una **lista de enlaces o marcadores** accesible directamente desde el visor e incluso generar páginas HTML, a partir de ella.

Para facilitar la tarea, cada día hay más sitios dedicados a la creación y mantenimiento de listas y directorios de recursos organizados según criterios geográficos o temáticos denominados generalmente “catálogos”. Pueden ser los puntos de partida ideales cuando buscamos información sobre un tema y queremos saber qué hay en la Red o buscamos algún servicio siguiendo criterios geográficos. A partir de las entradas del catálogo podemos ir descendiendo hasta encontrar lo que realmente nos interesa, simplemente siguiendo la estructura lógica que define el servicio.

De todos modos, el empleo de los catálogos deja de ser adecuado cuando lo que buscamos es más específico, ya que encontrarlo simplemente navegando puede resultar bastante difícil. Nos encontramos entonces con la necesidad de emplear herramientas para la búsqueda y selección de la información disponible

en la Red. Estamos hablando de lo que hemos denominado “buscadores de información”.

Distinguiremos dos métodos de búsqueda en función de la manera de interactuar con las herramientas:

- **Métodos activos**, que serán controlados directamente por nosotros, como los sistemas de búsqueda por patrones relacionados con el tema que buscamos o el acceso a bases de datos y, si lo anterior no resulta, la interacción con otras personas, preguntando directamente a alguien que creamos capacitado para respondernos, ya sea mediante el correo electrónico o utilizando grupos de noticias o listas de distribución que tengan relación con aquello que buscamos.
- **Métodos pasivos**, que serán sistemas automatizados que localizarán la información que queramos a partir de una descripción de nuestro objetivo. Se trata de emplear agentes informáticas que puedan automatizar muchas de las tareas anteriores. Es decir, marcas tus intereses y el sistema se encarga de mantenerte informado únicamente de las novedades sobre esos temas que aparezcan en ciertas listas, grupos o páginas. También puedes optar por sistemas más completos que pueden realizar todo lo anterior y actuar sobre la información contenida en su máquina, funcionando como verdaderos gestores de información.

## *Búsquedas en la Web*

Los sistemas de búsqueda más usados actualmente son los basados en páginas de hipertexto (servicio WWW) donde se introducen patrones o palabras clave a buscar. Estas páginas actúan como intermediarias entre el usuario y una base de datos almacenada en el servidor o accesible para éste. Algunos sistemas restringen la búsqueda al espacio WEB, pero la mayoría permiten buscar cualquier tipo de recursos accesible mediante un URL.

Los sistemas de búsqueda deben indexar en una base de datos propia parte de toda la información para no tener que recorrer toda la Red cada vez que se consulta. Los algoritmos utilizados en los programas de búsqueda se basan en estructurar la información de manera que optimice las búsquedas. El resultado es que los sistemas son prácticamente instantáneos, aunque contengan millones de entradas.

En poco tiempo han aparecido un gran número de sistemas de búsqueda basados en el WEB, con una presentación y posibilidades cada día más atrayentes. Los usuarios se percatan de su existencia por comentarios en la Red, en revistas o a partir de enlaces que aparecen en páginas muy visitadas. También ha contribuido mucho el hecho de que muchos navegadores permiten acceder directamente a algunos de ellos desde un submenú

Para incluir referencias a páginas personales, lo único que debes hacer es seguir las instrucciones que se suele indicar en algún lugar de las páginas de

búsqueda (generalmente mediante formularios). Así puedes ir dándote de alta en aquellos sistemas que te parezcan más interesantes; el problema es que enviar un resumen de tus páginas a todos los buscadores se puede hacer pesado.

En los buscadores que ofrecen el servicio gratuitamente, la causa de sustento económico es la popularidad, que se traduce a efectos comerciales en "audiencia", lo que les permite incluir publicidad en sus páginas, y cómo no, cobrar por ello.

Cuando terminas una nueva página puedes enviar un mensaje anunciándolo a las listas de correo o anunciándolo en un grupo de *news* y también puedes introducir la referencia en algunos de los motores de búsqueda, como www.yahoo.com o www.ozu.com .

## *Motores de búsqueda geográfica*

Es posible buscar información por aproximación geográfica. Este tipo de motor resulta útil si se busca un servicio en un país que aún no cuenta con un gran número de servidores Web. Este tipo de búsqueda se utiliza raramente salvo cuando se conoce con cierta exactitud el nombre del organismo buscado y especialmente su localización.

La interfaz muestra un mapa o lista de países y solicita la zona deseada. La búsqueda se realiza entonces por acercamiento progresivo, sobre mapas o sobre listas cada vez más refinadas (países, regiones, ciudades) hasta llegar a una compilación de los servicios disponibles.

Existe un gran número de este tipo de servicios que recogen geográficamente los recursos de la Web. Algunos están especializados en un continente mientras que otros son específicos de un país.

## *Motores de búsqueda temática*

Son servicios que intentan recoger los recursos de la Web clasificados por temas. La búsqueda se hace a través de rúbricas cada vez más precisas, para llegar a una lista de sedes pertenecientes a una categoría lo más precisa posible.

El Servicio Google así como Yahoo son el ejemplo más clásico de motor de búsqueda temática. En su página de bienvenida ofrece varias opciones:

***Por buscador (también llamado motor de búsqueda o search engine):***

Es el método más usado y recomendado. Es una página web que nos permite encontrar otras páginas web a partir de palabras, estas palabras pueden ser: el nombre de la empresa que estamos buscando, una palabra que indique el tema que queremos buscar, el servicio buscado, etc. Los buscadores más conocidos son:

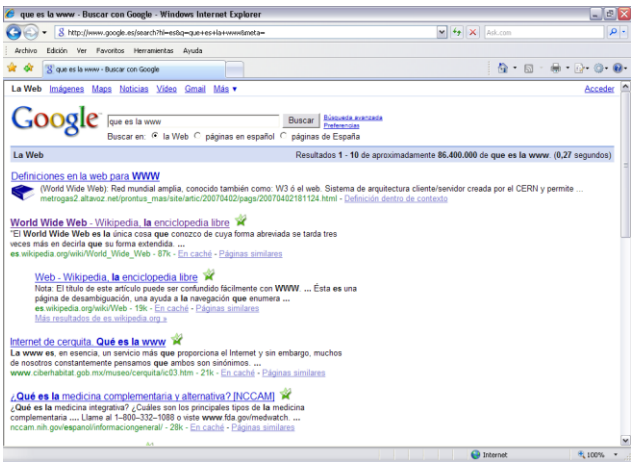
www.google.cat

www.yahoo.cat

Las páginas de inicio de estos dos buscadores son:

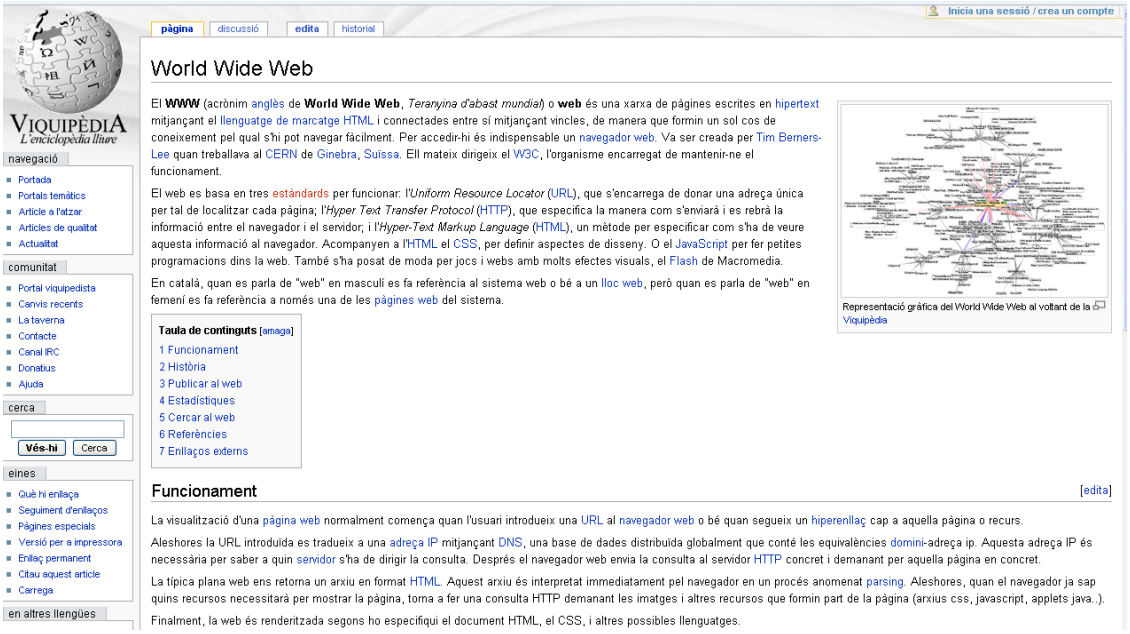
Este tipo de búsqueda es ideal para buscar información sobre un tema importante.

Una vez escritas las palabras en la casilla del tema buscado, se pulsa la tecla ENTER o bien el botón "Buscar con Google" o "Web". Como resultado, se nos mostrarán en el buscador muchas páginas que hablan de dicho tema. Cada página viene indicada con un título y una pequeña explicación, que nos puede dar una idea más concreta sobre lo que podemos encontrar en esa página sin necesidad de entrar a verla:




En caso de que realmente nos interese ver toda la información de la página escogida, no tenemos más que pulsar sobre su título, y automáticamente el navegador nos llevará a la página web deseada. El tiempo que tarde depende en gran medida de lo saturada que esté la máquina que sirve esa página a

Internet. Por ejemplo, si nos interesa el segundo enlace que hay en la lista anterior, le hacemos clic y veremos la nueva página con su dirección en la barra superior:




## *Por Palabras Clave:*

Si queremos obtener información sobre el edificio “Empire State Building” de Nueva York, lo mejor es buscarlo mediante palabras clave poniendo precisamente estas palabras, Empire State Building, en el buscador.

Podemos hacer lo que hace todo el mundo: buscamos nuestro nombre en Internet. Introducimos en la casilla de búsqueda nuestro nombre. Si especificamos demasiado, puede que no encontremos nada, pero si sólo ponemos un nombre y un apellido, seguramente encontremos páginas web donde aparezcan estos nombres, refiriéndose a personas que se llaman igual que nosotros, que pueden haber varias en el mundo.

Los buscadores nos permiten concretar el tipo de búsqueda a realizar. Cuando buscamos algo con varias palabras clave, el buscador puede pensar que queremos las páginas web en las que aparezca alguna de esas palabras, o todas ellas en la misma página, o todas ellas en el mismo orden en que las hemos escrito, y seguidas una detrás de otra. Todo esto se puede especificar antes de realizar la búsqueda.

Es muy normal que Aundo. Los buscadores más usados los puedes ver en www.alexa.com/site/ds/top_500. Aquí también podrás ver las webs con más tráfico en España o cualquier otro país.

### *Búsqueda avanzada*

Para obtener una búsqueda mucho más precisa podemos usar la opción de búsqueda avanzada de Google, que se encuentra en la parte derecha de la casilla de búsqueda:

Además de permitirte introducir los términos de tu búsqueda en el campo de búsqueda, Google ofrece un sinfín de opciones. Gracias a la Búsqueda avanzada, podrás buscar exclusivamente páginas que:

- contengan TODOS los términos de la búsqueda,
- contengan la frase exacta de la consulta,
- contengan al menos uno de los términos de la consulta,
- NO contengan ninguno de los términos de la consulta,
- estén redactadas en un idioma determinado,
- se hayan creado en un formato de archivo específico,
- se hayan actualizado en un periodo de tiempo determinado,
- pertenezcan a un dominio o sito web en particular,
- no contengan material para adultos.

### *Búsqueda de archivos*

Muchos sitios de Internet guardan un montón de archivos que no son accesibles a través de páginas web. La característica que tienen estos archivos es que se muestran en el navegador con la siguiente forma:

**Index of /mp3**

| Name | Last modified | Size | Description |
|---|---|---|---|
| Parent Directory | | - | |
| (V)Silvestre Dangond..> | 16-Jul-2008 21:50 | 5.4M | |
| (V)Silvestre Dangond..> | 16-Jul-2008 22:16 | 9.4M | |
| (V)Silvestre Dangond..> | 16-Jul-2008 21:50 | 4.8M | |
| (V)Silvestre Dangond..> | 16-Jul-2008 22:11 | 4.4M | |
| (V)Silvestre dangond..> | 16-Jul-2008 21:56 | 6.8M | |
| 01 - Las mujeres hay..> | 27-Oct-2006 10:14 | 4.5M | |
| 01. Me va tocar Olvi..> | 11-Sep-2008 16:18 | 227K | |
| 01 Caso cerrado- pil..> | 21-Apr-2007 21:24 | 10M | |
| 01 ENAMORADO DE TI.mp3 | 21-Oct-2006 19:33 | 6.3M | |
| 01El Mosaico del Agi..> | 11-Jan-2008 14:51 | 4.4M | |
| 01 Lo mas lindo que ..> | 18-Apr-2008 18:14 | 3.9M | |
| 01 POR QUE TE AMO ( ..> | 30-May-2007 21:10 | 4.2M | |
| 01Serealizaronmissue..> | 02-Feb-2007 05:47 | 1.8M | |
| 01 UN NOVIO COMO YO.mp3 | 31-May-2007 22:56 | 5.1M | |
| 01_No_Me_Atrevo.mp3 | 03-Feb-2007 18:03 | 5.8M | |
| 01 te pinto en mis s..> | 11-Jun-2007 00:31 | 496K | |
| 02 A mi me gusta la ..> | 18-Apr-2008 18:10 | 4.5M | |
| 02Larumbera-KillySan..> | 27-Jan-2007 18:39 | 3.4M | |
| 03 Me gusta, me gust..> | 13-Jun-2008 19:50 | 5.8M | |
| 04. Diferentes a Tod..> | 11-Sep-2008 16:18 | 0 | |
| 05. Si estubieras Aq..> | 11-Sep-2008 16:18 | 32K | |
| 05 Caja de Ilusiones..> | 18-Apr-2008 18:03 | 3.3M | |
| 05 Huele a Caribe.mp3 | 08-Nov-2007 05:34 | 3.9M | |
| 06 6. EL MISTERIOSO.mp3 | 11-May-2007 01:57 | 4.3M | |
| 08. No te Vallas - M..> | 11-Sep-2008 16:18 | 32K | |
| 08 El dueño del circ..> | 24-Sep-2007 18:12 | 3.4M | |
| 09 - Pendiente de Ti..> | 11-Nov-2007 16:05 | 5.8M | |

Fijaos que en la parte superior aparece la frase "*Index of*"; lo que hay que hacer es usar una búsqueda en Google que busque la frase "*Index of*" seguida de los tipos de archivos que queremos, separados por una pipa (|), que se obtiene pulsando las teclas ALT GR + 1; por ejemplo, una búsqueda de canciones en formato mp3 o vídeos mp4 y .avi de archivos sin página web sería: "*Index of*" mp3|mp4|avi:

O bien una búsqueda de documentos Word (.doc), PDFs (.pdf) o PowerPoints (.ppt), que traten de la enfermedad celíaca: "index of" pdf|doc|ppt "enfermedad celiaca":

En la mayoría de los casos suelen aparecer páginas web que en realidad son una lista de archivos.

Entonces basta con elegir un archivo y hacer clic en el archivo para reproducirlo. Si no se reproduce o no se descarga, entonces conviene ir a la *home* de la página web; para ello, borrar la dirección de la barra de direcciones del navegador para dejar solamente el nombre del dominio:

### *Buscador de buscadores*

Hay tantos buscadores, que se ha creado un buscador de buscadores en www.buscopio.net.

### *Buscadores semánticos*

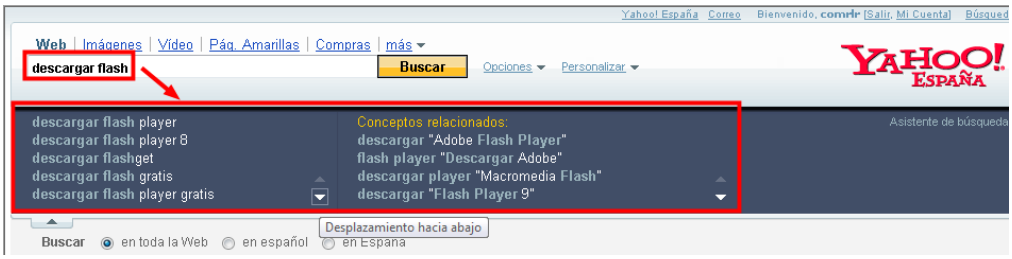
Hay también buscadores a los que les puedes hacer preguntas, llamados buscadores semánticos, que en vez de que tú escribas unas palabras clave, les haces preguntas como si fuesen un persona, como ww.ask.es, demos.bitext.com y muchos más. En inglès, www.hakia.com o www.lexxe.com.

### *Buscador de búsquedas*

¿Qué es lo que la gente busca por Internet y qué respuestas han obtenido? Míralo en www.answers.com.

### *Buscador con typing suggestions*

A medida que buscas, te sugiere palabras de búsqueda: www.cuil.com. Además tabula los resultados. Yahoo también tiene *typing suggestions* basadas en las búsquedas de otros usuarios. Esta característica es configurable y se puede desactivar. Por ejemplo, aquí tenemos las *typing suggestions* de Yahoo:




### *Bloqueo de webs peligrosas*

Los buscadores suelen bloquear las webs que se sabe que son peligrosas, pues pueden descargar virus, spyware, adware, troyanos y otros. Por ejemplo. esta es una página bloqueada en Yahoo:

## *Motores de búsqueda por palabras clave*

El servicio indexa previamente un cierto número de páginas de la Web de todo el planeta. Esta indexación se efectúa por texto completo y todas las palabras de todas las páginas referenciadas se convierten en entradas del índice, potencialmente objeto de la búsqueda.

La indexación previa de las páginas puede hacerse de dos formas distintas:

- **Declaración voluntaria del responsable de la sede Web** remota que indica al motor de búsqueda, rellenando un formulario, la existencia de un servicio. El motor indexa entonces todas la páginas del servicio así referenciado.
- **Robots que recorren de modo automático la Web y sus páginas de información**. Estos robots parten de un cierto número de páginas básicas y persiguen todos los enlaces de hipertexto en cada una de las páginas encontradas.

- Cada página encontrada se indexa seguidamente por texto completo. Cada indexación de página provoca una captura de las informaciones presentes en ellas y un almacenamiento del archivo obtenido en los discos duros del motor de búsqueda.
- Este servicio evita la declaración previa de un servicio, desde el momento en que un enlace de hipertexto apunta hacia él en cualquier parte del mundo. Sólo indexan información encontrada en las páginas HTML y no tiene en cuenta la información contenida en las bases de datos específicas de los organismos presentes en la Web, accesibles mediante formularios.

➔ **Generación del catálogo a mano:** no es realmente así; algunos de los buscadores trabajan sobre catálogos que son gestionados por una o varias personas. En este caso, el índice se genera a partir de entradas realizadas a mano, de modo que el que introduce una referencia puede indicar qué palabras se deben indexar.

## *Visión general de la tecnología de indexación*

Google es la única empresa dedicada a desarrollar el "motor de búsqueda perfecto", definido por su cofundador, Larry Page, como algo que "comprende exactamente lo que el usuario quiere decir y le entrega exactamente lo que está buscando". Con ese fin en mente, Google insiste en continuar innovando y se niega a aceptar las limitaciones de los modelos existentes. Por ello, desarrolló su propia infraestructura de servidores y la avanzada tecnología PageRank™, que cambió la manera de llevar a cabo las búsquedas.

Desde el principio, los programadores de Google reconocieron que, para proporcionar los resultados más rápidos y precisos, era necesaria una nueva configuración de servidores. A diferencia de la mayoría de los motores de búsqueda que emplean un grupo de servidores grandes que suelen ralentizarse cuando procesan picos de carga, Google utiliza equipos conectados para encontrar rápidamente la respuesta a cada consulta. Esa innovación permitió lograr tiempos de respuesta más veloces, mayor escalabilidad y menores costes. Es una idea que otros han copiado desde entonces, mientras que Google sigue puliendo su tecnología interna para hacerla cada vez más eficiente.

El *software* integrado en la tecnología de búsqueda de Google realiza una serie de cálculos simultáneos en tan sólo una fracción de segundo. Los motores de búsqueda tradicionales se basan, en gran parte, en la frecuencia con que una palabra aparece en una página web. Google, en cambio, emplea la tecnología PageRank™ para examinar toda la estructura de vínculos de la web y determinar qué páginas son las más importantes. A continuación, realiza un análisis de concordancia de hipertextos para establecer qué páginas son relevantes para la búsqueda específica que se esté procesando. Al combinar la importancia general con la relevancia específica respecto a una consulta en

particular, Google puede colocar los resultados más relevantes y fiables en primer lugar.

**Tecnología PageRank:** PageRank realiza una medición objetiva de la importancia que tienen las páginas web. Para ello, resuelve una ecuación que contiene más de 500 millones de variables y 2.000 millones de términos. En lugar de contar los vínculos directos, PageRank interpreta un vínculo de la Página A a la Página B como un voto que recibe la Página B de parte de la Página A. PageRank evalúa, de esa manera, la importancia que tiene una página determinada al contar la cantidad de votos que recibe.

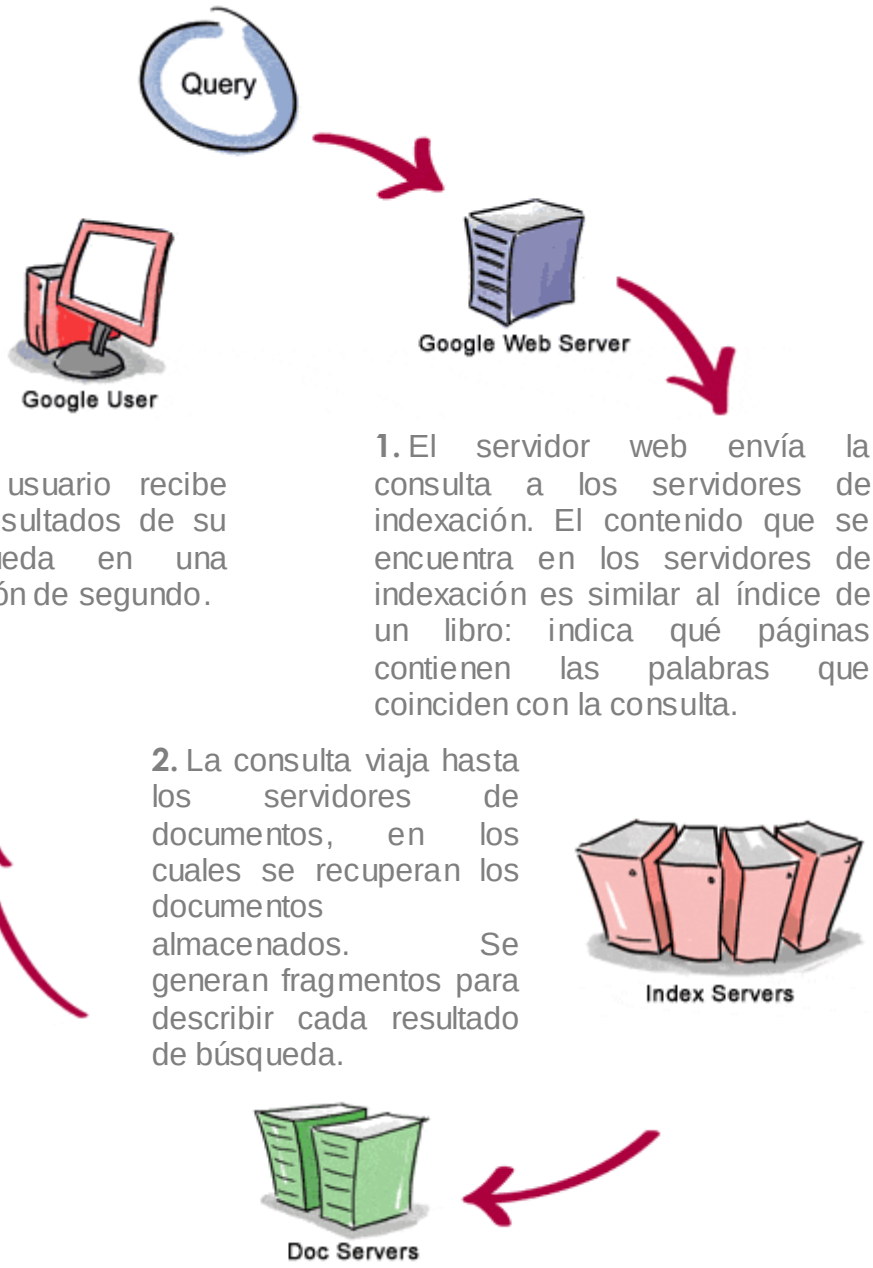
PageRank también considera la importancia de cada página que emite un voto, dado que a los votos procedentes de determinadas páginas se les otorga un valor mayor, incrementando así el valor de la página vinculada. Las páginas importantes reciben una valoración de PageRank más alta y aparecen en la parte superior de los resultados de búsqueda. La tecnología de Google emplea la inteligencia colectiva de la Web para determinar la importancia de una página. Los resultados se obtienen sin ningún tipo de participación humana; por este motivo, los usuarios han llegado a confiar en Google como fuente de información objetiva, libre de la manipulación que se genera cuando los sitios pagan por ocupar determinada posición en los resultados.

**Análisis de concordancia de hipertextos:** el motor de búsqueda de Google también analiza el contenido de cada página. Sin embargo, en lugar de explorar simplemente el texto de la página (que los editores de sitios pueden manipular mediante metaetiquetas), la tecnología de Google analiza todo el contenido de una página y tiene en cuenta también las fuentes, las subdivisiones y la ubicación precisa de cada palabra. Asimismo, Google analiza el contenido de páginas web vecinas para garantizar que los resultados encontrados son los más relevantes para la consulta del usuario.

Las innovaciones de Google no se limitan al escritorio. Para que los usuarios que acceden a la web a través de dispositivos portátiles puedan obtener resultados de búsqueda rápidos y precisos, Google desarrolló también la primera tecnología de búsqueda inalámbrica que traduce al momento el código HTML a formatos optimizados para WAP, i-mode, J-SKY y EZWeb. Actualmente, Google provee su tecnología inalámbrica a diferentes líderes del mercado, por ejemplo, a AT & T Wireless, Sprint PCS, Nextel, Palm, Handspring y Vodafone, entre otros.

## *La vida de una consulta de Google*

Una consulta de Google suele durar menos de medio segundo y, sin embargo, implica toda una serie de pasos que se deben completar antes de que la persona que está buscando información pueda ver los resultados.




## *Simple Search o búsqueda simple:*

- ➔ Search: permite efectuar búsquedas en páginas web o en Usenet.
- ➔ Display The Results: da lo opción de obtener los resultados de las búsquedas bajo distintos aspectos más o menos detallados.
- ➔ Formato de palabras clave:

- Si se escribe en minúscula, busca texto en mayúsculas y minúsculas en cualquier posición.
- Si escribe alguna letra en mayúsculas, busca la ocurrencia exacta.
- Los acentos siguen las mismas reglas que en los casos anteriores.
- Permite comodines con restricciones: detrás de la tercera letra, sustituye de cero a cinco letras, sólo minúsculas, las mayúsculas y cifras no se tienen en cuenta.
- Búsqueda de palabras adyacentes separadas por ";" o con comillas separadas por espacio.
- Signo + para AND y signo – para O.
- Pueden combinarse todas las posibilidades anteriores.

## *Búsquedas específicas:*

- Búsqueda por un enlace hipertexto, anchor:page.
- Búsqueda de applet Java, applet:javaclass10.
- Búsqueda en el nombre del servidor, host: nombre.org.
- Búsqueda de una imagen, image: logonodo50.gif.
- Búsqueda en el URL de un enlace, link:nodo50.org.
- Búsqueda únicamente en el texto, text:nodo50.
- Búsqueda por un título de página, title:homepage.
- Búsqueda en el interior de un URL, url:tintin.html.

## *Búsqueda de direcciones de correo*

Los principales servicios de búsqueda de direcciones de correo son:

- WHOWHERE: www.whowhere.com o spanish.whowhere.com
- FOUR11 DIRECTORY SERVICES: www.four11.com
- INTERNET ADRESS FINDER: www.iaf.net

## *Búsqueda de listas de distribución*

El principio de las listas de distribución se basa en la mensajería electrónica, pero la diferencia principal es que no se envía un mensaje a una persona sino a una lista de distribución, que representa un grupo de usuarios que se han abonado previamente.

Cuando alguien dirige un mensaje a una lista, la información se distribuye a todos los abonados de está, que pueden ser miles. Usted mismo recibirá en su buzón todos los mensajes enviados a una lista a la que se haya abonado.

Estas listas de mensajería están normalmente relacionadas con un tema concreto. Todo ámbito de reflexión es susceptible de ser objeto de una lista. Cualquier persona puede implementar una de estas listas sin ninguna restricción, contrariamente a los foros de discusión.

Una lista de distribución tiene dos direcciones de correo: una dirección administrativa, que sirve para abonos, bajas, etc., y una dirección para las propias discusiones. Para abonarse a una lista envíe un mensaje de este tipo:

"subscribe lista apellido nombre", donde lista es el nombre de la lista a la que desea abonarse.

## Búsqueda de servidores FTP

Los programas de FTP permiten acceder a distancia a ordenadores llamados servidores FTP, que ponen a disposición de la comunidad documentos o aplicaciones para su descarga. FTP permite conectarse a estos servidores mediante un *login* anónimo. Permite realizar transferencias desde el servidor remoto al disco local, y en algunas ocasiones en sentido inverso. Pero esto no es normal sobre todo por medidas de seguridad.

## Búsqueda de foros de discusión

Los *News* o Foros de discusión se encuentran a menudo en paralelo con las listas de distribución. Los dos procedimientos permiten el intercambio de información entre personas, en el interior de comunidades interesadas por un mismo tema de reflexión más o menos amplio.

Las listas de distribución se basan en el correo electrónico, mientras que los foros de discusión se basan en una red específica llamada Usenet.

En las listas de distribución los mensajes se almacenan en las máquinas de los abonados y son éstos los que deciden cuándo borrarlos. En los foros los mensajes se archivan en los servidores y se borrarán cuando el servidor lo crea necesario. Hay que tener en cuenta que varias veces al día todos los servidores se intercambian los mensajes para estar al día. Cada propietario de servidores de *News* decide cuáles son los temas y, por tanto, los grupos que difundirá. Hay aproximadamente unos 22.000 en todo el mundo.

Para obtener informaciones sobre el servicio *News*, por ejemplo: http://groups.google.com/ o http://groups.yahoo.com/

## Metodología de búsqueda

No toméis nunca una información encontrada en Internet al pie de la letra. La información que se encuentra es mayoritariamente gratuita y por ello tiene sus ventajas e inconvenientes. No existe un estándar en la definición de un documento en Internet: autor, fecha de creación, tamaño, fecha de actualización, et., todos estos datos son difíciles de encontrar si el creador de la información no los proporciona.

## Traducción de la consulta al lenguaje documental

Para explicar este punto os pedimos que os descarguéis el siguiente pdf (lenguaje_documental.pdf)

## Ejecución de la estrategia de busca

Valoración de los resultados

Sabemos que si un buscador es más eficaz que otro tiene que tener unos resultados de búsqueda tan depurados que el visitante logre encontrar aquello que busca en el menor tiempo posible; esto quiere decir en las primeras páginas de resultados.

Leyendo un interesante artículo de Bill Slawski en su Blog (http://www.seobythesea.com/ ), podemos aprender algo más acerca de los planteamientos futuros de buscadores en relación con una patente que ha adquirido Google y que combina novedosas técnicas para encontrar contenido duplicado y filtrarlo convenientemente.

No es la primera vez que abordamos este tema y las incógnitas que nos suscita. Saber discernir el contenido duplicado en la Red pasa por encontrar su autoría, conocer lo que el autor propone hacer con sus derechos, otorgarles la importancia por encima de copias piratas y no permitidas, etc. En el caso de informaciones comerciales, fichas de productos, catálogos, etc., hasta el momento parece que se valoraba la fecha de expansión en buscadores y su autoría. Sí sabemos que no sirve de nada posicionar un producto los primeros si luego no reunimos otros privilegios para el buscador que nos ayude a mantenernos en un posicionamiento óptimo. Vamos a quedar relegados en los resultados de búsqueda rápidamente por debajo de nuestros competidores.

## *Tratamiento de documentos cercanos a la duplicidad*

El buscador deberá discernir y tratar la información cuidadosamente en función del análisis que haga de la misma, teniendo en cuenta factores como la originalidad, tipo de copia y los derechos de autor.

- Nos encontramos con el mismo texto publicado en diferentes soportes (HTML, pdf, doc, xls, texto plano…) o en soportes para impresión o enfocados a teléfonos móviles.
- Compartir noticias y artículos en fuentes RSS publicadas en Blogs y otro tipo de webs.
- Utilización de páginas “espejo” con fines transparentes, como intentar evitar demoras de carga de webs o facilitar el uso en condiciones adversas.
- Detección de páginas que han violado los derechos de autor.
- Publicación del mismo contenido repetidamente en el mismo sitio web.

## *Los recientes esfuerzos de Google para combatir el contenido duplicado*

El año pasado, algunos empleados de Google hicieron un buen trabajo relacionado con el tema de conjuntos de procesos para la detección de duplicados y formas de valorarlos. Detección de Proximidad para la web duplicados Rastreo (pdf).

En el mencionado documento, uno de los procesos descritos en detalle fue desarrollado por Moses Charikar, un profesor de Princeton que trabajó para

Google en el pasado y que fue el inventor de una patente comprada por Google el pasado año en la que se hace referencia a una serie de métodos en relación con el tema de la duplicidad de contenidos.

- Encontrar archivos similares en una gran Red.
- Huellas digitales en documentos.
- Copiar mecanismos de detección de documentos digitales.
- Agrupación sintáctica en la Web.
- Similitud de técnicas de estimación de redondeos en algoritmos.
- Similitud con el sistema de búsqueda de estructuras de datos compactos.
- Métodos par la identificación de documentos versionados y plagiados.

La conclusión de estos estudios ha sido que ninguno de los algoritmos está funcionando perfectamente para ayudar a encontrar duplicidad de pares en el mismo sitio web, pero sí logra una alta precisión y acerca lo que podría incorporarse a las nuevas técnicas de Google en su continua mejora en su buscador.

Como conclusión, Bill Slawski apunta que el proceso descrito en esta nueva patente no introduce un nuevo método de identificación de contenido duplicado, pero sí aporta un nuevo enfoque en cuanto a sus métodos de detección. ¿Quién sabe si estas técnicas se utilizarán definitivamente por los ingenieros de Google para aportar más transparencia al buscador?