ÍNDICE

PRACTICA Nº4 - BUSCADORES

Buscadores

El primer buscador fue "Wandex", un índice (ahora desaparecido) realizado por la World Wide Web Wanderer. El primer motor de búsqueda de texto completo fue WebCrawler, que apareció en 1994. A diferencia de sus predecesores, éste permitía a sus usuarios una búsqueda por palabras en cualquier página web, lo que llegó a ser un estándar para la gran mayoría de los buscadores. WebCrawler fue también el primero en darse a conocer ampliamente entre el público. También apareció en 1994 Lycos .

Muy pronto aparecieron muchos más buscadores, como Excite, Infoseek, Inktomi, Northern Light y Altavista. De algún modo, competían con directorios populares tales como Yahoo!. Más tarde, los directorios se integraron o se añadieron a la tecnología de los buscadores para aumentar su funcionalidad.

En la actualidad el sector de los motores de búsquedas en Internet está dominado por Google, Yahoo! y Windows Live Search. Con ellos se puede encontrar cualquier contenido que esté disponible públicamente en la Red, a pesar de que podemos encontrar muchos otros, centraré mi estudio en Google como buscador mas utilizado y como muestra de utilización de un buscador, en segundo lugar haré una breve descripción de los otros dos mas populares, no obstante cualquiera de la siguiente lista puede resultar útil, sin olvidar los buscadores especializados:

- Alltheweb
- AltaVista
- Amfibi
- Ask
- Bing
- CerCAT
- Google
- Lycos
- Yahoo!
- Windows Live Search

Buscador Google

Google nació en 1998. Sergey Brin y Larry Page, de la Universidad de Stanford (Estados Unidos) estaban trabajando en un proyecto de clase para identificar patrones en la estructura de enlaces de la Red. Fue este estudio lo que les dio pie a diseñar un motor de búsqueda basado la estructura de los enlaces. Originalmente, el motor de búsqueda creado fue llamado Googol, haciendo referencia al número 10 elevado a 100, lo cual representa el número infinito de documentos de búsqueda en la Red. Tras presentar el proyecto a un inversor, los fundadores recibieron un cheque a favor de «Google», por lo que decidieron cambiarle el nombre. En la actualidad, Google es uno de los mejores motores de búsqueda que hay. Existen versiones en alemán, chino (simplificado), chino (tradicional), coreano, danés, español, finlandés, francés, holandés, inglés, italiano, japonés, noruego, portugués y sueco. Además, el motor reconoce e indexa documentos en formato PDF, RTF, PostScript, Word, Excel y PowerPoint, entre otros.

Clasificación de la información

Google clasifica la información mediante una red de ordenadores repartidos por todo el mundo y unos programas denominados arañas que van buscando información y capturando diferentes pantallas, estas arañas guardan el número de enlaces que apuntan a una página para definir su importancia, en función de la cantidad de enlaces considerará la página más o menos importante.

“PageRank” es el número que asigna Google a una página considerando un alto número de variables, no únicamente el apuntado anteriormente, el algoritmo que utiliza Google es secreto y va cambiando a fin de evitar que los programadores burlen estos criterios para situar sus páginas.
Como ejemplo tenemos el boicot a la SGAE en que al buscar la palabra ladrones salía en primer lugar.

En la actualidad vemos que aparece la noticia de que ladrones no lleva a SGAE ya que Google ha actuado contra el Google Bombing.

http://www.google.com/intl/es/corporate/tech.html

Visión general de la tecnología

Google es la única empresa abocada 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 de 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.

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.

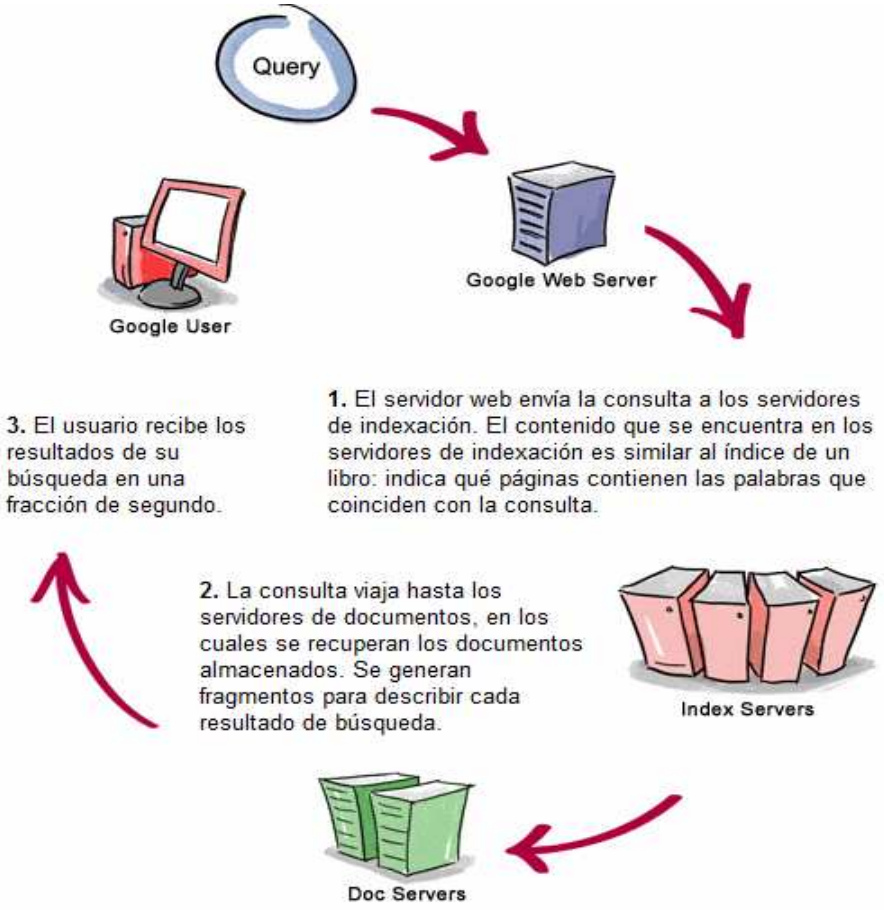


Ilustración 1 Proceso de una consulta en Google

Análisis de concordancia de hipertextos

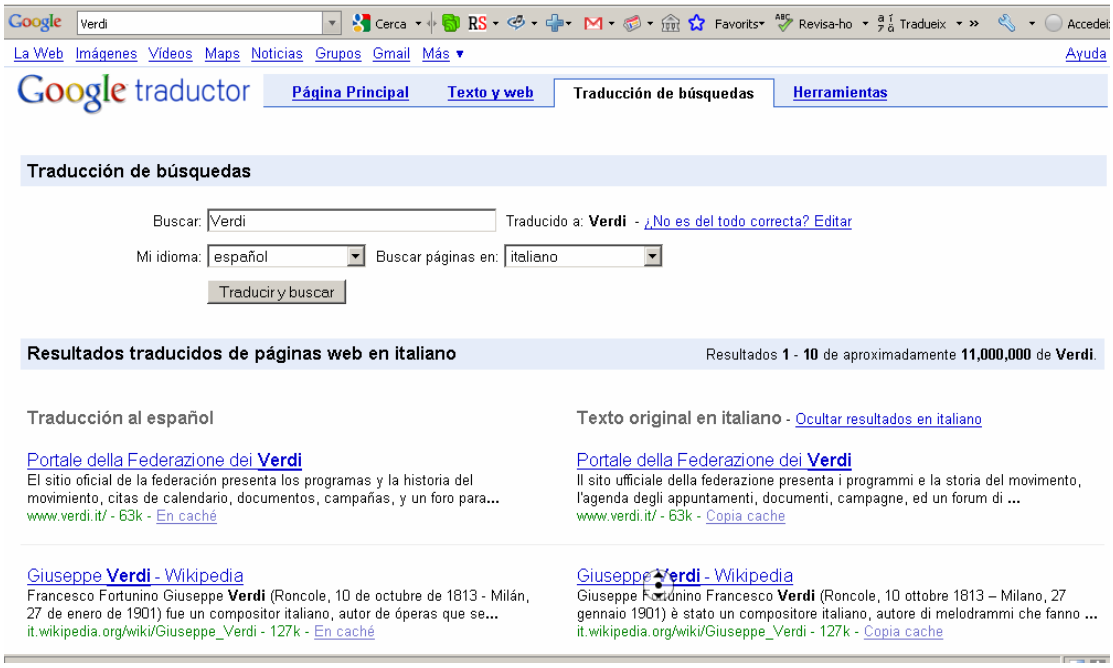
El motor de búsqueda de Google también analiza el contenido de cada página. Sin embargo, en lugar de explorar simplemente el texto de la página (que los editores de sitios pueden manipular mediante metaetiquetas), la tecnología de Google analiza todo el contenido de una página y toma en cuenta también 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.

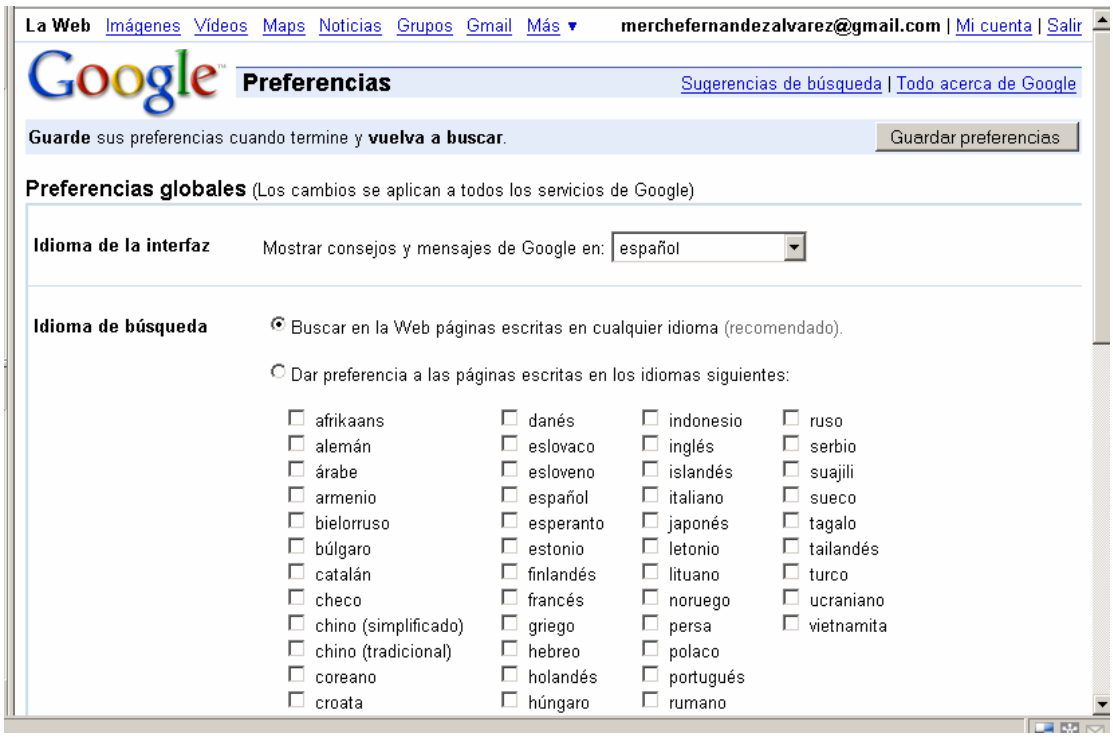
Herramientas del idioma

Permite buscar una información en otro idioma y traducir la página al español.

Una utilidad seria buscar información por ejemplo sobre Verdi en italiano



Preferencias de Google



Se debe guardar para que filtre adecuadamente las búsquedas.

Si buscamos la palabra educación nos devuelve 130.000.000 resultados en 0,12 segundos, de hecho no podemos acceder a todos los resultados.
Nuestro objetivo al realizar una búsqueda es ir limitando cada vez el número de resultados sin dejar fuera información valiosa. Nos fijaremos sobre todo en las direcciones proporcionadas en el caso de páginas importantes ofrece enlaces secundarios en el caso de la página del departamento de Educación de Cataluña el resultado ofrece enlaces "secundarios"

Departament d'*Educació*

- [Traducir esta página]

De la Generalitat de Catalunya. Informació institucional sobre normativa, estudis, centres d'ensenyament i tràmits del professorat.

*www20.gencat.cat/portal/site/**Educacio*** - En caché - Similares -

Borsa d'interins	Ajuts, beques i subvencions
Oposicions i concursos	Professorat
Centres i serveis educatius	Borsa de treball PAS

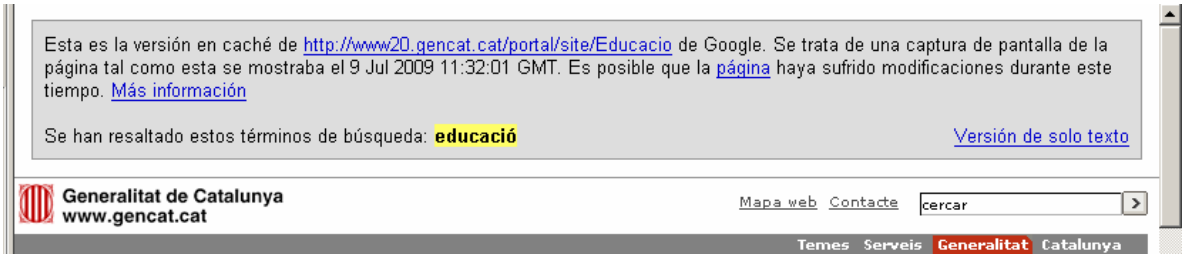
Más resultados de gencat.cat »

Al final de la página de nos permite acceder a otros conceptos, en este caso:

Búsquedas relacionadas con: **educació**

educació **social** educació **especial**

Si pulsamos en caché la página aparece con un encabezado informativo con una fecha determinada.



Una utilidad seria poder visitar páginas que cuando intentamos entrar no tenemos acceso, si accedemos a la página en caché mostrará el resultado del momento en que fue indexada la dirección a que accedemos.

Búsquedas dentro de las búsquedas

Podemos añadir “escoles” a la palabra “educació” por tanto si buscamos “educació escoles” el primer enlace es:

FP Educació Infantil'09

www.institutoaccess.com Proves Lliures+Sistema Presencial De L'Institut Access.+Informació a:

Se trata de un enlace patrocinado, es decir una empresa que ha pagado por el hecho de poder situarse en primer término. Google cobra por click..

Si buscamos telefónica nos encontramos en primer lugar los enlaces patrocinados de telefónica y de hecho segundo lugar de los enlaces no patrocinados la página oficial de telefónica.

Diversos estudios han llegado a la conclusión de que los enlaces más visitados son siempre los tres primeros por tanto hay un verdadero interés comercial en situarse en estos puestos.

Ver los archivos en HTML si en una búsqueda nos encontramos con un archivo de tipo PDF podemos ver su versión HTML para comprobar si me interesa sin necesidad de abrir el lector pdf que únicamente abriremos si nos interesa.

Búsqueda avanzada.

Además de permitir 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 período de tiempo determinado,
- pertenezcan a un dominio o sito web en particular,
- no contengan material para adultos.

Entre los operadores de Búsqueda avanzada figuran los siguientes:

- búsqueda de inclusión,
- búsqueda mediante el operador OR,
- búsqueda en dominios
- búsqueda de intervalos numéricos,
- otras funciones de la Búsqueda avanzada.

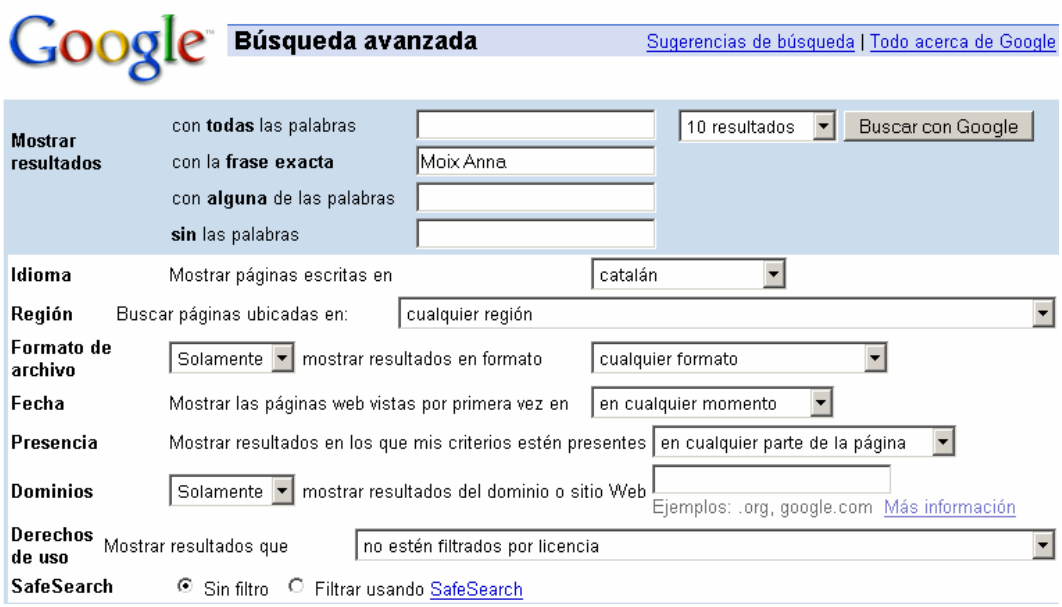
Google ignora palabras y caracteres comunes, como *dónde*, *el/la/los/las*, *cómo*, así como algunos dígitos y letras independientes, porque tienden a ralentizar la búsqueda sin mejorar los resultados. Google indicará si se ha excluido alguna palabra en la información detallada que aparecerá en la página de resultados, debajo del cuadro de búsqueda.

Si para obtener los resultados que deseas es imprescindible incluir un término común, puedes precederlo del signo "+". Asegúrate de incluir un espacio antes de dicho signo.

Por ejemplo, para asegurarte de que Google incluye "I" en una búsqueda de La Guerra de las Galaxias, Episodio I, especifica la consulta de la siguiente manera:

Por ejemplo información sobre Anna Moix escritora y utilizamos la búsqueda con todas las palabras no encontraremos información sobre ella y tendremos una serie de resultados sin valor.

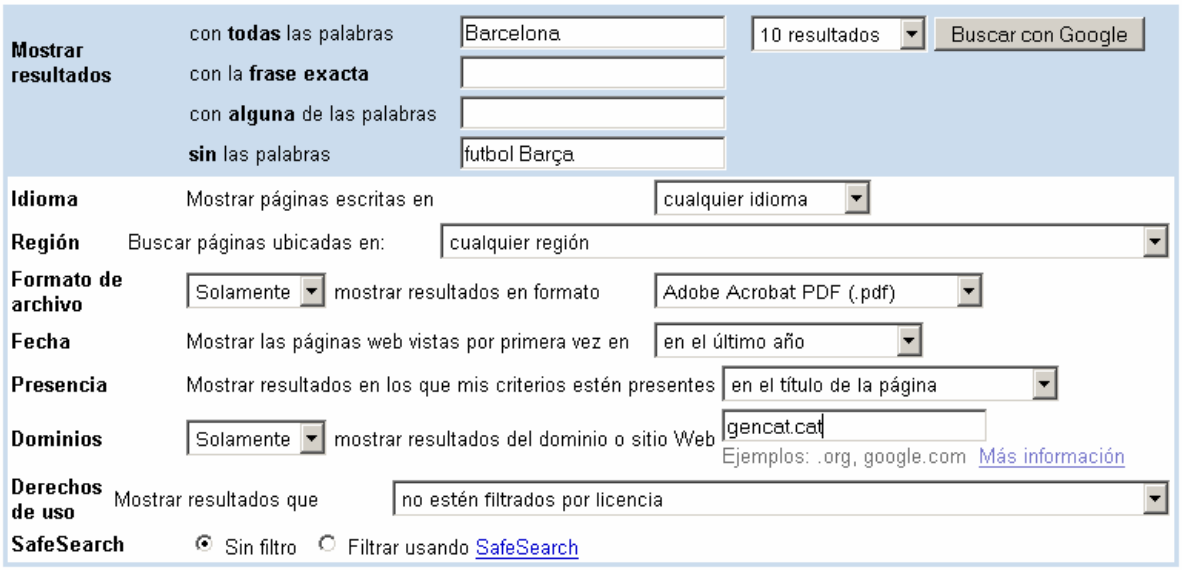
Si situamos la búsqueda en con la frase exacta.



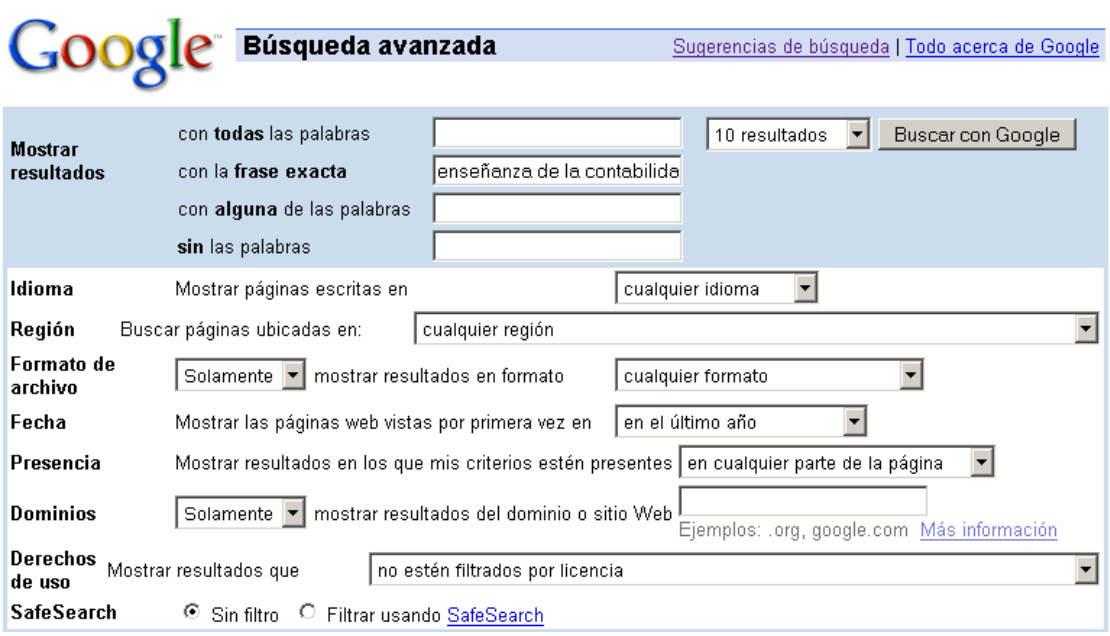
Encontramos la información buscada.

Si buscamos información sobre la ciudad de Barcelona encontraremos una mejor información si eliminamos la palabra “futbol” y Barça ya que muchas de las informaciones son referentes al club deportivo.

Si continuamos filtrando el resultado podemos añadir por ejemplo páginas en formato pdf del último año que se encuentren en el dominio gencat.cat



Si buscamos enseñanza de la contabilidad con los siguientes parámetros



Operadores en Google

Si buscamos “departament de educacio” obtenemos por defecto todos los resultados que obtenemos con el operador AND es decir páginas que tengan relación con las palabras “departament” y además las páginas relacionadas con “educació”

Carácter “

Si utilizamos el carácter " " la información es la referente a la suma de las dos palabras que buscamos "departament de educación" con lo que nuestra busqueda será.

Operadores + y –

Con el operador + realizamos una búsqueda parecida a la primera descrita sin operadores AND

Con el – restringimos la búsqueda

Si buscamos Barcelona obtendremos muchos resultados relacionados con el Barça si restringimos Barcelona-futbol

Los dos puntos : indican un rango de búsqueda

Buscando información Barcelona 1900:1910 encontraría información entre 1900 y 1910

El asterisco * sustituye cualquier carácter

Si buscamos Terenci Moix*Ana buscaria información relacionada con Terenci Moix y su hermana Ana aunque haya alguna palabra entre los dos nombres.

Búsqueda en una web Site:

Si búscamos las fiestas de Barcelona en la página de Barcelona, la sintaxis seria Fiestas site:BCN.cat

Búsqueda de un tipo de fichero concreto: Filetype

Buscaría las extensiones que le pedimos

Presupuesto de tesorería filetype:xls busca hojas de cálculo de presupuesto de tesoreria.

http://video.google.es/videoplay?docid=-2084919753106562775

video Google

http://www.youtube.com/watch?gl=ES&hl=es&v=CtBWv1sRcAI

futuro internet

Yahoo!

Yahoo! siempre ha sido considerado como un directorio, y como tal también se incluye en esta clasificación, en diciembre de 2002 compró el potente motor de búsqueda japonés Inktomi. El objetivo era apartar de sus páginas a Google, que hasta entonces era su motor de búsqueda, y entrar en la lucha por posicionarse entre los mejores puestos en el mercado de la búsqueda de información por Internet. Además, en 2003 adquirió el servicio de resultados de pago Overture, que tenía en ese momento acuerdos para servir sus listados a MSN Search y AOL Europa, entre otros.
Yahoo! Search permite realizar búsquedas en hasta treinta idiomas y tiene vanos buscadores especializados en diversos temas, entre ellos imágenes, vídeos, noticias y compras. Ofrece, además, al usuario la posibilidad de acceder al contenido caché de las páginas, muy útil en el caso de que el enlace esté inaccesible. Otra de sus grandes ventajas es que mantiene todos los servicios que daba antes como portal de Internet, que incluyen, entre otros, búsqueda de viajes, ofertas de empleo o anuncios personales. Además, sigue ofreciendo su servicio de personalización My Yahoo!. Entre algunos de los servicios que ofrece el buscador de Yahoo! se encuentran las siguientes:

Motor de búsqueda local «http://local.vahoo.com/results»

Ofrece a los usuarios la posibilidad de buscar números telefónicos, mapas, calificaciones y reseñas de una serie de servicios.

Yahoo! Image Search «http://images.search.yahoo.com/»

El buscador de imágenes de Yahoo! ofrece en sus listas de resultados una pequeña instantánea de cada una de las imágenes encontradas. Permite personalizar la búsqueda para obtener únicamente imágenes en blanco y negro.

Yahoo! Desktop Search «http://desktop.yahoo.com/»

Yahoo! Desktop Search localiza dentro de los ordenadores personales documentos en más de doscientos formatos: mensajes de correo de Outlook y Outlook Express, documentos en Word y Excel, presentaciones en PowerPoint, archivos en PDF y HTML, imágenes, archivos de vídeo y audio, e incluso

programas ejecutables o documentos y archivos comprimidos en formatos estándar como zip.

Yahoo! Video Search «http://video.search.yahoo.com»
Este buscador localiza archivos de vídeo en formato Avi, MPEG, MOV, RM o WMV.

Windows Live Search

Microsoft no se ha querido quedar fuera de esta batalla por conventirse en líder en la búsqueda de información por Internet y a principios de 2005 lanzó al mercado su propio motor de búsqueda, reemplazando así a la tecnología de Yahoo! que utilizaba hasta entonces en su sitio de MSN. Según la compañía, el MSN Search selecciona sus resultados a partir de una base de datos de más de 5000 millones de documentos y páginas web.
Una de las grandes ventajas del MSN Search es que ofrece un servicio de respuestas directas procedentes de los más de cuarenta mil artículos de su enciclopedia virtual Encarta, que se unen a los resultados generados por su motor de búsqueda.
El motor de Microsoft tiene diferentes versiones para los siguientes países: Alemania, Austria, Australia, Bélgica, Canadá (en inglés y francés), Dinamarca, España, Francia, Finlandia, India, Italia, Japón, Malasia, Holanda, Nueva Zelanda,

Metabuscadores

Los metabuscadores permiten realizar una búsqueda en varios buscadores a la vez. Uno de sus inconvenientes, además de un mayor tiempo de espera, es que no suele ser posible precisar la búsqueda, ya que cada uno de los motores que engloba tiene sus propias peculiaridades de búsqueda. Su método de funcionamiento es el siguiente: cuando el usuario realiza una búsqueda, el metabuscador la dirige a sus motores asociados, componiendo una lista de aciertos que representan, teóricamente, las mejores respuestas a la pregunta. Posteriormente, algunos ofrecen a posibilidad de ordenar por relevancia, entendiendo por relevancia el grado con que la web resultante concuerda con la búsqueda realizada por el usuario o con la información que necesita.
El cálculo de la relevancia que realiza el motor de búsqueda (normalmente es expresada en un porcentaje al lado de cada enlace proporcionado por el motor)

implica varios factores, entre los que se encuentran formatear los resultados de forma consistente, verificar la accesibilidad o eliminar enlaces muertos.

Tipos de metabuscadores:

Metabuscadores que no agrupan los resultados. Se debe revisar el listado resultante de cada uno de los motores en los que se ha realizado la búsqueda. En este caso hay muchas posibilidades de que existan webs duplicadas. Un ejemplo de este caso es Dogpile «http://www.dogplle.com».

Metabuscadores que agrupan los resultados. Son los más numerosos. Poseen la vetaja de que eliminan los duplicados. Un ejemplo es MetaCrawler «http:// www.metacrawler.com».

Algunos de los principales metabuscadores

COPERNIC

«HTTP://WWW.COPERNIC.COM»

Copernic, producto de la empresa canadiense Copernic Technologies, es un buscador múltiple que transfiere una ecuación de búsqueda a un conjunto de buscadores de manera simultánea, recupera las referencias pertinentes y las ordena según el grado medio de relevancia obtenido de cada uno de los buscadores.
La edición gratuita ofrece los servicios básicos de consulta y, con la intención de persuadir a los posibles compradores de las ediciones comerciales, da una idea bastante aproximada de las posibilidades del programa completo. Los tipos de Copernic comercializados son los siguientes:

Copernic Basic, la edición gratuita, permite consultar simultáneamente cerca de ochenta buscadores importantes, agrupados en siete categorías: el web (más una categoría opcional relacionada con un idioma o un país), grupos de noticias, direcciones de correo electrónico, compra de libros, compra de hardware y compra de software. A pesar de ser la edición reducida de un programa comercial, ofrece tantas o más posibilidades que las versiones completas de otros buscadores múltiples, como Lexibot, NetAttaché Pro o Internet EZ Search.
Copernic Plus, la edición comercial más económica, permite acceder a más de mil fuentes de información agrupadas en 90 categorías de búsqueda especializada, como enciclopedias, salud, multimedia, ciencias, negocios y fi-

nanzas, descarga de software, cine, artículos sobre las tecnologías, música, etcétera.
Copernic Pro, la edición comercial más completa, ofrece, además, otras prestaciones interesantes: actualización de búsquedas programadas, servicio de alerta y verificador ortográfico de las búsquedas.

DOGPILE

«HTTP://WWW.DOGPILE.COM»
Desarrollado por Aarón Flin, permite realizar las búsquedas en 23 buscadores a la vez con bastante velocidad. Además, ofrece la posibilidad de buscar específicamente en grupos de noticias, servidores FTP, etcétera. Dogpile admite operadores booleanos a la hora de refinar la búsqueda.

Vivísimo

«HTTP://VIVISIMO.COM»
Vivísimo no sólo realiza búsquedas múltiples entre muchos buscadores, sino que, además, organiza esos resultados automáticamente en categorías. Por ejemplo, si se busca 'Miguel Delibes' este buscador devuelve entre otros grupos 'libros', 'biografías', 'fotos', 'enciclopedias' y 'vídeos', según el tipo de datos que contengan las páginas web encontradas. Así, navegando por estos grupos, el usuario puede acceder a la información que desea de forma más precisa.
METACRAWLER «HTTP://METACRAWLER.COM»
Es uno de los metabuscadores más completos. Permite buscar en los siguientes buscadores: Opentext, Lycos, WebCrawler, Infoseek, Excite, Inktomi, Galaxy y Yahoo!.

MOTORES DE BÚSQUEDA ESPECIALIZADOS

http://cookieface.com.ar/2008/07/21/compilado-de-motores-de-busqueda-especializados/
Cleepr: motor de búsquedas de vídeos musicales.
BuscaTube: un metabuscador que agrupa videos de diferentes sitios de almacenamiento.
Copyscape: buscador de copy & paste de tu web.
Tweetscan: buscar en Twitter.
Buskka: buscador de Rapidshare, Megaupload y similares.

Twitter Search: buscar palabras y gente en Twitter.

Recetas: buscador de recetas de comidas.

Searchcube: buscador con resultados en forma de cubo.

Butacon: buscador de películas (descarga directa y online).

Njouba: buscador de archivos mp3, videos, imágenes, en directorios FTP, libros, en Rapidshare y Torrents.

Picitup: buscador visual de imágenes.

Wallpaper Search: Buscador de Wallpapers.

SingsBOX: buscador de música.

Pixolu: buscador de imágenes.

Iconfinder: buscador de íconos.

Scribd: buscador de libros digitales.

find:Design: buscador especializado en diseño.

Estanlibres: buscador de dominios libres

PleaseDressMe: buscador de remeras.

Sharech: buscador de enlaces de descarga directa (tipo Rapidshare).

Totme: buscador de enlaces de descarga directa (tipo Rapidshare).

PicFindr: buscador de fotos copyleft.

Plorf: buscador de mp3s.

Pdf Geni: buscador de libros en pdf.

Seeqpod: buscador de música.

Speckly: buscador de torrents.

Vozavi: motor de búsqueda de opioniones.

CiteSeerX: buscador de informatica y literatura científica.

Wuzam: buscador de mp3.

Wikia: buscador de Wikipedia.

YourSerials: buscador de numeros seriales:

Powerset: buscador de Wikipedia.

Wink: buscador de personas.

123People: buscador de personas.

Pipl: buscador de personas.

Peek You: buscador de personas.

LookTorrent: buscador de torrents.

AirMp3: buscador de Mp3.

lyricsfly: buscador de letras de canciones.

Carrot: buscador temático.

Domize: buscador de dominios disponibles.

Facesaerch: buscador de rostros.

Iconlook: buscador de íconos.

Mixturtle: buscador de música.

Natyo: buscador de clasificados (Mercado Libre, etc).

Kartoo: buscador de contenidos.

exalead: buscador de contenidos.

Oskope: buscador con asistente visual.

Samfind: motor de búsqueda personalizado.

2lingual: buscador + traductor.

Youtorrent: buscador de torrents legales.

IFACnet.com: buscador para contadores.

Pdf search: buscador de PDF.

Expedia: buscador para viajeros.

Blodico Plus: buscador de blogs en español.