Xarxa Punt TIC

MÒDUL 1 NIVELL BÀSIC
Recerca i recuperació d'informació a Internet
Unitat didàctica 3: El procés de recerca de la informació

plan avanza2 | Gobierno de España | Plan E | Ministerio de Industria, Turismo y Comercio | Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información | Generalitat de Catalunya

➔ ÌNDEX

Unitat didàctica 3

➔ El procés de recerca de la informació

Tipologia de cerques

Una part de les hores que passem davant l'ordinador buscant informació ens la podem estalviar aprenent com obtenir-la, i una altra gran part, examinant i porgant dades completament irrellevants per a nosaltres.

Les eines més utilitzades a Internet es basen en l'ús de patrons de cerca. Gairebé ningú accedeix a bases de dades utilitzant complicades sessions Telnet, normalment es fa mitjançant sistemes de consulta més potents i còmodes. L'experiència i habilitat per usar-los correctament és molt mes important si desitgem trobar dades que ens siguin útils.

La crisi de la sobreinformació és bastant recent i es deu a la incapacitat d'una sola persona de retenir ni tan sols una mínima part de la informació que li és útil. El problema s'agreuja quan la informació útil és una part ínfima de tota la que rep i inapreciable comparada amb tota la que té a la seva disposició.

Amb tanta informació disponible, el que hem de saber és quina ens interessa, on trobar-la, com obtenir-la i la millor manera d'administrar-la. Amb l'ajuda dels navegadors de WWW es faciliten les tasques de localització i obtenció, però saber què és el que hi ha, el que ens interessa i com utilitzar-ho són qüestions més complexes.

Hi ha diversos mètodes d'afrontar el problema; lluny de "divagar" per la Xarxa, que consisteix a anar d'enllaç en enllaç sense rumb fix, es pot intentar navegar utilitzant guies, llistes de recursos classificats, receptaris, etc. amb resultats molt diversos. Això és probablement el que estem fent des de fa temps.

Una vegada localitzades les pàgines que contenen referències a aquells ítems de major interès, és habitual que l'usuari en faci una llista. De fet, amb la majoria de navegadors de Web, un pot cremar-se una llista d'enllaços o marcadors accessible directament des del visor i fins i tot generar pàgines HTML a partir d'ela.

Per facilitar la tasca, cada dia hi ha més llocs dedicats a la creació i manteniment de llistes i directoris de recursos organitzats segons criteris geogràfics o temàtics denominats generalment catàlegs. Poden ser els punts de partida ideals quan busquem informació sobre un tema i volem saber què hi ha a la Xarxa o quan busquem algun servei seguint criteris geogràfics. A partir de les entrades del catàleg podem anar descendint fins a trobar el que realment ens interessa, simplement seguint l'estructura lògica que defineix el servei.

De totes maneres, la utilització dels catàlegs deixa de ser adequada quan el que busquem és més específic, ja que trobar-ho simplement navegant pot resultar bastant difícil. Ens trobem llavors amb la necessitat d'emprar eines per a la cerca i selecció de la informació disponible a la Xarxa. Estem parlant del que hem denominat cercadors d'informació.

Distingirem dos mètodes de recerca en funció de la manera d'interactuar amb les eines:

- **Mètodes actius**, que són controlats directament per nosaltres, com els sistemes de cerca per patrons relacionats amb el tema que busquem o l'accés a bases de dades i, si no resulta, la interacció amb altres persones, preguntant

directament a algú que creguem capacitat per a respondre'ns, ja sigui mitjançant el correu electrònic o utilitzant grups de notícies o llistes de distribució que tinguin relació amb allò que busquem.

- **Mètodes passius**, que són sistemes automatitzats que localitzessin la informació que vulguem a partir d'una descripció del nostre objectiu. Es tracta d'emprar agents informàtics que puguin automatitzar moltes de les tasques anteriors. És a dir, marques els teus interessos i el sistema s'encarrega de mantenir-te informat únicament de les novetats sobre aquests temes que apareguin en certes llistes, grups o pàgines. També pots optar per sistemes més complets que poden realitzar tot això i actuar sobre la informació continguda en la seva màquina, funcionant com veritables gestors d'informació.

Cerques en el Web

Els sistemes de cerca més utilitzats actualment són els basats en pàgines d'hipertext (servei WWW), en les quals s'introdueixen patrons o paraules clau a buscar. Aquestes pàgines actuen com a intermediàries entre l'usuari i una base de dades emmagatzemada en el servidor o accessible per a aquest. Alguns sistemes restringeixen la cerca a l'espai WEB, però la majoria permeten buscar qualsevol tipus de recursos accessible mitjançant un URL.

Els sistemes de cerca han d'indexar en una base de dades pròpia part de tota la informació per a no haver de recórrer tota la Xarxa cada vegada que es consulta. Els algorismes utilitzats en els programes de cerca es basen a estructurar la informació de manera que optimitzi les cerques. El resultat és que els sistemes són pràcticament instantanis, encara que continguin milions d'entrades.

En poc temps han aparegut un gran nombre de sistemes de cerca basats en el WEB, amb una presentació i unes possibilitats cada dia més atraients. Els usuaris coneixen la seva existència per comentaris a la Xarxa, en revistes o a partir d'enllaços que apareixen en pàgines molt visitades. També ha contribuït molt el fet que molts navegadors permeten accedir directament a alguns d'ells des d'un submenú

Per incloure referències a pàgines personals, l'única cosa que has de fer és seguir les instruccions que se solen indicar en algun lloc de les pàgines de cerca (generalment mitjançant formularis). Així pot anar donant-se d'alta en aquells sistemes que li semblin més interessants, el problema és que enviar un resum de les seves pàgines a tots els cercadors es pot fer pesat. En els cercadors que ofereixen el servei gratuïtament, la causa del manteniment econòmic és la popularitat, que es tradueix a efectes comercials en "audiència", la qual cosa els permet incloure publicitat a les seves pàgines i, evidentment, cobrar per això.

Quan acabes una nova pàgina pots enviar un missatge anunciant-ho a les llistes de correu o anunciant-ho en un grup de *news* i també pots introduir la referència en alguns dels motors de cerca com www.yahoo.com o www.ozu.com.

Motors de cerca geogràfica

És possible buscar informació per aproximació geogràfica. Aquest tipus de motor resulta útil si es busca un servei en un país que encara no compta amb un gran nombre de servidors Web. Aquest tipus de cerca s'utilitza rarament exceptuant quan es coneix amb certa exactitud el nom de l'organisme buscat i especialment la seva localització.

La interfície mostra un mapa o llista de països i sol·licita la zona desitjada. La cerca es realitza llavors per acostament progressiu, sobre mapes o sobre llistes cada vegada més refinades (països, regions, ciutats) fins a arribar a una compilació dels serveis disponibles. Existeixen un gran nombre d'aquest tipus de serveis que recullen geogràficament els recursos del Web. Alguns estan especialitzats en un continent i d'altres són específics d'un país.

Motors de cerca temàtica

Són serveis que intenten recollir els recursos del Web classificats per temes. La cerca es fa a través de rúbriques cada vegada més precises, per tal d'arribar a una llista de seus pertanyents a una categoria el més precisa possible.

El Servei Google i Yahoo són els exemples més clàssics de motor de cerca temàtica. A la seva pàgina de benvinguda ofereix diverses opcions:

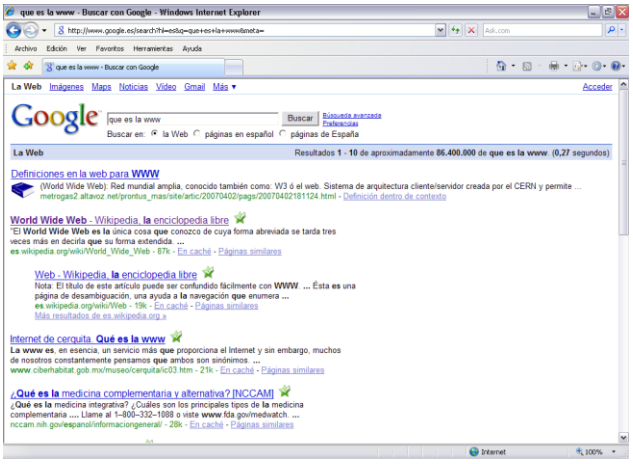
Per cercador (també anomenat motor de cerca o search engine):

És el mètode més usat i recomanat. És una pàgina web que ens permet trobar altres pàgines web a partir de paraules; aquestes paraules poden ser: el nom de l'empresa que estem buscant, una paraula que indiqui el tema que volem buscar, el servei buscat, etc. Els cercadors més coneguts són:

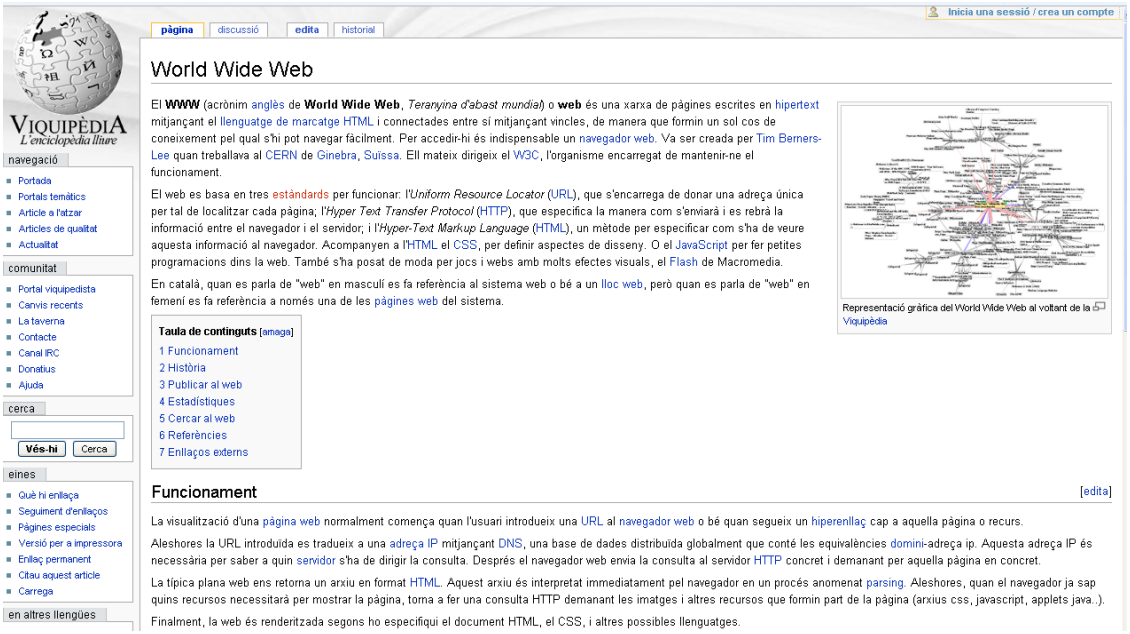
www.google.cat
www.yahoo.cat

Aquestes són les pàgines d'inici d'aquests cercadors:

Aquest tipus de recerca és ideal si estem buscant informació sobre un tema important. Una vegada escrites les paraules a la casella del tema buscat, es prem la tecla ENTER o bé el botó "Buscar con Google" o "Web". El cercador ens mostrarà moltes pàgines que parlen d'aquest tema. Cada pàgina ve indicada amb un títol i una petita explicació, que ens pot donar una idea més concreta sobre el que podem trobar en aquesta pàgina sense necessitat d'entrar-hi.



En cas que realment ens interessi veure tota la informació de la pàgina escollida, hem de prémer sobre el seu títol i automàticament el navegador ens durà a la pàgina web desitjada. El temps que tarda depèn sobretot de com estigui de saturada la màquina que serveix aquesta pàgina a Internet. Per exemple, si ens interessa el segon enllaç que hi ha a la llista anterior, li fem un clic i veurem la nova pàgina amb la seva adreça a la barra superior:



Per paraules clau:

Si volem obtenir informació sobre l'edifici Empire State Building de Nova York, el millor és buscar-la mitjançant paraules clau posant precisament aquestes paraules: Empire State Building en el cercador. Podem fer el que fa tothom: busquem el nostre nom a Internet, introduint a la casella de cerca el nostre nom. Si especifiquem massa, pot ser que no trobem res, però si només posem un nom i un cognom, segurament trobem pàgines Web on apareguin aquests noms, referint-se a persones que es diuen igual que nosaltres, que en poden haver vàries al món.

Els cercadors ens permeten concretar el tipus de cerca que realitzem. Quan busquem alguna cosa amb diverses paraules clau, el cercador pot pensar que volem les pàgines web en les quals aparegui alguna d'aquestes paraules, o totes elles a la mateixa pàgina, o totes elles en el mateix ordre que les hem escrit i seguides una rere l'altra. Tot això es pot especificar abans de realitzar la cerca.

És molt normal que quan realitzem una cerca per paraules trobem un resultat de 50.000 o 100.000 pàgines web que contenen aquesta paraula clau. Quan passa això podem concretar més la nostra cerca, afegint més paraules clau a la

cel·la de cerca, de manera que puguem reduir el nombre de resultats a 50 o 100, com a molt, per poder mirar-los d'un en un.

Exemples de cerques a Google: hi ha molts més cercadors a part de Google i Yahoo, encara que aquests siguin els que utilitza tothom. Els cercadors més utilitzats els pots veure a www.alexa.com/site/ds/top_500. Aquí també podràs veure els webs amb més tràfic a Espanya o a qualsevol altre país.

Cerca avançada

Per obtenir una cerca molt més precisa podem usar l'opció de cerca avançada de Google, que es troba a la part dreta de la casella de cerca:

A més de permetre't introduir els termes de la teva cerca al camp de cerca, Google ofereix infinitat d'opcions. Gràcies a la recerca avançada, podràs buscar exclusivament pàgines que:

- ➔ continguin TOTS els termes de la recerca,
- ➔ continguin la frase exacta de la consulta,
- ➔ continguin almenys un dels termes de la consulta,
- ➔ no continguin cap dels termes de la consulta,
- ➔ estiguin redactades en un idioma determinat,
- ➔ s'hagin creat en un format d'arxiu específic,
- ➔ s'hagin actualitzat en un període de temps determinat,
- ➔ pertanyin a un domini o lloc web en particular,
- ➔ no continguin material per a adults.

Cerca d'arxius

Molts llocs d'Internet guarden un munt d'arxius que no són accessibles a través de pàgines web. La característica que tenen aquests arxius és que es mostren al navegador amb la següent 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	

Fixeu-vos que a la part superior apareix la frase "*Index of*". El que cal fer és utilitzar una cerca a Google que busqui la frase "*Index of*" seguida dels tipus d'arxiu que volem, separats per una pipa (|), que s'obté prement les tecles ALT GR + 1. Per exemple, una cerca de cançons en format mp3 o vídeos mp4 i .avi d'arxius sense pàgina web seria: "*Index of*" mp3|mp4|avi:

O bé una cerca de documents Word (.doc), PDFs (.pdf) o PowerPoints (.ppt), que tractin de la malaltia celíaca: "index of" pdf|doc|ppt "enfermedad celiaca":

Normalment solen aparèixer pàgines web que en realitat són una llista d'arxius.

Llavors n'hi ha prou amb triar un arxiu i fer clic a l'arxiu per reproduir-lo. Si no es reprodueix o no es descarrega, llavors cal anar a la *home* de la pàgina web, per això s'ha d'esborrar l'adreça de la barra d'adreces del navegador per deixar només el nom del domini.

Cercador de cercadors

Hi ha tants cercadors, que s'ha creat un cercador de cercadors a www.buscopio.net.

Cercadors semàntics

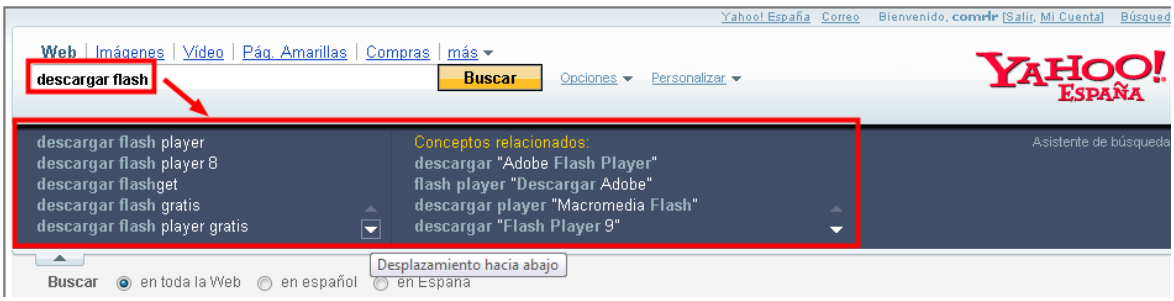
Hi ha també cercadors als quals se'ls pot fet preguntes, anomenats cercadors semàntics, que en comptes d'escriure tu unes paraules clau, els fas preguntes com si fossin persones. Alguns exemples són: ww.ask.es, demos.bitext.com. i molts altres, però en anglès, com www.hakia.com o www.lexxe.com.

Cercador de cerques

Què és el que la gent busca per Internet i quines respostes han obtingut? Mira-ho a www.answers.com.

Cercador amb typing suggestions

A mesura que teclejes la cerca que vols fer, el cercador et suggereix paraules de cerca: per exemple www.cuil.com. A més, tabula els resultats. Yahoo i Google també tenen *typing suggestions* basades en les cerques d'altres usuaris. Aquesta característica es pot configurar i es pot desactivar. Per exemple, aquí tenim les *typing suggestions* de Yahoo:



Bloqueig de webs perillosos

Els cercadors solen bloquejar els webs que se sap que són perillosos, ja que pots descarregar virus, *spyware*, *adware*, *troians* i d'altres. Per exemple, aquesta és una pàgina bloquejada a Yahoo:

Motors de cerca per paraules clau

El servei indexa prèviament un determinat nombre de pàgines del Web de tot el planeta. Aquesta indexació s'efectua per text complet i totes les paraules de totes les pàgines referenciades es converteixen en entrades de l'índex, potencialment objecte de la cerca.

La indexació prèvia de les pàgines pot fer-se de diferents maneres:

Declaració voluntària del responsable de la seu Web remota que indica al motor de cerca, emplenant un formulari, l'existència d'un servei. El motor indexa llavors totes la pàgines del servei referenciat d'aquesta manera.

Robots que recorren de manera automàtica el Web i les seves pàgines d'informació. Aquests robots parteixen d'un cert nombre de pàgines bàsiques i persegueixen tots els enllaços d'hipertext en cadascuna de les pàgines trobades.

Cada pàgina trobada s'indexa per text complet. Cada indexació de pàgina provoca una captura de les informacions que presenten i un emmagatzematge de l'arxiu obtingut en els discs durs del motor de cerca.

Aquest servei evita la declaració prèvia d'un servei, des del moment que un enllaç d'hipertext apunta cap a ell a en qualsevol lloc del món. Només indexen informació trobada a les pàgines HTML i no té en compte la informació continguda a les bases de dades específiques dels organismes presents al Web, accessibles mitjançant formularis.

Generació del catàleg a mà: no és realment així, però alguns dels cercadors treballen sobre catàlegs que són gestionats per una o diverses persones. En

aquest cas, l'índex es genera a partir d'entrades realitzades a mà, de manera que qui introdueix una referència pot indicar quines paraules s'han d'indexar.

Visió general de la tecnologia d'indexació

Google és l'única empresa abocada a desenvolupar el "motor de recerca perfecte", definit pel seu cofundador Larry Page com una cosa que "comprèn exactament el que l'usuari vol dir i li lliura exactament el que està buscant". Amb aquesta objectiu, Google insisteix a continuar innovant i es nega a acceptar les limitacions dels models existents. Per això, va desenvolupar la seva pròpia infraestructura de servidors i l'avançada tecnologia *PageRank*™, que va canviar la manera de realitzar les cerques.

Des del principi, els programadors de Google van reconèixer que, per proporcionar els resultats més ràpids i precisos, calia una nova configuració de servidors. A diferència de la majoria dels motors de cerca que utilitzen un grup de servidors grans que solen ralentir-se quan processen pics de càrrega, Google utilitza equips connectats per trobar ràpidament la resposta a cada consulta. Aquesta innovació va permetre assolir temps de resposta més ràpida, una escalabilitat i menors costos. És una idea que altres han copiat des de llavors, mentre que Google segueix polint la seva tecnologia interna per fer-la cada vegada més eficient.

El programari integrat a la tecnologia de cerca de Google realitza una sèrie de càlculs simultanis en tan sols una fracció de segon. Els motors de cerca tradicionals es basen, en gran part, en la freqüència amb què una paraula apareix en una pàgina web. Google, en canvi, utilitza la tecnologia *PageRank*™ per examinar tota l'estructura de vincles del Web i determinar quines pàgines són les més importants. A continuació, realitza una anàlisi de concordança d'hipertextos per establir quines pàgines són rellevants per a la cerca específica que s'estigui processant. En combinar la importància general amb la rellevància específica respecte d'una consulta en particular, Google pot col·locar els resultats més rellevants i fiables en primer lloc.

Tecnologia *PageRank:* *PageRank* realitza una mesura objectva de la importància que tenen les pàgines web. Per a això, resol una equació que conté més de 500 milions de variables i 2.000 milions de termes. En lloc de comptar els vincles directes, *PageRank* interpreta un vincle de la pàgina A a la pàgina B com un vot que rep la pàgina B de part de la pàgina A. *PageRank* avalua, d'aquesta manera, la importància que té una pàgina determinada comptant la quantitat de vots que rep.

PageRank també considera la importància de cada pàgina que emet un vot, ja que als vots procedents de determinades pàgines se'ls atorga un valor més gran, incrementant així el valor de la pàgina vinculada. Les pàgines importants reben una valoració de *PageRank* més alta i apareixen a la part superior dels resultats de cerca. La tecnologia de Google utilitza la intel·ligència col·lectiva del Web per determinar la importància d'una pàgina. Els resultats s'obtenen sense cap tipus de participació humana; per aquest motiu, els usuaris han arribat a confiar en Google com a font d'informació objectiva, lliure de la

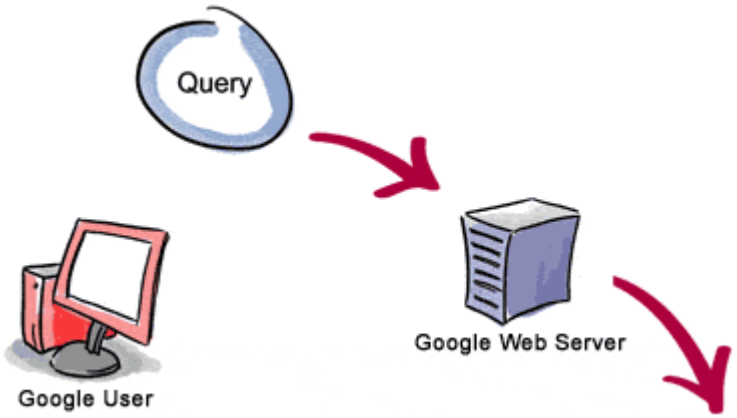
manipulació que es genera quan els llocs paguen per ocupar determinada posició en els resultats.

Anàlisi de concordança d'hipertextos: el motor de cerca de Google també analitza el contingut de cada pàgina. Tot i així, en lloc d'explorar simplement el text de la pàgina (que els editors de llocs poden manipular mitjançant metaetiquetes), la tecnologia de Google analitza tot el contingut d'una pàgina i té en compte també les fonts, les subdivisions i la ubicació precisa de cada paraula. Així mateix, Google analitza el contingut de pàgines web veïnes per garantir que els resultats trobats són els més rellevants per a la consulta de l'usuari.

Les innovacions de Google no es limiten a l'escriptori. Perquè els usuaris que accedeixen al Web a través de dispositius portàtils puguin obtenir resultats de cerca ràpids i precisos, Google va desenvolupar també la primera tecnologia de cerca sense fil que tradueix al moment el codi HTML a formats optimitzats per a WAP, i-mode, J-SKY i EZWeb. Actualment, Google proveeix la seva tecnologia sense fil a diferents líders del mercat, per exemple, a AT & T Wireless, Sprint PCS, Nextel, Palm, Handspring i Vodafone, entre d'altres.

La vida d'una consulta de Google

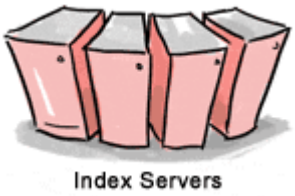
Una consulta de Google sol durar menys de mig segon i, tot i així, implica tota una sèrie de passes que s'han de completar abans que la persona que està buscant informació pugui veure els resultats.



3. L'usuari rep els resultats de la seva cerca en una fracció de segon.

1. El servidor web envia la consulta als servidors d'indexació. El contingut que es troba als servidors d'indexació és similar a l'índex d'un llibre: indica quines pàgines contenen les paraules que coincideixen amb la consulta.

2. La consulta viatja fins als servidors de documents, en els quals es recuperen els documents emmagatzemats. Es generen fragments per descriure cada resultat de cerca.



Simple Search o cerca simple:

Search: permet efectuar cerques a pàgines web o a Usenet.

Display The Results: dóna l'opció d'obtenir els resultats de les cerques sota diferents aspectes més o menys detallats.

Format de paraules clau:

- Si s'escriu en minúscules, busca majúscules i minúscules en qualsevol posició.
- Si s'escriu alguna lletra en majúscules, busca l'ocurrència exacta.
- Els accents segueixen les mateixes regles que en els casos anteriors.
- Permet comodins amb restriccions: darrere de la tercera lletra, substitueix de zero a cinc lletres, només minúscules, les majúscules i xifres no es tenen en compte.
- Cerca de paraules adjacents separades per ; o amb cometes separades per espai.
- Signe + per a AND i signe – per a O.
- Poden combinar-se totes les possibilitats anteriors.

Cerques específiques:

- Cerca per un enllaç hipertext, anchor:page.
- Cerca d'applet Java, applet:javaclass10.
- Cerca al nom del servidor, host: nombre.org.
- Cerca d'una imatge, image: logonodo50.gif.
- Cerca a l'URL d'un enllaç, link:nodo50.org.
- Cerca únicament al text, text:Nodo50.
- Cerca per un títol de pàgina, title:homepage.
- Cerca a l'interior d'un URL, url:tintin.html.

Cerca de direccions de correu

Els principals serveis de cerca de direccions de correu són:

WHOWHERE:www.whowhere.com o spanish.whowhere.com.
FOUR11 DIRECTORY SERVICES: www.four11.com.

INTERNET ADRESS FINDER: www.iaf.net.

Cerca de llistes de distribució

El principi de les llistes de distribució es basa en la missatgeria electrònica, però la diferència principal és que no s'envia un missatge a una persona sinó a una llista de distribució, que representa un grup d'usuaris que s'han abonat prèviament.

Quan algú dirigeix un missatge a una llista, la informació es distribueix a tots els seus abonats, que poden ser milers. Vostè mateix rebrà a la seva bústia tots els missatges enviats a una llista a la qual s'hagi abonat.

Aquestes llistes de missatgeria estan normalment relacionades amb un tema concret. Tot àmbit de reflexió és susceptible de ser objecte d'una llista. Qualsevol persona pot implementar una d'aquestes llistes sense cap restricció, contràriament als fòrums de discussió.

Una llista de distribució té dues adreces de correu: una adreça administrativa que serveix per a abonaments, baixes, etc., i una adreça per a les pròpies discussions. Per abonar-se a una llista enviï un missatge d'aquest tipus: "subscriu llista cognom nom", en el qual llista és el nom de la llista a la qual desitja abonar-se.

Cercador de servidors FTP

Els programes d'FTP permeten accedir a distància a ordinadors anomenats servidors FTP, que posen a la disposició de la comunitat documents o aplicacions per a descarregar-los. FTP permet connectar-se a aquests servidors mitjançant un *login* anònim. Permet realitzar transferències des del servidor remot al disc local i en algunes ocasions en sentit invers, però això no és normal sobretot per mesures de seguretat.

Cerca de fòrums de discussió

Els *News* o Fòrums de discussió es troben sovint en paral·lel amb les llistes de distribució. Els dos procediments permeten l'intercanvi d'informació entre persones, en comunitats interessades per un mateix tema de reflexió més o menys vast.

Les llistes de distribució es basen en el correu electrònic mentre que els fòrums de discussió es basen en una xarxa específica anomenades Usenet.

A les llistes de distribució els missatges s'emmagatzemen a les màquines dels abonats i són aquests els que decideixen quan esborrar-los. Als fòrums els missatges s'arxiven als servidors i s'esborraran quan el servidor vulgui. Cal tenir en compte que diverses vegades al dia tots els servidors s'intercanvien els missatges per estar al dia. Cada propietari de servidors de *News* decideix quins són els temes i, per tant, els grups que difondrà. N'hi ha aproximadament uns 22.000 a tot el món.

Per obtenir informació sobre el servei *News*, per exemple: http://groups.google.com/ o http://groups.yahoo.com/

Metodologia de cerca

No s'ha de prendre mai al peu de la lletra una informació d'Internet.

No prendre mai una informació oposada en Internet al peu de la lletra. La informació que hi ha és majoritàriament gratuïta i això té els seus avantatges i inconvenients. No existeix un estàndard en la definició d'un document a Internet: autor, data de creació, grandària, data d'actualització, tots aquestes dades són difícils de trobar si el creador de la informació no les proporciona.

Traducció de la consulta a llenguatge documental

Per explicar aquest punt us demanem que us descarregueu el pdf següent "Lenguaje documental" d'aquest curs.

Execució de l'estratègia de cerca

Valoració dels resultats

Sabem que si un cercador és més eficaç que un altre ha de tenir uns resultats de cerca tan depurats que el visitant aconsegueixi trobar allò que busca en el menor temps possible, això vol dir a les primeres pàgines de resultats.

Si llegim un interessant article de Bill Slawski al seu bloc (http://www.seobythesea.com/) podem aprendre alguna cosa més sobre els plantejaments futurs de cercadors en relació amb una patent que ha adquirit Google i que combina noves tècniques per trobar contingut duplicat i filtrar-lo convenientment.

No és la primera vegada que abordem aquest tema i les incògnites que ens suscita. Per saber destriar el contingut duplicat a la Xarxa cal trobar la seva autoria, conèixer el que l'autor proposa fer amb els seus drets, atorgar-los la importància per sobre de còpies pirates i no permeses, etc. En el cas d'informacions comercials, fitxes de productes, catàlegs, etc., fins al moment sembla que es valorava la data d'expansió en cercadors i la seva autoria. Sabem que no serveix de res posicionar un producte els primers si després no reunim altres privilegis per al cercador que ens ajudin a mantenir-nos en un posicionament òptim. Quedarem relegats en els resultats de recerca ràpidament per sota dels nostres competidors.

Tractament de documents propers a la duplicitat

El cercador haurà de destriar i tractar la informació curosament en funció de l'anàlisi que en faci, tenint en compte factors com l'originalitat, el tipus de còpia, els drets d'autor.

Ens trobem amb el mateix text publicat en diferents suports (HTML, pdf, doc, xls, text pla, etc.) o en suports per a impressió o enfocats a telèfons mòbils.

Compartir notícies i articles en fonts RSS publicades en blocs i altres tipus de webs.

Utilització de pàgines "mirall" amb fins transparents com intentar evitar demores de càrrega de webs o facilitar-ne l'ús en condicions adverses.

Detecció de pàgines que han violat els drets d'autor.

Publicació del mateix contingut repetidament al mateix lloc web.

Els esforços recents de Google per combatre el contingut duplicat

L'any passat alguns empleats de Google van fer un bon treball relacionat amb el tema de conjunts de processos per a la detecció de duplicats i formes de valorar-los. "Detección de Proximidad para la web duplicados Rastreo" (pdf).

Un dels processos descrits en detall al document esmentat va ser desenvolupat per Moses Charikar, un professor de Princeton que va treballar per a Google en el passat i que va ser l'inventor d'una patent comprada per Google l'any passat en la qual es fa referència a una sèrie de mètodes en relació amb el tema de la duplicitat de continguts.

- Trobar arxius similars en una gran Xarxa.
- Empremtes digitals en documents.
- Copiar mecanismes de detecció de documents digitals.
- Agrupació sintàctica en el Web.
- Similitud de tècniques d'estimació d'arrodoniments en algorismes.
- Similitud amb el sistema de recerca d'estructures de dades compactes.
- Mètodes per a la identificació de documents versionats i plagiats.

La conclusió d'aquests estudis és que cap dels algorismes està funcionant perfectament per ajudar a trobar duplicitat en el mateix lloc web, però sí que aconsegueix una alta precisió i podria incorporar-se a les noves tècniques de Google en la seva contínua millora del seu cercador.

Com a conclusió, Bill Slawski apunta que el procés descrit en aquesta nova patent no introdueix un nou mètode d'identificació de contingut duplicat, però sí aporta un nou enfocament pel que fa als seu mètodes de detecció. Qui sap si aquestes tècniques s'utilitzaran definitivament pels enginyers de Google per aportar més transparència al cercador?