Capítol 8 – Cerca de documents audiovisuals de química en el WWW

1. Consideracions inicials
2. Cerca d'informació a Internet
3. Terminologia relacionada amb la cerca d'informació a Internet
4. Funcionament dels serveis de cerca a Internet
5. Internet i el WWW
6. Tipologia i rànquing de cercadors
7. Resultats i discussió

8. Cerca de documents audiovisuals de química en el WWW

8.1. Consideracions inicials

L'ús del codi digital com a sistema per emmagatzemar o transmetre tota mena d'informació, siguin textos escrits, paraules, música, nombres o imatges, comporta una revolució en l'organització de totes les activitats relacionades amb la difusió de la informació i el coneixement. Al llarg de la història, cada tipus d'informació ha tingut un codi propi de transmissió: alfabets, ones sonores, etc., i ha creat també unes tecnologies pròpies: impremta, ràdio, fotografia, cinema, etc. absolutament diferents les unes de les altres. Joan Majó —president de la Information Society Forum de Brussel·les i de la Corporació Catalana de Ràdio i Televisió— defineix la societat digital actual com una societat del coneixement (Majó, 1998).

Avui dia, la convergència de codis implica la convergència de tecnologies. Una de les conseqüències previsibles d'aquest fet és la sobreabundància d'informació. La quantitat d'informació a l'abast es multiplica d'una manera constant i aparentment sense límit. El problema principal per als usuaris no és la manca d'informació sinó l'excés. Per tant, serà una habilitat imprescindible saber seleccionar, filtrar, ordenar, valorar i assimilar la informació per tal de convertir-la en coneixement útil (Majó, 1997).

A més, cal considerar que malgrat que el món és tridimensional i multimèdia, la informàtica ha estat sempre fonamentalment textual, i avui dia Internet és encara predominantment textual. Aquesta situació anirà canviant, d'una banda, gràcies als avenços en la digitalització de qualsevol tipus d'informació i, de l'altra, per l'augment de la banda ampla, que facilita la transmissió dels grans volums d'informació associats als documents audiovisuals digitalitzats. S'observa una tendència creixent a fer d'Internet un entorn multimèdia que incorpora imatges i sons, sense desplaçar el text. S'estan passant una part o una gran part dels fons audiovisuals creats en format analògic a format digital, i també s'estan creant molts documents directament en aquest format. La informació és cada vegada més audiovisual en qualsevol camp del coneixement.

L'aparició i l'ús generalitzat de les tecnologies de la informació i la comunicació en molts aspectes de la vida diària produeixen canvis de grans dimensions en l'entorn en què els sistemes educatius desenvolupen la seva tasca. L'any 1998 Roger Kaufman[38] comentava que l'última panacea en el camp de la tecnologia educativa era Internet i el World Wide Web. Els nous vehicles per obtenir ràpidament dades i informació d'una gent àmpliament dispersada són benvinguts i excitants. Es compara el que representa Internet per a l'educació i l'aprenentatge amb el que va suposar la premsa escrita per a la societat preindustrial; és a dir, fa que siguin més barats, més ràpids i millors. És la nova eufòria. És rara l'organització pública o privada que no estigui ja en el web (Kaufman, 1998). Evidentment, la visió de Kaufman no deixa de ser crítica respecte al protagonisme que s'atorga a aquest mitjà, i insisteix en la necessitat de considerar-lo com a tal i donar més importància als continguts que a la forma de transmetre'ls. Encara és més crític Pierre Lévy (1997) quan afirma que estem davant del segon diluvi a escala planetària; el primer va ser d'aigua, el segon és d'informació.

[38] Professor i director de l'Office for Needs Assessment & Planning, Learning Systems Institute, Florida State University.

A mesura que augmenta el volum d'informació disponible a través del WWW es fa patent la principal paradoxa de la informació: la seva abundància és un dels obstacles fonamentals a l'hora de consumir-la. Encara que sovint es diu que la informació és immaterial, el cert és que no tan sols no ho és sinó que el seu processament està sotmès a les limitacions més materials que es puguin imaginar. Com a conseqüència, la capacitat de processament de la informació dels éssers humans està tan sotmesa a restriccions com qualsevol altre procés material (Codina i Del Valle, 2001).

El volum d'informació es duplica en un espai de temps cada vegada més curt, la qual cosa provoca una sobreabundància d'informació. En uns estudis elaborats per la School of Information Management and Systems (SIMS)[39], de Berkeley, s'estima que la quantitat de nova informació emmagatzemada en paper, film i medis òptics i magnètics va créixer gairebé un 30% anualment entre el 1999 i el 2002. La que es produeix en suport paper és ínfima (menys del 0,003% del total), i majoritàriament s'emmagatzema després en suport digital. Així, una gran part ja neix en aquest format i és produïda des d'un ordinador. Segons aquests estudis, es consumeix cada cop una part menor de la informació que es produeix. Paradoxalment, aquesta sobreabundància pot tenir un efecte de desinformació en les persones o bé constituir un obstacle per al coneixement. L'excés pot causar impermeabilització, immunització i actituds no receptives enfront de la quantitat d'informació no buscada amb què ens trobem. Davant d'aquesta situació adquireix encara més valor pedagògic la capacitat amb què les persones accedeixen a la informació de forma intencional, eficaç i selectiva (Martínez i Bujons, 2001, p. 29).

Encara que la informació augmenta, el temps de què disposem per llegir-la, veure-la o sentir-la no varia, amb la qual cosa cada cop té més sentit «l'economia de l'atenció» (Cornella, 2001). Alfons Cornella anomena «infoxicació» l'excés d'informació que ens envolta pel fet que disposem de molta més informació de la que podem manejar. Afirma que gestionar la informació serà una part important en qualsevol activitat relacionada amb el coneixement. Així doncs, caldrà aprendre'n, ja que normalment es fa malament i de manera poc estructurada perquè no s'ha rebut cap mena de formació (Cornella, 2000).

Un dels grans inconvenients que té Internet com a recurs pedagògic es deriva precisament de la dificultat a l'hora de trobar la informació que es desitja. La quantitat d'informació que tenim a la nostra disposició és tan gran que trobar allò que realment ens interessa pot ser, en molts casos, una tasca molt complexa, desagraïda i plena d'obstacles. L'excés d'informació es converteix fàcilment en un problema bàsicament per dos motius:

1. És molt difícil localitzar la documentació rellevant per a una determinada necessitat d'informació.

2. Es tendeix a assimilar superficialment la informació.

Per resoldre aquesta situació cal tractar la informació a fi de poder-la seleccionar amb facilitat segons les demandes concretes.

[39] Es poden consultar a: http://sims.berkeley.edu/research/projects/how-much-info/ (2000, 2003)

8.2. Cerca d'informació a Internet

Internet ja no és un possible recurs que cal considerar en una investigació, sinó que s'ha erigit en un instrument necessari, encara que no suficient, per abordar qualsevol treball d'investigació amb independència del camp científic en què s'inclogui. En alguns aspectes, com ara la cerca d'informació, la relació amb altres investigadors o la difusió de treballs es fa imprescindible, ja que moltes de les fonts de partida tenen com a únic lloc de residència Internet (Martínez et al., 2001). Internet és la font d'informació més gran que existeix actualment. Hi ha productes de caràcter informatiu i documental, que abans es difonien de manera dispersa i material, creats en exclusiva per ser difosos a través d'aquest mitjà. Un dels principals problemes d'Internet és la localització de la informació d'interès, és a dir, la selecció. Per treure'n la màxima rendibilitat des d'un punt de vista informatiu i documental cal conèixer i utilitzar totes les eines que existeixen per localitzar recursos (Maldonado, 2001).

Cercar informació a través de la xarxa no és tan senzill com pot semblar a primer cop d'ull. Els subjectes més estratègics, els que fan cerques conscientment, intencionalment, de manera complexa i flexible, arriben a millors resultats (Fuentes, 2001). Cal definir amb la màxima exactitud i concreció possible l'objectiu, o els objectius, abans de fer la cerca. Cal conèixer els diferents instruments de recerca —directoris, cercadors o metacercadors— i quines són les opcions que ens permeten fer una cerca adequada.

Un cop localitzada la informació cal seleccionar-la, ja que generalment a partir d'una cerca a Internet s'obté un nombre molt elevat de documents i no és possible revisar-los tots, ja que es requereix d'una quantitat de temps molt elevada. Es pot dur a terme la selecció tenint en compte diversos criteris, entre els quals hi ha l'afinitat amb el que s'està localitzant, així com la fiabilitat o la veracitat de la font de la qual procedeix la informació.[40]

Cal que la persona que investiga tingui la solvència adequada per a l'aplicació dels criteris. Quant al rigor, a la fiabilitat i a la veracitat de la informació, pot servir d'orientació conèixer-ne l'origen i l'autoria, que sigui contrastable, mirar en quina freqüència s'actualitza, si conté un accés a la direcció electrònica de l'autor i/o patrocinador, l'actualitat temàtica, l'existència d'enllaços pertinents, la freqüència de consultes, la facilitat de navegació i la seguretat del sistema. També calen criteris per valorar la qualitat del contingut; cal saber què i per què es busca una informació.

8.3. Terminologia relacionada amb la cerca d'informació a Internet

La consulta no seqüencial d'un document digital s'anomena «navegació hipertextual», i els programes informàtics que la fan possible són els sistemes de gestió d'hipertextos (SGH). El procés de selecció de la informació adequada a les demandes dels usuaris s'anomena «recuperació de la informació» (RI), i els programes informàtics que en permeten l'automatització són els sistemes de recuperació de la informació (SRI).

[40] A Internet, concretament a l'espai WWW, qualsevol persona pot tenir-hi penjada una pàgina personal amb les informacions que hagi estimat oportú posar-hi sense cap mena de control sobre el seu contingut.

- *Robot*. És un programa d'ordinador o programari dissenyat per recórrer de forma automàtica l'estructura hipertext del web, amb la finalitat de crear automàticament bases de dades textuals a partir de documents HTML (la sigla prové de l'anglès *hypertext markup language*, 'llenguatge d'etiquetatge d'hipertext') distribuïts pels diversos servidors. També se'ls anomena *spiders*, *wanderers*, *crawlers*, *worms*, etc., denominacions que fan referència a la seva forma de treball (Vaquero, 1997).

- *Motor de cerca* (*search engine*, agent). És una eina web que localitza de forma ràpida informació existent a Internet i que està formada per tres elements: interfície, robot i base de dades. També se'ls denomina «cercadors». Permet a un usuari entrar una pregunta (*query*) i cercar una base de dades que cobreix una molt substancial part del contingut del web.

- *Índexs o directoris*. És una pàgina o full web on, sota un arbre de matèries, es troben organitzades les diferents pàgines existents a Internet. Els índexs o directoris es caracteritzen perquè les adreces que tenen són recopilades, organitzades i classificades de forma manual, generalment per universitaris que es dediquen a aquesta activitat.

A vegades costa fer distincions, ja que hi ha índexs que presenten la mateixa interfície que els motors de cerca, o sigui, tenen camps on s'introdueixen els termes pels quals es vol realitzar la cerca, i també hi ha motors que tenen la possibilitat de cercar a través de directoris. Per diferenciar els índexs dels motors cal llegir algun fitxer d'informació de l'eina o escriure un correu electrònic al responsable per sol·licitar-lo.

Els índexs, també anomenats «directoris temàtics», són grans bases de dades organitzades per professionals que les avaluen i les col·loquen en la categoria adequada. A més, un gestor de pàgines web fa de passarel·la entre la base de dades i l'usuari que consulta. La informació es presenta classificada en diversos grups conceptuals encapçalats per termes orientatius, i cada grup està dividit jeràrquicament en més subcategories, a través de les quals es va descendint a nivells d'especificitat fins a trobar el que es busca. Són molt bons per acostar-se per primera vegada a un tema. Com que s'actualitzen molt lentament, es poden obtenir documents obsolets o adreces que ja no existeixen; els motors poden ser útils per actualitzar la informació que s'obté a través dels índexs.

8.4. Funcionament dels serveis de cerca a Internet

Els servidors de cerca d'Internet són sistemes basats en la relació client-servidor que caracteritza Internet.

En la part del servidor consten de dos grans subsistemes: un programa per localitzar documents i una base de dades per gestionar la informació sobre ells. La base de dades es divideix en dues parts o subsistemes: un motor d'anàlisi i d'indexació, i un sistema de consulta o interrogació. Un quart programa fa de passarel·la (*gateway*) i serveix d'unió entre el servidor web que gestiona les peticions d'informació dels navegadors (entorns que fan servir els usuaris d'Internet) i la base de dades que crea i gestiona els índexs.

En la part del client hi ha el navegador, que és l'únic sistema que interactua directament amb l'usuari. El navegador recull les preguntes de l'usuari i les envia al servidor http (sigla que prové de l'anglès *hypertext transfer protocol*, 'protocol de transferència d'hipertext'), que les envia a la base de dades a través del programa passarel·la. El navegador també rep les respostes de documents web que envia el servidor i les presenta a l'usuari (Codina, 1997a).

Un servei de cerca conté, doncs, quatre parts:

- *Primera*: un programa que explora Internet (robot) per localitzar documents i adreces de documents.

- *Segona*: un sistema automàtic d'anàlisi i d'indexació dels documents localitzats pel robot.

- *Tercera*: un sistema d'interrogació. Aquesta part inclou el llenguatge de consulta que està a disposició de l'usuari per expressar les seves necessitats d'informació, així com un procediment per llançar consultes a l'índex creat pel sistema anterior.

  Les parts segona i tercera constitueixen la base de dades del servidor.

- *Quarta*: un programa que actua de passarel·la entre el servidor de documents HTML i la base de dades. Aquest programa s'activa quan el servidor rep la petició del client.

8.4.1 Com actuen els robots

Els robots o agents d'Internet són programes que exploren periòdicament i contínua Internet amb l'objectiu de localitzar nous documents. La seva activitat s'articula al voltant de la creació d'una llista d'adreces d'ubicacions web, de l'accés i de la lectura dels documents continguts en les adreces, així com del manteniment de la llista d'adreces. Per dur a terme aquestes accions, primer parteix d'una llista inicial d'adreces URL (sigla que prové de l'anglès *uniform resource locator* 'localitzador uniforme de recursos') que han estat compilades a mà pels administradors del robot. Posteriorment, aquesta llista es nodreix de les inscripcions voluntàries que realitzen els administradors d'altres ubicacions web que emplenen en línia uns documents que ofereixen els servidors de cerques, i dels documents localitzats pel robot (Codina, 1997a). Cada cop que un robot arriba a una pàgina comprova si ja l'havia visitat abans o si és nova. Si ja l'havia recollit, s'assegura que no hagi sofert modificacions i, en el cas que n'hi hagi, actualitza la informació que emmagatzemava sobre aquesta pàgina.

Si és la primera vegada que hi accedeix, agafa les dades de localització i la indexa. Des de cada pàgina visitada, el robot té accés a altres pàgines a través dels nodes que hi troba i rastreja els enllaços que proporciona la pàgina principal (Marcos, 1998b). En qualsevol cas, fent servir la mencionada llista d'adreces, els robots accedeixen una a una a les pàgines principals de cada ubicació web i llegeixen les noves adreces que troben, rastrejant les etiquetes del llenguatge HTML que senyalen enllaços cap a altres documents. Totes aquestes adreces passen a formar part de la llista original, que d'aquesta manera s'incrementa contínuament amb noves adreces.

De cadascuna de les pàgines llegides per aquest procediment copien una part o la totalitat del text que hi figura i l'envien a la base de dades. Com que un document HTML no consisteix en un conjunt de pàgines superposades, sinó que és una xarxa de nodes amb *n* nivells de profunditat, si un document web té *n* nivells, alguns robots exploren només el primer node del primer nivell, mentre que altres poden explorar diversos nodes del primer nivell i fins i tot continuar aprofundint fins a arribar a un nivell determinat. Tres sembla que és el màxim de profunditat a la qual arriben els cercadors actuals. Finalment, el robot executa accions de manteniment de la llista d'adreces, revisant-la periòdicament per observar els canvis que s'hi puguin produir, com ara noves pàgines, modificacions de les pàgines, URL canviades, etc.

8.4.2 Com funciona la base de dades

La base de dades és un programa de gestió documental semblant a les bases de dades documentals convencionals. Rep com a entrada el text dels documents localitzats pel robot i produeix com a sortida un índex invertit.[41] Molt sovint hi ha paraules úniques o bé les seves arrels, ja que molt sovint el fitxer invertit emmagatzema només les arrels de les paraules. Cal tenir en compte que solen excloure's del fitxer les anomenades «paraules buides» (*stopwords* o *noise words*), com en les bases de dades convencionals. Com que la freqüència de les actualitzacions és molt variable, des de vint-i-quatre hores fins a diverses setmanes, sempre hi ha un desfasament entre la realitat d'Internet i el que reflecteix l'índex, que se suma al d'actualització de les direccions URL ja emmagatzemades, per la qual cosa és habitual que un cert nombre d'adreces de la llista recuperada no siguin vàlides i generin un missatge d'error en intentar accedir-hi.

Alguns fitxers invertits guarden més informació que d'altres, com ara la situació o l'ordre relatiu de cada paraula dins del document, els termes que apareixen en zones o etiquetes especials del llenguatge HTML com ara els de la secció <*head*>, on, a més del títol, es pot incorporar informació referida a l'autoria o al tema del document mitjançant paraules clau. Solen fer una ponderació de les paraules que es troben en determinades parts dels documents, com el títol, les capçaleres principals i els primers paràgrafs, ja que el contingut fonamental sol estar concentrat en aquests punts. Els que van millor són els que recuperen informació de les etiquetes <*meta*> que són etiquetes HTML que ofereixen informació sobre el document, tant de tipus formal com de contingut. L'inconvenient és que no totes les pàgines web en tenen, ja que el seu ús no està regulat. A més, molts cercadors no les revisen perquè consideren que contenen informacions poc fiables.

8.4.3 Com calculen els cercadors la rellevància dels resultats

Encara que cada robot fa servir algoritmes diferents, tots tenen un principi de funcionament semblant. Com que els ordinadors no poden interpretar el sentit de les paraules però en canvi poden fer operacions aritmètiques i lògiques a gran velocitat, la base del càlcul de rellevància és el nombre de parells d'atributs comuns entre la necessitat d'informació i els documents, la qual cosa es tradueix, en la pràctica, en el nombre de paraules comunes entre cada document i la pregunta (Codina, 1997a).

---

[41] Un índex invertit entra paraules úniques, cadascuna de les quals amb totes les entrades que les contenen. Així, si la paraula *Internet* apareix 10.000 vegades en uns 2.000 documents, no hi ha 10.000 entrades sinó una sola, juntament amb les 2.000 direccions on apareix.

Els procediments emprats oscil·len entre mètodes certament simples i mètodes lleugerament sofisticats. Un exemple dels primers consisteix simplement en el sumatori del nombre de vegades que apareixen els termes de la pregunta en cada document (nombre d'ocurrències). D'altres ponderen aquests valors absoluts amb el nombre de termes diferents de la pregunta presents en el document, o bé donen més importància als termes que apareixen en seccions clau del document, com el títol o el primer paràgraf del document. Finalment, un altre dels recursos més emprats és atorgar un pes desigual als termes segons la seva capacitat de discriminació. Es ponderen més alt els termes de baixa freqüència global, és a dir, que apareixen poc en el conjunt de la base de dades però amb alta freqüència local, i es ponderen amb valors baixos els termes molt abundants globalment.

Quan es fa una cerca, el sistema examina només la cadena de caràcters seleccionada (paraula) sense tenir en compte possibles accepcions sinònimes, i selecciona els documents que la contenen encara que sigui fora de context o de forma metafòrica. Un cop identificats els documents potencialment rellevants, com més vegades apareguin les paraules de la pregunta en un document més rellevant serà considerat, i per tant més alt serà el lloc que ocupi en el rànquing que es mostra a l'usuari. També cal tenir present el fenomen de les homonímies, pel qual una paraula serveix per designar referents diferents. També són freqüents les falses coordinacions, és a dir, combinacions de paraules que es refereixen a conceptes diferents.

Els sistemes de cerca tenen una fe cega en el sistema de rellevància, que porta a la supersimplificació del llenguatge d'interrogació. No importa que els usuaris no puguin expressar les seves necessitats d'informació amb massa detall, ja que només entren les paraules o les frases que l'expressen, sense preocupar-se de com relacionar-les. Si cal, l'usuari pot entrar tots els sinònims que vulgui del mateix terme per assegurar-se que no perd cap document rellevant. Se suposa que, encara que es recuperin milers de documents, l'usuari només haurà de consultar els 10 o 20 primers, ja que el sistema d'ordenació s'encarregarà de col·locar primer els més rellevants.

En realitat, el cert és que la simple repetició d'un terme està lluny de ser un indici fiable de la rellevància d'un document. Molts autors no repeteixen els termes sinó que fan servir molts sinònims, i per tant la riquesa de vocabulari col·locaria el seu document en els últims llocs del rànquing.

8.4.4 La recuperació de la informació

La recuperació de la informació a Internet (*information retrieval*) es fa a través d'un diàleg entre l'agent que fa una cerca i el sistema de recuperació. Mitjançant una o diverses preguntes i respostes (equacions de cerca i resultats) s'optimitza la crida i es minimitza el soroll i el silenci o documents no recuperats. Es considera soroll els documents no rellevants que han aparegut com a resultat de la polisèmia dels termes de la cerca. El silenci pot ser per no haver inclòs tots els termes de cerca que codifiquen el concepte. Si el resultat no és òptim serà sotmès a filtratge, en el cas que contingui massa soroll, o a una ampliació de la cerca, en el cas que el silenci sigui excessiu, sense descartar un replantejament radical de la cerca (Pinto et al., 2002, p. 167).

El web és un gegantesc magatzem d'hiperdocuments confeccionats amb llenguatges d'etiquetatge que expressen la forma en què els navegadors han de presentar el contingut: colors, alguns aspectes de la maquetació, fonts, etc., i no pas el significat o la semàntica. A causa del gran i creixent nombre d'aquests recursos, els actuals motors de cerca són incapaços d'oferir taxes de precisió mínimament adequades en resultats, cosa que evidencia que les tècniques lexicoestadístiques no poden solucionar pel seu compte la problemàtica de la recuperació de la informació (Peis et al., 2003). De moment, els motors de cerca són molt eficaços per localitzar informació coneguda, però no ho són tant per descobrir nous recursos informatius la cerca dels quals pot produir un gran nombre de registres recuperats. Entre els registres recuperats poden ser molt pocs els rellevants, i no existeix la possibilitat de saber quants registres rellevants s'han deixat de recuperar (Pinto et al., 2002, p. 201).

Això és degut al fet que quan se cerca amb un motor no es passa per un procés de filtratge semblant al que utilitzen les bases de dades que contenen metadades o registres descriptius del contingut del document. Això genera una gran quantitat de soroll i de pèrdua d'informació molt habitual en aquesta mena de tecnologies. Per això, en un motor de cerca no es disposa del nivell de representació de la informació que permeti prendre decisions sense la necessitat d'examinar seqüencialment tot el conjunt de documents originals. Si únicament estem interessats a consultar documents de determinades característiques formals, no tenim més remei que examinar un per un tots els documents obtinguts tant si la llista és curta com si és molt llarga.

La major carència actual d'Internet és un sistema universal d'etiquetatge, de representació i d'estructuració de la informació que permeti la cerca i el processament automàtic més adequat a qualsevol tipus de pàgina web (Pinto et al., 2002, p. 203). Aquesta mancança s'intenta resoldre mitjançant el llenguatge XML (sigla que prové de l'anglès *extensible markup language*, 'llenguatge d'etiquetatge extensible') i la creació del que s'anomena «web semàntica», amb l'objectiu que les màquines siguin capaces de «comprendre» el contingut dels documents. D'aquesta manera, en el procés de recuperació de la informació l'usuari interrogaria un agent software que efectua tasques complexes d'associació i d'inferència de coneixement, i torna a l'usuari resultats precisos i contextualitzats. Per aconseguir-ho cal proporcionar semàntica a la web, i aquesta feina es podria dur a terme mitjançant l'elaboració d'ontologies i d'etiquetatges descriptius. Però no tothom hi està d'acord, ja que, malgrat els beneficis potencials que reportaria aquest ús pel que fa a la recuperació d'informació, la seva elaboració és molt més complexa que la de les actuals pàgines basades en el llenguatge HTML (Codina, 2003a). Un dels principals promotors de la web semàntica és Tim Berners-Lee, creador del WWW, que n'impulsa l'ús des del World Wide Web Consortium (W3C).[42]

8.4.5 El llenguatge de consulta

És la part del programa que agafa les preguntes de l'usuari, de vegades en llenguatge gairebé natural, i recorre l'índex de la base de dades per seleccionar-ne els documents més rellevants. La part més important del sistema de consultes és el llenguatge de cerca, mitjançant el qual l'usuari pot expressar —millor o pitjor— la seva necessitat d'informació, establir diversos filtres o condicions, limitar el nombre de respostes, revisar la pregunta i navegar per la llista de resultats.

[42] Es pot consultar a: http://www.w3.org/Consortium/

El rendiment global dels servidors de cerca depèn de tres aspectes:

1. L'eficàcia del robot per descobrir nous documents i mantenir la llista d'adreces.
2. La quantitat d'informació que guarda el fitxer invertit de cada document.
3. La versatilitat i la potència del llenguatge de cerca.

En concret, un robot pot ser molt exhaustiu i tenir ben localitzats el 90% o més dels documents publicats a Internet, però si la base de dades guarda poca informació respecte a aquests documents o el llenguatge d'interrogació és poc flexible, el conjunt com un tot proporciona un rendiment baix. En general, el llenguatge de consulta és, amb diferència, la part més dèbil dels cercadors actuals.

El tipus de consulta que es fa a una base de dades no pot ser molt àmplia, ja que en aquest cas es pot donar «soroll documental», que consisteix en la recuperació de documents no significatius en el context de la demanda; al contrari, en delimitar molt una cerca existeix el perill de no recuperar aquells documents que hi ha a la base i que són significatius, cosa que dóna com a resultat «silenci documental» (Bellveser et al., 1999).

8.4.6 Possibilitats de consulta als cercadors

Bàsicament hi ha dues modalitats de consulta: la simple i l'avançada.

- *Consulta simple*. Es fan cerques per paraules i/o frases sense relació entre elles. És la més emprada i la que s'ofereix per defecte quan s'entra en qualsevol cercador. Si es vol assegurar una taxa alta d'exhaustivitat, cal acumular tots els sinònims coneguts i entrar-los en la finestra de cerca separats per comes. Internament, de forma transparent a l'usuari quedaran vinculades per l'operador lògic OR: el sistema de consulta busca tots els documents que contenen una, dues o la totalitat de les paraules de la cerca.

  Generalment permet fer servir alguns operadors que solen ser comuns a tots els cercadors. Són operadors de requerir/excloure, cometes o truncaments. Quan es llança una pregunta de tipus frase (qualsevol combinació de paraules es considerada una frase) s'ha de posar entre cometes per delimitar-la.

- *Consulta avançada*. Facilita les mateixes opcions que la simple, a més d'incloure l'ús d'operadors booleans i altres possibilitats com controlar el rànquing o l'ordre dels resultats, l'elecció de la data i de l'idioma de les pàgines. Així es pot fer una cerca més precisa i amb menys soroll documental. Cal tenir en compte que la possibilitat d'obtenir molt soroll documental és molt elevada quan es consulten els cercadors de caràcter general, ja que emmagatzemen tota mena d'informació.

Molt sovint els cercadors no fan una distinció de caràcters (majúscula/minúscula, accents, dièresis, *ñ*, *ç*, etc.), de manera que, si no s'especifica, recuperen tots els documents independentment de la manera com s'hagi escrit la paraula.

Les operacions habituals que es poden efectuar en una cerca, a més de fer servir la paraula clau, són les següents (Montes, 1999; Codina, 1997b):

- *El truncament*. Es fa servir per recuperar termes que tenen una arrel semàntica comuna. Es representa amb el símbol (*) després de la posició de truncament de la paraula, i en alguns casos el símbol és ($) o (?); és una opció específica que es pot activar. És útil quan no se sap exactament com està indexada una paraula, també en el cas de singular/plural i de termes derivats. Alguns cercadors ho fan de manera automàtica. Si no es vol truncar una paraula cal afegir-hi un punt al final (.) o tancar el terme entre cometes (“ .”), depenent del cercador. El truncament pot ser, segons els sistemes, tant per la dreta (recuperació mitjançant l’arrel de la paraula) com per l’esquerra (recuperació de totes les paraules que tenen la mateixa terminació), o intercalant-lo dins la paraula. Per exemple, si s’escriu la paraula clau «biblio*ia» buscarà pàgines relacionades amb biblioteconomia, bibliografia, bibliometria, bibliotecologia, etc.

- *Operadors de requerir o excloure*. Es representen amb els signes (+) i (-). La presència indica l’obligatorietat de l’aparició d’un terme en tots els documents recuperats i la manera més habitual d’indicar aquesta condició és precedir el terme afectat del signe d’addició (+). De la mateixa manera, l’absència indica que el terme en qüestió no ha d’estar inclòs en cap dels documents recuperats i s’expressa precedint el terme afectat del signe (-). Funcionen de manera semblant als operadors booleans NO i Y, però no permeten agrupar termes emprant parèntesi. En alguns cercadors se substitueixen per text en forma d’opcions de menú semblants a: *must, contain/should, contain/must, not contain*, i normalment apareixen en l’opció de cerca avançada.

- *Operadors booleans*. També s’anomenen «operadors lògics», i serveixen per relacionar els termes de cerca. N’hi ha tres: AND (intersecció), OR (unió) i NOT (negació). La intersecció serveix per controlar que els documents recuperats continguin tots els termes de la cerca introduïts a l’equació. La unió porta a la recuperació de documents que continguin tant una de les condicions de cerca com totes les que s’hagin fet. L’exclusió o negació serveix per indicar que determinats termes no han d’aparèixer en cap dels documents recuperats. En una equació de cerca és freqüent que calgui més d’un operador de Boole, i llavors és imprescindible utilitzar parèntesis per acotar l’abast de cada operador.

  Els cercadors d’Internet, per defecte, interpreten sempre la presència de dos o més paraules com un cas d’OR: documents on apareguin qualsevol d’aquestes paraules o totes. Cada cop és més habitual que els creadors dels motors intentin simplificar l’ús de l’eina, i en lloc d’oferir els operadors directament ho fan mitjançant explicacions en el menú.

  Y = tots els termes s’han d’incloure en els documents (*all words*)
  O = alguns dels termes han d’aparèixer en els documents (*any words*)
  NO = els termes no han d’aparèixer en els documents

  Aquesta facilitat d’ús mitjançant menús desplegables té com a inconvenient que es perd la possibilitat d’agrupar entre parèntesis els termes i els operadors per crear subseqüències de cerca.

- *Frases exactes*. Cometes dobles o l'opció *exact phrase* en la consulta avançada d'alguns. És molt útil si es coneix el títol d'obres o de documents, els noms de teoremes, lleis, hipòtesis, objectes, persones, etc., o bé si es té la certesa que una frase concreta forma part d'un document.

- *Operadors de proximitat*. També s'anomenen «operadors sintàctics», i serveixen per limitar l'espai que es vol que hi hagi entre dos termes introduïts. Es poden buscar paraules que estiguin juntes, separades per diverses paraules o caràcters, que es trobin en la mateixa frase o paràgraf, i a més es pot especificar si s'ha de respectar l'ordre en què s'han introduït els termes. Trobem els operadors següents:

  - *ADJ (adjacència)*. Recupera els termes en la mateixa posició i en l'ordre establert. Amb "OADJ" el prefix *o-* indica que cal mantenir l'ordre especificat en la consulta. Amb "FAR" / "OFAR" es limita la cerca a un nombre determinat de termes de distància.

  - *NEAR (proximitat controlada)*. Recupera els termes pròxims però no necessàriament en l'ordre marcat. Es pot indicar el nombre de paraules que pot haver-hi entre l'aparició dels dos termes amb (NEAR/n), o bé "ONEAR" per mantenir l'ordre especificat en la consulta.

  - *WITH (paràgraf)*. Recupera termes de la cerca dins de la mateixa frase.

  - *SAME (camps)*. Recupera els termes de la cerca en el mateix camp.

  També es poden substituir aquests operadors per opcions de menú del tipus «totes les paraules (fins a *n* paraules, en qualsevol ordre) / (fins a *n* paraules ordenades)».

  S'ha de tenir en compte que els operadors booleans i de proximitat s'han d'escriure en majúscula en la majoria dels cercadors.

- *Operadors de comparació*. També s'anomenen «numèrics», i serveixen per indicar al sistema de cerca que els documents recuperats han de contenir una data anterior, igual o posterior a la indicada (>, <, =). Es fan servir per restringir les cerques per data de creació o d'última actualització de les pàgines web. Normalment, en les cerques s'utilitzen conjuntament els operadors lògics amb els sintàctics, o també amb els numèrics.

- *Ajudes de cerca*. Fan suggeriments respecte als termes de cerca. Molt sovint apareix l'opció «*Refine your search*», «*Search within results*» o «Tornar a consultar». També s'ofereix l'opció de revisar els llocs més visitats o millors en funció de la consulta amb «*Top n most visited sites*» i «*Recommended links*». Amb «*More like this*» i «*Find similar*» en prémer una pàgina, realitza una nova cerca el resultat de la qual seran documents de temàtica semblant a la pàgina activada. Finalment, amb «*This site only*» es mostren totes les pàgines, limitades a la consulta feta i que es deriven d'aquesta adreça.

- *Delimitadors*. Ofereixen la possibilitat de consultar les URL pels codis que es fan servir en les pàgines HTML, anomenats també «etiquetes».

Hi ha dues formes de fer-ho: escrivint el *comando* seguit de dos punts i el text que es consulta sense espais en blanc, per exemple *title*:video; o seleccionant l'opció més adequada d'un menú del cercador mateix. Les possibilitats són les següents:

- *Title*. Troba els documents que tinguin el text especificat en el títol o en la capçalera.
- *Anchor*. Localitza documents que continguin el text consultat en un hiperenllaç de la pàgina.
- *Link*. És semblant a l'anterior però especificant l'adreça concreta en lloc del text. Pot ser útil per saber quantes pàgines té un enllaç determinat a un lloc web.
- *Host*. Localitza totes les pàgines publicades per alguna empresa, organisme o institució consultant-ne l'URL.
- *URL*. Localitza pàgines que contenen en el seu URL el text consultat.
- *Domain*. Localitza les pàgines que contenen el text especificat en el domini del seu URL. Aquests dominis poden ser els dels països: .es, .it, .uk, etc., o bé els que s'utilitzen per designar organismes governamentals (.gov), educatius (.edu), comercials (.com), etc.
- *Image*. Localitza les pàgines que tenen alguna imatge amb un nom que coincideixi amb el de la consulta.

Generalment, les funcions avançades que permeten formes d'interrogació més complexes que la simple juxtaposició de paraules solen estar mig amagades per no intimidar l'usuari. Així, és possible convertir necessitats d'informació en preguntes ben plantejades, tot i que cap de les interfícies d'usuari actuals dels cercadors d'Internet permeten plantejar preguntes amb tota l'eficàcia necessària. Cal conformar-se amb alguna versió dèbil de la pregunta, en concret amb versions lligades a alguna de les moltes estructures superficials amb què es pot expressar aquesta necessitat d'informació.

- *Serveis de valor afegit*. A més de les opcions d'interrogació anteriors, hi ha aquests serveis, que poden ser de diversos tipus:
  - *Directori temàtic*. No tots en tenen, però és una pràctica cada cop més habitual oferir una selecció de documents organitzada en matèries.
  - *Personalització de les preferències en les opcions de consulta*. Va bé quan es fa servir freqüentment un cercador.
  - *Sistema d'ajuda sobre el propi motor*. Informa sobre la manera de fer correctament les cerques.
  - *Servei de pàgines blanques*. És una eina de localització de persones.

- *Servei de pàgines grogues*. És una eina de localització d'empreses.

- *Servei de notícies d'actualitat*. Recopila les notícies del dia, que es posen a la disposició dels usuaris des de la pantalla principal.
- *Últimes novetats a la xarxa i pàgines d'especial interès*. S'anomenen «*cool sites*» o «Top 5%», segons el cercador.

- *Mapes de llocs*. Guies de viatges, serveis de compres, reserves, etc.

- *Informació especialitzada*. Cotitzacions de borsa, meteorologia, horòscop, resultats esportius, turisme, etc.

8.5. Internet i el WWW

Quan es va imposar la flexibilitat del protocol HTTP i el llenguatge HTML de les pàgines web, en poc temps gairebé tota la informació disponible a Internet es va bolcar al WWW. Hi ha milions de documents accessibles mitjançant aquest sistema d'emmagatzematge d'informació, i són els cercadors i els directoris les eines capaces de localitzar informació sobre un assumpte proposat per l'usuari. Malgrat la gran quantitat d'informació que es pot trobar a Internet, l'organització anàrquica de la xarxa causa alguns problemes:

- Absència d'una llista exhaustiva de recursos per a temes concrets.

- Distribució irregular dels continguts. En alguns termes es pot trobar un excés d'informació i, en canvi, en d'altres la informació recuperada pot ser insuficient o fins i tot inexistent.

- Manca de control de qualitat del material recuperat. Qualsevol persona pot posar informació de tot tipus sense contrastar.

- No hi ha cap sistema d'indexació de continguts d'una manera estandarditzada.

- Les URL apareixen i desapareixen constantment. De moment, els canvis en les adreces dels ordinadors no són introduïts en els cercadors. El missatge «URL not found» comença a ser habitual quan se cerca a Internet.

8.5.1 Internet invisible o Infranet

La mida de l'anomenada «Internet dels continguts» i sobretot de l'espai web és un tema que es debat contínuament i que ha estat objecte de diversos treballs en els últims anys, ja que aquesta informació resulta clau per entendre la importància d'Internet com a font d'informació (Lawrence i Giles, 1999). En l'actualitat, els motors per feines cobreixen del 20 al 25% del total de l'espai web. Es denomina «Internet invisible» o «Infranet» (Cornella, 2000) el conjunt de recursos accessibles únicament a través d'algun tipus de passarel·la o formulari web, i que per tant no poden ser indexats de forma estructural pels robots dels cercadors.

El volum d'informació que això representa ha augmentat notablement durant els darrers anys, tant en termes quantitatius com qualitatius, i podria ser milers de vegades superior a tots els documents en format HTML. El motiu de la seva existència és estructural i no per falta de connectivitat. Molts administradors de webs impedeixen l'accés de robots per evitar problemes de saturació o per raons de confidencialitat. La major part de la informació de la Internet invisible està dipositada en bases de dades que constitueixen un dels sistemes d'accés a informacions altament estructurades i, en general, de qualitat (Maldonado, 2001, p. 161-178).

Segons Alfons Cornella, l'any 2000 es considerava, en l'últim estudi fet fins aquell moment, que si a la Internet de superfície hi havia uns 2.000 milions de pàgines, a la Internet profunda n'hi havia 500.000 milions de possibles. Quatre anys més tard, Ricardo Baeza-Yates afirmava que la web tenia almenys uns 4.000 milions de pàgines estàtiques i un nombre cent cops més gran de dinàmiques (les que es creen producte d'un clic o d'una consulta en un lloc web), a les quals calia afegir tota la web invisible, Intranet o pàgines amb accés restringit (Baeza-Yates, 2004). L'estructura de la web és complexa i evoluciona amb el temps. Hi ha sectors altament connectats i illes que només coneixen alguns cercadors. Un espai fosc és el conjunt d'accessos tancats als cercadors i als usuaris per *logins* i *passwords*. Per tant, es recupera el que els diferents llocs web posen a lliure disposició. La regió amb ratlles correspon a la zona on es mouen els cercadors web, que es correspon en una gran part a la zona pública estàtica i una mica a la dinàmica.

En l'actualitat s'estima que el nombre de pàgines amb informació semàntica, és a dir, la relacionada amb el contingut, constitueix menys del 5%, encara que s'espera que augmenti. Aquesta informació semàntica, basada en metadades, no és molt emprada, ja que existeix un tant per cent molt alt de pàgines que contenen informació no fidedigna o amb finalitats publicitàries.[43]

Tal com comenta Alfons Cornella, allò que es coneix d'Internet és el que està referenciat en algun cercador que, a través dels recorreguts automàtics que fa, busca noves pàgines i les col·loca a la seva base de dades. El problema és que aquests motors de cerca només poden veure la part oberta d'Internet, i no poden accedir a la informació que està darrera les bases de dades tradicionals, a les quals s'accedeix mitjançant un *password*, ja que sovint són de pagament. Trobar informació rellevant, és a dir, que correspongui a les nostres necessitats específiques, és cada vegada més difícil. Encara que es tinguin sistemes de cerca més potents sempre mostraran una petita part del que conté la web. Això suggereix que en el futur s'haurà d'anar cap a cercadors menys exhaustius, però que es concentrin a trobar el millor i el més adequat a una cerca específica.

Malauradament, la majoria de la informació interessant amb què s'acaba treballant és informació que es troba per casualitat. Això en anglès s'anomena *serendipity*, i expressa la idea que sovint ens topem amb informació sense voler, i saber manejar-la és, en aquests moments, un art. Trobar una informació, guardar-la i recuperar-la passat un temps per fer-la servir és una habilitat personal que costa molt d'adquirir, i s'està estudiant com convertir-la en un sistema automatitzat (Cornella, 2000).

[43] En anglès es coneix amb el nom de *spamming*. Consisteix en la tramesa indiscriminada d'un missatge a un gran nombre de bústies electròniques, generalment amb finalitats publicitàries.

L'elecció entre motors i directoris —i, dins de cada grup, la inclinació sobre algun d'aquests en concret— s'ha de fer coneixent el que se'n pot esperar. Amb els directoris, l'usuari no necessita formular amb termes la seva consulta sinó que n'hi ha prou que seleccioni els temes més afins a les seves necessitats. Aquest sistema assegura que tots els documents recuperats tractaran efectivament sobre el tema en què s'han classificat, però probablement se n'hauran perdut molts altres no seleccionats pels responsables del directori.

Amb els motors de cerca es poden fer consultes emprant les opcions pròpies dels sistemes de recuperació d'informació (RI). Si l'equació de cerca és correcta, en la majoria de casos el nombre de documents recuperats és excessiu. L'ordenació per rellevància ajuda l'usuari a seleccionar els documents que llegirà en primer lloc. En principi, ni els uns ni els altres són la solució ideal per assegurar una correcta i completa recuperació d'informació en el web. Segons el que es busqui, pot ser útil una combinació de tots dos (Marcos, 1998a). Per a una consulta molt concreta pot anar bé un motor, ja que fa una cerca exhaustiva en el web amb poca possibilitat de recuperar documents no rellevants. Per buscar informació sobre un tema general, convé començar la cerca per un directori, ja que proporciona documents rellevants sobre el tema i en les pàgines consultades es poden trobar enllaços a altres amb informació relacionada.

8.6. Tipologia i rànquing de cercadors

L'accés remot a la informació i la seva recuperació de manera selectiva i automatitzada són dos factors determinants per aconseguir una gestió més eficient de l'explosió d'informació que estem vivint. L'increment de la producció documental és estimulat per l'accés fàcil, selectiu i a distància proporcionat per Internet, que escurça el cicle de producció dels investigadors i augmenta la velocitat de creació de nous documents (Rovira, 1998, p. 96).

L'únic que creix més ràpid que Internet són els mateixos serveis de cerca. La llista de robots i de serveis de cerca i recuperació d'informació a Internet és extraordinàriament llarga i no para de créixer. Cada motor té unes característiques a l'hora d'emmagatzemar, indexar, recuperar i presentar la informació recollida pel robot, que el fan més o menys adequat com a eina de cerca i recuperació documental.

N'hi ha massa per conèixer-los i fer-los servir tots. Malgrat el nombre elevat, la majoria de consultes es fan en un nombre molt petit. Mitja dotzena concentra el 90% de la utilització, tant per raons d'eficàcia com d'amplitud de la cobertura.

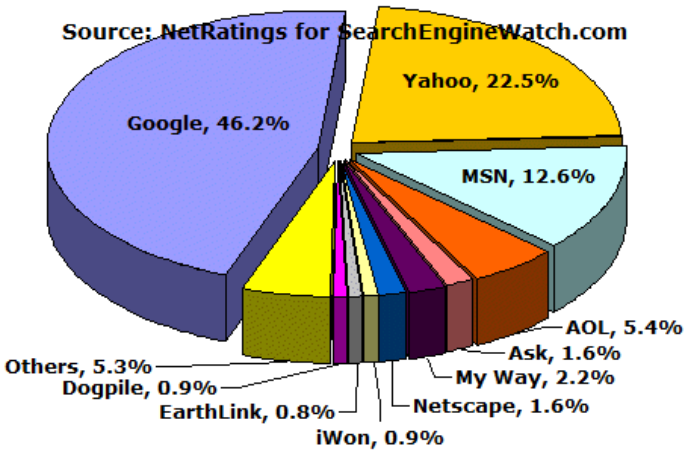
Segons un estudi estadístic fet per OneStat[44] fet el gener del 2003 sobre l'ús dels diversos cercadors, els set principals eren: Google (54,7%), Yahoo! (22,1%), MSN Search (9,5%), AOL Search (3,7%), Terra Lycos (2,8%), Altavista (2,5%) i Askjeeves (1,5%). El juliol del 2005, en un estudi fet per Nielsen//NetRatings,[45] els tres primers eren: Google (46%), Yahoo! (22%) i MSN (13%).

---

[44] OneStat. Number One Real-Time Intelligence Web Analytics. Es pot consultar a: http://www.onestat.com-/html/aboutus_pressbox17.html

[45] Nielsen//NetRatings: http://www.nielsen-netratings.com/. Es pot consultar a: http://direct.www.nielsen-netratings.com/pr/pr_050824.pdf

Al gràfic 8.6 es mostren les dades corresponents a les posicions del rànquing de cercadors, en l'últim estudi fet el juliol de 2005 als Estats Units.[46] Es pot veure que Yahoo! continua en segona posició darrere de Google. Yahoo!, que inicialment era únicament un directori, ha anat adquirint altres cercadors (Overture, Altavista, Alltheweb) per poder oferir opcions de cerca similars a Google.



Gràfic 8.6. Rànquing de cercadors. Juliol 2005.

Els cercadors actuals es poden catalogar tenint en compte diverses característiques:

- Respecte a la manera de fer les cerques:
  - *Motors de cerca*. Són els que cerquen per paraules.
  - *Índexs o directoris*. Són els que mostren la informació mitjançant índexs jeràrquics.
  - *Mixtos*. Són els que combinen els dos tipus anteriors. Cada vegada és més comú que els motors de cerca ofereixin un directori temàtic (elaborat per ells mateixos o comprat a algú altre), i els directoris fan servir sistemes de cerca a través del directori mitjançant paraules clau.
  - *Metacercadors*. Permeten llançar cerques simultànies a diferents cercadors. Tenen com a inconvenient el fet que generalment no permeten obtenir tota la potència de cerca de cada eina de cerca per separat, la qual cosa obliga l'usuari a fer una cerca per defecte o una consulta avançada però amb menys opcions (Martínez et al., 2001). Alguns efectuen la cerca simultàniament a les diferents eines de cerca amb les quals manté enllaç i altres només envien la pregunta a l'eina que se seleccioni.

[46] SearchEngineWatch. Es pot consultar a: http://searchenginewatch.com/reports/article.php/2156451

- *Multicercadors*. Són una mena de metacercadors, que proliferen en les pàgines personals perquè són molt fàcils de crear. Es basen en la idea de copiar la finestra de cerca de la pàgina principal de cada cercador en una mateixa pàgina web, a través de la qual s'accedeix directament a qualsevol cercador

- *Anells web*. Són un tipus especial d'índex temàtic molt útils per efectuar cerques ràpides i poc exigents sobre determinats temes. Són conjunts de pàgines web enllaçades que tracten un mateix tema organitzades mitjançant índexs. No admeten les cerques per operadors booleans, només per termes.

➢ Per àmbit geogràfic:

- *Globals*. Indexen tots els recursos accessibles en qualsevol ordinador del món.
- *Regionals*. Limiten la indexació de recursos a àmbits d'un país i a àrees lingüístiques.

➢ D'acord amb el contingut:

- *Generalistes*. També se'ls anomena «cercadors universals».

- *Selectius o temàtics*. Solen ser molt semblants als cercadors per índexs o per categories, però en les seves bases de dades només tenen informació sobre un tema concret. Són grans compiladors dels recursos existents a Internet sobre un tema específic, i ofereixen millors resultats que els cercadors no especialitzats o generalistes.

- *Assistents automàtics per a cerques*. Són programes especials que s'instal·len als ordinadors clients específics per realitzar cerques programades per l'usuari (*Intelligent Research Assistant*). Es connecten automàticament a Internet, realitzen les cerques especificades prèviament, desen els resultats i es desconnecten quan han acabat.

La presentació dels resultats també pot diferir d'un cercador a un altre. Les dades que solen aparèixer són: el títol de la pàgina, la direcció, el tant per cent de rellevància respecte a la consulta, la mida del document, un resum i els descriptors si el document té etiquetes meta, o si no n'hi ha les primeres línies del text. Alguns ofereixen una versió reduïda on consta únicament el títol de la pàgina, l'enllaç, una línia de resum, i a vegades el percentatge de rellevància.

L'ordre en què es mostren els resultats es fa generalment en funció de la rellevància que el motor assigna als documents segons la freqüència amb què apareixen els termes de la cerca. Cada cop es té més en compte en quina part del document apareixen els termes i es dóna més importància a aquells que els contenen en els camps del títol i resum, en els primers paràgrafs i en les capçaleres.

També es pot variar el nombre total de resultats que es mostren en una pàgina. N'hi ha que ofereixen una llista de termes relacionats amb els que s'han introduït, això proporciona opcions de cerca que potser a l'usuari no se li havien acudit.

8.7. Resultats i discussió.

8.7.1 Ús d'un cercador genèric: Google

S'ha escollit el motor de cerca Google com a primera opció per determinar el contingut de documents audiovisuals en el web a partir de les dades exposades en l'apartat 8.6, en què se'l considera el motor de cerca amb més quantitat de recursos revisats i un cercador pur que no inclou publicitat ni altres tipus de serveis en fer una consulta.

Va ser creat el 1998 per Larry Page i Servey Brin, dos estudiants de doctorat de la Universitat de Stanford. *Google* és un joc de paraules amb el terme *googol*, expressió creada per Milton Sirotta —nebot del matemàtic nord-americà Edward Kasner— per referir-se al nombre representat per un 1 seguit de 100 zeros. L'ús del terme per part de Google reflecteix l'objectiu de la companyia d'organitzar la immensa quantitat d'informació disponible a Internet.[47] A l'inici del 2004 el nombre de pàgines web revisades pel cercador era de prop de 8.000 milions. Per tant, s'ha considerat com una bona opció per cercar documents audiovisuals de química a Internet, concretament a la web.

La innovadora tecnologia de cerca de Google i el seu elegant disseny d'interfície d'usuari el diferencien de les màquines de cerca de primera generació. Fa servir la tecnologia *PageRank*, que assegura que els resultats més importants es mostren en primer lloc. Els creadors afirmen que està estructurat de manera que ningú pot comprar un lloc privilegiat en la llista ni alterar els resultats amb finalitats comercials, i per tant les cerques fetes es poden considerar honestes i objectives.

Google presenta dues pantalles de cerca: la simple, que s'ofereix per defecte tal com es mostra en la figura 8.7.1.1 i que constitueix la pantalla d'inici del cercador, i l'avançada, a la qual es pot accedir a través de la pantalla principal.

Figura 8.7.1.1. Google. Pantalla principal.

[47] Informació corporativa. Es pot consultar a: http://www.google.es/intl/es/corporate/

Dins les opcions bàsiques hi ha la possibilitat de cercar al web, o bé a imatges, a grups, a directoris o a notícies. Permet ajustar la cerca a pàgines que estiguin només en espanyol o bé ubicades en servidors d'Espanya, així com escollir l'idioma de consulta: català, gallec, euskara, etc. La funció «*Voy a tener suerte*» porta directament a la pàgina que més s'adapta a la cerca.

Google és considerat un cercador pur, ja que no ofereix altres tipus de serveis com ara correu electrònic, notícies, xat, serveis de compra, etc., i pot concentrar els esforços per millorar les tècniques de cerca i d'avaluació de les webs trobades. Els formats de documents dels quals es pot limitar una cerca són Adobe Acrobat (.pdf), Adobe Postcript (.ps), Microsoft Word (.doc), Microsoft Excel (.xls), Microsoft Powerpoint (.ppt) i en format de text enriquit (.rtf), tal com es mostra en la figura 8.7.1.2.

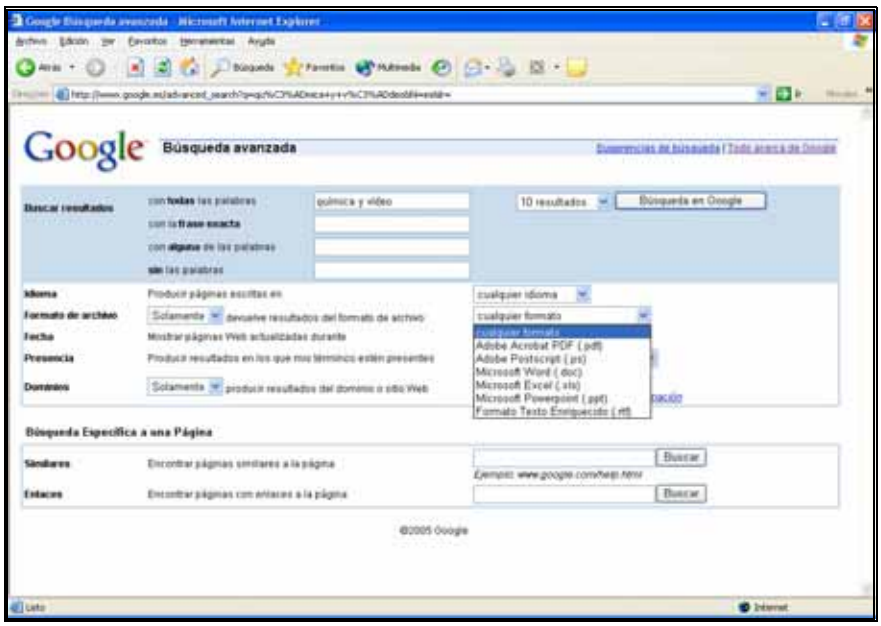


Figura 8.7.1.2. Google. Pantalla de cerca avançada.

La cerca «química» proporciona 2.390.000 resultats, cosa que sembla lògica donada la poca selectivitat del terme emprat en la cerca. En fer una cerca per «química y vídeo» s'obtenen 256.000 resultats, que continua sent un nombre massa elevat per poder revisar els documents.

La immensa majoria de resultats (se n'han arribat a revisar els 200 primers) obtinguts corresponen a documents textuals que parlen de vídeos però que no faciliten imatges de vídeos ni accés a cap vídeo.

La cerca és susceptible als accents i a l'ordre en què s'introdueixen els termes en la consulta, tal com es pot veure en les figures 8.7.1.3 i 8.7.1.4, on es mostren les pantalles corresponents a la primera pàgina de resultats corresponents a les cerques efectuades per «vídeo y química» i «video y quimica».

Figura 8.7.1.3.Google. Cerca «vídeo y química»: 399.000 resultats.

Figura 8.7.1.4. Google. Cerca «video y quimica»: 173.000 resultats.

La cerca «*chemistry*» llançada a tot el web produeix 27.100.000 resultats, alguns dels quals corresponen a adreces que contenen documents audiovisuals però que es troben submergides enmig d'un munt de pàgines que podríem qualificar de textuals. Aquest nombre tan elevat és degut a la poca discriminació de la cerca, amb una mescla de significats del mot. S'hi poden trobar fins i tot vídeos de sexe, ja que tot sovint sota la paraula *química* s'inclouen obres d'aquest tipus.

Una cerca de «*chemistry and video*» rebaixa la quantitat a uns 5.040.000 resultats, però amb una total mescla de gèneres, formats i significats de mot. Majoritàriament corresponen a pàgines web que no contenen vídeos i en les quals es poden trobar documents on es parla de l'ús del vídeo, de la creació d'algun vídeo, de catàlegs de vídeos, d'ofertes de compra, etc. En la figura 8.7.1.5 es mostra la primera pàgina de resultats de la cerca «*chemistry and video*».

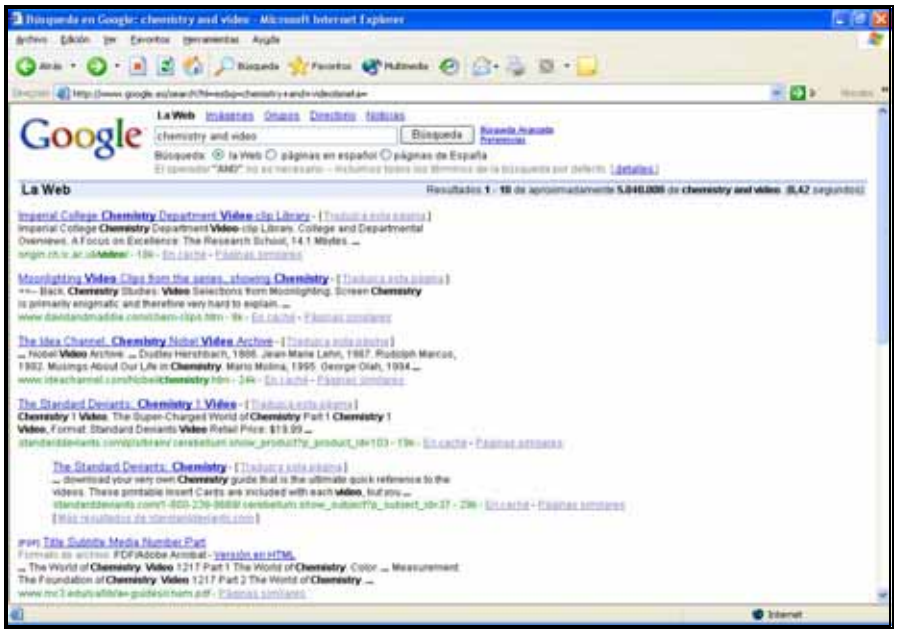


Figura 8.7.1.5. Google. Cerca «chemistry and video»: 5.040.000 resultats.

Evidentment, no és possible revisar aquesta quantitat de resultats per veure què es troba i esperar que, a més, s'adeqüi a les nostres necessitats acadèmiques.

Per tant, Google, que és un molt bon cercador per a documents basats en text i imatges, no és apropiat per efectuar cerques de documents audiovisuals. Per això es va plantejar la necessitat d'estudiar altres cercadors per poder determinar si n'hi havia cap que permetés localitzar documents audiovisuals en el web, en algun dels formats que es descriuen en la taula 9.3.2 de l'apartat 9.3 d'aquesta tesi.

8.7.2 Localització de cercadors

Partint dels resultats obtinguts en l'apartat 8.7.1 ha estat necessari conèixer quins cercadors existien actualment dins de les tipologies descrites en l'apartat 8.6. Per localitzar els cercadors s'han emprat diverses fonts:

- *Consulta a les biblioteques universitàries espanyoles*. Solen contenir algun recull de cercadors per orientar els usuaris respecte als que poden emprar. Es van consultar els reculls de les biblioteques següents:
  - Biblioteca de la Universitat de Barcelona.
  - Biblioteca de la Facultat de Farmàcia de la Universitat Complutense de Madrid.

- *Publicacions*. Hi ha diversos llibres que expliquen el funcionament d'Internet que contenen algun capítol dedicat a com es realitzen les cerques i donen informació sobre alguns cercadors. Entre les obres que s'han consultat hi ha les següents:
  - Ángeles Maldonado (coord.) (2001). *La información especializada en Internet. Directorio de recursos de interés académico y* profesional. Madrid: CINDOC (CSIC).
  - Miguel Ángel Cazorla, Otto Colomina, Patricia Compañ (1999). *Internet para universitarios*. Universitat d'Alacant.
  - Francisco José Maldonado, Paula Luna, Rodrigo Fernández, José Luis Salmerón (2001). *Internet para investigadores. Hacia la e-ciencia*. Universitat de Huelva.
  - Randolph Hock (2001). *The Extreme Searcher's Guide to Web Search Engines. A Handbook for the Serious Searcher*. New Jersey: CyberAge Books.

- *Cercadors de cercadors*. Empren un cercador especialitzat en la cerca d'altres cercadors. Es va escollir Buscopio[48] perquè conté un recull de cercadors en forma de directori on els manté agrupats en setze categories temàtiques dividides en 642 temes diferents, que contenen un total de 3.073 adreces. Dins la secció «Ciencias» hi ha els cercadors especialitzats en física i química. Únicament conté quatre adreces de cercadors especialitzats en química: ChemExper, ChemFinder, Chemie.De i Chemistry.org. Són de caràcter general i més aviat orientats cap al món empresarial. No n'hi cap especialitzat en la cerca de documents audiovisuals en el web.

[48] Buscopio©. Ricardo Fornas Carrasco: http://www.buscopio.net/

8.7.3 Breu història dels cercadors en general

Les eines de cerca a Internet tenen una molt breu història, menys d'una dècada. Abans que existissin els motors de cerca web hi havia el caos. Si es volia trobar alguna cosa a Internet se n'havia de conèixer l'adreça exacta (Hock, 2001).

La primera etapa significativa des d'aquest caos cap a un cert grau d'organització del contingut d'Internet va ser el desenvolupament dels *gophers*: col·leccions d'adreces d'Internet basades en servidors i ordenades en un format determinat. El terme *gopher* prové de la mascota de la Universitat de Minnesota, de la qual va emergir el primer *gopher*. Els *gophers* no estaven basats en HTML i bàsicament indexaven no molt més que els títols d'arxius i una molt breu descripció, però si sabies com trobar un *gopher* et podia permetre descarregar arxius seleccionats.

Els *gophers* van generar Archie, que buscava *gophers*, i Archie va generar Veronica, que buscava en tot el *gopherespai*. Veronica va originar Jughead. El llinatge *gopher* tenia poc més d'un parell d'anys quan va quedar enfosquit pel ràpid desenvolupament del World Wide Web, que permetia l'ús d'hiperlinks, cerques de text complet, navegadors geogràfics i altres tecnologies fàcils d'utilitzar i altament interactives, així com el desenvolupament dels motors de cerca web.

El primer motor de cerca web que va aparèixer va ser WebCrawler —procedent de la Universitat de Washington—, que va fer el seu debut públic l'abril de 1994. En un any va haver-hi tres competidors en escena: Lycos, Infoseek i OpenText. A finals de 1995 van aparèixer Altavista i Excite. Curiosament, molta, potser la major part, de l'actual tecnologia de cerca d'un investigador seriós està present en diversos graus en aquests primitius motors de cerca, incloent-hi característiques com ara operadors booleans, truncaments, etc.

Desgraciadament, cap va aprofitar la forta tecnologia de cerca que es podia trobar en serveis de bases de dades bibliogràfics tradicionals. A més, cap dels motors de cerca ni dels directoris web van aprofitar l'àmplia teoria i pràctica de classificació temàtica dels últims dos-cents anys. La causa cal buscar-la en el fet que els motors de cerca web van ser i són desenvolupats per al cercador aficionat, i no per als que desitgen ansiosament aprofitar plantejaments i tècniques més sofisticats.

HotBot va sorgir el 1996, Nothern Light l'any 1997. Google, que va aparèixer el 1998, va adquirir ràpidament popularitat tant entre els cercadors aficionats com entre els professionals, gràcies a la combinació del seu rànquing de resultats basat en la popularitat i de l'ús d'una interfície molt simple.

Mentrestant, la cursa per ser el motor de cerca més gran havia amainat fins que l'aparició de Fast Search el 1999 proclamant que tenia una base de dades de 200 milions de registres va reiniciar-la, i el gener del 2000 hi havia quatre motors que havien aconseguit els 200 milions de registres.

OpenText va desaparèixer a començaments de 1998. Segurament hi haurà més desaparicions durant els propers anys i l'aparició almenys d'un o dos motors de cerca grans.

Mentrestant, els canvis continuen, encara que alguns són superficials i es produeixen més com a part de la naturalesa del servei de portal que no pas en la cerca. Com la resta del món dels negocis, les companyies dels motors de cerca són extremadament susceptibles a les modes. Els anys 1999 i 2000 comporten l'aparició d'un corol·lari per al concepte de *portal* més subtil i menys anunciada: la incorporació de directoris. Primer s'inclouen en les pàgines d'inici dels portals amb l'esperança que la gent els faci servir. El 1999 s'incorpora el contingut d'aquests components addicionals a les pàgines de resultats. Les eines que reben l'atenció dels usuaris seran recordades, millorades, copiades i valorades.

El problema és que la gent que els fa servir —l'usuari típic— està poc interessada en característiques més sofisticades i orientades a la cerca. Per fer-ho evident només cal veure algun cercador, per exemple Lycos, que proporciona una interessant però de vegades depriment llista de cerques favorites. En una setmana típica, les 50 millors cerques n'inclouen 46 que són de les categories d'entreteniment, esports o jocs.

Això significa que cal afrontar la realitat que la major font d'obtenció de diners per part dels motors de cerca no és fent servir el web amb finalitats professionals. Les bones notícies és que l'audiència global augmenta, i per tant el nombre de gent que utilitza els motors de cerca amb propòsits professionals, per fer inversions, per augmentar l'alfabetització en tòpics com ciència, humanitats, negoci i medicina, potser també creixerà. Si el nombre de cerques més valuoses intel·lectualment augmenta, els productors de motors de cerca hauran de mantenir i millorar les característiques més serioses.

8.7.4 Estudi de les característiques d'alguns cercadors

S'han revisat un total de 107 cercadors de diversa índole: generals, específics, generals d'Espanya, de Catalunya i metacercadors. En el quadre 8.7.4 es mostra la llista dels cercadors revisats en ordre alfabètic.

S'han classificat segons l'organització temàtica emprada per Martínez i col·laboradors (2001). S'ha estructurat en cercadors per índex temàtic: internacionals, hispanoamericans, espanyols; especialitzats; anells web; cercadors per contingut: internacionals i espanyols; metacercadors; cercadors per aproximació geogràfica: internacionals, espanyols i catalans; cercadors intel·ligents; cercadors especialitzats: cercadors de cercadors, enllaçats a bases de dades d'accés gratuït, química, tecnologia de la informació i Internet, i empreses; portals i diversos: sense qualificar, portals d'empreses, cercadors per a finalitats diverses.

Per constatar la capacitat de cerca respecte al tema d'aquesta tesi es llança la mateixa cerca a tots els que es consideren apropiats i s'observa el nombre de resultats obtinguts, així com la rellevància pel que fa a la localització de documents audiovisuals relacionats amb la química a nivell universitari. A més, per saber amb quina es recuperen més resultats, es fan diverses proves emprant diferents paraules clau: audiovisual, vídeo, animació, multimèdia, etc.

Quadre 8.7.4. Llista de cercadors localitzats i revisats.

Cercador	*Adreça web*	*Característiques del cercador*
Cerca per índex temàtic: internacionals		
Yahoo!	http://search.yahoo.com/ http://www.yahoo.com/	
LookSmart	http://www.looksmart.com/	Comercial
MSN	http://search.msn.com/	
Dmoz - Open Directory Project	http://dmoz.org/	
Amfibi	http://www.amfibi.com/	Directori de l'ODP
Galaxy The Internet's First Searchable Directory	http://www.galaxy.com/ http://galaxy.einet.net/	Directori i cercador LOGIKA Corporation
WebBrain - The Smartest Way to See the Web!	http://www.webbrain.com/	Cercador visual
Cerca per índex temàtic: hispanoamericans		
Vindio	http://www.vindio.com/	
StarMedia-LatinGuía / Wanadoo	http://www.latinguia.com/	Cercador multimèdia
Hispavista	http://www.hispavista.com/	
Ubbi	http://ubbi.com/ http://www.buscador.clarin.com/home.asp	
Cerca per índex temàtic: espanyols		
Yahoo!	http://www.yahoo.es/	
Apali	http://www.apali.com/	
Terra	http://buscador.terra.es/	
Ozú	http://www.ozu.es/	
MSN	http://www.msn.es/default.asp	
Temáticos	http://www.tematicos.com/	
Biwe - El Buscador en Internet en Español	http://www.biwe.com/ http://www.biwe.es	
Eureka	http://www.eureka.creativeweb.es/	
Yupi	http://www.yupimsn.com/	
Telepolis	http://www.telepolis.com/	
Busco.com	http://www.busco.com/	
Vindio	http://www.vindio.com/	
Cerca per índex amb selecció especialitzada		
BUBL	http://www.bubl.ac.uk/ http://www.bubl.ac.uk/link/	
About	http://www.about.com	
University Libraries - University at Albany	http://library.albany.edu/internet/searchnet.html	
Infomine	http://infomine.ucr.edu/	
The WWW Virtual Library	http://vlib.org/Overview.html	
The Argus Clearinghouse The Internet's Premier Research Library	http://www.clearinghouse.net/	
lii.org - Librarian's Index to the Internet Information you can trust	http://www.lii.org/	Universitat de Califòrnia (Berkeley)
Cerca per anells web		
Anillo Web - Pàgina de los Anillos en Español	http://www.anilloweb.com	
WebRing Directory and Online Community	http://www.webring.org	
Cerca per contingut: internacionals		
Altavista	http://www.altavista.com	Yahoo!
Alltheweb	http://www.alltheweb.com	Yahoo!
Excite	http://www.excite.com/	
Northern Light	http://www.northernlight.com/engine.html	
Overture	http://www.overture.com/	Yahoo!
Go	http://www.go.com/	
Google	http://www.google.com	
HotBot	http://www.hotbot.lycos.es/ http://www.hotbot.es/	Lycos, Inc.
Lycos	http://www.lycos.com	
Teoma Search with Authority	http://www.teoma.com/ http://www.directhit.com/	Ask Jeeves Inc.
WebCrawler	http://www.webcrawler.com	
The WhatUseek Network	http://www.whatuseek.com/	Directori i cercador
Wanadoo	http://www.wanadoo.es/	Portal i cercador
Wisenut	http://www.wisenut.com/	LookSmart. Ltd. i cercador
Cerca per contingut: espanyols		
Sol	http://www.sol.es/	
Lycos	http://www.lycos.es	
Google	http://www.google.es/	
Hispavista	http://www.hispavista.com/	
El Inspector de Telepolis	http://www.telepolis.com	
Metacercadors		
Allonesearch	http://www.allonesearch.com/	
Cyber411	http://www.cyber411.com/search/	
Digisearch	http://www.digiway.com/digisearch/	

Dogpile	http://www.dogpile.com/	
Information	http://www.information.com/	Oversee.net Company
Metacrawler	http://www.metacrawler.com/	
Mamma	http://www.mamma.com/	
Metamonster	http://www.metamonster.com/	
Search	http://www.search.com/	
Vivisimo	http://www.vivisimo.com	
Profusion	http://www.profusion.com/	
Webtaxi	http://www.webtaxi.com/	
Cerca per aproximació geogràfica: internacionals		
Encuentrelo	http://encuentrelo.com/	
Cerca per aproximació geogràfica: espanyols		
Dónde? - Directorio Online de España	http://donde.uji.es/	
Cerca per aproximació geogràfica: catalans		
CampusRed	http://www.campusred.net/	Portal Fundación Telefónica
Canal 21	http://www.canal21.com/ http://www.canal21.com/portales/main_cataluniacat.htm	Portal per comunitat autònoma
Catalunya Online - El portal català d'Internet	http://www.catalunyaonline.com/	Portal
Cercat - Lincaweb - El Català al Món	http://www.cercat.com/ http://www.cercat.com/lincaweb/default.htm	Directori de recursos en català
Cercador.com Cercador universal en català	http://www.grec.net/cgibin/cercador.pgm	Interacció Editorial, SL Grup Enciclopèdia Catalana Cercador i directori
La Malla	http://lamalla.net	Diputació de Barcelona Portal
Racó Català	http://racocatala.com/	Cercador de notícies
Som-hi!!!	http://som-hi.com/index.cgi	Cercador i directori
VilaWeb Nosaltres.com - El Cercador de Vilaweb	http://www.vilaweb.com http://www.vilaweb.com/nosaltres/	Diari electrònic Cercador i directori
Cercadors intel·ligents		
AskJeeves	http://www.askjeeves.com/	
Cerca especialitzada: cercadors de cercadors		
Buscopio	http://www.buscopio.net/	
Fossick - The Web Search Alliance Directory	http://www.fossick.com	Fossick.com
Tematicos	http://www.tematicos.com/	
Cerca especialitzada: enllaç a bases de dades d'accés gratuït		
Internet Invisible	http://www.internetinvisible.com/	
Cerca especialitzada: química		
About - Chemistry	http://chemistry.about.com/	Dr. Anne Marie Helsmentine
Alkimistas	http://www.alkimistas.com	
AskaChe - The Ultimate Help Desk	http://www.askache.com/	
Chemedia	http://www.chemedia.com/	
ChemSpy	http://www.chemspy.com/	
ChemWeb	http://www.chemweb.com/	
LearnNet - RSC	http://www.chemsoc.org/networks/learnnet/index.htm	ChemSoc
PSIGate - Physical Sciences Information	http://www.psigate.ac.uk/	Portal científic
Rolf Claessen's Chemistry Index	http://www.claessen.net/	
Scirus - For Scientific Information Only	http://www.scirus.com/srapp/	
Sciseek - Science Online	http://www.sciseek.com/	Pay Per Clic (PPC)
RDN	http://www.rdn.ac.uk/	Internet Resources - Text Only
Eguiame	http://www.esade.es/guiame/	ESADE
Cerca especialitzada: tecnologia de la informació i Internet		
Search Internet	http://search.internet.com/	Jupitermedia Corporation
Cerca especialitzada: empreses		
Me guías	http://www.meguias.com/	Buscador d'empreses
Portals		
Ya.com - Internet Factory	http://www.ya.com/	Grupo T-Online
EducaWeb	http://www.educaweb.com/	Portal privat
Cibereduca	http://www.cibereduca.com/	Portal educatiu
Tiscali	http://search.tiscali.es/	Portal
Tecnociencia	http://www.tecnociencia.es/	MEC / CSIC / FECYT
Rediris	http://www.rediris.es/rediris/	Interconnexió Recursos Inf.
Universia - El portal de 842 universidades	http://www.universia.es/	Portal universitari
Portals: diversos		
FindWhat - Performance Driven Marketing	http://www.findwhat.com/	Empresa
BigFoot	http://www.bigfoot.com/	Empresa
United Learning	http://www.unitedlearning.com/	Venda
411.com	http://search411.com/	Subhastes
eBay - The World's Online Marketplace	http://pages.ebay.com/	Compres

8.7.5 Cercadors de documents audiovisuals

Els cercadors que es van considerar apropiats per localitzar documents audiovisuals en el web són els que mostren únicament adreces en les quals és possible veure un document audiovisual o enllaçar-hi. En el quadre 8.7.5.1 es mostren els cercadors trobats i en el quadre 8.7.5.2, les seves característiques de cerca.

Quadre 8.7.5.1. Cercadors amb els quals és possible realitzar una cerca de documents en format vídeo.

Cercador	*Anagrama del cercador*	*Adreça web*
Alltheweb	alltheweb find it all	http://multimedia.alltheweb.com/?avkw=fogg&cat=vid&cs=utf8&q=&_sb_lang=pref
Altavista	altavista	http://www.altavista.com/
Mamma	mamma The Mother of All Search Engines	http://www.mamma.com/
Wanadoo	wanadoo	http://www.wanadoo.es/
Terra	terra	http://buscador.terra.es/
MSN Search	msn	http://search.msn.es/ http://search.msn.com/
Dogpile	DOGPILE	http://www.dogpile.com/
Metacrawler	metacrawler	http://www.metacrawler.com/
Tiscali	tiscali.search	http://search.tiscali.es/

Amb els cercadors mostrats en la taula 8.7.5.1 es va llançar una cerca amb la paraula clau «chemistry», que era la que proporcionava el major nombre de resultats en les proves efectuades prèviament. El major nombre de resultats es va obtenir amb Alltheweb, Altavista, Terra, Wanadoo, Dogpile i Metacrawler, amb molt poca diferència entre ells, al voltant de 1.100 resultats (cerca feta el març del 2004). El metacercador Mamma proporciona bons resultats però en un nombre inferior als citats anteriorment. MSN i Tiscali donaven una quantitat de resultats molt inferior i molt allunyada dels anteriors, per la qual cosa es van descartar.

Quadre 8.7.4.2. Característiques dels cercadors revisats que permeten trobar documents audiovisuals.

Cercador	*Adreça (data de revisió)*	*Tipus*	*Cercador bàsic*	*Cercador avançat*	*Cerca DAV*
Altavista	http://www.altavista.com/ (9-03-2004)	Cercador Directori temàtic	Web Imagen MP3/Audio Vídeo Directorio	Paraula clau Idioma País Data Domini URL	Sí
Alltheweb	http://multimedia.alltheweb.com/?avkw=fogg&cat=vid&cs=utf-8&q=&_sb_lang=pref (8-03-2004)	Cercador	Web News Pictures Vídeo Audio FTP Files	Paraula clau Idioma Domini IP Media Types File format Data - Size	Sí
Dogpile	http://www.dogpile.com/ (31-03-2004)	Cercador	Web search Yellow pages White pages Web pages Images Audio Multimedia News Shopping	Paraula clau Data Domini Idioma	Sí
Mamma	http://www.mamma.com/ (1-04-2004)	Metacercador	Web News Images	Escollir cercador. Permet refinar la cerca. Suggereix opcions de cerca a partir de la primera cerca feta.	Sí
Metacrawler	http://www.metacrawler.com/ (29-04-2004)	Metacercador	Web Pages Images Audio Multimedia News Shopping	Paraula clau Data Domini Idioma Filtre d'adults Nombre de resultats per cercador	Sí
MSN Search	http://search.msn.es/ http://search.msn.com/ (13-03-2004)	Cercador Directori temàtic		Paraula clau País - Idioma Domini Tipus d'arxiu Media Types	Sí
Terra	http://buscador.terra.es/ (12-03-2004)	Cercador Portal	En español En Internet Compras Noticias Imágenes MP3 Vídeos Sonidos	Paraula clau Domini Idioma Limitar resultats ofensius	Sí
Tiscali	http://search.tiscali.es/ (17-03-2004)	Cercador	España Mundo Imágenes Vídeo MP3 Noticias	Vídeo: AVI - MPEG - Real QuickTime Streaming	Sí
Wanadoo	http://www.wanadoo.es/ (10-03-2004)	Cercador Directori temàtic Portal	Sitios Web Imágenes Vídeos MP3 Noticias Directorio		Sí

DAV = Documents audiovisuals.

8.7.6 Pantalles de cerca i resultats obtinguts amb els diversos cercadors de vídeo

S'ha observat un alt grau de coincidència dels resultats obtinguts emprant els cercadors Alltheweb, Altavista, Terra, Wanadoo, Dogpile i Metacrawler, que sembla que només s'explica pel fet que tots utilitzen la mateixa base de dades. De fet, en la cerca feta amb Metacrawler i Dogpile consta que l'origen dels resultats és el motor de cerca multimèdia Fast. Aquest enginy de cerca pertany a la companyia sueca Fast, que també era la propietària del motor de cerca Alltheweb. El seu gran valor rau en la capacitat de la seva cerca avançada: pàgines web, vídeos, mp3, *news*, etc.

El que difereix d'un cercador a l'altre és la forma de presentació dels resultats. Alguns com Alltheweb, Altavista i Wanadoo mostren una imatge de cadascun dels resultats, la qual cosa facilita enormement el coneixement del contingut, encara que sigui de manera orientativa, mentre que en els altres cercadors la informació presentada és únicament textual.

A continuació es descriuen més detalladament tres dels cercadors revisats que es consideren suficientment representatius dels resultats obtinguts.

a) Alltheweb

L'empresa sueca Fast era la propietària d'Alltheweb, que recentment ha estat adquirit per Yahoo! De moment s'ha substituït la base de dades d'Alltheweb per la de Yahoo!, i s'ha mantingut la de la cerca de documents audiovisuals: vídeo i àudio. Segons Serrano (2003), Alltheweb permet les opcions següents:

1. Cerca avançada de pàgines web. Busca pàgines amb Macromedia Flash, Java Applets, Javascript, Realvideo, Realaudio. Conté filtres de domini, i també per regions: Àsia, Europa, etc.; per IP(sigla que prové de l'anglès *Internet protocol*, 'protocol d'Internet'): si es coneix l'IP d'un servidor es pot buscar dins de totes les webs que hi té allotjades, filtres per dates d'actualització, mida del document, profunditat de la cerca dins del directori.

2. Cerca avançada de notícies: tipus de notícies (econòmiques, esportives, etc.), filtre per domini, ordenació de resultats per rellevància o data.

3. Cerca avançada d'imatges: format (jpeg, gif), tipus d'imatge (en color, escala de grisos, etc.), tipus de fons (transparent o no transparent).

4. Cerca avançada de vídeos: formats (AVI, MPEG, Real, QuickTime, etc.), buscar reproducció en temps real[49] o per descarregar.

5. Cerca avançada d'àudio.

6. Cerca avançada d'FTP: filtres per domini, per mida i per data, fins i tot permet aplicar truncaments (*wildcards*) o definir si busquem paraules amb majúscules o no (*case sensitive*).

[49] En anglès s'anomena *streaming* la reproducció de fitxers audiovisuals procedents d'una xarxa informàtica, sense haver d'esperar que es completi la descàrrega de les dades.

La pantalla de presentació destaca per la simplicitat: és molt càlida visualment i està molt cuidada. Dins la cerca simple només cal introduir una paraula clau i llançar la consulta tal com es mostra en la figura 8.7.6.1.

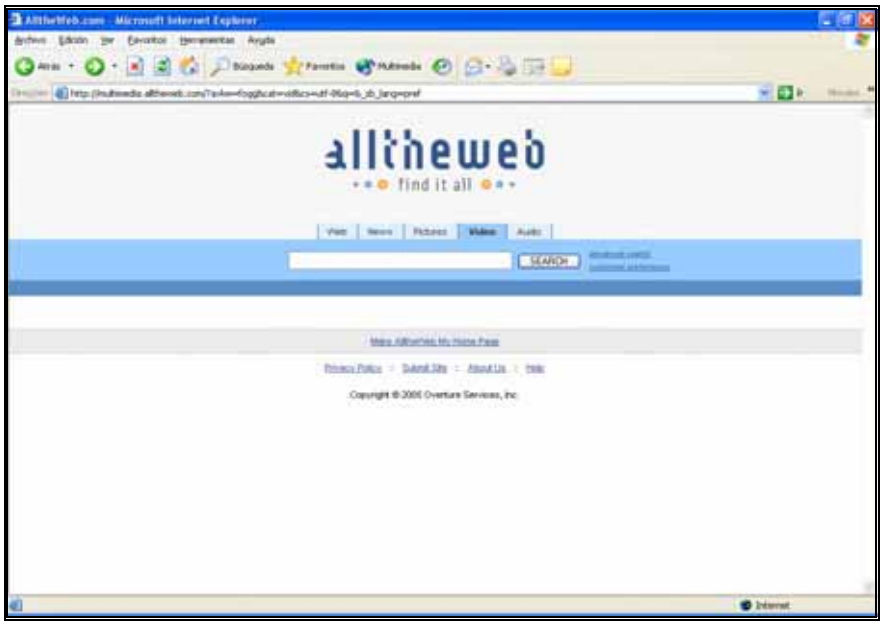


Figura 8.7.6.1. Alltheweb. Pantalla de la cerca simple de vídeo.

La pantalla de cerca avançada varia segons l'opció de cerca escollida. La més complexa i amb més possibilitats correspon a la de cerca de documents en el web. En la cerca avançada de vídeo es pot seleccionar el tipus de format (tots, AVI, AVI/DivX, MPEG, Real, QuickTime), si es vol descarregar o només mirar (*stream*) i l'activació del filtre d'arxius ofensius (On/Off), tal com es mostra en la figura 8.7.6.2.

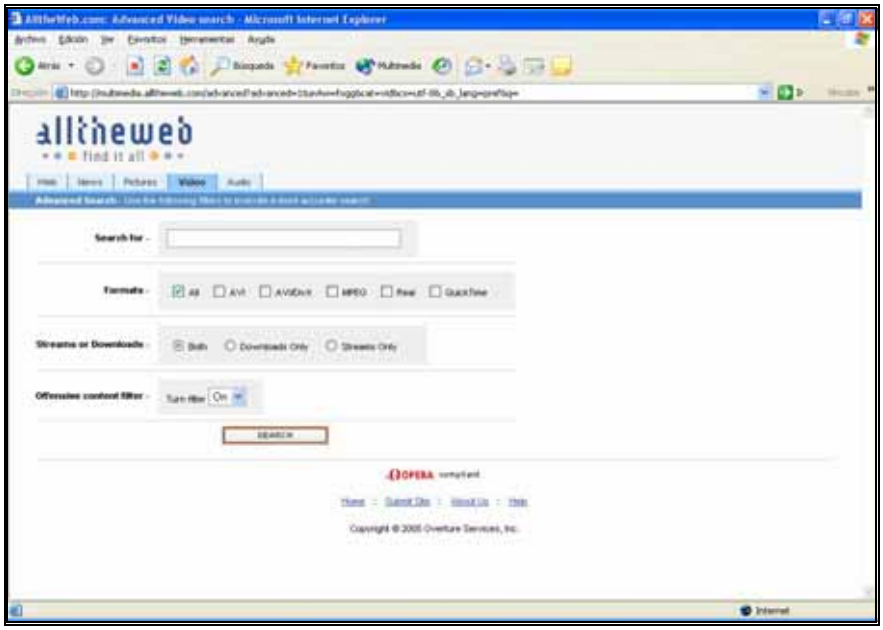


Figura 8.7.6.2. Alltheweb. Pantalla de cerca avançada de vídeo.

De cada resultat en mostra una imatge i en dóna el títol, una petita descripció relacionada amb l'origen del resultat i el vincle per poder-hi accedir, tal com es mostra en la figura 8.7.6.3. També proporciona el temps de durada (hores:minuts:segons), la mida (MB) i el format del document (MPEG, QuickTime, RealMedia, AVI, etc.).

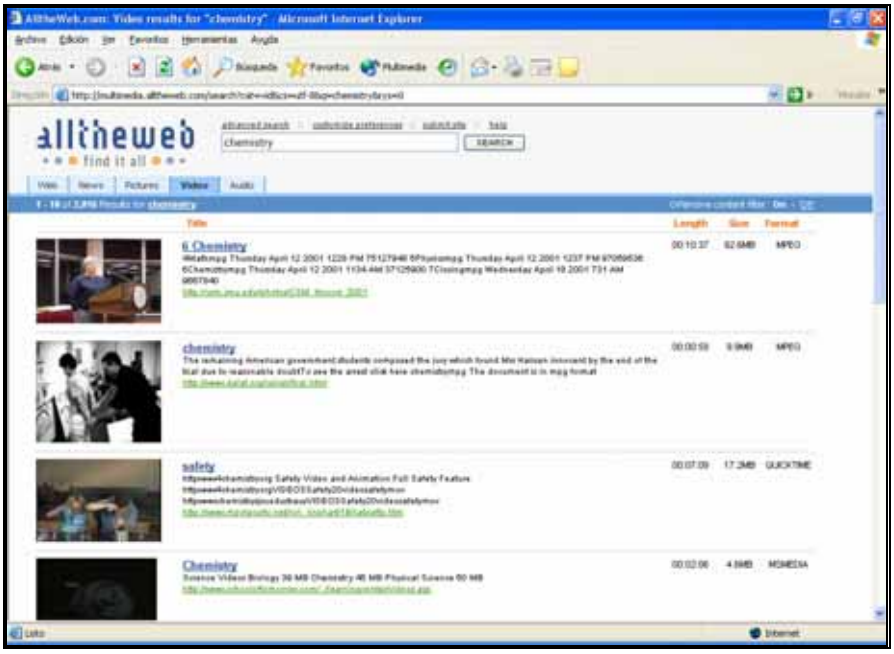


*Figura 8.7.6.3. Alltheweb. Pantalla de presentació dels resultats de la cerca de vídeo: «*chemistry*».*

b) Altavista

Altavista sorgeix el desembre de 1995 de la mà de Digital Equipment Corporation. Genera el major índex de pàgines d'aquell moment i durant uns quants anys manté el lideratge, malgrat el creixement espectacular de llocs web.

A finals de 1998, Altavista i Inktomi competeixen per disposar dels majors índexs. Un any més tard la lluita és entre Altavista, Nothern Light i Alltheweb (Fast Search). La caiguda d'aquests cercadors coincideix amb l'aparició de Google l'any 1998, i és lenta però continuada fins a arribar a la situació actual, en que està entre els deu primers cercadors però a molta distància de Google, que es manté en primer lloc.

Al començament era un cercador simple, després es va convertir en un portal de continguts que va incorporar nombrosos serveis: notícies, fòrums, traductors, directoris, etc., i va tornar a la sobrietat dels seus orígens. Ha anat cap a un model d'èxit, el de Google, o si es vol veure així, al model que exhibia a l'inici. La companyia també ha estat adquirida per Yahoo!

No disposa d'una pantalla específica de cerca avançada de vídeo. Dins de la cerca simple permet escollir el format (MPEG, AVI, QuickTime, Windows Media, Real, altres), així com la durada dels documents que es volen localitzar (totes les durades, > 1 minut, < 1 minut), tal com es mostra en la figura 8.7.6.4.

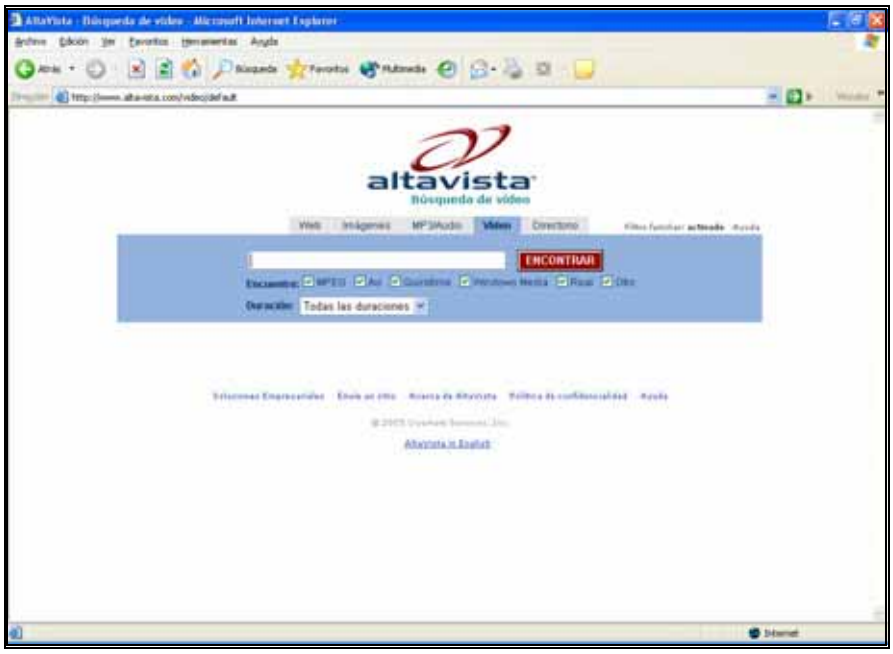


Figura 8.7.6.4. Altavista. Pantalla de cerca de vídeo.

De cada resultat en mostra una imatge, que a la vegada fa d'hipervincle i en dóna el títol, la mida i el temps de durada en segons, tal com es pot veure en la figura 8.7.6.5.

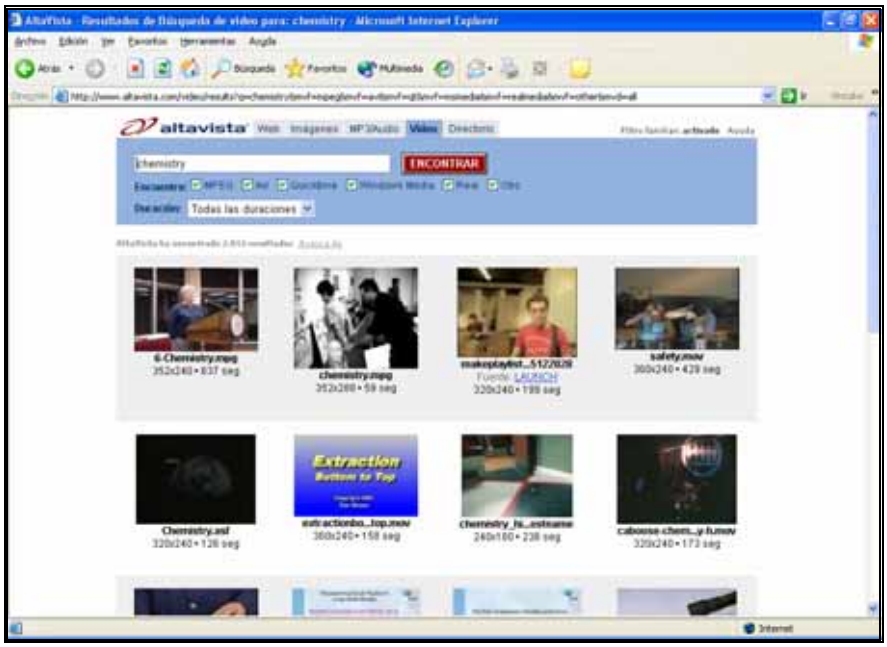


Figura 8.7.6.5. Altavista. Pantalla de presentació dels resultats de cerca de vídeos de «chemistry».

c) Mamma

És una empresa del Canadà creada el 1996 com a resultat d'una tesi de màster. És un metacercador intel·ligent. Cada cop que es fa una consulta, Mamma la llança de forma simultània a una diversitat de motors, directoris i altres continguts de la Internet profunda. En la figura 8.7.6.6 es mostra la pantalla d'inici del cercador *Mamma*.

Figura 8.7.6.6. Pantalla de cerca del metacercador Mamma.

Un cop compilats els resultats, elimina els duplicats i els mostra d'una manera uniforme ordenats segons la seva rellevància. De cada resultat en dóna el títol, que també fa d'hipervincle; una petita descripció extreta de la pàgina web on es troba el document audiovisual localitzat, i l'hipervincle juntament amb el cercador del qual procedeix la informació. No es mostra cap imatge en els resultats. En la figura 8.7.6.7 apareix la primera pàgina de resultats obtinguts en fer la cerca «video on chemistry».

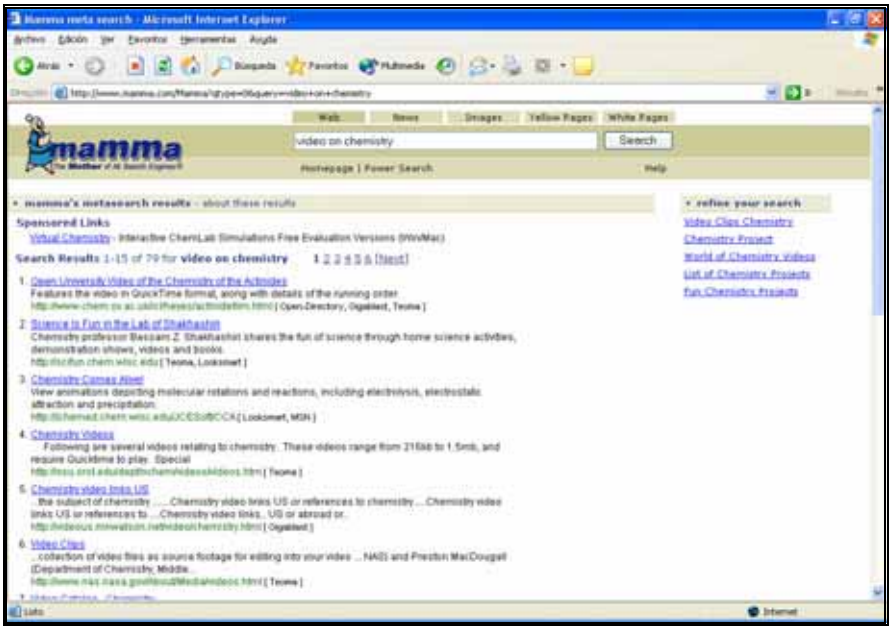


*Figura 8.7.6.7. Mamma. Pantalla de presentació dels resultats de la cerca «*video on chemistry*».*

8.7.7 Cercador emprat per localitzar documents audiovisuals en el Web

Un cop finalitzada la fase d'estudi dels diversos cercadors (107 en total), es va acotar a 9 el nombre de cercadors que podrien ser apropiats d'acord amb els objectius proposats, tal com s'ha comentat en l'apartat 8.7.4. D'aquests es va acabar seleccionant Alltheweb com el més adient, ja que és d'on es considera que procedeix la capacitat de cerca que permet la localització de documents audiovisuals en el WWW.

En el moment de fer la consulta (març-abril del 2004) i posterior revisió dels llocs web localitzats, el nombre de resultats obtinguts va ser de 1.109. Es van revisar tots els resultats, i es van descarregar i guardar en el disc dur de l'ordinador tots els documents audiovisuals localitzats a través dels enllaços facilitats pel cercador. En el capítol 9 es descriuen els recursos localitzats i s'analitzen des d'una perspectiva química i didàctica.

En el moment de presentar aquest treball d'investigació el nombre de resultats obtinguts amb el cercador Alltheweb ja és de l'ordre de 3000.

Els cercadors d'àmbit català i espanyol que van ser revisats no van proporcionar pràcticament cap enllaç a recursos audiovisuals que poguessin ser visualitzats a través d'Internet. En la majoria de casos els resultats obtinguts en les cerques corresponien a informacions de caràcter general relacionades amb ofertes de cursos per part d'entitats privades o universitàries, així com altres activitats acadèmiques. Es podria dir que la funció primordial de les pàgines web en aquests àmbits és promocional o informativa: cursos, gestions acadèmiques, etc. Les pàgines amb continguts relacionats amb alguna assignatura determinada tenen un format bàsicament textual. L'únic component visual correspon a la presència d'imatges estàtiques que acompanyen el text.

Per altra banda, Yahoo! que tal com s'ha comentat en l'apartat 8.7.6 va adquirir Alltheweb i Altavista, és de preveure que properament ofereixi la capacitat de cerca de vídeo entre les seves capacitats de cerca.

En el moment de presentar aquest treball d'investigació, ja és possible localitzar alguns recursos audiovisuals d'àmbit espanyol, mitjançant els cercadors comentats en l'apartat 8.7.5. En l'apartat 9.4.1 se'n farà una breu descripció.

8.7.8 Problemes que plantegen els motors de cerca

Com que no hi ha cap tipus de metainformació, no existeix un procés previ d'anàlisi documental: en utilitzar un motor de cerca, s'accedeix directament a la llista de documents, sense passar abans pel procés de filtratge propi de les bases de dades amb registres descriptius típiques de l'anàlisi documental. Això és el que genera la gran quantitat de soroll i de pèrdues d'informació tan habituals en aquesta mena de tecnologies.

No es disposa de la informació necessària per decidir si s'ha d'examinar seqüencialment tot el conjunt de documents originals. Si únicament ens interessa consultar documents de determinades característiques formals o de gènere, no tenim altre remei que examinar un per un tots els documents obtinguts tant si la llista és curta com si és molt llarga (Codina, 2003a, p. 46).

L'índex que facilita l'accés als documents obtinguts a través d'un motor de cerca es genera a partir del document original que forma part del context del document. En canvi, en una base de dades, l'índex que facilita l'accés a la informació es genera amb els termes que procedeixen no únicament del document original sinó també dels camps que formen part del conjunt de metadades. Actualment, els bancs de dades de documents audiovisuals relacionats amb la química, de lliure accés a través del web, són inèdits.

Un altre fet que s'ha constatat després de fer la cerca és que es recupera el que troben els enginys de cerca, però a vegades no buiden tots els arxius que hi ha i n'ofereixen només alguns.

D'altres vegades, quan es consulta un dels resultats mostrats pels cercadors en un lloc web d'una universitat, es troben altres documents o enllaços a altres recursos que els cercadors no han recollit, ja sigui dins de la mateixa pàgina web que es consulta, en una altra secció de la mateixa universitat o en altres universitats.

Els cercadors són una bona opció, però pot ser que encara que trobin molts recursos no localitzin el que es necessita i que, per altra banda, sí que existeix però en un lloc que no ha revisat cap cercador. Se sap que els cercadors visiten més les pàgines més consultades, que poden ser-ho per diversos motius. Els primers resultats de les cerques solen ser els més consultats, ja que es dóna per fet (tant per experiències personals adquirides en les diverses cerques fetes, com perquè així ho manifesten obertament els cercadors) que són els més rellevants, els més concordants amb la consulta feta a través de les paraules clau. Els resultats no són sempre els buscats, i pot ser que pàgines que contenen recursos molt vàlids no apareguin a la llista.

De fet, d'una manera que podríem anomenar manual es pot accedir a un nombre més ampli de recursos, si bé és una feina que podria no acabar mai, a causa del volum de pàgines web penjades a Internet. A més, sempre hi haurà un determinat «espai fosc» de pàgines noves o velles però no vinculades a altres i que són com illes dins de l'entramat.

Els recursos que trobem a les diferents universitats i que formen part dels cursos que s'estan impartint en línia normalment estan subjectes a qüestions de gestió economicoacadèmica, a causa dels pagaments de matrícules, i solen estar tancats per un accés que s'ha de validar amb un nom d'usuari i una contrasenya.

A aquest tipus de recursos no hi tenen accés els enginys de cerca, ni tampoc els usuaris que consulten el web. Per tant, es fa molt difícil, per no dir impossible, conèixer què hi ha realment i a on. Evidentment es pot esbrinar si visitem la universitat i demanem què s'està impartint i com. No cal dir que és molt difícil obtenir d'una forma sistematitzada i ràpida el que fa cada professor individual en les seves assignatures, quin tipus de recursos utilitza i, si penja materials en línia, quins són i com s'hi pot accedir. Per saber tot això cal una participació de les institucions educatives. A més, molt sovint el professorat és esquiu a aquesta mena de requeriments, ja que una col·laboració suposa un consum de temps i, en un cert sentit, viola el sentiment d'intimitat respecte al desenvolupament de l'activitat docent, que considera com a quelcom propi, com un tret individual i generalment no compartit.

Un dels problemes principals a l'hora d'emprar aquests enginys és que no n'hi ha cap que sigui exhaustiu. A més a més, cada un té una gramàtica o llenguatge d'interrogació propis, i com que n'hi ha tants es fa difícil conèixer i dominar la sintaxi de tots. Conèixer bé les característiques de l'enginy de cerca i el seu llenguatge d'interrogació pot millorar molt els resultats; així mateix, cada vegada és més habitual que els robots emprin tècniques més avançades d'indexació.

8.7.9 Consideracions finals

En només una dècada Internet ha experimentat un creixement exponencial, tant en la informació disponible com en els usuaris que hi accedeixen. Aquest procés d'expansió d'Internet ha portat a una reducció significativa del component universitari pel que fa als continguts d'Internet i a l'usuari final, però també han augmentat les possibilitats d'informació i de comunicació entre els investigadors. Internet no és cap panacea però s'està manifestant com una condició *sine qua non* de tota investigació. Possiblement és utòpic pretendre organitzar i ordenar tot l'univers Internet de manera uniforme: la immensa massa de materials existents a la xarxa, el seu creixement en progressió geomètrica, la diversitat de procedències i sobretot d'interessos dels organismes i de les persones que hi publiquen dificulten l'adopció d'un sistema comú de metadades. Se suposa que, pels recursos que poden interessar a comunitats concretes d'usuaris, sigui més fàcil arribar a acords que contribueixin a controlar bibliogràficament determinats espais d'Internet.

L'altra qüestió que cal considerar és el pervers volum de dades actualment presents en el web, que juntament amb el que s'afegeix cada dia hauria de fer sentir un grau de respecte pel que han aconseguit els motors de cerca web en un període de temps tan curt. L'accés encara que sigui elemental a centenars de pàgines de materials és una proesa que hauria d'inspirar molta més admiració que no pas decepció. Tot i que no tenen el nivell dels serveis oferts per les bases de dades comercials, cal tenir en compte que és un servei gratuït, que treballa amb dades desestructurades o almenys que tenen molt poca consistència estructural. L'única estructura «intel·lectual» està en els títols i en les etiquetes meta.

El 1999 Steve Lawrence i C.L. Giles estimaven un volum de 800 milions de pàgines i que els motors de cerca més grans cobrien menys d'una quarta part d'aquest material (Lawrence i Giles, 1999). Cal considerar que cobrir només una quarta part és força bo; de fet, els serveis tradicionals d'indexació mai han cobert aquests percentatges de material publicat. Serveis tan respectats com el Chemical Abstracts o el Pedagogical Abstracts no han pretès mai cobrir tot el que es publica en el camp de la química o la pedagogia. En poques paraules, cal aprofitar el que els motors, els directoris i altres sistemes de cerca poden cobrir, i buscar en més d'un enginy si es vol trobar tanta informació com es pugui sobre un tema (Hock, 2001).