Xarxa Punt TIC

MÒDUL 1 NIVELL AVANÇAT
Recerca i recuperació de la informació a Internet (avançat)
Problemàtica documental de la informació al Web

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

➔ ÍNDEX

➔ A. Problemàtica documental de la informació al Web

Amb l'aparició dels BBS (sigla de *Bulletin Board System*), que permetia accedir de manera rudimentària a dades remotes usant un mòdem, l'accés a documents des d'una computadora casolana es va simplificar bastant. Les boques de producció de documents electrònics es van multiplicar, així com la quantitat d'usuaris amb computadora que intercanviaven informació. Tot i així, cada BBS actuava de manera autònoma (dècada dels '80 i els primers anys '90).

I llavors va arribar el Web. Per a bé o per a malament. Per a bé, perquè Internet va connectar totes les màquines que produïen (i recaptaven) informació, perquè va abaratir i democratitzar la producció i recol·lecció d'informació i perquè la quantitat d'informació va créixer (y segueix creixent) en proporcions mai imaginades. Però també per a malament, perquè llavors la informació no estava tota en el mateix format, perquè no era necessàriament veritable i perquè podia estar "desactualitzada" o tenir errors.

En aquest context van aparèixer els cercadors d'Internet, per tal d'intentar posar una mica de sentit a tot aquest inabastable oceà d'informació *online*. Els cercadors van evolucionar ràpidament, intentant ajudar, cada vegada millor, a organitzar els milers de milions de documents que es produeixen.

Tipologia. Estructura

A grans trets, hi ha dos tipus de cercadors a Internet: **els directoris i els motors de cerca.**

Els directoris

Són cercadors organitzats a partir d'una jerarquia temàtica (taxonomia). El més conegut dels directoris és Yahoo a la seva pàgina http://es.dir.yahoo.com i Google a les pàgines http://www.google.com/dirhp?hl=ca o http://www.google.com/dirhp?hl=es. Es pot navegar un directori endinsant-nos en les seves categories i subcategories o ingressant una paraula clau que mostrarà les diferents categories i llocs on apareix aquesta paraula.

Hi ha directoris generalistes, com Yahoo!, i especialitzats, com Ariadna, cercador de recursos periodístics (http://www.periodismo.com/buscador/). Tot i que la majoria de cercadors generalistes imiten la taxonomia de Yahoo!, no hi ha un estàndard en aquest sentit. Tampoc hi ha cap tipus d'homogeneïtat o criteri comú entre els cercadors especialitzats. El directori mostra els resultats de la seva cerca basant-se només en el títol i la descripció del lloc. A més, els directoris inclouen llocs complets, no pàgines i seccions dins un lloc. Una altra de les característiques dels directoris és que les pàgines són revisades per

éssers humans i incorporades només si acompleixen els criteris de pertinència del cercador. Això fa que la quantitat de llocs dels directoris sigui petita en comparació amb tots els llocs que existeixen.

L'altre tipus de cercadors són el motors de cerca. El més conegut actualment és Google (www.google.cat). Els motors de cerca no tenen una taxonomia, es pot accedir als resultats només a partir d'una paraula clau. Tot el procés d'indexació de pàgines és automàtic, no hi ha persones revisant cada lloc. A diferència dels directoris, els motors de cerca busquen en tota la pàgina d'un lloc web, no es limiten al títol i a la descripció. Si bé no indexen totes les pàgines d'un lloc, tampoc es limiten a indexar una sola pàgina. La quantitat de pàgines indexades és enorme: en el moment d'escriure això, Google tenia indexades més de 8.000.000.000 pàgines web. Els motors de cerca tenen robots cercadors que exploren els llocs web i els incorporen a les seves bases de dades. Aquesta acció s'anomena indexar.

Actualment molts cercadors combinen la potència dels motors de cerca amb la lògica dels directoris. Si Yahoo! no troba resultats en el seu directori, mostra els resultats del seu motor de cerca. Però també els motors de cerca recorren als directoris: per a qui necessiti resultats més ordenats, Google utilitza les dades del directori Dmoz, conegut també com *Open Directory* (http://dmoz.org/).

Com funcionen

Robots

En una escala microscòpica, els robots dels segle XXI són invisibles i immaterials. Aquests robots es dediquen a fer una cosa clau: indexar les pàgines que es visiten (Piscitelli, 2005).

Avui la situació és molt diferent de fa quatre o cinc anys enrere. En aquell moment, la fe en els robots cercadors feia suposar que si cercadors com Altavista o Hotbot no trobaven allò que buscaven, era senzillament perquè aquesta informació no existia a la Xarxa.

Tanmateix, es va començar a donar importància a la qualitat per sobre de la quantitat. Això va determinar que era preferible indexar llocs de qualitat, abans que apilar la major quantitat de llocs possibles, acudint als cercadors. L'univers Web estava ple de pàgines que no valia la pena de visitar mai.

En aquell moment interessava l'àrea de l'aprenentatge robòtic i es va construir un robot anomenat Inquirus, capaç d'interrogar a altres robots sobre l'existència de documents que complissin certa estructura de cerca; aquest robot podia aportar un benefici secundari més valuós que el que es buscava originalment, estimant la mida real de la xarxa, un número que en aquell moment ningú coneixia amb certesa. Entre els resultats que va aconseguir l'Inquirus aplicat al cercador Hotbot, va descobrir, el 1997, que el Web comptava amb prop de 320 milions de documents (el doble del que es cria abans). I no només això, Hotbot es preava de ser el més exitós i exigent dels robots en aquella època, però, de sobte, es va veure devaluat quan es va descobrir que només indexava el 34% de tota el Web. Com a premi de consol va poder presumir de que als altres

robots els anava encara pitjor: Altavista només cobria un 28% y altres cercadors –com Lycos, que aviat va caure a les mans de Terra i Telefónica– a penes cobrien un 2% de la Xarxa.

El febrer de 1999, quan es va repetir el mateix exercici, els investigadors van trobar que la Xarxa havia crescut (tenia 800 milions de documents), però que la capacitat dels robots d'indexar llocs havia empitjorat.

Un cercador excel·lent de l'època, Northern Light, va ocupar llavors la *pole positon* cobrint el 16% del Web, però Altavista havia baixat al 15% i Hotbot ressenyava a penes l'11% de les pàgines existents. Mentrestant Google, que era un benjamí entre els pesos pesats, a penes veia llavors un 7,8% de les pàgines estimades. El juny de 2001, Google va cobrir per primera vegada 1.000 milions de documents, seguit de prop per Alltheweb. Avui, Google està prop d'arribar als 9.000 milions de documents.

Per molt impressionant que sigui la capacitat d'indexació dels motors, el Web creix infinitament més ràpid que la capacitat que tenen els motors d'analitzar-lo. A més, existeix el Web profund, que és almenys 550 vegades més gran que el que els robots poden arribar, fet pel qual l'asimetria entre el que és visible i el que existeix s'amplia molt més.

L'any 2000, sis de cada deu pàgines no havien estat visitades mai. Avui els resultats arriben a xifres d'entre vuit i nou de cada deu.

Els cercador en el seu rol de gatekeeper

La noció de *gatekeeping* (el porter o vigilant de l'accés) investiga la manera irregular en què les informacions circulen i es troben sotmeses a instàncies que les demoren o traven en algun punt de la cadena de comunicació, i la fluïdesa amb què circulen després les que aconsegueixen passar la barrera. Aquests llocs de demora o nusos que actuen com a barrera i filtre en la circulació de la informació serien eles *gatekeepers* o porters.

El concepte de *gatekeeper* va ser introduït pel psicòleg Kurt Lewin el 1947 mentre treballava en dinàmica de grups i va observar que la informació circulava d'una manera molt irregular, ja que en alguns moments podia interrompre's pels nusos o fluir de manera molt àmplia després de superar-los.

Cal imaginar el cercador com a un *gatekeeper*: en l'univers de totes les les pàgines del Web, el cercador té el poder d'orientar en el camí cap a la cerca de la informació.

Ja se sap que els cercadors no indexen totes les pàgines del Web. Aquí ja hi ha una primera selecció. El *gatekeeper* cercador deixa fora dels seus resultats una gran quantitat de contingut. La segona selecció està en la rellevància: el cercador defineix que determinades pàgines són més importants que d'altres. I es pot comprovar fàcilment que aquest criteri és subjectiu, encara que sigui automàtic, si es comparen els resultats dels diferents cercadors.

Com desafiar aquests criteris? El segon criteri es pot enganyar més fàcilment: utilitzant diversos cercadors i directoris es pot arribar a una "intersubjectivitat de resultats". Hi ha metacercadors com kartoo, turbo10, webcrawler, dogpile,

clusty entre d'altres, per exemple clusty (http://clusty.com/) mostre a l'usuari les millors posicions en què figura cada pàgina en els diferents cercadors.

Desafiar la lògica del cercador pel que fa als continguts que deixa fora dels seus resultats, porta a endinsar-se a l'anomenada Internet invisible.

Com buscar

El primer pas en una cerca és saber què és el que es busca. No necessàriament s'ha de saber amb precisió. Pot interessar, puntualment, trobar la bandera de Rússia o, més vagament, trobar legislació sobre jubilació privada a Amèrica Llatina.

Després de conceptualitzar el que es vol buscar, se sabrà cap on anar. Si la cerca és més específica, es començarà amb un motor de cerca que condueixi cap el lloc que es necessita. Si és més general, serà bo començar per un directori que agrupi tots els llocs comuns al tema que s'investigui.

Si el que es necessita és local o regional, s'haurà de restringir la cerca a aquells països o regions o, millor encara, consultar cercadors de la zona en qüestió.

En aquest sentit, si la temàtica és específica, s'haurà de partir d'un cercador generalista, buscar allà un directori temàtic i realitzar en el directori temàtic una segona cerca, més acotada.

Cada cercador té les seves regles (sintaxi) i per això és recomanable llegir la documentació i les pàgines d'ajuda per entendre bé les seves opcions de cerca.

La majoria dels cercadors accepten els operadors "booleans": AND, OR i AND NOT (aquest últim en alguns cercadors funciona posant només NOT o el signe -).

Per defecte, la majoria de cercadors funcionen amb l'operador AND o +. Això vol dir que posar en un cercador les paraules mapa argentina o mapa AND argentina o mapa + argentina és equivalent.

Posant la paraula OR, mostrarà els documents que continguin almenys una d'aquestes paraules. Per exemple, si es posa argentina OR uruguai mostrarà les pàgines que continguin la paraula argentina, les pàgines que continguin la paraula uruguai i també les que continguin les dues paraules. L'operador OR és també útil si no se sap com s'escriu una paraula (volswagen OR volkswagen), ja que portarà documents que continguin almenys una de les grafies.

L'operador NOT o el signe – exclou paraules de la pàgina de resultats. "Cindy Crawford" – sex – porn –adult –xxx –nude mostrarà documents que mencionin la supermodel, però no contingut pornogràfic. D'això se'n diu filtrar o refinar una cerca. També es pot fer servir si, per exemple, es vol informació només sobre Windows XP però no de Vista haurem de posar windows +XP –Vista).

Les cometes, com en l'exemple anterior, serveixen per indicar una frase exacte: termes que se sap que han d'anar junts, com el títol d'un llibre, d'una pel·lícula,

d'una cançó o d'un joc. S'ha de tenir la certesa que s'escriu de manera correcta, perquè si no ignorarà el que es demana.

Els operadors AND i OR s'han d'escriure en majúscules.

Tots aquests operadors poden ser molt útils si s'utilitzen de manera combinada. Així, per exemple, si volem trobar totes les pàgines en les quals aparegui mencionat Estats Units en català, excloent la grafia en anglès, es pot posar: **"Estats Units" OR EE.UU. OR EEUU -"United States" –USA.**

Cercadors i directoris

- Gigablast Inc. Gigablast <http://www.gigablast.com/>
- Periodismo.com. Ariadna <http://www.periodismo.com/buscador/>
- Google. Google <http://www.google.cat>
- Grub Buscador <http://www.grub.org>
- Ask.com <http://www.ask.com/>·
- IAC Search & Media. Excite <http://www.excite.com/>
- LookSmart, Ltd. Wisenut <http://www.wisenut.com/
- Lycos, Inc. Hotbot <http://www.hotbot.com/>
- Lycos, Inc. Lycos search <http://www.lycos.com/>
- Microsoft. MSN <http://www.msn.com>
- Overture Services, Inc. Alltheweb, find it all <http://www.alltheweb.com/>
- Overture Services, Inc. Altavista <http://www.altavista.com/>
- The New York Times Company. About <http://www.about.com/>
- Walt Disney Internet Group (WDIG). Go.com <http://go.com/>
- WebFile.com. Webfile <http://www.webfile.com/>
- Yahoo! Inc. Yahoo! <http://www.yahoo.com./>

Metacercadors

- ·Copernic Technologies, Inc. Copernic <http://www.copernic.com/>
- ·Digital Tsunami, Inc. Quickfindit <http://www.quickfindit.com/>
- ·Ez2find.com. ez2finf <http://ez2find.com/>
- ·InfoSpace, Inc. Dogpile <http://www.dogpile.com/>
- ·InfoSpace, Inc. Metacrawler <http://www.metacrawler.com/>
- ·Intelliseek, Inc. ProFusion <http://www.profusion.com/index.htm>
- ·Mamma, Inc. Mamma <http://www.mamma.com/>
- ·Surfboard BV. Ixquick <http://www.ixquick.com/>.
- Netscape Communications Corporation. DMOZ Open Directory Project <http://dmoz.org> [

Cercadors de cercadors

- ·Multibuscador.com <http://dir.multibuscador.com/>
- ·Buscopio <http://www.buscopio.net/esp/>

Eines de segona generació: classificació documental

Ens trobem amb un conjunt totalment nou d'eines, diferents a les anteriors perquè són *client-side*. Es tracta, per tant, de programes totalment independents que s'instal·len a l'ordinador client, fet que redunda en un major control i personalització de les seves funcions. El fet que, de vegades, algunes d'aquestes eines poden funcionar de forma autònoma respecte el client en el qual estiguin instal·lades, ha portat a que, incorrectament, es generalitzi el nom d'agent o *bot*, que pot identificar algunes d'elles però no totes.

En general, el conjunt resulta relativament heterogeni, la qual cosa permet construir una classificació molt descriptiva.

A més, com que alguns dels mecanismes són paral·lels als que existeixen com a servidors, aquesta segregació resulta especialment útil i admet anàlisis comparatives de prestacions. Tanmateix, el valor afegit d'alguns d'ells no es restringeix únicament a un increment de la capacitat d'automatització, sinó que ofereixen possibilitats totalment noves. Algunes de les opcions inèdites resulten impossibles d'implementar des d'un servidor.

Entre les novetats més singulars destaquem:

- La possibilitat d'extreure informació d'Internet invisible (infranet), el conjunt de registres de bases de dades o catàlegs de biblioteca accessibles mitjançant formularis web, però que no són indexats pels motors.
- L'ús dels veritables agents que, de manera autònoma, mitjançant mecanismes intel·ligents, poden recórrer la Xarxa, extreure informació i, fins i tot, "aprendre" amb ajuda de l'operador humà. La majoria dels programes revisats són productes comercials disponibles sota el sistema *Shareware* (avaluar abans d'adquirir), cosa que significa que es pot obtenir una còpia d'aquests programes, més o menys operativa, a la xarxa Internet. El preu no és excessivament car, i són precisament els programes més sofisticats els que costen més. Lamentablement, per a aquests tipus de programes a penes s'ofereix suport tècnic i alguns títols, a més, desapareixen ràpidament.

A continuació presentem una classificació comentada d'aquestes eines, utilitzant com a criteri sistematitzador les potencialitats i aplicacions documentals que tenen. Aquest criteri exclou altres programes, relativament nombrosos actualment, de vegades reunits sota la categoria d'"utilitats d'Internet", que són potencialment interessants. L'interès d'aquests programes, sobretot informàtic, pot ser més evident en un futur no gaire llunyà. Segons els usos documentals, distingim cinc grans grups per odre de complexitat:

- Clients Z39.50
- Bolcadors

- Metacercadors
- Indexadors
- Mapadors de port

A banda, també hi ha les eines canalitzadores, que tenen un caràcter mixt. Basades en la tecnologia *push*, podríem qualificar-les d'híbrides, ja que necessiten tant una instal·lació client com un servidor.

La incorporació d'aquest tipus de serveis als clients universals (Netscape y Explorer) ens ha dut finalment a excloure aquestes eines de la nostra classificació, on prèviament les consideràvem "bolcadors" sofisticats. Es pot estar al corrent de les principals novetats d'aquest tipus de programes visitant periòdicament algun dels principals dipòsits de *software* a Internet.