Xarxa Punt TIC

MÒDUL 1 NIVELL AVANÇAT
Recerca i recuperació de la informació a Internet (avançat)
El Web privat / el Web propietari / el Web realment invisible

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

➔ C. El Web privat / el Web propietari / el Web realment invisible

Diversos especialistes i entitats acadèmiques es dediquen a la tasca d'elaborar i mantenir pàgines concentradores de recursos web seleccionats per àrees d'especialitat (*subject guides*), que poden contenir recursos que no són recuperables amb un cercador comú. Aquests directoris anotats o guies temàtiques solen tenir una gran qualitat, ja que comprometen el prestigi dels autors i de les institucions involucrades. La selecció de recursos sol ser molt acurada i l'actualitzen freqüentment. De vegades, diferents institucions s'associen i formen "circuits" (*web rings*) per a l'elaboració cooperativa d'aquestes guies. Un exemple és *The WWW Virtual Library*.

Els directoris anotats o guies poden incloure, a més, algun mecanisme de cerca en les seves pàgines o al Web en general (Moreno Jiménez, 2004). No n'hi ha prou a conèixer la varietat d'eines de cerca disponibles al Web, sinó que fa falta una orientació sobre el seu funcionament, sobre quines estratègies s'han de seguir per traçar una ruta de cerca adient i sobre com escollir els millors instruments per a cada necessitat. D'això se n'ocupen els tutorials o programes d'aprenentatge. *How to Choose a Search Engine or Directory,* de la Universitat d'Albany, als Estats Units, i les guies de *SearchAbility* i de la Universitat de Leiden, a Holanda, *A Collection of Special Search Engines* orienten l'usuari en l'ampli món tant dels recursos especialitzats en el Web com de les maquinàries que permeten la seva localització.

Però més enllà de totes aquestes eines i recursos es troba el Web invisible.

Sherman i Price identifiquen quatre tipus de continguts invisibles en el Web:

- El Web opac (the opaque web).
- El Web privat (the private web).
- El Web propietari (the proprietary web).
- El Web realment invisible (the truly invisible web).

El Web opac consta d'arxius que podrien estar inclosos en els índexs dels motors de cerca, però no ho estan per alguna d'aquestes raons:

- Extensió de la indexació: per economia, no totes les pàgines d'un lloc són indexades en els cercadors.
- Freqüència de la indexació: Els motors de cerca no tenen capacitat per a indexar totes les pàgines existents; diàriament se n'afegeixen moltes, o es modifiquen o desapareixen i la indexació no es realitza al mateix temps.

- Nombre màxim de resultats visibles: tot i que els motors de cerca presenten un gran nombre de resultats de cerca, generalment limiten el nombre de documents que es mostren (entre 200 i 1.000 documents).
- URL desconnectats: les generacions més recents de cercadors, como Google, presenten els documents per rellevància, basada en el nombre de vegades que apareixen referenciats o lligats a altres documents. Si un document no té un enllaç en un altre serà impossible que la pàgina es descobreixi, ja que no haurà estat indexada.

El Web privat està format per pàgines web que podrien estar indexades en els motors de cerca, però són excloses deliberadament per alguna d'aquestes causes:

- Estan protegides per contrasenyes (*passwords)*.
- Contenen un arxiu "robots.txt" per evitar que les indexin.
- Contenen un camp "noindex" per evitar que el cercador indexi la part corresponent al cos de la pàgina.

El Web propietari inclou aquelles pàgines en les quals és necessari registrar-se per tenir accés al seu contingut, ja sigui de forma gratuïta o pagant. Es diu que almenys el 95% del Web profund conté informació d'accés públic i gratuït (Turner, 2003).

El Web realment invisible està format per pàgines que no poden ser indexades per limitacions tècniques dels cercadors, com ara: informació emmagatzemada en bases de dades relacionals, que no es pot extreure si no es realitza una petició específica. Una altra dificultat és l'estructura i el disseny variables de les bases de dades, així com dels diferents procediments de cerca.

Eines de cerca en el Web profund

Els motors de cerca han millorat el seu funcionament en els darrers anys i permeten un nivell de precisió més alt en les cerques i ofereixen els resultats de manera més útil per a l'usuari. Però encara hi ha molts cercadors que només poden recuperar directament la informació que es troba disponible al Web, però no la informació que s'ofereix a través del Web. Quan es va prendre consciència de la magnitud del Web que resultava invisible per les dificultats que presenten els motors de cerca per accedir a ella, aquests motors de cerca van incorporar funcionalitats addicionals per facilitar la cerca en l'anomenat Web profund. Així, han sorgit cercadors especialitzats en aquest segment del Web. Per afrontar una cerca al Web profund cal tenir en compte que els metacercadors poden presentar limitacions, respecte les possibilitats de cerca de cada cercador per separat. Per exemple, quan la cerca és sobre materials i formats especials, resulta més fàcil utilitzar les opcions de cerca avançada que tenen el cercadors i, si fos necessari, cal realitzar cerques successives en diversos cercadors o recórrer als directoris concentradors que tenen.

Els mecanisme utilitzats per localitzar recursos en el Web profund consisteixen, sobretot, en directoris de recursos especialitzats, principalment bases de dades disponibles gratuïtament a la Xarxa. El patrocini de les institucions acadèmiques en l'elaboració dels directoris, particularment dels que són anotats, garanteix la cobertura i la qualitat dels recursos compilats. Les guies de recursos especialitzats generalment estan elaborades per bibliotecaris i són una excel·lent erina de cerca i localització de recursos, a més de constituir un bon instrument d'aprenentatge en l'ús de la informació.

Les pàgines *How to Choose a Search Engine or Directory,* de la Universitat d'Albany (Estats Units) i les guies de *SearchAbility* i de la Universitat de Leiden (Holanda) *A Collection of Special Search Engines* inclouen els recursos d'informació i cerca en el web profund.

Finalment, els motors de pregunta dirigida (*directed query engines*) tenen la capacitat de realitzar cerques simultànies en diferents bases de dades al Web. Lexibot i el seu successor, Deep Query Manager, així com Distributed Explorer (Warnick i d'altres) i FeedPoint, són exemples d'aquests motors avançats de cerca.

Estratègies de cerca en el Web profund

A més de les estratègies ja explicades per a la cerca al Web, podem afegir-ne d'altres específiques per a la cerca al Web profund o invisible.

Per a la cerca d'informació especialitzada:

- ➔ Usar les eines de cerca en el Web profund si busquem informació acadèmica de qualitat.
- ➔ Usar cercadors regionals especialitzats per localitzar informació originada fora d'Estats Units o en idiomes diferents a l'anglès.
- ➔ Usar metacercadors per realitzar cerques en diferents cercadors especialitzats alhora.

Per realitzar cerques avançades:

- ➔ Usar les opcions avançades dels cercadors per localitzar imatges o arxius PDF o PostScript.
- ➔ Usar directoris concentradors de cercadors per realitzar cerques avançades successives en uns quants d'ells.

Per avaluar la informació disponible al Web:

- ➔ Usar directoris anotats per avaluar si els recursos disponibles al Web profund són útils per la cerca que estem realitzant.
- ➔ Usar directoris de bases de dades per saber quines d'elles poden oferir-nos informació útil per a la nostra cerca.

Per buscar informació a bases de dades:

- ➔ Usar guies, directoris o motors avançats si la informació que busquem pot estar en una base de dades.

No hi ha dubte que els actuals cercadors i directoris del Web estan millorant el seu funcionament. Més enllà dels detalls tècnics que el públic no pot veure, l'eficiència d'aquestes tecnologies ha augmentat i això s'aprecia en els resultats de les cerques. A mesura que aquestes eines es vagin fent més poderoses, disminuirà la necessitat d'elaborar guies i concentradors de recursos, i també la d'orientar en les estratègies de cerca i en l'ús i aprofitament dels recursos localitzats.

Observant els resultats obtinguts pels motors de cerca, es pot verificar que persisteix encara la pràctica de no indexar totes les pàgines d'un lloc per part dels robots. Per exemple, es pot tenir la referència d'una base de dades que està disponible a través d'un lloc web mitjançant un enllaç a ella que conté una de les pàgines del lloc, en canvi, pot ser que no aparegui la referència a la pàgina d'accés directe a aquesta base de dades en aquest lloc.

És evident que la freqüència de la indexació ha augmentat en alguns cercadors i fins i tot es realitza de forma diferenciada per a alguns recursos. Les pàgines que canvien més (la informació de la borsa, per exemple) serien visitades amb més freqüència pels robots que les que són més estables en el seu contingut.

El nombre màxim de resultats visibles no és un problema quan els cercadors presenten els resultats ordenats per rellevància, ja que sempre apareixeran primer els resultats que s'ajusten més a la cerca realitzada. Quan es pugui realitzar una cerca avançada i els criteris de rellevància combinin el nombre de "lligues" amb la freqüència de paraules, la presentació dels resultats no serà un obstacle per trobar la informació.

L'usuari sempre ha de tenir en compte que els cercadors són més apropiats quan la cerca és específica, és a dir, quan es coneixen dades sobre el que s'està buscant; mentre que és millor realitzar cerques temàtiques en els directoris. Els URL desconnectats podrien evitar-se si existís l'obligació de registrar, encara que fos d'una manera molt senzilla, totes les pàgines que es pengen al Web. Però la gran descentralització d'Internet fa pensar que això no passarà en un futur immediat.

El segment del Web privat no representa una pèrdua de gran valor, pel que fa a la informació que conté, ja que en general es tracta de documents exclosos deliberadament del circuit d'informació per la seva poca utilitat. En qualsevol cal, són els amos de la informació els que decideixen no fer-la disponible, la qual cosa vol dir que difícilment es podran trobar mecanismes legítims per franquejar aquesta barrera. A més, els arxius robots.txt serveixen per evitar que els robots caiguin en "forats negres", que els facin entrar en processos circulars interminables, minvant així l'eficiència del seu funcionament.

En un article recent de la OCLC Office for Research (O'Neill; Lavoie i Bennett) s'examinen les tendències pel que fa a la mida, el creixement i la

internacionalització del Web públic, és a dir, la part d'informació més visible i accessible per a l'usuari. Les principals conclusions de l'estudi són:

El creixement del Web públic mostra un estancament en els últims anys, perquè es creen menys llocs web i d'altres desapareixen. Però això no vol dir que no augmenti el volum d'informació, és a dir, el nombre de pàgines o el nombre de terabytes. Una altra possibilitat, que no es menciona en aquest estudi però que pot deduir-se de les restriccions per a l'accés a ells, és que alguns llocs web són accessibles mitjançant el pagament d'una subscripció o un altre mitjà de registrament.

El Web públic està dominat per continguts originats als Estats Units d'Amèrica, escrits en anglès. Això ens porta a pensar que probablement hi hagi més recursos invisibles a pàgines originades en d'altres països i en d'altres idiomes.

Alguns cercadors tradicionals com Altavista o Google han evolucionat i presenten ara la possibilitat de realitzar cerques per materials o formats especials. Així, Google ens permet realitzar cerques avançades per localitzar imatges. Per altra banda, el concentrador HotBot presenta la possibilitat de buscar per diferents formats, per localitzar imatges, àudio, vídeo, arxius PDF, Script i Shockwave/Flash. Aquestes opcions estan actives a HotBot per als cercadors Fast (Altheweb) i Inktomi (Pure Web Search), però no funcionen amb Teoma ni Google, tot i que, com vàrem dir, existeix aquesta possibilitat si es realitza la cerca directament des de el lloc de Google.

Aquestes cerques en materials especials, com imatges, àudio i vídeo, són possibles gràcies a una catalogació textual. Les cerques en documents que presenten formats PDF, Flash, etc. es poden realitzar perquè existeixen directoris d'aquests arxius. Així, el principal mitjà amb el qual es poden fer les cerques és el text. Per exemple, si volem recuperar imatges en blanc i negre, aquestes imatges han d'estar classificades d'aquesta manera a la base de dades. Això implica, lògicament, un procés manual. Una pàgina web que conté una imatge, sense cap informació textual sobre el seu contingut, no es podrà recuperar automàticament si no és per la seva extensió (".jpg", per exemple).

Com hem vist, la definició més genèrica del que constitueix el Web invisible o profund apunta als recursos que no poden ser recuperats mitjançant les eines comuns de cerca. Per verificar la visibilitat del Web profund, que ha estat identificat pels autors de *The Invisible Web*, Moreno Jiménez (2003) ha seleccionat a l'atzar deu recursos del seu *The Invisible Web Directory* i ha realitzat la cerca en un cercador, un directori, un metacercador i un agent metacercador avançat en la seva versió gratuïta. Els resultats d'aquesta prova senzilla apareixen en el següent quadre:

Recurs	MSN	Yahoo!	MetaCrawler	Copernic
Artcyclopedia	SI	SI	SI (6 cercadors)	SI (8 cercadors)

CRA Forsythe List	SI	SI	SI (3 cercadors)	SI (5 cercadors)
Current Films in the Work (BoxofficeHollywood Hot Set)	SI	SI	SI (3 cercadors)	SI (4 cercadors)
Employee Benefits INFOSOURCE	SI	SI	SI (2 cercadors)	SI (3 cercadors)
Hamnet	SI	SI	SI (4 cercadors)	SI (6 cercadors)
Infonation	SI	SI	SI (5 cercadors s)	SI (7 cercadors)
Jourlit	SI	SI	SI (3 cercadors)	SI (7 cercadors)
Scholarly Societies Project	SI	SI	SI (4 cercadors)	SI (6 cercadors)
Vessel Registration Query System	SI	SI	SI (2 cercadors)	SI (6 cercadors)
Who's who in American Art (AskArt)	SI	SI	SI (6 cercadors)	SI (8 cercadors)

Quadre. 15. Resultats de cerca de recursos de *The Invisible Web Directory.*

Tots els recursos seleccionats de *The Invisible Web Directory* són localitzables amb les actuals eines de cerca. A més, en els resultats s'observa que existeixen múltiples referències en altres pàgines, es a dir, que es tracta de pàgines "connectades". L'única dificultat per trobar-les consisteix, en alguns casos, en les paraules amb què es denomina el lloc o el recurs. Per exemple, en el *The Invisible WebDirectory* apareix "Vessel Query Registration System", en lloc de "Vessel Registration Query System", això fa que la cerca per totes les paraules tingui èxit, però la cerca per frase no. Igualment, la denominació de "Who's who in American Art" per al lloc de "AskArt" dificulta la cerca, però si es busca directament pel seu nom apareix en molts cercadors. La taula mostra a més com el solapament entre cercadors és variable.

Es pot dir que el contingut de les bases de dades que estan incloses en aquest directori és invisible, ja que cal realitzar les cerques directament en cadascuna d'elles. Però la veritat és que arribar fins la "porta" d'aquestes bases de dades resulta relativament senzill. El mateix fet que el directori hagi estat col·locat al Web, dóna una visibilitat millor als recursos inclosos, ja que els enllaços en el directori augmenten la possibilitat d'indexació d'aquestes pàgines. Llavors, podem dir que *The Invisible Web Directory* és un bon directori de recursos i bases de dades disponibles al Web, però no és un bon directori de recursos "invisibles".

En conclusió, el que realment segueix sent invisible al Web són:

- Les pàgines desconnectades.
- Les pàgines no classificades que contenen principalment imatges, àudio o vídeo.
- El contingut de les bases de dades “relacionals”.
- El contingut que es genera a temps real.

Però:

- És relativament senzill arribar fins la “porta” de les bases de dades amb contingut important.
- Existeixen motors avançats capaços de realitzar cerques directes simultànies a diferents bases de dades a la vegada; i, tot i que la majoria demanen pagament, també ofereixen versions gratuïtes; el contingut que es genera a temps real per validesa a molta velocitat, tret dels anàlisis històrics.
- És relativament senzill arribar fins la “porta” dels serveis que ofereixen informació a temps real; el contingut que es genera dinàmicament interessa únicament a alguns usuaris amb característiques específiques.
- És relativament senzill arribar fins la “porta” dels serveis que ofereixen contingut generat dinàmicament.