

<b>TENDINȚE NOI ~N BIBLIOTECONOMIE</b>
<b>Carmen Diaconescu: Reg\sirea informației pe Internet</b>

## Regăsirea informației pe Internet: metamotoarele de căutare

de Carmen DIACONESCU

***Résumé:** Retrouver l'information, ce n'était déjà pas très facile à faire avant le Net. Depuis, l'apparent chaos qui décrivait l'état de l'information numérisée semble encore plus difficile à dominer. Grâce au traitement de l'information - travail pris en charge, pour la plupart, par les professionnels de la documentation - le chaos se structure et y retrouver ce que l'on cherche devient possible. De l'indexation aux moteurs de recherche, et des moteurs aux métamoteurs, l'évolution a été rapide et l'Internet s'ouvre à chacun de nous comme un espace, sinon familier, du moins abordable.*

***(Internet; indexare; cuvinte-cheie; cuvinte-cheie compuse; motoare de căutare; metamotoare de căutare)***



### **Preambul**

Toată lumea știe ce este Internet-ul și mulți îl folosesc în diverse scopuri. Cel mai adesea imensa rețea este folosită la căutarea de informații. Or, pentru a căuta ceva, într-un mediu oarecare, trebuie asumată ideea că acel ceva există și că există și elemente care înlesnesc căutarea. În Internet, informație există (pentru că ea este mereu colectată) și există și acele elemente care dau acces la informație (provenite din prelucrarea acesteia). Stocată în baze de date cu structuri și funcționări dintre cele mai diverse, informația este totuși accesibilă fără a mai fi nevoie ca cel care o caută să cunoască bazele de date în care se va face căutarea. Totul e doar o chestiune de indexare: indexarea bazei de date. Creatorii de pe Internet au mers mai departe și au construit așa-numitele

motoare de căutare, softuri capabile să distribuie cererea de informație mai multor baze de date simultan și să colecteze răspunsurile pentru a oferi informația solicitată într-o formă unitară. La baza funcționării acestor motoare stă principiul indexării. La baza indexării stă principiul anticipării modului/modurilor în care va putea fi formulată cererea de informație. După experiențe îndelungate cu diverse limbaje formalizate menite să elimine barierele lingvistice în descrierea claselor în care ar putea fi împărțită cunoașterea umană, odată cu conturarea tot mai pregnantă a viitorului în domeniul căutării informației digitalizate (care se va face direct în text), indexarea cu cuvinte-cheie capătă o tot mai largă utilizare. Acest gen de indexare presupune existența unor tezaure de termeni, din care se alocă unui document

## TENDINȚE NOI ~N BIBLIOTECONOMIE

### Carmen Diaconescu: Reglsirea informației pe Internet

dat cei nimeriți să fie "cheie" pentru accesul la acesta.

Pentru înțelegerea lucrurilor legate de prelucrarea documentelor în vederea regăsirii informației pe Internet, vom face o mică incursiune în tehnicile de indexare mai des întâlnite. Referirile noastre vor avea în vedere acele dezvoltări care au făcut posibilă existența și utilizarea motoarelor de căutare.

Revenirea, în cele ce urmează, asupra unor lucruri deja amintite în acest segment introductiv nu are decât rolul de a nu fragmenta expunerea. Deci, după o secțiune dedicată trecerii în revistă a celor mai noi tendințe în indexare, vom expune câteva noutăți legate de motoarele de căutare.

#### Indexarea

Pentru a putea avea informația într-o bază de date și a o regăsi, este esențială introducerea datelor și procesarea lor (mai nou, cu ajutorul unei inteligențe artificiale) în urma unei analize competente. Introducerea datelor constituie o problemă ce depinde de multe criterii și care trebuie controlată de operatorii umani. Problemele care apar la introducerea datelor pot fi clasificate în trei mari categorii și anume probleme

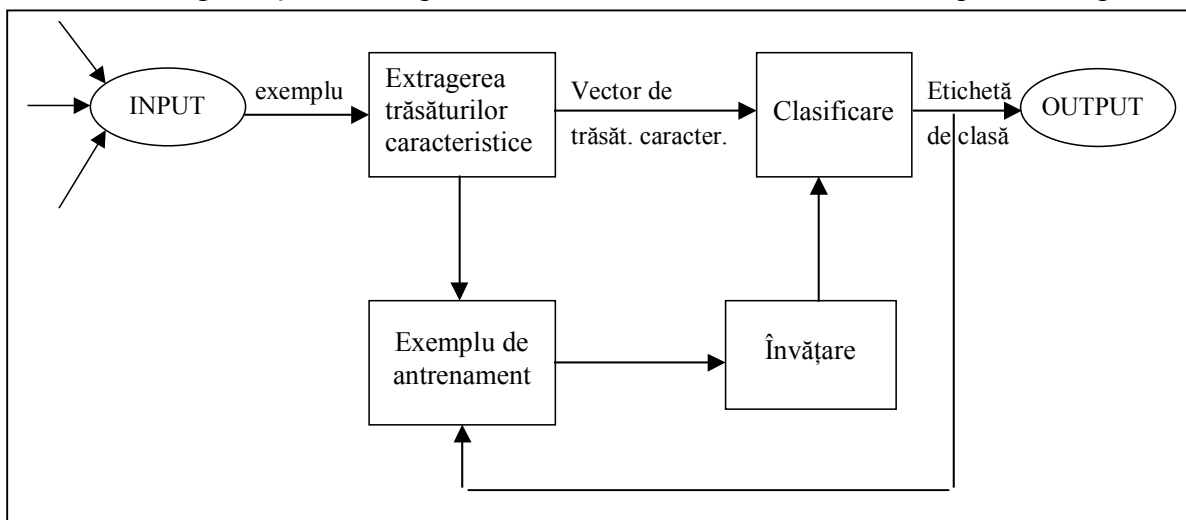
datorate: datelor prea multe, datelor prea puține, datelor disperate.

Procesarea datelor se referă la transformarea lor prin filtrare, ordonare, editare și prin îndepărtarea bruijelor la colectarea informației. Acestea se realizează prin vizualizare, eliminare, selecție, exemplu - pentru datele colectate, precum și la generarea de informații cu noi trăsături, obținute prin prelucrare (informații ce nu se pot măsura, date ce fuzionează, inducții constructive etc.).

Scopul general al sistemelor de clasificare este de a alocă un exemplu de input unei clase create. În sistemul de clasificare cu ajutorul algoritmilor de învățare, de exemplu, aceasta se realizează prin oferirea de către sistem a unui set de exemple de antrenament (training examples) cu o etichetă de clasă asignată pe parcursul procesului de "antrenament". Sistemul își acordă parametrii interni cu comportarea *input-output* dorită.

Iată, schematic, cum arată arhitectura unui sistem de clasificare prin învățare:

Problema care apare în legătură



## TENDINȚE NOI ~N BIBLIOTECONOMIE

### Carmen Diaconescu: Regăsirea informației pe Internet

cu regăsirea informației se referă la dificultatea pe care o întâmpină utilizatorul atunci când încearcă să formuleze cu acuratețe întrebarea pentru regăsirea documentelor dorite. Această problemă e legată de faptul că utilizatorul nu cunoaște lexiconul bazei de date și în final se obține un set de documente necorespunzătoare, nu tocmai relevant din punctul de vedere al căutării. Pentru a elimina aceste neajunsuri, bazele de date includ un tezaur care ajută la găsirea termenilor preciși. Datorită schimbării continue a bazelor de date pe Internet, și structurii neuniforme a acestora, tezaurul nu se poate actualiza manual. Tezaurul este însă o parte esențială a sistemului pentru căutare. Dacă nu este actualizat, apare bruiatul documentar. Pentru a ilustra această problemă să luăm un exemplu: dacă în motorul de căutare *Lycos* formulăm o cerere de informație după termenul "cluster", sistemul returnează un set de documente foarte mare (aprox. 10.000 de documente). Dacă însă căutăm după termenul "cluster analysis", numărul de documente se reduce la circa 100; deci, folosind cuvintele-cheie compuse situația se simplifică foarte mult.

În cazul tezaurului tipic, pentru fiecare termen se deschide o listă de cuvinte în care s-ar putea încadra chiar și termenii în relație cu termenul dat. Tezaurul poate fi construit manual sau automat. Cele automate sunt dependente de corpusul folosit la crearea lor, pentru că folosesc co-ocurența termenilor în document pentru a construi asociații de cuvinte. Aceasta determină și o actualizare în acest mediu care se schimbă perpetuu. Co-ocurența cuvintelor în interiorul propozițiilor poate realiza în mod automat crearea cuvintelor cheie compuse (o metodă mai veche). De

exemplu se alege un termen principal pe baza unei frecvențe de ocurență în text și se stabilesc asociațiile acestui termen cu alți termeni în aceeași propoziție. Din păcate, acest sistem este în mare măsură arbitrar și poate produce asociații fără sens. Noile metode generează dificultăți precum nevoia de a te raporta la relevanța în generarea cuvintelor cheie.

O altă metoda de generare a cuvintelor-cheie este cea care constă în folosirea frecvenței cuvintelor în documente. Există o listă a termenilor indexați din baza de date la care ne referim, listă care este generată automat prin tehnici simple de indexare. Se grupează apoi cuvinte înrudite semantic în așa-numite clase semantice. Dacă se face o căutare după un singur cuvânt-cheie, se face încadrarea acestuia într-o clasă, reducându-se plaja de căutare a asociațiilor fără sens (din punct de vedere semantic).

Cuvintele-cheie compuse sunt generate folosind informații din clasele semantice. Sistemul utilizează raportarea la o măsură de similaritate, pentru a aranja documentele într-o ordine în acord cu relevanța fiecărui document față de o întrebare.

Internetul conține milioane de pagini web și miliarde de cuvinte. În prezent nu există un catalog cu toate resursele *Internet*. Există totuși peste 100 de instrumente de căutare, dezvoltate de diferite companii și organizații, care încearcă să indexeze sau să clasifice aceste resurse. Acestea sînt motoarele de căutare.

#### **Motoarele de căutare**

Căutarea pe *Internet* se face în mai multe moduri.

## TENDINȚE NOI ~N BIBLIOTECONOMIE

### Carmen Diaconescu: Reglsirea informației pe Internet

1. Simplu - folosind un *browser* grafic standard (cum ar fi *Netscape*) și o adresă deja cunoscută (ex. *bcu-iasi.ro*, *bnf.fr*, *dntis.ro*, etc.). În acest caz se ajunge la pagina web dorită, de unde se navighează pînă la găsirea informației cerute. Informațiile căutate pot fi foarte greu accesibile în condiții "clasice", cînd nu există o bază documentară pentru un subiect.
2. Făcînd recurs la motoare de căutare care apelează mai multe baze de date, motoare de tipul: *Alta Vista*, *Infoseek*, *HotBot*, *Lycos*, *Yahoo*. Aceste tipuri de căutări devin din ce în ce mai curente și de aceea merită efortul de a fi cunoscute.
3. Folosind metamotoare de căutare. În prima linie sînt metamotoarele de căutare care interoghează simultan mai multe motoare pentru fiecare opțiune. Este cazul lui *Internet Sleuth*, unul din cele mai complete, care, printr-o simplă comandă, interoghează *Alta Vista*, *Excite*, *Ultraseek*, *HotBot*, *Infoseek*, *Lycos*, *Open Text*, *Yahoo*, *Webcrawler* sau *MataCrawler*, care lansează simultan *Yahoo*, *Excite*, *WebCrawler*, *Alta Vista*, *Lycos*, *Galaxy*.

Alte motoare mai modeste propun o căutare limitată (ex. *Savy Search* - care interoghează trei motoare în paralel).

Pentru a le descrie mai în detaliu "forța", vom aminti că, de exemplu, [www.isleuth.com](http://www.isleuth.com) - efectuează o căutare în 2000 de baze de date diferite; <http://guaraldi.cs.colostate.edu:2000> - permite interogarea a 26 de motoare diferite; <http://www.metafind.com/syntax.html> - are posibilitatea de triere în funcție de cuvinte-cheie, nume de domenii și e capabil și de ordonare alfabetică.

Aceste căutări, ce permit interogarea mai multor surse de informații, au un defect: sintaxa lor interogativă trebuie redusă la un singur cuvînt-cheie sau la operatori booleani clasici. Căutarea respectivă este departe de finețea căutării cu *Alta Vista*, ce permite un limbaj de interogare evoluat.

La acest nivel, simplul USE IT permite o căutare avansată. El permite un efect de lansare simultană a motoarelor de căutare și a indexului tematic. Mai mult, utilizatorii pot specifica timpii de căutare pentru fiecare motor selecționat (1-5 min) și numărul de răspunsuri maxim ce trebuie vizualizat (inferior sau superior lui 20). Este foarte probabil ca modalitatea aceasta de a pune niște criterii în căutare să ia amploare. Această aplicație de căutare a informației cu ajutorul metamotoarelor este foarte interesantă pentru că permite reperarea rapidă a resurselor interesante dintr-o mare masă de informație. De exemplu, dacă vrem să cunoaștem informații noi din robotica chirurgicală, în 5 minute, le putem afla cu *Internet Sleuth*.

Metamotoarele prezintă și un alt interes: posibilitatea de testare a motoarelor asupra unui subiect dat. În sfîrșit, grație interogării simultane a mai multor motoare de căutare este posibil să aflăm rapid relevanța lor asupra unor teme precise. Concret, este suficient să știm diferite cuvinte-cheie ale căutării într-un cîmp de cunoaștere pentru a lansa căutarea. Utilizatorii constată ce motoare de căutare le satisfac cerințele și duc la obținerea celor mai bune rezultate, iar în viitor se pot conecta direct, pentru o interogare profundă, beneficiind de toate posibilitățile oferite de limbajul de interogare al fiecăruia. Demarajul poate părea lent și greoi pe hîrtie, dar în

## TENDINȚE NOI ~N BIBLIOTECONOMIE

### Carmen Diaconescu: Reg\sierea informației pe Internet

realitate este simplu și plăcut, rapid și eficient.

Dacă prin aceste încercări nu reușim să descoperim motorul de căutare corespunzător așteptărilor noastre, o altă metodă constă în utilizarea acelor motoare care sînt deja clasificate în funcție de baza de date specifică unei teme sau zone geografice.

Fiecare motor de căutare lucrează pe o bază de date documentară din Internet, care este specializată pe sectoare geografice sau țări. Căutarea informației pe un domeniu deja clasificat este mult simplificată, iar interesul în utilizarea acelei informații crește. Așa se explică faptul că, de exemplu, <http://www.edirectory.com/> - propune la nivel mondial, căutarea motoarelor de căutare specifice fiecărei țări; <http://www.hj.se/hs/bib/miewww.index.html> - servește la căutarea motoarelor de căutare specifice fiecărei țări europene.

În același mod este posibil să identificăm motoare specializate pe anumite teme de căutare. Iată exemple în acest sens:

1. <http://www.isleut.com> - ce oferă aflarea motoarelor de căutare specifice fiecărui domeniu de activitate și furnizează o listă cu 21 de motoare de căutare specifice:

Sport  
 Politică  
 Sănătate  
 Informatică  
 Educație  
 Știri  
 Artă  
 Cultură  
 Învățămînt etc.,

motoare ce prezintă obiective, domenii de cercetare, tipuri de informații necesare.

2. <http://www.beaucoup.com> - identifică posibilitățile de căutare specifice unei teme.

Deci, aceste facilități permit căutarea și colectarea rapidă a informației și interogarea simultană a mai multor motoare. Pentru identificarea motoarelor de căutare, adaptate la cerințele de lucru off-line, trebuie să înțelegem că viteza de primire a informației de către utilizatorul comun este redusă. Limitarea se datorează modemului sau liniei de comunicație utilizate (viteza maximă suportată de o linie standard de telefon, de exemplu, este de 36 kbps). *Internetul* este foarte rapid (45Mbps, echivalent cu transmiterea *Operelor* lui Shakespeare într-o secundă) dar celelalte căi de comunicație au în medie 1,5 Mbps.

Pentru a economisi timp și din cauză că nu se poate elimina încetineala rețelei, informațiile se pot căuta și captura sub forma unor site-uri. Acest lucru se face folosind "uneltele" în continuă dezvoltare ale logicii aplicate *Internetului*. Ele permit, printr-o interogare simultană a mai multor motoare și printr-o înregistrare a rezultatelor pe discul dur, navigarea off-line în toată libertatea.

Acest mod de lucru care permite o economie notabilă de timp, funcționează cu navigatori clasici - *Internet Explorer* sau *Netscape*. Capturarea unui *site* complet implică existența unui spațiu suficient pe discul dur (cu titlu informativ, o pagină, în Franța, reprezintă 1-10 kocteți, iar un site 50-100 pagini). Este rațiunea pentru care majoritatea programelor arată nivelul maxim de ocupare a discului dur și permit fixarea programelor capturate. Aplicația poate deveni deosebit de interesantă în

## TENDINȚE NOI ÎN BIBLIOTECONOMIE

### Carmen Diaconescu: Reglșirea informației pe Internet

contextul în care informația este extrem de abundentă și se caracterizează printr-o dinamică incredibilă, când capturarea unui soft gratuit de pe Internet se poate face și într-un interval de timp când nu este necesară supravegherea operației (noaptea, de exemplu). Și, din nou, câteva exemple:

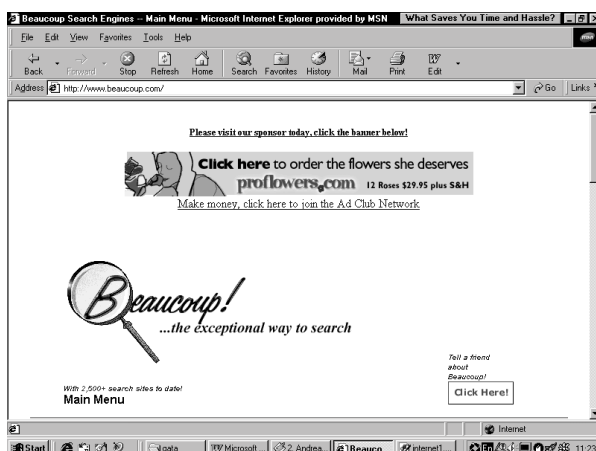
<http://memoweb.com> - program de capturare a unui site prezentat în Franța;  
<http://paperclip.com/getweb.htm> - permite înregistrarea unei pagini web pe discul dur, permițând o căutare ulterioară;  
<http://symantec.com/iff/> - care este complet, permițând o căutare în modul on-line și off-line.

Pentru a câștiga deci în eficacitate, trebuie să ne obișnuim cu aceste facilități și să le folosim în colectarea de informații. În acest context, să nu uităm că, pentru găsirea informațiilor, plecând de la baze de date și programe de căutare, este necesară înțelegerea și utilizarea eficientă a celor mai noi "oferte" ale explorării *Internetului*: motoarele și metamotoarele de căutare.

### Concluzii

### Referințe

- 1) G. Salton, *Automatic text processing*, New York, Addison-Wesley, 1989
- 2) G. Salton & C. Buckley, "Term weighting approaches in automatic text retrieval", in *Information Processing and Management*, 24/ 1988, pp. 513-523.
- 3) T. Saracevic & P. Kantor, "A study of information seeking and retrieving II: Users, questions and effectiveness", in *Journal of the American Society for Information Science*, 39 (3)/1988, pp. 177-196
- 4) D. Stoica, "Perspective biblio-logice Cuvintele-cheie", in *Biblos* 7/1998, pp. 21-26



**TENDINȚE NOI ÎN BIBLIOTECONOMIE**  
**Carmen Diaconescu: Reglșirea informației pe Internet**

