World Wide Web

Caracteristicile rețelei și ale motoarelor de căutare

decembrie 2018

Conținutul acestui curs

- Motoare de căutare web: scurt istoric, tehnologii de bază
- Estimarea nr. de documente indexate de motoarele de căutare web
- Eliminarea documentelor duplicat de către algoritmii de indexare

Caracteristici ale rețelei World Wide Web

1. Foarte mare (rețea globală, de nivel planetar)
2. Se extinde într-un mod necoordonat
3. Diversitate mare de utilizatori/participanți

Tehnologii pe care se bazează WWW:

HTML (Hypertext Markup Language)
HTTP (Hypertext Transfer Protocol)
URL (Universal Resource Locator)

Exemplu de URL:
`http://www.stanford.edu/home/atoz/contact/html`

- `http:` - protocol de transmitere a datelor
- `www.stanford.edu` - domeniu pt. o ierarhie de pagini web
- `/home/atoz/contact/html` - cale la o pagina web în ierarhia domeniului

Arhitectura browserelor web

Arhitectură client-server:

- comunicarea server-client prin protocolul `http` de transmitere asincronă a informațiilor de diverse feluri (text, imagini, fișiere video/audio etc.) formatate în limbajul `html`
- clientul (de obicei, un browser), transmite o cerere `http` la un server web:
  - se specifică un URL, de exemplu

    `http://www.stanford.edu/home/atoz/contact.html`
  - serverul transmite un răspuns codificat `html`, care conține
    - hiperlinkuri
    - conținut formatat cu taguri `html`.
- browserul are reguli de afișare a conținutului răspunsului, în funcție de formatarea `html`

Capabilitățile primelor browsere web

- Crearea ușoară de documente formatate `html`, și adresarea lor cu `url`
  - proces rapid de învățare
- Vizualizarea documentelor formatate html

Consecințe:

- publicarea masivă de informații diverse de către practic oricine $\Rightarrow$ WWW a devenit cea mai mare sursă de informații din lume
- Problemă nou ivită: WWW ca sursă enormă de informații necesită motoare performante de căutare a informațiilor de interes.

Descoperirea și accesarea informațiilor publicate, care sunt de interes pentru consumatori. Mai întâi au apărut 2 tipuri de motoare de căutare:

1. Căutare de cuvinte cheie în texte, bazată pe utilizarea de indecși inversați (calculați în prealabil) și mecanisme de calcul al gradului de importanță (engl. *ranking*). Exemple: Altavista, Excite și Infoseek
2. Căutare de documente organizate într-o ierarhie arborescentă de categorii. Exemplu: Yahoo!
   - **Avantaje:** metodă intuitivă de căutare, ușor de utilizat
   - **Dezavantaje:** clasificarea documentelor în categorii este un proces uman costisitor
     - de regulă, muncă făcută de editori experți
     - greu de efectuat pentru colecții mari de date (WWW)

- Bazat pe tehnici clasice de găsire a răspunsurilor la întrebări de căutare, precum
  - liste de indeși inversați
  - algoritmi de calcul al gradului de importanță al unui răspuns
- Volumul mare de documente publicate pe WWW ($10^6 \approx 10^7$) a determinat:
  - indexarea unui număr tot mai mare de documente, stocate și procesate pe un număr mare de mașini $\Rightarrow$ timpi de răspuns la întrebări < 1 secundă
  - calitatea și relevanța rezultatelor căutării lăsau de dorit:
    - $\Rightarrow$ inventarea unor algoritmi mai buni de calcul al gradului de relevanță (engl. *ranking algorithms*)
    - $\Rightarrow$ tehnici de detecție și evitare a spam-ului
    - $\Rightarrow$ tehnici de detecție al **gradului de autoritate** al unui document (de ex., pe baza site-ului web de unde provine)

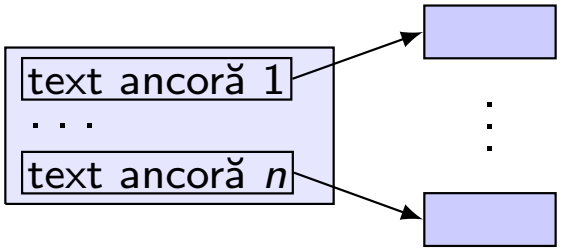
- Zeci de limbaje și mii de dialecte
  - extragerea și indexarea conținutului poate fi făcută doar cu algoritmi lingvistici avansați, pentru fiecare limbaj în parte
- Pagini web cu variații mari de structură, fonturi, și culori
  - unele pagini web nu au text indexabil, ci doau doar imagini
- Libertatea de a publica orice $\Rightarrow$ pagini ce conțin adevăruri, minciuni, contradicții, bănuieli, etc.
  - cum putem măsura gradul de încredere al unei pagini web?
- Întrebarea ,,cât de mare este WWW?" nu are un răspuns simplu:

  (+) putem estima nr. de pagini indexate de către un motor de căutare

  (?) cum recunoaștem și cum tratăm paginile web generate dinamic? (De ex., pagina generată dinamic despre zborul `AA129`)

Graful web

$G = (V, E)$ cu mulțimea de noduri V=pagini web statice; mulțimea E de arce=hiperlinkuri între pagini web.



Noţiuni auxiliare:

- link in al unei pagini A: un arc cu destinația A
- grad in al paginii A=nr. de arce cu destinația A
- link out al unei pagini A: un arc cu sursa A
- grad out al paginii A: nr. de arce cu sursa A

Se estimează că

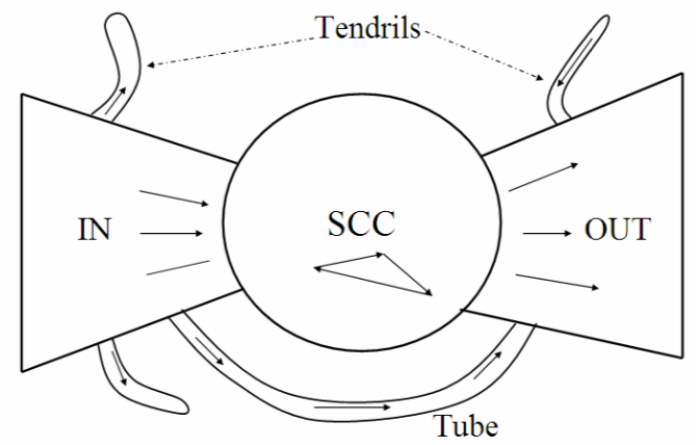
- În medie, **gradul in** al unui nod în G este între 8 și 15.
- Nr. de noduri n cu in-grad$(n) = i$ este proporțional cu $1/i^{\alpha}$, unde $\alpha \approx 2.1$

Proprietăți ale grafului web

Sunt 3 categorii majore de pagini web: IN, OUT și SCC, astfel încât

- există căi de la orice pagină IN la orice pagină SCC
- există căi de la orice pagină SCC la orice pagină OUT
- există căi de la orice pagină SCC la orice pagină SCC
- Nu există căi SCC$\rightsquigarrow$IN, și nici căi OUT$\rightsquigarrow$SCC
- nr. pagini IN $\approx$ nr. pagini OUT

$\Rightarrow$ graful web G are o structură de papion



Spam și tehnici de evitare a lui

Spam = tehnici de manipulare a conținutului paginilor web a.î. să aibe scor mare de importanță pentru căutări după anumite cuvinte cheie

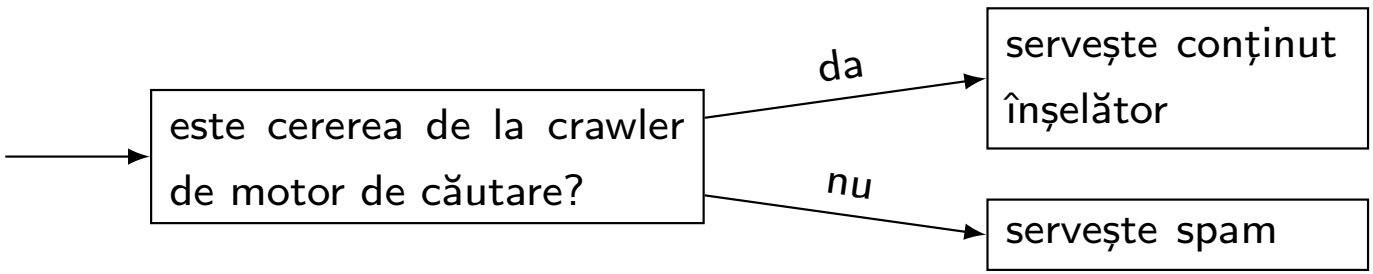
- activitate determinată, de obicei, de motive economice.

Exemplu de spam timpuriu:

- pagini web cu conținut repetat de anumiți termeni afișați în culoarea de background a paginii; create de vânzători ce fac reclamă la produsele lor, și de către agenții lor
- antidot: motoare de căutare care detectează nr. mare de repetiții de termeni în pagini web, și nu le atribuie un scor mare de importanță

Cloaking = tehnică mai sofisticată de creat spam:

- serverul web al creatorului de spam păcălește componenta **crawler** a motoarelor de căutare, verificând de la cine vine cererea de servire:



Producătorii de optimizatoare de motoare de căutare (SEO) oferă servicii de consultanță pentru clienții interesați să aibe pagini web cu scor mare de importanță pentru anumite cuvinte cheie

- se bazează pe descifrarea tehnicilor secrete ale motoarelor de căutare, care se folosesc pentru a calcula scorul de importanță al paginilor web
  - $\Rightarrow$ conflict perpetuu între producătorii de SEO și producătorii de motoare de căutare
  - $\Rightarrow$ numeroase motoare de căutare (începând cu Google) folosesc o tehnică, numită analiza link-urilor, pentru a combate spamming ca se bazează pe manipularea textului din pagini web

Reclama ca model economic

Branding=prezentarea de reclame (pe pagini web de pe site-uri populare) care produc impresii pozitive despre compania pentru care se face reclama.

Modele economice pentru reclamele făcute pe pagini web:

1. **cost per mil (CPM)**: costul suportat de companie pt. a afișa 100 de reclame
2. **cost per click (CPC)**: costul pt. numărul de click-uri pe reclama respectivă pt. a ajunge la pagina web a companiei pt. care se face reclama, de unde se pot efectua cumpărături

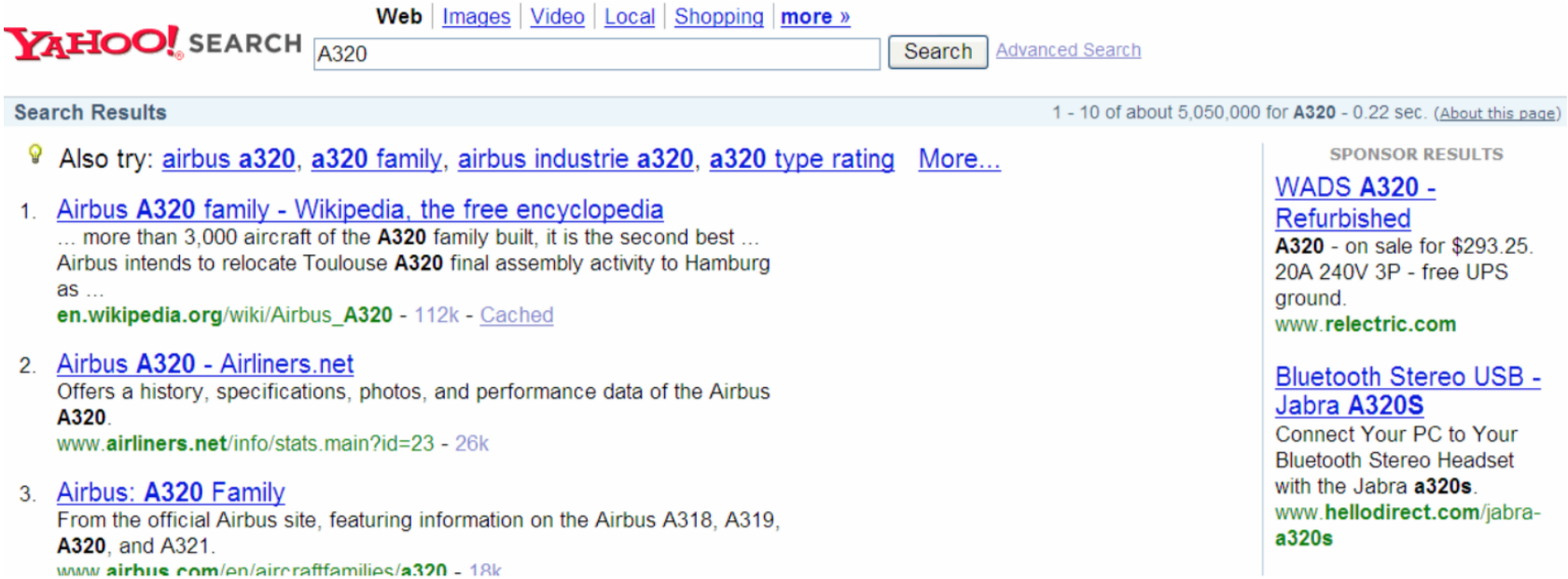
Modelul economic de căutare sponsorizată al companiei Goto (preluat apoi de majoritatea motoarelor de căutare):

- acceptă licitații de la companii pentru a le afișa paginile web ca reclame de căutare pentru anumiți termeni de căutare q
- afișarea reclamelor de căutare în ordinea descrescătoare a sumelor licitate
- Compania Goto era plătită după modelul economic CPC

Motoarele actuale de căutare afișează două tipuri de răspunsuri la o întrebare q:

- rezultate ale căutării algoritmice (sau căutare pură) ca răspuns primar la solicitarea utilizatorul
- rezultate ale căutării sponsorizate: de obicei acestea sunt afișate separat, în dreapta rezultatelor algoritmice
  - căutarea sponsorizată se bazează pe versiuni mai sofisticate ale modelulului Goto de reclamă sponsorizată

Tipuri de căutare
Exemplu ilustrat



Căutarea orientată pe cerințele utilizatorilor

Majoritatea utilizatorilor WWW sunt *neprofesioniști*:

- nu sunt interesați să folosească sintaxa sofisticată a limbajelor artificiale de interogare web (operatori booleeni, wildcard-uri, etc.); tind să folosească 2 sau 3 cuvinte de căutare

$\Rightarrow$ pentru a fi atractive și profitabile, motoarele de căutare s-au adaptat la cerințele utilizatorilor

Strategia de succes a lui Google:

1. A îmbunătățit precizia primelor rezultate afișate
2. A scurtat timpul de localizare a informațiilor de interes, prin afișarea rezultatelor în format text, cu foarte puține elemente grafice

DE REŢINUT: calitatea răspunsurilor unui motor de căutare la o întrebare q se estimează după două criterii:

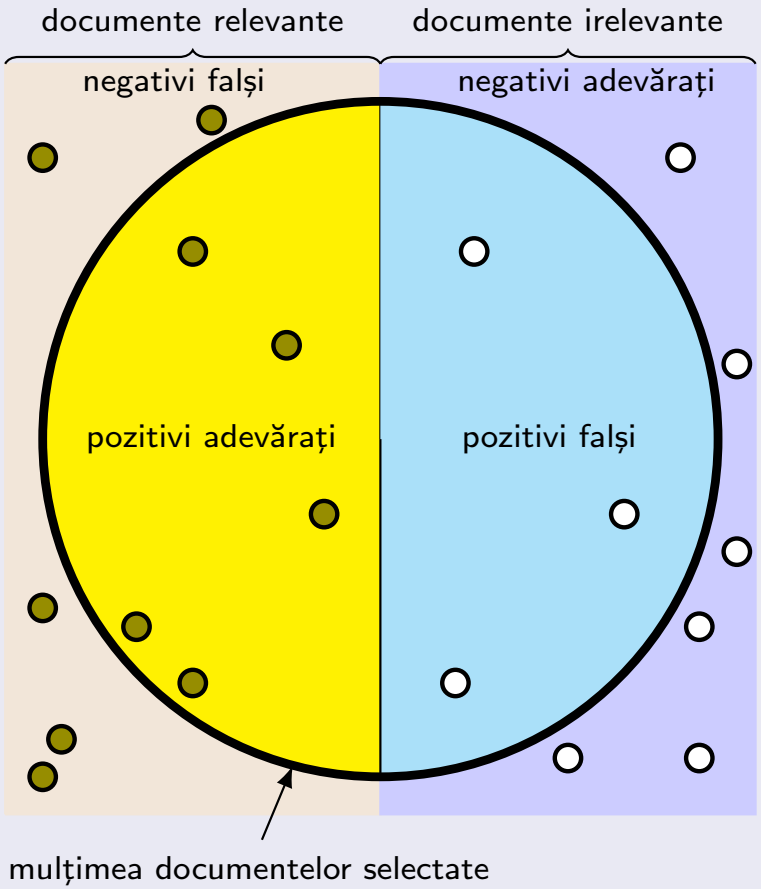
precizie: procentul de răspunsul relevante din totalul răspunsurilor primite

senzitivitate: procentul de răspunsuri relevante din totalul documenelor relevante

(vezi slide-ul următor)

Exemplu ilustrat

nr. total documente $N = 19$; nr. documente relevante $N_r = 10$;
nr documente selectate $N_s = 8$



: document relevant

: document irelevant

$$\text{precizie} = \frac{\text{pozitivi adevărați}}{N_s} = \frac{5}{8} = 0.625$$

$$\text{senzitivitate} = \frac{\text{pozitivi adevărați}}{N_r} = \frac{5}{10} = 0.5$$

Precizia măsoară **exactitatea** sau **calitatea** răspunsurilor.

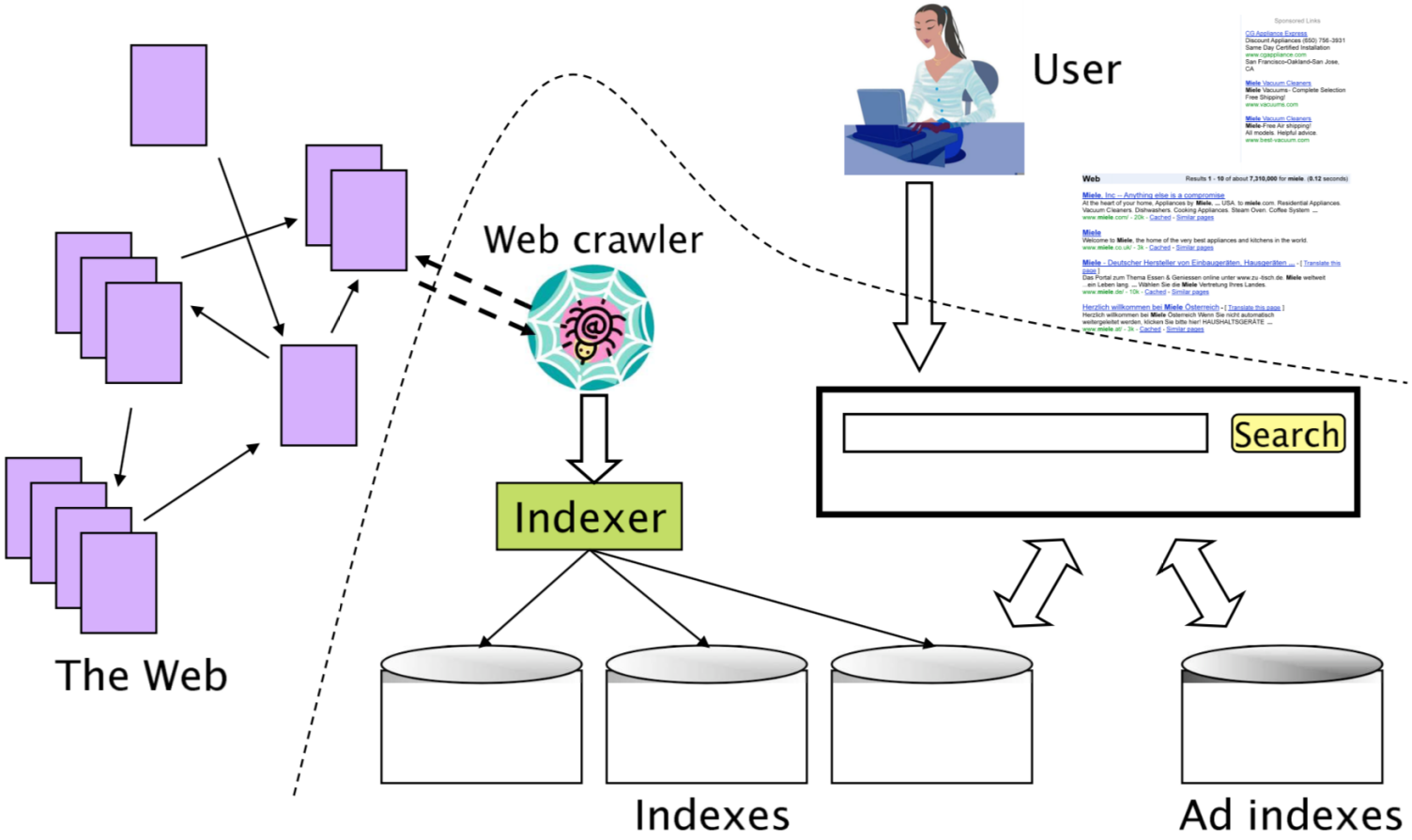
Senzitivitatea măsoară **completitudinea** sau
sau **cantitatea** răspunsurilor.

Tipuri de căutări ale utilizatorilor WWW

1. **Căutări informaționale**: au ca scop găsirea de informații generale despre un subiect larg (de exemplu, istoria imperiului roman); informațiile se extrag din mai multe pagini web.
2. **Căutări navigaționale**: se caută pagina web a unei anumite entități, de ex. pagina web a liniei aeriene KLM.
3. **Căutări tranzacționale**: au ca scop identificarea paginilor web ce ofera anumite servicii, de ex. cumpărarea unui produs, descărcarea unui fișier, efectuarea unei rezervări.

Motoarele de căutare trebuie să poată detecta tipul de căutare dorit de utilizator, și să își adapteze căutarea.

Structura generală a motoarelor moderne de căutare



REMARCĂ: Rolul și modul de funcționare al componentei numite crawler sunt descrise în alt curs.

Studii comparative ale motoarelor de căutare

După mărime =numărul de pagini web statice indexate

(-) Mărimea indexului variază rapid și nu este informație publică

(+) Există algoritmi de estimare a indexului.

Remarcă

Întrebarea: *„Care din motoarele de căutare A sau B are o mărime mai mare a documentelor indexate?”* nu are un răspuns precis, din mai multe motive:

- De regulă, se indexează doar o mică parte a conținutului paginilor web (câteva mii de cuvinte de la început)
- Indecșii sunt organizați pe nivele și în partiții distribuite pe mai multe mașini $\Rightarrow$ estimarea mărimii este un proces dificil

$\Rightarrow$ s-au inventat mai multe metode de măsurare aproximativă a mărimii indecșilor motoarelor de căutare.

Să se compare mărimile indeșilor motoarelor de căutare E_1 și E_2:

$$\frac{|E_1|}{|E_2|} = ? \quad \text{unde } |E_i| = \text{mărimea indexului motorului de căutare } E_i$$

Presupunem că

1. Fiecare motor E_i indexează o fracțiune de documente D_i publicate pe WWW
2. Fracțiunile de documente D_i sunt alese uniform și aleator.

Metoda experimentală capture-recapture este definită astfel:

- alege aleator a pagina din indexul lui E_1 și verifică dacă este în indexul lui $E_2 \Rightarrow x\%$ din paginile în E_1 sunt în E_2
- alege aleator a pagina din indexul lui E_2 și verifică dacă este în indexul lui $E_1 \Rightarrow y\%$ din paginile în E_2 sunt în E_1

Estimăm că $x\,|E_1| \approx y\,|E_2| \Rightarrow \frac{|E_1|}{|E_2|} \approx \frac{y}{x}$

- **Cazul simplu:** x și y sunt calculate de către cineva cu acces la indecșii motoarelor de căutare $\Rightarrow$ x și y se pot calcula ușor, alegând aleator documente din indecșii motoarelor E_1 și E_1.
- **Cazul dificil:** putem alege aleator pagini web doar *din exteriorul indecșilor* lui E_1 și E_2
  - alegerea aleatoare a unei pagini web din colecția totală D de documente publicate pe WWW este o problemă dificilă
  - tehnici aproximative de alegere aleatoare a unui document d din D:
    1. Căutări aleatoare
    2. Adrese IP aleatoare
    3. drumuri aleatoare (engl. *random walks*)
    4. interogări aleatoare (engl. *random queries*)

Căutarea aleatoare

- se pornește de la un log de căutare al unui grup (de ex., membrii unui grup de cercetare) care a fost de acord să i se inregistreze toate cererile de căutare
- Se alege aleator o cerere q din log și o pagină răspuns p din log pentru q
- Se trimite cererea q la motorul E_1 și se verifică dacă p apare în lista de răspunsuri $\Rightarrow x\%$ din paginile răspuns din log sunt indexate de către E_1
- Se trimite cererea q la motorul E_2 și se verifică dacă p apare în lista de răspunsuri $\Rightarrow y\%$ din paginile răspuns din log sunt indexate de către E_2

DEFICIENŢĂ: calculul lui x și y nu este aleator și uniform:

- depinde de interesele de căutare ale grupului ales.

Adrese IP aleatoare

- Se generează adrese IP aleatoare și se trimite o cerere la serverul web de la adresa aleasă $\Rightarrow$ se colectează toate paginile web de pe serverul respectiv
- Se combină colecțiile de pagini web culese de pe servere într-o colecție D
- se estimează

  x: ce procent de pagini din D sunt indexate de către E_1;

  y: ce procent de pagini din D sunt indexate de către E_2

Deficienţe

1. (determinată de virtual hosting): host-uri diferite de pagini web au aceeași adresă IP
2. Unele site-uri web au un nr. mic de pagini web $\Rightarrow$ documentele din D nu sunt alese aleator și uniform

Drumuri aleatoare (engl. *random walks*)

Tehnică adecvată pentru grafuri orientate puternic conexe ($\exists$ drum de la orice nod la orice alt nod)

- Un drum aleator care pornește de la o pagină arbitrar aleasă converge la o distribuție stabilă.

DEFICIENȚE:

- Graful web nu este graf orientat puternic conex
- Chiar și dacă ar fi, timpul necesar pt. a aproxima bine starea de distribuție stabilă este, în general mare și dificil de estimat.

Interogări aleatoare (engl. *random queries*)

Idee de bază: alegerea (aproape) uniform aleatoare a paginilor web indexate de un motor de căutare se face trimițând interogări aleatoare la motorul de căutare; de exemplu:

- din primele 100 răspunsuri ale lui E_1 la o interogare conjunctivă aleatoare q, alegem aleator o pagină p.
- verificăm dacă pagina p este indexată de E_2, astfel:
  - alegem 6-8 termeni cu frecvență joasă de apariții în p, și îi folosim pt. a construi o interogare conjunctivă pentru E_2

Evitarea indexării documentelor duplicate

- Estimare: aprox. 40% din paginile indexate web sunt duplicate ale altor pagini.
- Se dorește indexarea copiilor multiple ale aceluiași conținut
  - $\Rightarrow$ reducere a spațiului alocat pt. listele de indecși inversați, și a timpilor de preprocesare.

Tehnici de detecție a documentelor duplicate

Amprentarea (engl. *fingerprinting*)

- calculul, pt. fiecare pagină web, a unei amprente, care este un rezumat (digest) succint (de ex., 64 biți) al conț. de caractere al paginii respective.
- dacă două documente au aceeeași amprentă, se verifică dacă sunt duplicate.

DEFICIENŢE:

- Nu detectează ca fiind duplicate documentele care diferă prin modificarea câtorva caractere (engl. *near duplication*)
- Această deficiență poate fi evitată cu shingling

Pt. un întreg $k > 1$, mulțimea de k-shingles a unui document d este mț. de secvențe consecutive de 4 termeni care apar în d.

- Dacă d conține `a rose is a rose is a rose`, atunci mț. de 4-shingles a lui d este

  `a rose is a`
  `rose is a rose`
  `is a rose is`

Intuiție: două documente sunt aproape duplicate (engl. *near duplicates*) dacă mulțimile de *shingles* generate din ele sunt similare.

- avem nevoie de metode eficiente de calcul și comparare a mț. de shingles ale paginilor web.

Bibliografie

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Capitolul 19: *Web search basics* din An Introduction to Information Retrieval.
Ediție online (c) 2009 Cambridge UP.