5. Căutarea pe web

Puncte importante :

- *Rangul paginii*, pentru descoperirea celor mai "importante" pagini de web, utilizat de Google.
- *Indecşi şi autorităţi*, o evaluare mai detaliată a importanţei paginilor web utilizând o variantă a calculului de valori proprii utilizată pentru rangul paginii.

5.1. Rangul paginii

Intuitiv rezolvăm problema definiţiei "importanţei" recursiv : o pagina este importantă dacă pagini inportante conţin legături către ea.

Creăm o matrice stochastică a Internetului astfel :

1. Fiecare pagină i corespunde liniei i şi coloanei i a matricii.
2. Daca pagina j are n succesori (legături), atunci elementul i, j al matricii este $1/n$ daca pagina i este unul dintre aceşti succesori ai paginii j şi 0 altfel.

Intuiţia care stă în spatele acestei matrici este :

- Să ne imaginăm că iniţial fiecare pagină are o unitate de importanţă. La fiecare pas fiecare pagină îşi împarte importanţa între succesorii săi şi primeşte noi fracţiuni de importanţă de la predecesorii săi.
- Eventual, importanţa fiecărei pagini atinge o limită care este componenta corespunzătoare ei din vectorul principal de valori proprii al matricii.
- Această importanţă este de asemenea probabilitatea ca un navigator pe web, pornind de la o pagină aleatoare şi urmând legături aleator alese din fiecare pagină, să ajungă la pagina în discuţie după o lungă serie de legături.

Exemplul 5.1 : In 1839 Internetul consta din doar trei pagini : Netscape, Amazon si Microsoft. Legăturile între aceste trei pagini erau ca în Fig. 9.

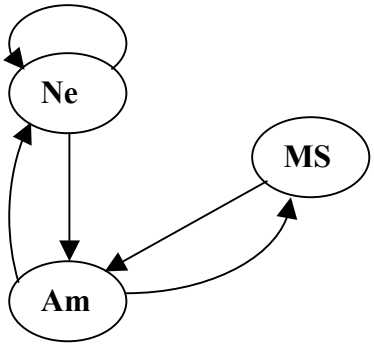

Figura 9: Internetul în 1839

Fie $[n, m, a]$ vectorul importanţei pentru cele trei pagini : Netscape, Microsoft respectiv Amazon. Atunci ecuaţia care descrie valorile asimptotice ale acestor trei variabile este :

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

De exemplu, prima coloană a matricii reflectă faptul că Netscape îşi divide importanţa între el însuşi şi Amazon. A doua coloană că Microsoft dă toată importanţa sa către Amazon.

Putem rezolva ecuaţii ca aceasta începând cu aserţiunea că n = m = 1 şi aplicând repetat matricea la estimarea curentă a acestor valori. Primele patru iteraţii dau următoarele estimări :

n	=	1	1	5/4	9/8	5/4
m	=	1	1/2	3/4	1/2	11/16
a	=	1	3/2	1	11/8	17/16

La limită, soluţia este $n = a = 6/5$; $m = 3/5$. Adică Netscape şi Amazon au fiecare aceeaşi importanţă si de două ori mai mare decât importanţa Microsoft (ei, asta se întampla în 1839).

- De notat că nu putem să obţinem niciodată valorile absolute ale lui n, m şi a ci doar raportul lor, de vreme ce aserţiunea iniţială că fiecare a pornit de la 1 a fost arbitrară.
- Deoarece matricea este stochastică (suma pe fiecare coloană este 1), procesul de *relaxare* de mai sus converge către vectorul principal de valori proprii.

5.2. Probleme cu grafuri reale ale Internetului

1. *Dead end*: o pagină care nu are succesori nu are către cine să-şi trimită importanţa. Eventual, toată importanţa "se va scurge" din Internet
2. *Capcane* : un grup de una sau mai multe pagini care nu au legături către pagini din afara grupului vor acumula eventual toată importanţa din Internet.

Exemplul 5.2 : Să presupunem că Microsoft încearcă să profite că este un monopol înlăturând toate legăturile din situl său. Noul Internet este ca in Fig. 10 iar matricea descriind tranziţiile este :

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

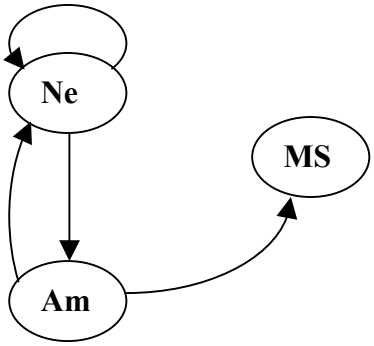

Figura 10: Microsoft devine dead end.

Primii patru paşi ai soluţiei iterative sunt :

n	=	1	1	3/4	5/8	1/2
m	=	1	1/2	1/4	1/4	3/16
a	=	1	1/2	1/2	3/8	5/16

Eventual, fiecare dintre n, m şi a devin 0; i.e. toată importanţa se scurge afară.

Exemplul 5.3 : Supărat de decizie, Microsoft decide să nu folosească decăt legături către el însuşi de acum încolo. Acum, Microsoft a devenit o capcană. Noul Internet este in Fig. 11, iar ecuaţia de rezolvat este:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

Primii patru paşi ai soluţiei sunt:

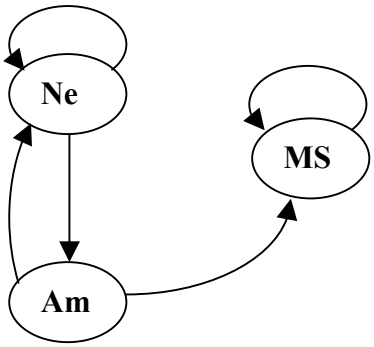

Figura 11: Microsoft devine o capcană.

n	=	1	1	3/4	5/8	1/2
m	=	1	3/2	7/4	2	35/16
a	=	1	1/2	1/2	3/8	5/16

Acum m converge la 3 iar $n = a = 0$.

5.3. Soluţia Google pentru dead end şi capcane

În loc de a aplica matricea direct, "taxăm" fiecare pagină cu o fracţiune din importanţa sa curentă şi distribuim importanţa taxată în mod egal tuturor paginilor.

Exemplul 5.4 : Dacă folosim o taxă de 20% ecuaţia din exemplul 5.3 devine:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = 0.8 \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

Soluţia acestei ecuaţii este $n = 7/11$; $m = 21/11$; $a = 5/11$.

- De notat că suma celor trei valori nu este 3 dar obţinem o distribuţie mult mai rezonabilă a importanţei decăt în Exemplul 5.3.

5.4. Procedee anti-spam la Google

"Spamming" este în acest context încercarea multor situri web de a părea că sunt despre un subiect care atrage navigatorii fără ca într-adevăr să fie despre acel subiect.

- Google, ca şi alte motoare de căutare, încearcă să potrivească cuvintele din cererile de căutare cu cuvinte din pagini web. Cu toate acestea, Google, spre deosebire de alte motoare de căutare, tinde să creadă ceea ce spun alţii în textul legăturilor despre o pagină web facând mai greu pentru aceasta să pară ca fiind despre ceva ce nu este.
- Utilizarea rangului paginii pentru a măsura importanţa în locul unei măsuri mult mai naive ca "numărul de legături către acea pagină" protejează de asemenea impotriva spamului. Masura naivă poate fi înşelată de un spammer care creează 1000 de pagini care se referă între ele în timp ce rangul paginii recunoaşte că nici una dintre acestea nu au importanţă reală.

5.5. Indecşi si autorităţi

Intuitiv, definim "index" şi "autoritate" într-un mod mutual recursiv: un index conţine legături către multe autorităţi iar o autoritate este referită de mulţi indecşi.

- Autorităţile pot fi pagini care oferă informaţii despre un subiect, e.g. pagina Quest despre proiectul IBM de data mining.
- Indecşii sunt pagini care nu furnizează informaţii ci spun unde se găsesc informaţii, e.g. pagina cursului CS345.
- Utilizează o formalizare matricială similară cu cea de la rangul paginii dar fără restricţia stochastică. Numărăm fiecare legătură ca 1, indiferent de câţi succesori sau predecesori are o pagină.
- Aplicarea repetată a matricii duce la divergenţă, dar putem introduce un factor de scalare pentru a ţine valorile calculate pentru gradul de "autoritate" sau de "indexare" pentru fiecare pagina între limite finite.

Definim matricea A ale cărei linii şi coloane corespund paginilor web având elementul $A_{ij} = 1$ dacă pagina i referă pagina j si 0 altfel.

- De notat că A^T, transpusa lui A, arată ca matricea utilizată pentru calculul rangului paginilor dar A^T are 1 acolo unde matricea pentru rang are fracţii.

Fie a si h doi vectori iar componenta lor i corespunde gradului de autoritate respectiv indexare a paginii i. Fie λ şi μ factorii de scalare corespunzători care vor fi calculaţi mai târziu. Atunci putem afirma că:

1. $h = \lambda A a$. Adică gradul de indexare al fiecărei pagini este suma gradelor de autoritate ale tuturor paginilor referite, scalată cu λ.
2. $a = \mu A^T h$. Adică gradul de autoritate al fiecărei pagini este suma gradelor de indexare ale tuturor paginilor care o referă, scalată cu μ.

Din (1) şi (2) putem deduce folosind substituţia, două ecuaţii care leagă vectorii a si h doar de ei înşişi:

$$a = \lambda \mu A^T A a \; ; \; h = \lambda \mu A A^T h$$

Ca urmare, putem calcula h şi a prin relaxare, obţinând vectorul principal de valori proprii al matricilor AA^T şi respectiv A^TA

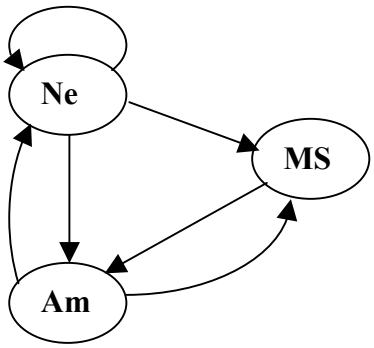

Figura 12: Internetul pentru exemplul 5.5

Exemplul 5.5 : Considerăm Internetul ca în Fig. 12. Matricile relevante sunt:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad AA^T = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \quad A^TA = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Dacă utilizăm $\lambda = \mu = 1$ şi considerăm că vectorii $h = [h_n, h_m, h_a]$ şi $a = [a_n, a_m, a_a]$ sunt iniţial fiecare [1, 1, 1], primele trei iteraţii ale ecuaţiilor pentru a şi h sunt:

$$
\begin{array}{llllll}
a_n & = & 1 & 5 & 24 & 114 \\
a_m & = & 1 & 5 & 24 & 114 \\
a_a & = & 1 & 4 & 18 & 84
\end{array}
$$

$$
\begin{array}{llllll}
h_n & = & 1 & 6 & 28 & 132 \\
h_m & = & 1 & 2 & 8 & 36 \\
h_a & = & 1 & 4 & 20 & 96
\end{array}
$$

De exemplu, vectorul a, scalat corespunzător, va converge către un vector în care $a_n = a_m$ si fiecare dintre aceste numere este mai mare ca a_a în raportul $1 + \sqrt{3} : 2$ sau aproximativ 1.36.