Cursul 14

Algoritmul PageRank

14.1 Construcţia lanţului Markov pe graful WEB. Algoritmul PageRank

Succesul extraordinar şi dominaţia motorului Google se datorează în principal algoritmului PageRank, care exploatează structura linkurilor din WWW pentru a determina un indice de popularitate al fiecărei pagini, independent de interogarea formulată de utilizator.

Documentele de pe WEB (paginile WEB) sunt identificate de aplicaţiile software ale motorului, numite roboţi sau *crawlere*. Documentele sunt apoi indexate. Modulul de indexare extrage cuvintele cheie, constituind aşa numitul sac de cuvinte. Un alt modul, numit *query module* (modulul de interogare), converteşte cererea formulată de utilizator, în limbaj natural, într-un vector cerere, cu care consultă indexul de conţinut şi extrage paginile relevante cererii. Modulul de ierarhizare ordonează descrescător aceste pagini în funcţie de coeficienţii de popularitate. PageRank-ul este un vector ale cărui coordonate sunt coeficienţii de popularitate ai paginilor WEB identificate de crawler. Acest vector este distribuţia de echilibru a unui lanţ Markov definit pe graful WEB.

Să definim mai întâi lanţul Markov ce stă la baza algoritmului `PageRank`. Considerăm $W = \{1, 2, \ldots, m\}$ mulţimea tuturor paginilor WEB, $H = (h_{ij})$ matricea de conectivitate a lui W sau matricea hyperlink:

$$h_{ij} = \begin{cases} 1, & \text{dacă există link în pagina i către pagina j,} \\ 0, & \text{dacă nu există link în pagina i către pagina j.} \end{cases}$$

Se spune că H este o matrice rară, căci are foarte multe zerouri (în medie, 3-10 elemente sunt nenule pe o linie). Suma elementelor de pe linia i a matricei H indică numărul de out-linkuri, adică numărul de linkuri din pagina i către alte pagini sau ea însăşi. Notăm această sumă cu $r_i = \sum_{j=1}^{m} h_{ij}$. r_i se numeşte ordinul ieşirilor din pagina i. Suma elementelor de pe coloana i a matricei hyperlink indică numărul de in-linkuri ale paginii i, adică numărul de linkuri către pagina i.

Larry Page şi Serghei Brin au definit un mers aleator pe graful WEB considerând că un surfer ajuns în pagina i alege cu aceeaşi probabilitate oricare din paginile către care

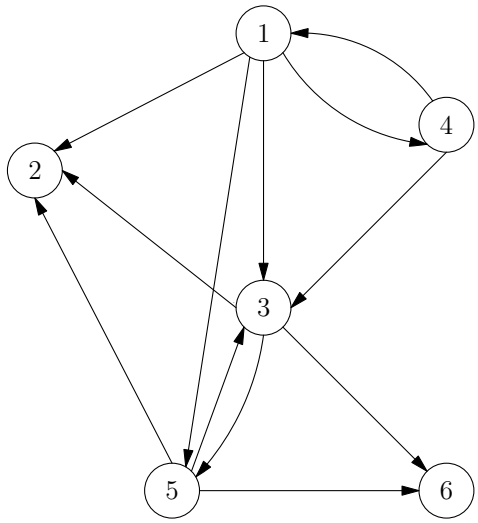

Fig.14.1: Graf orientat ilustrând linkurile între 6 pagini WEB.

aceasta are linkuri, prin urmare probabilitatea de a trece din pagina i în pagina j este:

$$p_{ij} = \begin{cases} \dfrac{1}{r_i}, & \text{dacă există link în pagina i către pagina j,} \\ 0, & \text{dacă nu există link în pagina i către pagina j.} \end{cases}$$

De exemplu, dacă

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

atunci ordinul de ieşire din pagina 5 este $r_5 = 3$, deci probabilitatea de a trece din pagina 5 în oricare din paginile $\{1, 2, \ldots, 6\}$ este $p_{5j} = h_{5j}/3$. Altfel spus, cu aceeaşi probabilitate de $1/3$ un surfer poate trece din pagina 5 în pagina 2, 3 sau 6.

Vom exemplifica construcţia propusă de L. Page şi S. Brin prin modelul simplu de reţea izolată de pagini WEB (reţea intranet) din Fig. 14.1.

Notăm cu $Q = (p_{ij})_{i,j=\overline{1,6}}$, matricea probabilităţilor de tranziţie definite mai sus. Se observă din structura grafului de conectivitate că paginile 2 şi 6 sunt pagini ce nu conţin linkuri către alte pagini. Acestea se numesc *dangling pages*. De exemplu, fişierele `pdf`, `ps` sau fişierele imagine sunt pagini dangling. Prin urmare, liniile 2 şi 6 din matricea de tranziţie au toate elementele nule şi, astfel, Q nu este o matrice stochastică, deci nu poate fi interpretată ca matricea de tranziţie a unui lanţ Markov cu spaţiul stărilor $\{1, 2, 3, 4, 5, 6\}$.

Q=

	1	2	3	4	5	6
1	0	1/4	1/4	1/4	1/4	0
2	0	0	0	0	0	0
3	0	1/3	0	0	1/3	1/3
4	1/2	0	1/2	0	0	0
5	0	1/3	1/3	0	0	1/3
6	0	0	0	0	0	0

Pentru a remedia această problemă, Page şi Brin au propus ca vector de probabilitate dintr-o pagină dangling i distribuţia uniformă, considerând

$$p_{ij} = 1/m,\ j = \overline{1,m}.$$

Adică, în mod artificial, se adaugă linkuri dintr-o pagină dangling către toate paginile WEB sau, echivalent, ajuns într-o pagină dangling, un navigator poate apoi alege cu o probabilitate uniformă orice pagină din WWW. Astfel, matricea stochastică obţinută din matricea Q este:

$\tilde{Q}$=

	1	2	3	4	5	6
1	0	1/4	1/4	1/4	1/4	0
2	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**
3	0	1/3	0	0	1/3	1/3
4	1/2	0	1/2	0	0	0
5	0	1/3	1/3	0	0	1/3
6	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**

Lanţul Markov definit de matricea stochastică $\tilde{Q}$ nu este în general ireductibil (adică nu există drum de linkuri între orice două pagini sau, echivalent, graful WEB nu este tare conex) şi pot exista traiectorii periodice, adică surferul, navigând conform matricei de tranziţie $\tilde{Q}$, ar putea fi prins, ca într-o capcană, într-o mişcare aleatoare ciclică. Din acest motiv, dar şi pentru că la un moment dat, şi în realitate, orice surfer renunţă să navigheze urmând linkurile din pagini, L. Page şi S. Brin au introdus ipoteza că doar cu probabilitatea $\alpha \in (0,1)$ surferul navighează conform matricei $\tilde{Q}$ şi cu probabilitatea $1-\alpha$ ignoră linkurile şi alege cu probabilitate uniformă oricare din paginile de pe WEB, introducând adresa URL în linia de comandă a browser-ului. Probabilitatea α se numeşte *factor de damping*. În lucrarea iniţială a fondatorilor Google, α era menţionat ca având valoarea 0.85. Cu această modificare matricea de tranziţie este:

$$G = \alpha\tilde{Q} + (1-\alpha)\underbrace{\begin{bmatrix} 1/m & 1/m & \dots & 1/m \\ 1/m & 1/m & \dots & 1/m \\ & & \vdots & \\ 1/m & 1/m & \dots & 1/m \end{bmatrix}}_{E}.$$

Matricea G se numeşte matricea Google, iar matricea E, de elemente identice $1/m$, se numeşte *matricea de teleportare*, deoarece surferul se teleportează din navigarea aleatoare urmând linkuri într-o "navigare artificială". Evident că şi matricea E este o matrice stochastică, iar G fiind o combinaţie convexă de astfel de două matrice este o matrice stochastică (vezi proprietăţile matricelor stochastice). Mai mult, $G(i,j) > 0, \forall\ i,j = \overline{1,m}$, deci matricea Google este ireductibilă şi aperiodică.

Se presupune că matricea Google este cea mai "uriaşă" matrice cu care se lucrează în vreo aplicaţie la ora actuală.

Lanţul Markov având:

• spaţiul stărilor constituit din mulţimea paginilor WEB, de cardinal m

• matricea de tranziţie de tipul G, cu α fixat

• distribuţia iniţială de probabilitate π_0 (distribuţia uniformă sau oricare alta)

este un lanţ ireductibil şi aperiodic, deci are o unică distribuţie de echilibru π, numită vectorul PageRank.

PageRank–ul, π, este limita şirului (π_n), definit prin $\pi_n^T = \pi_0^T G^n$. Limita este aceeaşi indiferent de distribuţia iniţială de probabilitate π_0, adică indiferent cu ce probabilitate surferul alege pagina din care începe navigarea. $\pi(j)$ se numeşte PageRank-ul paginii j şi reprezintă şansa asimptotică pe care o are pagina j de a fi vizitată de navigatorul aleator sau proporţia din timpul de navigare pe care surferul ar petrece-o vizitând pagina j. Deci, $\pi(j)$ este un indice de popularitate al paginii.

Când un utilizator introduce cuvinte cheie în bara de căutare, motorul Google caută paginile ce conţin cuvintele cheie şi le afişează în ordinea descrescătoare a PageRank-ului lor.

Remarcăm că PageRank-ul unei pagini este independent de interogarea formulată de utilizator. Ea depinde doar de structura grafului WEB şi se poate calcula offline. PageRank-ul se calculează la intervale regulate de timp. Până în 2008 se calcula lunar, dar acum se actualizează la intervale mai scurte de timp.

Vectorul PageRank se calculează numeric, folosind aşa numita metodă a puterii, adică se calculează recursiv. Pornind de la π_0 şi G, se determină distribuţiile (sau PageRank-ul la n paşi de navigare) $\pi_n^T = \pi_{n-1}^T G$. Se consideră că metoda a atins stadiul de convergenţă (adică s-a ajuns la echilibru) într-o etapă n în care $\|\pi_n - \pi_{n-1}\| < \varepsilon$, unde ε este un număr pozitiv foarte mic, prescris.

Pseudocodul algoritmului de calcul al PageRank-ului este:

```
1: function PageRank(G, m);
2:     π = [1/m, 1/m, ..., 1/m]; //Distributia initiala de probabilitate
3:     eps = 10^-7;
4:     do
5:         π' = π;
6:         π = π' * G;
7:     while (||π − π'|| >= eps);
8:     return π;
9: end function
```

S-a demonstrat că viteza de convergenţă a metodei puterii este aceeaşi cu rata de convergenţă a lui α^n, unde α este factorul de damping.

Implicaţii asupra PageRank-ului. Din punctul de vedere al vitezei de convergenţă ar fi preferabil un factor α cât mai apropiat de zero. În acest caz, ţinând seama că matricea Google este $G = \alpha\tilde{Q} + (1-\alpha)E$, ar rezulta că se acordă o pondere redusă, α, navigării conform linkurilor din graful WEB (cu modificarea pentru pagini dangling) şi o pondere mai mare navigării artificiale, conform matricei de teleportare E. Cu alte cuvinte, în acest caz PageRank-ul asociat nu ar reflecta popularitatea reală a paginilor WEB. De aceea o valoare rezonabilă, aşa cum a fost ea aleasă iniţial de Larry Page şi Serghei Brin, $\alpha = 0.85$, conduce la rezultate mai apropiate de realitate şi la o viteză de convergenţă suficient de bună (un reprezentant Google a declarat că metoda puterii converge după 100-200 de iteraţii). Dacă vreti să aflaţi PageRank-ul unor pagini WEB intraţi aici:

`http://www.prchecker.info/check_page_rank.php`

14.2 Pagerank-ul personalizat

Pentru o ierarhizare personalizată a paginilor WEB, matricea E se calculează luând în considerare vectorul personalizat w, care este un vector probabilist $w = [a_1, a_2, \ldots, a_m]^T$ ale cărui coordonate reprezintă probabilitatea ca surferul, ce iese din navigarea conform linkurilor, să aleagă pagina $1, 2, \ldots, m$ din WEB. Cu alte cuvinte, el nu alege o pagină în mod uniform, ci are anumite preferinţe, identificate de motor în decursul timpului. Astfel, matricea de teleportare va fi $E = ew^T$, unde $e = [1, 1, \ldots, 1]^T$, iar matricea Google corespunzătoare este

$$G = \alpha\tilde{Q} + (1-\alpha)ew^T.$$

Distribuţia de echilibru corespunzătoare este PageRank-ul personalizat.