Cursul 11

Lanţuri Markov. Definiţie, proprietăţi, simulare. Algoritmul PageRank-Google

Lanţurile Markov discrete se folosesc în modelarea şi simularea sistemelor în care se produc evenimente la momente discrete de timp $t = 0, 1, 2, \ldots, n, \ldots$. Astfel de lanţuri se folosesc în design-ul algoritmilor de rutare în reţele de calculatoare, al protocoalelor pentru reţelele wireless, în controlul mişcării roboţilor, ranking-ul paginilor WEB şi a echipelor sportive, în analiza protocoalelor de management a memoriei, studiul performanţei serverelor, în algoritmi randomizaţi, în machine learning (de ex în computational advertising), etc. Există chiar şi un limbaj de modelare pentru sisteme hardware/software în timp real, numit *POOSL, Parallel Object Oriented Specification Language*, care generează un lanţ Markov, ca model al sistemului.

Un sistem este reprezentat de o mulţime finită de stări (noduri), $S = \{1, 2, \ldots, m\}$ sau infinit numărabilă $S = \mathbb{N}$. Mulţimea S se numeşte spaţiul stărilor sau al nodurilor unei reţele. Schimbările de stare se produc la întâmplare, la momente discrete de timp $t = 0, 1, 2, \ldots, n, \ldots$. Noi vom discuta preponderent exemple în care la momente discrete de timp se produce o transmitere de informaţie de la un nod din reţea spre altul sau "un călător virtual" trece de la un nod spre altul. Nodul spre care se transmite informaţia (sau nodul în care trece călătorul) depinde de diverse circumstanţe şi astfel informaţia (călătorul) are o traiectorie aleatoare în reţea.

Fiecărui moment de timp $n \in \mathbb{N}$ i se asociază o variabilă aleatoare X_n ce ia valori în mulţimea nodurilor:

$$X_n = \begin{pmatrix} 1 & 2 & \ldots & m \\ \pi_n(1) & \pi_n(2) & \ldots & \pi_n(m) \end{pmatrix}$$

unde $\pi_n(i)$ este probabilitatea ca la momentul n informaţia să atingă (să ajungă) în nodul $i \in S$.

Definiţia 11.0.1 *Un şir de variabile aleatoare* (X_n), $n \in \mathbb{N}$, *definite pe acelaşi spaţiu de probabilitate,* $(\Omega, \mathcal{K}, P)$, *cu valori în mulţimea stărilor (nodurilor)* S, *defineşte un lanţ Markov discret dacă probabilitatea ca informaţia să treacă la momentul* $n+1$ *în nodul* j *ştiind că în momentele de timp anterioare se afla respectiv în nodurile* $s_0, s_1, \ldots s_{n-1}, i$, *este:*

$$P(X_{n+1} = j | X_0 = s_0, X_1 = s_1, \ldots X_n = i) = P(X_{n+1} = j | X_n = i) \tag{11.1}$$

$P(X_{n+1} = j|X_n = i)$ se citeşte, "probabilitatea ca informaţia să treacă în nodul j la momentul $n+1$, ştiind că se află în nodul i la momentul n".

Relaţia (11.1) se numeşte proprietate markoviană. Proprietatea markoviană caracterizează "lipsa parţială de memorie" a lanţului: cunoscând succesiunea de noduri, $s_0, s_1, \dots, s_{n-1}, i$, prin care a trecut informaţia până la momentul n, doar nodul în care se află în prezent, $(X_n = i)$, influenţează probabilitatea de trecere, în momentul următor, într-un alt nod, nu şi drumul parcurs până la momentul curent, n (cu alte cuvinte doar istoria recentă, nu şi cea trecută, influenţează evoluţia viitoare).

În continuare ne referim la lanţuri Markov cu o mulţime finită de stări. În mod normal probabilităţile $P(X_{n+1} = j|X_n = i)$ depind n, adică

$$P(X_{n+1} = j|X_n = i) = p_{ij}(n)$$

Un lanţ Markov discret (X_n) se numeşte *lanţ Markov omogen*, dacă probabilităţile condiţionate:

$$P(X_{n+1} = j|X_n = i)$$

nu depind de n.

În continuare în loc de trecerea informaţiei sau călătorului virtual, de la un nod la altul, spunem trecerea lanţului Markov.

Pentru un lanţ Markov omogen, notăm cu $p_{ij} = P(X_{n+1} = j|X_n = i)$, $i, j = \overline{1, m}$, probabilitatea ca la momentul $n+1$ lanţul Markov să treacă în nodul j, ştiind că la momentul n se afla în nodul i. p_{ij} se numeşte probabilitate de trecere într-un singur pas, din nodul i în nodul j, iar matricea Q, de elemente $Q(i, j) := p_{ij}$, $i, j = \overline{1, m}$, se numeşte *matricea de tranziţie a lanţului Markov*.

În concluzie, un lanţ Markov discret defineşte o lege de mişcare la întâmplare pe mulţimea nodurilor.

Matricea de tranziţie, $Q = (p_{ij})_{i,j \in S}$, are proprietăţile:

1) $p_{ij} \geq 0, \forall\ (i, j) \in S \times S$;

2) $\sum_{j=1}^{m} p_{ij} = 1$, $\forall\ i \in S$, adică suma elementelor de pe fiecare linie este 1.

O astfel de matrice se numeşte *matrice stochastică*, iar liniile ei, *vectori stochastici*. $p_{i1}, p_{i2}, \dots, p_{im}, i = \overline{1, m}$, indică probabilităţile ca din starea i sistemul să treacă respectiv în stările $1, 2, \dots, m$.

Mulţimea nodurilor, S, a unui lanţ Markov şi matricea de tranziţie definesc un graf orientat. Există arc orientat de la nodul i la j, dacă probabilitatea p_{ij} este nenulă. Graful astfel asociat se numeşte *graf de tranziţie al lanţului Markov*.

Exemplul 1. Fie $S = \{1, 2, 3\}$ mulţimea nodurilor unei reţele şi Q matricea de tranziţie de la un nod la altul:

$$Q = \begin{pmatrix} 0.4 & 0.5 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.4 & 0.4 & 0.2 \end{pmatrix}$$

Graful asociat este vizualizat în Fig. 11.1.

Proprietăţi ale matricilor stochastice

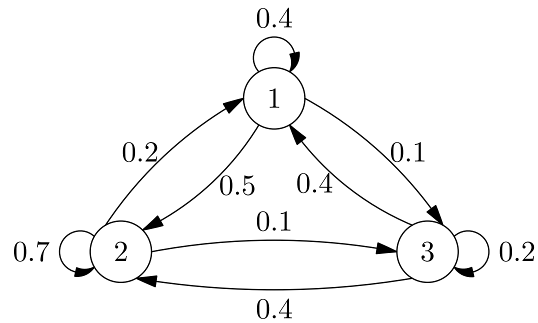

Fig.11.1: Graful de tranziţie al unui lanţ Markov. Pe fiecare arc este indicată probabilitatea de trecere între nodurile conectate de arc.

• Notăm cu $\mathbf{e}$ vectorul având toate coordonatele egale cu 1, $\mathbf{e} = [1, 1, \ldots, 1]^T$. Produsul $Q\mathbf{e}$ este:

$$\begin{pmatrix} p_{11} & p_{12} & \ldots & p_{1m} \\ \vdots & & & \\ p_{k1} & p_{k2} & \ldots & p_{km} \\ \vdots & & & \\ p_{m1} & p_{m2} & \ldots & p_{mm} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} p_{11} + p_{12} + \cdots + p_{1m} \\ \vdots \\ p_{k1} + p_{k2} + \cdots + p_{km} \\ \vdots \\ p_{m1} + p_{m2} + \cdots + p_{mm} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Această relaţie ne permite exprimarea concentrată a proprietăţii lui Q de a avea toate liniile vectori stochastici:

$$Q\mathbf{e} = \mathbf{e}$$

Această relaţie exprimă şi faptul că e este vector propriu al matricii Q, corespunzător valorii proprii $\lambda = 1$. În continuare o vom folosi ca relaţie de definiţie a unei matrici stochastice (subînţelegând că elementele ei sunt mai mari sau egale cu zero).

• Produsul a două matrici stochastice P, Q, este o matrice stochastică, deoarece $P\mathbf{e} = \mathbf{e}$ şi $Q\mathbf{e} = \mathbf{e}$ implică $(PQ)\mathbf{e} = P(Q\mathbf{e}) = P\mathbf{e} = \mathbf{e}$.

Ca o consecinţă a acestei proprietăţi avem că dacă Q este matricea de tranziţie a unui lanţ Markov, atunci şi Q^n, este matrice stochastică pe linii, $\forall\, n \in \mathbb{N}$.

• Dacă P, Q sunt matrici stochastice şi $\alpha \in (0, 1)$ atunci combinaţia convexă:

$$M = \alpha P + (1 - \alpha)Q$$

este matrice stochastică.

O realizare a lanţului Markov, (X_n), sau o observaţie asupra lanţului este un şir de noduri ce pot fi vizitate de lanţ, $(s_0, s_1, \ldots, s_n, \ldots)$, $s_k \in S$, şi se numeşte **traiectorie a lanţului**.

Pentru a putea analiza şi simula un lanţ Markov, trebuie precizată **distribuţia iniţială de probabilitate**, care dă probabilitatea ca traiectoria aleatoare a lanţului să pornească dintr-un nod i. Mai precis, distribuţia iniţială de probabilitate este un vector probabilist (vector cu coordonatele în [0,1] şi suma coordonatelor egală cu 1):

$$\pi_0 = [\pi_0(1), \pi_0(2), \ldots, \pi_0(m)]^T, \quad \pi_0(k) \geq 0, \quad \sum_{k=1}^{m} \pi_0(k) = 1,$$

unde $\pi_0(k) = P(X_0 = k)$ este probabilitatea ca la momentul $t = 0$ lanţul să pornească din nodul k. Cu alte cuvinte, distribuţia iniţială de probabilitate a lanţului Markov este distribuţia de probabilitate a variabilei aleatoare discrete X_0:

$$X_0 = \begin{pmatrix} 1 & 2 & \ldots & m \\ \pi_0(1) & \pi_0(2) & \ldots & \pi_0(m) \end{pmatrix} \tag{11.2}$$

Dacă distribuţia iniţială de probabilitate este , de exemplu, $\pi_0 = [0, 1, 0, \ldots, 0]$, atunci spunem că lanţul porneşte sigur (adică cu probabilitatea 1) din nodul 2.

Dacă pentru lanţul Markov din Exemplul 1 distribuţia iniţială de probabilitate este:

$$X_0 = \begin{pmatrix} 1 & 2 & 3 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}$$

atunci înseamnă că probabilitatea ca un mers (drum) aleator în mulţimea $S = \{1, 2, 3\}$ să pornească din starea 2 este 0.5.

Un lanţ Markov discret este simulat în mod iterativ. **Algoritmul de simulare a unui LM** este prototip pentru *clasa algoritmilor iterativi aleatori*. Având dat spaţiul stărilor $S = \{1, 2, \ldots, m\}$, distribuţia iniţială de probabilitate $\pi_0 = [\pi_0(1), \pi_0(2), \ldots, \pi_0(m)]^T$ şi matricea de tranziţie a $Q = (p_{ij})$, $i, j = \overline{1, m}$, a unui lanţ Markov (X_n), putem genera o traiectorie aleatoare $s_0, s_1, \ldots, s_N$, astfel:

• construim simulatorul unei variabile aleatoare discrete arbitrare:

$$Y = \begin{pmatrix} 1 & 2 & \ldots & m \\ \mathtt{p_1} & \mathtt{p_2} & \ldots & \mathtt{p_m} \end{pmatrix}$$

pe care îl notăm simbolic

$$\mathtt{simulator}(1, 2, \ldots, \mathtt{m}; \mathbf{p})$$

$\mathbf{p}$ este vectorul de probabilitate al variabilei aleatoare discrete ce ia valori în $\{1, 2, \ldots, m\}$. $\mathtt{j} = \mathtt{simulator}(1, 2, \ldots, \mathtt{m}; \mathbf{p})$ simbolizează faptul că simulatorul generează numărul (starea) j.

• Se generează starea iniţială s_0, din care porneşte traiectoria, simulând variabila aleatoare X_0 definită în (11.2).

Probabilităţile de trecere din starea $i = s_0$ în una din stările sistemului sunt date de elementele din linia i a matricii de tranziţie $Q = (p_{ij})$. Astfel starea la momentul $t = 1$, s_1, este o observaţie asupra variabilei aleatoare discrete,

$$T_1 = \begin{pmatrix} 1 & 2 & \ldots & m \\ Q_{i1} & Q_{i2} & \ldots & Q_{im} \end{pmatrix}$$

ce ia valorile $\{1, 2, \ldots, m\}$ cu probabilităţile din linia i a matricii de tanziţie, etc.

Algoritmul de generare a segmentului de traiectorie $\mathtt{s_0}, \mathtt{s_1}, \ldots, \mathtt{s_n}$, este atunci:

```
function LantMarkov(m, π0, Q, n)
    s0 = simulator(1, 2, ... m; π0);
    i = s0;
    for k = 1 : n
        p = Q[i, :]; //Q[i, :] = [linia i din matricea Q];
        sk = simulator(1, 2, ... m; p);
        i = sk;
    end for
    return s0, s1, ... sn;
end function
```

11.0.1 Analiza unui lanţ Markov

Pe lângă tranziţia într-un singur pas a unui lanţ Markov, suntem interesaţi şi de tranziţia dintr-un nod în altul, în n paşi. Fie

$$P(X_n = j|X_0 = i), \quad i, j \in S$$

probabilitatea ca lanţul să treacă din nodul iniţial i în nodul j după n paşi. Să arătăm că această probabilitate este dată de elementul din poziţia (i, j), a matricii de tranziţie la puterea n.

Propoziţie: $P(X_n = j|X_0 = i) = Q^n[i][j]$

Demonstraţie: Să arătămm mai întâi câ $P(X_2 = j|X_0 = i) = Q^2[i][j]$.

Evenimentul $(X_2 = j|X_0 = i) = \cup_{k=1}^m (X_2 = j, X_1 = k, X_0 = i)$. Deci

$P(X_2 = j|X_0 = i) = \sum_{k=1}^m P(X_0 = i, X_1 = k, X_2 = j) =$

$\sum_{k=1}^m \underbrace{P(X_0 = i)}_{=1} P(X_1 = k|X_0 = i) \underbrace{P(X_2 = j|X_1 = k, X_0 = i)}_{P(X_2=j|X_1=k)} = \sum_{k=1}^m Q[i][k]Q[k][j] = Q^2[i][j]$

Presupunem că egalitatea $P(X_{n-1} = j|X_0 = i) = Q^{n-1}[i][j]$ este adevărată. Pentru a o demonstra şi în cazul n, exprimăm evenimentul:

$$(X_n = j|X_0 = i) = \cup_{k=1}^m (X_n = j, X_{n-1} = k, X_0 = i)$$

Deci la fel ca în cazul n=2 avem:

$P(X_n = j|X_0 = i) = \sum_{k=1}^m P(X_n = j, X_{n-1} = k, X_0 = i) =$
$\sum_{k=1}^m P(X_0 = i)P(X_{n-1} = k|X_0 = i)P(X_n = j|X_{n-1} = k, X_0 = i) = \text{etc}$

că *matricea de tranziţie în n paşi*, este chiar Q^n – matricea de tranziţie într-un pas, ridicată la puterea n. □

Mai mult, $Q^n[i][j] = P(X_n = j|X_0 = i) = P(X_{n+k} = j|X_k = i)$, oricare ar fi k. Deci probabilitatea de a trece în n paşi din nodul i în nodul j nu depinde de momentul în care lanţul este în nodul i.

Pentru a putea face predicţii asupra traiectoriei aleatoare definite de lanţ să calculăm câteva probabilităţi ale unor evenimente de interes.

Propoziţia 11.0.1 *Probabilitatea ca lanţul să evolueze pe traiectoria $s_0, s_1, s_2, \ldots, s_n \in S$ este:*

$$P(X_0 = s_0, X_1 = s_1, \ldots X_{n-1} = s_{n-1}, X_n = s_n) = \pi_0(s_0)Q(s_0, s_1)\ldots Q(s_{n-1}, s_n). \tag{11.3}$$

Demonstraţie: Din formula condiţionării iterate (Vezi cursul relativ la probabilităţi condiţionate) şi a proprietăţii markoviene (11.1) avem:

$$\begin{aligned}
&P(X_0 = s_0, X_1 = s_1, \ldots X_{n-1} = s_{n-1}, X_n = s_n) = \\
&\pi_0(s_0) \cdot P(X_1 = s_1 | X_0 = s_0) \cdot P(X_2 = s_2 | X_0 = s_0, X_1 = s_1) \cdots \\
&P(X_n = s_n | X_0 = s_0, X_1 = s_1 \ldots X_{n-1} = s_{n-1}) = \qquad (11.4)\\
&\pi_0(s_0)P(X_1 = s_1 | X_0 = s_0) \cdot P(X_2 = s_2 | X_1 = s_1) \cdots P(X_n = s_n | X_{n-1} = s_{n-1}) = \\
&\pi_0(s_0)Q(s_0, s_1)\ldots Q(s_{n-1}, s_n).
\end{aligned}$$

□

Exemplul 2. Considerăm lanţul Markov din Exemplul 1, având distribuţia iniţială de probabilitate $\pi_0 = [0.2, 0.35, 0.45]$. Să se calculeze probabilitatea ca lanţul să evolueze din starea $s_0 = 2$, în starea $s_{11} = 3$ pe traiectoria: 2, 1, 3, 2, 1, 2, 3, 1, 3, 2, 1, 3.

Conform formulei deduse probabilitatea ca lanţul să aibă traiectoria 2, 1, 3, 2, 1, 2, 3, 1, 3, 2, 1, 3 este $P = \pi_0(2)Q(2,1)Q(1,3)Q(3,2)Q(2,1)Q(1,2)Q(2,3)\, Q(3,1)Q(1,3)Q(3,2)Q(2,1)Q(1,3)$. Calculul numeric al acestui produs conduce la: $P(X_0 = 2, X_1 = 1, X_2 = 3, X_3 = 2, X_4 = 1, X_5 = 2, X_6 = 3, X_7 = 1, X_8 = 3, X_9 = 2, X_{10} = 1, X_{11} = 3) = 8.9600000000e - 09$.

Distribuţia de probabilitate a variabilei de stare la momentul n

În definiţia lanţului Markov, (X_n), nu se precizează şi distribuţia de probabilitate a variabilei aleatoare de stare X_n. Variabila aleatoare X_n ia valorile $\{1, 2, \ldots, m\}$ şi evenimentul $(X_n = j)$ este evenimentul ca la momentul n traiectoria aleatoare să ajungă în nodul $j \in S$. Notăm cu $\pi_n(j) = P(X_n = j)$

Vom arăta că dacă se cunoaşte distribuţia iniţială de probabilitate, π_0, şi matricea de tranziţie, Q, a lanţului Markov (X_n), atunci putem determina şi distribuţia de probabilitate, $\pi_n = [\pi_n(1), \pi_n(2), \ldots, \pi_n(m)]^T$, a variabilei aleatoare X_n, $\forall n > 0$.

Propoziţia 11.0.2 *Distribuţia de probabilitate a stării la momentul n este:*

$$\pi_n^T = \pi_0^T Q^n$$

sau detaliat:

$$\left[\ \pi_n(1) \quad \pi_n(2) \quad \ldots \quad \pi_n(m)\ \right] = \left[\ \pi_0(1) \quad \pi_0(2) \quad \ldots \quad \pi_0(m)\ \right] Q^n$$

Observaţie Vectorii probabilişti π_0, π_n sunt matrici coloană, deci transpusele lor sunt matrici linie

Demonstraţie: Notăm cu A evenimentul $(X_n = j)$ şi cu H_i evenimentele ipoteze, $H_i = (X_0 = i)$, $i = 1, 2, \ldots, m$. Evident că $H_1, H_2, \ldots, H_m$ constituie o descompunere a evenimentului sigur în m evenimente mutual exclusive două câte două. Conform formulei probabilităţii totale avem:

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A|H_i) \tag{11.5}$$

Rescriem formula probabilităţii totale înlocuind A cu $(X_n = j)$ şi H_i cu $(X_0 = i)$:

$$P(X_n = j) = \sum_{i=1}^{m} P(X_0 = i)P(X_n = j|X_0 = i), \tag{11.6}$$

adică

$$P(X_n = j) = \sum_{i=1}^{m} \pi_0(i)P_{ij}^n = \sum_{i=1}^{m} \pi_0(i)Q^n(i, j). \tag{11.7}$$

Rezultă astfel că:

$$\pi_n(j) = \sum_{i=1}^{m} \pi_0(i)Q^n(i, j), \tag{11.8}$$

ceea ce ne conduce la relaţia matricială:

$$[\ \pi_n(1) \quad \pi_n(2) \quad \ldots \quad \pi_n(m)\ ] = [\ \pi_0(1) \quad \pi_0(2) \quad \ldots \quad \pi_0(m)\ ]Q^n \tag{11.9}$$

sau concentrat

$$\pi_n^T = \pi_0^T Q^n \tag{11.10}$$

$\square$

Din relaţia (11.9) rezultă că distribuţia de probabilitate a stării la momentul n, X_n se poate calcula recursiv pornind de la distribuţia iniţială π_0:

$$\begin{aligned} \pi_1^T &= \pi_0^T Q \\ \pi_2^T &= \pi_0^T Q^2 = \pi_1^T Q \\ &\vdots \\ {\pi_n}^T &= \pi_{n-1}^T Q \end{aligned} \tag{11.11}$$

Exemplul 3. Pentru lanţul Markov din Exemplul 1, având distribuţia iniţială de probabilitate, $\pi_0 = [0.2, 0.35, 0.45]$, distribuţiile de probabilitate ale stărilor $X_1, X_2, \ldots, X_{10}$, calculate conform relaţiei de recurenţă de mai sus, sunt:

$$\begin{aligned}
\pi_1 &= [0.33000000000 \quad 0.525000000000 \quad 0.145000000000];\\
\pi_2 &= [0.29500000000 \quad 0.590500000000 \quad 0.114500000000];\\
\pi_3 &= [0.28190000000 \quad 0.606650000000 \quad 0.111450000000];\\
\pi_4 &= [0.27867000000 \quad 0.610185000000 \quad 0.111145000000];\\
\pi_5 &= [0.27796300000 \quad 0.610922500000 \quad 0.111114500000];\\
\pi_6 &= [0.27781550000 \quad 0.611073050000 \quad 0.111111450000];\\
\pi_7 &= [0.27778539000 \quad 0.611103465000 \quad 0.111111145000];\\
\pi_8 &= [0.27777930700 \quad 0.611109578500 \quad 0.111111114500];\\
\pi_9 &= [0.27777808430 \quad 0.611110804250 \quad 0.111111111450];\\
\pi_{10} &= [0.27777783915 \quad 0.611111049705 \quad 0.111111111145];
\end{aligned}$$

Faptul că matricii de tranziţie, Q şi distribuţiei iniţiale de probabilitate, π_0, i se asociază un şir de vectori probabilişti (π_n), ne conduce la întrebările:

• În ce condiţii este şirul (π_n) convergent?

• Dacă şirul (π_n) este convergent la π, ce reprezintă limita sa, π?

• Fiecare distribuţie iniţială π_0 defineşte un alt şir π_n, cu $\pi_n^T = \pi_0^T Q^n$, deci ne aşteptăm ca în caz de convergenţă, să obţine mereu altă limită, pentru fiecare π_0 distinct sau poate aceeaşi limită.

Pentru a răspunde acestor întrebări detaliem câteva particularităţati ale matricii Q şi a transpusei sale, Q^T, precum şi ale şirului π_n.

• Dacă şirul (π_n) al distribuţiilor de stare este convergent, atunci limita sa este un vector probabilist π, adică un vector de coordonate mai mari sau egale ca zero şi suma coordonatelor este 1.

Demonstraţie: **Opţional** $\pi_n = (\pi_n(1), \pi_n(2), \dots, \pi_n(m))^T$ şi $\pi_n(i) \in (0,1)$, $\sum_{i=1}^m \pi_n(i) = 1$. Presupunem că şirul vectorial (π_n) este convergent la vectorul $\pi = (\pi_1, \pi_2, \dots, \pi_m)^T$. Se ştie din analiză că $\pi_n \to \pi$ dacă şi numai dacă coordonatele $\pi_n(i) \to \pi_i$, $\forall\ i = 1, 2, \dots, m$. Astfel şirul $S_n = \pi_n(1) + \pi_n(2) + \cdots + \pi_n(m) \to \pi_1 + \pi_2 + \cdots + \pi_m$. Dar $S_n = 1$ oricare ar fi n, deci şi limita sa este 1, adică $\pi_1 + \pi_2 + \cdots + \pi_m = 1$ (ceea ce evidenţiază că π este vector probabilist). □

Propoziţia 11.0.3 *Dacă şirul π_n al distribuţiilor de stare, la momentul n converge la π, atunci $\pi^T = \pi^T Q$.*

Demonstraţie: Dacă $\lim_{n\to\infty} \pi_n = \pi$ atunci şi $\lim_{n\to\infty} \pi_{n-1} = \pi$. Trecând la limită când $n \to \infty$ în relaţia:

$$\pi_n^T = \pi_{n-1}^T Q$$

obţinem:

$$\pi^T = \pi^T Q,$$

□

Definiţia 11.0.2 *O distribuţie de probabilitate π, pe spaţiul nodurilor unui lanţ Markov, cu proprietatea că $\pi^T = \pi^T Q$ se numeşte distribuţie invariantă, staţionară sau distribuţie de echilibru.*

Ce înseamnă asta? Din faptul că π este limita şirului π_n rezultă că pentru orice $\epsilon > 0$, există un rang $N(\epsilon)$ astfel încât pentru orice $n > N(\epsilon)$, avem $\|\pi_n - \pi\| < \epsilon$, adică dacă se calculează π_n, pentru n suficient de mare, cu ajutorul relaţiei $\pi_n^T = \pi_0^T Q^n$, atunci π_n aproximează destul de bine limita π şi deci de la un anumit rang N, avem că $\pi_{n+1}^T = \underbrace{\pi_n^T}_{\approx \pi^T} Q \approx \pi$ şi analog $\pi_{n+2} \approx \pi$, etc.

Dacă există distribuţia de echilibru π (ca limită a şirului π_n), $\pi = (\pi(1), \pi(2) \dots, \pi(m))^T$, atunci $\pi(j)$ reprezintă şansa asimptotică de a fi vizitat nodul j. Cu alte cuvinte dacă mişcarea aleatoare pe S continuă indefinit, şi şirul (π_n) este convergent, atunci de la un moment dat mişcarea aleatoare se stabilizează şi vizitează fiecare nod $j \in \{1, 2, \dots, m\}$ cu aceeaşi frecvenţă, $\pi(j)$.

În cazul exemplului (3) dacă calculaţi distribuţiile de probabilitate $\pi_{50}, \pi_{51}, \dots$ şi afişati coordonatele cu 15 zecimale veţi obţine

$$\pi_{50} = \pi_{51} = \cdots = \pi_{100} = [0.277777777778 \quad 0.611111111111 \quad 0.111111111111]$$

Observăm că $P(X_{50} = k) = \cdots = P(X_{100} = k)$, $\forall\, k = 1, 2, 3$. Orice distribuţie de probabilitate π_n, cu $n \geq 50$, am calcula cu 15 zecimale, am obţine distribuţii identice cu π_{50}.

adică după momentul $n = 50$ distribuţia este staţionară, nu se mai modifică în primele 15 zecimale şi deci sistemul a ajuns într-un echilibru. De-a lungul oricărei traiectorii, înregistrate începând cu momentul $n = 50$, $s_{50}, s_{51}, \dots s_{50+N}$, de orice lungime N+1, starea 1 este "vizitată" de lanţul Markov în proporţie de $100\pi(1)\% = 27.7777777778\%$, starea 2 de $100\pi(2)\% = 61.1111111111\%$, starea 3 în proporţie de $100\pi(m)\% = 11.1111111111\%$.

Mai mult observăm că pentru $n > 50$ norma diferenţei dintre două distribuţii consecutive $\pi_{n-1} - \pi_n$ este mai mică decât 10^{-15}, $\|\pi_n - \pi_{n-1}\| < 10^{-15}$.

Ne întrebăm în mod natural, ce anume proprietăţi ale lanţului Markov asigură convergenţa şirului de distribuţii (π_n)? Dacă lanţul Markov ar fi definit pe nodurile grafului WEB (ce constau din pagile WEB) şi mişcarea aleatoare s-ar face alegând din fiecare pagină cu o anumită probabilitate paginile către care există linkuri, atunci existenţa distribuţiei de echilibru ar permite caracterizarea popularităţii paginilor WEB cu ajutorul acestei distribuţii, π, $\pi(j)$ reprezentând frecvenţa asimptotică cu care un navigator aleator pe WEB ar vizita pagina j, adică popularitatea paginii j.

În continuare prezentăm condiţiile pe care trebuie să le îndeplinească graful de tranziţie al unui lanţ Markov pentru ca şirul distribuţiilor (π_n) să fie convergent. Particularităţile pe care le evidenţiem sunt cele care au inspirat modul de definire a navigarii aleatoare de către Larry Page şi Serghei Brin, fondatorii motorului de căutare Google şi autorii PageRank-ului.

Definiţia 11.0.3 *Un lanţ Markov pe $S = \{1, 2, \dots, m\}$ se numeşte lanţ ireductibil, dacă oricare ar fi două noduri $i, j \in S$, există $n > 0$, astfel încât $Q^n[i][j] > 0$, adică cu probabilitate nenulă lanţul poate trece într-un număr de paşi din nodul i în nodul j.*

Practic lanţul este ireductibil dacă şi numai dacă graful de tranziţie este tare conex (există drum de arce între orice două noduri).

Dacă matricea de tranziţie are toate elementele $Q[i][j] > 0$, $\forall\, i, j \in \overline{1, m}$, atunci lanţul Markov este ireductibil. Dacă există un $n > 1$ astfel încât $Q^n[i][j] > 0$, $\forall\, i, j \in \overline{1, m}$, lanţul Markov este de asemenea ireductibil (condiţia precedentă corespunde cazului $n = 1$).

Exemplul 4. Lanţul Markov având spaţiul stărilor $S = \{1, 2, 3, 4\}$ şi matricea de tranziţie:

$$Q = \begin{pmatrix} 0.35 & 0.65 & 0 & 0 \\ 0.45 & 0.55 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

nu este un lanţ ireductibil, aşa cum se poate observa mai simplu din graful asociat. Mulţimile de stări $\{1, 2\}$ şi $\{3, 4\}$ nu comunică între ele.

Observaţia 11.0.1 *Simplul fapt că $Q[i][j] = 0$ nu asigură că i nu comunică cu j. Cele două noduri nu comunică într-un pas, dar pot comunica în mai mulţi paşi, adică s-ar putea ca $Q^n[i][j] > 0$, pentru un $n > 1$.*

Un lanţ Markov poate avea şi *traiectorii periodice* (Fig.11.2), adică traiectorii în care succesiunea de noduri $s_0, s_1, \ldots, s_{T-1}, s_T = s_0$ se repetă indefinit într-o traiectorie a lanţului.

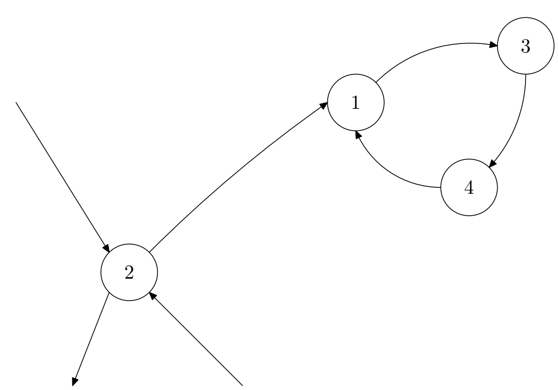

Fig.11.2: Graful de tranziţie al unui lanţ Markov ce are traiectoria periodică (1,3,4,1)

Din analiza grafului din figură rezultă că probabilităţile $Q^3[1][1] > 0$, $Q^6[1][1] > 0$ şi în general $Q^{3k}[1][1] > 0$, $\forall\, k \in \mathbb{N}$, ceea ce ilustrează că, cu o probabilitate pozitivă, o traiectorie ce porneşte din 1 se reîntoarce în 1 după 3, paşi, 6 paşi, sau mai general un multiplu de 3, paşi. Această proprietate indică că traiectoria ce porneşte din 1 este periodică. Analog pentru 3 şi 4.

Definiţia 11.0.4 *Perioada unui nod, i, este numărul*

$$\tau_i = c.m.m.d.c\{n \in \mathbb{N} \,|\, Q^n[i][i] > 0\}$$

În cazul exemplului nostru cel mai mare divizor comun al numerelor de forma $3k$, $k \in \mathbb{N}$ este 3, deci nodul 1 este periodic de perioadă 3 şi analog nodurile 3 şi 4.

Un nod i a cărui perioadă este 1, adică nodul pentru care cel mai mare divizor comun al numerelor n cu proprietatea că $Q^n[i][i] > 0$ este 1, se numeşte *nod aperiodic*, iar un lanţ Markov care are toate nodurile aperiodice se numeşte *lanţ Markov aperiodic*. Cel mai simplu exemplu de nod aperiodic este un nod i, pentru care $Q[i][i] > 0$.

Evident că dacă matricea de tranziţie are toate elementele de pe diagonala principală strict pozitive, $Q[i][i] > 0$, atunci lanţul este aperiodic.

Remarcăm că proprietăţile de ireductibilitate şi aperiodicitate ale unui lanţ Markov sunt proprietăţi ale matricii de tranziţie.

Se poate demonstra (algebric) următoarea proprietate:

Un lanţ Markov ireductibil are toate nodurile de aceeaşi perioadă, d. Prin urmare dacă un nod a unui lanţ ireductibil este aperiodic, atunci toate nodurile sunt aperiodice. Astfel pentru a arăta că un lanţ ireductibil este aperiodic este suficient să identificăm un nod i pentru care $Q[i][i] > 0$ (probabilitatea de trecere de la starea i la ea însăşi este nenulă), pentru a concluziona că nodul i este aperiodică şi deci toate stările lanţului sunt aperiodice. Dar există lanţuri ireductibile care nu au niciun nod i, astfel încât $Q[i][i] > 0$, adică graful asociat nu conţine nicio buclă, şi totuşi lanţul este aperiodic.

De exemplu pentru lanţul ce are matricea de tranziţie:

$$Q = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

graful asociat nu are nicio buclă. Lanţul este ireductibil şi nodul 4 aparţine unui ciclu de lungime 2 şi unuia de lungime 3. Deci c.m.md.c. (2,3)=1. Prin urmare nodul 4 este aperiodic şi cum lanţul este ireductibil, atunci toate nodurile sunt aperiodice.

Se demonstrează următorul rezultat, pe baza căruia s-a fundamentat şi algoritmul PageRank–Google:

Dacă un lanţ Markov pe S este ireductibil şi aperiodic, atunci oricare ar fi distribuţia iniţială de probabilitate, π_0, şirul distribuţiilor de probabilitate la momentul n, $(\pi_\mathbf{n})$, asociat, este convergent şi limita acestuia este vectorul propriu probabilist, π, care nu depinde de distribuţia iniţială (deci indiferent de distribuţia iniţială de probabilitate, şirurile asociate converg la aceeaşi limită π). Această limită este unica distribuţie de echilibru a lanţului Markov. Mai mult şi şirul $(\mathbf{Q}^\mathbf{n})$ este convergent la o matrice de rang 1 având fiecare linie

egală cu $\pi^{\mathbf{T}}$:

$$\lim_{n\to\infty} Q^n = \begin{pmatrix} \pi(1) & \pi(2) & \dots & \pi(m) \\ \pi(1) & \pi(2) & \dots & \pi(m) \\ \vdots & \vdots & \dots & \vdots \\ \pi(1) & \pi(2) & \dots & \pi(m) \end{pmatrix}$$

Remarcăm că faptul că limita π este distribuţie de echilibru, adică $\pi^T = \pi^T Q$ este echivalent, aplicând transpunerea cu:

$$Q^T \pi = \pi$$

adică π este vector propriu al matricii Q^T, corespunzător valorii proprii 1. Vezi Cursul de Algebră, relativ la popularitatea nodurilor unei reţele orientate. Acolo am vazut că vectorul ranking este vector propriu al matricii de adiacenţă, transpusă, corespunzător valorii proprii dominante. În cazul matricilor stochastice valoarea proprie dominantă este 1!!!!

Acest rezultat matematic a influenţat modul de definire al navigării aleatoare pe graful WEB, ca un lanţ Markov ireductibil şi aperiodic pe mulţimea paginilor WEB.

11.1 Construcţia lanţului Markov pe graful WEB. Algoritmul PageRank

Succesul extraordinar şi dominaţia motorului Google se datorează în principal algoritmului PageRank, care exploatează structura link-urilor 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 aceste pagini în ordinea descrescătoare a popularităţii lor. 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 precis lanţul Markov ce stă la baza algoritmului `PageRank`. Fie $W = \{1, 2, \dots, m\}$, mulţimea tuturor paginilor WEB, $H = (h_{ij})$ matricea de conectivitate a lui W, sau matricea hyperlink:

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

H este o matrice rară, adică cu foarte multe zerouri (în medie cu 3-10 elemente nenule pe o linie). Suma elementelor de pe linia i a matricii 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 pagină. Suma elementelor de pe coloana i a matricii 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 aceasta are linkuri, prin urmare probabilitatea de a trece din pagina i în pagina j este:

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

De exemplu dacă

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

atunci ordinul de ieşire din pagina 2 este $r_2 = 3$ şi deci probabilitatea de a trece din pagina 2 în oricare din paginile $\{1, 2, \ldots, 10\}$ este $p_{2j} = h_{2j}/3$, adică cu aceeaşi probabilitate de 1/3, un surfer poate trece din pagina 2 în pagina 1, 5 sau 8.

Vom exemplifica construcţia propusă de L. Page şi S. Brin prin modelul simplu, de reţea izolată, de pagini WEB (reţea intranet), din Fig.11.3. Notăm cu $Q = (p_{ij})$ matricea probabilităţilor de tranziţie $p_{ij} = h_{ij}/r_i$, $i, j = \overline{1, 6}$. Se observă din structura grafului de conectivitate că paginile 2 şi 6 sunt pagini ce nu conţin link-uri 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 situaţie, L. Page şi S. Brin au propus ca vector de probabilitate de tranziţie dintr-o pagină dangling, i, distribuţia uniformă $p_{ij} = 1/m$, $\forall, j = \overline{1, m}$. Adică, în mod artificial se adaugă link-uri 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:

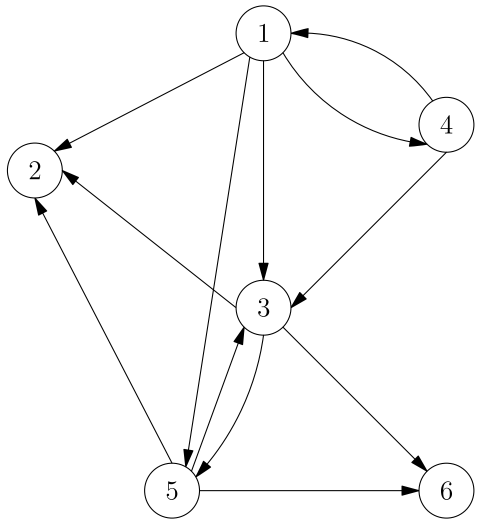

Fig.11.3: Graf orientat ilustrând link-urile între 6 pagini WEB.

$\tilde{Q}=$

	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 link-uri între orice două pagini sau echivalent, graful WEB nu este tare conex) şi pot exista traiectorii periodice, adică surferul navigând conform matricii 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, cei doi, L. Page şi S. Brin, au introdus ipoteza că doar cu probabilitatea $\alpha \in (0,1)$, surferul navighează conform matricii $\tilde{Q}$ şi cu probabilitatea complementară, $(1-\alpha)$, ignoră hyperlink-urile 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* şi 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 link-uri într-o "navigare artificială". Evident că şi matricea E este o matrice stochastică pe linii, iar G fiind o combinaţie convexă de astfel de două matrici este o matrice stochastică (vezi proprietăţile matricilor stochastice, date la pagina 3!!!). Mai mult, $G(i,j) > 0$, $\forall\ i,j = \overline{1,m}$, şi 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 la un moment dat, de cardinal m,

• matricea de tranziţie de tipul G, cu α fixat, şi

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

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

PageRank–ul, π, este este limita şirului (π_n), cu $\pi_n^T = \pi_0^T G^n$. 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, folsind aşa numita metodă a puterii, adică se calculează recursiv, pornind de la π_0 şi G, distribuţiile (sau PageRankul 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 $\|\pi_n - \pi_{n-1}\| < \epsilon$, unde ϵ este un număr pozitiv foarte mic, prescris.

Pseudocodul algoritmului de calcul al PageRank-ului este dat pe pagina următoare.

function PageRank($\mathtt{G}, \mathtt{m}$);
    $\pi = [\mathtt{1/m}, \mathtt{1/m}, \ldots, \mathtt{1/m}]$; **//Distributia initiala de probabilit**
    $\mathtt{eps} = 10^{-7}$;
    do
        $\pi' = \pi$;
        $\pi = \pi' * \mathtt{G}$;
    while $(\|\pi - \pi'\| >= \mathtt{eps})$;
    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 PageRankului. Din punctul de vedere al vitezei de convergenţă ar fi preferabil un factor, α, cât mai apropiat de zero. În acest caz însă 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 matricii 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 dupa 100-200 de iteraţii).

Dacă vreti să aflaţi Pagerankul unor pagini WEB intraţi aici:
`http://www.prchecker.info/check_page_rank.php`

11.2 Pagerankul personalizat

Pentru o ierarhizare personalizată a paginilor WEB, matricea de teleportare E, se calculează luând în considerare vectorul personalizat w, ce este un vector probabilist ale cărui coordonate $w = (a_1, a_2, \dots a_m)^T$, reprezintă probabilitatea ca surferul, ce iese din navigarea conform linkurilor, să aleagă pagina $1, 2, \dots, m$, 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 $e = (1, 1, \dots, 1)^T$, şi matricea Google, corespunzătoare:

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

iar distribuţia de echilibru corespunzătoare este Pagerankul personalizat.