Curs 10: Analiza linkurilor

Analiza hiperlinkurilor

Criteriu de analiză a grafului rețelei WWW

- Folosit la calculul scorului de importanță al unei pagini web pentru o cerere de informație q
- Inspirat din bibliometrie: calculul factorului de impact al publicațiilor științifice bazat pe analiza citărilor:
  - o citare a lucrării B în lucrarea A atribuie o valoare științifică la lucrarea B din partea autorului lucrării A
  - la fel, un hiperlink `<a href=`*URL-B*`>`*text-ancoră*`</a>` într-o pagină A crește scorul de importanță al paginii B cu o dovadă că pagina de la adresa *URL-B* are un conținut descris de termenii din *text-ancoră*.

Graful rețelei WWW

- Un hiperlink *text-ancoră* într-o pagină *A* reprezintă o sursă de informații: *text-ancoră* este o descriere concisă a conținutului paginii *B*.
- Un hiperlink de la *A* la *B* crește scorul de importanță al paginii *B* cu o referință din partea paginii *A*.
  - Există și excepții de la această regulă: de exemplu, multe pagini web ale unei corporații pot avea hiperlink-uri la o pagină *B* referitoare la drepturi de copyright.
    - ▶ scorul de importanță al paginii *B* nu trebuie să depindă de numărul de hiperlinkuri de la alte paginila *B*

    $\Rightarrow$ astfel de linkuri "interne" se omit din analiza hiperlinkurilor

Un fragment HTML precum

```
<a href="http://www.acm.org/jacm/">Journal of the ACM</a>
```

asociază textul ancoră `Journal of the ACM.` cu URL-ul `http://www.acm.org/jacm/`

- Adesea, textul ancoră din alte pagini web este o descriere succintă a conținutului paginii de la URL-ul din tag-ul `<a href="...">`. Uneori, această descriere nu există în textul paginii web:
  - De exemplu, pagina web `http://www.ibm.com` nu conține termenul `computer` deși IBM este cel mai mare producător de calculatoare
  - $\Rightarrow$ majoritatea motoarelor de căutare îl indexează $\Rightarrow$ vocabular extins cu termeni de căutare cu un indicator al faptului ce provin din text ancoră
- Uneori, analiza link-urilor ia în considerare o porțiune mai mare de text din jurul textului ancoră, numită text ancoră extins

Fiecărui termen din textul ancoră i se atribuie o pondere în calculul scorului de importanță ca răspuns la o întrebare q:

- Atribuirea ponderilor la termeni se face cu algoritmi de machine-learning
  - de obicei, pondeerea unui termen este proporțională cu frecvența aparițiilor
  - se evită atribuirea de ponderi mair la termeni care apar foarte des (de exemplu, termenii `Click` și `here` care apar foarte des în text ancoră)

Indexarea textului ancoră de către motoarele de căutare poate fi folosită la crearea de atacuri orchestrate cu spam:

- un site web își poate mări scorul de importanță la căutări după anumiți termeni, adăugând mult text ancoră cu termenii respectivi, care se referă la sine.

$\Rightarrow$ motoarele de căutare caută să detecteze și să evite aceste forme de abuz cu spam.

Atribuie un scor de importanță $scor(n) \in [0..1)$ la fiecare nod *nod* al grafului rețelei web. $scor(nod)$ se calculează luând în considerare doar graful de hiperlinkuri al rețelei web.

Scenariu pe care se bazează algoritmul PageRank

Un navigator WWW accesează o pagină web (=un nod al grafului rețelei WWW) și parcurge aleator acest graf:

- la fiecare moment dat, trece de la pagina curentă A la o pagină B aleasă aleator (cu probabilități egale) dintre paginile spre care există un link din A
- Dacă nu există link-uri de la pagina curentă A la nici o pagină, navigatorul efectuează un pas teleport: trece de la A la o pagină B aleasă aleator dintre nodurile grafului rețelei WWW. În particular, putem avea $B = A$.
  - Navigatorul poate introduce un URL oarecare în câmpul de adrese URL al unui browser

Fiecare nod al grafului rețelei web este vizitat cu o anumită frecvență: PageRank calculează aceste frecvențe, și definește scorul de importanță $\pi(n)$ al fiecărui nod n astfel:

- $\pi(n) \in [0..1]$ este frecvența de vizitare a nodului n de către un drum aleator al unui navigator.

Fie $G = (V, E)$ graful rețelei web cu $V = \{nod_1, nod_2, \ldots, nod_N\}$ și

$$A_G = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ a_{21} & a_{22} & \ldots & a_{2N} \\ \vdots & \ddots & \vdots & \\ a_{N1} & a_{N2} & \ldots & a_{NN} \end{pmatrix} \text{ unde } a_{ij} = \begin{cases} 1 & \text{dacă } \exists \text{ link de la } nod_i \text{ la } nod_j, \\ 0 & \text{altfel.} \end{cases}$$

Se presupune dată $\alpha \in (0..1)$ probabilitatea de efectuare a unui pas teleport la orice moment dat (o valoare tipică este pt. α este $\alpha = 0.1$)

- Se calculează matricea de probabilități $P = (P_{ij})$, unde P_{ij} :=probabilitatea de trecere din nodul nod_i în nod_j. Dacă $B_1, \ldots, B_r$ sunt paginile spre care există linkuri din A, atunci
  - probab. de trecere din A la o pagină $B \notin \{B_1, \ldots, B_r\}$ este α/N
  - probab. de trecere din A la o pagină $B_i \in \{B_1, \ldots, B_r\}$ este

$$\frac{1 - (N - r)\alpha/N}{r}$$

- Vectorul scorurilor de importanță $\vec{\pi} = (\pi(nod_1), \ldots, \pi(nod_N))$ este vectorul caracteristic al matricii P pt. valoarea caracterisică $\lambda = 1$: $\vec{\pi} \cdot P = \vec{\pi}$

Navigarea descrisă a rețelei WWW este un lanț Markov=un proces the pași discreți în timp care trece printr-un nr. finit de N stări (=nodurile grafului web). Lanțul Markov al algoritmului PageRank este caracterizat de matricea de probabilități de tranziție

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1N} \\ P_{21} & P_{22} & \dots & P_{2N} \\ \vdots & \ddots & \vdots & \\ P_{N1} & p_{N2} & \dots & P_{NN} \end{pmatrix}.$$

P este **matrice stohastică** deoarece, pentru toți $i, j \in \{1, 2, \dots, N\}$:

$$0 \leq P_{ij} \leq j \quad \text{și} \quad \sum_{j=1}^{N} P_{ij} = 1.$$

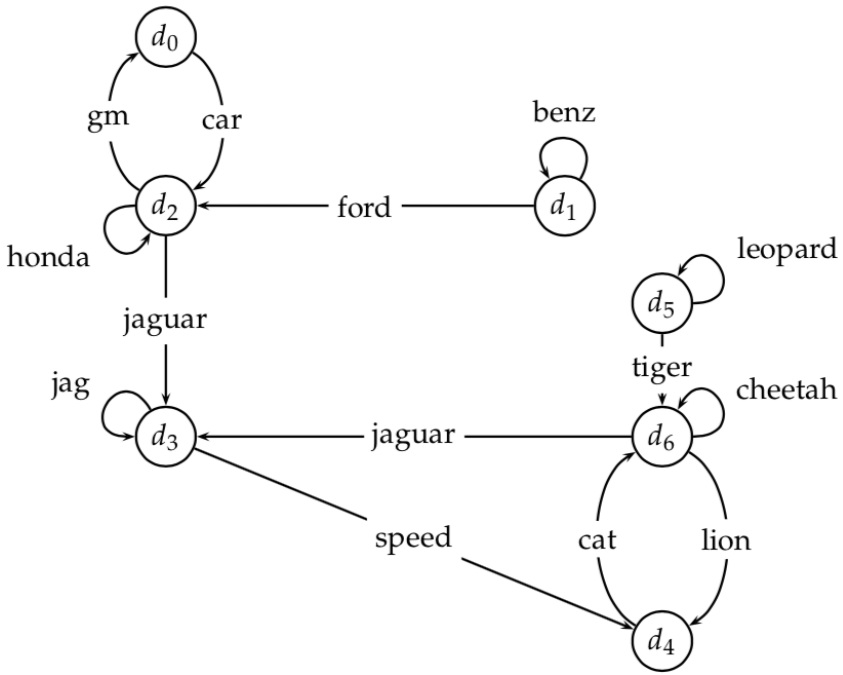
Dacă distribuția de probabilități la $t = 0$ a poziției navigatorului în unul din nodurile grafului web este vectorul $\vec{x} = (x_1, \dots, x_N)$ cu $0 \leq x_i \leq 1$ și $\sum_{i=1}^{N} x_i = 1$, atunci

1. distribuția de probabilități la $t = n$ (adică după n pași) este vectorul $\vec{x} \cdot P^n$
2. vectorul de distribuție de probabilități converge la un vector $\vec{\pi} = (\pi_1, \dots, \pi_N)$:

$$\lim_{n \to \infty} \vec{x} \cdot P^n = \vec{\pi} \text{ astfel încât } \vec{\pi} \cdot P = \vec{\pi}$$

Algoritmul PageRank atribuie scorurile de importanță $\pi_1, \pi_2, \dots, \pi_N$ nodurilor $nod_1, nod_2, \dots, nod_N$ are rețelei web.

Graful web ilustrat mai jos are fiecare arc anotat cu termenul care apare în textul ancoră al linkului corespunzător:



$$\Rightarrow A_G = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Algoritmul PageRank

Exemplu ilustrat pentru $\alpha = 0.14$

$$A_G = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \Rightarrow P = \begin{pmatrix} 0.02 & 0.02 & 0.88 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.45 & 0.45 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.31 & 0.02 & 0.31 & 0.31 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.45 & 0.45 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.88 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.45 & 0.45 \\ 0.02 & 0.02 & 0.02 & 0.31 & 0.31 & 0.02 & 0.31 \end{pmatrix}$$

Valorile succesive ale lui $\vec{x_{n+1}} = \vec{x_n} \cdot P$ pornind de la $\vec{x_0} = (1, 0, 0, 0, 0, 0, 0)$ sunt

$\vec{x_1}$	0.02	0.02	0.88	0.02	0.02	0.02	0.02
$\vec{x_2}$	0.27	0.023	0.298	0.287	0.0343	0.0286	0.05
$\vec{x_3}$	0.105	0.0322	0.352	0.243	0.158	0.0323	0.077
$\vec{x_4}$	0.121	0.034	0.225	0.247	0.147	0.034	0.192
$\vec{x_5}$	0.085	0.035	0.203	0.246	0.181	0.0345	0.216
$\vec{x_6}$	0.078	0.035	0.166	0.246	0.188	0.035	0.253
$\vec{x_7}$	0.068	0.035	0.150	0.246	0.198	0.035	0.269
$\vec{x_8}$	0.063	0.035	0.136	0.245	0.203	0.035	0.282
$\vec{x_9}$	0.059	0.035	0.129	0.245	0.206	0.035	0.29
...	...	...	...	...	...	...	...

$$\Rightarrow \vec{\pi} = \lim_{n \to \infty} \vec{x_n} = (0.05, 0.04, 0.11, 0.25, 0.04, 0.31)$$

Algoritmul PageRank pentru subiecte specifice

PageRank poate fi ajustat să ia în considerare preferințele de căutre ale navigatorului:

- pentru navigatori interesați de sport, operația de teleport se reduce la submulțime S de pagini despre sport

  $\Rightarrow$ un lanț Markov care vizitează o submulțime Y cu $S \subseteq Y$ a tuturor nodurilor grafului web
  $\Rightarrow$ vectorul $\vec{\pi_S} = (\pi_1, \pi_2, \ldots, \pi_M)$ de frecvențe de vizitare a nodurilor din Y de către surfer (M este nr. de noduri din Y)

  $\vec{\pi_S}$ se numește vectorul PakeRank specific pentru subiectul sport.

- alegerea unei pagini pentru un pas teleport poate fi neuniformă.

Aplicații:

- Se pot calcula vectori $\vec{\pi_S}$ de scor de importanță PageRank pentru mai multe subiecte S, de exemplu: știință, politică, religie, etc.
- Un motor de căutare poate lua în calcul preferințele de căutare ale unui utilizator: specificate explicit de utilizator, sau determinate din istoricul căutărilor efectuate.

PageRank personalizat

Model de calcul al scorurilor de importanță pentru navigatori cu o combinație de interese.

Exemplu: navigator interesat 60% în sport și 40% în politică

Procesul de navigare se modelează astfel: fiecare pas teleport alege o pagină despre sport cu probabilititea 60% și o pagină despre politică cu probabilitatea 40%

- Se poate demonstra că, pentru acest lanț Markov, vectorul de distribuție de probabilități de vizitare a nodurilor grafului web este $0.6 \cdot \vec{\pi_S} + 0.4 \cdot \vec{\pi_P}$, unde

  $\vec{\pi_S}$ este vectorul PageRank pentru subiectul sport
  $\vec{\pi_P}$ este vectorul PageRank pentru subiectul politică

Dacă probabilitatea de teleport este $\alpha = 10\%$, atunci utilizatorul efectuează 6% teleport la o pagină despre sport și 4% teleport la o pagină despre politică.

Bibliografie

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze:
Capitolul 21: *Link analysis* din An Introduction to Information Retrieval.
Ediție online (c) 2009 Cambridge UP.