Google Page Rank algoritam

Uvod

- Pretraživači Interneta
  - Google je prvi Internet pretraživač koji je uspješno izdvajao rezultate od smapova, tehnologija Page Rank
  - Trust Rank rješava problem link spam-a
  - Topic-sensitive Page Rank
  - HITS

Prvi pretraživači

- Pojam term – stringovi karaktera bez bjelina
- Termovi su organizovani u vidu inverznog indeksa
  - Struktura podataka koja za svaki term efikasno pronalazi stranice u kojima se pojavljuje
- Search query – stranice koje sadrže tražene pojmove pronalaze se u indeksu i prikazuju kao rezultat
  - Rezultat je uređen prema upotrebi termova u nađenim stranicama
  - Term u hederu stranice ili veliki broj pojavljivanja traženog terma favorizuje datu stranicu

How to fool search engine?

- Term spam
  - Nekoliko hiljada puta upisati riječ “movie”, obojati tekst kao pozadini da bi bio nevidljiv, pa će svaki upit koji sadrži riječ “movie” u rezultatu prikazati vašu stranicu
  - Izvršiti upit sa “movie” i sadržaj stranice koja je prva u rezultatu kopirati u vašu stranicu

Google Page Rank

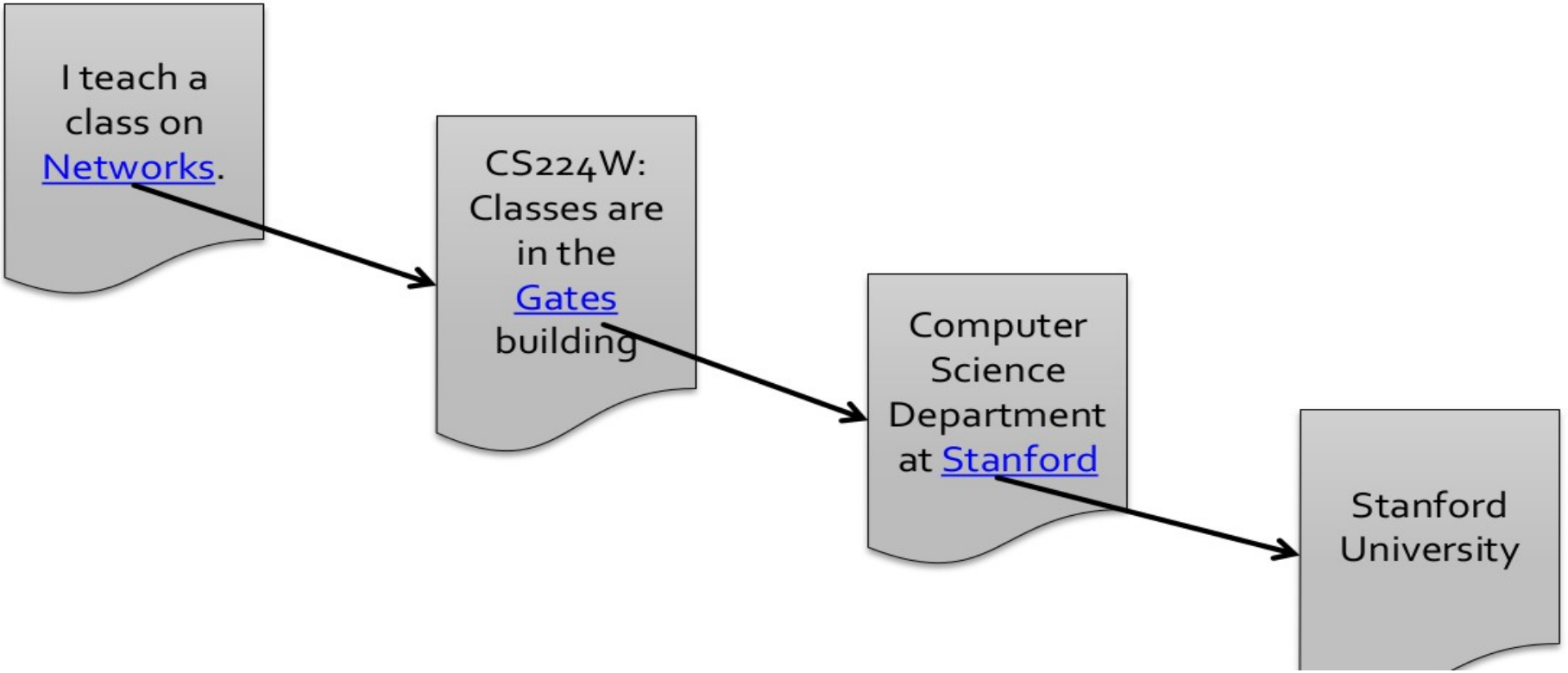
- Koncept “random surfer” - počevši od slučajno odabrane veb stranice, biraju se odlazni linkovi sa te stranice na slučajan način pa se prelazi na sljedeću, proces se ponavlja veliki broj puta, stranice sa velikim brojem posjeta bolje se rangiraju od onih koje se rijetko posjećuju
- Sadržaj stranice procjenjuje se i na osnovu termina koji se koriste “u blizini” linkova prema razmatranoj stranici

Jednostavni Page Rank

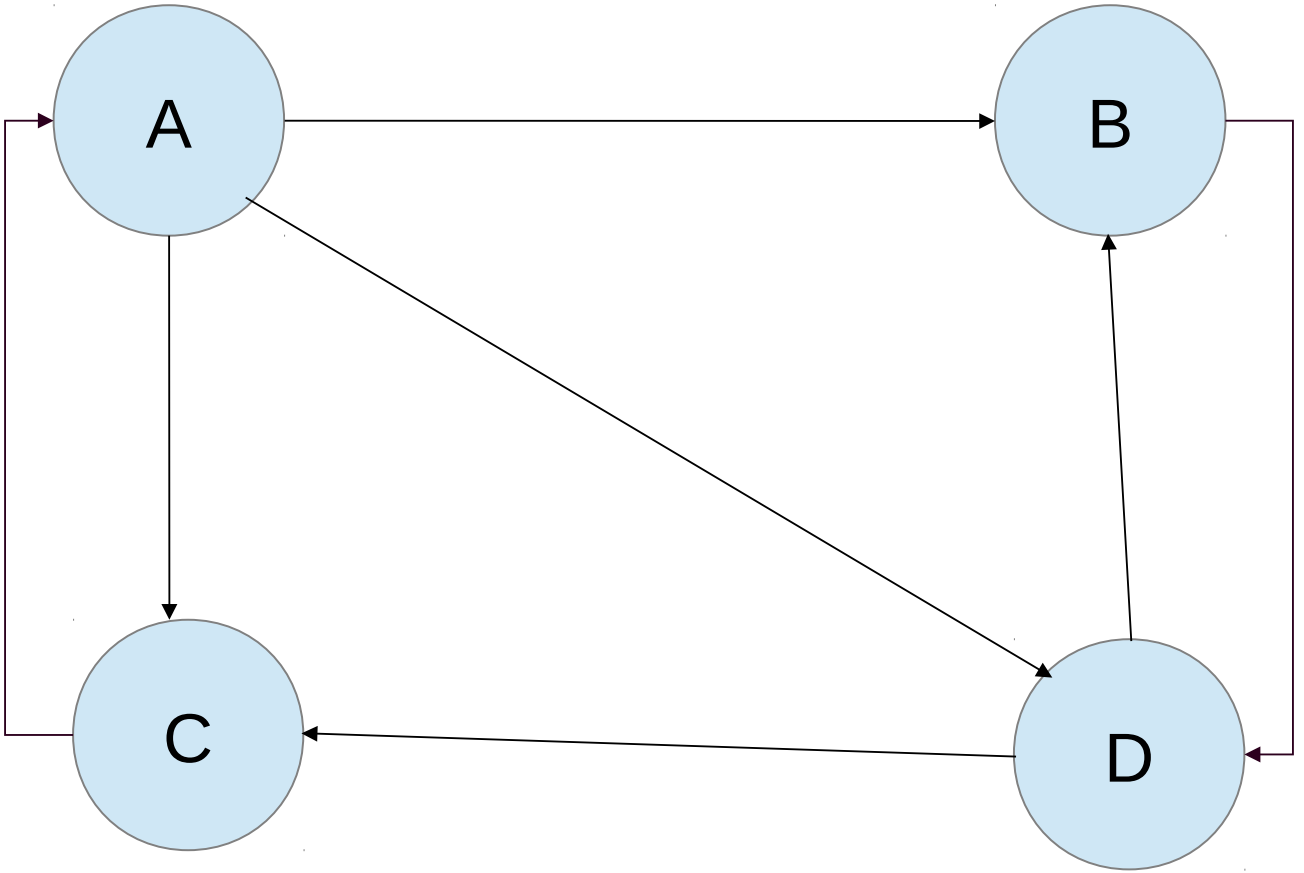
- Aproksimacija – izbrojati broj linkova ka razmatranoj stranici
  - Nije rješenje jer bi potencijalno bilo moguće napraviti veliki broj stranica koje sadrže link ka razmatranoj stranici, koja bi na taj način neopravdano bila rangirana jako visoko
- Kako Google Page Rank rješava
  - dodavanje termina “movie” u stranicu?
  - kreiranje velikog broja stranica koje ukazuju na “lažnu”?
  - kreiranje unakrsnih linkova između “lažnih” stranica?

Graf model Interneta

- Stranice su čvorovi, veza od p_1 do p_2 ako postoji link od stranice p_1 do stranice p_2
- Usmjereni graf



Primjer



Primjer (2)

- Neka “random surfer” počinje sa stranice A
  - Postoji linkovi ka stranicama B,C i D, pa će u sljedećem koraku biti na jednoj od ovih stranica sa vjerovatnoćom 1/3, vjerovatnoća da ostane na stranici A je 0
- Matrica prelaza M, sa n redova i n kolona ako imamo n stranica
  - Element m_{ij} je 1/k ako sa stranice j ima k linkova a jedan je prema stranici i, inače m_{ij} je 0
  - Sastaviti matricu prelaza za prethodni primjer

Page Rank definicija

- Raspodjela vjerovatnoća za lokaciju slučajnog obilaska Interneta data je vektorom, čiji je član j vjerovatnoća da se nalazimo na stranici j
  - Vjerovatnoće su Page Rank funkcija
- Ako obilazak započinjemo tako što slučajno odaberemo jednu od n stranica, početni vektor v_0 sadrži 1/n sa svaku komponentu
- Poslije prvog koraka raspodjela je Mv_0, poslije drugog M^2v_0, poslije koraka i je M^iv_0

Objašnjenje

- Za dati vektor raspodjele v, sljedeća raspodjela data je sa x = Mv
  - Vjerovatnoća x_i da u sljedećem koraku budemo na stranici i je

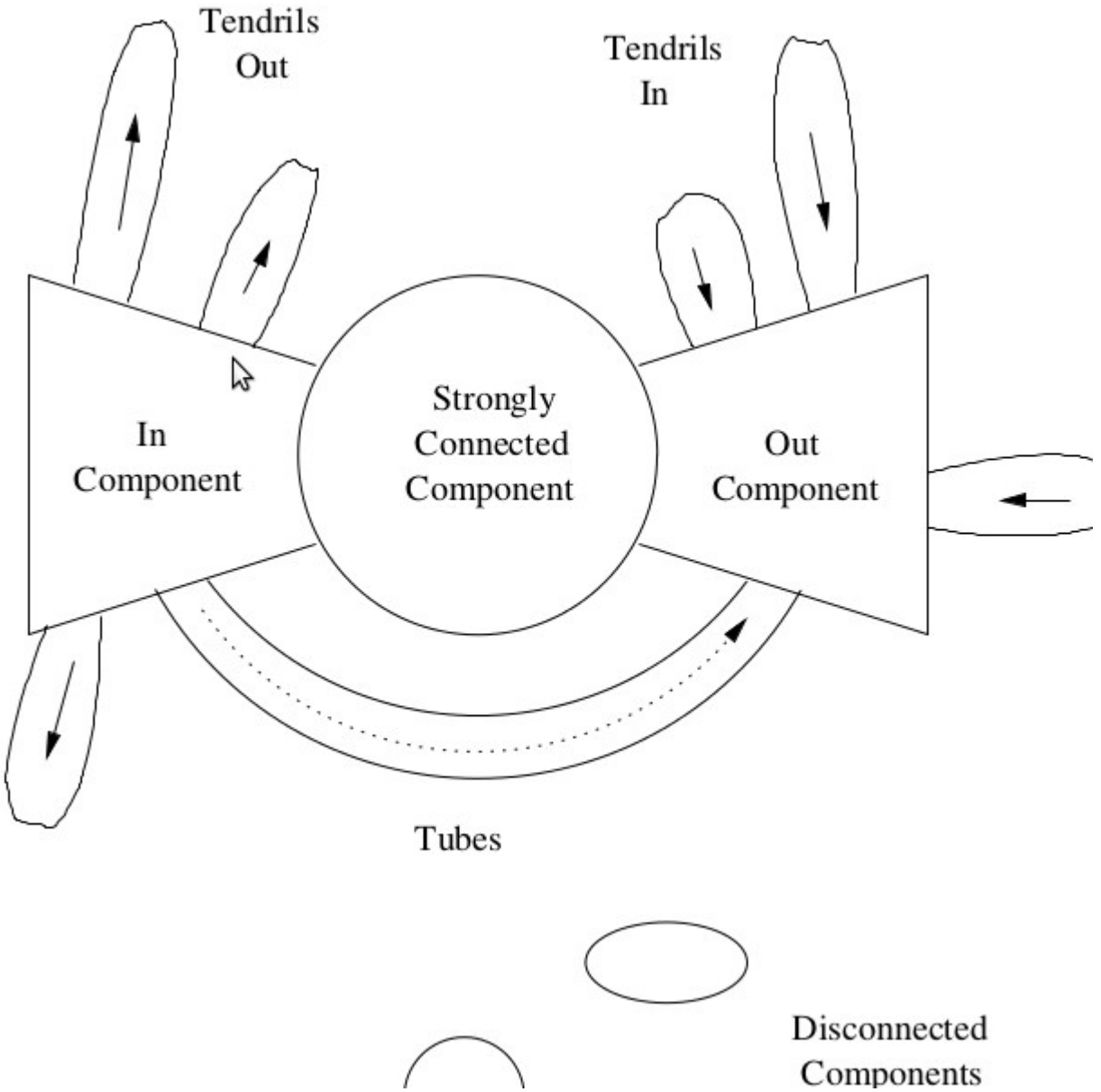
$$\sum_j m_{ij} \mathbf{v}_j .$$

  - Uslovi
    - Graf je povezan
    - Ne postoje čvorovi bez izlaznih linkova, dead ends
  - Suma u bilo kojoj koloni je 1
  - Maksimum za v = Mv

Zaključak

- Ideja Page Rank je da je stranica važna kolika je vjerovatnoća da slučajnim obilaskom ona bude posjećena
- Maksimum se može naći uzastopnim množenjem v_0 sa M sve dok se proizvod ne mijenja ili se malo mijenja
  - Dovoljno je 50-75 iteracija

Struktura Interneta



Struktura Interneta (2)

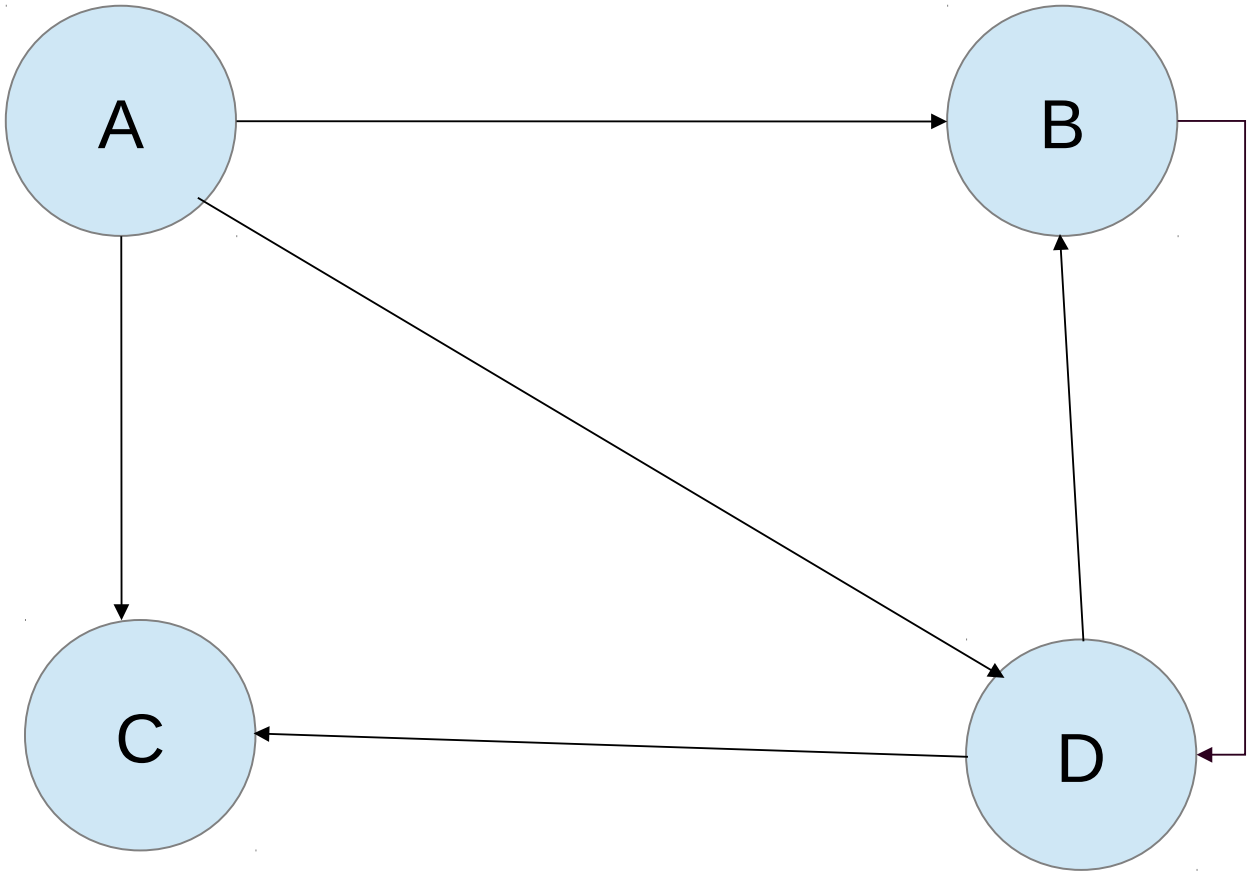
- Povezana komponenta SCC
- In component – stranice koje imaju linkove prema SCC, ali nijesu dostupne iz SCC
- Out component – stranice koje su dostupne iz SCC ali nemaju linkove ka SCC
- Tendrils
  - Stranice dostupne iz in component ali nemaju linkove na in component
  - Stranice sa linkovima na out component ali nedostupne za out component

Struktura Interneta (3)

- Tubes – stranice koje su dostupne iz in component i imaju linkove na out componenta, ali nijesu dostupne iz SCC i nemaju linkove na SCC
- Izolovane komponente koje su nedostupne iz i istovremeno nemaju linkove ka SCC, in i out komponentama

Dead end problem

- Stranica bez izlaznih linkova je dead end
- Zašto je dead end problem za Page Rank?

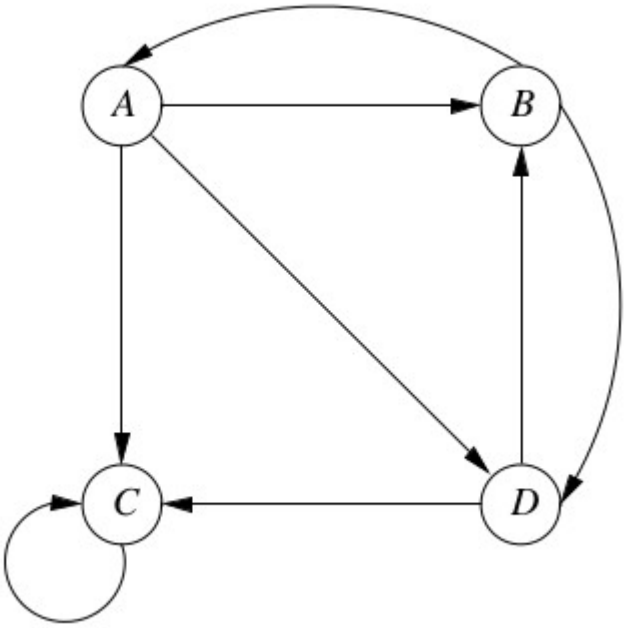


Rješenja za dead end

- Izbrisati dead end stranice i linkove ka njima
  - Kreiranje novih dead end stranica, rekurzivno brisanje
  - Konačno bi ostali SCC, in komponenta i eventualno djelovi izolovanih komponenti
  - Računanje page rank za stranice koje su izbrisane je redosljedom suprotnim od brisanja
- Taxation

Taxation i *spider trap*

- Spider trap je grupa stranica pri čemu svaka stranica ima izlazne linkove ali samo na neku od stranica iz grupe
  - Čitav Page Rank je dodijeljen stranici C



Rješenje - taxation

- Dozvoljeno je sa određenom vjerovatnoćom prelazak na slučajno odabranu stranicu, umjesto da se prati neki od linkova sa tekuće
- Formula
  - $v' = \beta M v + (1 - \beta)e/n$, gdje je $0.8<\beta<0.9$ konstanta, e je jedinični vektor, n je broj stranica
  - Rješava i problem sa dead end stranicama

Page Rank implementacija

- Potrebno je računati množenja matrica-vektor dok se rezultat mijenja, što je reda 50 puta
- Matrica prelaza je rijetka matrica, čuvaju se samo elementi različiti od nule

Reprezentacija matrice prelaza

- Prosječna Internet stranica sadrži 10 izlaznih linkova
  - Ako razmatramo graf sa 10^7 stranica, onda je samo jedan element od milion različit od nule
  - Čuvaju se lokacije ne-nula elemenata i njihova vrijednost, prostorna složenost je linearna u odnosu na broj ne-nula elemenata
  - Poboljšanje: kolona se predstavlja jednim brojem za out-degree i lokacijom ne-nula elementa

Reprezentacija matrice prelaza (2)

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

ORIGINAL

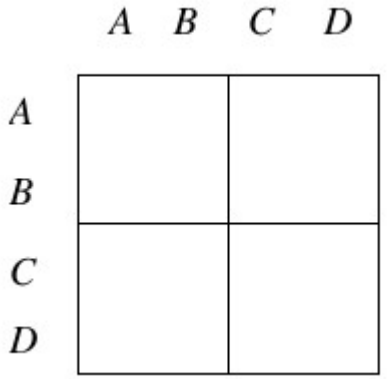
Source	Degree	Destinations
A	3	B, C, D
B	2	A, D
C	1	A
D	2	B, C

Efikasno množenje matrica-vektor

$$\begin{bmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \\ \mathbf{v}_3' \\ \mathbf{v}_4' \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix}$$

Reprezentacija blokova matrice

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$



Source	Degree	Destinations
A	3	B
B	2	A

(a) Representation of M_{11} connecting A and B to A and B

Source	Degree	Destinations
C	1	A
D	2	B

(b) Representation of M_{12} connecting C and D to A and B

Source	Degree	Destinations
A	3	C, D
B	2	D

(c) Representation of M_{21} connecting A and B to C and D

Source	Degree	Destinations
D	2	C

(d) Representation of M_{22} connecting C and D to C and D

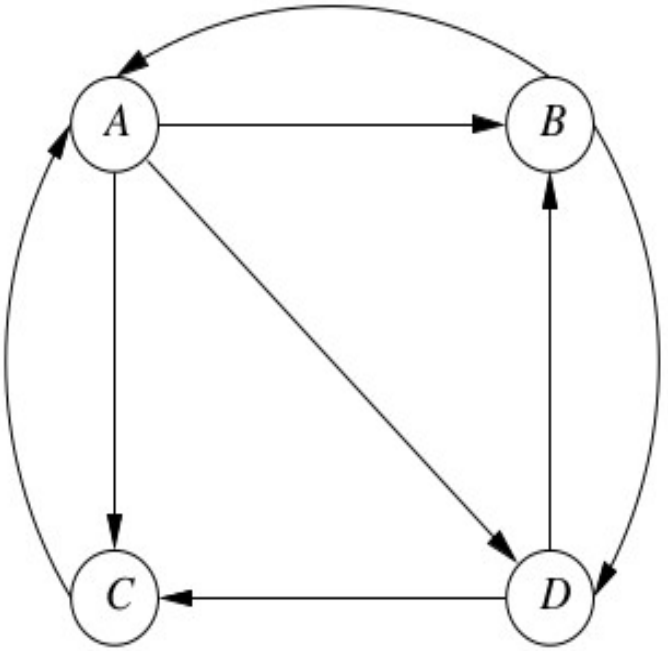
Topic sensitive Page Rank

- Težina, značaj stranice na osnovu njenog sadržaja, slučajna šetnja Internetom zaustavlja se na stranici sa određenom temom
- Primjer, upit sa riječju jaguar može da se odnosi na automobile, životinje, verziju MAC operativnog sistema itd.
    - Zadatak pretraživača da prepozna interesovanje osobe koja zadaje upit
    - Svaki korisnik ima svoj PageRank vektor, neizvodljivo zbog potrebnog memorijskog prostora
    - Topic sensitive metod kreira za mali skup tema po jedan Page Rank vektor, klasifikuje korisnike po temama
    - www.dmoz.org klasifikacija Internet stranica

Biased random walk

- Prepoznali smo stranice koje odgovaraju temi SPORT
- Kreiramo topic-sensitive PageRank za SPORT tako što pretpostavljamo da su korisnici zainteresovani samo za sportske stranice
  - Korisnici su na sportskim stranicama ili na stranicama koje su, u nekoliko linkova, dostupne sa sportskih stranica
  - Formula $v' = \beta M v + (1 - \beta) e_S / |S|$, gdje je S skup stranica sa određenom temom, e_S je vektor koji sadrži 1 za stranice koje su u S, inače sadrži 0

Primjer, beta = 0.8



$$\beta M = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$

Primjer, nastavak, S = {B, D}

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$\begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix} \begin{bmatrix} 2/10 \\ 3/10 \\ 2/10 \\ 3/10 \end{bmatrix} \begin{bmatrix} 42/150 \\ 41/150 \\ 26/150 \\ 41/150 \end{bmatrix} \begin{bmatrix} 62/250 \\ 71/250 \\ 46/250 \\ 71/250 \end{bmatrix} \cdots \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix}$$

Integracija topic sensitive PageRank

- Potrebno je da pretraživač
  - izabere teme, npr. DMOZ
  - izabere skup S za svaku od tema, i pomoću skupa S formira topic-sensitive PageRank vektore
  - na osnovu upita prepozna skup relevantnih tema
  - sortira rezultat upita na osnovu PageRank vektora sa prepoznatu temu

Prepoznavanje teme

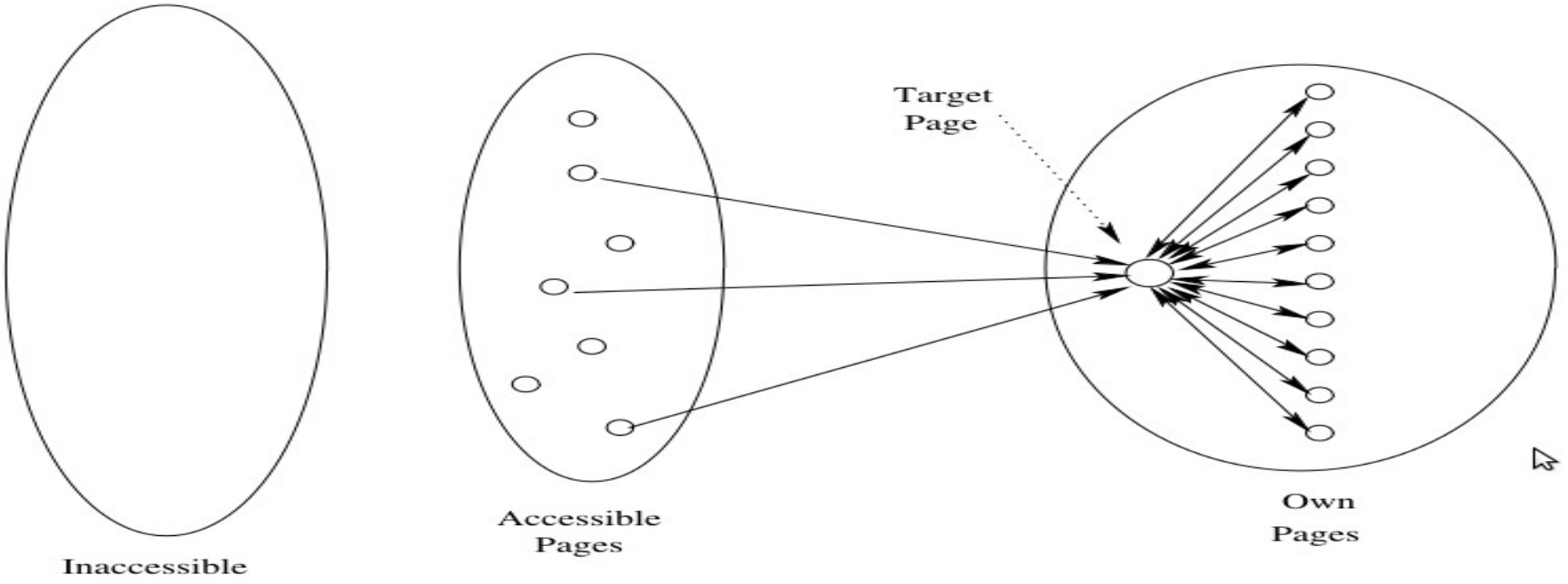
- Omogućiti da korisnik izabere temu iz menija
- Prepoznati temu na osnovu riječi iz stranica koje je korisnik posjetio ili na osnovu upita koje je postavio u posljednje vrijeme
- Prepoznati temu na osnovu raspoloživih informacija o korisniku, npr. sa facebook profila

Mapiranje riječi u teme

- Ideja je da se teme prepoznaju na osnovu riječi koje se jako često pojavljuju u dokumentima sa tom temom
  - Znamo frekvenciju riječi
  - Izdvojimo stranice koje su o određenoj temi
  - Izračunamo ponovo frekvencije i izdvojimo riječi čije su frekvencije značajno veće od polaznih
  - Stranica P se pridružuje temi sa kojom ima najveći Jakardov koeficijent

Link spam

- Link spam – vještačko uvećavanje PageRank-a
- Spam farm – kolekcija stranica čiji je cilj uvećanje PageRank-a neke određene stranice ili grupe stranica



Spam farm

- Inaccessible pages – nedostupne stranice
- Accessible pages – stranice koje ne kontrolišu spameri, ali mogu da utiču na njih
- Own pages – stranice koje spameri kontrolišu, sačinjavaju spam farmu zajedno sa linkovima od spolja, bez spoljašnjih linkova farma je neupotrebljiva
    - Spoljni link na blogovima ili sličnim sajtovima
- Target page – stranica za koju je cilj postići maksimalni PageRank, veliki broj supporting stranica sa linkovima u oba smjera

Spam farm analiza

- Neka je beta = 0.85, n je ukupan broj veb stranica, t je target stranica, m je broj supporting stranica, x je suma PageRank svih stranica sa linkovima ka t, y je PageRank za t
- PageRank za supporting stranice je

$\beta y/m + (1 - \beta)/n$

- PageRank y za t potiče od
  - Sume x
  - Supporting stranica $\beta(\beta y/m - (1 - \beta)/n)$
  - $(1 - \beta)/n$, što je zanemarljivo

Spam farm analiza (2)

- Rješavanjem po y dobija se

$$y = x + \beta m\left(\frac{\beta y}{m} + \frac{1-\beta}{n}\right) = x + \beta^2 y + \beta(1-\beta)\frac{m}{n}$$

$$y = \frac{x}{1-\beta^2} + c\frac{m}{n}$$

$$c = \beta(1-\beta)/(1-\beta^2) = \beta/(1+\beta)$$

Eliminacija link spam-a

- Loakcija spam farme, prepoznati stranicu sa linkovima ka velikom broju drugih stranica, pri čemu postoje i povratni linkovi, izbrisati ovu strukturu iz pretrage
    - Spameri mogu da konstruišu isti efekat ali koristeći drugačiju organizaciju spam farme
- TrustRank, automatski smanjuje PageRank za spam stranice
- Spam mass, identifikovanje spam stranica i njihovo brisanje ili drastično smanjivanje njihovog PageRank-a

Trust Rank za link spam

- Trust Rank je topic-sensitive PageRank, gdje je “topic” skup stranica za koje znamo da nijesu spam
  - Ideja: spam stranica može da ima linkove na stranice koji nijesu spam, obratno nije očekivano
  - Problem su blogovi ili slične stranice gdje svaki korisnik može da postavi svoje linkove, ovakve stranice ne mogu da se računaju u “sigurne”

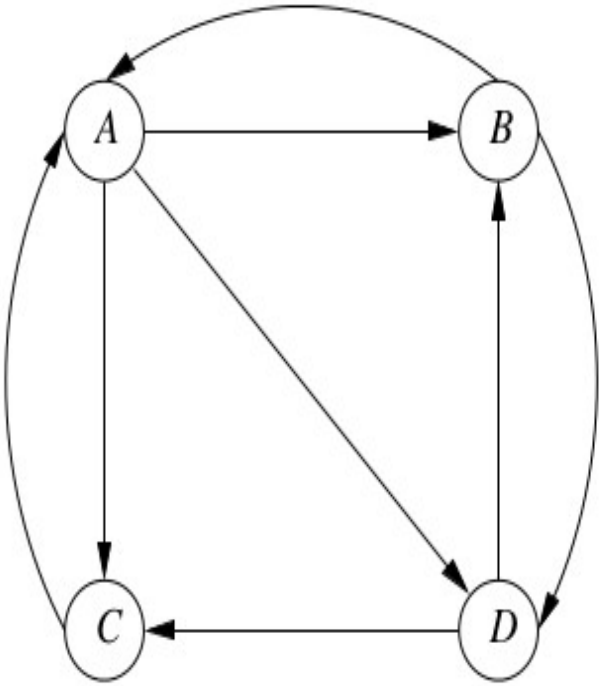
Implementacija Trust Rank

- Potrebno je definisati odgovarajući teleport skup, dva su načina:
  - Korisnik eksplicitno bira skup, moguće po PageRank vrijednosti
  - Izabrati kontrolisane domene, npr. edu ili ac, nije dobro mil ili gov (zašto?)

Spam mass

- Ideja je izračunati koji dio PageRank potiče od spam stranica
- Neka stranica p posjeduje PageRank r i TrustRank t, spam mass je (r – t)/r
    - Negativne ili male pozitivne vrijednosti za spam mass znače da stranica nije spam, vrijednosti blizu 1 znače da je vjerovatno spam
    - Eliminisati stranice koje su spam bez poznavanja strukture “farme” iz koje potiču

Primjer, teleport skup = B, D



Node	PageRank	TrustRank	Spam Mass
A	$3/9$	$54/210$	0.229
B	$2/9$	$59/210$	-0.264
C	$2/9$	$38/210$	0.186
D	$2/9$	$59/210$	-0.264

Koje su stranice spam?

Hubs and Authorities

- HITS je algoritam koji za razliku od PageRank nije faza pre-procesiranja upita, već se izvršava istovremeno sa realizacijom upita
- HITS daje “značaj” stranici iz dva izvora
  - Stranice koje daju informacije o temi su *authorities* i one su značajne
  - Stranice koje ne daju informacije o temi već ukazuju na stranice koje se odnose na određenu temu su takođe značaje, nazivaju se *hubs*
- Poređenje
  - PageRank: stranica je značajna ako je “linkovana” sa značajnih stranica
  - HITS: stranica je dobar hab ako linkuje dobre autoritete, stranica je dobar autoritet ako linkovana sa dobrih habova

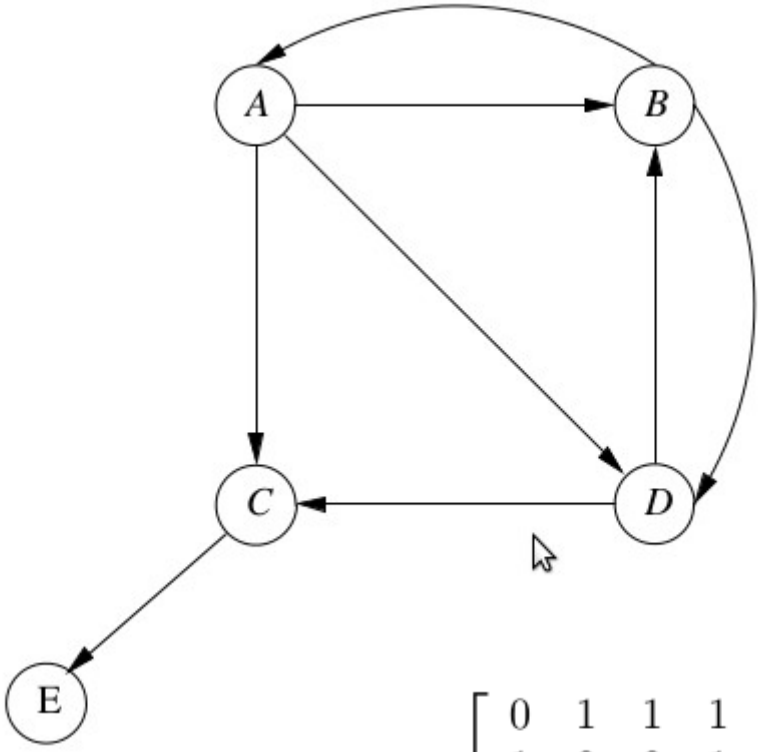
Formalizacija

- Stranice se rangiraju sa
  - *hubiness*, koliko je dobar hab
  - *authority*, koliko je dobar autoritet
- Uz pretpostavku da su stranice numerisane, imamo dva vektora h i a
  - Vrijednosti iz vektora su normalizovane ili u sumi daju 1
- Za stranicu p *hubiness* se procjenjuje sabiranjem autoriteta sljedbenika, *authority* se procjenjuje sabiranjem mjere hubiness za prethodnike

Formalizacija (2)

- Matrica linkova L, ako je ukupno n stranica, matrica L je $n \times n$, $L_{ij} = 1$ ako postoji link sa stranice i na stranicu j, inače je $L_{ij} = 0$, L^T je transponovana matrica od L, $L^T_{ij} = 1$ ako postoji link sa stranice j na stranicu i, inače je $L^T_{ij} = 1$
- L^T je slična matrici M iz PageRank algoritma, ali sa razlikom da umjesto 1 u L^T u matrici M je 1 / broj izlaznih linkova sa stranice koja odgovara posmatranoj koloni

Primjer



$$L = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad L^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Formalizacija (3)

- Mjera hubiness je proporcionalna sa sumom autoriteta sljedbenika, $h = \lambda L a$
- Mjera autority je proporcionalna sa sumom hubiness vrijednosti za prethodnike, $a = \mu L^T h$
- Zamjenom prve u drugu jednakost dobija se
  - $h = \lambda L \mu L^T h = \lambda \mu L L^T h$
  - $a = \mu L^T \lambda L a = \lambda \mu L^T L h$

Formalizacija (4)

1. neka je h vektor sa svim 1
2. $a = L^T h$, normalizovati da je najveći broj 1
3. $h = La$, normalizovati da je najveći broj 1
4. Ponavljati korake 3 i 4 do trenutka kada se vrijednosti vektora h i a ne mijenjaju značajno

Primjer

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1/2 \\ 1 \\ 1 \\ 1 \\ 1/2 \end{bmatrix} \quad \begin{bmatrix} 3 \\ 3/2 \\ 1/2 \\ 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1/2 \\ 1/6 \\ 2/3 \\ 0 \end{bmatrix}$$

$$\mathbf{h} \qquad L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

$$\begin{bmatrix} 1/2 \\ 5/3 \\ 5/3 \\ 3/2 \\ 1/6 \end{bmatrix} \quad \begin{bmatrix} 3/10 \\ 1 \\ 1 \\ 9/10 \\ 1/10 \end{bmatrix} \quad \begin{bmatrix} 29/10 \\ 6/5 \\ 1/10 \\ 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 12/29 \\ 1/29 \\ 20/29 \\ 0 \end{bmatrix}$$

$$L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

.....

$$\mathbf{h} = \begin{bmatrix} 1 \\ 0.3583 \\ 0 \\ 0.7165 \\ 0 \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} 0.2087 \\ 1 \\ 1 \\ 0.7913 \\ 0 \end{bmatrix}$$