PageRank algoritam na Wikipediji

Vuk Vuković 0119/17
Filip Vesović 0018/17

Verovatnoća i statistika (13S082VS)
Elektrotehnički fakultet,
Univerzitet u Beogradu

Jun 2019.

Uvod

- Prvi internet pretraživači
- Google - više stotina milijardi indeksiranih stranica
- Koje stranice prikazati korisniku za zadatu pretragu?
- Po kom redosledu sortirati stranice ukoliko je dostupno više rezultata?

PageRank algoritam

- Larry Page i Sergey Brin
- Studenti doktorskih studija, Univerzitet Stanford
- 1996. razvijaju PageRank algoritam kao deo istraživačkog projekta o internet servisu za pretragu
- Glavna ideja: rangiranje informacija na internetu po kredibilitetu (bitnosti stranica na kojima se one nalaze)
- Bitnost jedne stranice zavisi kako od stranica koje nju linkuju, tako i bitnosti tih stranica
- Nakon nekoliko godina osnivaju Google

PageRank algoritam danas

- Koristi se kao jedan od kriterijuma na osnovu kojih se rangiraju rezultati pretrage
- Google ovaj algoritam izvršava neprestano na velikom broju servera
- Primena algoritma na manji skup podataka
- Engleska verzija slobodne enciklopedije Wikipedia dosutpna za preuzimanje

Slučajni procesi

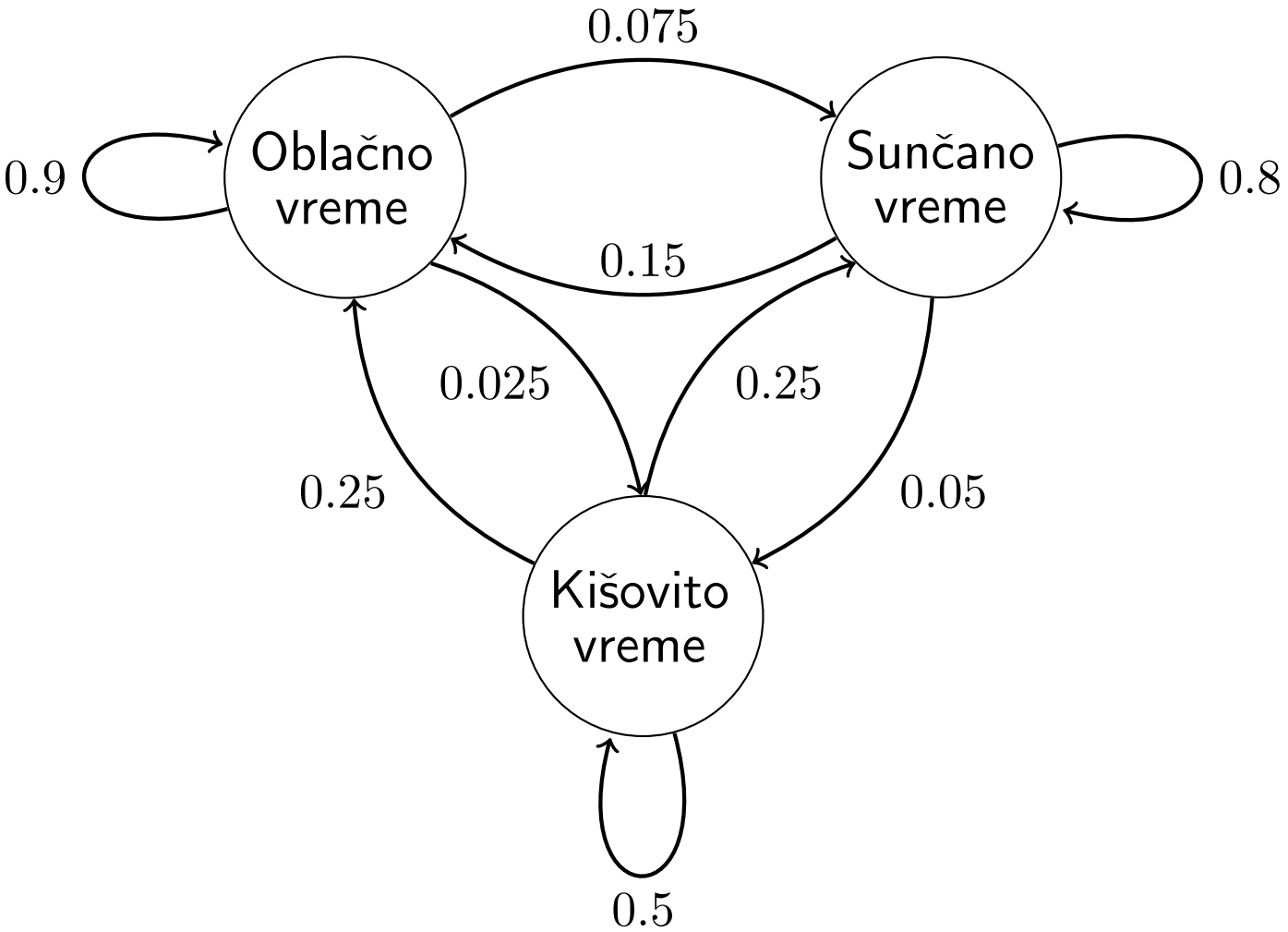
- Slučajni proces je familija slučajnih promenjivih $\{X_t\}_{t \in T}$
- Ako je skup T diskretan tada se radi o slučajnom procesu sa diskretnim vremenom
- Skup svih mogućih vrednosti slučajnih promenljivih X_t za $t \in T$ zove se skup stanja procesa X_t *(Merkle 2016)*

Markovski lanac

- Markovski lanac predstavlja slučajni proces sa diskretnim vremenom i konačnim skupom stanja
- U svakom Markovskom lancu važi da verovatnoća prelaska iz stanja X_k u stanje X_{k+1} ne zavisi od stanja u kojima je sistem bio u prethodnim trenucima, već isključivo od trenutnog stanja X_k (osobina Markova)

$$P(X_{k+1} = x | X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = P(X_{k+1} = x | X_k = x_k)$$

Primer (Markovski lanac)



Iterativni metodi

- Metodi koji na osnovu početne pretpostavke i zadatog algoritma, iterativno menjaju rešenje
- Konvergencija (sekvenca iteracija konvergira za zadate početne uslove)
- Striktnu matematička pozadina nije obrađena s obzirom da se radi o inženjerskom algoritmu pre svega zasnovanog na intuitivnim pretpostavkama

$$X_0, X_1, X_2, ..., X_k, ...X_\infty$$

PageRank algoritam

- Određivanje verovatnoća da će se nasumični korisnik pronaći na određenoj stranici posle (beskonačno) mnogo koraka
- Inicijalno, svaka stranica ima jednaku verovatnoću nalaska korisnika
- Iterativnim postupkom na osnovu raspodele verovatnoća u trenutku $t = k$ određuje se raspodela verovatnoća u narednom trenutku $t = k + 1$
- Nije moguće odrediti krajnje rešenje konvergencije ($t = \infty$)
- Prekid rada kada razlika između dve uzastopne raspodele padne ispod unapred definisane vrednosti

PageRank algoritam

Model Markovljevih lanaca koji PageRank koristi je sledeći *(Page et al., 1999)*:

- Svaka stranica predstavlja jedno od mogućih stanja
- Sa svake stranice moguće je preći na bilo koju drugu stranicu sa verovatnoćom δ
- Preostala verovatnoća $1 - \delta$ je uniformno raspodeljenja za prelazak na linkove te stranice ($\frac{1-\delta}{L}$, gde L predstavlja ukupan broj linkova na stranici)

Faktor odbacivanja

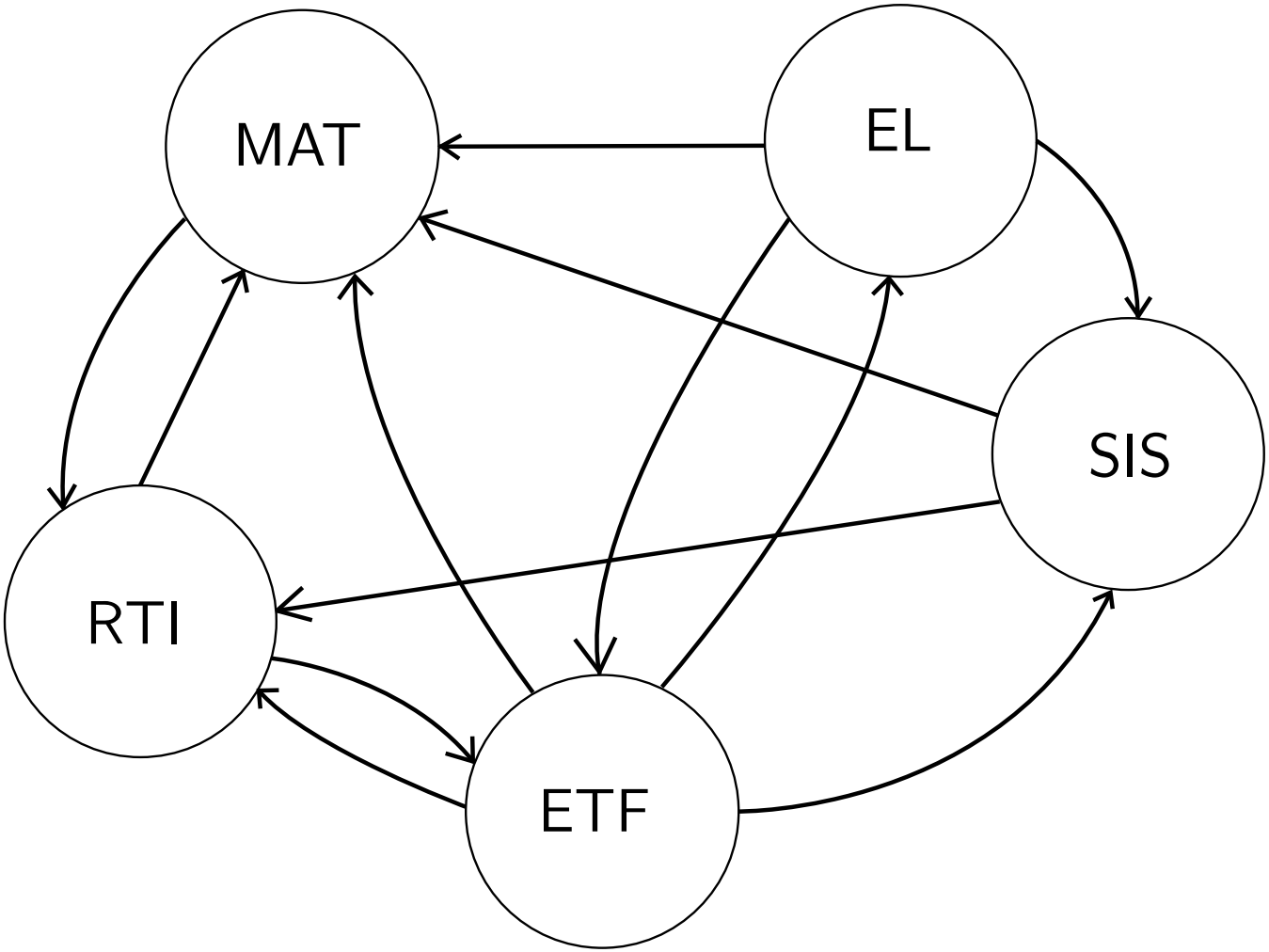
- Parameter δ se još naziva i stepenom odbacivanja (engl. *Damping factor)*
- Postojanje verovatnoće da korisnik umesto klika na neki od linkova na stranici pređe na drugu stranicu unošenjem adrese stranice direktno (ili pretragom)
- Neophodan za ispravan rad algoritma
- Stranice bez izlaznih linkova bi imale najveći PageRank (korisnici ih ne mogu napustiti)

Mera bitnosti (relevantnosti)

Smisao PageRank-a kao mere bitnosti stranica na internetu:

- Stranice koje imaju više linkova ka njima teže da imaju veći PageRank (više stranica se oslanja na nju, smtraju je bitnom i verodostojnom)
- Stranice koje imaju veći PageRank doprinose značajnije PageRank-u stranica koje linkuju nego one sa manjim (pokazatelj verodostojnosti jeste ukoliko se neka verodostojna stranica oslanja na nju)

PageRank primer



PageRank primer

	ETF	RTI	Matematika	SIS	Elektronika
0	0.2000	0.2000	0.2000	0.2000	0.2000
1	0.1667	0.3500	0.3167	0.1167	0.0500
2	0.1917	0.4167	0.2917	0.0583	0.0417
3	0.2222	0.3688	0.2993	0.0618	0.0479
4	0.2003	0.3858	0.2868	0.0715	0.0554
∞	0.2069	0.3793	0.2931	0.0690	0.0517

- Za koju stranicu bismo intuitivno rekli da će imati najveći PageRank?
- Da li se to poklapa sa rezultatima?

Algoritam

1: **procedure** PageRank
2: $\mathcal{P}(s) \leftarrow {}^{1}/_{N},\ \forall s \in \mathcal{S}$
3: **repeat**
4: $\mathcal{P}'(s) \leftarrow 0,\ \forall s \in \mathcal{S}$
5: **for each** $s \in \mathcal{S}$ **do**
6: **for each** $f \in in(s)$ **do**
7: $\mathcal{P}'(s) \leftarrow \mathcal{P}'(s) + \dfrac{(1-\delta)\mathcal{P}(f)}{outCnt(f)}$
8: $unused \leftarrow 0$
9: **for each** $s \in \mathcal{S}$ **do**
10: **if** $outCnt(s) = 0$ **then**
11: $unused \leftarrow unused + \mathcal{P}(s)$
12: **else**
13: $unused \leftarrow unused + \delta\mathcal{P}(s)$
14: **for each** $s \in \mathcal{S}$ **do**
15: $\mathcal{P}'(s) \leftarrow \mathcal{P}'(s) + {}^{unused}/_{N}$
16: $\Delta\mathcal{P} = \sum |\mathcal{P} - \mathcal{P}'|$
17: $\mathcal{P} \leftarrow \mathcal{P}'$
18: **until** $\Delta\mathcal{P} > \varepsilon$

Wikipedia

- Dozvoljava preuzimanje čitave engleske verzije (5,5 miliona članaka)
- XML datoteka od preko 70GB
- Svaki članak se sastoji od naslova, određenih tehničkih podataka i sadržaja
- Interni i eksterni linkovi (odbačeni eksterni, šablonski i kategorijski)

WIKIPEDIA
The Free Encyclopedia

Obrada podataka

- Program za obradu podataka napisan u C++
- Svaki članak opisan je na sledeći način:
  - Naziv članka
  - Automatski dodeljen identifikator
  - Niz identifikatora članaka koje posmatrani članak linkuje
- S obzirom na količinu podataka, trajanje obrade podataka i izdvajanja potrebnog sadržaja je oko 2.5 časa
- Obradom 72 GB dobijeno je oko 2.6 GB podataka nad kojima je rađen PageRank algoritam

Primena PageRank algoritma

- Kod samog algoritma takođe je napisan u C++
- Faktor odbacivanja koji je korišćen je $\delta = 0.1$
- 150 iteracija
- Krajna greška od $7 \cdot 10^{-12}$ za šta je bilo potrebno oko 15 minuta
- Vreme potrebno za računanje jedne iteracije je oko 6,5 sekundi
- Krajnji rezultat predstavlja sortirana lista imena članaka i njihovih verovatnoća od oko 500 MB

Prvih 20 stranica

1.	United States	11.	List Of Sovereign States
2.	World War II	12.	Association Football
3.	United Kingdom	13.	Italy
4.	France	14.	Canada
5.	Race And Ethnicity In The US	15.	Australia
6.	Germany	16.	The New York Times
7.	India	17.	London
8.	New York City	18.	English Language
9.	Catholic Church	19.	World War I
10.	China	20.	Russia

Tabela: Prvih 20 stranica po PageRank

Prvih 5 stranica po kategorijama

871.	C
1191.	Java
1317.	C++
1903.	Python
1946.	Javascript

Tabela: Prvih 5 programskih jezika po PageRank

243.	Microsoft
307.	Youtube
388.	Apple Inc
462.	Google
488.	IBM

Tabela: Prvih 5 tehnoloških kompanija po PageRank

Prvih 5 stranica po kategorijama

164.	Barack Obama
198.	Carl Linnaeus
216.	Elizabeth II
220.	Napoleon
247.	George W. Bush

Tabela: Prvih 5 ličnosti po PageRank

8.	New York City
17.	London
28.	Washington, D.C.
30.	Paris
49.	Los Angeles

Tabela: Prvih 5 gradova po PageRank

Prvih 5 stranica po kategorijama

93.	Mathematics
224.	Physics
286.	Economics
317.	Linguistics
327.	Philosophy

Tabela: Prvih 5 oblasti po PageRank

12.	Association Football
128.	Basketball
143.	American Football
195.	Cricket
245.	Ice Hockey

Tabela: Prvih 5 sportova po PageRank

Ostali rezultati

88.	Protein
127.	Moth
168.	Microsoft Windows
183.	Serbia
791.	Belgrade
868.	SFRY
12080.	Nikola Tesla
13061.	Novak Djokovic
265825.	Merkle Tree

Tabela: Tabela ostalih rezultata

Link ka prvih 10 000 članaka

Zaključak

- Demonstrirana implementaicija i primena PageRank algoritma na stranicama engleske verzije Wikipedije
- Obrađeno preko 70GB podataka
- Izdvojeni zanimljivi rezultati
- Dobijeni rezultati se poklapaju sa očekivanim (uz određena iznenađenja)
- Moguće unapređenje podrazumeva primenu napisanog programa na drugi skup podataka i dalju analizu

Literatura

- Page, L., Brin, S., Motwani, R., Winograd, T., 1999. *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.
- Merkle, M., 2016. *Verovatnoća i statistika za inženjere i studente tehnike.* Akademska Misao 2016.

Kraj

Hvala na pažnji!