PageRank algoritam na Wikipediji

Filip Vesović 0018/17, Vuk Vuković 0119/17,
Projekat iz predmeta Verovatnoća i statistika (13S082VS)
Elektrotehnički fakultet, Univerzitet u Beogradu

Sažetak—PageRank je algoritam koji služi za kvantitativno određivanje važnosti (uticaja) internet stranica. Zasnovan je na Markovskim lancima i modelu nasumičnog korisnika interneta (korisnika koji se nasumično kreće po stranicama klikovima na sadržane linkove). U ovom radu prikazani su teorijska osnova, pregled PageRank algoritma i rezultati algoritma izvršenog na stranicama engleske verzije internet enciklopedije Wikipedije.

1 Uvod

Još od nastanka prvih internet servisa za pretragu, pa sve do danas, kada najveći internet pretraživač, Google, broji više stotina milijardi indeksiranih stranica, osnovni problem je kako izabrati koje stranice prikazati korisniku za zadatu pretragu. Još jedan od problema predstavlja i kriterijum po kom je potrebno rasporediti prikazane stranice ukoliko je dostupno više rezultata pretrage.

Tadašnji studenti doktorskih studija na Univerzitetu Stanford, *Larry Page* i *Sergey Brin* su razvili PageRank algoritam 1996. godine kao deo istraživačkog projekta o internet servisu za pretragu. Glavna ideja je bila da se informacije na internetu rangiraju po kredibilitetu, tj. po 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, ova dva naučnika osnivaju Google, danas najveći i najkorišćeniji servis za pretragu interneta.

Čak i dan danas, PageRank algoritam se koristi kao jedan od mnogih kriterijuma na osnovu kojih Google rangira internet stranice. S obzirom na veliku količinu podataka koju je potrebno obraditi, Google ovaj algoritam izvršava neprestano na velikom broju servera. Iz tog razloga, prilikom izrade ovog projekta, odlučili smo se da ovaj algoritam primenimo na znatno manju količinu podataka koju je moguće obraditi u kućnim uslovima - najveću slobodnu enciklopediju, Wikipediju. Ipak, uzeti skup podataka je dovoljan da verno prikaže rezultate PageRank algoritma.

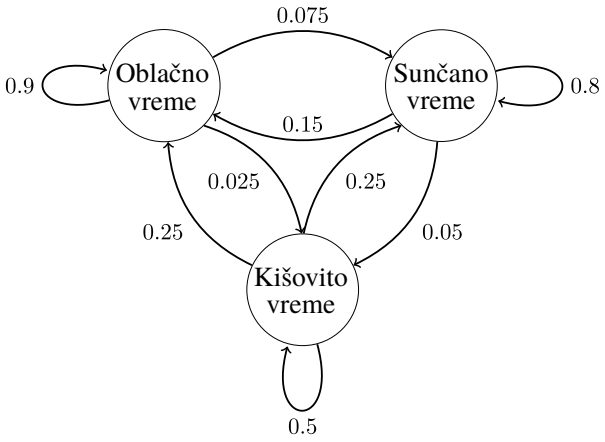
2 Markovski lanci

Slučajni proces je familija slučajnih promenjivih $\{X_t\}_{t\in T}$, gde je T neki beskonačan podskup realnih brojeva. Indeks t se najčešće interpretira kao vreme. 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 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).

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

Na slici 1 je prikazan primer Markovskog lanca sa 3 stanja koja predstavljaju sunčano, oblačno i kišovito vreme u toku jednog sata. Na usmerenim ivicama koje povezuju stanja su zadate verovatnoće za prelazak iz jednog u drugo stanje. Dati primer se može tumačiti na sledeći način: ukoliko je trenutno sunčano vreme, verovatnoća da će u narednom satu vreme biti sunčano je 0.8, oblačno 0.15, a kišovito 0.05.



Slika 1. Primer Markovski lanca za vremensku prognozu

3 Iterativni metodi

Iterativni metodi u matematici su metodi koji na osnovu početne pretpostavke i zadatog algoritma, iterativno menjaju rešenje. Iterativni metod je konvergentan, ukoliko njegova sekvenca iteracija konvergira za zadate početne uslove.

Iako pomenuti metodi imaju striktnu matematičku pozadinu koja obuhvata dokazivanje konvergencije, to u ovom radu nije obrađeno s obzirom da PageRank predstavlja inženjerski algoritam koji je samim tim zasnovan na intuitivnim i inženjerskim pretpostavkama.

4 PageRank

Ideja PageRank algoritma zasniva se na izračunavanju verovatnoća da se nasumični korisnik pronađe na određenoj stranici posle (beskonačno) mnogo koraka.

PageRank započinje izvršavanje dodeljujući svakoj stranici jednaku početnu verovatnoću za pronalskom slučajnog korisnika na toj stranici. Nakon toga se iterativnim postupkom na osnovu raspodele verovatnoća u trenutku $t = k$ određuje raspodela verovatnoća u narednom trenutku $t = k + 1$. Kako ovim postukom nije moguće odrediti krajnje rešenje konvergencije ($t = \infty$), algoritam prekida sa radom kada razlika između dve uzastopne raspodele padne ispod unapred definisane vrednosti.

Bitno je napomenuti da postoje i neiterativni metodi za određivanje rešenja PageRank algoritma, ali s obzirom na njihovu vremensku složenost oni nisu od praktičnog značaja.

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

1) Svaka stranica predstavlja jedno od mogućih stanja.
2) Sa svake stranice moguće je preći na bilo koju drugu stranicu sa verovatnoćom δ.
3) Preostala verovatnoća $1 - \delta$ je uniformno raspodeljenja na linkove te stranice, tj. verovatnoća da se pređe na stranicu čiji link se nalazi na trenutnoj stranici je $\frac{1-\delta}{L}$, gde L predstavlja ukupan broj linkova na stranici.

Parameter δ se još naziva i stepenom odbacivanja (engl. *Damping factor*). Ovaj parametar je neophodan za ispravan rad algoritma s obzirom da bi bez njegovog postojanja rezultat konvergencije bio da stranice koje nemaju izlazne linkove imaju najveći PageRank indeks (ostale stranice bi imale verovatnoću 0, slučajan korisnik nikada ne bi napustio ovakve stranice). Objašnjenje ovog parametra u modelu nasumičnog korisnika interneta je postojanje verovatnoće da korisnik umesto klika na neki od linkova na stranici pređe na drugu stranicu unošenjem adrese stranice direktno.
Smisao PageRank-a kao mere bitnosti stranica na internetu:

1) Stranice koje imaju više linkova ka njima teže da imaju veći PageRank. Što više linkova usmerava ka određenoj stranici, to znaci da se više stranica oslanja na nju, tj. smatra je za verodostojnu i bitnu.
2) Stranice koje imaju veći PageRank doprinose značajnije PageRank-u stranica koje linkuju nego one sa manjim. Ako stranicu linkuje neka bitna stranica, to je značajniji pokazatelj da je stranica bitna nego linkovanje od manje bitnih stranica.

Kao ilustraciju funkcionisanja PageRank algoritma prikazaćemo njegovo izvršavanje na hipotetičkom primeru povezanosti internet stranica katedri na Elektrotehničkom fakultetu (slika 2).

Zbog jednostavnosti, za faktor odbacivanja u ovom primeru uzeto je $\delta = 0$ (jedini prelazi koji su mogući su prelazi dati na slici). Početna vrednost verovatnoća je $\{0.2, 0.2, 0.2, 0.2, 0.2\}$.

U tabeli 1 prikazane su prve 4 iteracije, kao i krajanja iteracija (nakon konvergencije) za pomenuti

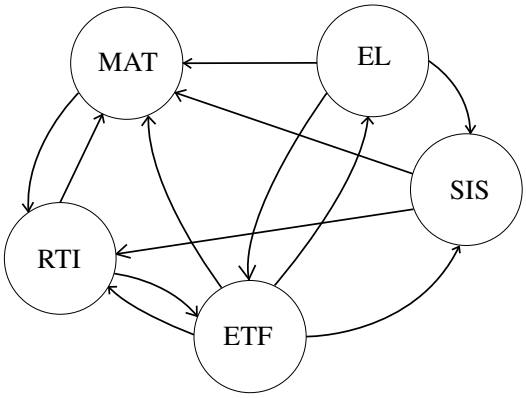
	ETF	RTI	Matematika	SIS	Elektronika
0	0.2000	0.2000	0.2000	0.2000	0.2000
1	0.1667	0.3500	0.3167	0.1167	0.0500
2	0.1917	0.4167	0.2917	0.0583	0.0417
3	0.2222	0.3688	0.2993	0.0618	0.0479
4	0.2003	0.3858	0.2868	0.0715	0.0554
∞	0.2069	0.3793	0.2931	0.0690	0.0517

Tabela 1. PageRank iteracije za primer povezanosti stranica ETF katedri

primer.

Iako je za pretpostaviti da će najbolje rangirana stranica biti stranica Katedre za primenjenu matematiku s obzirom na to da sve ostale stranice linkuju nju, rezultati pokazuju drugačije. Stranica Katedre za računarsku tehniku i informatiku je najbolje rangirana zbog toga što stranica MAT (koja je izuzetno relevantna zbog broja ulaznih linkova) linkuje isključivo nju.

Kako stranica SIS linkuje dve stranice (EL i ETF) sa verovatnoćama 0.2, gde EL ima tri linka, a ETF četiri linka, verovatnoću stranice SIS u prvoj iteraciji dobijamo po formuli $^{0.2}/_{3} + {}^{0.2}/_{4} = 0.1167$.



Slika 2. Hipotetički primer povezanosti ETF katedri

5 Algoritam

Pseudokod algoritma implementiranog u ovom radu prikazan je algoritmom 1.

Algoritam 1 PageRank algoritam
$\mathcal{S}$ - skup stranica
N - broj stranica
δ - stepen odbacivanja
$in(s)$ - skup stranica koje linkuju stranicu s
$outCnt(s)$ - broj linkova na stranici s
ε - kriterijum zaustavljanja

procedure PageRank
  $\mathcal{P}(s) \leftarrow {}^{1}/_{N}$, $\forall s \in \mathcal{S}$
  repeat
    $\mathcal{P}'(s) \leftarrow 0$, $\forall s \in \mathcal{S}$
    for each $s \in \mathcal{S}$ **do**
      for each $f \in in(s)$ **do**
        $$\mathcal{P}'(s) \leftarrow \mathcal{P}'(s) + \frac{(1-\delta)\mathcal{P}(f)}{outCnt(f)}$$
    $unused \leftarrow 0$
    for each $s \in \mathcal{S}$ **do**
      if $outCnt(s) = 0$ **then**
        $unused \leftarrow unused + \mathcal{P}(s)$
      else
        $unused \leftarrow unused + \delta\mathcal{P}(s)$
    for each $s \in \mathcal{S}$ **do**
      $\mathcal{P}'(s) \leftarrow \mathcal{P}'(s) + {}^{unused}/_{N}$
    $\Delta\mathcal{P} = \sum |\mathcal{P} - \mathcal{P}'|$
    $\mathcal{P} \leftarrow \mathcal{P}'$
  until $\Delta\mathcal{P} > \varepsilon$

6 Wikipedia

Wikipedia, kao najveća slobodna enclikopedija, dozvoljava preuzimanje celokupnog sadržaja engleske verzije sajta koji sadrži preko 5.5 miliona članaka. Svaki članak se sastoji od naslova, određenih tehničkih podataka i sadržaja.

Unutar sadržaja članaka nalaze se dva tipa linkova: interni linkovi, koji predstavljaju veze ka drugim člancima Wikipedije, kao i eksterni linkovi koji uglavnom predstavljaju reference ka raznim knjigama, časopisima, naučnim radovima i drugim sajtovima koji predstavljaju izvor informacija. U algoritmu predstavljenim u ovom radu korišćeni su isključivo interni linkovi, dok su svi eksterni linkovi bili odbačeni. Takođe su odbačeni specijalni linkovi koji vode ka kategorijama i šablonskim stranicama, s obzirom da su oni automatski linkovani na većini stranica, a prosečan korisnik ih ne posećuje.

Čitava engleska verzija sajta je dostupna u okviru jedne XML (engl. *Extensible Markup*

Language) datoteke. Ovako preuzeta datoteka teži preko 70 GB podataka. Kako bismo iskoristili pomenute podatke, bilo je potrebno obraditi ih na odgovarajući način i izdvojiti samo informacije potrebne za izvršavanje PageRank algoritma. Program za obradu napisan je u programskom jeziku C++. Kod programa za obradu moguće je pronaći na adresi u prilogu.

Svaki članak opisan je na sledeći način:

1) Naziv članka
2) Automatski dodeljen identifikator
3) Niz identifikatora članaka koje posmatrani članak linkuje

S obzirom na količinu podataka, trajanje obrade podataka i izdvajanja potrebnog sadržaja je oko 2.5 časa.

7 Primena PageRank algortima

Izdvajanjem naslova i linkova iz fajla veličine 72 GB dobijeno je oko 2.6 GB podataka nad kojima je rađen PageRank algoritam. Primena algoritma izvršena je u 150 iteracija, sa krajnjom greškom od $7 \cdot 10^{-12}$ za šta je bilo potrebno oko 15 minuta. Faktor odbacivanja koji je korišćen je $\delta = 0.1$. 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 podataka.

Sama primena PageRank algoritma nad obrađenim podacima takođe je napisana u programskom jeziku C++. Kod programa za primenu PageRank algoritma moguće je pronaći na adresi u prilogu.

8 Rezultati i diskusija

Prvih 20 stranica

U tabeli 2 prikazano je prvih 20 stranica sortiranih opadajuće po PageRank algoritmu.

Primećuje se da ovih 20 stranica pre svega čine države što se objašava time da veliki broj članaka linkuje povezane države (na primer odakle potiče poznata ličnost, kojoj državi pripada grad, lokacija događaja). Takođe se mogu primetiti i dva grada: New York i London. Pomalo iznenađujući rezultat je da su Katolička crkva i fudbal zauzeli 9. i 12. mesto na listi.

Kako stranica *USA* ima PageRank u vrednosti od 0.00138227, ovaj rezultat možemo tumačiti da bi za veliki broj nasumičnih korisnika, 1.3‰ bio očekivan deo korisnika koji su na stranici *USA*.

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

Tabela 2. Prvih 20 stranica po PageRank

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

Tabela 3. Prvih 5 programskih jezika po PageRank

Prvih 5 stranica po kategorijama

U tabelama 3, 4, 5, 6, 7 i 8 prikazani su po prvih 5 najbolje rangiranih stranica iz različitih kategorija, kao i njihov globalni rang. Primetimo da je *Carl Linnaeus* rangiran kao 2. po PageRank indeksu u tabeli poznatih ličnosti. Radi se o švedskom botaničaru koji je uveo nomeklaturu

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

Tabela 4. Prvih 5 tehnoloških kompanija po PageRank

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

Tabela 5. Prvih 5 ličnosti po PageRank

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

Tabela 6. Prvih 5 gradova po PageRank

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

Tabela 7. Prvih 5 oblasti po PageRank

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

Tabela 8. Prvih 5 sportova po PageRank

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

Tabela 9. Tabela zanimljivih rezultata

nazivanja organizama. Iako stranice organizama ne linkuju stranicu ovog naučnika, već stranice o pomenutim nomenklaturama, *Carl Linnaeus* je rangiran ovako visoko. To je iz razloga što njegov članak linkuju verodostojne stranice nomenklatura. Sličan ishod desio se u primeru na slici 2.

Ostali zanimljivi rezultati

U tabeli 9 prikazan je globalni rang još nekih članaka koji bi verovatno bili zanimljivi čitaocima.

Sortirana lista prvih 10000 članaka sa izračunatim PageRank indeksima može se pronaći na adresi u prilogu.

9 Zaključak

U ovom radu demonstrirana je primena PageRank algoritma na stranicama engleske verzije internet enkciklopedije Wikipedia. Izvrešena je obrada preko 70GB podataka (izdvajanje članaka zajedno sa linkovima koji vode ka ostalim stranicama). Prikazano je prvih 20 članaka raspoređeno po PageRank indeksu, kao i po 5 najbolje rangiranih stranica po određenim kategorijama. Dobijeni rezultati se poklapaju sa očekivanim. Ipak, moguće je primetiti i određene zanimljivosti za koje se ne bi pretpostavilo da budu tako visoko rangirane. Moguća unapređenja ovog projekta podrazumevaju pre svega primenu napisanog programa na još neki skup podataka na kojem bi se mogla izvršiti dalja analiza.

Literatura

[1] Page, L., Brin, S., Motwani, R., Winograd, T., 1999. *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.

[2] Merkle, M., 2016. *Verovatnoća i statistika za inženjere i studente tehnike.* Akademska Misao 2016.

Prilog

Izvorni kod programa za obradu i primenu PageRank algoritma, kao i lista prvih 10000 članaka sa njihovim verovatnoćama se može pronaći na sledećoj stranici:
https://github.com/FilipVesovic/WikiRank