PageRank i slučajni posjetitelj internetskih stranica

Tvrtko Tadić[1]

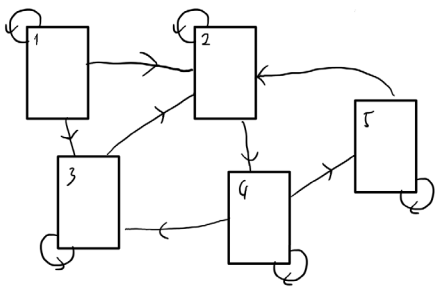
U dosadašnjim člancima podatci su uglavnom bili brojevi, no podatci mogu imati mnogo složeniju strukturu. Takve su, na primjer, web stranice širom interneta i problem njihova *rangiranja*. Glavni napredak u rangiranju web stranica postignut je 1996. godine kada su tada doktorski studenti Larry Page i Sergey Brin došli do ideje ***PageRanka*** nakon koje su osnovali danas vodeću tražilicu *Google*.

U ovom članku dajemo jednostavno objašnjenje tj. interpretaciju PageRanka za razumijevanje za koje je dovoljno osnovno poznavanje vjerojatnosti i nije potrebno znati pojmove linearne algebre.

Pojednostavljeni model Interneta

Za potrebe ovoga članka internet će imati 5 web stranica kao na slici 1. Pristup s više web stranica je identičan:

- Stranice ćemo nazvati brojevima 1, 2, 3, 4, 5.
- Kao na slici 1, ukoliko stranica i ima poveznicu (link) koja šalje posjetitelja na stranicu j, stavljamo strelicu između tih dviju stranica.



Slika 1. Web stranice i njihova povezanost

- Vezu bilježimo i->j. Tako se na stranici 1 nalazi poveznica na stranicu 2 i stranicu 3.
- Također, kako posjetitelj može odlučiti ostati na postojećoj stranici, stavit ćemo da svaka stranica ima poveznicu na samu sebe, tj. i->i.

[1]Tvrtko Tadić, Microsoft, Redmond

Slučajni posjetitelj web stranica

Slučajni *surfer* kreće obilaziti internet na idući način:

- U svakoj jedinici vremena surfer se nalazi na jednoj web stranici.
- Polaznu web stranicu (na kojoj se nalazi u prvoj vremenskoj jedinici) odabire s jednakom vjerojatnošću za sve stranice.
- U svakoj idućoj jedinici vremena surfer ima mogućnost prijeći na bilo koju stranicu za koju ima poveznicu ili ostane na postojećoj web stranici. Stranicu na koju prelazi (ili ostaje) surfer odabire s jednakom vjerojatnošću među izborom koji ima.

Pseudo kôd koji opisuje ovaj algoritam može se zapisati ovako:

unos: broj_vremenskih_jedinica, web_stranice_s_poveznicama

izlaz: posjećene_web_stranice

za trenutni_korak = 1 ... broj_vremenskih_jedinica **radi**
    ako (trenutni_korak == 1) **onda**
        posjećene_web_stranice[trenutni_korak]
        = slučajno_odaberi_stranicu(web_stranice_s_poveznicama);

    inače
        posjećene_web_stranice[trenutni_korak] =
        slučajno_odaberi_stranicu(stranice_na_koje_povezuje (posjećene_web_stranice
        [trenutni_korak - 1]) unija {posjećene_web_stranice[trenutni_korak - 1]});

Ako surfer provodi *puno vremena* obilazeći internet, ***koliki će postotak vremena provesti na svakoj od pojedinih stranica***?

Simulacija obilaska web stranica

Postupak posjećivanja web stranica koji smo opisali sada ćemo simulirati na računalu. (Za detalje računalne implementacije vidi dodatak na kraju članka.)

Pretpostavit ćemo da surfer provede $n = 10\,000$ jedinica vremena obilazeći internetske stranice.

Pogledajmo nekoliko realizacija ovog postupka:

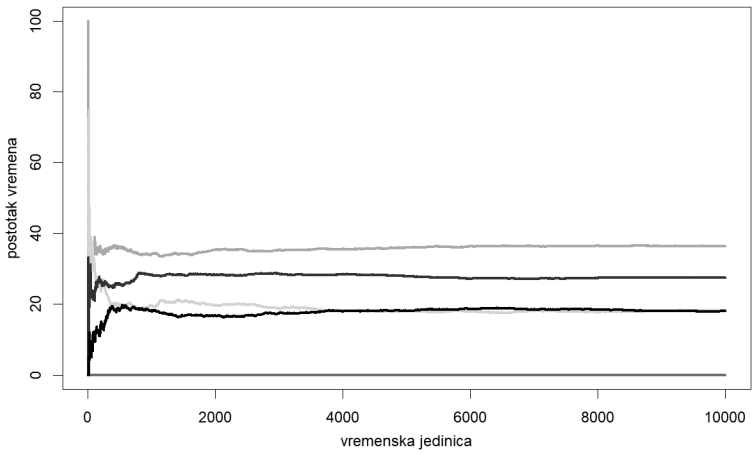
1	2	3	4	5
0.02 %	36.32 %	18.14 %	27.14 %	18.38 %
0 %	36 %	18.35 %	27.02 %	18.63 %
0 %	36.68 %	18.15 %	26.66 %	18.51 %
0 %	36.63 %	18.99 %	27.14 %	17.24 %
0 %	36.68 %	17.63 %	27.38 %	18.31 %

Tablica 1. Postotak vremena koje je slučajni surfer proveo na pojedinoj stranici u 10000 vremenskih jedinica

Iz tablice 1 može se izvesti nekoliko zaključaka:

- Brojevi u svakom stupcu ***otprilike*** **su isti**.
- Stranica 2 je ***najposjećenija***. Uočimo (sa Slike 1.) da na ovu stranicu upućuje i **najviše** drugih stranica – čak 3.
- Stranica 4 je ***druga najposjećenija***.
- Stranica 1 je ***najmanje posjećena*** i na nju nema poveznicu **ni jedna** od preostalih stranica.

U svakom trenutku bilježit ćemo postotke vremena koliko je (do tog trenutka) slučajni surfer proveo na pojedinoj stranici do neke vremenske jedinice. Slika 2. ilustrira razvoj situacije:



Slika 2. Postotak vremena proveden na pojedinoj stranici kroz vremenske jedinice

Slika 2. pokazuje nam da se postotci relativno brzo stabiliziraju. Ovo se može strogo matematički objasniti, tj. pokazati da se postotak vremena proveden na određenoj stranici relativno brzo približi određenoj vrijednosti. Možemo reći da se na stranice na kojima će slučajni surfer provesti najviše vremena mogu poredati ovako:

Rang posjećenosti	Stranica
1.	2
2.	4
3. – 4.	3, 5
5.	1

Tablica 2. Rang lista najposjećenijih stranica

Ovo rangiranje ideja je PageRanka.

Matematička pozadina

Opisani postupak spada u skupinu matematičkih slučajnih (stohastičkih) modela koje zovemo **Markovljevi lanci.** U takvim modelima s vremenom model mijenja stanje u kojem se nalazi po određenom **vjerojatnosnom pravilu**.

Jedan od glavnih rezultata teorije Markovljevih lanaca kaže da postotak vremena provedenog u jednom stanju (uz određene uvjete) teži konstantnoj vrijednosti. Nadalje, **konstantne granične vrijednosti** zadovoljavaju **sustav linearnih jednadžbi**.

Kada riješimo sustav linearnih jednadžbi, dobivamo da kroz dulje vrijeme udio vremenskih jedinica provedenih na pojedinoj web-stranici teži idućoj vrijednosti:

Web-stranica	**1**	**2**	**3**	**4**	**5**
Vrijednost kojoj teže frekvencije	0	$\frac{4}{11}$	$\frac{2}{11}$	$\frac{3}{11}$	$\frac{2}{11}$
Približna decimalna vrijednost	0.00	0.36	0.18	0.27	0.18

Tablica 3. Granične vrijednosti kojima teže frekvencije posjeta pojedinoj web-stranici

Ako usporedimo brojeve u tablicama 1 i 3, vidjet ćemo da su *približno* **iste**.

O Markovljevim lancima i njihovom graničnom ponašanju, te kako doći do jednadžbi koje nam daju ove vrijednosti, čitatelj može više naći u (Vondraček, 2013).

U stvarnom svijetu

Kao što je svima poznato, internet je velika mreža s milijardama web-stranica i još više veza između njih. Neke stranice imaju puno poveznica koje upućuju na njih, druge imaju puno manje. Zato iz praktičnih razloga ovaj način računanja PageRanka nije moguć. Rješava se sustav **velikog broja jednadžbi s puno nepoznanica** kako bi se izračunao PageRank. Detalje čitatelj može doznati u člancima (Horvat & Munđar, Rangiranje web stranica, 2017.) i (Horvat & Munđar, Rangiranje ekipa i prognoziranje ishoda u rukometu korištenjem PageRank algoritma, 2016.).

Ovo je najjednostavnija implementacija, no u svijetu internetskih pretraga stvari se mogu ozbiljno zakomplicirati. Izazov je prikupiti podatke o web-stranicama i vezama među njima. Uočite da smo u ovom članku govorili samo o rangiranju web stranica, ali ne i o *pretragama* njihova sadržaja. Znatiželjni čitatelj, s dubljim znanjem vjerojatnosnih pojmova i linearne algebre, može više naći u (Langville & Meyer, 2012.).

Patent

PageRank je patentiran, tj. zaštićena su autorska prava na njemu i ograničeno je njegovo iskorištavanje. Kako su Sergey Brin i Larry Page u trenutku otkrića bili stu-

denti na sveučilištu Stanford, **vlasnik patenta je sveučilište**, dok Google ima ekskluzivna prava na njegovo korištenje. Sveučilište je za uzvrat dobilo 1.8 milijuna dionica Googlea. Stanford je prodao te dionice 2005. godine za 336 milijuna dolara.

Zaključak

U ovom članku imali smo priliku vidjeti pomalo neuobičajen način obrade podataka. Podatci koje smo obradili su **web-stranice** i **veze** među njima. Postupak koji smo primijenili dao nam je određene vrijednosti na temelju kojih smo izvukli zaključak o ***odnosima*** između web-stranica – njihovo **rangiranje**. Pritom kao pozadina ovog načina rangiranja stoji ideja **koliko** će se **vremena provesti na pojedinoj web-stranici** ako ih počnemo obilaziti slučajno birajući poveznice. Sve vrijednosti mogu se izračunati rješavajući velike sustave linearnih jednadžbi.

Dodatak. Python implementacija

Ovdje ćemo kratko prikazati implementaciju slučajnog obilaska web stranica kao na slici 1. Implementaciju grafa sa slike 2 ostavljamo čitatelju za vježbu. Čitatelji kôd mogu preuzeti s web lokacije https://web.math.pmf.unizg.hr/~tvrtko/metodikaStatistike/clanak4

```
#unos paketa
import numpy as np #numericko racunanje
import random as rm #simualcije
import pandas #obrada podataka

#popis stranica
stranice = [1, 2, 3, 4, 5]

# vjerojatnosti da s pojedine web stranice
# predjemo na drugu
prijelazneVjerojatnosti = [
   [1/3, 1/3, 1/3, 0, 0], # web stranica 1
   [0, 1/2, 0, 1/2, 0], # web stranica 2
   [0, 1/2, 1/2, 0, 0], # web stranica 3
   [0, 0, 1/3, 1/3, 1/3], # web stranica 4
   [0, 1/2, 0, 0, 1/2] # web stranica 5
]

# implementacija obilaska
def slucajniObilazakWebStranica(pocetnaWebStarnica, brojVremen-
skihJedinica):

   listaPosjecenihStranica = [pocetnaWebStarnica]
   trenutnaWebStranica = pocetnaWebStarnica

for korak in range(0, brojVremenskihJedinica - 1):
```

```
     trenutnaWebStranica = np.random.choice(stranice,
p=prijelazneVjerojatnosti[trenutnaWebStranica - 1])
     listaPosjecenihStranica.append(trenutnaWebStranica);

  print("pocetna web stranica: " + str(pocetnaWebStarnica));
  print("broj vremenskih jedinica: " + str(brojVremenskihJedi-
  nica));
  print("zadnja web stranica posjecena: " + str(trenutnaWebStra-
  nica));
return listaPosjecenihStranica;

# slucajni odabir pocetne web-stranice
polaznaWebStranica = rm.choice(stranice);
# vrijeme obilaska
brojVremenskihJedinica = 10000;

posjeceneWebStranice =
slucajniObilazakWebStranica(polaznaWebStranica,     brojVremen-
skihJedinica)
relativnaFrekvencijaPosjecenihWebStranica = pandas.Series.value_
counts(posjeceneWebStranice) / brojVremenskihJedinica
relativnaFrekvencijaPosjecenihWebStranica

Ispis:

pocetna web stranica: 3
broj vremenskih jedinica: 10000
zadnja web stranica posjecena: 2

2    0.3675
4    0.2767
3    0.1807
5    0.1751
dtype: float64
```

Literatura:

1. Horvat, D., & Munđar, D. (2016.). Rangiranje ekipa i prognoziranje ishoda u rukometu korištenjem PageRank algoritma. *Poučak, 17*, str. 35-42. Preuzeto 19. 8. 2018. iz https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=266675
2. Horvat, D., & Munđar, D. (2017.). Rangiranje web stranica. *Osječki matematički list, 17*, str. 51-62.
3. Langville, A. N., & Meyer, C. D. (2012.). *Google's Page Rank and Beyond: The Science of Search Engine Rankings.* Princeton: Priceton University Press.
4. Vondraček, Z. (2013.). *Markovljevi lanci.* PMF-MO, Zagreb: online skripta. Dohvaćeno iz https://web.math.pmf.unizg.hr/~vondra/ml12-predavanja.html