O JEDNOJ PRIMENI KARAKTERISTIČNIH KORENA U SORTIRANJU PODATAKA

Ljubo Nedović

6. mart 2019

U ovoj sekciji je ilustrovana jedna primena karakterističnih vektora u rangiranju rezultata pretraživanja internet stranica, vidi [2].

Google je na scenu nastupio u kasnim devedesetim godinama. Od drugih internet pretraživača ga je u startu razlikovala jedna izuzetna stvar. Ostali pretraživači su, na korisnički upit, izlistavali internet stranice i sadržaje bez nekog naročitog reda i prioriteta, te su se željeni sadržaji morali tražiti na spisku sadržaja koji je sadržao mnoge linkove koje korisnika nisu interesovali. Za razliku od toga, Google je u velikoj meri na vrh spiska uglavnom stavljao sadržaje koje je korisnik zaista i tražio.

Internet pretraživači svoj zadatak obavljaju u sledeća tri koraka.

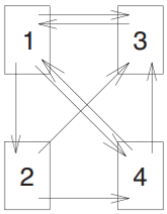
k1: Pretraživanje internet prostora (WEB) i pronalaženje svih stranica i sadržaja sa javnim pristupom.

k2: Indeksiranje pronađenih podataka (sadržaja) na takav način da se efikasno mogu pretraživati po ključnim rečima i frazama.

k3: Ocena i rangiranje značaja svakog pronađenog sadržaja na takav način da korisnik rezultata zadanog upita sa što većom verovatnoćom na vrhu spiska nađe saržaje koji ga zanimaju.

Postavlja se pitanje na koji način internet pretraživači rangiraju sadržaje na spisku sadržaja koji manje ili više odgovaraju želji korisnika. Za ovo rangiranje postoji više metoda. Google za rangiranje sadržaja koristi karakteristične korene, na način koji će biti ilustrovan u sledećem primeru. Ideja postupka je da se za rangiranje koristi „popularnost" pronađenih sadržaja, a ta „popularnost" se, motodom koja će biti prikazana, meri na osnovu povezanosti raznih sadržaja na internetu.

Sledi primer, gde su, radi ilustracije metode, posmatrana četiri internet sadržaja. Označimo ih sa 1, 2, 3, 4. Posmatrani sadržaji imaju i linkove jedni na druge, na način prikazan na sledećoj slici:



Mera važnosti svakog od sadržaja će biti nenegativan realan broj, koji se izračunava na osnovu broja linkova na posmatrani sadržaj, ali i mere važnosti svakog od linkova. Naime, ne vredi jednako svaki link na sadržaj koji rangiramo, već je od velikog značaja da li neki link dolazi od manje ili više „popularnog" sadržaja. Linkovi na neku posmatranu stranicu se nazivaju *povrati* ili *dolazni linkovi*, a linkovi sa posmatrane na druge stranice se nazivaju *odlazni linkovi*. Ovim linkovim sadržaji „glasaju" jedni za druge, pri čemu veći značaj treba da imaju „glasovi" onih sadržaja koji su „popularniji" od drugih. Na prikazanoj slici, najviše dolaznih linkova ima sadržaj 3, ali imajući u vidu sve opisane faktore, to još ne znači da će imati najveću kukupnu meru „popularnosti" nakon rangiranja, jer treba uračunati sve linkove i njihovu važnost. Sadržaji 1, 2, 3 i 4 (prizani na slici) imaju redom 2, 1, 3 i 2 dolazna linka, te bi imajući samo to u vidu bili po važnosti sortirani na sledeći način: 3, 4, 1, 2. Međutim, u meru „popularnosti" nekog sadržaja treba uračunati i sa kojim drugim sadržajima i na koji način je posmatrani sadržaj povezan. Nije svejedno da li na posmatrani sadržaj pokazuje neka važna ili neka nevažna stranica. Tako na prikazanoj slici vidimo da je sadržaj 1 važniji od sadržaja 4 jer na stranicu 1 pokazuje stranica 3 koja je važnija od stranice 2 koja pokazuje na stranicu 4.

Pri svemu tome, u algoritmu za izračunavanje mere važnosti neke stranice treba obezbediti mehanizam koji sprečava da neka stranica utiče na svoju „popularnost" tako što će se sama povezati sa velikim brojem stranica.

Pretpostavimo da u bazi svih internet sadržaja (stranica) imamo N sadržaja označenih sa x_i, $i \in \{1,\ldots,N\}$. Neka je X_i, $i \in \{1,\ldots,N\}$ nenegativan broj koji označava meru „važnosti" stranice x_i, i neka je n_i broj odlazećih linkova stranice x_i, $i \in \{1,\ldots,N\}$. Neka je $L_i \subseteq \{\mathsf{x}_1,\ldots,\mathsf{x}_N\}$ skup svih onih internet stranica koje imaju link ka stranici x_i. Imajući u vidu sve navedeno, formula za rangiranje „važnosti" stranice x_i glasi

$$\mathsf{X}_i = \sum_{\mathsf{x}_k \in L_i} \frac{1}{\mathsf{n}_k} \mathsf{X}_k.$$

Primenom navedene formule na primer prikazan slikom dobijamo sistem linearnih jednačina

$$\mathcal{S}: \quad \begin{array}{lcccccc} \mathsf{X}_1 & = & & & & \frac{1}{1}\mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 \\ \mathsf{X}_2 & = & \frac{1}{3}\mathsf{X}_1 & & & & & \\ \mathsf{X}_3 & = & \frac{1}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & & + & \frac{1}{2}\mathsf{X}_4 \\ \mathsf{X}_4 & = & \frac{1}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & & & \end{array}$$

Ovaj sistem jednačina možemo zapisati u matričnom obliku na sledeći način:

$$A\mathsf{X} = \mathsf{X},$$

gde je

$$A = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix},$$

a $\mathsf{X} = [\,\mathsf{X}_1\ \mathsf{X}_2\ \mathsf{X}_3\ \mathsf{X}_4\,]^T$ je vektor-kolona promenljivih iz sistema jednačina.

Problem rešavanja posmatranog sistema linearnih jednačina $\mathcal{S}$ se stoga sveo na određivanje karakterističnih vektora matrice A koji odgovaraju karakterističnom korenu $\lambda = 1$, jer jednakost $A\mathsf{X} = \mathsf{X}$ možemo zapisati kao $A\mathsf{X} = 1 \cdot \mathsf{X}$. Rešavajući sistem jednačina dobijamo

$$\begin{array}{cl}
 & \left.\begin{array}{rcrcrcrcl} -\mathsf{X}_1 & & & & \mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & - & \mathsf{X}_2 & & & & & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & - & \mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & & & - & \mathsf{X}_4 & = & 0 \end{array}\right. \\
\mathcal{S} \quad \Leftrightarrow & \\
 & \left.\begin{array}{rcrcrcrcl} -\mathsf{X}_1 & & & & \mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & - & \mathsf{X}_2 & & & & & = & 0 \\ \frac{4}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & - & 2\mathsf{X}_3 & & & = & 0 \\ -\frac{5}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & + & 2\mathsf{X}_3 & & & = & 0 \end{array}\right. \\
\overset{[1]}{\Leftrightarrow} & \\
 & \left.\begin{array}{rcrcrcrcl} -\mathsf{X}_1 & & & & \mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 & = & 0 \\ \frac{4}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & - & 2\mathsf{X}_3 & & & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & - & \mathsf{X}_2 & & & & & = & 0 \\ -\frac{1}{3}\mathsf{X}_1 & + & \mathsf{X}_2 & & & & & = & 0 \end{array}\right. \\
\overset{[2]}{\Leftrightarrow} & \\
 & \left.\begin{array}{rcrcrcrcl} -\mathsf{X}_1 & & & & \mathsf{X}_3 & + & \frac{1}{2}\mathsf{X}_4 & = & 0 \\ \frac{4}{3}\mathsf{X}_1 & + & \frac{1}{2}\mathsf{X}_2 & - & 2\mathsf{X}_3 & & & = & 0 \\ \frac{1}{3}\mathsf{X}_1 & - & \mathsf{X}_2 & & & & & = & 0 \end{array}\right. \\
\overset{[3]}{\Leftrightarrow} &
\end{array}$$

gde smo primenili sledeće transformacije sistema:

[1]: prva jednačina se oduzima od treće, i prva jednačina pomnožena sa 2 se dodaje na četvrtu;

[2]: treća jednačina se doda na četvrtu, a zatim se zamene druga i treća jednačina;

[3]: treća jednačina se doda na četvrtu.

Metodom zamene unatrag, za skup rešenja sistema, odnosno traženi skup karakterističnih vektora dobijamo

$$\mathsf{X}_1 = \alpha \in \mathbb{R},$$
$$\mathsf{X}_2 = \tfrac{1}{3}\alpha,$$
$$\mathsf{X}_3 = \tfrac{4}{6}\mathsf{X}_1 + \tfrac{1}{4}\mathsf{X}_2 = \tfrac{4}{6}\alpha + \tfrac{1}{4}\cdot\tfrac{1}{3}\alpha = \tfrac{3}{4}\alpha,$$
$$\mathsf{X}_4 = 2\mathsf{X}_1 - 2\mathsf{X}_3 = 2\alpha - 2\cdot\tfrac{3}{4}\alpha = \tfrac{1}{2}\alpha,$$

odnosno

$$V = \left\{\alpha\cdot\left(1,\frac{1}{3},\frac{3}{4},\frac{1}{2}\right) \mid \alpha\in\mathbb{R}\right\} = \{\alpha\cdot(12,4,9,6) \mid \alpha\in\mathbb{R}\}.$$

Ako svaku komponentu vektora $(12,4,9,6)$ podelimo sa zbirom $12+4+9+6=31$, dobićemo procentualno izražene komponente karakterističnih vektora:

$$\overline{\mathsf{X}}_1 = \tfrac{12}{31} \approx 0.387,$$
$$\overline{\mathsf{X}}_2 = \tfrac{4}{31} \approx 0.129,$$
$$\overline{\mathsf{X}}_3 = \tfrac{9}{31} \approx 0.290,$$
$$\overline{\mathsf{X}}_4 = \tfrac{6}{31} \approx 0.194.$$

Ove poslednje vrednosti interpretiramo kao procentualno izraženu popularnost posmatranih internet sadržaja. Dakle, Google svojim algoritmom sadržaje 1, 2, 3, 4 iz posmatranog primera sa slike rangira i sortira u redosledu: 1, 3, 4, 2. Možemo uočiti veliku razliku u odnosu na sortiranje 3, 4, 1, 2 koje je izvršeno samo po kriterijumu dolaznih linkova na posmatrane sadržaje. Ključna razlika u posmatranom primeru je ta da stranica 3, koja ima 3 dolazna linka, ima samo jedan izlazni link, na stranicu 1.

Krajnju našu procenu o tome koji algoritam je bolji, ipak donosimo na osnovu našeg ličnog iskustva i subjektivnog osećaja o tome koji algoritam „bolje pogađa naše želje". Postoje i drugi matematički modeli koji u manjoj ili većoj meri odgovaraju našim intuitivnim procenama o dobrim algoritmima za sortiranjem podataka po stepenu njihovih važnosti po određenim kriterijumima. Google-ov algoritam, zasnovan na teoriji karakterističnih korena matrice je stotinama miliona svojih zadovoljnih korisnika ubedljivo potvrdio svoju vrednost.

Razmatrani primer je specifičan po tome što je skup V karakterističnih vektora koji odgovaraju karakterističnom korenu 1 matrice A jednodimenzionalan, odnosno postoji tačno jedan odgovarajući karakteristični vektor sa zbirom komponenti jednakim jedinici. U opštem slučaju, može biti i više takvih karakterističnih korena, tj. prostor odgovarajućih karakterističnih korena može biti i višedimenzionalan. U takvom slučaju, algoritam za sortiranje treba još da donese odluku na koji način će da formira krajnje rangiranje prikazanih sadržaja.

Literatura

[1] Rade Doroslovački, *Principi algebre opšte, diskretne i linearne*, FTN Izdavaštvo, Novi Sad, 2015.

[2] Kurt Bryan and Tanya Leise, *The $25,000,000,000 Eigenvector: The Linear Algebra behind Google*, SIAM Rev. Volume 48, Issue 3, pp. 569–581, https://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf