

Mudrost mnoštva i primjene

Goran Kovačević

Sažetak

Mudrost mnoštva je fenomen koji se temelji na opažanju da je prosjek procjena veće grupe ljudi često vrlo točna procjena, ponekad čak i točnija od procjene nekog stručnjaka.

Objašnjenje je jednostavno: ljudi su različiti pa griješe na različite načine – neki precjenjuju, a neki podcjenjuju, pa se zajednički prosjek većih skupina ljudi obično stabilizira oko stvarne vrijednosti.

U radu se prikazuju dva jednostavna primjera: procjena mase vola i određivanje položaja nestale podmornice. Pokazuje se i kako se fenomen može matematički modelirati – primjenom teorema srednje vrijednosti za određeni integral. Zatim se razmatraju složenije mogućnosti primjene, u ocjenjivanju učenika i pri rangiranju internetskih stranica.

Ključni pojmovi: mudrost mnoštva, kolektivna procjena, prosjek, teorem srednje vrijednosti, ocjenjivanje učenika, rangiranje internetskih stranica, metoda najmanjih kvadrata

1 Procjena mase vola – Francis Galton (1906.)

Na seoskom sajmu u Plymouthu pogađala se masa vola nakon klanja i obrade. Sudjelovalo je 787 osoba, a bilo je 774 valjane procjene mase vola. Stvarna masa vola bila je 1198 funti, odnosno 543,404 kilograma.

Aritmetička sredina svih procjena tj. prosječna procjena mnoštva iznosila je 1197 funti, odnosno 542,95 kilograma. Dakle, prosječna procjena mase vola skupine od 774 osobe se od stvarne mase vola razlikovala samo za oko pola kilograma.

Galtonov eksperiment nije samo zanimljiva anegdota. U sljedećem odjeljku ćemo pomoću teorema o integralnoj srednjoj vrijednosti objasniti pojavu „mudrosti mnoštva“ u ovom eksperimentu.

2 Veza s teoremom o integralnoj srednjoj vrijednosti

Teorem o integralnoj srednjoj vrijednosti kaže da za realnu funkciju f koja je neprekidna na segmentu $[a, b] \subset \mathbb{R}$ postoji $c \in [a, b]$ takav da je

$$\frac{1}{b-a} \int_a^b f(x) dx = f(c).$$

Primjenom tog važnog teorema mogu se konstruirati slučajni uzorci za koje vrijedi mudrost mnoštva.

Pretpostavimo da se sve valjane procjene neke veličine čija je točna vrijednost m kreću u rasponu od p do P . Granične procjene p i P ne ovise o broju procjena i poznate su svim iskusnim procjeniteljima. Stoga možemo smatrati da svaka procjena pripada intervalu $[p, P]$ i da je $m \in \langle p, P \rangle$.

Uzmimo da su p i $P > p$ poznati brojevi i m bilo koji broj u intervalu $\langle p, P \rangle$.

Uvidom u geometrijsko značenje teorema o integralnoj srednjoj vrijednosti može se zaključiti da za proizvoljne realne brojeve a i $b > a$ postoji neprekidna strogo rastuća surjeksija $f: [a, b] \rightarrow [p, P]$ takva da je

$$\frac{1}{b-a} \int_a^b f(x) dx = m.$$

Budući da je $m \in \langle p, P \rangle$, zbog neprekidnosti funkcije f (odnosno po teoremu o međuvrijednostima), slijedi da postoji $c \in \langle a, b \rangle$ takav da je $f(c) = m$.

Definicija 1. Kažemo da je z **slučajan broj** u intervalu $[a, b]$ ako je vjerojatnost da z upadne u neki podinterval intervala $[a, b]$ proporcionalna duljini tog podintervala. Ako su vrijednosti neke varijable x isključivo slučajni brojevi iz $[a, b]$, onda kažemo da x ima **jednoliku ili uniformnu razdiobu** na $[a, b]$.

Za svaki $n \in \mathbb{N}$ formirat ćemo niz uzoraka (U_n) tako da U_1 sadrži samo jedan slučajni broj u intervalu $[a, b]$, a za svaki $n \in \mathbb{N}$ uzorak U_{n+1} dobijemo tako da uzorku U_n pridružimo još jedan slučajni broj u intervalu $[a, b]$. U dovoljno velikom uzorku se isti slučajni brojevi mogu pojaviti i više puta.

Za svaki $n \in \mathbb{N}$ uzorak U_n sadrži $n_1^{(n)}$ primjeraka broja $x_1^{(n)}$ kojemu se svaki put pridružuje (ista) procjena $y_1^{(n)} = f(x_1^{(n)})$, $n_2^{(n)}$ primjeraka broja $x_2^{(n)}$ kojemu se svaki put pridružuje (ista) procjena $y_2^{(n)} = f(x_2^{(n)})$, ..., sve do $n_{k^{(n)}}^{(n)}$ primjeraka broja $x_{k^{(n)}}^{(n)}$ kojemu se svaki put pridružuje (ista) procjena $y_{k^{(n)}}^{(n)} = f(x_{k^{(n)}}^{(n)})$, pri čemu je $n_1^{(n)} + n_2^{(n)} + \dots + n_{k^{(n)}}^{(n)} = n$ (za neki $k^{(n)} \in \{1, \dots, n\}$, gdje $k^{(n)}$

označava broj grupa takvih da istoj grupi pripadaju svi međusobno jednaki izabrani slučajni brojevi). Dakle, $x_i^{(n)} \in [a, b]$ zapravo predstavlja svaku osobu koja daje procjenu $y_i^{(n)} = f(x_i^{(n)})$.

Bez gubitka općenitosti, možemo pretpostaviti da je u svakom uzorku (tj. za svaki $n \in \mathbb{N}$):

$$x_1^{(n)} < x_2^{(n)} < \dots < x_{k^{(n)}}^{(n)},$$

odnosno

$$y_1^{(n)} < y_2^{(n)} < \dots < y_{k^{(n)}}^{(n)}.$$

Npr. uzorak od sljedećih 7 slučajno izabranih brojeva u intervalu $[0,1]$ (nakon uzlaznog sortiranja):

$$0,3, \quad 0,4, \quad 0,4, \quad 0,7, \quad 0,7, \quad 0,7, \quad 1$$

znači da smo dobili jednu procjenu $f(0,3)$, dvije procjene $f(0,4)$, tri procjene $f(0,7)$ i jednu procjenu $f(1)$. Tu je $k^{(7)} = 4$ i

$$n_1^{(7)} = 1, \quad n_2^{(7)} = 2, \quad n_3^{(7)} = 3, \quad n_4^{(7)} = 1,$$

$$x_1^{(7)} = 0,3, \quad x_2^{(7)} = 0,4, \quad x_3^{(7)} = 0,7, \quad x_4^{(7)} = 1,$$

$$y_1^{(7)} = f(0,3), \quad y_2^{(7)} = f(0,4), \quad y_3^{(7)} = f(0,7), \quad y_4^{(7)} = f(1).$$

Ako se u sljedećem slučajnom biranju odabere broj 0,5, nakon uzlaznog sortiranja dobije se uzorak:

$$0,3, \quad 0,4, \quad 0,4, \quad 0,5, \quad 0,7, \quad 0,7, \quad 0,7, \quad 1$$

u kojemu je $k^{(8)} = 5$ i

$$n_1^{(8)} = 1, \quad n_2^{(8)} = 2, \quad n_3^{(8)} = 1, \quad n_4^{(8)} = 3, \quad n_5^{(8)} = 1,$$

$$x_1^{(8)} = 0,3, \quad x_2^{(8)} = 0,4, \quad x_3^{(8)} = 0,5, \quad x_4^{(8)} = 0,7, \quad x_5^{(8)} = 1,$$

$$y_1^{(8)} = f(0,3), \quad y_2^{(8)} = f(0,4), \quad y_3^{(8)} = f(0,5),$$

$$y_4^{(8)} = f(0,7), \quad y_5^{(8)} = f(1).$$

Označimo s s_n prosjek n procjena dobivenih iz uzorka U_n . Tada je

$$s_n = \frac{1}{n} \sum_{i=1}^{k^{(n)}} n_i^{(n)} y_i^{(n)}.$$

Kako varijabla x funkcije f ima jednoliku razdiobu na $[a, b]$, jasno je da se niti jedna od mogućih procjena ne preferira u odnosu na ostale. Pretpostavimo da n raste, odnosno da se veličina uzorka izabranih slučajnih brojeva u intervalu $[a, b]$ povećava. Onda i $k^{(n)}$, koji predstavlja broj različitih procjena, (uglavnom strogo) raste. Događaj

$$n_i^{(n)} \approx \frac{n}{k^{(n)}}, \quad \forall i \in \{1, \dots, k^{(n)}\}$$

tada postaje sve vjerojatniji jer s porastom broja n raste i vjerojatnost da se sve različite procjene u prosjeku pojavljuju približno jednak broj puta.

Stoga i događaj

$$\frac{1}{n} \sum_{i=1}^{k^{(n)}} n_i^{(n)} y_i^{(n)} \approx \frac{1}{k^{(n)}} \sum_{i=1}^{k^{(n)}} y_i^{(n)},$$

odnosno događaj

$$s_n \approx \frac{1}{k^{(n)}} \sum_{i=1}^{k^{(n)}} y_i^{(n)}$$

s porastom broja n postaje sve sigurniji.

Također za dovoljno veliki n vrijedi: $x_1^{(n)} \approx a$, $x_{k^{(n)}}^{(n)} \approx b$ i $d_n \approx 0$,

gdje je $d_n = \max_{1 \leq i \leq k^{(n)}-1} (x_{i+1}^{(n)} - x_i^{(n)})$.

Prema tome, za dovoljno veliki n integral $\int_a^b f(x) dx$ je približno jednak donjoj Riemannovoj sumi funkcije f s čvorovima u točkama $x_1^{(n)}, \dots, x_{k^{(n)}}^{(n)}$ pa je tada

$$\frac{1}{b-a} \int_a^b f(x) dx \approx \frac{1}{b-a} \sum_{i=1}^{k^{(n)}-1} y_i^{(n)} (x_{i+1}^{(n)} - x_i^{(n)}).$$

S porastom broja n raste vjerojatnost da su čvorovi $x_1^{(n)}, \dots, x_{k^{(n)}}^{(n)}$ ravnomjerno raspoređeni unutar intervala $[a, b]$ pa se tako povećava i vjerojatnost događaja

$$x_{i+1}^{(n)} - x_i^{(n)} \approx \frac{b-a}{k^{(n)}-1}, \quad \forall i \in \{1, \dots, k^{(n)}-1\},$$

odnosno

$$\frac{1}{b-a} \int_a^b f(x) dx \approx \frac{1}{k^{(n)}-1} \sum_{i=1}^{k^{(n)}-1} y_i^{(n)} \approx \frac{1}{k^{(n)}} \sum_{i=1}^{k^{(n)}} y_i^{(n)}$$

Prema tome, za dovoljno veliki n je

$$s_n \approx m.$$

Kako je $\lim_{n \rightarrow \infty} d_n = 0$, možemo zaključiti da su procjene sve raznolikije kako n raste.

Dakle, za svaki dovoljno veliki uzorak slučajnih procjena (tj. uzorak slučajnih vrijednosti funkcije f) u intervalu $[p, P]$ je prosjek tih procjena približno jednak m .

Ako za neki dovoljno veliki $n \in \mathbb{N}$ prosjek procjena s_n ispadne jednak ili približno jednak m , onda se dogodila mudrost mnoštva.

Tako smo zapravo pokazali da za svaki član s_n niza (s_n) s dovoljno velikim indeksom n vrijedi mudrost mnoštva.

Prethodna analiza pokazuje da se mudrost mnoštva najčešće događa kada su ispunjeni sljedeći uvjeti:

- ima mnogo valjanih procjena
- sve valjane procjene su dosta raznolike/raznovrsne.

U ovom kontekstu funkcija f predstavlja raspodjelu procjena (npr. raspodjelu procjena mase vola u Galtonovom eksperimentu), a m stvarnu vrijednost (npr. točnu masu vola). Drugim riječima, rezultat Galtonovog eksperimenta nije slučajnost – matematički model pokazuje da je pri dovoljno velikom broju raznolikih procjena njihov prosjek uvijek blizu stvarne vrijednosti.

Primjer 1. *Odredimo neprekidnu strogo rastuću surjekciju*

$$f: [a, b] \rightarrow [p, P]$$

oblika

$$f(x) = \begin{cases} f_1(x) & \text{ako je } x \in [a, c] \\ f_2(x) & \text{ako je } x \in [c, b] \end{cases}$$

gdje su

$$f_1(x) = k_1x + l_1, \quad f_2(x) = k_2x + l_2,$$

$$i \ c \in \langle a, b \rangle \text{ takav da je } f(c) = m = \frac{1}{b-a} \int_a^b f(x) dx.$$

Treba odrediti konstante k_1, l_1, k_2, l_2 i c rješavanjem sustava:

$$\left\{ \begin{array}{l} f(a) = p \\ f(b) = P \\ f_1(c) = m \\ f_2(c) = m \\ \int_a^b f(x) dx = m(b-a) \end{array} \right. .$$

Iz $f(a) = p$ slijedi

$$k_1a + l_1 = p,$$

odnosno

$$l_1 = p - k_1a.$$

Iz $f(b) = P$ slijedi

$$k_2b + l_2 = P,$$

odnosno

$$l_2 = P - k_2b.$$

Onda iz jednadžbe $f_1(c) = m$ proizlazi

$$k_1c + p - k_1a = m,$$

odakle je

$$k_1 = \frac{m - p}{c - a}.$$

Slično, jednačba $f_2(c) = m$ daje

$$k_2c + P - k_2b = m,$$

tj.

$$k_2 = \frac{P - m}{b - c}.$$

Stoga je

$$l_1 = p - k_1a = p - \frac{m - p}{c - a}a = \frac{pc - ma}{c - a}$$

i

$$l_2 = P - k_2b = P - \frac{P - m}{b - c}b = \frac{mb - Pc}{b - c}.$$

Dakle,

$$f_1(x) = \frac{m - p}{c - a}x + \frac{pc - ma}{c - a}, \quad f_2(x) = \frac{P - m}{b - c}x + \frac{mb - Pc}{b - c},$$

a c određujemo iz zadnje jednačbe sustava:

$$\int_a^b f(x) = m(b - a),$$

odnosno

$$\int_a^c f_1(x)dx + \int_c^b f_2(x)dx = m(b - a).$$

Nakon integriranja dobivamo:

$$\frac{(m - p)(c + a)}{2} + pc - ma + \frac{(P - m)(b + c)}{2} + mb - Pc = m(b - a)$$

$$(m - p)(c + a) + (P - m)(b + c) + 2pc - 2Pc = 0$$

$$ma + pc - pa + Pb - Pc - mb = 0$$

$$(P - p)c = Pb - pa - m(b - a)$$

$$c = \frac{Pb - pa - m(b - a)}{P - p}.$$

Našli smo c pa se dalje lako određuju i sve ostale nepoznate konstante.

Budući da je $m \in \langle p, P \rangle$, iz formule za c se lako vidi da je $c \in \langle a, b \rangle$, pa je konstrukcija konzistentna (f je neprekidna strogo rastuća funkcija i vrijedi $f(c) = m$).

Primjenom ovakve (po dijelovima linearne) funkcije f se u GeoGebri lako generiraju slučajni uzorci brojeva u intervalu $[p, P]$ čija je prosječna vrijednost približno jednaka m .

Primjena matematičkih ideja pokazuje da mudrost mnoštva nije samo zanimljiv pokus sa sajma, već opći princip što se može vidjeti i u nešto drukčijem, ali životno važnom primjeru – potrazi za nestalom podmornicom.

3 Nestanak podmornice Scorpion

Američka nuklearna podmornica USS Scorpion je 1968. godine nestala negdje u Atlantskom oceanu. Klasične metode potrage nisu uspjele, a mornarica je saznala samo grubu procjenu položaja podmornice. Tada je kapetan John Craven odlučio promijeniti pristup. Umjesto da pita jednog „najpametnijeg“ stručnjaka, okupio je tim eksperata različitih profila: fizičare, inženjere, vojne analitičare, meteorologe, ... Nije poznato koliko je članova imao taj tim pa ćemo pretpostaviti da je brojio n članova, za neki dovoljno veliki broj n . Craven je svim članovima tima dao do tada poznate podatke o nestanku podmornice i zamolio svakog od njih ponaosob procjenu o njenom trenutnom položaju. Svaka osoba je dala svoju „pretpostavljenu lokaciju“ kao točku u prostoru.

Točku (x, y, z) možemo identificirati s trodimenzionalnim vektorom \vec{r} kojemu je početak u točki $(0,0,0)$, a kraj u točki (x, y, z) , tj. možemo pisati

$$\vec{r} = (x, y, z).$$

Za pretpostavljene lokacije:

$$\vec{r}_1 = (x_1, y_1, z_1), \quad \vec{r}_2 = (x_2, y_2, z_2), \dots, \quad \vec{r}_n = (x_n, y_n, z_n)$$

može se izračunati srednja vrijednost svih vektora (aritmetički prosjek):

$$\vec{r}_{\text{avg}} = \frac{1}{n} \sum_{i=1}^n \vec{r}_i = \left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i, \frac{1}{n} \sum_{i=1}^n z_i \right) = (\bar{x}, \bar{y}, \bar{z}).$$

To je točka za koju je mnoštvo (od n ljudi) – zajednički gledano – u prosjeku „vjerovalo“ da je lokacija podmornice. Iako nitko nije pogodio točnu lokaciju, prosjek njihovih procjena $(\bar{x}, \bar{y}, \bar{z})$ je bio udaljen samo 200 metara od mjesta na kojem je podmornica kasnije pronađena. Zbrajanjem pogrešaka u procjeni lokacije svih stručnjaka se umanjuje utjecaj greške svakog pojedinca. Iz tog razloga je prosječna procjena grupe eksperata bila izuzetno točna.

S vojnih operacija prelazimo u školske klupe. Naime, u sljedećem odjeljku vidjet ćemo kako se mudrost mnoštva može iskoristiti u obrazovanju.

4 Mudrost mnoštva u ocjenjivanju

Ocjenjivanje znanja učenika često se doživljava kao subjektivan proces. Mudrost mnoštva može osigurati pravedniji i objektivniji pristup.

Pretpostavimo da tražimo formulu za računanje konačne ocjene iz matematike svakog učenika jednog razreda s 30 učenika u sljedećem obliku:

$$Y = \alpha \cdot t + \beta \cdot z + \gamma \cdot a,$$

gdje je:

Y – ukupna ocjena promatranog učenika izražena u bodovima od 0 do 100,

t – prosječan broj bodova tog učenika na testovima, ako na svakom testu učenik može sakupiti od 0 do 100 bodova,

z – postotak riješenih zadataka

a – aktivnost, tj. postotak točnih odgovora na postavljena pitanja profesora na nastavi.

Parametri α , β i γ su nam za sada nepoznati. Nazivaju se težinskim koeficijentima, odnosno kraće težinama. Tražit ćemo da za te brojeve vrijedi:

$$\alpha, \beta, \gamma \geq 0$$

i

$$\alpha + \beta + \gamma = 1.$$

Odredit ćemo ih na sljedeći način. Izabrat ćemo određeni broj učenika razreda, npr. 10.

Zamolit ćemo sve učenike u tom razredu da iskreno procijene znanje svakog od izabranih učenika bodovima od 0 do 100. Za svakog od izabranih 10 učenika ćemo zbrojiti sve dobivene bodove i rezultat podijeliti s 30. Tako ćemo dobiti njihove prosječne procijenjene ukupne ocjene. Neka \hat{Y}_i označava prosječnu procijenjenu ukupnu ocjenu i -tog od izabranih 10 učenika za $i \in \{1, 2, \dots, 10\}$. Dakle, u ovom primjeru pretpostavljamo da svaki učenik daje procjenu znanja svojih kolega, pa konačnu procjenu pojedinog učenika dobivamo kao prosjek svih zaprimljenih ocjena. Takav postupak može se povezati s pojmom mudrosti mnoštva jer skup procjenitelja djeluje kao „kolektivni sud“ koji, u prosjeku, može biti točniji od pojedinačne procjene.

Ipak, ima samo 30 procjenitelja, a procjene vjerojatno nisu u potpunosti neovisne jer se ne mogu sa sigurnošću isključiti pristranosti učenika (tj. međusobna prijateljstva ili animoziteti). Osim toga, „najpravednija“ ocjena svakog učenika nije poznata pa ne možemo izravno provjeriti točnost kolektivne procjene.

Zato ovdje mudrost mnoštva ne treba shvatiti doslovno, nego u širem smislu – kao ideju da se kombiniranjem više različitih mišljenja smanjuje utjecaj pojedinačnih pogrešaka.

Težinske koeficijente α , β i γ možemo odrediti metodom najmanjih kvadrata. Ta metoda omogućuje da se težinski koeficijenti najbolje prilagode dobivenim kolektivnim procjenama. Traže se α , β i γ koji minimiziraju funkciju pogreške:

$$\text{Error}(\alpha, \beta, \gamma) = \sum_{i=1}^{10} [\hat{Y}_i - (\alpha t_i + \beta z_i + \gamma a_i)]^2$$

i zadovoljavaju uvjete

$$\alpha, \beta, \gamma \geq 0, \quad \alpha + \beta + \gamma = 1.$$

To je tzv. problem optimizacije, točnije problem minimuma kvadratnog programiranja. Ovaj problem nije previše kompliciran jer ima samo 3 nepoznanice. Koristeći uvjet $\alpha + \beta + \gamma = 1$, možemo jednu nepoznanicu izraziti pomoću preostalih pa tako dobiti problem s dvije nepoznanice. Ako se još i pojednostavni funkcija pogreške, dobije se sljedeći problem:

$$\text{Error}(\alpha, \beta) = C_1 \alpha^2 + C_2 \beta^2 + C_3 \alpha \beta + C_4 \alpha + C_5 \beta + C_6$$

s uvjetima

$$\alpha, \beta \geq 0, \quad \alpha + \beta \leq 1,$$

i poznatim konstantama C_1, \dots, C_6 . Primijetimo da konstante C_1, C_2 i C_6 ne mogu biti negativne. Pokažimo na jednom jednostavnom primjeru kako se rješava takav problem. U primjeru ćemo funkciju pogreške označiti sa z , a umjesto α i β koristit ćemo uobičajene oznake varijabli: x i y .

Primjer 2. *Odredimo najmanju vrijednost funkcije*

$$z = x^2 + y^2 + xy - x + y$$

u području D određenom nejednadžbama:

$$x, y \geq 0, \quad x + y \leq 1.$$

Zadano područje D je trokut (s vrhovima u točkama $(0,0)$, $(1,0)$ i $(0,1)$) koji je očito omeđen i zatvoren skup. Lako se vidi da je zadana funkcija diferencijabilna na \mathbb{R} , dakle i u svakoj točki područja D jer je $D \subset \mathbb{R}$. Stoga svaku svoju ekstremnu vrijednost poprima ili u nekoj svojoj stacionarnoj točki ili u graničnoj točki tog trokuta.

Tražimo stacionarne točke funkcije z . Određujemo njene parcijalne derivacije prvog reda:

$$z'_x = 2x + y - 1, \quad z'_y = 2y + x + 1$$

pa rješavamo sustav:

$$\begin{cases} z'_x = 0 \\ z'_y = 0 \end{cases}$$

Tako se dobije (jedina) stacionarna točka $M(1, -1)$. Kako ta točka ne pripada zadanom području, točka u kojoj zadana funkcija ima minimum mora pripadati rubu područja, tj. nekoj stranici trokuta. Nađimo najmanje vrijednosti funkcije z u svim stranicama trokuta.

Jedna stranica trokuta je dio pravca $y = 0$. Tada je $z = x^2 - x$ pa tražimo točku minimuma te funkcije (jedne varijable) na dijelu

$0 \leq x \leq 1$. Dobije se $z_{min1} = -\frac{1}{4}$ u točki $T_1\left(\frac{1}{2}, 0\right)$.

Druga stranica trokuta je dio pravca $x = 0$. Tu je $z = y^2 + y$ pa tražimo točku minimuma te funkcije (jedne varijable) na dijelu $0 \leq y \leq 1$. Dobije se $z_{min2} = 0$ u točki $T_2(0,0)$.

Treća stranica trokuta je dio pravca $x + y = 1$, odnosno $y = 1 - x$. Tada je

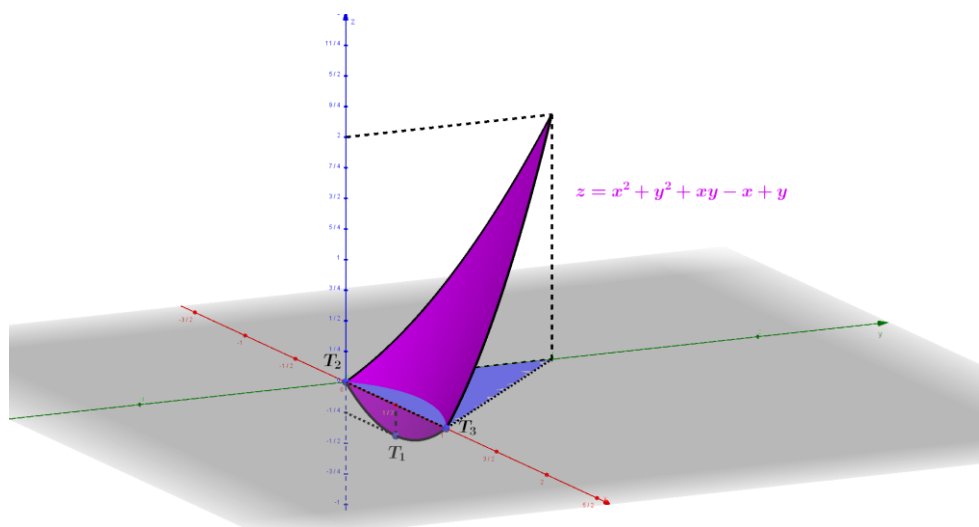
$$z = x^2 + (1 - x)^2 + x(1 - x) - x + 1 - x = x^2 - 3x + 2$$

pa tražimo točku minimuma te funkcije (jedne varijable) na dijelu $0 \leq x \leq 1$. Dobije se $z_{min3} = 0$ u točki $T_3(1,0)$.

Uspoređivanjem dobivenih minimuma zaključujemo da je optimalno rješenje našeg primjera:

$$z_{min} = -\frac{1}{4}$$

u točki $T_1\left(\frac{1}{2}, 0\right)$.



Slika 1. Grafički prikaz zadane plohe

Postoje računalni algoritmi koji relativno brzo rješavaju kompliciranije probleme optimizacije s više od 3 nepoznate težine, ali su poprilično složeni pa se njima u ovom radu nećemo baviti. Pokušajmo radije objasniti kvantitativno značenje težina α , β i γ . Zamislimo da pogađamo te brojeve. Ako damo previše važnosti testu (tj. ako uzmemo velik α), učenici s odličnim testovima koji nisu redovito pisali zadaće ili nisu bili aktivni na satu dobit će previsoku ukupnu ocjenu. Ako damo premalo važnosti zadaćama (mali β), učenici koji redovito rade zadaće neće dobiti priznanje za svoj rad pa neće biti zadovoljni s dobivenom ocjenom. Računalni algoritam traži takve α , β i γ da ocjenjivanje bude najpravednije moguće. Dakle, težine α , β i γ odražavaju stvarnu važnost pojedinog aspekta znanja onako kako ga razred u cjelini percipira.

Fenomen mudrosti mnoštva prisutan je i u digitalnom svijetu. Najpoznatiji primjer je Google, koji rangiranje stranica temelji upravo na kolektivnom ponašanju milijuna korisnika.

5 Rangiranje Google-ovih stranica

Mudrost mnoštva koristi se i u rangiranju Google-ovih stranica.

Naime, rangiranje se temelji na kolektivnom ponašanju milijuna korisnika – na koje linkove klikaju, koliko ostaju na stranicama, koje stranice često posjećuju. To je mudrost mnoštva jer se Google ne oslanja na ponašanje samo jednog surfera (npr. nekog stručnjaka za pojam koji se pretražuje), nego uzima u obzir „glasove“ svih surfera zajedno.

Opisat ćemo pojednostavljeni model rangiranja u kojem se uzima u obzir samo relevantnost, važnost i brzina učitavanja stranica.

Za svaku stranicu se, nakon upisanog pojma pretrage, računa tzv. „final score“ primjenom formule

$$Y = w_1 \cdot r + w_2 \cdot v + w_3 \cdot b,$$

gdje je:

Y – final score promatrane stranice nakon upisanog pojma pretrage,

r – relevantnost, tj. povezanost sadržaja stranice s pojmom pretrage,

v – važnost (PageRank) stranice,

b – prosječna brzina učitavanja stranice.

Što stranica ima veći final score, to će biti popularnija – na boljem mjestu (bliže vrhu) u konačnom poretku stranica. Dakle, matematika nam omogućuje da „izmjerimo“ popularnost i korisnost stranica na objektivan način. Opišimo najprije na koji način se određuju veličine r , v i b za svaku stranicu.

Prosječna brzina učitavanja b se lako može odrediti mjerenjem. Kod određivanja relevantnosti r bitan je broj klikova na promatranu stranicu, broj pojavljivanja traženog pojma na toj stranici, stopa napuštanja stranice, broj ponovnih posjeta, itd.

Ako se traženi pojam ne pojavljuje na promatranoj stranici, onda se uzima da je $r = 0$, a takva se stranica ne pojavljuje među rezultatima pretraživanja. Budući da se i brzina učitavanja b i relevantnost r procjenjuju na temelju velikog broja mjerenja i korisničkih interakcija, na njihove se vrijednosti može gledati kao na rezultat kolektivne procjene (tj. kao na mudrost mnoštva u digitalnom okruženju). Ostaje pokazati kako se određuje važnost stranice. U računanju važnosti koristi se linearna algebra. Naime, vjerojatnosti prijelaza između web-stranica možemo prikazati matricom. Stoga ćemo najprije definirati pojam matrice, a zatim i osnovne računске operacije s matricama.

5.1 Matrica

Definicija 2. Neka su m i n prirodni brojevi. **Matrica tipa $m \times n$** (nad skupom \mathbb{R}) je funkcija

$$A: \{1, \dots, m\} \times \{1, \dots, n\} \rightarrow \mathbb{R}.$$

Vrijednost funkcije $A(i, j)$ za neki $(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\}$ označavamo s a_{ij} . Sve vrijednosti a_{ij} obično se zapisuju u obliku pravokutne tablice s m redaka i n stupaca.

Drugim riječima, matrica tipa $m \times n$ je pravokutna tablica brojeva raspoređenih u m redaka i n stupaca. Svaku matricu označavamo velikim slovom, a njezine elemente istim, ali malim slovom s dva indeksa, pri čemu prvi indeks označava redni broj retka, a drugi redni broj stupca elementa. Tako zapis $A = [a_{ij}]$ označava matricu s elementima a_{ij} . Matrica je **pozitivna** ako su svi njezini elementi pozitivni brojevi.

Npr. $A = \begin{bmatrix} 1 & 4 & 7 \\ 3 & 2 & 5 \end{bmatrix}$ je pozitivna matrica tipa 2×3 čiji su elementi

$$\begin{aligned} a_{11} &= 1, & a_{12} &= 4, & a_{13} &= 7, \\ a_{21} &= 3, & a_{22} &= 2, & a_{23} &= 5. \end{aligned}$$

5.2 Aritmetika matrica

a) Zbrajanje matrica

Zbrajati se mogu samo matrice istog tipa. Ako su A i B matrice tipa $m \times n$, onda je i

$$C = A + B$$

matrica tipa $m \times n$. Element c_{ij} matrice C određujemo po formuli

$$c_{ij} = a_{ij} + b_{ij}.$$

Dakle, zbrajaju se elementi na istim pozicijama.

Primjer 3.

$$A = \begin{bmatrix} 2 & 3 & 7 \\ 1 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 & 4 \\ 2 & -1 & 1 \end{bmatrix};$$

$$C = A + B = \begin{bmatrix} 2+1 & 3+3 & 7+4 \\ 1+2 & 0+(-1) & 1+1 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 11 \\ 3 & -1 & 2 \end{bmatrix}.$$

b) Množenje matrice brojem

Matrica se množi brojem tako da se svaki element matrice pomnoži istim brojem.

Primjer 4.

$$A = \begin{bmatrix} 2 & 3 & 7 \\ 1 & 0 & 1 \end{bmatrix}, \quad \lambda = \frac{1}{2};$$

$$D = \lambda A = \frac{1}{2} \begin{bmatrix} 2 & 3 & 7 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \cdot 2 & \frac{1}{2} \cdot 3 & \frac{1}{2} \cdot 7 \\ \frac{1}{2} \cdot 1 & \frac{1}{2} \cdot 0 & \frac{1}{2} \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{2} & \frac{7}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

Definiramo još

$$-A = (-1) \cdot A.$$

c) Oduzimanje matrica

Oduzimati se mogu samo matrice istog tipa.

$$E = A - B = A + (-B) = A + (-1) \cdot B$$

Primjer 5.

$$\begin{aligned} A &= \begin{bmatrix} 2 & 3 & 7 \\ 1 & 0 & 1 \end{bmatrix}, & B &= \begin{bmatrix} 1 & 3 & 4 \\ 2 & -1 & 1 \end{bmatrix}; \\ E = A - B &= A + (-1) \cdot B = \begin{bmatrix} 2 & 3 & 7 \\ 1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 & -3 & -4 \\ -2 & 1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 2 + (-1) & 3 + (-3) & 7 + (-4) \\ 1 + (-2) & 0 + 1 & 1 + (-1) \end{bmatrix} \\ &= \begin{bmatrix} 2 - 1 & 3 - 3 & 7 - 4 \\ 1 - 2 & 0 - (-1) & 1 - 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 3 \\ -1 & 1 & 0 \end{bmatrix} \end{aligned}$$

d) Množenje matrica

Umnožak $A \cdot B$ matrica A i B postoji samo ako je broj redaka matrice B jednak broju stupaca matrice A . Za takve matrice A i B kažemo da su **ulančane**. Ako je A matrica tipa $m \times n$ i B matrica tipa $n \times k$, onda je $F = A \cdot B$ matrica tipa $m \times k$ čiji su elementi

$$f_{ij} = \sum_{l=1}^n a_{il} b_{lj} \quad \text{za } i \in \{1, \dots, m\}, j \in \{1, \dots, k\}.$$

Primjer 6. Odredimo matricu $F = A \cdot B$ ako je

$$\begin{aligned} \text{a)} \quad & A = \begin{bmatrix} 3 & -2 \end{bmatrix} \text{ i } B = \begin{bmatrix} 2 & 1 \end{bmatrix} \\ \text{b)} \quad & A = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 2 \end{bmatrix} \text{ i } B = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ 4 & 8 & 0 \end{bmatrix} \end{aligned}$$

Rješenje:

a) Matrica F nije definirana (tj. ne postoji) jer broj redaka matrice B nije jednak broju stupaca matrice A . Naime, matrica B ima jedan redak, a matrica A dva stupca.

b) $F = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \end{bmatrix}$, tj. F je matrica tipa 2×3 jer je A matrica tipa 2×3 , a B matrica tipa 3×3 ;

$$f_{11} = \sum_{l=1}^3 a_{1l}b_{l1} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} = 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 4 = 4,$$

$$f_{12} = \sum_{l=1}^3 a_{1l}b_{l2} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} = 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 8 = 9,$$

$$f_{13} = \sum_{l=1}^3 a_{1l}b_{l3} = a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33} = 1 \cdot (-1) + 0 \cdot 1 + 1 \cdot 0 \\ = -1,$$

$$f_{21} = \sum_{l=1}^3 a_{2l}b_{l1} = a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} = -1 \cdot 0 + 1 \cdot 1 + 2 \cdot 4 = 9,$$

$$f_{22} = \sum_{l=1}^3 a_{2l}b_{l2} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} = -1 \cdot 1 + 1 \cdot 0 + 2 \cdot 8 \\ = 15,$$

$$f_{23} = \sum_{l=1}^3 a_{2l}b_{l3} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} = -1 \cdot (-1) + 1 \cdot 1 + 2 \cdot 0 \\ = 2.$$

$$\text{Dakle, } F = \begin{bmatrix} 4 & 9 & -1 \\ 9 & 15 & 2 \end{bmatrix}.$$

Slijede najjednostavnije definicije elementarnih pojmova teorije grafova s praktičnim primjerima koje ćemo koristiti u nastavku ovog odlomka.

5.3 Što je graf?

Definicija 3. *Graf* je uređeni par $G = (V, E)$, gdje je V skup vrhova, a $E \subseteq \{u, v\} : u, v \in V, u \neq v\}$ skup bridova.

Definicija 4. *Usmjereni graf* je uređeni par $G = (V, E)$, gdje je V skup vrhova, a $E \subseteq V \times V$. Elemente skupa E nazivamo lukovima ili usmjerenim bridovima. Uređeni par $(u, v) \in E$ predstavlja luk usmjeren od vrha u prema vrhu v .

Ako je $(u, v) \in E$ i $(v, u) \notin E$, kratko pišemo

$$u \rightarrow v.$$

Ako je $(u, v) \in E$ i $(v, u) \in E$, kratko pišemo

$$u \rightleftarrows v.$$

Na primjer, u mreži web-stranica zapis $A \rightarrow B$ znači da stranica (vrh) A sadrži poveznicu na stranicu (vrh) B , dok stranica B nema poveznicu na stranicu A .

Definicija 5. Graf $G = (V, E)$ je **povezan** ako za svaka dva različita vrha $u, v \in V$ postoji put između njih, tj. konačan niz različitih vrhova

$$u = c_0, c_1, \dots, c_n = v,$$

takav da za svaki $i \in \{1, \dots, n\}$ vrijedi $\{c_{i-1}, c_i\} \in E$.

Definicija 6. Graf $G = (V, E)$ je **jako povezan** ako za svaka dva različita vrha $u, v \in V$ postoji usmjereni put između njih, tj. ako za svaka dva različita vrha $u, v \in V$ vrijedi:

1) postoji konačan niz različitih vrhova

$$u = c_0, c_1, \dots, c_n = v,$$

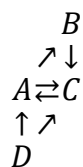
takav da za svaki $i \in \{1, \dots, n\}$ vrijedi $(c_{i-1}, c_i) \in E$

2) postoji konačan niz različitih vrhova

$$v = d_0, d_1, \dots, d_m = u$$

takav da za svaki $j \in \{1, \dots, m\}$ vrijedi $(d_{j-1}, d_j) \in E$.

Primjer 7.



je usmjereni graf čiji su vrhovi internetske stranice A, B, C, D s međusobnim vezama:

- A ima poveznice na B i C
- B ima poveznicu na C
- C ima poveznicu na A
- D ima poveznice na A i C .

Taj graf je povezan, ali nije jako povezan.

5.4 Matrica prijelaza

Definicija 7. Neka je $G = (V, E)$ usmjereni graf s n vrhova. **Matrica prijelaza** grafa G je matrica P tipa $n \times n$ čiji su elementi p_{ij} definirani formulom

p_{ij} = vjerojatnost da se u i -ti vrh stigne direktno iz j -tog vrha.

Ako za svaki $j \in \{1, \dots, n\}$ vrijedi

$$\sum_{i=1}^n p_{ij} = 1$$

i

$$p_{ij} \geq 0, \quad \forall i \in \{1, \dots, n\},$$

tada je P **stupčano stohastička matrica**.

Primjer 8. Za usmjereni graf iz prethodnog primjera je stranica A prvi vrh, stranica B drugi, stranica C treći, a stranica D četvrti vrh. Dakle, tu je $n = 4$.

- A ima poveznice na B i C pa je $p_{21} = p_{31} = 0,5$, a $p_{11} = p_{41} = 0$.
- B ima samo poveznicu na C pa je $p_{32} = 1$, a $p_{12} = p_{22} = p_{42} = 0$.
- C ima samo poveznicu na A pa je $p_{13} = 1$, a $p_{23} = p_{33} = p_{43} = 0$.
- D ima poveznice na A i C pa je $p_{14} = p_{34} = 0,5$, a $p_{24} = p_{44} = 0$.

Prema tome, matrica prijelaza P usmjerenog grafa iz prethodnog primjera je

$$P = \begin{bmatrix} 0 & 0 & 1 & 0,5 \\ 0,5 & 0 & 0 & 0 \\ 0,5 & 1 & 0 & 0,5 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Kao što vidimo, matrica prijelaza P je stohastička matrica koja opisuje pravila (vjerojatnosti) kretanja kroz mrežu – stupci su polazišta, a redci odredišta.

Definicija 8. **Vektor dimenzije n** je svaka matrica tipa $n \times 1$. Dakle, matrica

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

je vektor dimenzije n . Kažemo da je v_i **i -ta komponenta vektora V** (za $i = 1, \dots, n$). Vektor V je **stohastički vektor** ako vrijedi:

$$\sum_{i=1}^n v_i = 1, \quad v_i \geq 0, \forall i \in \{1, \dots, n\}.$$

Primjer 9. Slučajni surfer otvara internetski preglednik i na mini-webu može otvoriti samo četiri stranice (dakle, na webu nema drugih stranica osim te četiri). Situaciju u kojoj surfer prvim klikom sigurno otvara prvu stranicu opisuje stohastički vektor

$$V^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

čija i -ta komponenta ($v_i^{(1)}$) predstavlja vjerojatnost da se nakon prvog klika surfer nađe na i -toj stranici, za $i \in \{1, 2, 3, 4\}$.

Situaciju u kojoj drugim klikom svaku stranicu surfer otvara s jednakom vjerojatnošću opisuje stohastički vektor

$$V^{(2)} = \begin{bmatrix} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{bmatrix}$$

čija i -ta komponenta ($v_i^{(2)}$) predstavlja vjerojatnost da se nakon drugog klika surfer nađe na i -toj stranici (za $i \in \{1, 2, 3, 4\}$).

Na sličan se način, pomoću odgovarajućih stohastičkih vektora, mogu opisati situacije nakon bilo kojeg broja klikova.

5.5 Markovljev lanac

Definicija 9. Neka je (X_n) niz slučajnih varijabli sa zajedničkim konačnim skupom vrijednosti S , pri čemu X_n označava stanje nekog sustava u n -tom trenutku za svaki $n \in \mathbb{N}$. Niz (X_n) zove se **Markovljev lanac** ako ima **Markovljevo svojstvo**, tj. ako za svaki $n \in \mathbb{N}$ i bilo koji izbor stanja $s, s_1, \dots, s_n \in S$ vrijedi:

$$\begin{aligned} Pr(X_{n+1} = s | X_n = s_n, X_{n-1} = s_{n-1}, \dots, X_1 = s_1) \\ = Pr(X_{n+1} = s | X_n = s_n), \end{aligned}$$

je nepovezani lanac.

Teorem 1. *Ako je P pozitivna (stupčano stohastička) matrica prijelaza nekog povezanog lanca, onda postoji jedinstvena stacionarna raspodjela vjerojatnosti po stanjima sustava, tj. jedinstveni stohastički vektor V takav da je $P \cdot V = V$.*

Napomena. *Vektor V nazivamo stacionarnim stohastičkim vektorom. Taj vektor opisuje dugoročnu raspodjelu vjerojatnosti po stanjima sustava, tj. njegova i -ta komponenta (v_i) je vjerojatnost da se sustav u nekom trenutku, nakon mnogo prijelaza iz jednog stanja u drugo, nađe u i -tom stanju.*

Primjer 11. *U PageRank algoritmu, sve web-stranice (nekog mini-weba ili cijelog weba) čine usmjereni graf, odnosno Markovljev lanac u kojemu slučajni surfer predstavlja sustav, a stanja sustava su web-stranice koje slučajni surfer posjećuje. Pretpostavimo da lanac ima n stranica i da je P pozitivna stohastička matrica prijelaza tipa $n \times n$ koja opisuje vjerojatnosti kretanja kroz mrežu. Ako je $v_i^{(0)}$ vjerojatnost da se slučajni surfer u početnom trenutku surfanja nalazi na i -toj stranici za $i \in \{1, \dots, n\}$, onda tu početnu situaciju možemo opisati stohastičkim vektorom*

$$V^{(0)} = \begin{bmatrix} v_1^{(0)} \\ v_2^{(0)} \\ \vdots \\ v_n^{(0)} \end{bmatrix}.$$

Situaciju nakon prvog klika možemo opisati stohastičkim vektorom
$$V^{(1)} = P \cdot V^{(0)},$$

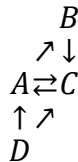
čija je i -ta komponenta ($v_i^{(1)}$) vjerojatnost da se slučajni surfer nakon prvog klika nalazi na i -toj stranici za $i \in \{1, \dots, n\}$. Itd.

*Po Teoremu 1, postoji jedinstveni stohastički vektor V za kojega vrijedi $P \cdot V = V$. Vektor V na jedinstven način opisuje **stacionarnu raspodjelu vjerojatnosti po stanjima (web-stranicama)** – situaciju koja se više ne mijenja, tj. ne ovisi o daljnjem načinu surfanja slučajnog surfera, pa tako određuje važnost svake stranice. Za svaki i , i -ta komponenta vektora V zapravo predstavlja omjer vremena koje surfer provede na i -toj stranici i ukupnog (dosta dugog) vremena surfanja, odnosno vjerojatnost da se slučajni surfer nađe na i -toj stranici nakon dugoročnog surfanja. Vjerojatnije stranice, tj. stranice s većim omjerom, su važnije.*

Primjer 12. *Matrica prijelaza*

$$P = \begin{bmatrix} 0 & 0 & 1 & 0,5 \\ 0,5 & 0 & 0 & 0 \\ 0,5 & 1 & 0 & 0,5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

iz primjera s usmjerenim grafom



opisuje pravila nepovezanog lanca jer npr. sa stranice A ne možemo doći na stranicu D. Matrica P ipak ima jedinstveni stacionarni stohastički vektor

$$V = \begin{bmatrix} 0,4 \\ 0,2 \\ 0,4 \\ 0 \end{bmatrix}$$

To ne znači da svaka matrica prijelaza koja opisuje pravila nepovezanog lanca ima jedinstveni stacionarni stohastički vektor. Npr. matrica prijelaza

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

opisuje pravila lanca s nepovezanim grafom



(u mini-webu sa stranicama A, B, C i D). U ovom slučaju stohastički vektor V za kojega vrijedi $P \cdot V = V$ nije jedinstven. Naime, za svaki stohastički vektor V oblika

$$V = \begin{bmatrix} \alpha \\ \alpha \\ \frac{1}{2} - \alpha \\ \frac{1}{2} - \alpha \end{bmatrix}, \quad \alpha \in \left[0, \frac{1}{2}\right]$$

vrijedi $P \cdot V = V$.

Problemi, kao što su stranice bez linkova i skup(ovi) stranica koje linkaju samo međusobno, se stalno pojavljuju na stvarnom webu. Stoga Google, umjesto matrice P (koja je tipa $n \times n$ i nije pozitivna), za matricu prijelaza uvijek uzima stupčano stohastičku matricu

$$P' = 0,85P + 0,15E,$$

gdje je $E = [e_{ij}]$ matrica tipa $n \times n$ takva da je $e_{ij} = \frac{1}{n}$ za sve $i, j \in \{1, \dots, n\}$. Matrica P' je uvijek pozitivna pa postoji mala vjerojatnost da eventualnim nehotimičnim klikom slučajni surfer skoči na bilo koju stranicu mreže (ma kako beznačajna ona bila) ili da ostane na stranici na kojoj se trenutno nalazi. Onda, po Teoremu 1, postoji jedinstveni stohastički vektor V takav da je $P' \cdot V = V$. Taj stacionarni vektor V određuje PageRank svake stranice. Naravno, zanima nas kako naći taj vektor. O tome detaljnije u nastavku rada. Najprije ćemo dokazati nekoliko jednostavnih tvrdnji.

Tvrdnja 1. *Ako je P stupčano stohastička matrica tipa $n \times n$ i Q stohastički vektor tipa $n \times 1$, onda je i vektor $U = P \cdot Q$ stohastički vektor (tipa $n \times 1$).*

Dokaz. Neka je $P = [p_{ij}]$ stupčano stohastička matrica tipa $n \times n$. Tada za svaki $j \in \{1, \dots, n\}$ vrijedi:

$$p_{ij} \geq 0, \quad \forall i \in \{1, \dots, n\}$$

i

$$\sum_{i=1}^n p_{ij} = 1.$$

Neka je $Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}$ stohastički vektor tipa $n \times 1$ (što znači da je $q_j \geq 0, \forall j \in \{1, \dots, n\}$ i $\sum_{j=1}^n q_j = 1$). Tada je

$$U = P \cdot Q = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

pri čemu za komponente vektora U vrijedi:

$$u_i = \sum_{j=1}^n p_{ij} q_j, \quad \forall i \in \{1, \dots, n\}.$$

Za svaki $i \in \{1, \dots, n\}$ je $u_i \geq 0$ jer je u_i jednak zbroju umnožaka nenegativnih brojeva.

Nadalje je

$$\sum_{i=1}^n u_i = \sum_{i=1}^n \left(\sum_{j=1}^n p_{ij} q_j \right) = \sum_{j=1}^n \left(\underbrace{\sum_{i=1}^n p_{ij}}_{=1, \forall j} \right) q_j = \sum_{j=1}^n q_j = 1.$$

Dakle, i vektor U je stohastičan.

□

Za vektor $T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$ definiramo $\|T\|_1$ sa

$$\|T\|_1 = \sum_{i=1}^n |t_i|.$$

Za stohastički vektor T vrijedi $\|T\|_1 = 1$.

Tvrđnja 2. Ako je P stupčano stohastička matrica tipa $n \times n$ onda za svaka dva stohastička vektora Q i T tipa $n \times 1$ vrijedi

$$\|P \cdot Q - P \cdot T\|_1 \leq \|Q - T\|_1.$$

Dokaz.

$$U = P \cdot Q - P \cdot T = P \cdot (Q - T) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

pri čemu za komponente vektora U vrijedi:

$$u_i = \sum_{j=1}^n p_{ij}(q_j - t_j), \quad \forall i \in \{1, \dots, n\}.$$

Stoga je

$$\begin{aligned} \|P \cdot Q - P \cdot T\|_1 &= \|U\|_1 = \sum_{i=1}^n |u_i| = \sum_{i=1}^n \left| \sum_{j=1}^n p_{ij}(q_j - t_j) \right| \leq \\ &\leq \sum_{i=1}^n \sum_{j=1}^n p_{ij} |q_j - t_j| = \sum_{j=1}^n \left(\underbrace{\sum_{i=1}^n p_{ij}}_{=1, \forall j} \right) |q_j - t_j| = \sum_{j=1}^n |q_j - t_j| = \|Q - T\|_1. \end{aligned}$$

□

Tvrđnja 3. Neka je P stupčano stohastička matrica tipa $n \times n$ i $P' = dP + (1-d)E$ za bilo koji $d \in (0,1)$, gdje je $E = [e_{ij}]$ matrica tipa $n \times n$ takva da je $e_{ij} = \frac{1}{n}$ za sve $i, j \in \{1, \dots, n\}$. Tada za svaka dva stohastička vektora Q i T tipa $n \times 1$ vrijedi:

$$\|P' \cdot Q - P' \cdot T\|_1 \leq d \|Q - T\|_1.$$

Dokaz.

$$\begin{aligned} \|P' \cdot Q - P' \cdot T\|_1 &= \|[dP + (1-d)E] \cdot Q - [dP + (1-d)E] \cdot T\|_1 \\ &= \|dP \cdot (Q - T) + (1-d)(E \cdot Q - E \cdot T)\|_1 \end{aligned}$$

Kako je

$$E \cdot Q = E \cdot T = \frac{1}{n} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

onda primjenom Tvrdnje 2 dobivamo

$$\|P' \cdot Q - P' \cdot T\|_1 = \|dP \cdot (Q - T)\|_1 = d\|P \cdot (Q - T)\|_1 \leq d\|Q - T\|_1.$$

□

P' je pozitivna matrica pa, po Teoremu 1, postoji jedinstveni stohastički vektor V tipa $n \times 1$ takav da vrijedi:

$$P' \cdot V = V.$$

Neka je $V^{(0)}$ bilo koji stohastički vektor tipa $n \times 1$ i $(V^{(m)})$ niz stohastičkih vektora definiran rekurzivnom formulom

$$V^{(m)} = P' \cdot V^{(m-1)}, \quad \forall m \in \mathbb{N}.$$

Po Tvrdnji 3, za svaki $m \in \mathbb{N}$ i svaki $d \in \langle 0, 1 \rangle$ vrijedi

$$\|P' \cdot V^{(m-1)} - P' \cdot V\|_1 \leq d\|V^{(m-1)} - V\|_1,$$

odnosno

$$\|V^{(m)} - V\|_1 \leq d\|V^{(m-1)} - V\|_1. \quad (*)$$

Stoga je

$$\begin{aligned} \|V^{(m)} - V\|_1 &\leq d\|V^{(m-1)} - V\|_1 \leq d^2\|V^{(m-2)} - V\|_1 \leq \dots \\ &\leq d^m\|V^{(0)} - V\|_1. \end{aligned}$$

Iz

$$\lim_{m \rightarrow \infty} (d^m\|V^{(0)} - V\|_1) = \|V^{(0)} - V\|_1 \lim_{m \rightarrow \infty} d^m = \|V^{(0)} - V\|_1 \cdot 0 = 0$$

onda slijedi

$$\lim_{m \rightarrow \infty} \|V^{(m)} - V\|_1 = 0.$$

Tvrdnja 4. Za proizvoljne vektore $Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}$ i $T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$ vrijedi:

$$|\|Q\|_1 - \|T\|_1| \leq \|Q - T\|_1.$$

Dokaz. Znamo da za svaka dva realna broja r_1 i r_2 vrijedi:

$$||r_1| - |r_2|| \leq |r_1 - r_2|.$$

Stoga je za svaki $i \in \{1, \dots, n\}$

$$||q_i| - |t_i|| \leq |q_i - t_i|$$

pa je

$$\sum_{i=1}^n ||q_i| - |t_i|| \leq \sum_{i=1}^n |q_i - t_i| = \|Q - T\|_1.$$

S druge strane,

$$\left| \|Q\|_1 - \|T\|_1 \right| = \left| \sum_{i=1}^n |q_i| - \sum_{i=1}^n |t_i| \right| = \left| \sum_{i=1}^n (|q_i| - |t_i|) \right| \leq \sum_{i=1}^n \left| |q_i| - |t_i| \right|.$$

Kombiniranjem dobivenih nejednakosti

$$\left| \|Q\|_1 - \|T\|_1 \right| \leq \sum_{i=1}^n \left| |q_i| - |t_i| \right|$$

i

$$\sum_{i=1}^n \left| |q_i| - |t_i| \right| \leq \|Q - T\|_1$$

dobivamo

$$\left| \|Q\|_1 - \|T\|_1 \right| \leq \|Q - T\|_1.$$

□

Tvrđnja 5. Za svaki $m \in \mathbb{N}_0$ i svaki $d \in \langle 0,1 \rangle$ vrijedi:

$$\|V^{(m+1)} - V\|_1 \leq \frac{d}{1-d} \|V^{(m+1)} - V^{(m)}\|_1.$$

Dokaz. Izaberimo bilo koji $m \in \mathbb{N}_0$ i bilo koji $d \in \langle 0,1 \rangle$. Ako u Tvrđnju 4 uvrstimo vektore $Q = V^{(m)} - V$ i $T = V^{(m+1)} - V$ dobivamo

$$\left| \|V^{(m)} - V\|_1 - \|V^{(m+1)} - V\|_1 \right| \leq \|V^{(m+1)} - V^{(m)}\|_1. \quad (**)$$

Iz nejednakosti (*) slijedi sljedeći niz međusobno ekvivalentnih nejednakosti:

$$\begin{aligned} \|V^{(m+1)} - V\|_1 &\leq d \|V^{(m)} - V\|_1 \\ -\|V^{(m+1)} - V\|_1 &\geq -d \|V^{(m)} - V\|_1 \\ \|V^{(m)} - V\|_1 - \|V^{(m+1)} - V\|_1 &\geq (1-d) \|V^{(m)} - V\|_1. \end{aligned}$$

Kako je $(1-d) \|V^{(m)} - V\|_1 \geq 0$, onda je i

$$\|V^{(m)} - V\|_1 - \|V^{(m+1)} - V\|_1 \geq 0.$$

Stoga iz nejednakosti (**) dobivamo:

$$\|V^{(m)} - V\|_1 - \|V^{(m+1)} - V\|_1 \leq \|V^{(m+1)} - V^{(m)}\|_1,$$

odnosno

$$\|V^{(m)} - V\|_1 \leq \|V^{(m+1)} - V\|_1 + \|V^{(m+1)} - V^{(m)}\|_1. \quad (***)$$

Dakle,

$$\begin{aligned} \|V^{(m+1)} - V\|_1 &\stackrel{(*)}{\leq} d \|V^{(m)} - V\|_1 \\ &\stackrel{(***)}{\leq} d \left(\|V^{(m+1)} - V\|_1 + \|V^{(m+1)} - V^{(m)}\|_1 \right), \end{aligned}$$

tj.

$$\|V^{(m+1)} - V\|_1 \leq d \|V^{(m+1)} - V\|_1 + d \|V^{(m+1)} - V^{(m)}\|_1,$$

odakle slijedi

$$(1-d) \|V^{(m+1)} - V\|_1 \leq d \|V^{(m+1)} - V^{(m)}\|_1,$$

odnosno

$$\|V^{(m+1)} - V\|_1 \leq \frac{d}{1-d} \|V^{(m+1)} - V^{(m)}\|_1.$$

□

Posebno, za $d = 0,85$ (algoritam PageRank) vrijedi:

$$\|V^{(m+1)} - V\|_1 \leq \frac{17}{3} \|V^{(m+1)} - V^{(m)}\|_1.$$

Ako je

$$M = \max_{1 \leq i \leq n} |v_i^{(m+1)} - v_i^{(m)}|,$$

gdje je $v_i^{(m+1)} - v_i^{(m)}$ i -ta komponenta vektora $V^{(m+1)} - V^{(m)}$, onda je

$$\|V^{(m+1)} - V^{(m)}\|_1 \leq nM$$

pa je

$$\|V^{(m+1)} - V\|_1 \leq \frac{17}{3} nM.$$

Ako za izabrani $\varepsilon > 0$ vrijedi $M < \frac{3\varepsilon}{17n}$, onda je

$$\|V^{(m+1)} - V\|_1 < \varepsilon$$

pa za svaki $i \in \{1, \dots, n\}$ vrijedi

$$|v_i^{(m+1)} - v_i| < \varepsilon,$$

odnosno

$$v_i \in \left\langle v_i^{(m+1)} - \varepsilon, v_i^{(m+1)} + \varepsilon \right\rangle,$$

gdje je v_i i -ta komponenta vektora V , a $v_i^{(m+1)}$ i -ta komponenta vektora $V^{(m+1)}$.

Broj d se naziva **faktor prigušenja** (eng. **damping factor**) matrice P . Objasnimo zašto je Google za određivanje PageRanka izabrao baš vrijednost $d = 0,85$.

Kad bi se uzeo premali d , matrica P' bi se znatno razlikovala od matrice P , a to nije poželjno jer matrica P najbolje opisuje ponašanje svih korisnika na mreži. Naime, moguće je da se u formiranju matrice P ne uzimaju u obzir samo linkovi na svim web-stranicama, već i statistički podatci o navikama surfanja. Tako matricu P zapravo definira mudrost mnoštva svih korisnika, ne samo web-dizajnera koji pri izradi web-stranica postavljaju poveznice na druge web-stranice, nego i mrežne aktivnosti svih surfera.

S druge strane, d ne smije biti preblizu broja 1 jer bi to značajno usporilo konvergenciju niza vektora $(V^{(m)})$ prema vektoru V .

Stoga se $d = 0,85$ zaista čini kao vrlo dobar izbor.

Primjer 13. Matrica P' za matricu $P = \begin{bmatrix} 0 & 0 & 1 & 0,5 \\ 0,5 & 0 & 0 & 0 \\ 0,5 & 1 & 0 & 0,5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ iz

prethodnih primjera je $P' = 0,85P + 0,15E$, gdje je

$$E = \begin{bmatrix} 0,25 & 0,25 & 0,25 & 0,25 \\ 0,25 & 0,25 & 0,25 & 0,25 \\ 0,25 & 0,25 & 0,25 & 0,25 \\ 0,25 & 0,25 & 0,25 & 0,25 \end{bmatrix}.$$

Tako dobijemo:

$$P' = \begin{bmatrix} 0,0375 & 0,0375 & 0,8875 & 0,4625 \\ 0,4625 & 0,0375 & 0,0375 & 0,0375 \\ 0,4625 & 0,8875 & 0,0375 & 0,4625 \\ 0,0375 & 0,0375 & 0,0375 & 0,0375 \end{bmatrix}.$$

P' je pozitivna stupčano stohastička matrica pa postoji jedinstveni stohastički vektor V tipa 4×1 takav da je $P' \cdot V = V$.

Nađimo aproksimaciju $V^{(m+1)}$ vektora V za koju vrijedi

$$\|V^{(m+1)} - V\|_1 < 10^{-10}.$$

Iz uvjeta

$$M < \frac{3\varepsilon}{17n}$$

za $\varepsilon = 10^{-10}$ i $n = 4$ dobivamo kriterij zaustavljanja računanja iteracija $V^{(m)}$:

$$M < \frac{3}{680000000000} \approx 4,411764705882353 \times 10^{-12}.$$

Pretpostavimo da početnu situaciju opisuje stohastički vektor

$$V^{(0)} = \begin{bmatrix} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{bmatrix}.$$

Nakon 50 iteracija (tj. 50 množenja s matricom P') dobivamo

$$V^{(50)} \approx \begin{bmatrix} 0,379734313172914 \\ 0,198887083097552 \\ 0,383878603729534 \\ 0,0375 \end{bmatrix}$$

i

$$M = \max_{1 \leq i \leq n} |v_i^{(50)} - v_i^{(49)}| \approx 3,305467011216479 \times 10^{-12} < 4,411764705882353 \times 10^{-12}.$$

Dakle,

$$\|V^{(50)} - V\|_1 < 10^{-10}$$

pa za svaki $i \in \{1,2,3\}$ komponenta v_i vektora V zadovoljava uvjet

$$v_i \in \left\langle v_i^{(50)} - 10^{-10}, v_i^{(50)} + 10^{-10} \right\rangle,$$

gdje je $v_i^{(50)}$ i -ta komponenta vektora $V^{(50)}$.

Četvrta komponenta vektora V je izračunata bez greške jer je

$$v_4^{(m)} = 0,0375, \quad \forall m \in \mathbb{N}.$$

Stoga je

$$V = \begin{bmatrix} 0,379734313 \dots \\ 0,19888708 \dots \\ 0,383878603 \dots \\ 0,0375 \end{bmatrix}.$$

Prema dobivenom stacionarnom stohastičkom vektoru V zaključujemo da slučajni surfer našeg mini-weba s 4 stranice najviše vremena provodi na stranici C ($\approx 0,3838786t$), a najmanje na stranici D ($0,0375t$), gdje je t ukupno vrijeme surfanja. To ustvari znači da se nakon 50 i više klikova oko 38,38% svih surfera nalazi na stranici C , a samo 3,75% njih na stranici D , pa je ukupni poredak stranica po važnosti:

C

A

B

D .

Napomena. Isti stacionarni stohastički vektor V bismo dobili za bilo koji drugi početni stohastički vektor $V^{(0)}$. Dakle, V (tj. poredak stranica) ovisi isključivo o matrici P' . Kako P' ovisi samo o matrici P (jer je za veći n svaki element matrice $0,15E$ vrlo blizu nule), zaključujemo da V ovisi isključivo o matrici P koja predstavlja mudrost mnoštva svih korisnika interneta zajedno.

Google koristi „paukove“ (web crawleri) koji neprestano obilaze web stranice i na svakoj stranici bilježe sve linkove na druge stranice. Tako se gradi veliki usmjereni graf weba u kojem su vrhovi stranice, a usmjereni bridovi linkovi. Taj graf sadrži na milijarde vrhova i često se ažurira. Google ne računa važnost svih stranica iznova svaki put kad netko nešto pretražuje. Važnost se računa periodično offline kada se puno stranica promijeni ili se pojavi mnogo novih stranica. Da bi

računanje važnosti s velikim matricama bilo izvedivo moraju se koristiti razne prikladne numeričke metode. Dobiveni rezultati se spremaju i kasnije koriste u rangiranju rezultata.

Vidjeli smo da se mudrost mnoštva zapravo koristi kod određivanja svake od veličina: r , v i b .

Za težinske koeficijente w_1 , w_2 i w_3 vrijedi:

$$\sum_{i=1}^3 w_i = 1, \quad w_1, w_2, w_3 \geq 0.$$

Ti koeficijenti nisu javno dostupni. Kako ih Google određuje nije poznato. Moguće je da se pritom koriste kolektivne procjene kao i u metodi određivanja težina u prethodnom odjeljku.

Navedenim primjerima smo pokazali snagu matematike u stvarnom svijetu. Iza svake pretrage na webu stoji linearna algebra (matrice) i teorija vjerojatnosti, a mudrost mnoštva korisnika i web-dizajnera zajedno određuje konačni poredak stranica. Svaki klik svakog korisnika, koliko god izgledao beznačajno, sudjeluje u oblikovanju tog poretka. Rezultate jedne Googleove pretrage određuju milijarde glasova (klikova) – prava mudrost mnoštva.

6. Zaključak

Mudrost mnoštva je zanimljiv i koristan fenomen. Preko primjera u ovom radu opisane su mogućnosti njegove primjene u različitim situacijama – na sajmovima, u vojnim operacijama, u obrazovanju te pri internetskom pretraživanju. Teorem srednje vrijednosti za određeni integral poslužio nam je kao matematički model kojim se može objasniti mudrost mnoštva, a pokazali smo i kako se metoda najmanjih kvadrata može koristiti u složenijim primjenama tog fenomena. Poznata je poslovice „Znanje je moć“ engleskog filozofa Francisa Bacona. Ovaj nas fenomen uči da snaga nije samo u znanju pojedinca, nego i u zajedničkom razmišljanju. Kada se različite perspektive i iskustva udruže, moguće je doći do preciznijih i pouzdanijih rješenja koja su ponekad čak i bolja od onih koje bi ponudio najbolji stručnjak. Upravo zato mudrost mnoštva nije samo zanimljiva znanstvena ideja, nego i praktična lekcija o vrijednosti suradnje – u znanosti, obrazovanju i svakodnevnom životu.

Ipak, važno je naglasiti da mudrost mnoštva ne vrijedi u svim situacijama. Naime, kolektiv će dati dovoljno točnu prosječnu procjenu samo ako su njegovi članovi dovoljno međusobno različiti i neovisni. Nadalje, fenomen se osobito dobro očituje tek kod većih kolektiva jer kod manjih grupa u kolektivnoj prosječnoj procjeni često ne dolazi do umanjivanja utjecaja pogrešaka pojedinaca.

Većina ideja i primjera prikazanih u ovom radu ima izvorište u izvrsnoj knjizi Surowieckog [1]. Detaljnije o matricama i nekim zanimljivim primjenama zainteresirani čitatelj može pročitati u udžbeniku [2]. Teorem 1 u ovom radu slijedi iz Perronovog, odnosno Perron-Frobeniusovog teorema. Dokazi tih teorema se nalaze u diplomskom radu [3]. Na kraju navodimo popis ne previše komplicirane literature primjerene znatiželjnom srednjoškolcu ili studentu:

- [4] daje jednostavan uvod u vjerojatnost i Markovljeve lance s puno primjera iz svakodnevnog života,
- [5] je pristupačna knjiga s intuitivnim objašnjenjima i vizualnim prikazima,
- [6] predstavlja dobar most između osnovnog razumijevanja i ozbiljnije teorije,
- [7] na popularno-znanstveni način jednostavnim primjerima prikazuje matematičke ideje, posebno Markovljeve lance i Googleov algoritam PageRank.

Literatura

- [1] J. Surowiecki, *The Wisdom of Crowds*, Anchor Books, New York, 2005.
- [2] Grupa autora, *Mathematics Theory, Simulations and Exercises for Maritime Students*, University of Split, Faculty of Maritime Studies, Split, 2024.
- [3] M. Puček, *Perron-Frobeniusov teorem*, diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Zagreb, 2020.
- [4] S. M. Ross, *Introduction to Probability Models*, Academic Press, Amsterdam, 2014.
- [5] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation*, Wiley, Hoboken, 2017.
- [6] J. R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, 1997.
- [7] D. Sumpter, *The Ten Equations that Rule the World*, Allen Lane, London, 2020.

Goran Kovačević
Sveučilište u Splitu, Pomorski fakultet
Ruđera Boškovića 37, Split
E-mail adresa: gkovacev@pfst.hr