Matematika delovanja spletnih iskalnikov

Rok Požar

Univerza na Primorskem

Fakulteta za matematiko, naravoslovje in informacijske tehnologije

Famnitovi izleti v matematično vesolje

Januar 2015

Zgodovina iskanja informacij

Zgodovina iskanja informacij, stari Grki in Rimljani

Informacije zbrane na papirusovih zvitkih

povzetki kot značke na dokumentih,
povzetki kot ustna sporočila: grške igre (5. st. pr. n. št.)

Prvi knjižnični katalog

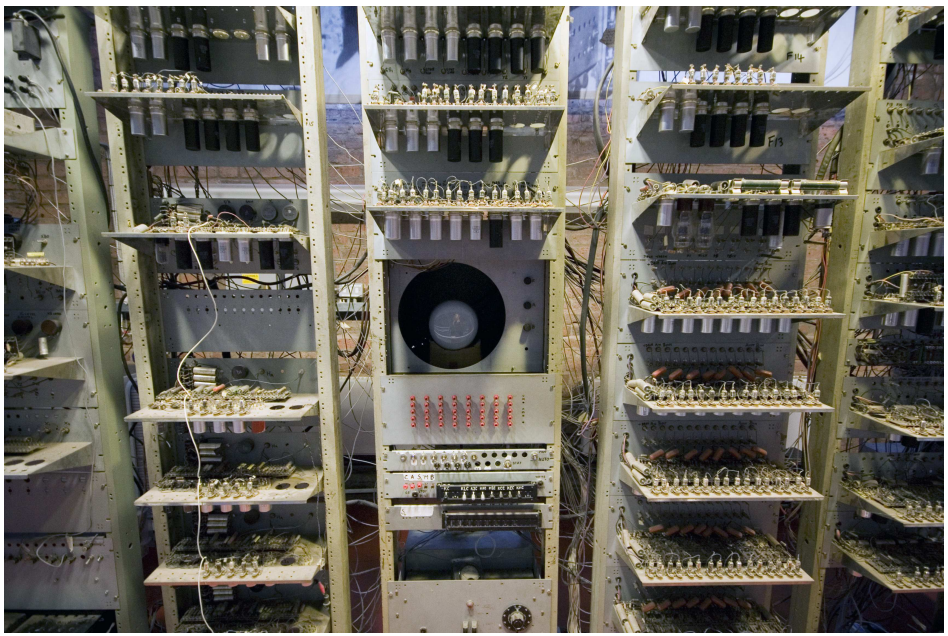
knjižnica v Aleksandriji (2. st. pr. n. št.),
skoraj 500 000 zvitkov razvrščenih po vsebini
Pinakes: tablice z naslovom in bibliografskimi podatki o avtorju

Zgodovina iskanja informacij, čas Guthenberga

Tisk s premičnimi črkami (1450)

hierarhičen sistem klasifikacije: skupine s podobno tematiko,
nasveti knjižnjičarja,
urejanje po naslovu in avtorju

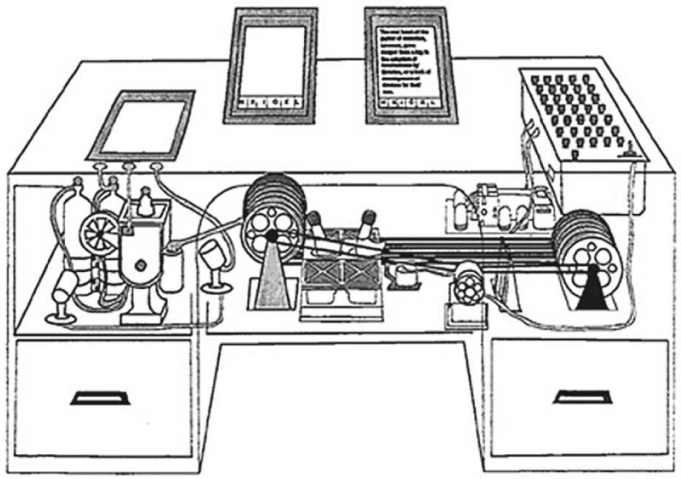
Zgodovina iskanja informacij, začetek digitalne dobe



Digitalni računalnik (1940-1950)

prvi računalniški iskalni sistemi,
avtomatično povpraševanje po knjigah in člankih (Cornel SMART sistem)

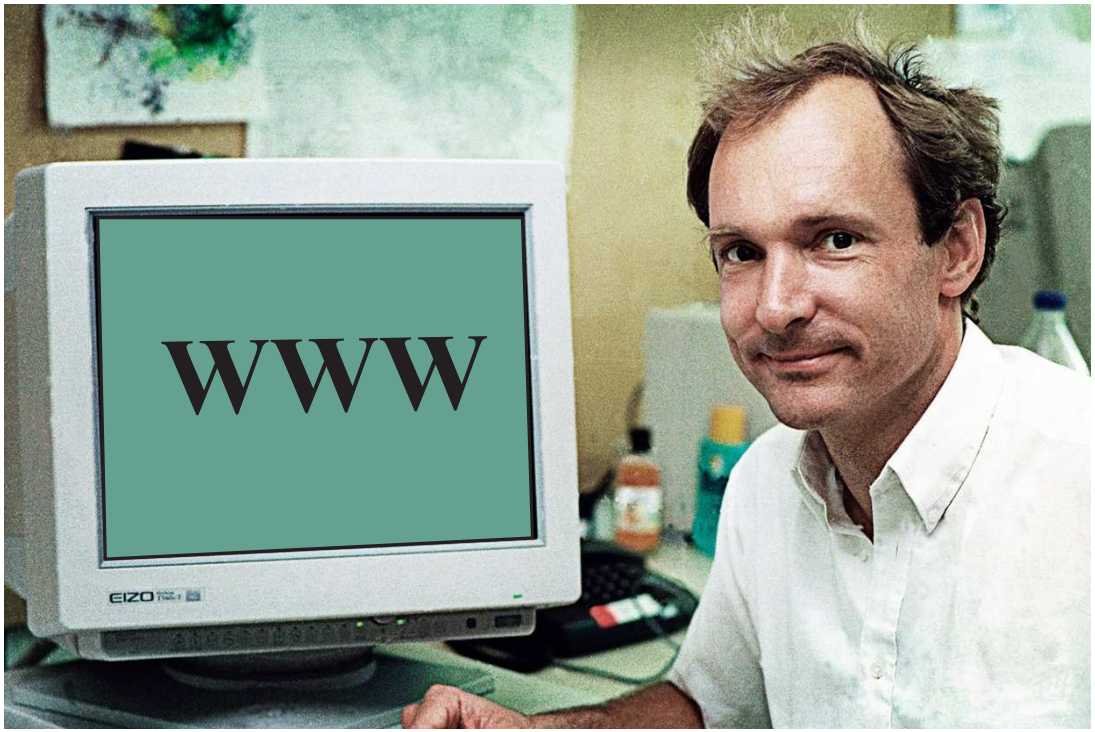
Zgodovina iskanja informacij, Vannevar Bush



Futuristični sestav *Memex*

esej: ,,As we may think“ (1945),
pristop kazal ni ustrezen,
,,človeško razmišljanje ne deluje tako, s tekoče misli preskoči na naslednjo“

Zgodovina iskanja informacij, Tim Berners-Lee



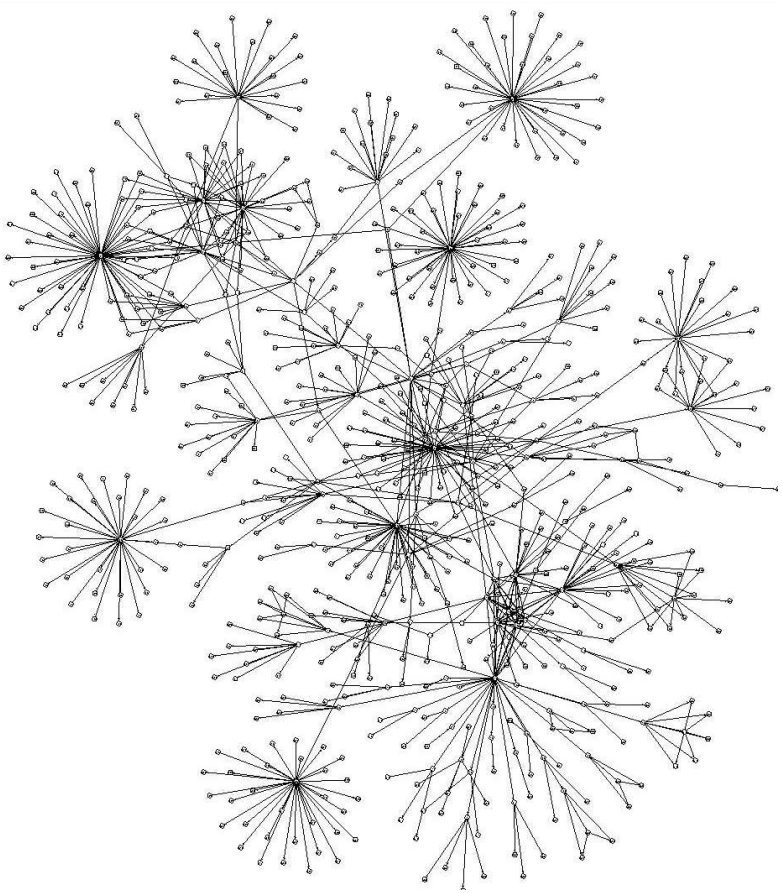
Ustanovitelj svetovnega spleta (World Wide Web, 1989)

povezovanje dokumentov na podlagi hiperpovezav in hiperteksta,
revolucija shranjevanja in dostopnosti informacij

Spletno iskanje kot iskanje igle v kupu sena

pregled nad spletom je temeljil na kazalih: razvrščanje po temah (Yahoo)

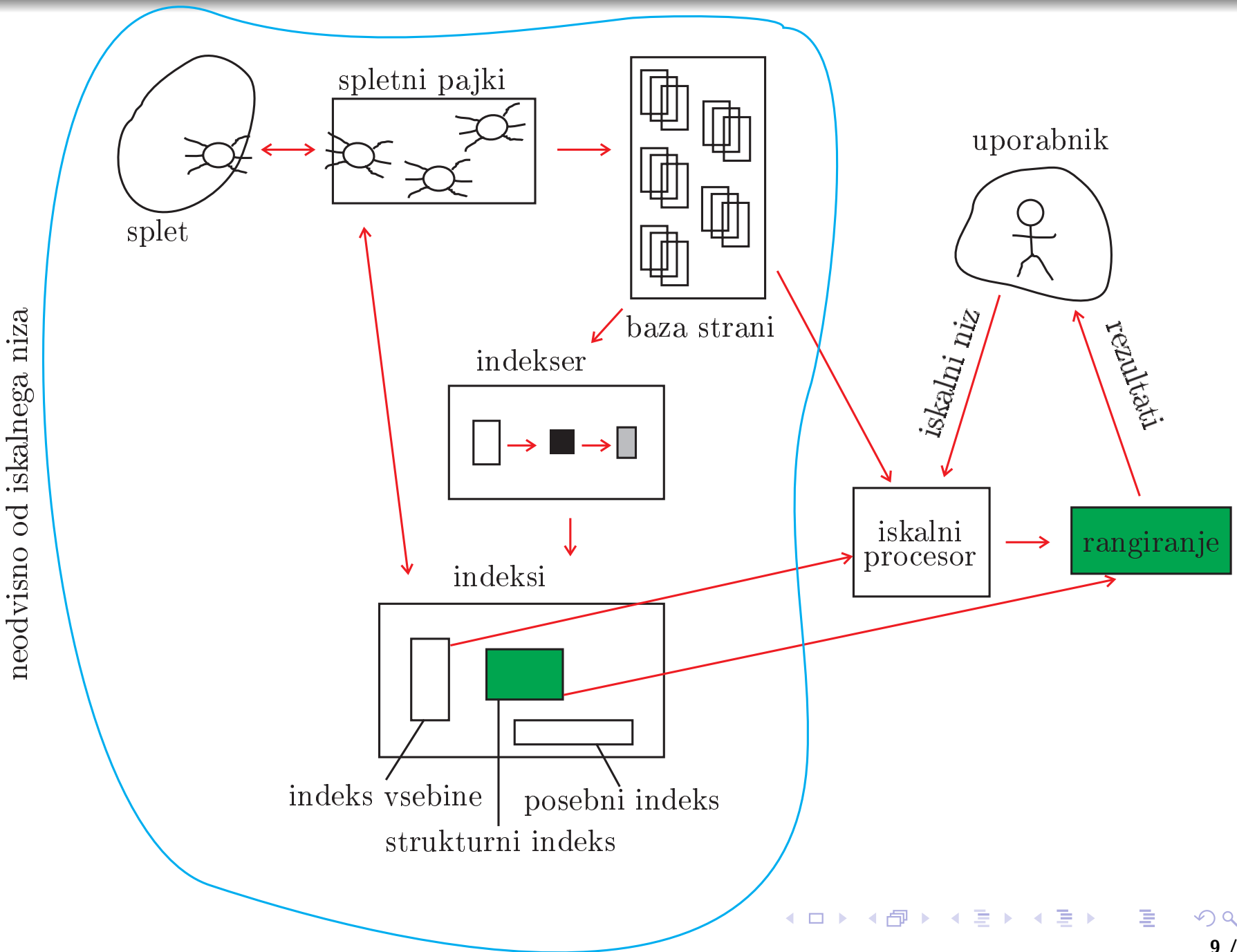
Zgodovina iskanja informacij, analiza hiperpovezav



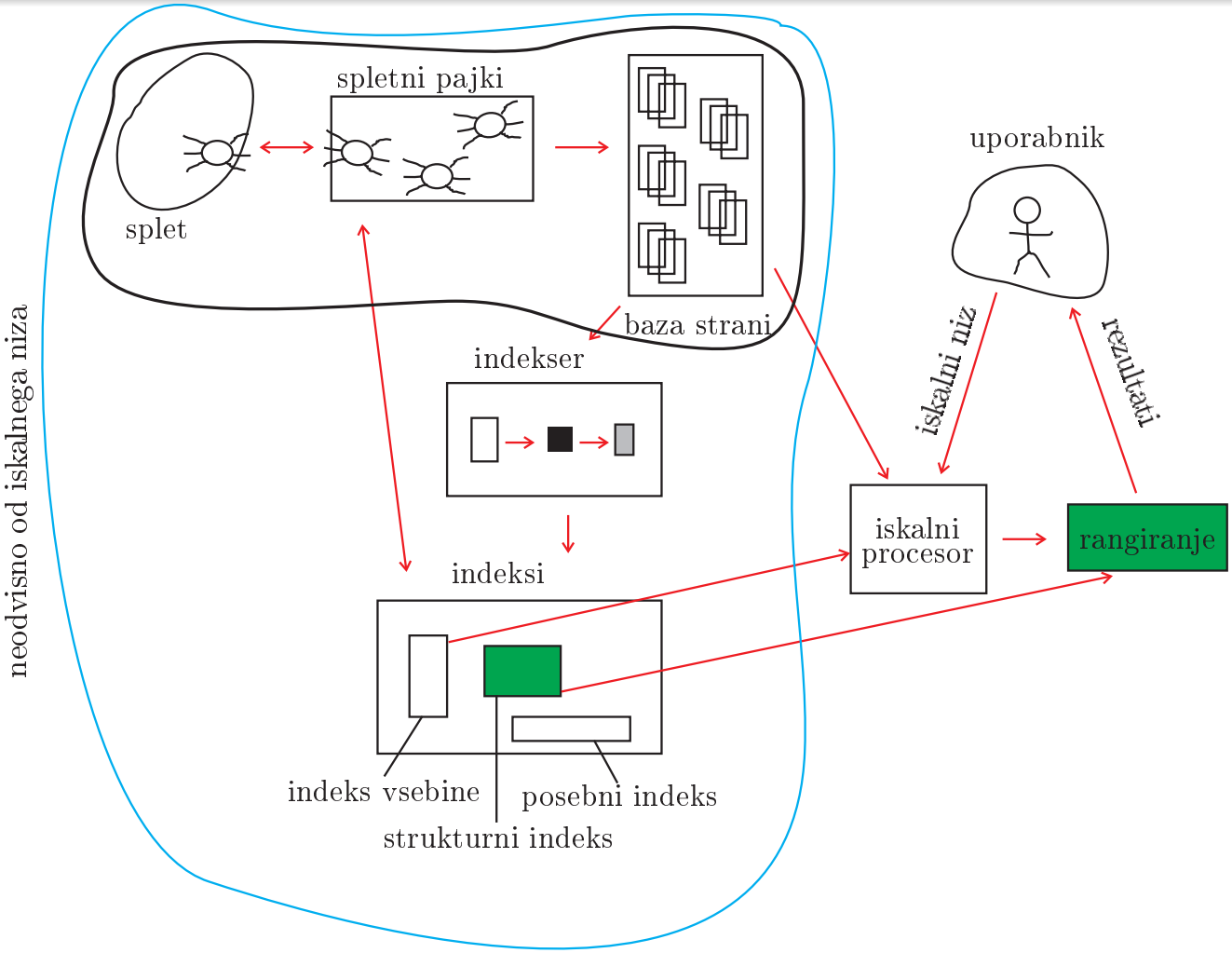
Študij strukturnih lastnosti svetovnega spleta (1998)

spletno iskanje strmo napreduje (Google)

Kako deluje spletni iskalnik?



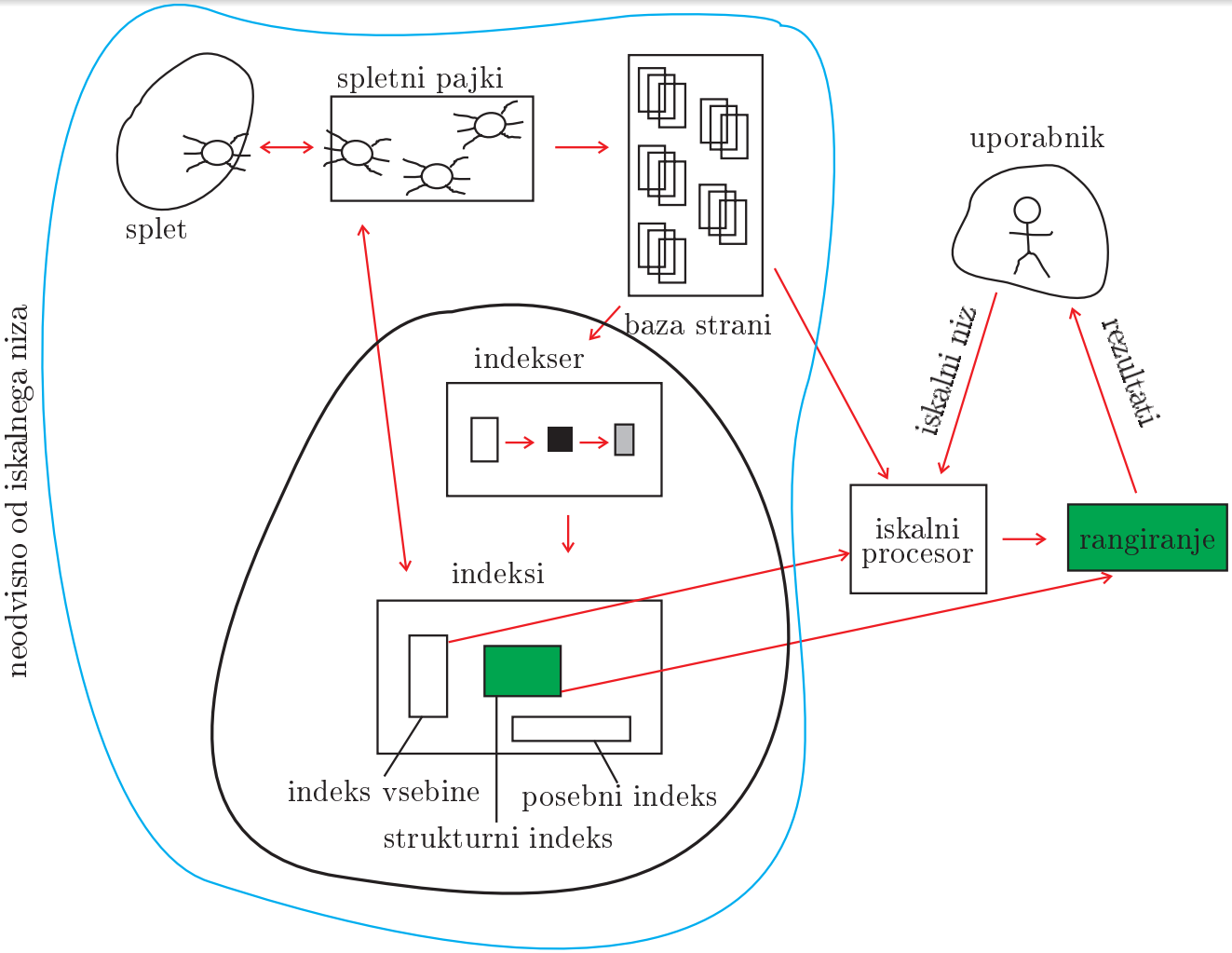
Kopiranje podatkov s spleta



Spletni pajki

iščejo in zapisujejo vsebino spleta v svojo bazo,
strani ostanejo v bazi do indeksiranja,
popularne strani so v bazi shranjene dlje

Indeksiranje spletnih strani



Indekser

izlušči le važne informacije,
naredi skrčen opis strani,
opis shrani v različne indekse

Indeksi

Indeks vsebine

tekstna vsebina, shranjena v obliki obrnjenih seznamov

$$\begin{aligned}
&\texttt{abak} - 3, 117, 3961\\
&\quad\vdots\\
&\texttt{abonent} - 3, 5, 19, 101, 117, 367, 3961\\
&\texttt{otrok} - 3, 31, 56, 94, 367, 673, 909, 11114\\
&\quad\vdots\\
&\texttt{zrak} - 344521
\end{aligned}$$

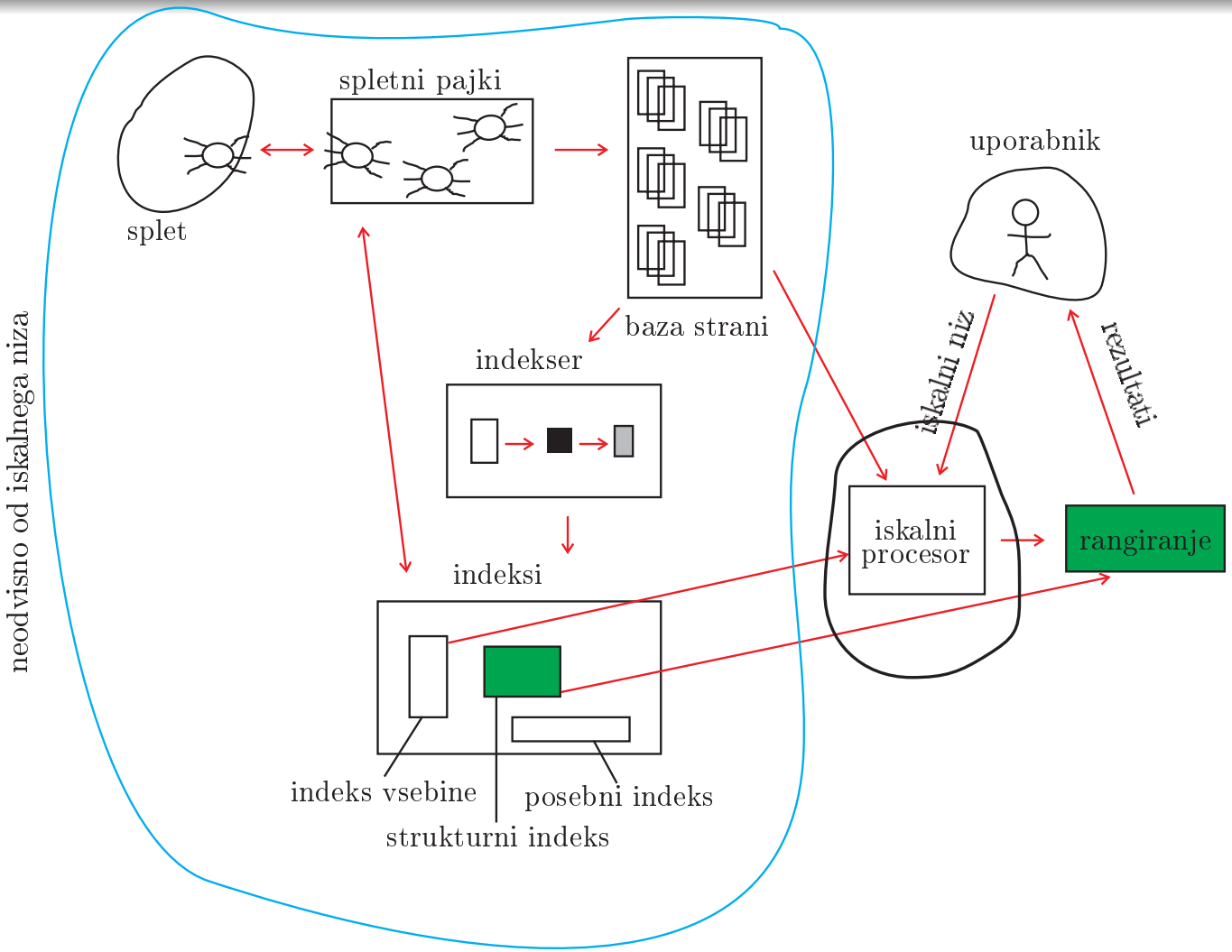
Strukturni indeks

struktura hiperpovezav med spletnimi stranmi, shranjena v skrčeni oblik, pajki ga včasih uporabljajo za iskanje novih strani

Posebni indeks

slike, pdf dokumenti,...

Iskanje relavantnih strani



Iskalni procesor

pretvori iskalni niz v sebi razumljiv jezik,
poveže se z indeksom vsebine in obrnjenim seznamom,
poišče **relavantne strani** (tiste, na katerih se pojavi iskalni niz)

Rangiranje relavantnih strani

Ocena glede na vsebino

upošteva, kje se nahaja iskalni niz (naslov, povzetek ali jedro),
upošteva relativne ferkvence pojavitev ključnih besed,
izračuna se na podlagi indeksa vsebine in obrnjenega sezanma,
odvisna od iskalnega niza

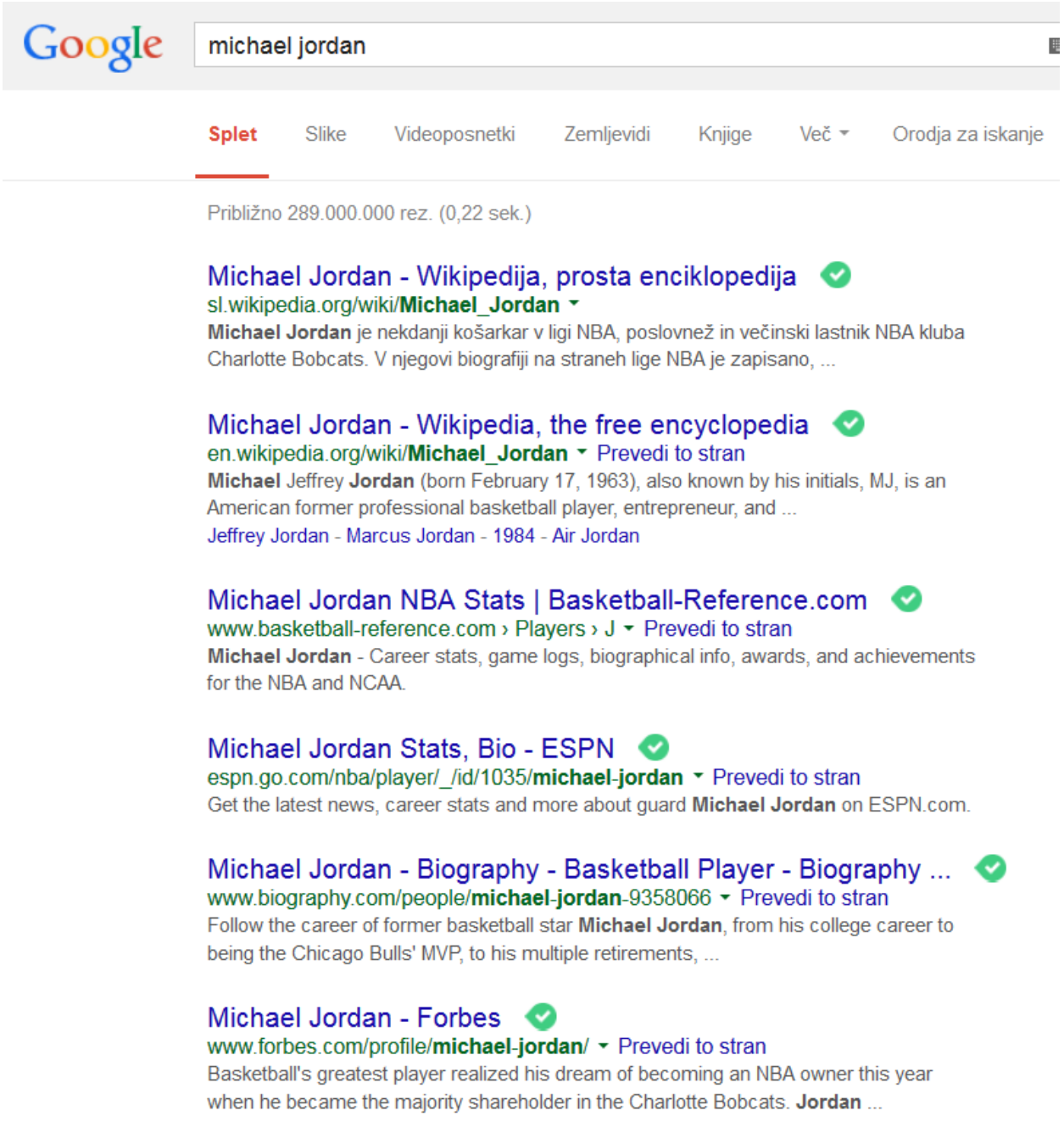
Ocena glede na popularnost

upošteva analizo strukturnih lastnosti hiperpovezav,
izračuna se iz strukturnega indeksa,
ponavadi neodvisna od iskalnega niza

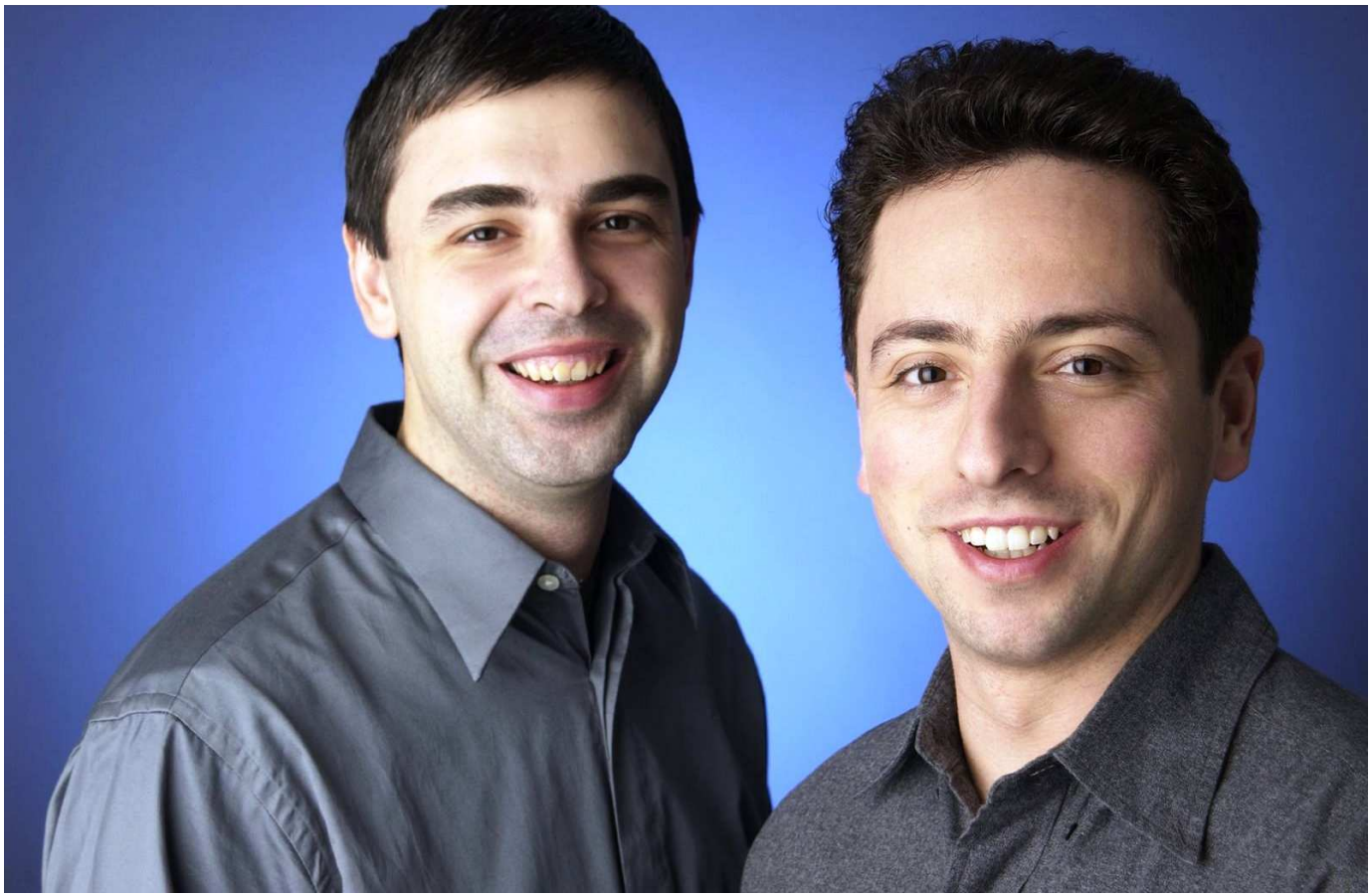
Skupna ocena = vsebinska ocena + ocena popularnosti

rezultat je urejen seznam relavantni strani,
bolj ,,verjetne“ strani so višje na seznamu

Rangiranje strani po popularnosti



Google



PageRank algoritem

Larry Page in Sergey Brin
članek *The anatomy of large-scale hypertextual Web search engine*

Svetovni splet kot (usmerjen) graf

Spletne strani, vhodne in izhodne povezave

poveži eno stran z drugo stranjo

Spletni graf

vozlišča = spletne strani
usmerjene povezave = hiperpovezave

Kaj vpliva na oceno popularnosti strani?

Hiperpovezave kot glasovi na volitvah

stran i kaže na stran j $\Leftrightarrow$ obstaja usmerjena povezava od i do j

Več glasov večja popularnost

število vseh strani, ki kažejo na njo

Pomemben status glasovalca

popularnost strani, ki kaže na njo

Glas volilca se sorazmerno uteži glede na število oddanih glasov

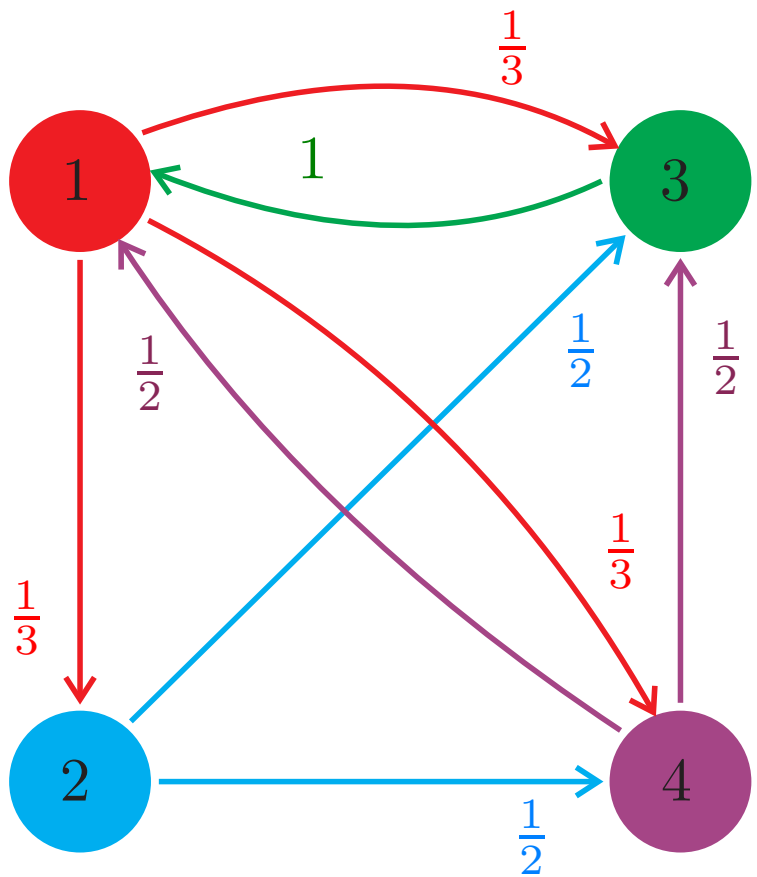
število izhodnih povezav strani, ki kaže na njo

Popularnost strani je določena s popularnostjo strani, ki kažejo na njo.

Štetje povezav

Popularnost strani je določena s popularnostjo strani, ki kažejo na njo:

$$
\begin{aligned}
P_1 &= P_3 + \frac{1}{2}P_4 \\
P_2 &= \frac{1}{3}P_1 \\
P_3 &= \frac{1}{3}P_1 + \frac{1}{2}P_2 + \frac{1}{2}P_4 \\
P_4 &= \frac{1}{3}P_1 + \frac{1}{2}P_2
\end{aligned}
$$



Ocena popularnosti, I

Popularnost strani je določena s popularnostjo strani, ki kažejo na njo:

$$P_1 = P_3 + \frac{1}{2}P_4$$

$$P_i = \sum_{j \in V_i} \frac{1}{N_j} P_j$$

pomembnost strani i

strani j, ki kažejo na stran i

število izhodnih povezav iz vozlišča j

pomembnost strani j

Ocena popularnosti, II

Prepišimo v primerno matematično obliko:

$$
\begin{aligned}
P_1 &= 0P_1 + 0P_2 + P_3 + \frac{1}{2}P_4 \\
P_2 &= \frac{1}{3}P_1 + 0P_2 + 0P_3 + 0P_4 \\
P_3 &= \frac{1}{3}P_1 + \frac{1}{2}P_2 + 0P_3 + \frac{1}{2}P_4 \\
P_4 &= \frac{1}{3}P_1 + \frac{1}{2}P_2 + 0P_3 + 0P_4
\end{aligned}
$$

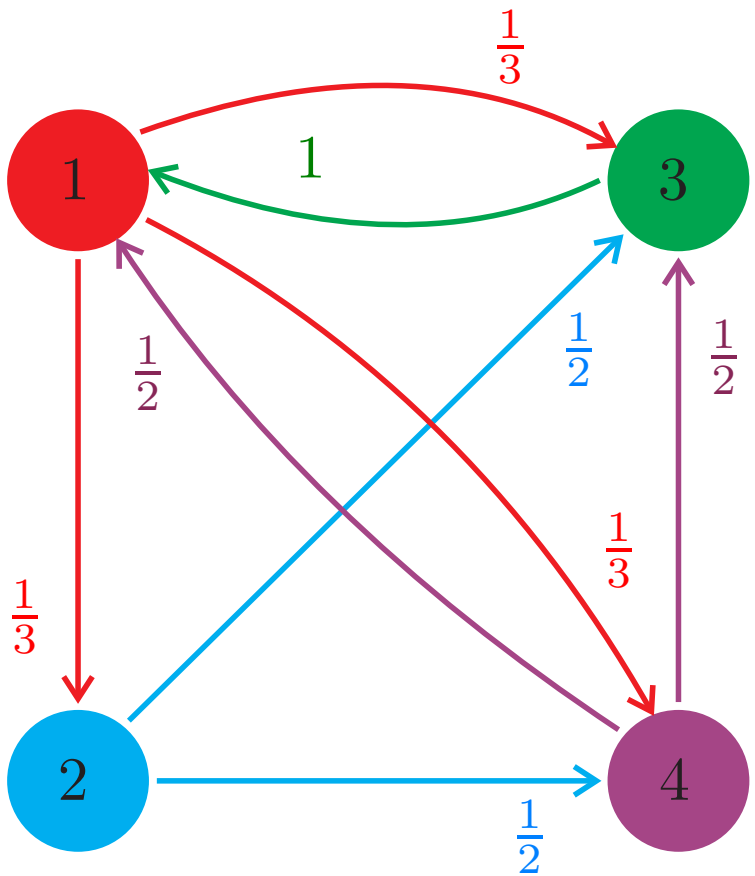
Ocena popularnosti kot lastni vektor matrike

In še malo bolj primerno obliko:

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{bmatrix}$$

ali

$$x = \mathbf{H}x$$



element j v stolpcu i = verjetnost prehoda iz vozlišča i v vozlišče j

Uporaba iterativne metode

Iščemo: x, da velja $x = \mathbf{H}x$

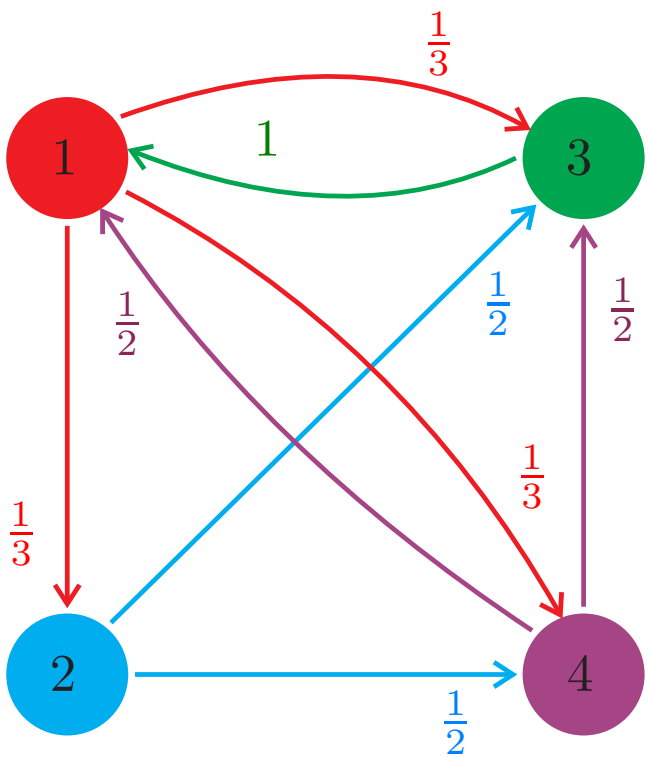
Potenčna metoda:

Vzemi začetni približek $x^{(0)}$

Ponavljaj

$$x^{(k+1)} := \mathbf{H}x^{(k)}$$

dokler „ne konvergira“



$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

$$x^{(0)} = \begin{bmatrix} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{bmatrix} \quad x^{(1)} = \begin{bmatrix} 0,37 \\ 0,08 \\ 0,33 \\ 0,20 \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} 0,43 \\ 0,12 \\ 0,27 \\ 0,16 \end{bmatrix}$$

$$x^{(3)} = \begin{bmatrix} 0,35 \\ 0,14 \\ 0,29 \\ 0,20 \end{bmatrix} \quad x^{(4)} = \begin{bmatrix} 0,39 \\ 0,11 \\ 0,29 \\ 0,19 \end{bmatrix} \quad x^{(5)} = \begin{bmatrix} 0,39 \\ 0,13 \\ 0,28 \\ 0,19 \end{bmatrix}$$

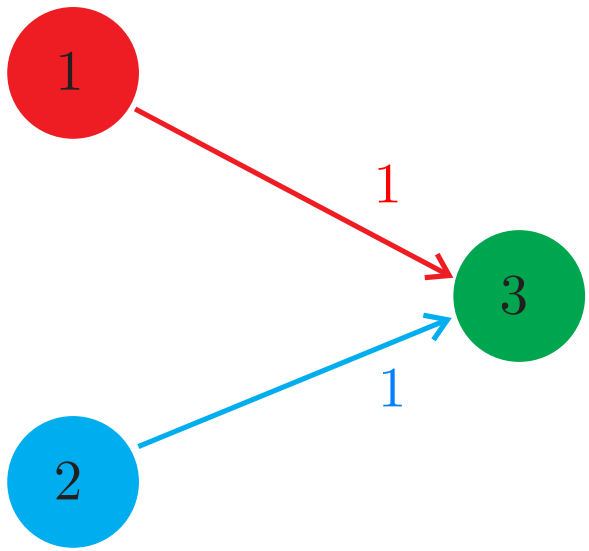
$$x^{(6)} = \begin{bmatrix} 0,38 \\ 0,13 \\ 0,29 \\ 0,19 \end{bmatrix} \quad x^{(7)} = \begin{bmatrix} 0,38 \\ 0,12 \\ 0,29 \\ 0,19 \end{bmatrix} \quad x^{(8)} = \begin{bmatrix} \mathbf{0,38} \\ \mathbf{0,12} \\ \mathbf{0,29} \\ \mathbf{0,19} \end{bmatrix}$$

Problemi pri reševanju z iterativno metodo, I

Ali konvergira k čemu smiselnemu v kontekstu ocene popularnosti?

Izolirana vozlišča brez izhodnih povezav (H ni stohastična):

pdf dokumenti, slike, tabele



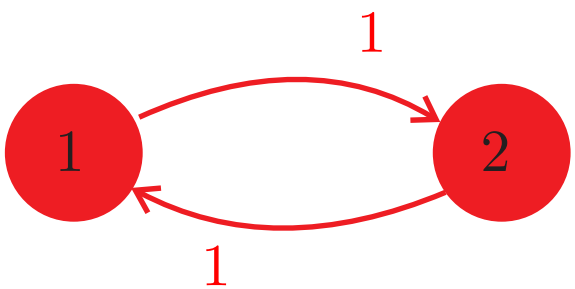
$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \quad x^{(0)} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \quad x^{(1)} = \begin{bmatrix} 0 \\ 0 \\ \frac{2}{3} \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

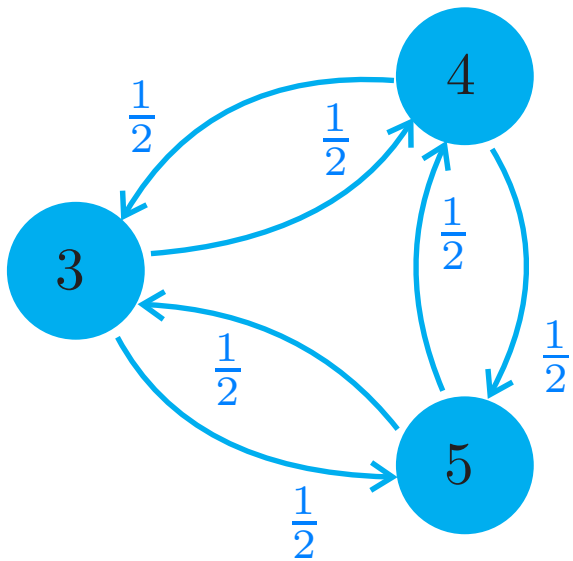
Problemi pri reševanju z iterativno metodo, II

Ali je rešitev enolična?

Graf ima več komponent (H je razcepna):

večina spletni strani kaže samo na prgišče ostalih strani





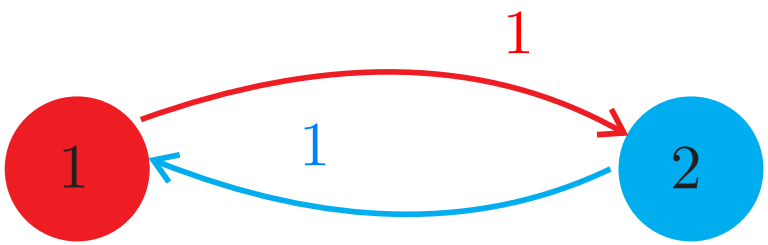
$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Problemi pri reševanju z iterativno metodo, III

Ali je konvergenca odvisna od izbire začetnega približka?

Cikli (H je periodična):

medsebojne združbe, ki glasujejo samo med sabo

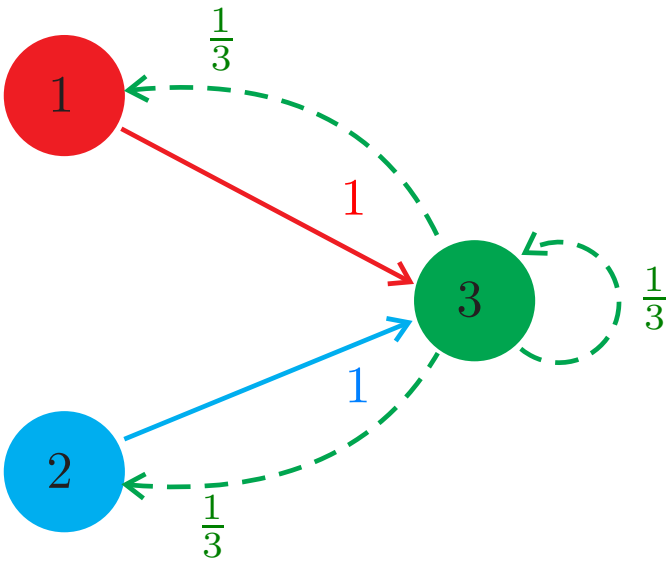


$$\mathbf{H} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$x^{(0)} = \begin{bmatrix} 0,5 \\ 0,5 \end{bmatrix} \quad x^{(1)} = \begin{bmatrix} \mathbf{0,5} \\ \mathbf{0,5} \end{bmatrix}$$

$$x^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad x^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Popravimo ničelne stolpce



postavi elemente 3. stolpca na $\frac{1}{3}$

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \\ 1 & 1 & \frac{1}{3} \end{bmatrix}$$

Zamenjaj matriko H z matriko S $= \mathbf{H} + \begin{bmatrix} \frac{a_1}{n} & \cdots & \frac{a_n}{n} \\ \vdots & \ddots & \vdots \\ \frac{a_1}{n} & \cdots & \frac{a_n}{n} \end{bmatrix}$

$a_i = 1$, če vozlišče i nima izhodnih povezav, in 0 sicer

Odpravimo še nepovezanost in cikle

Zamenjaj matriko S z matriko $\mathbf{G} = \alpha \mathbf{S} + (1-\alpha) \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$

„teleportacijski“ koeficient $0 \leq \alpha < 1$ (običajno $\alpha = 0,85$)

G je stohastična, nerazcepna in aperiodična

enoličen dominanten lastni vektor:

$$x = \mathbf{G}x, \qquad x \geq 0, \qquad \sum_{i=1}^{n} x_i = 1$$

x_i je ocena popularnosti za spletno stran i

G ima vse elemente neničelne

velikosti več deset milijard

Toda matrika H ima večino elementov ničelnih

$$x^{(k+1)} = \mathbf{G}x^{(k)} = \alpha\mathbf{H}x^{(k)} + \alpha \begin{bmatrix} \frac{a_1}{n} & \cdots & \frac{a_n}{n} \\ \vdots & \ddots & \vdots \\ \frac{a_1}{n} & \cdots & \frac{a_n}{n} \end{bmatrix} x^{(k)} + (1-\alpha) \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} x^{(k)}$$

Implementacija

niti **G** niti **S** nista eksplicitno shranjeni,
pri računanju se matrika **G** ne spreminja,
v praksi za izračun zadostuje 50-100 iteracij

Viri

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

A. Borodin, G.O. Roberts, J.S. Rosenthal, P. Tsaparas. Link Analysis Ranking Algorithms Theory And Experiments, 2004
http://cs-people.bu.edu/evimaria/cs565/lar.pdf

J. Khoury. How is it made? Google Search Engine, 2010.
http://aix1.uottawa.ca/ jkhoury/

A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, 2006.

R. S. Wills. Google's PageRank: The Math Behind the Search Engine, 2006.
http://www.cems.uvm.edu/ tlakoba/AppliedUGMath/other_Google/Wills.pdf

Hvala za pozornost!