Razvoj programske potpore za web i pokretne uređaje

- predavanja -

2021./2022.

13. Tražilice weba

Creative Commons

- slobodno smijete:
  - **dijeliti** — umnožavati, distribuirati i javnosti priopćavati djelo
  - **prerađivati** djelo
- pod sljedećim uvjetima:
  - **imenovanje:** morate priznati i označiti autorstvo djela na način kako je specificirao autor ili davatelj licence (ali ne način koji bi sugerirao da Vi ili Vaše korištenje njegova djela imate njegovu izravnu podršku).
  - **nekomercijalno:** ovo djelo ne smijete koristiti u komercijalne svrhe.
  - **dijeli pod istim uvjetima:** ako ovo djelo izmijenite, preoblikujete ili stvarate koristeći ga, preradu možete distribuirati samo pod licencom koja je ista ili slična ovoj.

U slučaju daljnjeg korištenja ili distribuiranja morate drugima jasno dati do znanja licencne uvjete ovog djela.
Od svakog od gornjih uvjeta moguće je odstupiti, ako dobijete dopuštenje nositelja autorskog prava.
Ništa u ovoj licenci ne narušava ili ograničava autorova moralna prava. *Tekst licence preuzet je s http://creativecommons.org/*

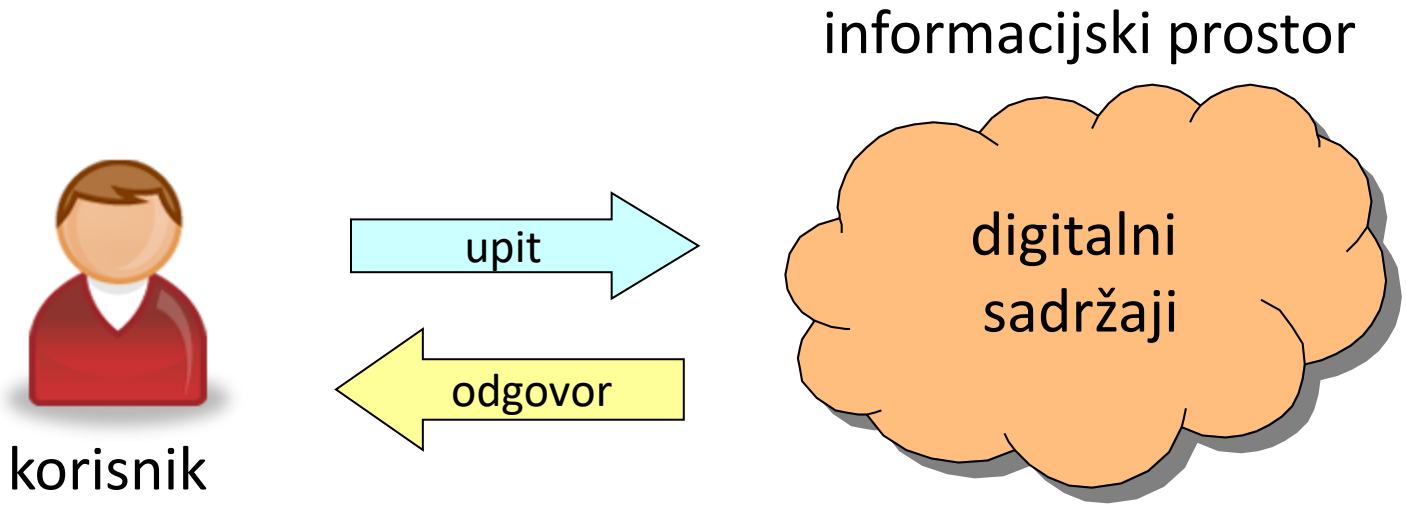
Sadržaj predavanja

- Pretraživanje digitalnog sadržaja
- Pretraživanje tekstualnog sadržaja (“klasični IR”)
  - Booleov model
  - vektorski prostorni model
- Pretraživanje i WWW
  - razlike u odnosu na “klasični IR”
  - arhitektura tražilice weba
  - rangiranje (PageRank)

Pretraživanje digitalnog sadržaja

engl. *information retrieval*

- pronaći dokumente iz informacijskog prostora koji zadovoljavaju informacijske potrebe korisnika (tj. relevantni su upitu kojim korisnik izražava svoje potrebe za informacijama)



Pojmovi

- informacijski prostor čini kolekcija dokumenata
- kolekcija je konačni skup višemedijskih dokumenata (npr. tekst, audio, video)
- upit je formalni iskaz koji definira korisnik, njime izražava svoje potrebe za informacijama prilikom pretraživanja
- odgovor je skup dokumenata koji sustav za pretraživanje nalazi relevantnim za neki upit
  - skup dokumenata je najčešće rangirana lista, prvi dokument je najrelevantniji
- Kada je dokument relevantan za dani upit?
  - kada zadovoljava korisničke potrebe za informacijama

Zadaće sustava za pretraživanje sadržaja

- generiranje strukturiranog prikaza dokumenata
  - izdvajanje značajnih svojstava iz dokumenata, npr. riječi iz teksta (rječnik) ili složeni postupci za video/audio
- generiranje strukturiranog prikaza upita iz korisničkog upita
- usporedba strukturiranog prikaza upita i dokumenata te generiranje odgovora
  - rangiranje dokumenata na temelju relevantnosti (*relevance*) za dani upit
  - sličnost (*similarity*) je mjera kojom se ocjenjuje relevantnost dokumenta za neki upit, uspoređuje sličnost dokumenta i upita

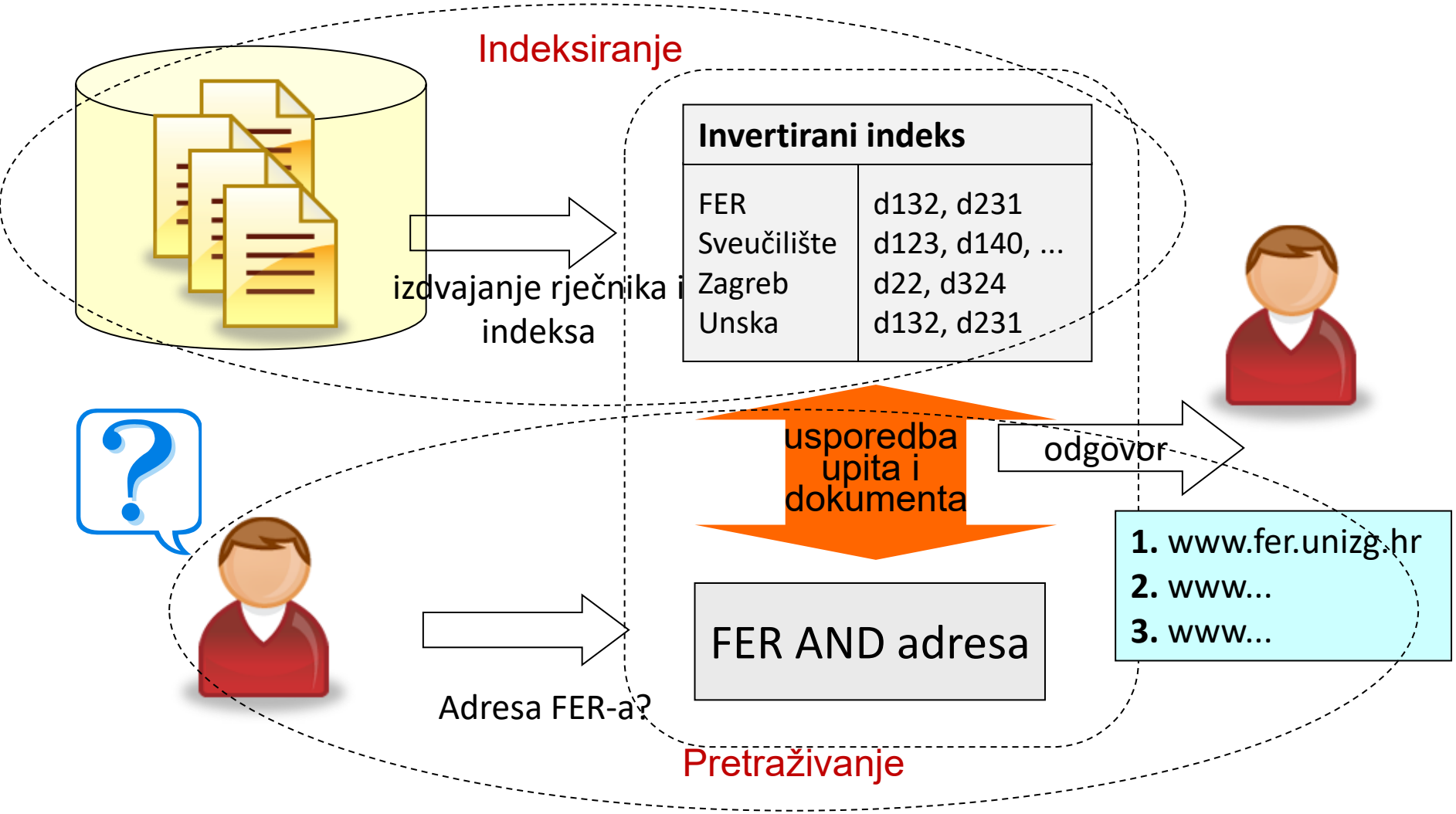
Sadržaj predavanja

- Pretraživanje digitalnog sadržaja
- Pretraživanje tekstualnog sadržaja (“klasični IR”)
  - Booleov model
  - vektorski prostorni model
- Pretraživanje i WWW
  - razlike u odnosu na “klasični IR”
  - arhitektura tražilice weba
  - rangiranje (PageRank)

Pretraživanje tekstualnog sadržaja

- potreba za informacijama izražava se najčešće u tekstualnom obliku
  - pretraživanje tekstualnih dokumenata u digitalnim knjižnicama
  - pretraživanje weba
- koriste se riječi iz dokumenata kao značajna svojstva za interpretaciju konteksta
  - značajno pojednostavljenje jer se npr. ignorira jezična gramatika, značenje riječi i slično
  - ovo pojednostavljenje se pokazalo uspješnim
  - dodatno se uzimaju u obzir poveznice među dokumentima za rangiranje u slučaju tražilica weba (primjer PageRank / Google)

Sustav za pretraživanje tekstualnog sadržaja



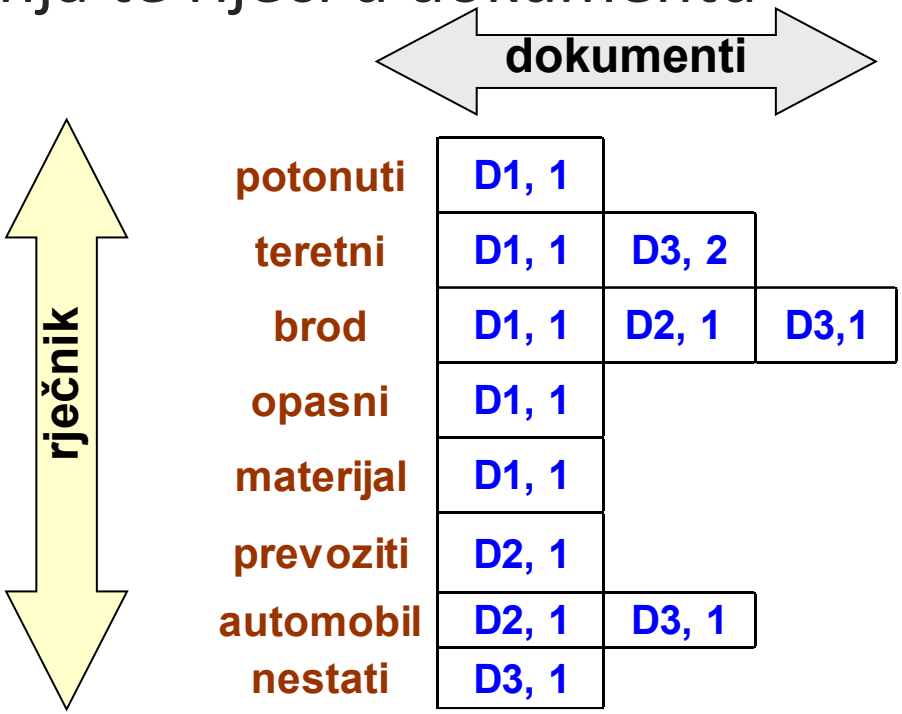
Pojmovi (1)

- indeksni termin (riječ) – ključna riječ ili grupa povezanih riječi koje imaju svoje značenje ili se pojavljuju u dokumentu
- rječnik – skup riječi koje se pojavljuju u tekstualnoj kolekciji
- upit – podskup riječi iz rječnika
- indeksiranje – izdvajanje rječnika i invertiranog indeksa iz kolekcije
- korjenovanje (*stemming*) - postupak svođenja različitih oblika neke riječi na njenu osnovu, kako bi se poboljšalo pretraživanje

Pojmovi (2)

Invertirani indeks

- povezuje svaku riječ iz rječnika s listom dokumenata u kojima se pojavljuje te s brojem pojavljivanja te riječi u dokumentu



Modeli za pretraživanje tekstualnog sadržaja

- cilj - pronaći podskup dokumenata koji su relevantni za dani upit i pridijeliti mjeru sličnosti dokumenta i upita
- model pretraživanja uključuje
  - strukturu prikaza dokumenta
  - strukturu prikaza upita
  - funkciju za usporedbu sličnosti upita i dokumenta
- kvaliteta modela ovisi o tome koliko dobro generirani odgovori zadovoljavaju korisničke potrebe za informacijama

Primjer

- Kolekcija od 3 dokumenta

  D1: Potonuo teretni brod s opasnim materijalom.
  D2: Brod prevozi automobile.
  D3: Nestao teretni automobil s teretnog broda.

- Upit

  Q: teretni AND brod AND (NOT automobil)

Matrica: riječi x dokumenti

	D1	D2	D3
potonuti	1	0	0
teretni	1	0	1
brod	1	1	1
opasni	1	0	0
materijal	1	0	0
prevoziti	0	1	0
automobil	0	1	1
nestati	0	0	1

1 - riječ se pojavljuje u dokumentu

0 - riječ se ne pojavljuje u dokumentu

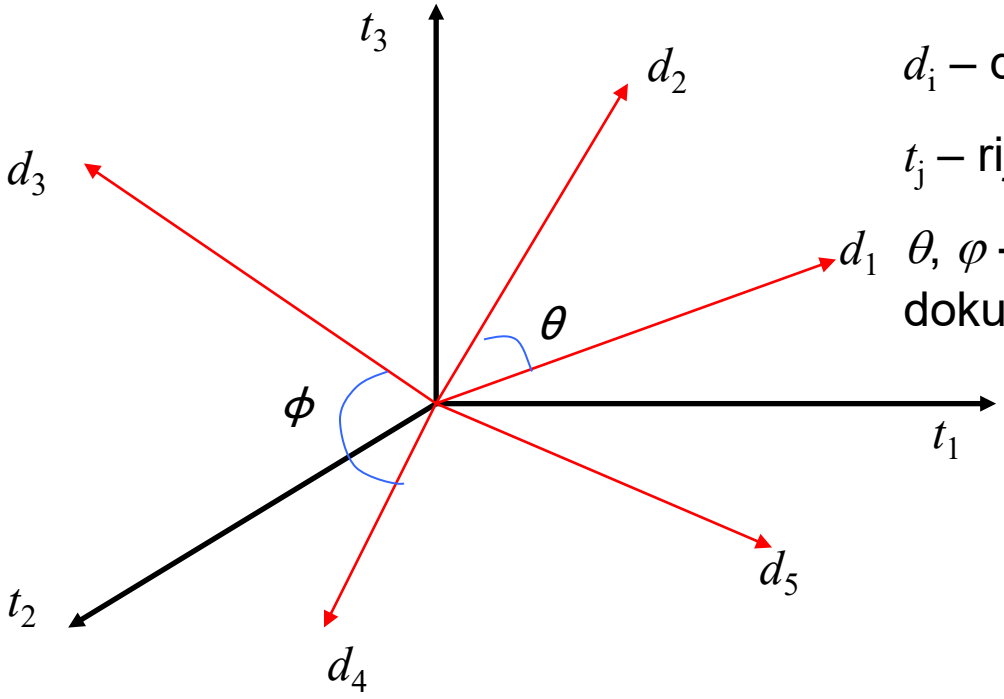
Q: teretni AND brod AND (NOT automobil) = 101 AND 111 AND (NOT 011) = 101 AND 100 = 100

Odgovor: **D1**

Booleov model

- prethodni primjer koristi Booleov model za pretraživanje koji se temelji na Booleovoj algebri
- dokument se promatra kao logička tvrdnja
  - 1 – riječ se pojavljuje u dokumentu
  - 0 – riječ se ne pojavljuje u dokumentu
- upit se formira kao Booleov izraz koristeći Booleove operatore (AND, OR, NOT)
  - dokument odgovara zadanom upitu samo onda kada su svi uvjeti upita ispunjeni
- nema rangiranja dokumenata
  - odgovor je skup dokumenata
  - dokument ili zadovoljava upit ili ne (nema rangiranja vezano uz relevantnost dokumenta za zadani upit)

Vektorski prostorni model



d_i – dokument iz kolekcije

t_j – riječ iz rječnika

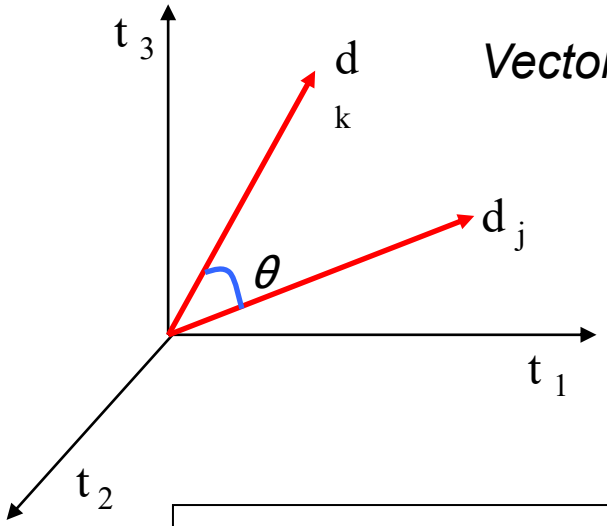
θ, φ - kut među dokumentima

Primjer 3-dimenzionalnog vektorskog prostora

Pretpostavka: Dokumenti koji su "bliže" u vektorskom prostoru semantički su slični ("govore o sličnim stvarima").

Rangiranje dokumenata

- Za rangiranje dokumenata u odgovoru na upit koristi se mjera *sličnosti* dokumenta i upita
- sličnost dokumenata d_j i d_k računa se kao kosinus kuta među njihovim vektorima



Vector space model

$$sim(d_j, d_k) = \cos(\theta) = \frac{\vec{d}_j \bullet \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|}$$

$$sim(d_j, d_k) = \frac{\sum_{i=1}^{m} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{m} w_{i,j}^2} \sqrt{\sum_{i=1}^{m} w_{i,k}^2}}$$

vektori dokumenata d_j i d_k

$\vec{d}_j = (w_{1j}, w_{2j}, \ldots w_{mj}), w_{ij} > 0 \text{ if } t_i \in d_j$

$\vec{d}_k = (w_{1k}, w_{2k}, \ldots w_{mk}), w_{ik} > 0 \text{ if } t_i \in d_k$

w_{ij} je težinski faktor vezan uz riječ t_i u dokumentu d_j

Upit se razmatra kao kratki dokument!

Težinski faktor

Kako odrediti težinski faktor w_{ij} vezan uz riječ t_i?

- težinski faktor w_{ij} vezan uz riječ t_i određuje se najčešće kao tf x idf

$$w_{ij} = tf(i,j) \cdot idf(i) = tf(i,j) \cdot \log\left(\frac{N}{df_i}\right)$$

- *$tf\ (i, j)$ – term frequency*
  - broj pojavljivanja riječi t_i u dokumentu d_j
- *$idf\ (i)$ – inverse document frequency*
  - N – veličina kolekcije (broj dokumenata)
  - df_i – broj dokumenata kolekcije u kojima se pojavljuje t_i

Vektorski prostorni model (primjer)

- neka imamo zadan upit Q i kolekciju dokumenata koja se sastoji od dokumenta D1,D2 i D3. Upit i dokumenti definirani su kao:

  Q: teretni automobil **(upit)**

  D1: Potonuo teretni brod s opasnim materijalom.

  D2: Brod prevozi automobile.

  D3: Nestao teretni automobil s teretnog broda.

- broj dokumenata u kolekciji N=3
- ako je riječ pojavljuje u samo jednom dokumentu idf=log(3/1)=0,477
- ako se riječ pojavljuje u dva dokumenta idf=log(3/2)=0,176
- ako se riječ pojavljuje u svim dokumentima idf=log(3/3)=0

Vektorski prostorni model (primjer)

- računamo za svaku riječ koja se pojavljuje bilo u upitu ili u dokumentu inverznu frekvenciju *idf*

	D1	D2	D3	Q
potonuti	0,477	0	0	0
teretni	0,176	0	0,176	0,176
brod	0	0	0	0
opasni	0,477	0	0	0
materijal	0,477	0	0	0
prevoziti	0	0,477	0	0
automobil	0	0,176	0,176	0,176
nestati	0	0	0,477	0

Q: Preuzeti vrijednost za riječi iz upita, ostale riječi = 0

Vektorski prostorni model (primjer)

- računamo za svaku riječ težinski faktor w_{ij}

	D1	D2	D3	Q
potonuti	0,477	0	0	0
teretni	0,176	0	0,352	0,176
brod	0	0	0	0
opasni	0,477	0	0	0
materijal	0,477	0	0	0
prevoziti	0	0,477	0	0
automobil	0	0,176	0,176	0,176
nestati	0	0	0,477	0

Riječ teretni se pojavljuje 2 puta u D3.

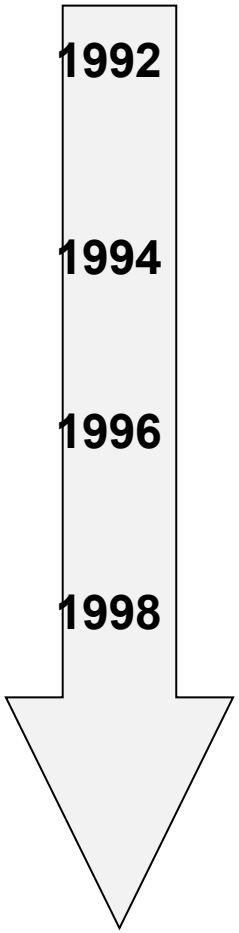
Rezultat: **1. sim(Q,D3) = 0,6037**
2. sim(Q,D2) = 0,2448
3. sim(Q,D1) = 0,1473

Veći iznos sim(Q,D) znači manji kut između Q i D!

Sadržaj predavanja

- Pretraživanje digitalnog sadržaja
- Pretraživanje tekstualnog sadržaja (“klasični IR”)
  - Booleov model
  - vektorski prostorni model
- Pretraživanje i WWW
  - razlike u odnosu na “klasični IR”
  - arhitektura tražilice weba
  - rangiranje (PageRank)

Malo povijesti...



Počeci weba

preglednici

Imenici

- Yahoo

Prve tražilice

- InfoSeek, Lycos, Altavista, Excite, Inktomi, …

Preporod web tražilica

- Google

Razlike u odnosu na “klasični IR”

- Kolekcija
  - veličina, dinamične promjene dokumenata
  - velike razlike u kvaliteti dokumenata
  - velika količina “duplikata”
  - velika količina sadržaja na webu nije indeksirana (*deep Web*)
- Korisnici
  - postavljaju kratke upite (najčešće 2 do 3 riječi)
  - neprecizno definirane potrebe za informacijama

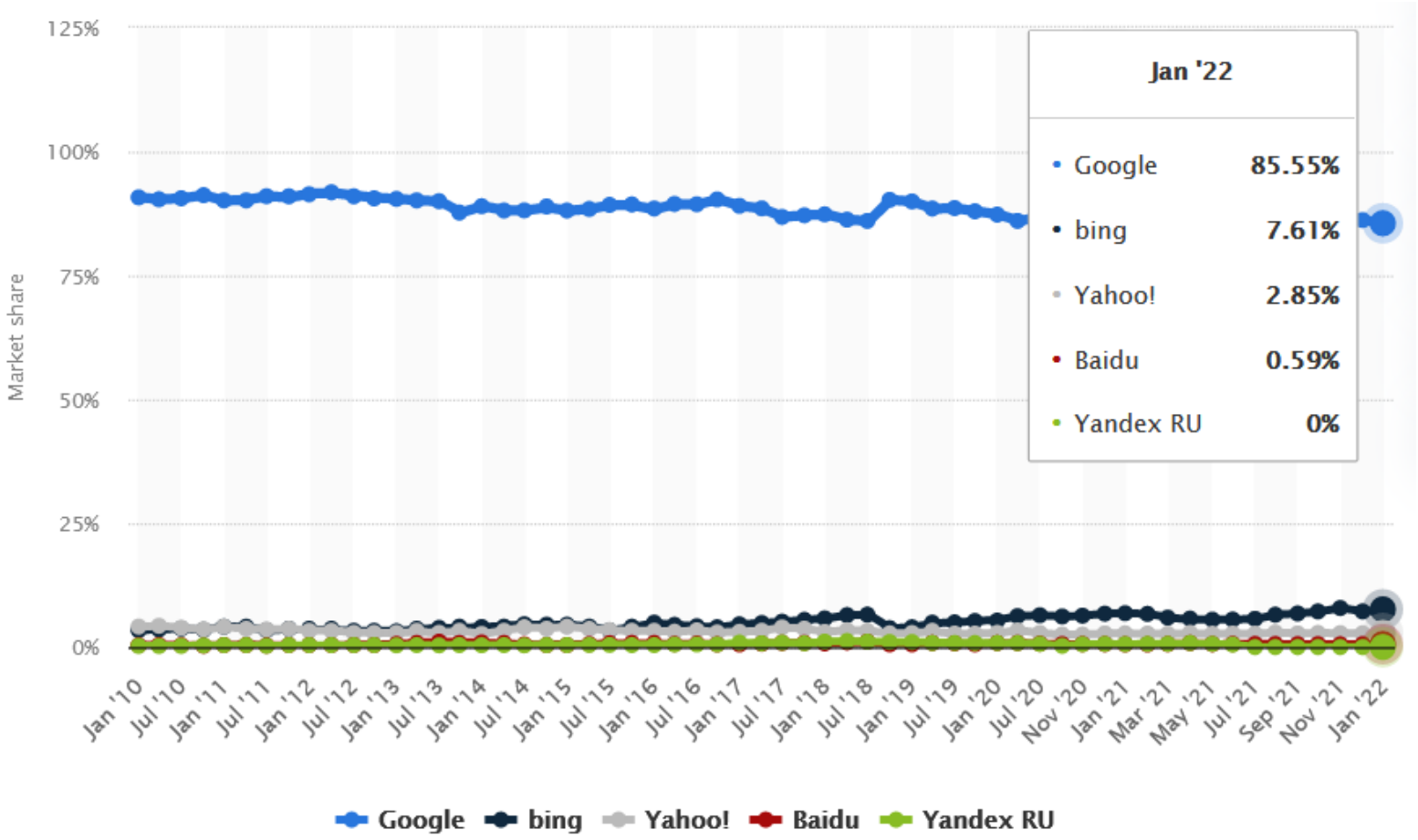
Zahtjevi

- jednostavno korisničko sučelje
- kratko vrijeme odziva
- korisnici definiraju kratke upite (2 do 3 riječi), koriste česte riječi u upitima
- rangiranje rezultata je iznimno važno
  - većina korisnika ne gleda rezultate nakon prve stranice s odgovorima, tražilica weba je optimizirana da korisnici pronađu odgovore već među prvih 10
- Search Engine Optimization (SEO): „*the process of making your site better for search engines*”, https://developers.google.com/search/docs/beginner/seo-starter-guide

Definiranje upita

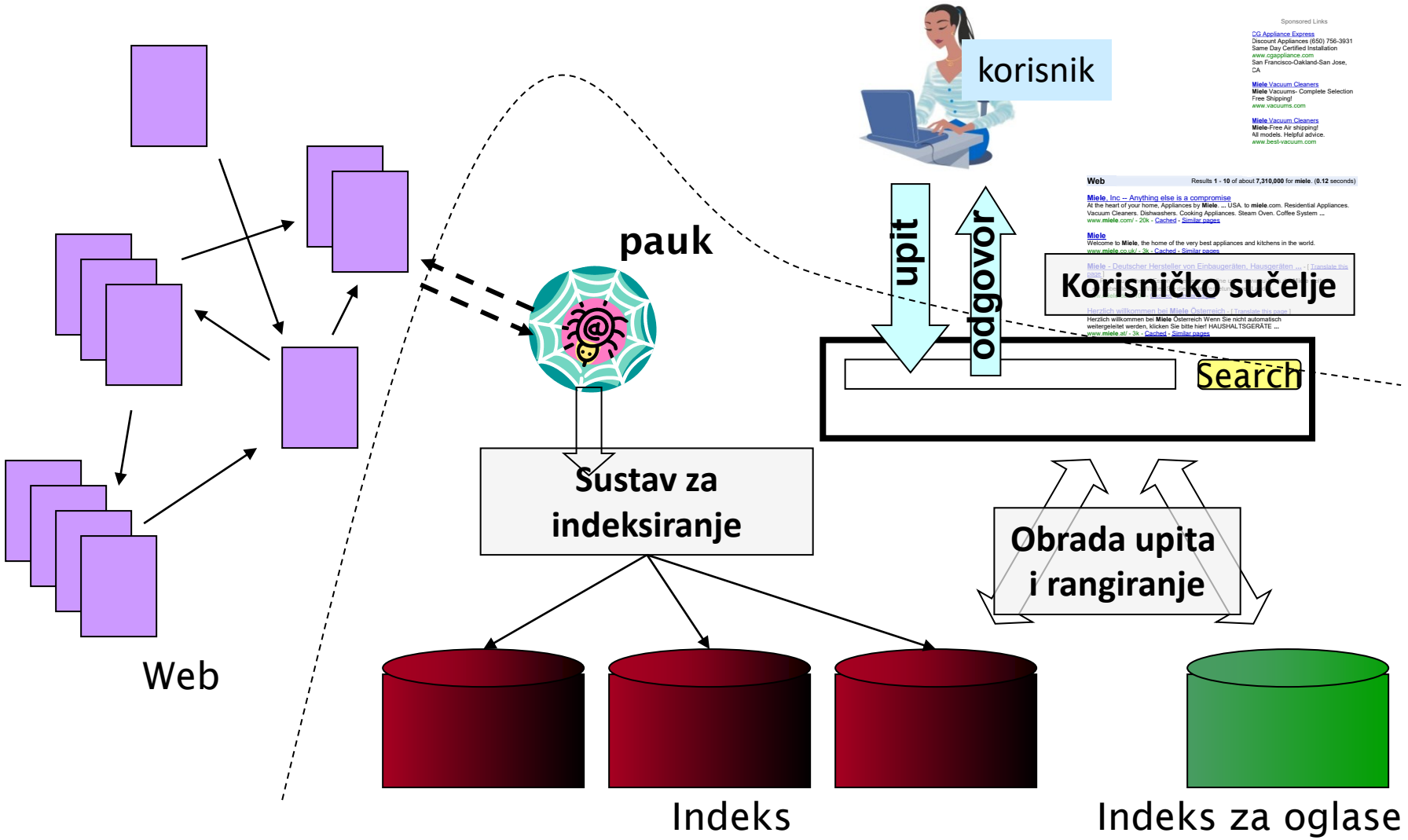
- uporaba malih i velikih slova (ignoriraju se velika slova)
  New York Times == new york times
- uporaba fraza
  "adresa fakulteta elektrotehnike i računarstva"
  "NASA space shuttle program"
- uporaba logičkih operatora (OR, NOT)
  San Francisco Giants 2004 OR 2005
  jaguar NOT car – Bing
- kontrola ključnih riječi (+, -) – Google, Bing
  +film +noir -"pinot noir"
  +python –monty
- kontrola resursa
  adresa FER site:fer.unizg.hr

Popularnost tražilica weba (globalni udio na tržištu, desktop)



Izvor: Statista

Arhitektura tražilice weba



Dijelovi tražilice

- Pauk (engl. *spider*, *crawler*)
  - program koji stvara kolekciju tako što posjećuje poznate web stranice, analizira sadržaj stranice te prati ugrađene poveznice
  - slika weba koju danas pretražujemo poznatim tražilicama stara je oko mjesec dana
  - Što je s vijestima, blogovima?
- Sustav za indeksiranje
  - kreira raspodijeljeni invertirani indeks
- Sustav za obradu upita i rangiranje
  - implementira model za pretraživanje
  - česte riječi se ignoriraju, a ostale svode na korijenski oblik (*stemming*)

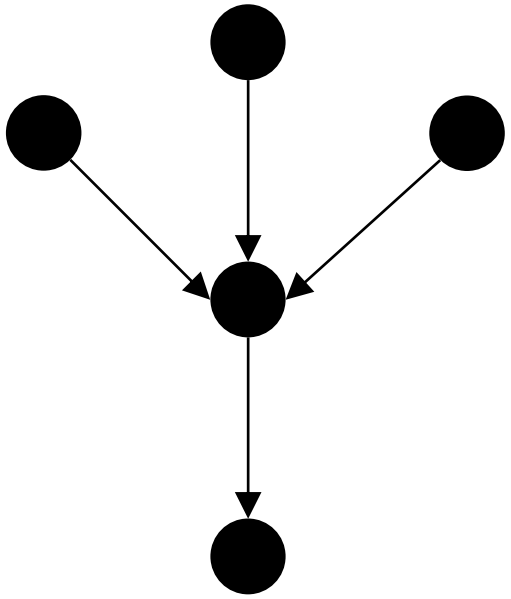
Podsjetimo se…

- WWW kao globalni višemedijski informacijski sustav
- Informacijski prostor weba čine
  - *informacijski izvori* ili ***resursi*** (*resource*): upravo smo vidjeli kako ćemo pretraživati tekstne resurse
  - međusobno povezani **poveznicama** (*hyperlink*): možemo li njih kako iskoristiti za pronalaženje relevantnih odgovora?

Rangiranje: kako poboljšati rezultate upita?

PageRank

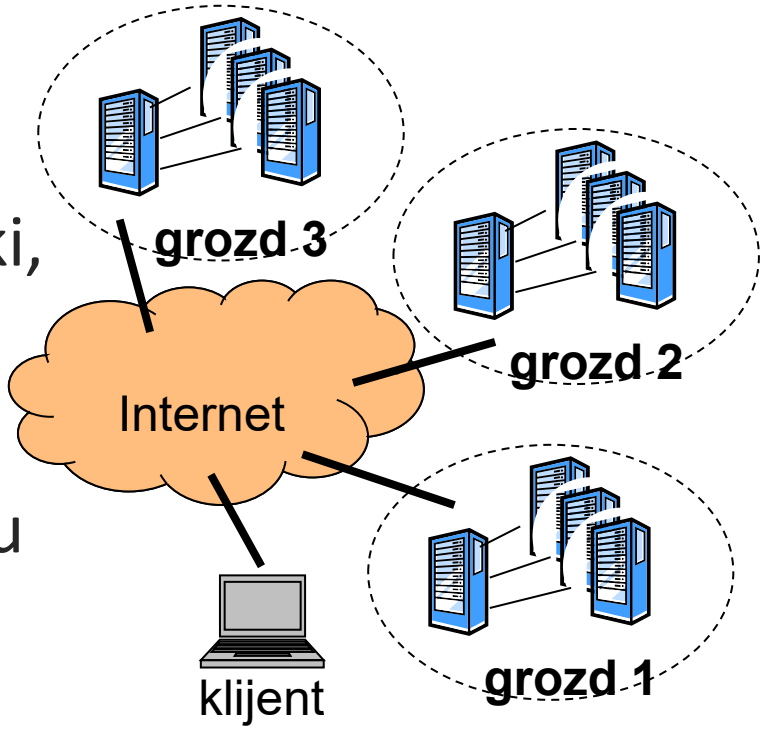
- algoritam koji je učinio Google najpopularnijom tražilicom
- modelira web usmjerenim grafom
- koristi ulazne i izlazne poveznice radi rangiranja relevantnih stranica s obzirom na njihovu popularnost
- neovisan o upitu



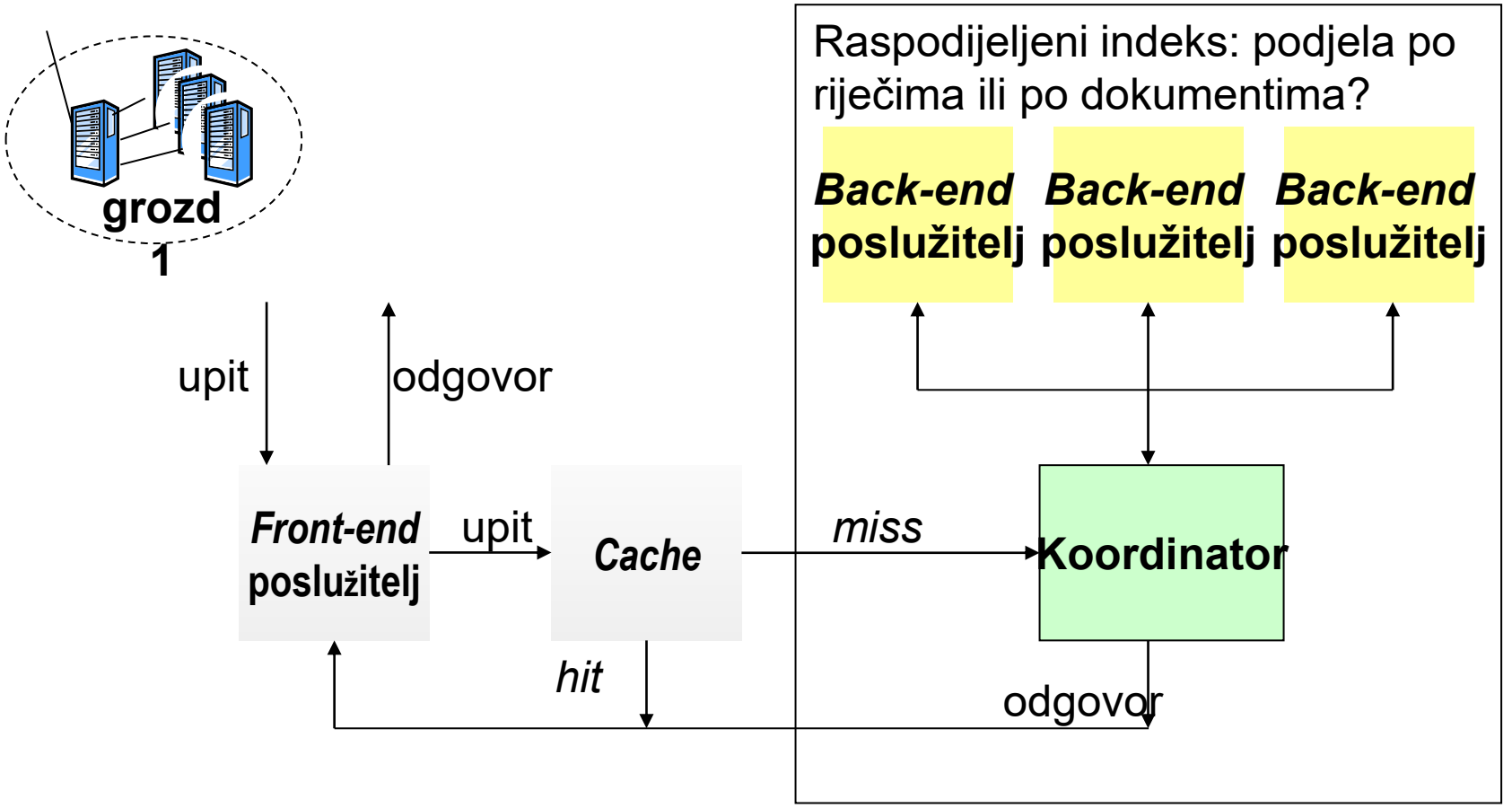
Ako vektorski model rangira 2 stranice jednako, PageRank će dati prioritet popularnijoj stranici, tj. stranica s više ulaznih poveznica dobiva viši PageRank pogotovo ako te “dolazne” stranice imaju veliki PageRank.

Kako poboljšati performance pretraživanja?

- Vrijeme odziva je ključni parametar kvalitete za svaku web tražilicu!
- Tražilica i za ovu svrhu koristi raspodijeljenu arhitekturu
- Indeksiranje se obavlja periodički, obnavlja se indeks na čvorovima za posluživanje
- Tražilica je replicirana više puta u različitim podatkovnim centrima



Arhitektura tražilice u grozdu računala



Pitanja za učenje i ponavljanje

1. Objasnite pojam relevantnosti dokumenta za korisnički upit.
2. Skicirajte i objasnite komponente tražilice weba.
3. Objasnite kako se računa sličnost između dva dokumenta ako se koristi vektorski prostorni model.
4. Objasnite težinske faktore *tf* i *idf*. Kako se primjenjuju u vektorskom prostornom modelu za pretraživanje tekstualnog sadržaja?
5. Objasnite strukturu invertiranog indeksa u donjoj tablici. Navedite veličinu rječnika i kolekcije dokumenata.
   Koji dokumenti se mogu pojaviti o odgovoru na upit „a AND c AND d“ ako se koristi
   a) Booleov model za pretraživanje tekstualnog sadržaja,
   b) vektorski prostorni model za pretraživanje tekstualnog sadržaja.

a	$d_1, d_3, d_4, d_8, d_{10}, d_{15}$
b	$d_2, d_7, d_8, d_{10}, d_{11}, d_{13}, d_{14}$
c	$d_1, d_3, d_4, d_5, d_6, d_9, d_{10}$
d	$d_3, , d_4, d_6, d_{12}$

Za više informacija

- Kolegij Umrežene višemedijske usluge
  - 5. semestar preddiplomskog studija
- Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto: Modern Information Retrieval - the concepts and technology behind search, Second edition Pearson Education Ltd., Harlow, England 2011
- In-depth guide to how Google Search works
- I vratimo se na početni upitnik