Diplomski studij

Informacijska i komunikacijska tehnologija:

Telekomunikacije i informatika

Računarstvo

Programsko inženjerstvo i informacijski sustavi
Računarska znanost

Ak.god. 2009./2010.

Raspodijeljeni sustavi

12.

Raspodijeljeno pretraživanje informacija

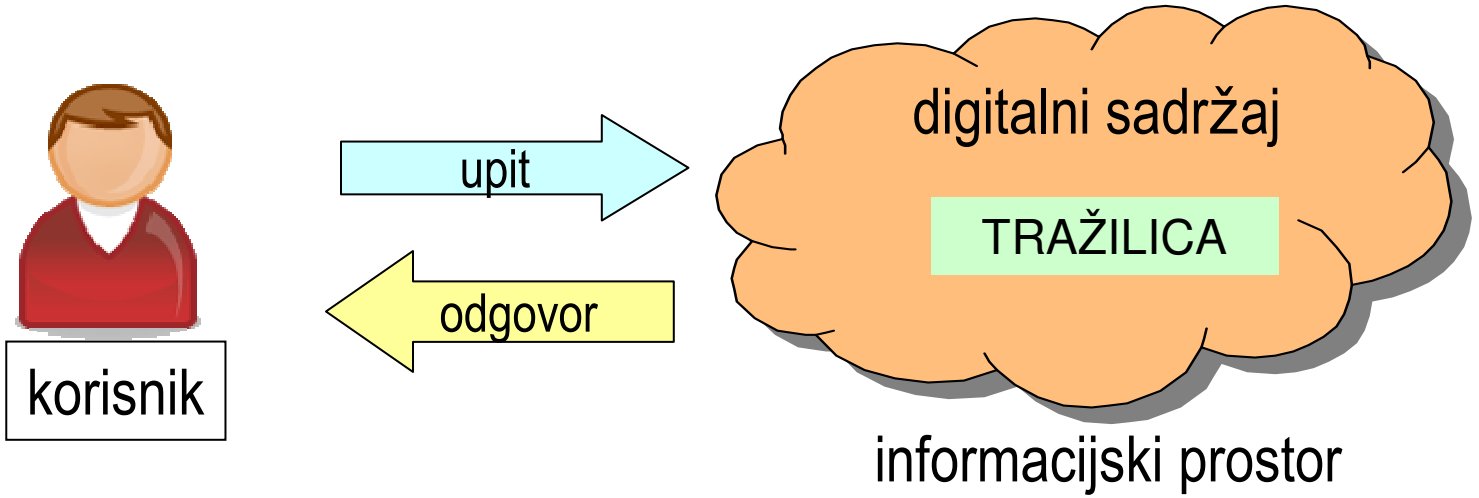
Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

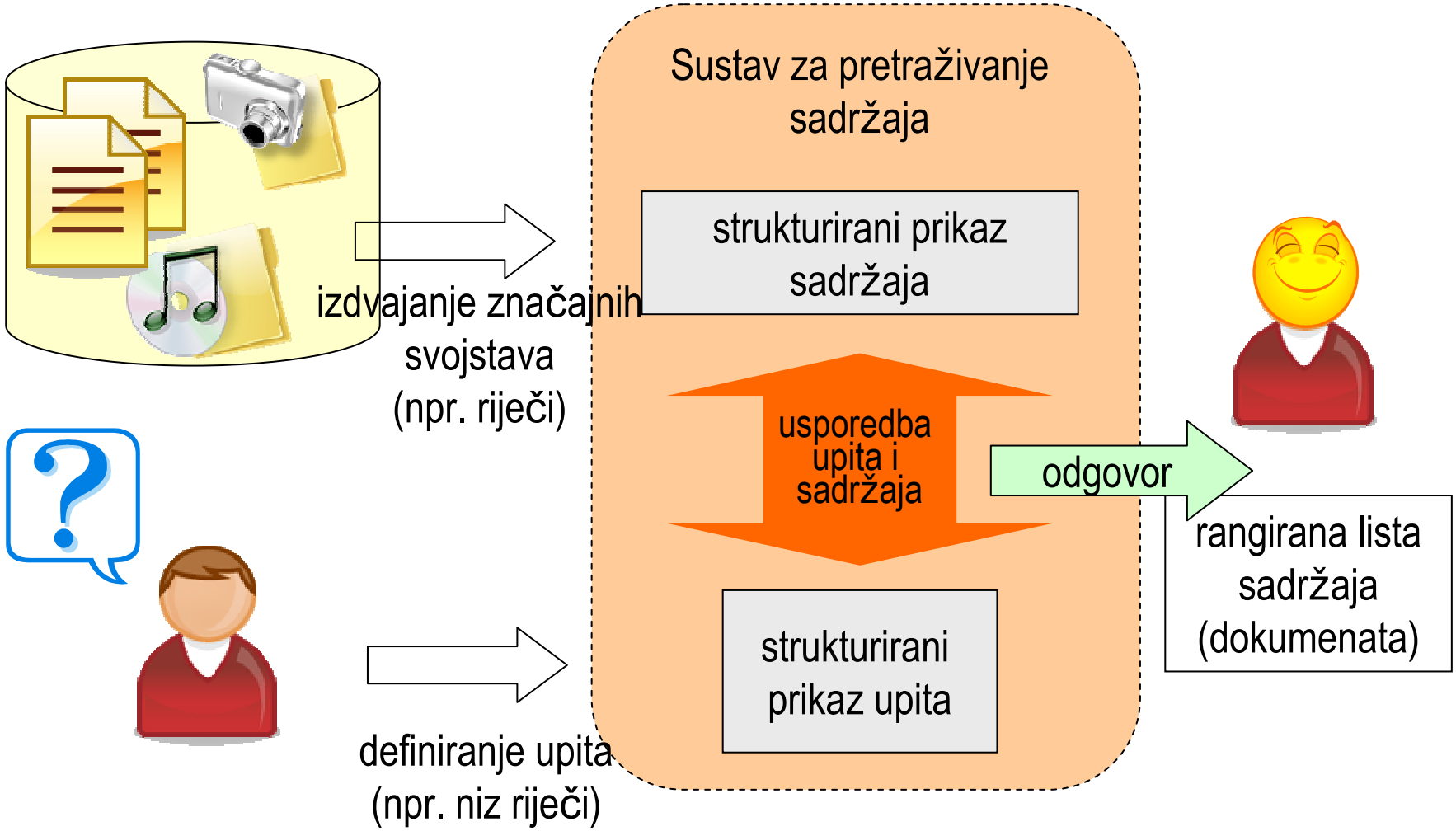
Pretraživanje sadržaja

PODSJETIMO SE: Višemedijske usluge

- pronaći sadržaj iz informacijskog prostora koji zadovoljava informacijske potrebe korisnika
- zadaća tražilice: pronaći sadržaj koji je relevantan za korisnički upit



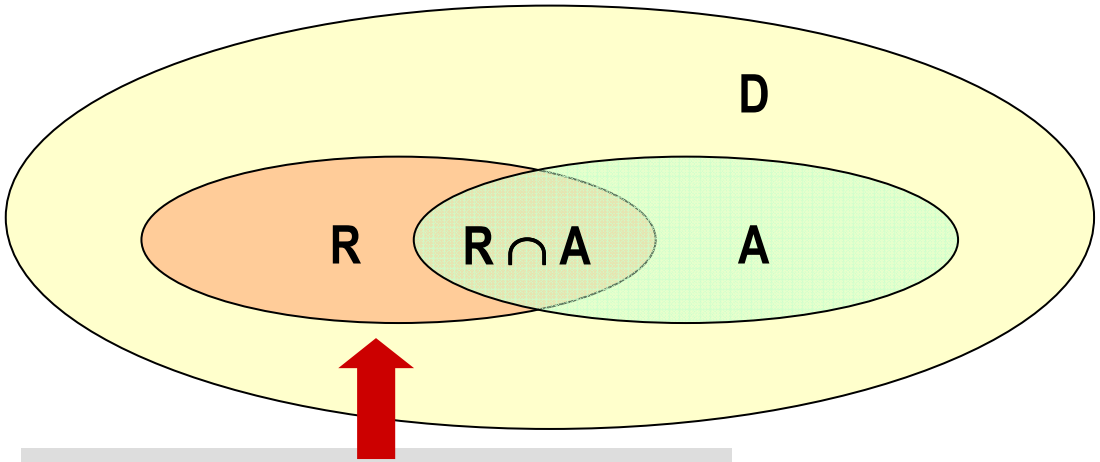
Sustav za pretraživanje sadržaja (tražilica)



Objašnjenje pojmova

- 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

Ocjena kvalitete tražilice



Idealna tražilica pronalazi sve dokumente iz skupa R!

D – kolekcija dokumenata

R – skup relevantnih dokumenata

A – skup dokumenata iz odgovora

R ∩ A – relevantni dokumenti iz odgovora

- dokument iz kolekcije je relevantan ili nije relevantan za neki upit
- Kako odlučiti koji su dokumenti iz kolekcije relevantni za neki upit?
  - jedino korisnik (ekspert) može odlučiti o relevantnosti dokumenta za neki upit
- cilj: povećati R ∩ A, idealno A = R

Ocjena kvalitete tražilice: odziv i preciznost

- Odziv (engl. *recall*)
  - postotak relevantnih dokumenata iz odgovora u odnosu na ukupni broj relevantnih dokumenata u kolekciji

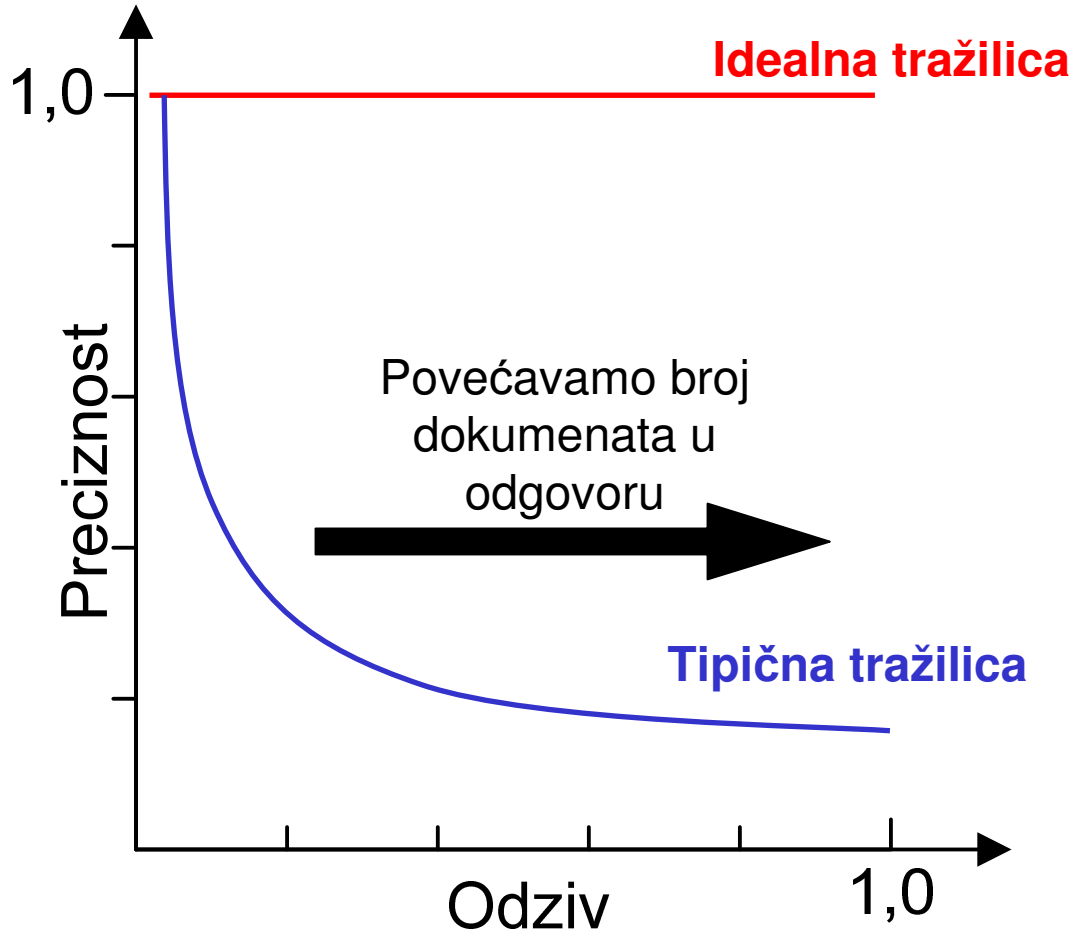
$$Recall = \frac{|A \cap R|}{|R|}$$

- Preciznost (engl. *precision*)
  - postotak relevantnih dokumenata iz odgovora u odnosu na ukupni broj dokumenata u odgovoru

$$Precision = \frac{|A \cap R|}{|A|}$$

- Odziv i preciznost se obično računa za "top-k" rezultata iz odgovora (precision@k, recall@k)
- Za manji broj dokumenata u odgovoru se pretpostavlja bolja preciznost, a time manji odziv
- Odziv će uvijek biti 1 ako su u odgovoru svi dokumenti iz kolekcije
- Idealna tražilica ima preciznost = 1

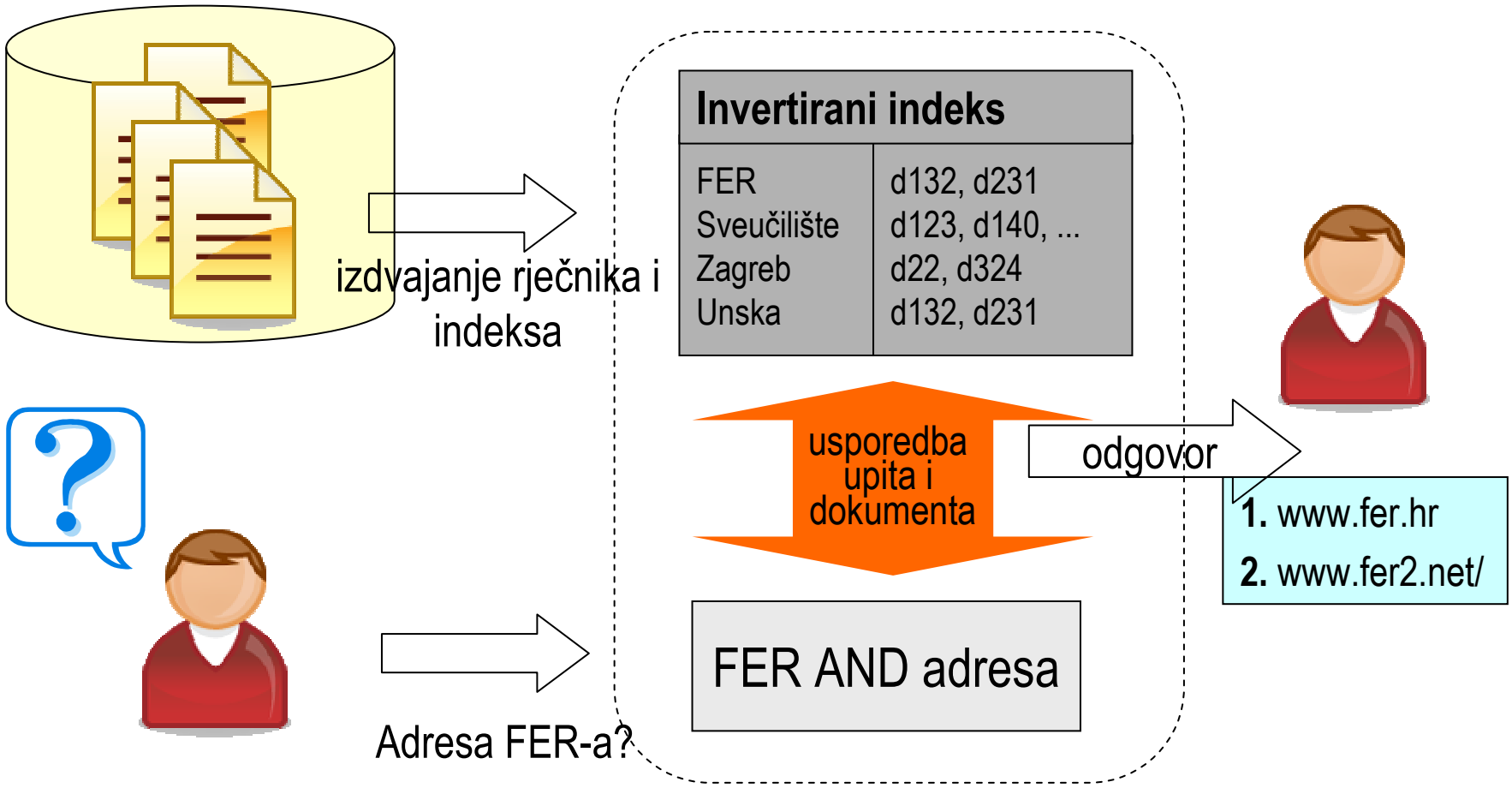
Primjer grafa preciznost/odziv



Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

Sustav za pretraživanje tekstualne kolekcije



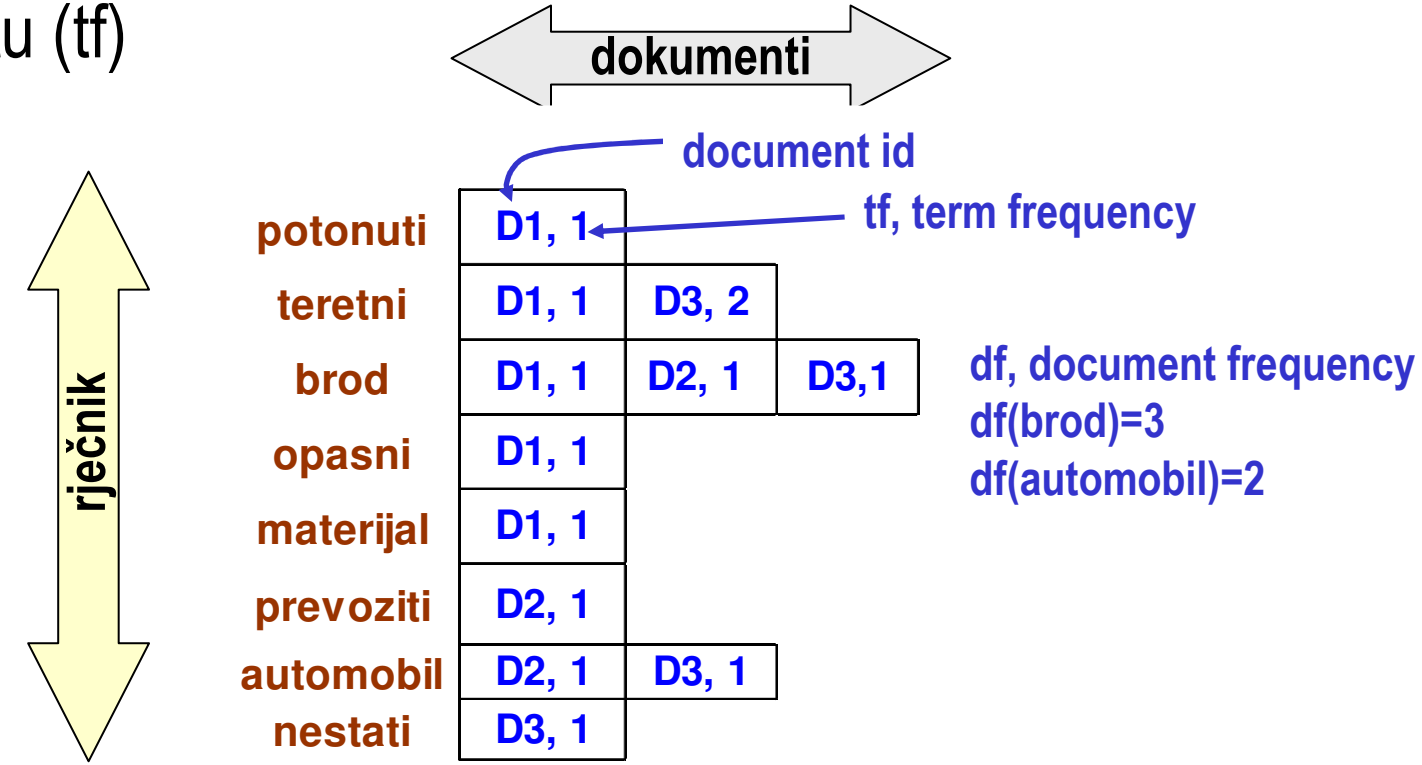
Pojmovi (1)

- 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

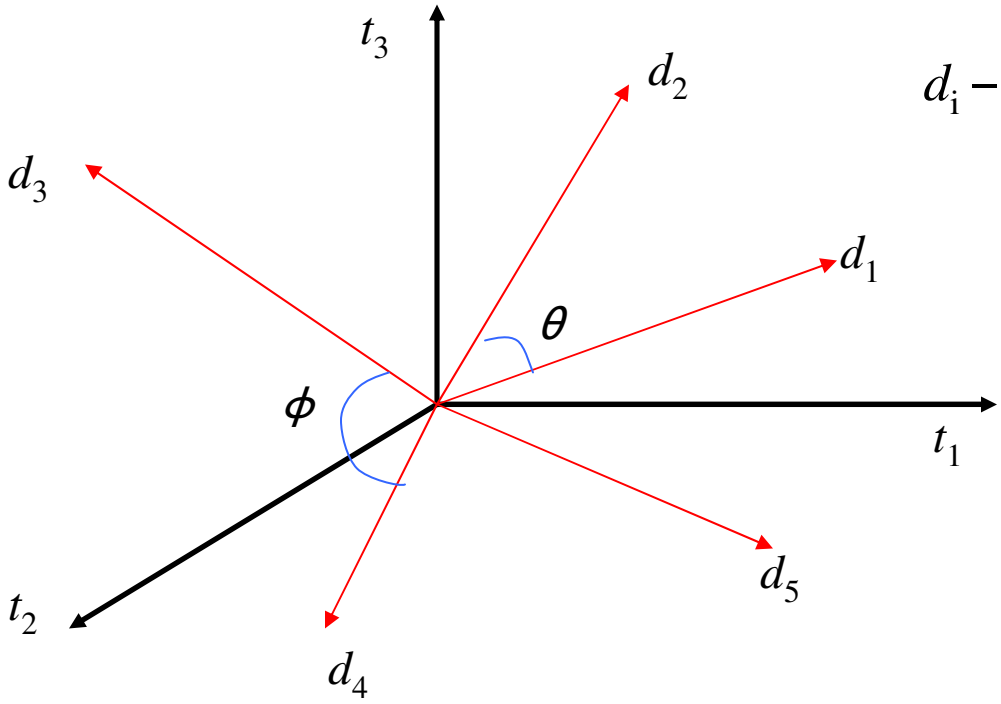
Pojmovi (2)

Invertirani indeks

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



Vektorski prostorni model



d_i – dokument iz kolekcije

t_j – riječ iz rječnika

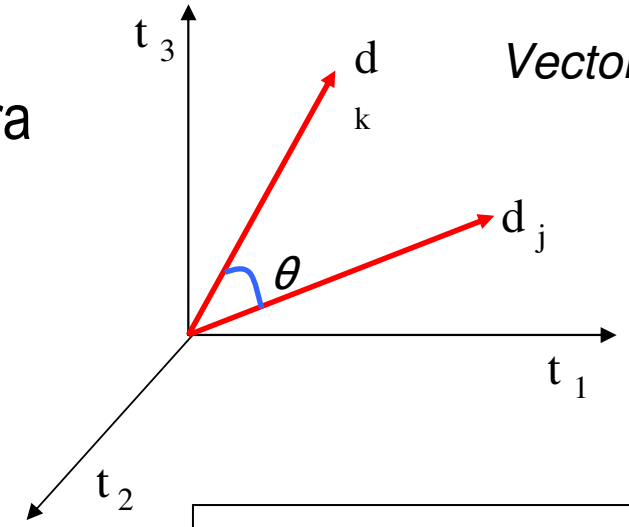
θ, φ - kut među dokumentima

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 d=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!

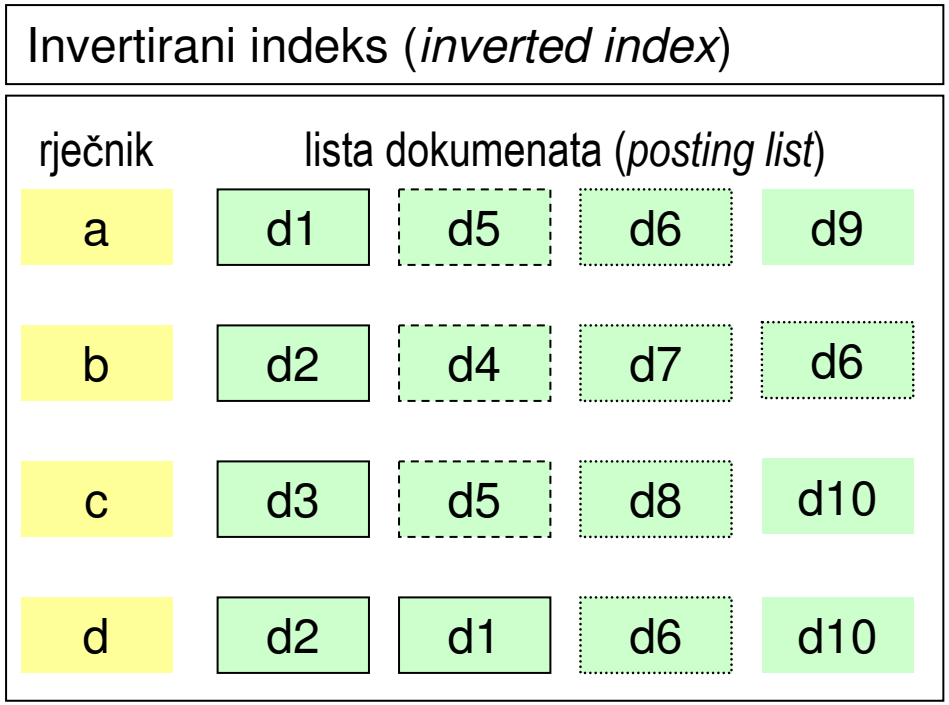
Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

Raspodiljene tražilice

- Okruženje
  - velika kolekcija dokumenata, veliki broj korisnika, visoki intenzitet korisničkih upita
- Raspodijeljenost
  - kolekcija i/ili indeks je raspodijeljen na veći broj računala
  - dijeljenje indeksa prema dokumentima ili riječima
- Replikacija
  - replikacija cjelokupnog sustava
  - u skladu s intenzitetom korisničkih upita, zadovoljavajuće vrijeme odziva

Dijeljenje indeksa prema dokumentima



Dijeljenje indeksa prema dokumentima, svaki čvor zadužen za particiju dokumenata

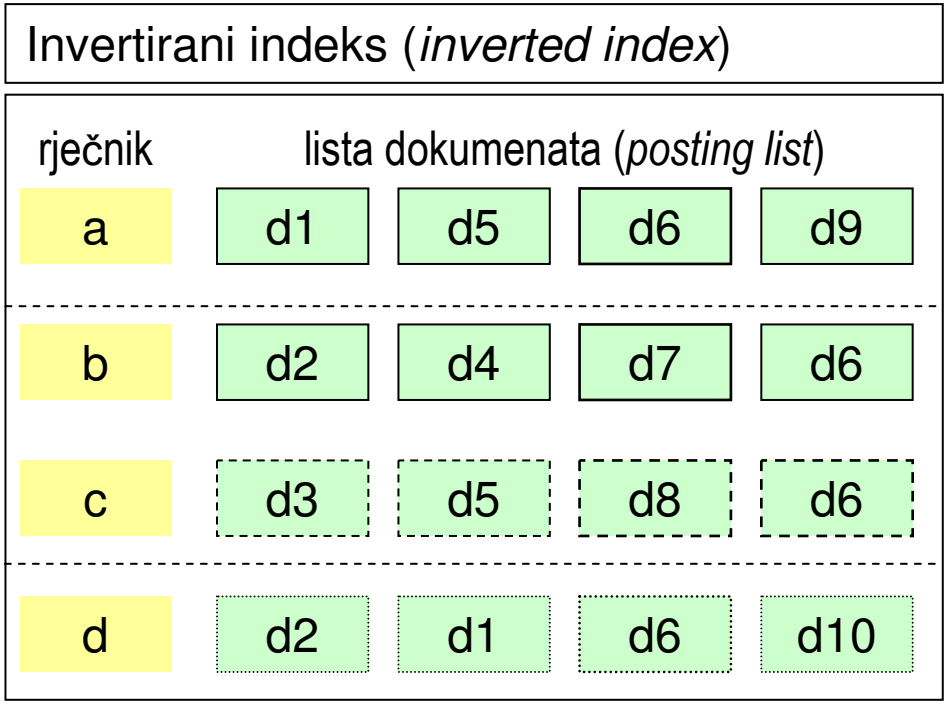
čvor 1: d1, d2, d3

čvor 2: d4, d5

čvor 3: d6, d7,d8

čvor 4: d9, d10

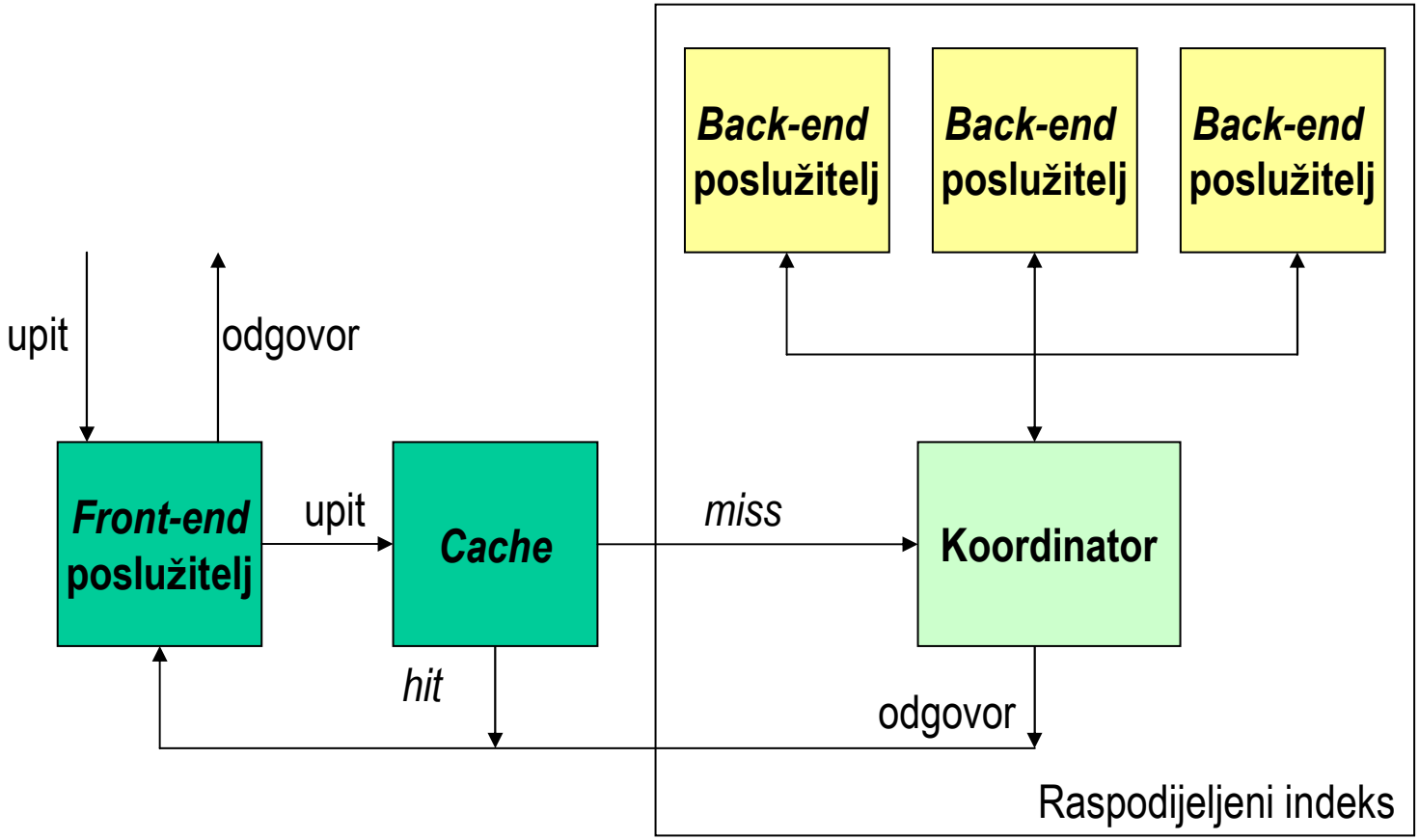
Dijeljenje indeksa prema riječima



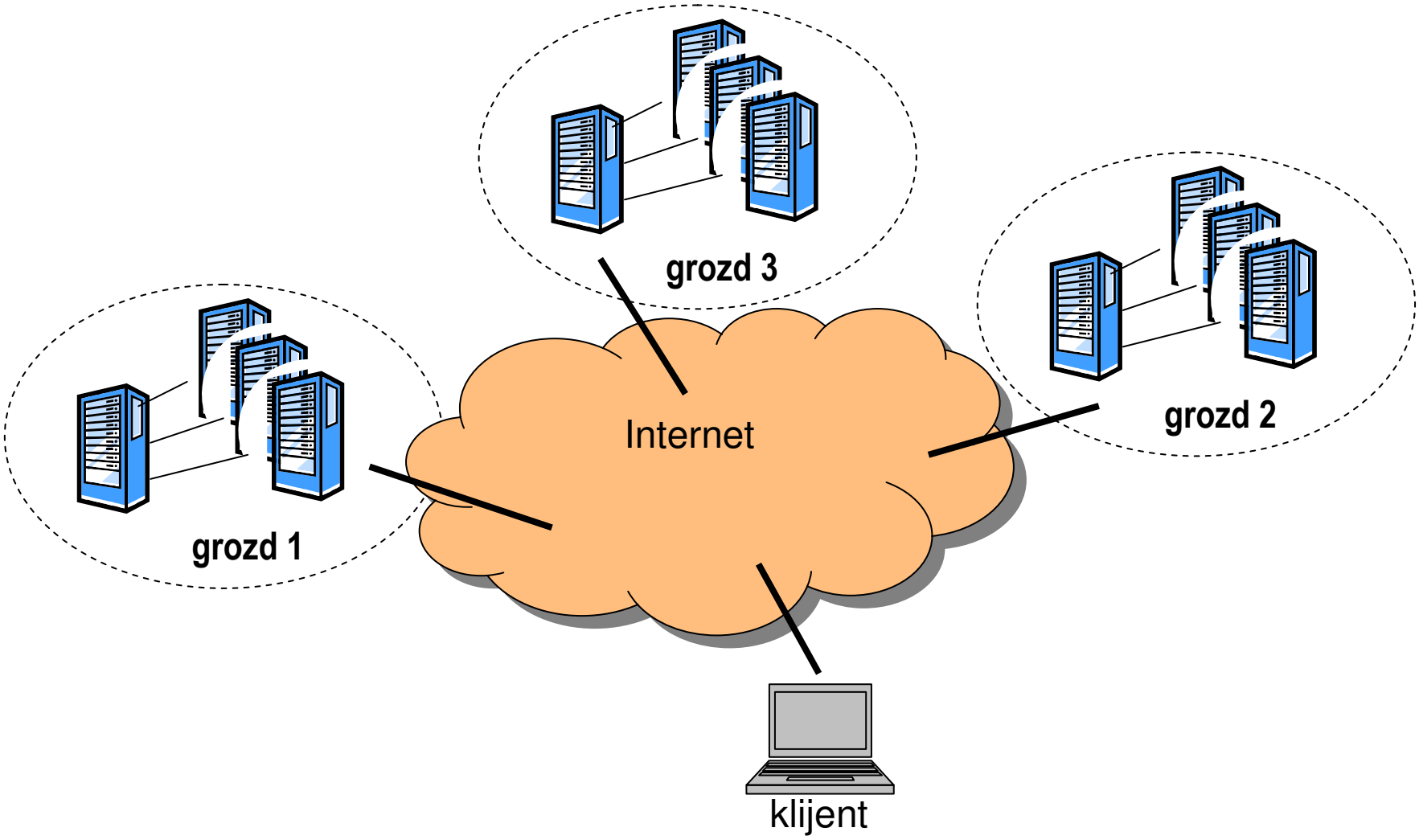
Dijeljenje indeksa prema riječima iz rječnika, svaki čvor zadužen na skup riječi iz rječnika

čvor 1: a
čvor 2: b, c
čvor 3: d

Arhitektura tražilice u grozdu računala



Arhitektura tražilice u spletu računala

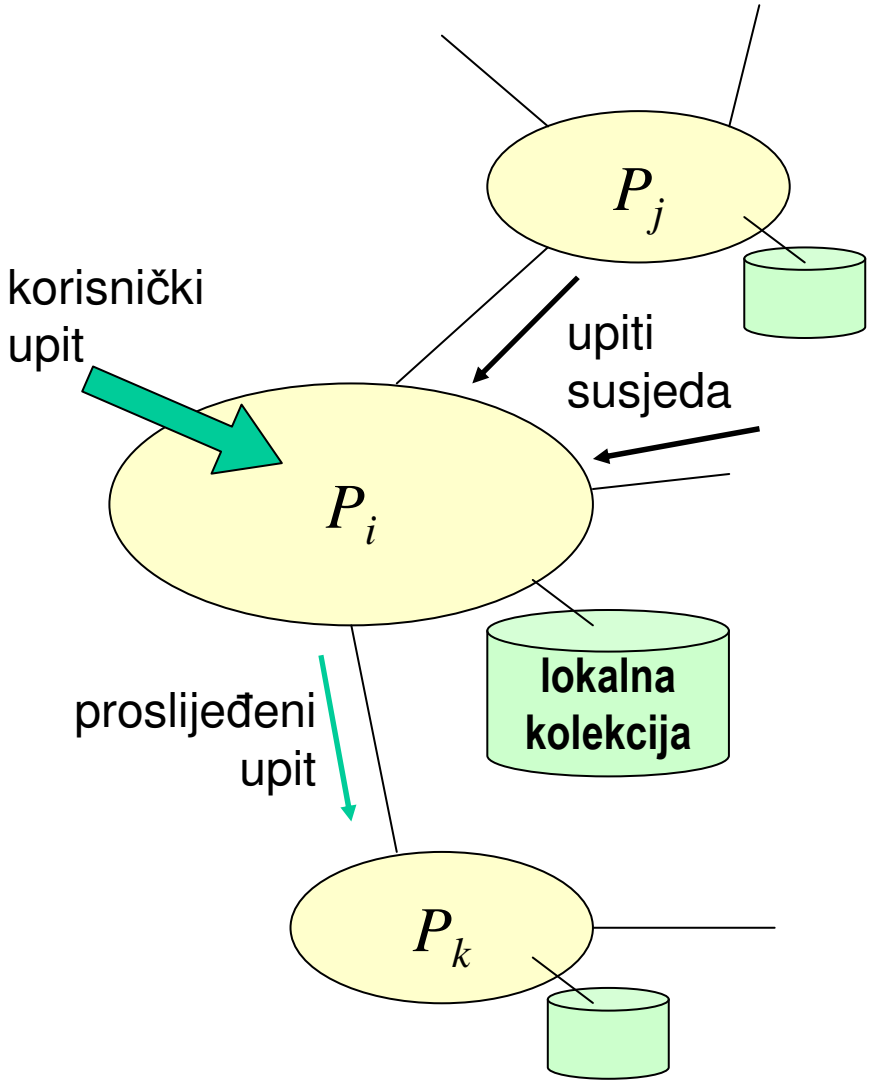


Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

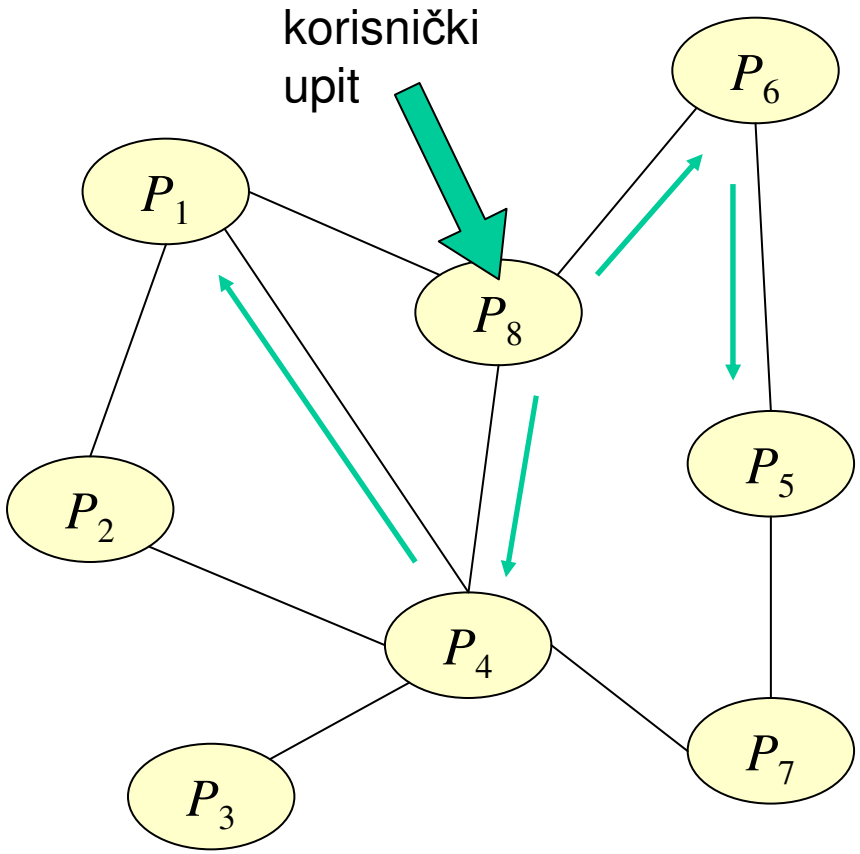
Pretraživanje u mrežama P2P

- Mreže P2P inherentno podržavaju pretraživanje
- Svaki peer pohranjuje, indeksira i dijeli sadržaj (lokalna kolekcija dokumenata)
- Svaki peer implementira funkciju pretraživanja
  - prima korisničke upite, ali i upite susjednih peerova
  - generira odgovor na primljeni upit na temelju lokalne i/ili globalne kolekcije dokumenata (unija svih lokalnih kolekcija), prosljeđuje upit susjedima



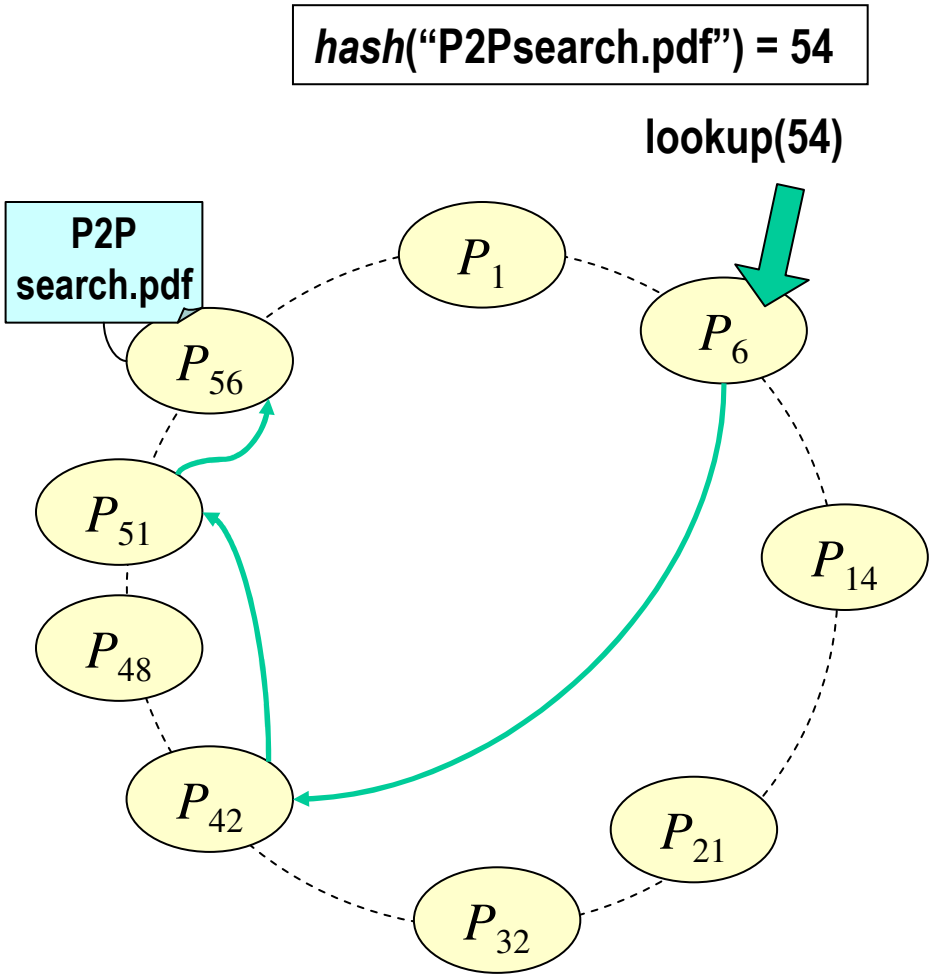
Pretraživanje u nestrukturiranim mrežama P2P

- podržavaju proizvoljnu strukturu upita
  - u skladu s funkcionalnošću tražilice peera
  - primjeri upita: niz riječi, booleov izraz, regularni izraz, semantički upit
- problem: usmjeravanje upita u mreži peerova i pronalaženje peerova koji mogu dati kvalitetan odgovor na upit



Pretraživanje u strukturiranim mrežama P2P

- raspodijeljena hash tablica
  - povezuje vrijednost atributa (*hash* kod) i sadržaj (npr. ime datoteke)
- podržava samo jednostavne upite (*exact-match queries*)
- skalabilno pretraživanje, no ograničena funkcionalnost



Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

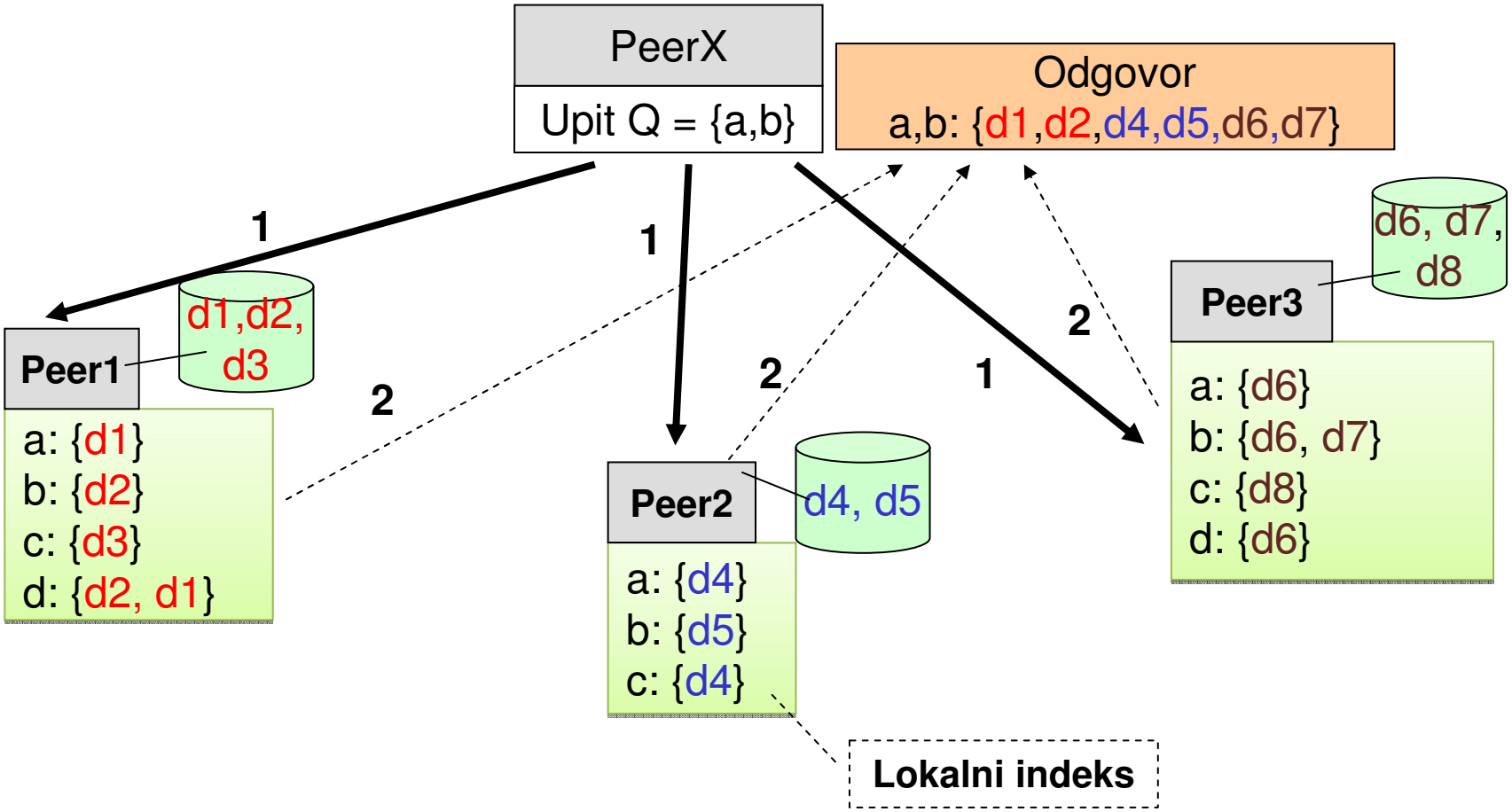
Motivacija

- Postojeće tražilice u grozdu/spletu računala dostižu gornju granicu u smislu skalabilnosti
  - Npr. $20*10^9$ web stranica ≈ 100 terabyte tekstualnih dokumenata
    - za održavanje indeksa (25 terabyte) treba oko 3.000 računala u 1 grozdu računala (x *c* radi redundancije i zadovoljavajućeg vremena odziva)
    - 10.000 upita u sekundi ⟶ *c*=10, treba ukupno 30.000 računala
  - količina sadržaja na webu eksponencijalno raste
  - današnje tražilice indeksiraju mali postotak dostupnog sadržaja
  - postoji velika količina privatnog sadržaja (npr. NASA) koje su zatvorene za tražilice kao Google ili Yahoo
  - problem indeksiranja sadržaja koji se dinamički mijenja (npr. vijesti, blog)

P2P mreže kao web tražilice?

- Ideja: koristiti mreže P2P za izgradnju web tražilice ili specijalizirane tražilice za određeno znanstveno područje
- Izgraditi sustav bez velikih početnih ulaganja
- Sustav može potencijalno obuhvatiti milijune čvorova, svaki čvor doprinosi vlastite dokumente u kolekciju, ali i resurse računala
- Onemogućuje zlouporabu rezultata pretraživanja
- Omogućuje organizaciju društvene mreže za pretraživanje i dijeljenje znanja
- Kako organizirati i koristiti postojeće mreže P2P kao web tražilice?

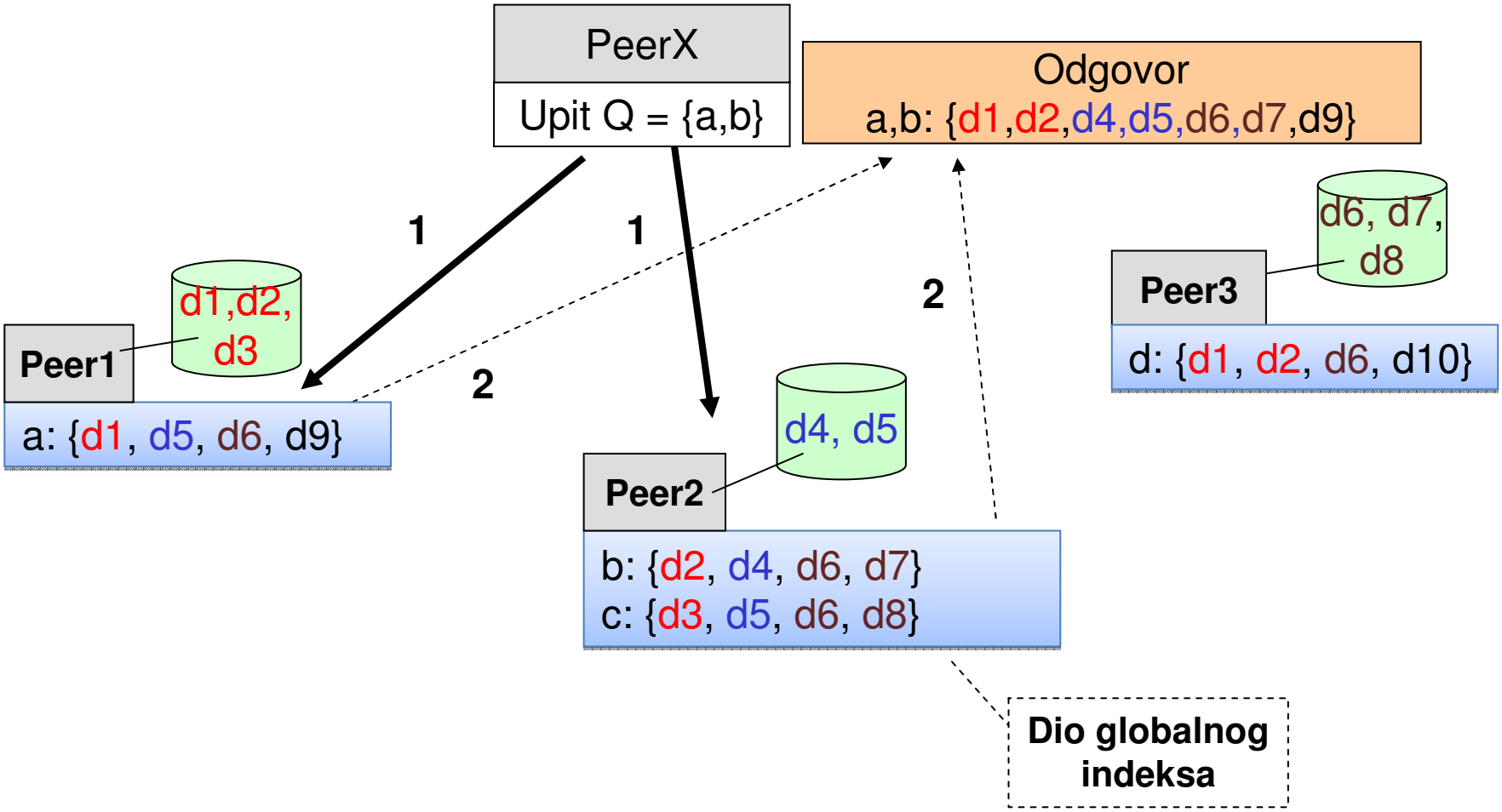
Dijeljenje indeksa prema dokumentima (1)



Dijeljenje indeksa prema dokumentima (2)

- pogodna organizacija indeksa za nestrukturirane mreže P2P
  - svaki čvor obrađuje upit neovisno o ostalim čvorovima kao samostalna tražilica nad lokalnom kolekcijom dokumenata
- glavni nedostatak: upit mora procesirati svaki čvor u mreži ili dio čvorova (kako pronaći “kvalitetne” čvorove i relevantne dokumente?)
  - neskalabilno rješenje zbog broja generiranih zahtjeva tijekom obrade upita, raste s $O(n)$ gdje je n broj čvorova
  - veličina poruke koja prenosi odgovor je relativno mala (samo skup dokumenata koji čine odgovor iz lokalne kolekcije)
- jednostavno je održavanje informacije o lokalnim dokumentima, ali ne i o globalnoj kolekciji
  - nemoguće je izračunati globalni tf*idf jer nemamo podatke za N ili df
- problem integriranja odgovora, čvorovi mogu koristiti različite modele za ocjenu relevantnosti svojih lokalnih dokumenata

Dijeljenje indeksa prema riječima (1)



Dijeljenje indeksa prema riječima (2)

- pogodna organizacija indeksa za strukturirane mreže P2P
  - upit obrađuju samo oni čvorovi koji su zaduženi za ključne riječi iz upita
  - npr. ključ(peer) = hash(“a”)
  - za upit od *k* riječi potrebno je kontaktirati najviše *k* čvorova
- podaci o ključnim riječima odnose se na globalnu kolekciju pa je moguće izračunati globalni tf*idf jer svaki peer ima informaciju o df(*t*), a *N* možemo procijeniti
- generira se dodatni promet radi organizacije indeksa u mreži peerova
- skalabilno rješenje u smislu broja generiranih poruka po upitu
  - raste s $O(k)$
- neskalabilno rješenje zbog veličine poruke koja prenosi listu dokumenata vezanu uz ključnu riječ iz upita
  - raste s $O(\sqrt{D})$, gdje je *D* veličina globalne kolekcije dokumenata izražena ukupnim brojem riječi (ova kompleksnost se dobije iz tzv. Heapovog zakona)

- Niti jedno od predložena 2 rješenja ne daje zadovoljavajuće performanse:
  - rezultati pretraživanja su nezadovoljavajuće kvalitete za nestrukturiranu organizaciju mreže
  - i u nestrukturiranim i u strukturiranim mrežama generira se značajan promet (nije skalabilan)
- Stoga se oblikuju posebne tražilice (P2P-IR) koje pokušavaju smanjiti generirani mrežni promet a da pri tome ne utječu na kvalitetu odgovora tražilice

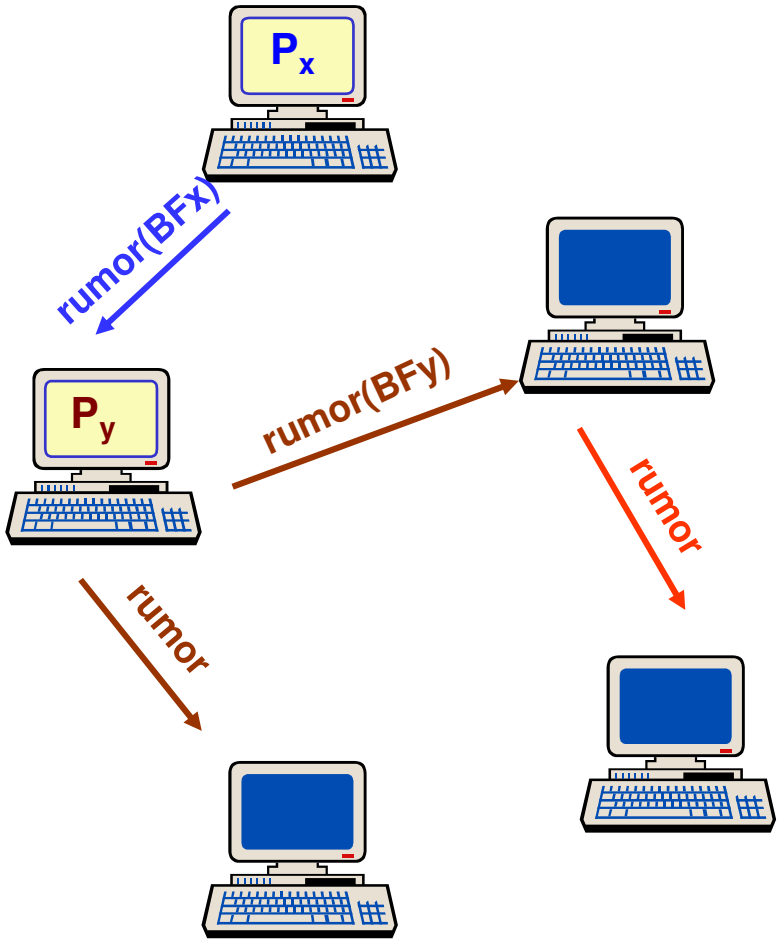
Sadržaj predavanja

- Pretraživanje sadržaja: osnovni pojmovi
- Pretraživanje tekstualnih kolekcija dokumenata
- Arhitektura raspodijeljene tražilice u grozdu/spletu računala
- Pretraživanje u mrežama P2P
- Pretraživanje tekstualne kolekcije dokumenata u mrežama P2P
- Primjeri tražilica temeljenih na mrežama P2P

PlanetP

Rutgers University

- koristi nestrukturiranu mrežu P2P
- peerovi međusobno razmjenjuju informacije o svom lokalnom indeksu, koristi se algoritam poznat pod nazivom “gossiping”
  - na slučajan odaberi susjeda i proslijedi mu informaciju o lokalnom indeksu
- informacija o lokalnom indeksu kodirana je pomoću Bloom filtra



Bloom filtar (1)

- Niz bitova duljine m koji omogućuje provjeru je li riječ dio rječnika ili nije
  - koriste se *hash* funkcije koje određene bitove bloom filtra postavljaju u 1
  - pomoću istih *hash* funkcija se provjerava članstvo u skupu
  - veličina bloom filtra << veličina kodiranog rječnika
  - postoji mala vjerojatnost za "*false positive*" (zaključujemo da je riječ dio rječnika, premda nije)
- Za skup $S = \{x_1, x_2, x_3, \ldots x_n\}$ gdje je $x_i \in U$, bloom filtar daje odgovor na sljedeće pitanje

$$y \in S \text{ ? (odgovor T ili F)}$$

Bloom filtar (2)

Niz bitova duljine m inicijalno se postavlja u 0.

B

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Primijeni funkciju *hash* na x_j iz S k puta. Ako je $H_i(x_j) = a$, postavi $B[a] = 1$.

B

0	1	0	0	1	0	1	0	0	1	1	1	0	1	1	0

Za provjeru je li y u S, provjeri $B[H_i(y)]$ za i=1…k. Svi bitovi moraju biti 1.

B

0	1	0	0	1	0	1	0	0	1	1	1	0	1	1	0

Moguće je da je svih k bitova 1, iako y nije element iz S.

B

0	1	0	0	1	0	1	0	0	1	1	1	0	1	1	0

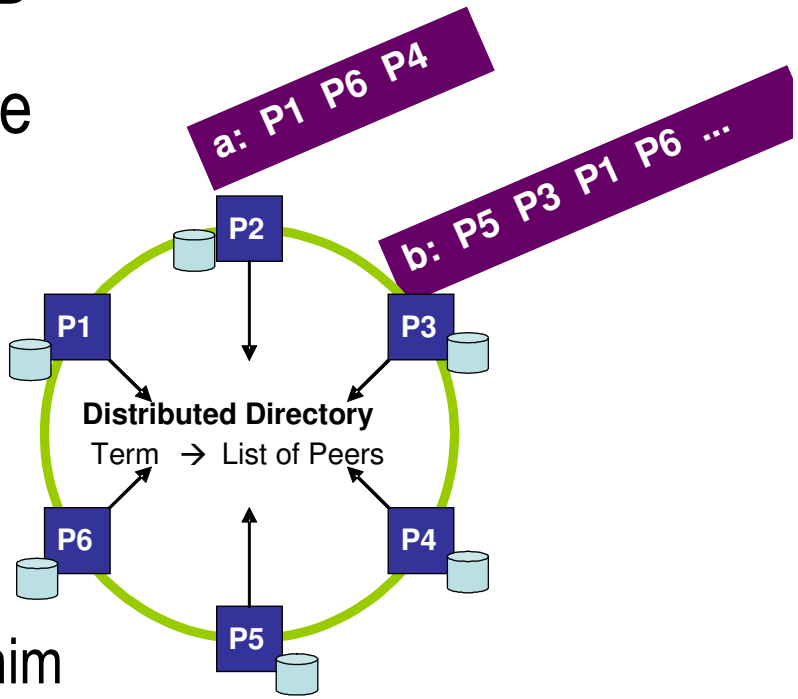
PlanetP: svojstva

- Svaki peer koristi posebnu heurističku funkciju za ocjenu “kvalitete” pojedinog peera za dani upit na temelju informacije o njegovom rječniku te odlučuje koliko će peerova kontaktirati tijekom pretraživanja
- Upitna je kvaliteta odgovora jer nije moguće ocijeniti kvalitetu lokalne kolekcije peera
- Nije moguće izračunati sličnost upita i kolekcije peera, nedostaju podaci za računanje tf*idf
- Rješenje je skalabilno za manje mreže (do 1000 peerova) u smislu generiranog prometa

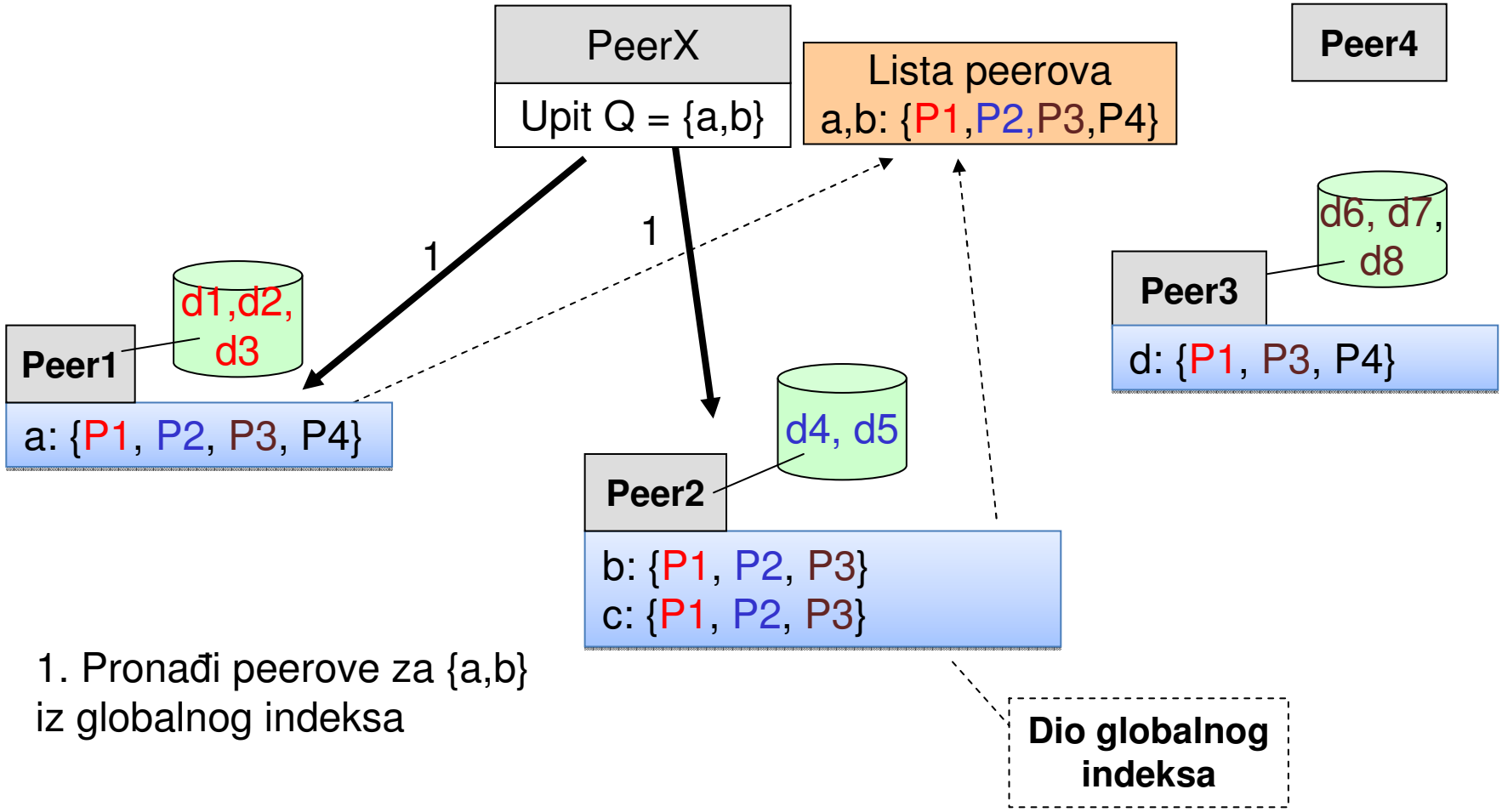
Minerva

Max-Planck Institute for Informatics, Saarbrücken

- *Peer Index* u strukturiranoj mreži P2P
- Izgrađuje indeks na nivou peera, a ne dokumenta i time smanjuje promet vezan uz indeksiranje kolekcije
- Upit se obrađuje u 2 koraka
  1. korak: pronađi peerove koji indeksiraju relevantne dokumente
  2. korak: pošalji originalni upit identificiranim peerovima te integriraj primljene odgovore

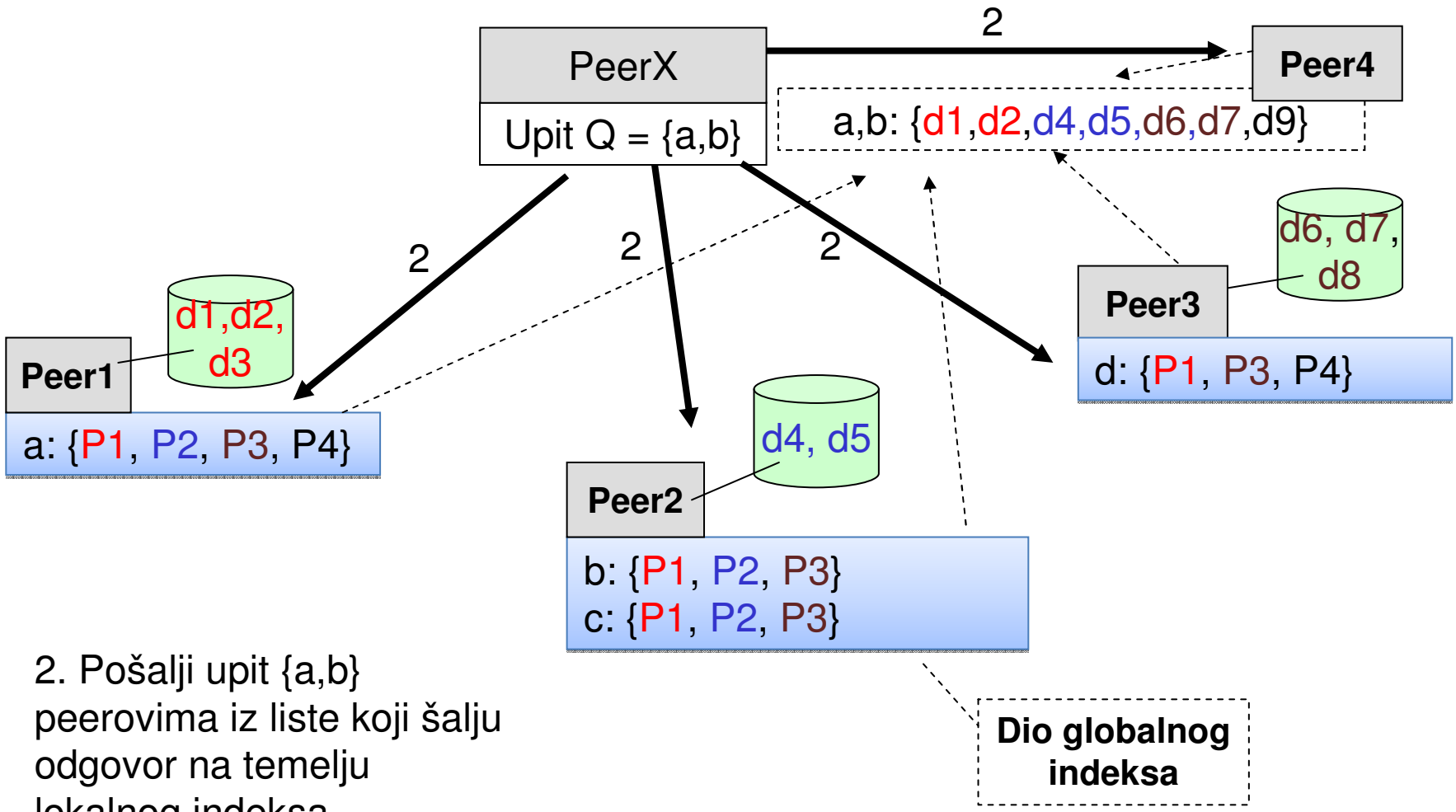


Minerva: 1. korak



1. Pronađi peerove za {a,b}
iz globalnog indeksa

Minerva: 2. korak



2. Pošalji upit {a,b} peerovima iz liste koji šalju odgovor na temelju lokalnog indeksa.

Minerva: svojstva

- rješenje je skalabilno u smislu generiranog prometa jer u koraku 1 prikuplja listu peerova, a nakon toga kontaktira mali skup peerova
  - broj peerova je značajno manji od broja dokumenata
- vrijeme odziva je povećano jer se na upit odgovara u 2 koraka
- upitna je kvaliteta odgovora jer ovisi značajno o ocjenama kolekcije pojedinog peera
  - koriste se posebni modeli za izračun relevantnosti kolekcije peera za dani upit, npr. CORI

AlvisP2P

Tražilica razvijena u okviru istraživačkog projekta ALVIS, EU FP6 (2004-2006)

http://globalcomputing.epfl.ch/alvis/

Daljnji razvoj: suradnja EPFL i FER

- Tražilica za pretraživanje tekstualne kolekcije dokumenata
  - koristi strukturiranu mrežu P2P za izgradnju i održavanje indeksa dokumenata
  - indeksira skupove riječi radi smanjenja mrežnog prometa tijekom obrade upita
- Koristi sljedeće činjenice vezane uz pretraživanje weba
  - korisnički upiti su kratki (u prosjeku 2 do 3 riječi)
  - riječi iz upita su “česte riječi” (riječi koje se često pojavljuju u tekstualnim dokumentima)
  - korisnike zanima mali broj kvalitetnih odgovora (preciznost je važnija od odziva)

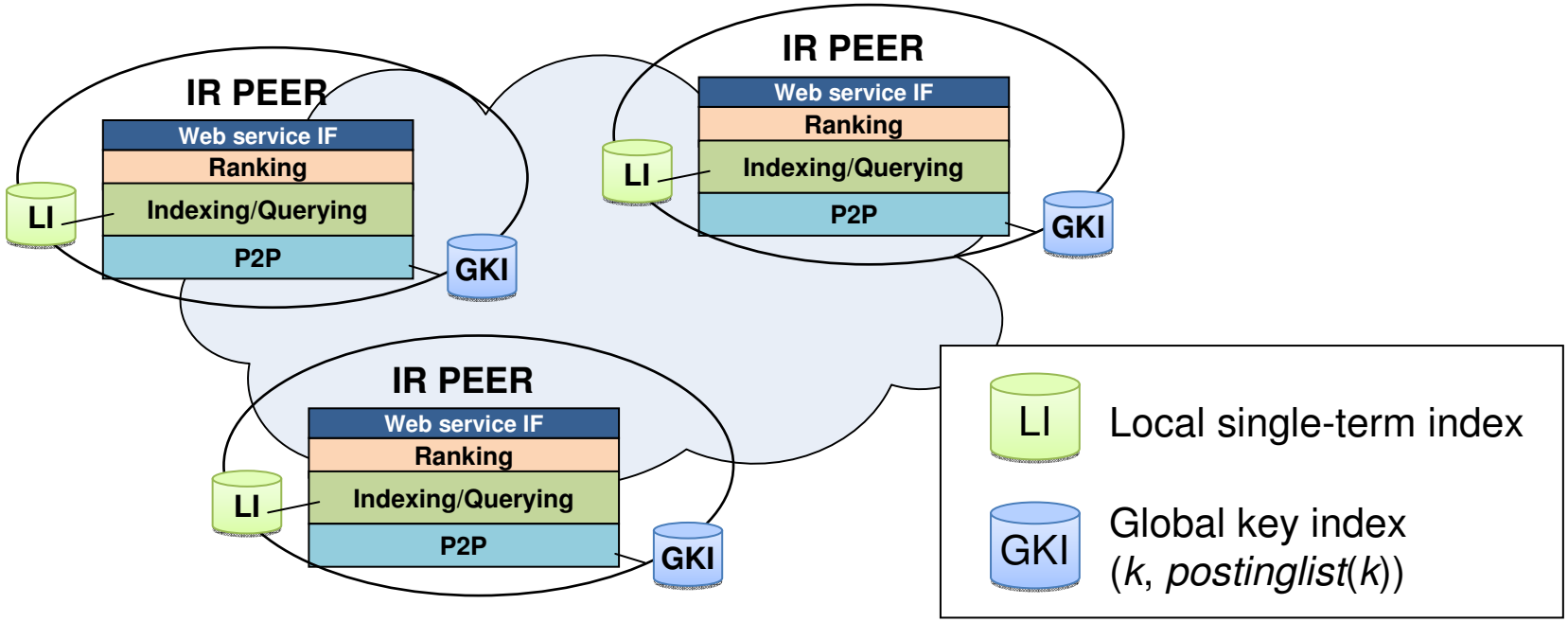
Arhitektura tražilice AlvisP2P

Strukturirana mreža P2P s *N* čvorova

Globalna kolekcija dokumenata $\mathcal{D}$ je podijeljena među peerovima, a svaki peer

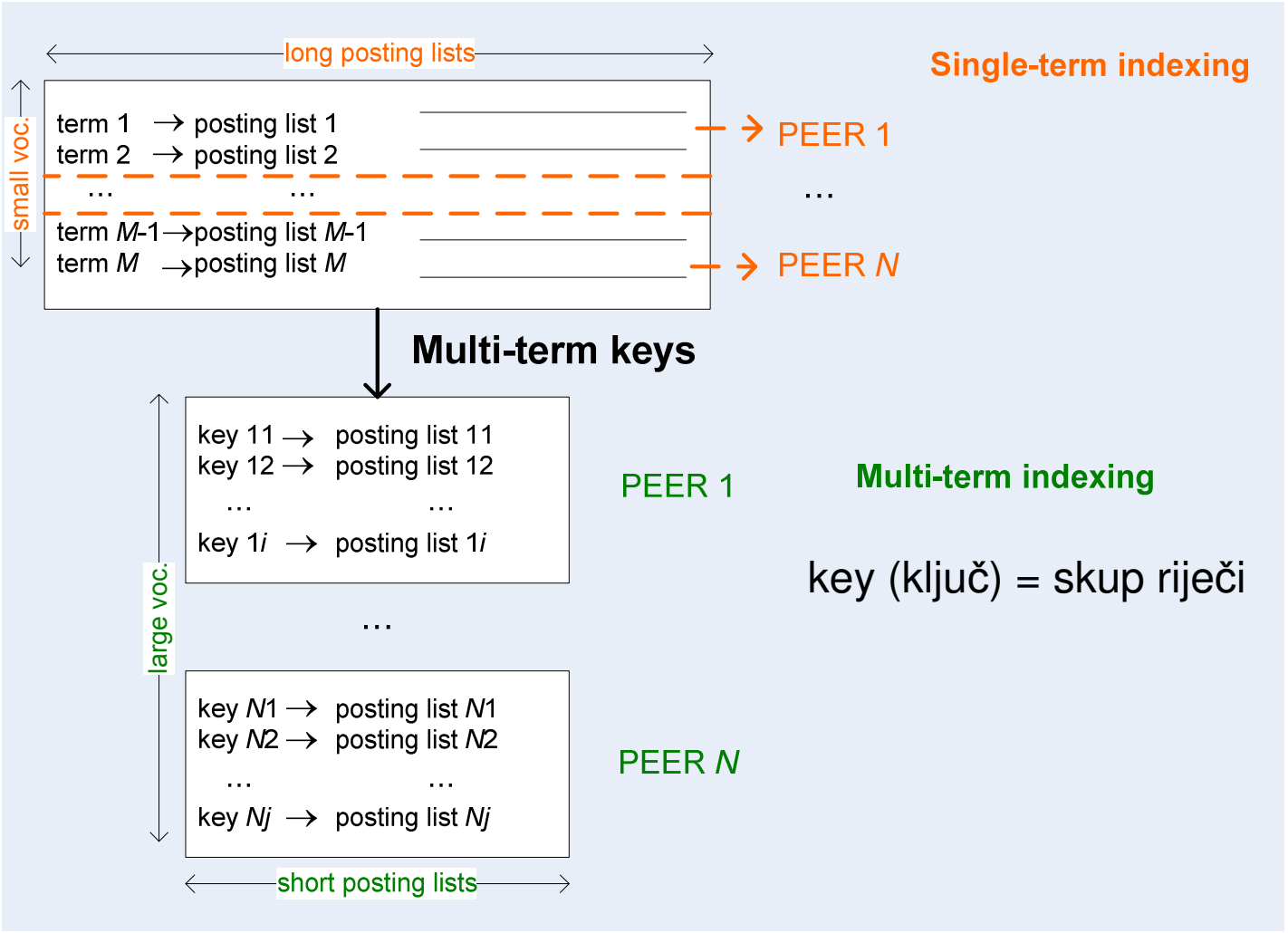
a) indeksira lokalnu kolekciju $\mathcal{D}$ (P_i) i unosi parove (*k*, *postinglist*(*k*)) u globalni indeks

b) održava dio globalnog indeksa (zadužen je za podskup ključeva)



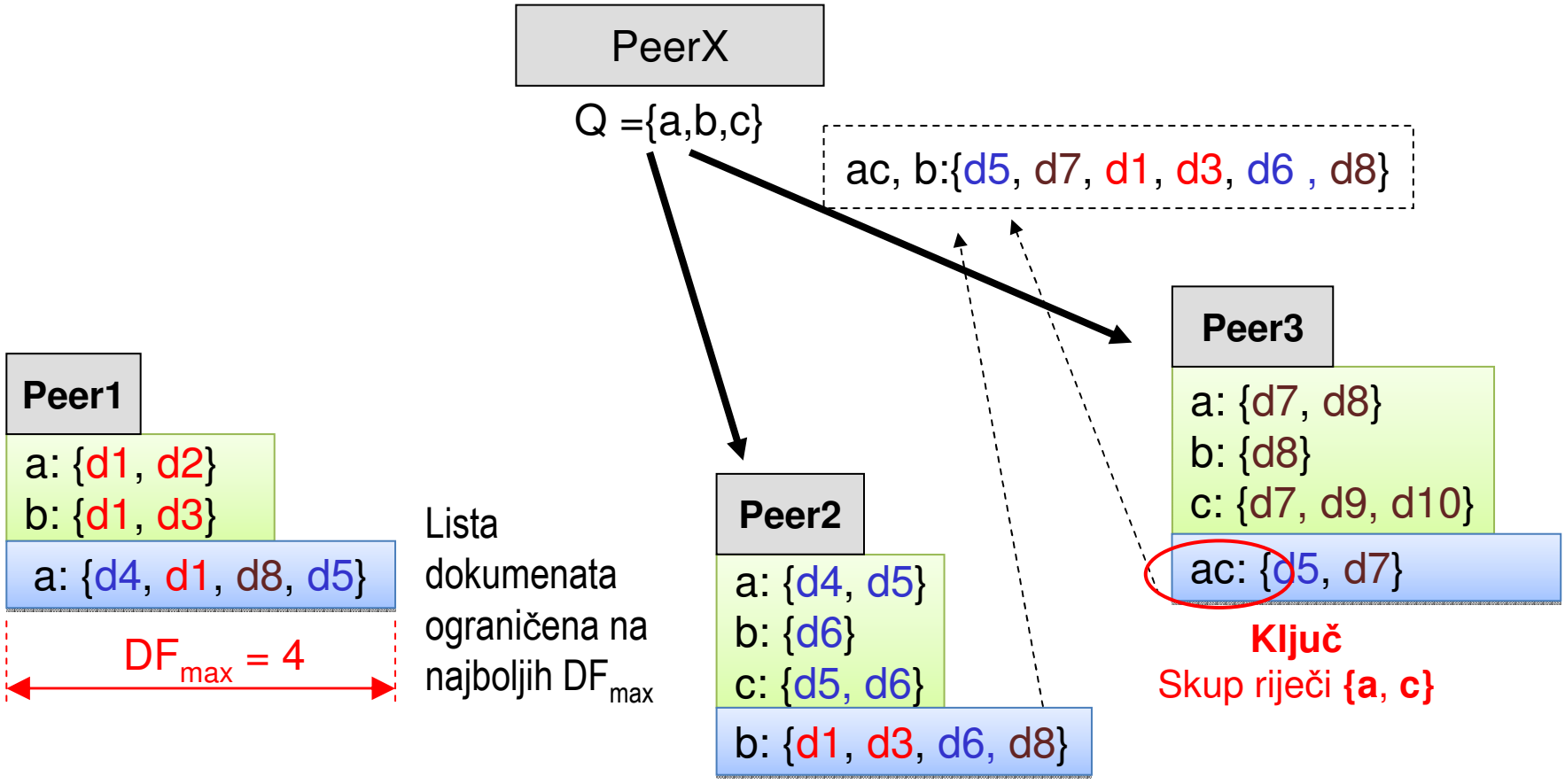
Indeksiranje skupa riječi (1)

Zavod za telekomunikacije



Indeksiranje skupa riječi (2)

Zavod za telekomunikacije



Promet tijekom pretraživanja ograničen je parametrom DF_{max} i brojem riječi iz upita!

Indeksiranje skupa riječi (3)

Kako odabrati skupove riječi za indeksiranje tako da kvaliteta odgovora (odziv i preciznost) bude zadovoljavajuća, a da veličina rječnika ne raste eksponencijalno?

- Indeksiranje pomoću *Highly Discriminative Keys* (HDKs)
- Indeksiranje na temelju upita (*Query-driven indexing*)

Indeksiranje pomoću HDKs

Non-Discriminative Keys (NDKs)

- t_1 je NDK iff
  - t_1 se pojavljuje u više od DF_{max} dokumenata kolekcije
- Lista dokumenata je ograničene veličine i sadrži samo najboljih DF_{max} dokumenata

Highly-Discriminative Keys (HDKs)

- e.g., (t_1, t_2) je HDK iff:
  - t_1 & t_2 se pojavljuje u manje od **DF_{max}** dokumenata kolekcije (diskriminativan ključ)
  - t_1 i t_2 su NDK (redundantnost)
  - t_1 i t_2 su u dokumentu udaljeni najviše ***w*** (udaljenost)
  - broj riječi koje čine ključ ograničena je s **s_{max}**

AlvisP2P: svojstva

- Indeksiranje pomoću HDKs se izvodi na temelju dokumenata kolekcije
  - veličina rječnika ključeva raste linearno s rastom veličine kolekcije korištenjem prethodno navedenih filtara
  - veličina rječnika i dalje je značajno veća od "običnog" rječnika (*single-term*)
- Indeksiranje na temelju upita, *Query-driven indexing* (QDI)
  - skup riječi čini ključ ako je HDK i pojavljuje se kao podskup riječi u upitu
  - na ovaj način se značajno smanjuje broj riječi u rječniku ključeva
- Eksperimenti pokazuju sljedeće:
  - tražilica AlvisP2P ima zadovoljavajuću kvalitetu odgovora
  - generirani promet tijekom indeksiranja je skalabilan (značajno manji za QDI)
  - performanse tražilice tijekom pretraživanja su izrazito dobre jer su odgovori na upite s više riječi već pripremljeni u indeksu, a generira se ograničen promet u mreži P2P

Dodatne informacije

- Kolegiji na FER-u
  - Umrežavanje sadržaja, 3. semestar
  - Sadržaj kolegija
    - Organizacija mreže posredničkih spremišta na webu
    - Mreže P2P i primjena za niz aplikacija
    - IPTV i video usluge
    - Obrada podataka u mrežama senzora
    - Trenutno poručivanje i prisutnost

Literatura

- Baeza-Yates, R.; Castillo, C.; Junqueira, F.; Plachouras, V.; Silvestri, F., "Challenges on Distributed Web Retrieval,", *IEEE 23rd International Conference on Data Engineering* (ICDE 2007), pp.6-20, 15-20 April 2007.
- **PlanetP**: F.M. Cuenca-Acuna, C. Peery, R.P. Martin, T.D. Nguyen, PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities, Proc. HPDC, 2003.
- **Minerva**: M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer, Improving collection selection with overlap awareness in P2P search engines, Proc. SIGIR, 2005.
- **AlvisP2P**: G. Skobeltsyn, T. Luu, I. Podnar Žarko, M. Rajman, K. Aberer, Query-Driven Indexing for Scalable Peer-to-Peer Text Retrieval, Future Generation Computer Systems, 25(2009), pp.88-99