Preddiplomski studij

Računarstvo

Modul:
Telekomunikacije i informatika

Višemedijske usluge

Pretraživanje informacija na WWW-u (1. dio)

Ak.god. 2007./2008.

24.04.2008.

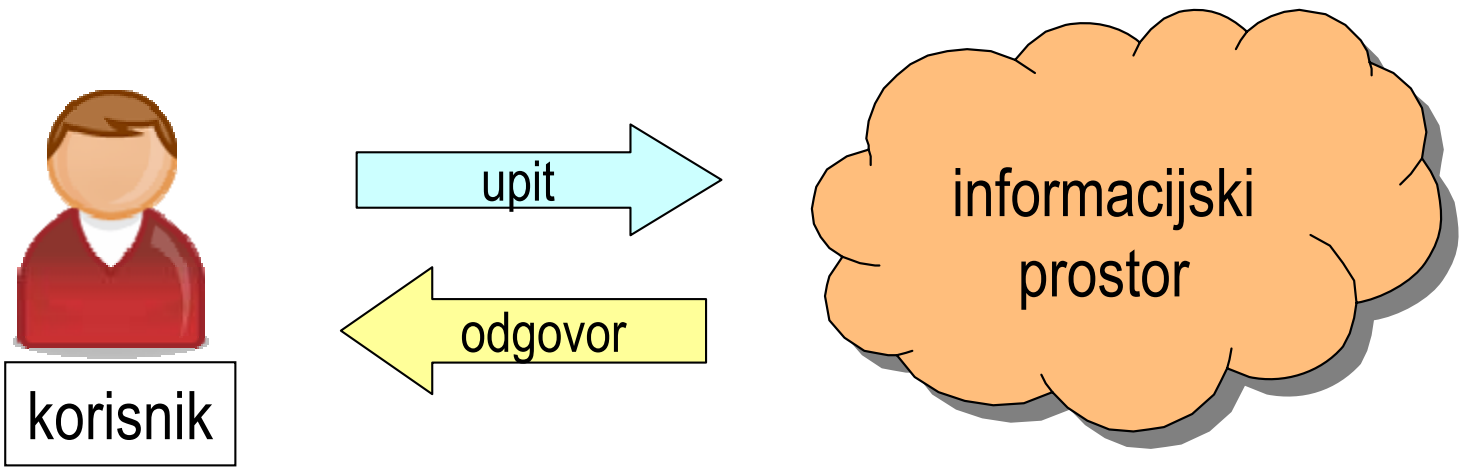
Sadržaj predavanja

- Pretraživanje informacija
  - sustavi za pretraživanje informacija
  - modeli i ocjena kvalitete modela
  - odziv i preciznost
- Pretraživanje tekstualnog sadržaja
  - Booleov model
  - vektorski prostorni model

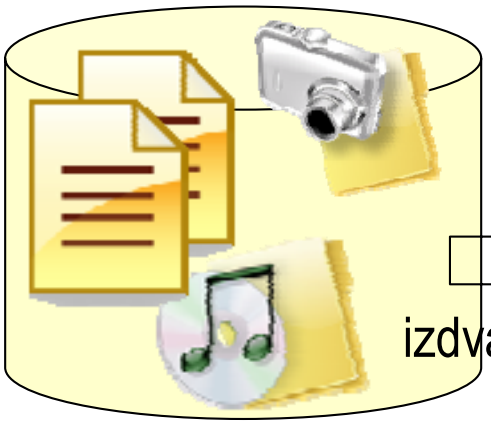
Pretraživanje informacija

engl. *information retrieval*

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



Sustav za pretraživanje informacija



izdvajanje značajnih svojstava (npr. riječi)

strukturirani prikaz dokumenata

definiranje upita (npr. niz riječi)

strukturirani prikaz upita

usporedba upita i dokumenta

odgovor

rangirana lista dokumenata

Pojmovi

- informacijski prostor čini kolekcija dokumenata
- kolekcija je konačni skup dokumenata različitog tipa (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

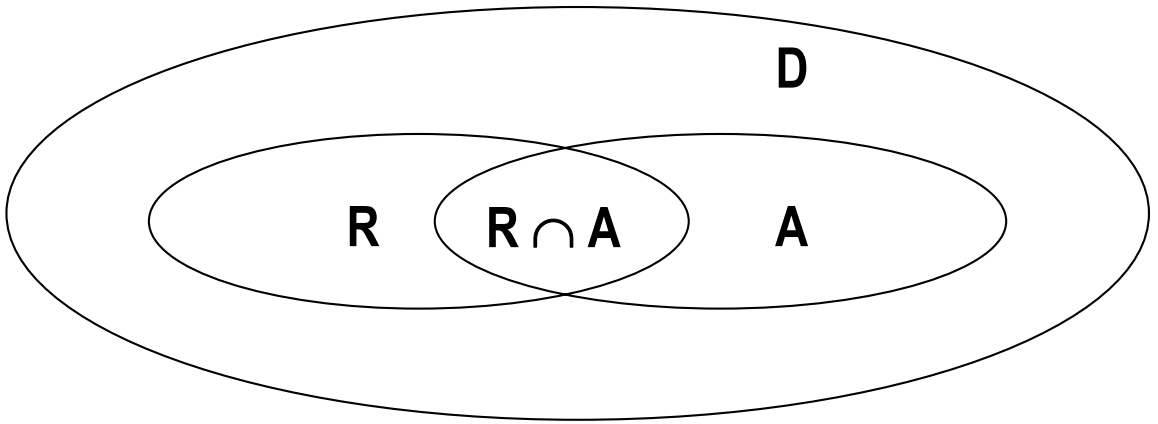
Zadaće sustava za pretraživanje informacija

- generiranje strukturiranog prikaza dokumenata
  - izdvajanje značajnih svojstava iz dokumenata, npr. riječi iz teksta (jednostavno) 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 (engl. *relevance*) za dani upit
  - sličnost (engl. *similarity*) je mjera koja određuje relevantnost dokumenta za neki upit, uspoređuje sličnost dokumenta i upita

Modeli za pretraživanje sadržaja

- cilj - pronaći podskup dokumenata koji su relevantni za dani upit
- 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

Ocjena kvalitete modela



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

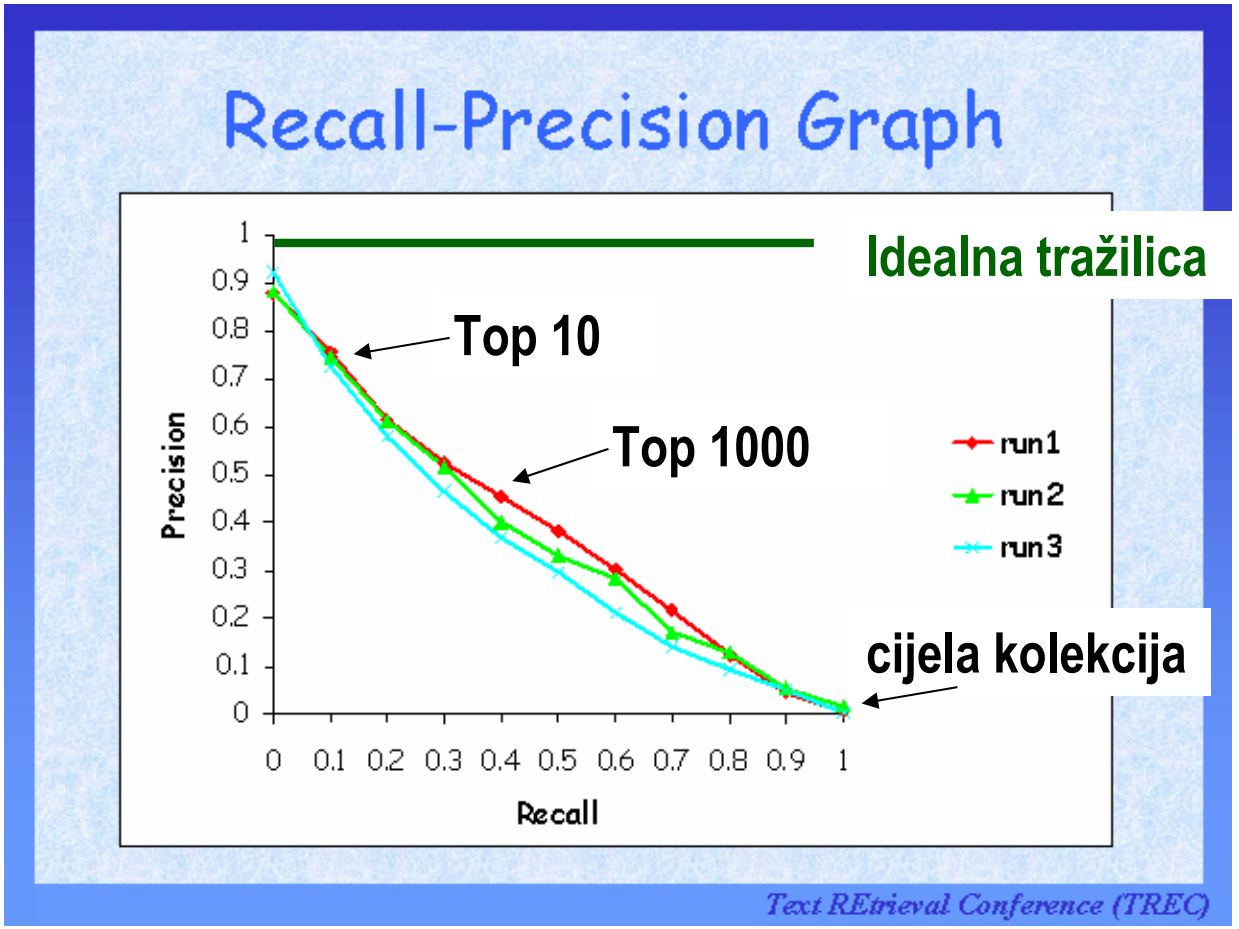
Odziv i preciznost

- 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

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



-veća preciznost znači manji odziv

-odnos preciznosti i odziva se može kontrolirati brojem dokumenata u odgovoru

-odziv će uvijek biti 100% ako je odgovor cijela kolekcija

-idealna tražilica ima preciznost 100%

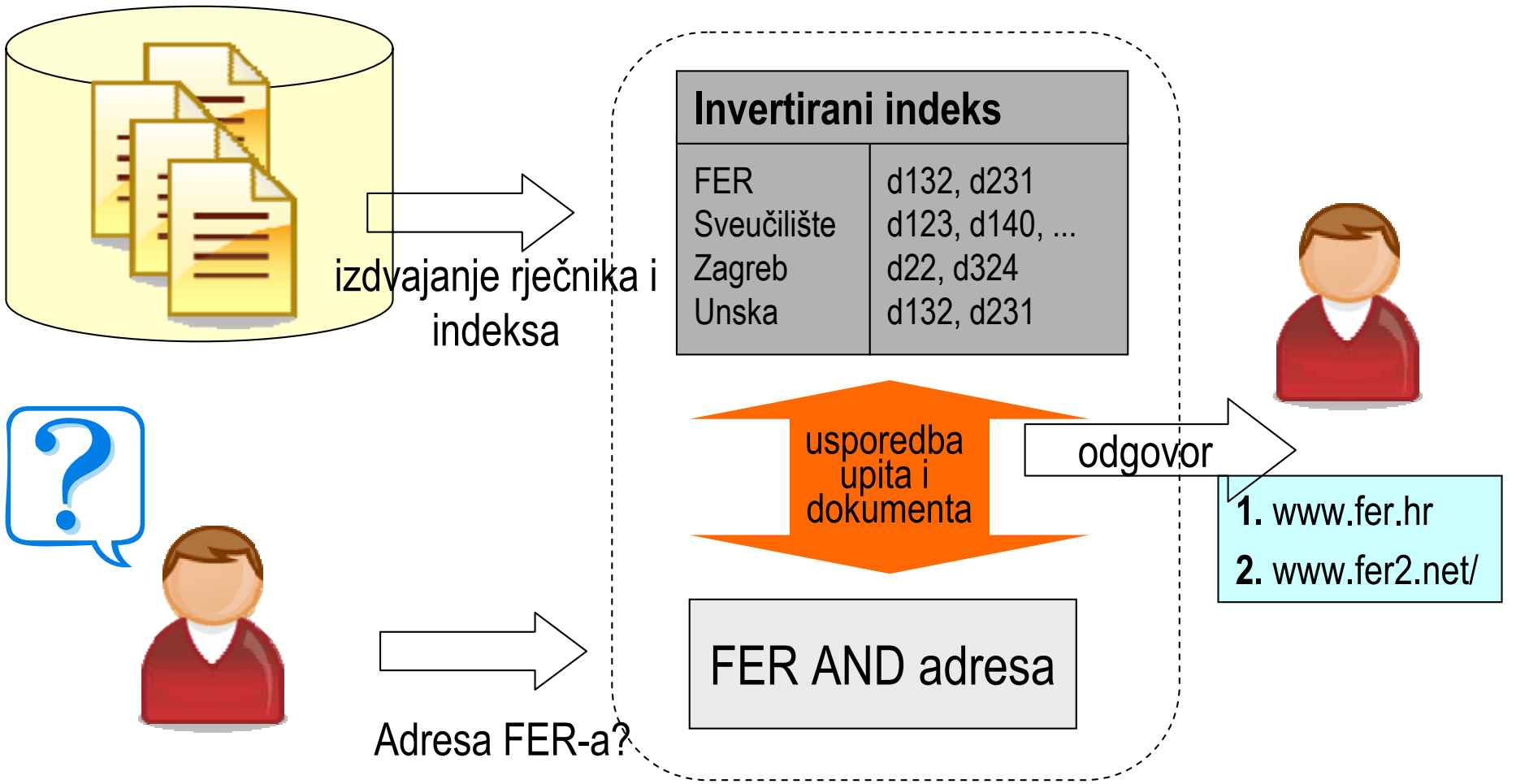
Sadržaj predavanja

- Pretraživanje informacija
  - sustavi za pretraživanje informacija
  - modeli i ocjena kvalitete modela
  - odziv i preciznost
- Pretraživanje tekstualnog sadržaja
  - Booleov model
  - vektorski prostorni model

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 (primjer PageRank / Google)

Sustav za pretraživanje tekstualnog sadržaja



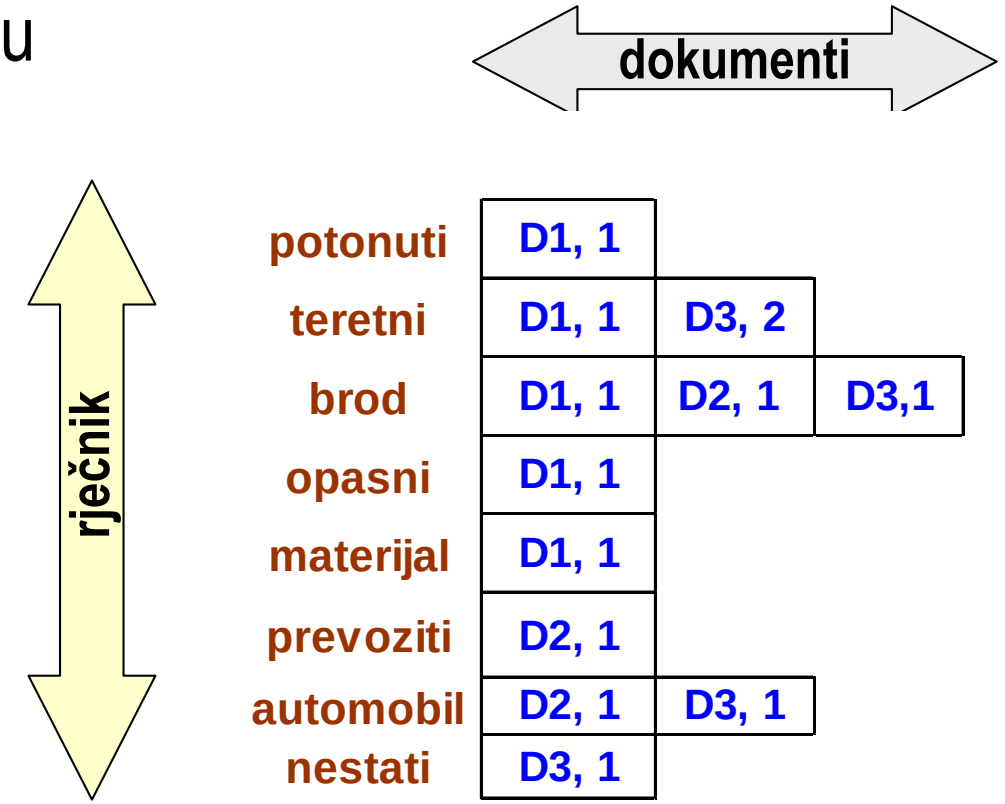
Pojmovi (1)

- 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 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



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**

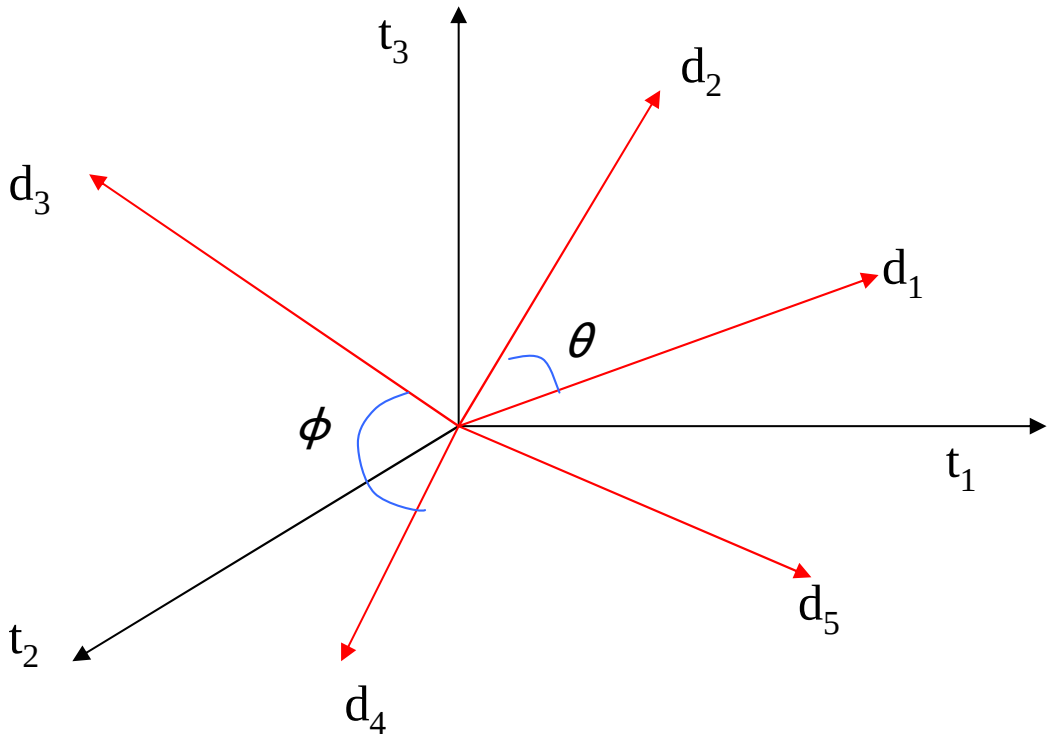
Booleov model

- prethodni primjer koristi Booleov model koji se temelji na Boolevoj 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
  - dokument ili zadovoljava upit ili ne (nema rangiranja vezano uz relevantnost dokumenta za zadani upit)

Vektorski prostorni model

- dokumenti i upiti prikazuju se kao težinski vektori u *m*-dimenzionalnom vektorskom prostoru (*m* je veličina rječnika kolekcije)
- sličnost upita i dokumenta
  - mjera kojom se određuje relevantnost dokumenta za neki upit
  - u odgovoru se mogu pojaviti i dokumenti koji ne sadrže sve riječi iz upita
- rangiranje dokumenata na temelju izračunate sličnosti
- danas jedan od najraširenijih modela

Vektorski model



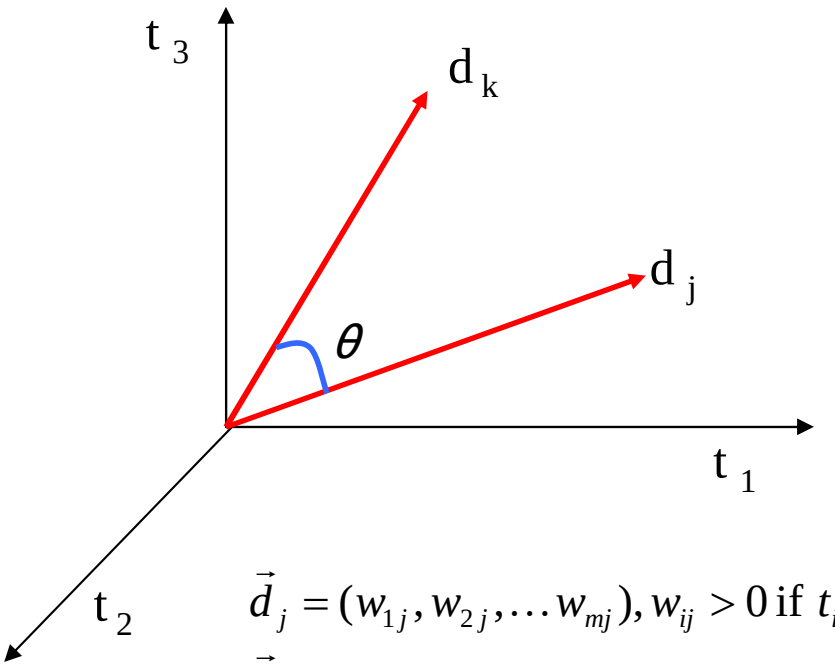
Pretpostavka: Dokumenti koji su "bliže" u vektorskom prostoru semantički su slični (govore o sličnim stvarima).

Sličnost dokumenata

- udaljenost među vektorima d_j i d_k računa se kao kosinus kuta među njima

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

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



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

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

Upit se razmatra kao kratki dokument!

Težinski faktor (1)

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

- *tf (i, j) – term frequency*
  - broj pojavljivanja riječi t_i u dokumentu d_j
  - Dokumenti su slični ako sadrže iste riječi. Što je veći broj pojavljivanja riječi u dokumentu, to je dokument relevantniji za upit koji sadrži tu riječ.
  - Što je s čestim riječima koje se pojavljuju u svim dokumentima?
- *idf (i) – inverse document frequency*
  - uzima u obzir koliko se često riječ t_i pojavljuje u dokumentima kolekcije

$$idf(i) = \log\left(\frac{N}{df_i}\right)$$

N – veličina kolekcije (broj dokumenata)

df_i – broj dokumenata kolekcije u kojima se pojavljuje t_i

Težinski faktor (2)

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

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

- za težinu je osim broja pojavljivanja riječi u dokumentu značajna i informacija koji postotak dokumenata kolekcije sadrži traženu riječ
  - ako se riječ često pojavljuje u samo 1 dokumentu kolekcije, onda je taj dokument najrelevantniji za upit
  - česte riječi koje se pojavljuju u svim dokumentima imaju težinu 0

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

  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

Vektorski prostorni model (zaključak)

- Model se pokazao dobrim za općenite kolekcije uz težinski faktor *tf* x *idf*
  - postoji niz alternativnih načina za računanje težinskih faktora
- Prednosti
  - težinski faktori poboljšavaju kvalitetu odgovora
  - relevantni dokumenti ne moraju sadržavati sve riječi iz upita
- Nedostaci
  - pretpostavlja se neovisnost indeksiranih riječi