

# Semantičko pretraživanje informacija u tekstualnim dokumentima

Jasminka Dobša  
Fakultet organizacije i informatike  
Varaždin  
*jasminka.dobsa@foi.hr*



---

# Dubinska analiza teksta - Text mining

- Semantičko pretraživanje informacija u tekstualnim dokumentima je zadatak discipline dubinske analize teksta ili rudarenja tekstualnih podataka (engl. text mining)
  - sastavni je dio discipline dubinske analize podataka ili rudarenja podataka (engl. data mining)
  - bavi se sadržajno utemeljenom obradom nestrukturiranih tekstualnih dokumenata i izdvajanjem korisne informacije iz njih

---

# Sadržaj

- Modeli za predstavljanje tekstualnih dokumenata (model vektorskog prostora)
- Proces indeksiranja i evaluacija pretraživanja informacija
- Problemi kod semantičkog pretraživanja informacija
- Mogućnosti za rješavanje tih problema
  - Naglasak na metodama snižavanja dimenzije vektorskog prostora: konceptualno indeksiranje
- Operativni problem kod snižavanja dimenzije vektorske reprezentacije dokumenata

---

# Modeli

- Zadatak pretraživanja informacija: vratiti kao rezultat pretraživanja na postavljen korisnički upit što više dokumenata relevantnih za korisnički upit i pri tome vratiti što manje dokumenata koji nisu relevantni
- Matematički modeli za predstavljanje tekstualnih dokumenata:
  - Vjerojatnosni model
  - Logički model
  - Model vektorskog prostora (MVP) ili model vreće riječi (engl. bag of words)
- U MVP tekstualni su dokumenti predstavljeni u visoko dimenzionalnom vektorskom prostoru
  - Dimenzija prostora ovisi o broju indeksnih pojmova
- MVP se implementira formiranjem matrice pojmova i dokumenata

# Matrica pojmov i dokumenata

- Matrica pojmov i dokumenata je matrica tipa  $m \times n$  gdje je  $m$  broj pojmova, a  $n$  je broj dokumenata
- **Redak matrice pojmov i dokumenata = pojam**
- **Stupac matrice pojmov i dokumenata = dokument**

Slika 1. Matrica pojmov i dokumenata

$$A = \begin{array}{cccc} & d_1 & d_2 & \dots & d_n & & \\ & \downarrow & \downarrow & & \downarrow & & \\ \left[ \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right] & \leftarrow p_1 & \leftarrow p_2 & \vdots & \leftarrow p_m \end{array}$$

# Upit

- Korisnički upit je predstavljen u istom obliku kao i dokumenti ( $m$ -dimenzionalni vektor)
- Mjera sličnosti između upita  $q$  i dokumenta  $a_j$  je kosinus kuta između vektorskih reprezentacija upita i dokumenta

$$\cos(\mathbf{a}_j, \mathbf{q}) = \frac{\mathbf{a}_j^T \mathbf{q}}{\|\mathbf{a}_j\|_2 \|\mathbf{q}\|_2}$$

---

# Proces indeksiranja

- Prethodna obrada teksta obuhvaća postupke kao što su:
  - ❑ leksička obrada teksta
  - ❑ uklanjanje stop riječi (članovi, veznici, prijedlozi i slične riječi koje nemaju diskriminacijsku vrijednost kod pretraživanja)
  - ❑ svođenje riječi na korijensku ili osnovnu formu
  - ❑ izbor indeksnih pojmova

# Indeksni pojmovi koji se sastoje od više riječi

- Identifikacija sintagmi i njihovo uvođenje kao indeksnih pojmova (npr. *godišnji odmor*)

- **n-grami**

D. Mladenić, M. Grobelnik, Word sequences as features in text-learning, Proceedings of the 7th Electornical and Computer Science Conference, Ljubljana, Slovenija, 1998.

- **fraze promjenjive dužine** (engl. flexible length phrases)

D. Radošević, J. Dobša, D. Mladenić, Z. Stapić, M. Novak, Genre document classification using flexible length phrases, *Proceedings of 17th International Conference on Information and Intelligent Systems*,, Varaždin, Hrvatska, 2006., 231.-234.

D. Radošević, J. Dobša, D. Mladenić, Flexible length phrases in document classification, *Procedings of the 28th International Conference of Information Technology Interfaces*, Cavtat/Dubrovnik, Hrvatska, 2006., 457.-462.

# Evaluacija pretraživanja informacija

- Mjere evaluacije:
  - Odaziv
  - Preciznost
  - Prosječna preciznost

- **Odaziv**

$$recall_i = \frac{r_i}{r_n}$$

- **Preciznost**

$$precision_i = \frac{r_i}{i}$$

- $r_i$  je broj relevantnih dokumenata između  $i$  najviše rangiranih dokumenata
- $r_n$  je ukupan broj relevantnih dokumenata u zbirci dokumenata
- **Prosječna preciznost** – prosječna preciznost na više nivoa odaziva (obično 11)

# Problemi kod pretraživanja informacija

- U klasičnom MVP sličnost između dokumenata i upita ispituje se leksički
- **Problem kod pretraživanja informacija predstavljaju**
  - **Sinonimi** – mogu biti razlog slabog odaziva
  - **Višeznačnice** – mogu biti razlog slaboj preciznosti pretraživanja
  - Kod pretraživanja informacija na Web-u nije moguće naći sve relevantne dokumente
- Neke od tehnika za semantičko pretraživanje:
  - proširivanje korisničkog upita
  - kod dokumenata na web-u: korištenje strukture poveznica (engl. link) između dokumenata
  - konceptualno indeksiranje dokumenata

---

# Primjer

- Zbirka od 15 dokumenata (naslovi knjiga ili članaka)
  - 9 iz područja dubinske analize (tekstualnih) podataka
  - 5 iz područja linearne algebre
  - 1 kombinacija ta dva područja (primjena linearne algebre u području dubinske analize podataka)
- Lista indeksnih pojmova je formirana
  - od pojmova sadržanih u barem 2 dokumenta
  - izbačeni su pojmovi sadržani u listi stop pojmova
  - pojmovi su svedeni na svoj osnovni oblik

# Dokumenti 1/2

D1	Survey of <u>text mining</u> : <u>clustering</u> , <u>classification</u> , and <u>retrieval</u>
D2	Automatic <u>text</u> processing: the transformation <u>analysis</u> and <u>retrieval</u> of <u>information</u> by computer
D3	Elementary <u>linear algebra</u> : A <u>matrix</u> approach
D4	<u>Matrix algebra</u> & its <u>applications</u> statistics and econometrics
D5	Effective databases for <u>text</u> & <u>document</u> management
D6	<u>Matrices</u> , <u>vector spaces</u> , and <u>information retrieval</u>
D7	<u>Matrix analysis</u> and <u>applied linear algebra</u>
D8	Topological <u>vector spaces</u> and <u>algebras</u>

# Dokumenti 2/2

D9	<u>Information retrieval</u> : <u>data</u> structures & <u>algorithms</u>
D10	<u>Vector spaces</u> and <u>algebras</u> for chemistry and physics
D11	<u>Classification</u> , <u>clustering</u> and <u>data analysis</u>
D12	<u>Clustering</u> of large <u>data</u> sets
D13	<u>Clustering algorithms</u>
D14	<u>Document</u> warehousing and <u>text mining</u> : techniques for improving business operations, marketing and sales
D15	<u>Data mining</u> and knowledge discovery

---

# Upiti

- Q1: Data mining
  - Relevantni dokumenti : Svi dokumenti vezani uz dubinsku analizu podataka (D1, D2, D5, D9, D11, D12, D13, D14, D15)
  
- Q2: Using linear algebra for data mining
  - Relevantan dokument: D6

## Rezultati pretraživanja u MVP

Upit Q1		Upit Q2	
Dokument	Skalarni produkt	Dokument	Skalarni produkt
D15	1.4142	D15	1.4142
D12	0.7071	D3	1.1547
D14	0.5774	D7	0.8944
D9	0.5000	D12	0.7071
D11	0.5000	D4	0.5774
D1	0.4472	D8	0.5774
D2	0	D10	0.5774
D3	0	D14	0.5774
D4	0	D9	0.5000
D5	0	D11	0.5000
D6	0	D1	0.4472
D7	0	D2	0
D8	0	D5	0
D10	0	D6	0
D13	0	D13	0

# Modifikacija upita u MVP

- Modifikacija upita korištenjem povratne informacije korisnika
- Iterativni postupak:
  - korisniku se predstavljaju dokumenti relevantni za njegov upit (temeljem algoritma za pretraživanja)
  - korisnik među vraćenim dokumentima bira relevantne (skup  $D_r$ ) i nerelevantne (skup  $D_n$ )
  - težine indeksnih pojmova u vektoru upita se korigiraju

- Neka je početna reprezentacija korisničkog upita u MVP dana vektorom

$$\mathbf{q} = (q_1, q_2, \dots, q_m)^T$$

čija i-ta komponenta predstavlja težinu i-tog indeksnog pojma

- Modificirani upit tada ima oblik:

$$\mathbf{q}_{\text{mod}} = \alpha \mathbf{q} + \frac{\beta}{|D_r|} \sum_{\mathbf{a}_j \in D_r} \mathbf{a}_j - \frac{\gamma}{|D_n|} \sum_{\mathbf{a}_j \in D_n} \mathbf{a}_j$$

pri čemu su  $\alpha$ ,  $\beta$  i  $\gamma$  konstante za podešavanje ( $>0$ )

- U slučaju upita  $Q_1$  (Data mining)
  - tražilica kao rezultat pretraživanja vraća dokumente D15, D12, D14, D9, D11 i D1
  - svi ti dokumenti su relevantni
  - to nisu svi relevantni dokumenti: kao rezultat pretraživanja nisu vraćeni relevantni dokumenti D2, D5 i D13
- Modificirani upit imat će oblik ( $\alpha=1, \beta=1$ )

$$\mathbf{q}_{\text{mod}} = \mathbf{q} + \frac{1}{6}(\mathbf{a}_{15} + \mathbf{a}_{12} + \mathbf{a}_{14} + \mathbf{a}_9 + \mathbf{a}_{11} + \mathbf{a}_1)$$

- U modificiranom upitu pored pojmova *data* i *mining* težine različite od 0 imaju i pojmovi *text*, *clustering*, *classification*, *retrieval*, *analysis*, *document* i *algorithm*

---

# Pretraživanje informacija na Web-u

- Web-stranice predstavljaju posebnu vrstu dokumenata: pored teksta one sadrže i poveznice (linkove) na druge stranice
- Web stranice se skupljaju s Web-a korištenjem posebnih programa, tzv. puzilica
  - Kreću od neke adrese i skupljaju stranice sukcesivno korištenjem poveznica (linkova) na druge adrese
- Kako indeksirati Web-stranice?
  - (Attardi i suradnici, 1999): Za danu Web-stranicu formira se **opisnik stranice**
  - (Fuhr i suradnici, 1999): Uključivanjem teksta na danoj Web-stranici i teksta na stranicama na koje ta stranica upućuje

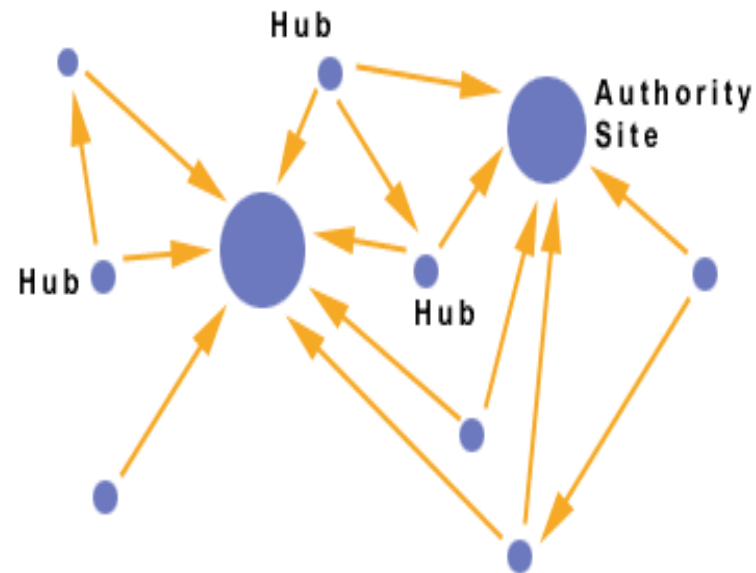
---

# Mjere za vrednovanje Web-stranica

- Mjera odaziva u slučaju Web-a gubi na važnosti jer nije moguće niti približno procijeniti broj relevantnih Web-stranica za određeni korisnički upit
- Zbog toga su razvijene mjere za vrednovanje Web-stranica
  - Web-stranice koje će biti više vrednovane preferirati će se toku pretraživanja
- 1996. godine razvijena su dva algoritma za pretraživanje informacija na Web-u
  - PageRank (prototipni algoritam za tražilicu Google)- Larry Page i Sergey Brin
  - HITS (Hyperlink Induced Topis Search) – Jon Kleinberg

# Mjere za vrednovanje Web-stranica: HITS algoritam

- HITS algoritam koristi mjere
  - **Utjecaja** (engl. authority value): stranica s visokim utjecajem je ona na koju mnogo stranica referencira, tj. stranica s velikim brojem ulaznih poveznica
  - **Informativnosti** (engl. hub value): stranica visoke informativnosti je stranica koja referencira na utjecajne stranice



# Mjere za vrednovanje Web-stranica: PageRank algoritam

- Algoritam PageRank koristi mjeru **statusa** (engl. prestige)
- Mjera statusa Web-stranice se definira kao zbroj statusa svih stranica koje referenciraju na tu stranicu

Larry Page i Sergej Brin



# Mjera statusa 1/2

**E** matrica susjedstva za graf Web-stranica

$E(k, j) = 1$  ako stranica  $d_k$  referencira na stranicu  $d_j$ , a nula inače

**s** - statusni vektor koji sadrži statusne vrijednosti svih stranica

Početna vrijednost statusnog vektora je

$$\mathbf{s} = (1, 1, \dots, 1)$$

Za postojeći statusni vektor **s** novi statusni vektor  $\mathbf{s}_1$  definira se **s**

$$\mathbf{s}_1 = \mathbf{E}^T \mathbf{s}$$

---

# Mjera statusa 2/2

- Konačna vrijednost statusnog vektora je fiksna točka iterativnog preslikavanja

$$\mathbf{s} \leftarrow \mathbf{E}^T \mathbf{s}$$

- Fiksna točka je jednaka svojstvenom vektoru matrice  $\mathbf{E}^T$  koji od odgovara njenoj najvećoj svojstvenoj vrijednosti

---

# Konceptualno indeksiranje

- Konceptualnim se indeksiranjem nastoje riješiti problemi sinonima i višeznačnica
- Dokumenti se predstavljaju novim značajkama ili karakteristikama (eng. features) – reparametrizacija
- Dvije tehnike konceptualnog indeksiranja
  - **Latentno semantičko indeksiranje (LSI)**
  - **Konceptno indeksiranje (CI)**
- Ovim se tehnikama dokumenti predstavljaju u vektorskom prostoru koji je često puno niže dimenzije nego reprezentacija u MVP

---

# Latentno semantičko indeksiranje (LSI)

- Predstavljeno 1990. godine
  - S. Deerwester, S. Dumas, G. Furnas, T. Landauer, R. Harsman: *Indexing by latent semantic analysis*, J. American Society for Information Science, 41, 1990, pp. 391-407
- Metoda dodavanja novih dokumenata i pojmova predstavljena 1995. godine
  - M. W. Berry, S.T. Dumas, G.W. O'Brien: *Using linear algebra for intelligent information retrieval*, SIAM Review, 37, 1995, pp. 573-595
- Temelji se na spektralnoj analizi matrice pojmova i dokumenata

# Dekompozicija singularnih vrijednosti

- Za svaku matricu  $A$  tipa  $m \times n$  postoji **dekompozicija singularnih vrijednosti** (engl. **singular value decomposition, SVD**)

$$A = U \Sigma V^T$$

$U$  ortogonalna matrica tipa  $m \times m$  čiji stupci su lijevi singularni vektori matrice  $A$

$\Sigma$  dijagonalna matrica na čijoj dijagonali su singularne vrijednosti matrice  $A$  u padajućem redoslijedu

$V$  ortogonalna matrica tipa  $n \times n$  čiji su stupci desni singularni vektori matrice  $A$

# Krnja dekompozicija singularnih vrijednosti

- Za metodu LSI koristi se **krnja dekompozicija singularnih vrijednosti (engl. truncated SVD)**

$$A_k = U_k \Sigma_k V_k^T$$

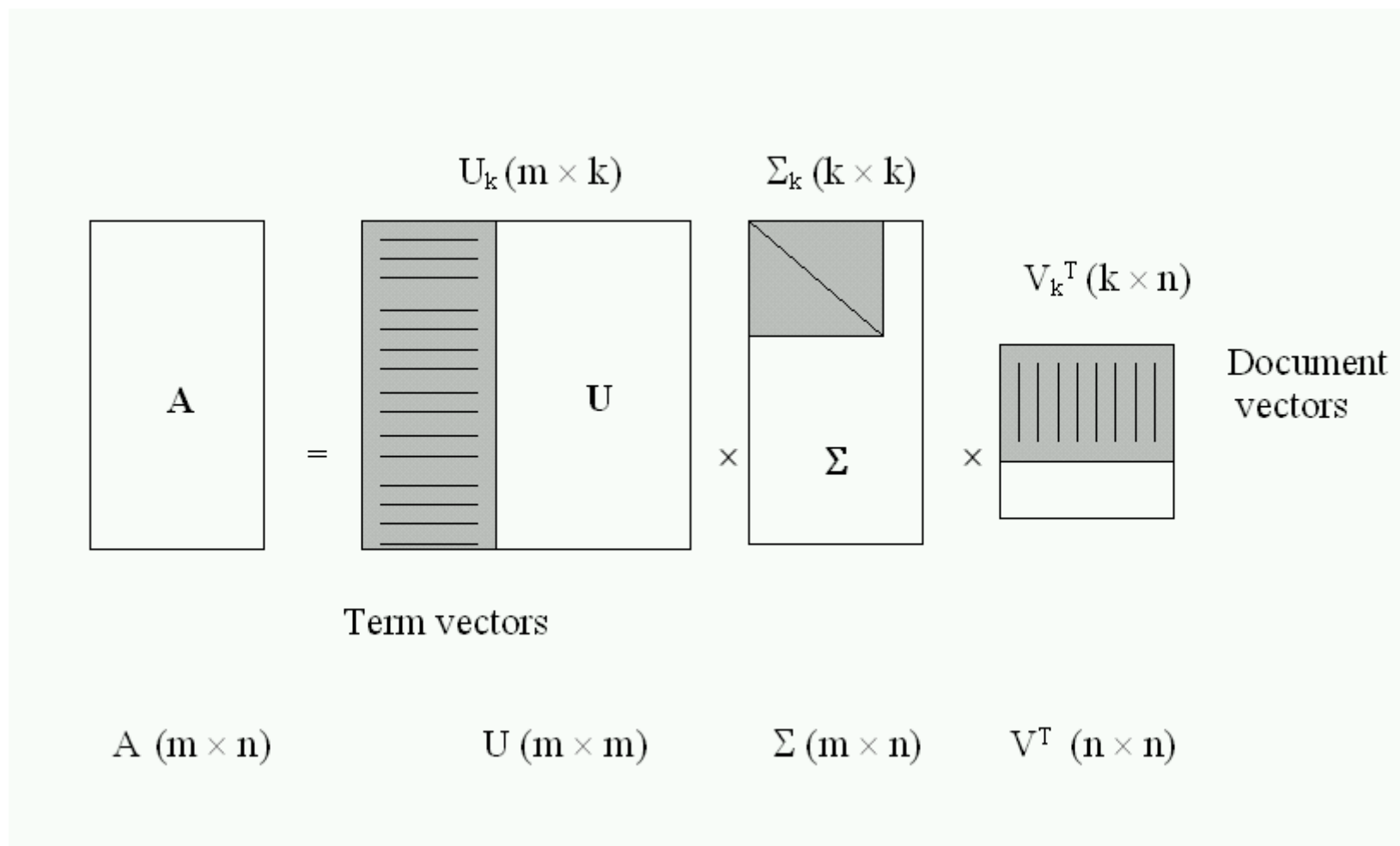
$U_k$  je matrica tipa  $m \times k$  čije stupce čini prvih  $k$  lijevih singularnih vektora matrice  $A$

$\Sigma_k$  je matrica tipa  $k \times k$  na čijoj je dijagonali vodećih  $k$  singularnih vrijednosti matrice  $A$

$V_k$  je matrica tipa  $n \times k$  čije stupce čini prvih  $k$  desnih singularnih vektora matrice  $A$

- Redci matrice  $U_k$  = indeksni pojmovi
- Redci matrice  $V_k$  = dokumenti

# Uloga matrica u krnjoj SVD



---

# Reprezentacije indeksnih pojmova i dokumenata

- Dokumenti su predstavljeni kao projekcije na prvih  $k$  svojstvenih vektora matrice  $AA^T$
- $AA^T$  je matrica sličnosti pojmova
- Svojstveni vektori koji odgovaraju najvećim svojstvenim vrijednostima sadrže informaciju o najvažnijim uzorcima varijabilnosti podataka
- Svojstveni vektori koji odgovaraju manjim svojstvenim vrijednostima su u smjeru manje varijabilnosti podataka koja može biti zanemarena ili se može interpretirati kao šum u podacima
- LSI je modificirana aplikacija metode glavnih komponenata za slučaj predstavljanja tekstualnih dokumenata

---

# Konceptno indeksiranje (CI)

- Indeksiranje korištenjem **konceptne dekompozicije (CD)**
- Konceptna dekompozicije je predstavljena 2001. godine

I.S.Dhillon, D.S. Modha: *Concept decomposition for large sparse text data using clustering*, Machine Learning, 42:1, 2001, pp. 143-175

# Konceptna dekompozicija

- **Prvi korak:** grupiranje (engl. clustering) matrice pojmova i dokumenata  $A$  u  $k$  grupa
- Algoritmi za grupiranje:
  - Sferični algoritam  $k$  srednjih vrijednosti (engl. spherical k-means algorithm)
  - Neizraziti algoritam  $k$  srednjih vrijednosti (engl. fuzzy k-means algorithm)
- Centroidi grupa = **konceptni vektori**
- **Konceptna matrica je matrica čiji stupci su konceptni vektori**

$c_j$  – centroid  $j$ -te grupe      $C_k = [c_1 \ c_2 \ \dots \ c_k]$

- **Drugi korak:** projekcija matrice na prostor razapet konceptnim vektorima
- **Konceptna dekompozicija**  $D_k$  matrice pojmova i dokumenata  $A$  je aproksimacija matrice  $A$  konceptnom matricom u smislu najmanjih kvadrata

$$D_k = C_k Z$$

$Z$  - rješenje problema najmanjih kvadrata

$$Z = (C_k^T C_k)^{-1} C_k^T A$$

- Redci matrice  $C_k =$  pojmovi
- Stupci matrice  $Z =$  dokumenti

---

# Primjer

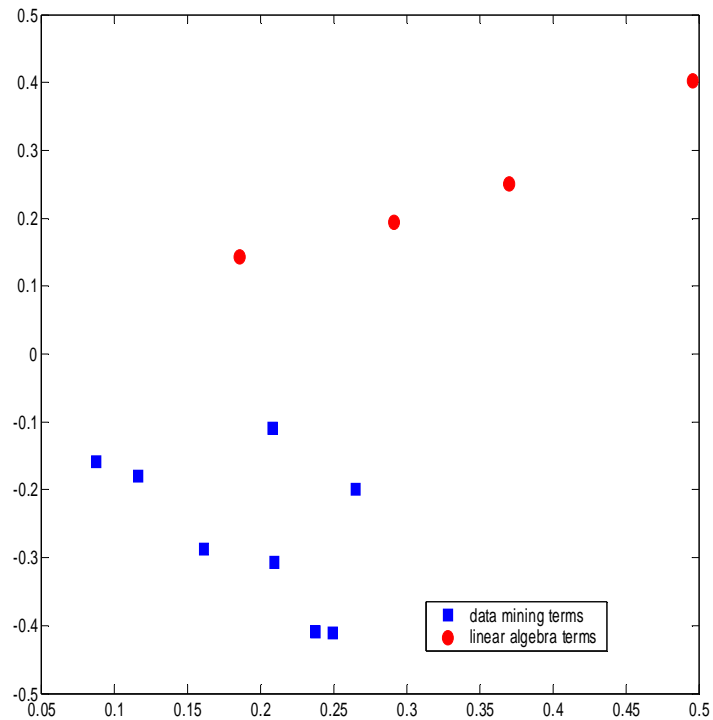
- Zbirka od 15 dokumenata (naslovi knjiga i članaka)
  - 9 iz područja dubinske analize podataka
  - 5 iz područja linearne algebre
  - 1 kombinacija ta dva područja (primjena linearne algebre u području dubinske analize podataka)
- Lista indeksnih pojmova je formirana
  - od pojmova sadržanih u barem 2 dokumenta
  - izbačeni su pojmovi sadržani u listi stop pojmova
  - pojmovi su svedeni na svoj osnovni oblik
- Na matricu pojmova i dokumenata je primjenjena
  - LSI metoda ( $k=2$ )
  - Metoda konceptnog indeksiranja ( $k=2$ )

# Indeksni pojmovi

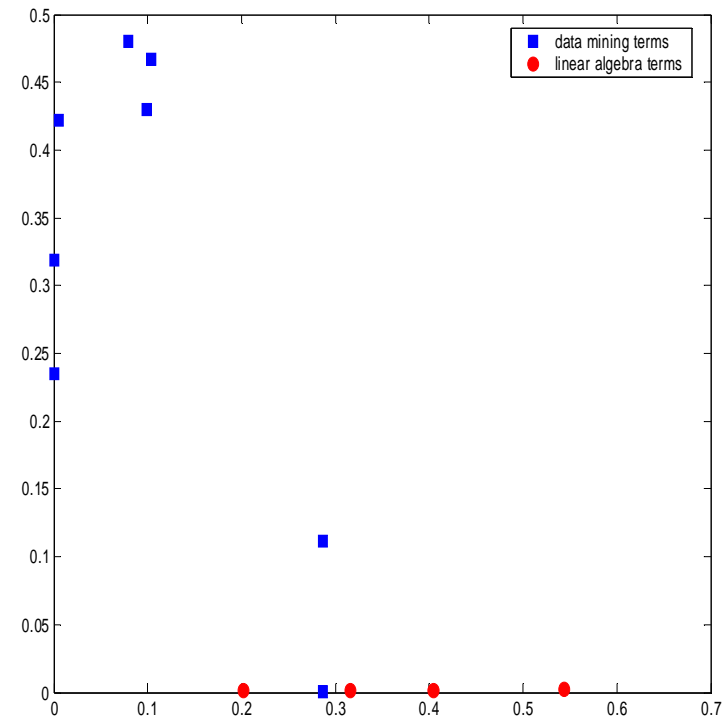
Pojmovi vezani uz dubinsku analizu (tekstualnih) podataka	Pojmovi vezani uz linearnu algebru	Neutralni pojmovi
Text	Linear	Analysis
Mining	Algebra	Application
Clustering	Matrix	Algorithm
Classification	Vector	
Retrieval	Space	
Information		
Document		
Data		

# Projekcije pojmova

## SVD



## Konceptna dekompozicija



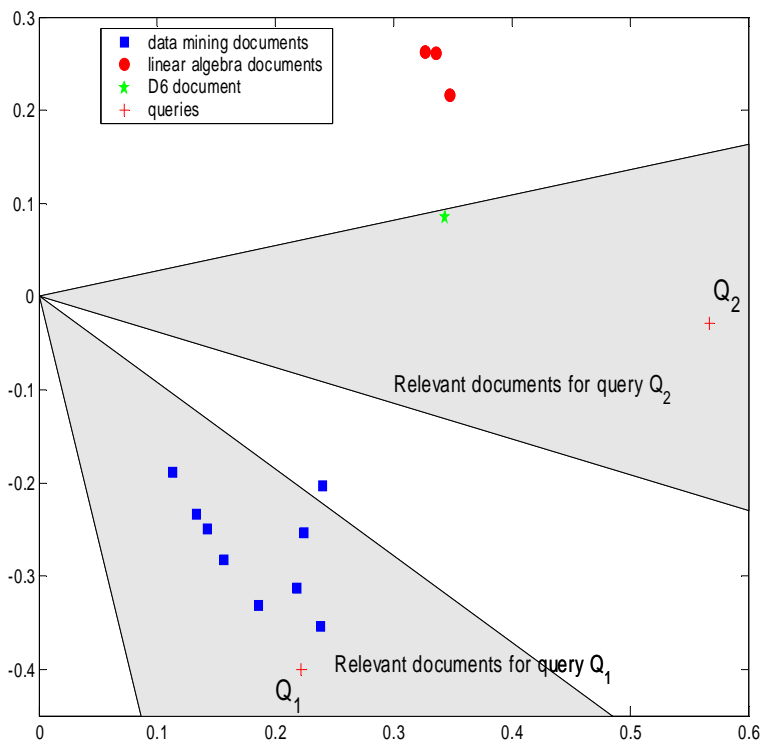
---

# Upiti

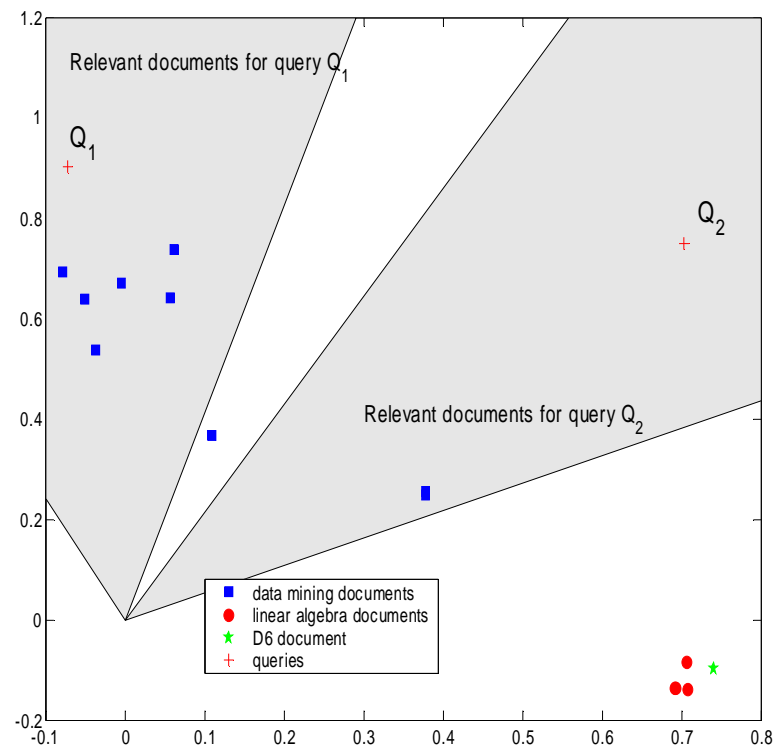
- Q1: Data mining
  - Relevantni dokumenti : Svi dokumenti vezani uz dubinsku analizu podataka
  
- Q2: Using linear algebra for data mining
  - Relevantan dokument: D6

# Projekcije dokumenata

## SVD



## Konceptna dekompozicija



---

# Eksperimentalne zbirke dokumenata

## ■ MEDLINE

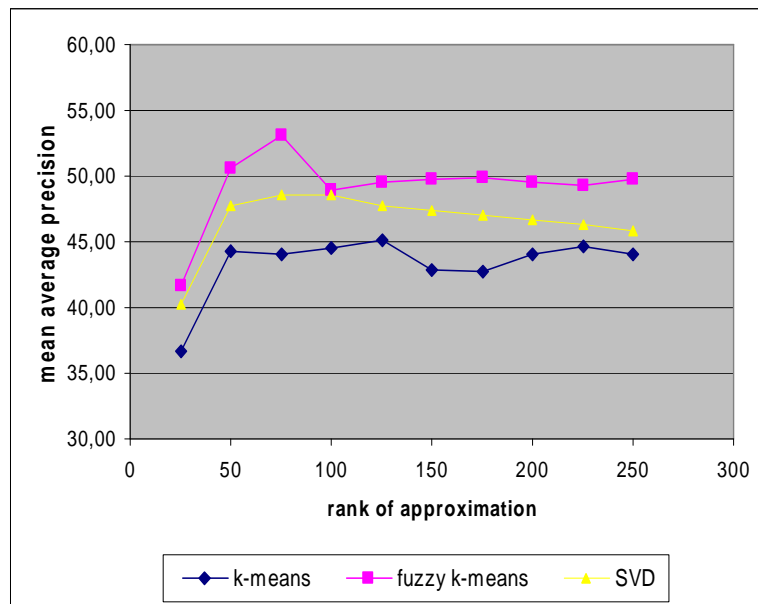
- Sažeci medicinskih znanstvenih članaka
- 1033 dokumenata
- 30 upita

## ■ CRANFIELD

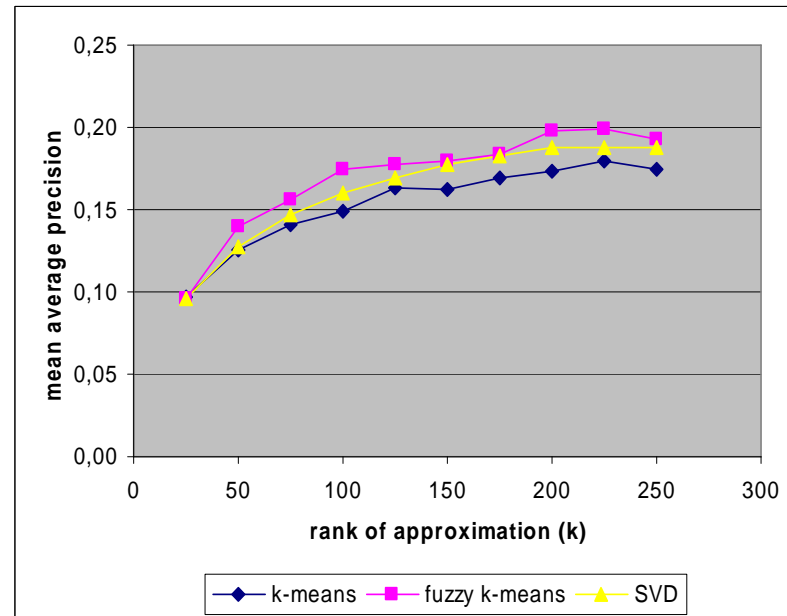
- Sažeci znanstvenih članaka iz područja aeronautike
- 1400 dokumenata
- 225 upita

# Preciznost pretraživanja na eksperimentalnim zbirkama dokumenata

- Srednja prosječna preciznost pretraživanja kod MVP:
  - MEDLINE : 43,54
  - CRANFIELD : 20,89



**MEDLINE**



**CRANFIELD**

## Problem kod konceptualnog indeksiranja: dodavanje novih dokumenata kod metoda LSI i CI

- Zbirke dokumenata su dinamičke: u njih se neprestano dodaju novi dokumenti ili se izbacuju stari
- Ako su dokumenti predstavljeni u prostoru snižene dimenzije kao kod metoda LSI i KI dodavanje reprezentacija novih dokumenata je problem
  - vektori na koje se vrši projekcija (singularni vektori, konceptni vektori) izračunati su na temelju cijele zbirke dokumenata i kod dodavanja novih dokumenata trebalo bi ponovo preračunavati SVD dekompoziciju ili konceptnu dekompoziciju
- Rješenje problema je u aproksimativnim reprezentacijama dodanih dokumenata

J. Dobša, B. Dalbello-Bašić, Approximate representation of textual documents in the concept space, *Informatica*, Vol. 31, No. 1, 2007., 21.-27.

---

# Algoritam za izračun aproksimativnih reprezentacija kod CI

- Da li je kod dodavanja aproksimativnih reprezentacija novih dokumenata potrebno dodavati nove pojmove?
- Razvijene su dvije metode za aproksimativno dodavanje novih dokumenata u konceptni prostor :
  - Projekcija dodanih dokumenata na postojeće konceptne vektore (**Metoda A**)
  - Projekcija dodanih dokumenata na postojeće konceptne vektore proširene u dimenzije dodanih pojmova (**Method B**)

# Proširena matrica pojmov a i dokumenata

Matrica

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix}$$

je proširena matrica pojmov a i dokumenata , pri čemu je  $\mathbf{A}_1$  matrica početnih dokumenata u prostoru početnih pojmov a

$\mathbf{A}_2$  matrica dodanih dokumenata u prostoru početnih pojmov a

$\mathbf{A}_3$  matrica početnih dokumenata u prostoru dodanih pojmov a

$\mathbf{A}_4$  matrica dodanih dokumenata u prostoru dodanih pojmov a

---

# Proširena konceptna matrica

- Neka je  $\mathbf{C}_1$  konceptna matrica čiji su stupci konceptni vektori matrice  $\mathbf{A}_1$
- Neka je  $\mathbf{C}_2$  matrica koja se sastoji od produžetaka konceptnih vektora u dimenzije dodanih pojmova
- Tada je

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix}$$

prošireni konceptni vektor

# Reprezentacije dokumenata u konceptnom prostoru

## Metoda A:

- Reprezentacije početnih dokumenata su dane s (1)
- Reprezentacije dodanih dokumenata su dane s (3)

## Metoda B:

- Reprezentacije početnih dokumenata su dane s(1)
- Reprezentacije dodanih dokumenata su dane s (3)+ $\alpha$ (4), gdje je  $\alpha \geq 1$

$$\begin{aligned} & \left( \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} \\ & \approx \begin{bmatrix} (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_1^T \mathbf{A}_1 + (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_2^T \mathbf{A}_3 & \vdots \\ (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_1^T \mathbf{A}_2 + (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_2^T \mathbf{A}_4 & \end{bmatrix} \\ & = [(1)+(2) \quad \vdots \quad (3)+(4)]. \end{aligned}$$

---

# Pokus

- Zbirka MEDLINE: 1033 dokumenata (sažetaka medicinskih članaka) i 35 upita
- Dokumenti su na slučajan način razdijeljeni u dva dijela: početni dokumenti i dodani dokumenti
- Početna lista indeksnih pojmova: sve riječi sadržane u najmanje dva dokumenta početne zbirke koje nisu na listi stop riječi
- Lista za cijelu zbirku ima 5940 indeksnih pojmova

# Rezultati pokusa

% dodanih dokumenata	Broj dodanih dok.	Broj dodanih pojmova	MAP Metoda A	MAP Metoda B, $\alpha=1.0$	MAP Metoda B, $\alpha=2.0$
0	0	0	54.99	54.99	54.99
10	104	456	51.98	52.20	52.37
20	208	753	54.96	55.10	55.23
30	311	1264	51.90	51.78	52.03
40	414	1673	50.84	50.60	51.64
50	517	2089	48.64	47.99	48.64
60	620	2696	44.26	44.08	45.49
70	723	3282	43.59	41.86	42.70
80	826	4024	39.87	40.56	43.74

---

# Diskusija rezultata

- Rezultati za Metodu B  $\alpha=2.0$  nisu signifikantno bolji od rezultata za Metodu A (t-test za zavisne uzorke,  $p=0.05$ )
- Dodavanje novih indeksnih pojmova ne popravlja rezultate: **reprezentacija tekstualnih podataka sadrži puno redundancije**
- Vrijednosti MAP za aproksimativne reprezentacije su 3-4% slabiji nego rezultati koji se postižu ponovnim preračunavanjem konceptne dekompozicije kada je postotak dodanih dokumenata do 40%
- MAP bitnije pada kada je broj dodanih dokumenata jednak ili veći od broj početnih dokumenata

---

**Hvala na pažnji!**