Semantičko pretraživanje informacija u tekstualnim dokumentima

Jasminka Dobša
Fakultet organizacije i informatike
Varaždin

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
  - Naglasak na model vektorskog prostora
- Proces indeksiranja i evaluacija pretraživanja informacija
- Problem sinonima i višeznačnica
- Mogućnosti za rješavanje tih problema
  - Naglasak na metodama snižavanja dimenzije vektorskog prostora: konceptualno indeksiranje
- Operativni problem kod snižavanja dimenzije vektorske reprezentacije dokumenata
- Smjernice za dalji rad

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 pojmova i dokumenata

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

Slika 1. Matrica pojmova i dokumenata

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$d_1 \quad d_2 \quad d_n$ (arrows ↓ ↓ ↓ above the columns)

$\leftarrow p_1$, $\leftarrow p_2$, $\vdots$, $\leftarrow p_m$ (arrows to the rows)

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

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

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

Algoritmi za svođenje riječi na osnovnu ili korijensku formu

- Postupak svođenja riječi na korijensku formu svodi se na uklanjanje prefiksa i sufiksa
- Važniju ulogu imaju sufiksi: većina algoritma za svođenje riječi na korijensku formu svodi se na uklanjanje sufiksa
- Najpoznatiji algoritam za engleski jezik je Porterov algoritam
- Na Fakultetu organizacije i informatike u tijeku je izrada algoritma za svođenje riječi na osnovni oblik (lemu) koji će se bazirati na principu strojnog učenja (Radošević, Dobša): takav algoritam za hrvatski jezik još ne postoji
- Algoritam za lematizaciju postoji za slovenski jezik (Mladenić, 2002)

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. fleksible 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**
- Neke od tehnika za rješavanje ovog problema su:
  - proširivanje korisničkog upita
  - konceptualno indeksiranje dokumenata
  - kod dokumenata na web-u: korištenje strukture poveznica (engl. link) između dokumenata

Primjer

- Zbirka od 15 dokumenata (naslovi knjiga)
  - 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 text mining: clustering, classification, and retrieval
D2	Automatic text processing: the transformation analysis and retrieval of information by computer
D3	Elementary linear algebra: A matrix approach
D4	Matrix algebra & its applications statistics and econometrics
D5	Effective databases for text & document management
D6	Matrices, vector spaces, and information retrieval
D7	Matrix analysis and applied linear algebra
D8	Topological vector spaces and algebras

Dokumenti 2/2

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

1. korisniku se predstavljaju dokumenti relevantni za njegov upit (temeljem algoritma za pretraživanja)
2. korisnik među vraćenim dokumentima bira relevantne (skup D_r) i nerelevantne (skup D_n)
3. težine indeksnih pojmova u vektoru upita se korigiraju

- Neka je početna reprezentacija korisničkog upita u MVP dana vektorom

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

- 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 α, β i γ 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 (α=1,β=1)

$$\mathbf{q}_{\text{mod}} = \mathbf{q} + \frac{1}{6}\left(\mathbf{a}_{15} + \mathbf{a}_{12} + \mathbf{a}_{14} + \mathbf{a}_{9} + \mathbf{a}_{11} + \mathbf{a}_{1}\right)$$

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

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*×*n* postoji **dekompozicija singularnih vrijednosti (engl. singular value decomposition, SVD)**

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

U ortogonalna matrica tipa *m*×*m* čiji stupci su lijevi singularni vektori matrice *A*

Σ dijagonalna matrica na čijoj dijagonali su singularne vrijednosti matrice *A* u padajućem redoslijedu

V ortogonalna matrica tipa *n*×*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)**

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

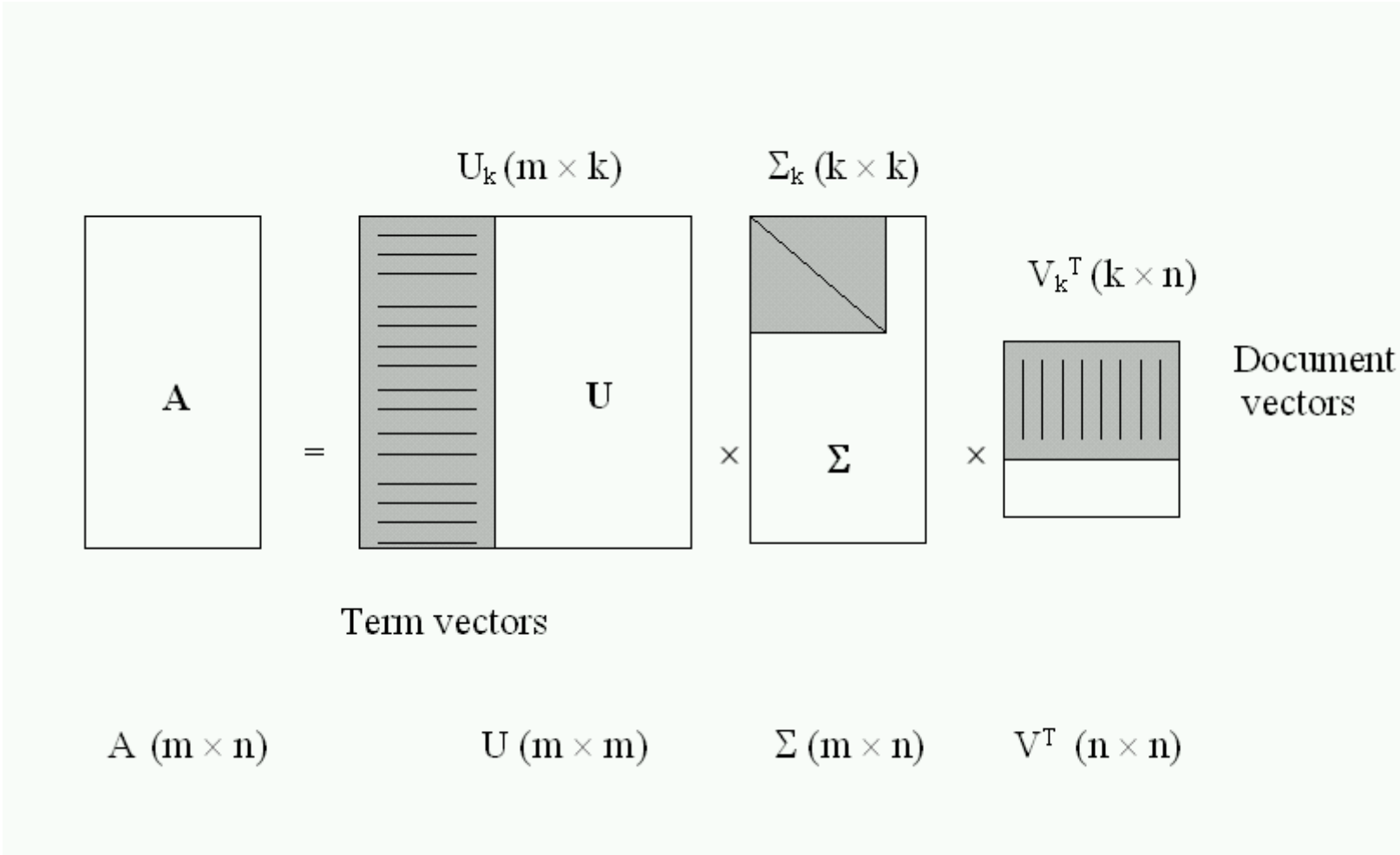
U_k je matrica tipa $m \times k$ čije stupce čini prvih k lijevih singularnih vektora matrice A

Σ_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 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 1/2

- **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)
- Sferični algoritam *k* srednjih vrijednosti je varijanta klasičnog algoritma k srednjih vrijednosti koja koristi činjenicu da su reprezentacije dokumenata i upita jedinične norme
- Centroidi grupa = **konceptni vektori**
- **Konceptna matrica je matrica čiji stupci su konceptni vektori**

$$C_k = \lfloor c_1 \quad c_2 \quad \cdots \quad c_k \rfloor$$

c_j – centroid j-te grupe

Konceptna dekompozicija 2/2

- **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

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

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

Primjer

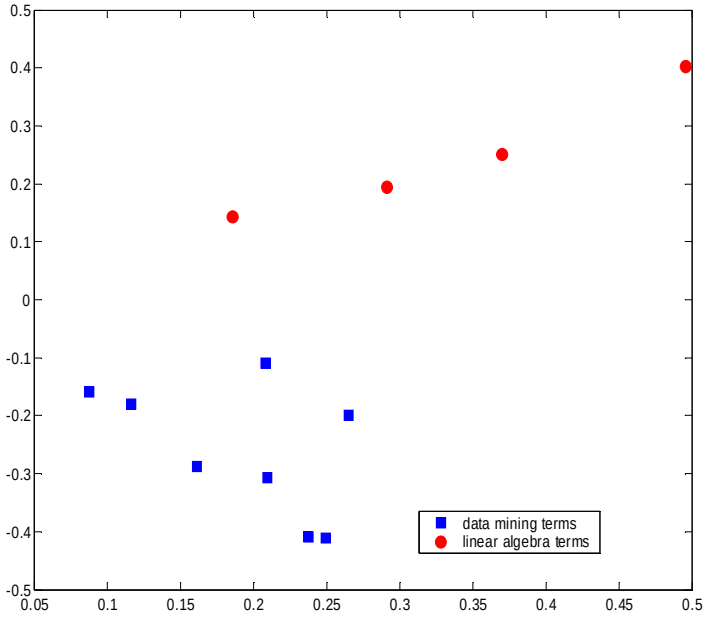
- Zbirka od 15 dokumenata (naslovi knjiga)
  - 9 iz područja dubinske analize podataka
  - 5 iz područja linearne algebre
  - 1 kombinacija ta dva područja (primjena linearne algebre u području dubinske analize podataka)
- Lista indeksnih pojmova je formirana
  - od pojmova sadržanih u barem 2 dokumenta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

Indeksni pojmovi

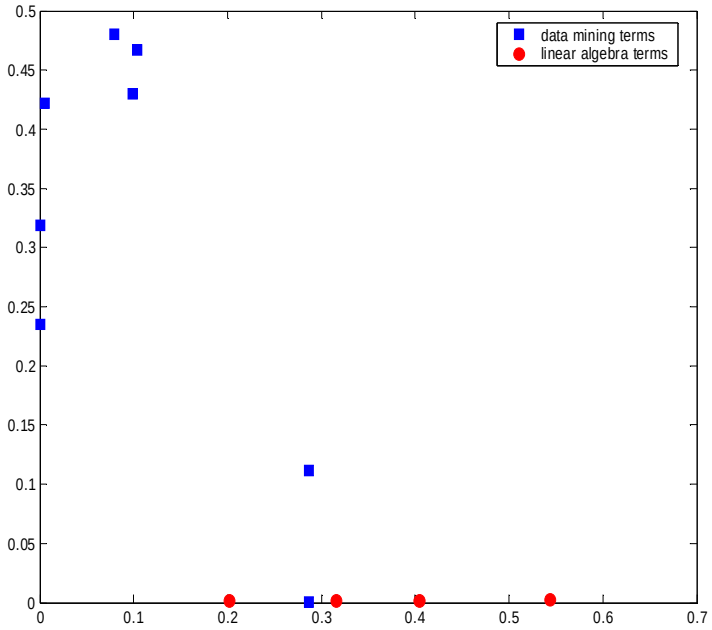
Pojmovi vezani uz dubinsku analizu (tekstualnih) podataka	Pojmovi vezani uz linearnu algebru	Neutralni pojmovi
Text	Linear	Analysis
Mining	Algebra	Application
Clustering	Matrix	Algorithm
Classification	Vector	
Retrieval	Space	
Information		
Document		
Data		

Projekcije pojmova

SVD



Konceptna dekompozicija

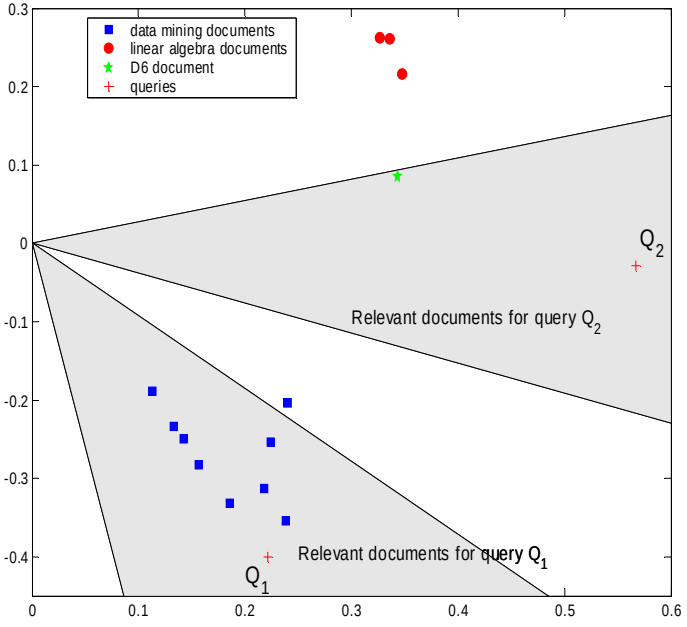


Upiti

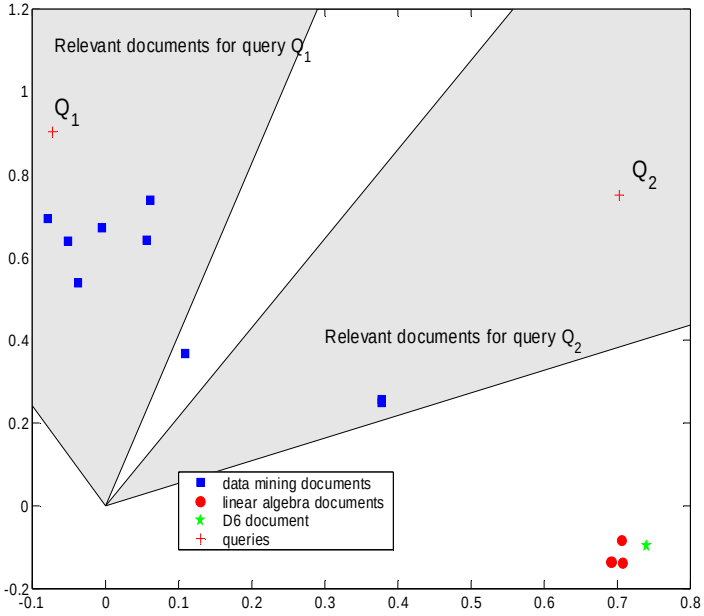
- Q1: Data mining
  - Relevantni dokumenti : Svi dokumenti vezani uz dubinsku analizu podataka
- Q2: Using linear algebra for data mining
  - Relevantan dokument: D6

Projekcije dokumenata

SVD



Konceptna dekompozicija

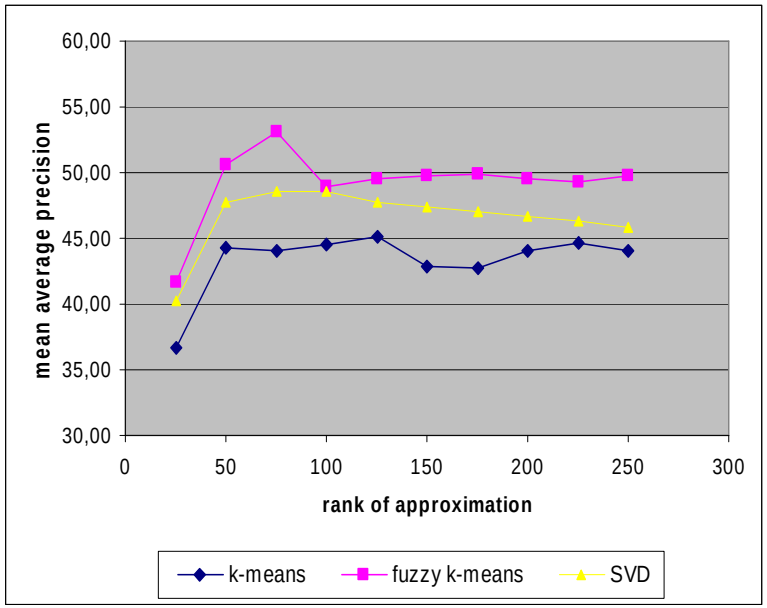


Eksperimentalne zbirke dokumenata

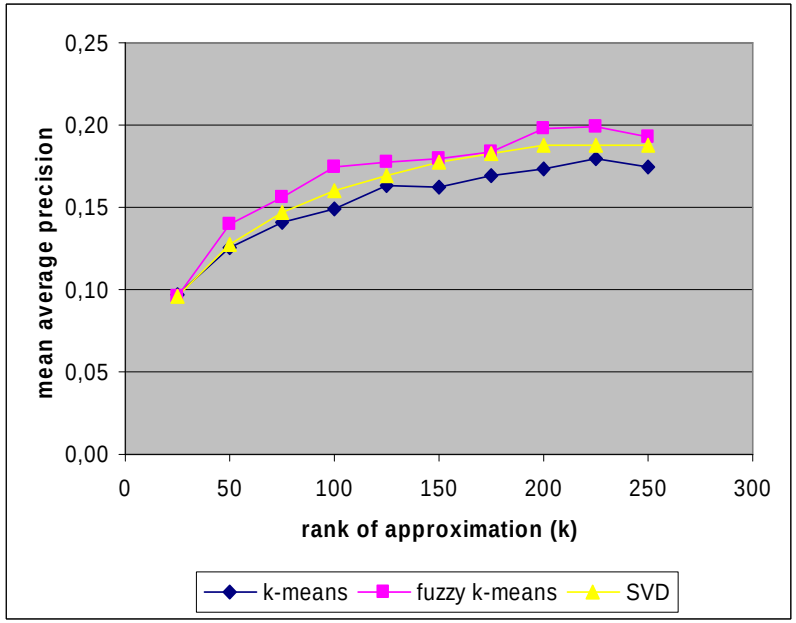
- MEDLINE
  - Sažeci medicinskih znanstvenih članaka
  - 1033 dokumenata
  - 30 upita
- CRANFIELD
  - Sažeci znanstvenih članaka iz područja aeronautike
  - 1400 dokumenata
  - 225 upita

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 KI

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

Buduća istraživanja

- Testiranje i nadogradnja algoritma za lematizaciju (zbirka www.hr)
- Razvoj i nadogradnja postojećih algoritama za snižavanja dimenzije vektorske reprezentacije dokumenata
- Razvoj algoritama za dodavanje novih dokumenata u prostoru snižene dimenzije
- Na Fakultetu organizacije i informatike je razvijen sustav za automatsku klasifikaciju tekstualnih dokumenata
  - Sustav implementira MVP
  - Prilagođen je hijerarhiji hrvatskih web dokumenata www.hr
  - Nadogradnja tog sustava za zadatak pretraživanja informacija