Primjena SVD rastava matrice u dohvatu informacija i kompresiji slike*

Zvonimir Ivančević†, Slobodan Jelić‡

Sveučilište Josipa Jurja Strossmayera, Odjel za matematiku
Trg Lj. Gaja 6, 31000 Osijek

9. travnja 2009.

Sažetak. U ovom radu ilustrirana je primjena dekompozicije matrice na singularne vrijednosti (SVD) u aproksimaciji matricom nižeg ranga. Kompresiju slike provodimo uzimajući u obzir samo prvih k singularnih vrijednosti u SVD rastavu matrice. Dohvat informacija objašnjen je u okviru modela indeksiranja skrivene semantike. Baza podataka odredena je matricom čiji stupci predstavljaju dokumente. Matrica baze podataka zamijenjena je aproksimacijom nižeg ranga, a pronalaženje relevantnih dokumenata na osnovu korisničkog upita svodi se na računanje kosinusa kuta izmedu vektora upita i vektora dokumenta. Što je kut manji, dokument je relevantniji za dani upit.

Ključne riječi: rastav matrice na singularne vrijednosti, kompresija slike, dohvat informacija, model vektorskog prostora, indeksiranje skrivene semantike, RGB slika, monokromatska slika, rang matrice

Abstract. In this paper the application of singular value decomposition (SVD) in low rank approximation of matrix is given. Image compression is performed by taking only first k singular values in SVD of matrix. Illustration of information retrieval is based on the concept of latent semantic indexing model (LSI). Database is determined by matrix whose columns represent documents. Database matrix is substituted with low rank approximation matrix, while searching for relevant documents with respect to user's query is based on calculating the cosine of angle between query and document vectors. Lesser is the angle, a document is more relevant for given query.

Key words: singular value decomposition, image compression, information retrieval, vector space model, latent semantic indexing, RGB image, monochrome image, rank of matrix

AMS Mathematical Classifications (2000): 15A03, 15A18, 68P20, 68U10, 94A08

*Seminarski rad iz kolegija Matematički praktikum, voditelj **prof. dr. sc. Rudolf Scitovski** i **doc. dr. sc. Kristian Sabo**

†e.mail: zivancev@mathos.hr
‡e.mail: sjelic@mathos.hr

1. Uvod

1.1. Kompresija slike i SVD

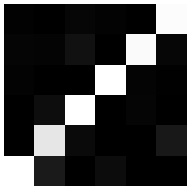
Kompresija slike je postupak sažimanja gdje nova slika zauzima manji prostor uz lošiju kvalitetu ali tako da svi bitni dijelovi ostanu uočljivi. Kompresija se javlja zbog potrebe uštede prostora, a možemo ju definirati i kao smanjenje broja pixela[1] koji služe za predstavljanje slike.

1.1.1. Način reprezentacije slike u računalu

Računalo prikazuje sliku kao matricu kod koje svaki element predstavlja intenzitet pojedinačnog pixela. Ukupan broj pixela korištenih za prikaz slike[2], naziva se *razlučivost zaslona*. Slika na računalu može biti prikazana na različite načine, pa stoga razlikujemo **crno-bijelu**, **monokromatsku** i **RGB sliku**[3]. Na Slici 1.1 prikazana je matrica crno-bijele slike veličine 6×6 pixela.

6	0	0	0	0	0	1
5	0	0	0	0	1	0
4	0	0	0	1	0	0
3	0	0	1	0	0	0
2	0	1	0	0	0	0
1	1	0	0	0	0	0
	1	2	3	4	5	6

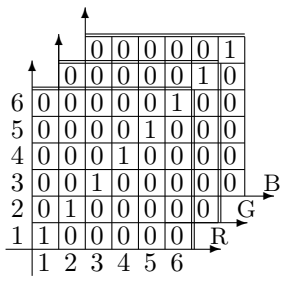
(a) Matrični prikaz slike



(b) Slika

Slika 1.1: Prikaz matrice za crno-bijelu sliku 6×6 pixela

Crno-bijela slika. Kod crno-bijele slike intenzitet pixela može biti 0 ili 1 (gdje 1 znači potpuno bijela, a 0 potpuno crna).
Monokromatska slika. Ako je intenzitet pixela između 0 i 1, onda kažemo da je slika monokromatska. Intenziteti pixela su različite nijanse sive boje (između crne i bijele).
RGB slika. RGB slika je troslojna, predstavljena pomoću tri matrice (vidi Sliku 1.2). U svaku od matrica pohranjuje se intenzitet pixela za jednu od tri boje (crvena, zelena i plava). Ovakav način prezentacije zovemo aditivni RGB model jer se boje dobivaju zbrajanjem pojedinih komponenti. Primjerice, vrijednost $(255, 255, 0)$ se preslikava u žutu boju najjačeg intenziteta, $(255, 255, 255)$ se preslikava u bijelu, a $(0, 0, 0)$ se preslikava u crnu.



Slika 1.2: Prikaz matrice za RGB sliku 6×6 pixela

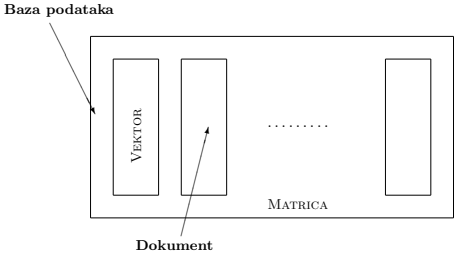
[1] eng. picture element
[2] iskazuje se kao broj vodoravnih puta broj okomitih zaslonskih točaka (npr. 1024 x 768, 1200 x 1024 i dr)
[3] eng. **R**ed **G**reen **B**lue

1.2. Dohvat informacija

U posljednje dvije decenije pojavom interneta i naglim razvojem informacijskih tehnologija razvilo se i posebno područje informacijske znanosti koje se bavi *dohvatom informacija*[4]. Osnovni zadatak znanosti o dohvatu informacija jeste razviti modele i metode za pretraživanje dokumenata u velikim bazama podataka, pronalaženje relevantnih informacija u dokumentima na osnovu zadanog upita, itd.

1.2.1. Model vektorskog prostora

U *modelu vektorskog prostora*[5] svaki dokument predstavljen je vektorom. Svaki element vektora predstavlja određenu važnost ili prisutnost pojma u opisu semantike tog dokumenta. Baza podataka je skup dokumenata opisanih određenim pojmovima. Možemo ju predstaviti matricom kod koje svaki vektor stupac prestavlja određeni dokument (vidi Sliku 1.3). Broj stupaca matrice jednak je broju dokumenata u bazi. Broj redaka matrice jednak je broju pojmova. *Indeksiranje* u određenom dokumentu predstavlja označavanje samo onih pojmova koji imaju ključnu važnost u opisu semantike tog dokumenta. Kako se indeksiranje provodi, ovisi o konkretnom modelu vektorskog prostora koji se koristi.



Slika 1.3: Model vektorskog prostora

1.2.2. Indeksiranje skrivene semantike

Sada ćemo formalno definirati matricu baze podataka.

Definicija 1.1 (**Matrica baze podataka**). *Neka je* $\mathbf{A} \in \mathbb{R}^{p\times d}$*, gdje je* d *ukupan broj dokumenata a* p *ukupan broj pojmova u bazi podataka. Matrica*

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2d} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{id} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{p1} & a_{12} & \dots & a_{pj} & \dots & a_{pd} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_j & \dots & \mathbf{a}_d \end{bmatrix}, \tag{1.1}$$

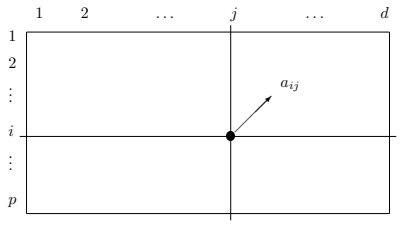
gdje je

$$a_{ij} = \begin{cases} 1 & , \quad \textit{ako } i\textit{-ti pojam opisuje } j\textit{-ti dokument} \\ 0 & , \quad \textit{inače} \end{cases}, \tag{1.2}$$

naziva se ***matrica baze podataka****. Vektori stupci* $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d$ *matrice* $\mathbf{A}$ *predstavljaju dokumente baze podataka.*

[4]eng. Information Retrieval
[5]eng. The Vector Space Model

Model *indeksiranja skrivene semantike*[6] predstavlja jednu varijantu modela vektorskog prostora u kome je matrica baze podataka aproksimirana matricom nižeg ranga. Vektorski prostor stupaca matrice baze podataka zamijenjen je jednim svojim potprostorom. Smanjenje ranga interpretiramo kao uklanjanje nepotrebnih informacija iz baze podataka.



Slika 1.4: Model indeksiranja skrivene semantike - LSI model

Način indeksiranja opisan u Definiciji 1.1 nije jedini i nužan. Općenito, elementi matrice baze podataka određuju važnost pojedinog pojma u dokumentu. Često su to relativne frekvencije pojma u dokumentu. Što je učestalost pojma veća, on je važniji u opisu semantike dokumenta. Osim relativnih frekvencija postoje i drugi načini indeksiranja kao što je dodjeljivanje lokalnih i globalnih težina pojmovima (vidi primjerice [3], [4]).

Primjer 1.1. *U Tablici 1 dana je baza podataka koja se sastoji od* 12 *naslova preuzetih iz kataloga* Gradske i sveučilišne knjižnice Osijek[7]. *Dokumente ćemo indeksirati pomoću ključnih riječi koje su podcrtane u nazivu svakog dokumenta. Odgovarajuća matrica baze podataka dana je u Tablici 2. Primjerice, dokument D6 s naslovom Kognitivno-bihevioralna terapija za psihijatrijske probleme sadrži ključne pojmove kognitivni, bihevioralni i terapija. To znači da će u matricu baze podataka na presjeku 6. stupca sa prvim, sedmim i jedanaestim retkom stojati jedinice (prema (1.2) iz Definicije 1.1). Na analogan način indeksiramo i ostale dokumente.*

Tablica 1: Različiti naslovi publikacija preuzeti iz kataloga GSKOS

Oznaka	Naslov
D1	Gramatika i predočavanje
D2	Klinička farmacija i terapija
D3	Kognitivna psihologija
D4	Kognitivna terapija
D5	Kognitivna znanost
D6	Kognitivno-bihevioralna terapija za psihijatrijske probleme
D7	Osnove kognitivne terapije
D8	Patologija i terapija tvrdih zubnih tkiva
D9	Psiholingvistika i kognitivna znanost u hrvatskoj primijenjenoj lingvistici
D10	Snaga intuicije
D11	Život bez boli
D12	Život bez droge

Osnovni cilj primjene LSI modela je olakšano pretraživanje baze podataka. Pretraživanje se vrši na osnovu korisničkog upita. *Korisnički upit* sastoji se od ključnih riječi za pretraživanje dokumenata. Vratimo se Primjeru 1.1. Korisnika mogu zanimati samo dokumenti koji imaju veze sa `kognitivnom terapijom`, tj. korisnički upit će sadržavati pojmove `kognitivni` i `terapija`.

[6] eng. Latent Semantic Indexing, dalje u tekstu LSI
[7] GSKOS - `http://baza.gskos.hr/anew/search.html`

Tablica 2: 18×12 matrica baze podataka iz Tablice 1

Pojmovi	Dokumenti											
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
Bihevioralni	0	0	0	0	0	1	0	0	0	0	0	0
Bol	0	0	0	0	0	0	0	0	0	0	1	0
Droga	0	0	0	0	0	0	0	0	0	0	0	1
Farmacija	0	1	0	0	0	0	0	0	0	0	0	0
Gramatika	1	0	0	0	0	0	0	0	0	0	0	0
Intuicija	0	0	0	0	0	0	0	0	0	1	0	0
Kognitivni	0	0	1	1	1	1	1	0	0	0	0	0
Lingvistika	0	0	0	0	0	0	0	0	1	0	0	0
Osnova	0	0	0	0	0	0	1	0	0	0	0	0
Patologija	0	0	0	0	0	0	0	1	0	0	0	0
Predočavanje	1	0	0	0	0	0	0	0	0	0	0	0
Primijenjeni	0	0	0	0	0	0	0	0	1	0	0	0
Psiholingvistika	0	0	0	0	0	0	0	0	1	0	0	0
Psihologija	0	0	1	0	0	0	0	0	0	0	0	0
Terapija	0	1	0	1	0	1	1	1	0	0	0	0
Tkiva	0	0	0	0	0	0	0	1	0	0	0	0
Znanost	0	0	0	0	1	0	0	0	0	0	0	0
Život	0	0	0	0	0	0	0	0	0	0	1	1

U LSI modelu upit je predstavljen vektorom[8] $\mathbf{q} \in \mathbb{R}^t$ gdje je t ukupan broj pojmova u bazi. Na mjestu pojmova koje sadržava upit nalazi se 1, inače stoji 0. U našem primjeru, upitu `kognitivna terapija` odgovara vektor

$$\mathbf{q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}^T. \tag{1.3}$$

Kada korisnik zada upit, on želi pronaći dokument u bazi koji najbolje odgovara tom upitu. Općenito, postoje različiti pristupi u određivanju najrelevantnijeg dokumenta. LSI model ima veoma jasan geometrijski pristup u rješavanju ovog problema. Dokument i upit predstavljeni su vektorima. To nam omogućuje da odredimo *kut između svakog dokumenta i upita. Dokument je relevantniji za korisnički upit što je kut između njih manji.* Kosinus kuta između vektora $\mathbf{a}_j$ i vektora upita $\mathbf{q}$ možemo odrediti prema formuli

$$\cos\theta_j = \frac{\mathbf{a}_j^T\mathbf{q}}{\|\mathbf{a}_j\|\|\mathbf{q}\|} = \frac{\sum_{i=1}^p a_{ij}q_i}{\sqrt{\sum_{i=1}^p a_{ij}^2}\sqrt{\sum_{i=1}^p q_i^2}}, \tag{1.4}$$

za $j = 1, 2, \ldots, d$, gdje je $\|\cdot\|$ Euklidska norma definirana kao $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^p x_i^2}$ za bilo koji $\mathbf{x} \in \mathbb{R}^p$.

2. SVD rastav matrice

SVD matrice[9] ima primjenu u rješavanju neuvjetovanog problema najmanjih kvadrata, procjene ranga matrice, itd. Važnost ove dekompozicije jeste u tome što ona otkriva strukturu matrice. Na osnovu SVD-a možemo odrediti rang matrice, njezinu jezgru i sliku, uvjetovanost, itd. U našem slučaju najvažnija je činjenica da SVD omogućava aproksimaciju nižeg ranga. U sljedećem teoremu dana je definicija i egzistencija SVD-a.

[8] upit je u LSI modelu isto što i dokument, samo ga definira korisnik

[9] eng. SVD - Singular Value Decomposition

Teorem 2.1 (Dekompozicija na singularne vrijednosti - SVD)**.** *Ako je* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *realna matrica, onda postoje ortogonalne matrice* $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m\times m}$ *i* $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m] \in \mathbb{R}^{n\times n}$ *te dijagonalna matrica* $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_s \end{bmatrix},$$

gdje je $s = \min\{m, n\}$ *i* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_s \geq 0$*, takve da je*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T.$$

Dokaz:

Vidi [8], 70 str., Teorem 2.5.2

□

Realne brojeve $\sigma_1, \sigma_2, \ldots, \sigma_s$ zovemo **singularne vrijednosti** matrice $\mathbf{A}$, matricu $\mathbf{\Sigma}$ singularnom matricom, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ **lijevim singularnim vektorima**, a $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ **desnim singularnim vektorima**.

Teorem 2.2. *Neka je* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *i* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *dekompozicija na singularne vrijednosti dana u Teoremu 2.1. Neka je* $r = r(\mathbf{A}) \leq s = \min\{m, n\}$*. Definiramo*

$$\mathbf{A}_k = \sum_{i=1}^{k} \mathbf{u}_i \sigma_i \mathbf{v}_i^T \tag{2.1}$$

gdje je $k < r$*. Tada je*

$$\min_{r(B)=k} \|\mathbf{A} - \mathbf{B}\|_2^2 = \|\mathbf{A} - \mathbf{A}_k\|_2^2 = \sigma_{k+1}.$$

Dokaz:

Vidi [8], 71 str.

□

U ovom radu, osnovni cilj jeste prikazati dvije primjene aproksimacije proizvoljne matrice $\mathbf{A}$ matricom nižeg ranga. U Teoremu 2.2 eksplicitno je određena najbolja aproksimacija ranga k i pri tome je u dana pogreška aproksimacije u smislu 2-matrične norme (vidi [6], 22. str., Lema 1.7). Budući da su singularne vrijednosti na dijagonali matrice $\mathbf{\Sigma}$ raspoređene u padajućem poretku, jasno je da će greška biti manja što je rang aproksimacije veći. Koristeći takvu ocjenu greške, možemo definirati i relativno smanjenje veličine matrice kao

$$\varrho_k = \frac{\|\mathbf{A} - \mathbf{A}_k\|_2}{\|\mathbf{A}\|_2} = \frac{\sigma_{k+1}}{\sigma_1}. \tag{2.2}$$

3. Primjena SVD rastava u kompresiji slike

Činjenica iz Teorema 2.2 može se upotrijebiti u svrhu kompresije slike.

(a) Monokromatska slika polarne lisice

(b) RGB slika polarne lisice

Slika 3.1: Slike polarne lisice

Monokromatska slika arktičke lisice (Slika 3.1a) zapisana je u obliku matrice veličine 283×192. Također možemo vidjeti i RGB sliku arktičke lisice (Slika 3.1b) koja je zapisana u obliku matrice $283 \times 192 \times 3$. Aproksimirajući matricu monokromatske slike koristeći redom prvih 3, 5, 10, 12, 15 i 20 singularnih vrijednosti u Formuli 2.1 dobivamo sljedeće kompresirane slike (Slika 3.2).

(a) $k = 3$

(b) $k = 5$

(c) $k = 10$

(d) $k = 12$

(e) $k = 15$

(f) $k = 20$

Slika 3.2: Kompresija monokromatske slike polarne lisice za $k = 3, 5, 10, 12, 15, 20$

Sa Slike 3.2 za $k = 20$ i $k = 15$ kompresije su gotovo identične originalu, a za njihov prikaz je potreban značajno manji broj pixela (vidi Tablicu 3). Za kompresiju ranga 20 potrebno je 9520 pixela što je $17,52\%$ od veličine originala.

Tablica 3: Broj pixela kompresirane monokromatske slike polarne lisice za $k = 3, 5, 10, 12, 15, 20$

	Original	$k = 3$	$k = 5$	$k = 10$	$k = 12$	$k = 15$	$k = 20$
crno-bijela slika (broj pixela)	54336	1428	2380	4760	5712	7140	9520

Aproksimirajući matricu slike u boji (RGB slike) koristeći redom prvih 3, 5, 10, 12, 15 i 20 singularnih vrijednosti u Formuli 2.1 dobivamo sljedeće kompresirane slike (Slika 3.3).

U Tablici 4 dane su veličine kompresiranih monokromatskih i RGB slika.

(a) $k = 3$ (b) $k = 5$ (c) $k = 10$

(d) $k = 12$ (e) $k = 15$ (f) $k = 20$

Slika 3.3: Kompresija RGB slike polarne lisice za $k = 3, 5, 10, 12, 15, 20$

Tablica 4: Broj pixela kompresirane RGB slike polarne lisice za $k = 3, 5, 10, 12, 15, 20$

	Original	$k = 3$	$k = 5$	$k = 10$	$k = 12$	$k = 15$	$k = 20$
RGB slika	8,53 KB	6,51 KB	6,95 KB	7,48 KB	7,71 KB	7,93 KB	8,21 KB
Monokromatska slika	7,51 KB	5,18 KB	5,59 KB	6,17 KB	6,43 KB	6,62 KB	6,97 KB

4. Primjena SVD rastava u dohvatu informacija

Neka je $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ rastav na singularne vrijednosti iz Teorema 2.1. Vektor upita $\mathbf{q}$ uspoređivat ćemo sa stupcima aproksimacije $\mathbf{A}_k$ matrice baze podataka $\mathbf{A}$, koja je dana u (2.1). Istu relaciju možemo zapisati i na sljedeći način

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T, \tag{4.1}$$

gdje je $\mathbf{U}_k = [\mathbf{u}_1\mathbf{u}_2\dots\mathbf{u}_k] \in \mathbb{R}^{m\times k}$ matrica koja se sastoji od prvih k stupaca matrice $\mathbf{U}$, $\mathbf{\Sigma}_k \in \mathbb{R}^{k\times k}$ dijagonalna matrica koja na dijagonali ima samo prvih k singularnih vrijednosti matrice $\mathbf{A}$ i $\mathbf{V}_k = [\mathbf{v}_1\mathbf{v}_2\dots\mathbf{v}_k] \in \mathbb{R}^{n\times k}$ matrica koja se sastoji od prvih k stupaca matrice $\mathbf{V}$.

Umjesto vektora a_j za $j = 1, 2, \dots, d$ u formuli (1.4) koristit ćemo j-ti stupac matrice $\mathbf{A}_k$ koji se može zapisati kao $\mathbf{A}_k\mathbf{e}_j$ gdje je $\mathbf{e}_j = [0, \dots, 1, \dots, 0]^T$ j-ti vektor kanonske baze od $\mathbb{R}^p$. Koristeći (4.1) u (1.4) dobivamo da je

$$\cos\theta_j = \frac{(\mathbf{A}_k\mathbf{e}_j)^T\mathbf{q}}{\|\mathbf{A}_k\mathbf{e}_j\|_2\|\mathbf{q}\|_2} = \frac{(\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T\mathbf{e}_j)^T\mathbf{q}}{\|\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T\mathbf{e}_j\|_2\|\mathbf{q}\|_2}. \tag{4.2}$$

Propozicija 4.1. *Neka je $(\mathbb{R}^n, +, \cdot)$ normirani vektorski prostor s normom $\|\cdot\|_n$ za koju vrijedi $\|\mathbf{x}\|_n = \sqrt{\mathbf{x}^T\mathbf{x}}$, $\mathbf{x} \in \mathbb{R}^n$ i $(\mathbb{R}^m, +, \cdot)$ normirani vektorski prostor s normom $\|\cdot\|_m$ za koju vrijedi $\|\mathbf{y}\|_m = \sqrt{\mathbf{y}^T\mathbf{y}}$, $\mathbf{y} \in \mathbb{R}^m$. Tada vrijede sljedeće tvrdnje*

a) Ako je $\mathbf{Q} \in \mathbb{R}^{m\times n}$ matrica takva da je $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n \in \mathbb{R}^{n\times n}$, onda je $\|\mathbf{Q}\mathbf{x}\|_m = \|\mathbf{x}\|_n$ za svaki $\mathbf{x} \in \mathbb{R}^n$.

b) Ako je $\mathbf{V} \in \mathbb{R}^{m\times n}$ matrica takva da je $\mathbf{V}\mathbf{V}^T = \mathbf{I}_m \in \mathbb{R}^{m\times m}$, onda je $\|\mathbf{V}^T\mathbf{y}\|_n = \|\mathbf{y}\|_m$ za svaki $\mathbf{y} \in \mathbb{R}^m$.

Dokaz:

a) Neka je $\mathbf{Q} \in \mathbb{R}^{m\times n}$ matrica takva da je $\mathbf{Q}^T\mathbf{Q} = \mathbf{I} \in \mathbb{R}^{n\times n}$ i $\mathbf{x} \in \mathbb{R}^n$. Tada vrijedi

$$\|\mathbf{Q}\mathbf{x}\|_m^2 = (\mathbf{Q}\mathbf{x})^T(\mathbf{Q}\mathbf{x}) = \mathbf{x}^T\mathbf{Q}^T\mathbf{Q}\mathbf{x} = \mathbf{x}^T\mathbf{I}_n\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_n^2,$$

odakle tvrdnja slijedi zbog proizvoljnosti od $\mathbf{x}$.

b) Dokazuje se analogno a).

□

Budući da je $\mathbf{U} \in \mathbb{R}^{p\times p}$ ortogonalna matrica, $\mathbf{U}_k \in \mathbb{R}^{p\times k}$ zadovoljava uvjet iz tvrdnje a) u Propoziciji 4.1. Odatle slijedi da relaciju u (4.2) možemo zapisati kao

$$\cos\theta_j = \frac{\mathbf{e}_j^T \mathbf{V}_k \mathbf{\Sigma}_k (\mathbf{U}_k^T \mathbf{q})}{\|\mathbf{\Sigma}_k \mathbf{V}_k^T \mathbf{e}_j\|_2 \|\mathbf{q}\|_2}.$$

Uvođenjem supstitucije $\mathbf{s}_j = \mathbf{\Sigma}_k \mathbf{V}_k^T \mathbf{e}_j$ za $j = 1, \ldots, d$ dobivamo da je

$$\cos\theta_j = \frac{\mathbf{s}_j^T (\mathbf{U}_k^T \mathbf{q})}{\|\mathbf{s}_j\|_2 \|\mathbf{q}\|_2}. \tag{4.3}$$

Sada ćemo dati geometrijsku interpretaciju formule (4.3).

- *Geometrijska interpretacija vektora* $\mathbf{s}_j$ *za* $j = 1, \ldots, d$

  Relaciju
  $$\mathbf{U}_k \mathbf{s}_j = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \mathbf{e}_j = \mathbf{A}_k \mathbf{e}_j$$
  možemo zapisati i na sljedeći način
  $$\mathbf{A}_k \mathbf{e}_j = \mathbf{U}_k \mathbf{s}_j = \sum_{i=1}^{k} s_{ij} \mathbf{u}_i^{(k)},$$
  gdje je $\mathbf{u}_i^{(k)}$ i-ti stupac matrice $\mathbf{U}_k$, a s_{ij} je i-ta komponenta vektora $\mathbf{s}_j$. Odatle vidimo da $\mathbf{s}_j$ sadrži koordinate j-tog stupca matrice $\mathbf{A}_k$ u potprostoru razapetom sa stupcima matrice $\mathbf{U}_k$. Stupci matrice $\mathbf{U}_k$ čine jednu ortonormiranu bazu u prostoru stupaca matrice $\mathbf{A}_k$.

- *Geometrijska interpretacija vektora* $\mathbf{U}_k\mathbf{q}$

  Na sličan način interpretiramo vektor $\mathbf{U}_k\mathbf{q}$. Najprije uočimo da je $\mathbf{U}_k\mathbf{U}_k^T\mathbf{q}$ projekcija vektora upita na prostor stupaca matrice $\mathbf{A}_k$.

  $$\mathbf{U}_k \mathbf{U}_k^T \mathbf{q} = \sum_{i=1}^{k} ((\mathbf{u}_i^{(k)})^T \mathbf{q}) \mathbf{u}_i^{(k)}.$$

  Odatle vidimo da vektor $\mathbf{U}_k^T\mathbf{q}$ sadrži koordinate projekcije vektora upita $\mathbf{q}$ na potprostor razapet stupcima matrice $\mathbf{U}_k$.

Iz prethodnih razmatranja jasno je da se vektor upita zamjenjuje projekcijom na vektorski prostor stupaca matrice $\mathbf{A}_k$ (gdje stupci od $\mathbf{U}_k$ čine jednu ortonormiranu bazu) a vektor j-og dokumenta zamjenjuje se j-im stupcem matrice $\mathbf{A}_k$ koji je u (4.3) prikazan kao linearna kombinacija stupaca matrice $\mathbf{U}_k$. Takav pristup računanja omogućava sljedeće:

- značajnu uštedu resursa[10] kod velikih baza podataka
- Vektori $\mathbf{s}_j$ računaju se samo jednom za cijelu bazu podataka
- Nije potrebno računati sve singularne vrijednosti, lijeve i desne singularne vektore već samo prvih k.

[10] računalna memorija i vrijeme

Primjer 4.1. *Neka je zadana baza podataka matricom iz Tablice 2. Pretpostavimo da korisnik želi pronaći sve dokumente čiji je sadržaj vezan za kognitivnu terapiju te da je njegov upit predstavljen vektorom* $\mathbf{q}$ *iz (1.3).*

Na Slici 4.1 predstavljen je rezultat pretraživanja na originalnoj matrici baze podataka iz Tablice 2. Najrelevantniji dokument je onaj koji ima najveću ocjenu iz (4.3). U našem slučaju to je dokument s nazivom `Kognitivna terapija`. Prilikom određivanja relevantnih dokumenata za zadani upit, određuje se granična vrijednost ocjene[11]. Ako je ocjena veća od te vrijednosti, dokument je relevantan za dani upit, inače smatramo ga irelevantnim. Primjerice, uzmemo li da je granična vrijednost 0.5, relevantni dokumenti za upit `kognitivna terapija` bit će dokumenti s nazivom `Kognitivna terapija`, `Kognitivno-bihevioralna terapija za psihijatrijske probleme` i `Osnove kognitivne terapije`.

`Kognitivna terapija`	1
`Kognitivno-bihevioralna terapija za psihijatrijske probleme`	0.8165
`Osnove kognitivne terapije`	0.8165
`Klinička farmacija i terapija`	0.5
`Kognitivna psihologija`	0.5
`Kognitivna znanost`	0.5
`Patologija i terapija tvrdih zubnih tkiva`	0.40825
`Snaga intuicije`	1.6188e-017
`Psiholingvistika i kognitivna znanost u hrvatskoj primijenjenoj lingvistici`	7.2139e-018
`Gramatika i predočavanje`	0
`Život bez droge`	-2.3797e-016
`Život bez boli`	-5.0218e-016

Slika 4.1: Rezultati upita `kognitivna terapija` na originalnoj matrici baze podataka

Usporedimo li dobivene rezultate i rezultatima sa Slike 4.2 gdje je matrica baze podataka zamijenjena aproksimacijom ranga 9, vidjet ćemo da su i dodatni dokumenti ocijenjeni kao relevantni. Ovdje ne možemo sa sigurnošću tvrditi da li postoji stvarna semantička veza između upita i dokumenata `Kognitivna psihologija`, `Kognitivna znanost` i `Klinička farmacija i terapija`. To je jedno od svojstava modela indeksiranja skrivene semantike. Jasno je da eventualne greške u određivanju relevantnih dokumenata dolaze zbog aproksimacije matrice baze podataka. Općenito je nemoguće odgovoriti koja aproksimacija[12] je najbolja. Koristeći (2.2) dobivamo da je

$$\varrho_9 = \frac{\sigma_{10}}{\sigma_1} = \frac{1}{2.9745} = 0.3362,$$

odnosno, aproksimacijom ranga 9 dolazi do smanjenja u veličini matrice za 33.62%

U sljedećem primjeru vidjet ćemo da takvo smanjenje veličine za određene upite smanjuje relevantnost vraćenih dokumenata.

Primjer 4.2. *Pretpostavimo da korisnik iz baze podataka dane u Tablici 1 želi preuzeti dokumente koji imaju veze sa* `bihevioralnom terapijom`.

Odgovarajući vektor upita (slično kao u (1.3)) je

$$\mathbf{w} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}^T. \tag{4.4}$$

Iako je

$$\varrho_6 = \frac{\sigma_7}{\sigma_1} = \frac{1}{2.9745} = 0.3362,$$

vidimo da je `Klinička farmacija i terapija` vraćen kao najrelevantniji dokument za upit w. Međutim, stvarni najrelevantniji dokument je `Kognitivno-bihevioralna terapija za psihijatrijske probleme`. U ovom slučaju vidimo da je aproksimacija rangom 6 neprihvatljiva iako uzrokuje isti gubitak informacija (33.62%).

[11] u engleskoj literaturi uglavnom se navodi kao *cut-off* ili *threshold*

[12] ovdje se misli na rang aproksimacije

Kognitivna terapija	1
Osnove kognitivne terapije	0.88806
Kognitivno-bihevioralna terapija za psihijatrijske probleme	0.88806
Kognitivna psihologija	0.55489
Kognitivna znanost	0.55489
Klinička farmacija i terapija	0.5429
Patologija i terapija tvrdih zubnih tkiva	0.43487
Snaga intuicije	1.8644e-017
Psiholingvistika i kognitivna znanost u hrvatskoj primijenjenoj lingvistici	7.5769e-018
Gramatika i predočavanje	0
Život bez droge	-3.3855e-016
Život bez boli	-5.9468e-016

Slika 4.2: Rezultati upita `kognitivna terapija` na aproksimaciji ranga 9 matrice baze podataka

Klinička farmacija i terapija	0.9616
Osnove kognitivne terapije	0.83169
Kognitivno-bihevioralna terapija za psihijatrijske probleme	0.83169
Kognitivna terapija	0.81266
Patologija i terapija tvrdih zubnih tkiva	0.548
Kognitivna psihologija	0.086524
Kognitivna znanost	0.086524
Život bez droge	7.403e-016
Život bez boli	7.3153e-016
Gramatika i predočavanje	0
Psiholingvistika i kognitivna znanost u hrvatskoj primijenjenoj lingvistici	-4.1303e-017
Snaga intuicije	-4.1303e-017

Slika 4.3: Rezultati upita `bihevioralna terapija` na aproksimaciji ranga 6 matrice baze podataka

Literatura

[1] Ioannis Antonellis and Efstratios Gallopoulos, *Exploring term-document matrices from matrix models in text mining*, 2006.

[2] Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup, *Matrices, vector spaces, and information retrieval*, SIAM Rev. **41** (1999), no. 2, 335–362.

[3] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review **37** (1995), no. 4, 573–595.

[4] Michael W. Berry and Ricardo D. Fierro, *Low-rank orthogonal decompositions for information retrieval applications*, Numerical Linear Algebra with Applications **3** (1996), 301–328.

[5] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science **41** (1990), 391–407.

[6] James Demmel, *Applied numerical linear algebra*, Society for Industrial and Applied Mathematics, 1997.

[7] C. Ding and J. Ye, *2-dimensional singular value decomposition for 2d maps and images*, Proc. of SIAM Int. Conf. on Data Mining (SDM 2005), 2005.

[8] Gene H. Golub and Charles F. Van Loan, *Matrix computations (johns hopkins studies in mathematical sciences)*, The Johns Hopkins University Press, October 1996.

[9] S. K. M. Wong, Ziarko Wojciech, and Patric C. N. Wong, *Generalized vector space model in information retrieval*, 1985, pp. 18–25.