SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mate Mihaljević

TENZORSKE REPREZENTACIJE HIPERGRAFA I PRIMJENE U ANALIZI DRUŠTVENIH MREŽA

Diplomski rad

Voditelj rada:
Izv. prof. dr. sc. Luka Grubišić

Zagreb, Rujan, 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Roditeljima, Jozi i Dražani, za svo strpljenje i podršku.

Sadržaj

Uvod

U ovom radu će se dati pregled tenzora višeg reda te pokazati njihova primjena na analizu društvene mreže Twitter.

U prvom poglavlju iznesene su definicije i notacija vezana uz tenzore kao i povijesni pregled. Također, navedene su osnovne operacije nad tenzorima. U drugom dijelu prvog poglavlja obraćena je pažnja na određivanje ranga tenzora i različitost svojstava matričnog i tenzorskog ranga. Na kraju prvog poglavlja dotaknuta je PARAFAC dekompozicija tenzora te je naveden trenutno radni algoritam za određivanje PARAFAC dekompozicije tenzora tzv. ALS algoritam.

Nakon odrađene teorijske pripreme u prvom poglavlju, u drugom poglavlju je modeliran grafa društvene mreže, oblikovani su tenzori te primjenjena PARAFAC dekompozicija za rangiranje autoriteta, prvo na jednostavnom primjeru, a zatim na složenijem. Središnji dio drugog poglavlja je posvećen TweetRank modelu, dok je u završnom dijelu iznesen kratak pregled algoritama povezanih sa TweetRank modelom.

Poglavlje 1

Tenzor

1.1 Definicije i notacija

U ovom poglavlju izložit ćemo pregled tenzora višeg reda i jedne njihove dekompozicije poznate kao PARAFAC dekompozicija. Iako je bilo aktivno istraživanje dekompozicija i modela tenzora u posljednja 4 desetljeća (odnosno, primjena dekompozicija na nizove podataka kako bi se dobila i objasnila njihova svojstva), vrlo malo ili nimalo od tog rada je objavljeno u matematičkim časopisima.

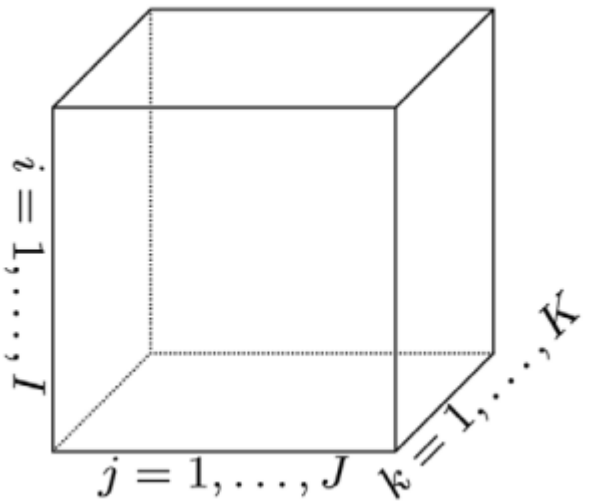
Definicija 1.1.1. *Tenzor je višedimenzionalni niz. Formalnije, N-smjerni tenzor ili tenzor N-tog reda je element tenzorskog produkta N vektorskih prostora gdje svaki od njih ima vlastiti koordinatni sustav.*

Definicija 1.1.2. *Red tenzora je broj dimenzija, također znanih kao smjerovi ili modovi.*

Tenzor trećeg reda ima tri indeksa kao što možemo vidjeti na Slici 1.1. Tenzor prvog reda je vektor, tenzor drugog reda je matrica, a tenzori trećeg ili višeg reda se nazivaju tenzori višeg reda.

Korištena notacija je konzistentna, koliko je to moguće, s terminologijom prijašnjih publikacija na području dekompozicije tenzora te je ujedno bliska za primjenu matematičarima. Ovdje korištena notacija je veoma slična onoj koju predlaže Kiers [30].

Vektori (tenzori reda jedan) se označavaju podebljanim malim slovima npr. $\mathbf{a}$. Matrice (tenzori reda dva) se označavaju podebljanim velikim slovima npr. $\mathbf{A}$. Tenzori višeg reda se označavaju podebljanim *Euler script* slovima npr. $\mathcal{X}$. Skalari se označavaju malim slovima npr. a. I-ti element vektora $\mathbf{a}$ je označen sa a_i, element (i, j) matrice $\mathbf{A}$ je označen s a_{ij}, a element (i, j, k) tenzora $\mathcal{X}$ trećeg reda je označen s x_{ijk}. Indeksi su obično u rasponu od 1 do njihove najveće vrijednosti npr. $i = 1, \ldots, I$. N-ti element u nizu je obilježen s eksponentom u zagradama npr. $\mathbf{A}^{(n)}$ označava n-tu matricu u nizu.



Slika 1.1: Tenzor trećeg reda: $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$.

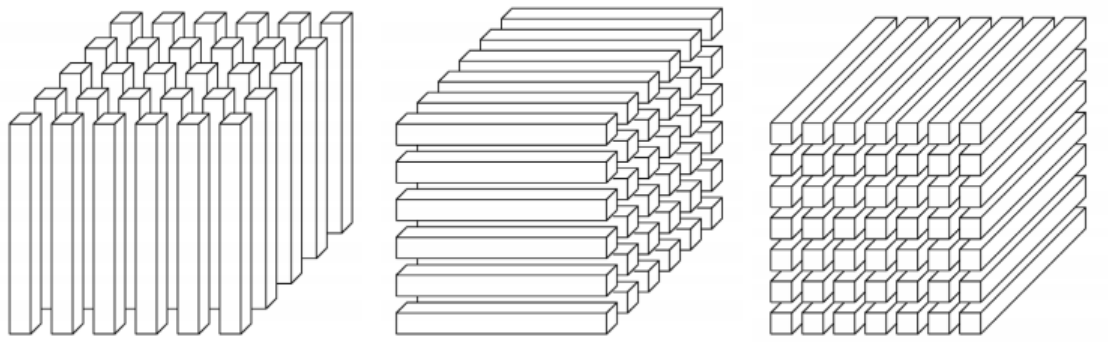
Fiksiranjem podskupa indeksa formiraju se podnizovi. Za matrice to su retci i stupci. Dvotočka se koristi za označavanje svih elemenata moda, tj. smjera. Prema tome, j-ti stupac od $\mathbf{A}$ je označen s $\mathbf{a}_{:j}$, a i-ti redak matrice $\mathbf{A}$ je označen s $\mathbf{a}_{i:}$. Alternativno, j-ti stupac matrice, $\mathbf{a}_{:j}$, se može označavati kompaktnije na sljedeći način $\mathbf{a}_j$.

Definicija 1.1.3. *Tenzorsko vlakno je jednodimenzionalni fragment tenzora dobiven fiksiranjem svih indeksa osim jednoga.*

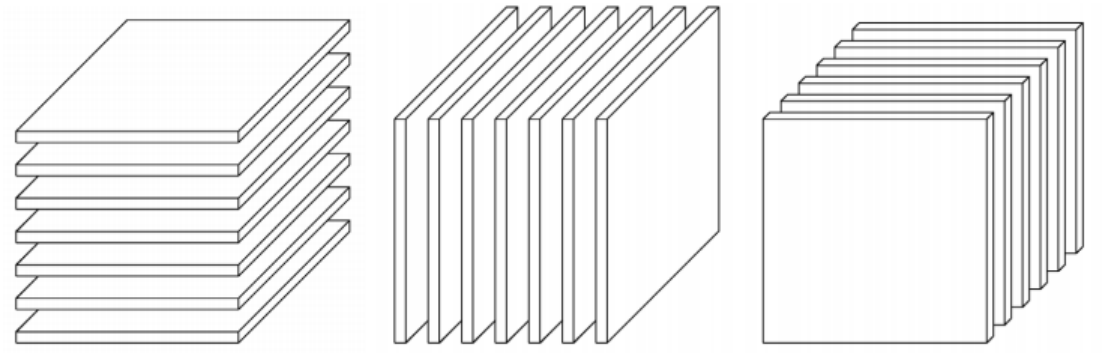
Tenzorska vlakna su analogoni višeg reda matričnih redaka i stupaca. Matrični stupac je tenzorsko vlakno moda 1, a matrični redak je tenzorsko vlakno moda 2. Tenzori trećeg reda imaju stupčana, retčana i poprečna tenzorska vlakna. Označavamo ih redom $\mathbf{x}_{:jk}$, $\mathbf{x}_{i:k}$, $\mathbf{x}_{ij:}$, vidi Sliku 1.2. Kada tenzorska vlakna izvučemo iz tenzora uvijek pretpostavljamo da su orijentirana kao stupčani vektor.

Definicija 1.1.4. *Presjeci tenzora su dvodimenzionalni fragmenti tenzora dobiveni fiksiranjem svih osim dvaju indeksa tenzora.*

Slika 1.3 prikazuje horizontalne, lateralne i frontalne presjeke tenzora $\mathcal{X}$ trećeg reda označene redom s $\mathbf{X}_{i::}$, $\mathbf{X}_{:j:}$, $\mathbf{X}_{::k}$.



Slika 1.2: Stupčana($\mathbf{x}_{:jk}$), retčana($\mathbf{x}_{i:k}$) i poprečna($\mathbf{x}_{ij:}$) tenzorska vlakna tenzora trećeg reda



Slika 1.3: Horizontalni($\mathbf{X}_{i::}$ ili $\mathbf{X}_i$), lateralni($\mathbf{X}_{:j:}$ ili $\mathbf{X}_j$) i frontalni($\mathbf{X}_{::k}$ ili $\mathbf{X}_k$) tenzorski presjeci tenzora trećeg reda

Definicija 1.1.5. *Norma tenzora* $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ *je kvadratni korijen sume svih kvadrata njegovih elemenata, tj.*

$$||\mathcal{X}|| = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1 i_2 \ldots i_N}^2}.$$

Tenzorska norma je analogna Frobeniusovoj matričnoj normi, koja se za matricu $\mathbf{A}$ označava sa $||\mathbf{A}||$.

Definicija 1.1.6. *Skalarni produkt dvaju tenzora istih dimenzija* $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ *je suma*

umnožaka njihovih elemenata, tj.

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1 i_2 \ldots i_N} y_{i_1 i_2 \ldots i_N}.$$

Odmah slijedi $\langle \mathcal{X}, \mathcal{X} \rangle = \|\mathcal{X}\|^2$.

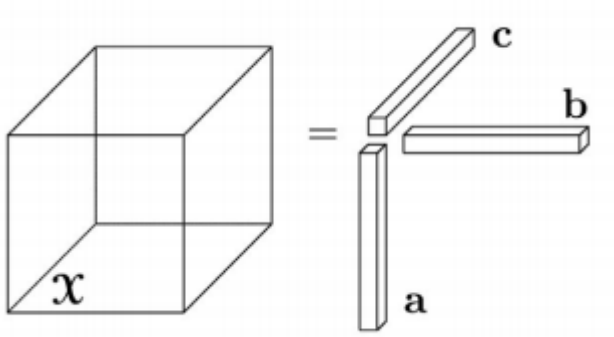
Definicija 1.1.7. *N-smjerni tenzor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ je ranga jedan ako se može zapisati kao vanjski produkt N vektora, tj.*

$$\mathcal{X} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \cdots \circ \boldsymbol{a}^{(N)}.$$

Simbol "$\circ$" predstavlja vanjski vektorski produkt. To znači da je svaki element tenzora produkt odgovarajućih elemenata vektora:

$$x_{i_1 i_2 \ldots i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N} \text{ za sve } 1 \leq i_n \leq I_n.$$

Slika 1.4 ilustrira $\mathcal{X} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$, tenzor trećeg reda ranga jedan.



Slika 1.4: Tenzor trećeg reda ranga jedan, $\mathcal{X} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$. Element (i, j, k) tenzora $\mathcal{X}$ je dan s $x_{ijk} = a_i b_j c_k$.

Definicija 1.1.8. *Tenzor nazivamo kubičnim ako je svaki njegov mod iste veličine, tj. $\mathcal{X} \in \mathbb{R}^{I \times I \times \cdots \times I}$. Kubični tenzor nazivamo supersimetričnim ako njegovi elementi ostanu konstantni upotrebom bilo koje permutacije indeksa.*

Npr. tenzor trećeg reda $\mathcal{X} \in \mathbb{R}^{I \times I \times I}$ je supersimetričan ako

$$x_{ijk} = x_{ikj} = x_{jik} = x_{jki} = x_{kij} = x_{kji} \text{ za sve } i, j, k = 1, \ldots, I.$$

Tenzori mogu biti (djelomično)simetrični u 2 ili više modova. Npr. trosmjerni tenzor $\mathcal{X} \in \mathbb{R}^{I\times I\times K}$ je simetričan u modovima jedan i dva ako su frontalni presjeci simetrični, tj. ako

$$\mathbf{X}_k = \mathbf{X}_k^T \text{ za sve } k = 1, \ldots, K.$$

Definicija 1.1.9. *Tenzor $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ je dijagonalan ako je $x_{i_1 i_2 \cdots i_N} \neq 0$ jedino ako je $i_1 = i_2 = \cdots = i_N$.*

Definicija 1.1.10. *Matricizacija (transformiranje tenzora u matricu) je proces prebacivanja elemenata N-smjernog niza u matricu.*

Npr. 2×3×4 tenzor se može prikazati kao 6×4 matrica ili 3×8 matrica korištenjem permutacija elemenata. U ovom radu promatramo samo poseban slučaj matricizacije *n*-moda jer je to jedina relevantna forma za nastavak rada. Općenitija obrada matricizacije se može pronaći u Kolda [32]. *N*-mod matricizacija tenzora $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ se označava s $\mathbf{X}_{(n)}$ te određuje da tenzorska vlakna *n*-moda budu stupci rezultirajuće matrice. Formalna notacija je nespretna. Tenzorski element $(i_1, i_2, \ldots, i_N)$ unosi se na mjesto matričnog elementa (i_n, j) gdje

$$j = 1 + \sum_{\substack{k=1\\k\neq n}}^{N} (i_k - 1)J_k \text{ pri čemu je } J_k = \prod_{\substack{m=1\\m\neq n}}^{k-1} I_m.$$

Koncept je lakše razumjeti na primjeru. Neka su frontalni presjeci tenzora $\mathcal{X} \in \mathbb{R}^{3\times4\times2}$

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix} \tag{1.1}$$

Tada su tri *n*-mod matricizacije

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix},$$

$$\mathbf{X}_2 = \begin{bmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{bmatrix},$$

$$\mathbf{X}_2 = \begin{bmatrix} 1 & 2 & 3 & \cdots & 11 & 12 \\ 13 & 14 & 15 & \cdots & 23 & 24 \end{bmatrix}.$$

Različita literatura ponekad različito slaže stupce n-mod matricizacije; usporedi s De Lathauwer i ostali [15] i Kiers [30]. Generalno, određena permutacija stupaca nije važna dok god smo konzistentni tokom računanja; vidi Kolda [32].

Na kraju, primjetimo da je moguće vektorizirati tenzor. U gornjem primjeru vektorizirana verzija je

$$vec(\mathcal{X}) = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 23 \\ 24 \end{bmatrix}.$$

Tenzore možemo međusobno množiti, iako su notacija i simboli za to mnogo složeniji nego za matrice. Za potpuni pregled množenja tenzora pogledajte Bader i Kolda [3]. Ovdje ćemo razmotriti samo n-mod tenzorski produkt, tj. množenje tenzora matricom (ili vektorom) u modu n.

Definicija 1.1.11. *N-mod (matrični) produkt tenzora $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ s matricom $\boldsymbol{U} \in \mathbb{R}^{J \times I_n}$ se označava s $\mathcal{X} \times_n \boldsymbol{U}$ i veličine je $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$. Po elementima imamo:*

$$(\mathcal{X} \times_n \boldsymbol{U})_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_N} u_{j i_n}.$$

Svako n-mod tenzorsko vlakno se množi matricom $\mathbf{U}$. Ideja se može izraziti u pojmovima nematriciziranih tenzora:

$$\mathcal{Y} = \mathcal{X} \times_n \mathbf{U} \Longleftrightarrow \mathbf{Y}_{(n)} = \mathbf{U}\mathbf{X}_{(n)}.$$

N-mod produkt tenzora s matricom je povezan s promjenom baze u slučaju kad tenzor definira multilinearni operator. Pogledajmo sljedeći primjer. Neka je $\mathcal{X}$ tenzor definiran kao u 1.1 i neka je $\mathbf{U} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$. Tada je produkt $\mathcal{Y} = \mathcal{X} \times_1 \mathbf{U} \in \mathbb{R}^{2 \times 4 \times 2}$ pri čemu su

$$\mathbf{Y}_1 = \begin{bmatrix} 22 & 49 & 76 & 103 \\ 28 & 64 & 100 & 136 \end{bmatrix}, \mathbf{Y}_2 = \begin{bmatrix} 130 & 157 & 184 & 211 \\ 172 & 208 & 244 & 280 \end{bmatrix}.$$

Navodimo nekoliko činjenica povezanih sa n-mod matričnim produktima. Redoslijed množenja je nevažan za različite modove u nizu množenja, tj.

$$\mathcal{X} \times_m \mathbf{A} \times_n \mathbf{B} = \mathcal{X} \times_n \mathbf{B} \times_m \mathbf{A} (m \neq n).$$

Ako su modovi isti imamo

$$\mathcal{X} \times_n \mathbf{A} \times_n \mathbf{B} = \mathcal{X} \times_n (\mathbf{B}\mathbf{A}).$$

Definicija 1.1.12. *N-mod (vektorski) produkt tenzora* $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *s vektorom* $\boldsymbol{v} \in \mathbb{R}^{I_n}$ *se označava s* $\mathcal{X} \bar{\times}_n \boldsymbol{v}$*. Rezultat je reda* $N-1$*, tj. veličine je* $I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N$*. Po elementima imamo*

$$(\mathcal{X} \bar{\times}_n \boldsymbol{v})_{i_1 \cdots i_{n-1} i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_N} v_{i_n}.$$

Ideja je izračunati skalarni produkt svakog n-mod tenzorskog vlakna s vektorom $\mathbf{v}$. Pogledajmo sljedeći primjer. Neka je tenzor $\mathcal{X}$ definiran kao u 1.1 te neka je $\mathbf{v} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}^T$. Tada

$$\mathcal{X} \bar{\times}_2 \mathbf{v} = \begin{bmatrix} 70 & 190 \\ 80 & 200 \\ 90 & 210 \end{bmatrix}.$$

Kada je u pitanju n-mod vektorsko množenje, prednost je bitna jer se red međurezultata mijenja. Drugim riječima,

$$\mathcal{X} \bar{\times}_m \mathbf{a} \bar{\times}_n \mathbf{b} = (\mathcal{X} \bar{\times}_m \mathbf{a}) \bar{\times}_{n-1} \mathbf{b} = (\mathcal{X} \bar{\times}_n \mathbf{b}) \bar{\times}_m \mathbf{a} \text{ za } m < n.$$

Nekoliko matričnih produkata je važno u odjeljcima koji slijede, stoga ih ovdje definiramo.

Definicija 1.1.13. *Kroneckerov produkt matrica* $\boldsymbol{A} \in \mathbb{R}^{I \times J}$ *i* $\boldsymbol{B} \in \mathbb{R}^{K \times L}$ *se označava sa* $\boldsymbol{A} \otimes \boldsymbol{B}$*. Rezultat je matrica veličine* $(IK) \times (JL)$ *definirana s*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1J}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2J}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\boldsymbol{B} & a_{I2}\boldsymbol{B} & \cdots & a_{IJ}\boldsymbol{B} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \boldsymbol{a}_1 \otimes \boldsymbol{b}_2 & \boldsymbol{a}_1 \otimes \boldsymbol{b}_3 & \cdots & \boldsymbol{a}_J \otimes \boldsymbol{b}_{L-1} & \boldsymbol{a}_J \otimes \boldsymbol{b}_L \end{bmatrix}.$$

Definicija 1.1.14. *Khatri-Raov [45] produkt je Kroneckerov produkt matrica sa jednakim brojem stupaca. Neka su date matrice* $\boldsymbol{A} \in \mathbb{R}^{I \times K}$ *i* $\boldsymbol{B} \in \mathbb{R}^{J \times K}$*, njihov Khatri-Raoov produkt se označava s* $\boldsymbol{A} \odot \boldsymbol{B}$*. Rezultat je matrica veličine* $(IJ) \times K$ *definirana s*

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_K \otimes \boldsymbol{b}_K \end{bmatrix}.$$

Ako su $\mathbf{a}$ i $\mathbf{b}$ vektori, onda su Khatri-Raoov i Kroneckerov produkt identični, tj. $\mathbf{a} \otimes \mathbf{b} = \mathbf{a} \odot \mathbf{b}$.

Definicija 1.1.15. *Hadamardov produkt je matrični produkt po elementima. Neka su date matrice* $\boldsymbol{A}$ *i* $\boldsymbol{B}$ *istih dimenzija* $I \times J$*. Njihov Hadamardov produkt se označava s* $\boldsymbol{A} * \boldsymbol{B}$*.*

Rezultat je također veličine $I \times J$ i definiran je s

$$\boldsymbol{A} * \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1J}b_{1J} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2J}b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \cdots & a_{IJ}b_{IJ} \end{bmatrix}.$$

Ovi matrični produkti imaju sljedeća svojstva [45][51] koja će se pokazati korisnim u nastavku:

$$\begin{aligned} (\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) &= \mathbf{AC} \otimes \mathbf{BD}, \\ (\mathbf{A} \otimes \mathbf{B})^{\dagger} &= \mathbf{A}^{\dagger} \otimes \mathbf{B}^{\dagger}, \\ \mathbf{A} \odot \mathbf{B} \odot \mathbf{C} &= (\mathbf{A} \odot \mathbf{B}) \odot \mathbf{C} = \mathbf{A} \odot (\mathbf{B} \odot \mathbf{C}), \\ (\mathbf{A} \odot \mathbf{B})^{T}(\mathbf{A} \odot \mathbf{B}) &= \mathbf{A}^{T}\mathbf{A} * \mathbf{B}^{T}\mathbf{B}, \\ (\mathbf{A} \odot \mathbf{B})^{\dagger} &= ((\mathbf{A}^{T}\mathbf{A}) * (\mathbf{B}^{T}\mathbf{B}))^{\dagger}(\mathbf{A} \odot \mathbf{B})^{T}. \end{aligned} \tag{1.2}$$

Ovdje $\mathbf{A}^{\dagger}$ označava Moore-Penroseov pseudoinverz od $\mathbf{A}$ [20].

Kao primjer korisnosti Kroneckerovog produkta pogledajmo sljedeće. Neka je $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ i $\mathbf{A}^{(n)} \in \mathbb{R}^{J_n \times I_n}$ za sve $n \in \{1, \ldots, N\}$. Tada za bilo koji $n \in \{1, \ldots, N\}$ imamo

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_N \mathbf{A}^{(N)} \Longleftrightarrow \mathbf{Y}_{(n)} = \mathbf{A}^{(n)} \left(\mathbf{A}^{N} \otimes \cdots \otimes \mathbf{A}^{(n+1)} \otimes \mathbf{A}^{(n-1)} \otimes \cdots \otimes \mathbf{A}^{1} \right)^{T}.$$

Za dokaz tvrdnje vidi Kolda [32].

1.2 Rang tenzora i CANDECOMP/PARAFAC dekompozicija

Hitchcock [25][26] je 1927. godine predložio ideju višemjesnog oblika tenzora, odnosno, izražavanje tenzora kao konačne sume tenzora ranga jedan; Cattell [11][12] je 1944. godine predložio ideju paralelne proporcionalne analize i ideju višestrukih osi za analizu. Koncept je postao popularan nakon svog trećeg uvođenja, 1970. godine u psihometrijskoj zajednici, u obliku CANDECOMP (eng. *canonical decomposition*) po Carroll i Chang [10] i PARAFAC (eng. *parallel factors*) po Harashman [21]. Mi se referiramo na CANDECOMP/PARAFAC dekompoziciju kao CP po Kiers [30]. Möcks [39] neovisno otkriva CP u kontekstu istraživanja mozga i naziva ga topografski komponentni model.

CP dekompozicija faktorizira tenzor u sumu komponenata tenzora ranga jedan. Npr. neka je dan tenzor trećeg reda $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, želimo ga aproksimirati kao

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{1.3}$$

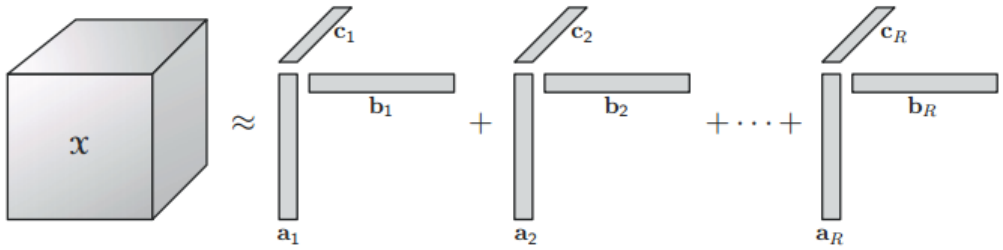
gdje je R pozitivni cijeli broj i $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, $\mathbf{c}_r \in \mathbb{R}^K$ za $r = 1, \ldots, R$, tj. želimo odrediti

$$\min_{\hat{\mathcal{X}}} \|\mathcal{X} - \hat{\mathcal{X}}\| \text{ uz } \hat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$

Po elemntima, aproksimaciju 1.3 možemo zapisati kao

$$x_{ijk} \approx \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \text{ za } i = 1, \ldots, I, \; j = 1, \ldots, J, \; k = 1, \ldots, K.$$

To je ilustriano Slikom 1.5.



Slika 1.5: CP dekompozicija trosmjernog tenzora

Matrice rastava se odnose na kombinaciju vektora iz komponenenata ranga jedan, odnosno, $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_R \end{bmatrix}$ i slično za $\mathbf{B}$ i $\mathbf{C}$. Koristeći ove definicije, aprksimacijski problem možemo zapisati na sljedeći način

$$\min_{\hat{\mathbf{A}}} \|\mathbf{X}_{(1)} - \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^T\|_F, \text{ gdje je } \hat{\mathbf{A}} = \mathbf{A} \cdot diag(\boldsymbol{\lambda}),$$

tj. 1.3 se može aproksimirati u matričnoj formi:

$$\begin{aligned} \mathbf{X}_{(1)} &\approx \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^T, \\ \mathbf{X}_{(2)} &\approx \hat{\mathbf{B}}(\mathbf{C} \odot \mathbf{A})^T, \\ \mathbf{X}_{(3)} &\approx \hat{\mathbf{C}}(\mathbf{B} \odot \mathbf{A})^T. \end{aligned} \tag{1.4}$$

Trosmjerni model je ponekad zapisan u uvjetima frontalnih presjeka od $\mathcal{X}$ (vidi Sliku 1.3):

$$\mathbf{X}_k \approx \mathbf{A}\mathbf{D}^{(k)}\mathbf{B}^T \text{ , gdje } \mathbf{D}^{(k)} \equiv diag(\mathbf{c}_{k:}) \text{ za } k = 1, \ldots, K.$$

Analogone aproksimacije možemo napisati i za horizontalne i lateralne presjeke. Generalno, izraze po presjecima nije lako proširiti za više od tri dimenzije. Slijedeći Kolda [32] (također vidi Kruskal [34]) CP model se može sažeto prikazati kao

$$\mathcal{X} \approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$

Često je korisno pretpostaviti da su stupci od $\mathbf{A}$, $\mathbf{B}$ i $\mathbf{C}$ normalizirani na duljinu 1 s težinama pohranjenim u vektor $\boldsymbol{\lambda} \in \mathbb{R}^R$ stoga imamo

$$\mathcal{X} \approx [\![\boldsymbol{\lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \tag{1.5}$$

Rad je usredotočen na trosmjerni slučaj koji se široko primjenjuje i koji je dovoljan za mnoge potrebe. Za općeniti tenzor N-tog reda, $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, CP dekompozicija je

$$\mathcal{X} \approx [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}]\!] \equiv \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \mathbf{a}_r^{(N)},$$

gdje je $\boldsymbol{\lambda} \in \mathbb{R}^R$ i $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ za $n = 1, \ldots, N$. U ovom slučaju n-mod matricizirana verzija je data s

$$\mathbf{X}_{(n)} \approx \mathbf{A}^{(n)} \boldsymbol{\Lambda} (\mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)})^T,$$

gdje je $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda})$.

Rang tenzora

Definicija 1.2.1. *Rang tenzora $\mathcal{X}$, u oznaci rang $(\mathcal{X})$, je najmanji broj tenzora ranga jedan koji generiraju $\mathcal{X}$ kao njihov zbroj [25][34]. Drugim riječima, to je najmanji broj komponenata u egzaktnoj CP dekompoziciji gdje egzaktan znači da vrijedi jednakost u 1.5.*

Hitchcock [25] prvi predlaže ovu definiciju ranga 1927. godine, Kruskal [34] to čini 50 godina nakon neovisno o njemu. Egzaktna CP dekompozicija sa $R = rang(\mathcal{X})$ komponenata se naziva rang dekompozicija.

Definicija tenzorskog ranga je analogna definiciji matričnog ranga, ali svojstva matričnog i tenzorskog ranga su prilično različita. Razlika je da rang realnog tenzora može biti različit nad $\mathbb{R}$ i nad $\mathbb{C}$. Npr. neka je $\mathcal{X}$ tenzor čiji su frontalni presjeci definirani s

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ i } \mathbf{X}_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Tablica 1.1: Maksimalni rang trosmjernih tenzora nad $\mathbb{R}$

Dimenzije tenzora	Maksimalni rang	Citat
$I \times J \times 2$	$\min\{I, J\} + \min\{I, J, \lfloor \max\{I, J\}/2 \rfloor\}$	[27][36]
$3 \times 3 \times 3$	5	[36]

Ovo je tenzor ranga tri nad $\mathbb{R}$ i ranga dva nad $\mathbb{C}$. Rang dekompozicija nad $\mathbb{R}$ je $\mathcal{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ gdje

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \text{ i } \mathbf{C} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}.$$

Rang dekompozicija nad $\mathbb{C}$ sadrži sljedeće matrice

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & i \end{bmatrix}, \mathbf{B} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \text{ i } \mathbf{C} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}.$$

Ovaj primjer je uzet iz Kruskal [35]. Dokaz da je ranga tri nad $\mathbb{R}$ i metodologija za računanje faktora mogu se naći u Ten Berge [46].

Sljedeća bitna razlika između matričnog i tenzorskog ranga je ta da ne postoji neposredni algoritam za utvrđivanje ranga za određeni dati nam tenzor; problem je NP-težak [23].

Još jedna posebnost tenzora je vezana uz maksimalni i tipični rang.

Definicija 1.2.2. *Maksimalni rang je najveći dostižan rang. Tipični rang je proizvoljni rang koji se pojavljuje sa vjerojatnošću većom od nule.*

Za skup matrica dimenzija $I \times J$, maksimalni i tipični rang su identični i jednaki su $\min\{I, J\}$. Za tenzore mogu biti različiti. Poznata je samo slaba gornja ograda za maksimalni rang proizvoljnog tenzora trećeg reda $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ [36]:

$$rank(\mathcal{X}) \leq \min\{IJ, IK, JK\}. \tag{1.6}$$

Tablica 1.1 pokazuje poznate maksimalne rangove za tenzore određenih dimenzija. Najopćenitiji rezultat je za tenzore trećeg reda sa samo dva presjeka. Tablica 1.2 prikazuje neke poznate formule za tipični rang određenih tenzora trećeg reda nad $\mathbb{R}$. Za više pogledaj Bader i Kolda [4].

Računanje odrezane CP dekompozicije

Kao što smo već spomenuli ne postoji konačan algoritam za određivanje ranga tenzora [36][23]; posljedično, prvi problem koji se javlja u računanju CP dekompozicijeje je kako

Tablica 1.2: Tipični rang trosmjernih tenzora nad $\mathbb{R}$

Dimenzije tenzora	Tipični rang	Citat
$2 \times 2 \times 2$	$\{2, 3\}$	[36]
$3 \times 3 \times 2$	$\{3, 4\}$	[35] [46]
$5 \times 3 \times 3$	$\{5, 6\}$	[48]
$I \times J \times 2$ sa $I \geq 2J$ (veoma visok)	$2J$	[49]
$I \times J \times 2$ sa $J < I < 2J$ (visok)	I	[49]
$I \times I \times 2$ (kompaktan)	$\{I, I + 1\}$	[46][49]
$I \times J \times K$ sa $I \geq JK$ (veoma visok)	JK	[47]
$I \times J \times K$ sa $JK - J < I < JK$ (visok)	I	[47]
$I \times J \times K$ sa $I = JK - J$ (kompaktan)	$\{I, I + 1\}$	[47]

odabrati broj komponenata ranga jedan. Većina procedura isprobava više CP dekompozicija sa različitim brojem komponenata dok jedna ne bude dobra. Idealno, ako su podatci bez šuma i imamo proceduru za računanje CP dekompozicije sa danim brojem komponenata tada možemo napraviti taj izračun za $R = 1, 2, 3, \ldots$ komponenata i stati na prvoj vrijednosti R-a koja daje točnost od sto posto. U nastavku rada ćemo vidjeti da ne postoji savršena procedura za određivanje CP dekompozicije za dani broj komponenata.

Uz pretpostavku da je broj komponenata fiksiran postoje mnogi algoritmi za računanje CP dekompozicije. U radu ćemo obratiti pažnju na aktualni radni algoritam za CP dekompoziciju: metodu alternirajućih najmanjih kvadrata ALS (eng. *Alternating Least Squares*) predloženu od Caroll i Chang [10] i Harshman [21]. Zbog lakše prezentacije, promatramo metodu samo u slučaju tenzora trećeg reda. Cijeli algoritam je prikazan za N-smjerni tenzor na Slici 1.6.

procedure CP-ALS($\mathcal{X}$,R)
  inicijaliziraj $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ for $n = 1, \ldots, N$
  repeat
    for $n = 1, \ldots, N$ **do**
      $\mathbf{V} \leftarrow \mathbf{A}^{(1)\mathsf{T}}\mathbf{A}^{(1)} * \cdots * \mathbf{A}^{(n-1)\mathsf{T}}\mathbf{A}^{(n-1)} * \mathbf{A}^{(n+1)\mathsf{T}}\mathbf{A}^{(n+1)} * \cdots * \mathbf{A}^{(N)\mathsf{T}}\mathbf{A}^{(N)}$
      $\mathbf{A}^{(n)} \leftarrow \mathbf{X}^{(n)}(\mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)})\mathbf{V}^{\dagger}$
      normaliziraj stupce od $\mathbf{A}^{(n)}$ (spremi norme kao $\boldsymbol{\lambda}$) ;
    end for
  until prestalo poboljšavanje ili je dostignut maksimalan broj iteracija
  vrati $\boldsymbol{\lambda}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}$
end procedure

Slika 1.6: ALS algoritam za računanje CP dekompozicije sa R komponenata za tenzor N-tog reda $\mathcal{X}$ dimenzije $I_1 \times I_2 \times \cdots \times I_N$.

Neka je $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ tenzor trećeg reda. Želimo izračunati CP dekompoziciju sa R komponenata koja najbolje aproksimira $\mathcal{X}$, tj. želimo naći

$$\min_{\hat{\mathcal{X}}} \|\mathcal{X} - \hat{\mathcal{X}}\| \text{ uz } \hat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\lambda; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]. \tag{1.7}$$

ALS pristup fiksira **B** i **C** za rješavanje **A**, zatim fiksira **A** i **C** za rješavanje **B**, nakon toga fiksira **A** i **B** za rješavanje **C** te nastavlja ponavljajući cijelu proceduru dok se ne zadovolji neki kriterij zaustavljanja.

Fiksiranje svih matrica osim jedne reducira problem na linearni problem najmanjih kvadrata. Npr. pretpostavimo da su **B** i **C** fiksirani. Tada, iz 1.4, možemo zapisati minimizacijski problem u matričnoj formi kao

$$\min_{\hat{\mathbf{A}}} \|\mathbf{X}_{(1)} - \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^T\|_F, \text{ gdje je } \hat{\mathbf{A}} = \mathbf{A} \cdot diag(\boldsymbol{\lambda}).$$

Najbolje rješenje je dano s

$$\hat{\mathbf{A}} = \mathbf{X}_{(1)} \left[(\mathbf{C} \odot \mathbf{B})^T\right]^{\dagger}.$$

Budući da pseudoinverz Khatri-Raoovog produkta ima posebnu formu kao u 1.2, rješenje možemo zapisati na sljedeći način:

$$\hat{\mathbf{A}} = \mathbf{X}_{(1)}(\mathbf{C} \odot \mathbf{B})(\mathbf{C}^T\mathbf{C} * \mathbf{B}^T\mathbf{B})^{\dagger}.$$

Prednost ove verzije jednakosti je da jedino trebamo izračunati pseudoinverz $R \times R$ matrice umjesto da računamo pseudoinverz $JK \times R$ matrice. Ova verzija nije uvijek preporučljiva s obzirom na moguću pojavu numeričkog lošeg uređaja. Naposljetku, normaliziramo stupce od $\hat{\mathbf{A}}$ da dobijemo **A**. Drugim riječima, neka je $\lambda_r = \|\hat{\mathbf{a}}_r\|$ i $\mathbf{a}_r = \hat{\mathbf{a}}_r/\lambda_r$ za $r = 1, \ldots, R$.

Cijela procedura ALS za N-smjerni tenzor je prikazana na Slici 1.6. Ona pretpostavlja da je broj komponenata, R, CP dekompozicije određen. Faktor matrice se mogu inicijalizirati na bilo koji način, kao što je nasumično ili postavljanjem

$$\mathbf{A}^{(n)} = R \text{ vodećih lijevih svojstvenih vektora od } \mathbf{X}_{(n)} \text{ za } n = 1, \ldots, N.$$

Za više o inicijalizaciji vidi [7][45]. U svakoj unutarnjoj iteraciji se mora izračunati pseudoinverz matrice **V** (vidi Sliku 1.6), ali on je veličine samo $R \times R$. Iteracije se ponavljaju dok neka kombinacija uvjeta zaustavljanja ne bude zadovoljena. Mogući uvjeti zaustavljanja su: 1) mali ili nikakav napredak u ciljnoj funkciji, 2) mala ili nikakva promjena u faktor matricama, 3) ciljna vrijednost nula ili blizu nule, 4) prekoračenje unaprijed zadanog broja iteracija.

ALS metodu je lako shvatiti i implementirati, ali može joj trebati mnogo iteracija za konvergenciju. Štoviše, nemamo garanciju konvergencije u globalni minimum ili barem u stacionarnu točku od 1.7. Jedino imamo garanciju konvergencije u rješenje gdje ciljna funkcija od 1.7 prestaje padati. Konačno rješenje može jako ovisiti o početnim pretpostavkama.

Faber, Bro i Hopke [19] uspoređuju ALS sa šest drugih metoda od kojih nijedna nije bolja od ALS metode ako uzmemo u obzir kvalitetu rješenja, iako je ASD (eng. *Alternating Slicewise Diagonalization*) metoda [29] priznata kao alternativa kad se uzme u obzir vrijeme računanja. Tomasi i Bro [50] uspoređuju ALS i ASD metode sa četiri druge te sa tri varijante koje koriste Tucker-bazirano sažimanje i na tako smanjenom polju računaju CP dekompoziciju; vidi [8]. U ovoj usporedbi se nalaze i prigušena Gauss-Newton (dGN, eng. *domped Gauss-Newton*) varijanta te varijanta nazvana PMF3 od Paatero [40]. Obje metode, dGN i PMF3, optimiziraju sve faktor matrice simultano. Za razliku od prijašnjeg istraživanja ASD metoda se smatra inferiornom obzirom na druge alternirajuće metode. dGN i PMF3 metode su generalno superiornije od ALS ako promatramo konvergencijska svojstva, ali su memorijski i vremenski skuplje.

Poglavlje 2

TweetRank

2.1 Motivacija ili tko koga slijedi

Rangiranje autoriteta je komponenta od velike važnosti za široki rang aplikacija društvene mreže kao što su tematsko usmjerenje, višeslojno pretraživanje ili preporuke kontakata. Zajednice na mreži (kao Twitter ili Facebook) daju vrlo ograničene statistike o korisničkim odnosima kao što je broj kontakata u korisnikovoj kontakt listi, broj registriranih promatrača korisnikovih postova itd. Iako ovi brojevi adekvatno odražavaju korisnikovu popularnost u društvenoj zajednici, njihova interpretacija u usmjerenijem kontekstu postaje veoma teška. Kao stvaran primjer za ovaj problem možemo promatrati dva stvarna i usto prilično popularna korisnika u Twitter zajednici: *timerners_lee* (20 tisuća pratitelja) i *parishilton* (1.7 milijuna pratitelja). Početkom 2010. godine, oba korisnika su u isto vrijeme komentirala lansiranje nove Apple iPad tehnologije postovima (tvitovima) kao što je prikazano na Slici 2.1.

timberners_lee	*parishilton*
Following: 59	Following: 272
Followers: 20.692	Followers: 1.709.116
RT @janl: Apple:	I Love my new I-Pad.
"Preparing your web content for iPad:	So much fun!
2. Use W3C standard web technologies." #w3c	Technology rocks!
9:46 AM Mar 21st via TweetDeck	about 2 h ago via UberTwitter

Slika 2.1: Stvarni tvitovi korisnika *timberners_lee* i *parishilton* na Twitteru

Iz čisto špekulacijske perspektive, možemo donijeti zaključak da je korisnik *parishilton* jači Twitter autoritet od korisnika *timerners_lee* na temu novih tehnologija. Dublja

analiza postova obaju korisnika odmah pokazuje suprotno. Dok su doprinosi korisnika *timerners_lee* točno usmjereni na nove istraživačke/tehnološke aspekte (*Web Science*, *Linked Open Data*, *Semantic Web*), korisnik *parishilton* je radije posvećen temama kao što su poznate osobe, stil života ili samopromocija.

Potreba za kontekstualizacijom doprinosa dovodi do specifičnih mehanizama značajnih za pojedine platforme te do samoorganiziranja stvaranjem vokabularnih proširenja. Najpoznatiji oblik jednostavne kontekstualizacije sadržaja je tagiranje, tj. korištenje kratkih izraza za označavanje sadržaja. U Twitter zajednici primjer za to su tzv. *hashtagovi* (karakteristične, kontekst indikativne riječi unutar tekstualnog sadržaja obilježene prethodećim *hash* znakom) koji su postali prilično popularni. Npr. pojam *#w3c* iz teksta sa Slike 2.1 je jedan *hashtag*. U ovom konkretnom slučaju, zajednica je uspostavila praksu da se koristi *#w3c* za označavanje postova srodnih W3C konzorciju. Ovo je klasičan primjer novonastale semantike u modernim društvenim mrežama.

Osnovni scenarij za laganu kontekstualizaciju korisničkih odnosa može biti organiziran na sličan način. Konkretno, možemo pretpostaviti da neki korisnici društvene mreže izričito iskazuju zanimanje za neku temu (recimo da je to jedan *hashtag h*). U nastavku ćemo se osvrnuti na takve grupe korisnika kao na jedan *h*-kandidatni skup. Rangiranje autoriteta za *h*-kandidatne skupove može biti temeljeno na kontekstualiziranim listama pratitelja. Iz idejne perspektive, ovaj pristup možemo promatrati kao poseban slučaj suradničkog glasanja usmjerenog na *h*. Zapravo, možemo ograničiti opseg liste pratitelja i prividno ukloniti sve korisnikove unose koji nisu izravno u *h*-kandidatnom skupu (tj. nevažni su u kontekstu *h*). Drugim riječima, jednostavno reduciramo liste pratitelja u kontakte iz *h*-kandidatnog skupa. Dobivene kardinalnosti reduciranih lista pratitelja daju prirodno rangiranje za preporuke kontakata (koga slijediti) u kontekstu promatranog *hashtaga*.

Za promatrani Twitter scenarij, ova funkcionalnost je zaista ponuđena od strane vanjskog poslužitelja *wefollow.com*. Tijekom registracije korisnici specificiraju ključne riječi koje su im od interesa (tj. specificiraju *hashtagove*) te daju podatke o svojim računima portal poslužitelju (osiguravanje pristupa listi korisnika koje korisnik prati). Nakon toga, korisnici mogu pristupiti rangiranoj listi kontakt preporuka za odabranu ključnu riječ (sugestija koga slijediti). Iako precizna organizacija od *wefollow.com* nije javna, generirana lista može se reproducirati s iznimno viskom točnošću koristeći *h*-kandidatne skupove. Slika 2.2 prikazuje top 20 preporuka vezanih uz ključnu riječ *semanticweb* po *wefollow.com* i po suradničkom glasanju. Implementacija suradničkog glasanja koristi Twitterove vlastite API-e (eng. *Application Programming Interface*) za prikupljanje javno dostupnih korisničkih informacija, npr. liste pratitelja. Male razlike u rang pozicijama mogu se objasniti činjenicom da Twitter profili nekoliko članova *wefollow.com* za ključnu riječ *semanticweb* (6 profila od 242 promatrana) nisu otvoreni za javnost.

Očiti nedostatak predložene strategije je njeno ograničenje na proaktivne korisnike koji

#	wefollow.com preporuke	*h*-skup preporuke	#followers ukupno	u *h*-skupu	*H*-skup preporuke	#followers ukupno	u *H*-skupu
1	tommyh	PaulMiller	2,215	82	timberners_lee	20,694	252
2	jahendler	jahendler	909	76	timoreilly	1,428,425	200
3	ivan_herman	tommyh	738	76	jahendler	910	185
4	PaulMiller	ivan_herman	680	69	LeeFeigenbaum	273	176
5	opencalais	opencalais	1,797	68	danbri	1,781	160
6	danja	danja	1,313	59	kidehen	1,806	159
7	CaptSolo	juansequeda	988	57	ivan_herman	680	153
8	juansequeda	CaptSolo	1,224	52	PaulMiller	2,216	150
9	sclopit	gothwin	685	50	tommyh	737	142
10	gothwin	robocrunch	3,679	49	novaspivack	7,896	139
11	robocrunch	alexiskold	4,321	48	johnbreslin	1,933	128
12	kristathomas	kristathomas	1,499	48	w3c	10,672	128
13	kendall	kendall	1,694	45	mimasnews	198	127
14	bobdc	andraz	2,065	44	iand	1,094	123
15	phclouin	sclopit	513	42	rww	1,045,511	123
16	brown2020	cjmconnors	466	39	terraces	632	119
17	alexiskold	gkob	399	37	mhausenblas	449	118
18	andraz	dorait	2,541	35	opencalais	1,799	112
19	cjmconnors	phclouin	266	35	ldodds	621	110
20	ontoligent	openamplify	1,387	34	semanticnews	843	102

Slika 2.2: Preporuke koga slijediti, *h=#semanticweb*

koriste novu uslugu i eksplicitno se ”pretplaćuju” za *hashtagove* od interesa. Korisnici koji se nisu ”pretplatili” nisu dio *h*-kandidatnog skupa i na taj način ne mogu biti preporučeni. Zanemaruje se njihova stvarna važnost u kontekstu *h*. Ovo ograničenje se može izbjeći drugim strategijama konstrukcije *h*-kandidatnog skupa. Možemo uzeti u obzir ”aktivne” korisnike u smislu *h* (npr. tražeći *h* u njihovim nedavnim postovima). U nastavku ćemo obratiti pažnju prema korisnicima koji aktivno koriste željeni *hashtag* u određenom vremenskom razdoblju (npr. 4 tjedna) te će oni biti *H*-kandidatni skup. U praksi, *H*-kandidatni skup se može izravno dobiti uz pomoć uobičajenih API-a pretraživanjem po ključnoj riječ pri čemu je *h* upit. Drugi dio Slike 2.2 prikazuje odgovarajuće preporuke koga slijediti za *hashtag h=#semanticweb* na Twitteru. To se može promatrati kao *H*-kandidatni skup koji dobiva znatno veći značaj uključivanjem jako relevantnih korisnika iz područja *h=#semanticweb* koji nemaju eksplicitnu registraciju na wefollow.com.

Upoznate strategije omogućuju bilježenje korisničkih odnosa iz uobičajenih prikaza društvene mreže u graf strukture gdje čvorovi predstavljaju korisnike, a bridovi odnose koji međusobno povezuju korisnike. Prema tome, algoritmi bazirani na grafovima za rangiranje autoriteta kao što su PageRank [6], HITS [31] ili SALSA [37] mogu se primjeniti na društvene mreže. Umjesto ocjena Web stranica promatrat ćemo korisničke ocjene u

društvenom okruženju uzimajući u obzir jedan ili više kriterija, npr. hub i autoritet ocjene u HITS. Te ocjene odražavaju značaj (centralnost) pojedinih korisnika u društvenoj mreži i na taj se način mogu iskoristiti za procjenu relevantnosti npr. u preporukama koga slijediti.

Možemo donijeti dva važna zapažanja o rangiranju autoriteta u društvenim grafovima. Prvo, računalni modeli standardnih algoritama za mrežnu analizu uzimaju u obzir samo strukturne informacije, tj. samo povezanost čvorova grafa. Dodatna semantika povezanosti, npr. znanje o različitim tipovima odnosa, nije korištena. Drugo, mnogo je slučajeva u kojima se koristi vokabular koji se preklapa, preopširan je ili proturječno opisuje slične probleme. Dakle, možemo očekivati preopširnost u smislu postojanja različitih *hashtagova* sa veoma sličnim (ili povezanim) značenjem, npr. *h1=#semanticweb*, *h2=#RDF* i *h3=#ontology*. Obično algoritmi za rangiranje autoriteta ne pružaju podršku za pronalazak ovakvih grupa.

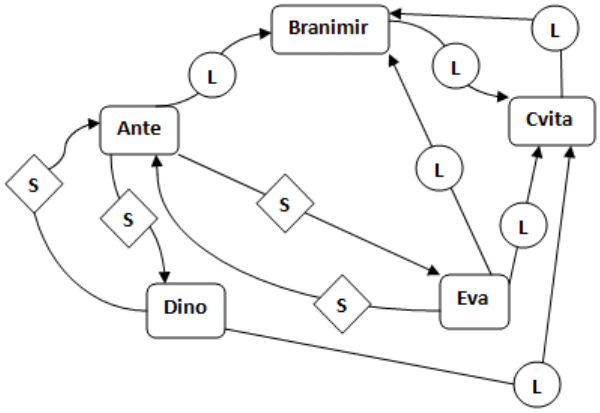
2.2 Društvena mreža kao tenzor

U nastavku ćemo se upoznati sa uspješnim TweetRank pristupom za rangiranje autoriteta u zajednicama društve mreže. Doradit ćemo formalnu notaciju društvenih grafova i tenzora, upoznati faktorizaciju tenzora za fino rangiranje autoriteta i pokazati stvarne primjere sa rezultatima.

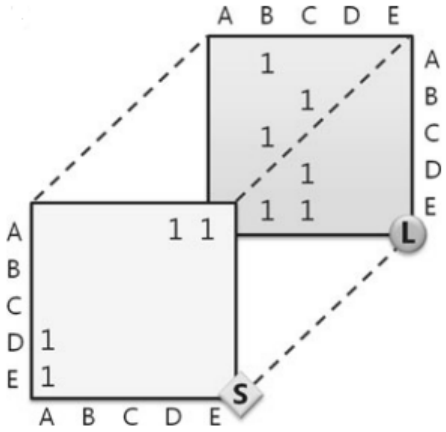
TweetRank model

Definiramo graf društvene mreže kao graf $G = (V, L, E, linkType)$, gdje je V skup korisnika u zajednici, L je skup literala (*hashtagova*) i E je skup relacija među korisnicima u V. Funkcija $linkType : E \longrightarrow L$ vraća oznaku iz L koja povezuje dva korisnika. Slika 2.3 pokazuje pojednostavljeni graf društvene mreže koji sadrži pet korisnika (A=Ante, B=Branimir, C=Cvita, D=Dino, E=Eva), dva *hashtaga* kao literale (*lifestyle*=L, *semanticweb*=S) i deset relacija dvaju različitih tipova: *slijedi-lifestyle* i *slijedi-semanticweb*. Precizna semantika takvih poveznica je specifična obzirom na aplikaciju; u našem jednostavnom primjeru pretpostavljamo da je korisnik X povezan sa korisnikom Y bridom tipa Z ako i samo ako 1) X slijedi Y i 2) oba su nedavno koristili *hashtag* Z u svojim postovima. Npr. graf izražava da Ante slijedi Branimira po svojstvu *lifestyle*.

Mi prikazujemo grafove društvene mreže kao trodimenzionalni tenzor $\mathcal{T}$ gdje svaki od njegovih presjeka predstavlja matricu susjedstva za jednu relaciju tipa L. Slika 2.4 prikazuje tenzor nastao iz jednostavnog grafa sa Slike 2.3. Prva matrica susjedstva $\mathbf{T}_{::1}$ modelira povezanost po svojstvima *semanticweb*. Ulaz >0 odgovara postojanju poveznice po tom svojstvu, prazni ulazi se smatraju nulama. Druga matrica $\mathbf{T}_{::2}$ modelira veze po svojstvu *lifestyle*. Npr. činjenica da Ante slijedi Branimira po svojstvu *lifestyle* rezultira time da je $t_{122} = 1$ u tenzorskoj reprezentaciji.



Slika 2.3: Jednostavni graf društvene mreže

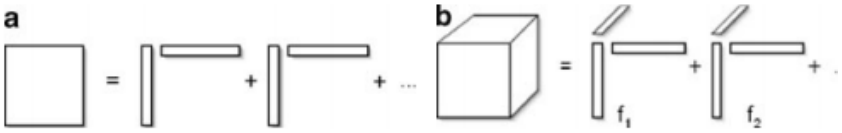


Slika 2.4: Tenzorska reprezentacija

PARAFAC dekompozicija za rangiranje autoriteta

Graf društvene mreže može biti opisan matricom susjedstva. Na mrežni graf matrice **M** može se primjeniti dobro znanu metodu rangiranja autoriteta kao HITS [31]. HITS definira problem rangiranja autoriteta kroz međusobno jačanje ocjena huba i autoriteta za vrhove grafa (zajednica korisnika u našem slučaju). Ocjena autoriteta (važnosti) svakog vrha je definirana kao suma ocjena huba njegovih prethodnika. Analogno, ocjena huba (povezanosti) svakog vrha je definirana kao suma ocjena autoriteta njegovih sljedbenika.

Primjenom SVD dekompozicije na matricu susjedstva dobivamo ocjene huba i autoriteta za vrhove grafa za svaku svojstvenu vrijednost od $\mathbf{M}$ što se može interpretirati kao dodjeljivanje ranga različitim nama zanimljivim temama. Formalno, ovom metodom, je neka proizvoljna matrica $\mathbf{M} \in \mathbb{R}^{k\times l}$ podijeljena u tri matrice $\mathbf{U} \in \mathbb{R}^{k\times m}$, $\mathbf{S} \in \mathbb{R}^{m\times m}$ i $\mathbf{V} \in \mathbb{R}^{l\times m}$. Matrice $\mathbf{U}$ i $\mathbf{V}$ predstavljaju izlazne i ulazne veze s obzirom na glavni čimbenik sadržan u $\mathbf{S}$. Matricu $\mathbf{M}$ možemo zapisati kao sumu matrica ranga jedan $\mathbf{M} = \sum_{k=1}^{m} \mathbf{S}_k \cdot \mathbf{U}_k \circ \mathbf{V}_k$. Ova dvosmjerna dekompozicija daje ocjene autoriteta i huba (vidi Sliku 2.5 a)) [4].



Slika 2.5: a)Dekompozicija matrice, b)Dekompozicija tenzora

Modeliranje nekoliko tipova veze odvojenim matricama rezultira veoma prorijeđenim i nepovezanim matricama. Umjesto toga, tenzorski model primjenjen u TweetRank-u omogućuje reprezentaciju svih matrica susjedstva uključujući informacije o povezanosti između tipova veze. Metoda tenzorske dekompozicije kao PARAFAC može tada otkriti do prije skrivene ovisnosti.

Ove metode se smatraju ekvivalentnima matričnim dekompozicijama višeg reda. Tenzorska dekompozicija PARAFAC kao prednost ima robusnost i učinkovitost računanja. Ove prednosti su zbog njenog jedinstvenog skaliranja i permutiranja dobivenih komponent matrica [22]. PARAFAC ulazne tenzore transformira u tzv. Kruskalove tenzore, u sumu tenzora ranga jedan. Posljedično, u TweetRank metodi iz posebnih tenzora ranga jedan dobivenih dekompozicijom dijelimo autoritet i hub ocjene za posebne skrivene aspekte analiziranih podataka. U kontekstu rada usredotočimo se na trosmjerne tenzore koji reprezentiraju povezanost između čvorova grafa (korisnika) zajedno sa semantikom korisničkih odnosa.

Formalno, tenzor $\mathcal{T} \in \mathbb{R}^{k\times l\times m}$ je rastavljen PARAFAC dekompozicijom u komponentne matrice $\mathbf{U} \in \mathbb{R}^{k\times n}$, $\mathbf{V} \in \mathbb{R}^{l\times n}$, $\mathbf{Z} \in \mathbb{R}^{m\times n}$ i n glavnih (centralnih) faktora (pf) λ_i u silaznom poretku. Tenzor $\mathcal{T}$ možemo zapisati kao Kruskalov tenzor $\mathcal{T} \approx \sum_{k=1}^{n} \lambda_k \cdot \mathbf{U}_k \circ \mathbf{V}_k \circ \mathbf{Z}_k$, gdje λ_k označava k-ti centralni faktor, $\mathbf{U}_k$, $\mathbf{V}_k$, $\mathbf{Z}_k$ označava k-ti stupac matrice te $\circ$ označava vanjski produkt [4]. Matrice $\mathbf{U}$, $\mathbf{V}$ i $\mathbf{Z}$ daju odnos i-te dimenzije i centralnih faktora. Jednako kao SVD, PARAFAC dekompozicija izvodi skrivene ovisnosti povezane sa glavnim faktorima te izražava dimenzije tenzora preko veza sa glavnim faktorima. Ovisno o broju glavnih faktora PARAFAC dekompozicija može biti bez gubitaka. Za rang tenzora $\mathcal{T} \in \mathbb{R}^{k\times l\times m}$ trećeg reda znamo slabu gornju ogradu (vidi 1.6). Nema pravog načina za procjenu optimalnog broja glavnih faktora za prikladnu dekompoziciju, ali postoji nekoliko indikatora.

Npr. analiza ostataka ili konzistencija jezgre [9]. PARAFAC tenzorska dekompozicija izvodi ocjene autoriteta i hubova i usto dodatne ocjene za važnost tipova veze (vidi Sliku 2.5 b)). Tenzor $\mathcal{T}$ iz Odlomka 2.2 kombinira informacije o tome tko koga slijedi s našim objašnjenjima sljedbeničkih veza. Dakle, PARAFAC dekompozicija će dati matricu $\mathbf{U}$ sa subjekt-glavni faktor odnosom, $\mathbf{V}$ s objekt-glavni faktor odnosom i $\mathbf{Z}$ sa svojstvo-glavni faktor odnosom. Drugim riječima, $\mathbf{U}$ sadrži hub ocjene kao povezanost korisnika i glavnih faktora, $\mathbf{V}$ sadrži ocjene autoriteta kao povezanost korisnika i glavnih faktora, $\mathbf{Z}$ sadrži ocjene povezanosti literala (*hashtagova*) i glavnih faktora. U usporedbi s HITS-om najveći ulaz od $\mathbf{U}_1$ odgovara najboljem hubu za prvi glavni faktor, a najveći ulaz od $\mathbf{V}_1$ najboljem autoritetu.

Primjer rangiranja

Primjenom navedene faktorizacije i analize na graf ilustriran Slikom 2.3 dobivamo rezultate prikazane na Slici 2.6 u prva četiri stupca. Na Slikama 2.7 i 2.8 se vidi izvršni kod u Matlab okruženju gdje je korišten *MATLAB tensor toolbox* [1]. Prepoznajemo dvije grupe, jednu gdje *hashtag lifestyle* ima visoku ocjenu i jednu gdje *hashtag semanticweb* ima visoku ocjenu. Autoritativni izvori za svaku grupu se razlikuju. Cvita i Branimir imaju visoke ocjene s obzirom na *lifestyle*. Dino i Eva su najveći autoriteti kad je u pitanju *semanticweb*. Primjena HITS algoritma sa Slike 2.9 u Payton okruženju rezultira rangiranjem prikazanim u petom i šestom stupcu na Slici 2.6. HITS rangiranje odgovara rangiranju baziranom na ulaznim stupnjevima izvora. Rangiranje dobiveno PARAFAC dekompozicijom je različito od HITS rezultata jer provode rangiranje uzimanjem u obzir različitih vidova znanja o podatcima.

2.3 Implementacija

Nakon upoznavanja teorijske pozadine TweetRank-a, objasnit ćemo implementaciju u primjenjiv sistem. Opisane su tri glavne komponente TweetRank arhitekture koje ukratko opisuju proces (1) prikupljanje podataka i transformacija u tenzorski model, (2) predobrada podataka, (3) PARAFAC dekompozicija za TweetRank. Na kraju dajemo primjer korištenja koji je preuzet iz [44] zajedno sa razultatima i analizom. Sizov i ostali [44] su koristili programsko okruženje Java[2].

[1] Besplatno dostupan na `http://www.sandia.gov/~tgkolda/TensorToolbox/index-2.5.html`

[2] Okvir za implementaciju je dostupan kao *open source* paket na `http://west.uni-koblenz.de/Research`

PARAFAC					HITS		ULAZNI STUPANJ	
Ocjena	Hashtag	Ocjena	Korisnik		Ocjena	Korisnik	Stupanj	Korisnik
Grupa 1					0.6190	Cvita	3	Branimir
1.0000	*lifestyle*	0.7125	Cvita		0.5402	Branimir	3	Cvita
		0.7017	Branimir		0.5001	Ante	2	Ante
Grupa 2					0.1671	Dino	1	Dino
1.0000	*semanticweb*	0.6976	Dino		0.1671	Eva	1	Eva
		0.6976	Eva					
		0.1636	Ante					

Slika 2.6: PARAFAC i HITS

```
T is a tensor of size 5 x 5 x 2
    T(:,:,1) =
         0    0    0    1    1
         0    0    0    0    0
         0    0    0    0    0
         1    0    0    0    0
         1    0    0    0    0
    T(:,:,2) =
         0    1    0    0    0
         0    0    1    0    0
         0    1    0    0    0
         0    0    1    0    0
         0    1    1    0    0
```

Slika 2.7: Tenzor u Matlabu

Prikupljanje podataka i transformacija

Prvi procesni korak za rangiranje podataka iz društvene mreže je njihovo prikupljanje. Cilj ovog koraka je konstrukcija grafa semantičkih relacija između relevantnih korisnika platforme $G = (V, L, E, linkType)$. Za mnoge platforme društvenih mreža prikupljanje relevantnih podataka za konstrukciju kandidatnog skupa korisnika V može biti izravno im-

```
>> X=parafac_als(T,2)

CP_ALS:
 Iter  1: fit = 3.401520e-01 fitdelta = 3.4e-01
 Iter  2: fit = 3.618262e-01 fitdelta = 2.2e-02
 Iter  3: fit = 3.661575e-01 fitdelta = 4.3e-03
 Iter  4: fit = 3.672102e-01 fitdelta = 1.1e-03
 Iter  5: fit = 3.674633e-01 fitdelta = 2.5e-04
 Iter  6: fit = 3.675246e-01 fitdelta = 6.1e-05
 Final fit = 3.675246e-01
X is a ktensor of size 5 x 5 x 2
        X.lambda = [ 1.9999    1.4142 ]
        X.U{1} =
                0.3482   0.9865
                0.3589   0.0000
                0.3481   0.0000
                0.3589   0.1157
                0.7071   0.1157
        X.U{2} =
                0.0017   0.1636
                0.7017  -0.0009
                0.7125  -0.0028
               -0.0009   0.6976
               -0.0009   0.6976
        X.U{3} =
                0.0018   1.0000
                1.0000   0.0012
```

Slika 2.8: PARAFAC dekompozicija

plementirano u postojećoj platformi - specijalne API funkcije. Npr. Twitter API podrška je dostupna za sve glavne programske jezike. Ovaj API može koristiti proizvoljni vlasnik računa na platformi sa izvjesnim performansnim ograničenjima (u smislu broja upita po satu). Pristup većeg razmjera API funkcijama može se odobriti nakon pojedinačnog zahtjeva.

Proces prikupljanja podataka počinje sa specificiranjem interesnih pojmova. Ti pojmovi se proslijede platformi, specifičnoj API funkciji za pretragu baziranu na ključnoj riječi koja vraća odgovarajuće postove zajedno sa metapodatcima (uobičajeno uključujući autorov ID). ID-ovi od autora se tada vade i dodaju u V. Nakon toga, prethodnici i sljedbenici svih korijenskih korisnika, u smislu sadržajno povezanih relacija, se također učitavaju

```
def hits(graph, root_set=[], max_iterations=5, min_delta=0.0001):
    base_set = []
    if not root_set:
        base_set = graph.nodes()
    else:
        base_set = []
        for node in root_set:
            for nod in graph.node_incidence[node]:
                base_set.append(nod)
        for node in root_set:
            for nod in graph.node_neighbors[node]:
                base_set.append(nod)
        base_set.extend(root_set)
        base_set = set(base_set)
    auth = dict.fromkeys(base_set, 1)
    hub = dict.fromkeys(base_set, 1)
    def normalize(dictionary):
        norm = sum((dictionary[p] for p in dictionary))
        return {k: v / norm for (k, v) in dictionary.items()}
    i = 0
    for i in range(max_iterations):
        for p in base_set:
            auth[p] = sum((hub.get(q, 0) for q in graph.node_incidence[p]))
        auth = normalize(auth)
        old_hub = dict()
        for p in base_set:
            old_hub[p] = hub[p]
            hub[p] = sum((auth.get(r, 0) for r in graph.node_neighbors[p]))
        hub = normalize(hub)
        delta = sum((abs(old_hub[k] - hub[k]) for k in hub))
        if delta <= min_delta:
            return (hub, auth)
    return (hub, auth)
```

Slika 2.9: HITS algoritam

i dodaju u V. Ovaj korak je obično podržan od strane platformskih API funkcija za detalje korisničkih profila. Npr. Twitter API osigurava određene funkcije za pronalazak prethodnika i sljedbenika datog korisnika: lista sljedbenika (korisnici koji promatraju postove datog korisnika) i lista korisnika koje dati korisnik slijedi. Daljnje širenje skupa V možemo postići prolazeći kroz relacije slijedi i slijeđen do određene zadane maksimalne dubine (pronalazak sljedbenika od sljedbenika).

U sljedećem koraku, učitava se za svakog korisnika iz $u \in V$ broj njegovih nedavnih postova. U Twitter API ova funkcionalnost je izravno ponuđena kao dio opsežne potpore za pretragu. Svaki post se nakon toga rastavi u pojmove $t \in L$. Za konstrukciju bridova u E, pretpostavljamo da je korisnik $u_1 \in U$ povezan s korisnikom $u_2 \in U$ bridom $e \in E$ označenim $linkType(e) = t \in L$ ako je u_1 poznat kao prethodnik od u_2 u smislu platforme (u_1 slijedi u_2) i oba su često koristili pojam t u svojim postovima. U Twitter slučaju, uobičajena praksa laganog označavanja sadržaja je upotreba *hashtagova*. Iz tog razloga, prije slanja inicijalnog upita Twitteru dodaje se *hash* znak ispred svakog pojma koji nas zanima te ne razmatramo pojmove bez *hash* znaka u vraćenim postovima. Konačno, graf je transformiran u tenzorsku reprezentaciju za faktorizacijsku analizu.

Važno podešavanje parametara u okviru uključuje broj potrebnih postova za inicijali-

zaciju liste korisnika, maksimalan broj prethodnika/sljedbenika koje treba uzeti u obzir za svakog korisnika, broj nedavnih postova po korisniku koji će biti obrađivani i filtracijski kriterij za uklanjanje potencijalno nevažnih korisnika i pojmova. Za Twitter je instanciran okvir sljedećim postavkama (empirijski procijenjeno u seriji sličnih eksperimenata) 300 korisnika za inicijalni skup, do 500 izravnih prethodnika/sljedbenika po korisniku (smnjivanje njihovog broja slučajnim uklanjanjem suvišnih unosa kada je to potrebno), 100 nedavnih postova po korisniku za analizu, svaki *hashtag* $t \in L$ trebao bi biti korišten od strane barem 10 različitih korisnika, svaki korisnik $u \in V$ trebao bi koristiti u svojim postovima barem 3 *hashtaga* iz L.

Daljnji predprocesni korak je davanje težine prikupljenim korisničkim relacijama za daljnje uklanjanje negativnih učinaka dominacije. Pojačavaju se relacije na temelju njihove *hashtag* učestalosti tako da se postovi sa manje učestalim *hashtagovima* pojačavaju jače od uobičajenih relacija. Kao posljedica, susjedski pokazatelji u tenzoru imaju sljedeće svojstvo:

$$t_{xyz} = \begin{cases} 1 + \log \frac{\alpha}{links(z)}, & x, y \in V \\ & links(z) =\mid \{e \in E \mid linkType(e) = z\} \mid \\ & \alpha = links(x) \mid \forall t \in L, links(x) \geq links(t) \\ 0, & \text{inače} \end{cases}$$

Vrijednost α označava broj relacija u kojima sudjeluje najdominantniji *hashtag*. Funkcija $links(t)$ ($links : L \longrightarrow \mathbb{N}_0$) vraća broj relacija iz E induciranih *hashtagom* $t \in L$.

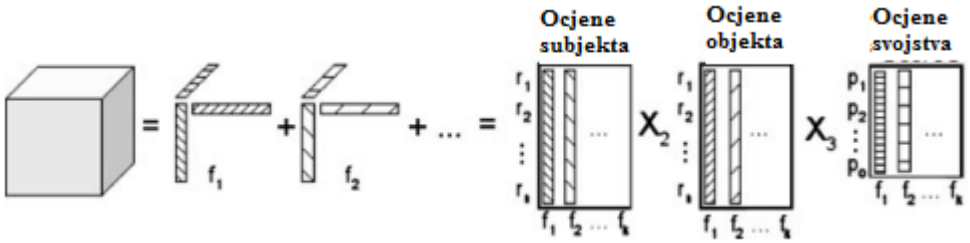
Naglašavamo da su implementirani koraci predobrade podataka vrijedni za oblikovanje rang analize u cjelini. Znakovito je da jednostavne metode za rangiranje autoriteta, primjerice prebrojavanje ocjena ulaznih veza po resursu i predikatu, imaju više koristi od predobrade, nego što to imaju mnogo složenije metode kao npr. PARAFAC.

PARAFAC dekompozicija za TweetRank

Analiza provodi PARAFAC dekompoziciju tenzora kao što je modelirano i kreirano u prethodnom koraku. Integrirni su postojeći softverski paketi [3] u tu svrhu. Kao što je ranije navedeno broj faktora za PARAFAC dekompoziciju je presudan za kvalitetu rezultata analize. Određivanje optimalnog broja faktora je još otvoren problem. Objavljene su neke heuristike za određivanje pogodnog broja faktora npr. CONCORDIA [9]. Određivanje faktora primjenjeno u TweetRank modelu je razvijeno u ovom radu.

Rezultat analize je Kruskalov tenzor [4] koji aproksimira orginalni tenzor. Kao što je ilustrirano na Slici 2.10 rezultirajući vektori za prvu (redak), drugu (stupac) i treću dimenziju su prikazani sa tri matrice. Stupci svake od matrica odgovaraju ocjenama izračunatim za različite faktore $pf_1 \cdots pf_n$ (glavni faktori). Analogno SVD-u, vrijednosti u stupčanim

vektorima odgovaraju ocjenama autoriteta, odnosno, ukazuju na važnost resursa uzimajući u obzir njegov ulazni stupanj. Vrijednosti u retčanim vektorima odgovaraju hub ocjenama, odnosno, ukazuju na važnost resursa uzimajući u obzir njegov izlazni stupanj. Za temeljitu analizu veze između SVD dekompozicije i njene interpretacije za analiziranje povezanosti pogledajte [31]. Vrijednosti u vektorima treće dimenzije pokazuju važnost izraza s obzirom na hub i autoritet korisnika. Koristeći ove oznake, interpretirane su hub ocjene kao indikativne za važnost korisnika kao promatrača drugih korisnika. Obratno, autoritet ocjene pokazuju važnost korisnika kao subjekta kojeg se promatra. Kao što su modelirali-rani izrazi u postovima, njihova važnost za glavne faktore se može vidjeti u vektorima treće dimenzije.



Slika 2.10: Rezultat analize

Primjer korištenja

Za demonstraciju funkcionalnosti prikazanog okvira razmotren je upit *semanticweb* iz probnog Twitter scenarija. Pokrenuta je konstrukcija tematski usmjerenog društvenog grafa slanjem Twitteru zahtjeva za pretragom baziranog na ključnoj riječi $q = \#semanticweb$. Nakon koraka proširenja (dodavanje prethodnika, sljedbenika, preuzimanja postova prikupljenih dosad od svih korisnika) i zajedničke predobrade dobiven je društveni graf $G = (V, L, E, linkType)$ s $| V |= 1323$ korisnika, $| L |= 175$ *hashtagova*, $| E |= 17190$ korisničkih relacija.

Nakon toga, tenzorska dekompozicija sa $f = 15$ PARAFAC faktora omogućuje ocjene autoriteta i huba za korisnike te *hashtag* značaj za svaki faktor. Slika 2.11 preuzeta iz [44] prikazuje pet najznačajnijih *hashtagova* i korisnika za neke faktore ove dekompozicije (poredani redom po *hashtag* značaju i autoritetu korisnika).

Mogu se izvesti neka važna zapažanja o rezultatima prikazanim na Slici 2.11. Prvo, ocjenjivanje autoriteta za promatrani faktor *semanticweb* je usko povezano sa rezultatima nastalim jednostavnijim mehanizmima rangiranja (npr. preporuka H-kandidatnog skupa).

Ocjena	Hashtag	Ocjena	Korisnik	Ocjena	Hashtag	Ocjena	Korisnik
Faktor 1	("semanticweb")			**Faktor 4**	("programming")		
0.147	semantics	0.238	timberners_lee	0.111	programming	0.183	cjmconnors
0.125	business	0.143	PEPublishing	0.053	analytics	0.178	DublinCore
0.106	lod	0.142	jahendler	0.047	semantic	0.108	spirinet
0.091	semweb	0.138	timoreilly	0.046	microdata	0.098	AskAaronLee
0.054	semanticweb	0.097	semanticnews	0.040	startups	0.096	GeoffWigz
Faktor 2	("web tools")			**Faktor 5**	("securiry&privacy")		
0.266	java	0.419	SCMagazine	0.163	security	0.285	BLSocSci
0.251	php	0.143	HTML5watcher	0.130	web20	0.278	socialwendy
0.220	http	0.128	opencalais	0.083	privacy	0.162	pedantic_web
0.214	joomla	0.102	hadoop	0.067	apps	0.154	drthinkmore
0.199	javascript	0.097	LSIstorage	0.054	china	0.136	linuxhoundhost
Faktor 3	("multimedia")			**Faktor 6**	("social media")		
0.049	music	0.265	Beyond15	0.212	foaf	0.484	jwolfnbaa
0.044	video	0.185	junglejar	0.172	socialmedia	0.165	rdfQuery
0.032	semweb	0.172	CSS3	0.118	facebook	0.130	CreativeCustoms
0.023	iphone	0.134	davidstack	0.109	rdfa	0.078	websciencetrust
0.021	innovation	0.125	emtacl	0.054	webscience	0.064	virtualrooms

Slika 2.11: TweetRank rezultati za upit *semanticweb* na Twitteru

Rezultati su u potpunosti bazirani na nedavnim korisničkim postovima i trenutnim korisničkim relacijama. Dakle, bilo koja nuspojava dugoročnog korisničkog profiliranja (kao što je privremena korisnička aktivnost na određenu temu, ali veoma davno) neće imati utjecaja na trenutne preporuke. Nakon glavnog faktora (recimo *semanticweb* jezgre) dekompozicija dohvaća teme drugog reda povezane sa *semanticweb* i reflektiranih u Twitter postovima, npr. *web tools*, *multimedia*, *security & privacy*, *social media* i *programming*. U tom smislu, raznolikost i strukturiranje rezultata preporuke su značajno povećani. Kao rezultat, korisnik može bolje identificirati stvarni cilj od vlastitog interesa povezan sa *semanticweb* i tada slijedi najbolje ocijenjene korisnike u kontekstu te teme.

2.4 Povezani rad

Iz idejne perspektive, dva naslova možemo vidjeti kao usko povezana s TweetRank modelom: rangiranje autoriteta za mrežni sadržaj i rangiranje polustrukturiranih podataka bazirano na grafu. Ovaj odlomak daje kratak pregled tih područja i izdvaja TweetRank od drugih postojećih rješenja.

Vrednovanje web stranica

PageRank [6], HITS [31] i SALSA [37] su istaknuti algoritmi bazirani na analizi poveznica za rangiranje web stranica. PageRank razvija model slučajne šetnje među web stranicama gdje je stalna vjerojatnost prolaska kroz određene stranice interpretirana kao mjera njihove važnosti: HITS se temelji na ideji međusobnog jačanja između ocjena važnosti (autoritet) i povezanosti (hub) web stranica. SALSA-u možemo promatrati kao složenije hibridno rješenje koje integrira ideje PageRank i HITS algoritma kombinacijom obiju uputa prelaska poveznica za konstrukciju modela grafa. Konceptualna generalizacija ove vrste metoda je data u [17]. Za razliku od TweetRank modela, ova familija metoda daje neprirodne mehanizme za izražavanje i iskorištavanje semantike veza/relacija.

Kontekstualizacija modela grafa može se postići kroz različite prilagodbe spomenutih modela. Moguće prilagodbe uključuju razna prilagođena davanja težina bridovima grafa (npr. na temelju pojave određenih pojmova u web dokumentima [41][43], klasifikaciji sadržaja [16][24], strukturnim svojstvima kao povezivanje *u domeni-izvan domene* [5]) ili zajedničko vjerojatnosno modeliranje za sadržaj i povezanost web stranica [14]. U suprotnosti sa TweetRank modelom, ova rješenja su dizajnirana za web okruženje i ne uvode razlikovanje semantike veze. Rješenje predstavljeno u [33] koristi za rangiranje web autoriteta reprezentaciju višeg reda hiperpovezanog grafa označavajući bridove grafa fiksiranim tekstom hiperpoveznica. Ova metoda je usko povezana sa TweetRank modelom, ali se bavi potpuno drugačijom postavkom problema (poveznice i vrhovi u web grafu nasuprot korisničkih relacija u društvenoj mreži).

Drugu vrstu kontekstualizacije za modele rangiranja autoriteta možemo promatrati na području personalizacije pretrage npr. Eirinaki i Vazirgiannis [18] predstavljaju modifikaciju PageRank algoritma za računanje personaliziranih preporuka web stranica dajući put posjećenih stranica. Njihov model zahtjeva pristup web server evidenciji koja pruža statistike o putevima pregledanim od strane drugih korisnika. BrowseRank [38] je dodatni primjer pristupa rangiranju stranica koji zahtjeva prikupljanje statistika o ponašanju korisnika kao što je vrijeme provedeno na web stranici. Generalizirani algoritam za personalizirano rangiranje autoriteta je opisan u [28].

Naš TweetRank pristup je dizajniran za scenarij kontekstno orijentirane preporuke za kontakt u okruženju društvene mreže. Prezentirani pristup je konceptualno općenitiji i ne oslanja se na korisničke profile i evidentirane upite. Detaljne statističke informacije o

prijašnjim interakcijama korisnika često nisu dostupne kroz API društvene mreže (jedan od razloga je čuvanje privatnosti). Ova informacija se može lako integrirati u TweetRank model ako je potrbno.

Rangiranje polustrukturiranih podataka

ObjectRank dodaje prijenosne težine autoritetima za različite tipove veza u PageRank algoritmu. Te težine utječu na slučajnu šetnju potencijalnih korisnika i dodjeljuju ih stručnjaci za domenu. Beagle++ [13] je proširenje za Beagle desktop tražilicu koja stavlja ObjectRank u RDF među podatke o desktop objektima za poboljšanje njihovog rangiranja u desktop scenarijima pretraživanja. TweetRank model također razmatra semantiku relacija. To je pristup za brzo računanje rangova za korisnike i korisničke grupe kao odgovor na *hashtag* upit. Ne oslanja se na ručno dodane težine veze te se bazira na generaliziranom HITS algoritmu umjesto PageRank algoritma.

Anyanwu i Sheth prezentiraju okvir za odgovaranje na upite s uvažavanjem tzv. semantičkih udruženja [2]. Semantičko udruženje predstavlja semantičku sličnost između putova koji povezuju različite resurse u RDF modelu. Aleman-Meza i ostali [1] su prezentirali i razvili metode za rangiranje semantičkih udruženja. Kao nastavljeni rad od Anyanwu i Sheth [2], prezentirane metode gađaju identifikaciju sličnih resursa kako bi je primjenili na scenarijima poput prevencije fatalnih pogrešaka. Njihov pristup uključuje kriterij rangiranja uzimajući u obzir strukturu grafa i korisnički kontekst. Korisnički kontekst je definiran statički odabirom ontologije koncepata koji se smatraju reprezentativnim za korisnički kontekst. Ramakrishnan i ostali prezentiraju heuristike za davanje težine uzorcima grafa što povezuju dva čvora u grafu uzimajući u obzir razlike s bridovima danim od strane RDF grafova koji uključuju *schema* informacije kodirane kao RDFS ontologije [42]. Raniji pristupi analizi uzoraka grafa prezentiraju metode s pretpostavkom da postoji samo jedan tip bridova. Uz prezentaciju heuristika, prezentiraju procjenu svojeg ciljnog pitanja koja heuristika rezultira kvalitetnijim uzorcima.

Bibliografija

[1] B. Aleman-Meza, C. Halaschek-Wiener, I. B. Arpinar, C. Ramakrishnan i A. P. Sheth, *Ranking complex relationships on the semantic web*, IEEE Internet Computing **9** (2005), 37–44.

[2] K. Anyanwu i A. P. Sheth, *The p operator: Discovering and ranking associations on the semantic web*, SIGMOND Record **31** (2002), 42–47.

[3] B. W. Bader i T. G. Kolda, *Algorithm 862: MATLAB tensor classes for fast algorithm prototyping*, ACM Trans. Math. Software **32** (2006), 635–653.

[4] _____, *Tensor Decompositions and Applications*, SIAM Review **51** (2009), 455–500.

[5] K. Bharat i M. R. Henzinger, *Improved Algorithms for Topic Distillation in a Hyperlinked Environment*, 21st Annual International ACM SIGIR Conference (1998), 104–111.

[6] S. Brin i L. Page, *The anatomy of large-scale hypertextual web search engine*, Seventh International World-Wide Web Conference (WWW 1998) (1998).

[7] R. Bro, *Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications*, Ph.D. thesis, University of Amsterdam (1998).

[8] R. Bro i C. A. Andersson, *Improving the speed of multi-way algorithms: Part II. Compression*, Chemometrics and Intelligent Laboratory Systems **42** (1998), 105–113.

[9] _____, *The n-way toolbox for matlab*, Chemometrics and Intelligent Laboratory Systems **52** (2000), 1–4.

[10] J. D. Carroll i J. J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decompositionn*, Psychometrika **35** (1970), 283–319.

[11] R. B. Cattell, *Parallel proportional profiles and other principles for determining the choice of factors by rotation*, Psychometrika **9** (1944), 267–283.

[12] ———, *The three basic factor-analytic research design–their interrelations and derivatives*, Psych. Bull. **49** (1952), 452–499.

[13] P. A. Chirita, S. Ghita, W. Nejdl i R. Paiu, *Beagle++: Semantically enhanced searching and ranking on the desktop*, ESWC (2006).

[14] D. A. Cohn i T. Hofmann, *The missing link – a probabilistic model of document content and hypertext connectivity*, 13th Conference on Advances in Neural Information Processing Systems (NIPS) (2000), 430–436.

[15] L. De Lathauwer, B. De Moor i J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl. **21** (2000), 1253–1278.

[16] M. Diligenti, M. Gori i M. Maggini, *Web Page Scoring System for Horizontal and Vertical Search*, 11th International World Wide Web onference (WWW) (2002), 508–516.

[17] C. H. Q. Ding, X. He, P. Husbands, H. Zha i H. D. Simon, *PageRank, HITS and a Unified Framework for Link Analysis*, 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2002), 353–354.

[18] M. Eirinaki i M. Vazirgiannis, *Usage-based pagerank for web personalization*, Data Mining, IEEE International Conference (2002), 130–137.

[19] N. K. M. Faber, R. Bro i P. K. Hopke, *Recent developments in CANDECOMP/PARAFAC algorithms: A critical review*, Chemometrics and Intelligent Laboratory Systems **65** (2003), 119–137.

[20] G. H. Golub i C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press (1996).

[21] R. A. Harashman, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics **16** (1970), 1–84.

[22] R. A. Harashman i M. E. Lundy, *Parafac: Parallel factor analysis*, Computational Statistics and Data Analysis **18** (1994), 39–72.

[23] J. Hastad, *Tensor rank is NP-complete*, J. Algorithms **11** (1990), 644–654.

[24] T. H. Haveliwala, *Topic-sensitive PageRank*, 11th International World Wide Web onference (WWW) (2002), 517–526.

[25] F. L. Hitchcock, *The expression of a tensor or a polyadic as a sum of products*, J. Math. Phys. **6** (1927), 164–189.

[26] ———, *Multiple invariants and generalized rank of a p-way matrix or tensor*, J. Math. Phys. **7** (1927), 39–79.

[27] J. Jaja, *Optimalevaluation of pairs of bilinear forms*, SIAM J. Comput. **8** (1979), 443–462.

[28] G. Jeh i J. Widom, *Scaling Pesonalized Web Search*, 12th International World Wide Web Conference(WWW) (2003), 271–279.

[29] J. Jiang, H. Wu, Y. Li i R. Yu, *Three-way data resolution by alternating slice-wise diagonalization (ASD) method*, J. Chemometrics **14** (2000), 15–36.

[30] H. A. L. Kiers, *Towards a standardized notation and terminology in multiway analysis*, J. Chemometrics **14** (2000), 105–122.

[31] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, J. ACM **46** (1999), 604–632.

[32] T. G. Kolda, *Multilinear Operators for Higher-Order Decompositions*, Tech. Report SAND2006-2081 (2006).

[33] T. G. Kolda, B. W. Bader i J. P. Kenny, *Higher-Order Web Link Analysis Using Multilinear Algebra*, 5th IEEE International Conference on Data Mining (ICDM) (2005), 242–249.

[34] J. B. Kruskal, *Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra Appl. **18** (1977), 95–138.

[35] ———, *Statement of Some Current Results about Three-Way Arrays*, manuscript, AT& T Bell Laboratories (1983).

[36] ———, *Rank, decomposition, and uniqueness for 3-way and N-way arrays*, in Multiway Data Analysis, R. Coppi and S. Bolasco, eds. (1989), 7–18.

[37] R. Lampel i S. Moran, *SALSA: the Stohastic Approach for Link-Structure Analysis*, ACM Transactions on Information Systems (TOIS) **19** (2001), 131–160.

[38] Y. T. Liu, B. Gao, T. Y Liu, Y. Zhang, Z. Ma, S. He i H. Li, *Browserank: letting web users vote for page importance*, SIGIR (2008), 451–458.

[39] J. Möcks, *Topographic components model for event-related potentials and some biophysical considerations*, IEEE Trans. Biomed. Engrg. **35** (1988), 482–484.

[40] P. Paatero, *A weigted non-negative least squares algorithm for three-way PARAFAC factor analysis*, Chemometrics and Intelligent Laboratory Systems **38** (1997), 223–242.

[41] D. Rafiei i A. O. Mendelzon, *What is this Page known for? Computing Web Page Reputations*, Computer Networks **33** (2000), 823–835.

[42] C. Ramakrishnan, W. H. Milnor, M. Perry i A. P. Sheth, *Discovering informative connection subgraphs in multi-rational graphs*, SIGKDD Explor. Newsl. **7** (2005), 56–63.

[43] M. Richardson i P. Domingos, *The inteligent surfer: Probabilistic Combination of Link and Content Information in PageRank*, 14th Conference on Advances in Neural Information Processing Systems (NIPS) (2001), 1441–1448.

[44] S. Sizov, S. Staab i T. Franz, *Analysis of Social Networks by Tensor Decomposition*, Handbook of Social Network Technologies and Applications (B. Furht, ur.), Springer, str. 45–58.

[45] A. Smilde, R. Bro i P. Geladi, *Multi-Way Analysis: Applications in the Chemical Sciences*, (2004).

[46] J. M. F. Ten Berge, *Kruskal's polynomial for* $2 \times 2 \times 2$ *arrays and a generalization to* $2 \times n \times n$ *arrays*, Psychometrika **65** (2000), 631–636.

[47] _____, *The typical rank of tall three-way arrays*, Psychometrika **65** (2000), 525–532.

[48] _____, *Partial uniqueness in CANDECOMP/PARAFAC*, J. Chemometrics **18** (2004), 12–16.

[49] J. M. F. Ten Berge i H. A. L. Kiers, *Simplicity of core arrays in three-way principal component analysis and the typical rank of* $p \times q \times 2$ *arrays*, Linear Algebra Appl. **294** (1999), 169–179.

[50] G. Tomasi i R. Bro, *A comparison of algorithms for fitting the PARAFAC model*, Comput. Statist. Data Anal. **50** (2006), 1700–1734.

[51] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math. **123** (2000), 85–100.

Sažetak

U ovom radu iznesene su osnovne definicije vezane uz tenzore. Detaljno je opisana notacija i operacije nad tenzorima. Dotaknuto je određivanje ranga tenzora kao i razlike u usporedbi svojstava tenzorskog s matričnim rangom (rang tenzora može biti različit nad $\mathbb{R}$ i nad $\mathbb{C}$, problem određivanja ranga tenzora je NP-težak). Nakon toga rad se bavi određivanjem PARAFAC dekompozicije tenzora trećeg reda, dok je za tenzore višeg reda iznesen pseudokod na Slici 1.7.

Također, ovom radu je predstavljen TweetRank model, novi pristup rangiranju autoriteta u zajednici društvenih mreža. Koncepcijski, TweetRank je dodatak metodama za rangiranje autoriteta znanih iz web pretraživanja kao što su PageRank i HITS. Ovaj pristup iskorištava novi reprezentacijski model za društvene grafove temeljen na trodimenzionalnim tenzorima. To nam omogućuje da prirodnim putem iskoristimo semantiku korisničkih relacija. Primjenom PARAFAC tenzorske dekompozicije identificiramo autoritativne izvore u društvenoj mreži kao i grupe semantički usko povezanih pojmova od interesa. Iz tog razloga TweetRank model možemo smatrati sljedećim korakom prema učinkovitijoj i djelotvornijoj tehnologiji pretraživanja/preporučivanja za društvenu mrežu.

Summary

In this work we presented basic definitions related with tensors. Specified description is made for notation and tensor operations. We also observed determination of tensor rank and differences in comparation between property of tensor and matrix rank (tensor rank can be different over $\mathbb{R}$ and over $\mathbb{C}$, the problem of determining tensor rank is NP-hard). After that this work deals with determining PARAFAC tensor decomposition of third-order tensors, while for the higher-order tensors is presented pseudocode on Figure 1.7.

Also, in this work we presented TweetRank, a novel approach for authority ranking in Social Web communities. Conceptually, TweetRank is a correspondent to authority ranking methods known from Web retrieval, such as PageRank or HITS. This approach exploits the novel representational model for social graphs, based on 3-dimensional tensors. This allows us to exploit in the natural way the available semantics of user relations. By applying the PARAFAC tensor decomposition we identify authoritative sources in the social network as well as groups of semantically coherent terms of interest. Therefore, TweetRank can be seen as a next step towards efficient and effective search/recommendation technology for the Social Web.

Životopis

Rođen sam 9. kolovoza 1988. godine u Livnu, BiH. Godinu dana nakon rođenja selim u Sinj gdje sam kasnije pohađao Osnovnu školu fra Pavla Vučkovića. 2003. godine upisujem Gimnaziju Dinka Šimunovića u Sinju koju završavam 2007. godine. Iste godine upisujem preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Nakon završenog preddiplomskog studija, 2012. godine upisujem diplomski sveučilišni studij Računarstvo i matematika na istom odsjeku.