TOPHITS algoritam

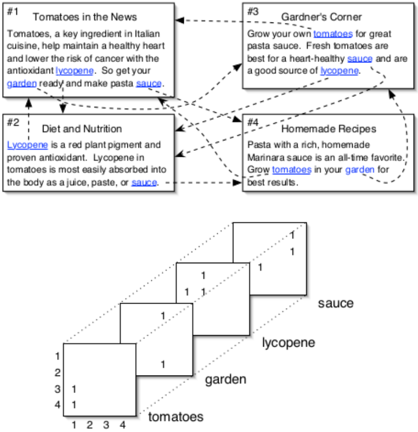
Bruno Fabulić, Helena Marciuš, Dora Parmač

8. ožujka 2021.

Sadržaj

1. HITS algoritam
2. TOPHITS algoritam
3. PARAFAC dekompozicija
   - Pohlepni PARAFAC
   - Alternirajući najmanji kvadrati za PARAFAC
4. TOPHITS i upiti

- Povećanjem *Weba*, sve je važnije analizirati strukturu linkova uzimajući u obzir i kontekst
- TOPHITS algoritam - generalizacija HITS algoritma, dodaje treću dimenziju za oblikovanje tenzora susjedstva koji uključuje tekst poveznice



HITS algoritam

- Internet prikazujemo usmjerenim grafom G - *web* stranice su čvorovi, poveznice između stranica su bridovi
- Postoji brid između čvora i i čvora j ako postoji poveznica sa stranice i na stranicu j
- Svakoj *web* stranici pridružuje se autoritet-vrijednost a_i te hub-vrijednost h_i
- Neka je n broj *web* stranica. Autoritet-vrijednost i hub-vrijednost stranice i definira se sljedećim formulama:

$$a_i = \sum_{j \to i} h_j \qquad h_i = \sum_{i \to j} a_j, \quad i = 1, \dots n \tag{1}$$

- Ove se vrijednosti iterativno računaju na sljedeći način:

$$h_i^{(t+1)} = \sum_{i \to j} a_j^{(t)} \qquad a_i^{(t+1)} = \sum_{j \to i} h_j^{(t+1)}, \quad i = 1, \dots n \tag{2}$$

HITS algoritam

- Usmjerenom grafu G pridružujemo matricu susjedstva A:

$$\mathrm{A}_{ij} = \begin{cases} 1, & \texttt{ako } i \to j \\ 0, & \texttt{inače} \end{cases} \tag{3}$$

- Jednakosti (2) mogu se zapisati kao

$$\mathrm{h}^{(t+1)} = \mathrm{A}\mathrm{a}^{(t)} \qquad \mathrm{a}^{(t+1)} = \mathrm{A}\mathrm{h}^{(t+1)} \tag{4}$$

gdje su h i a vektori hub, tj. autoritet vrijednosti

- Koristeći SVD dekompoziciju, možemo pronaći aproksimaciju od A ranga p:

$$\mathrm{A} \approx \sum_{i=1}^{p} \sigma^{(i)} \mathrm{u}^{(i)} \circ \mathrm{v}^{(i)} \tag{5}$$

- Vektori hub i autoritet vrijednosti u iteracijama (4) konvergiraju dominantnim singularnim vektorima:

$$\mathrm{h}^{(\mathrm{t})} \to \mathrm{u}^{(1)} \qquad \mathrm{a}^{(\mathrm{t})} \to \mathrm{v}^{(1)} \tag{6}$$

TOPHITS algoritam

- Kombinira *anchor* text s *hyperlink* strukturom Interneta
- Osim hub i autoritet vrijednosti, računa i *topic* vrijednost za pojmove u *anchor* tekstu
- Može se računati iterativno. Ako je n broj *web* stranica i m broj pojmova. Tada vrijedi

$$
\begin{aligned}
h_i^{(t+1)} &= \sum_{i \xrightarrow{\text{k}} j} a_j^{(t)} t_k^{(t)} \\
a_j^{(t+1)} &= \sum_{i \xrightarrow{\text{k}} j} h_i^{(t+1)} t_k^{(t)} \\
t_k^{(t+1)} &= \sum_{i \xrightarrow{\text{k}} j} a_j^{(t+1)} h_i^{(t+1)}
\end{aligned}
\tag{7}
$$

TOPHITS algoritam

- Za prikazivanje veza između *web* stranica koristi se tenzor $\mathbf{A}$ reda 3:

$$\mathrm{A}_{ijk} = \begin{cases} 1, & \text{ako } i \to j \text{ s anchor tekstom } k \\ 0, & \text{inače} \end{cases} \tag{8}$$

- Jednakosti (7) mogu se zapisati kao

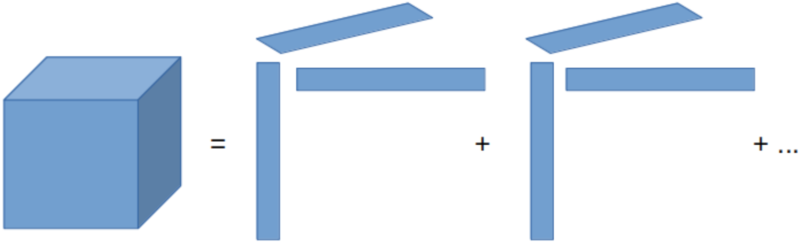
$$\begin{aligned} \mathrm{h}^{(t+1)} &= \mathbf{A} \times_2 \mathrm{a}^{(t)} \times_3 \mathrm{t}^{(t)} \\ \mathrm{a}^{(t+1)} &= \mathbf{A} \times_1 \mathrm{h}^{(t+1)} \times_3 \mathrm{t}^{(t)} \\ \mathrm{t}^{(t+1)} &= \mathbf{A} \times_1 \mathrm{h}^{(t+1)} \times_2 \mathrm{a}^{(t+1)} \end{aligned} \tag{9}$$

- $\times_n$ - množenje u modu n. Npr., za $\mathrm{h} = \mathbf{A} \times_2 \mathrm{a} \times_3 \mathrm{t}$ vrijedi $\mathrm{h}_i = \sum_{i=1}^{n} \sum_{k=1}^{m} \mathrm{A}_{ijk} \mathrm{a}_j \mathrm{t}_k,\ i = 1, \ldots n$

TOPHITS algoritam

- Za izračunavanje vektora hub, autoritet i topic vrijednosti, može se koristiti PARAFAC dekompozicija
- Pomoću PARAFAC dekompozicije, računamo aproksimaciju ranga p tenzora $\mathbf{A}$:

$$\mathbf{A} \approx \sum_{i=1}^{p} \sigma^{(i)} \mathrm{u}^{(i)} \circ \mathrm{v}^{(i)} \circ \mathrm{w}^{(i)} \tag{10}$$



TOPHITS algoritam

- Može se implementirati algoritam koji uz odgovarajuće uvjete daje aproksimaciju tako da iteracije (9) konvergiraju prema dominantnim singularnim vektorima:

$$\mathrm{h}^{(t)} \to \mathrm{u}^{(1)}, \quad \mathrm{a}^{(t)} \to \mathrm{v}^{(1)}, \quad \mathrm{t}^{(t)} \to \mathrm{w}^{(1)} \tag{11}$$

- Najveća vrijednost u $\mathrm{w}^{(1)}$ definira domminantnu temu, a najveće vrijednosti u $\mathrm{u}^{(1)}$ i $\mathrm{v}^{(1)}$ dominantan hub i autoritet za dominantnu temu.

PARAFAC dekompozicija

- Cilj: za tenzora reda N pronaći dekompoziciju tako da se tenzor može zapisati kao suma produkta vektora
- Neka je $\mathcal{X}$ tenzor reda N dimenzija $I_1 \times I_2 \times \cdots \times I_N$ i $R > 0$ željeni rang
- Tražimo matrice $\mathrm{U}^{(n)}$ dimenzija $I_n \times R$ za $n = 1, \ldots, N$ i vektor težina λ duljine R tako da

$$\mathcal{X} \approx \lambda[[\mathrm{U}^{(1)}, \mathrm{U}^{(2)}, \ldots, \mathrm{U}^{(N)}]] = \sum_{r=1}^{R} \lambda_r \mathrm{u}_r^{(1)} \circ \mathrm{u}_r^{(2)} \circ \cdots \circ \mathrm{u}_r^{(N)} \quad (12)$$

- Rješavamo sljedeći problem minimizacije:

$$\min \left\| \mathcal{X} - \lambda[[\mathrm{U}^{(1)}, \mathrm{U}^{(2)}, \ldots, \mathrm{U}^{(N)}]] \right\|^2 \quad (13)$$

s obzirom na $\lambda \in \mathbb{R}^R, \; \mathrm{U}^{(n)} \in \mathbb{R}^{I_n \times R}, \; n = 1, \ldots, N$

Pohlepni PARAFAC - pseudokod

ulaz: tenzor $\mathbf{X}$ dimenzija $I_1 \times I_2 \times \cdots \times I_N$, rang $R > 0$
for r = 1, . . . , R **do**
  inicijaliziraj $v^{(n)}$ kao vektor jedinica duljine I_n, $n = 1, \ldots, N$
  do
    for $n = 1, \ldots, N$ **do**
      $\mathrm{z}^{(n)} = \mathrm{v}^{(1)} \otimes \cdots \otimes \mathrm{v}^{(n-1)} \otimes \mathrm{v}^{(n+1)} \otimes \mathrm{v}^{(N)}$
      $\mathrm{w} = \mathrm{X}_{(n)}\mathrm{z}^{(n)} - \sum_{i=1}^{r-1} \left(\mathrm{u}_i^{(n)} \prod_{\substack{m=1 \\ m \neq n}}^{N} (\mathrm{v}^{(n)})^T \mathrm{u}_i^{(m)} \right)$
      $\lambda_r = \|\mathrm{w}\|$
      $\mathrm{v}^{(n)} = \mathrm{w}/\lambda_r$
    end for
  while fit ceases to improve
  $\mathrm{u}_r^{(n)} = \mathrm{v}^{(n)}$, $n = 1, \ldots, N$
end for
izlaz: $\lambda \in \mathbb{R}^R$, $\mathrm{U}^{(n)} \in \mathbb{R}^{I_n \times R}$, $n = 1, \ldots, N$

Pohlepni PARAFAC

- Vanjska petlja (po r) računa $\{\mathrm{u}_r^{(1)}, \ldots, \mathrm{u}_r^{(N)}\}$
- Za svaki $r, r = 1, \ldots, N$, unutarnja petlja metodom alternirajućih najmanjih kvadrata minimizira

$$\left\| \left(\mathbf{X} - \sum_{i=1}^{r-1} \lambda_i \mathrm{u}_i^{(1)} \circ \cdots \circ \mathrm{u}_i^{(N)} \right) - \left(\mathrm{v}^{(1)} \circ \cdots \circ \mathrm{v}^{(N)} \right) \right\| \tag{14}$$

po vektorima $\mathrm{v}^{(1)}, \ldots, \mathrm{v}^{(N)}$

- Tada je $\mathrm{u}_r^{(n)} = \mathrm{v}^{(n)}, \; n = 1, \ldots, N$

Alternirajući najmanji kvadrati za PARAFAC - pseudokod

ulaz: tenzor $\mathbf{X}$ dimenzija $I_1 \times I_2 \times \cdots \times I_N$, rang $R > 0$
Inicijaliziraj $\mathrm{U}^{(n)}$ za $n = 1, \ldots, N$
do
  for $n = 1, \ldots, N$ **do**
    $\mathrm{Z}^{(n)} = U^{(1)} \odot \ldots U^{(n-1)} \odot U^{(n+1)} \odot \cdots \odot U^{(N)}$
    $\mathrm{Y}^{(n)} = (\mathrm{U}^{(1)T}\mathrm{U}^{(1)} * \ldots \mathrm{U}^{(n-1)T}\mathrm{U}^{(n-1)} * \mathrm{U}^{(n+1)T}\mathrm{U}^{(n+1)} * \cdots * \mathrm{U}^{(N)T}\mathrm{U}^{(N)})^{-1}$
    $\mathrm{V} = \mathrm{X}_{(n)}\mathrm{Z}^{(n)}\mathrm{Y}^{(n)}$
    for $r = 1, \ldots, R$ **do**
      $\lambda_r = \|\mathrm{v}_r\|$
      $\mathrm{u}_r^{(n)} = \mathrm{v}_r / \lambda_r$
    end for
  end for
while fit ceases to improve
izlaz: $\lambda \in \mathbb{R}^R$, $\mathrm{U}^{(n)} \in \mathbb{R}^{I_n \times R}$, $n = 1, \ldots, N$

Alternirajući najmanji kvadrati za PARAFAC

- U svakoj iteraciji računamo matricu $\mathrm{U}^{(n)} \equiv \mathrm{V}$, a ostale su fiksirane, tj. rješavamo minimizacijski problem:

$$\min_{V} \left\| \mathbf{X} - [[\mathrm{U}^{(1)}, \ldots, \mathrm{U}^{(n-1)}, \mathrm{V}, \mathrm{U}^{(n+1)}, \ldots, \mathrm{U}^{(N)}]] \right\|^2 \tag{15}$$

što se može zapisati kao

$$\min_{V} \left\| \mathrm{X}_n - \mathrm{V}\mathrm{Z}^{(n)T} \right\|^2 \tag{16}$$

- Rješenje gornjeg problema je

$$\mathrm{V} = \mathrm{X}_{(n)} \left(\mathrm{Z}^{(n)T} \right)^{\dagger} \tag{17}$$

- Može se pokazati

$$\left(\mathrm{Z}^{(n)T} \right)^{\dagger} = \mathrm{Z}^{(n)} \mathrm{Y}^{(n)T} \tag{18}$$

Inicijalizacija $\mathrm{U}^{(n)}$

- Pohlepni PARAFAC - inicijaliziramo matrice $\mathrm{U}^{(n)}$ kao matrice dobivene pohlepnim PARAFAC algoritmom
- Nasumična inicijalizacija - matrice $\mathrm{U}^{(n)}$ inicijaliziramo nasumičnim vrijednostima
- HOSVD inicijalizacija - za svaki mod $\mathrm{X}_{(n)}$ izračunamo SVD dekompoziciju, $\mathrm{X}_{(n)} = \mathrm{U}_n \Sigma_n \mathrm{V}_n^T$, i inicijaliziramo matrice $\mathrm{U}^{(n)} = \mathrm{U}_n, \ n = 1, \dots, N$

TOPHITS + PARAFAC

- Tenzor $\mathbf{A}$ reda 3 dimenzija $N \times N \times M$ i $R > 0$
- PARAFAC dekompozicijom dobivamo matrice $\mathrm{U}^{(1)} \equiv \mathrm{H}$ i $\mathrm{U}^{(2)} \equiv \mathrm{A}$ dimenzija $N \times R$ te $\mathrm{U}^{(3)} \equiv \mathrm{T}$ dimenzije $M \times R$, te vektor $\lambda \in \mathbb{R}^R$ tako da vrijedi

$$\mathbf{A} \approx \lambda[[\mathrm{H}, \mathrm{A}, \mathrm{T}]] = \sum_{r=1}^{R} \lambda_r \mathrm{h}_r \circ \mathrm{a}_r \circ \mathrm{t}_r \tag{19}$$

- Svaka trojka $\{\mathrm{h}_r,\ \mathrm{a}_r,\ \mathrm{t}_r\}$ određuje grupu hubova, autoriteta i pojmova; λ_r definira težinu grupe
- $\{\mathrm{h}_1,\ \mathrm{a}_1,\ \mathrm{t}_1\}$ određuju najbolji hub, autoritet i pojam za dominantnu temu

TOPHITS i upiti

- Kako pronaći sve stranice povezane termom ili skupom termova?
- Neka je q vektor duljine M (M je broj termova) definiran s

$$q_k = \begin{cases} 1, \text{ ako je term } k \text{ u upitu} \\ 0, \text{ inače} \end{cases} \tag{20}$$

- Pomoću ovog vektora možemo pronaći podudarne grupe i skupove autoriteta za zadani upit

Pronalaženje podudarne grupe

- Možemo identificirati grupiranja relevantna za dani upit
- Računamo vektor s koji sadrži vrijednost svakog grupiranja:

$$s = \mathbf{\Lambda}\mathrm{T}^T\mathrm{q}, \quad \mathbf{\Lambda} = diag(\lambda) \tag{21}$$

- s_r daje vrijednost r-te grupe, a što je vrijednost veća, grupa je relevantnija za dani upit
- Alternativno, možemo konstruirati vektor upita $\hat{\mathrm{q}} \in \mathbb{R}^N$ na temelju web stranica te izračunati rezultate grupe kao

$$\hat{s} = \mathbf{\Lambda}\mathrm{A}^T\hat{\mathrm{q}}, \quad \mathbf{\Lambda} = diag(\lambda) \tag{22}$$

Pronalaženje skupa autoriteta

- Možemo vratiti tradicionalnu ranginaru listu vjerojatnosti - kombiniramo informacije dobivene TOPHITS-om da bi vratili skup kombiniranih autoriteta ili hubova
- Ako definiramo s kao u (21), tada su kombinirani autoriteti definirani s

$$a^* = \mathrm{A}s = \sum_{r=1}^{R} s_r a_r \tag{23}$$

- Sortiranjem a^* dobivamo rangiranu listu autoriteta
- Analogno računamo kombinirane hubove:

$$h^* = \mathrm{H}s = \sum_{r=1}^{R} s_r h_r \tag{24}$$