Tenzor link analysis

Bruno Fabulić, Helena Marciuš, Dora Parmać

5. ožujka 2021.

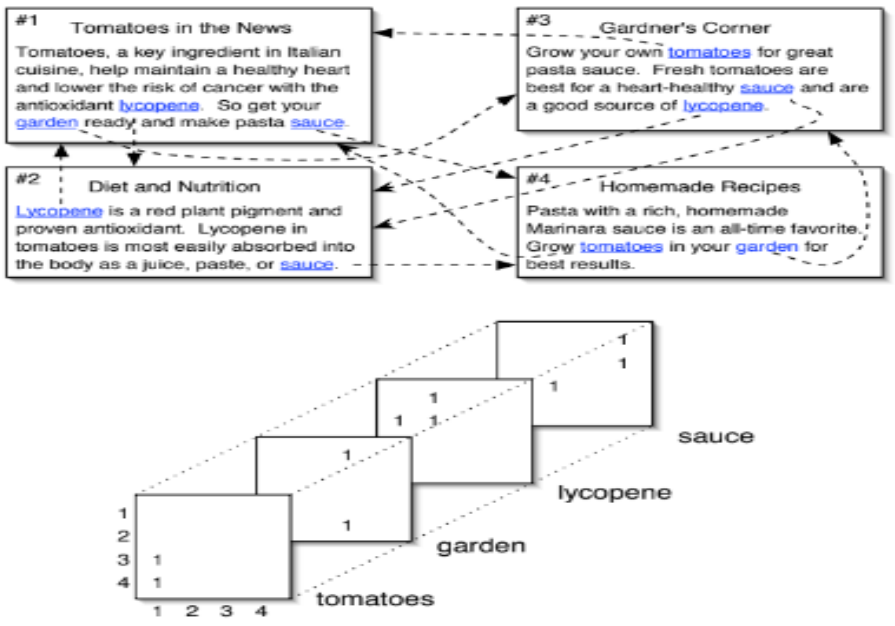
Sadržaj

1 Uvod

Kako se veličina Weba povećava, postaje sve važnije analizirati strukturu linka uzimajući u obzir i kontekst. Multilinearna algebra pruža novi alat za obuhvaćanje teksta poveznice i drugih informacija koji se koriste u drugih metodama analize poveznica kao što je HITS algoritam.

Algoritam kojeg proučavamo naziva se Topical HITS ili TOPHITS. On je generalizacija Kleinbergovog HITS modela. TOPHITS dodaje 3. dimenziju za oblikovanje tenzora susjedtstva koji uključuje i tekst poveznice. Ova dodatna informacija pruža način za obuhvaćanje konteksta uz izračunavanje autoriteta i hubova, što je postignuto 3D Parallel Factors (PARAFAC) dekompozicijom, više-dimenzionalni analogon SVD dekompozicije. Dodavanje 3. dimenzije znatno poboljšava primjenjivost HITS-a, zato što se TOPHITS analiza može izvoditi unaprijed i offline. Također, TOPHITS otkriva latentno grupiranje stranice i informacija.

U ovom seminaru, opisat ćemo brži matematički algoritam za izračunavanje TOPHITS algoritma na podatcima.



Slika 1: TOPHITS analizira 3D tenzor koji reprezentira kolekciju web stranica

2 HITS i TOPHITS

Mnoge metode za analizu weba, poput PageRanka i HITS-a bazirane su na matrici susjedstva grafa kolekcije web stranica. PageRank rezultati dani su ulazima glavnog svojstvenog vektora Markovljeve matrice vjerojatnosti. S druge strane, HITS računa hubove i autoritete za svaki vrh i oni odgovaraju glavnom lijevom i desnom singularnom vektoru matrice susjedstva. Zanimljivo svojstvo HITS-a, što ne dijeli s PageRankom jest da se višestruki parovi singularnih vektora mogu razmatrati.

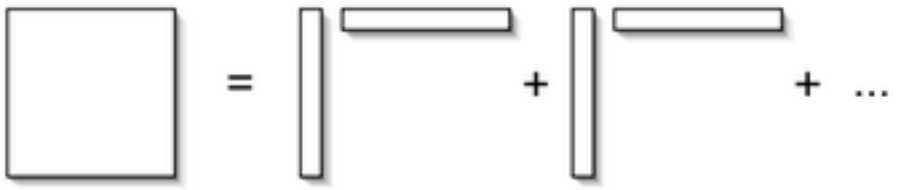
Promotrimo kolekciju I web stranica. U HITS-u, IxI matrica susjedstva $\boldsymbol{X}$ definirana je kao:

$$x_{i,j} = \begin{cases} 1 & \text{ako stranica i pokazuje na stranicu j} \\ 0 & \text{inače} \end{cases} \tag{1}$$

HITS metoda može biti objašnjena na sljedeći način: koristi SVD matričnu dekompoziciju za računanje ranga R aproksimacije od $\boldsymbol{X}$:

$$\boldsymbol{X} \approx \boldsymbol{H\Sigma A^T} \equiv \sum_{\boldsymbol{r=1}}^{\boldsymbol{R}} \boldsymbol{\sigma_r h_r \circ a_r} \tag{2}$$

Vrijedi $\Sigma = diag\{\sigma_1 \dots \sigma_R\}$ i pretpostavljamo $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R > 0$. Matrice $\boldsymbol{H} i \boldsymbol{A}$ su veličine $I \times R$ i imaju ortonormalna polja. To možemo zamisliti kao aproksimaciju matrice $\boldsymbol{X}$ sumom R RANK-1 OUTER PRODUCTS, kako je prikazano na (2) Glavni par singularnih vektora $\boldsymbol{h_1}$ i $\boldsymbol{a_1}$ pruža hub i aurotitet rezultate za dominantnu temu u kolekciji web stranica. Drugim riječima, stranice koje imaju najveće rezultate $\boldsymbol{h_1}$ su najbolji hubovi za dominantnu temu, a stranice koje imaju najveće rezultate u $\boldsymbol{a_1}$ su najbolji autoriteti.



Slika 2: SVD dekompozicija

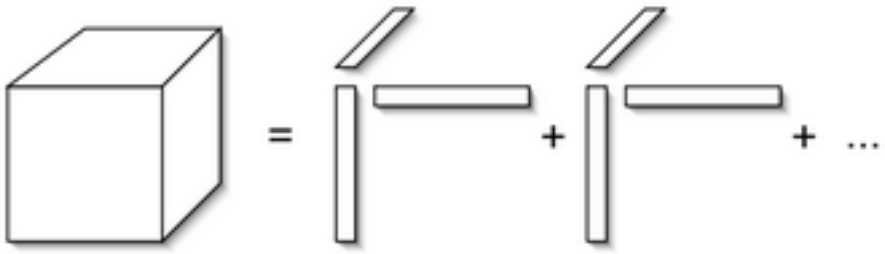
Ilustrirajmo sada TOPHITS algoritam. Tu je $I \times I \times K$ tenzor susjedstva $\boldsymbol{\mathcal{X}}$ definiran kao:

$$x_{i,j,k} = \begin{cases} 1 & \text{ako stranica i pokazuje na stranicu j koristeći term k} \\ 0 & \text{inače} \end{cases} \tag{3}$$

TOPHITS koristi PARAFAC model za generiranje R-ranga aproksimacije:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\lambda}[\![\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{T}]\!] \equiv \sum_{\boldsymbol{r=1}}^{\boldsymbol{R}} \boldsymbol{\lambda_r} \boldsymbol{h_r} \circ \boldsymbol{a_r} \circ \boldsymbol{t_r} \tag{4}$$

Ovdje pretpostavljamo $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_R$. Matrice $\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{T}$ imaju polja duljine 1, ali suprotno od SVD-a, nisu generalno ortonormalne. PARAFAC dekompozicija aproksimira tenzor $\boldsymbol{\mathcal{X}}$ sumom R ranga 1 OUTER PRODUCTS, kao što je prikazano na (3)



Slika 3: PARAFAC dekompozicija

Glavna trojka PARAFAC vektora $\boldsymbol{h_1}$, $\boldsymbol{a_1}$, $\boldsymbol{t_1}$ pruža hub i aurotitet rezultate za dominantnu temu u kolekciji web stranica (na isti način kao što je prethodno opisano kod HITS-a), te term rezultate. Drugim riječima, termi koji imaju najveće rezultate $\boldsymbol{t_1}$ su najdeskriptivniji termi.

3 TOPHITS model

Ideja iza TOPHITSa je sljedeća: pretpostavimo da analiziramo kolekciju I web stranica imajući ukupno K termova u tekstu povezncia svih poveznica. Tada je $IxIxK$ tenzor susjedstva $\boldsymbol{\mathcal{X}}$ definiran kao u (3). Primijetio da je tenzor $\boldsymbol{\mathcal{X}}$ rijedak zato što većina stranica pokazuje samo na na nekoliko drugih stranica u kolekciji, a svaki link koristi samo nekoliko termova. Stoga je realno za očekivati da je broj ne nul elemenata u $\boldsymbol{\mathcal{X}}$ jednak $O(I)$.

Uzimajući u obzir $R > 0$ (broj različitih grupiranja podataka), TOPHITS algoritam pronađe matrice $\boldsymbol{HiA}$, veličina IxR, i matricu $\boldsymbol{T}$ veličine KxR, TO YIELD. Svaka trojka $\{\boldsymbol{h_r}, \boldsymbol{a_r}, \boldsymbol{t_r}\}$, za $r = 1 \dots R$ definira grupiranje hubova, autoriteta i terma uzimajući u obzir ulaze s najvećom vrijednosti u svakom vektoru; vrijednost λ_r definira težinu grupiranja. (Bez smanjenja općenitosti, pretpostavljamo da su stupci naših matrica normalizirani.)

3.1 PARAFAC model

3d dekompozicija simultano je predložena sa strane Harshmana, koristeći ime "Parallel Factors ili PARAFAC" te sa strane Carroll i Chang, koristeći ime "Canonical Decomposition ili CANDECOMP". Cilj je prikazati vektor reda N kao sumu vektora vanjskog produkta kao što je prikazano na slici 3.

Matematički, problem je zadan na sljedeći način. Pretpostavimo da imamo tenzor $\boldsymbol{\mathcal{X}}$ veličine $I_1 x I_2 x \dots x I_n$ i željeni rang aproksimacije R. Tada želimo naći matrice $\boldsymbol{U}^n$ veličine $I_n x R$, za svaki $N = 1, \dots, N$ i težinski vektor $\boldsymbol{\lambda}$ duljine R tako da vrijedi:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\lambda}[\![\boldsymbol{U^{(1)}}, \boldsymbol{U^{(2)}}, \dots, \boldsymbol{U^{(N)}}]\!] \tag{5}$$

Kruskalov operator $[\![.]\!]$ je skraćeni zapis sume ranga svih vanjskih produkata stupaca. Drugim riječima,

$$\lambda[[\mathrm{U}^{(1)}, \mathrm{U}^{(2)}, \dots, \mathrm{U}^{(N)}]] \equiv \sum_{r=1}^{R} \lambda_r \mathrm{u}_r^{(1)} \circ \mathrm{u}_r^{(2)} \circ \dots \circ \mathrm{u}_r^{(N)} \tag{6}$$

Bez smanjenja općenitosti, prepostavljamo da je $\left\| \boldsymbol{u}_r^{(n)} \right\| = 1$ za sve $r = 1, \dots, R$ i $n = 1, \dots, N$. Povrh toga, preuredimo završno rješenje tako da vrijedi $\lambda_1 \geq \lambda_2 \geq \dots \lambda_R$.

Cilj nam je riješiti problem minimizacije:

$$\min \left\| \mathrm{X} - \lambda[[\mathrm{U}^{(1)}, \mathrm{U}^{(2)}, \dots, \mathrm{U}^{(N)}]] \right\|^2 \tag{7}$$

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

U slučaju TOPHITSA, $\boldsymbol{\lambda}$ je tenzor reda 3, tj $N = 3$ i vrijedi:

$$\boldsymbol{H} \equiv \boldsymbol{U^{(1)}}, \boldsymbol{A} \equiv \boldsymbol{U^{(2)}}, \boldsymbol{T} \equiv \boldsymbol{U^{(3)}} \tag{8}$$

3.2 Pohlepni PARAFAC

Notacija $\boldsymbol{X}_{(n)}$ reprezentira n-to UNFOLDING tenzora $\boldsymbol{\mathcal{X}}$. Drugim riječima, $\boldsymbol{X}_{(n)}$ predstavlja reorganizirane ulaze $\boldsymbol{\mathcal{X}}$ u matricu veličine $I_n \times J$, gdje je $J = \prod_{k=1,k\neq n}^{N} I_k$ tako da su "vlakna" u dimenziji n uređena kao stupci matrice. Matematički imamo:

$$[\boldsymbol{X}_{(n)}]_{i,j} = x_{i_1,i_2,\ldots,i_N}$$
$$i = i_n, j = 1 + \sum_{k=1,k\neq n}^{N} (i_n - 1) \prod l = 1, l \neq n^{k-1} I_l$$
$$i \leq i \leq I_n, 1 \leq j \leq J \tag{9}$$

3.2.1 Pohlepni PARAFAC - pseudokod

Algorithm 1 Greedy PARAFAC

in: Tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \cdots \times I_N$.
in: Desired rank $R > 0$.
for $r = 1, \ldots, R$ **do** {outer loop}
  Set $\mathbf{v}^{(n)}$ to be a vector of all ones of length I_n for $n = 1, \ldots, N$.
  repeat {middle loop}
    for $n = 1, \ldots, N$ **do** {inner loop}

$$\text{Set } \mathbf{w} = \mathbf{X}_{(n)}\mathbf{z}^{(n)} - \sum_{i=1}^{r-1} \left(\mathbf{u}_i^{(n)} \prod_{\substack{m=1 \\ m \neq n}}^{N} (\mathbf{v}^{(n)})^{\mathsf{T}} \mathbf{u}_i^{(m)} \right) \text{ where } \mathbf{z}^{(n)} \equiv \mathbf{v}^{(1)} \otimes \cdots \otimes \mathbf{v}^{(n-1)} \otimes \mathbf{v}^{(n+1)} \otimes \cdots \otimes \mathbf{v}^{(N)}.$$

    Set $\lambda_r = \|\mathbf{w}\|$.
    Set $\mathbf{v}^{(n)} = \mathbf{w}/\lambda_r$.
    end for
  until the fit ceases to improve or the maximum number of middle-loop iterations has been exceeded.
  Set $\mathbf{u}_r^{(n)} = \mathbf{v}^{(n)}$ for $n = 1, \ldots, N$.
end for
out: $\boldsymbol{\lambda} \in \mathbb{R}^R$ and $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R}$ for $n = 1, \ldots, N$.

1. 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

2. $$\left\| \left(\mathrm{X} - \sum_{i=1}^{r-1} \lambda_i \mathrm{u}_i^{(1)} \circ \cdots \circ \mathrm{u}_i^{(N)} \right) - \left(\mathrm{v}^{(1)} \circ \cdots \circ \mathrm{v}^{(N)} \right) \right\| \qquad (10)$$

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

3.3 Alternirajući najmanji kvadrati za PARAFAC

Češći pristup rješavanja PARAFAC dekompozicija je metoda alternirajućih najmanjih kvadrata. U svakoj unutarnjoj iteraciji, računamo cijelu ntu matricu $\boldsymbol{U^{(n)}}$ dok sve ostale matrice držimo fiksne. V koji se računa nakon svake unutarnje iteracije je rješenje danog optimizacijskog problema:

$$\min_V \left\| \mathrm{X} - [[\mathrm{U}^{(1)}, \ldots, \mathrm{U}^{(n-1)}, \mathrm{V}, \mathrm{U}^{(n+1)}, \ldots, \mathrm{U}^{(N)}]] \right\|^2 \qquad (11)$$

što se matrično može zapisati kao

$$\min_{V} \left\| \mathrm{X}_n - \mathrm{V}\mathrm{Z}^{(n)T} \right\|^2 \tag{12}$$

Matrica $\boldsymbol{Z}^{(n)}$ je veličine $J \times R$ i definirana kao **??** Rješenje najmanjih kvadrata za 12 uključuje pseudo inverz matrice $\boldsymbol{Z}^{(n)}$:

$$\boldsymbol{V} = \boldsymbol{X}_{(n)}(Z^{(n)T})^{\dagger} \tag{13}$$

Prikladno, pseudo inverz $\boldsymbol{Z}^{(n)}$ ima specijalnu strukturu. Neka je $\boldsymbol{Y}^{(n)}$ simetrična $R \times R$ matrica kao u (2.7). Tada se može pokazati da je:

$$(Z^{(n)T})^{\dagger} = \boldsymbol{Z}^{(n)}\boldsymbol{Y}^{(n)T} \tag{14}$$

Dakle rješenje za 12 je dano (2.5), pa računanje $\boldsymbol{U}_n$ zapravo svodimo na računanje inverza specijalne $R \times R$ matrice $\boldsymbol{Y}^{(n)}$.

3.3.1 Alternirajući najmanji kvadrati za PARAFAC - pseudokod

Algorithm 2 Alternating Least Squares (ALS) for N-way arrays

in: Tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \cdots \times I_N$.
in: Desired rank $R > 0$.
Initialize $\mathbf{U}^{(n)}$ for $n = 1, \ldots, N$ (see §2.4).
repeat {outer loop}
  for $n = 1, \ldots, N$ **do** {inner loop}

(2.5) Set $\mathbf{V} = \mathbf{X}_{(n)}\mathbf{Z}^{(n)}\mathbf{Y}^{(n)}$,

(2.6) where $\mathbf{Z}^{(n)} \equiv \sum_{r=1}^{R} \mathbf{u}_r^{(1)} \otimes \ldots \otimes \mathbf{u}_r^{(n-1)} \otimes \mathbf{u}_r^{(n+1)} \otimes \ldots \otimes \mathbf{u}_r^{(N)}$,

(2.7) and $\mathbf{Y}^{(n)} \equiv \left(\mathbf{U}^{(1)\mathsf{T}}\mathbf{U}^{(1)} * \cdots * \mathbf{U}^{(n-1)\mathsf{T}}\mathbf{U}^{(n-1)} * \mathbf{U}^{(n+1)\mathsf{T}}\mathbf{U}^{(n+1)} * \cdots * \mathbf{U}^{(N)\mathsf{T}}\mathbf{U}^{(N)}\right)^{-1}$.

    for r=1,...,R **do** {Assign $\mathbf{U}^{(n)}$}
      Set $\lambda_r = \|\mathbf{v}_r\|$
      Set $\mathbf{u}_r^{(n)} = \mathbf{v}_r/\lambda_r$.
    end for
  end for
until the fit ceases to improve or the maximum number of outer iterations is exceeded.
out: $\boldsymbol{\lambda} \in \mathbb{R}^R$ and $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R}$ for $n = 1, \ldots, N$.

3.3.2 Inicijalizacija

- 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 $\mathrm{U}^{(n)} = \mathrm{U}_n,\ n = 1, \ldots, N$

3.4 Specijalna konsideracija za razbacane podatke

Kao što smo već spomenuli, tenzor $\boldsymbol{\mathcal{X}}$ je veoma rijedak. Posljedično, njegova reprezentacija $\boldsymbol{X}_n$ je rijetka matrica (ima iste ne nul elemente, ali preoblikovane). Matrica $\boldsymbol{Z}^n$ iz (2.6) se ne bi trebala formirati eksplicitno jer bi onda bila gusta matrice veličine $I_n \times J$ gdje je $J = \prod_{k=1,k\neq n}^{N} I_N$. Umjesto toga, računanje

$$\boldsymbol{X}_{(n)}\boldsymbol{Z}^{(n)} \tag{15}$$

potrebno za (2.5) mora biti izračunato posebno, iskorištavanjem svojstava Kroneckerove strukture produkta u $\boldsymbol{Z}^{(n)}$ da bi se očuvala rijetkost. Završni rezultat je veličine $I_n \times R$ stoga sse može pohraniti u gustu matricu. Jedna od metoda prikazana je sljedećim algoritmom:

Algorithm 3 Computing the sparse product $\mathbf{X}_{(n)}\mathbf{Z}^{(n)}$

in: Tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \cdots \times I_N$ with Q nonzeros. Let the index of the qth nonzero be $(k_{1_q}, k_{2_q}, \ldots, k_{N_q})$ and its value be given by v_q.
in: Index n and matrices $\mathbf{U}^{(m)}$ for $1 \leq m \leq N, m \neq n$.
for $r = 1 \ldots, R$ **do**
  for $q = 1, \ldots, Q$ **do**
    Compute $w_q = v_q \prod_{\substack{m=1 \\ m \neq n}}^{N} u^{(m)}_{k_{m_q}, r}$
  end for
  for $i = 1, \ldots, I_n$ **do** {Compute rth column of $\mathbf{P}$}
    Set $p_{ir} = \sum_{\substack{q=1 \\ k_{n_q} = i}}^{Q} w_q$.
  end for
end for
out: $\mathbf{P} = \mathbf{X}_{(n)}\mathbf{Z}^{(n)}$

4 TOPHITS i upiti

Jednom kada smo izračunali TOPHITS model ranga R

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\lambda}[\![\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{T}]\!] \tag{16}$$

možemo ga iskoristiti za razumijevanje podataka na razne načine. Gledajući najveću vrijednost svake trojke $\{\boldsymbol{h_r}, \boldsymbol{a_r}, \boldsymbol{t_r}\}$ pruža grupiranje hubova, autoriteta, termova web stranice, a λ_r pruža relativnu težinu grupiranja.

Jedno pitanje koje smatramo osnovnim pitanjem web pretraživanja: pronaći sve stranice povezane s određenim termom ili skupom termova. Promotrimo vektor upita g duljine K (gdje je K broj termova) kao:

$$q_k = \begin{cases} 1 & \text{ako je term k u upitu} \\ 0 & \text{inače} \end{cases} \tag{17}$$

Primijetimo da nema razloga restringirati se na upite o termovima. Također se možemo baviti problem pronalaženja web stranica i/ili termova povezanih s određenom stranicom ili skupom stranica.

4.1 Pronalaženje podudarne grupe

Umjesto vraćanja liste rangiranih stranica, TOPHITS omogućuje opciju identificiranja grupiranja relevantnih za dani upit. Možemo kreirati vektor grupe $\boldsymbol{s}$ duljine R koji sadrži vrijednosti svakog grupiranja, na temelju matrice $\boldsymbol{T}$ iz PARAFAC modela:

$$\boldsymbol{s} = \boldsymbol{\Lambda T^T q}, \boldsymbol{\Lambda} = diag(\boldsymbol{\lambda}) \tag{18}$$

Ulaz s_r daje rezultat-vrijednost r-te grupe, a grupiranja s većim rezultatima smatraju se više relevantnima.

Alternativno, možemo konstruirati vektora upita na temelju web stranica, $\hat{\boldsymbol{q}} \in R^l$ te izračunati rezultate grupe kao:

$$\hat{\boldsymbol{s}} - \boldsymbol{\Lambda A^T \hat{q}}, \boldsymbol{\Lambda} = diag(\boldsymbol{\lambda}) \tag{19}$$

4.2 Pronalaženje skupa autoriteta

Također je moguće vratit tradicionalnu rangiranu listu vjerojatnosti. Možemo kombinirati sve informacije TOPHITS modela da bi vratili skup rangiranih autoriteta i/ili hubova. Definirajmo $\boldsymbol{s}$ kao u 18. Tada su kombinirani autoriteti dani s:

$$a^* = As = \sum_{r=1}^{R} s_r a_r \tag{20}$$

Sortiranje ulaza $\boldsymbol{a^*}$ vraća rangiranu listu autoriteta. Na isti način, kombinirani hubovi dani su:

$$h^* = Hs = \sum_{r=1}^{R} s_r h_r \tag{21}$$

5 Implementacija