Introduction to Information Retrieval

2. seminar

Vector space model, link analysis

University of Pannonia

Tamás Kiezer, Miklós Erdélyi

Review (1)

- Document processing workflow
  - Parsing
  - Tokenization
  - Stopword removal
  - Stemming
  - Inverted file building (indexing)

Review (2): Vector Space Model

- Documents are mapped into term vector space
- Dimensions represent a weight for one term
- Queries are treated like documents
- Documents are ranked by their similarity to the query

Review (3): weighting methods

- Binary weighting: $w_{ij} = \begin{cases} 1 & if \quad t_i \quad occurs \ in \quad D_j \\ 0 & otherwise \end{cases}$
- Frequency weighting: $w_{ij} = f_{ij}$.
- Max-normalized (*max-tf*): $w_{ij} = \frac{f_{ij}}{\max\limits_{1 \le k \le n} f_{kj}}$
- Length-normalized (*norm-tf*): $w_{ij} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{n} f_{kj}^2}}$
- Term frequency inverse document frequency $w_{ij} = f_{ij} \times \left(\log \frac{m}{F_i} \right)$
- Length normalized term frequency inverse document frequency *(norm-tf-idf)*: $w_{ij} = \frac{f_{ij} \times \left(\log \frac{m}{F_i} \right)}{\sqrt{\sum_{k=1}^{n} \left(f_{kj} \times \left(\log \frac{m}{F_k} \right) \right)^2}}$

Similarity measures

- Cosine measure:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}}$$

- Dice's coefficient:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = 2 \cdot \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sum_{i=1}^{n} (w_{ij} + q_i)}$$

- Jaccard's coefficient:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sum_{i=1}^{n} \left(\frac{w_{ij} + q_i}{2^{w_{ij} q_i}} \right)}$$

Exercise: TD matrix, VSM model

- Build the TD matrix for the following collection:

D_1 = " Infant and Toddler First Aid "

D_2 = " Babies and Children's Room (For your Home)"

D_3 = " Child Safety at Home "

D_4 = "Your Baby's Health and Safety: From Infant to Toddler "

D_5 = " Baby Proofing Basics "

D_6 = "Your Guide to Easy Rust Proofing "

D_7 = "Beanie Babies Collectors Guide "

t_1 = "Baby"

t_2 = "Child"

t_3 = "Guide"

t_4 = "Health"

t_5 = "Home"

t_6 = "Infant"

t_7 = "Proofing

t_8 = "Safety"

t_9 = "Toddler"

- Use binary weighting, then *norm-tf*
- Rank documents to the query "child infant home proofing safety" with the learnt similarity measures

Solution

Doc.	Similarity values (Rank)		
	Cosine	**Jaccard**	**Dice**
D1	0.316 **(4.)**	0.092 **(4.)**	0.174 **(4.)**
D2	0.516 **(2.)**	0.142 **(2.)**	0.26 **(2.)**
D3	0.775 **(1.)**	0.224 **(1.)**	0.39 **(1.)**
D4	0.4 **(3.)**	0.094 **(3.)**	0.178 **(3.)**
D5	0.316 **(4.)**	0.092 **(4.)**	0.174 **(4.)**
D6	0.316 **(4.)**	0.092 **(4.)**	**0.174** **(4.)**
D7	0	0	0

Review: PageRank

- *Idea*: a Web page's importance is determined by the importance of the pages linking to it
- Extended version's interpretation: random surfer
- Simple formula: $R_i = \sum_{W_j \in B_i} \frac{R_j}{L_j}$
- Calculation by power method: $M'' = \alpha M' + (1-\alpha) M$ (with teleportation)

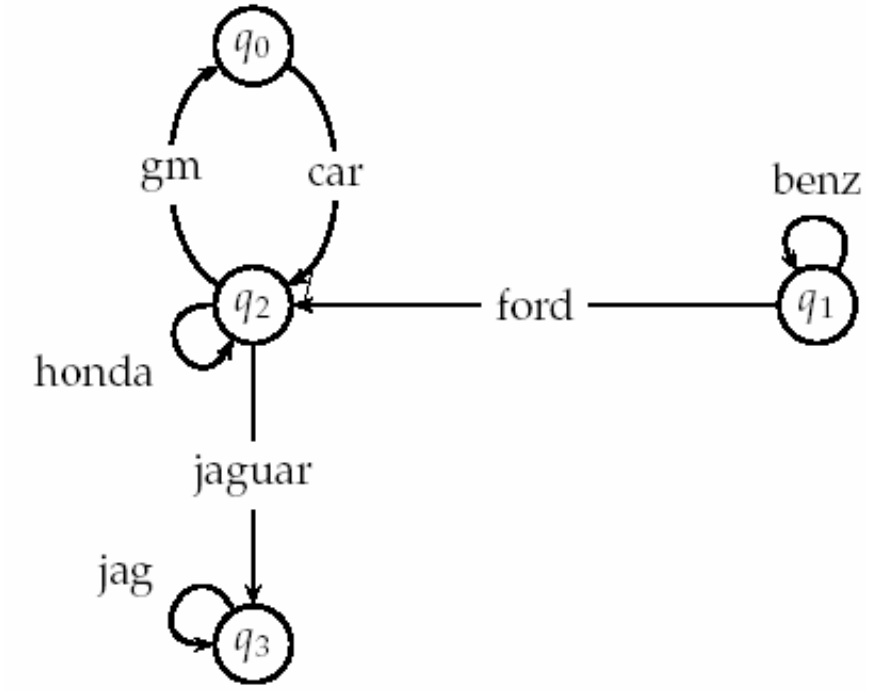
$$R_0 = [\frac{1}{N}, \ldots, \frac{1}{N}]^T$$

$$M'' \times R_{i-1} = R_i$$

$$m_{ij} = \begin{cases} \frac{1}{L_j}, \text{ if } W_j \to W_i \\ 0, \text{ otherwise} \end{cases} \quad m_{ij}' = \begin{cases} \frac{1}{N}, \text{ if } L_j = 0 \\ 0, \text{ otherwise} \end{cases}$$

Exercise: determine PR using power method

- Given the following graph:



- α = 0.15

Solution

- M'' =

0.0375	0.0375	0.8875	0.0375
0.0375	0.4625	0.4625	0.0375
0.3206	0.0375	0.3206	0.3206
0.0375	0.0375	0.0375	0.8875

- R = [0.5 0.5 0.5 0.5]

Review: HITS

- HITS = *Hyperlink Induced Topic Search*
- A *hub* is a page pointing to many *authority* pages
- An *authority* is a page pointed to by many *hub* pages
- Definition:

$$x^{\langle p \rangle} \leftarrow \sum_{q:(q,p)\in E} y^{\langle q \rangle}$$

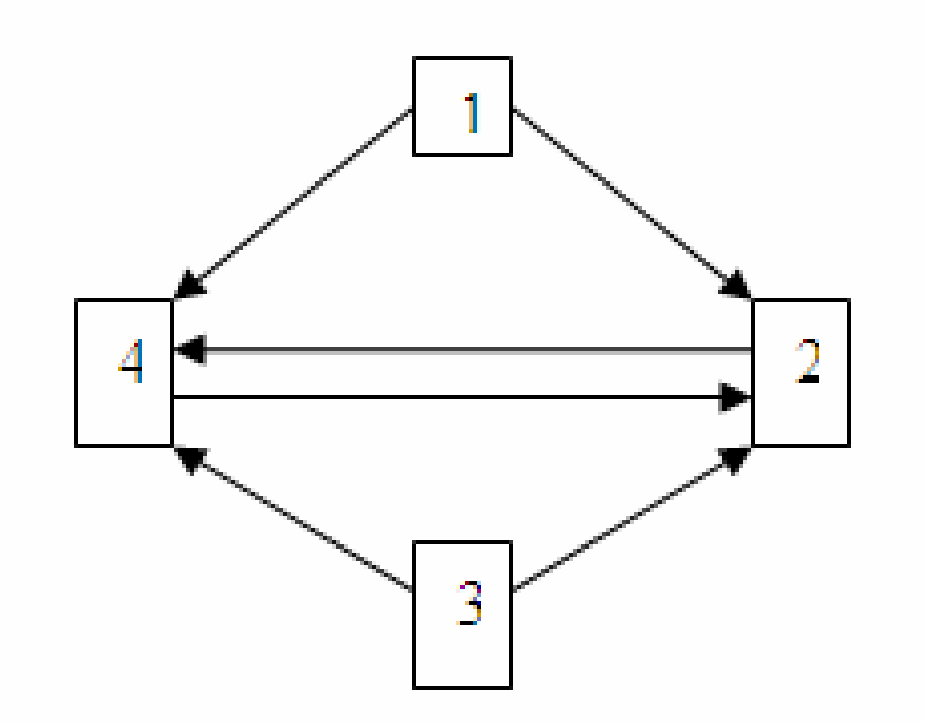
$$y^{\langle p \rangle} \leftarrow \sum_{q:(p,q)\in E} x^{\langle q \rangle}$$

- Calculation by power iteration:

$x_0 = [1,\ldots,1]^T$, $y_0 = [1,\ldots,1]^T$, then: $x_{i+1} = M^T y_i, \quad y_{i+1} = M x_{i+1}$

Exercise: applying HITS

- Calculate the hub and authority scores for the nodes of the following mini graph:



Solution

- x = [0 0.7071 0 0.7071]
- y = [0.6325 0.3162 0.6325 0.3162]

Questions?