Latent Semantic Analysis

Hongning Wang
CS@UVa
VS model in practice

- Document and query are represented by **term vectors**
  - Terms are not necessarily **orthogonal** to each other
    - Synonymy: car v.s. automobile
    - Polysemy: fly (action v.s. insect)

<table>
<thead>
<tr>
<th>Access</th>
<th>Document</th>
<th>Retrieval</th>
<th>Information</th>
<th>Theory</th>
<th>Database</th>
<th>Indexing</th>
<th>Computer</th>
<th>REL</th>
<th>MATCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>R</td>
<td>M</td>
</tr>
<tr>
<td>Doc 2</td>
<td></td>
<td></td>
<td></td>
<td>x*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>R</td>
<td>M</td>
</tr>
<tr>
<td>Doc 3</td>
<td>x</td>
<td></td>
<td>x*</td>
<td>x*</td>
<td></td>
<td></td>
<td>x</td>
<td>R</td>
<td>M</td>
</tr>
</tbody>
</table>

*Query: “IDF in computer-based information look-up”*
Choosing basis for VS model

• A concept space is preferred
  – Semantic gap will be bridged
How to build such a space

• Automatic term expansion
  – Construction of thesaurus
    • WordNet
  – Clustering of words

• Word sense disambiguation
  – Dictionary-based
    • Relation between a pair of words should be similar as in text and dictionary’s description
  – Explore word usage context
How to build such a space

• Latent Semantic Analysis
  – Assumption: there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval
  – It means: the observed term-document association data is contaminated by random noise
How to build such a space

• Solution
  – Low rank matrix approximation

Imagine this is *true* concept-document matrix
Imagine this is our observed term-document matrix
Random noise over the word selection in each document
Latent Semantic Analysis (LSA)

• Low rank approximation of term-document matrix $C_{M \times N}$
  
  – Goal: remove noise in the observed term-document association data
  
  – Solution: find a matrix with rank $k$ which is closest to the original matrix in terms of Frobenius norm

$$\hat{Z} = \arg\min_{Z | \text{rank}(Z) = k} \| C - Z \|_F$$

$$= \arg\min_{Z | \text{rank}(Z) = k} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (C_{ij} - Z_{ij})^2}$$
Basic concepts in linear algebra

• Symmetric matrix
  – $C = C^T$

• Rank of a matrix
  – Number of linearly independent rows (columns) in a matrix $C_{M \times N}$
  – $\text{rank}(C_{M \times N}) \leq \min(M, N)$
Basic concepts in linear algebra

• Eigen system
  – For a square matrix \( C_{M \times M} \)
  – If \( Cx = \lambda x \), \( x \) is called the right eigenvector of \( C \) and \( \lambda \) is the corresponding eigenvalue

• For a symmetric full-rank matrix \( C_{M \times M} \)
  – We have its eigen-decomposition as
    - \( C = Q\Lambda Q^T \)
    - where the columns of \( Q \) are the orthogonal and normalized eigenvectors of \( C \) and \( \Lambda \) is a diagonal matrix whose entries are the eigenvalues of \( C \)
Basic concepts in linear algebra

• Singular value decomposition (SVD)

\[ C_k = U \Sigma_k V^T \]

– We define \( C_{M \times N}^k = U_{M \times k} \Sigma_{k \times k} V_{N \times k}^T \)

  • where we place \( \Sigma_{ii} \) in a descending order and set \( \Sigma_{ii} = \sqrt{\lambda_i} \) for \( i \leq k \), and \( \Sigma_{ii} = 0 \) for \( i > k \)
Latent Semantic Analysis (LSA)

• Solve LSA by SVD

\[ C_k = U \Sigma_k V^T \]

1. Perform SVD on document-term adjacency matrix
2. Construct \( C_{M \times N}^k \) by only keeping the largest \( k \) singular values in \( \Sigma \) non-zero

Map to a lower dimensional space
\[ C_k = U \Sigma_k V^T \]

- \( D_{M \times M} = C_{M \times N} \times C_{M \times N}^T \)
  - \( D_{ij} \): document-document similarity by counting how many terms co-occur in \( d_i \) and \( d_j \)
  - \( D = (U \Sigma V^T) \times (U \Sigma V^T)^T = U \Sigma^2 U^T \)
    - Eigen-decomposition of document-document similarity matrix
    - \( d_i \)'s new representation is then \( (U \Sigma^{1/2})_i \) in this system (space)
    - In the lower dimensional space, we will only use the first \( k \) elements in \( (U \Sigma^{1/2})_i \) to represent \( d_i \)
  - The same analysis applies to \( T_{N \times N} = C_{M \times N}^T \times C_{M \times N} \)
Geometric interpretation of LSA

- $C^k_{M \times N}(i, j)$ measures the relatedness between $d_i$ and $w_j$ in the $k$-dimensional space.
- Therefore
  - As $C^k_{M \times N} = U_{M \times k} \Sigma_{k \times k} V_{N \times k}^T$.
  - $d_i$ is represented as $\left( U_{M \times k} \Sigma^{2}_{k \times k} \right)_i$.
  - $w_j$ is represented as $\left( V_{N \times k} \Sigma^{2}_{k \times k} \right)_j$. 
Latent Semantic Analysis (LSA)

- Visualization

- Graph theory

- HCI

<table>
<thead>
<tr>
<th>Titles</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>Human machine interface for Lab ABC computer applications</td>
</tr>
<tr>
<td>c2</td>
<td>A survey of user opinion of computer system response time</td>
</tr>
<tr>
<td>c3</td>
<td>The EPS user interface management system</td>
</tr>
<tr>
<td>c4</td>
<td>System and human system engineering testing of EPS</td>
</tr>
<tr>
<td>c5</td>
<td>Relation of user-perceived response time to error measurement</td>
</tr>
<tr>
<td>m1</td>
<td>The generation of random, binary, unordered trees</td>
</tr>
<tr>
<td>m2</td>
<td>The intersection graph of paths in trees</td>
</tr>
<tr>
<td>m3</td>
<td>Graph minors IV: Widths of trees and well-quasi-ordering</td>
</tr>
<tr>
<td>m4</td>
<td>Graph minors: A survey</td>
</tr>
</tbody>
</table>
What are those dimensions in LSA

- Principle component analysis
Latent Semantic Analysis (LSA)

• What we have achieved via LSA
  – Terms/documents that are closely associated are placed near one another in this new space
  – Terms that do not occur in a document may still close to it, if that is consistent with the major patterns of association in the data
  – A good choice of concept space for VS model!
LSA for retrieval

• Project queries into the new document space
  \[ \tilde{q} = qV_{N \times k} \Sigma_{k \times k}^{-1} \]
  • Treat query as a pseudo document of term vector
  • Cosine similarity between query and documents in this lower-dimensional space
LSA for retrieval

q: "human computer interaction"

Graph theory

HC1

Titles

c1: Human machine interface for Lab ABC computer applications

c2: A survey of user opinion of computer system response time

c3: The EPS user interface management system

c4: System and human system engineering testing of EPS

c5: Relation of user-perceived response time to error measurement

m1: The generation of random, binary, unordered trees

m2: The intersection graph of paths in trees

m3: Graph minors IV: Widths of trees and well-quasi-ordering

m4: Graph minors: A survey
Discussions

• Computationally expensive
  – Time complexity $O(MN^2)$

• Empirically helpful for recall but not for precision
  – Recall increases as $k$ decreases

• Optimal choice of $k$

• Difficult to handle dynamic corpus

• Difficult to interpret the decomposition results

We will come back to this later!
LSA beyond text

- Collaborative filtering
  - User item matrix stores for each user the rating for the items

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>\ldots</th>
<th>$i_{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>\ldots</td>
<td>1</td>
</tr>
<tr>
<td>$u_2$</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
<td>5</td>
</tr>
<tr>
<td>$u_3$</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
<td>4</td>
</tr>
<tr>
<td>$u_4$</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>\ldots</td>
<td>2</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>$u_{\ell}$</td>
<td>2</td>
<td>\ldots</td>
<td>4</td>
<td>\ldots</td>
<td>4</td>
<td>\ldots</td>
<td>1</td>
</tr>
</tbody>
</table>

Predicting unknown ratings
LSA beyond text

- Eigen face
LSA beyond text

• Cat from deep neuron network

One of the neurons in the artificial neural network, trained from still frames from unlabeled YouTube videos, learned to detect cats.
What you should know

• Assumption in LSA
• Interpretation of LSA
  – Low rank matrix approximation
  – Eigen-decomposition of co-occurrence matrix for documents and terms
• LSA for IR