PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

Prof. Chris Clifton
7 September 2020
Latent Semantic Indexing
Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

PURDUE UNIVERSITY
Department of Computer Science

Retrieve Concepts, not Terms

- Problem: Query is necessarily an incomplete representation of information needed
  - Terms known to querier
  - Exact information presumably unknown
- Idea: Retrieve similar concepts, not similar terms
- Challenge: What is the space of concepts?
  - How do we map document to concept?
  - How does user specify concept?

6

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

Dual space of terms and documents

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	1	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

PURDUE UNIVERSITY
Department of Computer Science

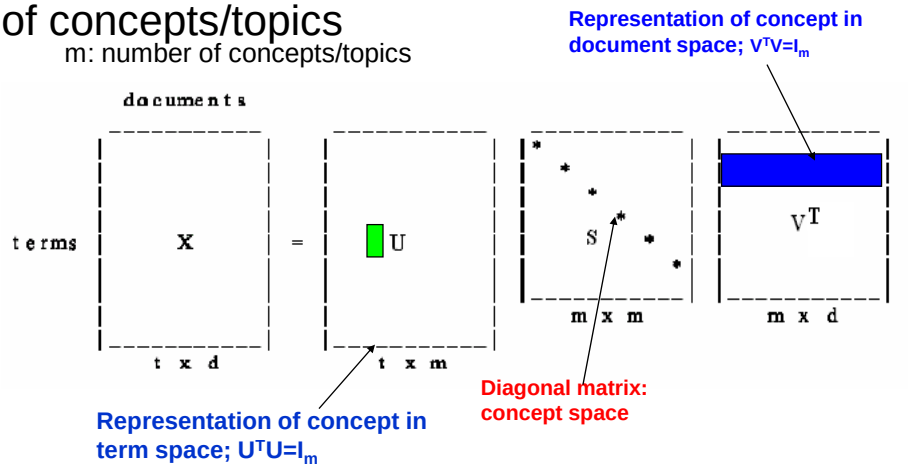
Retrieval Models: Latent Semantic Indexing

- Latent Semantic Indexing (LSI): Explore correlation between terms and documents
  - Two terms are correlated (may share similar semantic concepts) if they often co-occur
  - Two documents are correlated (share similar topics) if they have many common words
- Associate each term and document with a small number of semantic concepts/topics

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

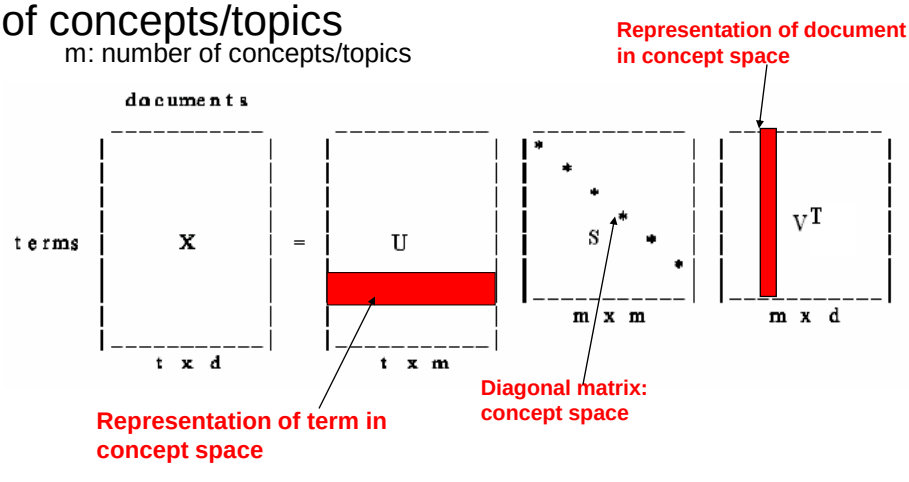
- Use singular value decomposition (SVD) to find a small set of concepts/topics
  m: number of concepts/topics



PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

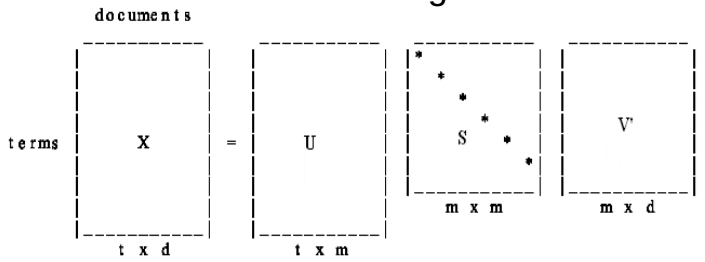
- Use singular value decomposition (SVD) to find a small set of concepts/topics
  m: number of concepts/topics



PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- Properties of Latent Semantic Indexing

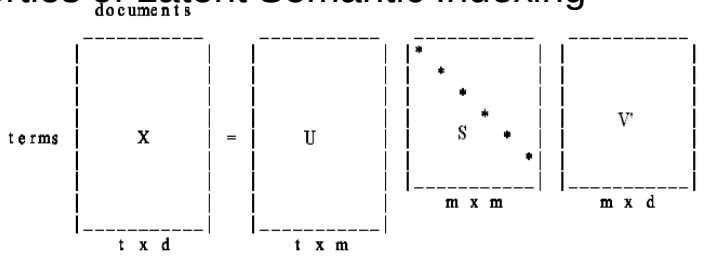


- Diagonal elements of S as Sk in descending order, the larger the more important
- $\hat{x}_k = \sum_{i \le k} u_k S_k v'_k$ is the rank-k matrix that best approximates X, where Uk and V'k are the column vector of U and V'

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

Other properties of Latent Semantic Indexing



- The columns of U are eigenvectors of XX^T
- The columns of V are eigenvectors of X^TX
- The singular values on the diagonal of S, are the positive square roots of the nonzero eigenvalues of both SS^T and S^TS

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

$$\begin{pmatrix} -0.3467 & -0.1369 \\ -0.3467 & -0.1369 \\ -0.6215 & -0.0987 \\ -0.4544 & -0.0327 \\ -0.3329 & -0.0049 \\ -0.0452 & 0.5225 \\ -0.2245 & 0.4859 \\ -0.0452 & 0.5225 \\ -0.0401 & 0.4118 \end{pmatrix} \times \begin{pmatrix} 3.1395 & 0 \\ 0 & 2.3912 \end{pmatrix} \times \begin{pmatrix} -0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\ -0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \end{pmatrix}$$

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

$$\begin{pmatrix} -0.3467 & -0.1369 \\ -0.3467 & -0.1369 \\ -0.6215 & -0.0987 \\ -0.4544 & -0.0327 \\ -0.3329 & -0.0049 \\ -0.0452 & 0.5225 \\ -0.2245 & 0.4859 \\ -0.0452 & 0.5225 \\ -0.0401 & 0.4118 \end{pmatrix} \times \begin{pmatrix} 3.1395 & 0 \\ 0 & 2.3912 \end{pmatrix} \times \begin{pmatrix} -0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\ -0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \end{pmatrix}$$

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

$$\begin{pmatrix} -0.3467 & -0.1369 \\ -0.3467 & -0.1369 \\ -0.6215 & -0.0987 \\ -0.4544 & -0.0327 \\ -0.3329 & -0.0049 \\ -0.0452 & 0.5225 \\ -0.2245 & 0.4859 \\ -0.0452 & 0.5225 \\ -0.0401 & 0.4118 \end{pmatrix} \times \begin{pmatrix} 3.1395 & 0 \\ 0 & 2.3912 \end{pmatrix} \times \begin{pmatrix} -0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\ -0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \end{pmatrix}$$

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

$$\begin{pmatrix} -0.3467 & -0.1369 \\ -0.3467 & -0.1369 \\ -0.6215 & -0.0987 \\ -0.4544 & -0.0327 \\ -0.3329 & -0.0049 \\ -0.0452 & 0.5225 \\ -0.2245 & 0.4859 \\ -0.0452 & 0.5225 \\ -0.0401 & 0.4118 \end{pmatrix} \times \begin{pmatrix} 3.1395 & 0 \\ 0 & 2.3912 \end{pmatrix} \times \begin{pmatrix} -0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\ -0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \end{pmatrix}$$

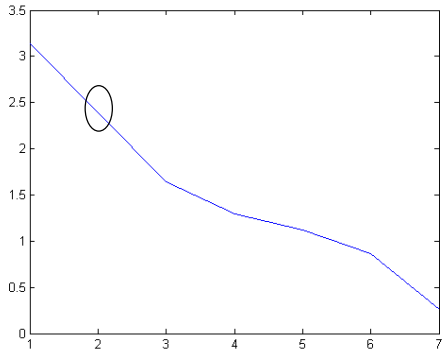
PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- Importance of Concepts

Importance of Concept

Reflects Error of Approximating X with small S



Size of S_k

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- SVD representation
  - Reduce high dimensional representation of document or query into low dimensional concept space
  - SVD tries to preserve the Euclidean distance of document/term vector

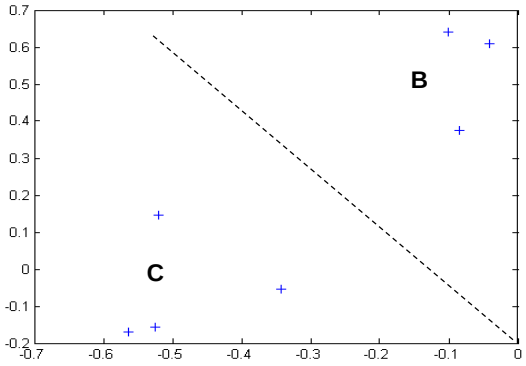
	C1	C2
information	1	1
retrieval	1	1
machine	1	1
learning	0	1
system	1	0
protein	0	0
gene	0	0
mutation	0	0
expression	0	0

	C1	C2
Concept 1	-0.5248	-0.1578
Concept 2	-0.5635	-0.1695

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- SVD Representation



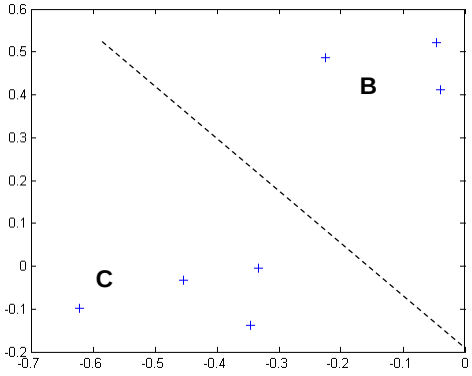
Representation of the documents in two dimensional concept space

19

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- SVD Representation



Representation of the terms in two dimensional concept space

20

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- Retrieval with respect to a query
- Map (fold-in) a query into the representation of the concept space

$$\vec{q}'^{T} = \vec{q}^{T} U_k Inv(S_k)$$

- Use the new representation of the query to calculate the similarity between query and all documents
  - Cosine Similarity

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

Query: Machine Learning Protein

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1

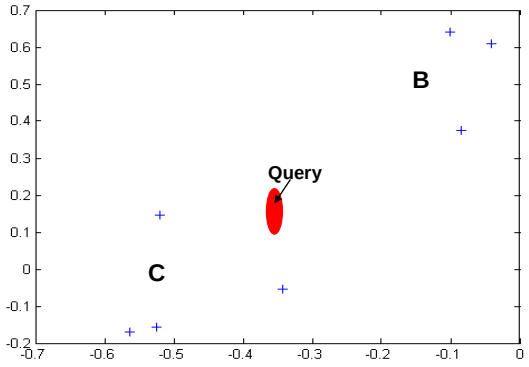
Representation of the query in the term vector space:

$[0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 0]^T$

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

- Representation of the query in the latent semantic space (2 concepts): $\vec{q}'^{T} = \vec{q}^{T} U_k Inv(S_k) = [-0.3571\ 0.1635]^{T}$



PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

Comparison of Retrieval Results in term space and concept space

	C1	C2	C3	C4	B1	B2	B3
information	1	1	0	0	0	0	0
retrieval	1	1	0	0	0	0	0
machine	1	1	1	1	0	0	0
learning	0	1	1	1	0	0	0
system	1	0	1	0	0	0	0
protein	0	0	0	0	0	1	1
gene	0	0	1	0	1	1	0
mutation	0	0	0	0	0	1	1
expression	0	0	0	0	1	0	1
Query Similarity in term space	0.29	0.58	0.58	0.82	0	0.33	0.33
Query Similarity in concept space	0.75	0.75	0.98	0.83	0.61	0.55	0.48

Query: Machine Learning Protein

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Latent Semantic Indexing

Problems with latent semantic indexing

- Difficult to decide the number of concepts
- There is no probabilistic interpretation for the results
- The complexity of the LSI model obtained from SVD is costly

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Outline

- Retrieval Models
- Exact-match retrieval method
  - Unranked Boolean retrieval method
  - Ranked Boolean retrieval method
- Best-match retrieval
  - Vector space retrieval method
  - Latent semantic indexing
  - Probabilistic retrieval models