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?
Retrieval Models: Latent Semantic Indexing

**Dual space of terms and documents**

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>information</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>retrieval</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>machine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>expression</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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
Retrieval Models: Latent Semantic Indexing

• Use singular value decomposition (SVD) to find a small set of concepts/topics

  \( \text{m: number of concepts/topics} \)

  \[
  \begin{align*}
  \text{representation of document in concept space} & = U \\
  \text{representation of term in concept space} & = S \\
  \text{diagonal matrix: concept space} & = \mathbf{V}' \mathbf{V} = \mathbf{I}_m
  \end{align*}
  \]
Retrieval Models: Latent Semantic Indexing

- Properties of Latent Semantic Indexing
  - Diagonal elements of $S$ as $S_k$ in descending order, the larger the more important
  - $\tilde{x}_k = \sum_{i \leq k} u_k s_k v_k'$ is the rank-$k$ matrix that best approximates $X$, where $U_k$ and $V_k'$ are the column vector of $U$ and $V'$

- 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 $S^TS$
### Retrieval Models: Latent Semantic Indexing

#### Table 1

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>info</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ret</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mas</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lea</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sys</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gen</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mut</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>exp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Matrix Multiplication

\[
\begin{pmatrix}
-0.3467 & -0.3467 \\
-0.6215 & -0.0987 \\
-0.4544 & -0.0327 \\
-0.3329 & -0.0049 \\
-0.0452 & 0.5225 \\
-0.2245 & 0.4859 \\
-0.0452 & 0.5225 \\
-0.0401 & 0.4118 \\
\end{pmatrix}
\begin{pmatrix}
3.1395 \\
0 \\
2.3912 \\
\end{pmatrix}
\times
\begin{pmatrix}
-0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.6415 \\
-0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \\
\end{pmatrix}
\]

---

### Retrieval Models: Latent Semantic Indexing

#### Table 1

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>info</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ret</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mas</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lea</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sys</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gen</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mut</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>exp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Matrix Multiplication

\[
\begin{pmatrix}
-0.3467 & -0.3467 \\
-0.6215 & -0.0987 \\
-0.4544 & -0.0327 \\
-0.3329 & -0.0049 \\
-0.0452 & 0.5225 \\
-0.2245 & 0.4859 \\
-0.0452 & 0.5225 \\
-0.0401 & 0.4118 \\
\end{pmatrix}
\begin{pmatrix}
3.1395 \\
0 \\
2.3912 \\
\end{pmatrix}
\times
\begin{pmatrix}
-0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.6415 \\
-0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \\
\end{pmatrix}
\]

---

© 2019 Christopher W. Clifton
Retrieval Models: Latent Semantic Indexing

\[
\begin{pmatrix}
-0.3467 & -0.1369 \\
-0.3467 & -0.1369 \\
-0.6215 & -0.0987 \\
-0.4544 & -0.0327 \\
-0.3329 & -0.0049 \\
-0.0452 & 0.5225 \\
-0.2245 & 0.4859 \\
-0.0452 & 0.5225 \\
-0.0401 & 0.4118 \\
\end{pmatrix}
\begin{pmatrix}
3.1395 \\
0 \\
2.3012 \\
\end{pmatrix}
\begin{pmatrix}
0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\
-0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \\
\end{pmatrix}
\]

Retrieval Models: Latent Semantic Indexing

\[
\begin{pmatrix}
-0.3467 & -0.1369 \\
-0.3467 & -0.1369 \\
-0.6215 & -0.0987 \\
-0.4544 & -0.0327 \\
-0.3329 & -0.0049 \\
-0.0452 & 0.5225 \\
-0.2245 & 0.4859 \\
-0.0452 & 0.5225 \\
-0.0401 & 0.4118 \\
\end{pmatrix}
\begin{pmatrix}
3.1395 \\
0 \\
2.3012 \\
\end{pmatrix}
\begin{pmatrix}
0.5248 & -0.5635 & -0.5202 & -0.3427 & -0.0843 & -0.1003 & -0.0415 \\
-0.1578 & -0.1695 & 0.1462 & -0.0550 & 0.3754 & 0.6402 & 0.6092 \\
\end{pmatrix}
\]
Retrieval Models: 
Latent Semantic Indexing

• Importance of Concepts

\[ \text{Importance of Concept} \]
\[ \text{Reflects Error of Approximating } X \text{ with small } S \]

Size of \( S_k \)

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

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>info</td>
<td>1.0</td>
</tr>
<tr>
<td>retrieval</td>
<td>1.0</td>
</tr>
<tr>
<td>machine</td>
<td>1.0</td>
</tr>
<tr>
<td>learning</td>
<td>0.1</td>
</tr>
<tr>
<td>system</td>
<td>0.0</td>
</tr>
<tr>
<td>protein</td>
<td>0.0</td>
</tr>
<tr>
<td>gene</td>
<td>0.0</td>
</tr>
<tr>
<td>mutation</td>
<td>0.0</td>
</tr>
<tr>
<td>expression</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept 1</td>
<td>-0.5248</td>
</tr>
<tr>
<td>Concept 2</td>
<td>-0.5635</td>
</tr>
</tbody>
</table>
Retrieval Models: Latent Semantic Indexing

- SVD Representation

Representation of the documents in two dimensional concept space

Retrieval Models: Latent Semantic Indexing

- SVD Representation

Representation of the terms in two dimensional concept space
Retrieval Models: Latent Semantic Indexing

• Retrieval with respect to a query
• Map (fold-in) a query into the representation of the concept space
  \[ \tilde{q}^T = \tilde{q}^T U_k Inv(S_k) \]
• Use the new representation of the query to calculate the similarity between query and all documents
  – Cosine Similarity

Query: Machine Learning Protein

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>information</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>retrieval</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>machine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>expression</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Representation of the query in the term vector space:

\[ [0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0]^T \]
Retrieval Models: Latent Semantic Indexing

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

Comparison of Retrieval Results in term space and concept space

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>information</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>retrieval</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>machine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>expression</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Query Similarity</td>
<td>0.29</td>
<td>0.58</td>
<td>0.58</td>
<td>0.82</td>
<td>0</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>in term space</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Similarity</td>
<td>0.75</td>
<td>0.75</td>
<td>0.98</td>
<td>0.83</td>
<td>0.61</td>
<td>0.55</td>
<td>0.48</td>
</tr>
<tr>
<td>in concept space</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Query: Machine Learning Protein
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

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