INF 141
IR METRICS
LATENT SEMANTIC ANALYSIS
AND INDEXING

Crista Lopes
Outline

- Precision and Recall
- The problem with indexing so far
- Intuition for solving it
- Overview of the solution
- The Math
How to measure

- Given the enormous variety of possible retrieval schemes, how do we measure how good they are?
Standard IR Metrics

- **Recall**: portion of the relevant documents that the system retrieved (blue arrow points in the direction of higher recall)

- **Precision**: portion of retrieved documents that are relevant (yellow arrow points in the direction of higher precision)
Definitions

\[ \text{precision} = \frac{| \{\text{relevant documents} \} \cap \{\text{retrieved documents} \}|}{| \{\text{retrieved documents} \}|} \]

\[ \text{recall} = \frac{| \{\text{relevant documents} \} \cap \{\text{retrieved documents} \}|}{| \{\text{relevant documents} \}|} \]
Definitions

True positives  
False negatives  
True negatives  
False positives

Precision = \frac{tp}{tp + fp}  
Recall = \frac{tp}{tp + fn}

(same thing, different terminology)
Doc1 = A comparison of the newest models of cars (keyword: car)
Doc2 = Guidelines for automobile manufacturing (keyword: automobile)
Doc3 = The car function in Lisp (keyword: car)
Doc4 = Flora in North America

Query: “automobile”

Retrieval scheme A

\[
\text{Precision} = \frac{tp}{tp + fp} \quad \text{Precision} = \frac{1}{1} = 1
\]

\[
\text{Recall} = \frac{tp}{tp + fn} \quad \text{Recall} = \frac{1}{2} = 0.5
\]
Example

Doc1 = A comparison of the newest models of cars (keyword: car)
Doc2 = Guidelines for automobile manufacturing (keyword: automobile)
Doc3 = The car function in Lisp (keyword: car)
Doc4 = Flora in North America

Query: “automobile”

Retrieval scheme B

\[
\text{Precision} = \frac{tp}{tp + fp} \quad \text{Precision} = \frac{2}{2} = 1
\]
\[
\text{Recall} = \frac{tp}{tp + fn} \quad \text{Recall} = \frac{2}{2} = 1
\]

Perfect!
Example

Doc1 = A comparison of the newest models of cars (keyword: car)
Doc2 = Guidelines for automobile manufacturing (keyword: automobile)
Doc3 = The car function in Lisp (keyword: car)
Doc4 = Flora in North America

Query: “automobile”

Retrieval scheme C

\[
\text{Precision} = \frac{tp}{tp + fp}
\]
\[
\text{Recall} = \frac{tp}{tp + fn}
\]

Precision = 2/3 = 0.67
Recall = 2/2 = 1
Example

- Clearly scheme B is the best of the 3.
- A vs. C: which one is better?
  - Depends on what you are trying to achieve

- Intuitively for people:
  - Low precision leads to low trust in the system — too much noise!
    (e.g. consider precision = 0.1)
  - Low recall leads to unawareness
    (e.g. consider recall = 0.1)
F-measure

- Combines precision and recall into a single number

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

More generally,

\[
F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}
\]

Typical values:
- \( \beta = 2 \)  \( \Rightarrow \) gives more weight to recall
- \( \beta = 0.5 \)  \( \Rightarrow \) gives more weight to precision
F-measure

\[
\text{F (scheme A)} = \frac{2 \times (1 \times 0.5)}{1+0.5} = 0.67
\]

\[
\text{F (scheme B)} = \frac{2 \times (1 \times 1)}{1+1} = 1
\]

\[
\text{F (scheme C)} = \frac{2 \times (0.67 \times 1)}{0.67+1} = 0.8
\]
Test Data

- In order to get these numbers, we need data sets for which we know the relevant and non-relevant documents for test queries
  - Requires human judgment
Outline

- The problem with indexing so far
- Intuition for solving it
- Overview of the solution
- The Math

Part of these notes were adapted from:
http://www.slidefinder.net/I/Introduction_Latent_Semantic_Analysis_Melanie/26158812
Indexing so far

- Given a collection of documents:
  - retrieve documents that are relevant to a given query
- **Match terms in documents to terms in query**
- Vector space method
  - term (rows) by document (columns) matrix, based on occurrence
  - translate into vectors in a vector space
    - one vector for each document + query
  - cosine to measure distance between vectors (documents)
    - small angle $\rightarrow$ large cosine $\rightarrow$ similar
    - large angle $\rightarrow$ small cosine $\rightarrow$ dissimilar
Two problems

- **synonymy**: many ways to refer to the same thing, e.g. car and automobile
  - Term matching leads to poor recall

- **polysemy**: many words have more than one meaning, e.g. model, python, chip
  - Term matching leads to poor precision
Two problems

Synonymy
Will have small cosine
but are related

Polysemy
Will have large cosine
but not truly related
Solutions

- Use dictionaries
  - Fixed set of word relations
  - Generated with years of human labour
  - Top-down solution

- Use latent semantics methods
  - Word relations emerge from the corpus
  - Automatically generated
  - Bottom-up solution
Dictionaries

- WordNet
  - [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/)
  - Library and Web API
Latent Semantic Indexing (LSI)

- First non-dictionary solution to these problems
- developed at Bellcore (now Telcordia) in the late 1980s (1988). It was patented in 1989.
LSI pubs


LSI (Indexing) vs. LSA (Analysis)

- LSI: the use of latent semantic methods to build a more powerful index (for info retrieval)
- LSA: the use latent semantic methods for document/corpus analysis
Basic Goal of LS methods

Given $N \times M$ matrix

<table>
<thead>
<tr>
<th>Term</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>...</th>
<th>$D_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term_1</td>
<td>$\text{tdidf}_{1,1}$</td>
<td>$\text{tdidf}_{1,2}$</td>
<td>$\text{tdidf}_{1,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{1,M}$</td>
</tr>
<tr>
<td>Term_2</td>
<td>$\text{tdidf}_{2,1}$</td>
<td>$\text{tdidf}_{2,2}$</td>
<td>$\text{tdidf}_{2,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{2,M}$</td>
</tr>
<tr>
<td>Term_3</td>
<td>$\text{tdidf}_{3,1}$</td>
<td>$\text{tdidf}_{3,2}$</td>
<td>$\text{tdidf}_{3,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{3,M}$</td>
</tr>
<tr>
<td>Term_4</td>
<td>$\text{tdidf}_{4,1}$</td>
<td>$\text{tdidf}_{4,2}$</td>
<td>$\text{tdidf}_{4,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{4,M}$</td>
</tr>
<tr>
<td>Term_5</td>
<td>$\text{tdidf}_{5,1}$</td>
<td>$\text{tdidf}_{5,2}$</td>
<td>$\text{tdidf}_{5,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{5,M}$</td>
</tr>
<tr>
<td>Term_6</td>
<td>$\text{tdidf}_{6,1}$</td>
<td>$\text{tdidf}_{6,2}$</td>
<td>$\text{tdidf}_{6,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{6,M}$</td>
</tr>
<tr>
<td>Term_7</td>
<td>$\text{tdidf}_{7,1}$</td>
<td>$\text{tdidf}_{7,2}$</td>
<td>$\text{tdidf}_{7,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{7,M}$</td>
</tr>
<tr>
<td>Term_8</td>
<td>$\text{tdidf}_{8,1}$</td>
<td>$\text{tdidf}_{8,2}$</td>
<td>$\text{tdidf}_{8,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{8,M}$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term_N</td>
<td>$\text{tdidf}_{N,1}$</td>
<td>$\text{tdidf}_{N,2}$</td>
<td>$\text{tdidf}_{N,3}$</td>
<td>...</td>
<td>$\text{tdidf}_{N,M}$</td>
</tr>
</tbody>
</table>
Basic Goal of LS methods

<table>
<thead>
<tr>
<th>Concept</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>...</th>
<th>$D_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{1,1}$</td>
<td>$v_{1,2}$</td>
<td>$v_{1,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{1,M}$</td>
</tr>
<tr>
<td>$v_{2,1}$</td>
<td>$v_{2,2}$</td>
<td>$v_{2,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{2,M}$</td>
</tr>
<tr>
<td>$v_{3,1}$</td>
<td>$v_{3,2}$</td>
<td>$v_{3,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{3,M}$</td>
</tr>
<tr>
<td>$v_{4,1}$</td>
<td>$v_{4,2}$</td>
<td>$v_{4,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{4,M}$</td>
</tr>
<tr>
<td>$v_{5,1}$</td>
<td>$v_{5,2}$</td>
<td>$v_{5,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{5,M}$</td>
</tr>
<tr>
<td>$v_{6,1}$</td>
<td>$v_{6,2}$</td>
<td>$v_{6,3}$</td>
<td>...</td>
<td>...</td>
<td>$v_{6,M}$</td>
</tr>
</tbody>
</table>

$K=6$

Squeeze terms such that they reflect **concepts**
Query matching is performed in the concept space too
Dimensionality Reduction: Projection

- Latent Semantic Indexing always reduces the number of dimensions
Dimensionality Reduction: Projection
How can this be achieved?

- Math magic to the rescue
- Specifically, linear algebra
- Specifically, matrix decompositions
- Specifically, Singular Value Decomposition (SVD)
- Followed by dimension reduction
  - Honey, I shrunk the vector space!
Singular Value Decomposition

A = UΣVT

(also A = TSDT)

Dimension Reduction

~A = ~U~ Σ ~VT

\[
\begin{align*}
A_{\text{r}} &= \begin{bmatrix}
U & \Sigma & V^T
\end{bmatrix}
\end{align*}
\]
SVD

- $A = TSD^T$ such that
  - $TT^T = I$
  - $DD^T = I$
  - $S =$ all zeros except diagonal (singular values);
    singular values decrease along diagonal
SVD examples

- [http://people.revoledu.com/kardi/tutorial/LinearAlgebra/SVD.html](http://people.revoledu.com/kardi/tutorial/LinearAlgebra/SVD.html)
- [http://users.telenet.be/paul.larmuseau/SVD.htm](http://users.telenet.be/paul.larmuseau/SVD.htm)
- Many libraries available
Truncated SVD

- SVD is a means to the end goal.
- The end goal is dimension reduction, i.e. get another version of A computed from a reduced space in $TSD^T$
- Simply zero S after a certain row/column $k$
What is \( \sum \) really?

- Remember, diagonal values are in decreasing order

```
64.9  0  0  0  0
 0 29.06 0  0  0
 0  0 18.69 0  0
 0  0  0  4.84 0
 0  0  0  0  0
```

- Singular values represent the strength of latent concepts in the corpus. Each concept emerges from word co-occurrences. (hence the word “latent”)

- By truncating, we are selecting the k strongest concepts
  - Usually in low hundreds

- When forced to squeeze the terms/documents down to a k-dimensional space, the SVD should bring together terms with similar co-occurrences.
SVD in LSI

\[
A_k = U \Sigma V^T
\]

- **Term x Document Matrix**
- **Term x Factor Matrix**
- **Singular Values Matrix**
- **Factor x Document Matrix**

Dimensions:
- \(m \times n\)
- \(m \times r\)
- \(r \times r\)
- \(r \times n\)
Properties of LSI

- The computational cost of SVD is significant. This has been the biggest obstacle to the widespread adoption to LSI.
- As we reduce \( k \), recall tends to increase, as expected.
- Most surprisingly, a value of \( k \) in the low hundreds can actually increase precision on some query benchmarks. This appears to suggest that for a suitable value of \( k \), LSI addresses some of the challenges of synonymy.
- LSI works best in applications where there is little overlap between queries and documents.
Retrieval with LSI

- Query is placed in factor space as a pseudo-document
- Cosine distance to other documents
Retrieval with LSI – Example

**HCI**

- 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

**Graph theory**

- 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
Example – Term-Document Matrix

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
<td>0</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>survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
SVD in LSI

\[ A_k = U \Sigma V^T \]

- Term x Document Matrix
- Term x Factor Matrix
- Singular Values Matrix
- Factor x Document Matrix

Dimensions:
- \( m \times n \)
- \( m \times r \)
- \( r \times r \)
- \( r \times n \)
Online calculator

http://www.bluebit.gr/matrix-calculator/

Note from here on:
The result of this calculator does not match exactly the numbers shown in the LSI paper...
Some -/+ are the opposite...
But bottom line is not affected.
**SVD – The T (term) Matrix**

<table>
<thead>
<tr>
<th>Term</th>
<th>T Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>-0.221 -0.113 0.289 -0.415 -0.106 -0.341 -0.523 0.060 0.407</td>
</tr>
<tr>
<td>interface</td>
<td>-0.198 -0.072 0.135 -0.552 0.282 0.496 0.070 0.010 0.109</td>
</tr>
<tr>
<td>computer</td>
<td>-0.240 0.043 -0.164 -0.595 -0.107 -0.255 0.302 -0.062 -0.492</td>
</tr>
<tr>
<td>user</td>
<td>-0.404 0.057 -0.338 0.099 0.332 0.385 -0.003 0.000 -0.012</td>
</tr>
<tr>
<td>system</td>
<td>-0.644 -0.167 0.361 0.333 -0.159 -0.207 0.166 -0.034 -0.271</td>
</tr>
<tr>
<td>response</td>
<td>-0.265 0.107 -0.426 0.074 0.080 -0.170 -0.283 0.016 0.054</td>
</tr>
<tr>
<td>time</td>
<td>-0.265 0.107 -0.426 0.074 0.080 -0.170 -0.283 0.016 0.054</td>
</tr>
<tr>
<td>EPS</td>
<td>-0.301 -0.141 0.330 0.188 0.115 0.272 -0.033 0.019 0.165</td>
</tr>
<tr>
<td>survey</td>
<td>-0.206 0.274 -0.178 -0.032 -0.537 0.081 0.467 0.036 0.579</td>
</tr>
<tr>
<td>trees</td>
<td>-0.013 0.490 0.231 0.025 0.594 -0.392 0.288 -0.255 0.225</td>
</tr>
<tr>
<td>graph</td>
<td>-0.036 0.623 0.223 0.001 -0.068 0.115 -0.160 0.681 -0.232</td>
</tr>
<tr>
<td>minors</td>
<td>-0.032 0.451 0.141 -0.009 -0.300 0.277 -0.339 -0.678 -0.183</td>
</tr>
</tbody>
</table>

\[ A = \mathbf{TS}_D^T \]
SVD – Singular Values Matrix

\[
\begin{bmatrix}
3.341 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 2.542 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 2.354 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 1.645 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 1.505 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.306 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.846 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.560 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.364
\end{bmatrix}
\]

\[
A = TSD^T
\]
### SVD – The $D^T$ (document) Matrix

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.197</td>
<td>-0.606</td>
<td>-0.463</td>
<td>-0.542</td>
<td>-0.279</td>
<td>-0.004</td>
<td>-0.015</td>
<td>-0.024</td>
<td>-0.082</td>
</tr>
<tr>
<td>-0.056</td>
<td>0.166</td>
<td>-0.127</td>
<td>-0.232</td>
<td>0.107</td>
<td>0.193</td>
<td>0.438</td>
<td>0.615</td>
<td>0.530</td>
</tr>
<tr>
<td>0.110</td>
<td>-0.497</td>
<td>0.208</td>
<td>0.570</td>
<td>-0.505</td>
<td>0.098</td>
<td>0.193</td>
<td>0.253</td>
<td>0.079</td>
</tr>
<tr>
<td>-0.950</td>
<td>-0.029</td>
<td>0.042</td>
<td>0.268</td>
<td>0.150</td>
<td>0.015</td>
<td>0.016</td>
<td>0.010</td>
<td>-0.025</td>
</tr>
<tr>
<td>0.046</td>
<td>-0.206</td>
<td>0.378</td>
<td>-0.206</td>
<td>0.327</td>
<td>0.395</td>
<td>0.349</td>
<td>0.150</td>
<td>-0.602</td>
</tr>
<tr>
<td>-0.077</td>
<td>-0.256</td>
<td>0.724</td>
<td>-0.369</td>
<td>0.035</td>
<td>-0.300</td>
<td>-0.212</td>
<td>0.000</td>
<td>0.362</td>
</tr>
<tr>
<td>-0.177</td>
<td>0.433</td>
<td>0.237</td>
<td>-0.265</td>
<td>-0.672</td>
<td>0.341</td>
<td>0.152</td>
<td>-0.249</td>
<td>-0.038</td>
</tr>
<tr>
<td>0.014</td>
<td>-0.049</td>
<td>-0.009</td>
<td>0.019</td>
<td>0.058</td>
<td>-0.454</td>
<td>0.762</td>
<td>-0.450</td>
<td>0.070</td>
</tr>
<tr>
<td>0.064</td>
<td>-0.243</td>
<td>-0.024</td>
<td>0.084</td>
<td>0.262</td>
<td>0.620</td>
<td>-0.018</td>
<td>-0.520</td>
<td>0.454</td>
</tr>
</tbody>
</table>

$$A = T S D^T$$
### SVD with $k=2$

**$T S_2 =$**

<table>
<thead>
<tr>
<th>term</th>
<th>$-0.738$</th>
<th>$-0.287$</th>
<th>$0.000$</th>
<th>$0.000$</th>
<th>$0.000$</th>
<th>$0.000$</th>
<th>$0.000$</th>
<th>$0.000$</th>
<th>$0.000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interface</td>
<td>$-0.662$</td>
<td>$-0.183$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>computer</td>
<td>$-0.802$</td>
<td>$0.109$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>user</td>
<td>$-1.350$</td>
<td>$0.145$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>system</td>
<td>$-2.152$</td>
<td>$-0.425$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>response</td>
<td>$-0.885$</td>
<td>$0.272$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>time</td>
<td>$-0.885$</td>
<td>$0.272$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>EPS</td>
<td>$-1.006$</td>
<td>$-0.358$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>survey</td>
<td>$-0.688$</td>
<td>$0.697$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>trees</td>
<td>$-0.043$</td>
<td>$1.246$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>graph</td>
<td>$-0.120$</td>
<td>$1.584$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>minors</td>
<td>$-0.107$</td>
<td>$1.146$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
</tbody>
</table>

*Distribution of topics over terms*
SVD -- A
t

\[ S_2 D^T = \]

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.658</td>
<td>-2.025</td>
<td>-1.547</td>
<td>-1.811</td>
<td>-0.932</td>
<td>-0.013</td>
<td>-0.050</td>
<td>-0.080</td>
<td>-0.274</td>
</tr>
<tr>
<td>-0.142</td>
<td>0.422</td>
<td>-0.323</td>
<td>-0.590</td>
<td>0.272</td>
<td>0.491</td>
<td>1.113</td>
<td>1.563</td>
<td>1.347</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Distribution of topics over documents
Query

“Human computer interaction”

Query is placed at the centroid of the query terms in the concept space
Plot in the first 2 dimensions

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>-0.221</td>
<td>-0.113</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interface</td>
<td>-0.198</td>
<td>-0.072</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computer</td>
<td>-0.240</td>
<td>0.043</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>user</td>
<td>-0.404</td>
<td>0.057</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>system</td>
<td>-0.644</td>
<td>-0.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>response</td>
<td>-0.265</td>
<td>0.107</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>-0.265</td>
<td>0.107</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPS</td>
<td>-0.301</td>
<td>-0.141</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>survey</td>
<td>-0.206</td>
<td>0.274</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trees</td>
<td>-0.013</td>
<td>0.490</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graph</td>
<td>-0.036</td>
<td>0.623</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minors</td>
<td>-0.032</td>
<td>0.451</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ T_{1,2} \]

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>-0.197</td>
<td>-0.606</td>
<td>-0.463</td>
<td>-0.542</td>
<td>-0.279</td>
<td>-0.004</td>
<td>-0.015</td>
<td>-0.024</td>
<td>-0.082</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>-0.056</td>
<td>0.166</td>
<td>-0.127</td>
<td>-0.232</td>
<td>0.107</td>
<td>0.193</td>
<td>0.438</td>
<td>0.615</td>
<td>0.530</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ D^T_{1,2} \]

2d plot just for illustration purposes.
In practice it’s still a 9d space
FIG. 1. A two-dimensional plot of 12 Terms and 9 Documents from the sample TM set. Terms are represented by filled circles. Documents are shown as open squares, and component terms are indicated parenthetically. The query ("human computer interaction") is represented as a pseudo-document at (q,13) based on a weighted Term-Term comparison. The dotted area represents the region whose points are within a cosine of 

\[ \text{cosine} = \frac{a \cdot b}{\|a\| \|b\|} \]  

where \( a \) and \( b \) are vectors representing the documents and terms respectively.
Centroid (geometric center)

\[ C = \frac{x_1 + x_2 + \ldots + x_n}{N} \]

\[ C_q = [-0.23 \ 0.035] \]