

Probabilistic Latent Semantic Analysis

Thomas Hofmann

Presentation by

Ioannis Pavlopoulos & Andreas Damianou

for the course of

Data Mining & Exploration

Outline

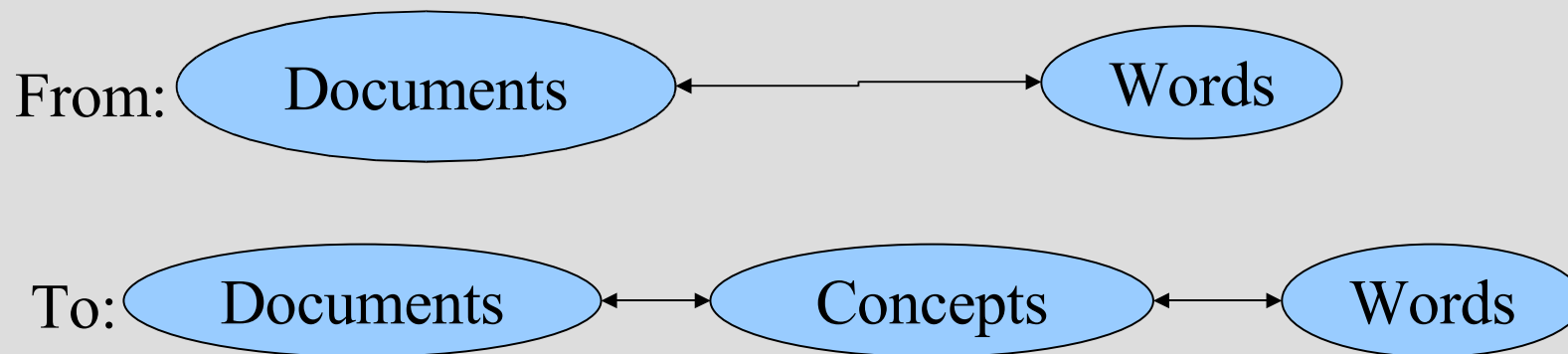
- **Latent Semantic Analysis**
 - Need
 - Overview
 - Drawbacks
- **Probabilistic Latent Semantic Analysis**
 - Solution to drawbacks of LSA
 - Comparison with LSA and document clustering
 - Model Construction
- **Evaluation of PLSA**

Need for Latent Semantic Analysis

- Applications
 - Compare documents in the semantic (concept) space
 - Relations between terms
 - Compare documents across languages
 - *Given*: Bag of words → *Find*: matching documents in the semantic space
- Problems addressing
 - Synonymy
ex: buy - purchase
 - Polysemy
ex: book (verb) - book (noun)

LSA Overview

- Capturing the meaning among words
- Addressing polysemy and synonymy
- Key Idea
 - Dimensionality reduction of word-document co-occurrence matrix
 - Construction of Latent Semantic space

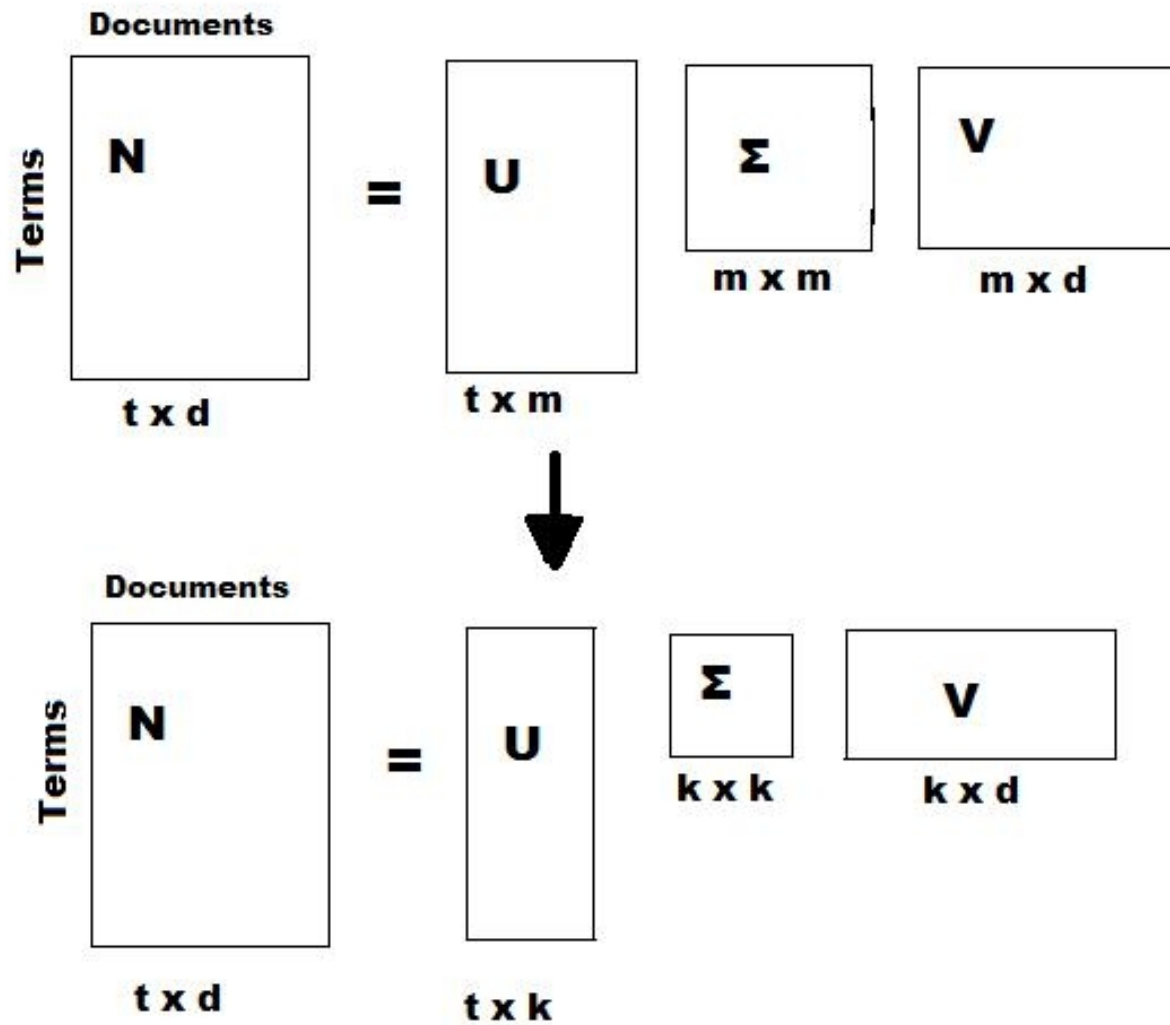


LSA may classify documents together even if they don't have common words!

LSA Concept

- Singular Value Decomposition (SVD)
- Given N which is the word-document co-occurrence matrix, compute:
- $N = U\Sigma V^t$ where:
 - Σ is the diagonal matrix with the singular values of N
 - U, V two orthogonal matrices

LSA SVD



LSA Concept

- Dimensionality Reduction
 - Keep the K – largest singular values which show the dimensions with the greatest variance between words and documents
 - Discarding the lowest dimensions is supposed to be equivalent to reducing the "noise"
 - Terms and documents are converted to points in a K -Dimensional latent space
- Results **do not introduce well defined probabilities** and thus, are difficult to interpret

Probabilistic LSA

Overview

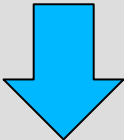
- Implemented to address:
 - Automated Document Indexing
- Same concept to LSA
 - Dimensionality Reduction
 - Construction of a latent space

BUT.....

- Sound Statistical foundations
 - Well defined probabilities
 - Explicable results

Probabilistic LSA

Aspect Model

- Generative model based on the Aspect model
 - Latent variables \mathbf{z} are introduced and relate to documents \mathbf{d} .
 - $|\mathbf{z}| \ll |\mathbf{d}|$, as the same \mathbf{z}_i may be associated with more than one documents
- 
- \mathbf{z} performs as a bottleneck and results in dimensionality reduction

Probabilistic LSA Model

$$P(d, w) = P(d)P(w|d), \text{ where}$$

$$P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z)P(z|d) .$$

Multinomial
Mixtures

Multinomials

Mixing
weights

- Joint probability shows the probability of a word w to be inside a document d
- *Word distributions* are combinations of the factors $P(w|z)$ and the mixing weights $P(z|d)$

Probabilistic LSA Model

- Conditional Independence assumption
 - Documents and Words are independent given z
- Thus, equivalently:

$$P(d, w) = \sum_z P(z)P(d|z)P(w|z)$$



Probabilistic LSA

Model fitting

- Expectation Maximization
- Standard procedure for latent variable models
- E-step: Compute the posteriors for the latent variables \mathbf{z}

$$P(\mathbf{z}|d, w) = \frac{P(\mathbf{z})P(d|\mathbf{z})P(w|\mathbf{z})}{\sum_{\mathbf{z}' \in \mathcal{Z}} P(\mathbf{z}')P(d|\mathbf{z}')P(w|\mathbf{z}')}$$

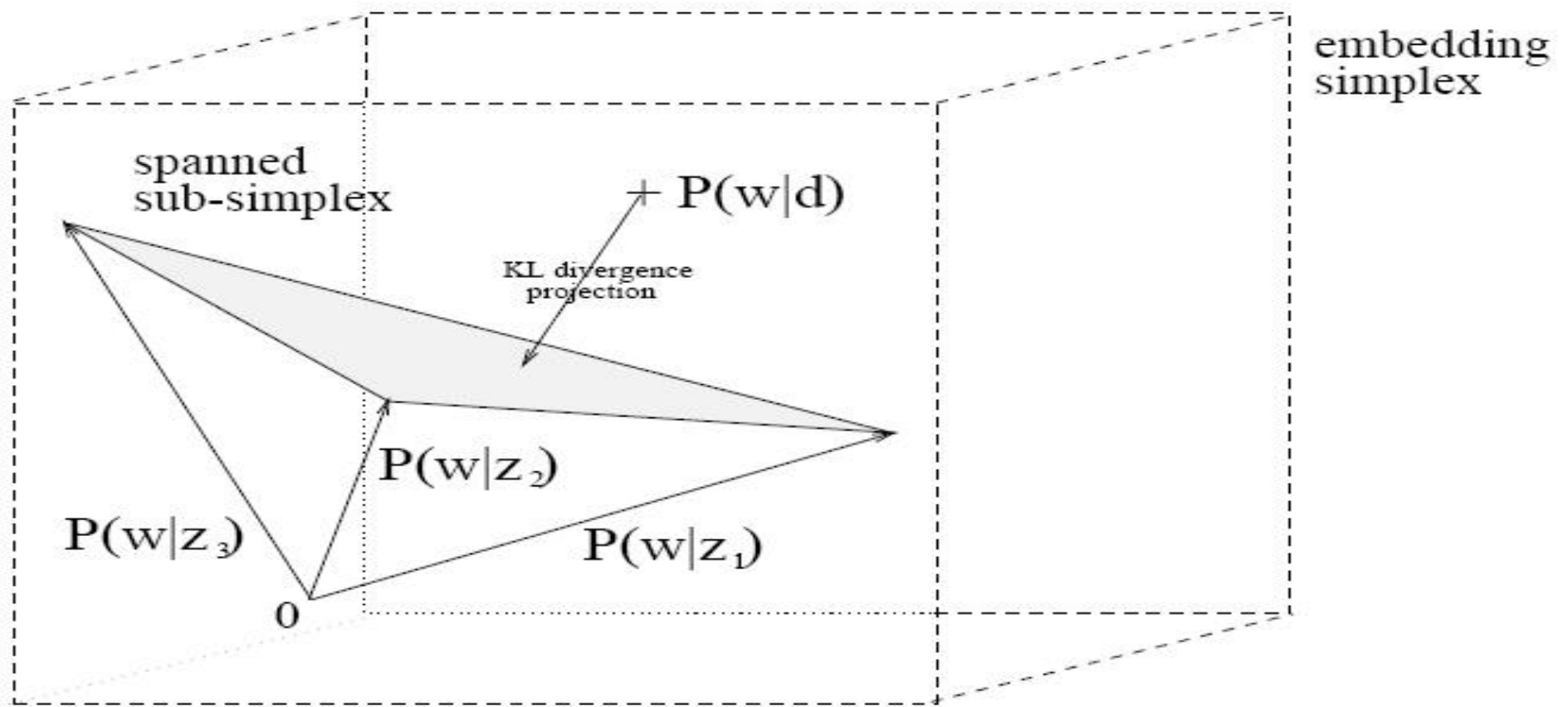
- M-step: Update the parameters

$$P(w|\mathbf{z}) \propto \sum_{d \in \mathcal{D}} n(d, w)P(\mathbf{z}|d, w),$$

$$P(d|\mathbf{z}) \propto \sum_{w \in \mathcal{W}} n(d, w)P(\mathbf{z}|d, w),$$

$$P(\mathbf{z}) \propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d, w)P(\mathbf{z}|d, w)$$

Probabilistic LSA *Space*



Sub-simplex dimensionality $\leq K-1 \ll D-1$

Tempered EM

- Avoid overfitting training data
- Introduce a regularization term β

Tempered EM - Concept

- Add a term $\beta < 1$ in the E step.

$$P_{\beta}(z|d, w) = \frac{P(z) [P(d|z)P(w|z)]^{\beta}}{\sum_{z'} P(z') [P(d|z')P(w|z')]^{\beta}}.$$

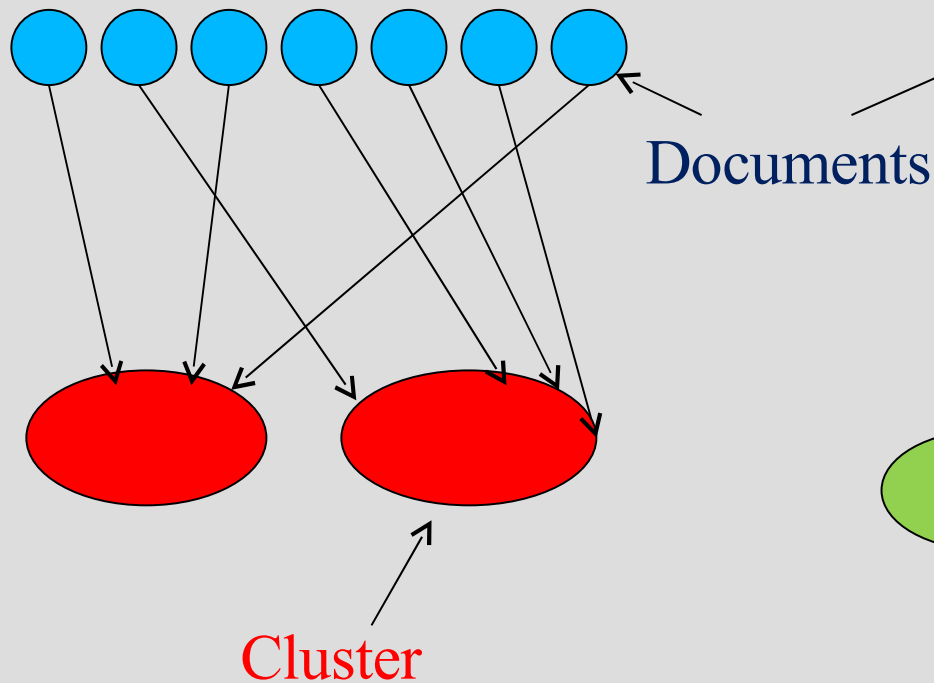
- Used to dampen probabilities in M step.
- Accelerate model fitting procedure compared to other methods (ex. annealing)
- Perform EM iterations and then decrease β until performance on held-out data deteriorates.

PLSA vs LSA

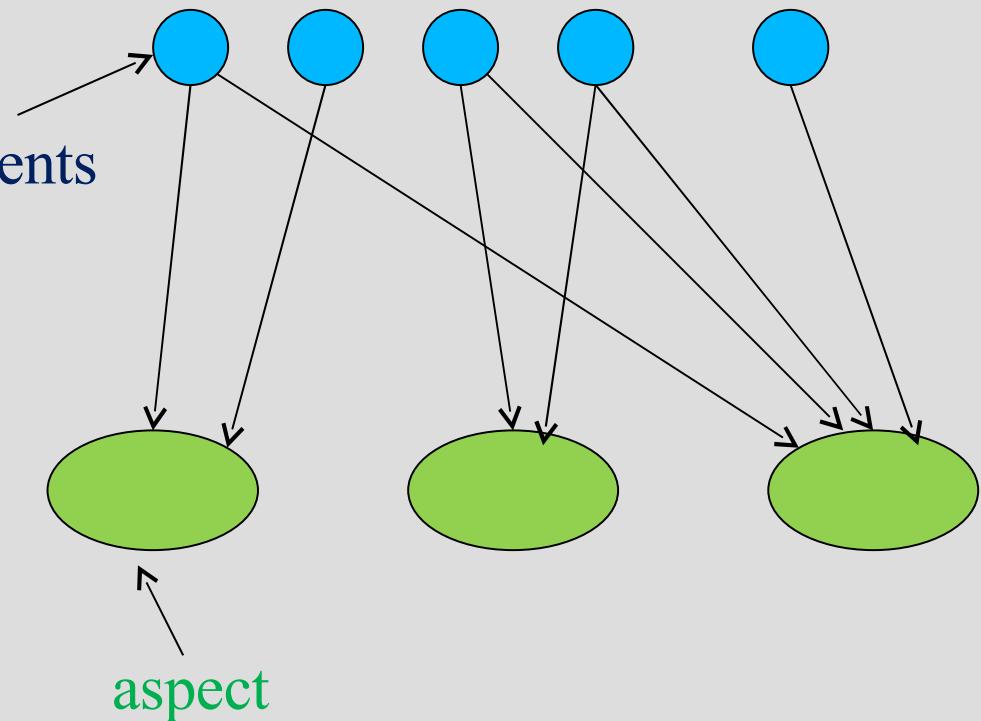
- Great PLSA advantages on the modeling side
 - Well defined probabilities
 - Interpretable directions in the Probabilistic Latent Semantic space as multinomial word distributions
 - Better model selection and complexity control (TEM)
- Important LSA drawbacks in the same side
 - Not defined properly normalized probabilities
 - No obvious interpretations of LS space directions
 - Selection of dimensions based on ad-hoc heuristics
- Potential computational advantage of LSA over PLSA (SVD vs EM which is an iterative method)

Aspect Model vs Clusters

Document Clustering



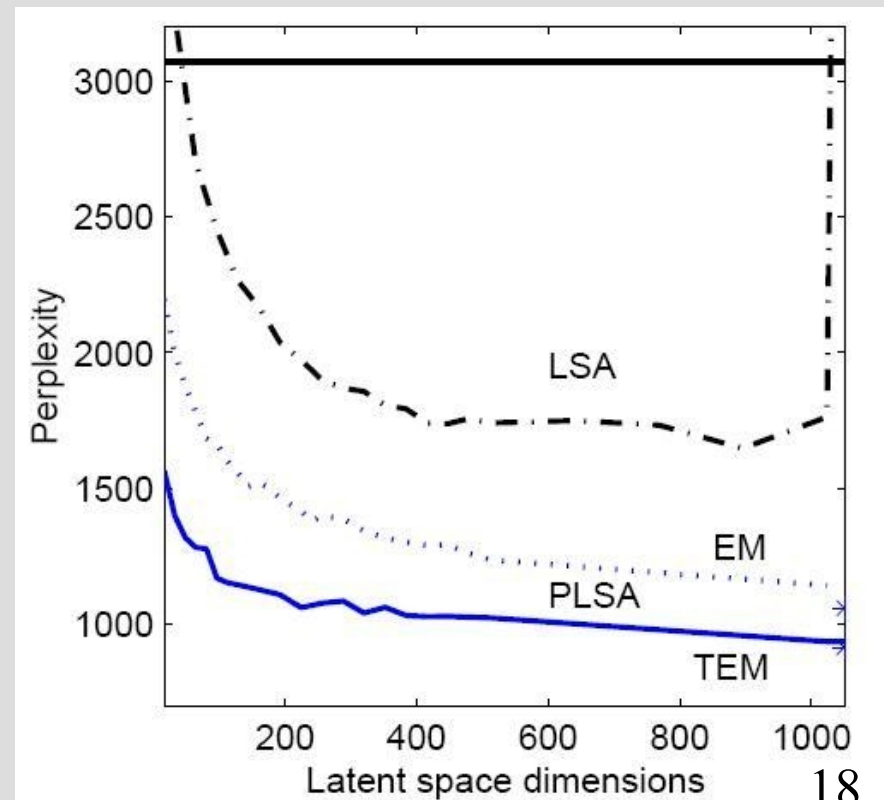
Aspect Model



PLSA: Documents are not related to a single cluster
➔ flexibility, effective modeling

Evaluation *perplexity*

- Perplexity: Measures how “well” a prob. distribution can make predictions.
- Low perplexity → more certain predictions, better model
- PLSA evaluation method:
 - Extract probabilities from LSA
 - Unigram model as baseline
- PLSA evaluation results
 - PLSA better than LSA
 - TEM better than EM
 - PLSA allows $|Z| > \text{rank}(N)$
(N is the co-oc. Matrix)



Evaluation

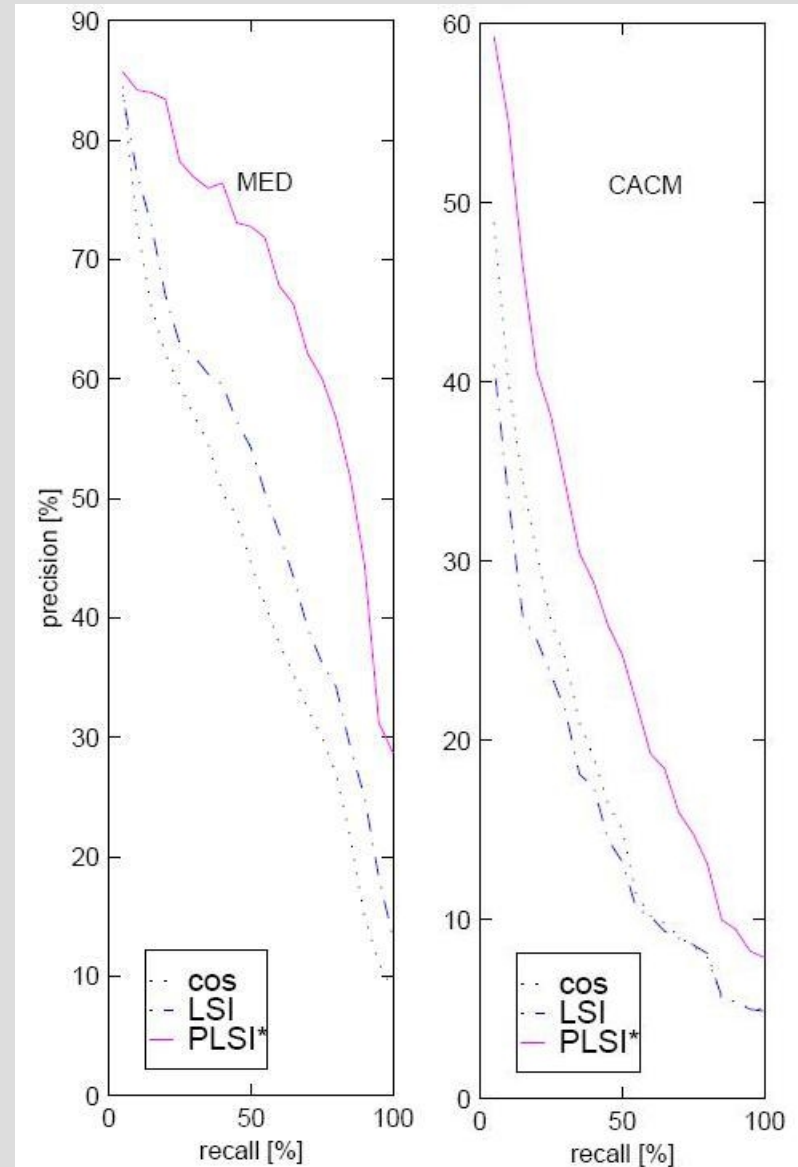
Automatic Indexing

- Given a short document (query q) find the most relevant documents
- Baseline term matching $s(d,q)$: cosine scoring method combined with term frequencies
- LSA: Linear combination of $s(d,q)$ and the one derived from the latent space
- PLSA: Evaluation of similarities of $P(z|d)$ & $P(z|q)$

Evaluation

Precision & Recall

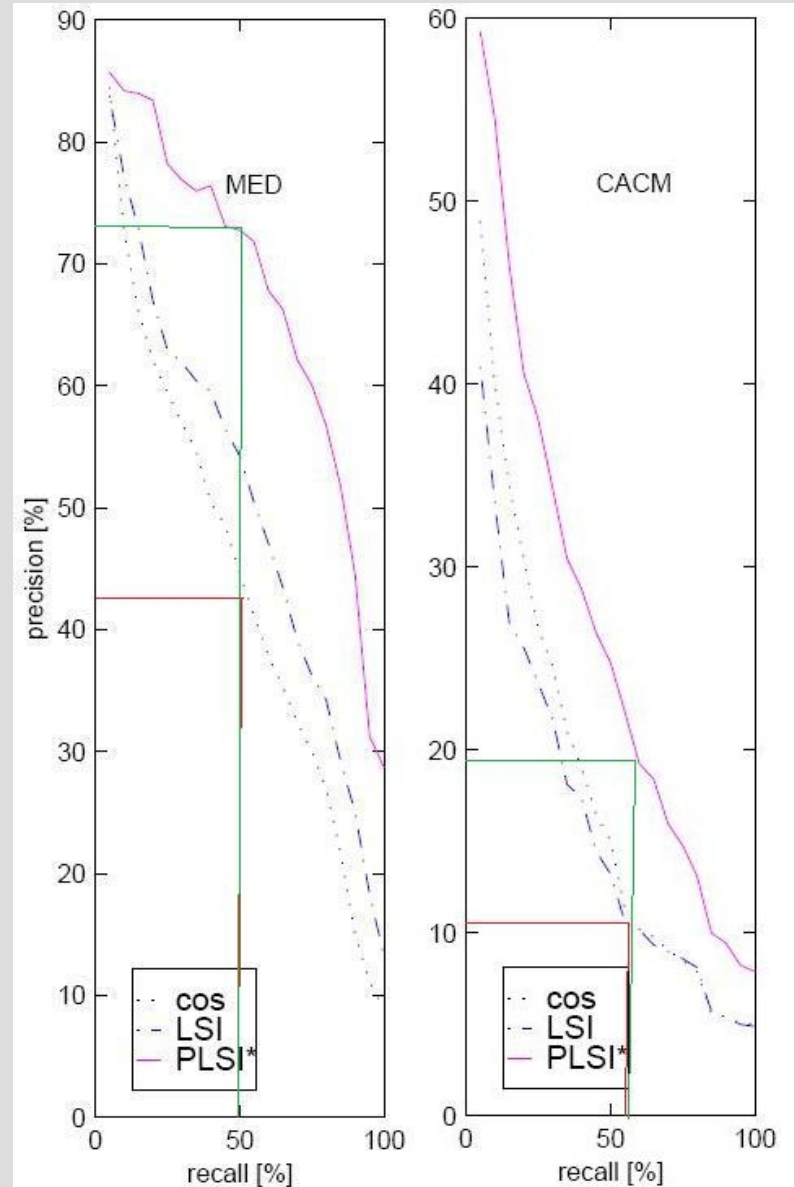
- Precision & Recall:
*Popular measures in
Information Retrieval.*



Evaluation

Precision & Recall

For intermediate values of recall, the precision of PLSA is almost 100% better than the baseline method!!!



Evaluation

Polysemy

- Results show advantage of PLSA over polysemy

“segment 1”	“segment 2”	“matrix 1”	“matrix 2”
imag SEGMENT texture color tissue brain slice cluster mri volume	speaker speech recogni signal train hmm source speakerind. SEGMENT sound	robust MATRIX eigenvalu uncertainti plane linear condition perturb root suffici	manufactur cell part MATRIX cellular famili design machinepart format group

Conclusion

- Documents are represented as vectors of word frequencies
- There is no syntactic relation or word-ordering but co-occurrences still provide useful semantic insights about the document topics
- PLSA is a generative model based on this idea.
- It can be used to extract topics from a collection of documents
- PLSA significantly outperforms LSA thanks to its probabilistic basis.

References

- D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, 2003, pp. 993-1022.
- T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” *Machine Learning*, vol. 42, Jan. 2001, pp. 177-196.
- T. Hofmann, “Probabilistic latent semantic analysis,” *In Proc. of Uncertainty in Artificial Intelligence, UAI’99*, 1999, pp. 289--296.
- DEERWESTER, S., DUMAIS, S., LANDAUER, T., FURNAS, G., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Info. Sci.* 41, 391-407.