Probabilistic Latent Semantic Analysis

Hofmann (1999)

Presenter: Mercè Vintró Ricart

February 8, 2016

Outline

Background

Topic models: What are they? Why do we use them?

Latent Semantic Analysis (LSA)

Methodology

The Aspect Model

Training the model: EM Algorithm.

Evaluation

Perplexity

Information Retrieval

Topic models

Background

- What is a topic?

  The **subject matter** of a text. It captures what it is about.

- Why do we want to extract topics?

  Important for many **text mining tasks**: search result organization, document clustering, passage segmentation, etc.

- How do we do that?

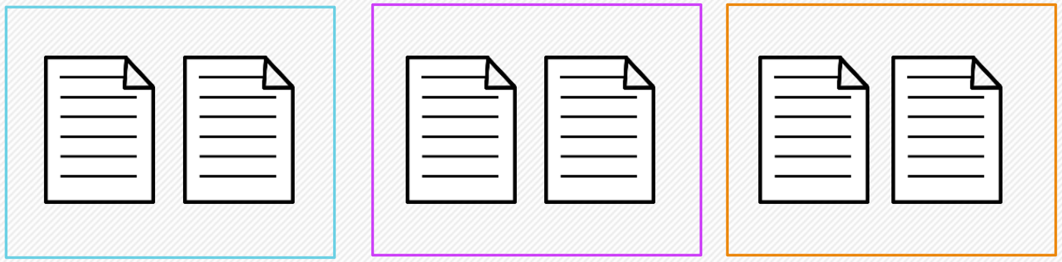
  Use **topic models** to discover hidden topic-based patterns.

Topic models

Background

Text

Politics Sport Technology



Images

Dogs

Wolves

Latent Semantic Analysis (LSA)

- Technique for extracting and representing the **contextual-usage meaning of words**.

- Mapping from high-dimensional count vectors to a lower dimensional representation:

    1. Write frequencies as a **term-document matrix**
    2. Perform **Singular Value Decomposition** (SVD) of the matrix

Latent Semantic Analysis (LSA)

1. Term-document matrix

Doc 1: I have a fluffy cat.
Doc 2: I see a fluffy dog.

	I	have	a	fluffy	cat	see	dog
Doc 1	1	1	1	1	1	0	0
Doc 2	1	0	1	1	0	1	1

Latent Semantic Analysis (LSA)

Background

2. Singular Value Decomposition (SVD)

$$\mathbf{N} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t \xrightarrow{\text{LSA}} \tilde{\mathbf{N}} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^t$$

$\mathbf{U}$ Orthogonal matrix containing the **left singular vectors**.

$\mathbf{V}$ Orthogonal matrix containing the **right singular vectors**.

$\mathbf{\Sigma}$ Diagonal matrix containing the **square roots of eigenvalues from U or V** in descending order.

$\tilde{\mathbf{N}}$ **LSA approximation** of N.

LSA and topics

- Documents with **similar topical content** tend to be close in the latent semantic space.

- Documents which share no terms with each other directly but which do **share many terms with another one** are similar in the latent semantic space.

From LSA to PLSA

Strengths of LSA

- Fully automatic construction
- Representationally simple

Weaknesses of LSA

- No generative model
- Many ad-hoc parameters
- Polysemous words

PLSA to the rescue!

Probabilistic Latent Semantic Analysis (PLSA)

Aspect model

- Latent variable model
- The data can be expressed in terms of:

documents $d \in \mathcal{D} = \{d_1, \cdots, d_N\}$

words $w \in \mathcal{W} = \{w_1, \cdots, w_M\}$

observed variables

topics $z \in \mathcal{Z} = \{z_1, \cdots, z_K\}$ latent variables

Probabilistic Latent Semantic Analysis (PLSA)

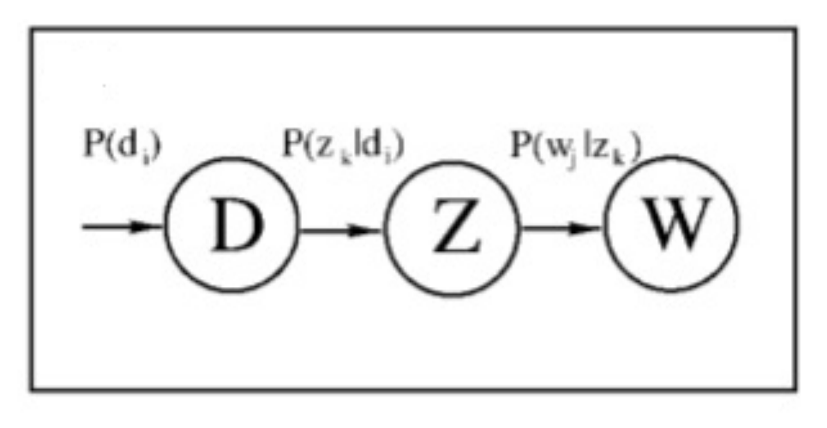
Methodology

Aspect model

- Conditional independence assumption:

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

- Graphical model representation of the aspect model:



Probabilistic Latent Semantic Analysis (PLSA)

Methodology

Aspect model

Product rule

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

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

Conditional independence assumption

$$\Downarrow$$

$$p(d,w) = \underbrace{p(d)}_{\text{Probability of a document}} \sum_{z} \underbrace{p(w|z)}_{\text{Probability of a word given a topic}} \underbrace{p(z|d)}_{\text{Probability of a topic given a document}}$$

Probabilistic Latent Semantic Analysis (PLSA)

Methodology

The EM Algorithm

- E-step

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

The posterior probabilities for the latent variables are computed

- M-step

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

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

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

The parameters are updated

PLSA: Relation to LSA

- The model can be equivalently parameterized by $P(d,w) = \sum_z \underline{P(z)}\,\underline{P(d|z)}\,\underline{P(w|z)}$

- The joint probability P(w,d) can be interpreted as $P = \underline{U}\,\underline{\Sigma}\,\underline{V}^T$

  - U Contains the document probabilities, $\boxed{\text{P(d|z)}}$
  - Σ Diagonal matrix of the prior probabilities of the topics, $\boxed{\text{P(z)}}$
  - V Contains the word probabilities, $\boxed{\text{P(w|z)}}$

PLSA: Polysemy

- The word stems are the 10 most probable words in the distribution P(w|z) in descending order.
- *Segment* is identified as a polysemous word.
  Topic 1: "Image region"
  Topic 2: "Phonetic segment"

Topic 1 **Topic 2**

"segment 1"	"segment 2"
imag	speaker
SEGMENT	speech
texture	recogni
color	signal
tissue	train
brain	hmm
slice	source
cluster	speakerind.
mri	SEGMENT
volume	sound

PLSA: Some limitations

Methodology

- The number of parameters grows linearly with the size of training documents

⇩

The model is **prone to overfitting**

⇩

Tempered EM

- **Not a well-defined** generative model

⇩

Latent Dirichlet Allocation

Perplexity

- Compare the **predictive performance** of PLSA and LSA.

- Perplexity

  - Measure commonly used in language modelling to assess the **generalization performance of a model**.
  - A **lower value** of perplexity indicates better performance.

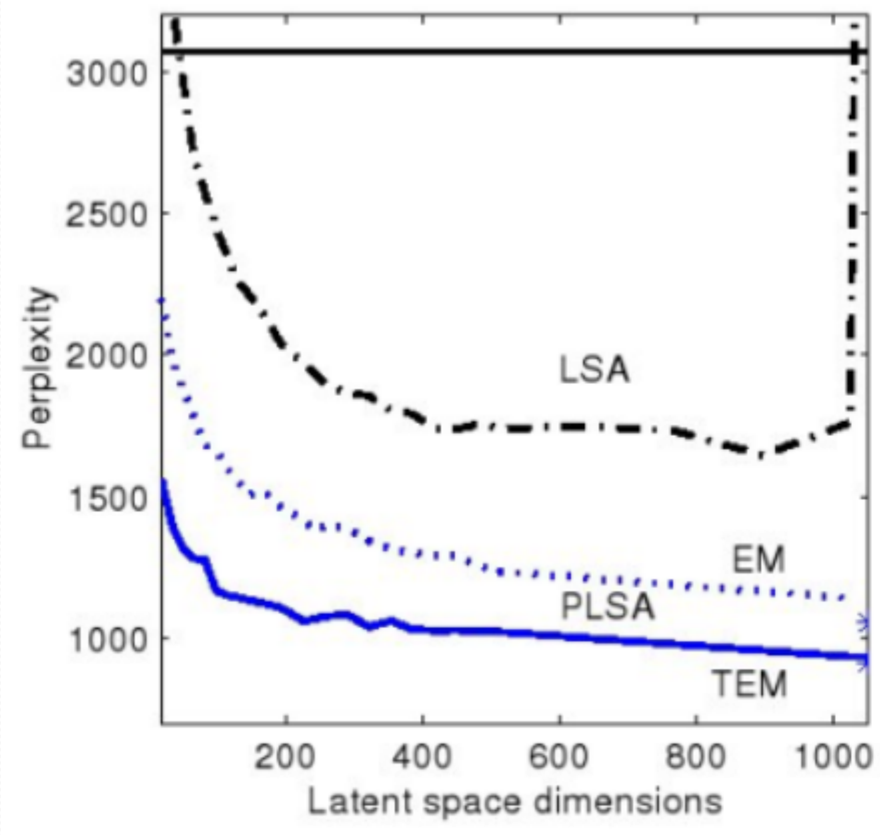
- Two data sets used

  MED: information retrieval test collection with 1033 documents

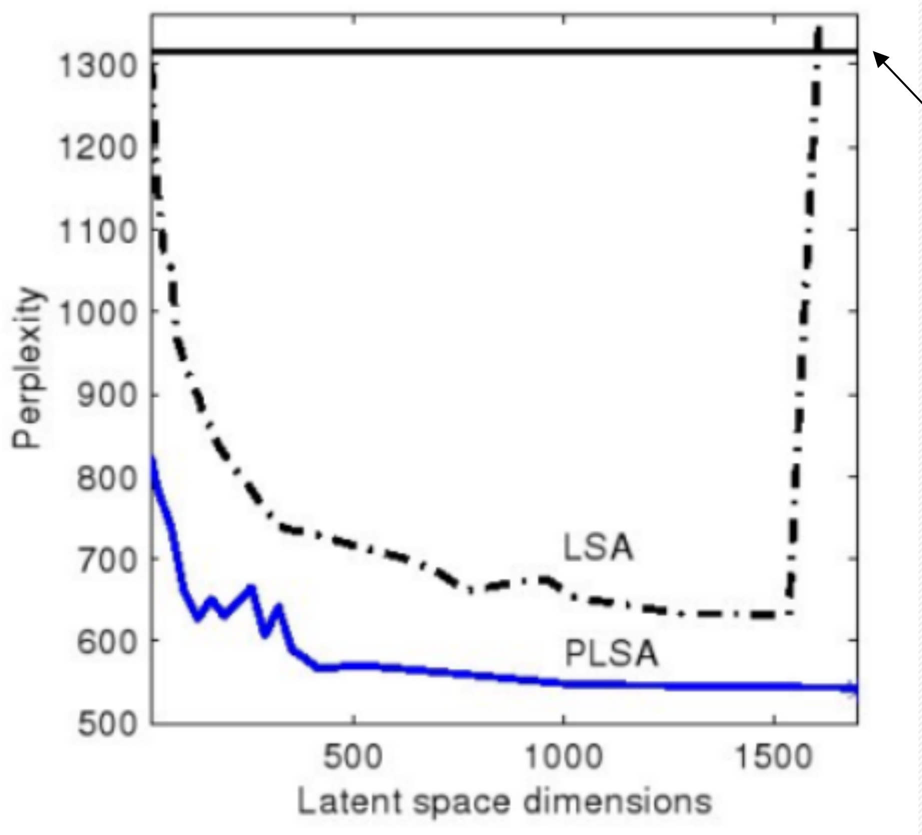
  LOB: dataset with noun-adjective pairs

Perplexity

MED data



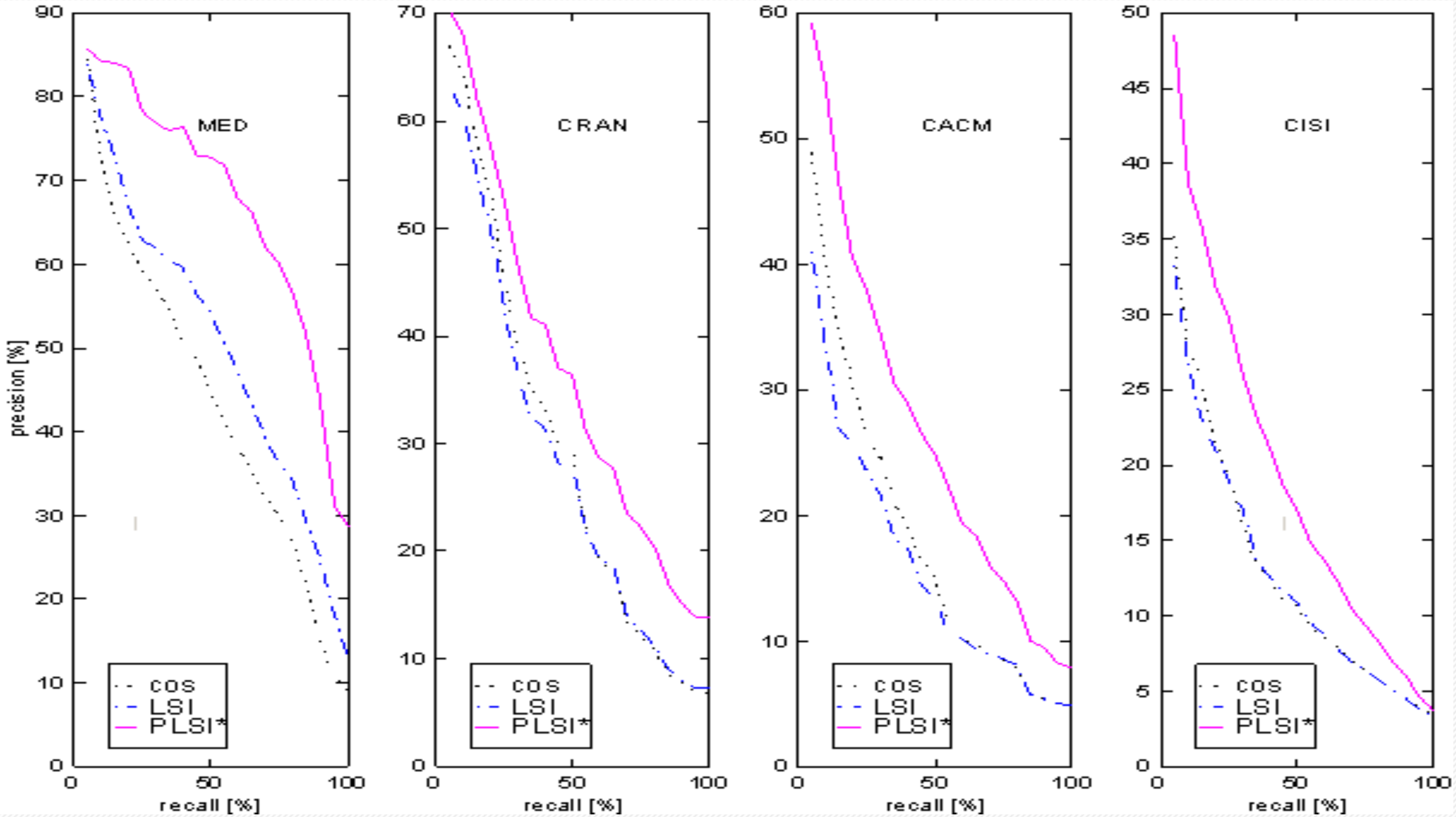
LOB data



Upper
baseline

Information Retrieval

Evaluation



Summary

- LSA can provide useful semantic insights about documents, but it **lacks a sound statistical foundation**.
- PLSA is a **probabilistic variant** of LSA.
- Used to **extract topics** from a collection of documents.
- The model evaluation shows that **PLSA significantly outperforms LSA**.
- Prone to **overfitting** (Tempered EM),
- **Not a well-defined** generative model.

Thank you!
Any questions?