

707.009  
Foundations of Knowledge Management  
„Latent Semantic Analysis“

How can we uncover semantic relationships from natural language text?

**Markus Strohmaier**

Univ. Ass. / Assistant Professor  
Knowledge Management Institute  
Graz University of Technology, Austria

e-mail: [markus.strohmaier@tugraz.at](mailto:markus.strohmaier@tugraz.at)  
web: <http://www.kmi.tugraz.at/staff/markus>

## Slides in part based on

- Slides of Melanie Martin  
“An Introduction to Latent Semantic Analysis”
  
- „An Introduction to Latent Semantic Analysis”  
Thomas K Landauer, Peter W. Foltz, Darrell Laham  
Link: <http://lsa.colorado.edu/dp1.LSAintro.pdf>

# Overview

Today's Agenda:

## **Latent Semantic Analysis**

- **Motivation & Approach**
- **Examples**
- **Evaluation**

# Wissensorganisation – Zwei Herangehensweisen

## Formale vs. inhaltliche Struktur

Viele Informationen liegen in unstrukturierten Freitexten (Inhalt ohne Struktur) vor. Aussagekräftig aber schlecht auswertbar

### Zwei Herangehensweisen:

- Verwendung einer standardisierten Sprache a **priori** (stark formalisiert)
- Interpretation der heterogenen Sprache a **posteriori** (NLP, ...)

Taxonomien,  
Ontologien,  
Semantische  
Netze

Schlüsselwort-  
extraktion,  
Folksonomies



### Examples:

<http://delicious.com/?view=tags>

<http://dir.yahoo.com/>

<http://www.dmoz.org/>

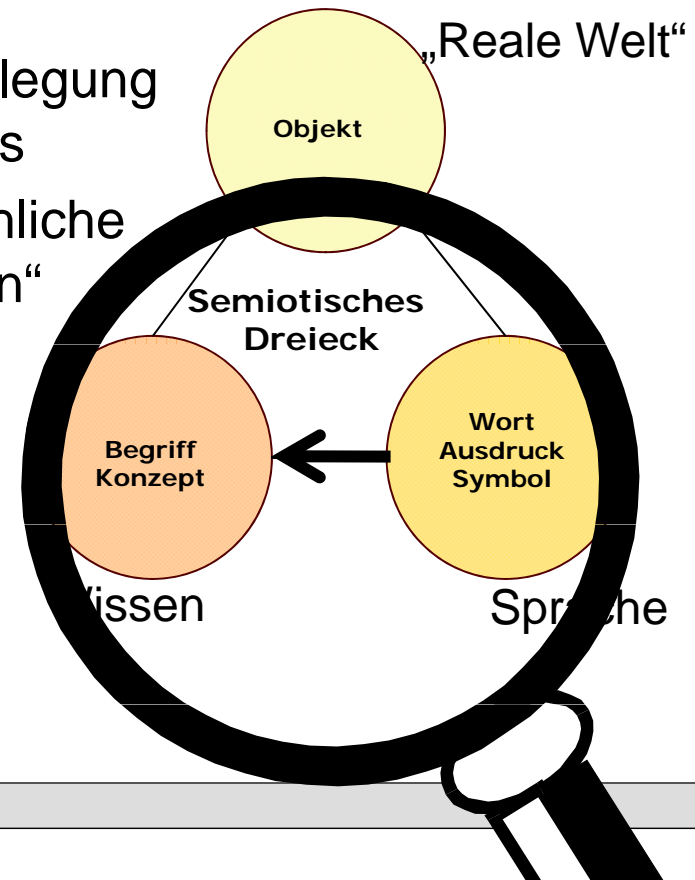
# Was sind Konzeptsysteme?

Konzeptsysteme sind Systeme von unterscheidbaren *Konzepten*, die mittels *Relationen* in Beziehung zueinander gesetzt werden und in einer natürlicheren *Sprache* formuliert werden können



**Zielsetzung:** Entwicklung und Festlegung eines gemeinsamen Verständnisses

**Repräsentationssysteme:** menschliche Sprache, Logik, „Computersprachen“



# What is LSA?

*LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse.*

**It is not** a traditional natural language processing or artificial intelligence program;

**it uses no** humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

**Instead:**

LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears.

# What is LSA?

The LSA mechanism that solves the problem consists simply of accommodating a very large number of local co-occurrence relations (between the right kinds of observational units) simultaneously in a space of the right dimensionality.

(1a)  $NP_0$  such as  $\{NP_1, NP_2 \dots, (and \mid or)\} NP_n$

are such that they imply

(1b) for all  $NP_i, 1 \leq i \leq n, hyponym(NP_i, NP_0)$

Thus from sentence (S1) we conclude

$hyponym(\text{"Bambara ndang"}, \text{"bow lute"}).$

## A look back: Hearst Patterns (!)

(S1) The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.

Hypothetically, the optimal space for the reconstruction has the same dimensionality as the source that generates discourse, that is, the human speaker or writer's semantic space.

# Excursus

Introduction to Information Retrieval

<http://informationretrieval.org>

IIR 18: **Latent Semantic Indexing**

(see additional slides)

Hinrich Schütze

Institute for Natural Language Processing, Universität Stuttgart

2009.07.21

# What is LSA?

In SVD, a rectangular matrix is decomposed into the product of three other matrices.

- One component matrix describes the original row entities as vectors of derived orthogonal factor values,
- another describes the original column entities in the same way, and
- the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed.

*There is a mathematical proof that any matrix can be so decomposed perfectly, using no more factors than the smallest dimension of the original matrix.*

# LSA

Idea (Deerwester et al):

“We would like a representation in which a set of terms, which by itself is **incomplete** and **unreliable** evidence of the relevance of a given document, is replaced by some other set of entities which are more reliable indicants.

We take advantage of the implicit higher-order (or latent) structure in the association of terms and documents to reveal such relationships.”

# LSA

## Implementation: four basic steps

- term by document matrix (more generally **term by context**) tend to be sparse
- convert matrix entries to weights, typically:
  - $L(i,j) * G(i)$ : local and global
  - $a_{ij} \rightarrow \log(\text{freq}(a_{ij}))$  divided by entropy for row ( $-\sum (p \log p)$ , over  $p$ : entries in the row)
    - weight directly by estimated importance in passage
    - weight inversely by degree to which knowing word occurred provides information about the passage it appeared in

# LSA

## Four basic steps

- Rank-reduced Singular Value Decomposition (SVD) performed on matrix
  - all but the  $k$  highest singular values are set to 0
  - produces  $k$ -dimensional approximation of the original matrix (in least-squares sense)
  - this is the “semantic space”
- Compute similarities between entities in semantic space (usually with cosine)

# LSA

## SVD

- unique mathematical decomposition of a matrix into the product of three matrices:
  - two with orthonormal columns
  - one with singular values on the diagonal
- tool for dimension reduction
- similarity measure based on co-occurrence
- finds optimal projection into low-dimensional space

# A Small Example

To see how this works let's look at a small example

This example is taken from:

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (1990). "Indexing by latent semantic analysis." *Journal of the Society for Information Science*, 41(6), 391-407.

Slides are from a presentation by Tom Landauer and Peter Foltz

# A Small Example

## Technical Memo Titles

- c1: *Human machine interface* for 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

- m1: The generation of random, binary, ordered 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

	c1	c2	c3	c4	c5	m1	m2	m3	m4
<b>human</b>	1	0	0	1	0	0	0	0	0
<b>interface</b>	1	0	1	0	0	0	0	0	0
<b>computer</b>	1	1	0	0	0	0	0	0	0
<b>user</b>	0	1	1	0	1	0	0	0	0
<b>system</b>	0	1	1	2	0	0	0	0	0
<b>response</b>	0	1	0	0	1	0	0	0	0
<b>time</b>	0	1	0	0	1	0	0	0	0
<b>EPS</b>	0	0	1	1	0	0	0	0	0
<b>survey</b>	0	1	0	0	0	0	0	0	1
<b>trees</b>	0	0	0	0	0	1	1	1	0
<b>graph</b>	0	0	0	0	0	0	1	1	1
<b>minors</b>	0	0	0	0	0	0	0	1	1

Bag of Words /  
no knowledge of e.g. POS tags:

# A Small Example - 2

	<b>c1</b>	<b>c2</b>	<b>c3</b>	<b>c4</b>	<b>c5</b>	<b>m1</b>	<b>m2</b>	<b>m3</b>	<b>m4</b>
<b>human</b>	1	0	0	1	0	0	0	0	0
<b>interface</b>	1	0	1	0	0	0	0	0	0
<b>computer</b>	1	1	0	0	0	0	0	0	0
<b>user</b>	0	1	1	0	1	0	0	0	0
<b>system</b>	0	1	1	2	0	0	0	0	0
<b>response</b>	0	1	0	0	1	0	0	0	0
<b>time</b>	0	1	0	0	1	0	0	0	0
<b>EPS</b>	0	0	1	1	0	0	0	0	0
<b>survey</b>	0	1	0	0	0	0	0	0	1
<b>trees</b>	0	0	0	0	0	1	1	1	0
<b>graph</b>	0	0	0	0	0	0	1	1	1
<b>minors</b>	0	0	0	0	0	0	0	1	1

$$r(\text{human.user}) = -.38 \quad r(\text{human.minors}) = -.29$$

## *A Small Example - 4*

$\{U\} =$

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

# *A Small Example - 5*

$\{S\} =$

3.34

2.54

2.35

1.64

1.50

1.31

0.85

0.56

0.36

## *A Small Example - 6*

$\{V\} =$

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

# A Small Example - 7

	c1	c2	c3	c4	c5	m1	m2	m3	m4
<b>human</b>	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
<b>interface</b>	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
<b>computer</b>	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
<b>user</b>	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
<b>system</b>	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
<b>response</b>	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
<b>time</b>	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
<b>EPS</b>	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
<b>survey</b>	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
<b>trees</b>	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
<b>graph</b>	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
<b>minors</b>	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

1  
0

Spearman Rank correlation:

$$r(\text{human.user}) = .94$$

$$r(\text{human.minors}) = -.83$$

# A Small Example - 2 reprise

	c1	c2	c3	c4	c5	m1	m2	m3	m4
<b>human</b>	1	0	0	1	0	0	0	0	0
<b>interface</b>	1	0	1	0	0	0	0	0	0
<b>computer</b>	1	1	0	0	0	0	0	0	0
<b>user</b>	0	1	1	0	1	0	0	0	0
<b>system</b>	0	1	1	2	0	0	0	0	0
<b>response</b>	0	1	0	0	1	0	0	0	0
<b>time</b>	0	1	0	0	1	0	0	0	0
<b>EPS</b>	0	0	1	1	0	0	0	0	0
<b>survey</b>	0	1	0	0	0	0	0	0	1
<b>trees</b>	0	0	0	0	0	1	1	1	0
<b>graph</b>	0	0	0	0	0	0	1	1	1
<b>minors</b>	0	0	0	0	0	0	0	1	1

Spearman Rank correlation:

$$r_{\text{human.user}} = -.38 \quad r_{\text{human.minors}} = -.29$$

# Correlation

## Raw data

	<i>c1</i>	<i>c2</i>	<i>c3</i>	<i>c4</i>	<i>c5</i>	<i>m 1</i>	<i>m 2</i>	<i>m 3</i>
<i>c2</i>	- 019							
<i>c3</i>	0.00	0.00						
<i>c4</i>	0.00	0.00	0.47					
<i>c5</i>	- 033	0.58	0.00	- 031				
<i>m 1</i>	- 017	- 030	- 021	- 016	- 017			
<i>m 2</i>	- 026	- 045	- 032	- 024	- 026	0.67		
<i>m 3</i>	- 033	- 058	- 041	- 031	- 033	0.52	0.77	
<i>m 4</i>	- 033	- 019	- 041	- 031	- 033	- 017	0.26	0.56

<b>0.02</b>	
<b>- 030</b>	<b>0.44</b>

Correlations in first-two dimension space

<i>c2</i>	0.91							
<i>c3</i>	1.00	0.91						
<i>c4</i>	1.00	0.88	1.00					
<i>c5</i>	0.85	0.99	0.85	0.81				
<i>m 1</i>	- 085	- 056	- 085	- 088	- 045			
<i>m 2</i>	- 085	- 056	- 085	- 088	- 044	1.00		
<i>m 3</i>	- 085	- 056	- 085	- 088	- 044	1.00	1.00	
<i>m 4</i>	- 081	- 050	- 081	- 084	- 037	1.00	1.00	1.00

Mean Spearman Corr.

0.92	
-0.72	1.00

# Evaluation – Synonym Detection

*How would you go about designing an evaluation test?*

It is claimed that LSA, on average, represents words of similar meaning in similar ways.

When one compares words with similar vectors as derived from large text corpora, the claim is largely but not entirely fulfilled at an intuitive level. Most very near neighbors appear closely related in some manner.

In one scaling (an LSA/SVD analysis) of an encyclopedia, “physician,” “patient,” and “bedside” were all close to one another,  $\cos > .5$ .

# Evaluation – Synonym Detection

The TOEFL vocabulary test consists of items in which

- the question part is usually a single word, and
- there are four alternative answers, usually single words, from which the test taker is supposed to choose the one most similar in meaning.

## Practice Questions

1. Receptors for the sense of smell are located at the top of the nasal cavity.  
A. upper end  
B. inner edge  
C. mouth  
D. division
2. Passenger ships and aircraft are often equipped with ship-to-shore or air-to-land radio telephones.  
A. highways  
B. railroads  
C. planes  
D. sailboats
3. Dotting the marshy expanse of the Florida Everglades are little islands known locally as hummocks.  
A. generally  
B. to all  
C. in that area  
D. occasionally

# Evaluation – Synonym Detection

How would you go about designing an evaluation test?

LSA's knowledge of synonyms was assessed with a standard vocabulary test.

The 80 item test was taken from retired versions of the **Educational Testing Service** (ETS) and the **Test of English as a Foreign Language** (TOEFL: for which we are indebted to Larry Frase and ETS).

**LSA was trained** by running the SVD analysis on a large corpus of representative English. In various studies, collections of *newspaper text* from the Associated Press news wire and *Grolier's Academic American Encyclopedia* (a work intended for students), and a *representative collection of children's reading* have been used.

# Evaluation – Synonym Detection

In one experiment:

- trained on a total of 4.5 million words of text,
- roughly equivalent to what a child would have read by the end of eighth grade.
- resulted in a vector for each of 60 thousand words.

# Evaluation – Synonym Detection

## LSA Approach:

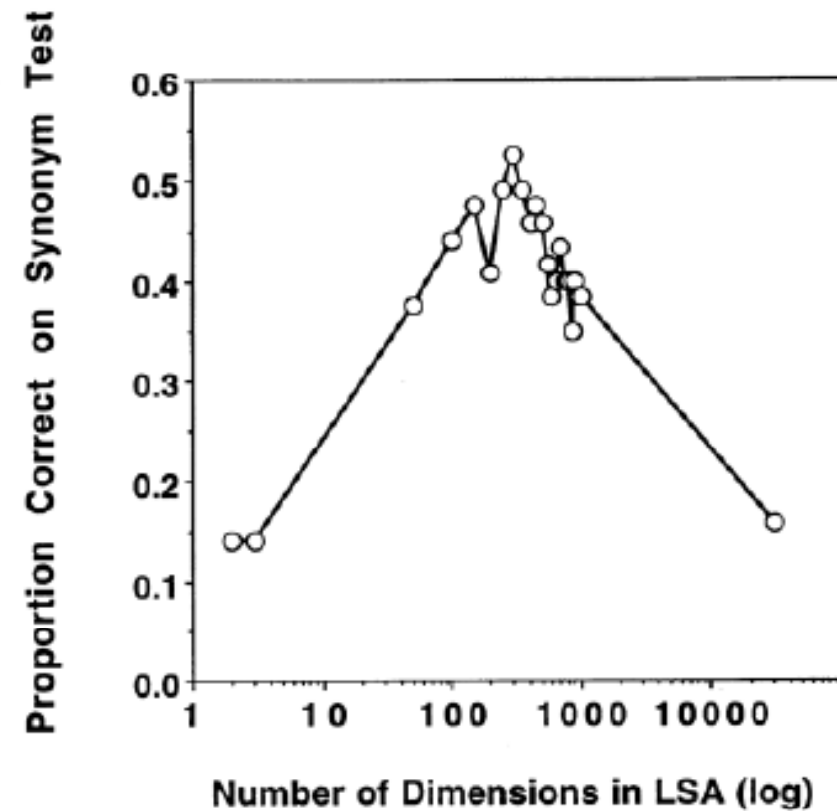
- To simulate human performance, the cosine between the question word and each alternative was calculated, and
- the LSA model chose the alternative closest to the stem

## Result:

- LSA got 65% correct, identical to the average score of a large sample of students applying for college entrance in the United States from non-English speaking countries.

# Evaluation – Synonym Detection

## Influence of Number of Dimensions



# Summary

## Some Issues

- SVD Algorithm complexity  $O(n^2k^3)$ 
  - $n$  = number of terms
  - $k$  = number of dimensions in semantic space (typically small ~50 to 350)
  - for stable document collection, only have to run once
  - dynamic document collections: might need to rerun SVD, but can also “fold in” new documents

# Summary

## Some issues

- Finding optimal dimension for semantic space
  - The number of dimensions retained in LSA is an empirical issue. Because the underlying principle is that the original data should not be perfectly regenerated but, rather, an optimal dimensionality should be found that will cause correct induction of underlying relations
  - run SVD once with big dimension, say  $k = 1000$ 
    - then can test dimensions  $\leq k$
  - in many tasks 150-350 works well, still room for research

# Summary

## Some issues

- SVD assumes normally distributed data
  - term occurrence is not normally distributed
  - matrix entries are weights, not counts, which may be normally distributed even when counts are not

# Summary

Has proved to be a valuable tool in many areas of NLP as well as IR

- summarization
- cross-language IR
- topics segmentation
- text classification
- question answering
- more

# Summary

## Ongoing research and extensions include

- Probabilistic LSA (Hofmann)
- Iterative Scaling (Ando and Lee)
- Psychology
  - model of semantic knowledge representation
  - model of semantic word learning

# Some History

## The first papers about LSI:

- Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285.
- Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R.A. (1990) "Indexing by latent semantic analysis." Journal of the Society for Information Science, 41(6), 391-407.
- Foltz, P. W. (1990) "Using Latent Semantic Indexing for Information Filtering". In R. B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA, 40-47.

Any questions?

**See you tomorrow!**