Investigation of Latent Semantic Analysis for Clustering of Czech News Articles

Michal Rott, Petr Červa

Laboratory of Computer Speech Processing

4. 9. 2014

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Idea of article clustering

Presumptions:

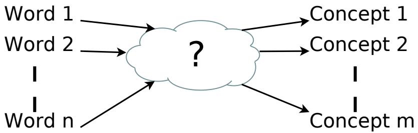
- A news article has only one main topic
- Topic is identified by combination of terms
- Documents with similar topic belong to same cluster

Problems:

- Ambiguous meaning of words
- Language inflection - e.g.: Czech nouns 14 forms
- Author's creativity generates "noise"
- Similarity of articles

What is the concept?

Concepts are represented as patterns of words.



Author's selection of words generates
a noise in word-concept relation.

The LSA solves task of reducing this noise.

Adopted LSA–based method

Idea: similar documents have similar concepts

Algorithm:

1. Preprocess input documents
2. Create term-document matrix **A**
3. Generate concept space
4. Reduce concept space
5. Extract vectors for documents from document matrix
6. Create hierarchy of similarity of the documents

Creation of term-document matrix

term-document matrix A

$$
\begin{array}{cc}
 & \text{documents} \\
\rotatebox{90}{terms} & \begin{bmatrix}
a_{11} & \cdots & a_{1d} & \cdots & a_{1|D|} \\
\vdots & & \vdots & & \vdots \\
a_{t1} & \cdots & a_{td} & \cdots & a_{t|D|} \\
\vdots & & \vdots & & \vdots \\
a_{|T|1} & \cdots & a_{|T|d} & \cdots & a_{|T||D|}
\end{bmatrix}
\end{array}
$$

$$a_{td} = TF(t \in d) * IDF(t)$$

$$IDF(t) = log \frac{|D_b|}{|\{d_b \in D_b : t \in d_b\}|}$$

Latent semantic analysis

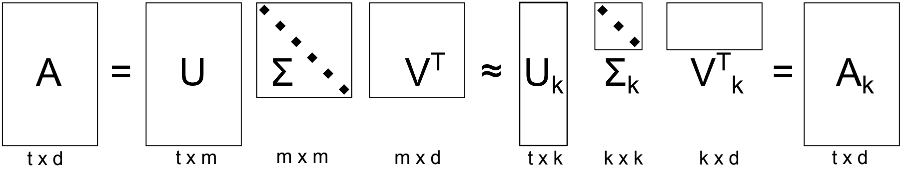
$$A = \overset{U}{\begin{bmatrix} \begin{bmatrix} & u_1 & \end{bmatrix} \\ \vdots \\ \begin{bmatrix} & u_m & \end{bmatrix} \end{bmatrix}} \cdot \overset{\Sigma}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}} \cdot \overset{V^T}{\begin{bmatrix} \begin{bmatrix} \\ v_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ v_m \\ \\ \end{bmatrix} \end{bmatrix}}$$

$\vec{u_1} \cdots \vec{u_m}$ are eigenvectors of AA^T
$\sigma_1 \cdots \sigma_m$ are singular values of $A^T A$
$\vec{v_1} \cdots \vec{v_m}$ are eigenvectors of AA^T

`http://www.alglib.net/`

Reduction of concept space

- Reduction during composition:
  - Lemmatisation
  - Remove stop words
  - Synonym replacement
  - Lemma occurrence threshold
- Low–rank approximation of concept space



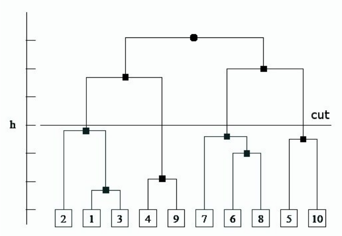
Hierarchical clustering

initialization: every document is in its own centroid

1. find the most similar centroids
2. merge them and calculate new centroid coordinates
3. repeat step 1 and 2, until only one centroid exists
4. cut dendrogram on demanded number of clusters

optimal:
don't build the whole dendrogram
stop condition:

- number of cluster reached
- distance of clusters is too height



Example — concept space

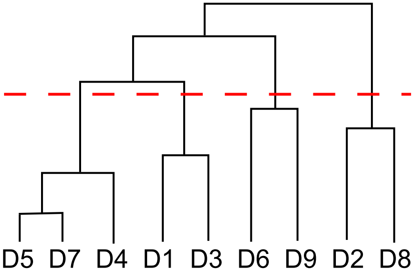
term-document matrix A

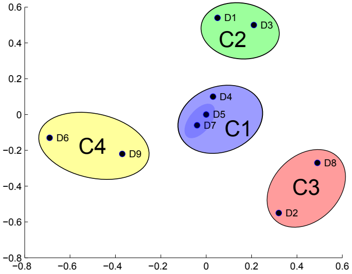
Lemmas	D1	D2	D3	D4	D5	D6	D7	D8	D9
průvodce	1					1			
investice	1	1	1	1	1	1	1	1	1
obchod	1		1						
zásoba	1		1					1	
hlupák		1						1	
kniha			1	1					
hodnota				1	1				
otec						1			1
bohatý						2			1
majetek							1		1
skutečný							1		1

SVD $\rightarrow V_K^T$

D1	D2	D3	D4	D5	D6	D7	D8	D9
-0.42	-0.31	-0.36	-0.20	-0.18	-0.44	-0.18	-0.45	-0.30
0.05	0.32	0.21	0.03	0	-0.69	-0.04	0.49	-0.37
0.54	-0.55	0.50	0.10	0	-0.13	-0.06	-0.27	-0.22

Example — clustering





$C1 = \{D4, D5, D7\}$	D4: kniha, hodnota D7: majetek, skutečný	D5: hodnota	Common: investice
$C2 = \{D1, D3\}$	D1: průvodce	D3: kniha	Common: investice, obchod, zásoba
$C3 = \{D2, D8\}$	D8: zásoba		Common: investice, hlupák
$C4 = \{D6, D9\}$	D6: průvodce	D9: majetek, skutečný	Common: investice, bohatý, otec

Metrics used

Count–based document similarity

- Euclidean distance
- Normalized Euclidean distance (Mahalanobis distance)
- Cosine similarity

Evaluation metrics:

- Internal metrics

$$\textbf{Silhouette} = \frac{1}{N}\sum_{i=1}^{N}\frac{n_i - c_i}{max\{c_i, n_i\}}$$

Error Sum of Squares, Davies–Bouldin

- External matrics

$$\textbf{Rand index} = \frac{TP+TN}{TP+FP+FN+TN}$$

Purity, F-measure, Jaccard index, etc.

Experiments

1. Comparison of similarity measures
2. Term-document matrix vs. document matrix
3. Dimension reduction
4. Influence of the preprocessing

Data description

Test sets:

1. Query–based
   - 20 documents with key-phrase "Česká národní banka"
   - 5 clusters
   - 10 annotators created reference clusters
   - 194 words in average per document
2. Category–based
   - 100 documents
   - 10 clusters
   - publisher category is used as cluster annotation
   - 151 words in average per document

Comparison of similarity measures

test set	query-based		category-based	
	RI	SI	RI	SI
reduction 80 %	2 dimensions		10 dimensions	
euclidean	**0.565**	**0.691**	0.235	-0.438
norm. euclidean	0.543	-0.473	0.235	-0.438
cosine	0.500	-0.645	**0.733**	-0.350
reduction 50 %	7 dimensions		33 dimensions	
euclidean	0.429	**0.764**	0.233	-0.102
norm. euclidean	0.431	-0.311	0.233	-0.216
cosine	**0.651**	-0.230	**0.752**	-0.273
reduction 20 %	13 dimensions		64 dimensions	
euclidean	0.446	**0.449**	0.235	0.115
norm. euclidean	0.431	0.081	0.237	0.039
cosine	**0.635**	-0.224	**0.714**	-0.163

cosine similarity yields best results

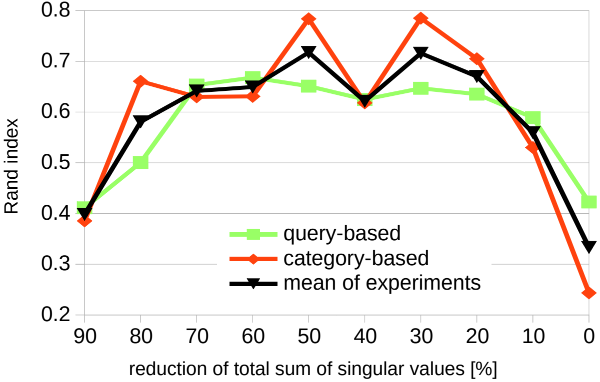
Comparison of the use of $\boldsymbol{A}_K$ and $\boldsymbol{V}_K^T$

Rand index	$\boldsymbol{A}_K$	$\boldsymbol{V}^T$
query-based	0.621	0.635
category-based	0.624	0.714

matrix size	$\boldsymbol{A}_K$	$\boldsymbol{V}^T$
query-based	2762×20	7×20
category-based	8556×100	33×100

$\boldsymbol{V}^T$ yields better results and takes less computation during similarity comparison

Experiment with dimension reduction



recommended reduction 70 - 20 %

reduction 50 % is used in other experiments

Influence of the preprocessing module

Lemmatisation module: Hunspell (tested Hunspell, FMorph and our FST analyzer)
Stop list: cca. 400 words, source: FST analyzer and Nanodictate modeling set
Synonym substitution: 7443 different groups from Wiktionary and Thesaurus

test set	query-based (mean of RI)	category-based (RI)
original input text	0.610	0.752
preprocessed text	0.651	0.785

improvement caused by preprocessing 0.037 on average

Accuracy of human cluster annotations

Query-based dataset = 20 documents, 5 clusters, 10 annotators

	min RI	mean RI	max RI
query-based	0.816	0.871	0.908

- difference between mean of annotators RI and automatic method RI is 0.086
- minimum and maximum RI difference is 0.092

Conclusion

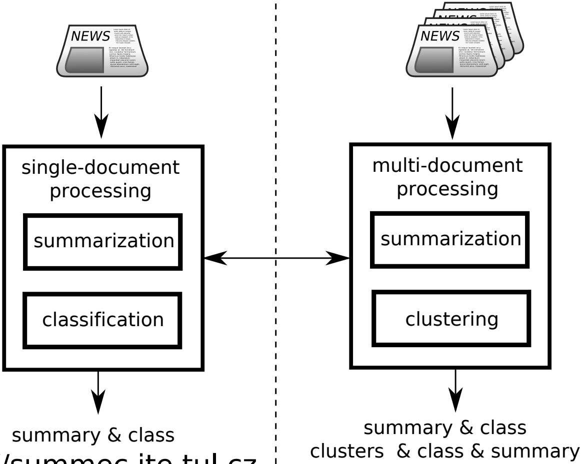
Best found configuration:

- $\boldsymbol{V}^T$ matrix gives better results
- Reduction 30-70 % of sum of all singular values
- Cosine similarity yields
- Preproccesing of the text is not necessary, but useful

Future work:

- More documents generate better combination of weighted lemma occurrences - generate vectors by LSI.
- expand to Slovak, Polish and Croatian language

News articles processing system



http://summec.ite.tul.cz

single-document summarization: TSD2013

It's done

Thank you for attention!