Search within a collection of documents

Mathematical Modelling

Nik Jenič, Tian Ključanin, Maša Uhan

Problem Introduction

- Finding relevant documents according to our search

Solution

- LSI - Latent Semantic Indexing

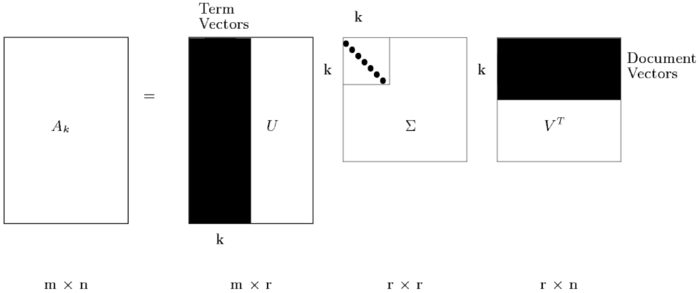

Figure: Mathematical representation of A_k

Optimization

- Giving words different weights
- Different ways of calculating the weights

$$a_{ij} = L_{ij} \cdot G_i$$

$$L_{ij} = \log(1 + f_{ij}), \quad G_i = 1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log n}, \quad p_{ij} = \frac{f_{ij}}{g_{f_i}}$$

Additional Improvements to the Solution

- Adding new documents without recalculation of SVD
- Adding new words without recalculation of SVD

Results

- Unoptimized Solution:

	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
k=**10**	642.0	620.0	583.8	539.1	478.3	394.5	294.7	163.4	55.38
k=**50**	395.5	349.9	282.6	221.0	170.1	116.9	93.88	73.63	30
k=**100**	436.0	397.2	322.9	251.9	185.2	122.6	100.7	70	22
k=**250**	545.4	513.8	451.6	360.4	257.2	193.3	127.7	83	26
k=**500**	649.4	626.7	571.4	471.2	354.9	240.6	156	79	22
k=**750**	649.4	626.7	571.4	471.2	354.9	240.6	156	79	22
k=**1000**	649.4	626.7	571.4	471.2	354.9	240.6	156	79	22

Results

- Optimized Solution:

Discussion

- NE PUSTIT PRAZNO

References

- Source for Figure 1: M. W. Berry, S.T. Dumais, G.W. O'Brien, Michael W. Berry, Susan T. Dumais, and Gavin. Using linear algebra for intelligent information retrieval. SIAM Review, 37:573–595, 1995