TENSOR BASED TEXT REPRESENTATION: A NEW DIMENSION IN IR

Ioannis Antonellis and Efstratios Gallopoulos

Department of Computer Engineering & Informatics, University of Patras, Greece

{antonell, stratis}@ceid.upatras.gr

Abstract

We investigate the basics of and experiment with a tensor based document representation in Information Retrieval. Most documents have an inherent hierarchical structure that renders desirable the use of flexible multidimensional representations such as those offered by tensor objects. We focus on the performance of special instances of a Tensor Model, in which documents are represented using second-order (matrices) and third-order tensors. We exploit the local-structure encapsulated by the proposed representation to approximate these tensors using high order singular value and nonnegative tensor decompositions and assemble the results to form the final term-document matrix. We analyze the spectral properties of this matrix and observe that topic identification is enhanced by deploying k-plane clustering. Our results provide evidence that tensor based models can be particularly effective for IR, offering an excellent alternative to traditional VSM and LSI especially for text collections of multi-topic documents.

Motivation

- Document vectors have no memory on how they were constructed
- Any document vector can be decomposed in an unlimited number of ways as linear combination of sufficiently many but almost any vectors

Basic idea:

Exploit meaningful decompositions of each document vector

↓

Store those decompositions into a Tensor Object $\mathcal{D}$

↓

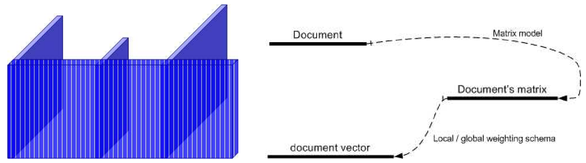
Apply a procedure of local denoising within $\mathcal{D}$ resulting on a new object $\widehat{\mathcal{D}}$

↓

Produce final document vector using $\widehat{\mathcal{D}}$ instead of $\mathcal{D}$

Related Work: Matrix Space Models (MSMs) [1]

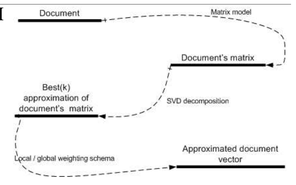
- **Approach:** Exploit document's intrinsic hierarchical structure: Sentences, Paragraphs, Sections, ... $\rightsquigarrow$ "extracts"
- **Observation:** Every document is
  - sum of vectors resulting from document's terms appearing at one, selected level of the document's hierarchical structure
  - algebraic product of a term-by-extract matrix (tem) with the "'all 1's" vector e



IR technique based on MSM: denoising within the tem matrix

Algorithm: Construct pseudo-tdm based on MSM
Input: Document collection $\{D_1, ..., D_m\}$
Output: Pseudo-tdm A
I. For each document D_j:
1. Prepare tsm S
2. Select $k' \leq \mathrm{rank}(S)$
3. $a_j = \mathrm{best}_{k'}(S)e^{(t)}$

II. Assemble pseudo-tdm $A := [a_1, ..., a_m]$



- eliminate polysemy and synonymy within each document
- effective for multi-topic documents

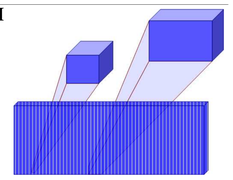
Tensor based Models (TSMs)

- **Approach:** Exploit many levels of document's intrinsic hierarchical structure
- **Observation:** Every document is
  - sum of vectors resulting from document's terms appearing at the selected level of the document's hierarchical structure
  - these vectors can be further decomposed into a number of vectors corresponding to the next level of hierarchy's analysis
  - recursively we can apply this decomposition procedure for all hierarchy levels
  - resulting objects are $(n+1)$-way tensors (for a document's hierarchy with n levels) $\rightsquigarrow$ e.g. (sections $\times$ sentences $\times$ terms)

IR technique based on TSMs: denoising within the tensor object

Algorithm: Construct pseudo-tdm based on TSM
Input: Document collection $\{D_1, ..., D_m\}$
Output: Pseudo-tdm A
I. For each document D_j:
1. Prepare tensor $\mathcal{T}$
2. Select $k' \leq \min\_\text{n-rank}(\mathcal{T})$
3. $a_j = \mathrm{unfold}(\mathrm{DENOISE}_{k'}(\mathcal{T}))e^{(t)}$

II. Assemble pseudo-tdm $A := [a_1, ..., a_m]$



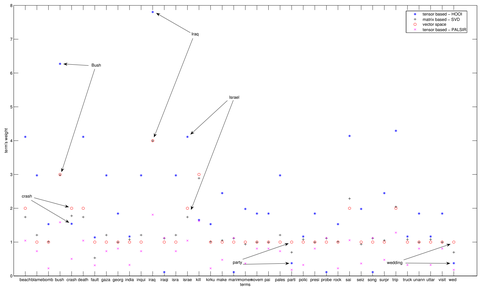
- As function $\mathrm{DENOISE}_{k'}$ we experiment with High Order Singular Value Decomposition (HOSVD) [3], High Order Orthogonal Iteration (HOOI) [3] and Projected Alternating Least Squares with Initialization and Regularization (PALSIR) [5]
- best precision/recall obtained using HOOI
- HOOI results on document vectors with clustered coefficients $\rightsquigarrow$ we deploy k-plane clustering [4] for topic identification using the produced document vector

Tensor based text representation: Example

Multi-topic document:

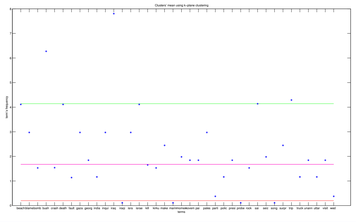
MSM representation:

TSM representation:



Note horizontal clusters that HOOI produces

Experiments for topic identification with k-plane clustering



keywords from Cluster 1: president, visit, unannounced ...

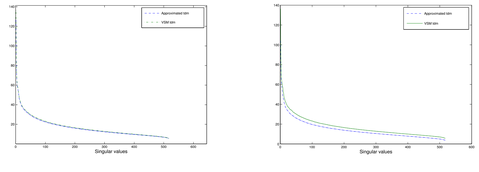
keywords from Cluster 2: palestinian, beach, gaza ...

keywords from Cluster 3: wedding, party, ...

Spectral properties of pseudo-tdms based on MSMs

Theorem 1 *Let $A \in \mathbb{R}^{m \times n}$ and write $A = [A_1, A_2]$ where $A_1 \in \mathbb{R}^{m \times n_1}$ and $A_2 \in \mathbb{R}^{m \times n_2}$. Then for any k_1, k_2, we have*

$$\sigma_i\left([\mathrm{best}_{k_1}(A_1)\, e^{(n_1)}, \mathrm{best}_{k_2}(A_2)\, e^{(n_2)}]\right) \leq \sigma_i\left([A_1 e^{(n_1)}, A_2 e^{(n_2)}]\right) \quad (1)$$



Experimental evaluation of MSM models

- MEDLINE dataset, artificial datasets to test the perfomance of the method when applied to multi-topic documents (MED_1, MED_2,..., MED_10), Documents of MED_i dataset contain i MEDLINE documents
- Tensors and tdm's constructed using add-ons to the Text to Matrix Generator MATLAB tool (TMG)[5] and Matlab tensor toolbox [2].

MED #	VSM	New	LSI $k=20$	LSI $k=100$
1	2(7%)	9(30%)	9(30%)	7(23%)
2	3(10%)	8(27%)	9(30%)	9(30%)
3	4(13%)	3(10%)	14(47%)	8(27%)
4	4(13%)	7(23%)	15(50%)	4(13%)
5	4(13%)	6(20%)	13(43%)	5(17%)
6	3(10%)	7(23%)	11(37%)	7(23%)
7	4(13%)	9(30%)	13(43%)	2(7%)
8	3(10%)	5(17%)	14(47%)	7(23%)
9	2(67%)	7(23%)	9(30%)	10(33%)
10	6(20%)	9(30%)	13(43%)	1(3%)

Number of queries that each method answers with greater precision for MEDLINE using 5 singular triplets.

References

[1] I. Antonellis and E. Gallopoulos, Exploring term document matrices from matrix models in text mining, In Proc. Text Mining Workshop, SIAM Data mining Conference 2006.

[2] Brett W. Bader and Tamara G. Kolda, MATLAB Tensor Classes for Fast Algorithm Prototyping, ACM Tran. Math. Software, to appear.

[3] L. De Lathauwer, J. Vandewalle Dimensionality Reduction in Higher-Order Signal Processing and Rank-$(R_1, R_2, \ldots, R_N)$ Reduction in Multilinear Algebra, Lin. Alg. Appl., Vol. 391, pp. 31-55, 2004.

[4] P. S. Bradley, O. L. Mangasarian, k-Plane Clustering, J. of Global Optimization 16, No 1, 2000, pp. 23-32.

[5] C. Boutsidis, E. Gallopoulos, P. Zhang, R. Plemmons, PALSIR: A new approach to Nonnegative Tensor Factorization, MMDS 2006, poster

[6] D. Zeimpekis, E. Gallopoulos, TMG Software, http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG