Combinatorial PCA for Text Classification

Andrei V. Anghelescu and Ilya B. Muchnik

site-visit presentation

26 February 2004

DIMACS, Rutgers University

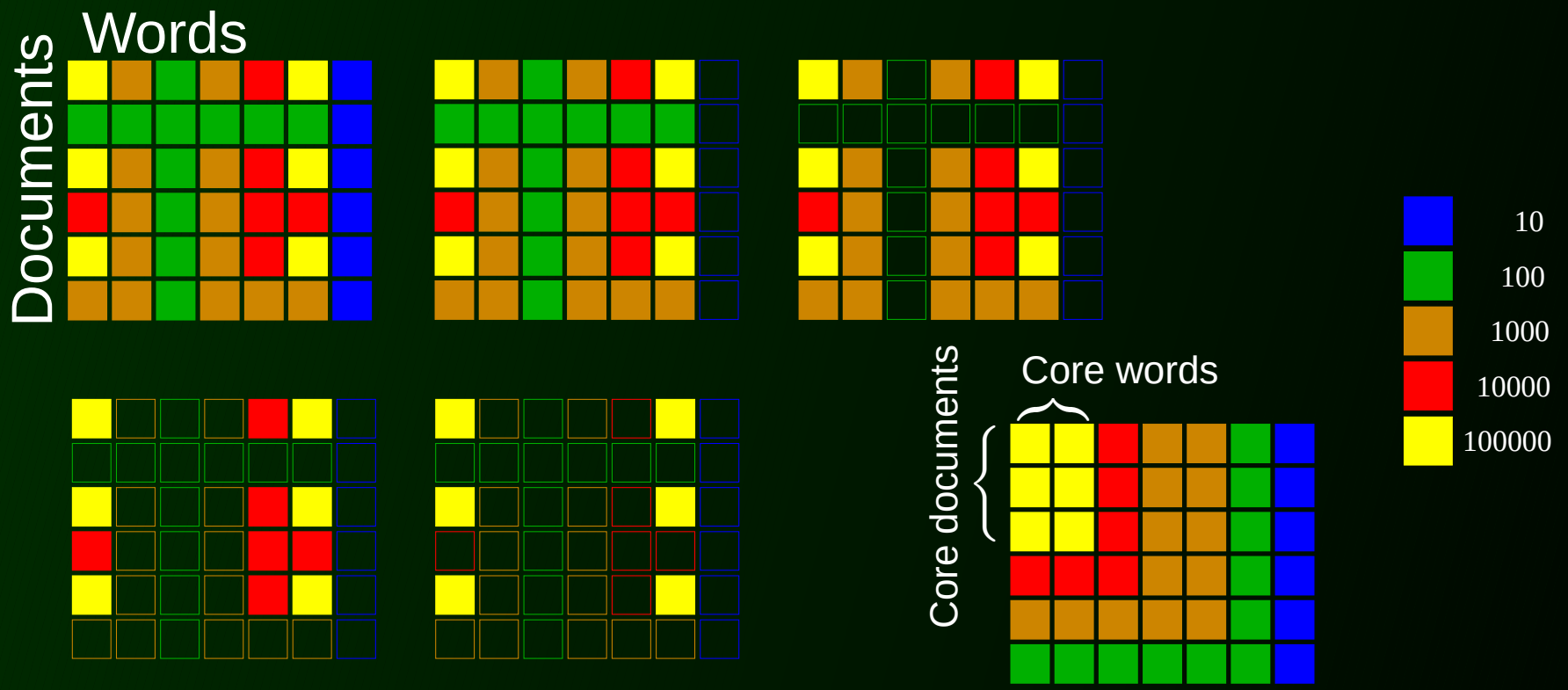
Synopsis

- simultaneously cluster documents and words
- find documents and words with high values

Road-map:

- current accomplishments: feature (word) selection
- *future work: document representation using*
  - *word concepts (clusters of words)*
  - *discriminative word concepts*

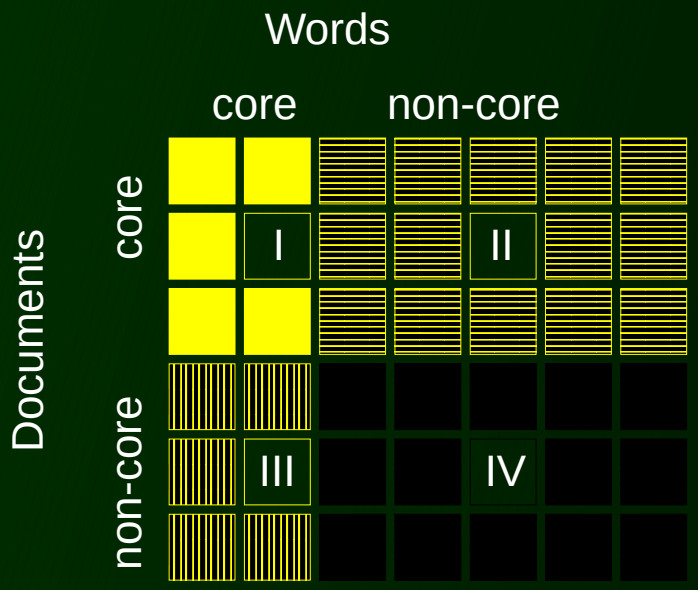
How it works...


Words
Documents
Core words
Core documents
10
100
1000
10000
100000

Summary of results (Reuters Corpus)

- data: 23,149 documents (47,152 distinct words) for training, 4,060 documents for testing
- 1,500 words selected from 47,152 ($\Rightarrow$ fewer words by a factor of 30)
- used 1-NN, linear SVM and our tuned version of SVM
- on average, F_1 scores in the reduced space classification were better than in the original space; 1-NN: $0.38 \rightarrow 0.41$; SVM: $0.40 \rightarrow 0.41$; aiSVM: $0.42 \rightarrow 0.44$
- improvement is consistent across various normalization schemes

Core-structure of the data



- natural partition of data into four quadrants
- correlated structures of words induced by the partition

#words	#doc	I	II	III	IV
647	1478	0.36	0.09	0.00	0.06
1026	2634	0.22	0.06	0.00	0.00
1509	4368	0.14	0.04	0.00	0.00
2026	6372	0.10	0.03	0.00	0.00
2513	8350	0.07	0.02	0.00	0.00
3021	10581	0.06	0.02	0.00	0.00
3519	12671	0.05	0.01	0.00	0.00
4057	14654	0.04	0.01	0.00	0.00
4502	16269	0.03	0.01	0.00	0.00
47152	23149	0.01	0.00	0.00	0.00

Average word frequencies in the four quadrants, for different sizes of the core cluster.

Future work (if funding is available)

- current results use only the extracted group of frequent words
- intermediate results provide simultaneous clusterings of documents and words
- use these intermediate results to develop novel methods of document representation
- two novel methods of document representation in:
  1. spaces of concepts (clusters of words), using corresponding document clusters for concept interpretation
  2. spaces of discriminative word concepts

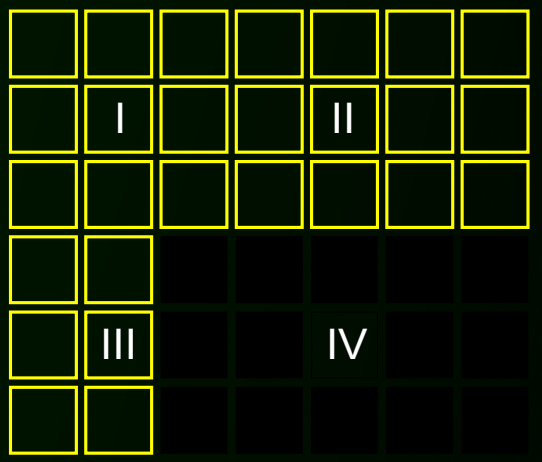
The word concept approach

Extensive prior work on concept-based models (Rendell, Setiono, Quinlan, Zheng):

- often accompanied by new semantic insights (but also creates the problem of concept interpretation)
- increased robustness of document representation
- dimensionality reduction

Novel method: *shelling word concepts*

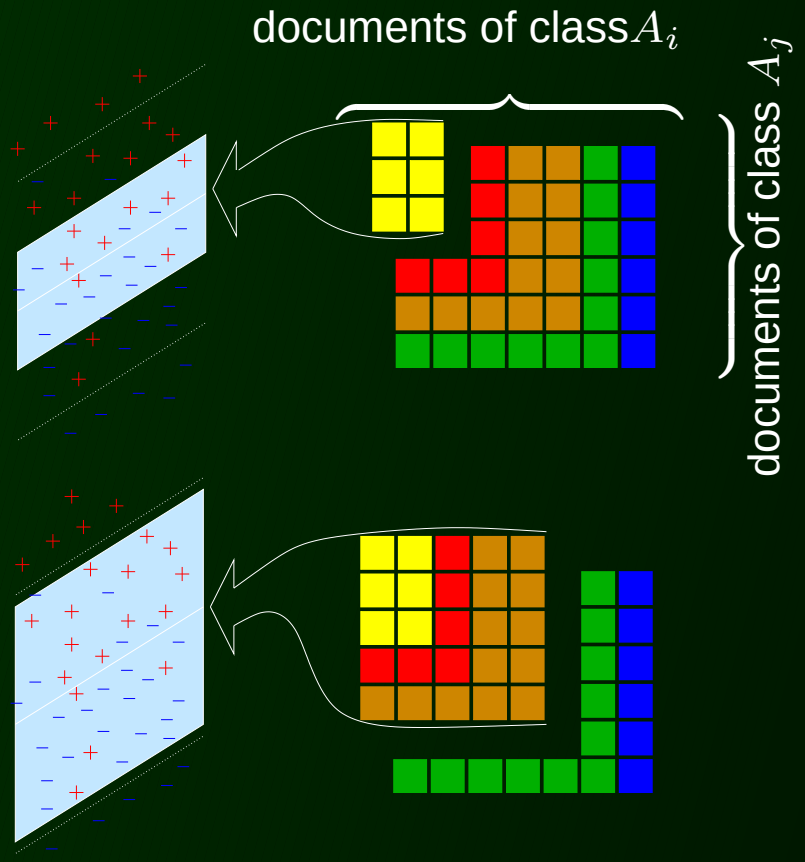
1. calculate the core cluster of the data matrix (documents $\times$ words)
2. use core words as concept and documents as its interpretation
3. *remove the core sub-matrix from data* (the basic shelling step)
4. repeat from step 1, until data is exhausted

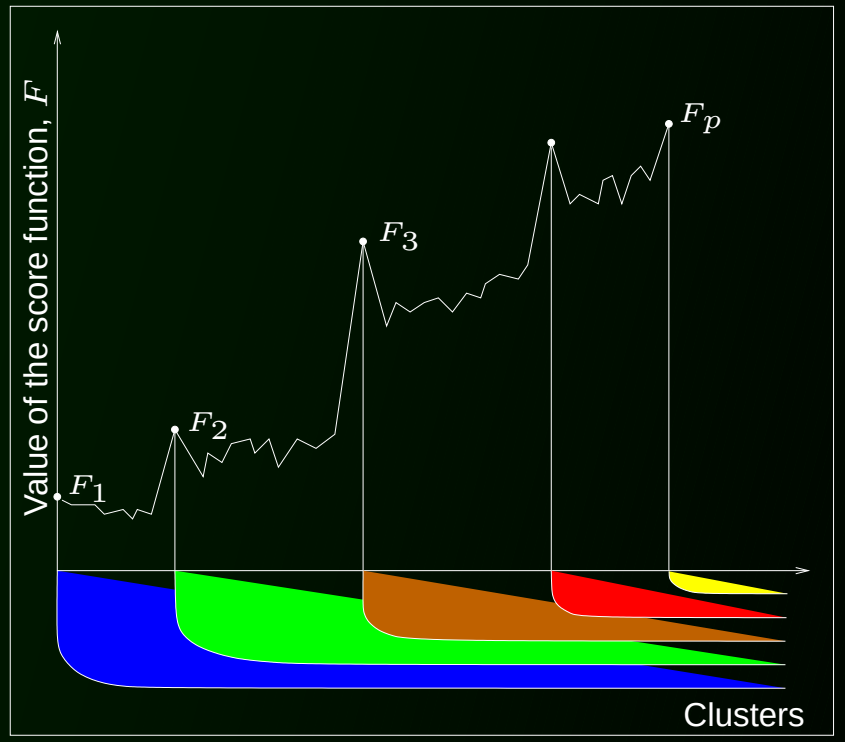


Properties of our method

- word concepts are associated with documents, which constitute a basis for interpretation
- automatic determination of the number of clusters
- fast and scalable

Discriminative concepts





- Paradigm of maximization of margin between document classes
- Input data: matrix of similarities between documents of different classes

Discriminative concepts (cont'd)

- assume a k class problem (with classes $A_1, \ldots, A_k$)
- build a similarity table for every pair of classes (A_i, A_j)
- use cPCA on such tables of similarities
- each core generates two sets of documents, D_i and D_j: $D_i \subseteq A_i, D_j \subseteq A_j$

Discriminative concepts (cont'd)

- generate concepts from the words present in these documents
  1. words in D_i but not in D_j
  2. words in D_j but not in D_i
- name such pairs of concepts *discriminative concepts*
- represent documents in the space of $\binom{k}{2}$ discriminative concepts

Discriminative concepts (cont'd)

Advantages:

- concepts have direct relationship with classifier construction
- ideally matches the SVM paradigm
- fast
- unique solution
- no parameters

Conclusions

- cPCA: a combinatorial procedure for simultaneous clustering of documents and words
- we presented the current results
- ... there is much more to obtain from the structure induced by cPCA: *novel representations* based on
  1. word concepts with associated documents as interpretation
  2. discriminative word concepts, to maximize the performance of SVM classifiers