Novel representations and methods in text classification

Manuel Montes, Hugo Jair Escalante

Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico.

http://ccc.inaoep.mx/~mmontesg/

http://ccc.inaoep.mx/~hugojair/

{*mmontesg, hugojair*}*@inaoep.mx*

7th Russian Summer School in Information Retrieval
Kazan, Russia, September 2013

Novel represensations and methods in text classification

CONCEPT-BASED REPRESENTATIONS FOR TEXT CATEGORIZATION

Outline

- The bag-of-concepts representation
- Distributional term representations (DTRs)
  - Document occurrence representation (DOR)
  - Term co-occurrence representation (TCOR)
- Using DTRs in short-text classification
- Random indexing
  - Definition and computation
  - Incorporating syntactic information
  - Results on text classification
- Final remarks

Bag-of-words representation

- Many text classification methods adopt the BoW representation because its **simplicity** and **efficiency**.
- Under this scheme, documents are represented by collections of terms, each term being an independent feature.
  - Word order is not capture by this representation
  - There is no attempt for understanding documents' content

Main problems

- BoW **ignores all semantic information**; it simply looks at the surface word forms
  - Polysemy and synonymy are big problems
- BoW tend to produce very **sparse representations**, since terms commonly occur in just a small subset of the documents
  - This problem is amplified by lack of training texts and by the **shortness** of the documents to be classified.

We need representations at **concept level**!

Bag-of-concepts

- Addresses the deficiencies of the BoW by considering the relations between document terms.
- BoC representations are based on the intuition that the meaning of a document can be considered as the **union of the meanings of their terms.**
- The meaning of terms is related to their usage; it is captured by their **distributional representation**.
  - Document occurrence representation (DOR)
  - Term co-occurrence representation (TCOR)

Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations: an experimental comparison. *Thirteenth ACM international conference on Information and knowledge management* (CIKM '04). New York, NY, USA, 2004

Document occurrence representation (DOR)

- DOR representation is based on the idea that the semantics of a term may be view as a function of the **bag of documents** in which the term occurs.

Representation of terms

	d_1	d_2	...	d_n
t_1				
t_2				
:		$w_{i,j}$		
t_m				

$$w_j = \langle w_{1,j}, \ldots, w_{N,j} \rangle$$
$$0 \leq w_{k,j} \leq 1$$

$$w_{k,j} = df(d_k, t_j) \cdot log \frac{|T|}{N_k}$$

$$df(d_k, t_j) = \begin{cases} 1 + \log(\#(d_k, t_j)) & if(\#(d_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

Intuitions about the weights

$$w_{k,j} = df(d_k, t_j) \cdot log\frac{|T|}{N_k}$$

$$df(d_k, t_j) = \begin{cases} 1 + \log(\#(d_k, t_j)) & if(\#(d_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

- DOR is a **dual version of the BoW** representation, therefore:
  - The more frequently t_i occurs in d_j, the more important is d_j for characterizing the semantics of t_i
  - The more distinct the words d_j contains, the smaller its contribution to characterizing the semantics of t_i.

Representing documents using DOR

- DOR is a word representation, not a document representation.
- Representation of documents is obtained by the weighted **sum of the vectors** from their terms.

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j}$$

Word representation

Word–Document matrix

	d_1	d_2	…	d_n
t_1				
t_2				
:		$w_{i,j}$		
t_m				

Document representation

Document–Document matrix

	d_1	d_2	…	d_n
d_1				
d_2				
:		$w_{i,j}$		
d_n				

Term co-occurrence representation (TCOR)

- In TCOR, the meaning of a term is conveyed by the terms commonly co-occurring with it; i.e. terms are represented by the **terms occurring in their context**

Representation of terms

	t_1	t_2	...	t_m
t_1				
t_2				
:		$w_{i,j}$		
t_m				

$$w_j = \langle w_{1,j}, \ldots, w_{N,j} \rangle$$
$$0 \leq w_{k,j} \leq 1$$

$$w_{k,t} = tff(t_k, t_j) \cdot log \frac{|T|}{T_k}$$

$$tff(t_k, t_j) = \begin{cases} 1 + \log(\#(t_k, t_j)) & if(\#(t_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

Intuitions about the weights

$$w_{k,t} = tff(t_k, t_j) \cdot log\frac{|T|}{T_k}$$

$$tff(t_k, t_j) = \begin{cases} 1 + \log(\#(t_k, t_j)) & if(\#(t_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

- TCOR is the kind of representation traditionally used in WSD, therefore:
  - The more words t_k and t_j co-occur in, the more important t_k is for characterizing the semantics of t_j
  - The more distinct words t_k co-occurs with, the smaller its contribution for characterizing the semantics of t_j.

Representing documents using TCOR

- TCOR, such as DOR, is a word representation, not a document representation.
- Representation of documents is obtained by the weighted sum of the vectors from their terms.

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j}$$

Word representation

Word–Word matrix

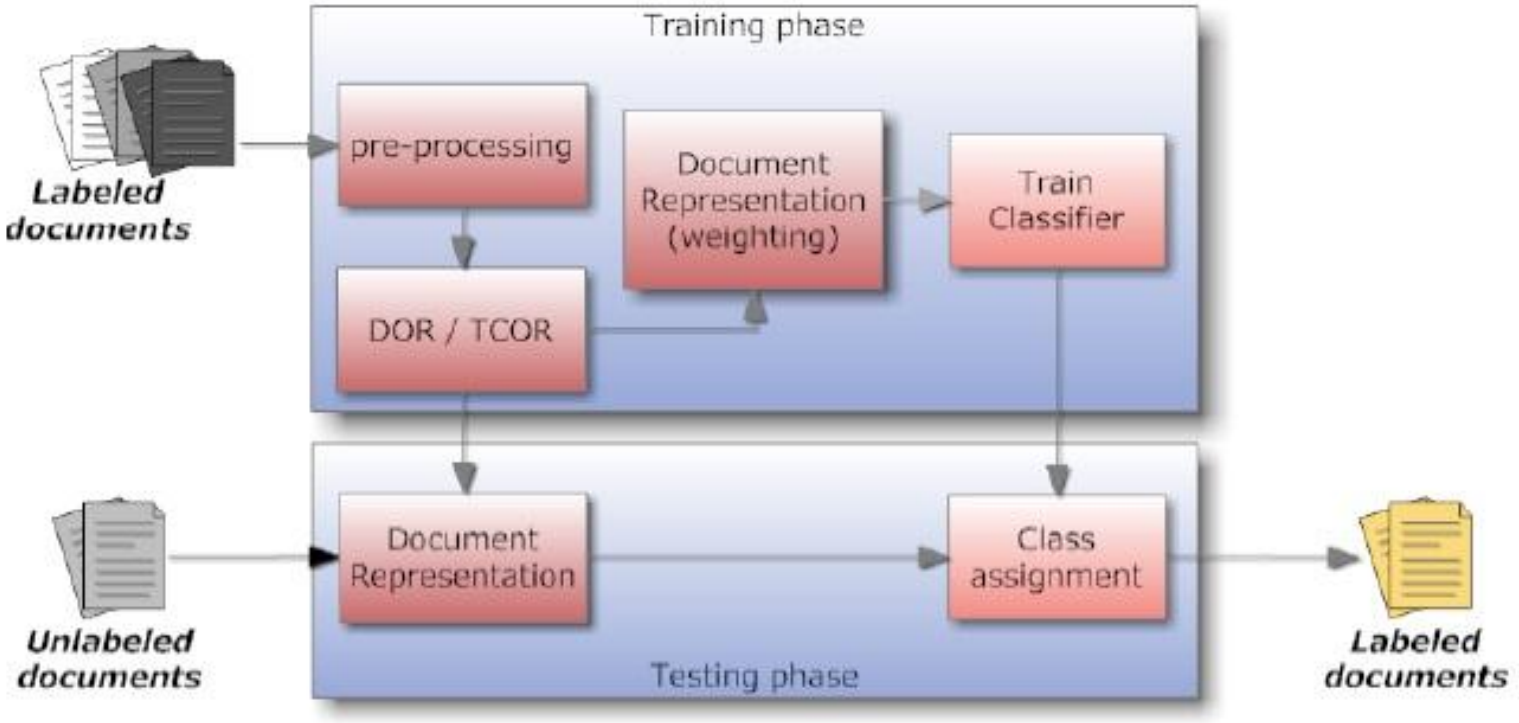
	t_1	t_2	...	t_m
t_1				
t_2				
:				
t_m				

Document representation

Document–Word matrix

	t_1	t_2	...	t_m
d_1				
d_2				
:				
d_n				

DOR/TCOR for text classification



Juan Manuel Cabrera, Hugo Jair Escalante, Manuel Montes-y-Gómez. Distributional term representations for short text categorization. *14th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2013). Samos, Greece, 2013.

Experiments

- **Short-text categorization** based on distributional term representations (DOR and TCOR)
  - They **reduce the sparseness** of representations and alleviates, to some extent, the low frequency issue.
- Our experiments aimed to:
  - Verify the difficulties of the BoW for effectively representing the content of short-texts
  - Assess the added value offered by concept-based representations over the BoW formulation

Evaluation datasets

- We assembled two types of collections:
  - Whole documents for training and test
  - Whole documents for training and **titles for test**

Feature	Train	Test (DD)	Test-Reduced (DT)
Vocabulary size	14,865	8,760	3,676
Number of Documents	4,559	2,179	2,179
Average terms per document	40.9	39.2	6.6

Reuters-R8

Feature	Regular (DD)	Reduced (DT)
Vocabulary size	1136	206
Number of Documents	48	48
Average terms per doc.	60.3	5.85

EasyAbstracts

Feature	Regular (DD)	Reduced (DT)
Vocabulary size	813	180
Number of Documents	48	48
Average terms per doc.	45.06	4.8

Cicling 2002

Short-text classification with BoW

R8

	Boolean			TF			TFIDF		
	DD	DT	Decrease	DD	DT	Decrease	DD	DT	Decrease
AdaBoost	0.64	0.18	-72.74%	0.64	0.18	-72.74%	0.64	0.18	-72.74%
Knn1	0.69	0.39	-43.98%	0.47	0.34	-27.53%	0.47	0.34	-27.53%
Naive Bayes	0.87	0.66	-24.16%	0.82	0.34	-58.97%	0.82	0.34	-59.13%
RandomForest	0.80	0.54	-32.21%	0.80	0.57	-29.02%	0.82	0.74	-10.46%
SVMLineal	**0.91**	**0.83**	-7.85%	0.90	0.73	-19.29%	0.90	0.70	-22.59%

EasyAbstract

AdaBoost	0.41	0.27	-34.34%	0.40	0.25	-37.70%	0.40	0.25	-37.70%
Knn1	0.21	0.11	-46.14%	0.14	0.09	-38.74%	0.14	0.09	-38.74%
Naive Bayes	0.70	0.40	-42.89%	0.74	0.35	-53.09%	0.79	0.37	-52.93%
RandomForest	0.57	0.24	-57.82%	0.49	0.22	-54.34%	0.53	0.19	-64.01%
SVMLineal	**0.69**	**0.59**	-15.64%	0.90	0.16	-82.05%	0.85	0.30	-64.67%

CICLing

AdaBoost s	0.36	0.27	-22.76%	0.36	0.27	-22.76%	0.31	0.20	-35.32%
Knn1	0.29	0.10	-65.62%	0.14	0.16	10.62%	0.13	0.09	-31.31%
Naive Bayes	0.43	0.33	-23.50%	0.43	0.39	-10.50%	0.37	0.14	-61.30%
RandomForest	0.40	0.25	-38.01%	0.31	0.30	-1.10%	0.22	0.12	-46.91%
SVMLineal	0.45	0.35	-21.14%	**0.54**	**0.48**	-11.91%	0.21	0.14	-35.52%

Conclusions (1)

- Acceptable performance was obtained when regular-length documents were considered
  - **SVM** obtained the best results for most configurations of data sets and weighting schemes
- The performance of most classifiers dropped considerably when classifying short documents
  - The **average decrement of accuracy was of 38.66%**
- Results confirm that the **BoW representation is not well suited for short-text classification**

Using DOR/TCOR for short-text classification

R8

Weigth	Boolean			TF			TFIDF		
Classifiers	BOW	DOR	TCOR	BOW	DOR	TCOR	BOW	DOR	TCOR
AB	0.175	0.645	0.668	0.175	0.632	0.651	0.175	0.591	0.667
KNN	0.386	**0.899**	**0.897**	0.337	**0.908**	**0.902**	0.337	**0.746**	**0.754**
NB	0.656	**0.881**	**0.893**	0.336	**0.874**	**0.886**	0.336	**0.785**	**0.854**
RF	0.543	**0.786**	**0.774**	0.565	**0.805**	**0.823**	0.736	**0.798**	**0.819**
SVM	0.834	**0.930**	**0.891**	0.728	**0.928**	**0.901**	0.699	**0.897**	**0.784**

EasyAbstract

AB	0.268	0.185	0.201	0.255	0.272	0.245	0.250	0.263	0.292
KNN	0.114	**0.600**	**0.482**	0.086	**0.666**	**0.712**	0.086	**0.571**	**0.541**
NB	0.402	**0.568**	**0.586**	0.345	**0.603**	**0.590**	0.370	**0.578**	**0.603**
RF	0.239	**0.495**	**0.332**	0.223	**0.507**	**0.582**	0.192	**0.588**	**0.550**
SVM	0.585	0.660	**0.639**	0.161	**0.728**	**0.733**	0.301	**0.622**	**0.589**

CICLIng2002

AB	0.274	0.188	0.244	0.274	0.129	0.224	0.199	0.201	0.232
KNN	0.099	**0.450**	**0.395**	0.156	**0.478**	**0.399**	0.089	**0.493**	**0.44**
NB	0.332	**0.473**	0.415	0.386	**0.426**	0.471	0.143	**0.506**	**0.399**
RF	0.249	0.184	0.369	0.304	0.279	0.374	0.119	**0.418**	**0.291**
SVM	0.354	**0.526**	0.414	0.48	0.504	0.502	0.135	**0.528**	**0.442**

Conclusions (2)

- **DOR and TCOR clearly outperformed BoW** for most configurations.
  - In 62 out of the 90 results the improvements of DTRs over BoW were statistically significant
- In average, results obtained with DOR and TCOR were very similar.
  - DOR is advantageous over TCOR because it may result in document representations of much lower dimensionality.

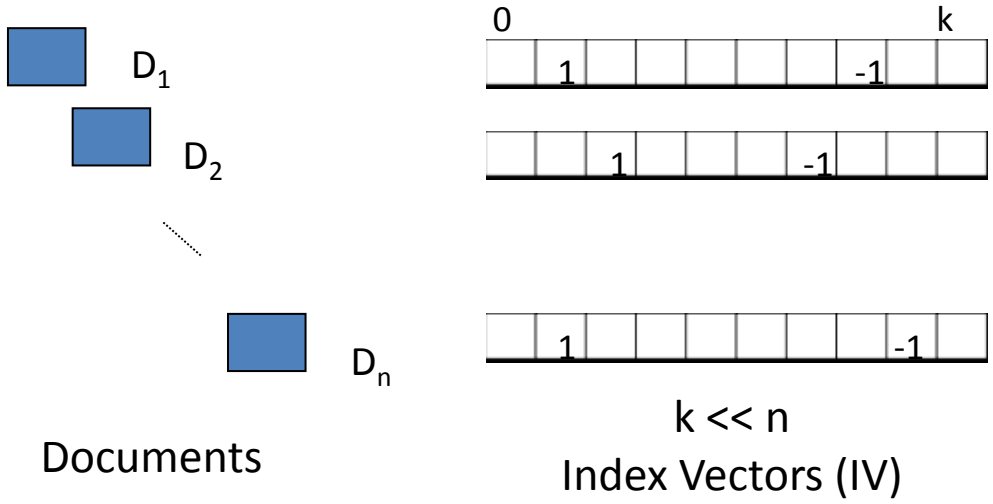
Bag of concepts by random indexing

- BoC approaches tend to be **computationally expensive**.
- They are based on a **co-occurrence matrix** of order ***w*×*c***; *w* = terms, and *c* = contexts (terms or documents)
- Random indexing produce these context vectors in a more computationally efficient manner: the co-occurrence matrix is replaced by a **context matrix** of order ***w*×*k***, where ***k* << *c***.

Magnus Sahlgren and Rickard Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. *20th international conference on Computational Linguistics* (COLING '04). Stroudsburg, PA, USA, 2004.

Random indexing procedure (1)

- *First step*: a unique **random representation** known as "index vector" is assigned to each context.
  - A context could be a document , paragraph or sentence
  - Vectors are filled with -1, 1 and 0s.

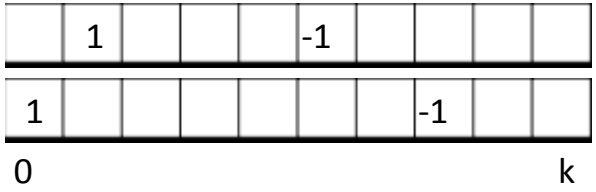


Random indexing procedure (2)

- *Second step*: index vectors are used to produce **context vectors** by scanning through the text

D_1: Towards an Automata Theory of Brain

D_2: From Automata Theory to Brain Theory



The context vector for brain

- *Third step*: build **document vectors** by adding their terms' context vectors.

d_i: "From *Automata* *Theory* to *Brain* *Theory*"
CV_1 CV_2 CV_3 CV_2

d_i will be represented as the weighted sum of these vectors:

$a_1CV_1+a_2CV_2+a_3CV_3+a_2CV_2$ a_1, a_2, a_3 are idf-values

Limitations of BoC representations

- BoC representations **ignore the large amount of syntactic data** in the documents not captured implicitly through term context co-occurrences
- Although BoC representations can successfully model some synonymy relations, since different words with similar meaning will occur in the same contexts, they **can not model polysemy** relations.
- Solution: a representation that encodes both the semantics of documents, as well as the **syntax** of documents

Random indexing with syntactic information

- Multiplicative bidding procedure:
  - For each PoS tag, generate a unique **random vector for the tag** of the same dimensionality as the term context vectors.
  - For each term context vector, we perform **element-wise multiplication** between that term's context vector and its identified PoS tag vector to obtain our combined representation for the term.
  - Finally, document vectors are created by **summing the combined term vectors**.

Jonathan M. Fishbein and Chris Eliasmith. Methods for augmenting semantic models with structural information for text classification. *30th European conference on Advances in information retrieval* (ECIR'08). Glasgow, UK, 2008.

An alternative procedure

- Circular convolution procedure:
  - For each PoS tag, generate a unique **random vector for the tag** of the same dimensionality as the term context vectors
  - For each term context vector, perform **circular convolution**, which binds two vectors :

term $\underline{A} = (a_0, a_1, \ldots, a_{n-1})$

tag $\underline{B} = (b_0, b_1, \ldots, b_{n-1})$

term-tag $\underline{C} = (c_0, c_1, \ldots, c_{n-1})$

$$\underline{C} = \underline{A} \otimes \underline{B}$$

$$c_j = \sum_{k=0}^{n-1} a_k b_{j-k}$$

  - Finally, document vectors are created by **summing the combined term vectors**

Circular convolution as binding operation

- Two properties that make it appropriate to be used as a binding operation:
  - The expected similarity between a convolution and its constituents is zero, thus **differentiating the same term acting as different parts of speech** in similar contexts.
    - Gives high importance to syntactic information
  - Similar semantic concepts (i.e., term vectors) bound to the same part-of-speech will result in similar vectors; therefore, usefully **preserving the original semantic model**.
    - Preserves semantic information

Results on text classification

- The goal of the experiment was to demonstrate that **integrating PoS** data to the text representation is **useful for classification** purposes.
- Experiments on the 20 Newsgroups corpus; a linear SVM kernel function was used; all context vectors were fixed to 512 dimensions

Syntactic Binding Method	$\mathcal{F}_1$ Score
BoC (No Binding)	56.55
Multiplicative Binding	57.48
Circular Convolution	**58.19**

Final remarks

- BoC representations constitute a viable supplement to word based represention.
- Not too much work in text classification and IR
  - Recent experiments demonstrated that TCOR, DOR and random indexing results outperform those from traditional BoW; in CLEF collections improvements have been around 7%.
- Random indexing is efficient, fast and scalable; syntactic information is easily incorporated.

References

- Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter. **Distributional word clusters vs. words for text categorization**. *Journal of Machine Learning Research.* March 2003.
- Juan Manuel Cabrera, Hugo Jair Escalante, Manuel Montes-y-Gómez. **Distributional term representations for short text categorization**. *14th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2013). Samos, Greece, 2013.
- Jonathan M. Fishbein and Chris Eliasmith. **Methods for augmenting semantic models with structural information for text classification**. *30th European conference on Advances in information retrieval* (ECIR'08). Glasgow, UK, 2008.
- Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. **Distributional term representations: an experimental comparison**. *Thirteenth ACM international conference on Information and knowledge management* (CIKM '04). New York, NY, USA, 2004.
- Magnus Sahlgren and Rickard Cöster. **Using bag-of-concepts to improve the performance of support vector machines in text categorization**. *20th international conference on Computational Linguistics* (COLING '04). Stroudsburg, PA, USA, 2004.
- Dou Shen, Jianmin Wu, Bin Cao, Jian-Tao Sun, Qiang Yang, Zheng Chen, and Ying Li. **Exploiting term relationship to boost text classification**. *18th ACM conference on Information and knowledge management* (CIKM '09). New York, NY, USA, 2009.
- Peter D. Turney and Patrick Pantel. **From Frequency to Meaning: Vector Space Models of Semantics**. *Journal of Artificial Intelligence Research,* vol. 37, 2010.

Novel represensations and methods in text classification

CONCISE SEMANTIC ANALYSIS

Outline

- Bag of concepts (again!)
- Concise semantic analysis
- Concise semantic analysis for author profiling
- Meta-features for authorship attribution
- Other bag-of-concept approaches

Bag of concepts (again)

- Under the bag-of-concepts formulation a document is represented as a vector in the space of concepts

- Concepts can be defined/extracted in different ways

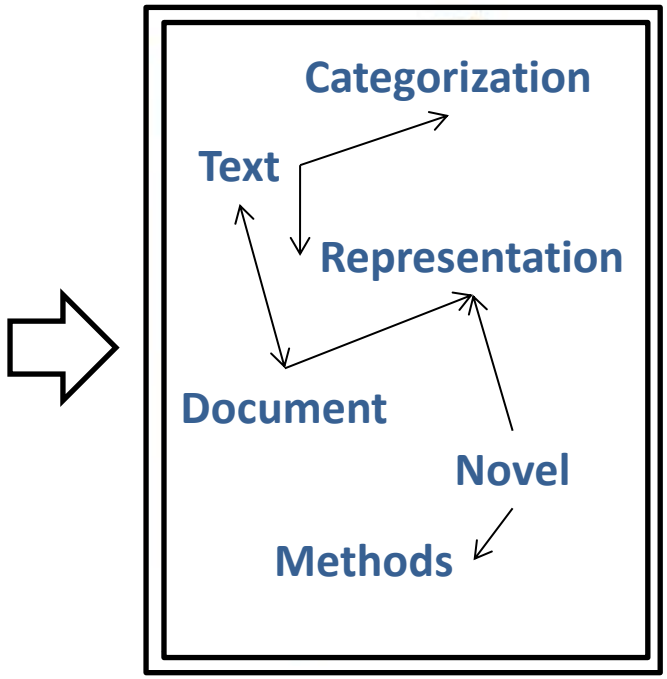
Novel representations and methods in text classification

Manuel Montes-y-Gómez & Hugo Jair Escalante

Two core components of any classification system are the adopted representation for documents and the classification model itself. This tutorial deals with recent advances and developments on both components. The default representation for documents in text classification is the bag-of-words(BOW), where weighting schemes similar to those used in information retrieval are adopted.Whereas this representation has proven to be very helpful for thematic text classification, in novel, non-thematic text classification problems (e.g., authorship attribution, sentiment analysis and opinion mining, etc.), the standard BOW can be outperformed by other advanced representations.

This course is focused on three document representations that have proved to be useful for capturing more information than the raw occurrence of terms in documents as in BOW. The considered representations are: locally weighted BOW, distributional term representations,concise representations and graph-based representations. Likewise, the tutorial covers recent developments in the task of building classification models. Specifically, we consider contextual classification techniques and full model selection methods. The former approach is focused in the design of classifiers that consider the neighborhood of a document for making better predictions.

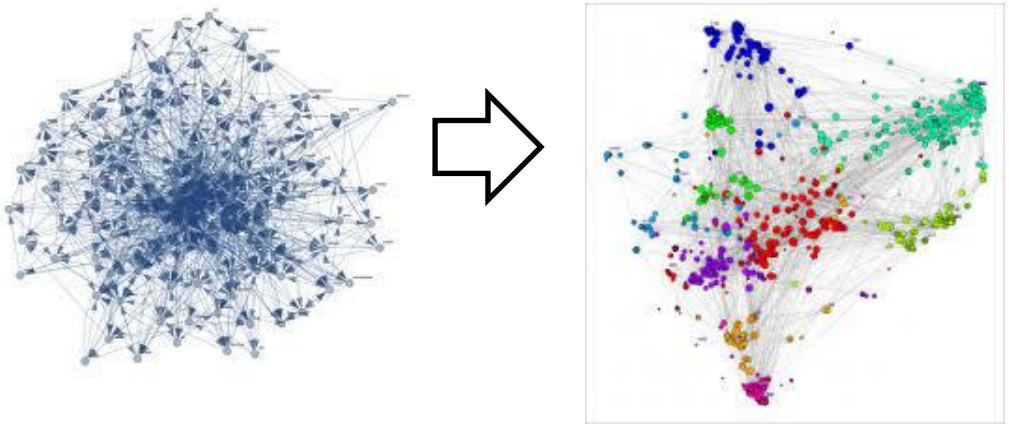
The latter formulation focuses in the development of automatic methods for building classification systems, that is, black box tools that receive as input a data set and return a very effective classification model.



Bag of concepts (again)

- So far we have seen representations that are unsupervised: *techniques proposed for other tasks than classification and that do not take into account information of labeled examples*

- Can we define concepts that take into account information from labeled documents?



Bag of concepts (again)

- Supervised bag-of-concepts: encode inter-class and/or intra-class information into concepts

- A simple (yet very effective) approach for building concepts/features in a supervised fashion:
  - **Concise semantic analysis:** *associate a concept with a class*

Concise semantic analysis

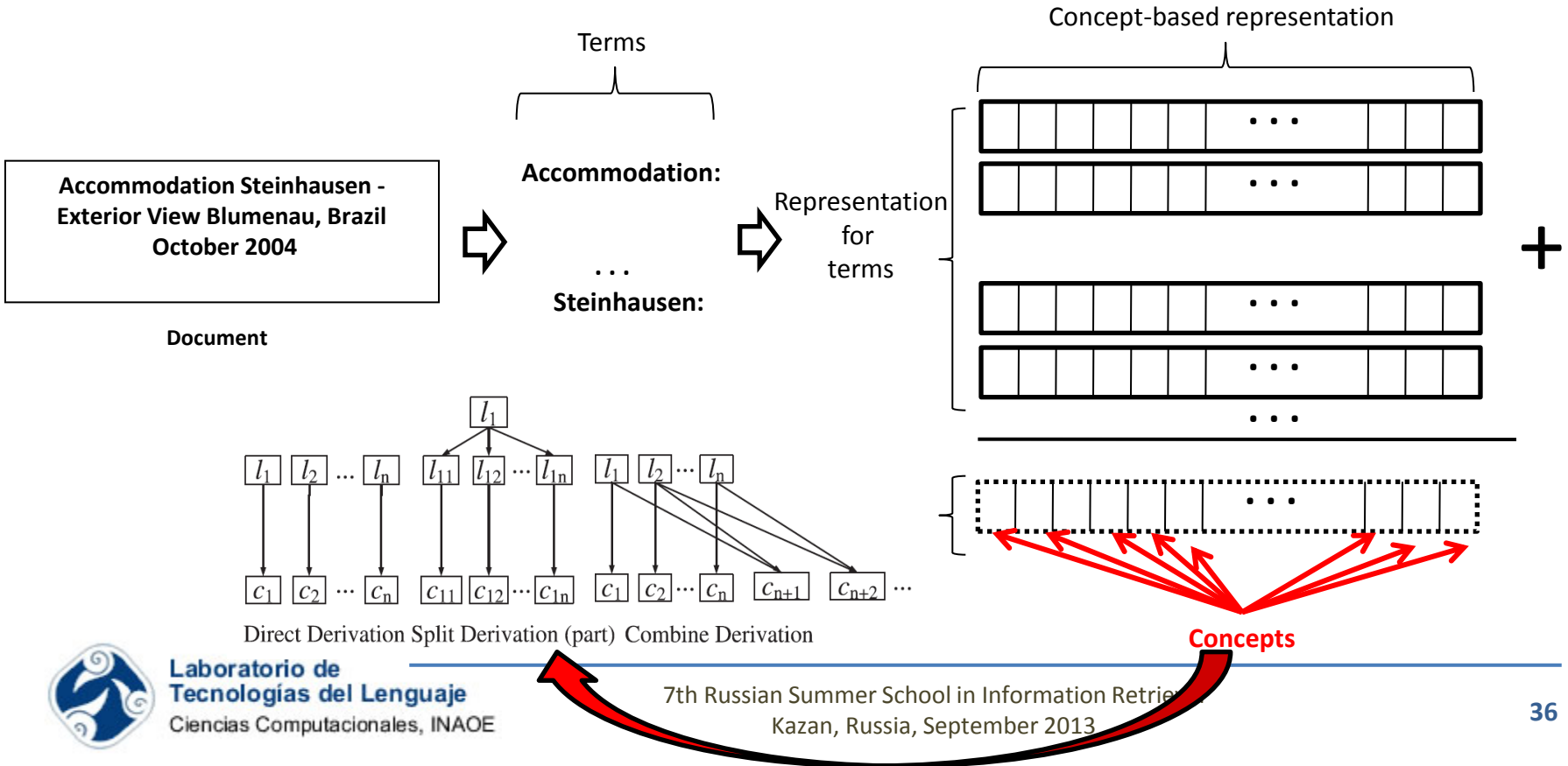
- **Underlying idea**: to associate a concept with each class of the categorization problem.

- After all, in text categorization classes are usually tied with *words/topics/concepts*.
  - E.g., when classifying news into thematic classes: *politics, religion, sports, etc.*

- Implicitly CSA reduces dimensionality, sparseness and incorporate document-term and term-class relationships

Concise semantic analysis

- Two stage process (as before, e.g., with DOR/TCOR)
  - Represent a term in the space of concepts
  - Represent documents in the space of concepts



Concise semantic analysis

- Stage 1: Represent a term in the space of concepts

$$w(t_i, C_j) = \sum_k H(d_k, C_i) \frac{\log(1 + tf(d_k, t_i))}{\log(1 + length(d_k))}$$

$$H(d_k, C_i) = \begin{cases} 1 & d_k \quad belongs - to \quad C_i \\ 0 & - otherwise - \end{cases}$$

term-concept matrix

	C_1	C_1	...	C_k
t_1				
t_2				
:		$w_{i,j}$		
$t_{\|V\|}$				

$w(t_i, C_j)$

Intuition: $w(t_i, C_j)$ Measures the association of term t_i with class C_j

Concise semantic analysis

- Stage 2: Represent documents in the space of concepts:

$$w_d(d_k, C_j) = \sum_{t_j \in d_k} w(t_j, C_i) \times tfrf(t_j, d_k)$$

$$tfrf(t_j, d_k) = tf(d_k, t_j) \times \log\left(2 + \frac{a}{c}\right)$$

a: # Positive documents that contain t_j

c: # Documents that contain t_j

document-concept matrix

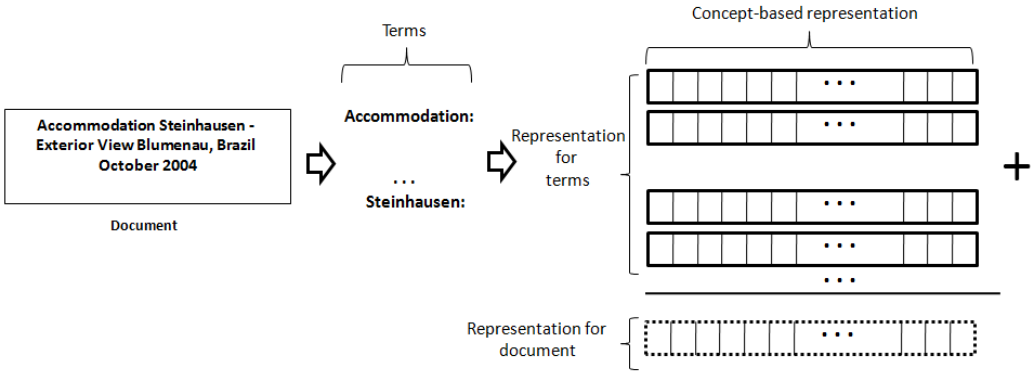
	C_1	C_1	...	C_k
d_1				
d_2				
:		$w_{d(i,j)}$		
d_N				

$w(t_i, C_j)$

Intuition: $w_d(d_k, C_j)$ **Measures the association of terms in the document with class C_j**

Concise semantic analysis

- CSA for text categorization:



Corpora	Representation	SVM		KNN	
		Mi-F_1	Ma-F_1	Mi-F_1	Ma-F_1
Reuters Top10	BOW	0.9272	0.8900	0.8404	0.8259
	CSA(DD)	0.9267	0.8457	0.9280	0.8510
	CSA(CD)[a]	0.9250	0.8380	0.9292	0.8536
20-NG	BOW	0.8081	0.8030	0.6913	0.6900
	CSA(DD)	0.8390	0.8335	0.8381	0.8327
Tancorp-12	BOW	0.9483	0.9172	0.9035	0.8478
	CSA(DD)	0.9294	0.8760	0.9324	0.8893
	CSA(SD)	0.9388	0.8988	0.9348	0.8961
Tancorp-60	BOW	0.7782	0.7493	0.7848	0.7001
	CSA(DD)	0.8345	0.7537	0.8322	0.7693

A variant of CSA for Author Profiling

- The Author Proling (AP) task consists in knowing as much as possible about an unknown author, just by analyzing a given text
  - How much can we conclude about the author of a text simply by analyzing it?

- Applications include business intelligence, computer forensics and security

- Unlike Authorship Attribution (AA), on the problem of AP we does not have a set of potential candidates. Instead of that, the idea is to exploit more general observations (socio linguistic) of groups of people talking or writing

A variant of CSA for Author Profiling

- Initially some works in AP have started to explore the problem of detecting gender, age, native language, and personality from written texts

- AP can be approached as a single-label multiclass classification problem, where profiles are the classes to discriminate

- From the point of view of text classification, we would have a set of training documents, labeled according to a category (e.g., *man and woman*).

A variant of CSA for Author Profiling

- Stage 1: Represent a term in the space of concepts (one concept per profile)

$$w(t_i, C_j) = \sum_k H(d_k, C_i) \frac{\log(1 + tf(d_k, t_i))}{\log(1 + length(d_k))}$$

$$\overline{w}(t_i, C_j) = \sum_k H(d_k, C_i) \log\left(1 + \frac{tf(d_k, t_i)}{length(d_k)}\right)$$

term-concept matrix

	C_1	C_1	...	C_k
t_1				
t_2				
:		$w_{i,j}$		
$t_{\|V\|}$				

$\overline{w}(t_i, C_j)$

Intuition: $\overline{w}(t_i, C_j)$ Measures the association of term t_i with class C_j

A variant of CSA for Author Profiling

- Stage 2: Represent documents in the space of concepts (one concept per profile):

$$w_d(d_k, C_j) = \sum_{t_j \in d_k} w(t_j, C_i) \times tfrf(t_j, d_k)$$

$$\overline{w}_d(d_k, C_j) = \sum_{t_j \in d_k} \frac{w(t_j, C_i)}{length(d_k)} \times tf(d_k, t_j)$$

document-concept matrix

	C_1	C_1	...	C_k
d_1				
d_2				
:		$w_{d(i,j)}$		
d_N				

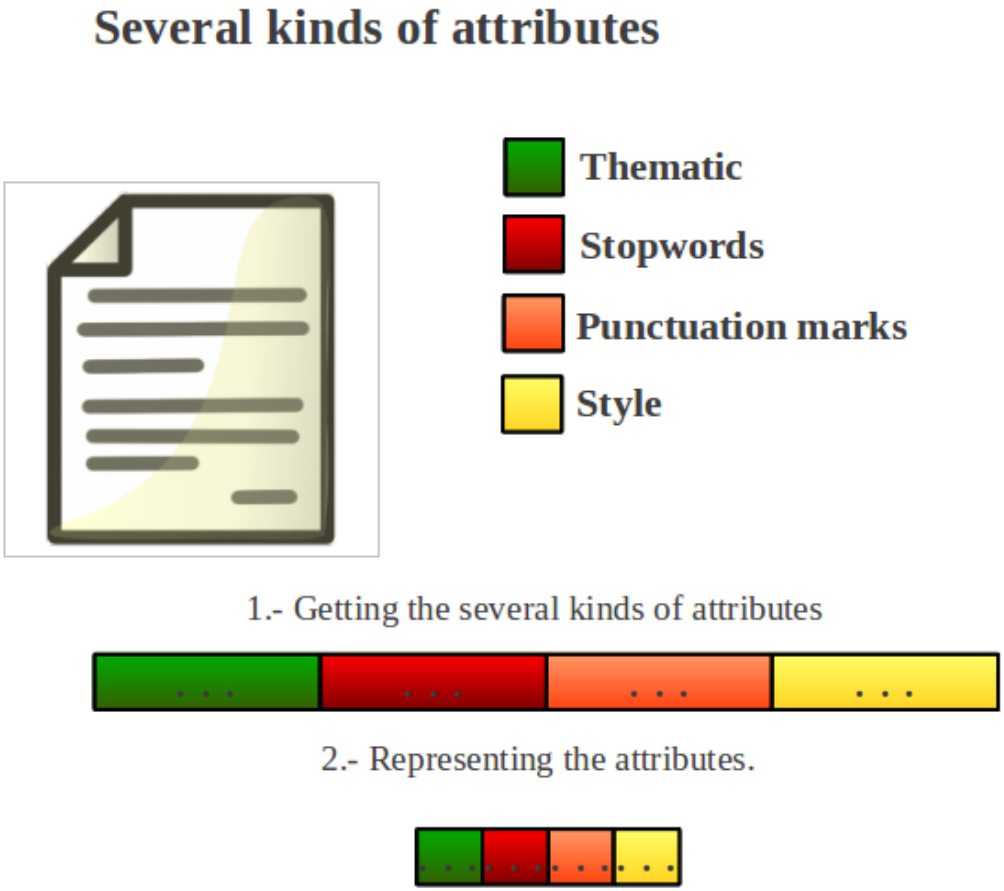
$\overline{w}(t_i, C_j)$

Intuition: $\overline{w}_d(d_k, C_j)$ **Measures the association of terms in the document with class C_j**

A variant of CSA for Author Profiling

- AP is not a thematic task and therefore the use of simple word occurrences may not work. Thus we also considered non-thematic attributes as well

Concepts were generated for each modality separately and we concatenated the space of multimodal concepts



PAN13's Author Profiling task

- The proposed representation was evaluated in the PAN AP 2013, track:
  - Two corpora English and Spanish
  - To determine the gender (male or female) and age (13-17, 23-27, 33-47) for each document.

	Instances	Blogs	Vocabulary
English	236,000	413, 564	180, 809, 187
Spanish	75,900	125,453	21,824,190

- A linear SVM classifier was used for classification

CSA in AP at PAN-13

- Official results:

Submission	Accuracy			Adult			Predator			Runtime
	Total	Gender	Age	Gender	Age	Both	Gender	Age	Both	(incl. Spanish)
meina13	0.3894	0.5921	0.6491	6	8	6	72	41	41	383821541
pastor13	0.3813	0.5690	0.6572	1	8	0	72	32	32	2298561
mechti13	0.3677	0.5816	0.5897	2	6	2	52	29	20	1018000000
santosh13	0.3508	0.5652	0.6408	9	9	9	69	32	29	17511633
yong13	0.3488	0.5671	0.6098	6	1	1	28	30	17	577144695
ladra13	0.3420	0.5608	0.6118	9	9	9	72	33	33	1729618
ayala13	0.3292	0.5522	0.5923	3	2	1	53	34	26	23612726
gillam13	0.3268	0.5410	0.6031	1	4	0	72	30	30	615347
kern13	0.3115	0.5267	0.5690	9	9	9	47	35	25	18285830
haro13	0.3114	0.5456	0.5966	0	8	0	69	44	41	9559554
aditya13	0.2843	0.5000	0.6055	0	0	0	72	40	40	3734665
hidalgo13	0.2840	0.5000	0.5679	0	0	0	72	40	40	3241899
farias13	0.2816	0.5671	0.5061	4	2	1	55	34	26	24558035
jankowska13	0.2814	0.5381	0.4738	1	0	0	72	44	44	16761536
flekova13	0.2785	0.5343	0.5287	4	4	4	61	39	34	18476373
weren13	0.2564	0.5044	0.5099	1	0	0	71	40	39	11684955
ramirez13	0.2471	0.4781	0.5415	9	0	0	12	40	9	64350734
jimenez13	0.2450	0.4998	0.4885	6	2	1	27	31	14	3940310
moreau13	0.2395	0.4941	0.4824	4	4	2	33	39	19	448406705
baseline	0.1650	0.5000	0.3333	–	–	–	–	–	–	–
patra13	0.1574	0.5683	0.2895	5	4	1	55	17	12	22914419
cagnina13	0.0741	0.5040	0.1234	4	7	4	24	9	8	855252000

CSA in AP at PAN-13

- Official results:

Submission	Accuracy			Runtime
	Total	Gender	Age	(incl. English)
santosh13	0.4208	0.6473	0.6430	17511633
pastor13	0.4158	0.6299	0.6558	2298561
haro13	0.3897	0.6165	0.6219	9559554
flekova13	0.3683	0.6103	0.5966	18476373
ladra13	0.3523	0.6138	0.5727	1729618
jimenez13	0.3145	0.5627	0.5429	3940310
kern13	0.3134	0.5706	0.5375	18285830
yong13	0.3120	0.5468	0.5705	577144695
ramirez13	0.2934	0.5116	0.5651	64350734
aditya13	0.2824	0.5000	0.5643	3734665
jankowska13	0.2592	0.5846	0.4276	16761536
meina13	0.2549	0.5287	0.4930	383821541
gillam13	0.2543	0.4784	0.5377	615347
moreau13	0.2539	0.4967	0.5049	448406705
weren13	0.2463	0.5362	0.4615	11684955
cagnina13	0.2339	0.5516	0.4148	855252000
hidalgo13	0.2000	0.5000	0.4000	3241899
farias13	0.1757	0.4982	0.3554	24558035
baseline	0.1650	0.5000	0.3333	–
ayala13	0.1638	0.5526	0.2915	23612726
mechti13	0.0287	0.5455	0.0512	1018000000

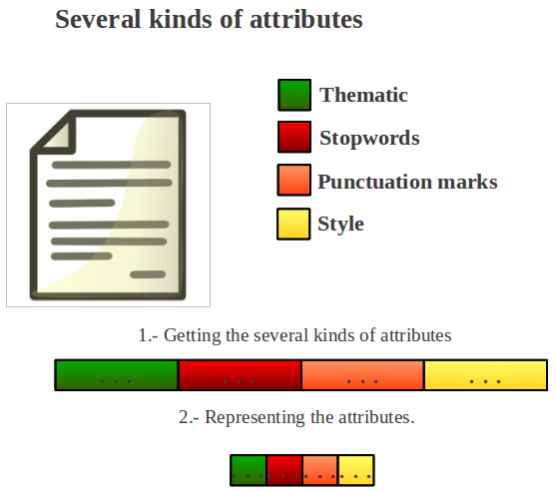
Conclusions: CSA for AP

- The proposed representation obtained the best result for the PAN'13-AP track (avg. performance was evaluated)
- Besides our proposal is much more efficient than most formulations (it was less efficient than two other approaches)
- Too much potential on the use of CSA for non-thematic tasks
- There are other ways of expanding the concept space to encode useful information

Multimodal concepts

- Multimodal information resulted very effective for AP

- Are there other ways of encoding multimodal information into the document representation?
  - **Basic approaches:** late fusion / early fusion of multimodal attributes
  - **Meta-features based concepts:** associate a feature/concept with the similarity to multimodal prototypes



Multimodal meta-features

- **Idea:** to extract features from first-level attributes to derive a multimodal representation that can improve the performance of first-level attributes
  - **First level attributes:** attributes/representations obtained from text directly (e.g., BoW, ngrams, POS-based features, stylistic)

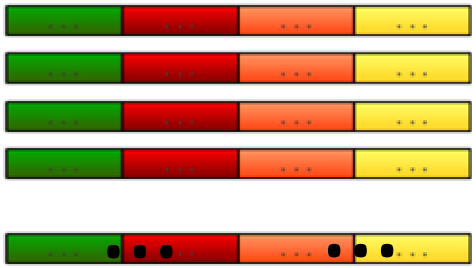
- Process:
  - Extract first-level attributes
  - Obtain multimodal prototypes
  - Meta-features: Estimate the similarity of documents to each prototype

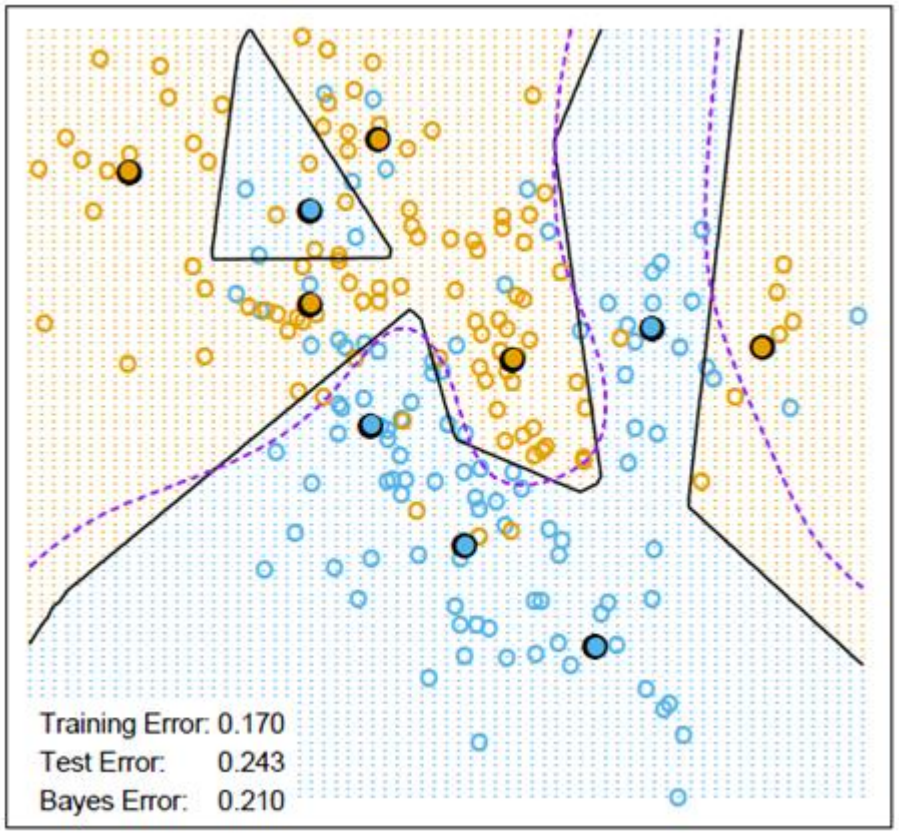
Multimodal meta-features

- First level attributes are clusted (e.g. via k-means)

For each modality k-clusters are generated

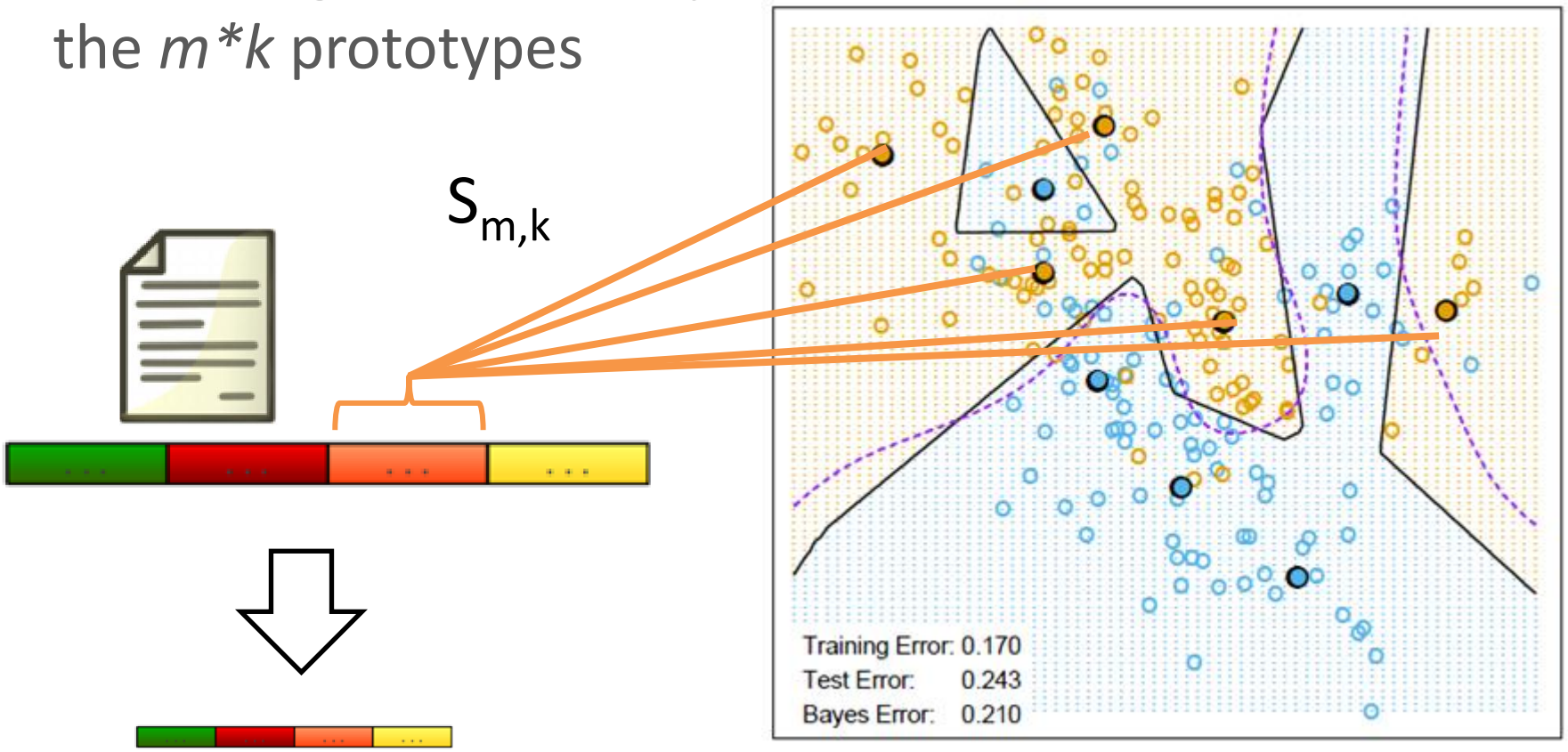
The centers of each cluster define multimodal prototypes



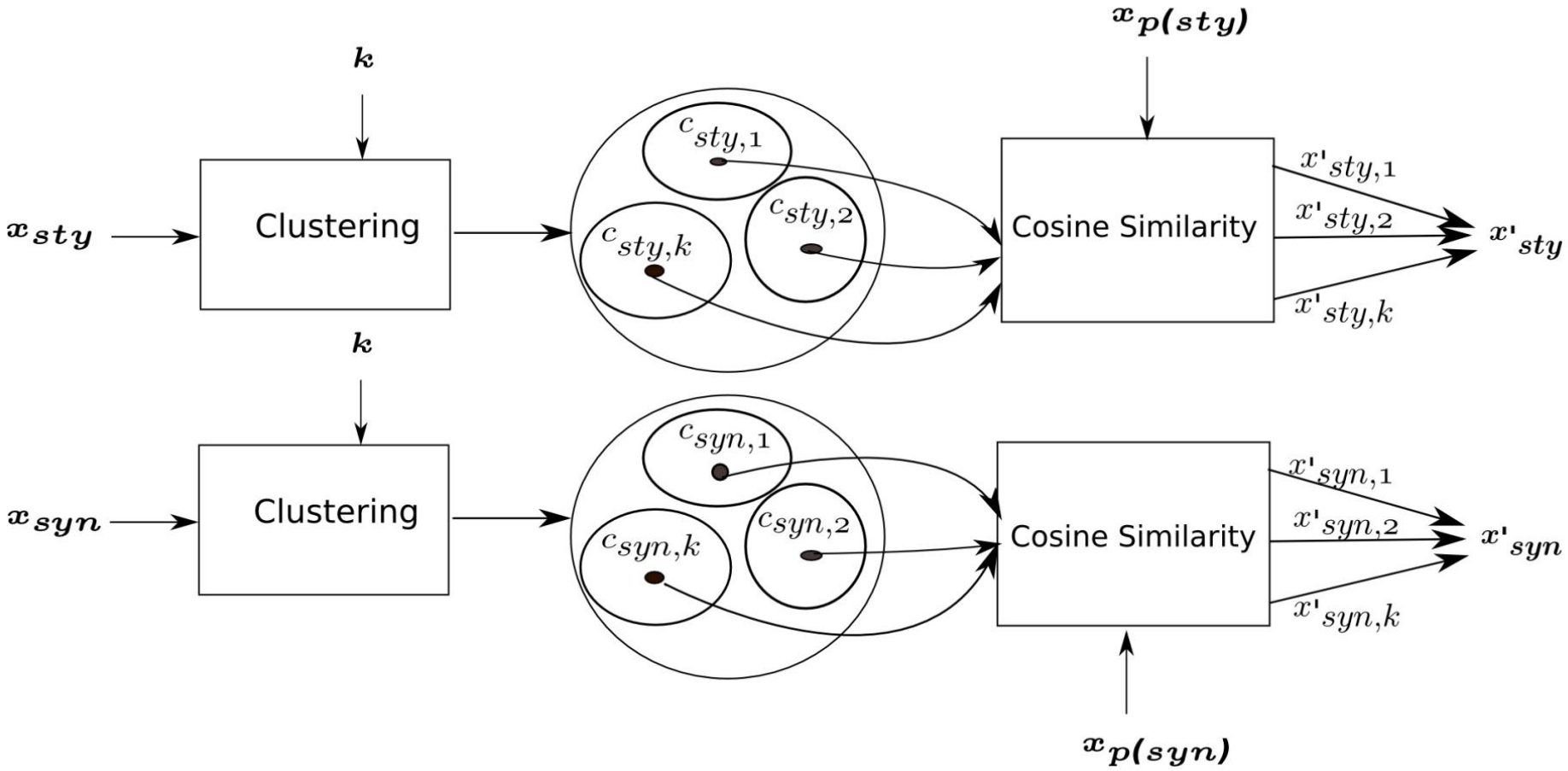


Multimodal meta-features

- Meta-features for a document are derived by estimating the similarity of the document to each of the *m*k* prototypes



Multimodal meta-features



Multimodal meta-features

- Application to authorship attribution. First level features:

Modality	First Level Features (FLF)
Stylistic	Total number of sentences^{+}
	Average number of tokens per sentence
	Percentage of words without vowel^{+}
	Average number of punctuations per sentence
	Percentage of contractions
	Total number of balanced parenthesis^{+}
	Percentage of two consecutive punctuation marks
	Percentage of three consecutive punctuation marks
	Total number of alphabetic characters
	Average number of tokens with at least a capitalized letter per sentence^{+}
	Toal number of sentence initial words with first letter capitalized
	Total number of quotations*
Syntactic	Top 1000 POS tag unigrams
	Top 1000 POS tag bigrams
	Top 1000 POS tag trigrams
	Top 1000 Grammatical relations from the dependency parses
Semantic	Top 1000 bag-of-words
Perplexity	All the perplexity values from character 4-grams*

Multimodal meta-features

- Some experimental results

dataset	#auth	MSMF+FLF (5fcv)	Benchmark Comparison
CHE	5	74.30 (79.00)	**75.47** [19]
CHE	10	**77.96** (76.07)	77.38 [19]
CHE	20	**72.48** (71.79)	71.42 [19]
CHE	50	**67.00** (65.09)	63.79 [19]
CHE	100	**63.61** (63.50)	62.10 [19]
Football	3	91.11 (92.75)	**93.34** CNG-WPI [5] 91.11 PCFG-E [17]
Business	6	86.66 (86.29)	**91.11** PCFG-E [17] 80.00 CNG-WPI [5]
Travel	4	90.00 (86.70)	**91.67** PCFG-E [17] 73.33 CNG-WPI [5]
Cricket	4	91.11 (95.59)	**95.00** PCFG-E [17] 90.00 CNG-WPI [5]
Poetry	6	63.63 (78.29)	**87.27** PCFG-E [17] 85.45 CNG-WPI [5]
CCAT	10	78.80 (84.20)	**86.4** BOLH Diffusion Kernel [6] 79.40 Char n-grams SVM [21] 78.00 STM-Asymmetric cross [16] 73.60 CNG-WPI [5]
CCAT	50	69.48 (76.12)	**74.04** Char n-grams SVM [8]

Multimodal meta-features

Dataset	#Author	Feature Set	Modality			
			Semantic	*Perplexity*	*Syntactic*	*Stylistic*
CHE	5	MSMF	38.02	23.95	34.89	54.86
		FLF	45.86	16.36	36.42	59.44
		MSMF+FLF	46.40	41.09	36.87	**65.11**
CHE	10	MSMF	30.75	40.73	23.02	60.70
		FLF	45.86	16.36	36.42	60.70
		MSMF+FLF	46.40	40.64	36.87	**65.10**
CHE	20	MSMF	31.46	14.73	22.17	56.05
		FLF	39.01	14.06	30.13	52.92
		MSMF+FLF	39.83	14.88	29.67	**60.42**
CHE	50	MSMF	30.50	15.57	19.68	51.45
		FLF	35.36	15.34	25.38	45.88
		MSMF+FLF	37.07	15.54	25.35	**54.04**
CHE	100	MSMF	31.21	14.84	20.72	50.93
		FLF	32.46	14.84	23.70	45.27
		MSMF+FLF	32.87	15.02	24.20	**52.09**

Football	3	MSMF	80.00	75.55	60.00	44.44
		FLF	86.66	**91.11**	82.22	64.44
		MSMF+FLF	86.66	77.77	82.22	73.33
Business	6	MSMF	73.33	77.77	40.00	32.22
		FLF	80.00	63.33	73.33	57.77
		MSMF+FLF	80.00	**83.33**	73.33	53.33
Travel	4	MSMF	80.00	**86.66**	36.66	35.00
		FLF	76.66	76.66	81.66	43.33
		MSMF+FLF	76.66	85.00	81.66	46.66
Cricket	4	MSMF	73.33	91.66	58.33	63.33
		FLF	80.00	61.66	90.00	66.66
		MSMF+FLF	80.00	**91.66**	91.66	80.00
Poetry	6	MSMF	40.00	**80.00**	27.27	18.18
		FLF	34.54	52.72	40.00	18.18
		MSMF+FLF	34.54	78.18	43.63	20.00

CCAT	10	MSMF	68.40	68.60	28.80	24.60
		FLF	74.80	51.00	74.00	33.20
		MSMF+FLF	**76.00**	69.60	73.60	31.80
CCAT	50	MSMF	57.92	56.92	15.96	13.28
		FLF	62.76	34.00	55.20	10.96
		MSMF+FLF	**66.08**	57.56	55.36	14.60

Conclusions Metafeatures for AA

- Multimodal and similarity-based metafeatures aim to represent a document by its similarity with documents of the same and different classes
- Metafeatures (combined with first-level features) resulted very effective for AA
- Higher improvements were observed when more authors are involved in the problem
- Metafeatures can be considered a type of concepts

Relation with other approaches

- **Latent semantic indexing:** Concepts are derived via SVD, concepts are the *principal components* of the term-document matrix

- **Topic models:** Concepts are probability distributions over words, they can be obtained in different ways (pLSI, LDA, etc.)

- **Deep learning:** Concepts are the outputs of hierarchical neural networks that aimed to reconstruct documents

Final remarks

- Concept-based representations encode useful information into the representation of terms/documents

- This useful information can be defined in different ways, depending on what we want to emphasize or characterize from documents
  - Term/document occurrence/co-occurrence
  - Inter-intra class information
  - Multimodal similarity
  - ….