Analyzing and mining multimedia collections

How (can) statistics help?

Guillaume Gravier

UMR IRISA

Multimedia data: what and what for?

Multimedia data = all kind of digital data related to media, i.e., intended for communication of a(n informative) message towards human

A wide **variety of sources**

→ TV, radio, newspapers, web media; photo agencies; video/photo sharing sites; social media; etc.

A wide **variety of material**

→ texts, posts and tweets; audio, speech and music; images; videos; etc.

A wide **variety of application domains**

→ media asset management; Internet portals; connected TV; on-line courses; etc.

Multimedia data are...

1. **semantic**

   → created by humans for humans
   → make sense only in context
   → variability, heterogeneity and multimodality

   ⇒ *robust and versatile machine interpretation*

2. **very very large scale**

   → 100 h of videos / min on YouTube!
   → 58 billion tweets / day

   ⇒ *efficient processing and data organization*

3. **unstructured and disconnected**

   → implicit structure not known ⇒ limited interpretation capabilities
   → lack of organization and links within collections

   ⇒ *collection structuring methodology*

Multimedia content processing: the big data era

From content description...

- all-purpose descriptors for texts, images, sounds and videos
  → lemma/stem, MFCC, SIFT, VLAD, Fisher vector, . . .
- supervised machine learning to detect (semantic) concepts
  → SVM, HMM, CRF, MKL, DNN, . . .
- information retrieval to search and rank documents at scale

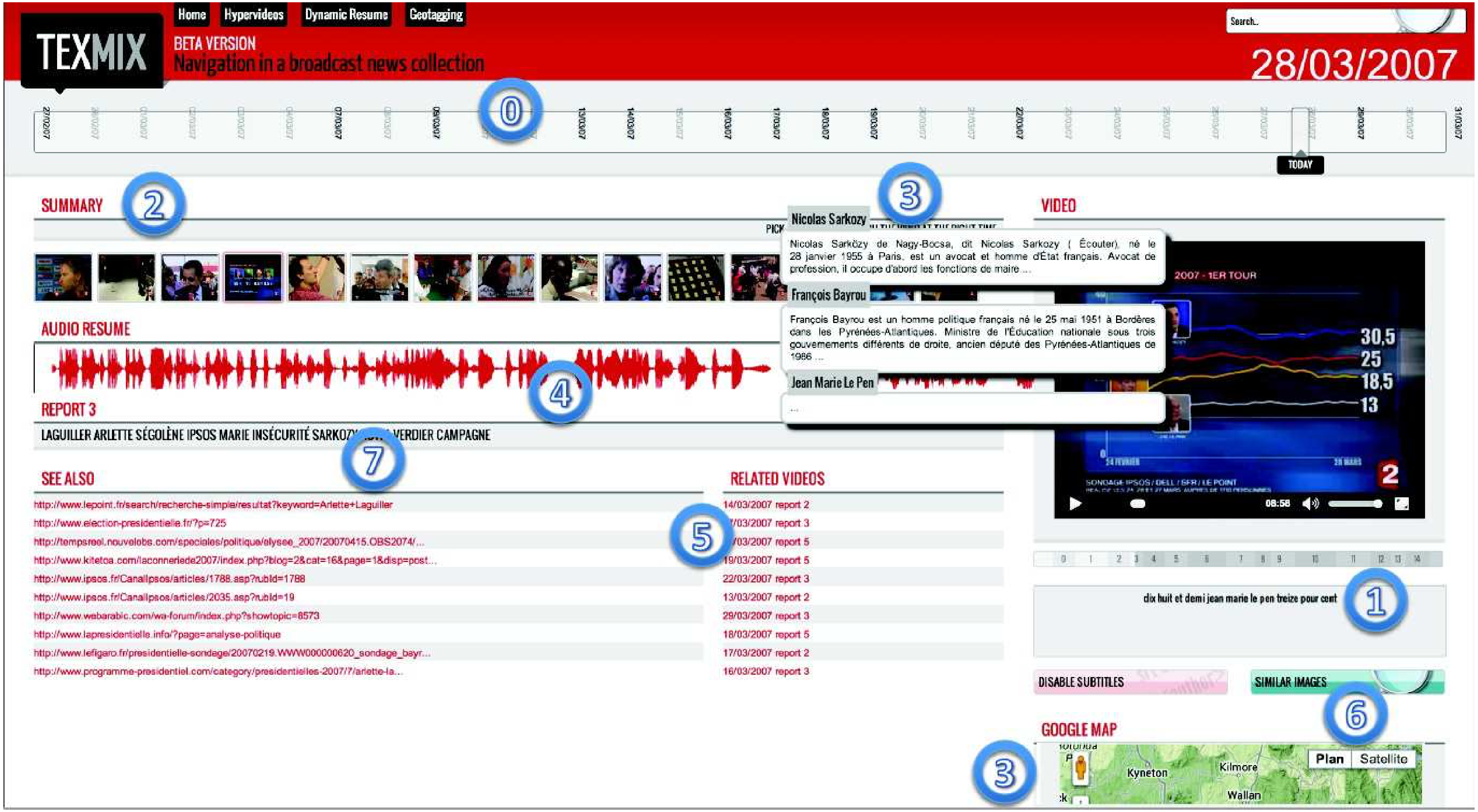
... to analytics

Data analytics can be defined as the application of [] data processing techniques to discover patterns, extract knowledge and gain insights from large-scale, typically multi-source data collections that may contain structured, unstructured and semi-structured data. — excerpt from BDVA Strategic Research and Innovation Agenda

Current limits and challenges

- semantic interpretation still far from perfect
  - → multimodality, structured I/O, confidence, annotation, etc.
- multimedia pattern mining at scale is lacking
  - → object discovery, audio pattern discovery, etc.
- multimedia data collections are unstructured
  - → multimedia data warehouses, graph structures, etc.
- usage and expectations are still unclear
  - → practical use of technology, acceptability, evaluation, etc.
- privacy and security are challenged by analytics and data agregation

Texmix: description (and browsing)



extension within industrial FUI project

NEXGEN TV

LIMAH: (description and) browsing

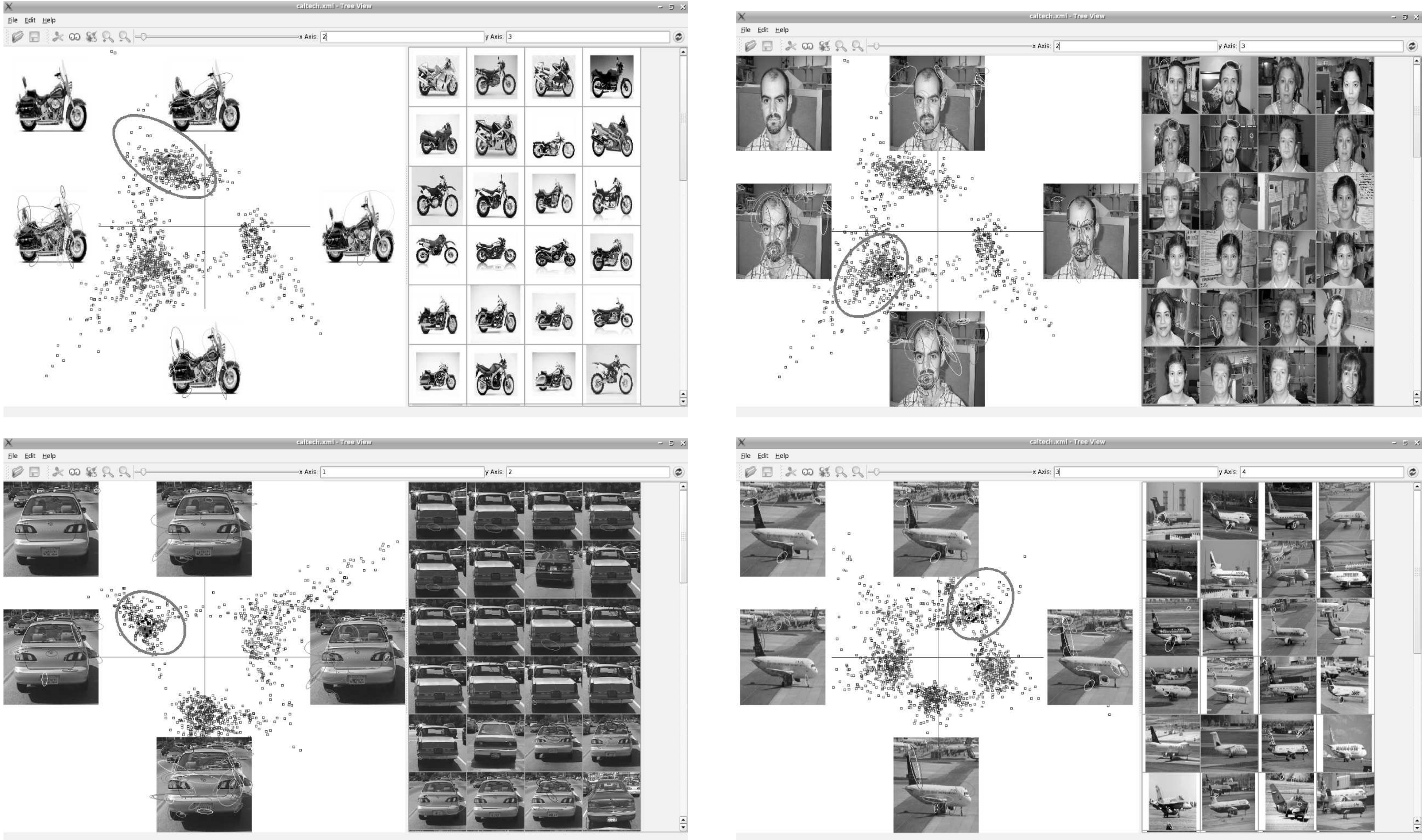
Probabilistic models are (almost) everywhere

- Image indexing and description
  $\rightarrow$ *Fisher vector, image/tag co-occurrence analysis, PCA, CCA, FCA, etc.*
- Speech recognition
  $\rightarrow$ *hidden Markov acoustic models, n-gram language models*
- Keyword extraction, named entity recognition
  $\rightarrow$ *term frequency analysis, conditional random fields*
- Topic segmentation
  $\rightarrow$ *mutual information (IM^3), language model, burst detection, etc.*
- Content matching
  $\rightarrow$ *term frequency, latent Dirichlet allocation, cross correlation analysis, etc.*
- Collection organization, navigation
  $\rightarrow$ *data visualization, graph analysis, etc.*

Deep learning now replacing good old probabilistic models

$\rightarrow$ Is deep learning statistics? If not, can both be combined?

Example: images and factorial correspondance analysis



[Nguyen-Khang Pham *et al.*. Intensive use of factorial correspondence analysis for large scale content-based image retrieval. Advances in Knowledge Discovery and Management, 2010]

Another example: HMM for video structuring

- four generic scenes

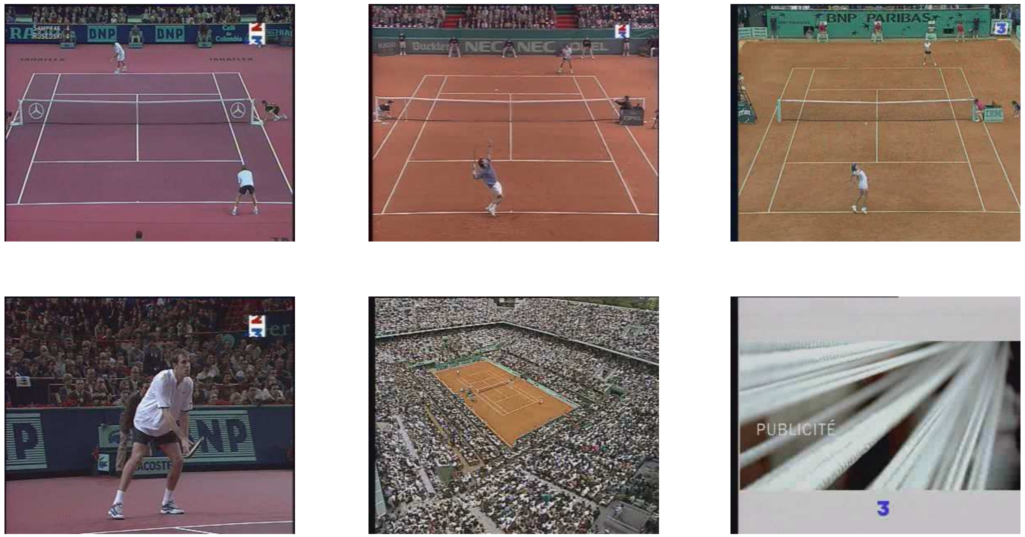
  → missed serve + rallye, rallye, replay, other

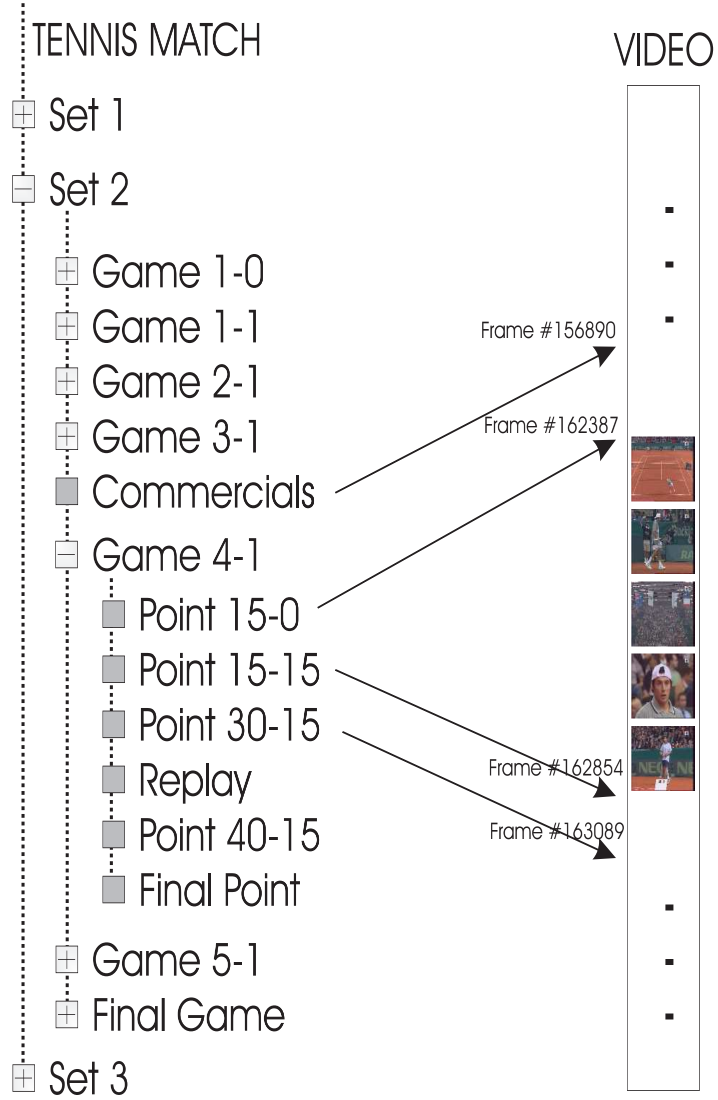
- *prior* knowledge

  → editing rules, tennis rules

- *posterior* knowledge

  → images & soundtrack



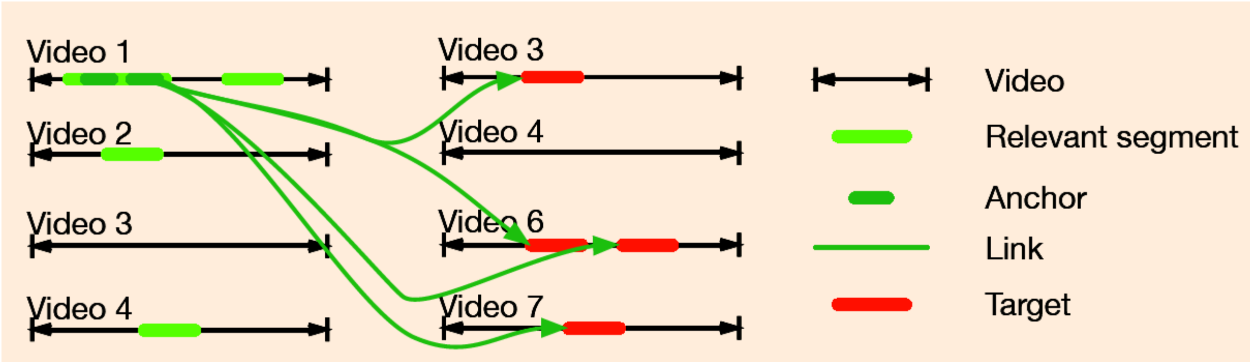


[Ewa Kijak *et al.*. Audiovisual integration for tennis broadcast structuring. Multimedia Tools and Applications, 2006]

What in media hyperlinking?

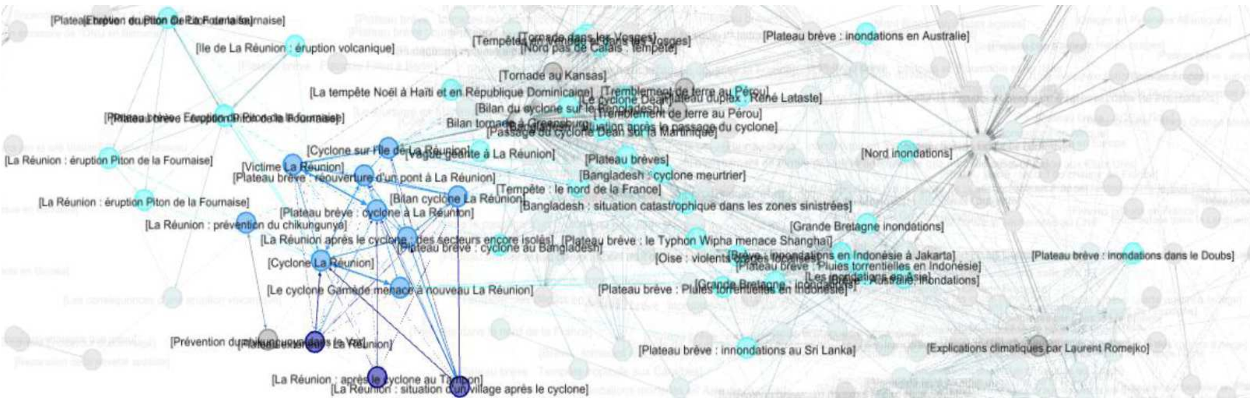
The Mediaeval/TRECVid hyperlinking task scenario:

1. anchor detection: finding potential anchors in the videos
2. fragment linking: creating a ranked list of segments related to the anchors



[M. Eskevich *et al.* Multimedia information seeking through search and hyperlinking. ICMR, 2013]

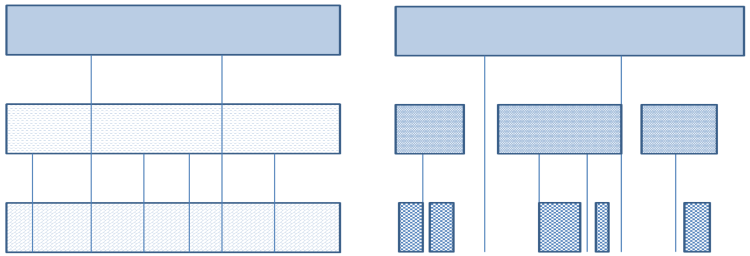
anchoring + hyperlinking = organizing a collection for analytics



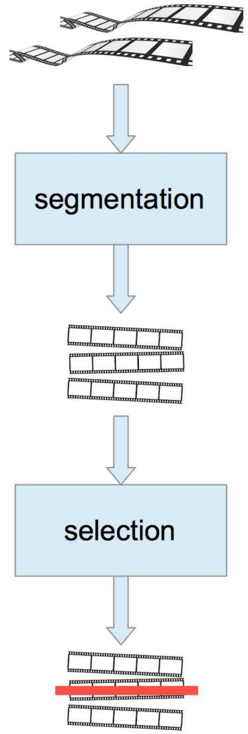
A two step language-based hyperlinking framework

Topic segmentation based on transcripts

- ASR-adapted linear topic segmentation
  [Guinaudeau *et al.*, Mediaeval 12; Şimon *et al.*, EMNLP 13]
- hierarchical topic segmentation
  [Guinaudeau *et al.*, Mediaeval 13; Şimon *et al.*, SLAM 13]
- fragmentation



[Şimon *et al.*, RANLP 15]

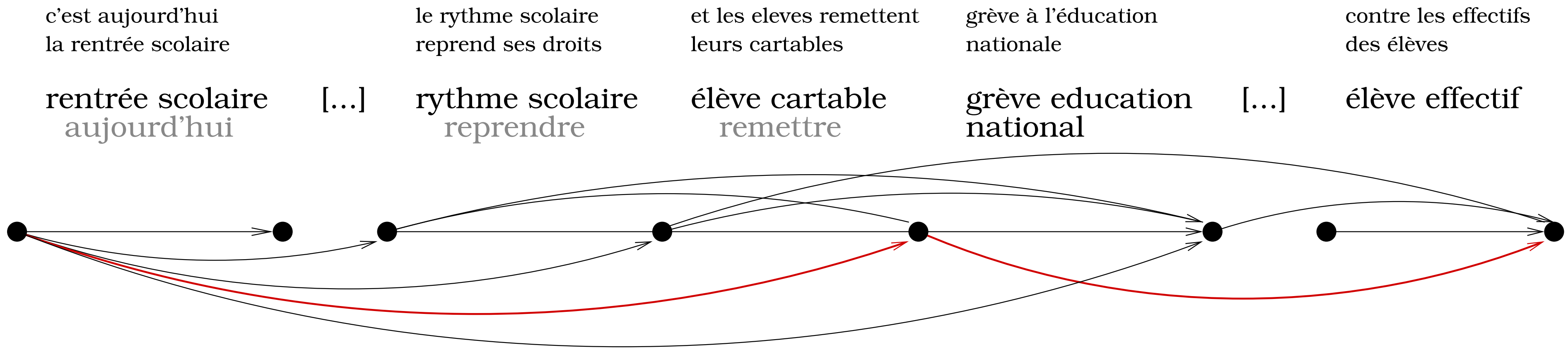


Standard language-based ranking

	reference			ASR		
system	lin+bow	lin+ngram	hie+bow	lin+bow	lin+ngram	hie+bow
P	0.31	0.42	0.26	0.20	0.33	0.19
P_{tol}	0.25	0.41	0.26	0.14	0.30	0.17

Mediaeval 2013 [Şimon *et al.*, SLAM 14]

Graph-based topic segmentation



vertex values = lexical cohesion = generalized probability

$$\hat{s} = \underset{s_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} \ln(P[w_{a_i}^{b_i} | S_i]) - \alpha \ln(n)$$

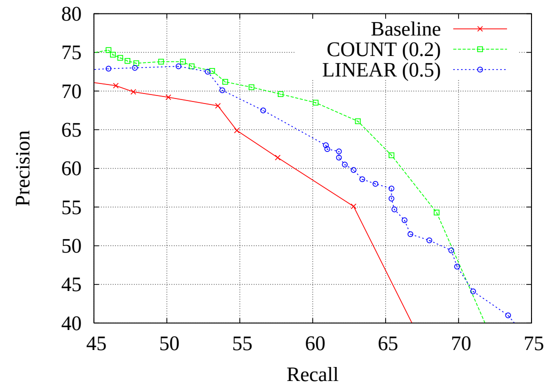
$$\Delta_i = \left\{ P_i(u) = \frac{C_i(u) + 1}{z_i}, \; \forall \, u \in V_K \right\} \qquad \ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_i]$$

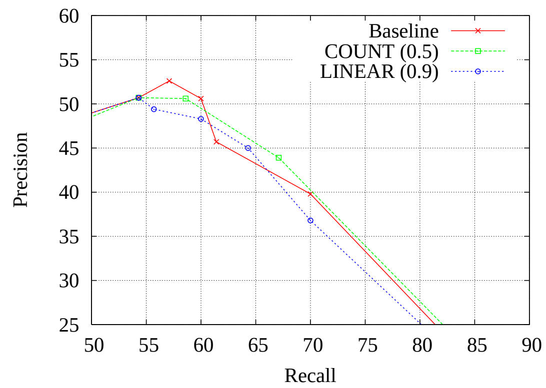
Graph-based topic segmentation (cont'd)

Improving language modeling

- using confidence measures
- adding semantic relations to cohesion

$$C_i''(u) = C_i(u) + \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u)$$





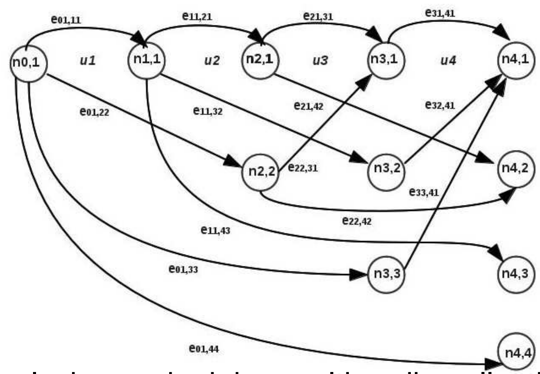
- using interpolated language models

$$\ln P[S_i; S_i, T] = \sum_{j=1}^{n_i} \ln(\lambda P[w_j^i; \Delta_i] + (1-\lambda) P[w_j^i; \Delta_t])$$

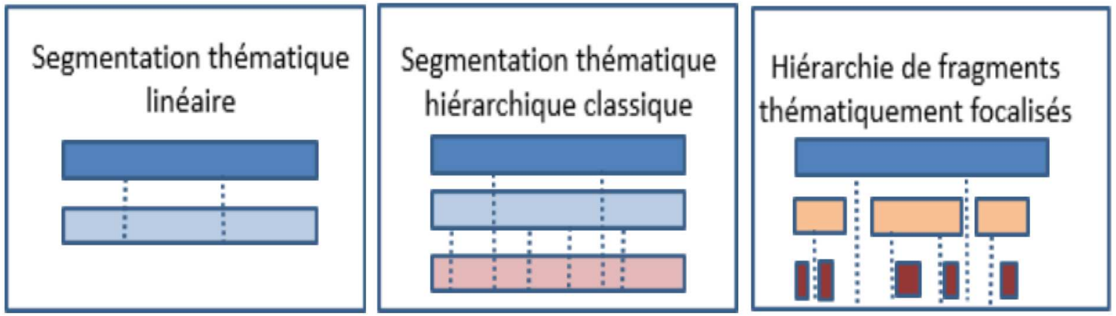
Remove independence assumption

$$P[W|S_1^m] = P[W|S_1] \prod_{i=2}^{m} P[W|S_i, S_{i-1}]$$

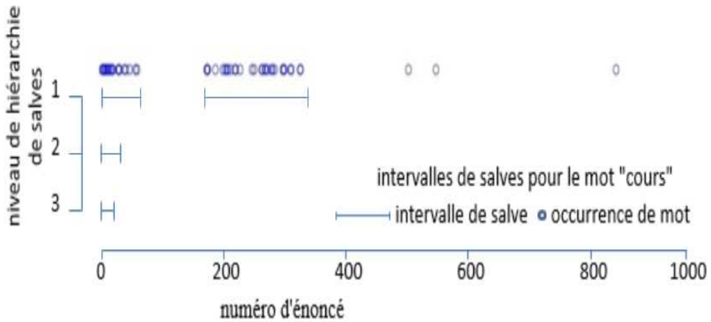
where $\ln P[W|S_i, S_{i-1}] = \ln P[W_i|S_i] - \lambda\Delta(W_i, W_{i-1})$



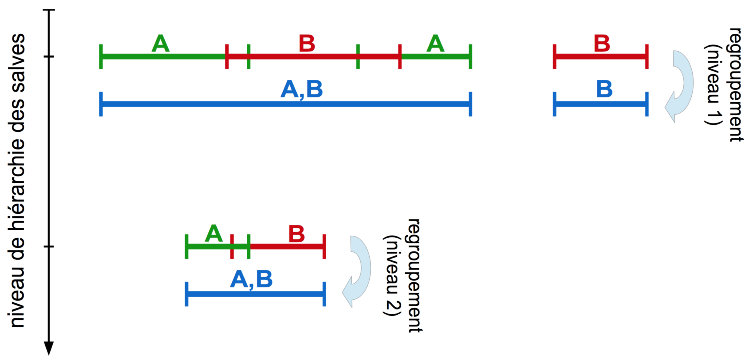
From segmentation to fragmentation

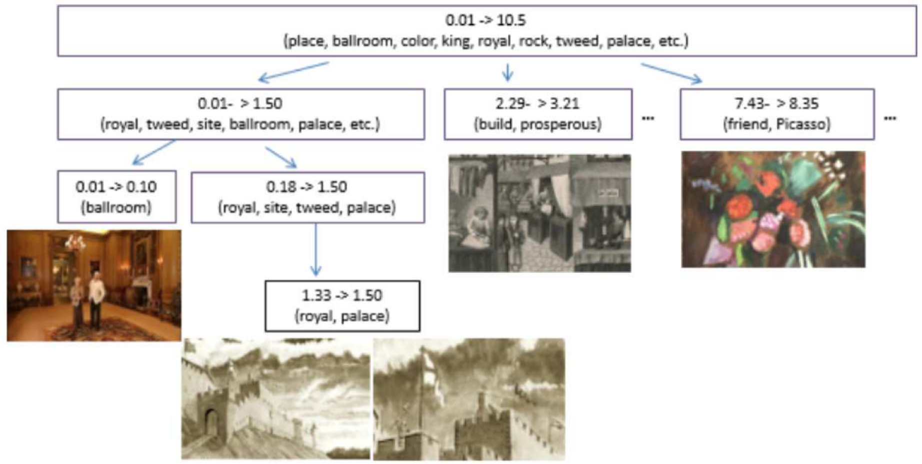


From burst analysis ...



... to topic fragments





UMR IRISA

Topic models for content matching

Principle: Explain documents in a collection as a mixture of K topics, where each word is assigned to a topic.

I eat fish and vegetables.
Fishes are pets.
My kitten eats fish.
[source: wikipedia]

Hierarchy with 10 levels, trained independently on BBC collection transcripts

- level 1, $K_1 = 50$, broad topics z_i^1 $(i \in [1, K_1])$
- level 10, $K_{10} = 1,700$, fine-grain topics z_i^{10}

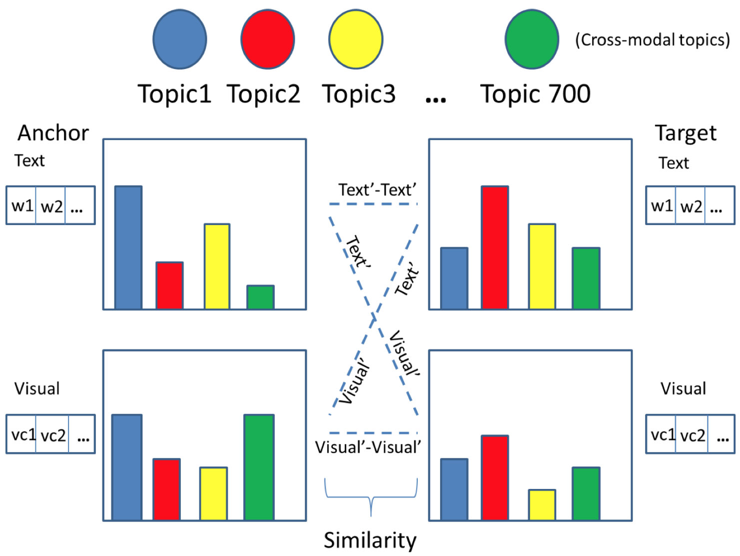
year	number of topics (K)				
	direct	50	150	300	700
2013	0.25	**0.44**	**0.34**	**0.35**	**0.34**
2014	0.19	0.18	**0.25**	**0.26**	0.21

Target reranking task, using targets from all MediaEval participants: relevance after 15 s of the top-10 targets [Şimon *et al.*, SLAM 15]

[combination into hierarchical stuff goes here]

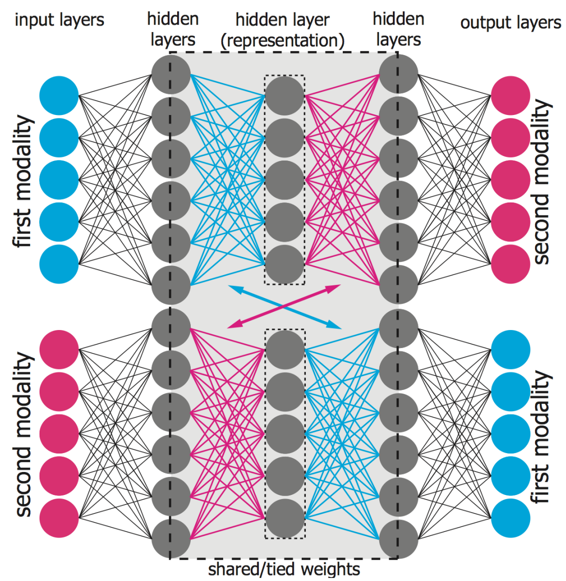
Multimodality to improve diversity

Bimodal extension of LDA mapping words and visual concepts via topic-specific distributions



[Bois *et al.*, TRECVid 15]

Crossmodal matching with symmetrical bi-directionnal auto-encoders



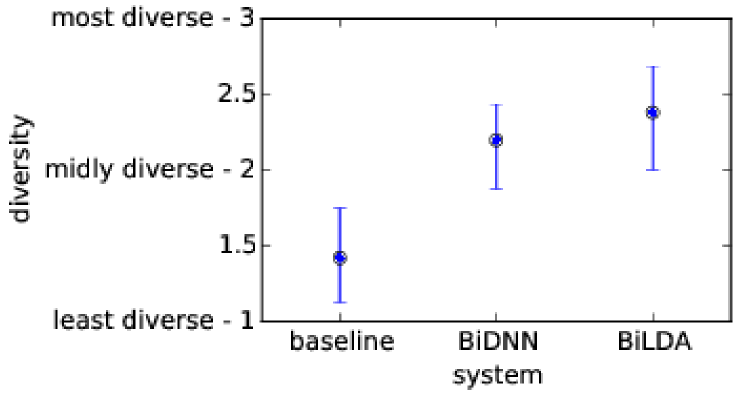
[Vukotić *et al.*, ICMR 16]

Multimodality to improve diversity (cont'd)

Measuring diversity of the top-5 relevant targets found by each approach

	transcripts			concepts		
	n_u	$\overline{d}_a$	$\overline{d}_i$	n_u	$\overline{d}_a$	$\overline{d}_i$
baseline	29.8	0.51	0.61	35.6	0.61	0.71
BiDNN	40.8	0.20	0.12	46.7	0.42	0.31
BiLDA	40.0	0.25	0.16	38.0	0.48	0.41

Intrinstic measures of diversity



Perceived diversity (25 subjects)

[Bois *et al.*, submitted to MMM 17]

Legend:

n_u = number of unique key words/concepts in top-5 targets

$\overline{d}_i$ = average similarity among the top-5 targets

$\overline{d}_a$ = average similarity between anchor and top-5 targets

Acknowledgements

Many thanks to all who contributed to the results presented here (in alphabetical order): Laurent Amsaleg, Rémi Bois, Morgan Bréhinier, Sébastien Campion, Vincent Claveau, Camille Guinaudeau, Ewa Kijak, Sien Moens, Pascale Sébillot, Ronan Sicre, Anca-Roxana Şimon, Arnaud Touboulic, and Vedran Vukotić.

And very likely others that I might have forgotten.

Work presented here benefited from the financial support of: BPI France (Quaero), CominLabs & ANR (LIMAH), EIT ICT Labs (OpenSEM) and Région Bretagne (ARED).