Structuring temporal sparse data with application to opinion mining

Julien Velcin
University of Lyon – ERIC Lab

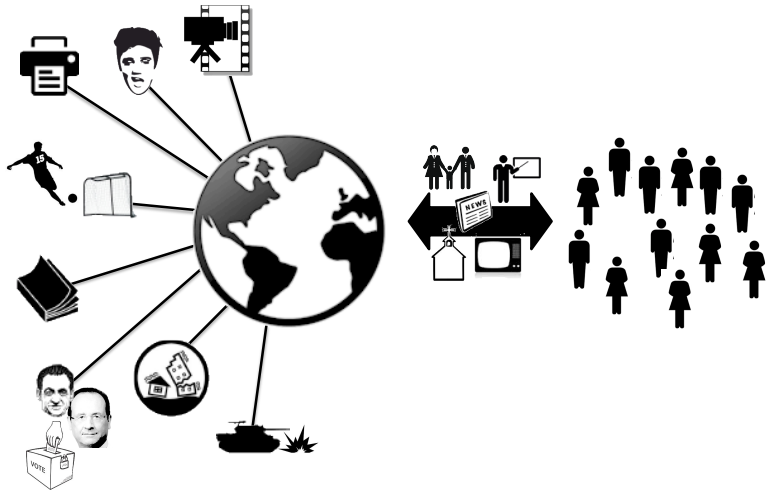
Joint work with Y.M. Kim, A. Hasnat, S. Bonnevay, J. Jacques and more…

1st Lyon-Columbia Research Workshop
ISFA, June 27, 2016

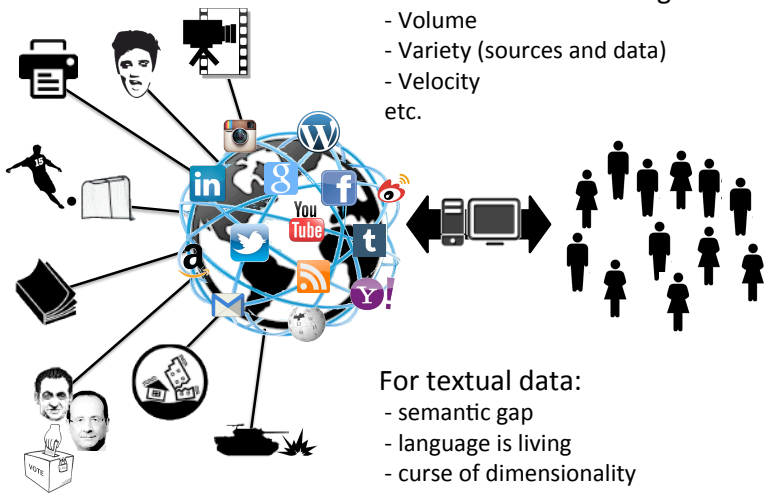
Outline of the talk

Outline of the talk

Studying representations



Nowadays with Internet



Chance and curse of big data:
- Volume
- Variety (sources and data)
- Velocity

etc.

For textual data:
- semantic gap
- language is living
- curse of dimensionality

Representing ≈ categorizing

- Philosophy, logic
  - Necessary and Sufficient Conditions [Aristotle]
  - Family resemblance [Wittgenstein,1958]
- Psychology, linguistics
  - Cognitive représentations and prototypes [Rosch,1973]
  - Linguistic categories [Lakoff,1987]
- Sociology
  - Social representations [Lippmann,1922] [Moscovici,1961]

➔ Data Science

Key idea to take home

Machine learning (weakly supervised clustering) can help for studying representations

What?	How?	Who?	When
↓	↓	↓	↓
topic learning	opinion mining	SN analysis, role detection	time-aware models

Representation and sparseness

Image of a movie

Title	Type	Plot	Actors	Rythm	Originality	etc.
Tomorrowland	Sci-fi	(…)	G. Clooney, H. Laurie…			
o	+	+	o	- -	+	

"New Disney rather desappointing. But **I like so much sci-fi movies** I couldn't miss it."

"**Ambitious and visually stunning**, this movie…"

"The film **stars George Clooney, Hugh Laurie**, Britt Robertson, and Raffey Cassidy"

"Tomorrowland' **forgettable** look into future"

"**Like the whole plot** but obviously **too long for kids**"

"How do you spell **boring**? T-O-M-O-R-R-O-W-L-A-N-D. "

Outline of the talk

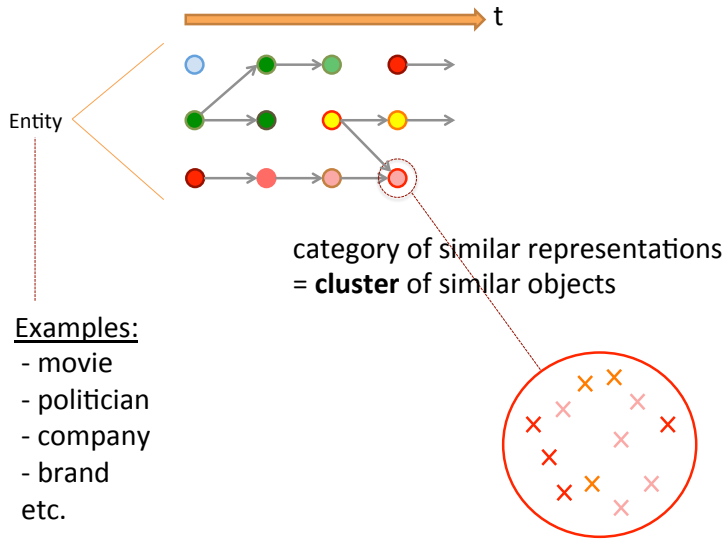
Part I: Representation(s) and Categorization(s)

Part II: Evolutionary Clustering for Sparse Data

Part III: Application to the ImagiWeb Project

Part IV: Conclusion and Future Work

Temporal evolution of entities



Examples:
- movie
- politician
- company
- brand

etc.

Sparse matrix as input

description features

Author	Time	f_1	f_2	f_3	f_4	f_5	f_6	...	f_{n-1}	f_n
pseudo1	t1		1				2		1	
pseudo1	t2		1				1			
pseudo1	t3				2					2
pseudo2	t1		3	1					1	
pseudo3	t1			3						
pseudo3	t2			2						
pseudo3	t3			2						
pseudo4	t3	3				1				
pseudo5	t3				3					2

Some state of the art

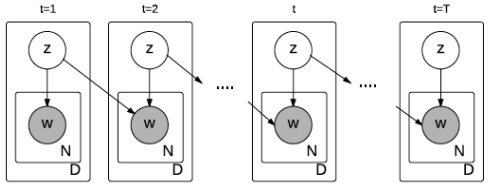
- Taking time into account
  - incremental clustering [Aggarwal,2003] [Labroche,2014]
  - evolutionary clustering [Chakrabarti,2006] [Chi,2007]
  - monitoring cluster evolution [Spiliopoulou,2006,2013]
- Dealing with sparse data
  - mixture models [Dempster,1977]
  - topic models [Hofmann,1999] [Blei,2003]
  - default clustering [Velcin,2005]

Our objective

- Analyze temporal sparse data using clustering
  - identify group of users who use similar descriptions
  - track entity's image over time
  - detect and interpret temporal changes
- Test on real data within ImagiWeb project
  - **case study 1:**
    image of French politicians given by Twitter users
  - **case study 2:**
    image of a big national company about nuclear energy given by bloggers

Model 1: Temporal Mixture Model

- **TMM** = probabilistic generative model [Kim,2015]



- What's new?
  - retrospective approach: the recent past matters
  - no Dirichlet prior, in opposite to most topic models
- Parameters to estimate: $\Theta = \{z, \pi, \phi\}$
- Optimization by Expectation-Maximization (EM)

Model 2: Parametric link approach

- **MM-Plink** = MM + linear link between (t-1) and (t) + model selection using BIC
- Relation between the parameters μ_{t-1} and μ_t:

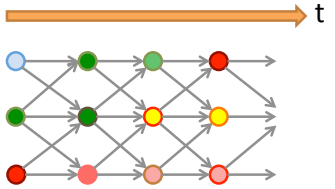
$$\mu^t = \Phi(\delta\Phi^{-1}(\mu^{t-1}) + \gamma)$$

cluster at time t — link parameter (mult.) — cluster at time t — link parameter (add.)

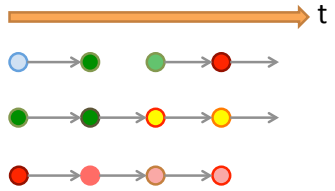
- Clustering estimated with classic EM
- Different combinations tested for (δ,γ):
  (1,0) = no change, (0,$\gamma_{j,k}$) = totally new clusters,
  (δ,γ) = same global change, etc.

Differences between the two models

model 1: TMM



model 2: MM-Plink
+ interpretation



Outline of the talk

Part I: Representation(s) and Categorization(s)

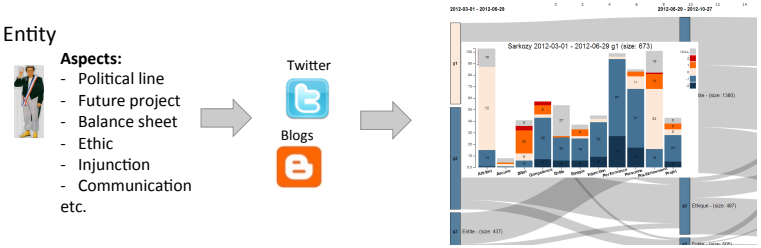
Part II: Evolutionary Clustering for Sparse Data

Part III: Application to the ImagiWeb Project

Part IV: Conclusion and Future Work

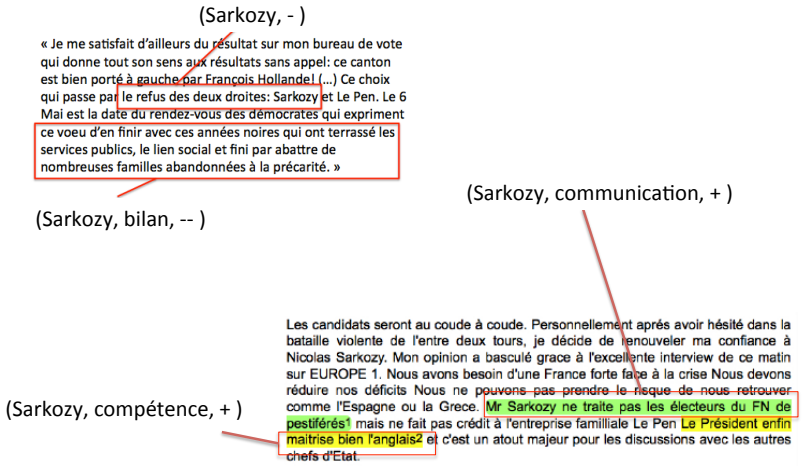
ImagiWeb project

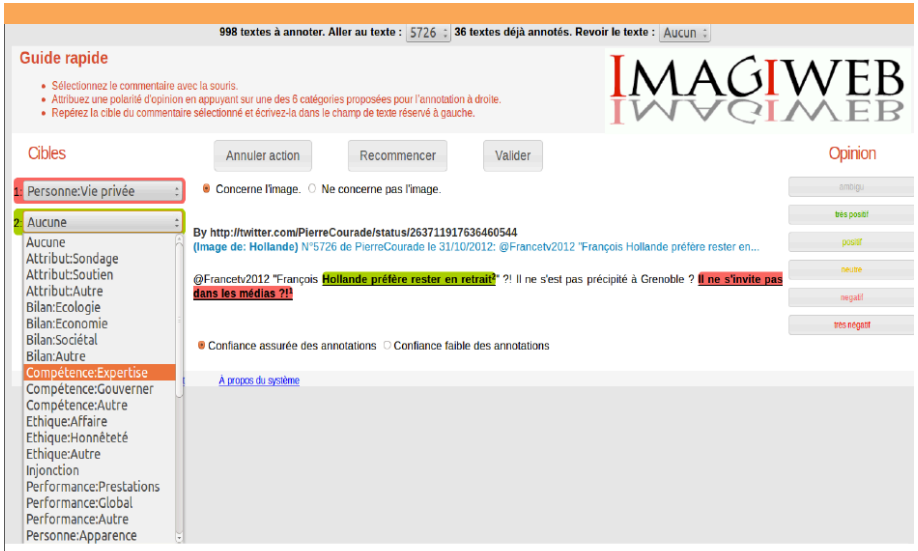
- Studying the image (representation) of entities emitted from the social media and its evolution over time [Velcin,2014]



- Granted by the ANR for 3 years (2012-2015)
- Needs complementary skills: NLP, machine learning, software engineering, analysis of public opinion, semiology…
- 6 partners : ERIC (management), CEPEL, LIA, AMI Software, EDF R&D, Xerox Research Centre Europe (XRCE)

Design of a full annotation scheme





JSON : {pertinence:concerne,confiance:assuree,ambigu:{},trespositif:{},positif:{2:"Hollande préfère rester en retrait"},neutre:{},negatif:{1:"Il ne s'invite pas dans les médias"},tresnegatif:{}}

Automatic annotation

La France est une république indivisible, démocratique, laïque et sociale, voilà mon engagement. #FH2012 → (Ethique, + +)

Geste fort du président #Hollande qui participera ce jeudi à la journée des mémoires, de la traite, de l'esclavage et de leurs abolitions. → (Positionnement, +)

Pourquoi j'aime bien Mélenchon et je voterai Hollande http://t.co/TVM8RwoH via @********** → (Injonction, +)

#Delanoë "ce qui me frappe ds la campagne de #Hollande c son honnêteté intellectuelle alors que #Sarkozy dit tout et n importe quoi" → (Ethique:Honnêteté, +)

@aut-1154 Neuilly sur Seine 61100 habitants , France 65000 000 .Votez Hollande. → (Injonction, +)

@****** Hollande n'a aucun charisme ! Il fait honte à la France et aux Français ! → (Personne:Charisme, -)

Sympatisch, ce Hollande. Et cultivé avec ça. On a parlé saucisses toute la soirée. → (Personne:Charisme, -)

Je savais qu'Hollande était un gros mou de socialiste. Mais là si ce n'est pas du reniement ou du renoncement ?#Libertédeconscience → (Ethique:Honnêteté, -)

François Hollande : le mensonge c'est maintenant: C'est cela un président . Il y a pas comme un léger bug → (Ethique:Honnêteté, - -)

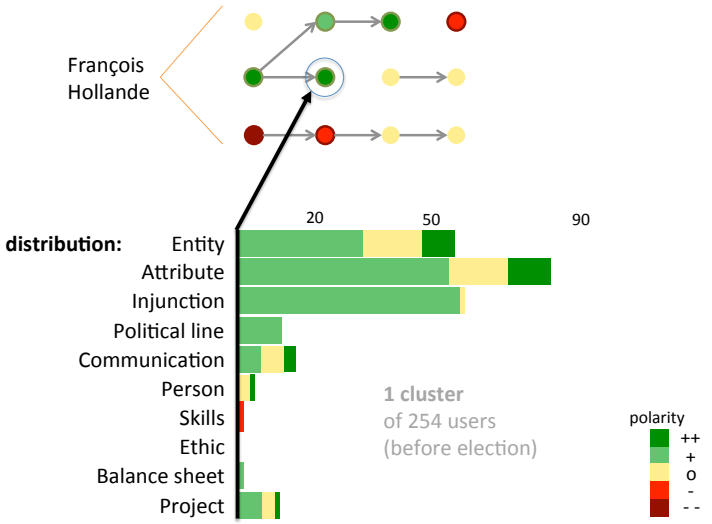
Copé appelle Hollande à "reprendre en main" son gouvernement "incompétent" http://t.co/IPanwi5r via @LePoint → (Compétence, -)

Extracting and monitoring images

E.g.
with politicians:

Author	Time	(a_1,++) (Entité, + +)	(a_1,+) (Entité, +)	(a_1,o) (Entité, o)	(a_1,-) (Entité, -)	(a_1,- -) (Entité, - -)	(a_2,++) (Attribut, + +)	...	(a_p,-) (Projet, -)	(a_p,- -) (Projet, - -)
pseudo1	t1		1				2		1	
pseudo1	t2		1				1			
pseudo1	t3				2					2
pseudo2	t1		3	1					1	
pseudo3	t1			3						
pseudo3	t2			2						
pseudo3	t3			2						
pseudo4	t3					1				
pseudo5	t3				3					2

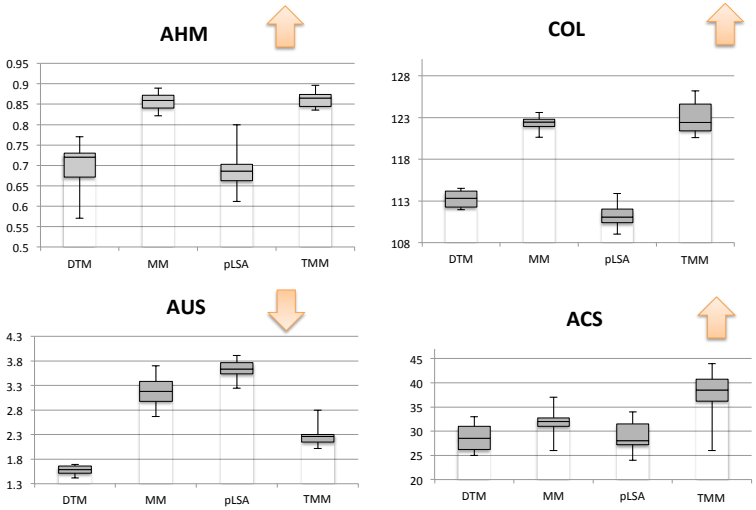
Testing model-based evolutionary clustering



Quantitative results of TMM

- Tested on a subset of tweets (entity FH, before and after election, k=9)
- Comparison with:
  - Dynamic Topic Model (DTM) [Blei,2006]
  - Simple Mixture Model (MM)
  - Probab. Latent Semantic Analysis (pLSA) [Hofmann,1999]
- Internal critera:
  - Co-ocurrence level (**COL**)
  - Average Unsmoothness (**AUS**)
  - Average Homogeneity (**AHM**)
  - Author Consistency Sum (**ACS**)

Quantitative results of TMM

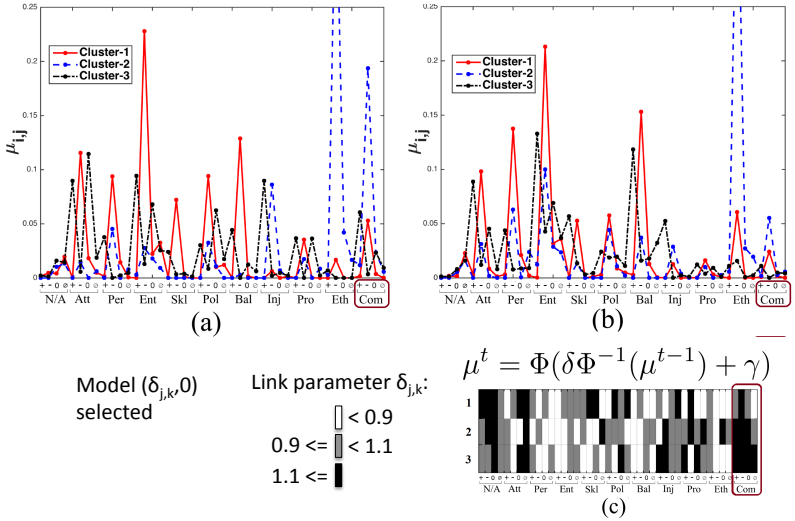


Quantitative results of MM-Plink

- Tested on a subset of tweets (entity FH, 3 time periods, k=3, total of ~3000 observations)
- Comparison between:
  - Simple Mixture Model (MM)
  - Temporal Mixture Model (TMM) [Kim,2015]
  - Parametric-link MM (MM-Plink) = our new proposal
- Additional criterion: Average Perplexity (**APL**)

	MM	TMM	MM-Plink
APL	1668.83	**1039.89**	1303.58
AUS	4.25	1.88	**1.42**
AHM	0.8	**0.87**	0.83
COL	59	**62**	61
ACS	63	**87**	82

Towards and understanding of evolution



Model ($\delta_{j,k}$,0) selected

Link parameter $\delta_{j,k}$:

< 0.9

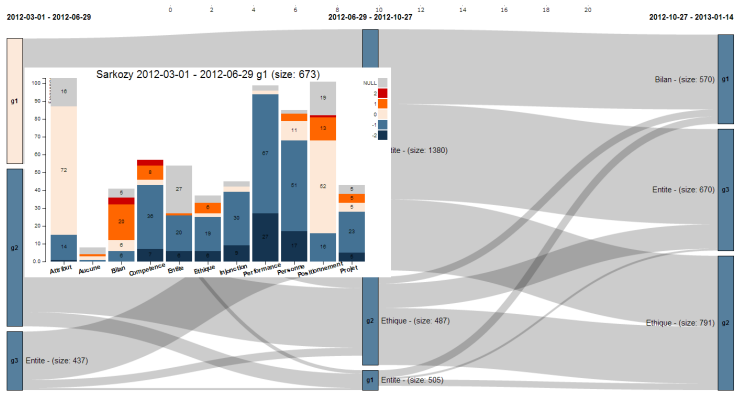
0.9 <= < 1.1

1.1 <=

Integrated into the final prototype

With TMM:

MM-Plink soon…



Outline of the talk

Part I: Representation(s) and Categorization(s)

Part II: Evolutionary Clustering for Sparse Data

Part III: Application to the ImagiWeb Project

Part IV: Conclusion and Future Work

Conclusion

- New models for evolutionary clustering
  - dedicated to sparse data
  - taking temporal transition into account
  - trying to add more interpretation of the evolution process
- applied to social media analysis
  - extraction and monitoring opinionated images
- in close collaboration with social sciences
  - joint work with specialists in political studies and semiologists (all along the process)

Future work

- From the methodology point of view:
  - going farther into the interpretation process
  - more comparisons needed (see MONIC [Spiliopoulou,2006] for instance)
  - testing non-parametric approaches
  - looking for change points in the timeline
- For the ImagiWeb project:
  - testing TMM and MM-Plink on the rest of the available data and 2nd case study
  - more (qualitative) evaluation needed
  - qualifying users' groups using additional variables

[THANK YOU]