

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

2

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

3

Outline of the talk

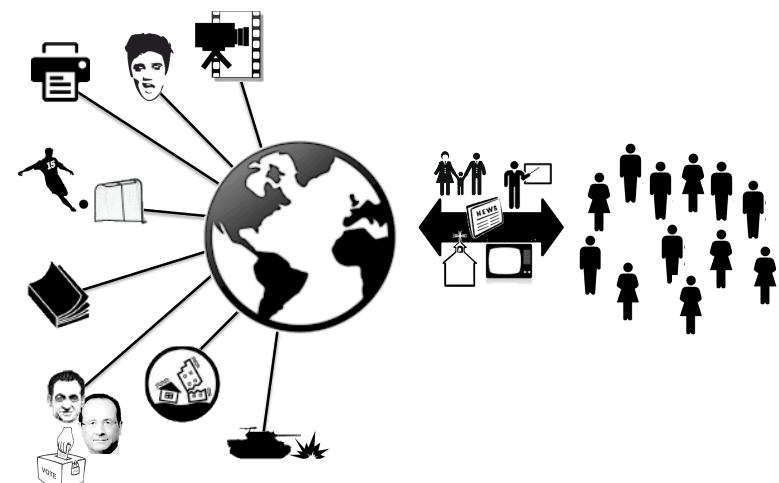
Part I: Representation(s) and Categorization(s)

Part II: Evolutionary Clustering for Sparse Data

Part III: Application to the ImagiWeb Project

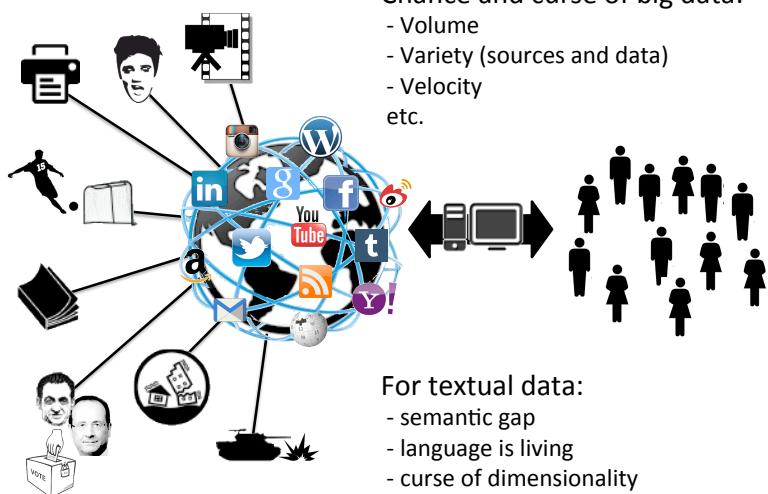
Part IV: Conclusion and Future Work

Studying representations



4

Nowadays with Internet



5

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

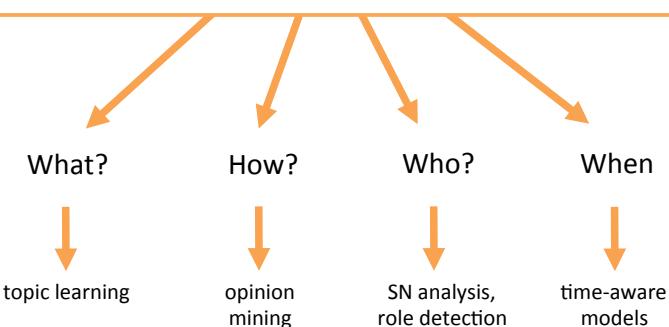
- Social representations
[Lippmann,1922] [Moscovici,1961]

→ Data Science

6

Key idea to take home

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



7

Representation and sparseness

Image of a movie

Title	Type	Plot	Actors	Rythm	Originality	etc.	Image
Tomorrowland	Sci-fi	(...)	G. Clooney, H. Laurie...				
0	+	+	0	--	+		

"New Disney rather disappointing. 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. "

8

Outline of the talk

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

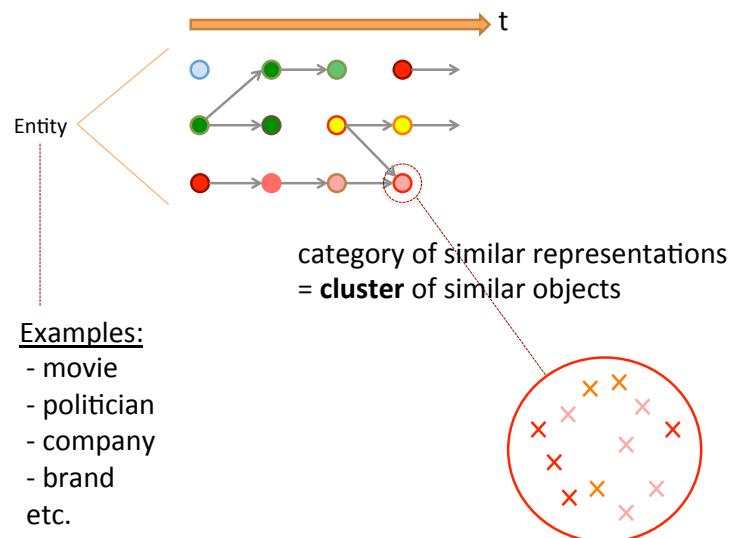
Part II: Evolutionary Clustering for Sparse Data

Part III: Application to the ImagiWeb Project

Part IV: Conclusion and Future Work

9

Temporal evolution of entities



10

Sparse matrix as input

Author	Time	description features								
		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

11

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 [Vercin,2005]

12

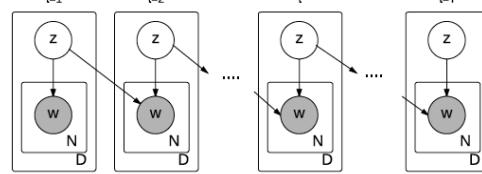
Our objective

- 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

13

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)

14

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 :

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

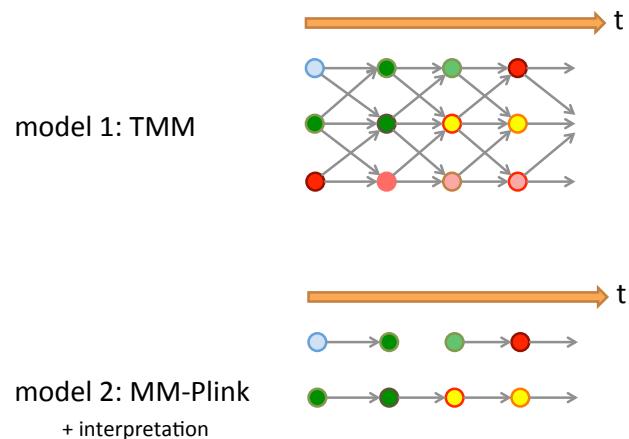
↑ cluster at time t ↑ cluster at time t ↑ link parameter (add.)

link parameter (mult.)

- 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.

15

Differences between the two models



16

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

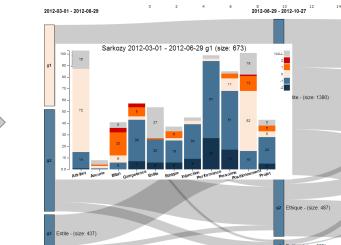
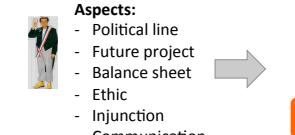
17

ImagiWeb project

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

Entity

- Aspects:**
- Political line
 - Future project
 - Balance sheet
 - Ethic
 - Injunction
 - Communication
 - etc.



- 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)

18

Design of a full annotation scheme

(Sarkozy, -)

« Je me satisfais d'ailleurs du résultat sur mon bureau de vote qui donne tout son sens aux résultats sans appel: ce canton est bien porté à gauche par François Hollande (...) Ce choix qui passe par le refus des deux droites: Sarkozy et Le Pen. Le 6 Mai est la date du rendez-vous des démocrates qui expriment ce vœu d'en finir avec ces années noires qui ont terrassé les services publics, le lien social et fini par abattre de nombreuses familles abandonnées à la précarité. »

(Sarkozy, bilan, --)

(Sarkozy, communication, +)

(Sarkozy, compétence, +)

Les candidats seront au coude à coude. Personnellement après avoir hésité dans la bataille violente de l'entre deux tours, je décide de renouveler ma confiance à Nicolas Sarkozy. Mon opinion a basculé grâce à l'excellente interview de ce matin sur EUROPE 1. Nous avons besoin d'une France forte face à la crise. Nous devons réduire nos déficits. Nous ne **peuvons pas prendre le risque de nous retrouver comme l'Espagne ou la Grèce**. Mr Sarkozy ne traite pas les électeurs du FN de **pestiférés**¹ mais ne fait pas crédit à l'entreprise familiale Le Pen. Le Président enfin maîtrise bien l'**anglais**² et c'est un atout majeur pour les discussions avec les autres chefs d'Etat.

19

998 textes à annoter. Aller au texte : 5726 : 36 textes déjà annotés. Revoir le texte : Aucun :

IMAGIWEB

Guide rapide

- Sélectionnez le commentaire avec la souris.
- Atribuez une polarité d'opinion en appuyant sur une des 6 catégories proposées pour l'annotation à droite.
- Repérez la cible du commentaire sélectionné et écrivez-la dans le champ de texte réservé à gauche.

Cibles	Annuler action	Recommander	Valider
1 Personne/Vie privée			
2 Aucune			
Aucune			
Attribut:Sondage			
Attribut:Soutien			
Attribut:Autre			
Bilan:Ecologie			
Bilan:Economie			
Bilan:Sociétal			
Bilan:Autre			
Compétence:Expertise			
Compétence:Gouverner			
Compétence:Autre			
Ethique:Affaire			
Ethique:Honnêteté			
Ethique:Autre			
Injunction			
Performance:Prestations			
Performance:Global			
Performance:Autre			
Personne:Apparence			

Opinion

- ambigu
- tres positif
- positif
- neutre
- negatif
- tres negatif

By http://twitter.com/PierreCourade/status/263711917636460544
(Image de: Hollande) N°5726 de PierreCourade le 31/10/2012: @FranceTV2012 "François Hollande préfère rester en...
@FranceTV2012 "François Hollande préfère rester en retrait" ? Il ne s'est pas précipité à Grenoble ? **Il ne s'invite pas dans les médias** ?!"

Concerne l'image. Ne concerne pas l'image.

Confiance assurée des annotations Confiance faible des annotations

A propos du système

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

20

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/lPanwi5r> via @LePoint → (Compétence, -)

21

Extracting and monitoring images

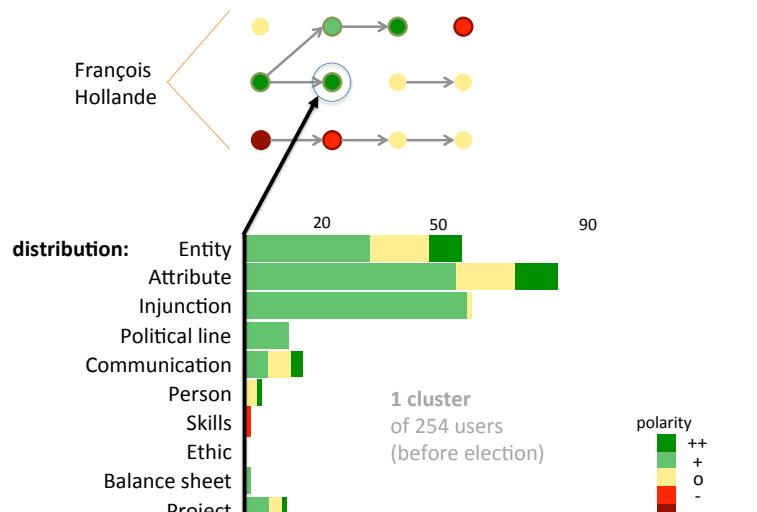
E.g.
with politicians:

(Entité, +)	(Entité, +)	(Entité, o)	(Entité, -)	(Entité, -)	(Attribut, +)	(Projet, -)
-------------	-------------	-------------	-------------	-------------	---------------	-------------

Author	Time	(a ₁ ,++)	(a ₁ ,+)	(a ₁ ,o)	(a ₁ ,-)	(a ₁ ',-)	(a ₂ ,++)	...	(a _p ,-)	(a _p ',-)
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

22

Testing model-based evolutionary clustering



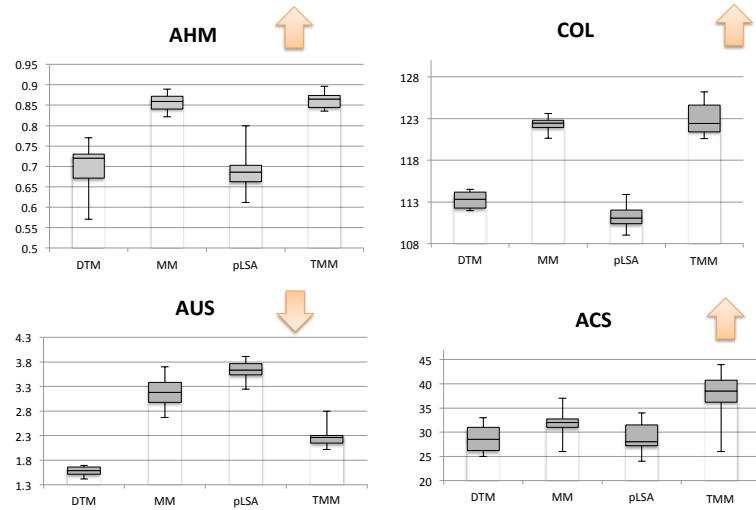
23

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 criteria:
 - Co-occurrence level (COL)
 - Average Unsmoothness (AUS)
 - Average Homogeneity (AHM)
 - Author Consistency Sum (ACS)

24

Quantitative results of TMM



25

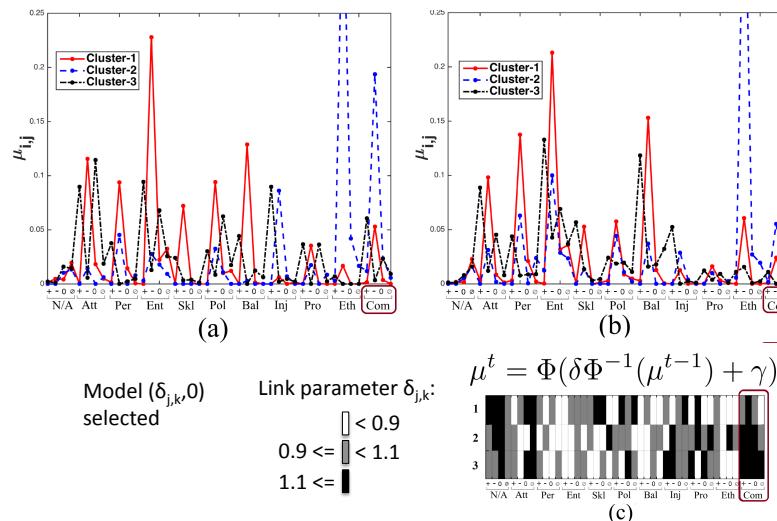
Quantitative results of MM-Plink

- 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	<u>1303.58</u>
AUS	4.25	<u>1.88</u>	1.42
AHM	0.8	0.87	<u>0.83</u>
COL	59	62	<u>61</u>
ACS	63	87	<u>82</u>

26

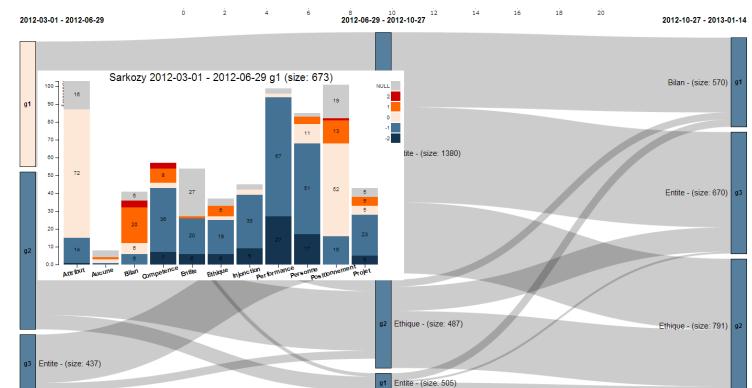
Towards and understanding of evolution



27

Integrated into the final prototype

With TMM:



28

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

29

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)

30

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

31

[THANK YOU]

32