

Analyse diachronique des données textuelles : mise en comparaison des méthodes LDA et des méthodes basées sur le clustering et les graphes de contraste

Jean-Charles LAMIREL

Synalp – LORIA

Université de DALIAN - CHINE

ISTEX-R WP1



Founding : ANR-10-IDEX-0004-02

EGC-GAST-2019

Presentation plan

- ❖ **Change detection challenge in text data**
- ❖ **A new combined approach for change detection**
- ❖ **Principle of application for diachronic analysis of heterogeneous text collection in the ISTEEX-R WP1 context**
- ❖ **Novel approaches based on contrast graphs**
- ❖ **Some results comparison with LDA**

Introduction

Text mining and change detection challenges

Text mining is a machine learning domain which raises difficult challenges mainly because of high dimensional data and large datasets :

- ❖ Facing with learning or mining model evaluation,
- ❖ Facing with distance ambiguities or inefficiency,
- ❖ Facing with multiple data representations,
- ❖ Facing with synthetizing and visualizing mining results.

Facing with potentially evolving data is even a more complex but strategic problem with many application :

- ❖ Opinion tracking and sentiment analysis,
- ❖ Technological and scientific surveys,
- ❖ Remote people control, ...

Diachronic analysis

Usual tools for topic identification [Guille et al. 2016]

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

LDA [Blei et al. 03]

considers that :

- ▶ Underlying topic of a corpus of document are characterized by a multinomial distribution of word present in a document
- ▶ A document is itself a composition of the extracted topics

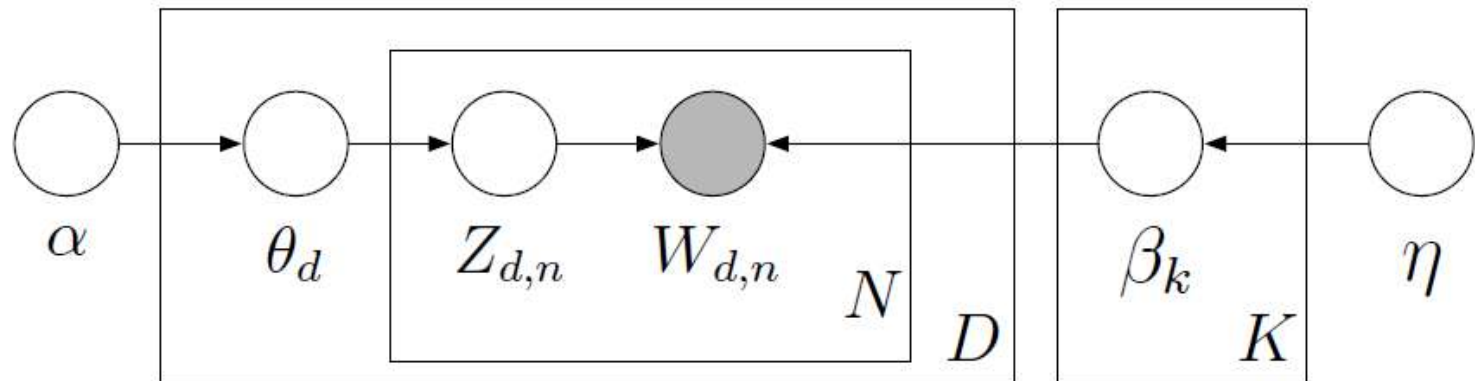
Problems

- ▶ Generated topics can strongly vary in quality and generality depending on process initialization condition or sampling

Topics changes can be highlighted using data timestamps.

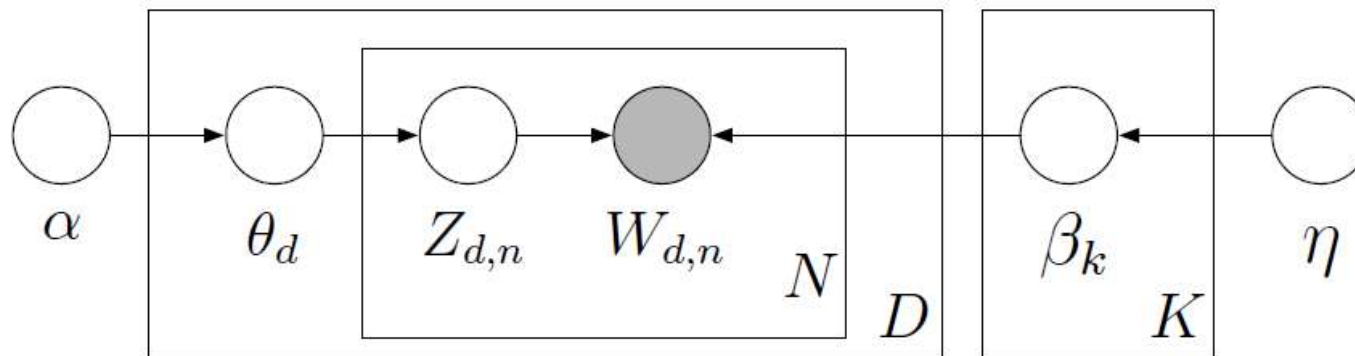
- ▶ Not scalable to the document level.

LDA inference principle



- From a collection of documents, infer
 - Per-word topic assignment $z_{d,n}$
 - Per-document topic proportions θ_d
 - Per-corpus topic distributions β_k
- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

LDA inference techniques



Approximate posterior inference algorithms

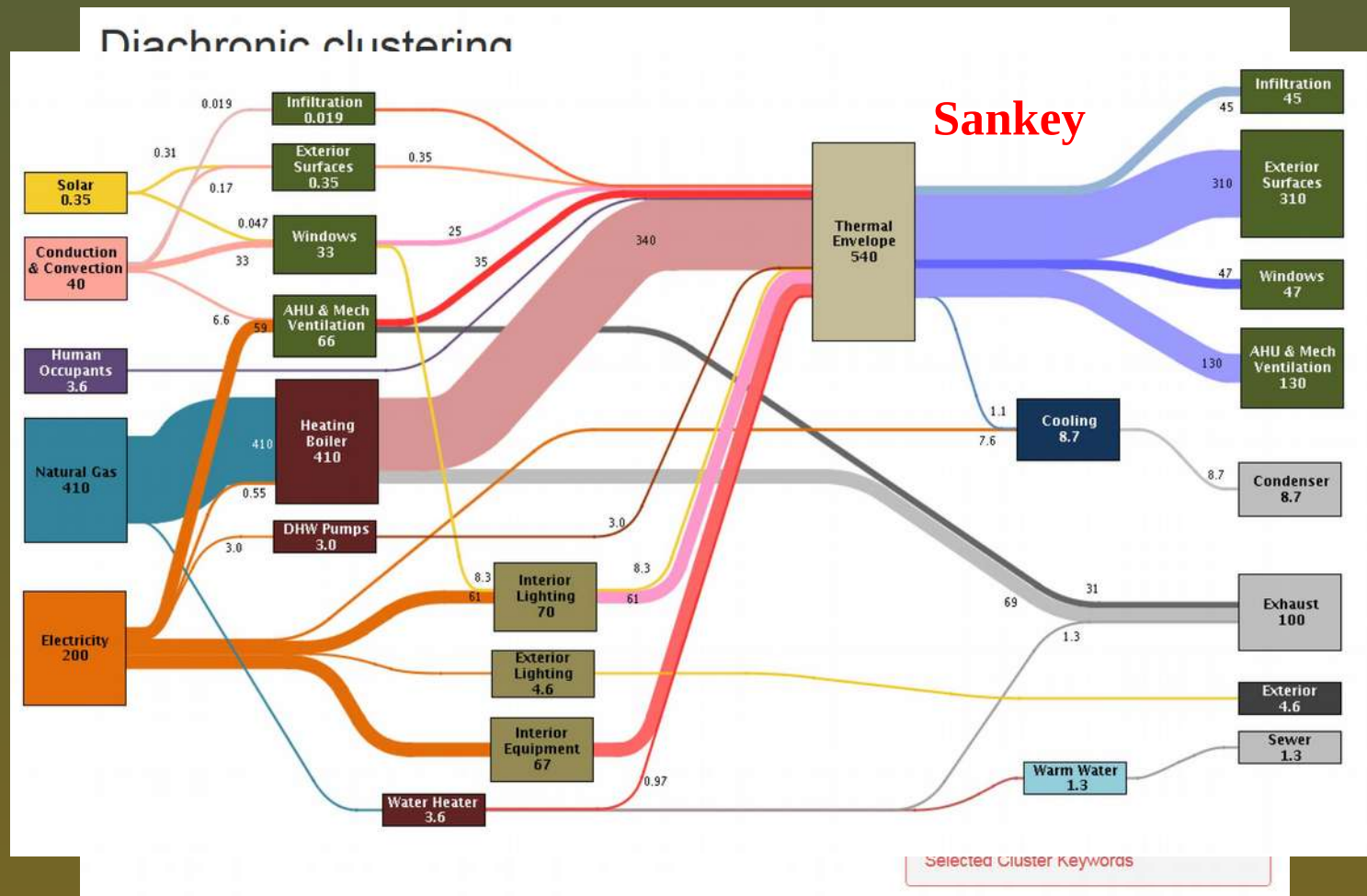
- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

For comparison, see Mukherjee and Blei (2009) and Asuncion et al. (2009).

Diachronic analysis

Visualization of changes

- ❖ Visualization of change is still an open problem and many methods are experienced



**A new reliable approach for
diachronic analysis**

How to find efficient alternative to LDA

- ❖ A combination of techniques is used to substitute to LDA and cope with its known problems
 - ◆ A new metric based on feature maximization is exploited in all steps,
 - ◆ Neural clustering is used for topic detection,
 - ◆ Unsupervised Bayesian reasoning is exploited for detection of changes.
- ❖ Variation of former techniques help to :
 - ◆ Period detection,
 - ◆ Optimal model identification,
 - ◆ (Automatic paper summarization and metadata extraction).

The method has not parameters, computation time is low and granularity of topics is homogeneous.

The feature maximization metric

[Lamirel 08]

Let us consider a partition C resulting from a grouping method applied on a set of data D represented with a set of descriptive features F , feature maximization is a metric which favors groups (i.e. clusters or classes) with maximum *Feature F-measure* which represents the harmonic mean between :

Feature Recall

$$FER_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f}$$

$$\equiv P(c|f)$$

Feature Precision

$$FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{\substack{f' \in F^c, d \in c \\ f \in F^c, d \in c}} W_d^{f'}}$$

$$\equiv P(f|c)$$

A maximized cluster feature is a feature whose *Feature F-measure* is **maximized** by the group members (i.e. data).

A simple example

- ❖ We consider a sample of **Men (M)** and **Women (F)** for which we measure **Hair_length** and **Shoes_size** and **Nose_size**

Shoes_size	Hair_length	Nose_size	Class
9	5	5	M
9	10	5	M
5	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

A simple example

- ❖ We compute the Feature Recall (FR) and the Feature Precision (FP) and the Feature F-measure (FF) for each class and each feature and each class

Shoes_size	Hair_length	Nose_size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

$$FR(S,M) = 27/43 = 0.62$$

$$FP(S,M) = 27/78 = 0.35$$

$$FF(S,M) = \frac{2(FR(S,M) \times FP(S,M))}{FR(S,M) + FP(S,M)} = 0.48 = 0.48$$

Recall is scale independent,
Precision is not.

A simple example

- ❖ We compute the average marginal values of Feature F-measure by feature (local) and the overall Feature

	$F(x, M)$	$F(x, F)$																	
			<table border="1"> <thead> <tr> <th></th> <th>$F(x, M)$</th> <th>$F(x, F)$</th> <th>$\overline{F(x, \cdot)}$</th> </tr> </thead> <tbody> <tr> <td>Hair_length</td> <td>0.39</td> <td>0.66</td> <td>0.53</td> </tr> <tr> <td>Shoes_size</td> <td>0.48</td> <td>0.22</td> <td>0.35</td> </tr> <tr> <td>Nose_size</td> <td>0,3</td> <td>0,24</td> <td>0,27</td> </tr> </tbody> </table>		$F(x, M)$	$F(x, F)$	$\overline{F(x, \cdot)}$	Hair_length	0.39	0.66	0.53	Shoes_size	0.48	0.22	0.35	Nose_size	0,3	0,24	0,27
	$F(x, M)$	$F(x, F)$	$\overline{F(x, \cdot)}$																
Hair_length	0.39	0.66	0.53																
Shoes_size	0.48	0.22	0.35																
Nose_size	0,3	0,24	0,27																
Hair_length	0.39	0.66	0.53																
Shoes_size	0.48	0.22	0.35																
Nose_size	0,3	0,24	0,27																

The features whose Feature F-measure is under the global Feature F-measure average are removed

⇒ **Nose_size is removed**

The remaining (i.e. selected) features whose F-measure is over marginal average in one class are considered as active in this class

⇒ **Shoes_size is active in Men class**

⇒ **Hair_length is active in Women class**

$\overline{F(x, \cdot)}$
0.38
0.38

A simple example

- ❖ The contrast factor highlights the degree of activity/passivity of selected features relatively to their marginal Feature F-measure average in the different classes

	F(x, M)	F(x, F)	F(x, .)
Hair_length	0.39	0.66	0.53
Shoes_size	0.48	0.22	0.35

	C(x, M)	C(x, F)
Hair_length	0.39/0.53	0.66/0.53
Shoes_size	0.48/0.35	0.22/0.35

	C(x, M)	C(x, F)
Hair_length	0.74	1.25
Shoes_size	1.37	0.63

The contrast can be seen as a function that will tend to:

1. Overlength the Hairs of Women
2. Oversize the Shoes of Men
3. Underlength the Hairs of Men
4. Undersize the Shoes of Women

Classification with FMC

Deft challenge [JADT 2014]

- ❖ Dataset of extracts of talk of CHIRAC et MITTERAND presidents:
 - ▶ 73255 sentences of Chirac,
 - ▶ 12320 sentences of Mitterrand.
- ❖ Best results till now on that dataset : 88% accuracy (almost 16850 bilateral errors) by LIA.
- ❖ Result with feature maximization : 99,999% accuracy (12 unilateral errors)
 - ▶ Extra-light NLP preprocessing,
 - ▶ No lemmatization is needed,
 - ▶ Stop words are kept and proof to be useful for analysis.

CHIRAC

1.930810 partenariat
1.858265 dynamisme
1.811123 exigence
1.775048 compatriotes
1.769069 vision
1.768280 honneur
1.763166 asie
1.762665 efficacité
1.745192 saluer
1.743871 soutien
1.737269 renforcer
1.715155 concitoyens
1.709736 réforme
1.703412 devons
1.695359 engagement
1.689079 estime
1.671255 titre
1.669899 pleinement
1.662398 cœur
1.661476 ambition
1.654876 santé
1.640298 stabilité
1.632421 amitié
1.628630 accueil
1.622473 publics
1.616558 diversité
1.614945 service
1.612488 valeurs
1.610123 détermination
1.601097 réformes
1.592938 état
.....

MITTERAND

1.881835 douze
1.852007 est-ce
1.800091 eh
1.786760 quoi
1.777568 -
1.758319 gens
1.747909 assez
1.741650 capables
1.716491 penser
1.700678 bref
1.688314 puisque
1.672872 on
1.662164 états
1.620722 parle
1.618184 fallait
1.604095 simplement
1.589586 entendu
1.580018 suite
1.572140 peut-être
1.571393 espère
1.560364 parlé
1.550856 dis
1.549594 cela
1.538523 existe
1.535598 façon
1.529225 pourrait
1.525645 là
1.525508 chose
1.523575 époque
1.522290 production
1.519365 trouve
.....

Classification with FMC

Dickens-Collins controversy (stylometry)

DICKENS	
1,227361	coming
1,04304	heart
1,197376	going
1,531862	boy
1,073734	hands
1,379369	cried
1,10153	men
1,491942	gentleman
1,261003	street
1,316684	dear
1,022143	love
1,175788	like
1,240113	young
1,04989	light
1,151491	seemed
1,194106	dark
1,16963	happy
1,130386	know
1,584914	indeed
1,507332	fire
1,50468	often
1,513448	great
1,455893	pretty

DICKENS
CHILDNESS (IDF only)

1,480477	delighted
1,477346	laugh
1,480015	observed
1,782888	boots
1,638954	blessed
1,146102	walk
1,57313	piece
1,585134	played
1,562992	rolling
1,49908	sing
1,463912	horses
1,439286	worked
1,436957	sun
1,4481	comfortable
1,454421	touching
1,389128	teach
1,461329	pleasant
1,531391	shadows
1,476144	windows
1,396892	pains
1,298137	youre
1,405204	raising
1,238744	wall

COLLINS	
1,294836	speaking
1,534484	answered
1,546532	waiting
1,250605	heard
1,575393	servant
1,232137	interest
1,064008	woman
1,288278	led
1,332932	left
1,299098	feel
1,123745	remember
1,069823	met
1,246964	open
1,321961	will
1,257266	look
1,064291	means
1,419852	husband
1,390256	doctor
1,538165	present
1,193449	suddenly
1,348551	herself
1,376969	truth
1,174412	letter

COLLINS
CHILDNESS (IDF only)

1,454853	proceeding
1,341769	explain
1,459178	anxiety
1,374366	notted
1,567781	discovery
1,519614	decide
1,486814	events
1,542078	informed
1,53803	approached
1,449153	eagerly
1,709461	confession
1,429975	accepted
1,450273	necessity
1,464363	evidence
1,416902	address
1,41972	capable
1,464345	patience
1,540817	sadly
1,386233	entering
1,416557	importance
1,685569	resoluton
1,364304	alarm
1,45935	possessed

The method provide exhaustive and precise results and allows fine-grained analysis modulation.

Feature maximization

Exploitation in clustering [Lamirel 12]

IGNGF clustering is a parameter-free incremental neural clustering method exploiting feature maximization in substitution to standard distances (Euclidean, cosine, ...)

- ❖ Combines clustering and feature selection/explanation capabilities

(pseudo-symbolic behavior)

- ❖ Shown to outperform both state-of-the-art symbolic (FCA) and numeric (Spectral Clustering) methods in complex problems:
 - Clustering of French verbs using syntactic-semantic features [Lamirel 12]

- ❖ Can be used to highlight latent classes in classification problems: classification of institutional websites using communication signatures [Lamirel 13]

C6- 14(14) [197(197)]

Prevalent Label — = AgExp-Cause

0.341100	G-AgExp-Cause
0.274864	C-SUJ:Ssub,OBJ:NP
0.061313	C-SUJ:Ssub
0.042544	C-SUJ:NP,DEOBJ:Ssub

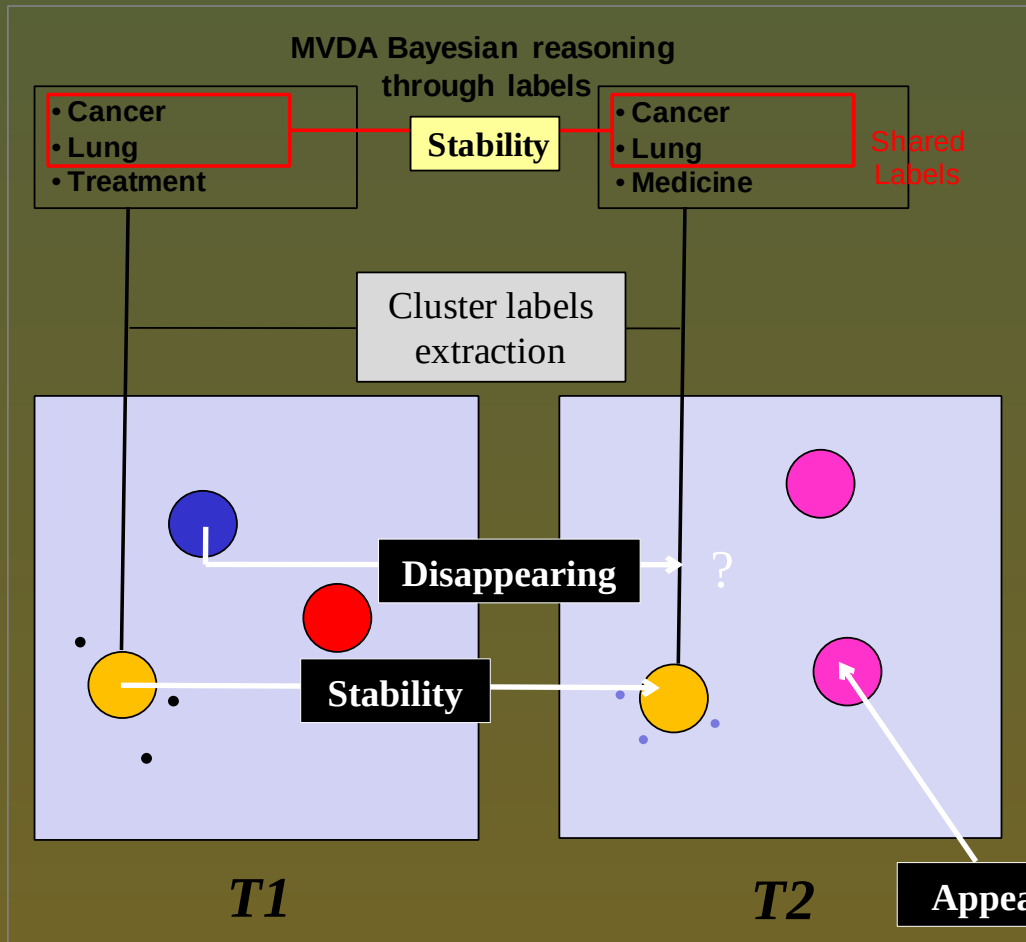
0.017787	C-SUJ:NP,DEOBJ:VPinf
0.008108	C-SUJ:VPinf,AOBJ:PP

[**déprimer 0.934345 4(0)]	[affliger 0.879122 3(0)]
[éblouir 0.879122 3(0)]	[choquer 0.879122 3(0)]
[décevoir 0.879122 3(0)]	[décontenancer 0.879122 3(0)]
[décontracter 0.879122 3(0)]	[désillusionner 0.879122 3(0)]
[**ennuyer 0.879122 3(0)]	[fasciner 0.879122 3(0)]
[**heurter 0.879122 3(0)]	...

F-max cluster labeling can be exploited with any clustering method.

Diachronic analysis exploiting feature maximization

Research topic changes tracking [Lamirel 12]



Multiple functions of the MVDA model are exploited :

- ▶ Time subperiods associated to viewpoints
- ▶ Optimized clustering using neural methods and unbiased quality measures
- ▶ High performance (F-max based) labeling techniques
- ▶ Adapted Bayesian reasoning on labels

F-max cluster labeling provides dimensionality reduction (feature selection) .

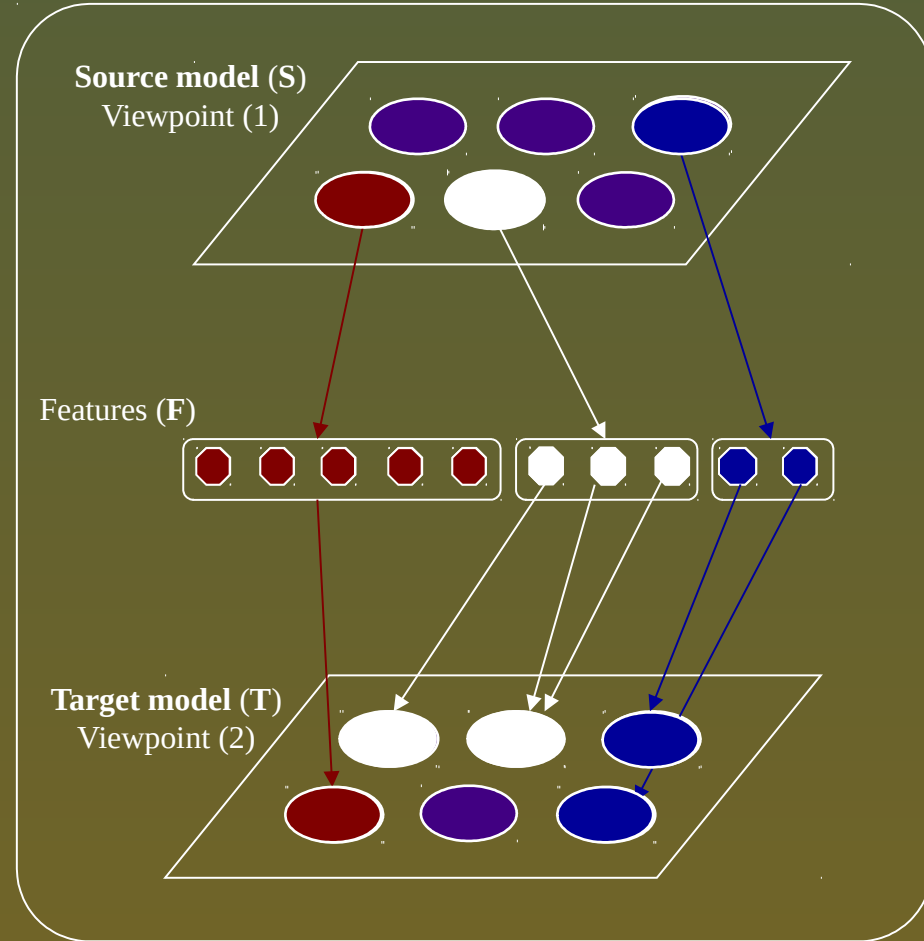
Diachronic analysis exploiting feature maximization

MVDA Paradigm

- ❖ MVDA paradigm relies on Bayesian reasoning
- ❖ Bayesian network is generated in an unsupervised way
- ❖ Uses clustering models shared data to perform cluster comparisons
- ❖ Applicable with any clustering method

Bayesian network model

:



Diachronic analysis exploiting feature maximization

MVDA Paradigm




$$PRG : S_k \rightarrow T_{i_k}$$

$$PC = \frac{1}{\bar{S}} \sum_{k; S_k \neq \emptyset} \frac{\sum_i P(act|T_{i_k})}{D_k + 1}$$

$$\bar{S} = |S_k \in S | S_k \neq \emptyset, S_k \in act|$$

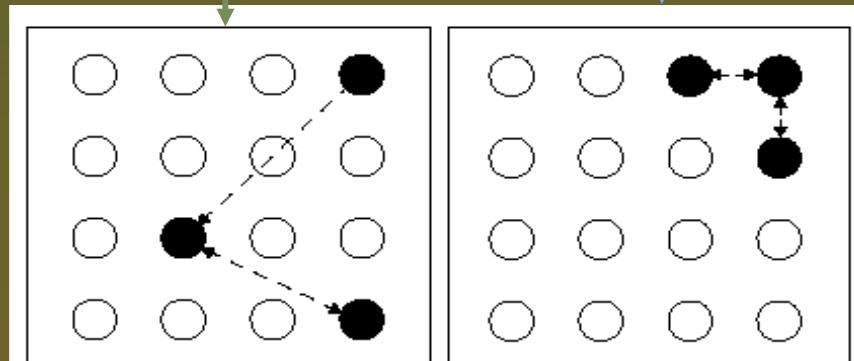
Strong focalization

PC measure

-  Evaluate the global similarity between the models (i.e. partitions) (1 = perfect).
-  Based on the average distance between the generated activities.
-  **Non symmetrical measure.**

Weak focalization

Activation of a class on the source viewpoint

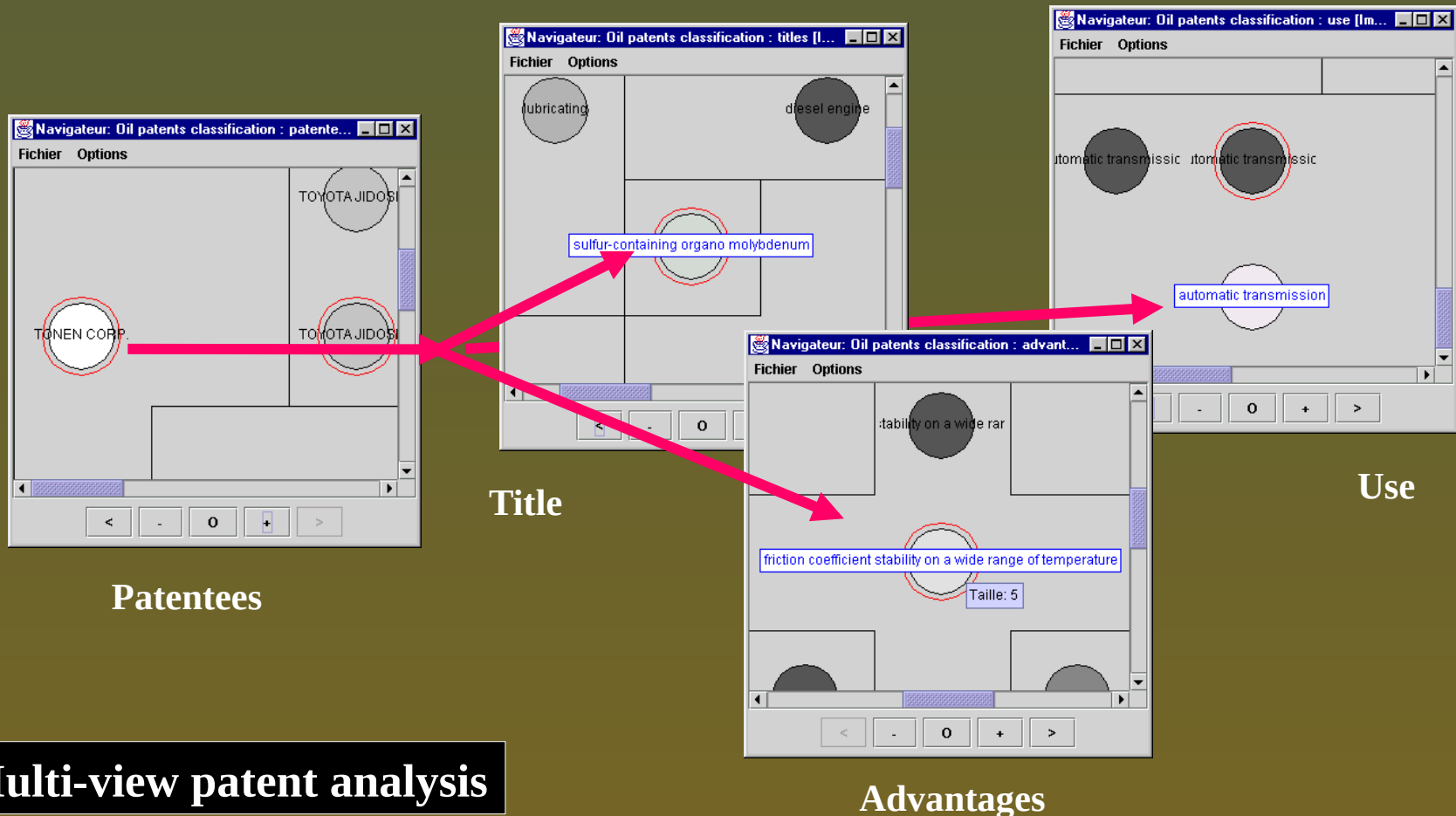


Potential activity profile on the target viewpoint

Diachronic analysis exploiting feature maximization

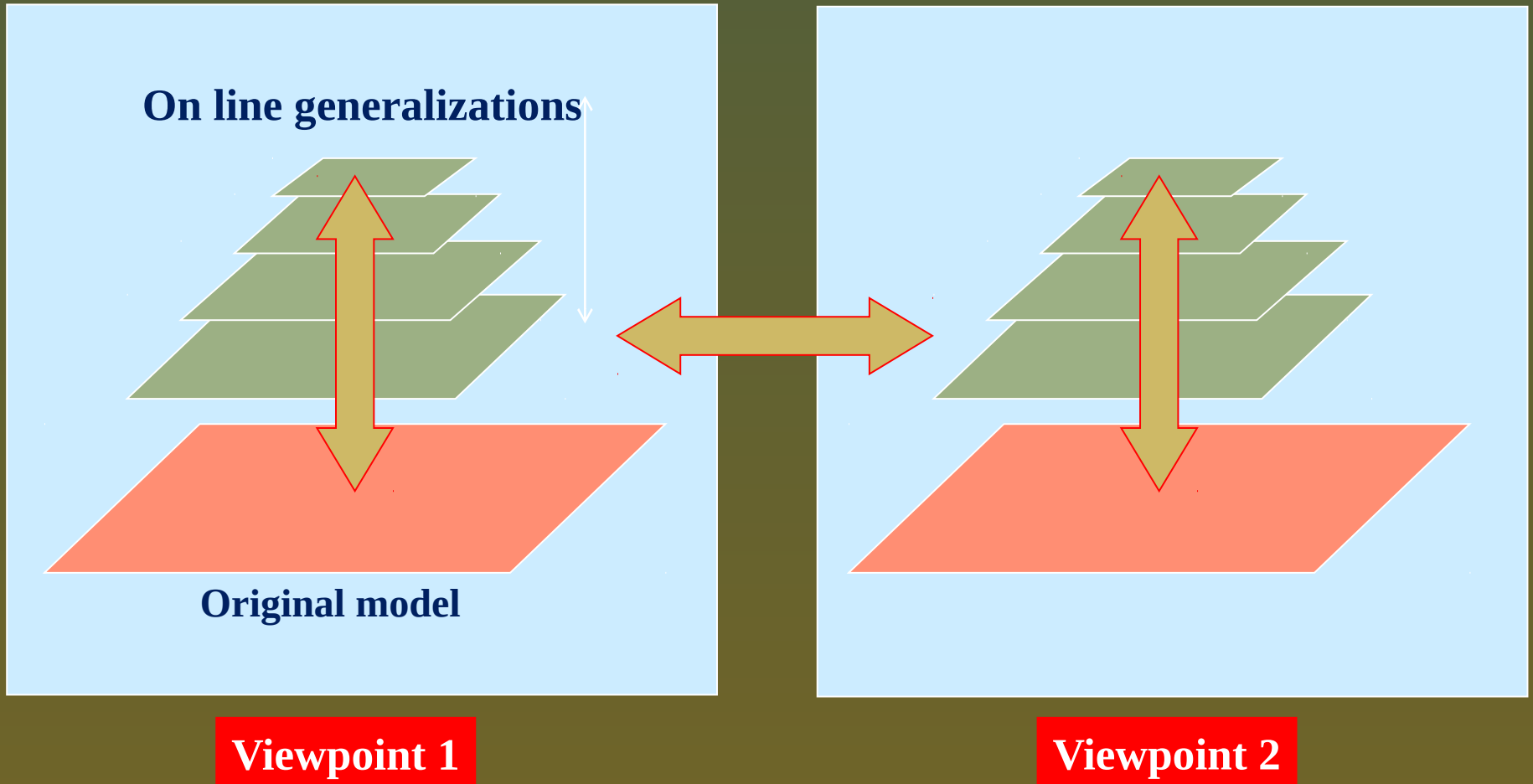
MVDA Paradigm

MDVA paradigm allows to exploit flexible communication between knowledge sources. New sources can be added on line.



Diachronic analysis exploiting feature maximization

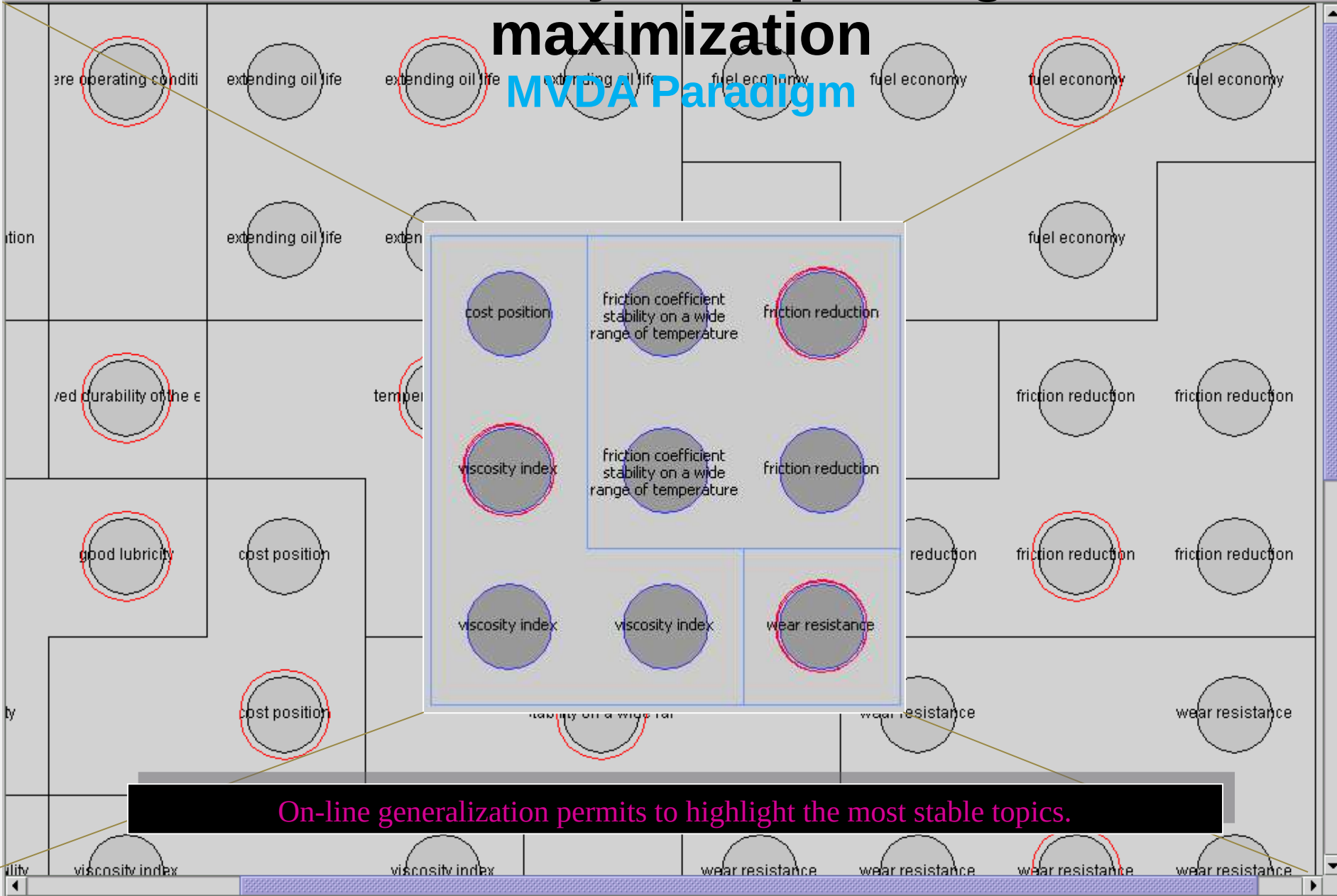
MVDA Paradigm



MVDA paradigm supports on-line generalization and inter-level unsupervised Bayesian reasoning.

Diachronic analysis exploiting feature maximization

MVDA Paradigm



On-line generalization permits to highlight the most stable topics.

Diachronic analysis exploiting feature maximization

MVDA Paradigm on labels and time periods

- ❖ Only clusters labels with labeling F-measure over average are considered for comparison
- ❖ Comparison is performed using an adaptation of MVDA Bayesian reasoning with :

$$P(t|s) = \frac{\sum_{l \in L_s \cap L_t} L_t - F(l)}{\sum_{l \in L_t} L_t - F(l)}$$

where L_x represent the set of labels associated to the cluster x , and $L_x \cap L_y$ represent the common labels, which can be called the label **matching kernel**, between the cluster x and the cluster y .

Diachronic analysis exploiting feature maximization

MVDA Paradigm on labels and time periods

The **similarity** between a cluster s of the source period and a cluster t of the target period is established using :

- ❖ The average matching probabilities $P_A(x)$ of a period cluster
- ❖ The global average activity A_x generated by a period model on the model of the alternative period and its standard deviation σ_x

Similarity is found if :

1) $P(t|s) > P_A(s)$ et $P(t|s) > A_s + \sigma_s$

2) $P(t|s) > P_A(s) + \sigma_s$ et $P(t|s) > A_s + \sigma_s$

Cluster splitting, cluster merging, vanishing clusters, appearing clusters events can be deduced from former similarity rules.

Diachronic analysis exploiting feature maximization

MVDA Paradigm on labels and time periods

source cluster: 27 [28/12] target cluster: 37 [16/8]

- Stable labels - similarity kernel

f1: 0.048624[27]
f1: 0.050103[27]

f2: 0.025228[37]
f2: 0.251873[37]

Microchannel plates (***)
Photon counting (***)

- Highly dominant (or peculiar) labels in source period

f1: 0.139294[27]
f1: 0.099508[27]
f1: 0.078299[27]
f1: 0.076769[27]

f2: 0.000000[-1]
f2: 0.000000[-1]
f2: 0.000000[-1]
f2: 0.000000[-1]

Photocathodes
Plasma diagnostic
Tokamak devices
Photomultipliers

- Highly dominant (or peculiar) labels in target period

f1: 0.000000[-1]
f1: 0.000000[-1]
f1: 0.000000[-1]

f2: 0.042154[37]
f2: 0.038080[37]
f2: 0.033203[37]

Quantum cryptography
Lidar
Quantum dot

Label is absent or not significant in period 1

Label Feature F-measures in source period (1)

Label Feature F-measures in target period (2)

Label names (***) = matching kernel labels

Cluster infos

Matching kernel (core labels)

Dominant labels in period 1

Dominant labels in period 2

Matching reports highlight some temporal correspondence between topics belonging to different periods.

Diachronic analysis exploiting feature maximization

MVDA Paradigm on labels and time periods

source cluster: 27	[28/12]	target cluster: 37	[16/8]
- Stable labels - similarity kernel			
f1: 0.048624[27]	f2: 0.025228[37]	Microchannel plates (***)	
- Highly dominant (or prevalent) labels in source period			
f1: 0.139294[27]	f2: 0.000000[-1]	Photocathodes	
f1: 0.099508[27]	f2: 0.000000[-1]	Plasma diagnostic	
f1: 0.078299[27]	f2: 0.000000[-1]	Tokamak devices	
f1: 0.076769[27]	f2: 0.000000[-1]	Photomultipliers	
- Highly dominant (or prevalent) labels in target period			
f1: 0.050103[27]	f2: 0.251873[37]	Photon counting (***)	
f1: 0.000000[-1]	f2: 0.042154[37]	Quantum cryptography	
f1: 0.000000[-1]	f2: 0.038080[37]	Lidar	
f1: 0.000000[-1]	f2: 0.033203[37]	Quantum dot	

Highlighted labels with high (italic) or middle high (others) Feature F-measure difference between periods

Label migrated from matching kernel to period 2

Rem : blue color codes are used for highlighted labels in source period (1) and red color codes are used for for highlighted labels in target period (2)

Labels migration consists in affecting matching kernel labels with high F-measure difference between periods to the most relevant period (using statistical tests).

Diachronic analysis exploiting feature maximization

Typical results – IST-PROMTECH dataset [Lamirel 12]

TIME PERIOD	NBR GROUPS	NBR MATCH	NBR DISAP	NBR APPE	NBR SPLI	NBR MERG
1996-1999	43	33	10	-	7	-
2000-2003	50	38	-	12	-	3

Label matching kernels permit to identify global topic temporal matches :

- ▶ Small temporal changes can be identify in the context of the global topic temporal matches,
- ▶ Big temporal changes can be associated to topics that do not participate to matching kernels.

Here splitting count is more important than merging count indicating a diversification in the field of optoelectronics.

Diachronic analysis exploiting feature maximization

Typical results – IST-PROMTECH dataset [Lamirel 12]

```
source cluster 12 [12/7]
- Stable labels Theory to practice
f1: 0.25911[23] f2: 0.129486[ 2] Conducting polymers (***)
f1: 0.086864[23] f2: 0.129486[ 2] Conducting polymers (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.034510[23] f2: 0.000000[-1] Experimental study

- Highly dominant (or peculiar) labels in target period
f1: 0.072006[23] f2: 0.206426[ 2] Polymer films (***)
f1: 0.054435[23] f2: 0.114637[ 2] Polymer blends (***)
f1: 0.000000[-1] f2: 0.039558[ 2] Spin-on coating
f1: 0.000000[-1] f2: 0.028204[ 2] Polymerization
```

```
source cluster 24 [20/8]
- Stable labels New component
f1: 0.03837[15] f2: 0.000000[-1] Diamond

- Highly dominant (or peculiar) labels in source period
f1: 0.043265[15] f2: 0.000000[-1] MIS structure
f1: 0.026522[15] f2: 0.000000[-1] Diamond

- Highly dominant (or peculiar) labels in target period
f1: 0.061132[15] f2: 0.222402[24] Amorphous semiconductors (***)
f1: 0.054647[15] f2: 0.131473[24] Hydrogen (***)
f1: 0.000000[-1] f2: 0.067403[24] Selenium
f1: 0.000000[-1] f2: 0.039028[24] Plasma CVD coatings
```

```
source cluster 29 [29/7]
- Stable labels Theory to practice
f1: 0.035721[14] f2: 0.000000[-1] Laser (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.148633[14] f2: 0.057783[14] Semiconductor laser (***)
f1: 0.078080[14] f2: 0.033436[14] Laser diodes (***)
f1: 0.026498[14] f2: 0.000000[-1] Surface
f1: 0.026027[14] f2: 0.000000[-1] Waveguide laser

- Highly dominant (or peculiar) labels in target period
f1: 0.000000[-1] f2: 0.068895[14] Light sources
f1: 0.000000[-1] f2: 0.039487[14] Laser beam applications
f1: 0.000000[-1] f2: 0.029637[14] Vertical cavity laser
f1: 0.000000[-1] f2: 0.025024[14] VCSEL
```

```
source cluster 7 [7/13]
- No stable labels Vocabulary change

- Highly dominant (or peculiar) labels in source period
f1: 0.266901[24] f2: 0.068167[33] Optical fabrication (***)
f1: 0.045998[24] f2: 0.000000[-1] Integrated circuit technology
f1: 0.042258[24] f2: 0.000000[-1] Interference filter
f1: 0.041773[24] f2: 0.000000[-1] Semiconductor technology

- Highly dominant (or peculiar) labels in target period
f1: 0.077799[24] f2: 0.213749[33] Optical design techniques (***)
f1: 0.000000[-1] f2: 0.055834[33] Aberrations
f1: 0.000000[-1] f2: 0.039636[33] Ray tracing
```

```
source cluster 16 is vanishing
f1: 0.141849[16] f2: 0.000000[-1] Optical fiber
f1: 0.078762[16] f2: 0.000000[-1] Fiber laser
f1: 0.060706[16] f2: 0.000000[-1] Acoustooptical device
f1: 0.049628[16] f2: 0.000000[-1] Ring laser
```

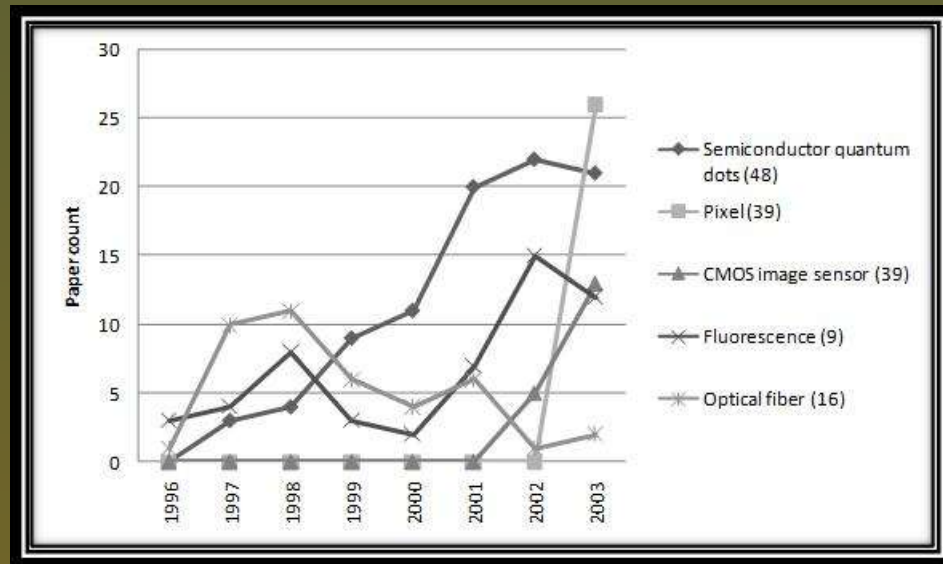
```
target cluster 9 is appearing
f1: 0.035520[ 5] f2: 0.160462[ 9] Fluorescence
f1: 0.000000[-1] f2: 0.082686[ 9] Phosphorescence
f1: 0.063888[ 1] f2: 0.105132[ 9] Exciton
```

```
target cluster 39 is appearing
f1: 0.000000[-1] f2: 0.144184[39] Pixel
f1: 0.000000[-1] f2: 0.110076[39] CMOS image sensors
f1: 0.000000[-1] f2: 0.077578[39] Chip
f1: 0.000000[-1] f2: 0.060044[39] High sensitivity
```

Diachronic analysis exploiting feature maximization

Validation techniques

CLUSTER REF.	TOPIC MAIN KEYWORDS	FEATURE F-MEASURE DIFFERENCE BETWEEN PERIODS	PAPER COUNT IN PERIOD 1 (1996-1999)	PAPER COUNT IN PERIOD 2 (2000-2003)
16	Optical fiber	0.14	28	13
9	Fluorescence	0.12	18	36
39	CMOS image sensors	0.11	0	18
39	Pixel	0.14	0	26
48	Semiconductor quantum dots	0.23	16	74



Validation is performed by querying database with topics main labels (keywords).

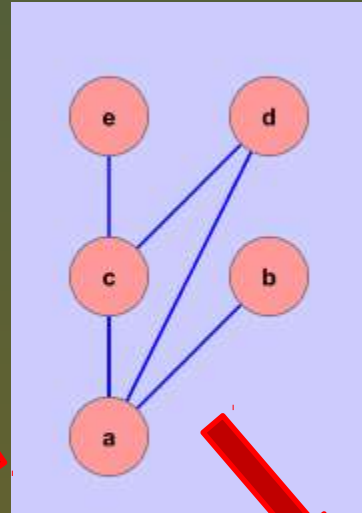
Contrast graphs

Principle

- ❖ Contrast graphs are bipartite graphs based on the relations between a set of features S and a set of labels L [Lamirel 2013].
- ❖ Theoretically, the set of labels L could represent any kind of information to which features can be related with and the set of features S is a subset of a global feature set F (i.e. the original feature space on which rely the data of a dataset) that has been obtained through a feature selection process, like feature maximization.
- ❖ In the case of the use of feature maximization, the weight $c(u,v)$ of an edge (u,v) , $u \in S$, $v \in L$ represents the contrast of³²

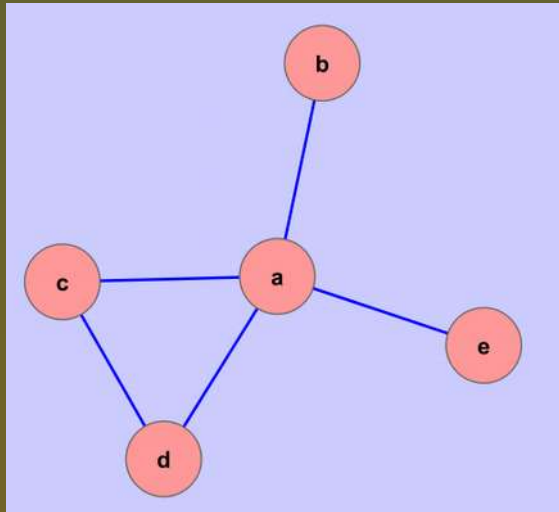
Building contrast graphs

Node S	Node T	Edge Weight
a	b	1
a	c	4
a	d	4
a	e	10
c	d	4

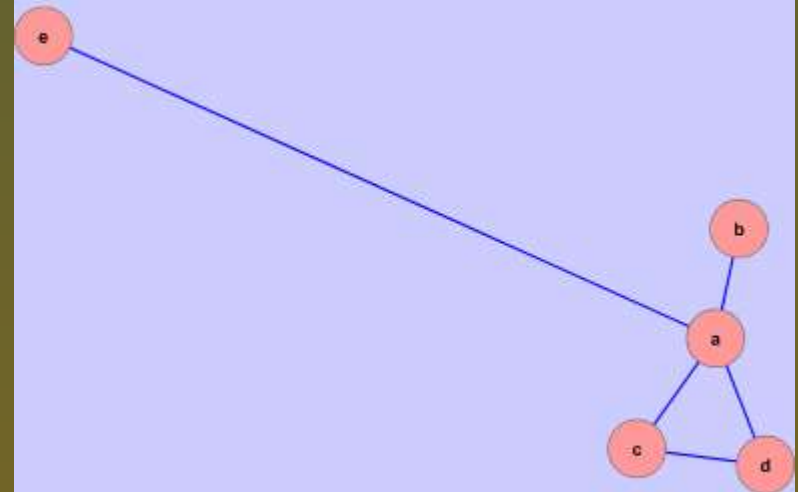


Contrast issued from feature maximization metric is used for materializing force.

Force directed graph

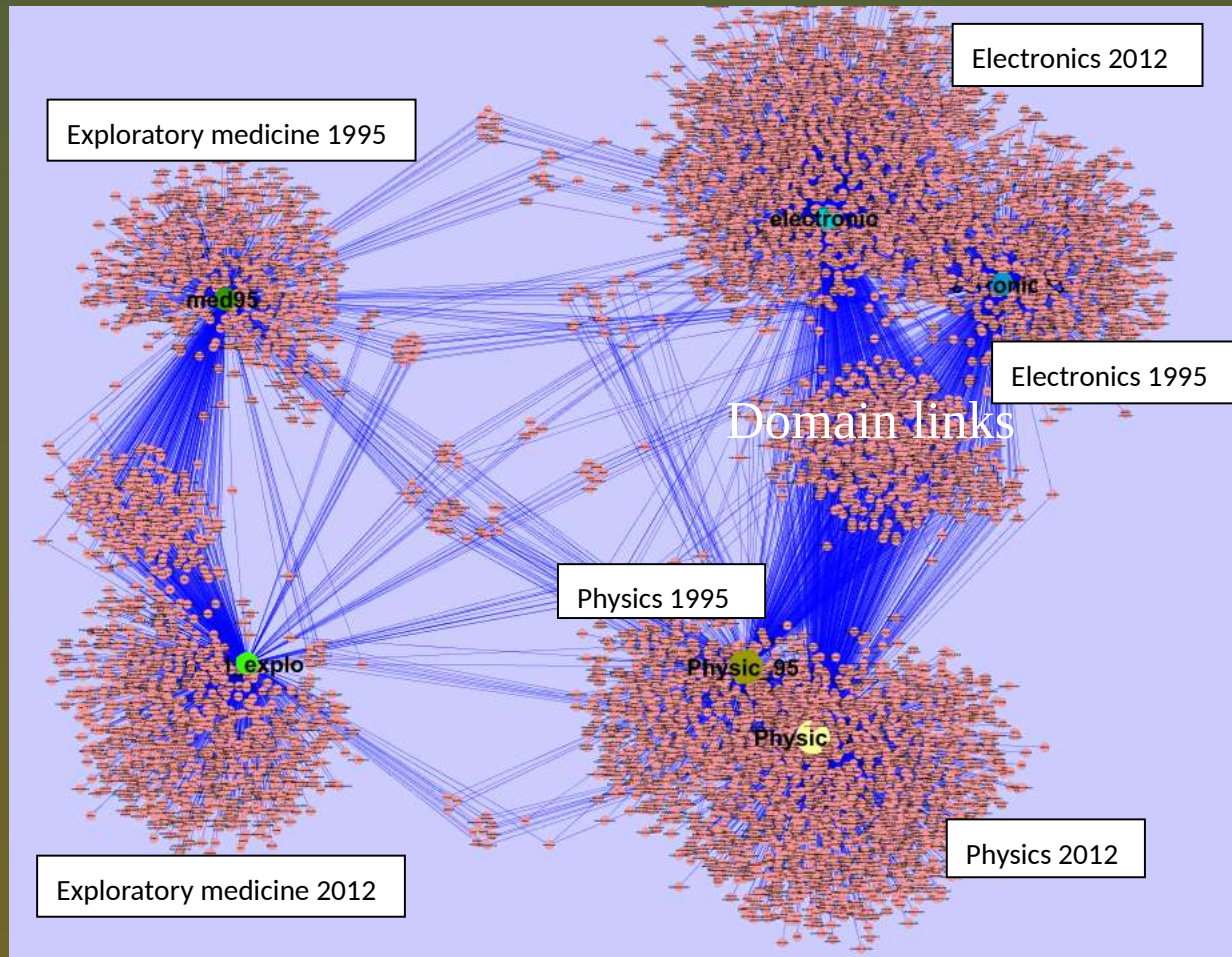


Force directed weighted graph



Building contrast graphs

Experiences on social network analysis



Strong links of different types can be highlighted

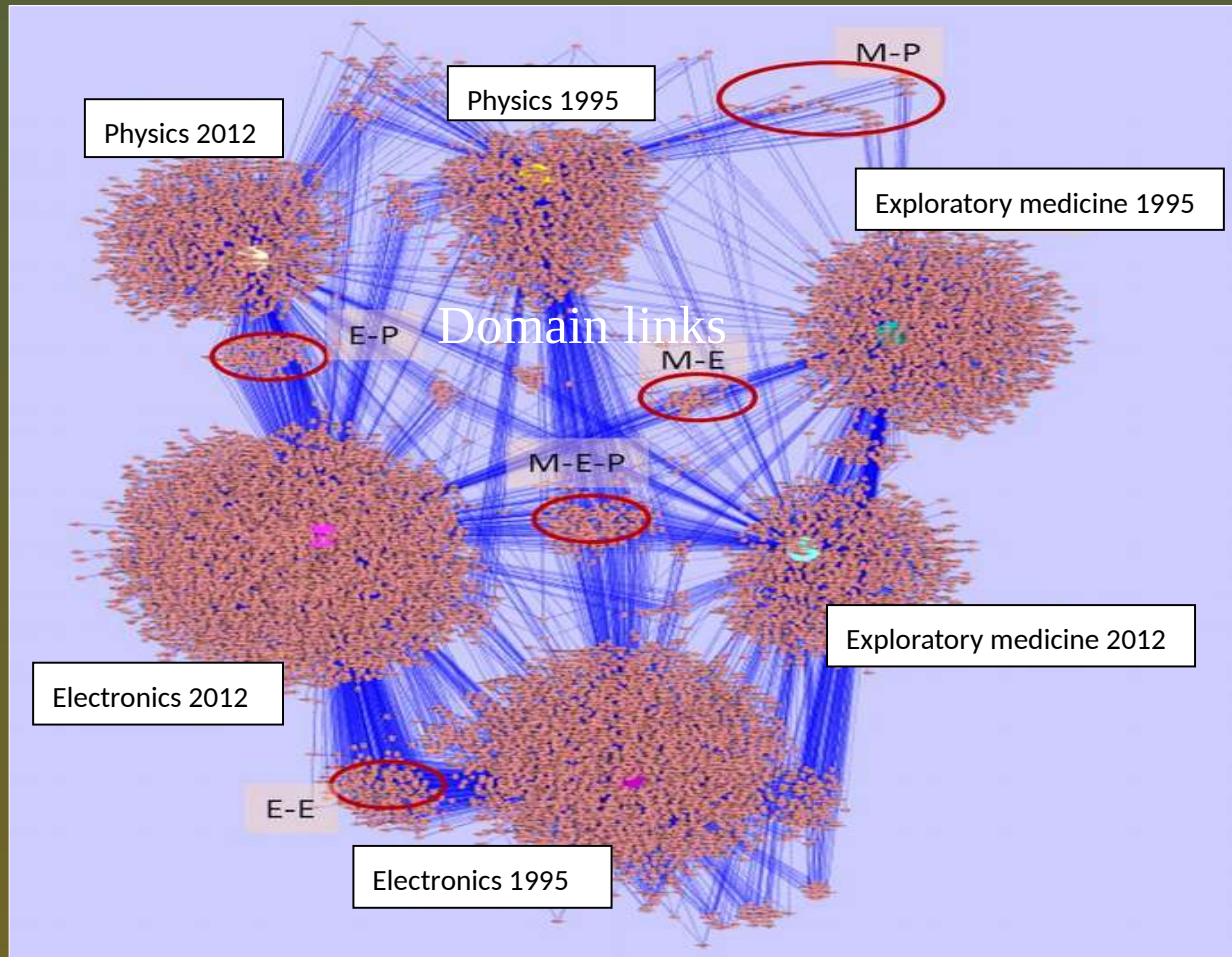
Multidisciplinary links

- **Electronics – Physics**
- Electronics-Medicine
- Electronic-medicine

Diachronic links
Medicine-Medicine

Building contrast graphs

Experiences on social network analysis



Authors graph with contributor-topic links highlight the transmitters of knowledge

1- Between disciplines

2 - Between periods

Can be used to analyze and exploit social tagging

Cluster quality evaluation

Pending problems

- ❖ Most of the quality indexes are based on Euclidean distance,
- ❖ Behavior of indexes is analyzed on low dimensional problem and results are often contradictory [Liu et al 2011],
- ❖ Min-square error optimization have been proven to be unable to solve complex clustering problem [Lamirel 2011],
- ❖ Min-square and Euclidean distance based indexes are unable to produce optimal results in high dimensional context (CH and DB) [Kassab et al. 2006] [Ghribi et al. 2010],
- ❖ Most of the realistic problems are not low dimensional problems with well-shaped clusters with more or less low overlap,
- ❖ Clustering methods are imperfect and error-prone,
- ❖ Indexes results depends on the methods [Lamirel 2004].

Quality evaluation using full-feature maximization (Principle)

- ❖ A good clustering model should be able to maximize sum of positive contrast in clusters (≈ generic intra-clusters inertia):

$$PC_k = \frac{1}{k} \sum_{k=1}^n \frac{1}{n_k} \sum_{f \in S_k} G_c(f)$$

- ❖ A more complete approach could combine summation of positive and summation of invert of negative contrast
- ❖ A more complete approach could combine summation of positive and summation of invert of negative contrast (≈ generic intra-cluster and inter-cluster inertia):

(≈ generic intra-cluster and inter-cluster inertia):

$$EC_k = \frac{1}{k} \sum_{k=1}^n \left(\frac{\frac{|S_k|}{n_k} \sum_{f \in S_k} G_c(f) + \frac{|\overline{S}_k|}{n_k} \sum_{h \in \overline{S}_k} \frac{1}{G_c(h)}}{|S_k| + |\overline{S}_k|} \right)$$

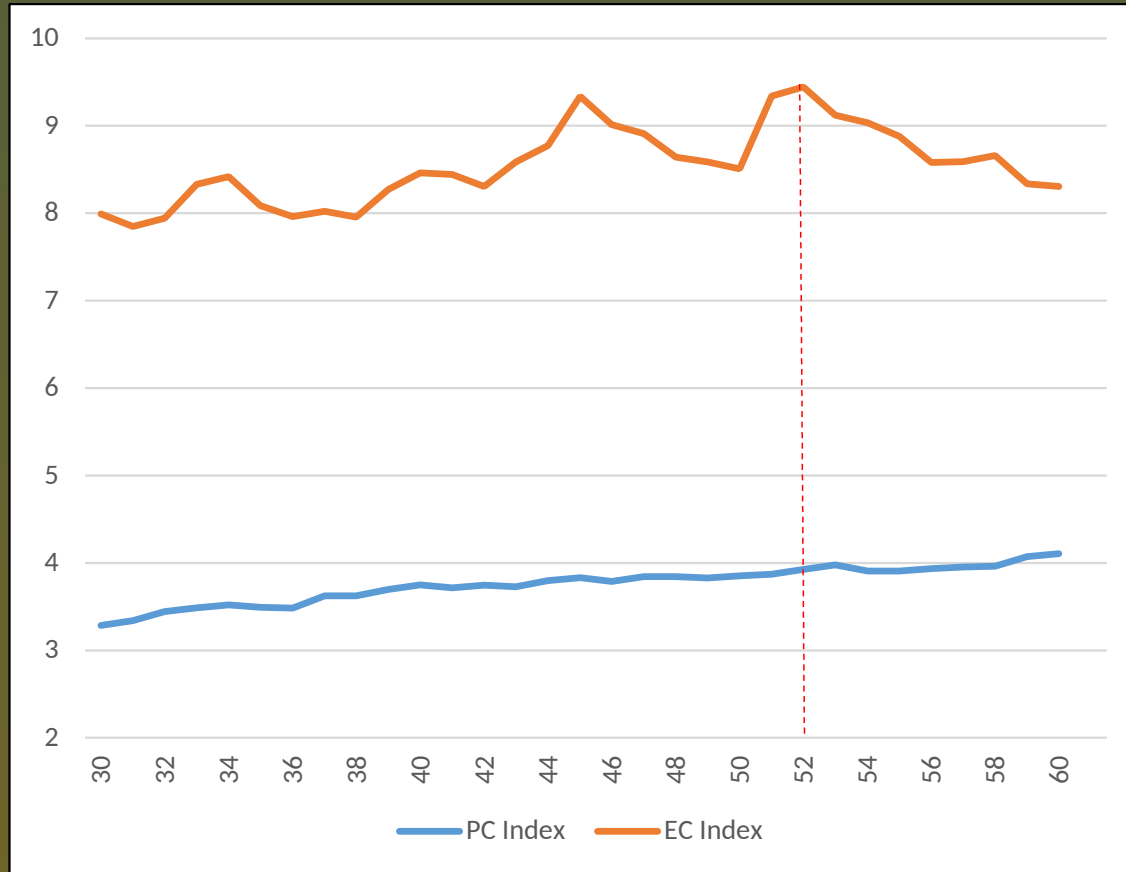
Quality evaluation using feature maximization (Results)

	IRIS	IRIS-B	WINE	PEN	ZOO	VRBF	R8	R52	Number of correct matches
DB	2	5	5	7	8	-out-	4	-out-	2/8
CH	2	3	6	8	4	-out-	6	-out-	2/8
DU	1	1	8	17	8	2	-out-	-out-	1/8
SI	4	2	7	14	4	-out	-out-	54	2/8
XB	2	7	-out-	19	-out-	23	-out-	-out-	0/8
PC	3	3	4	9	7	18	-out-	-out-	4/8
EC	3	3	4	9	7	15	6	52	7/8
MaxP	3	3	5	11	10	12-16	6	45-55	
Method	K-means	K-means	GNG	GNG	IGNGF	IGNF	IGNGF	IGNGF	

EC, and more especially PC, clearly outperform other approaches in realistic optimal clustering model evaluation.

Computation time is low: EC = 125s – SI = 43000s on R52

Quality evaluation using full-feature maximization (Index divergence case)



An index is divergent if it does not find the optimal model in a reasonable range around ground truth.

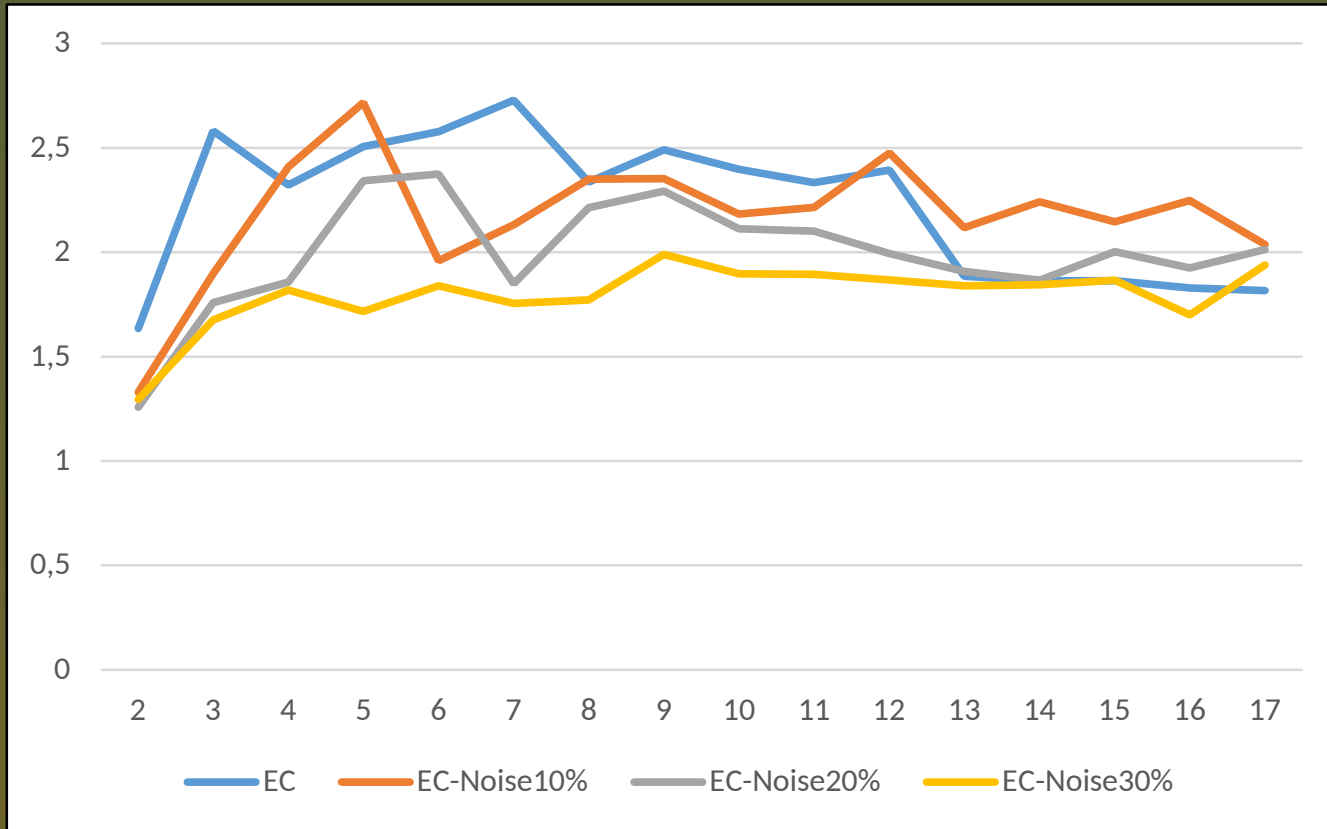
Quality evaluation using feature maximization (Resistance to noise)

- ❖ Data of clusters are migrated in a random way to other clusters to different fixed amount for all model sizes,
- ❖ Indexes are recalculated on noised models to look for potential variation in their behavior,
- ❖ This experiment highlights robustness to weak clustering results.

	ZOO	ZOO Noise 10%	ZOO Noise 20%	ZOO Noise 30%	Number of correct matches
DB	8	4	3	3	1/4
CH	4	5	3	3	0/4
DU	8	2	2	2	1/4
SI	14	-out-	-out-	-out-	0/4
XB	-out-	-out-	-out-	-out-	0/4
PC	6	4	11	9	1/4
EC	7	5	6	9	2/4
MaxP	10	7	10	10	
Method	IGNGF	IGNGF	IGNGF	IGNGF	IGNGF

EC, and more especially PC, never get “out of work” even when noise is increasing to a significant extent.

Quality evaluation using feature maximization (Resistance to noise)



PC index behavior is smoothing (i.e. degrading) progressively with noise.

Quality evaluation using feature maximization (Validation)

Criteria 1: number of matches

Criteria 2:

$$QMA = \sum_{i,j \in M} |S_{ij}| * \frac{P(i|j) + P(j|i)}{2}$$

where M represents the set of couple of clusters for which a match is detected.

In temporal matching process, hypothesis is that an accurate model selection will produce the larger number of matches, with matching kernels of the largest sizes and with the highest matching probability.

We consequently exploit two complementary criteria for the evaluation of the behavior of the indexes.

Quality evaluation using feature maximization (Validation)

	Opt. Period P1	Opt. Period P3	Opt Period P3	Number of temporal matches	QMA evaluation criteria
DB	-out-	-out-	-out-	0	0
CH	3	4	4	5	15.26
DU	14	20	-out-	6	8.60
SI	-out-	-out-	-out-	0	0
XB	-out-	-out-	-out-	0	0
PC	23	323	-out-	9	10.61
EC	10	6	8	13	27.20

Temporal matching results confirm the better performance of EC index as compared to other indexes.

Specific challenges on ISTEEX data for diachronic analysis

ISTEEX-R WP1

- ❖ ISTEEX data are issued from different editors, and there is no standardization of metadata or even no available metadata in some cases
- ❖ The exploited method must be able to tackle with large collection in an unsupervised way (time efficiency + a few of even no parameters)
- ❖ Overall time period lengths including stable topics can vary over time
- ❖ Visualization of diachronic changes is still on open problem

A first subset of ~10000 papers related to health care is extracted to perform a feasibility study on diachronic analysis (and other tasks of ISTEEX-R project)

A paper sample

CLINICAL AND RESEARCH REPORTS

Psychogeriatric Services at Certified Home Health Agencies

Case Reports and Guidelines for Psychiatric Consultants

Gary J. Kennedy, M.D.
Nelly Katsnelson, M.D.
Leila Laitman, M.D.
Ernesto Alvarez, M.S.W.

Because of the unmet mental health needs of older persons in the community, Medicare-certified home health agencies are increasingly taking the role of health providers. Here the authors review their experience and argue that the pathology seen in home mental health care situations is similar to that seen by specialized mental health outreach teams. Also the relations between the home care team and the psychiatric consultant require skillful management even when the team are mental health specialists. The authors offer guidelines for psychiatric consultants, given the extent to which home care services survive in a volatile, cost-contained environment. (American Journal of Geriatric Psychiatry 1995; 3:339-347)

Considerable information is available on specialized outreach programs for

mentally ill elderly patients,¹⁻⁷ but these programs are not in the context of Medicare-certified home health agencies, which have traditionally limited their services to nursing, social work, and physical or occupational therapy.⁸ Existing studies of outreach to mentally ill elderly patients are descriptive, with few controlled comparisons of interventions, personnel, or outcomes that might be used to establish the indications, benefits, cost offsets, or critical aspects of team composition. Nonetheless, they demonstrate a compelling need and document a variety of practical interventions and viable team configurations.

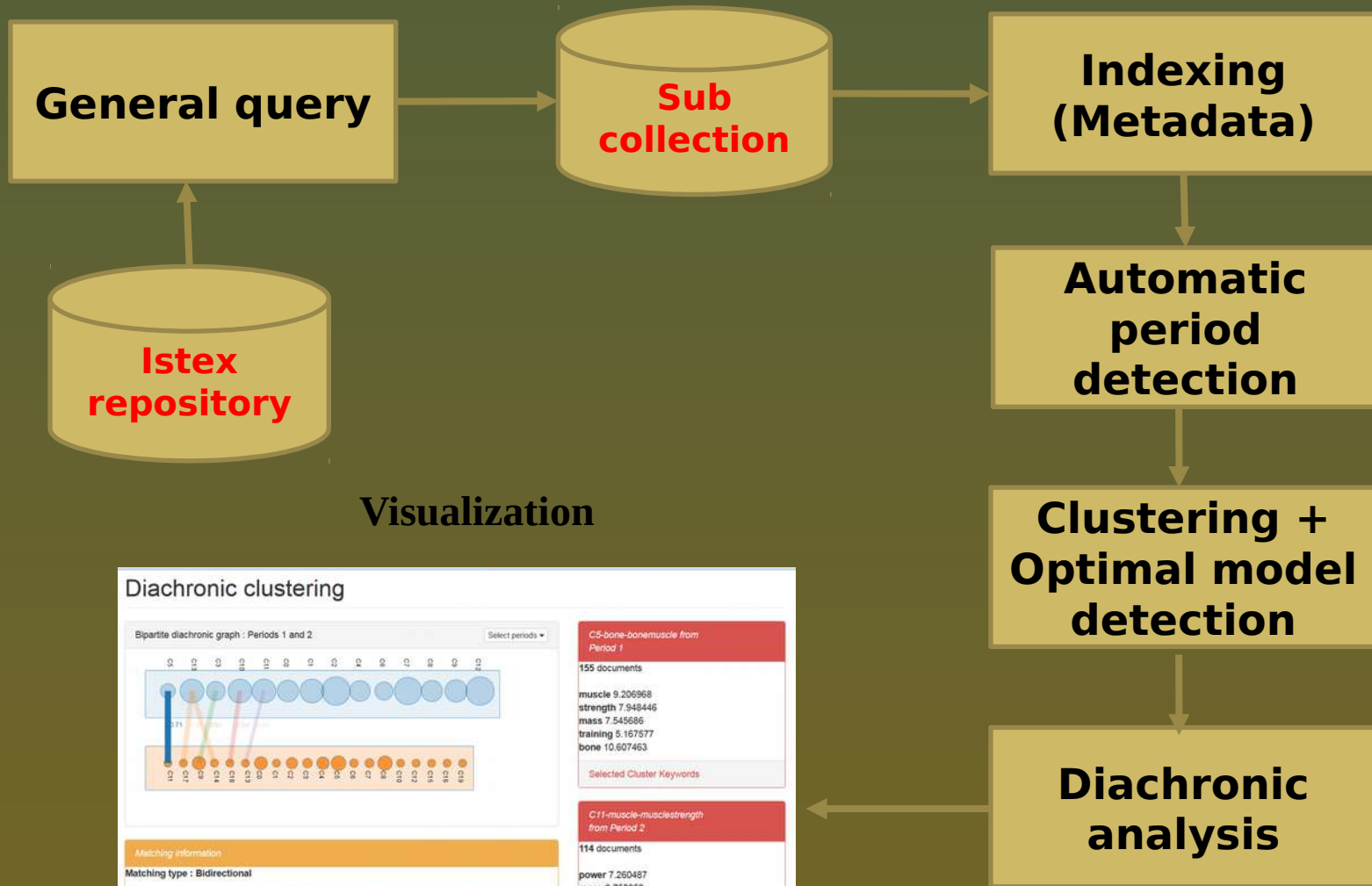
Less than 5% of older community residents in need of mental health services receive care.⁹ Because of the stigma of mental illness and biases about the efficacy of mental health services in old age, older adults are less likely to be offered and to accept a referral for psychiatric care.¹⁰ Also, older adults are reluctant to seek services for fear that disclosing their impairments would sacrifice their liberty.¹¹ Case-finding by social service agencies is also inadequate in that area agencies on aging and the mental health delivery systems lack systematic linkage.¹² As a result, Medicare-certified home health agencies are likely to encounter substantial unmet mental health needs among their older clients.

Among nonpsychiatric home care studies, measures of cost offsets, mortality, functional status, cognition, and rates of nursing home admission yield equivocal results. Failure to target appropriate patients and to manage care and the care team may account for the observation that more home care means more cost without much improvement in the older person's functioning.¹³

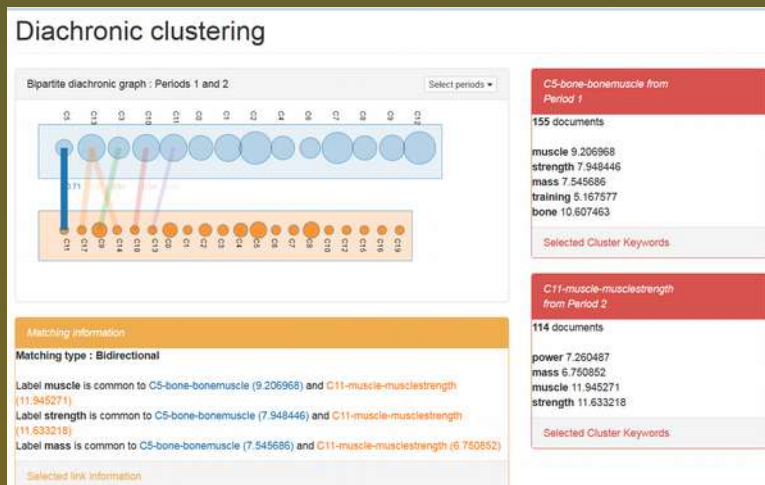
Each paper has an associated XML description file (metadata).

Received May 4, 1994; revised January 9, 1995; accepted March 14, 1995. From Montefiore Medical Center/Albert Einstein College of Medicine. Address correspondence to Dr. Kennedy, Department of Psychiatry, Montefiore Medical Center, 111 East 210th Street, Bronx, NY 10467.
Copyright © 1995 American Association for Geriatric Psychiatry

Overall view of the methodology



Visualization



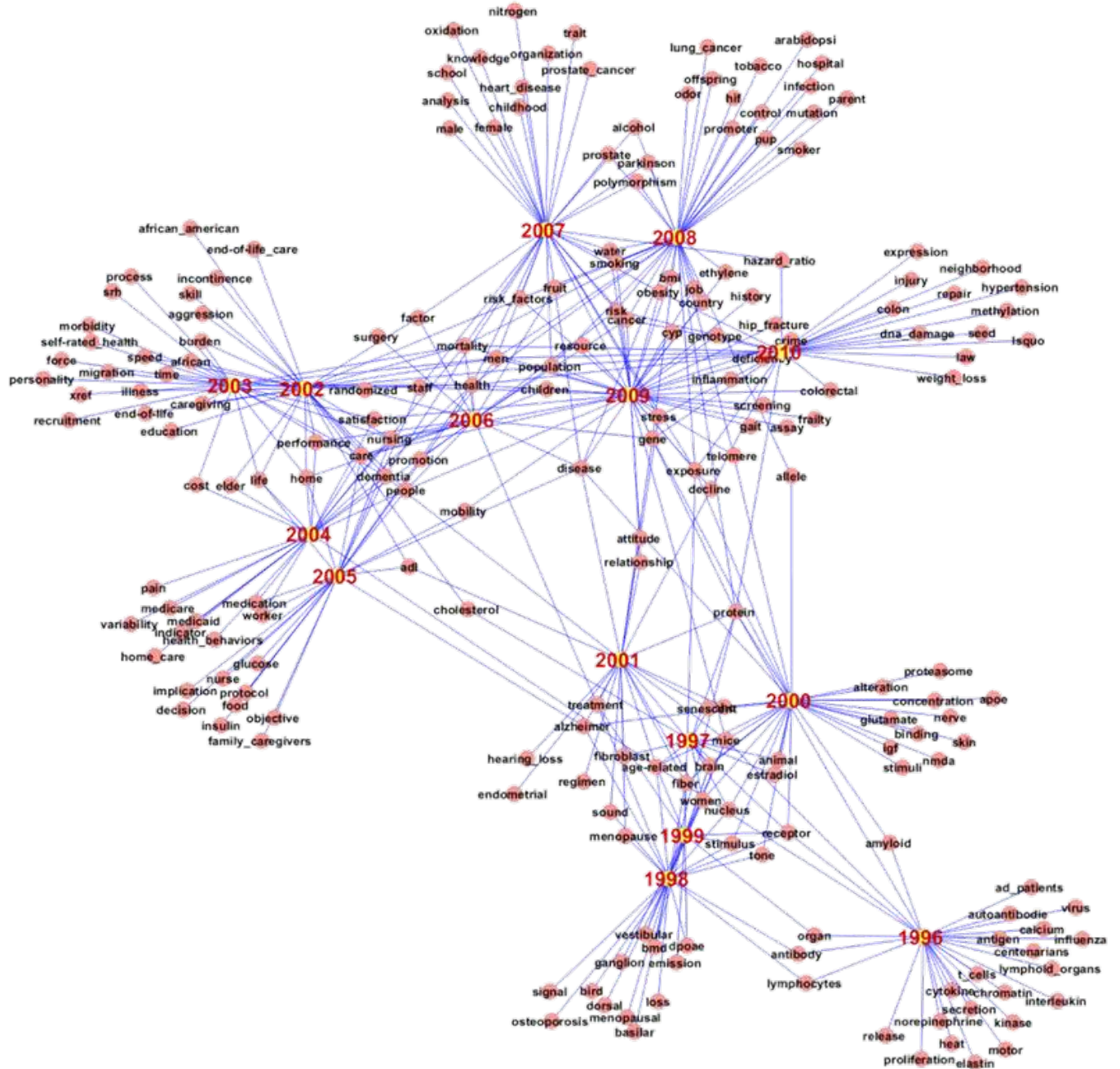
Data indexing

- ❖ 1) From title + abstracts
- ❖ **2) From full text ****
- ❖ **Part-of-Speech method **** (Python) :
 - ❖ Tokenization
 - ❖ Tagging
 - ❖ Lemmatization
 - ❖ Stop words list
 - ❖ Inadequate strings cleansing
- ❖ Comparison with state-of-art indexing methods (Termsuite, Rake, ...)

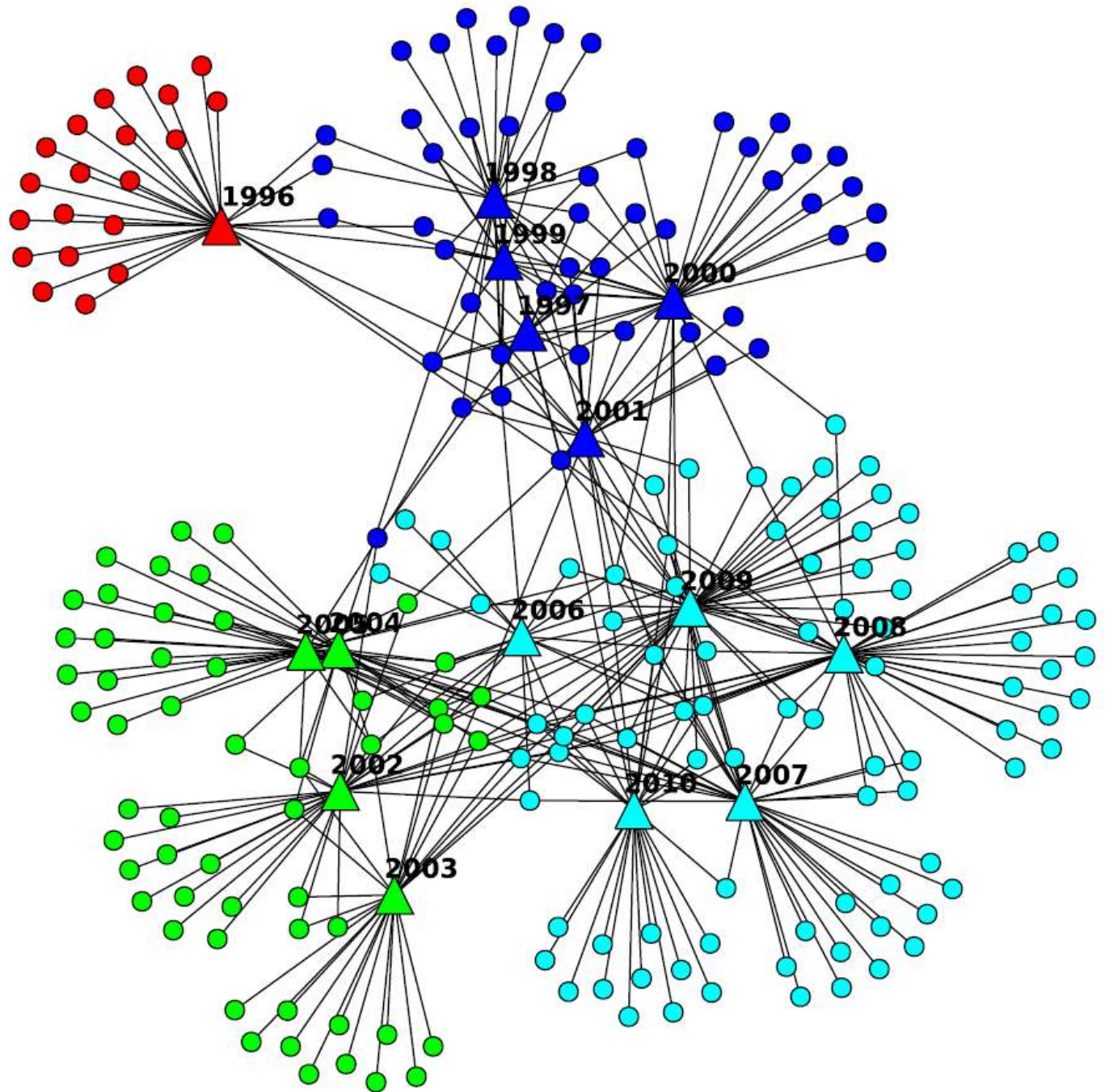
Automatic meta-period detection

- ❖ From global dataset and on the basis of a Year-Terms matrix, feature selection based on feature maximization is applied and a contrast graph is build up
- ❖ The state-of the art graph partitioning methods are tested on the graph and their parameters are adjusted :
 - ❖ FastGreedy
 - ❖ SpinGlass
 - ❖ MCL
 - ❖ **Walktrap (adjustment of walkstep) ** [Pons et al. 2005]**
 - ❖ ...

Filtered
bipartite
Year-Terms
graph
(for a readable
representation).



Bipartite
Year-Terms
graph
partitioned by
periods.

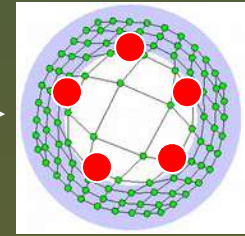
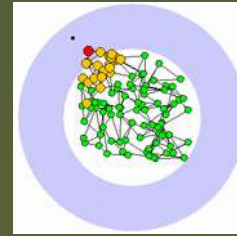


Clustering and detection of changes

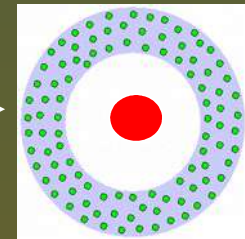
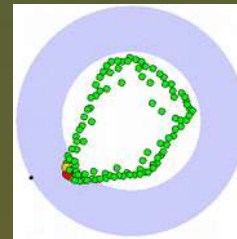
- ❖ Many clustering methods are tested :
 - ❖ K-means
 - ❖ SOM
 - ❖ **GNG ****
 - ❖ IGNGF
 - ❖
- ❖ Optimal model of each period is highlighted using feature maximization metrics
- ❖ Detection of changes is performed using unsupervised Bayesian reasoning on features and contrast measures

Neural Clustering Methods

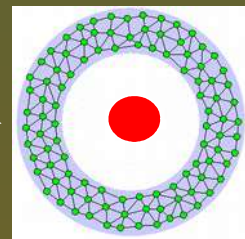
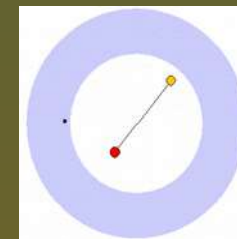
- ❖ Kohonen Self Organizing Map (SOM) [Kohonen 82]
fixed topology



- ❖ Neural Gas (NG) [Martinetz 91]
free topology



- ❖ Growing Neural Gas (GNG) [Fritzke 95]
free topology + periodically changing neuron count



- Neural clustering methods are less dependant to initial condition than classical clustering methods
 - “Incremental” versions of SOM or GNG exist :
 - ISOM [Merk 03];
 - IGNG [Prud 05], I²GNG [Hamz 08].
- ⇒ **Without specific adaptation**, original SOM method is the only one to provide **embedded visualization mechanism**.

Clustering and detection of changes (1)

2	2	*1.610*	-1.521-	75	37.50	4.992	75	37.50	-2.008-
3	3	*2.046*	-1.943-	292	97.33	19.311	446	148.67	-8.549-
4	4	*2.366*	-2.121-	361	90.25	18.449	675	168.75	-9.754-
5	5	*2.649*	-2.232-	416	83.20	17.549	879	175.80	-10.246-
6	6	*2.851*	-2.326-	492	82.00	14.304	1087	181.17	-10.572-
7	7	*3.123*	-2.408-	548	78.29	14.463	1302	186.00	-10.892-
8	8	*3.404*	-2.476-	620	77.50	14.587	1517	189.62	-11.073-
9	9	*3.542*	-2.502-	685	76.11	16.266	1708	189.78	-12.326-
10	10	*3.749*	-2.548-	756	75.60	17.705	1905	190.50	-13.399-
11	11	*3.884*	-2.509-	837	76.09	16.236	2119	192.64	-12.349-
12	12	*4.117*	-2.577-	881	73.42	15.855	2348	195.67	-12.232-
13	13	*4.187*	-2.573-	952	73.23	13.088	2536	195.08	-10.218-
14	14	*4.425*	-2.592-	1000	71.43	14.289	2741	195.79	-11.162-
15	15	*4.537*	-2.630-	1068	71.20	14.988	2917	194.47	-11.676-
16	16	*4.740*	-2.668-	1090	68.12	14.085	3112	194.50	-11.124-
17	17	*5.010*	-2.707-	1171	68.88	13.397	3439	202.29	-10.682-
18	18	*5.124*	-2.694-	1248	69.33	12.076	3623	201.28	-9.672-
19	19	*5.223*	-2.726-	1282	67.47	12.251	3783	199.11	-9.840-
20	20	*5.240*	-2.694-	1383	69.15	12.453	3937	196.85	-9.916-
21	21	*5.509*	-2.710-	1450	69.05	12.232	4186	199.33	-9.782-
22	22	*5.582*	-2.739-	1489	67.68	12.871	4327	196.68	-10.277-
23	23	*5.683*	-2.754-	1519	66.04	12.318	4507	195.96	-9.907-
24	24	*5.685*	-2.708-	1600	66.67	12.697	4690	195.42	-10.156-
25	25	*5.825*	-2.744-	1671	66.84	12.597	4866	194.64	-10.079-

5.417783 muscle
5.327384 strength
5.217635 exercise
4.168652 power
3.841000 mass
3.776102 training
2.814531 body
2.502115 performance

2.466132 week
2.371438 weight
1.897246 participant
1.857157 gerontol
1.775870 month
1.752784 woman
1.688477 measurement
1.677586 baseline
1.648806 biol
1.637920 function
1.625222 test
1.599973 decrease
1.586033 animal
1.529848 colleague
1.528399 age-related
1.489449 disability
1.482194 group
1.458563 adult
1.434127 change
1.403684 increase

Optimal model of each period is highlighted and feature with over-average contrast of the clusters of optimal model are isolated.

Clustering and detection of changes (1)

2	2	*1.610*	-1.521-	75	37.50	4.992	75	37.50	-2.008-
3	3	*2.046*	-1.943-	292	97.33	19.311	446	148.67	-8.549-
4	4	*2.366*	-2.121-	361	90.25	18.449	675	168.75	-9.754-
5	5	*2.649*	-2.232-	416	83.20	17.549	879	175.80	-10.246-
6	6	*2.851*	-2.326-	492	82.00	14.304	1087	181.17	-10.572-
7	7	*3.123*	-2.408-	548	78.29	14.463	1302	186.00	-10.892-
8	8	*3.404*	-2.476-	620	77.50	14.587	1517	189.62	-11.073-
9	9	*3.542*	-2.502-	685	76.11	16.266	1708	189.78	-12.326-
10	10	*3.749*	-2.548-	756	75.60	17.705	1905	190.50	-13.399-
11	11	*3.884*	-2.509-	837	76.09	16.236	2119	192.64	-12.349-
12	12	*4.117*	-2.577-	881	73.42	15.855	2348	195.67	-12.232-
13	13	*4.187*	-2.573-	952	73.23	13.088	2536	195.08	-10.218-
14	14	*4.425*	-2.592-	1000	71.43	14.289	2741	195.79	-11.162-
15	15	*4.537*	-2.630-	1068	71.20	14.988	2917	194.47	-11.676-
16	16	*4.740*	-2.668-	1090	68.12	14.085	3112	194.50	-11.124-
17	17	*5.010*	-2.707-	1171	68.88	13.397	3439	202.29	-10.682-
18	18	*5.124*	-2.694-	1248	69.33	12.076	3623	201.28	-9.672-
19	19	*5.223*	-2.726-	1282	67.47	12.251	3783	199.11	-9.840-
20	20	*5.240*	-2.694-	1383	69.15	12.453	3937	196.85	-9.916-
21	21	*5.509*	-2.710-	1450	69.05	12.232	4186	199.33	-9.782-
22	22	*5.582*	-2.739-	1489	67.68	12.871	4327	196.68	-10.277-
23	23	*5.683*	-2.754-	1519	66.04	12.318	4507	195.96	-9.907-
24	24	*5.685*	-2.708-	1600	66.67	12.697	4690	195.42	-10.156-
25	25	*5.825*	-2.744-	1671	66.84	12.597	4866	194.64	-10.079-

5.417783 muscle
5.327384 strength
5.217635 exercise
4.168652 power
3.841000 mass
3.776102 training
2.814531 body
2.502115 performance

2.466132 week
2.371438 weight
1.897246 participant
1.857157 gerontol
1.775870 month
1.752784 woman
1.688477 measurement
1.677586 baseline
1.648806 biol
1.637920 function
1.625222 test
1.599973 decrease
1.586033 animal
1.529848 colleague
1.528399 age-related
1.489449 disability
1.482194 group
1.458563 adult
1.434127 change
1.403684 increase

Optimal model of each period is highlighted and feature with over-average contrast of the clusters of optimal model are isolated.

Clustering and detection of changes (2)



2	2.103293
3	4.894217
4	4.661046
5	4.283290
6	4.116885
7	4.485136
8	4.229456
9	4.093945
10	4.053818
11	4.025391
12	4.059147
13	3.866609
14	3.577087
15	3.776318
16	3.537132
17	3.608063
18	3.642788
19	3.514834
20	3.430515
21	3.327520
22	3.456252
23	3.239336
24	3.192178
25	3.271146

Davis-
Bouldin
(behaves in
a convex
way on the
considered
interval)



Termsuite
(provides
mostly
mixed
general
terms)

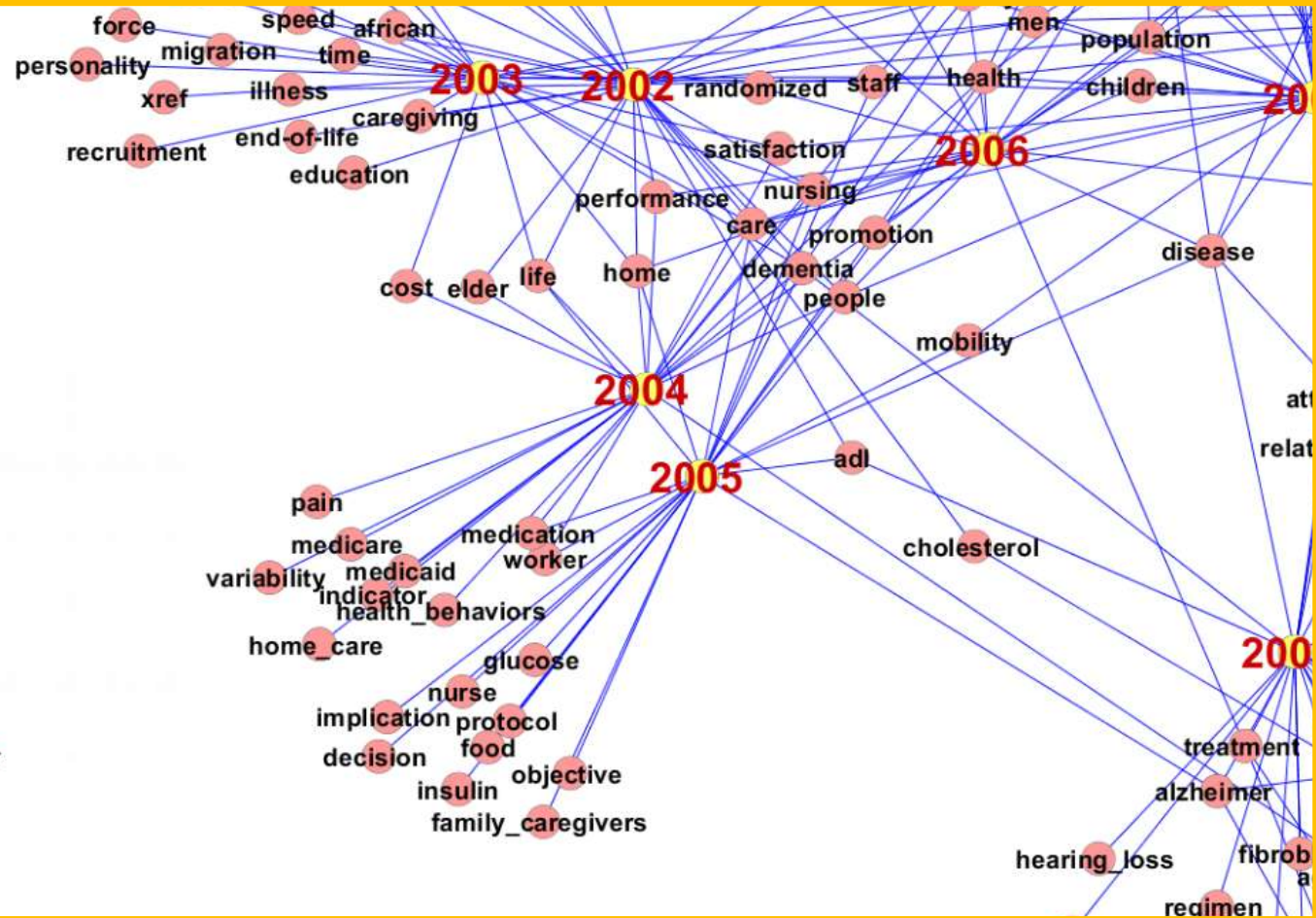
```
4.724185 americans
4.350677 gap
4.307484 cell
4.244061 circumstances
4.232458 health_condition
4.044825 instrument
3.659395 future
3.653448 threshold
3.189270 orientation
2.608994 problem
2.606182 frame
2.476067 protocol
2.188425 quantitative
*****
*****
2.042368 code
1.885890 absence
1.830976 admission
1.793681 term
1.759934 allele
1.686220 flexor
1.550164 mellitus
1.520950 fast
1.503231 flow
1.472898 medium
1.436958 scores
1.430196 editor
1.419317 paper
1.403544 notion
1.368368 understanding
1.295318 compute
1.262428 glucose
1.193671 church
1.159139 home_resident
1.138192 sign
1.091506 technology
1.076663 cope
1.036130 transportation
1.032563 hospitalize
1.013571 writing
1.001360 public_health
```



Bad indexing as well as bad quality indexes
produce unmanageable results.

On going work on F-max metric

Contrast graph (cont.)



Contrast graphs identify essential relations between categories and features also highlighting highly central features.

On going research on contrast graph

Another approach : direct use of contrast graph for diachrony

- ❖ Contrast graphs can be directly used to highlight diachronic paths
- ❖ The principle is to interpret the result of the clustering process (topic extraction) as a contrast graph in which only the salient properties of the clusters are considered
- ❖ The salient properties of the clusters are identified by the feature maximization process
- ❖ The optimal clustering model is detected by the use of quality measures based on feature maximization process as well



Founding : ANR-10-IDEX-0004-02

On going research on contrast graph

Another approach : direct use of contrast graph for diachrony

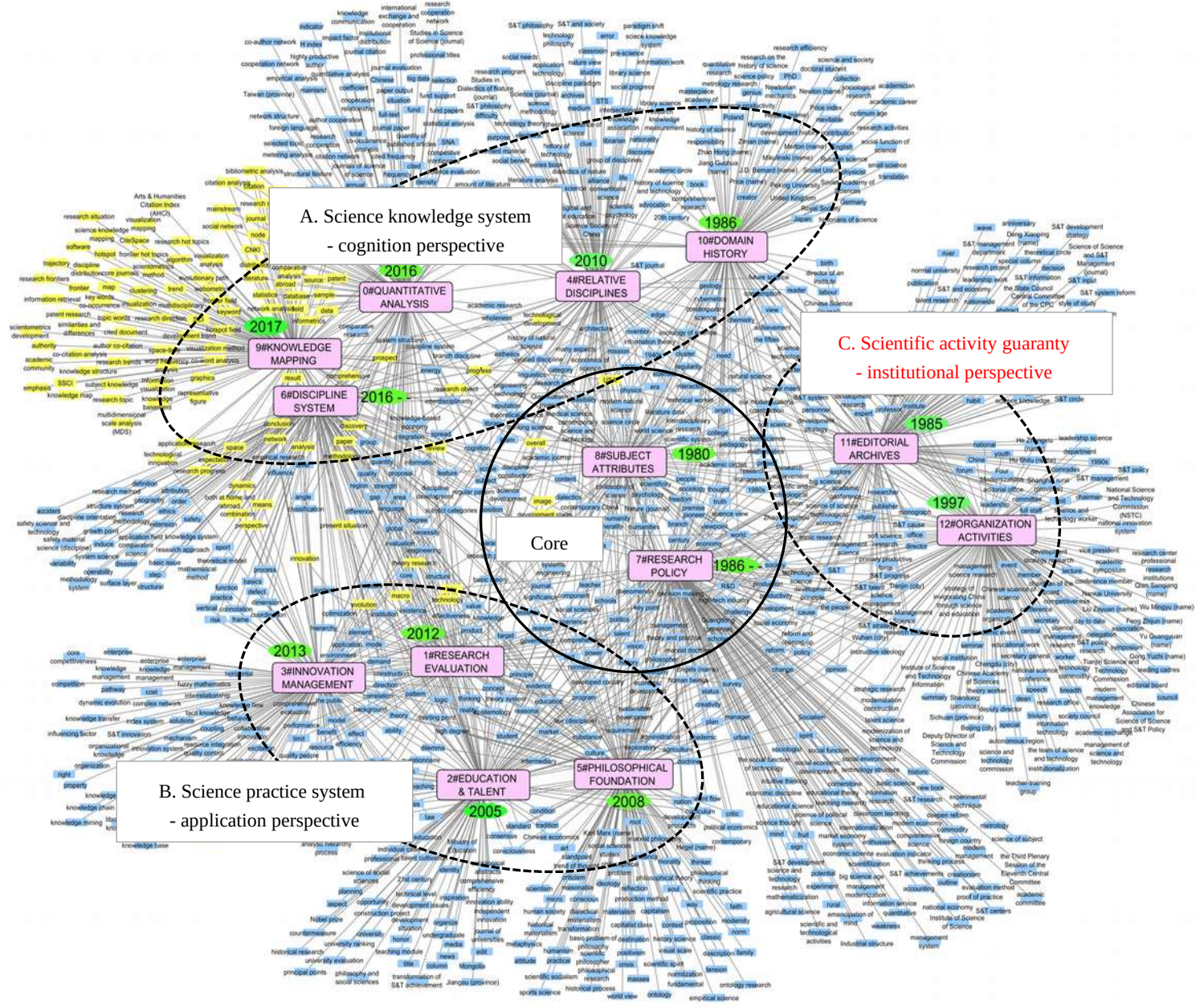
- ❖ Temporal tags of the clusters data (documents) can be used to identify the most influent periods in the clusters
- ❖ The older or root topics that are densely connected to the others tend to appear at the center of the obtained graph, the younger topics tend to appear at the periphery
- ❖ Further analysis of the temporal tags in the clusters can be performed to evaluate the influence of the clusters over time
- ❖ Such methodology has been applied for the study the 40 years old history of Science of Science in China [Chinese journal paper & upcoming Scientometrics paper]



IGTXX P 1011

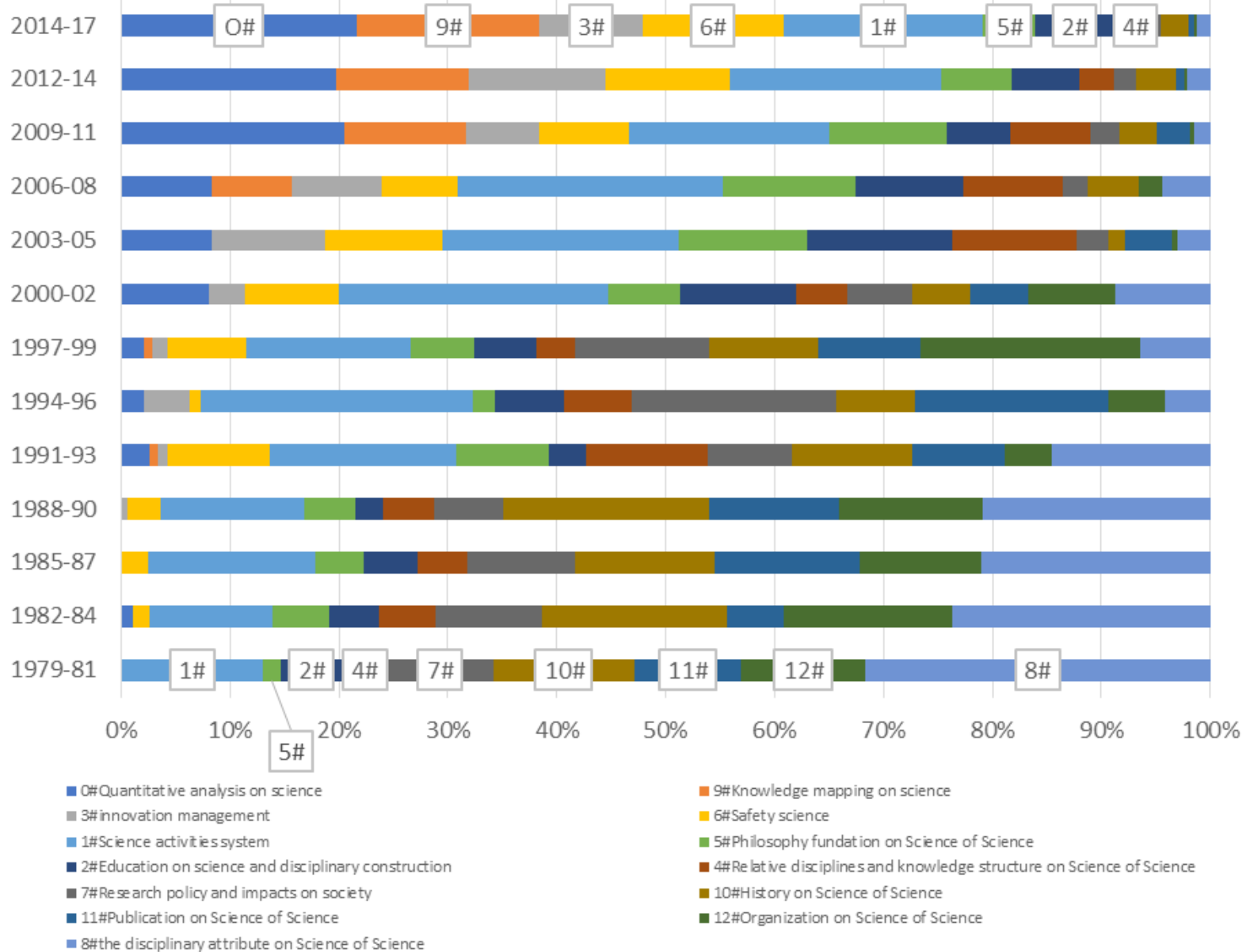
Funding : ANR-10-IDEX-0004-02

On going research on contract graph



04-02

On going research on contract graph

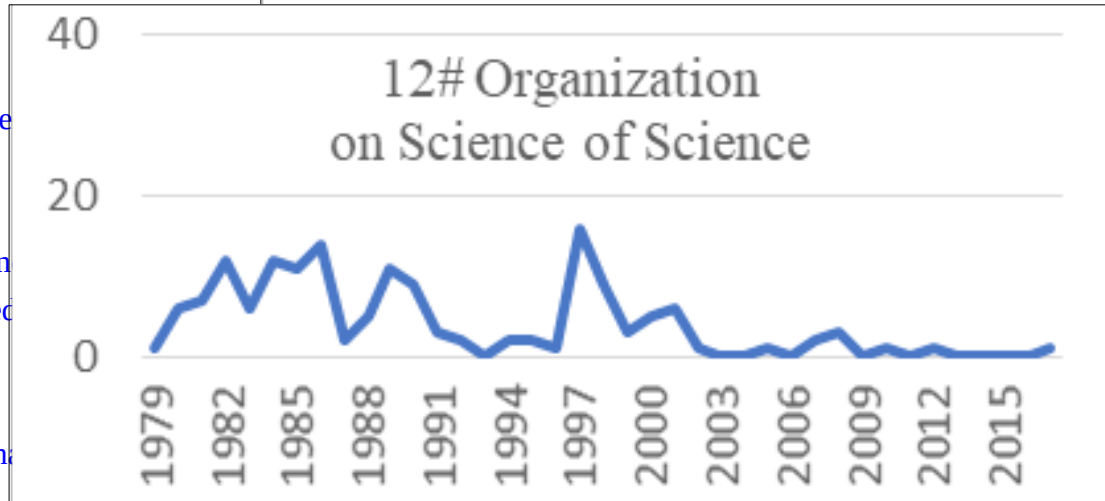


On going research on contrast graph ... : research topics and their evolution

- 5.376770 theme ,
- 5.030978 research hot topics ,
- 4.827424 literature ,
- 4.734794 software ,
- 4.697236 frontier ,
- 4.595268 development tre
- 4.401170 research topic ,
- 4.342141 hotspot ,
- 4.159228 both at home an
- 3.989917 science knowled
- 3.873852 international ,
- 3.801943 expectation ,
- 3.648744 visualization an
-
- 3.130080 trajectory ,
- 2.930680 clustering ,
- 2.872097 research trends ,
- 2.855751 topic words ,
- 2.779299 distribution ,
-,

40
20
9# Knowledge mapping on science

40
20
0
12# Organization on Science of Science



Chinese
that the

IST

method provide them with very fine
grained description of domain history,
evolution and domain topics interaction

Comparison with LDA

Topic extraction capabilities

- ❖ Topic extraction capabilities is a critical point in the diachronic analysis process
- ❖ Weak or improper topic extraction can lead to non interpretable or trivial results and put further diachronic analysis step into problems
- ❖ The more complex is this task, the higher is the difference between the methods
- ❖ Using a 10000 patents dataset related to IA and extracted from USPTO we manage to retrieve the main topic of IA by the comparative use of LDA and combination of GNG (ISTEX P-WP1) cluster labeling and optimal model maximization (cluster detection)



Founding : ANR-10-IDEX-0004-02

Comparison with LDA

Topic extraction capabilities

- Cluster 0: Autonomous and distributed systems
- Cluster 1: Control (industrial processes and vehicles)
- Cluster 2: Image processing
- Cluster 3: Diagnosis systems
- Cluster 4: Neural networks and neuro-inspired approaches and circuits
- Cluster 5: Prediction and forecasting (supervised learning – type 1)
- Cluster 6: Natural language processing (NLP)
- Cluster 7: Man/machine interfaces
- Cluster 8: Clustering and knowledge discovery (unsupervised learning)
- Cluster 9: Recommendation systems and personalized user's services and assistance
- Cluster 10: Genetic and evolutionary algorithms
- Cluster 11: Relevance evaluation and scoring/ranking
- Cluster 12: Classification (supervised learning - type 2)

NO.	Term	topic
T1	client.web.mobile.page.server.ontology.content.concept.knowledge, information.dialog.system.emotion.task.request.tag.statement, personality.data.context.application.device.message.service.engine, repository	Mobile Internet
T2	data.value.parameter.state.plurality.signal.time.device.system, network.unit.function.information.probability.module.event, resource.control.algorithm.component.configured.embodiment, score.sensor.performance.pattern	Signal processing and resource allocation
T3	spike.signal.unit.neuron.data.node.stream.synaptic.circuit.value, plurality.network.neural.deployment.character.word.keyword, variant.document.transition.theme.candidate.image.category, recipient.fuzzy.object	Neural networks Yes – but noise
T4	signature.markov.evidence.wireless.query.violation.validated.cache, sparse.posterior.splitting.application.categorical.queries.identity, destination.ui.recursively.search.merchant.fabric.optimizer	Verification and Application
T5	treatment.advertisement.website.topic.proximity.click.suggestion, request.rating.document.article.evolution.user.indexed.therapy, content.search.described.category.queries.compiling.genome.query, impression	Recommendation and search
	relevance.unstructured.sample.composite.scope.disparate.bound, harvest.fragment.signal.parametric.monte.carlo.taxonomy.framework, diverse.objects.generative.network.trie.data.customizable.clean, cached.minimize	Problem solving method

In the 10 topic it found, LDA method identify 2 only main topics (including one with high noise) among main topics of AI domain. Other found topics were either minor topics or a mix of general terms and noise.

Our proposed method put all the 12 main topics of AI into evidence with very low noise in their descriptions.

WP1

Founding : ANR-10-IDEX-0004-02

EGC-GAST-2019

Comparison with LDA

Topic extraction capabilities

NO.	Term	topic
T1	client,web,mobile,page,server,ontology,content,concept,knowledge,information,dialog,system,emotion,task,request,tag,statement,personality,data,context,application,device,message,service,engine,repository	Mobile Internet
T2	data,value,parameter,state,plurality,signal,time,device,system,network,unit,function,information,probability,module,event,resource,control,algorithm,component,configured,embodiment,score,sensor,performance,pattern	Signal processing and resource allocation
T3	spike,signal,unit,neuron,data,node,stream,synaptic,circuit,value,plurality,network,neural,deployment,character,word,keyword,variant,document,transition,theme,candidate,image,category,recipient,fuzzy,object	Neural networks Yes - but noise
T4	signature,markov,evidence,wireless,query,violation,validated,cache,sparse,posterior,splitting,application,categorical,queries,identity,destination,ui,recursively,search,merchant,fabric,optimizer	Verification and Application
T5	treatment,advertisement,website,topic,proximity,click,suggestion,request,rating,document,article,evolution,user,indexed,therapy,content,search,described,category,queried,compiling,genome,query,impression	Recommendation and search
T6	relevance,unstructured,sample,composite,scope,disparate,bound,harvest,fragment,signal,parametric,monte,carlo,taxonomy,framework,diverse,objects,generative,network,trie,data,customizable,clean,sched,minimize	Problem solving method

Cluster 4: Neural networks and neuro-inspired approaches and circuits

8.726535 neuron
 8.407532 gate
 7.735760 semiconductor
 7.689034 coupling
 7.525440 summing
 6.760579 cell
 6.605940 interconnected
 6.460970 circuit
 6.432827 synaptic
 6.344198 layer
 6.295610 charge
 6.184274 floating
 6.139169 interconnecting
 6.096223 sum
 6.089437 amplifier
 5.783809 differential
 5.763820 spike
 5.657769 synapse
 5.586802 connected
 5.582337 realized
 5.537529 layern
 5.480876 electrode
 5.438975 couple
 5.415030 intermediate
 5.389476 synapsis
 5.382641 activation
 5.373843 array
 5.276979 analog

The comparison clearly highlights topic precision of the proposed method as well as topic imprecision (i.e. low resolution capabilities) of LDA.

Founding : ANR-10-IDEX-0004-02

EGC-GAST-2019

Conclusion

- ❖ We present a fully unsupervised approach for diachronic analysis of large collection of heterogeneous text data. The approach is parameter-free and incremental and has superior performance as compared to state-of the art approaches (computation time, flexibility, stability of results, scalability, ...)
- ❖ We successfully propose a simplified variant of our original approach using contrast graphs. The new approach proved to be powerful enough to produce very relevant results on complex “real life” data

Conclusion

- ❖ Comparison of the topic extraction capabilities of the method with the ones of LDA clearly high a very significant difference with sufficiently complex data to deal with, such difference being critical for further accurate diachronic analysis
- ❖ Contrast graph are multiscale knowledge and parameter free approach that has also powerful application at the document level for automatic summarization and meta data extraction
- ❖ Feature maximization approach can also be successfully transposed to community detection in complex graph analysis

Contact and questions

emails:

lamirel@loria.fr, Pascal.Cuxac@inist.fr

Some references :

- 1) Lamirel J.-C. : A new diachronic methodology for automatizing the analysis of research topics dynamics : an example of application on optoelectronics research, *Scientometrics* 93(1): 151-166 (2012).
- 2) Lamirel J.C., Cuxac P., Chivukula A.S., Hajlaoui K. : Optimizing text classification through efficient feature selection based on quality metric. *Journal of Intelligent Information Systems*, May 2014, p.1-18, Springer.
- 3) Lamirel J.-C., Cuxac P. : ***New quality indexes for optimal clustering model identification with high dimensional data***, Proceedings of ICDM-HDM'15 - International Workshop on High Dimensional Data Mining, Atlantic City, USA, November 2015.