Les nouveaux paradigmes scientifiques : alternance de citations et d'oublis. Étude automatique sur une volumineuse bibliothèque numérique. Exemple de l'astrophysique.

Jean-Charles LAMIREL

LORIA – SYNALP, Vandœuvre-lès-Nancy, France
IUT Robert Schuman, Université de Strasbourg

Les challenges

- But : explorer un corpus de données scientifiques en mesurant les changements de sujets incluant la récurrence de sujets,
- Travailler sur un corpus de données issues de la base mutti-éditeurs ISTEX gérée par l'INIST,
- Mettre en place des techniques a la frontière de l'état de l'art :
  - Distances de compromis entre la généralité et la discrimination

  ⇒ Théorie de la maximisation des traits,
  - Travailler avec des vues multiples et des mécanismes de généralisation en ligne

  ⇒ Paradigme MVDA,
  - Intégrer la visualisation

  => Approche Diachronic'Explorer,
  - Intégrer les informations produites par les entités nommées dans le processus (en cours),
- Travailler à titre d'exemple sur des données du domaine de l'astronomie (en cours).

Métrique de maximisation des traits (MT)

[Lamirel et al. 2016]

Considérons un ensemble de données *D* représentées par un ensemble de variables (ou traits) *F*, et un ensemble de classes *C* résultant d'une méthode de regroupement. La métrique de maximisation des traits favorise les classes avec une valeur maximale de F-mesure de trait définie comme la moyenne harmonique entre :

Rappel de trait

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f} \equiv P(c|f)$$

Dominance de trait

$$FD_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F^c, d \in c} W_d^{f'}} \equiv P(f|c)$$

Une **variable de classe maximisée** est une variable dont la F-mesure de trait est **maximisée** par les membres de la classe (les données).

Un exemple illustratif simple

- Nous calculons le rappel de trait (FR), la dominance d'étiquetage (FD) et la F-mesure de trait (FF) pour chaque classe et chaque variable.

Taille Chaussures	Longueur cheveux	Taille Nez	Classe
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

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

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

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

Un exemple illustratif simple

- Nous calculons les valeurs moyennes de F-mesure de trait pour chaque variable (locale) et la F-mesure globale pour chaque classe et chaque variable de chaque classe.

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Longueur cheveux	0.39	0.66	0.53
Taille chaussures	0.48	0.22	0.35
Taille nez	0,3	0,24	0,27

$\overline{F(.,.)}$
0.38

Les variables dont la F-mesure est inférieure à la F-mesure globale sont retirées

⇒ La taille du nez est retirée

Les variables restantes (sélectionnées) dont la F-mesure est supérieure à la moyenne marginale dans une classe sont considérées actives dans cette classe

=> Taille des chaussures est active dans la classe Homme

=> Longueur des cheveux est active dans la classe Femme

Un exemple illustratif simple

- Le facteur de contraste met en évidence le degré d'activité/passivité des variables sélectionnées par rapport à leur F-mesure moyenne marginale dans les différentes classes.

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Long. cheveux	0.39	0.66	0.53
Taille chauss.	0.48	0.22	0.35

	C(x,M)	C(x,F)
Long. cheveux	0.39/0.53	0.66/0.53
Taille chauss.	0.48/0.35	0.22/0.35

Le contraste peut-être considéré comme une fonction qui aura tendance à:

1. Augmenter la longueur des cheveux des femmes
2. Augmenter la taille des chaussures des hommes
3. Diminuer la longueur des cheveux des hommes
4. Diminuer la taille des chaussures des femmes

	C(x,M)	C(x,F)
Long. cheveux	0.74	1.25
Taille chauss.	1.37	0.63

Propriétés de la maximisation des traits

- Mesure sans biais,
- Approche sans paramètres,
- Capacités explicatives en classification,
- Capacités de synthèse et d'extraction de sujets en cas de combinaison avec le clustering (approche très compétitive avec des méthodes connues comme LDA [Blei et al. 2003),
- Capacités d'analyse diachronique en combinaison avec d'autres paradigmes comme MDVA (…. suite).

Expériences sur des données textuelles de référence (20 newsgroups)

rec.autos	rec.motorcycles	sci.electronics	comp.graphics	sci.space
14.40 car	14.89 ride	13.69 circuit	12.22 imag	14.51 launch
14.01 ford	14.09 dog	13.42 wire	11.67 viewer	14.24 moon
10.54 auto	13.40 dod	11.89 outlet	11.03 graphic	13.53 mission
9.98 alarm	11.62 helmet	10.35 ham	9.88 render	12.87 space
9.79 shift	9.39 lean	9.92 concret	9.69 manipul	12.41 solar
9.74 mileag	9.23 chain	9.13 relai	8.50 pub	12.29 planet
9.20 oil	9.22 cage	9.03 neutral	8.34 plot	11.81 satellit
8.64 gear	9.11 cit	8.82 tone	8.34 gif	11.79 atmospher
7.96 tire	8.89 drink	8.81 ground	7.77 crop	11.17 sky
7.87 transmiss	8.49 newbi	8.61 led	7.74 format	10.82 henri

sci.med	talk.politics.guns	talk.politics. mideast	soc.religion. christian	alt.atheism
14.67 patient	14.43 gun	12.07 occupi	10.66 bless	8.10 keith
13.24 medic	10.18 cdt	11.19 villag	10.54 rutger	7.84 societ
12.60 doctor	10.08 accident	11.02 soldier	9.58 sin	7.78 moral
12.00 food	9.51 revolv	10.75 territori	9.05 marri	7.34 belief
11.65 medicin	9.18 compound	9.62 israel	8.93 church	7.30 vice
11.56 treatment	8.57 semi	9.14 border	8.76 spirit	7.20 instinct
11.45 clinic	8.38 fire	9.00 shout	8.35 mari	6.93 evolut
11.41 infect	8.19 raid	8.50 turkei	8.22 easter	6.74 jon
11.30 cure	8.13 assault	8.45 arab	8.07 pope	6.74 bake
10.93 Gordon	8.07 weapon	8.39 greec	8.06 geneva	6.72 speci

La méthode a des capacités additionnelles d'extraction de sujets.

ANALYSE DU DISCOURS
(Corpus DEFT 2005)

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
.....

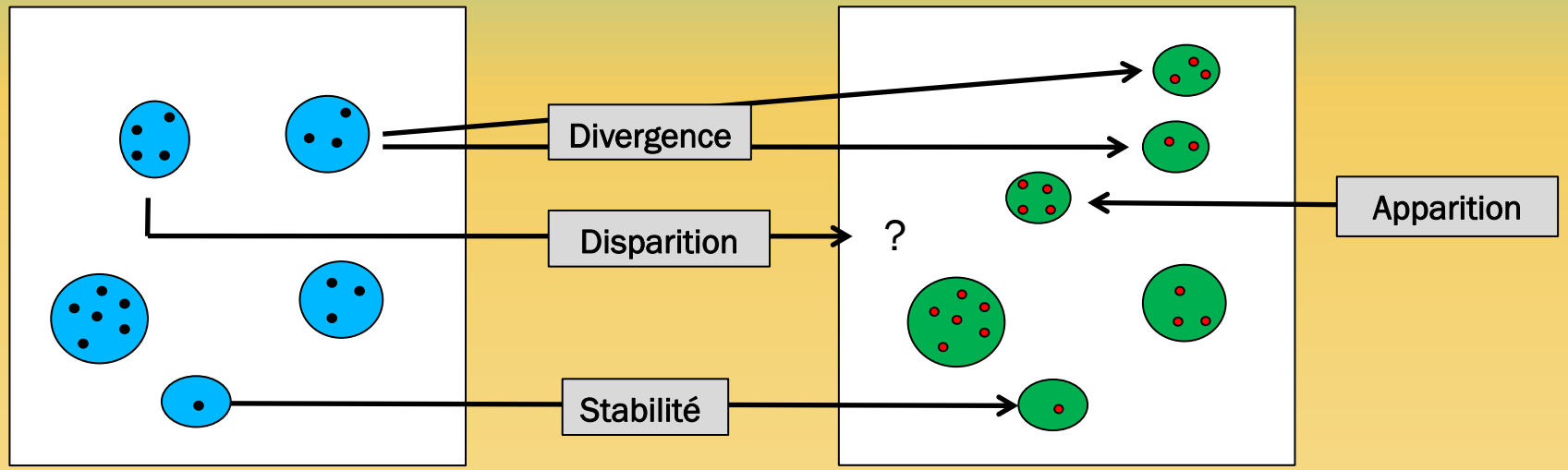
MITTERAND

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 étais
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
.....

ANALYSE DIACHRONIQUE DE LA RECHERCHE

Buts :

Automatiser le processus d'analyse par pas de temps (analyse diachronique) de l'évolution des thèmes de recherches en exploitant les capacités du paradigme MVDA et celles de la maximisation des traits et **revisiter les "premiers résultats" obtenus en mode semi-supervisé par le projet IST PROM-TECH [Lamirel 2012].**

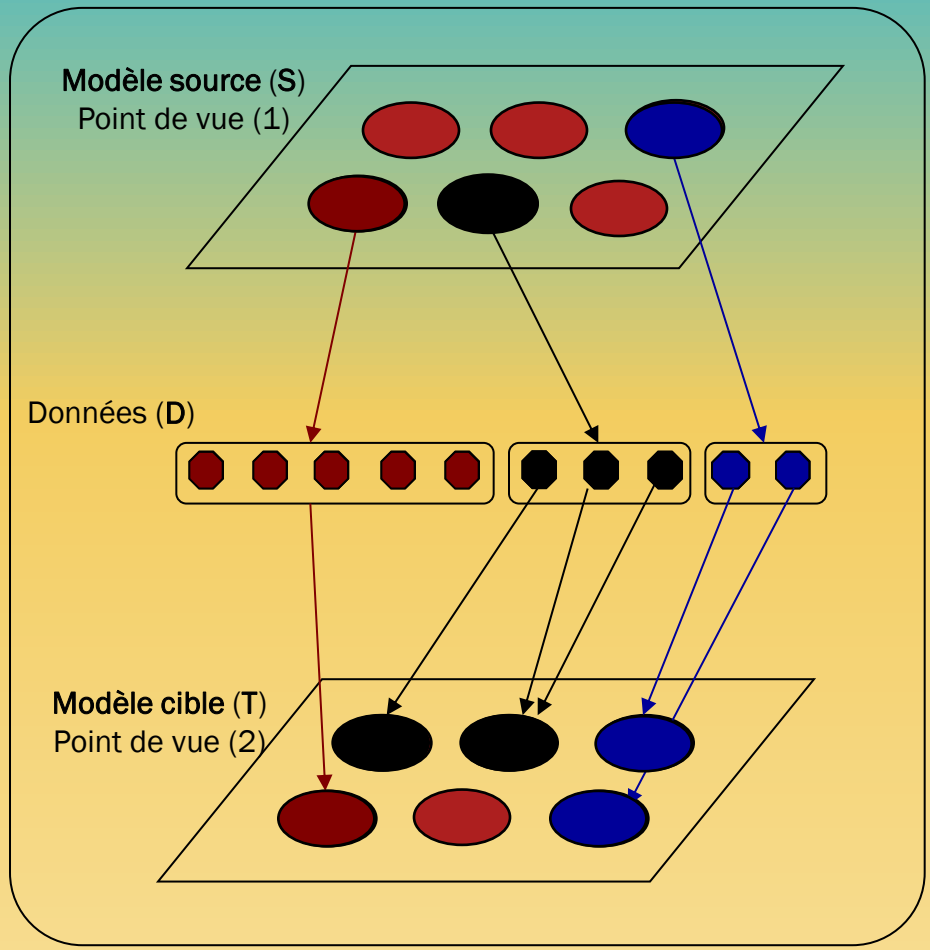


Première expérience basée sur un corpus de référence contenant approx. 4000 notices PASCAL relatives à la recherche en **optoélectronique** durant la période 1996-2003, et originellement divisé en 2 sous-périodes. **=> 4000 dimensions**

LE PARADIGME MVDA

Déduction thématique

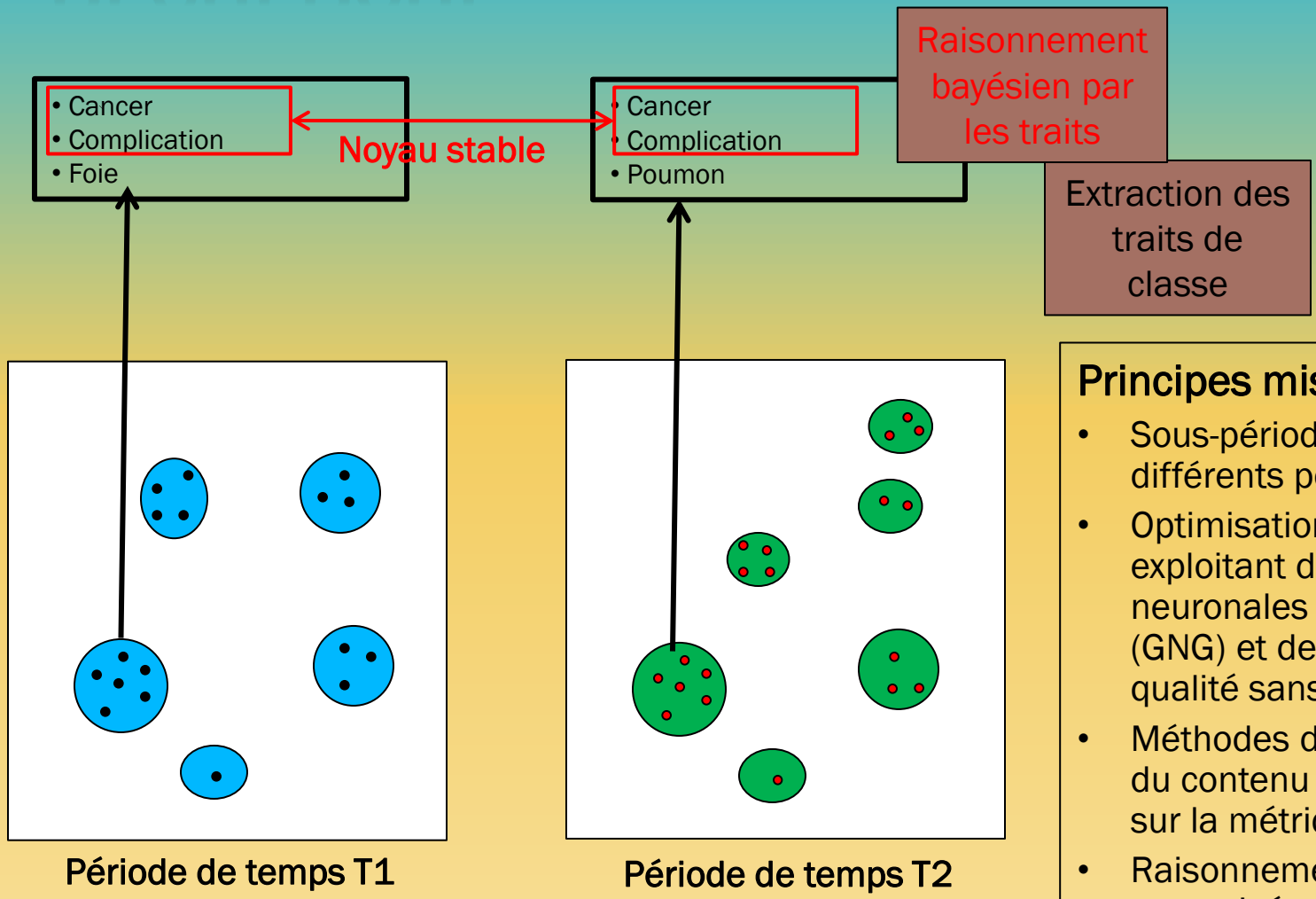
- Le paradigme MVDA repose sur le raisonnement bayésien,
- Le réseau bayésien est généré de manière non supervisée,
- Les données partagées entre les modèles sont exploitées pour la comparaison des classes,
- Applicable aux résultats de toute méthode de classification non supervisée,
- Supporte la généralisation.



Modèle de réseau bayésien :

$$P(act_m \mid T_j, Q) = \frac{\sum_{d \in act_m, T_j} Sim(d, S_i)}{\sum_{d \in T_j} Sim(d, S_i)}$$

ANALYSE DIACHRONIQUE DE LA RECHERCHE



Principes mis en œuvre :

- Sous-périodes associées à différents points de vue,
- Optimisation du clustering en exploitant des méthodes neuronales à topologie libre (GNG) et des indices de qualité sans biais,
- Méthodes de caractérisation du contenu des classes basée sur la métrique F-max,
- Raisonnement bayésien non supervisé adapté aux étiquettes.

MDVA

Raisonnement sur les étiquettes (traits) (1)

- Seules les variables (traits) de classes dont la F-mesure de trait est supérieure à la F-mesure moyenne du modèle sont exploitées dans la comparaison.
- La comparaison est opérée en utilisant une adaptation du raisonnement bayésien original du modèle MVDA avec :

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

où L_x représente l'ensemble des traits associés à la classe *x, et* $L_x \cap L_y$ représente les traits communs, également appelé **noyau stable**, entre la classe *x* et la classe *y*.

MDVA

Raisonnement sur les étiquettes (traits) (2)

La **similarité** entre une classe s de la période source et une classe *t* de la période cible est établie en exploitant :

- La probabilité moyenne $P_A(x)$ de correspondance d'une classe *x* d'une période avec les classes de l'autre période,
- L'activité globale A_p générée par un modèle de période *p* sur le modèle de la période alternative et son écart-type $б_p$.

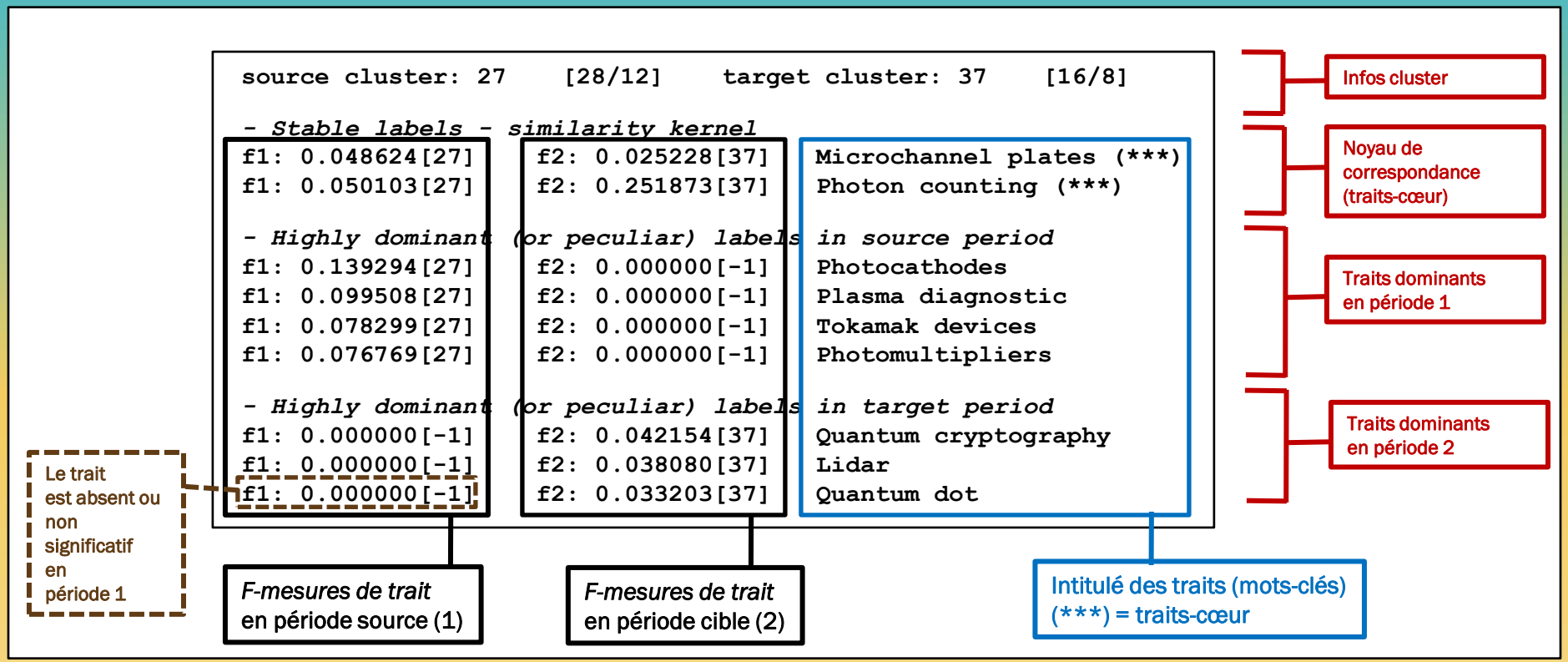
Une similarité est trouvée si :

1) $P(t|s) > P_A(s)$ et $P(t|s) > A_s + б_s$
2) $P(s|t) > P_A(t)$ et $P(s|t) > A_t + б_t$

Les événements d'**éclatement de classes**, de **fusion de classes**, de **disparation de classes**, et, d'**apparition de classes** peuvent être déduits des règles précédentes.

ANALYSE DIACHRONIQUE

Construction de rapports de correspondance



Deux types de rapports sont produits :
similitude (incl. variations contextuelles)
et **divergence** (changements importants).

ANALYSE DIACHRONIQUE

Résultats de correspondance

```
source cluster: 23    [19/10]    target cluster: 2     [12/7]
- Stable labels - similarity kernel
f1: 0.259231[23]     f2: 0.313356[ 8] Optical 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
                     2: 0.028204[ 2] Polymerization
```

Théorie vers pratique

```
source cluster: 15    [22/9]     target cluster: 24    [20/8]
- Stable labels - similarity kernel
f1: 0.038370[15]     f2: 0.044230[24] Silicon compound (***)

- 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
                     : 0.039028[24] Plasma CVD coatings
```

Nouveau composant

```
source cluster: 14    [18/6]     target cluster: 14    [29/7]
- Stable labels - similarity kernel
f1: 0.035721[14]     f2: 0.041813[14] Surface emitting 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
                     0.029637[14] Vertical cavity laser
                     0.025024[14] VCSEL
```

Théorie vers pratique

```
source cluster: 24    [23/9]     target cluster: 33    [27/13]
- No stable labels

- 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
                     6[33] Ray tracing
```

Changement de vocabulaire

```
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
```

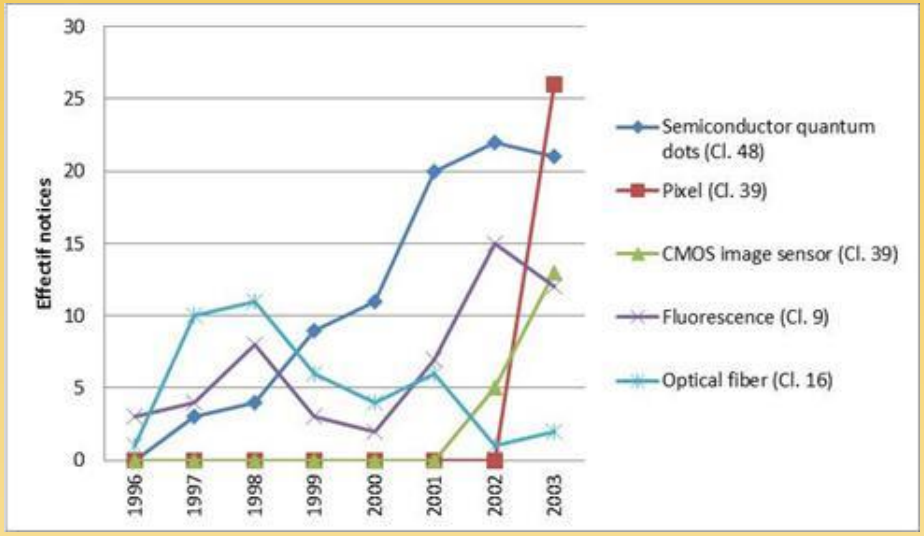
```
target cluster 39 is appearing
f1: 0.000000[-1]     f2: 0.144184[39] Pixel
f1: 0.000000[-1]     f2: 0.110076[39] CMOS image sensors
f1: 0.000000[-1]     f2: 0.077578[39] Chip
f1: 0.000000[-1]     f2: 0.060044[39] High sensitivity
```

ANALYSE DIACHRONIQUE

Validation

NUMERO CLUSTER (ID) THEME	MOTS-CLES DOMINANYT DU THEME	DIFFERENCE DE F-MESURE D'ETIQUETAGE ENTRE PERIODES	EFFECTIF NOTICES DU THEME EN PERIODE 1 (1996-1999)	EFFECTIF NOTICES DU THEME EN PERIODE 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

Les changements caractérisés par la méthode sont corrélés avec la variation du nombre de publications.

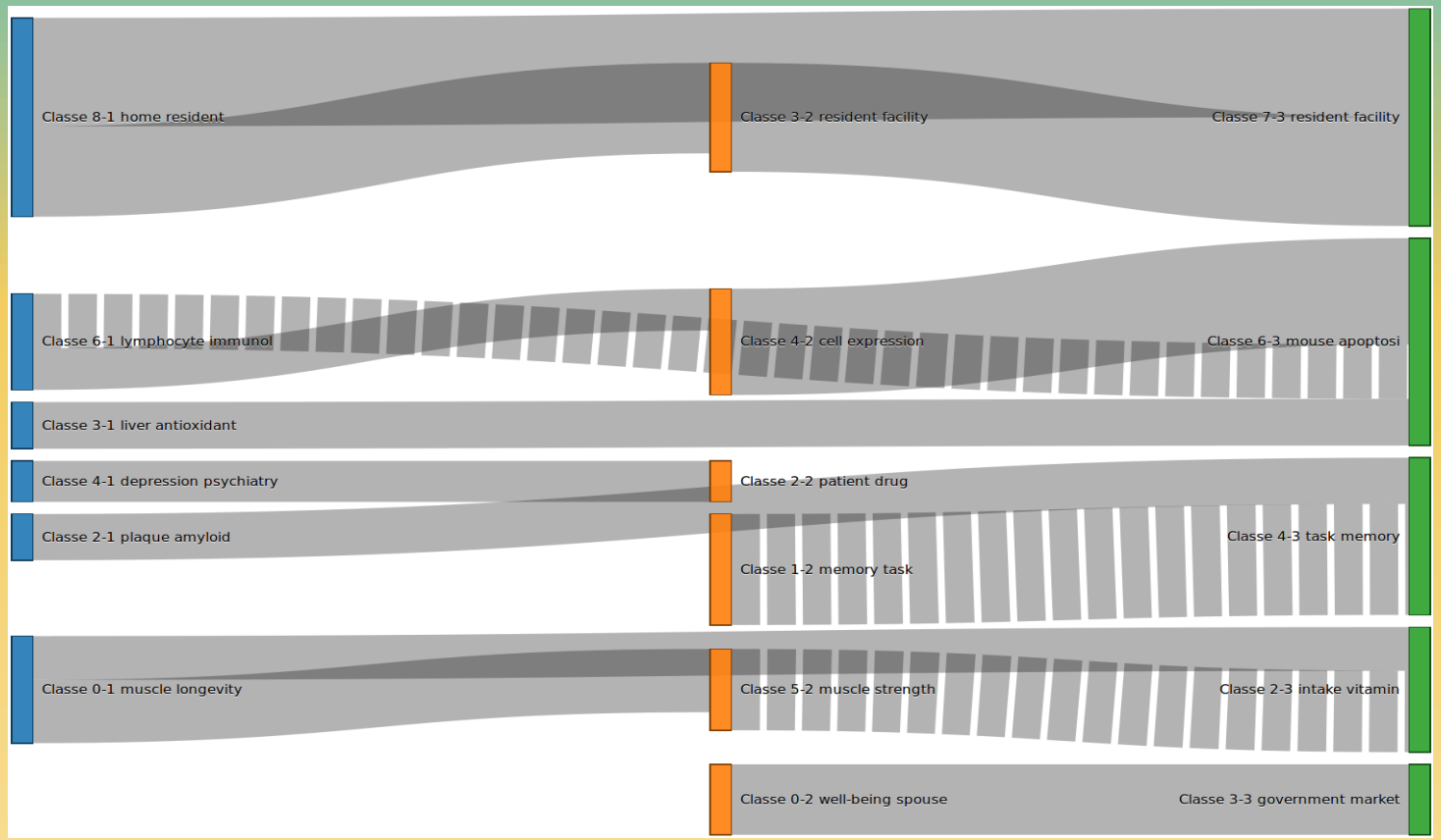


L'APPROCHE DIACHRONIC'EXPLORER

- Approche développée spécifiquement pour l'analyse diachronique des collections ISTEX [Dugué et al. 2016],
- Exploite la combinaison de la maximisation des traits, du clustering et le paradigme MDVA,
- Propose une large batterie d'outils de visualisation, y compris des outils originaux (graphes de contrastes, …)

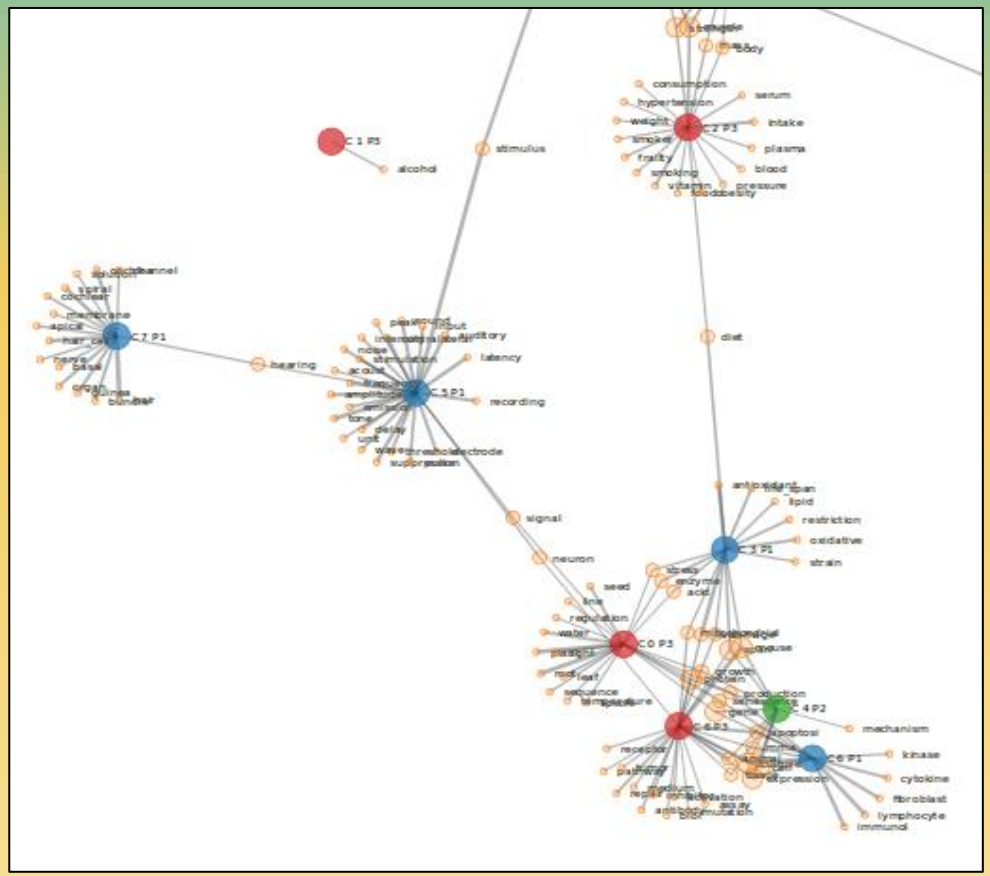
L'APPROCHE DIACHRONIC'EXPLORER

Une adaptation de la représentation Sankey est utilisée pour la visualisation globale des changements diachroniques :



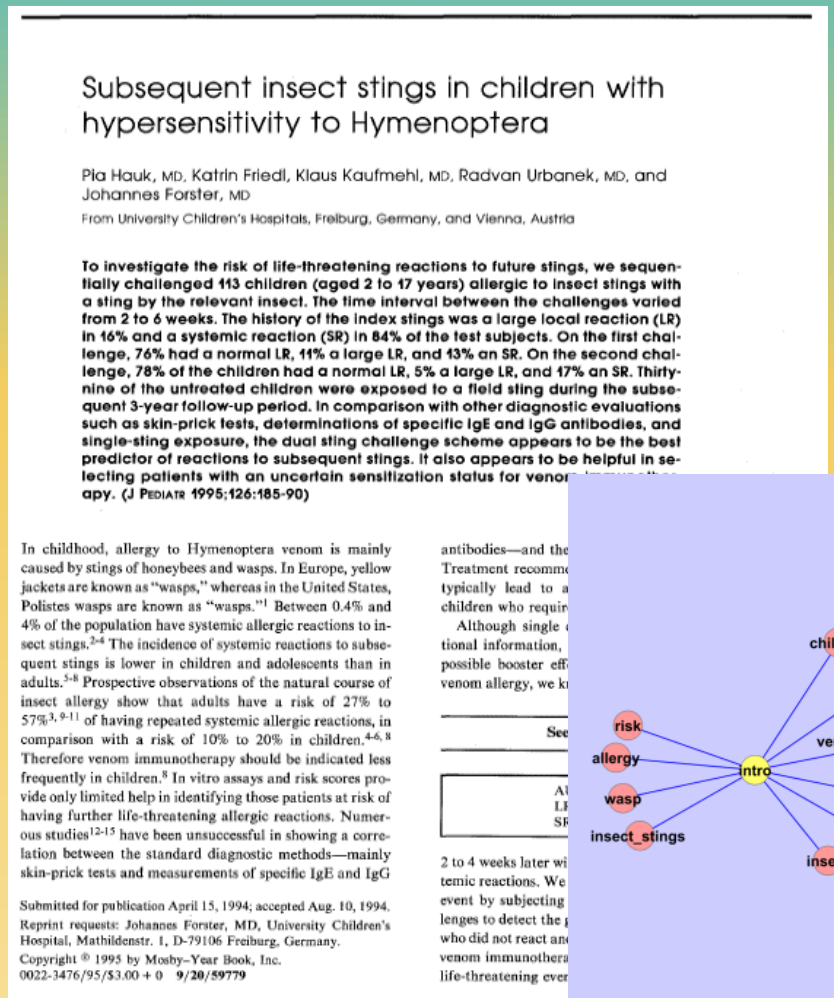
L'APPROCHE DIACHRONIC'EXPLORER

Les graphes de contrastes permettent également de mieux comprendre les dépendances diachroniques des sujets :



Extraction automatique de métadonnées par MT

Le pouvoir de synthèse de la métrique F-max est très important car elle permet de mettre en évidence la structure d'un texte (et ses métadonnées descriptives) par un mécanisme simple de compétition entre blocs.



Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD, Radvan Urbanek, MD, and Johannes Forster, MD

From University Children's Hospitals, Freiburg, Germany, and Vienna, Austria

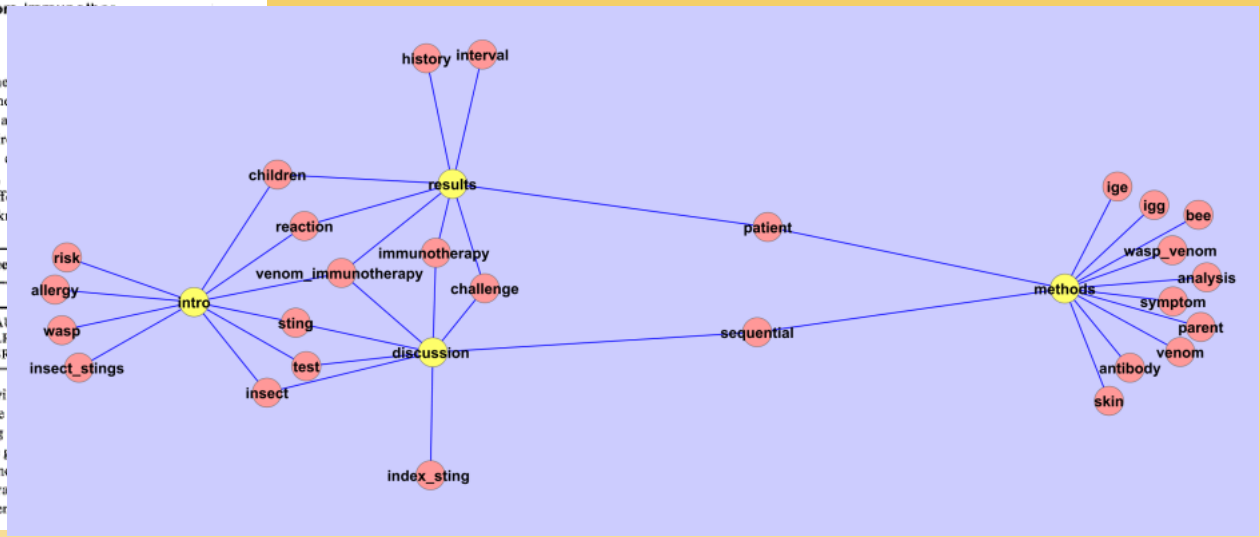
To investigate the risk of life-threatening reactions to future stings, we sequentially challenged 113 children (aged 2 to 17 years) allergic to insect stings with a sting by the relevant insect. The time interval between the challenges varied from 2 to 6 weeks. The history of the index stings was a large local reaction (LR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal LR, 11% a large LR, and 13% an SR. On the second challenge, 78% of the children had a normal LR, 5% a large LR, and 17% an SR. Thirty-nine of the untreated children were exposed to a field sting during the subsequent 3-year follow-up period. In comparison with other diagnostic evaluations such as skin-prick tests, determinations of specific IgE and IgG antibodies, and single-sting exposure, the dual sting challenge scheme appears to be the best predictor of reactions to subsequent stings. It also appears to be helpful in selecting patients with an uncertain sensitization status for venom immunotherapy. (J Pediatr 1995;126:185-90)

In childhood, allergy to Hymenoptera venom is mainly caused by stings of honeybees and wasps. In Europe, yellow jackets are known as "wasps," whereas in the United States, Polistes wasps are known as "wasps."[1] Between 0.4% and 4% of the population have systemic allergic reactions to insect stings.[2-4] The incidence of systemic reactions to subsequent stings is lower in children and adolescents than in adults.[5-8] Prospective observations of the natural course of insect allergy show that adults have a risk of 27% to 57%[3, 9-11] of having repeated systemic allergic reactions, in comparison with a risk of 10% to 20% in children.[4-6, 8] Therefore venom immunotherapy should be indicated less frequently in children.[8] In vitro assays and risk scores provide only limited help in identifying those patients at risk of having further life-threatening allergic reactions. Numerous studies[12-15] have been unsuccessful in showing a correlation between the standard diagnostic methods—mainly skin-prick tests and measurements of specific IgE and IgG antibodies—and the

Submitted for publication April 15, 1994; accepted Aug. 10, 1994.

Reprint requests: Johannes Forster, MD, University Children's Hospital, Mathildenstr. 1, D-79106 Freiburg, Germany.

0022-3476/95/$3.00 + 0 9/20/59779



Cas des données astromoniques

- Corpus d'env. 500000 articles sur le thème général de l'astronomie de issues de la base ISTEX,
- Période couvrant 189 ans,
- Les données sont étiquetées par les entités nommées, noms de lieu, nom de personnes, dates
  => Unitex/CacSys [Maurel et al. 2016].
- Identification des articles les plus cités,
- Analyse du contenu par extraction automatique des métadonnées et isolement de sujet centraux (ex: big bang, théorie des cordes ….),
- Extraction du contexte des citations (phrases dans lesquelles les citations apparaissent) [Al Zied et al. 2017],
- Mesure du cumul de contraste/période généré sur les sujets centraux par le contexte des citations,
- Visualisation des variabilités temporelles,
- Mise en parallèle avec une analyse directe basée sur le clustering et sur MVDA.

Cas des données astromoniques

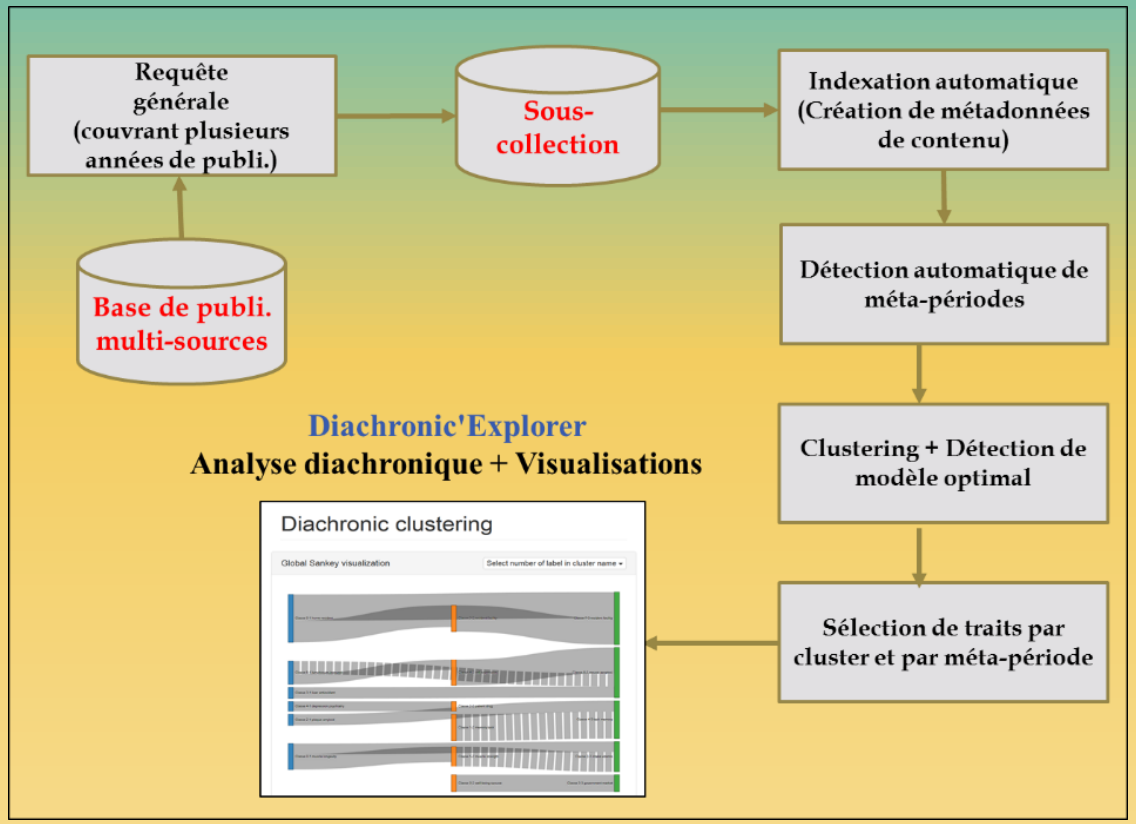
Un exemple d'annotations dans un article :

Conclusion

- Nous avons présenté une méthodologie permettant d'analyser les données de manière diachronique à partir des données étiquetées par des entités nommées,
- Le but est d'analyser les effets de récurrence liés aux nouveaux paradigmes scientifiques,
- Le principe général de l'approche repose sur des techniques récemment expérimentées avec succès,
- Les données étiquetées restent cependant en cours de traitement,
- Un problème important est celui du repérage cohérent des citations dont la syntaxe varie en fonction des périodes de temps.

Annexe : Analyse diachronique et navigation dans les données multisources ISTEX-R - WP1

La figure présente le déroulement de l'approche Diachronic'Explorer complète jusqu'à la visualisation



La méthode ne présente pas les inconvénients des méthodes d'extraction de sujets usuelles, comme LDA (Blei et al. 2003) : sujets imprécis et dépendants du processus d'optimisation utilisé, non applicabilité à l'échelle des documents, ...

L'indexation automatique peut être remplacée par le processus d'extraction de métadonnées basé sur la maximisation des traits.