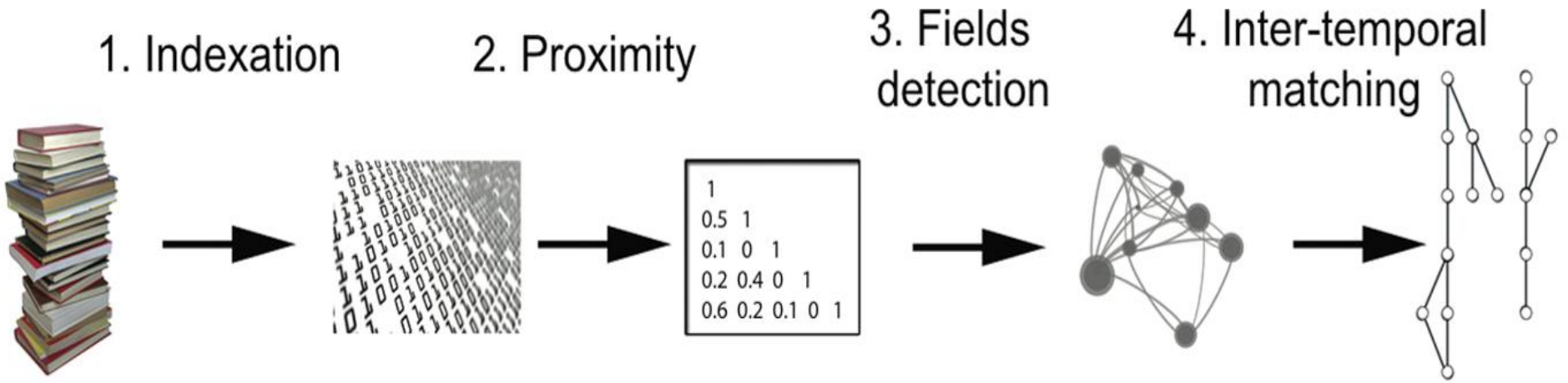
# Construction de réseaux « phylomémétiques »

## Axe 3 : Dynamicité

- Partenaires :
  - LIP6 : B. Amann, H. Naacke
  - IRISA : D. Gross-Amblard, Z. Miklos
  - CAMS-EHESS : D. Chavalarias, ….
- Recrutements :
  - Juan Pablo Stocca (sept-déc. 2013), ingénieur
  - Alexis Guichard (avril-sept, 2014), stage M2R/UPMC
  - Jonathan Lajus (avril-août, 2014), stage M2R/ENS

# Objectif

- Extraction d'une « Carte de l'évolution des Sciences » à partir d'un corpus scientifique (WoS)




**Réseau phylomémétique**

David Chavalarias and Jean-Philippe Cointet. Phylomemetic Patterns in Science Evolution The Rise and Fall of Scientic Fields. PLoS ONE, 8(2) :e54847+, February 2013.

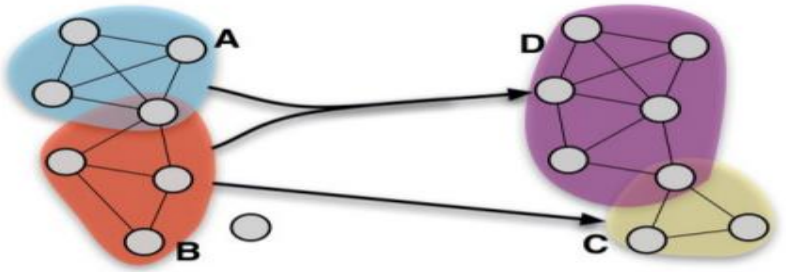
# Defis « Masse de Données »

- Passage à l'échelle
- Application actuelle :
  - 5 000 termes /  100 000 articles
- Application visée :
  - Web Of Science :
    - 12k journaux depuis 1990
    - 300 000 termes / 30 M articles
  - Analyse interactive et « temps-réel »
- Défis :
  - Complexité des phases clusterisation et analyse temporelle
  - Taille des résultats intermédiaires

# Approche en trois étapes

**3. Analyse temporelle**

**Concepts**
(cliques, clusters)




**2. Clustering**
**(fouille de texte/graphe)**

Co-ocurrences
de **termes**
(phrases)

**1. Extraction de termes**

**Corpus**
(articles)

2010

2011

# Phase « Clusterisation »

Juan-Pablo Stocca / Jonathan Lajus

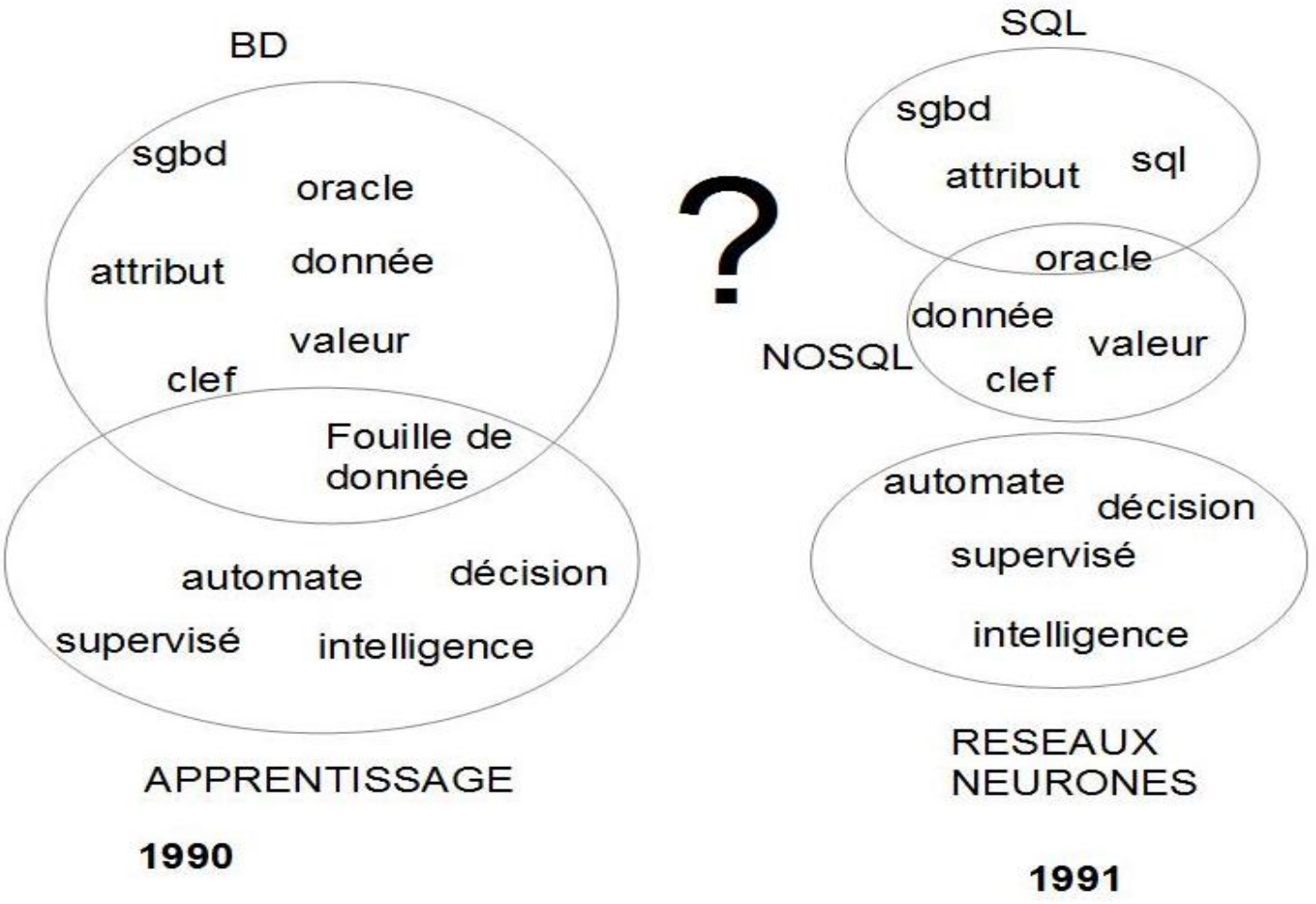
Extraction de concepts (ensembles de termes cohérentes) dans un grand corpus (3M documents)

- Génération d'un index inversé
- Génération de graphe de co-occurrences *G(année)*
- Extraction de concepts de *G(année)*
  - Percolation (polynomiale) de cliques maximales (NP complet)
    - Tomita et. al.
  - Optimisation de modularité
    - Louvain : partitionnement, taille de partitions normalisée

# Génération de graphe de co-occurrences

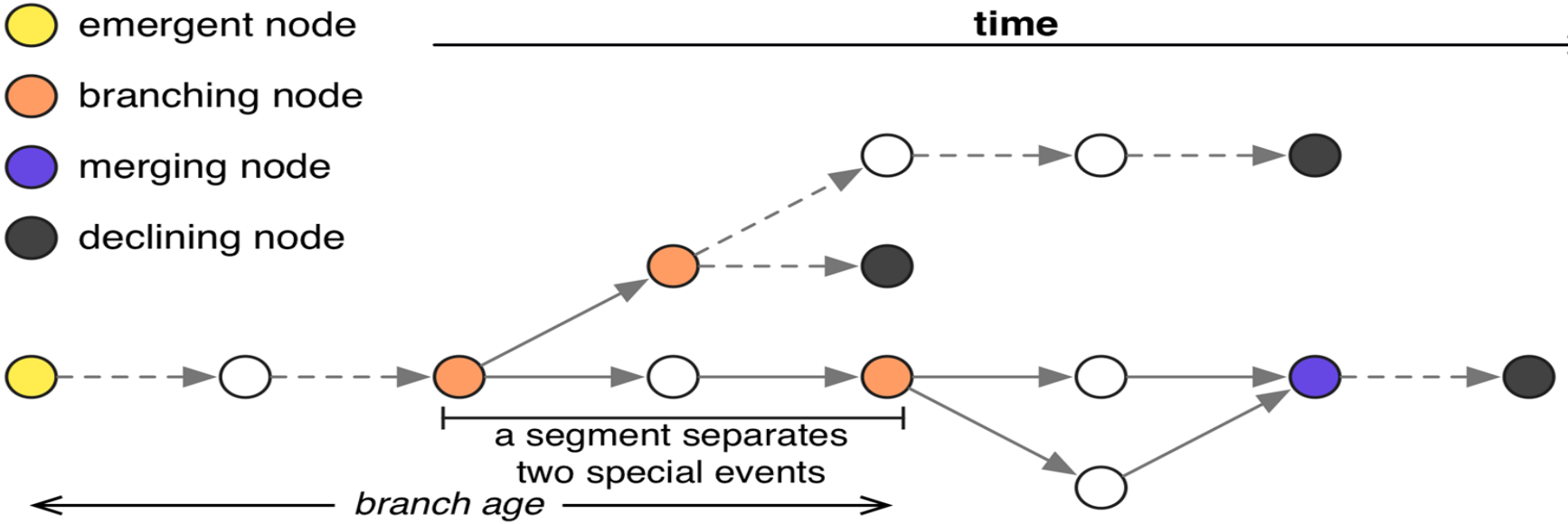
- Occurrences O = { (doc, terme) } : scripte Python
- Co-occurrences G = { (terme, terme) } : Spark / Hadoop
- Analyse de G(1990):
  - 865k sommets, 290M arêtes
  - Composantes connexes : 5, taille min=2, taille max = 290k
  - Arêtes par sommet: avg=336k, max=757k
- Résultats :
  - Tomita et.al. : difficilement parallélisable
  - Louvain: valeurs de modularité très/trop faibles
- Solution :
  - Filtrage d'arêtes et de termes (stopwords, degré de nœuds, Pagerank, …)
  - Plus d'expériences nécessaires …

# Phase « Alignement de concepts »



# Phylomémie

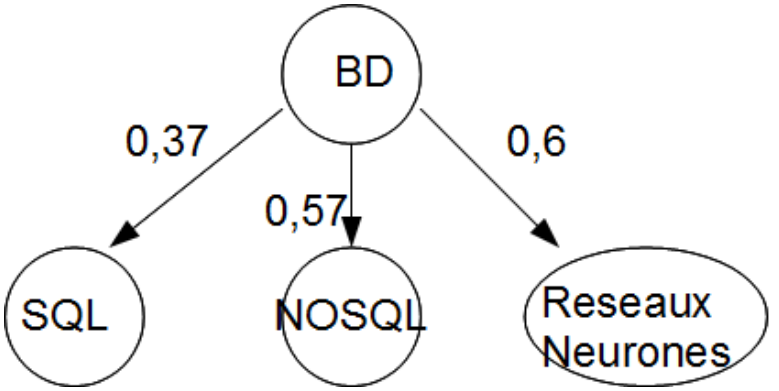
- Inspiré des arbres phylogénétiques

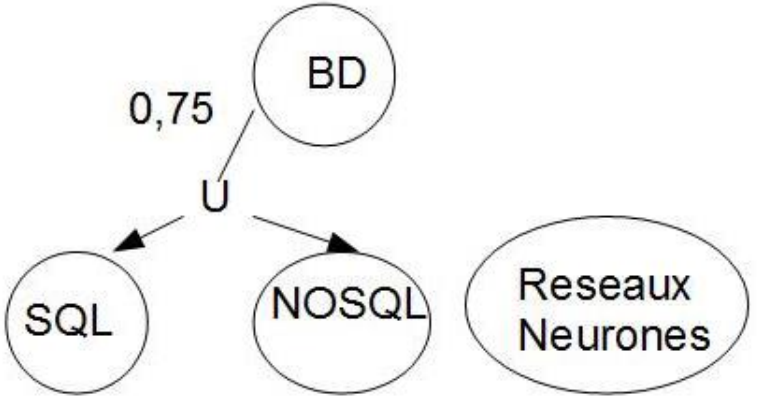



David Chavalarias and Jean-Philippe Cointet. Phylomemetic Patterns in Science Evolution
The Rise and Fall of Scientic Fields. PLoS ONE, 8(2) :e54847+, February 2013.

# Alignement par similarité

- Similarité Jaccard
- Filtrage par seuil s0
- Union de deux clusters
- Approche ascendante et descendante
- Optimisation : élaguer l'espace de recherche
  - Somme des similarités supérieure à la meilleure similarité

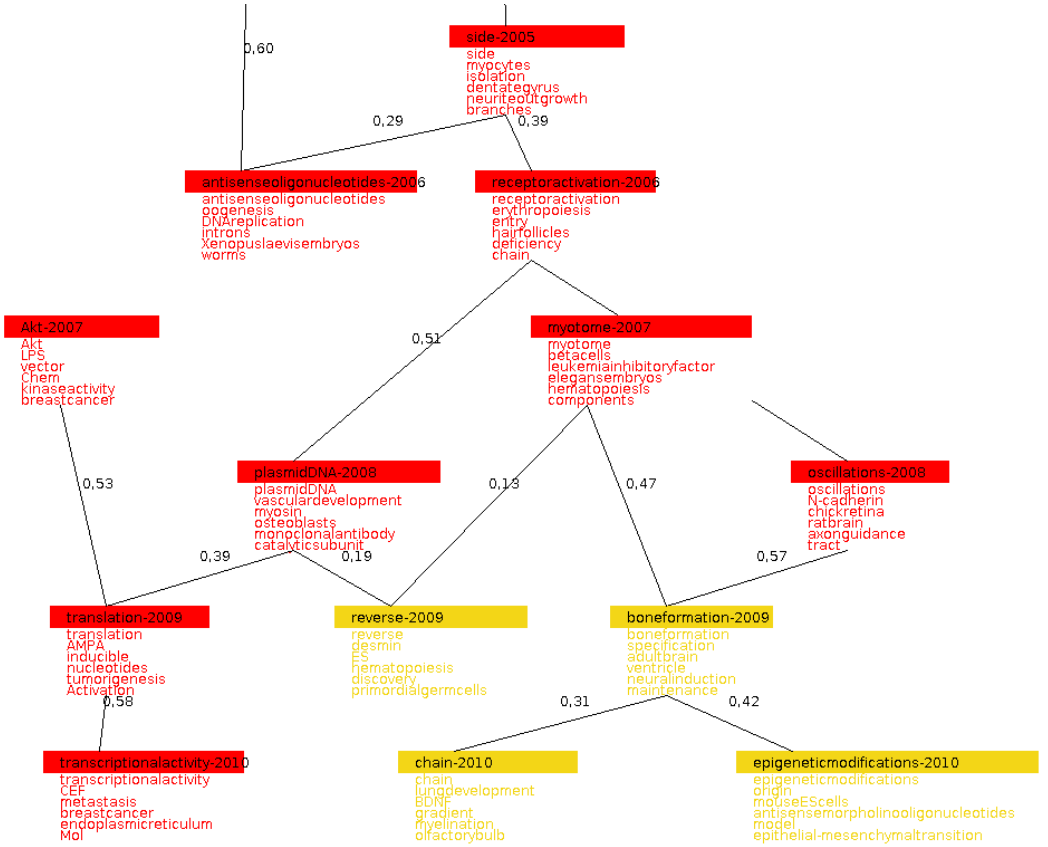



$$\frac{A \cap (B \cup C)}{A \cup (B \cup C)} \leq \frac{A \cap B}{A \cup B} + \frac{A \cap C}{A \cup C}$$

# Identification de catégories

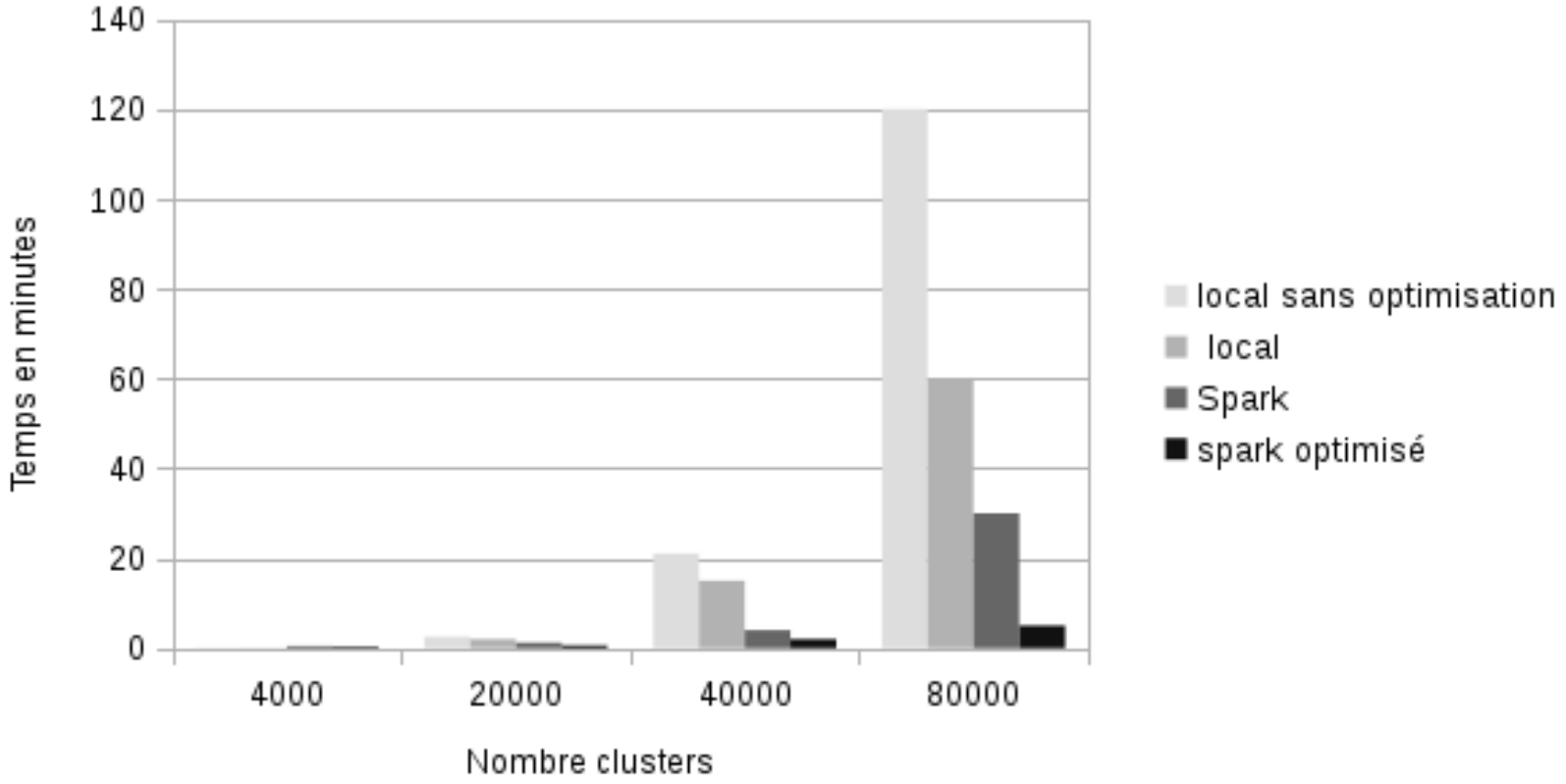
- Catégories de termes :
  - Hérités
  - Recombinés
  - Émergents
- Différencier les termes d'un concept
- Trouver un nom :
  - Aux domaines
  - Aux concepts



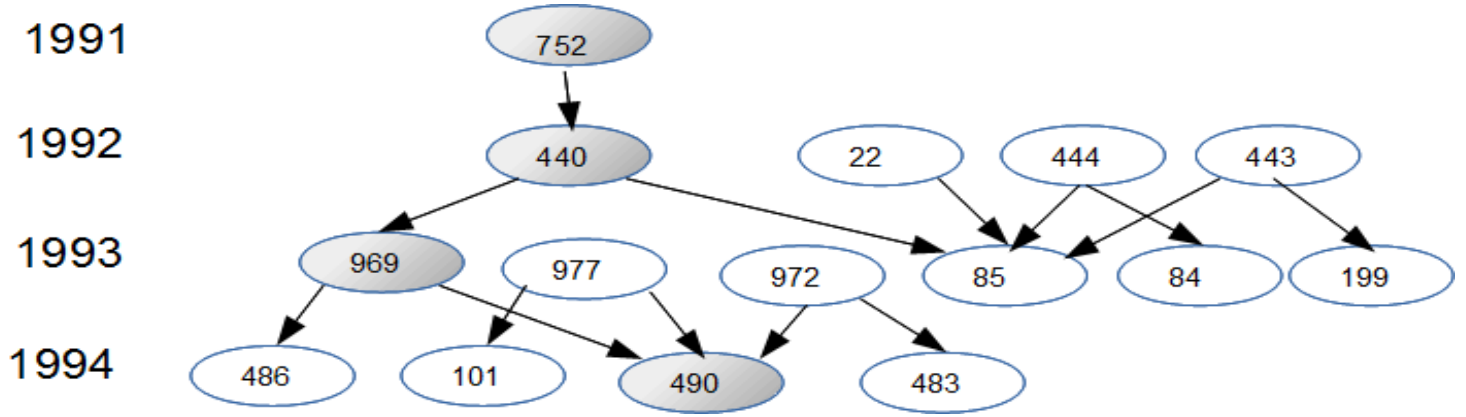
# Spark

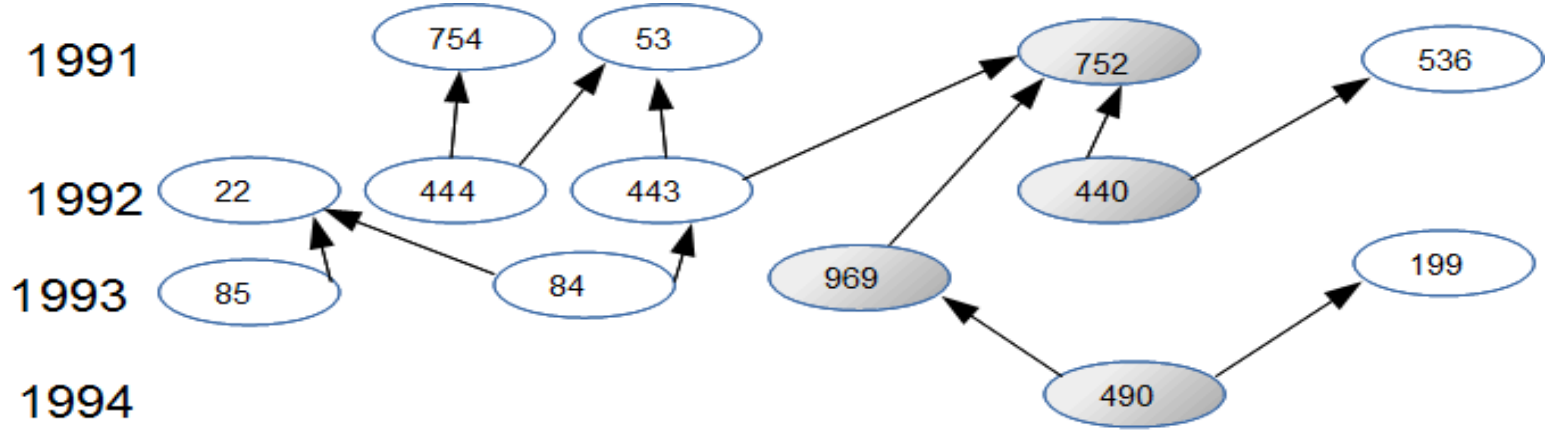
- Programmation parallèe + données distribuées
  - Algèbre sur les RDDs (Resilient Distributed Datasets) : selection, union, join, …
  - Implantation Map-Reduce (Hadoop)
- Algorithme :
  - Cluster année t : Ci, …
  - Cluster année t+1 : Cj,Ck, …
  - Produit cartésien : (Ci, Cj), (Ci, Ck), …
  - Jointure : (Ci, ( Cj, Ck) ), …
  - Filtrage par le seuil s1 pour le niveau de précision

# Expérimentations : temps d'exécution

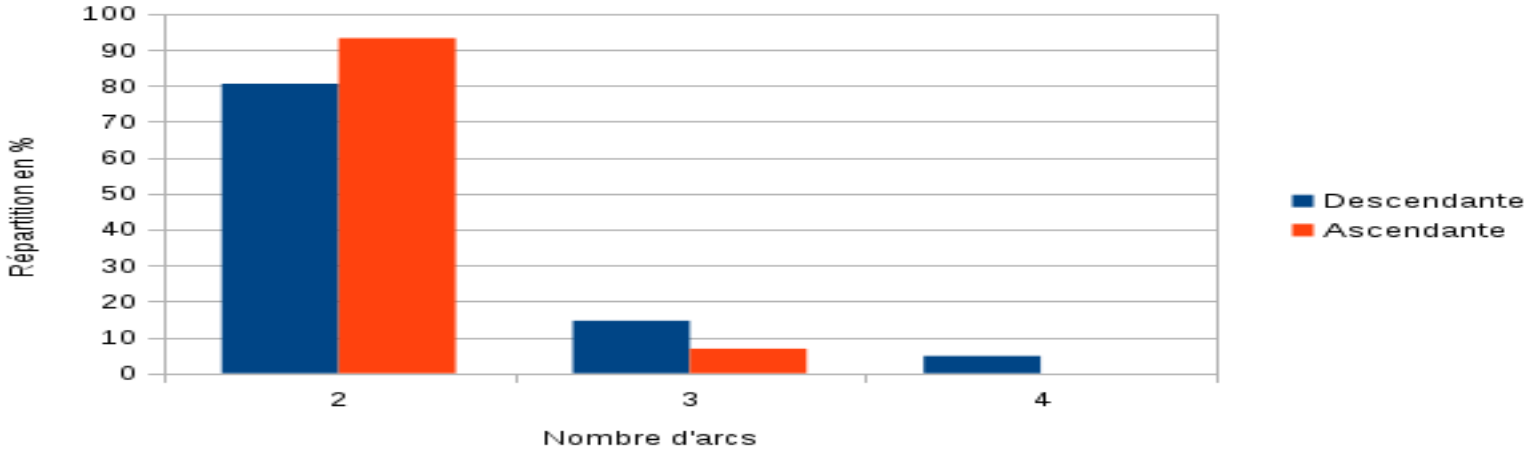


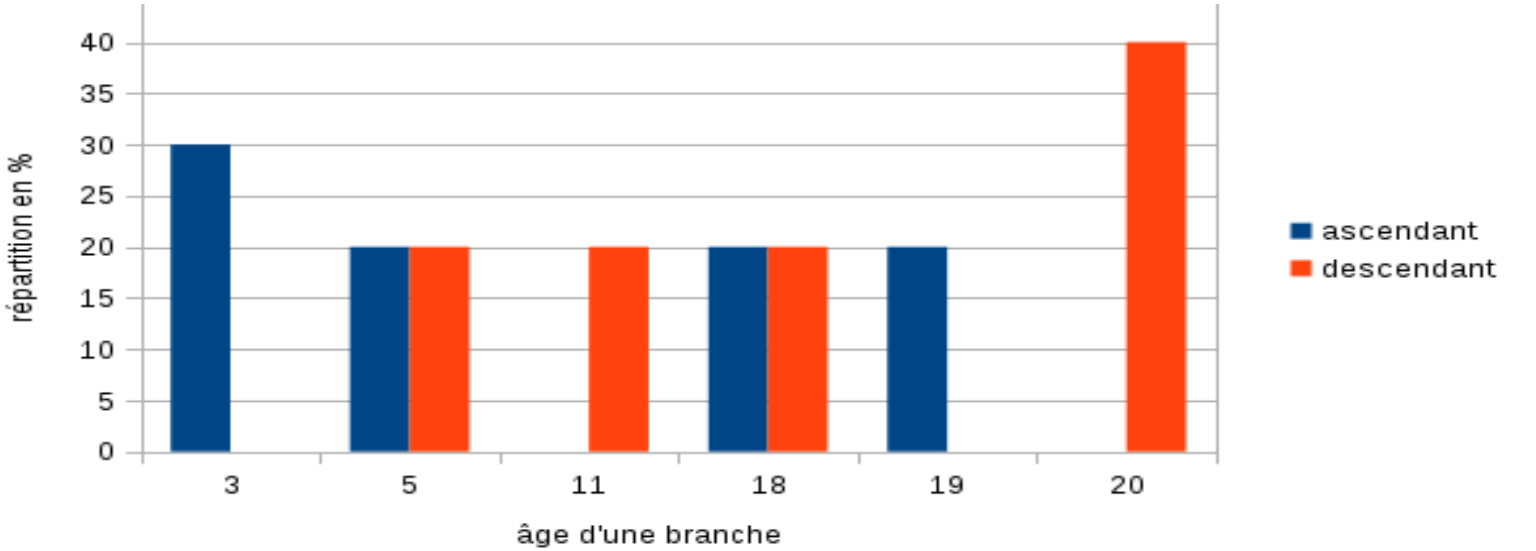
# Expérimentations : Impacts sur les transitions



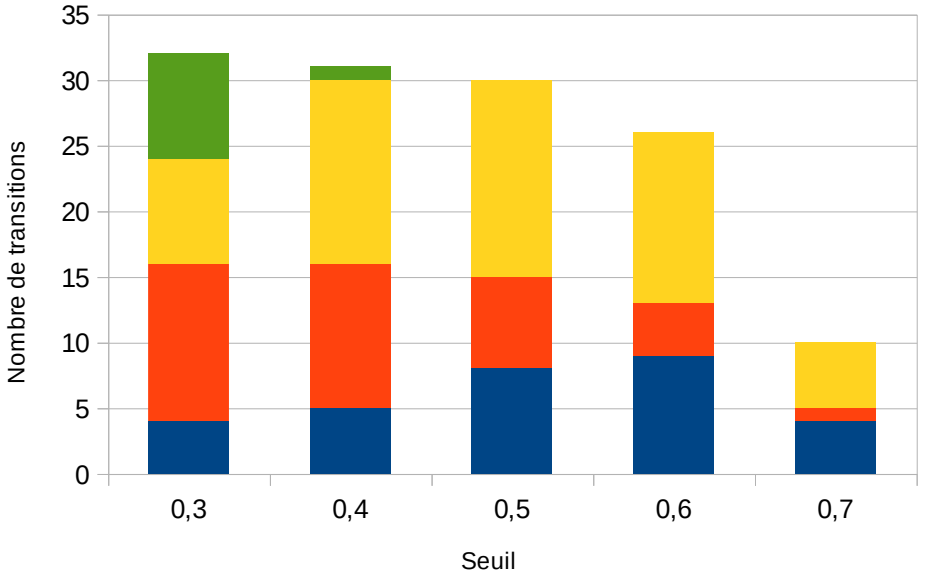



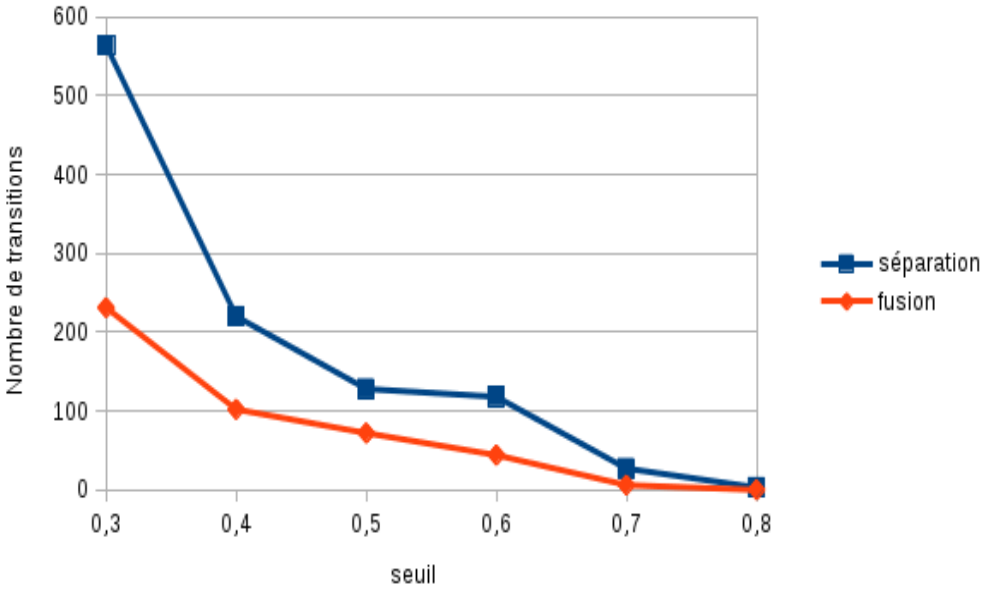
# Expérimentations : Impacts sur les transitions



# Expérimentations : Impacts sur les seuils



# Travaux et résultats

- Travaux effectués : « étude de faisabilité »
  - Formalisation du problème
  - Réalisation Spark/Map-Reduce des phases Extraction et Clustering
  - Expérimentations sur un cluster Hadoop à Rennes et au LIP6
- Travaux futurs / perspectives :
  - Etudier d'autres critères / approches de clustering
  - Expérimentations qualité/performance (WoS complet)
  - Analyse interactive / incrémental / re-clustering / visualisation
  - Projet ANR en cours de soumission

# Merci

# Travaux similaires

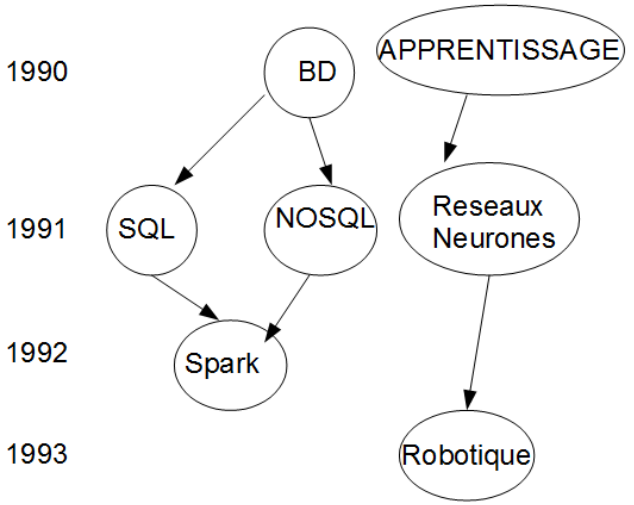
- Deux approches dans la littérature :
  - Détection des changements de clusters :
    - Système MONIC ( Modeling and Monitoring Cluster) *
  - Clustering spatio-temporel :
    - Système SOAPs ( Spatial Object Association Patterns) **

* Myra Spiliopoulou and Irene Ntoutsi. The monic framework for cluster transition detection.

** Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. A generalized framework for mining spatio-temporal patterns in scientic data. In Patterns in Scientic Data, ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pages 716721, 2005.

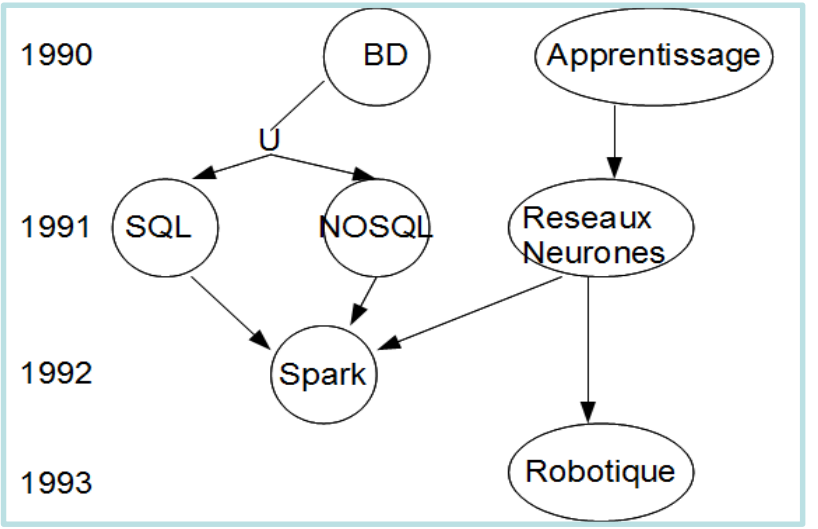
# Détection de transitions externes

- 6 transitions : émergence, disparition, fusion, séparation, ré-émergence, reconduction.
- Filtrage par seuil :
  - S0, similarité significative
  - S1,  résolution du résultat

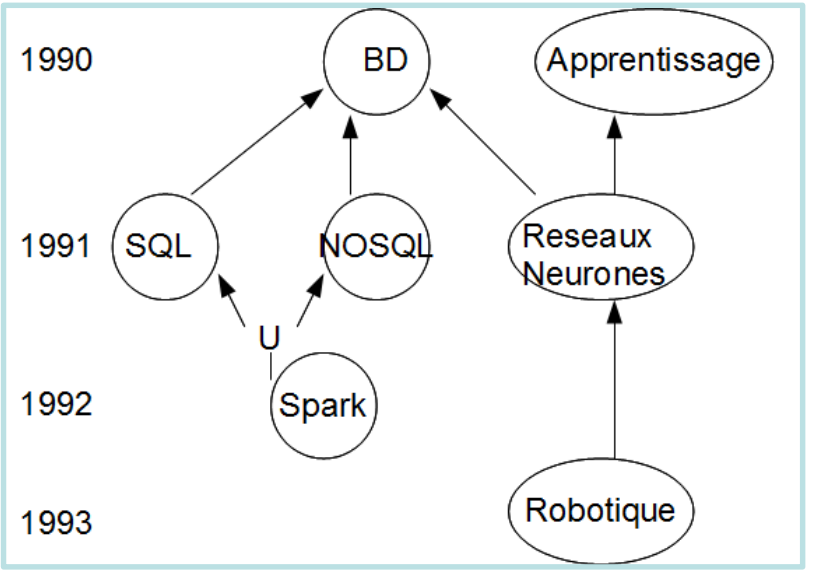


# Approche ascendante et descendante

Approche descendante




Approche ascendante



# Modèle et optimisation

- Calcul d'union : améliorer la complexité
- Solution : élaguer l'espace de recherche
  - Trier la liste des candidats par ordre décroissant
  - Somme des similarités supérieure à la meilleure similarité

$$\frac{A \cap (B \cup C)}{A \cup (B \cup C)} \leq \frac{A \cap B}{A \cup B} + \frac{A \cap C}{A \cup C}$$

# Conclusion

- Définition d'un modèle générique
- Ajout d'une transition de ré-émergence
- Définition de deux analyses d'alignement
- Optimisation pour l'union
- Solution Spark pour le passage à l'échelle
- Programmation d'un logiciel de visualisation avec seuil paramétrable : version Java 8 (spark), Java 7, Scala