ERIC
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

Structuration semi-supervisée des données complexes

Marian-Andrei RIZOIU

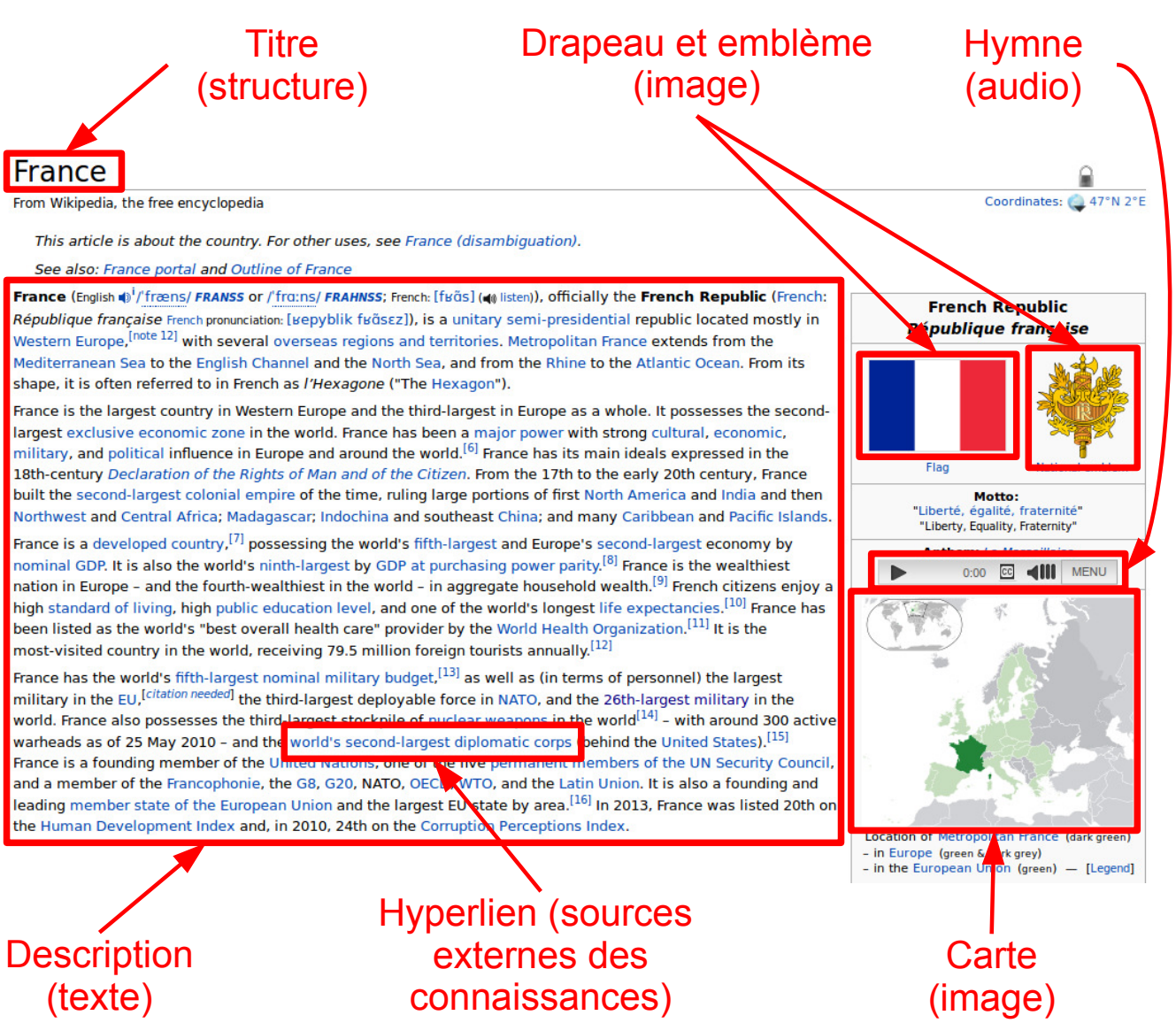
Directeurs de thèse

Stéphane LALLICH **Julien VELCIN**

24 Juin 2013

Laboratoire ERIC, Université Lumière Lyon 2, Lyon, France

Le contexte du travail - les données complexes



Area	
- Total[note 2]	674,843 km² (41st) 260,558 sq mi
- Metropolitan France	
- IGN[note 3]	551,695 km² (47th) 213,010 sq mi
- Cadastre[note 4]	543,965 km² (47th) 210,026 sq mi
Population	
(2012)	
- Total[note 2]	65,350,000[2] (19th)
- Metropolitan France	63,460,000[1] (22nd)
- Density[note 5]	116/km² (89th) 301/sq mi

Indicateurs (format numérique)

Spécificités :

- Différentes types des données
- Information additionnelle
- Dimension temporelle/dynamique
- Grande dimensionnalité
- Sources diverses et distribuées

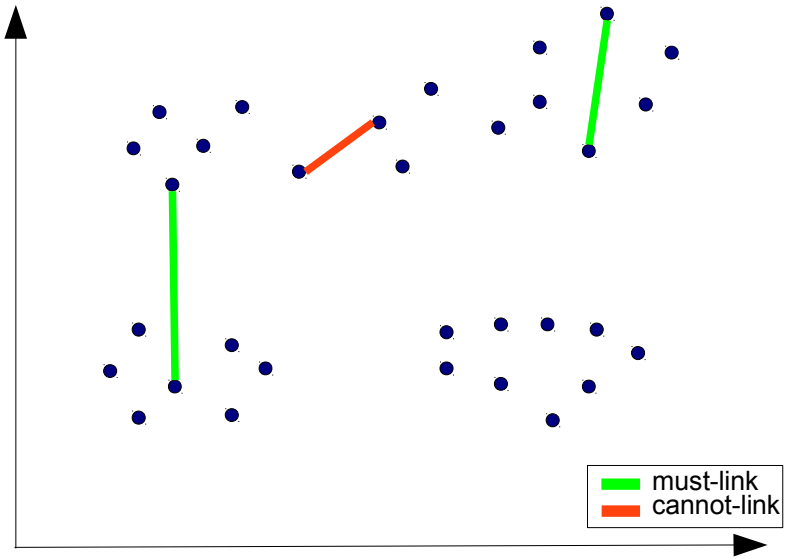
Notre approche

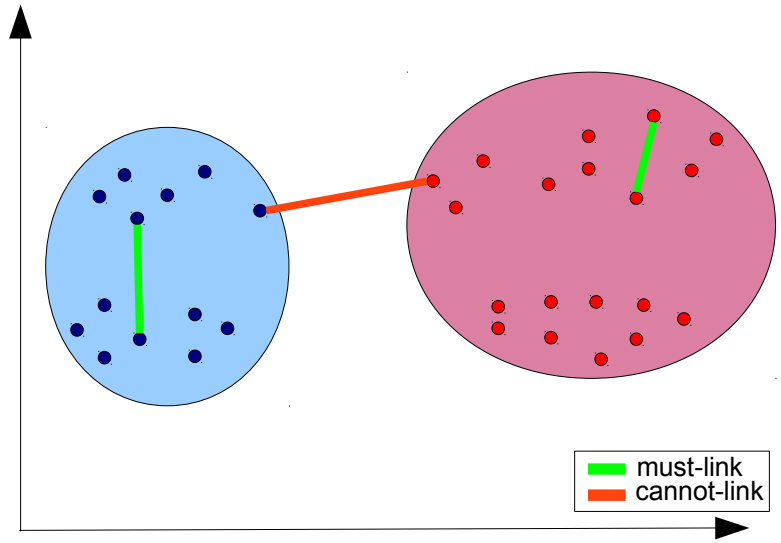
Objectif général : extraire des connaissances à partir de données complexes, souvent dans un contexte non-supervisé ;

ajouter de la sémantique dans l'analyse de données ;

utiliser les connaissances (supervision) disponibles.

Intégrer les connaissances : clustering semi-supervisé [WAG00]





Enjeux de recherche

Utiliser la sémantique pour analyser les données complexes.

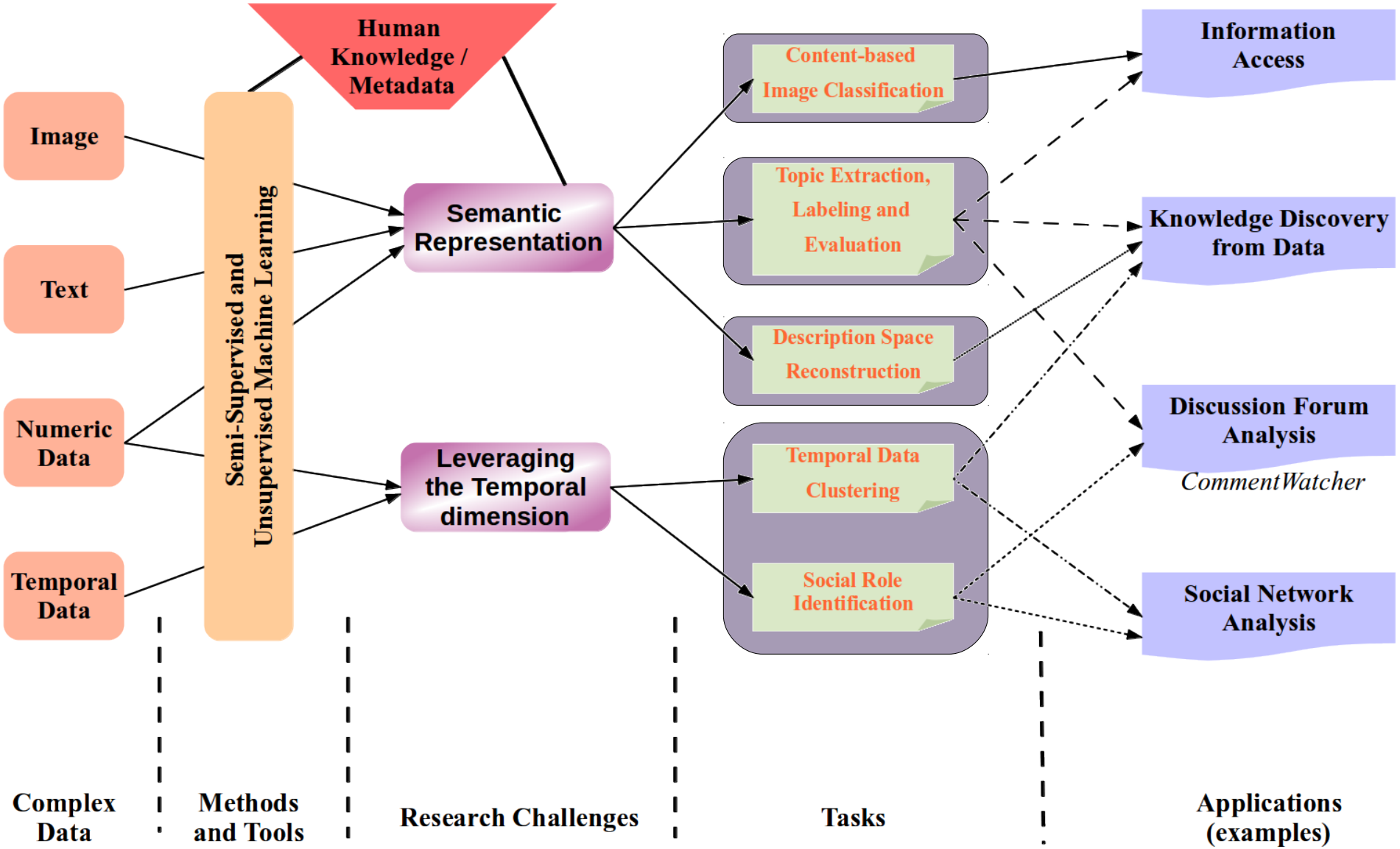
- ➜ plonger les données dans un espace de représentation capable de capturer la sémantique sous-jacente aux données
- ➜ injecter des connaissances externes dans les algorithmes d'apprentissage automatique

Prendre en compte la dimension temporelle des données complexes.

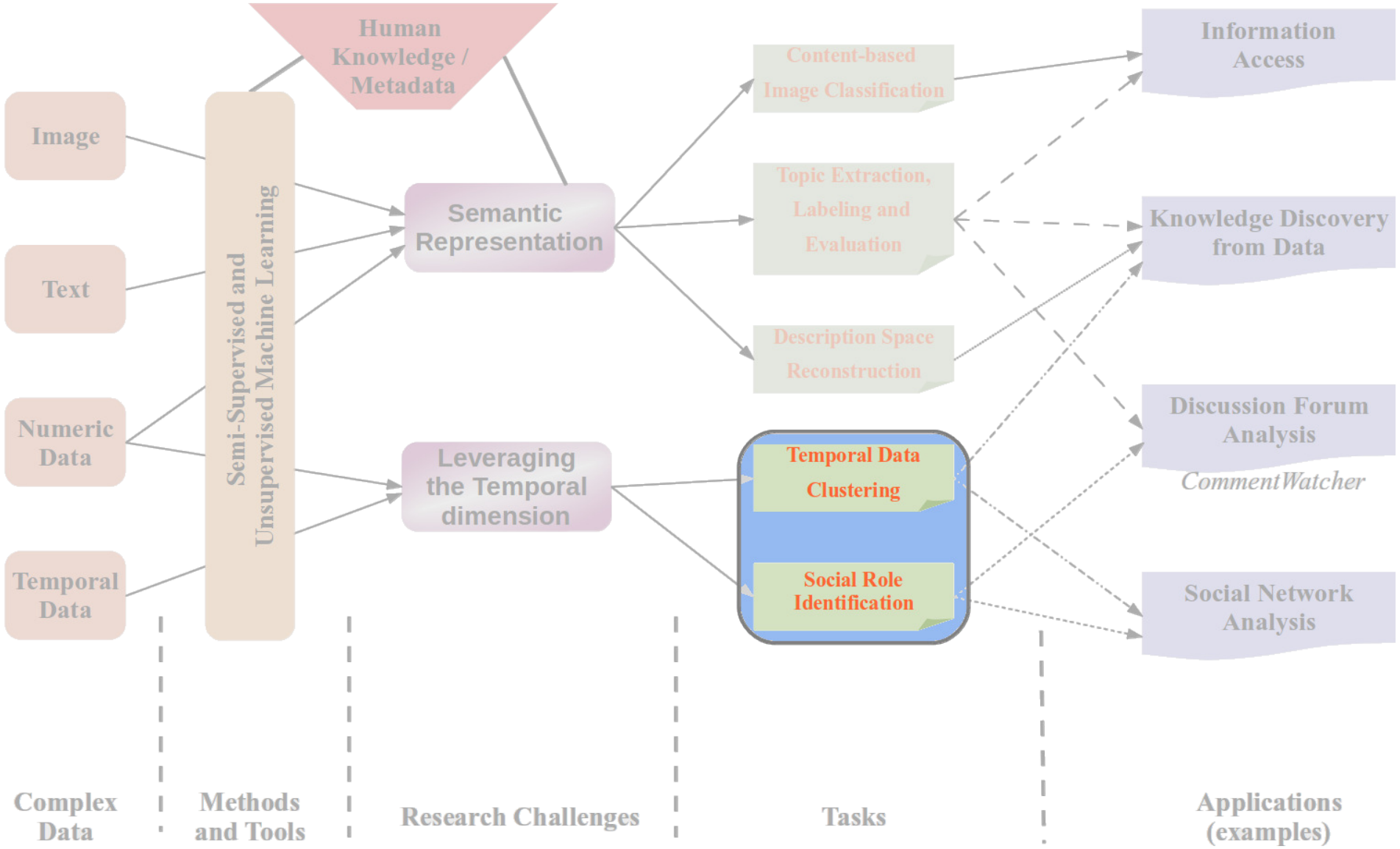
Tâches spécifiques :

- ➜ *détection des évolutions typiques*
- ➜ *reconstruction sémantique de l'espace de représentation des données*
- ➜ *intégration des informations externes dans la description numérique des images.*

Schéma des travaux

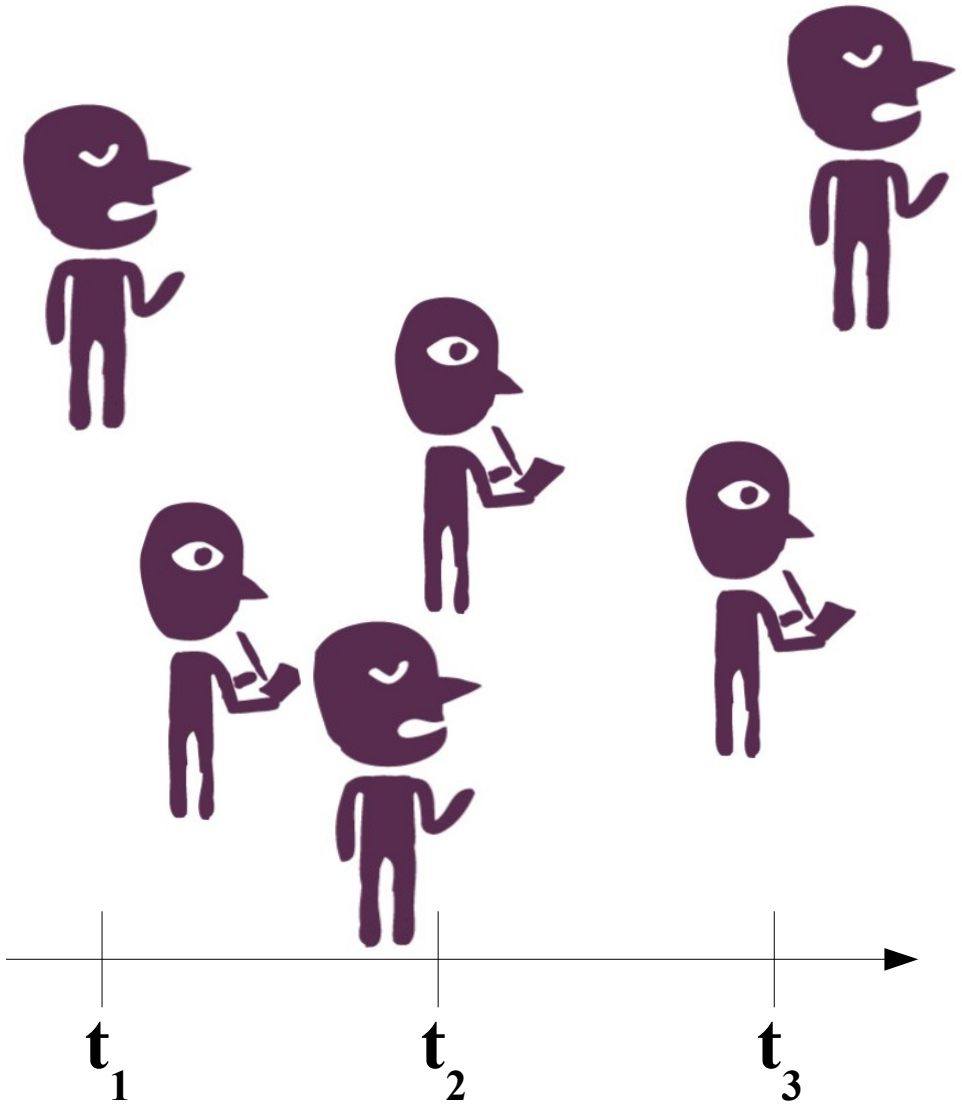


Partie I. Détection des évolutions typiques



Jeux de données:

Les valeurs des attributs descriptifs (x^d) pour plusieurs entités (φ) sont enregistrées à plusieurs dates (t)

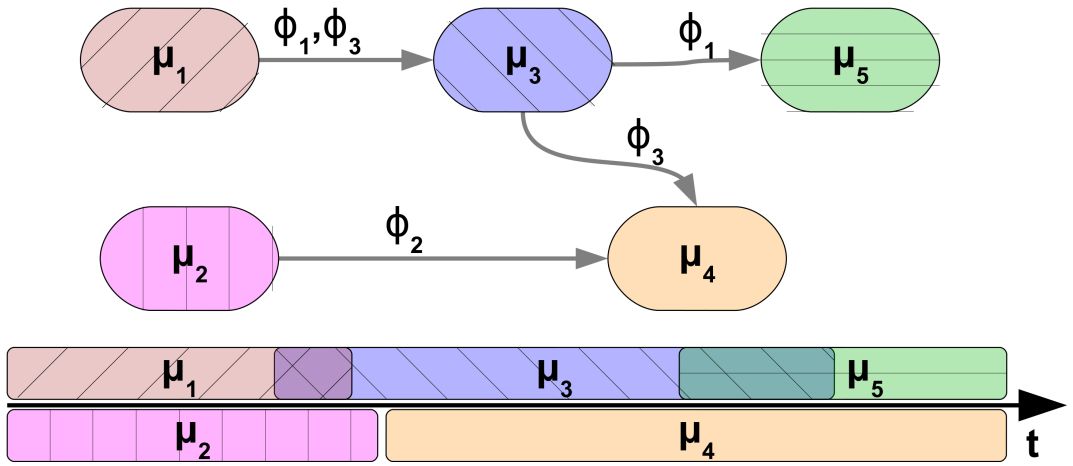


φ_1	t_1	x_1^d
	t_2	x_2^d
	t_3	x_3^d
φ_2	t_1	x_4^d
	t_2	x_5^d
	t_3	x_6^d

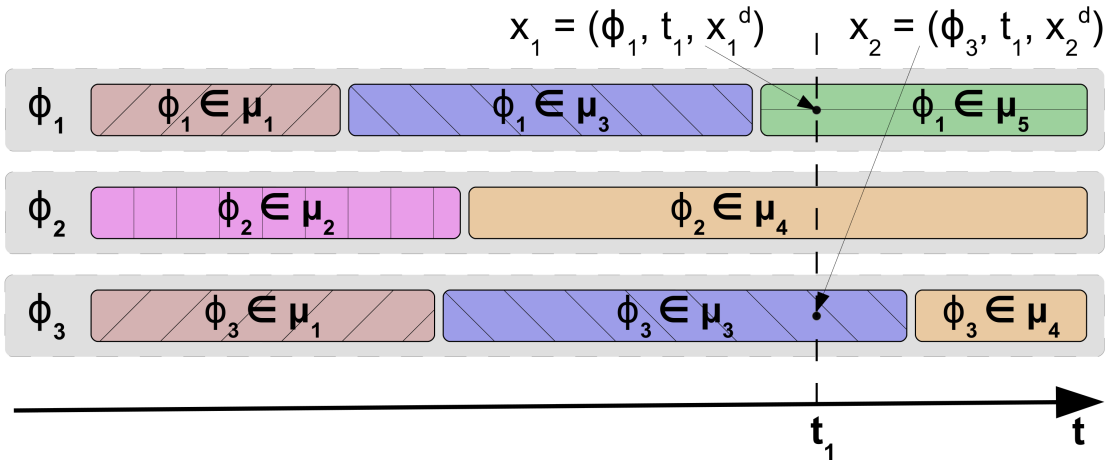
Enjeu de recherche : Utiliser la dimension temporelle dans l'analyse de données complexes

Tâche d'apprentissage : Détecter les motifs d'évolutions typiques pour les entités du jeu de données

a) phases parmi lesquelles passent les entités durant le temps



b) trajectoires des entités parmi les phases



Enjeu de recherche : Utiliser la dimension temporelle dans l'analyse de données complexes

Tâche d'apprentissage : Détecter les motifs d'évolutions typiques pour les entités du jeu de données

Difficultés :

- Les phases d'évolution ou les critères qui définissent les phases ne sont pas connus à l'avance
- Les phases doivent regrouper des observations similaires du point de vue descriptif et temporel
- Comment modéliser le temps dans l'algorithme de découverte des phases ?
- Les algorithmes de clustering temporel typiques [KIS10] généralement traitent des séries temporelles entières, nous travaillons au niveau d'observations.

Solution proposée (1)

Un algorithme de clustering temporel avec contraintes, les clusters obtenues servent comme phases d'évolution.

La partition obtenue doit assurer :

- la cohérence descriptive des clusters
- la cohérence temporelle des clusters

→ Mesure de dissimilarité temporelle (1)

- la segmentation contiguë des observations qui appartiennent à une entité

→ Contraintes de contiguïté (2)

Algorithme inspiré des K-Means. La fonction objective à minimiser est :

$$J = \sum_{\mu_j \in M} \sum_{x_i \in C_j} \left(\underbrace{\| x_i - \mu_j \|_{TA}}_{1} + \underbrace{\sum_{(x_k \notin C_j) \wedge (x_k^{\varphi} = x_i^{\varphi})} w(x_i, x_k)}_{2} \right)$$

Solution proposée (2) Les contributions

Mesure de dissimilarité temporelle → Distance dans l'espace descriptif et temporel

composante descriptive / composante temporelle

$$\|x_i - x_j\|_{TA} = 1 - \left(1 - \gamma_d \frac{\|x_i^d - x_j^d\|^2}{\Delta x_{max}^2}\right)\left(1 - \gamma_t \frac{|x_i^t - x_j^t|^2}{\Delta t_{max}^2}\right)$$

$$\gamma_d = \begin{cases} 1+\alpha, & si\ \alpha \leqslant 0 \\ 1, & si\ \alpha > 0 \end{cases}$$

$$\gamma_t = \begin{cases} 1, & si\ \alpha \leqslant 0 \\ 1-\alpha, & si\ \alpha > 0 \end{cases}$$

Propriétés :

- $\|x_i - x_j\|_{TA} \in [0,1], \forall\ x_i, x_j \in X$
- $\|x_i - x_j\|_{TA} = 0 \Leftrightarrow x_i^d = x_j^d \wedge x_i^t = x_j^t$
- $\|x_i - x_j\|_{TA} = 1 \Leftrightarrow \left(\gamma_d = 1 \wedge \|x_i^d - x_j^d\| = \Delta x_{max}\right) \vee \left(\gamma_t = 1 \wedge |x_i^t - x_j^t| = \Delta t_{max}\right)$

Solution proposée (2) Les contributions

Mesure de dissimilarité temporelle → Distance dans l'espace descriptif et temporel

Segmentation contiguë → Contraintes semi-supervisées MUST-LINK → **Fonction de pénalisation** dépendante du temps

Fonction de pénalisation

$$w(x_i, x_j) = \beta * e^{\frac{-1}{2}\left(\frac{|x_i^t - x_j^t|}{\delta}\right)^2}$$

$$pour\ x_i^\varphi = x_j^\varphi$$

(2 observations d'une même entité φ)

Solution proposée (2) Les contributions

Mesure de dissimilarité temporelle → Distance dans l'espace descriptif et temporel

Segmentation contiguë → Contraintes semi-supervisées MUST-LINK → **Fonction de pénalisation** dépendante du temps

L'algorithme TDCK-Means → Temporal-Driven Constrained K-Means

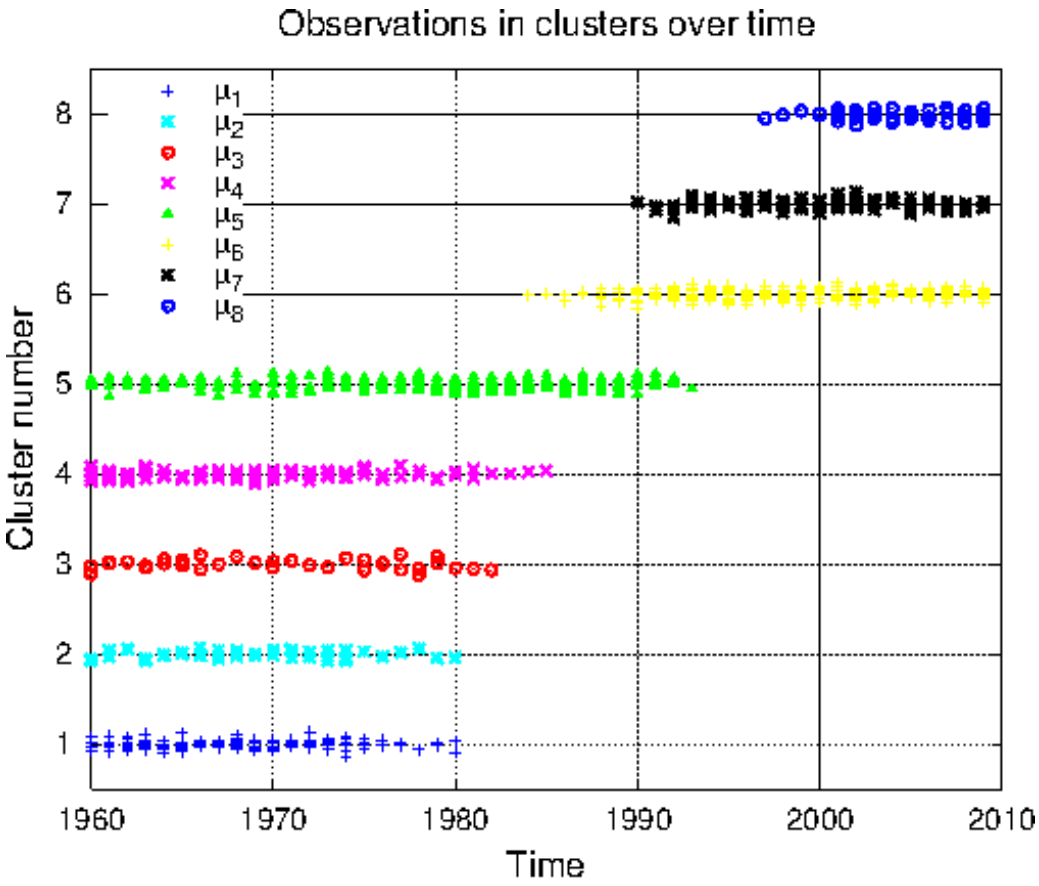
Inspiré des K-Means. Recalcule de façon itératif les centroïdes et l'appartenance des observations aux clusters.

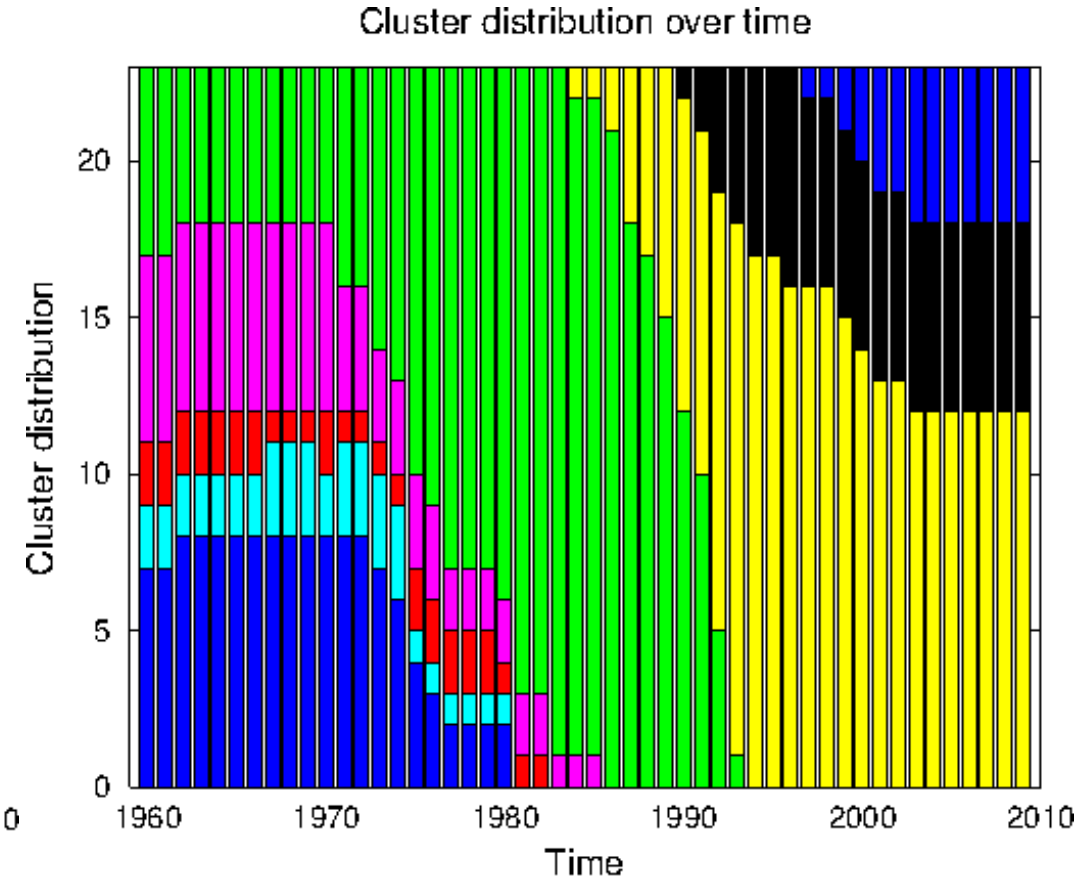
Utilise la mesure de dissimilarité temporelle et la fonction de pénalisation pour la segmentation contiguë

Centroïdes : (μ_j^t, μ_j^d)

Expérimentations et résultats

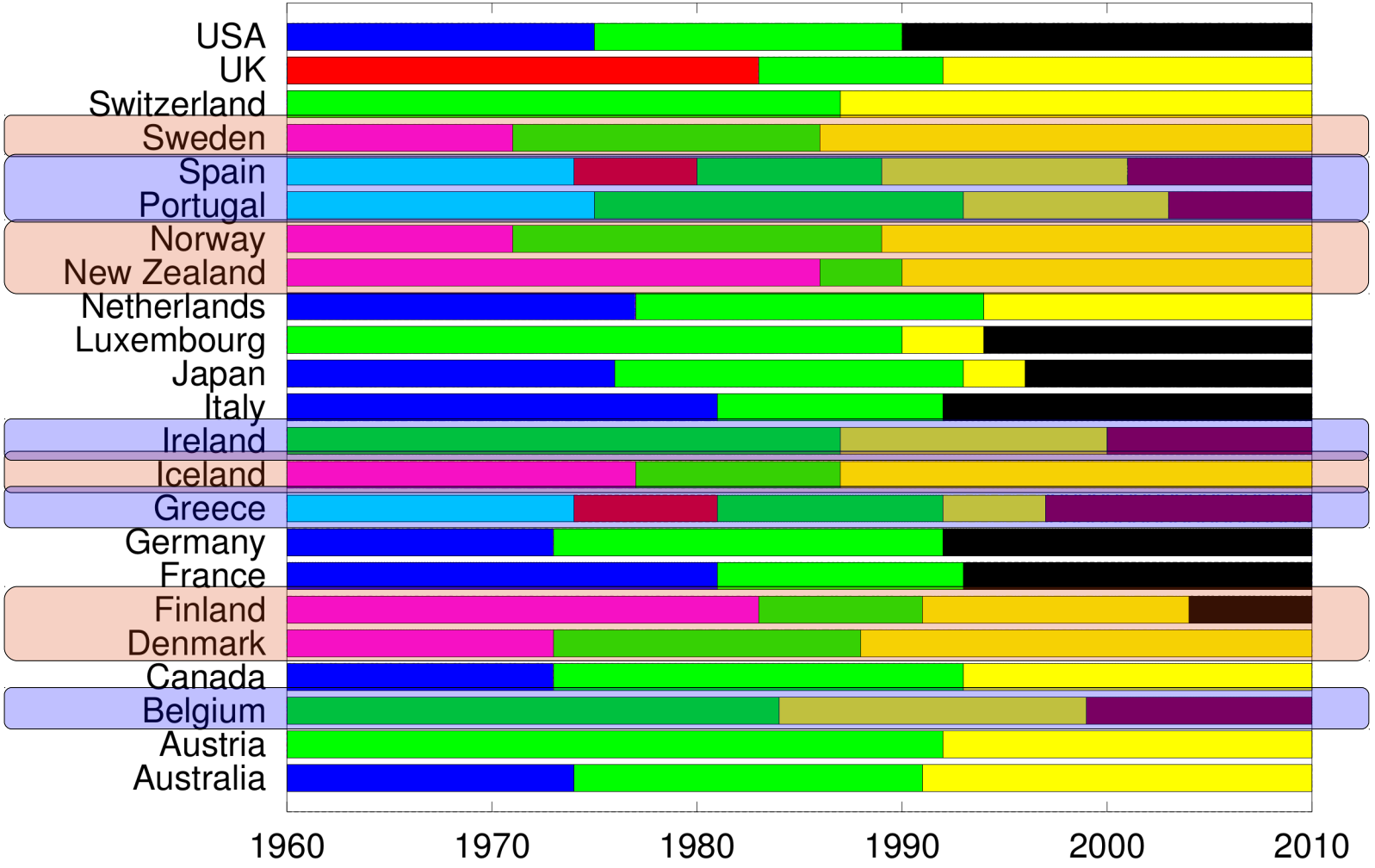
Compared Political Dataset I [ARM11] :
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.





Expérimentations et résultats

Compared Political Dataset I [ARM11] :
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.



Expérimentations et résultats

Compared Political Dataset I [ARM11] :
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.

Un exemple de graphe d'évolutions :
(construction postérieure au clustering)

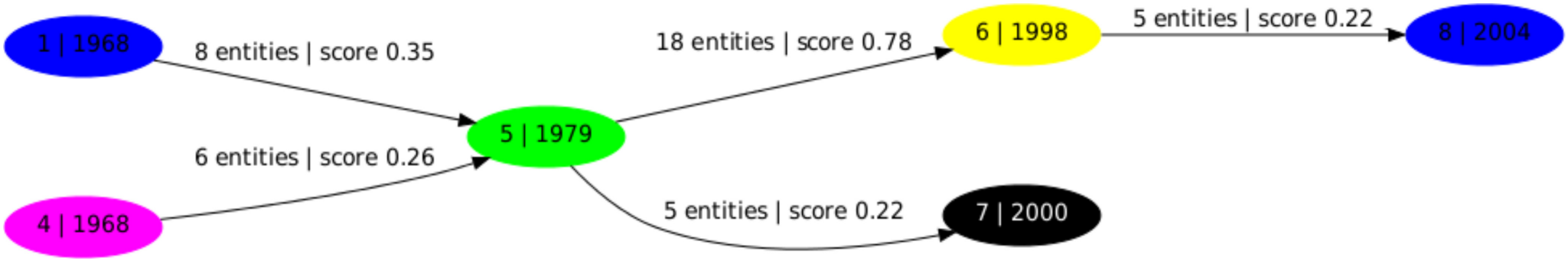
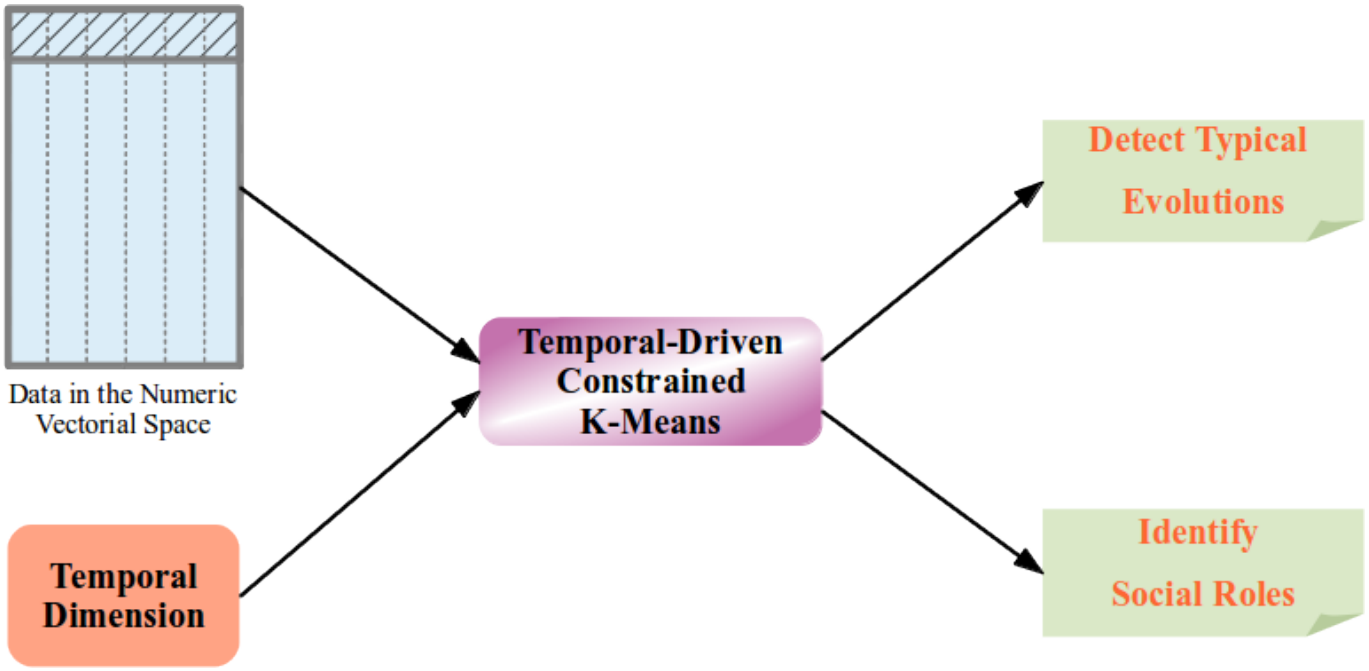


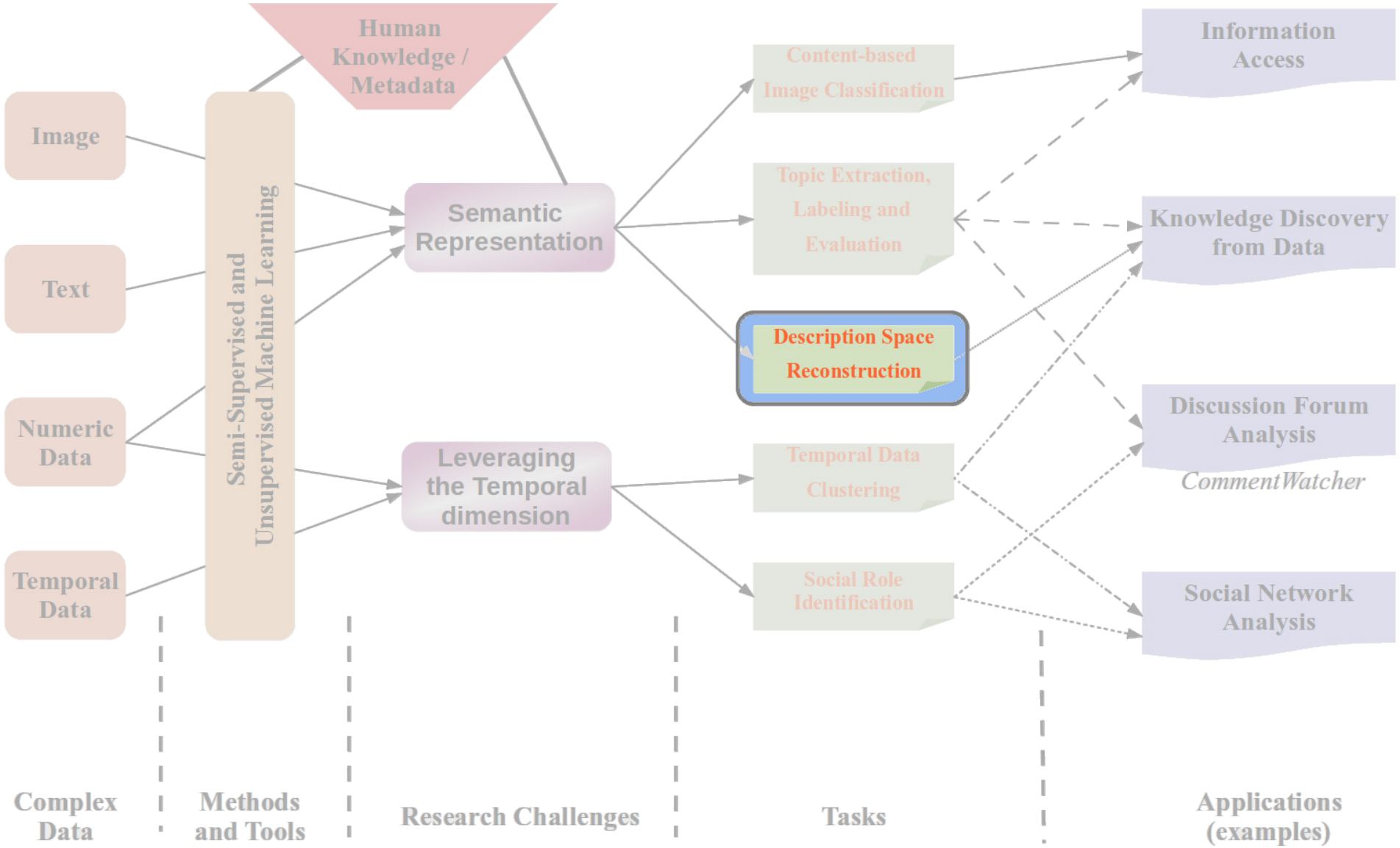
Schéma conceptuel



Publications :

- M.A. Rizoiu, J. Velcin, S. Lallich. *Structuring typical evolutions using Temporal-Driven Constrained Clustering*. In **International Conference on Tools with Artificial Intelligence (ICTAI'12)**, pp. 610–617. IEEE, 2012. **Best Student Paper Award**
- M.A. Rizoiu, J. Velcin, S. Lallich. *How to use Temporal Driven Constrained Clustering to detect typical evolutions*. **International Journal of Artificial Intelligence Tools**, 2013. (under review)

Partie II. Reconstruction de l'espace de représentation



Les données :

Représentation matricielle attribut-valeur
Les valeurs sont **booléennes**

	a_1	a_2	a_3	a_4	a_5	
φ_1		0	0			x_1^d
φ_2		1	0			x_2^d
φ_3		0	0			x_3^d
φ_4		0	0			x_4^d
φ_5		1	1			x_5^d
φ_6		1	1			x_6^d

Les défis :

- corrélation entre les attributs
- mauvaise représentation des données.

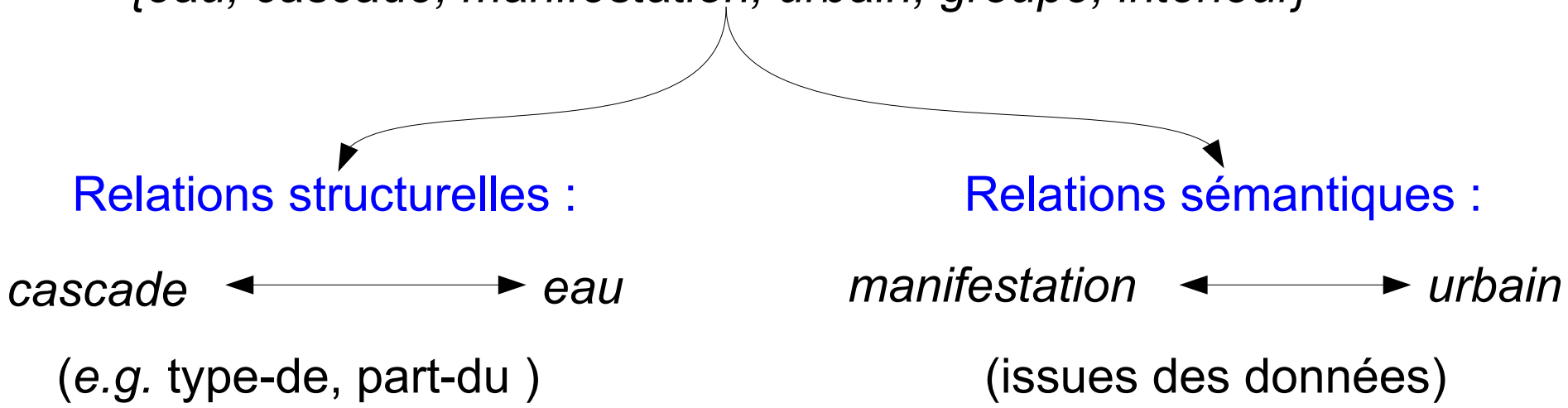
Enjeu de recherche : Utiliser la sémantique dans l'analyse des données

Tâches d'apprentissage :

- ➜ Construire un espace de représentation avec une sémantique enrichie
- ➜ Réduire les corrélations entre les attributs
- ➜ Découvrir des informations manquantes sur les relations entre les attributs

Étant donné l'ensemble des attributs :

{eau, cascade, manifestation, urbain, groupe, intérieur}



Enjeu de recherche : Utiliser la sémantique dans l'analyse des données

Tâches d'apprentissage :

- ➜ Construire un espace de représentation avec une sémantique enrichie
- ➜ Réduire les corrélations entre les attributs
- ➜ Découvrir des informations manquantes sur les relations entre les attributs

{groupes, rue, bâtiment, intérieur}

groupes ∧ $\overline{\textit{rue}}$ ∧ *intérieur*

groupes ∧ *rue* ∧ bâtiment

Enjeu de recherche : Utiliser la sémantique dans l'analyse des données

Tâches d'apprentissage :

- ➜ Construire un espace de représentation avec une sémantique enrichie
- ➜ Réduire les corrélations entre les attributs
- ➜ Découvrir des informations manquantes sur les relations entre les attributs

Difficultés :

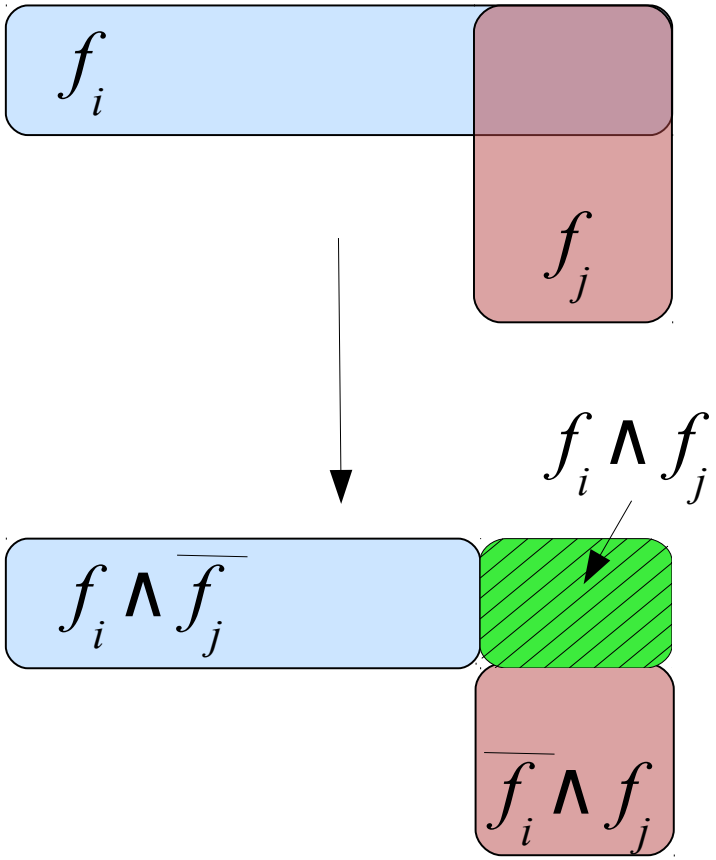
- ➜ Comment capturer les liens sémantiques entre les attributs ?
- ➜ Comment construire des nouveaux attributs compréhensibles ?
- ➜ La littérature sur la construction des attributs propose généralement plus des algorithmes supervisés [ZHE98]. Comment construire les nouveaux attributs de façon non-supervisée ?

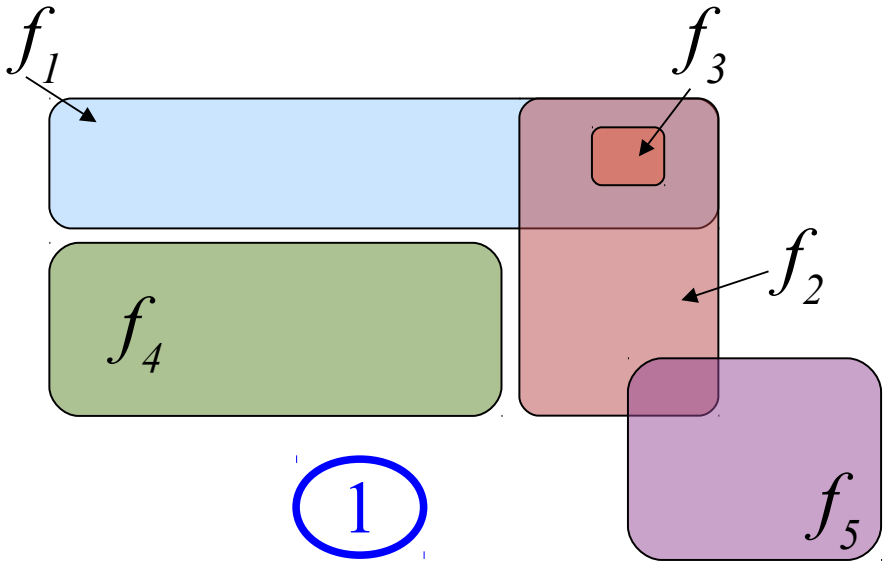
La solution proposée : Remplacer les paires d'attributs très corrélées par des conjonctions de littéraux (attributs ou leur négations)

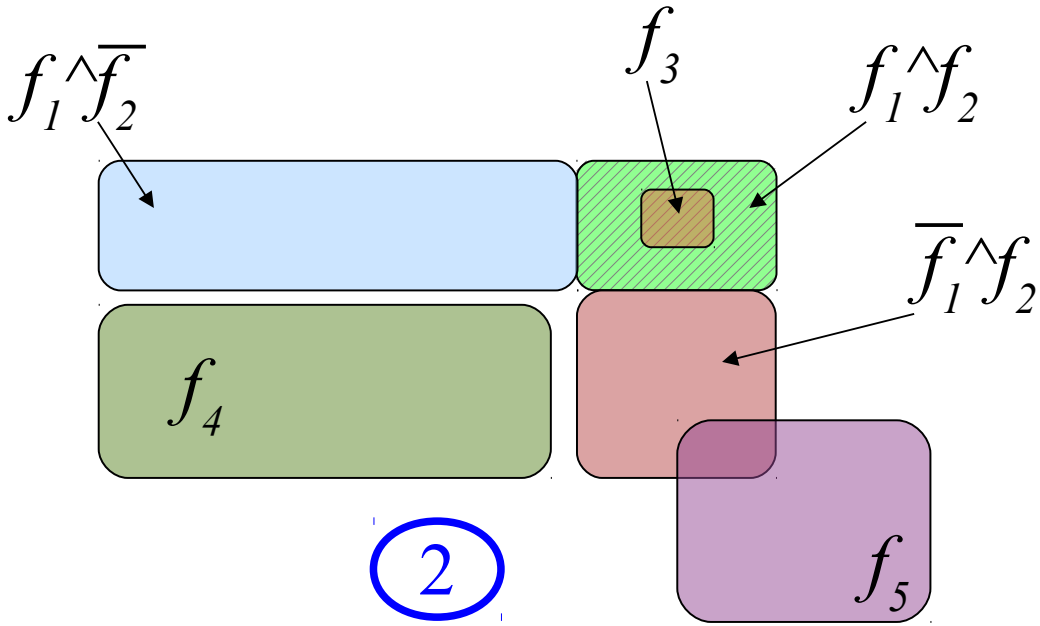
$$\{f_i, f_j\} \longrightarrow \{f_i \wedge f_j, f_i \wedge \overline{f_j}, \overline{f_i} \wedge f_j\}$$

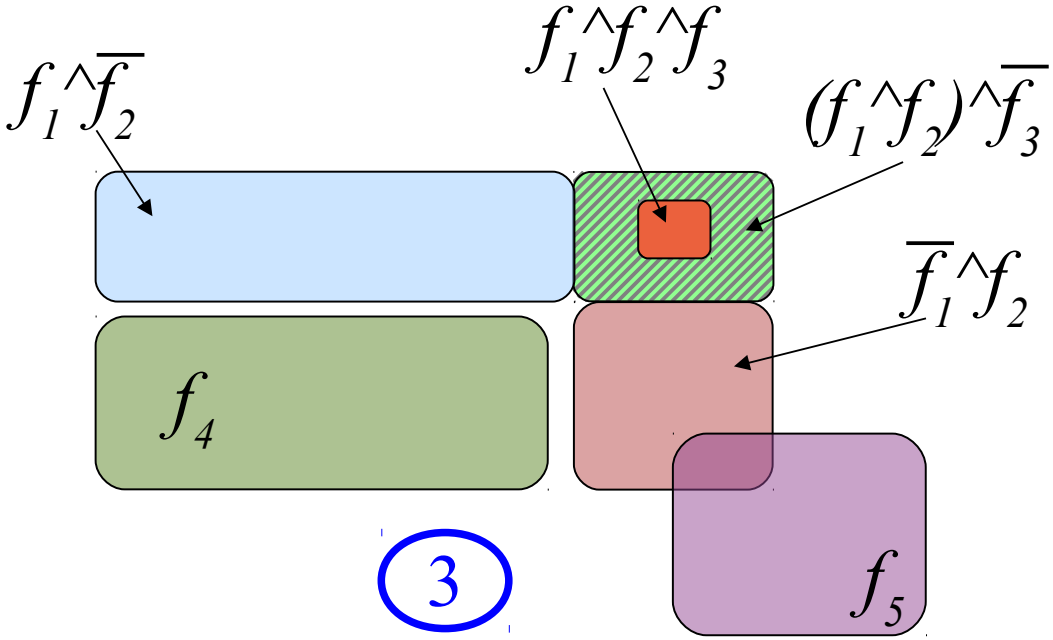
Algorithme *uFC* :

- Chercher des paires d'attributs corrélés (coefficient de Pearson)
- Construire de nouveaux attributs
- Éliminer les anciens attributs et les attributs sans support




f_1
f_3
f_2
f_4
f_5
1


$f_1 \wedge \overline{f_2}$
f_3
$f_1 \wedge f_2$
$\overline{f_1} \wedge f_2$
f_4
f_5
2


$f_1 \wedge \overline{f_2}$
$f_1 \wedge f_2 \wedge f_3$
$(f_1 \wedge f_2) \wedge \overline{f_3}$
$\overline{f_1} \wedge f_2$
f_4
f_5
3

Évaluation d'un ensemble d'attributs - deux mesures contraires :

Overlapping Index : basée sur la formule de Poincaré, évalue la corrélation totale de la collection des attributs

→ $OI(F) \in [0,1], doit\ être\ minimisé$

Complexité : nombre d'attributs par rapport au nombre maximal qui peut être construit ;

→ A chaque itération, le nombre des attributs augmente

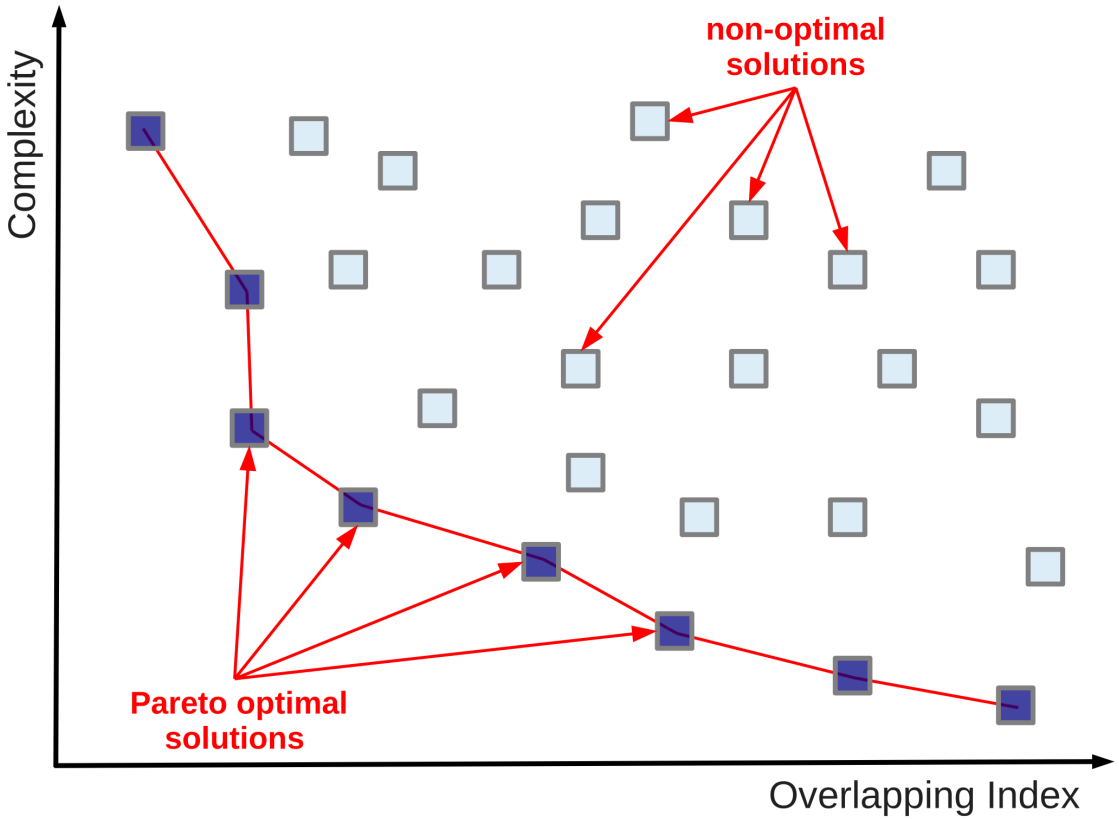
→ La longueur moyenne des attributs augmente

→ $C_0(F) \in [0,1], doit\ être\ minimisée$

Le compromis entre les deux critères opposés

- Optimisation simultanée des 2 critères opposés
- On utilise la notion d'optimalité Pareto
- Pas de solution unique, mais un ensemble de solutions – **Front de Pareto** [SAW85]

Idée :
Faire varier les 2 paramètres et projeter les solutions dans l'espace (OI, C_0)



Expérimentations et résultats

Trois jeux de données : **UCI Spect Heart** et les annotations associées avec 2 jeux de données image : **LabelMe** [RUS08], **Hungarian**

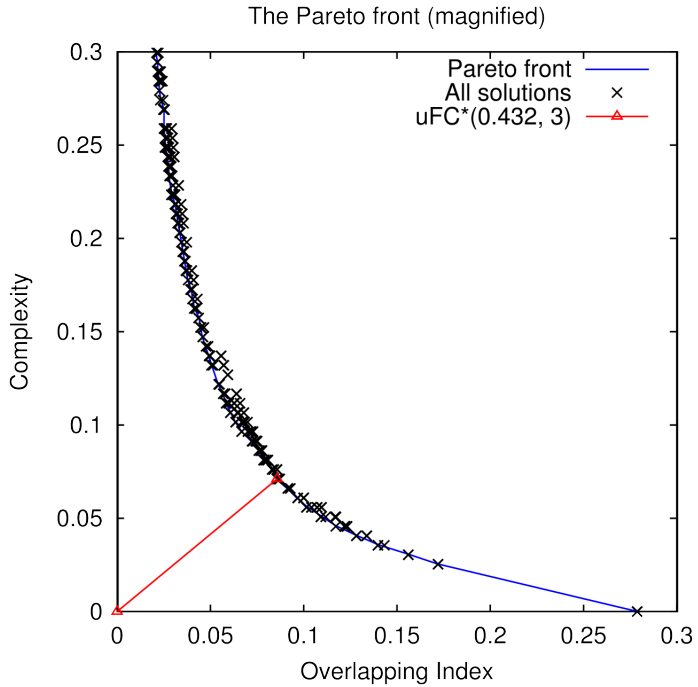
Heuristique pour choisir les paramètres :

« Closest-point » : choisir sur le front Pareto le point où le gain de score de recouvrement *(OI)* et la perte de complexité *(C_o)* et sont relativement égaux.

Le set reconstruit des attributs :

groupes ∧ *rue* ∧ *intérieur*
groupes ∧ *rue* ∧ *intérieur*
groupes ∧ *rue* ∧ *intérieur*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *panorama*
ciel ∧ *bâtiment* ∧ *panorama*

ciel ∧ *bâtiment* ∧ *panorama*
groupes ∧ *rue* ∧ *personne*
groupes ∧ *rue* ∧ *personne*
groupes ∧ *rue* ∧ *personne*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
eau ∧ *cascade*
arbre ∧ *forêt*
gazon
statue



Expérimentations et résultats

Trois jeux de données : **UCI Spect Heart** et les annotations associées avec 2 jeux de données image : **LabelMe** [RUS08], **Hungarian**

{ groupe, ciel, arbre, bâtiment, rue }

{ ciel ∧ bâtiment ∧ arbre ∧ $\overline{forêt}$, ciel ∧ groupe ∧ rue }

{ groupe, eau, cascade, ciel, arbre, gazon, forêt }

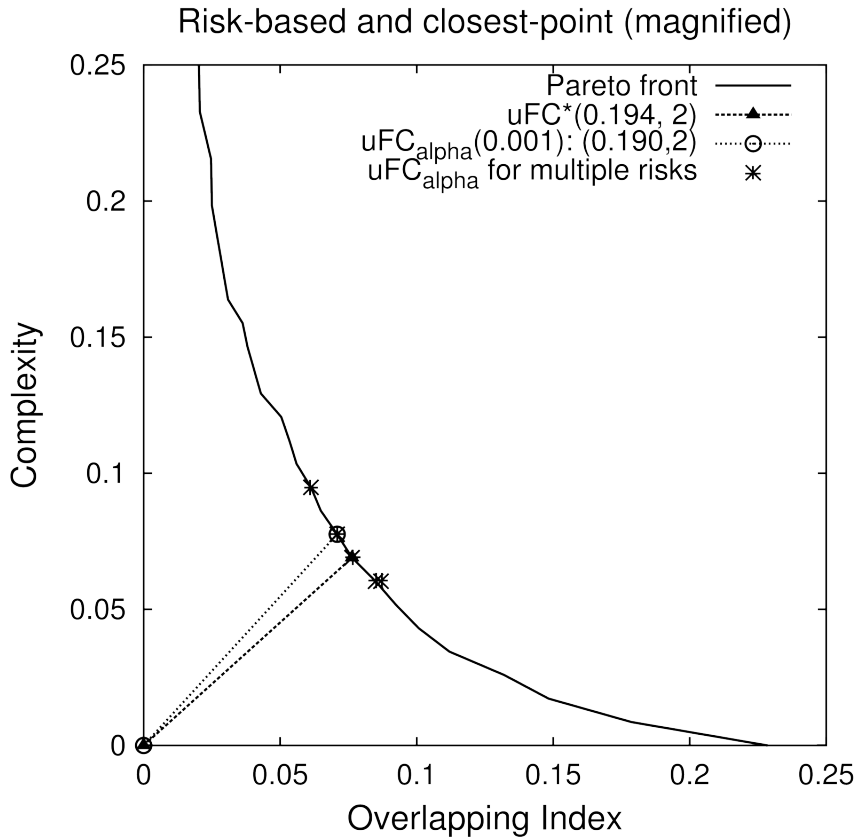
{ groupes ∧ $\overline{rue}$ ∧ $\overline{intérieur}$, eau ∧ cascade ∧ arbre ∧ forêt, ciel ∧ $\overline{bâtiment}$ ∧ $\overline{panorama}$ }

Expérimentations et résultats

Trois jeux de données : **UCI Spect Heart** et les annotations associées avec 2 jeux de données image : **LabelMe** [RUS08], **Hungarian**

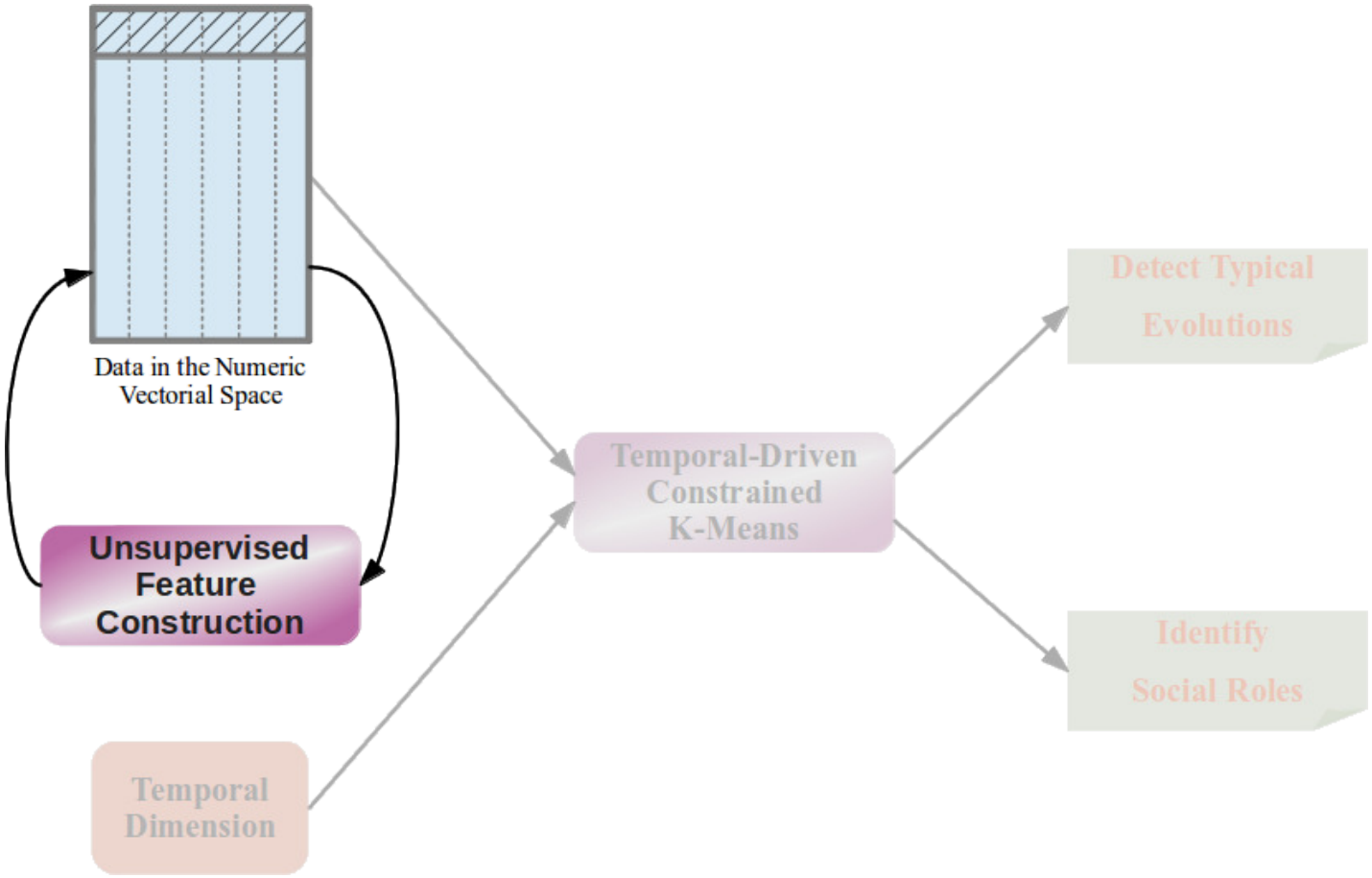
L'heuristique « risk-based » : éviter de couteuses exécutions multiples

- → Un des paramètres est choisi en utilisant des tests statistiques
- → Le deuxième est transformé dans une condition d'arrêt des itérations



Nous utilisons le front de Pareto pour évaluer les solutions choisies par rapport à la solution "closest-point"

Schéma conceptuel

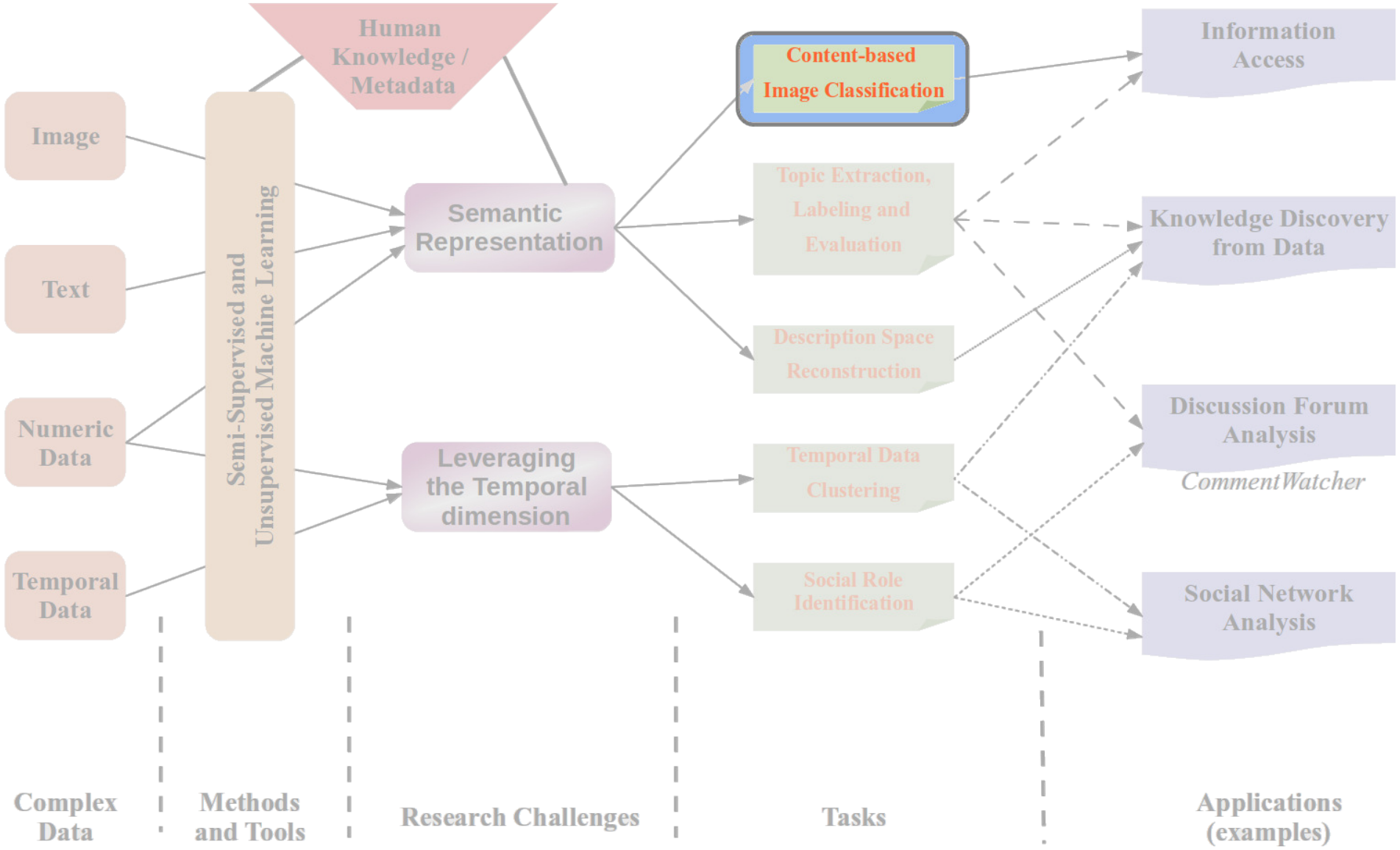


Publications :

- M.A. Rizoiu, J. Velcin, S. Lallich. *Unsupervised Feature Construction for Improving Data Representation and Semantics*. **Journal of Intelligent Information Systems**, vol. 40, no. 3, pages 501–527, 2013.

Partie III. Analyse de données image :

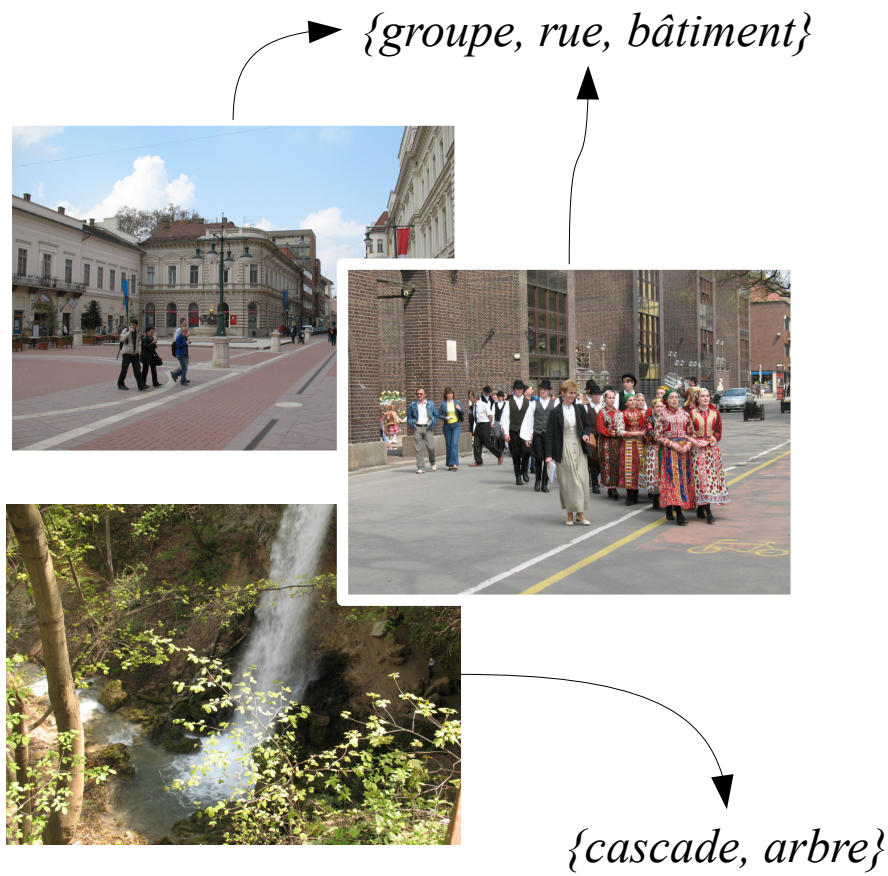
Représentation numérique avec une sémantique enrichie



Les données :

Une collection d'images, dont certaines sont annotées (ex. Picasa, réseaux sociaux *etc.*)

Annotations non positionnelles



Les défis :

- les caractéristiques de bas niveau captent peu d'information sur la sémantique
- prendre en compte les connaissances supplémentaires existantes

Enjeu de recherche :

Utiliser la sémantique dans l'analyse des données images

Tâches d'apprentissage :

- ➜ construire une représentation numérique des images avec une sémantique enrichie
- ➜ utiliser les annotations expertes pour enrichir la sémantique
- ➜ améliorer les performances d'une classification basée sur le contenu

Hypothèse :

la création d'une représentation basées sur une sémantique enrichie permet d'obtenir des performances en apprentissage plus élevées

Proposition :

Utiliser les annotations pour enrichir la représentation numérique des images

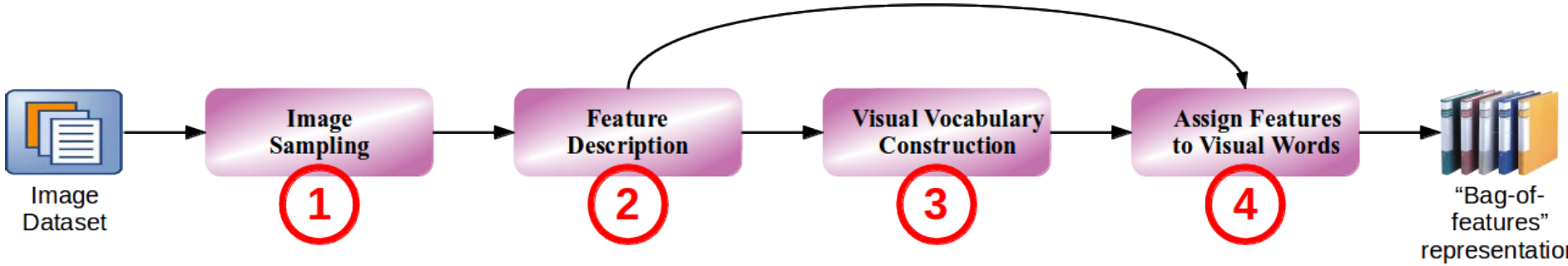
Construire une représentation numérique pour les images :

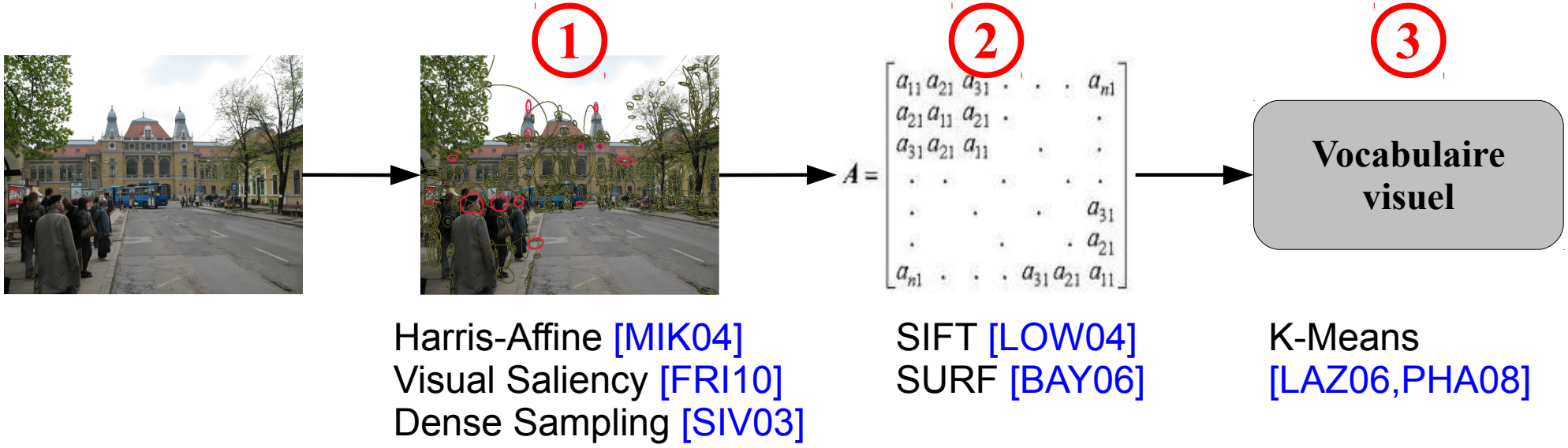
« sac-de-caractéristiques »

Concept de mot visuel : Une abstraction visuelle qui est prédictive pour un certain objet sémantique

Construire une représentation numérique pour les images :

Schéma de construction d'une représentation « *sac-de-caractéristiques* » :

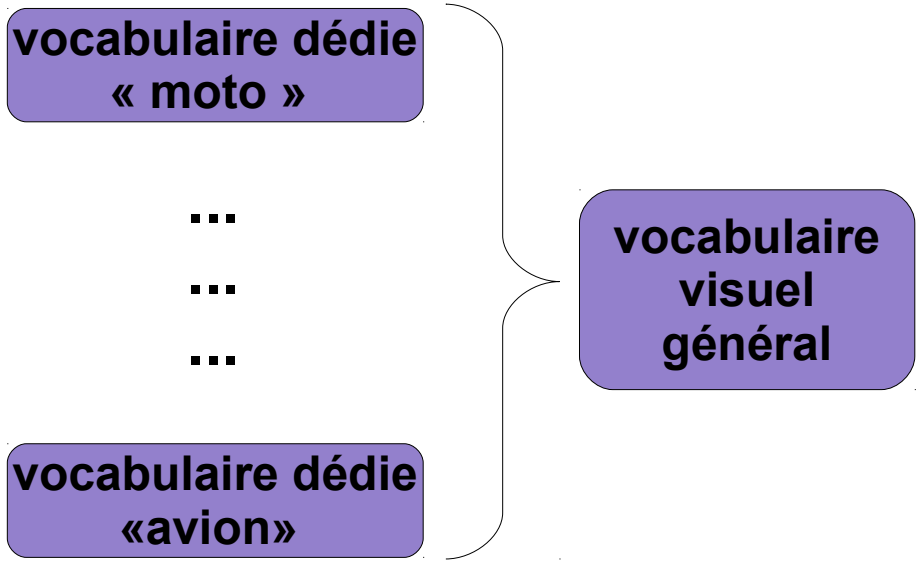




La tâche : Enrichir la sémantique de la représentation des images en utilisant les annotations dans la construction du vocabulaire visuel.

La solution proposée : Augmenter le rapport entre les caractéristiques pertinentes et le bruit.

Vocabulaires dédiés : Utiliser les annotations pour construire un vocabulaire visuel dédié pour chaque tag.



La tâche : Enrichir la sémantique de la représentation des images en utilisant les annotations dans la construction du vocabulaire visuel.

La solution proposée : Augmenter le rapport entre les caractéristiques pertinentes et le bruit.

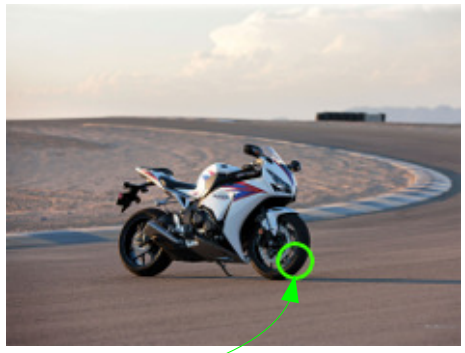
Filtrage des caractéristiques : Filtrer les caractéristiques qui sont susceptibles de ne pas être reliées à l'objet annoté

2 collections auxiliaires

Enlever les caractéristiques plus proches des exemples négatifs

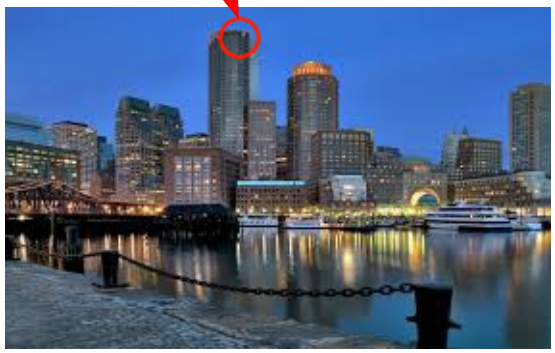
« moto »

Exemples positifs



« moto »

Exemples négatifs



«bâtiment»

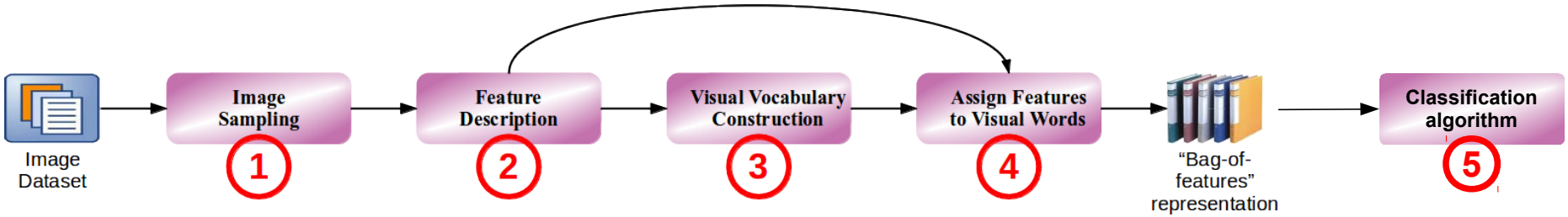
Expérimentations et résultats

Caltech101 [FEI07], RandCaltech101 [KIN10]

Protocole expérimental

Le but : Mesurer le gain relatif de performance obtenu en enrichissant la sémantique

La tâche : Classification d'images basée sur le contenu



	Dataset	model	filt+ model	random+ km
Clustering	Caltech101	13,96%	**15,69%**	4,36%
	Caltech101-3	6,58%	**7,36%**	2,73%
	RandCaltech101	20,49%	**26,27%**	12,07%
SVM	Caltech101	5,98%	12,02%	**12,05%**
	Caltech101-3	4,71%	**5,24%**	1,90%
	RandCaltech101	5,89%	**15,20%**	13,21%

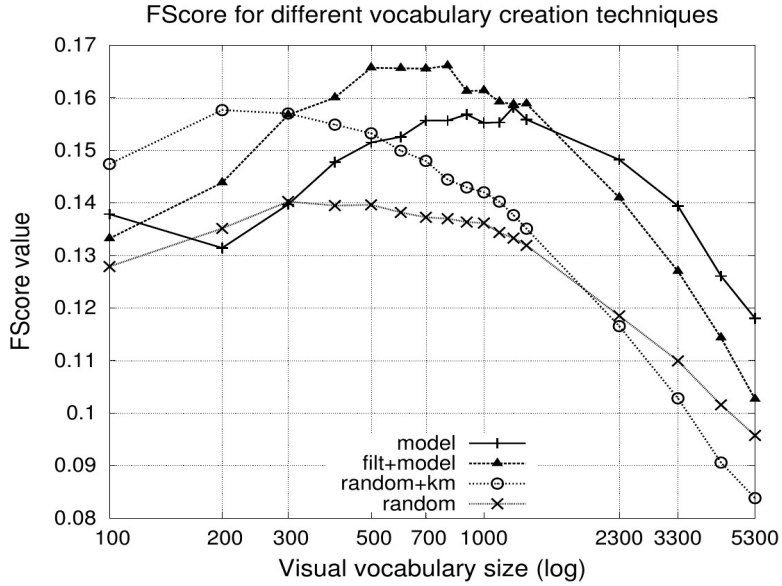
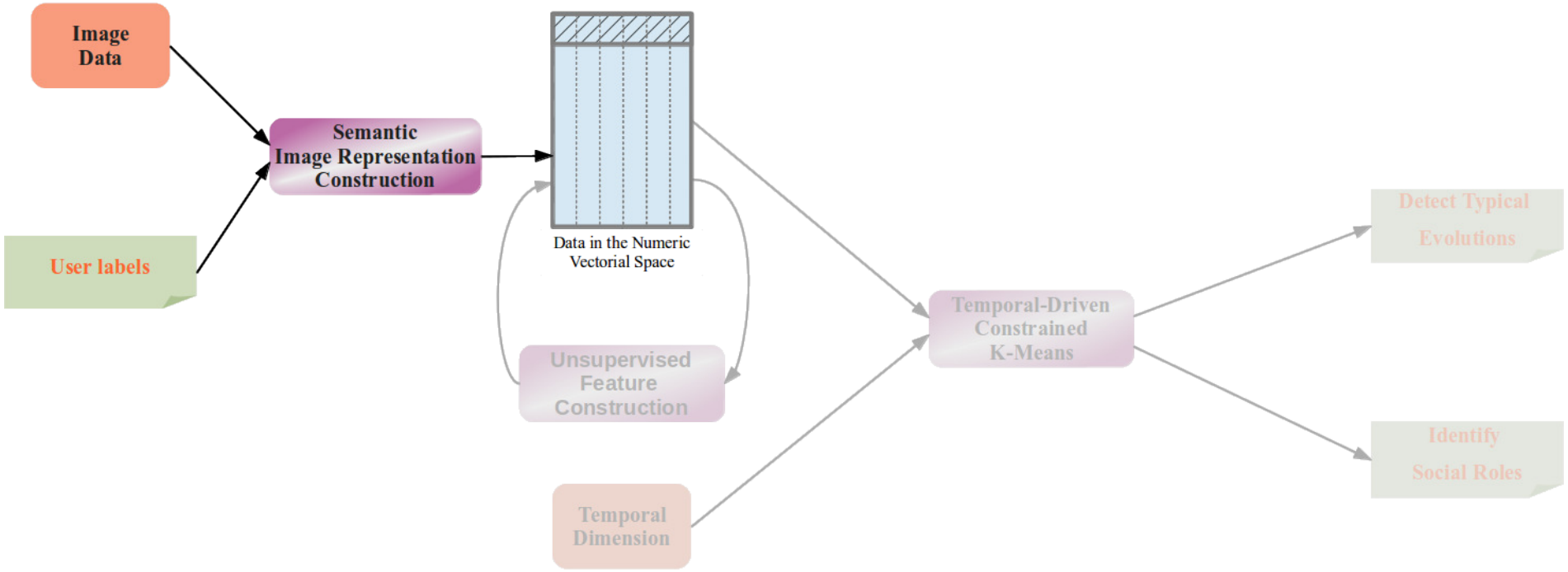


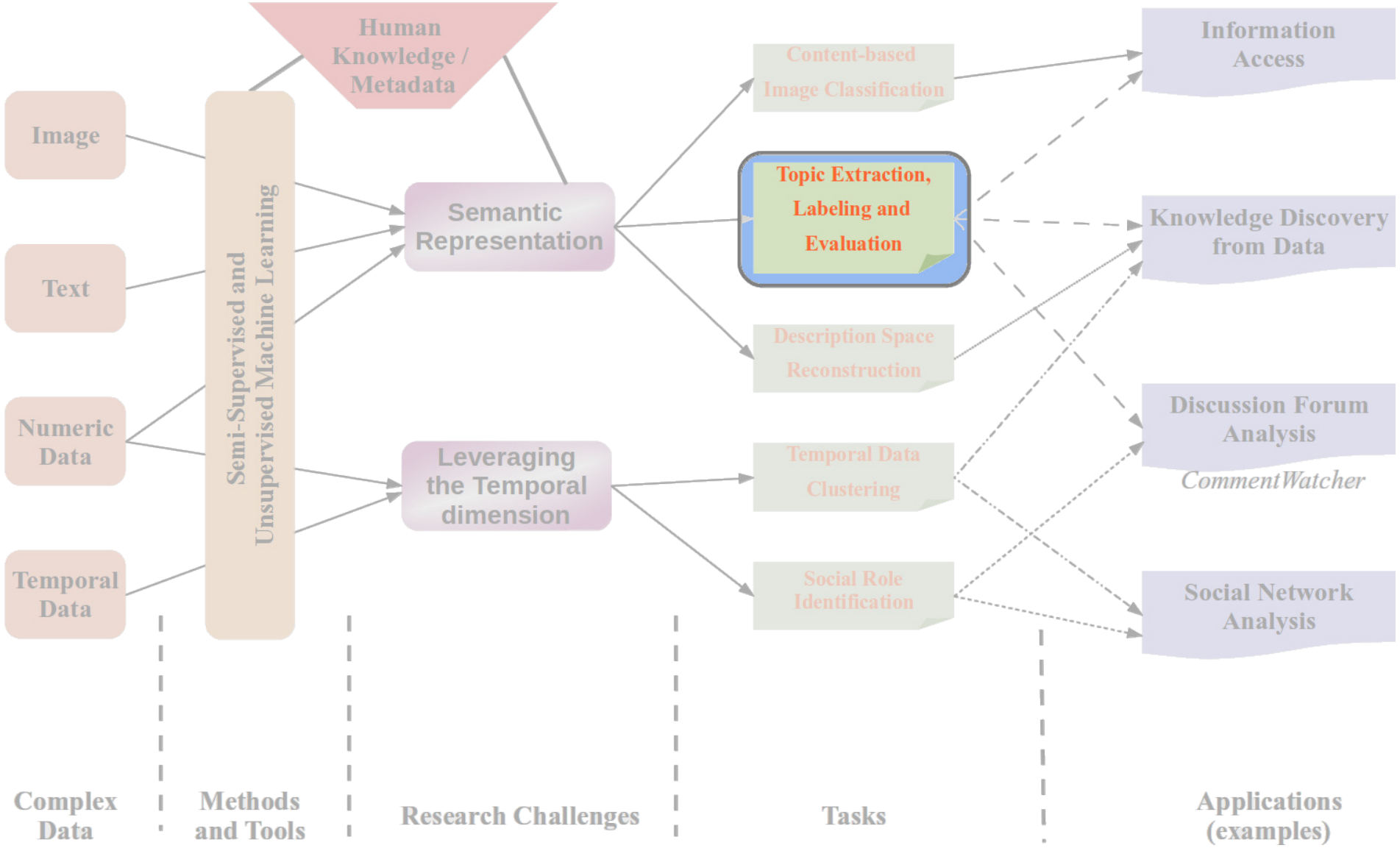
Schéma conceptuel



Publications :

- M.A. Rizoiu, J. Velcin, S. Lallich. *Semantic-enriched Visual Vocabulary Construction in a Weakly Supervised Context*. **Intelligent Data Analysis**, 2013. (under review)

Partie IV. Analyse de données textuelles : Extraction, étiquetage et évaluation des *thématiques*



Les données : Une collection de textes en langage naturel, souvent issus de l'internet.

Les défis :

- ➜ grands volumes de données ;
- ➜ besoin de résumer les « idées » principales : *les thématiques*
- ➜ la plupart de la littérature évalue les thématiques à l'aide des mesures statistiques, sans prendre en compte la sémantique

ex. indice de perplexité [WAL09]

Enjeu de recherche :

Utiliser la sémantique dans l'analyse des données textuelles

Tâches d'apprentissage :

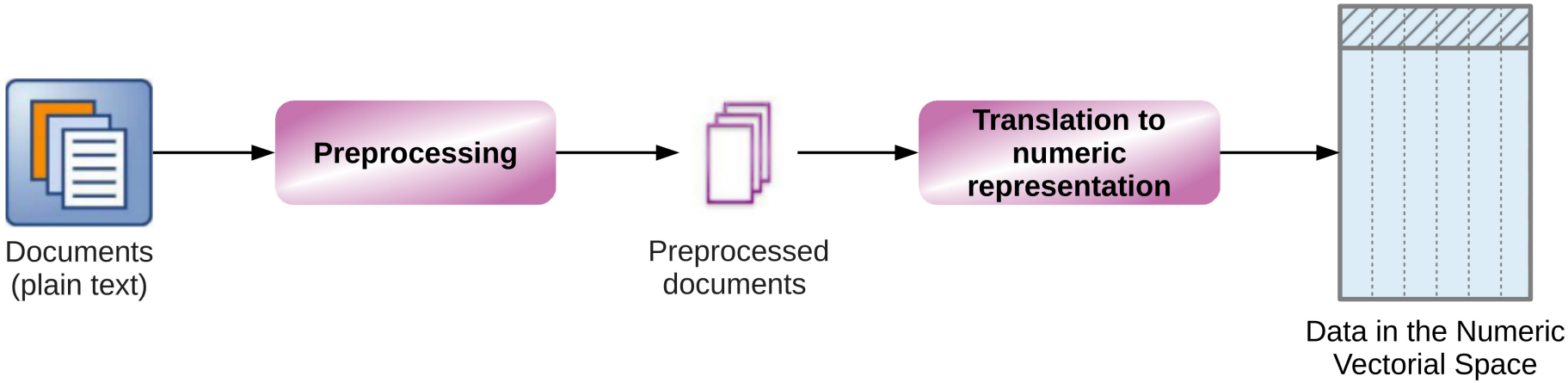
- extraction des thématiques
- étiquetage des thématiques avec des noms compréhensibles pour un être humain
- utilisation des connaissances sémantiques dans l'évaluation de ces thématiques.

Dimension appliquée :

- travaux intimement liés au différents projets de recherche (CRTT-ERIC, ImagiWeb *etc.*) ;
- implémentation dans le logiciel **CommentWatcher** ;

Solution proposée (1) : Extraction des thématiques à l'aide du clustering recouvrant

Plonger les textes dans un espace numérique : « *sac-de-mots* » :



Regroupement des documents à l'aide du clustering recouvrant : OKM [CLE08]

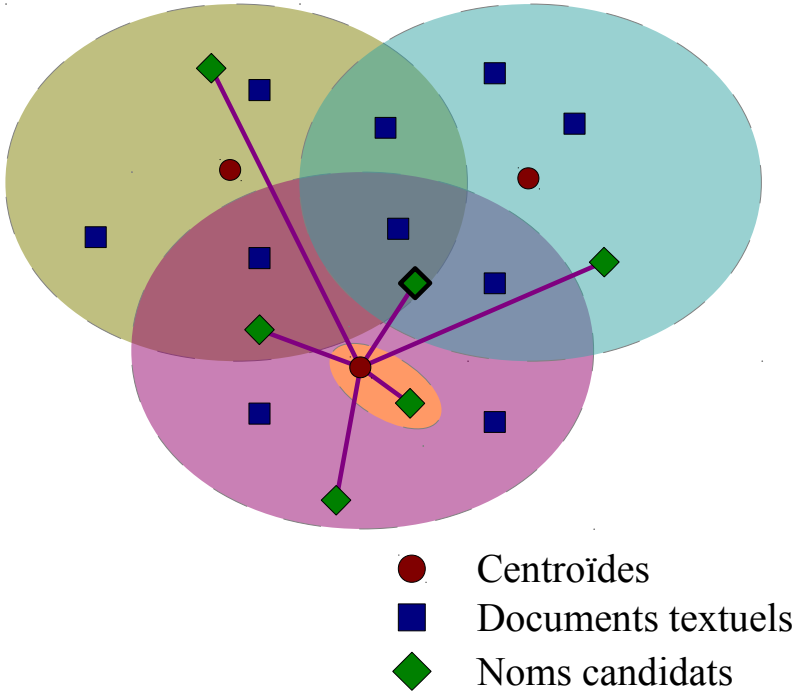
- Une extension des KMeans qui autorise un document à appartenir à plusieurs clusters ;

Solution proposée (2) : Étiquetage des thématiques

- Extraire des expressions complètes fréquentes à partir du texte original ;

  tableaux de suffixes **[MAN93]**.

- Injecter les expressions comme des pseudo-documents et calculer la similarité.



Solution proposée (3) : Évaluer la cohésion sémantique des thématiques

Hypothèse sous-jacente : Les mesures statistiques ne parviennent pas totalement à émuler le jugement humain [CHA09]

Idée : Relier une distribution statistique de fréquences à une structure sémantique (ex. WordNet [MIL95])

Solution proposée (3) : Évaluer la cohésion sémantique des thématiques

Alignement des thématiques :

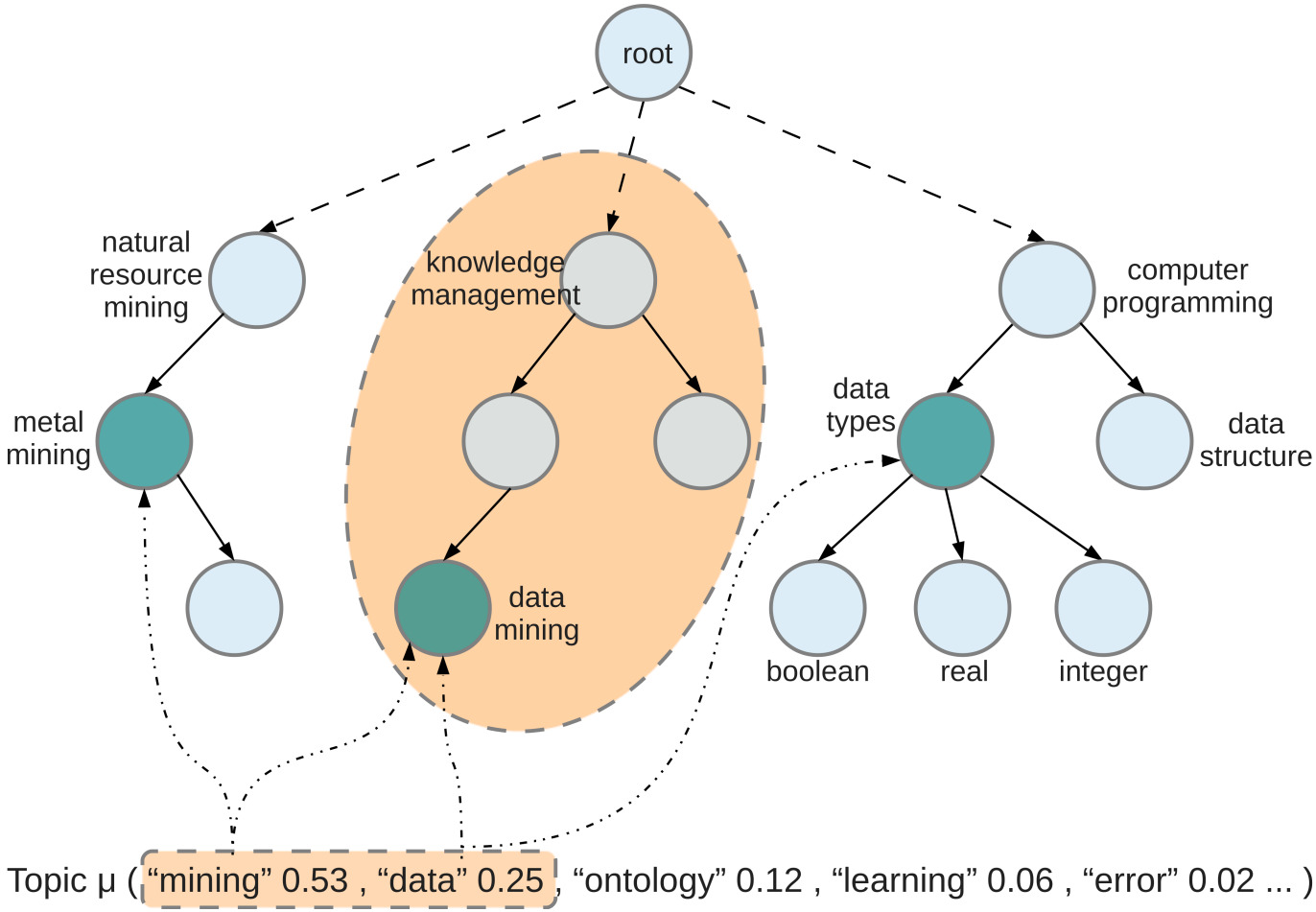
Déterminer le sous-arbre le plus spécifique qui contient au moins un sens pour chacun des mots les plus représentatifs de la thématique.

Mesures :

- couverture
- spécificité

Évaluation d'une thématique :

$$\varphi(\mu, c) = \omega_{spec}\, spec(\mu, c) + \omega_{cov}\, cov(\mu, c)$$



Expérimentations et résultats

Reuters, **Suall11**

Le forum « Y a-t-il trop de commémorations en France? », sur www.liberation.fr

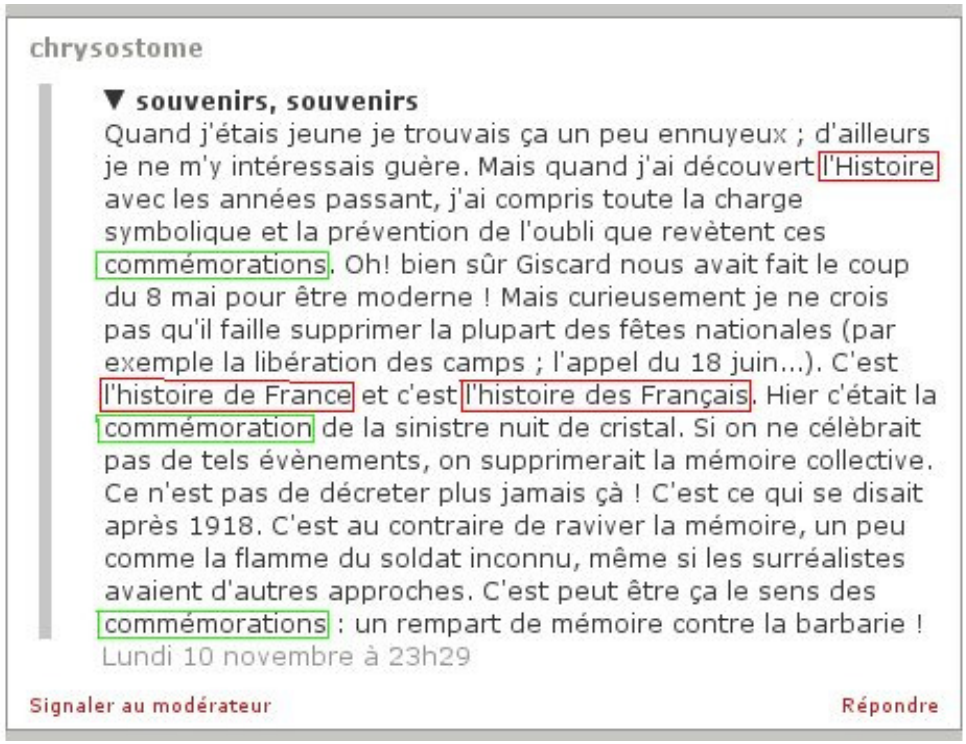
Corpus économique extrait à partir du site d'Associated Press

```
--> Iteration no 11:
---> Objective function value: 189.154
---> Partitions:
----> Cluster 0 [101]: .............
----> Cluster 1 [90]: .............
----> Cluster 2 [128]: ..... texte_81 .............
----> Cluster 3 [192]: ..................... texte_81 ....

Result - Cluster description:

-> Centroid[0]: "jours fériés"
-> Centroid[1]: "travailler plus pour gagner"
-> Centroid[2]: "commémoration"
-> Centroid[3]: "histoire de france"
```

Exemple de sortie du logiciel d'extraction de thématiques, inclus dans **CommentWatcher**



chrysostome

▼ **souvenirs, souvenirs**

Quand j'étais jeune je trouvais ça un peu ennuyeux ; d'ailleurs je ne m'y intéressais guère. Mais quand j'ai découvert l'Histoire avec les années passant, j'ai compris toute la charge symbolique et la prévention de l'oubli que revètent ces commémorations. Oh! bien sûr Giscard nous avait fait le coup du 8 mai pour être moderne ! Mais curieusement je ne crois pas qu'il faille supprimer la plupart des fêtes nationales (par exemple la libération des camps ; l'appel du 18 juin...). C'est l'histoire de France et c'est l'histoire des Français. Hier c'était la commémoration de la sinistre nuit de cristal. Si on ne célèbrait pas de tels évènements, on supprimerait la mémoire collective. Ce n'est pas de décreter plus jamais çà ! C'est ce qui se disait après 1918. C'est au contraire de raviver la mémoire, un peu comme la flamme du soldat inconnu, même si les surréalistes avaient d'autres approches. C'est peut être ça le sens des commémorations : un rempart de mémoire contre la barbarie !

Lundi 10 novembre à 23h29

Signaler au modérateur Répondre

Document « texte_81 » sur le site du forum

Expérimentations et résultats

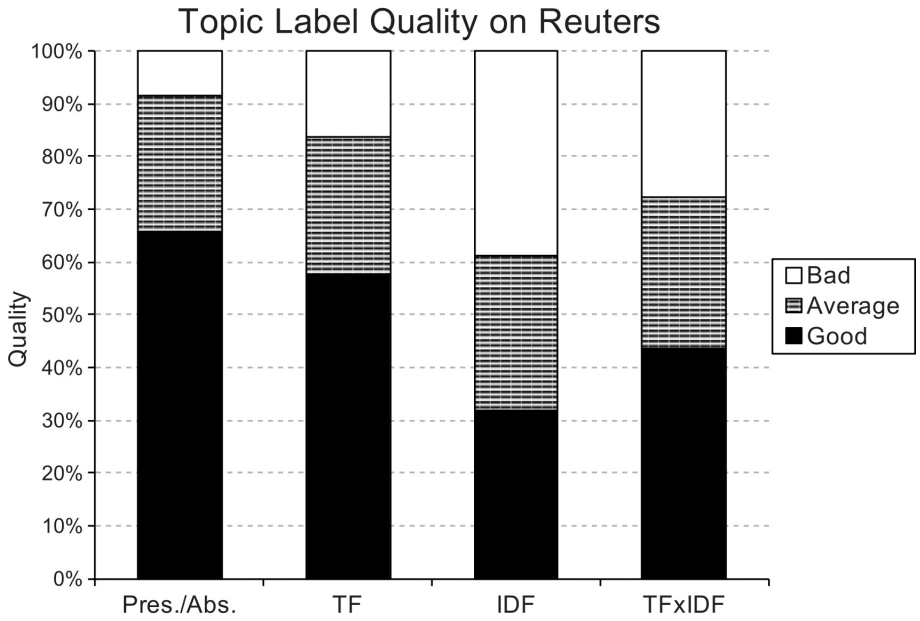
Reuters, **Suall11**

Le forum « Y a-t-il trop de commémorations en France? », sur www.liberation.fr

Corpus économique extrait à partir du site d'Associated Press

Protocole :

Basé sur des experts, inspiré de la littérature [CHA09]

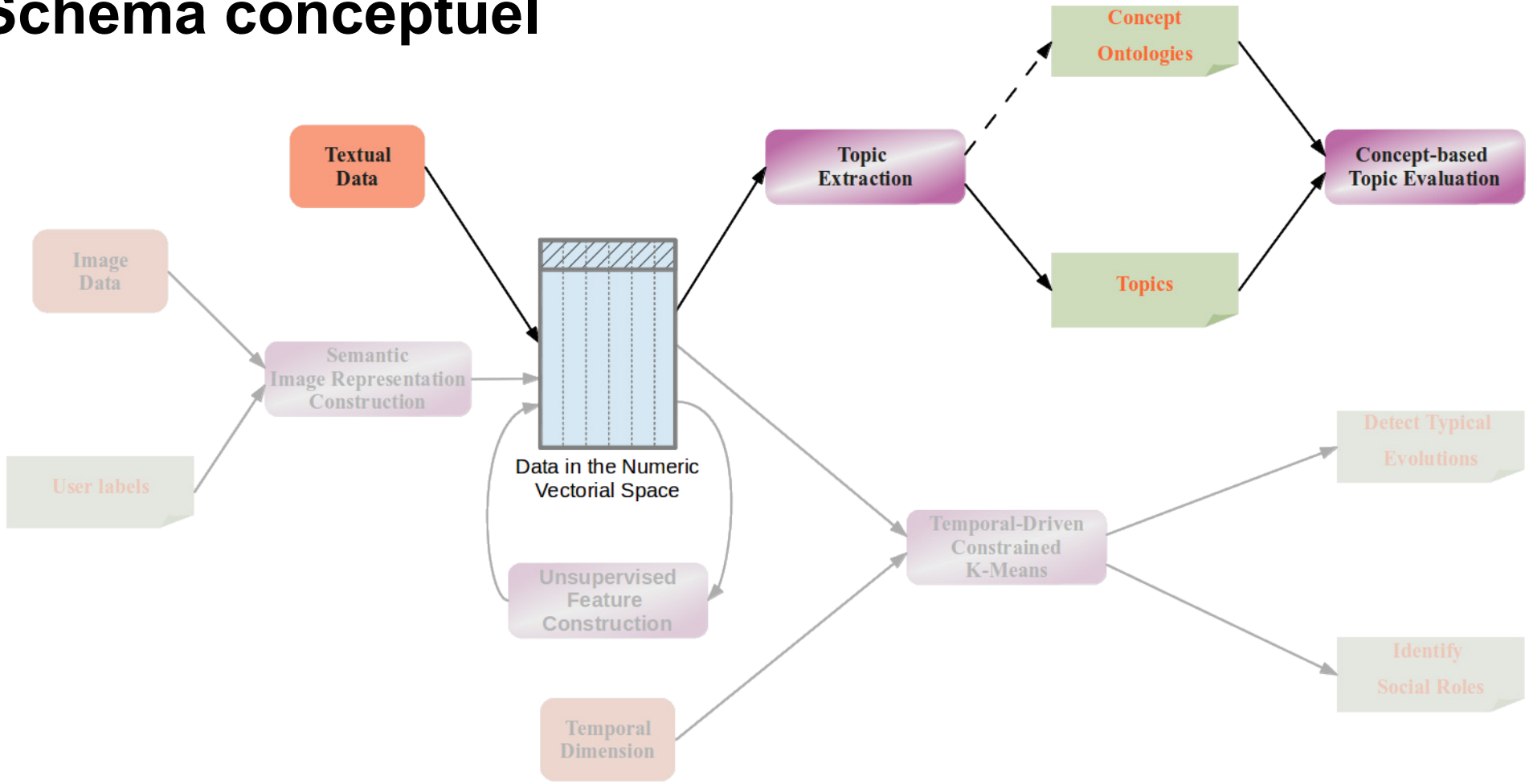


Évaluation de noms des thématiques

Dataset	$\overline{hit}_+$	$\overline{hit}_-$	Avantage rel. $\overline{hit}$
AP	**0,69**	0,65	6,93 %
Suall11	**0,75**	0,59	28,55 %

Évaluation de l'alignement des thématiques aux sous-arbres des concepts

Schéma conceptuel



Publications :

- C. Musat, J. Velcin, S. Trausan-Matu, M.A. Rizoiu. *Improving topic evaluation using conceptual knowledge*. In **International Joint Conference on Artificial Intelligence (IJCAI'11)**, pp. 1866–1871, 2011.
- C. Musat, J. Velcin, M.A. Rizoiu,S. Trausan-Matu. *Concept-based Topic Model Improvement*. In **International Symposium on Methodologies for Intelligent Systems (ISMIS'11)**, pp. 133–142, 2011.
- M.A. Rizoiu, J. Velcin, J.H. Chauchat. *Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes*. In **Extraction et Gestion des Connaissances (EGC'10)**, pp. 561–572, 2010.

Conclusions

Deux enjeux de recherche

1) utiliser la sémantique pour analyser les données complexes

2) prendre en compte la dimension temporelle des données complexes

Contributions les plus importantes

- ➜ utiliser la dimension temporelle et descriptive des données dans un algorithme de clustering, afin de détecter des évolutions typiques
- ➜ utiliser la sémantique des données pour améliorer l'espace de représentation et reconstruire les attributs
- ➜ enrichir la sémantique de la représentation des images en utilisant des connaissances additionnelles sous forme d'annotations
- ➜ utiliser un alignement des thématiques aux sous-arbres des concepts et évaluer la cohésion sémantique des thématiques

Travaux à court terme

- Appliquer l'algorithme TDCK-Means à une autre problématique d'apprentissage : la **détection de rôles sociaux dans les communautés en ligne**

- Construire simultanément une structure de graphe entre les clusters temporels obtenus

- Construction temporelle des attributs

  Prendre en compte la dimension temporelle dans la reconstruction sémantique de l'espace de description

  Détecter des corrélations avec un certain délai de temps δ

Travaux futurs

- Adapter l'algorithme de construction de représentation des images à l'annotation incomplète en utilisant l'algorithme de construction des attributs

- Déterminer automatiquement les valeurs des paramètres de TDCK-Means (α, β, δ, γ), en utilisant une approche inspirée de l'optimisation multi-critère à l'aide des algorithmes génétiques [ZHA07]

- Implémenter notre proposition d'évaluation sémantique de thématiques dans CommentWatcher

Publications pendant la thèse				
Journaux internationaux	Conférences internationales	Chapitres de livre internationaux	Ateliers internationaux	Conférences nationales
1 (*JIIS'13*)	**3** (*ICTAI'12, IJCAI'11, ISMIS'11*)	**1** (*Ontology Learning '11*)	**1** (*IJCAI'13*)	**1** (*EGC'10*)

Développement de logiciels académiques open-source :

CommentWatcher : analyser les discussions sur des forums en ligne
Modules : récupération des messages des forums, extraction et visualisation de thématiques, visualisation des réseaux sociaux

CKP : extraction et étiquetage des thématiques

Je vous remercie pour votre attention !

Bibliographie

[WAG00] Kiri Wagstaff and Claire Cardie. Clustering with Instance-level Constraints. In International Conference on Machine Learning, Proceedings of the Seventeenth, pages 1103–1110, 2000.

[ZHE98] Zijian Zheng. Constructing conjunctions using systematic search on decision trees. Knowledge-Based Systems, vol. 10, no. 7, pages 421–430, 1998.

[SAW85] Y. Sawaragi, H. Nakayama and T. Tanino. Theory of multiobjective optimization, volume 176. Academic Press New York, 1985.

[RUS08] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy and William T. Freeman. LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, vol. 77, no. 1, pages 157–173, 2008.

[MIK04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. International Journal of Computer Vision, vol. 60, no. 1, pages 63–86, 2004.

[FRI10] Simone Frintrop, Erich Rome and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP), vol. 7, no. 1, page 6, 2010.

[KIS10] Slava Kisilevich, Florian Mansmann, Mirco Nanni and Salvatore Rinzivillo. Spatio-temporal clustering. Data mining and knowledge discovery handbook, pages 855–874, 2010.

[SIV03] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In Computer Vision, Proceedings of the Ninth IEEE International Conference on, ICCV 2003, pages 1470–1477. IEEE, 2003.

[LOW04] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pages 91–110, 2004.

[BAY06] Herbert Bay, Tinne Tuytelaars and Luc Van Gool. Surf: Speeded up robust features. Computer Vision–ECCV 2006, pages 404–417, 2006.

[LAZ06] Svetlana Lazebnik, Cordelia Schmid and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2169–2178. IEEE, 2006.

[KIN10] Teemu Kinnunen, Joni Kristian Kamarainen, Lasse Lensu, Jukka Lankinen and Heikki Kälviäinen. Making Visual Object Categorization More Challenging: Randomized Caltech-101 Data Set. In 2010 International Conference on Pattern Recognition, pages 476–479. 2010.

[FEI07] Li Fei-Fei, Rob Fergus and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, vol. 106, no. 1, pages 59–70, 2007.

[ARM11] Klaus Armingeon, David Weisstanner, Sarah Engler, Panajotis Potolidis, Marlène Gerber and Philipp Leimgruber. Comparative Political Data Set 1960-2009. Institute of Political Science, University of Berne., 2011.

[WAL09] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov and David Mimno. Evaluation methods for topic models. In International Conference on Machine Learning, Proceedings of the 26th Annual, pages 1105–1112. ACM, 2009.

[MAN93] Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. SIAM Journal on Computing, vol. 22, no. 5, pages 935–948, 1993.

[CLE08] Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1–4. IEEE, 2008.

[MIL95] George A. Miller. WordNet: a lexical database for English. Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995.

[ZHA07] Qingfu Zhang and Hui Li. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. Evolutionary Computation, IEEE Transactions on, vol. 11, no. 6, pages 712–731, 2007.

[CHA09] Jonathan Chang, Jonathan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In Advances in Neural Information Processing Systems, Proceedings of the 23rd Annual Conference on, volume 31 of NIPS 2009, 2009.

[PHA08] Pham, N.K., Morin, A., Gros, P., Le, Q.T.: Factorial correspondence analysis for image retrieval. In: Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on. pp. 269–275. IEEE (2008)