LABORATOIRE INFORMATIQUE D'AVIGNON

ERIC
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

Le contexte sémantique et la dimension temporelle dans l'analyse de données complexes

Marian-Andrei RIZOIU

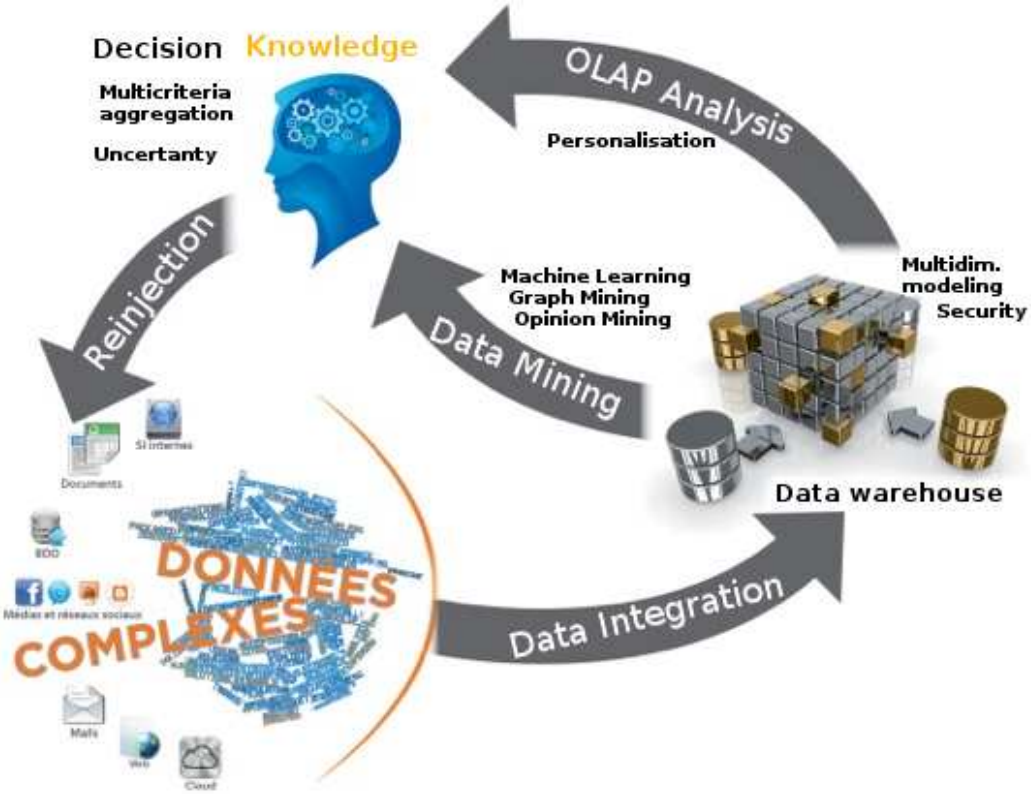
Laboratoire ERIC, Lyon

Avignon, France

13 Mars 2014

Contexte

Laboratoire ERIC



Université Lumière

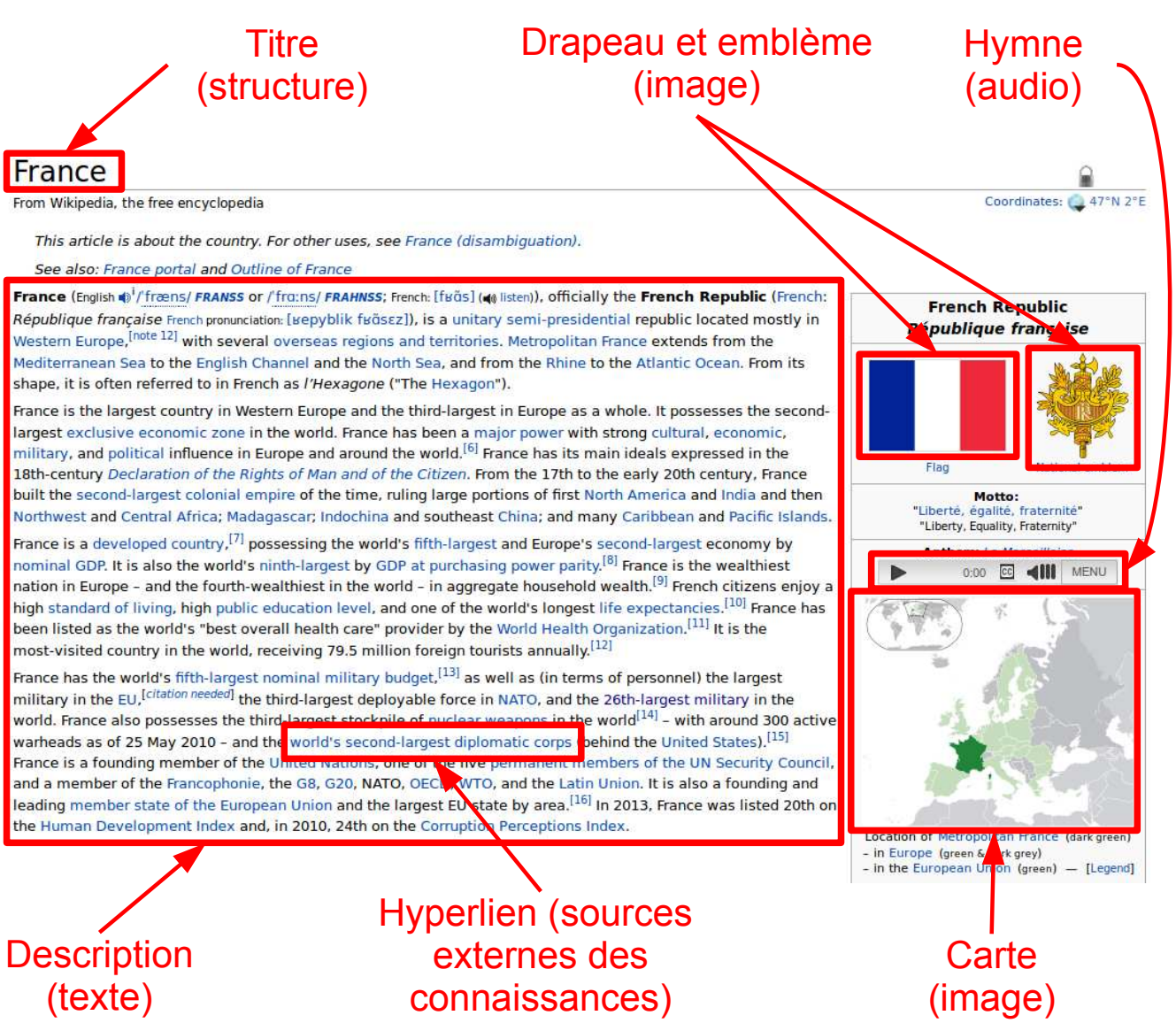
Sciences Humaines et Sociales

(Sociologie, Psychologie, Linguistique, Histoire, etc.)

Projets de recherche multidisciplinaire

- *Identifier des patrons à partir des textes mortuaires (historiens)*
- *Analyser des débats publiques : forums en ligne et média traditionnel (sciences sociales)*
- *Discoures sur la médecine nucléaire: évolution diachronique et diastratique (linguistes)*
- *Evolution de l'images des politiciens et des entreprises à travers le média social (sciences politiques)*
- *Détecter des rôles sociaux dans des réseaux sociaux enferré à partir des discussions sur des forum en ligne (Technicolor)*

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

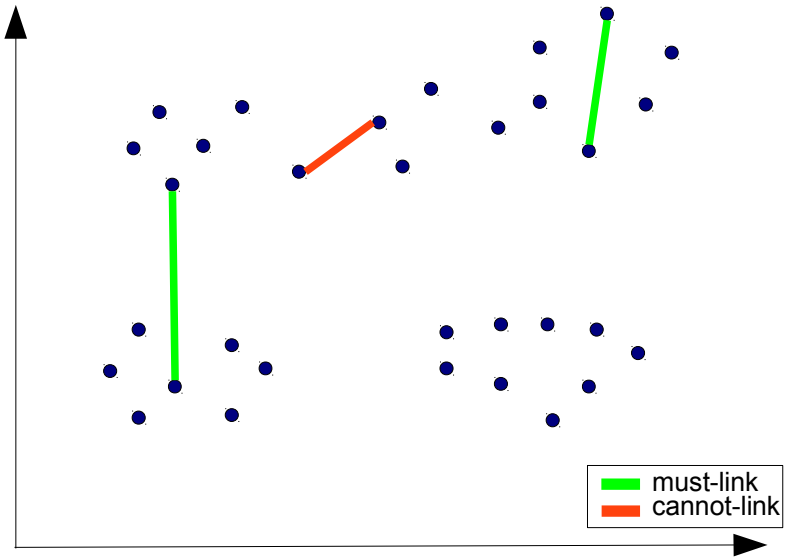
Notre approche

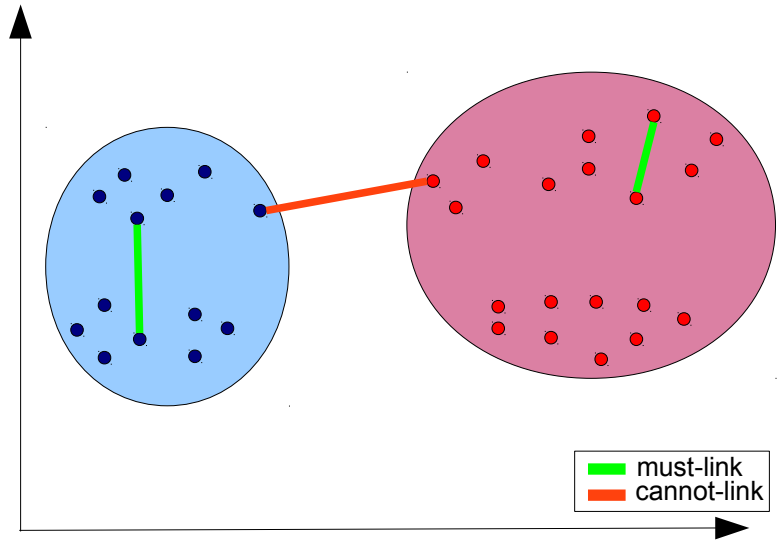
Objectif général : extraire des connaissances à partir de données complexes, souvent dans un contexte non-supervisé ;
(e.g., construction des profils utilisateurs, campagne marketing)

ajouter de la sémantique dans l'analyse de données ;
(e.g., améliorer l'interpretabilité des résultats)

utiliser les connaissances (supervision) disponibles.

Intégrer les connaissances : clustering semi-supervisé [WAG00]





Enjeux de recherche

Utiliser la sémantique pour analyser les données complexes.

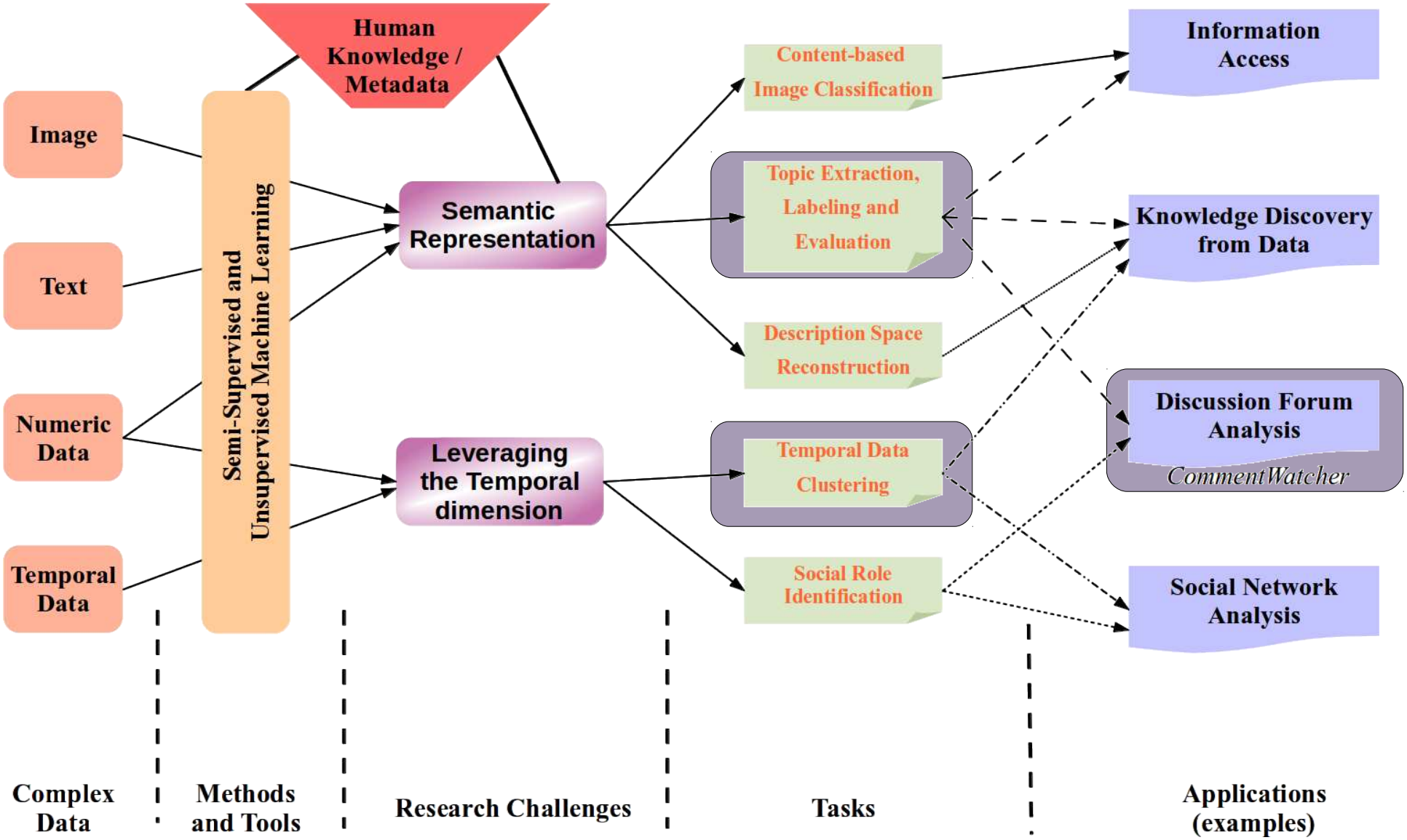
- ➜ plonger les données dans un espace de représentation capable de capturer la sémantique sous-jacente aux données
- ➜ injecter des connaissances externes dans les algorithmes d'apprentissage automatique

Prendre en compte la dimension temporelle des données complexes.

Tâches spécifiques :

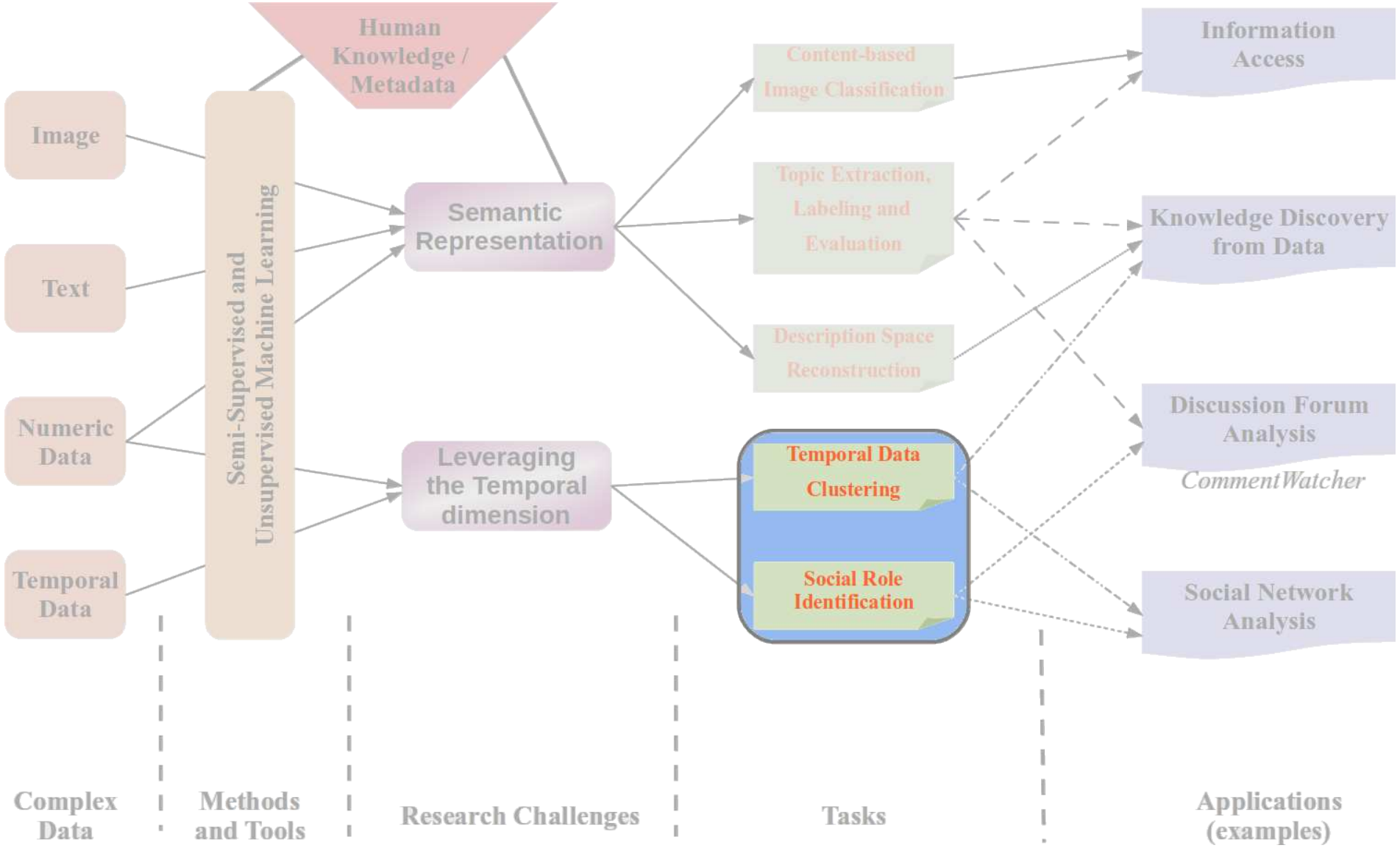
- ➜ *détection des évolutions typiques*
- ➜ *reconstruction sémantique de l'espace de représentation des données*
- ➜ *intégration des informations externes dans la description numérique des images.*

Schéma des travaux



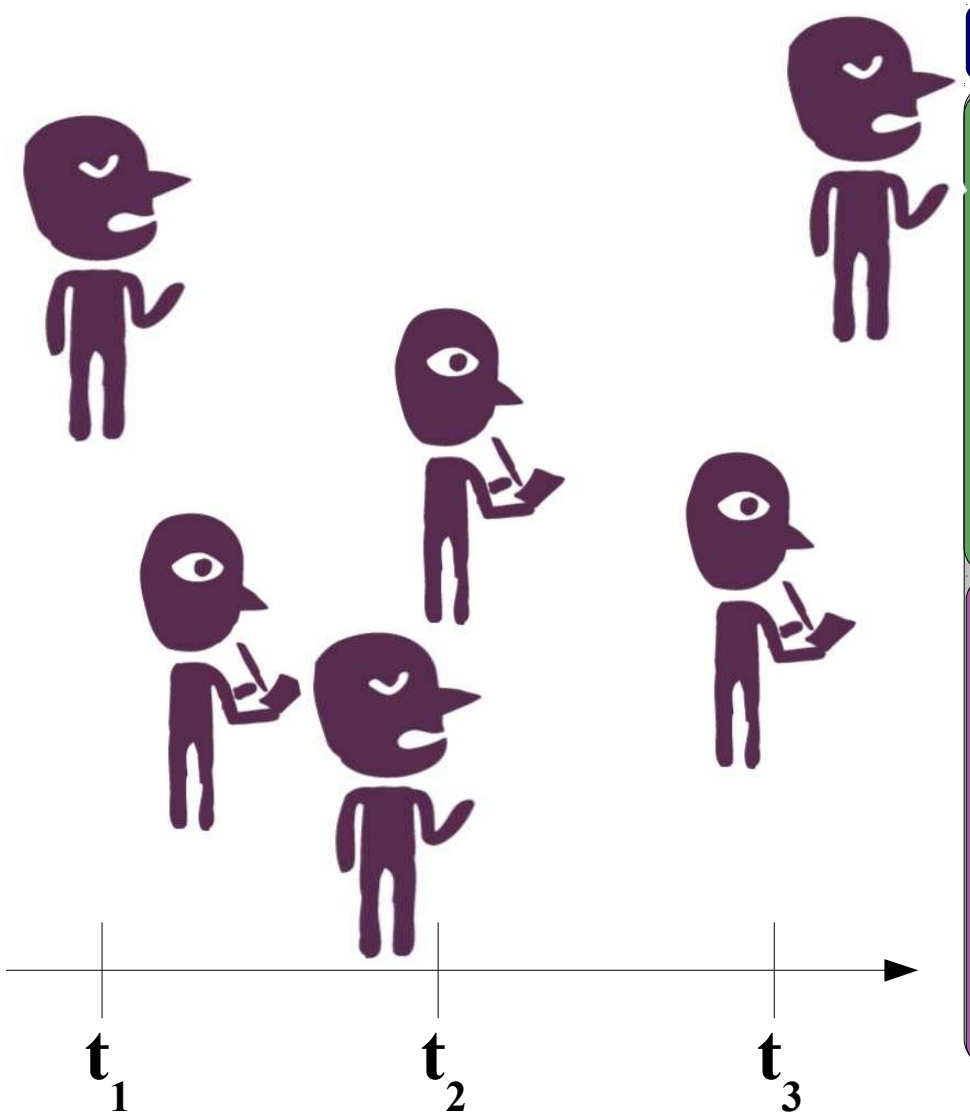
Détection des évolutions typiques :

Prendre en compte l'effet temporel



Jeux de données:

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

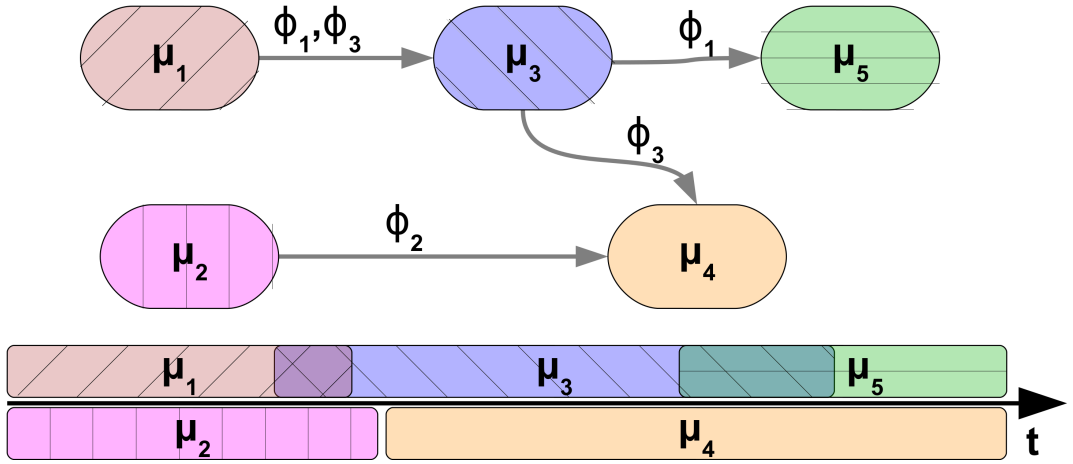


φ_1	t_1	x_1^d
	t_2	x_2^d
	t_3	x_3^d
φ_2	t_1	x_4^d
	t_2	x_5^d
	t_3	x_6^d

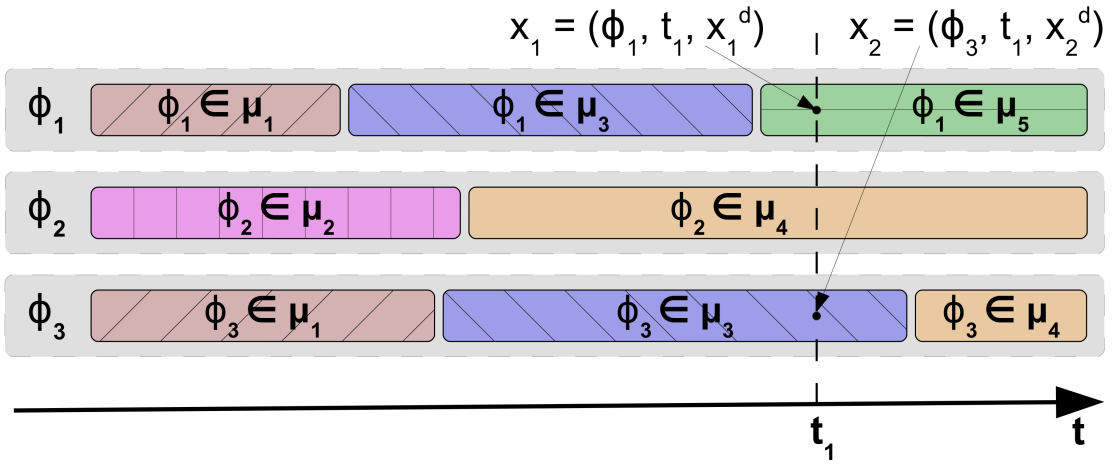
Enjeu de recherche : Utiliser la dimension temporelle dans l'analyse de données complexes

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 :

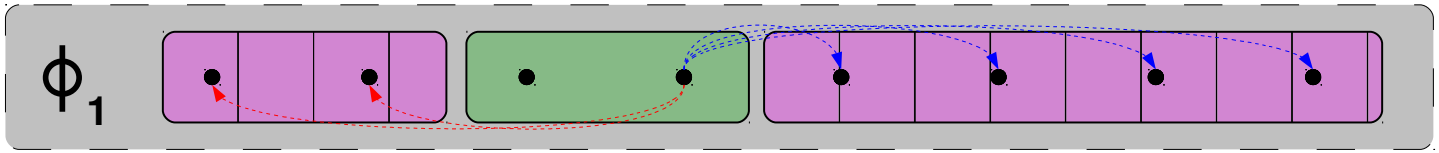
$$J = \sum_{\mu_j \in M} \sum_{x_i \in C_j} \left(\underbrace{\| x_i - \mu_j \|_{TA}}_{1} + \underbrace{\sum_{(x_k \notin C_j) \wedge (x_k^{\varphi} = x_i^{\varphi})} w(x_i, x_k)}_{2} \right)$$

Solution proposée (2) Les contributions

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

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

Fonction de pénalisation



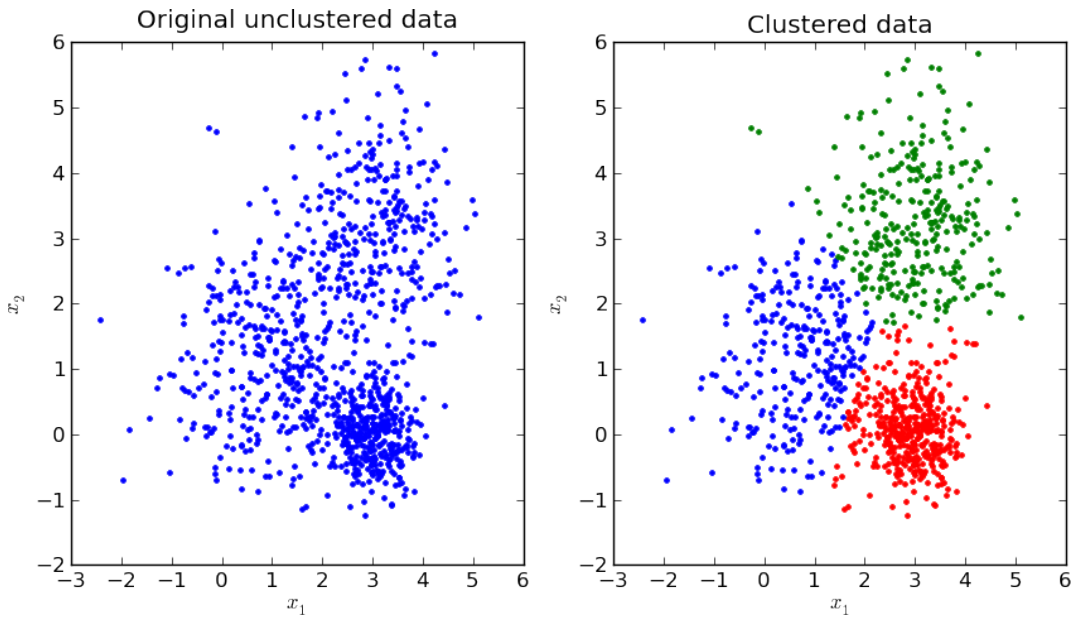
Solution proposée (2) Les contributions

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

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

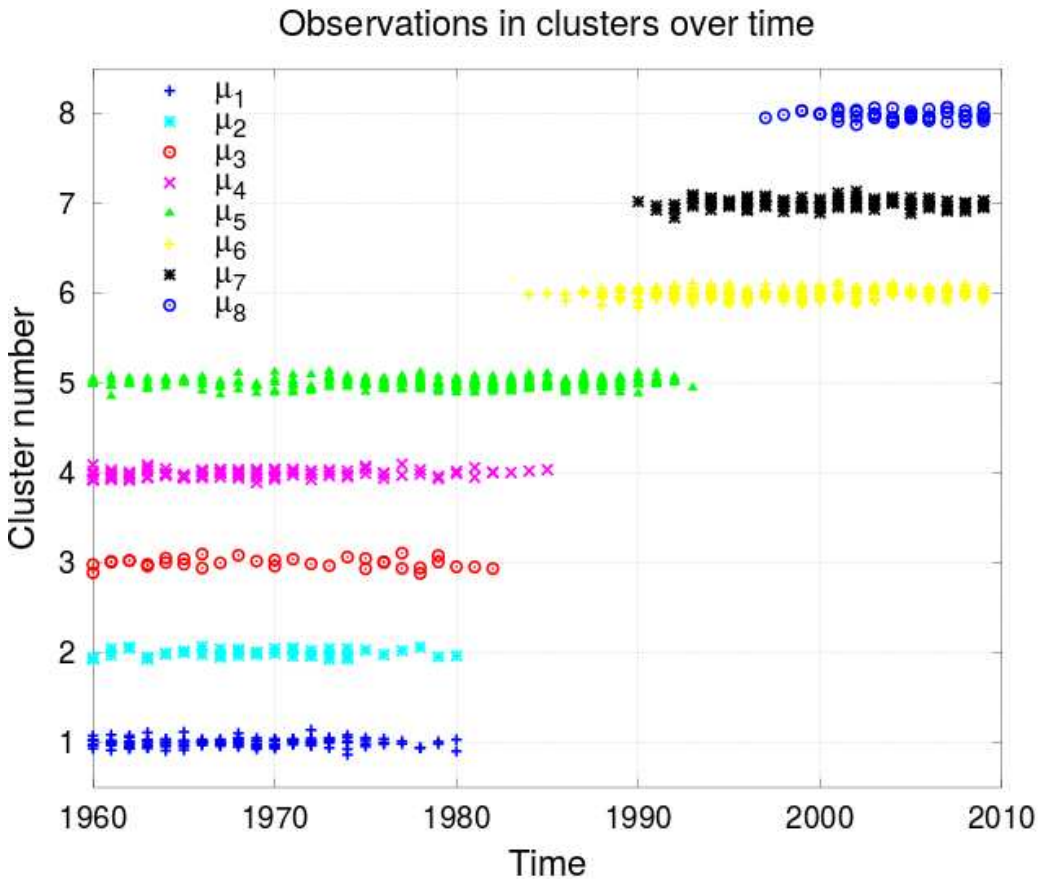
L'algorithme TDCK-Means

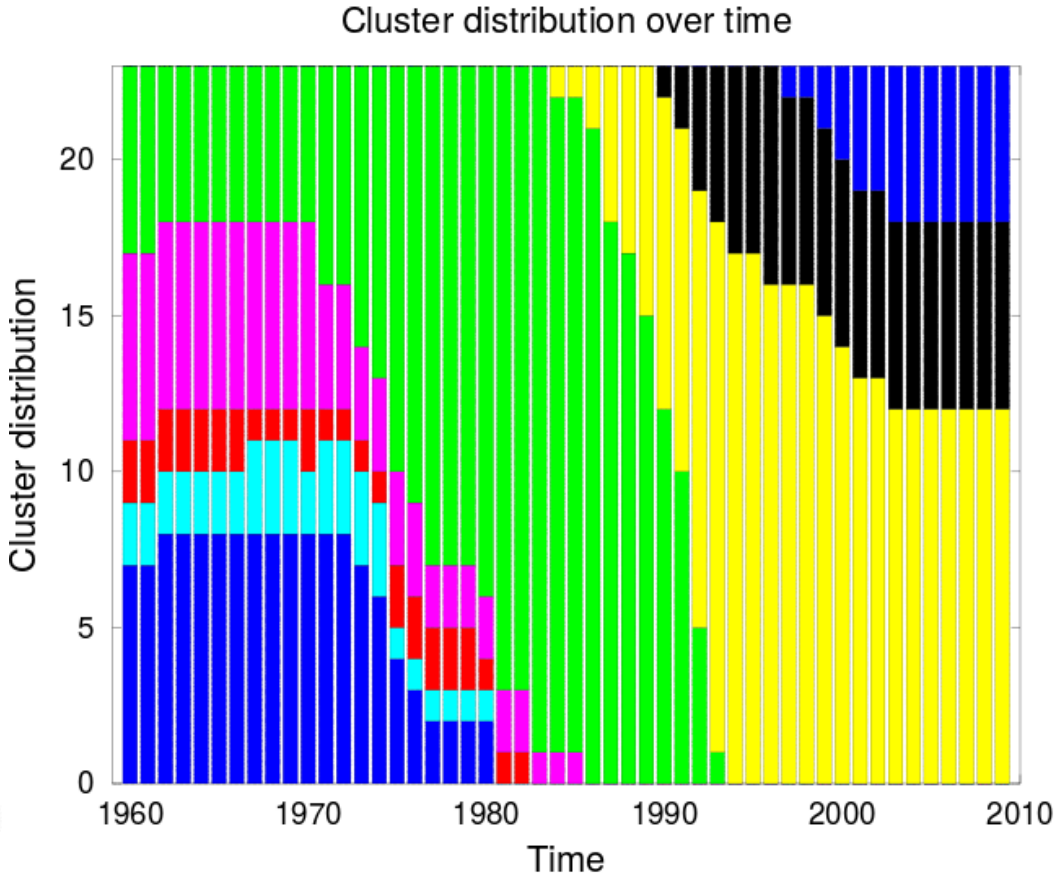
- Inspiré des K-Means.
- Mesure de *dissimilarité temporelle* et la *fonction de pénalisation*



Expérimentations et résultats

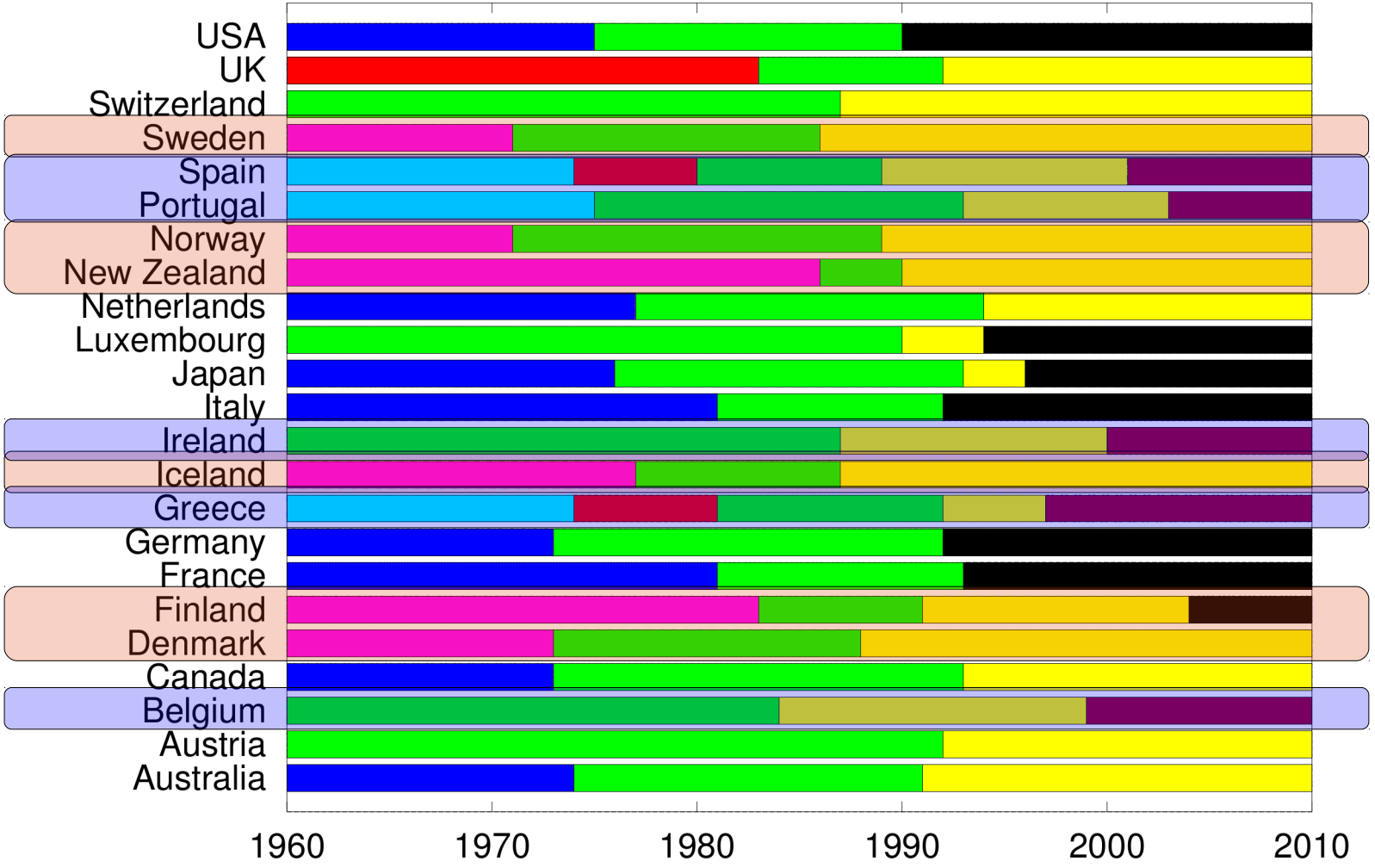
Compared Political Dataset I [ARM11] :
23 pays, 60 années, 207 variables politiques, démographiques, sociales et économiques.





Expérimentations et résultats

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



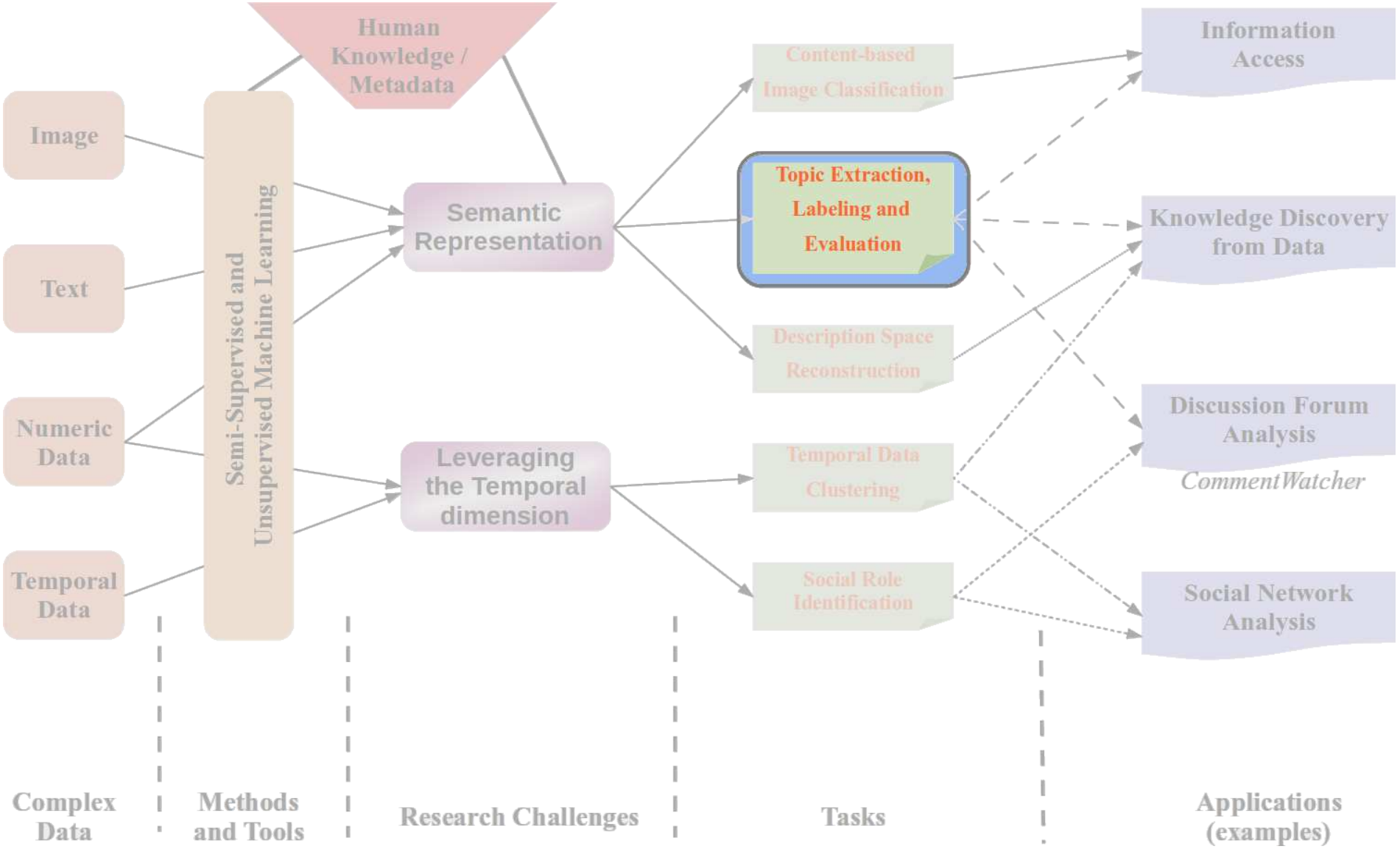
Expérimentations et résultats

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)

Analyse de données textuelles :

Extraction, étiquetage et évaluation des *thématiques*



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

Projets avec i) un startup, ii) des sociologues, iii) des linguistes

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]

Tâches d'apprentissage :

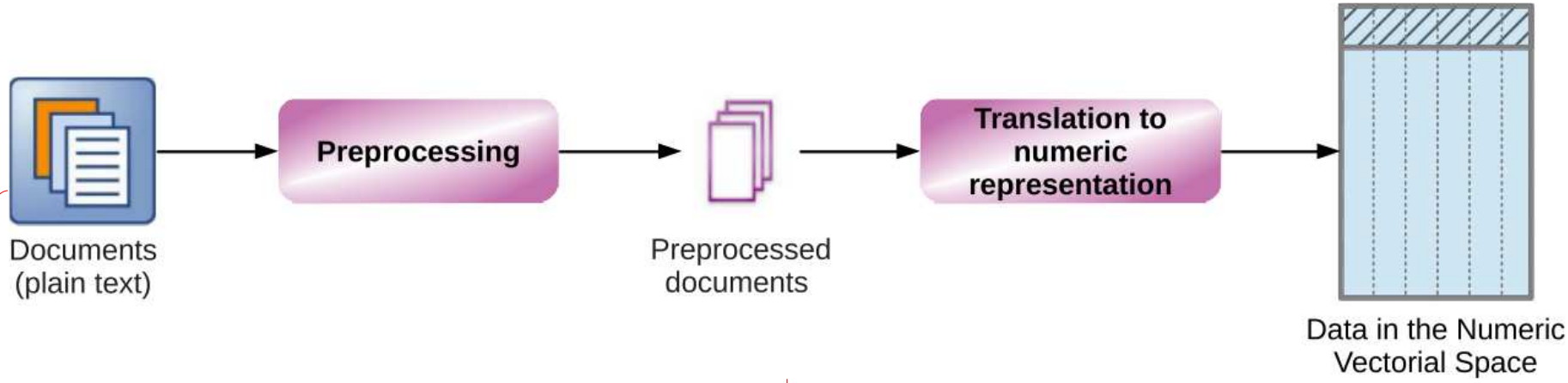
- extraction des thématiques
- étiquetage des thématiques avec des noms compréhensibles pour un être humain
- utilisation des connaissances sémantiques dans l'évaluation de ces thématiques.

Dimension appliquée :

- Demande forte de la part des scientifiques des **Sciences Humaines et Sociales** *(Sociologie, Psychologie, Linguistics, Histoire, etc.)*
- implémentation dans le logiciel **CommentWatcher** ;

Solution proposée (1) : Une alternative aux modèles graphiques (*e.g.*, LDA [BLE03]): le clustering textuel

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



enlever de « mots outils »
stemming
lemmatisation
prétraitement spécifique
(*e.g., Twitter)*

présence / absence
fréquence des termes
TFxIDF

Solution proposée (2) : Extraction et étiquetage des thématiques

I. Extraction des thématiques à l'aide du clustering recouvrant

OKM [CLE08] Une extension des KMeans qui autorise un document à appartenir à plusieurs clusters

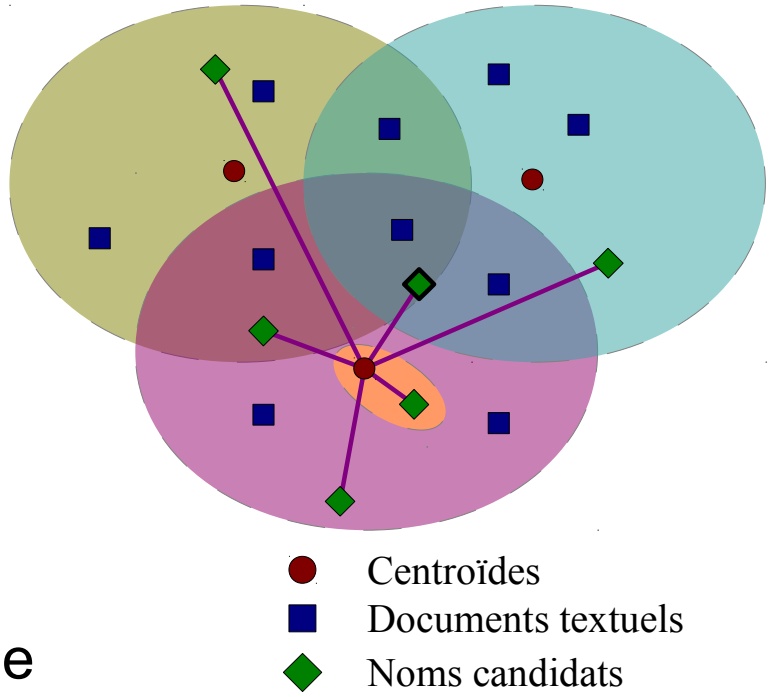
- ➜ construire une partition avec recouvrement
- ➜ les centroids sont des abstractions de leur clusters: *les thématiques*

II. Étiquetage des thématiques

- ➜ extraire des expressions complètes fréquentes à partir du texte original

  tableaux de suffixes [MAN93].

- ➜ injecter les expressions comme des pseudo-documents
- ➜ calculer la similarité et choisir le plus proche candidat



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

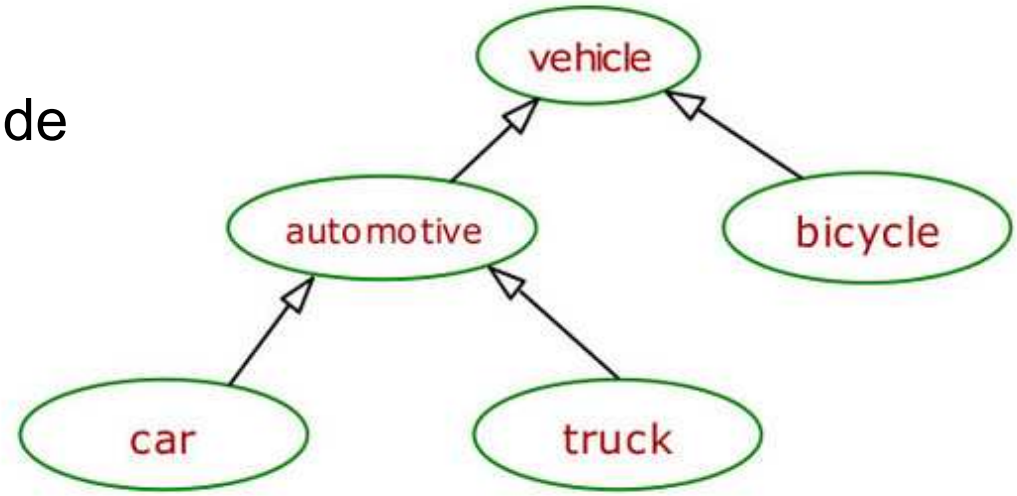
Hypothèse sous-jacente : Les mesures statistiques ne parviennent pas totalement à émuler le jugement humain [CHA09]

Idée : Relier une distribution statistique de fréquences à une structure sémantique

Utiliser les termes les plus pertinents attachés à une thématique

WordNet [MIL95]

- ressource linguistique, hiérarchie de concepts
- un concept regroupe ensemble un set de synonymes
- **polysemie**: un mot peut avoir plusieurs significations, dépendent du contexte



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

Alignement des thématiques :

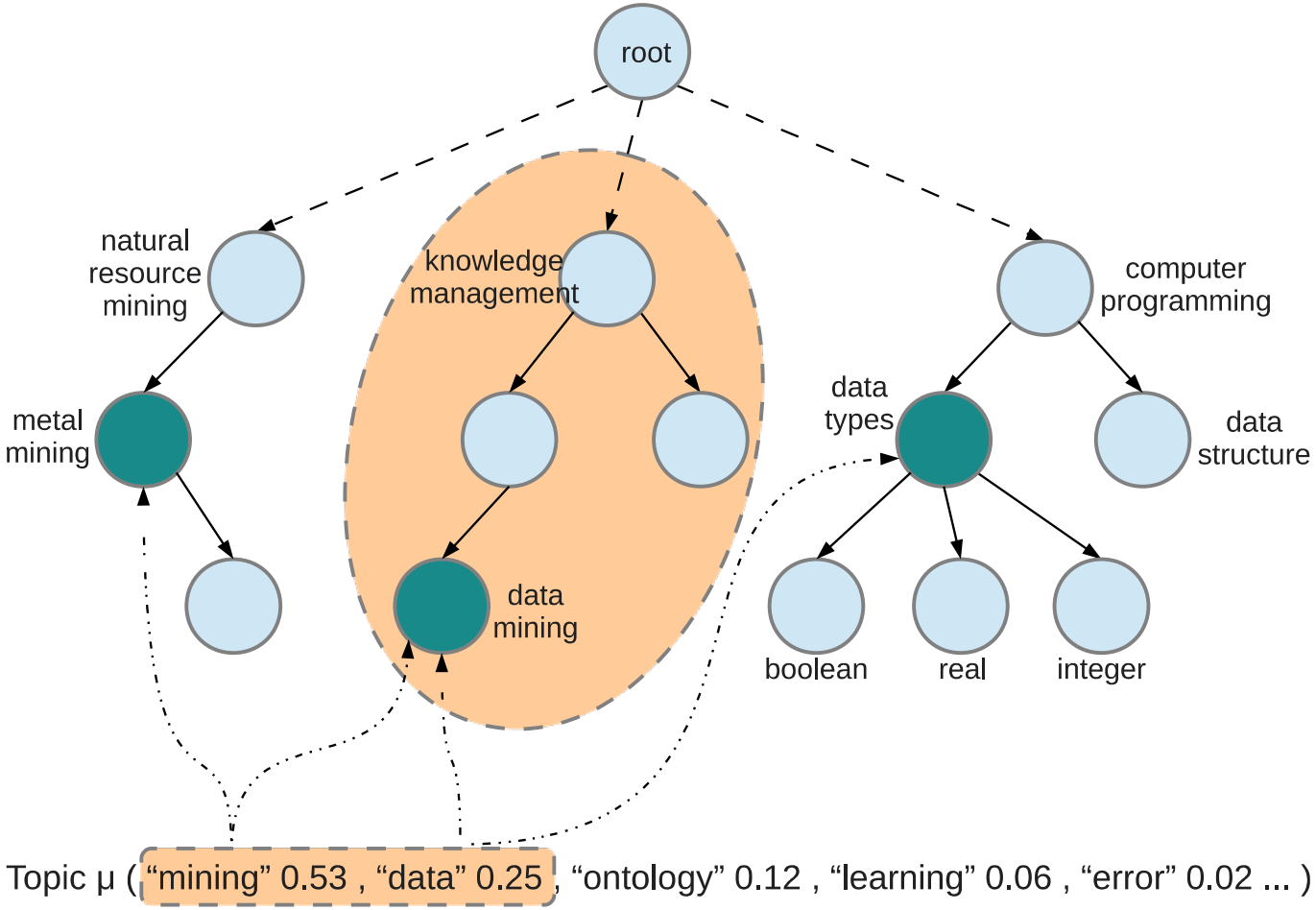
Déterminer le sous-arbre le plus spécifique qui contient au moins un sens pour chacun des mots les plus représentatifs de la thématique.

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 revêtent 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

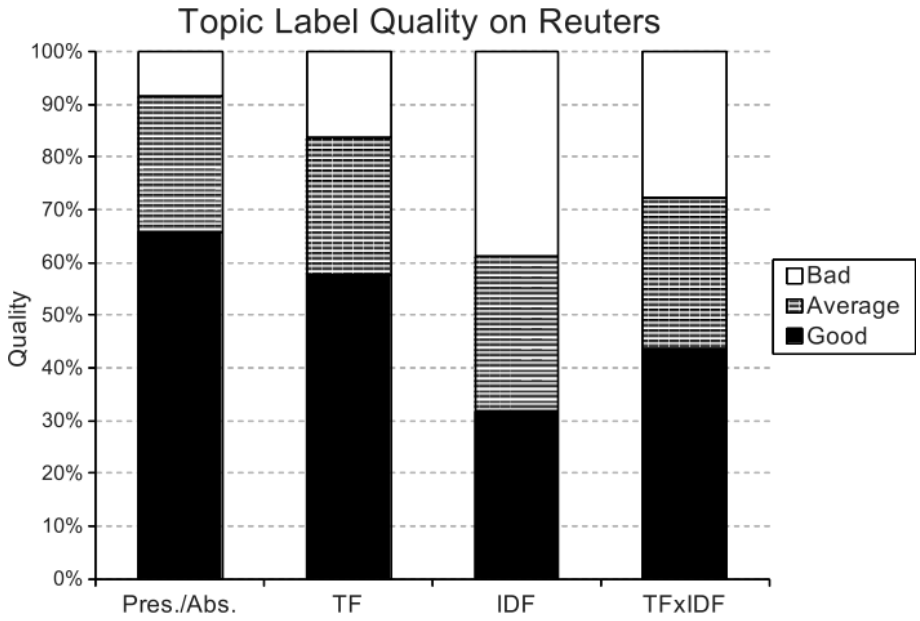
Reuters, **Suall11**

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]



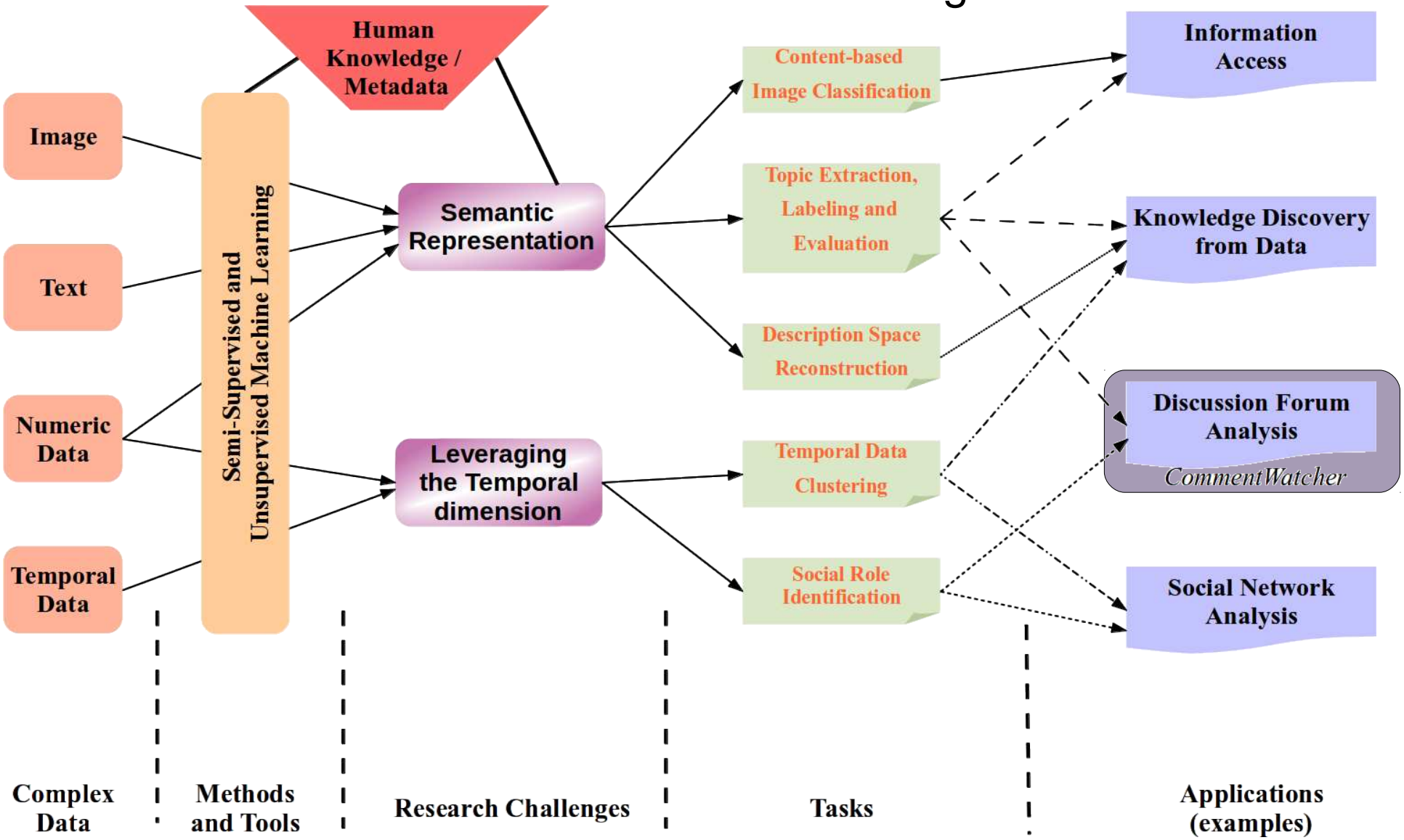
Évaluation de noms des thématiques

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

Dimension appliqué :

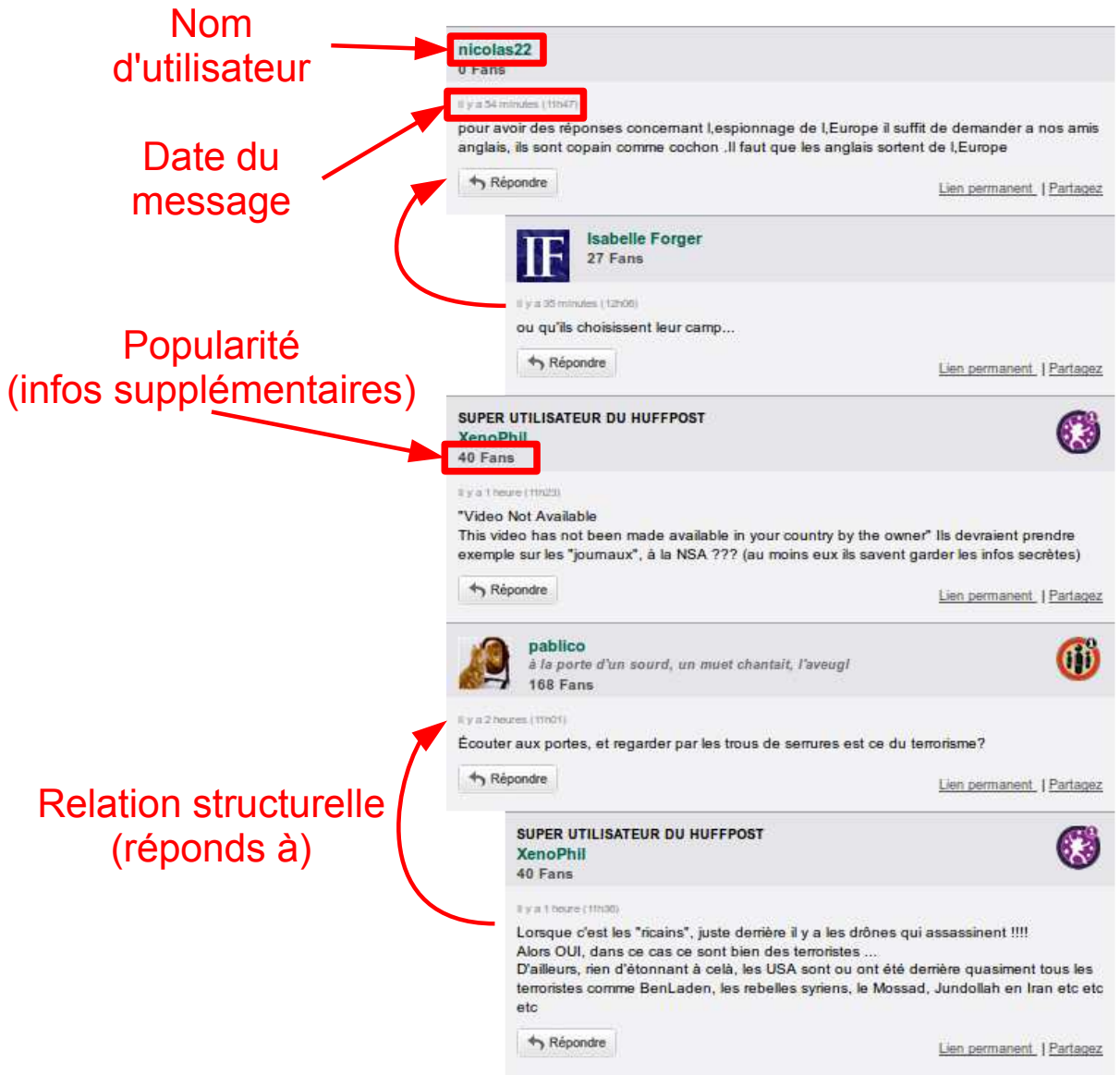
Logiciel d'analyse de discussions sur des forums en ligne



Le contexte du travail - les forums de discussion

Difficultés :

- La plupart des outils ne traitent pas l'aspect réseau social des données forum [AME12, GUI13]
- Manque de jeux de données issues de forums
- Structure des sites qui change constamment
- Problème de licence sur le contenu des forums



Objectif général :

deux types d'utilisateurs

l'analyste des forums : comprendre les discussions entre les utilisateurs et leurs thématiques

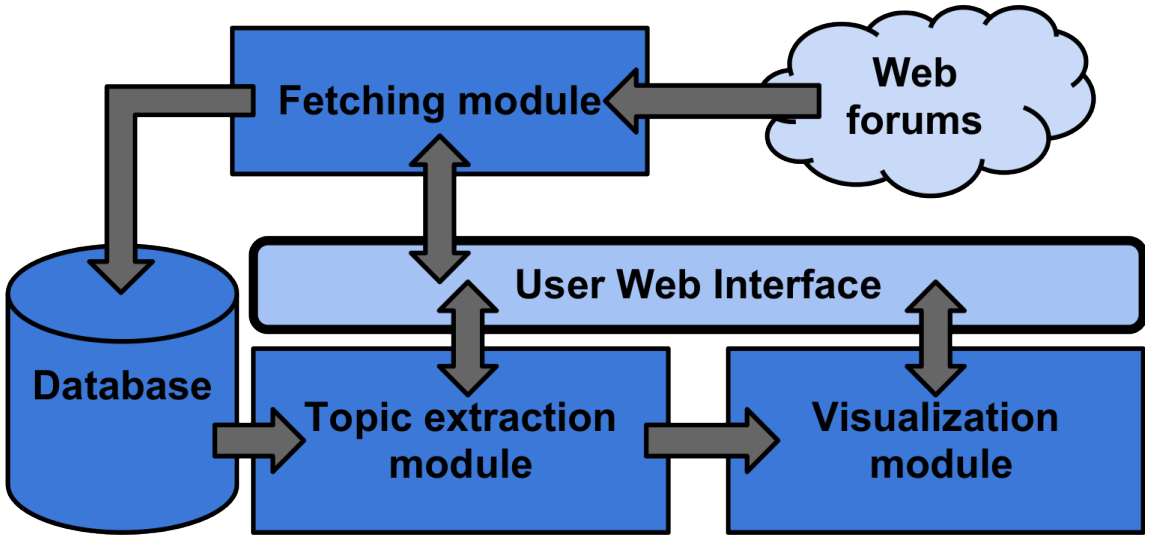
le chercheur : construire des jeux de données forums, analyser les évolutions des thématiques de discussion

Notre proposition : CommentWatcher

Plateforme Web opensource (GPLv3)

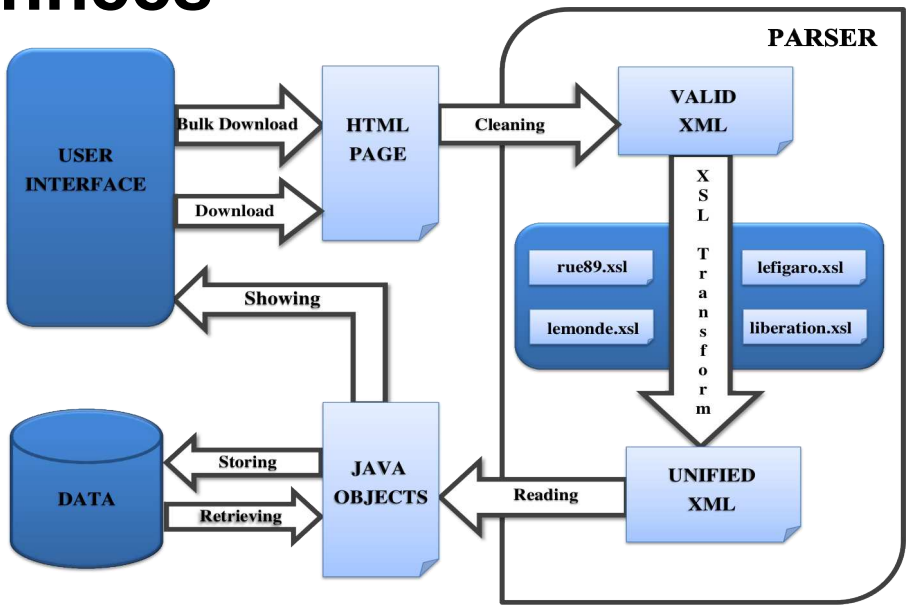
4 tâches :

- Récupération des données à partir d'Internet
- Extraction de thématiques
- Visualisation de thématiques comme un nuage d'expressions et l'évolution temporelle
- Visualisation du réseau social sous-jacent



Module I. Récupération des données

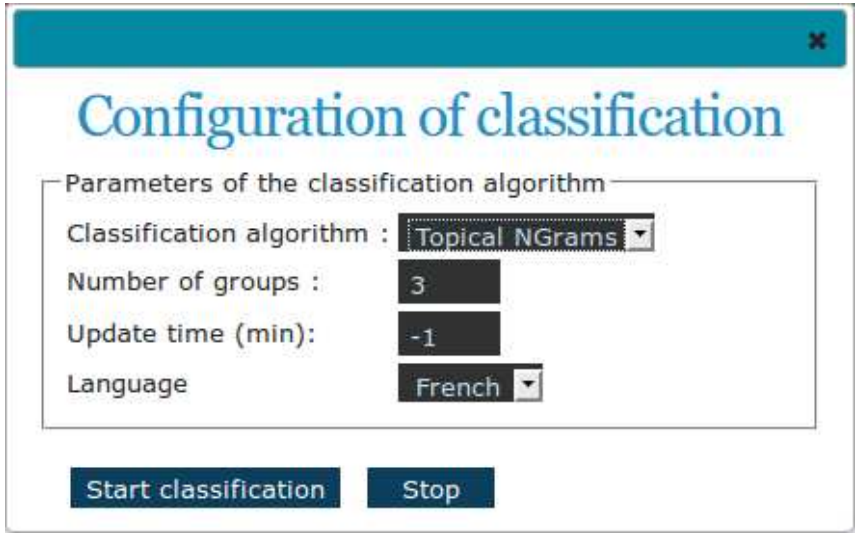
- Méta-parseur, indépendant de la structure des pages web
- Support pour de nouveaux sites via des fichiers de définition
- Recherche des forums supportés via une requête, en utilisant l'API Bing
- Téléchargement « en masse » des forums



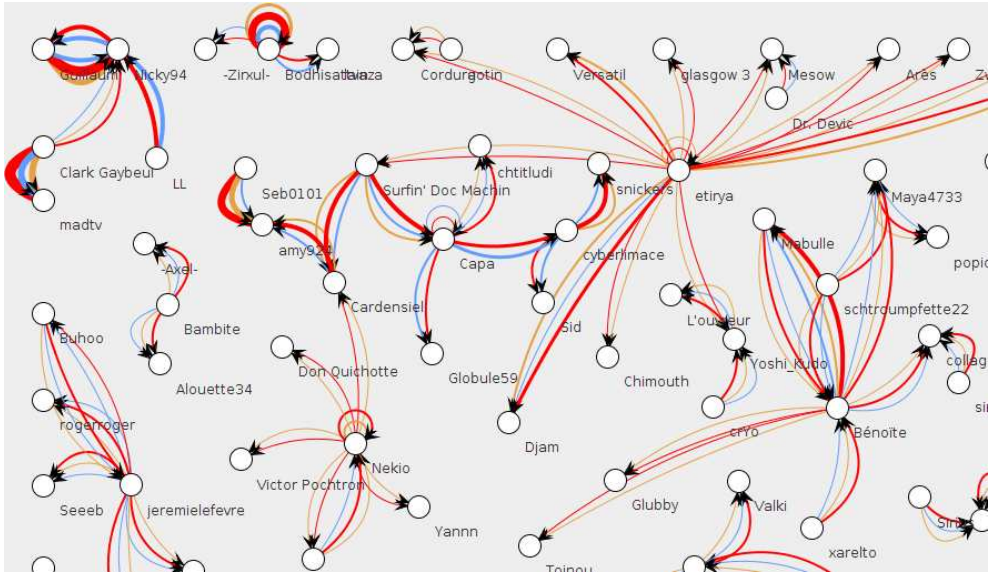
Module II. Extraction des thématiques

3 algorithmes supportés :

- Topical Ngrams (suite Mallet [MCC02])
- CKP [RIZ10]
- Dynamic Topic Models [BLE06] **(en développement)**



Module III. Visualiseurs



- Nuage d'expressions pour chaque thématique
- Évolution temporelle par forum et par site
- Évolution de la popularité d'une thématique

- Réseau social modélisé comme un multigraphe
- **Nœuds :** les utilisateurs ; **Arcs :** les messages associés à des thématiques
- Basé sur la relation de citation

Travaux en cours

Faire évoluer automatiquement un ensemble de cibles

Projet ImagiWeb

Objectif du projet : analyser les images (représentations) qui circulent sur Internet

Partenaires

- ami ENTERPRISE INTELLIGENCE SOFTWARE — éditeur de logiciel pour la veille
- CEPEL — sociologues spécialistes sciences politiques
- EDF — utilisateurs, étude sémiologique
- eric — datamining, apprentissage automatique
- LABORATOIRE INFORMATIQUE D'AVIGNON — fouille de textes et d'opinion, recherche d'information
- xerox Research Centre Europe — TAL, extraction d'information

Problématique d'apprentissage : Faire évoluer d'une façon automatique les cibles utilisées pour annoter

Motivation :

Actuellement, les textes (tweets, blogs) sont annotés en utilisant des cibles statiques, définies par des expertes.

Ex : L'ensemble des cibles politiques semble particulièrement adapté pour la période avant les élections

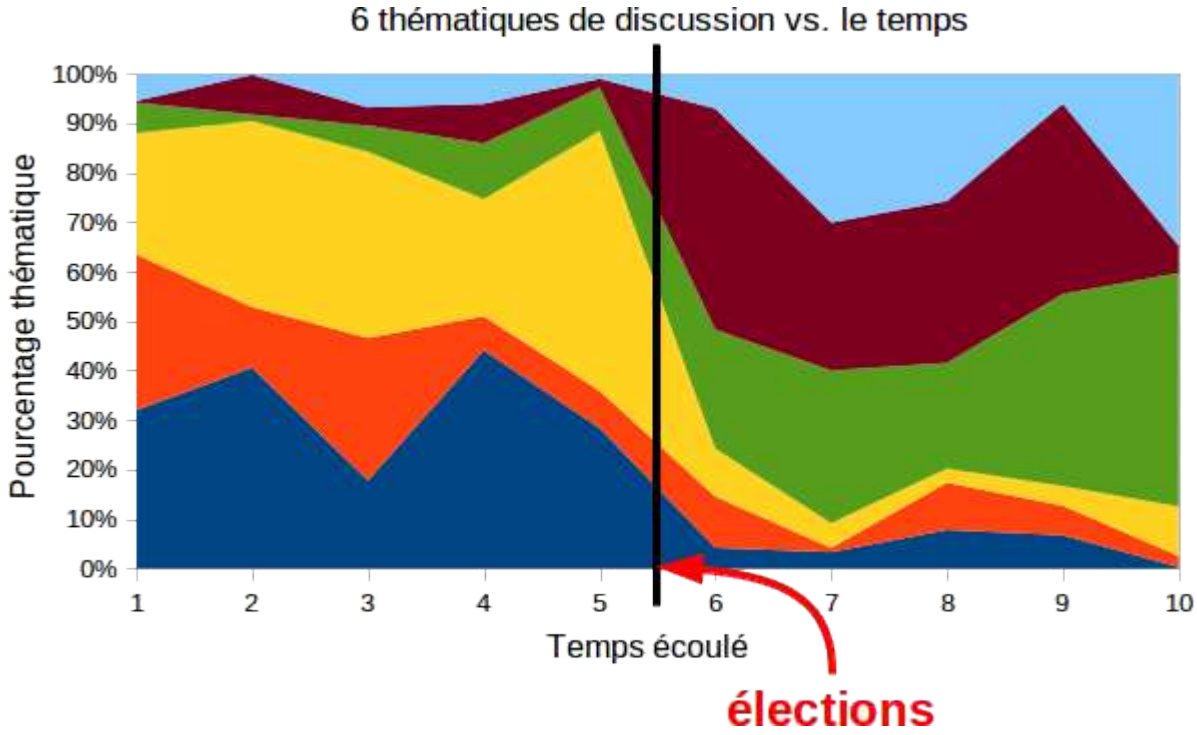
Cibles : Injonction / Soutien, Attribut / Sondage

Problématique d'apprentissage : Faire évoluer d'une façon automatique les cibles utilisées pour annoter

Hypothèse :

Avec l'écoulement de temps, les sujets de discussion entre internautes changent. Par conséquent, les cibles d'annotation doivent évoluer.

Nécessité de faire évoluer ou proposer des nouvelles cibles, d'un façon automatique



Problématique d'apprentissage : Faire évoluer d'une façon automatique les cibles utilisées pour annoter

Les données :

- Le texte de tweets et blogs, qui évolue temporellement

> • *Dernière ligne gauche! Pour virer la droite! Un seul vote @fhollande le seul qui ne promet pas ce qu'il ne pourra pas faire ! #FH2012 !*
> • *RT @Vanneur: Francois Hollande demande a Aqmi de liberer les otages avant qu il soit trop tard... Traduction vous les tuez, on vous tue. #mali*
> • *@reineroro

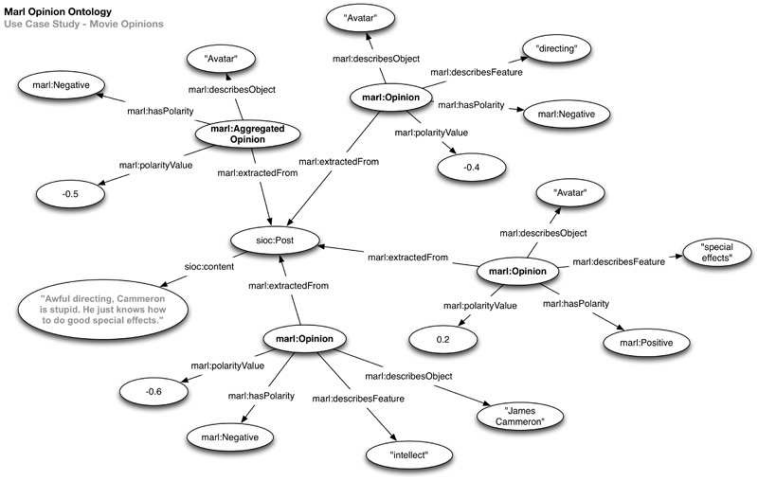
Le nouveau slogan avec @Hollande #larnaquecestmaintenant !

*
> • *Hollande dit « dégage » à Boris Boillon, le « Sarko boy » de Tunis http://t.co/2VJpTEAk*

- L'ensemble original de cibles, proposées par des expertes

> *Attribut / Sondage ; Attribut / Soutien/Non-soutien ; Bilan / Écologie ; Bilan / Économie ; Bilan / Societal ; ...*

- Un ensemble de concepts dans une base de connaissances



La base de connaissances :

Dbpedia – extraite automatiquement à partir de Wikipédia (en français)

dbpedia-owl:birthDate	1955-01-28 (xsd:date)
dbpedia-owl:birthName	Nicolas Paul Stéphane Sarközy de Nagy-Bocsa
dbpedia-owl:birthPlace	dbpedia-fr:17e_arrondissement_de_Paris dbpedia-fr:Paris
dbpedia-owl:child	dbpedia-fr:Jean_Sarkozy
dbpedia-owl:nationality	dbpedia-fr:Nationalité_française
dbpedia-owl:occupation	dbpedia-fr:Avocat_(métier) dbpedia-fr:Nicolas_Sarkozy__1 dbpedia-fr:Nicolas_Sarkozy__2 dbpedia-fr:Nicolas_Sarkozy__3
dbpedia-owl:party	dbpedia-fr:Rassemblement_pour_la_République dbpedia-fr:Union_pour_un_mouvement_populaire dbpedia-fr:Union_des_démocrates_pour_la_République
dbpedia-owl:religion	dbpedia-fr:Catholicisme
dbpedia-owl:spouse	dbpedia-fr:Cécilia_Attias dbpedia-fr:Carla_Bruni-Sarkozy dbpedia-fr:Marie-Dominique_Culioli
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/d/d7/Flickr_-_europeanpeoplesparty_-_EPP_Summit_October_2010
dbpedia-owl:thumbnailCaption	Nicolas Sarkozy, au sommet du Parti populaire européen, le .
dbpedia-owl:university	dbpedia-fr:Université_Paris_Ouest_Nanterre_La_Défense
dbpedia-owl:wikiPageExternalLink	http://infokiosques.net/imprimersans2.php3?id_article=295 http://www.404brain.info/NEWExpression/ExpressionEnginePB/images/uploads/Serge.Portelli.Ruptures.FRENCH.pdf http://www.editions-zones.fr/spip.php?id_article=116&page=lyberplayer http://www.assemblee-nationale.fr/12/tribun/fiches_id/2680.asp http://www.ina.fr/recherche/recherche?search=nicolas+sarkozy&vue=Video http://www.amisdenicolassarkozy.fr http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/le-conseil-constitutionnel/les-membres-du-conseil/
dbpedia-owl:wikiPageID	675918 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	95629613 (xsd:integer)
dbpedia-owl:wikiPageWikiLink	dbpedia-fr:Catégorie:Naissance_dans_le_département_de_la_Seine dbpedia-fr:Catégorie:Ancien_conseiller_régional_d'Île-de-France dbpedia-fr:Catégorie:Député_européen_élu_en_France_1999-2004 dbpedia-fr:Alain_Madelin dbpedia-fr:Catégorie:Ministre_français_des_Finances dbpedia-fr:Catégorie:Personnalité_de_l'Union_pour_un_mouvement_populaire dbpedia-fr:Catégorie:Député_de_la_IXe_législature_de_la_Ve_République dbpedia-fr:Catégorie:Député_de_la_Xe_législature_de_la_Ve_République

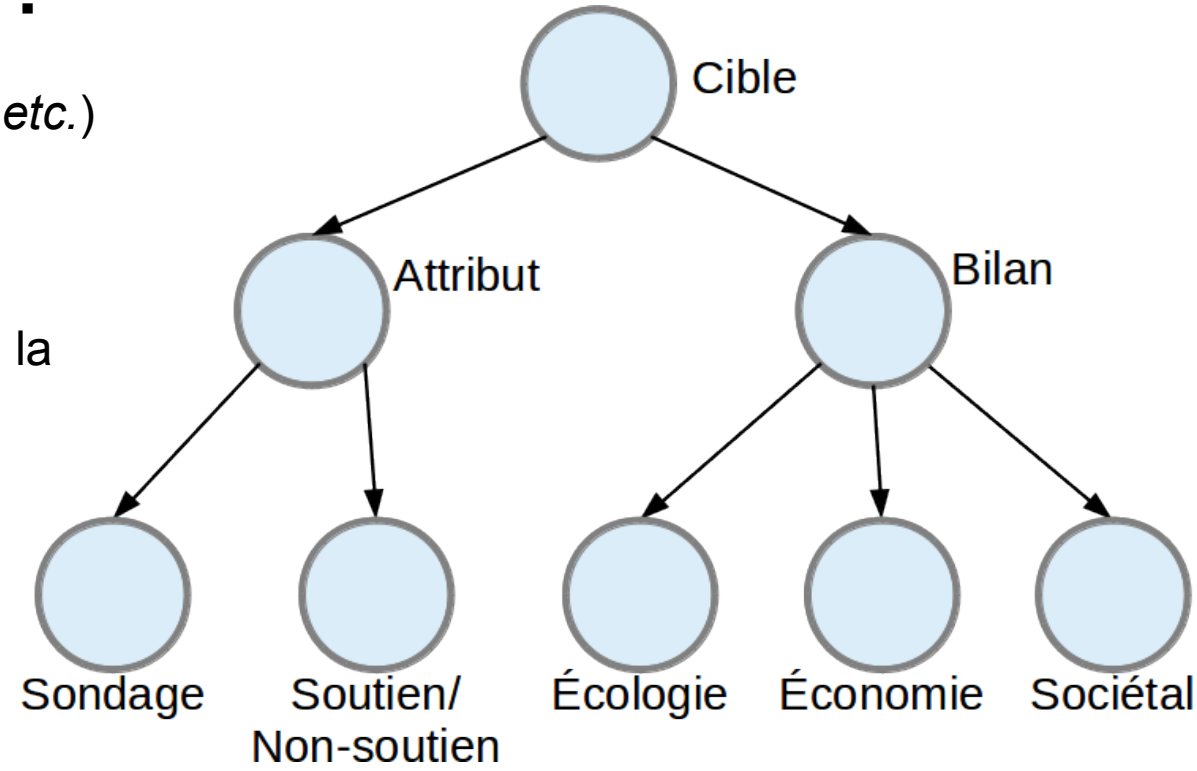
*L'article **Wikipédia** sur Nicolas Sarkozy*

*L'entité Nicolas Sarkozy dans **Dbpedia***

La base de connaissances : **Dbpedia** – extraite automatiquement à partir de Wikipédia (en français)

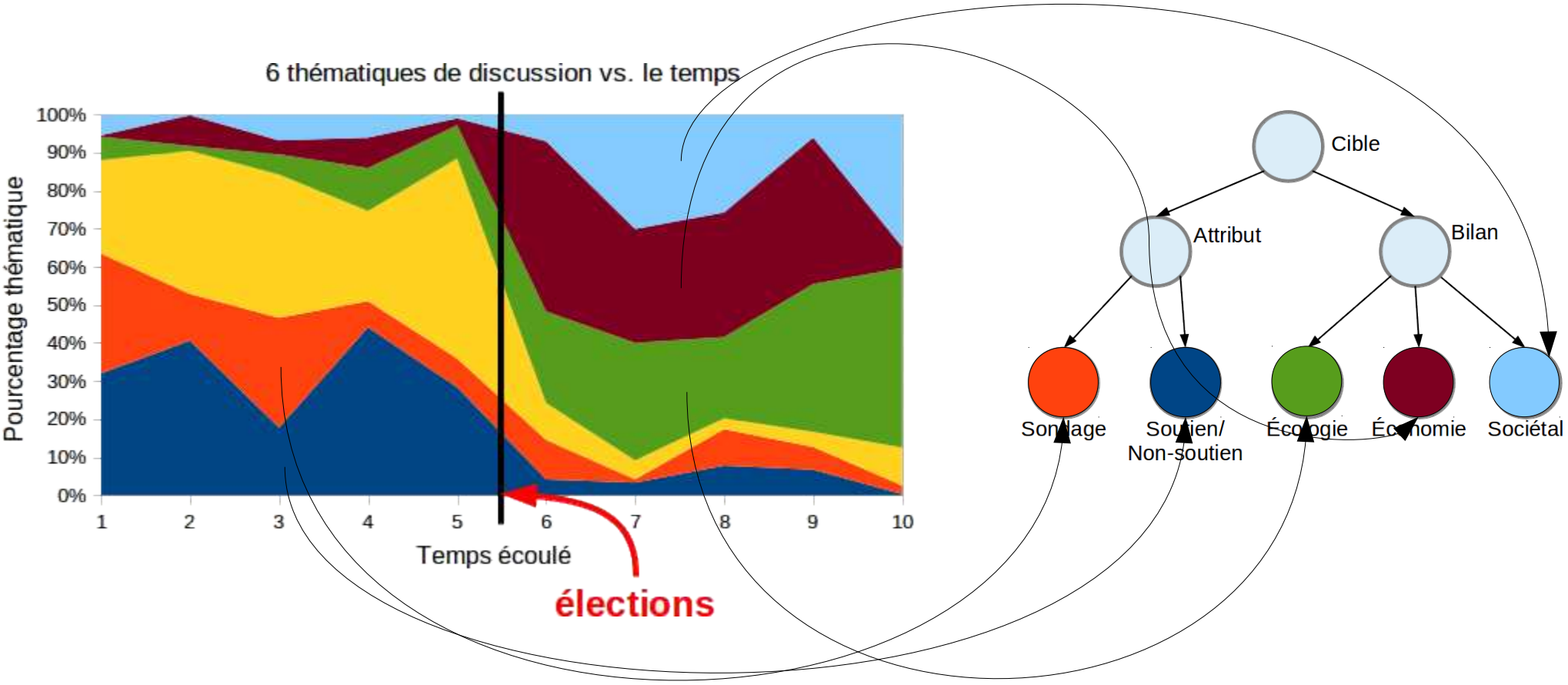
Des connaissances sur :

- Des lieux (villes, pays, continentes *etc.*)
- Des événements (*ex. les guerres mondiales, des attentats*)
- Des concepts (*ex.* le bien et le mal, la philosophie *etc.*)
- ...



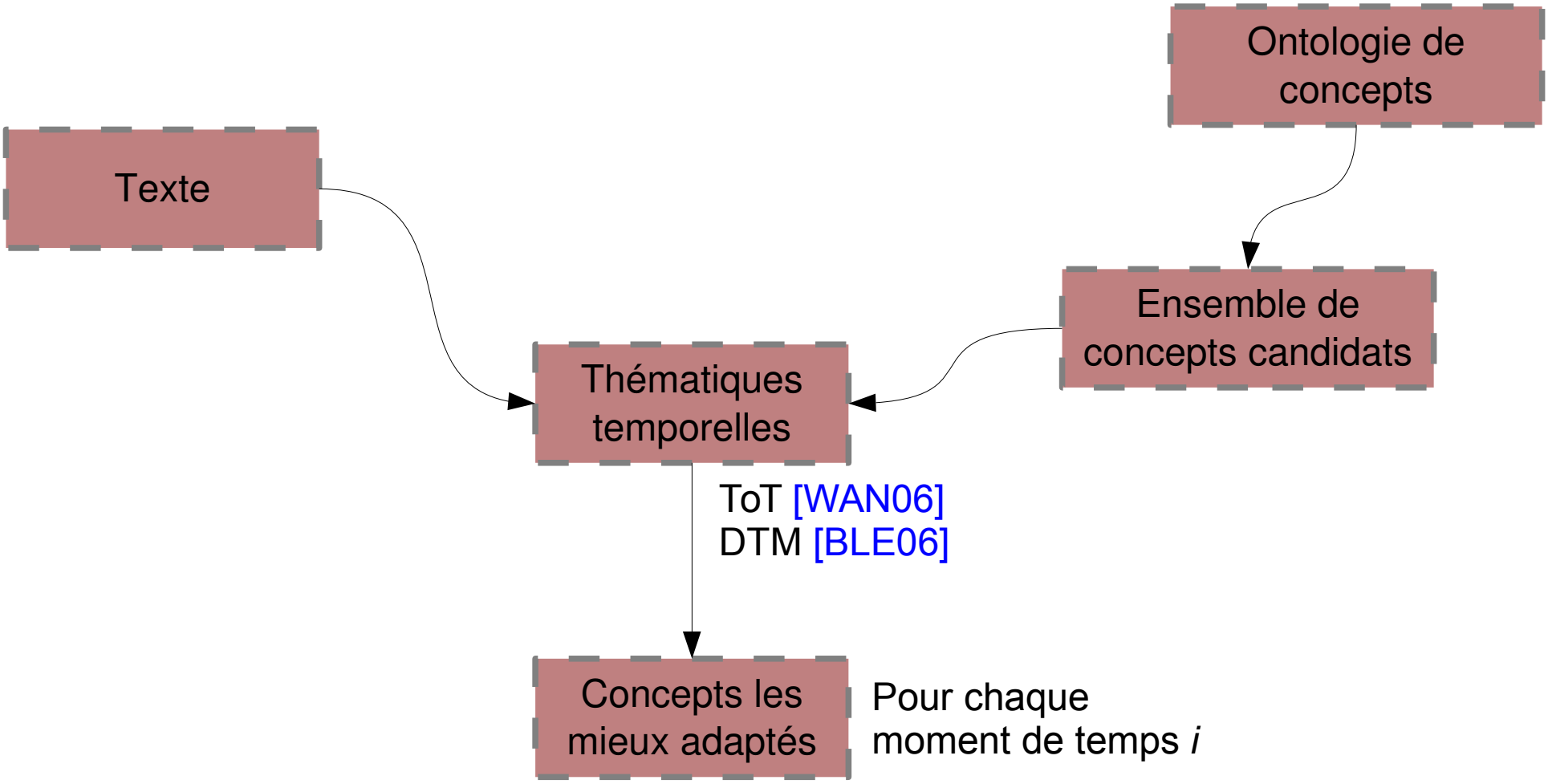
Peuvent être organisées comme une hiérarchie

L'idée : Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.



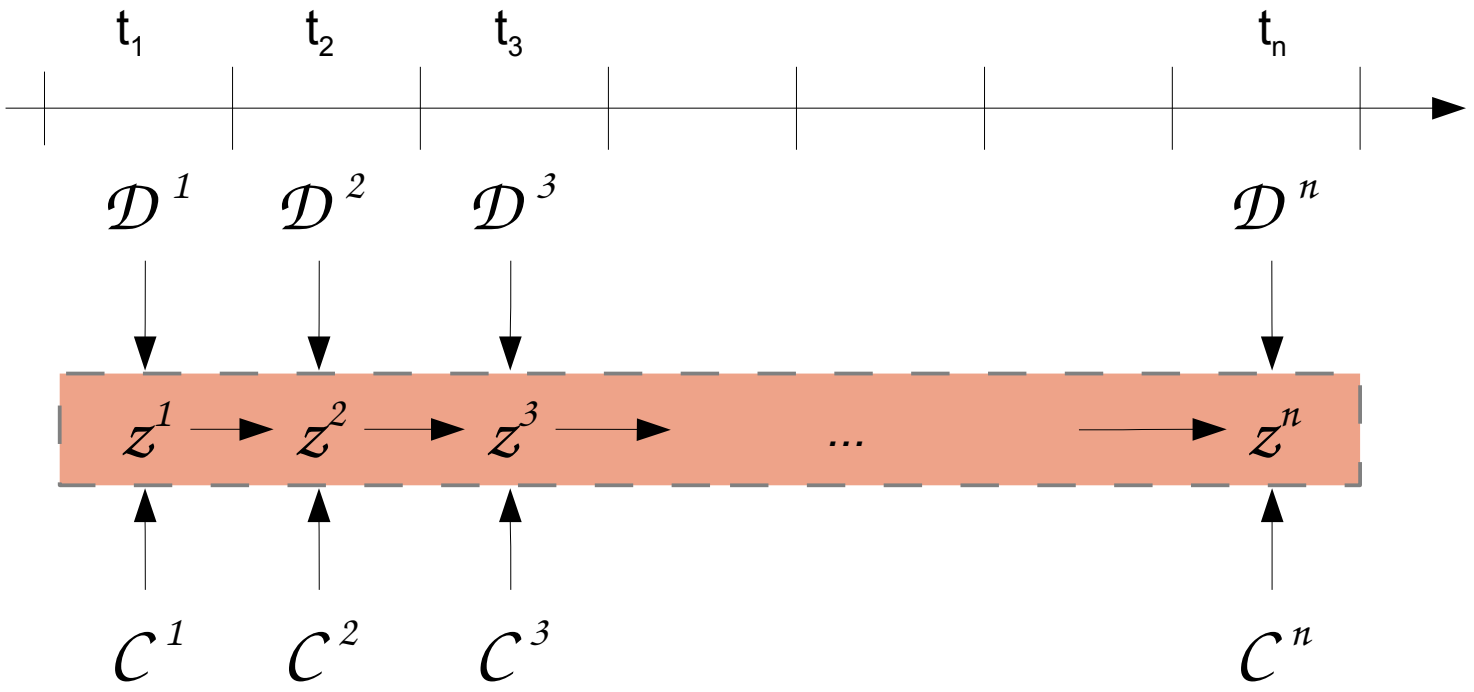
L'idée : Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Faire le lien entre le texte et les concepts à travers l'extraction temporelle de thématiques :



L'idée : Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Faire le lien entre le texte et les concepts à travers l'extraction temporelle de thématiques :



Temporal axis

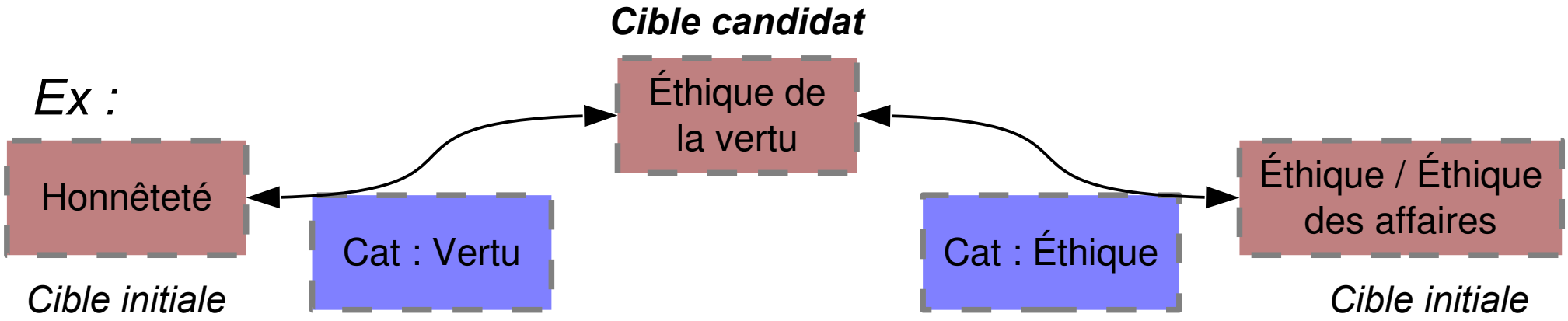
Sets of documents in each timeslice

Topical representation spaces

Sets of best fitted concepts

L'idée : Faire évoluer automatiquement les cibles en utilisant les concepts de **Dbpedia**.

Peupler l'ensemble de concepts candidats – liaison sémantique basée sur les connexions dans **Dbpedia** et la similarité textuelle



Conclusions

- exemples de travaux en liaison avec la **dimension temporelle** de données complexes (*e.g.*, détection d'évolutions typiques) et l'intégration des connaissances sémantiques dans la représentation des données
- exemples de travaux qui adressent en partie le **caractère multimodale** de données (image et texte) issue du Web (*e.g.*, réseaux sociaux en ligne, microblogging, articles de presse)
- des informations externes, souvent en liaison avec le **Web Sémantique**, peuvent être utilisées pour amélioré l'analyse de données et l'extraction des connaissances (*e.g.*, construction des profiles utilisateur)
- des autres travaux en cours sur des **problématiques de vie privé** : la perte inhérente d'intimité en ligne et l'inférence de traits privés à partir des comportements publiques.

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,

[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)