Modélisation unifiée du document et de son domaine pour une indexation par termes-clés libre et contrôlée

Adrien Bougouin Florian Boudin Beatrice Daille

Laboratoire LINA,
Université de Nantes

JEP-TALN-RECITAL 2016, Paris, France

Introduction I

- **Termes clés** : mots ou expressions polylexicales qui caractérisent le contenu jugé le plus important d'un document
- La tâche d'indexation par termes-clés consiste à identifier automatiquement les termes clés des documents
  - Indexation libre : fournit des termes-clés apparaissant dans le document
  - Indexation contrôlée : fournit des termes-clés apparaissant dans un vocabulaire contrôlée
- La plupart des travaux existant portent sur l'indexation libre
  - Pourtant, l'indexation réalisée par des indexeurs professionnels inclut aussi bien des termes-clés libres que contrôlés !

Introduction II

- L'objectif de ce travail est de proposer une méthode d'indexation par termes-clés (hybride) libre et contrôlée
  - Extension d'une méthode d'indexation libre (TopicRank, TAL 55-1)

Plan

Approche proposée I

TopicRank

1. Sélection des termes-clés candidats : `/(N|A)+/`
2. Groupement des candidats en sujets
3. Création d'un graphe de sujets
4. Ordonnancement des sujets dans le graphe
5. Extraction du "meilleur" candidat de chaque sujet

- Avantages de la méthode TopicRank :
  - Évite la redondance dans les termes-clés extraits
  - Capture avec précision les relations entre les sujets
  - Méthode non supervisée au niveau de l'état-de-l'art

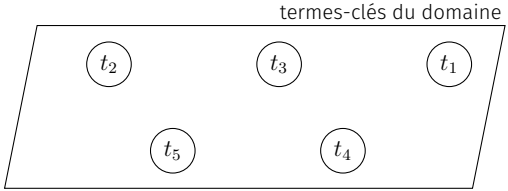
Approche proposée II

- Comment étendre TopicRank pour permettre l'indexation contrôlée ?
- Modification des étapes de construction du graphe, d'ordonnancement et de sélection des termes clés
  - Unifier le graphe de sujets à un graphe des termes-clés du domaine (vocabulaire contrôlé)
  - Ordonner conjointement sujets et termes-clés du domaine
- Nouvelle méthode : TopicRankSpe

TopicRankSpe I

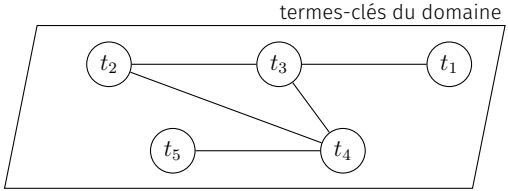
1. Création du graphe du domaine :

TopicRankSpe I



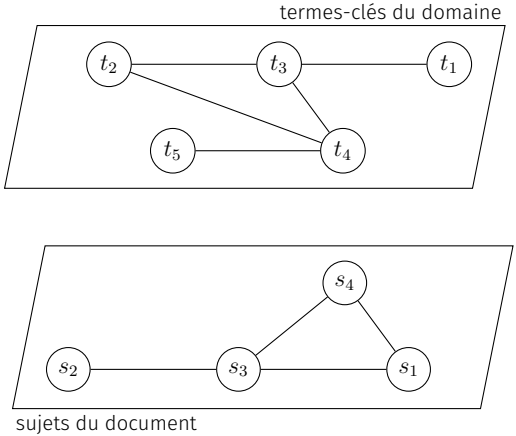
1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé

TopicRankSpe I



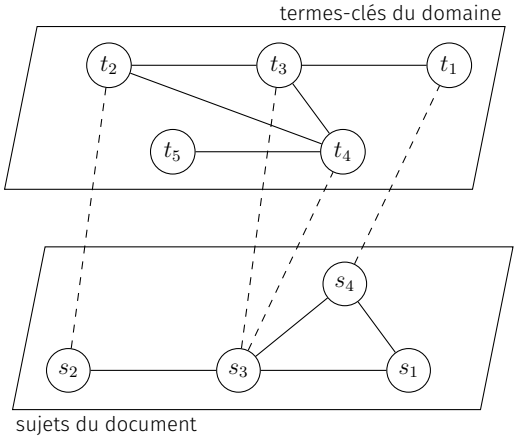
1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés

TopicRankSpe I



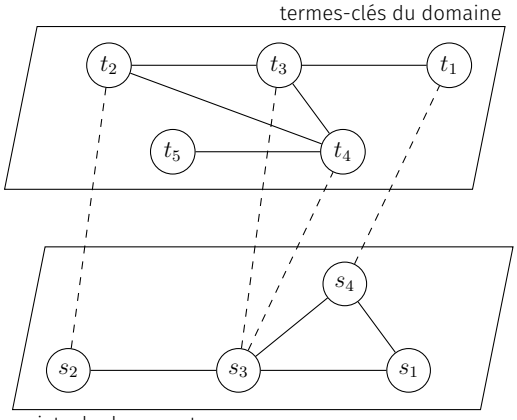
1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :

TopicRankSpe I



1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :
   - $t_i \subseteq s_j$
     $\Rightarrow$ lien domaine/document

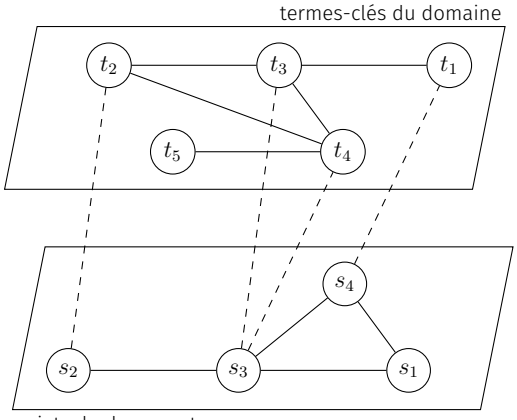
TopicRankSpe I



$$G = (N, A)$$
$$N = \{s_1..s_n\} \cup \{t_1..t_m\},\ A \subseteq N \times N$$

1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :
   - $t_i \subseteq s_j$
     $\Rightarrow$ lien domaine/document

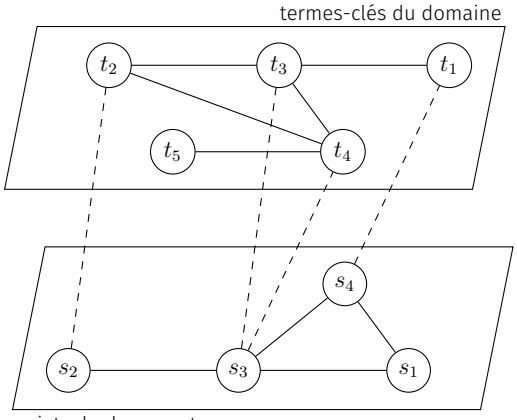
TopicRankSpe I



$$G = (N, A)$$
$$N = \{s_1..s_n\} \cup \{t_1..t_m\},\ A \subseteq N \times N$$

1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :
   - $t_i \subseteq s_j$
     $\Rightarrow$ lien domaine/document
3. Ordonnancement conjoint :

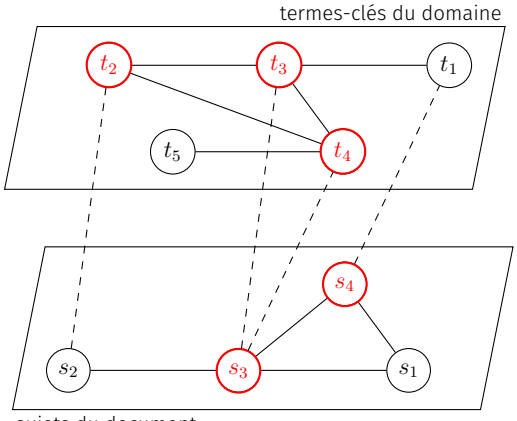
TopicRankSpe I



$$G = (N, A)$$
$$N = \{s_1..s_n\} \cup \{t_1..t_m\},\ A \subseteq N \times N$$

1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :
   - $t_i \subseteq s_j$
     $\Rightarrow$ lien domaine/document
3. Ordonnancement conjoint :
   - $S(n_i) = (1 - \lambda)\ R_{\text{ext}}(n_i) + \lambda\ R_{\text{int}}(n_i)$

TopicRankSpe I



$$G = (N, A)$$
$$N = \{s_1..s_n\} \cup \{t_1..t_m\},\ A \subseteq N \times N$$

1. Création du graphe du domaine :
   - termes-clés d'entraînement
     $\Rightarrow$ vocabulaire contrôlé
   - termes-clés assignés ensemble $\Rightarrow$ sémantiquement liés
2. Unification au graphe de sujets :
   - $t_i \subseteq s_j$
     $\Rightarrow$ lien domaine/document
3. Ordonnancement conjoint :
   - $S(n_i) = (1 - \lambda)\ R_{\text{ext}}(n_i) + \lambda\ R_{\text{int}}(n_i)$

TopicRankSpe II

- Avantages de la méthode TopicRankSpe :
  - Pas besoin de vocabulaire contrôlé ; les termes-clés déjà assignés sont utilisés pour construire le graphe du domaine
- Indexations libre et contrôlée réalisées conjointement (se renforçant mutuellement)

Plan

Paramètres expérimentaux I

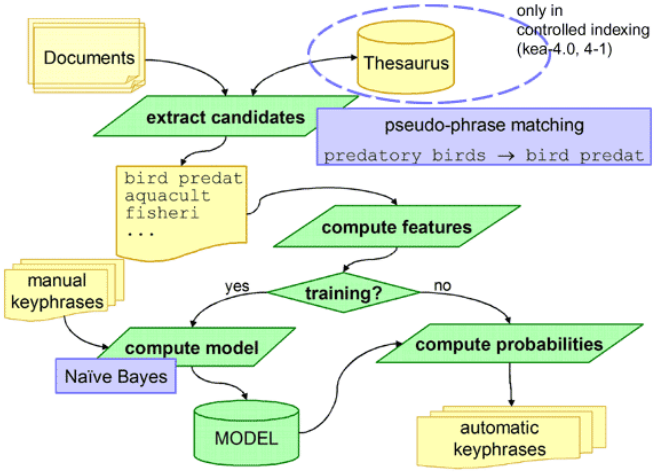
- 4 collections de notices bibliographiques en français et en domaines de spécialité

Collection	Notices		Termes-clés		
	Quantité	Mots moy.	Quantité moy.	"À assigner"	Mots moy.
Linguistique					
Appr.	515	160,5	8,6	60,6 %	1,7
Test	200	147,0	8,9	62,8 %	1,8
Sciences de l'info.					
Appr.	506	105,0	7,8	67,9 %	1,8
Test	200	157,0	10,2	66,9 %	1,7
Archéologie					
Appr.	518	221,1	16,9	37,0 %	1,3
Test	200	213,9	15,6	37,4 %	1,3
Chimie					
Appr.	582	105,7	12,2	75,2 %	2,2
Test	200	103,9	14,6	78,8 %	2,4

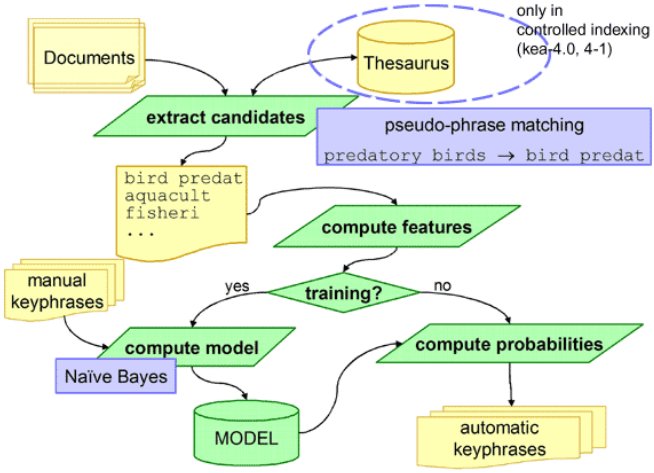
Paramètres expérimentaux II

- Mesures d'évaluation :
  - Précision (P), Rappel (R) et f1-mesure (F) calculés sur les N-meilleurs termes-clés
- Méthodes de référence :
  - TopicRank
  - Tf$\times$Idf
  - Kea++ (méthode supervisée)
- Afin de mesurer l'efficacité de l'ordonnement conjoint de TopicRankSpe :
  - TopicRankSpe$_{libre}$
  - TopicRankSpe$_{contrôlé}$

KEA++

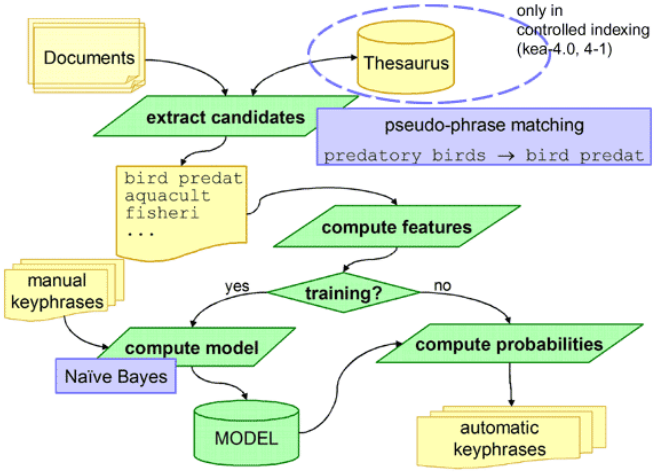

only in
controlled indexing
(kea-4.0, 4-1)
Documents
Thesaurus
extract candidates
pseudo-phrase matching
predatory birds → bird predat
bird predat
aquacult
fisheri
...
compute features
manual
keyphrases
yes
training?
no
compute model
Naïve Bayes
compute probabilities
MODEL
automatic
keyphrases

KEA++



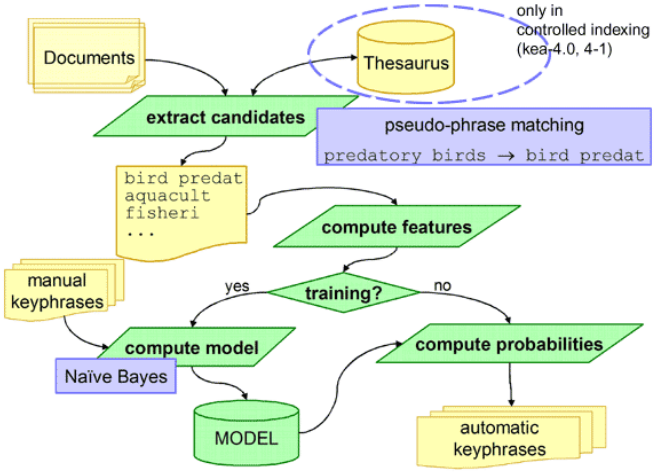
- Entrées :
  - document
  - thésaurus

KEA++



- Entrées :
  - document
  - thésaurus
- Candidats :
  - termes clés assignés

KEA++



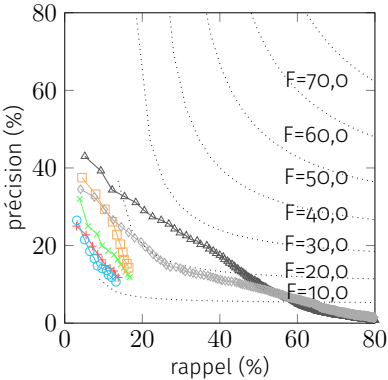
- Entrées :
  - document
  - thésaurus
- Candidats :
  - termes clés assignés
- Classification :
  - TF-IDF
  - $1^{\text{ère}}$ position
  - taille
  - degré sémantique

Résultats I

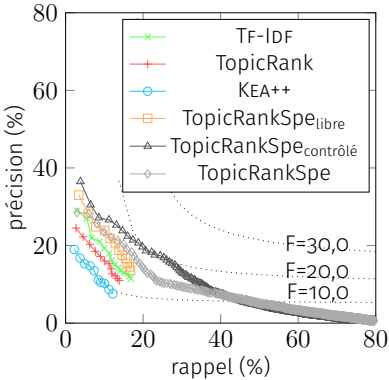
Méthode	Linguistique			Sciences de l'info.			Archéologie			Chimie		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	13,3	15,8	14,2	13,5	14,2	13,4	28,2	19,2	22,3	15,8	12,3	13,2
TopicRank	11,8	13,8	12,5	12,2	12,8	12,2	29,9	20,3	23,7	14,6	11,5	12,3
KEA++	11,6	13,0	12,1	9,5	10,2	9,6	23,5	16,2	18,8	11,4	8,5	9,2
TopicRankSpe$_{\text{libre}}$	14,3	16,5	15,1	15,4	15,9	15,2‡	36,7	24,6	28,8†	15,8	12,1	13,1
TopicRankSpe$_{\text{contrôlé}}$	**24,5**	**28,3**	**25,8**	**19,7**	**19,8**	**19,2**‡	**47,8**	**32,3**	**37,7**†	**20,0**	**14,8**	**16,3**†
TopicRankSpe	18,8	21,9	19,9	17,3	17,7	17,0‡	38,3	25,7	30,1†	17,2	13,4	14,4‡

TABLE – Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, TopicRankSpe$_{libre}$, TopicRankSpe$_{contrôlé}$ et TopicRankSpe appliqués aux collections Termith. † et ‡ indiquent une amélioration significative vis-à-vis des méthodes de référence, à 0,001 et 0,05 pour le t-test de Student, respectivement.

Résultats II

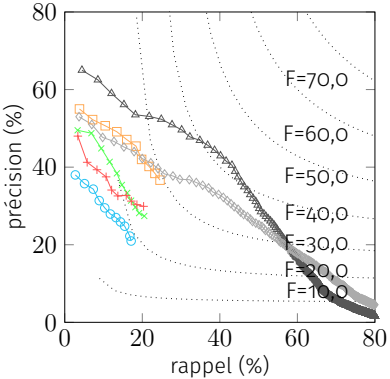


(a) Linguistique

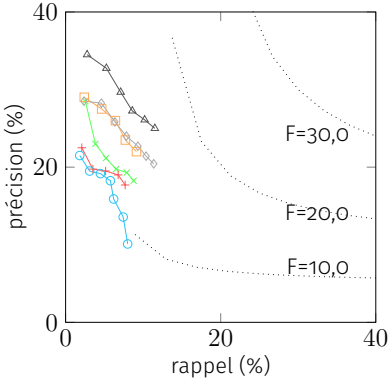


(b) Sciences de l'info.

Résultats III



(c) Archéologie



(d) Chimie

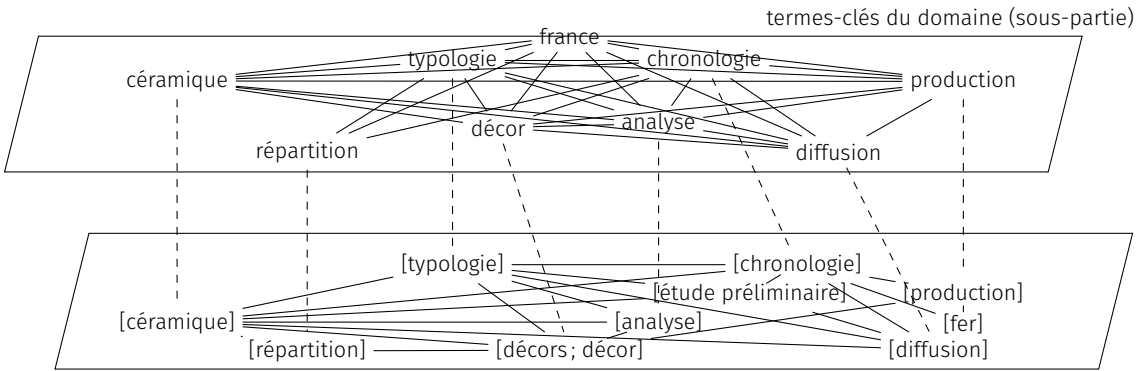
Exemple I

Étude préliminaire de la céramique non tournée micacée du bas Languedoc occidental : typologie, chronologie et aire de diffusion

L'étude présente une variété de céramique non tournée dont la typologie et l'analyse des décors permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le décor effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de fouilles anciennes menées sur les oppidums du Cayla à Mailhac (Aude) et de Mourrel-Ferrat à Olonzac (Hérault). La carte de répartition fait état d'échanges ou de commerce à l'échelon macrorégional rarement mis en évidence pour de la céramique non tournée. S'il est difficile de statuer sur l'origine des décors, il semble que la production s'insère dans une ambiance celtisante. La chronologie de cette production se situe dans le deuxième âge du Fer. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence : distribution ; mourrel-ferrat ; olonzac ; le cayla ; mailhac ; micassé ; céramique non-tournée ; celtes ; production ; echange ; commerce ; cartographie ; habitat ; oppidum ; site fortifié ; fouille ancienne ; identification ; décor ; analyse ; répartition ; diffusion ; chronologie ; typologie ; céramique ; etude du matériel ; hérault ; aude ; france ; europe ; la tène ; age du fer.

Exemple II


termes-clés du domaine (sous-partie)
france
typologie
chronologie
céramique
production
décor
analyse
répartition
diffusion
[typologie]
[chronologie]
[étude préliminaire]
[production]
[céramique]
[analyse]
[fer]
[répartition]
[décors ; décor]
[diffusion]
sujets du document (sous-partie)

Exemple III

Sortie de TopicRank : décors ; céramique ; chronologie ; typologie ; production ; fin ; étude préliminaire ; fer ; deuxième âge ; aire.

Sortie de TopicRankSpe : céramique ; décors ; typologie ; chronologie ; production ; étude préliminaire ; diffusion ; analyse ; france ; répartition.

Plan

Discussion

- TopicRankSpe : méthode à base de graphe pour l'indexation par termes-clés libres et contrôlés
- Meilleures performances que l'état-de-l'art
- La connaissance du domaine, encodées par le graphe de termes-clés du domaine, joue un rôle prépondérant dans l'amélioration des résultats