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

Adrien Bougouin, Florian Boudin, Béatrice Daille

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

Adrien Bougouin Florian Boudin Béatrice Daille
LINA – UMR CNRS 6241, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
<prenom.nom>@univ-nantes.fr

RÉSUMÉ

Dans cet article, nous nous intéressons à l'indexation de documents de domaines de spécialité par l'intermédiaire de leurs termes-clés. Plus particulièrement, nous nous intéressons à l'indexation telle qu'elle est réalisée par les documentalistes de bibliothèques numériques. Après analyse de la méthodologie de ces indexeurs professionnels, nous proposons une méthode à base de graphe combinant les informations présentes dans le document et la connaissance du domaine pour réaliser une indexation (hybride) libre et contrôlée. Notre méthode permet de proposer des termes-clés ne se trouvant pas nécessairement dans le document. Nos expériences montrent aussi que notre méthode surpasse significativement l'approche à base de graphe état de l'art.

ABSTRACT

Unified document and domain-specific model for keyphrase extraction and assignment

This paper focuses on document indexing from keyphrases as performed by professional indexers. From an analysis of indexers working at Digital Libraries, we propose a graph-based method that combines both document information and domain-specific knowledge to perform both keyphrase extraction and assignment (free and controlled indexing). Apart from begin able to assign keyphrases that do not necessarily appear within documents, our experiments show that our approach outperforms the state-of-the-art graph-based approach.

MOTS-CLÉS : Indexation libre ; extraction de termes-clés ; indexation contrôlée ; assignement de termes-clés ; domaine de spécialité ; méthode à base de graphe ; ordonnancement conjoint.

KEYWORDS: Free indexing ; keyphrase extraction ; controlled indexing ; keyphrase assignment ; specific domain ; graph-based method ; graph co-ranking.

1 Introduction

Les termes-clés caractérisent le contenu d'un document. Ceux-ci sont plus communément appelés mots-clés, mais afin d'éviter toute ambiguïté concernant leur forme syntagmatique, un terme clé pouvant indifféremment être un mot simple ou une expression multi-mots, nous préférons l'appellation « terme-clé » et réservons l'usage de « mot-clé » aux seuls mots simples. La tâche d'indexation par termes-clés consiste à identifier automatiquement les termes-clés d'un document. Dans la littérature, elle se divise en deux catégories : l'indexation libre, qui fournit des termes-clés apparaissant dans le contenu du document, et l'indexation contrôlée, qui fournit des termes-clés appartenant à un vocabulaire contrôlé et n'apparaissant pas nécessairement dans le document. Utile pour de nombreuses tâches,

telles que la recherche d'information (Jones & Staveley, 1999), le résumé automatique (D'Avanzo & Magnini, 2005) et la classification de document (Han *et al.*, 2007), l'indexation par termes-clés fait l'objet de nombreux travaux (Hasan & Ng, 2014). Toutefois, la majorité des travaux existants s'intéresse principalement à l'indexation libre.

Contrairement aux travaux de la littérature, l'indexation par termes-clés réalisée par des indexeurs professionnels inclut aussi bien des termes-clés libres que contrôlés. Si nous prenons l'exemple des ingénieurs documentaliste de l'Inist (Institut de l'information scientifique et technique), les pratiques d'indexation manuelle respectent cinq règles qui impliquent une indexation (hybride) libre et contrôlée :

1. Conformité : les termes-clés doivent être conformes au contenu du document et au langage documentaire utilisé dans son domaine ;
2. Exhaustivité : les termes-clés doivent représenter tous les aspects importants du document, même lorsque ceux-ci sont implicites ;
3. Homogénéité : les termes-clés des documents d'un même domaine doivent être cohérents et identiques lorsqu'ils représentent le même concept ;
4. Spécificité : les termes-clés doivent décrire le contenu d'un document au niveau le plus spécifique et peuvent parfois être accompagnés de termes-clés plus génériques afin de restituer le contenu du document dans son domaine ;
5. Impartialité : les termes-clés ne doivent pas être représentatifs d'un jugement apporté par l'indexeur.

Si les critères de conformité et d'homogénéité sont en faveur d'une indexation contrôlée, les principes d'exhaustivité et de spécificité nécessitent l'usage simultané d'une indexation libre et d'une indexation contrôlée.

Dans cet article, nous présentons une méthode à base de graphe pour l'indexation par termes-clés (hybride) libre et contrôlée. Pour ce faire, les termes-clés candidats extraits du document et les termes du domaine sont modélisés dans deux graphes que nous unifions. Les termes-clés candidats et les termes du domaine sont ensuite ordonnés par importance par un processus d'ordonnancement conjoint. Les plus importants sont proposés comme termes-clés.

Le reste de cet article est organisé comme suit. Dans un premier temps, nous présentons brièvement l'état de l'art des méthodes d'indexation automatique par termes-clés. Dans un second temps, nous présentons notre nouvelle méthode à base de graphe. Enfin, nous présentons les résultats de notre travail, puis concluons et présentons quelques perspectives.

2 Indexation par termes-clés

Dans cette section, nous présentons les méthodes de la littérature pour l'indexation automatique par termes-clés. Tout d'abord, nous nous intéressons à l'indexation libre, puis à l'indexation contrôlée.

2.1 Indexation libre

L'indexation libre est l'approche la plus employée pour l'indexation automatique par termes-clés. Les méthodes de la littérature utilisant cette approche appliquent diverses techniques (Hasan & Ng, 2014),

du simple ordonnancement statistique de termes-clés candidats (Salton *et al.*, 1975), à la classification binaire de ces mêmes termes-clés candidats (Witten *et al.*, 1999), en passant par un ordonnancement à base de graphe des mots du document (Mihalcea & Tarau, 2004). Comme notre travail repose sur la technique de l'ordonnancement à base de graphe, nous ne décrivons que les méthodes relevant de cette dernière catégorie.

Depuis les travaux fondateurs de Mihalcea & Tarau (2004) avec TextRank, la technique d'ordonnancement à base de graphe pour l'indexation automatique par termes-clés est très populaire. L'idée originale de cette technique est de représenter le document sous la forme d'un graphe de cooccurrences de mots, puis d'ordonner les mots par importance au sein du graphe. Ensuite, les k mots les plus importants (mots-clés), sont utilisés pour extraire les termes-clés, qui ne sont autres que les expressions du document qui contiennent uniquement des mots-clés.

L'ordonnancement par importance au sein du graphe repose sur le principe de la recommandation, ou du vote. Soit le graphe $G = (N, A)$, un graphe non-orienté constitué de nœuds représentant chacun un mot et un ensemble d'arêtes reliant deux nœuds si les mots qu'ils représentent cooccurrent dans le document. Un nœud n_i est d'autant plus important qu'il est connecté à beaucoup d'autres nœuds et que les nœuds auxquels il est connecté sont eux mêmes importants :

$$\text{Importance}(n_i) = (1 - \lambda) \times \lambda \sum_{n_j \in A(n_i)} \frac{\text{Importance}(n_j)}{|A(n_j)|} \quad (1)$$

où $A(n_i)$ est l'ensemble des nœuds connectés au nœud n_i et λ est un facteur de lissage fixé à 0,85 par Brin & Page (1998).

Dans la continuité des travaux de Mihalcea & Tarau (2004), Wan & Xiao (2008) ont proposés de pondérer les arêtes selon le nombre de cooccurrences $c_{i,j}$ entre les mots des nœuds n_i et n_j, de sorte que ceux qui cooccurrent le plus fréquemment se transfèrent plus d'importance (voir l'équation 2).

$$\text{Importance}(n_i) = (1 - \lambda) \times \lambda \sum_{n_j \in A(n_i)} \frac{c_{i,j} \times \text{Importance}(n_j)}{\sum_{n_k \in A(n_j)} c_{j,k}} \quad (2)$$

Plus récemment, Bougouin *et al.* (2013), ont cherché à tirer profit des sujets abordés dans le document. Ils considèrent comme sujet les groupes de termes-clés candidats ayant un nombre suffisant de mots en commun. Ces sujets sont ensuite ordonnés, puis les termes-clés sont extraits des sujets les plus importants (un terme-clé par sujet). Notre travail étend celui de Bougouin *et al.* (2013), nous revenons donc plus en détail sur cette méthode dans la section 3.

2.2 Indexation contrôlée

Contrairement à l'indexation libre, l'indexation contrôlée nécessite un vocabulaire spécifique au domaine du document analysé. Elle a pour objectif d'indexer les documents de manière homogène (un seul terme-clé par concept, quels que soient les documents) et spécifique au domaine (respectueuse du langage du domaine). Dans ce but, l'indexation contrôlée est plus difficile, car elle doit être capable de trouver des termes-clés qui ne sont pas obligatoirement présents dans le contenu du document.

Medelyan & Witten (2006) ont proposé KEA++, une méthode d'indexation par termes-clés contrôlée qui assigne les termes-clés à partir d'un thésaurus (vocabulaire contrôlé). Dans un premier temps, les

termes du thésaurus sont projetés dans le document et ceux présents dans le document sont retenus comme termes-clés candidats. Dans un second temps, les termes-clés candidats sont classés en tant que « terme-clé » ou « non terme-clé » avec un classifieur naïf bayésien et trois traits statistiques : le TF-IDF du terme-clé candidat (Witten *et al.*, 1999), sa première position (Witten *et al.*, 1999) et le nombre de relations d'association qu'il entretien avec les autres au sein du thésaurus.

Au cours de la campagne d'évaluation BioASQ (Partalas *et al.*, 2013), l'indexation contrôlée par termes-clés a été formulée en un problème de classification multi-étiquettes multi-classes. Les étiquettes sont les termes-clés du document et les classes les entrées d'un vocabulaire contrôlé. Le problème de classification multi-étiquettes multi-classes est généralement considéré comme plusieurs problèmes de classification binaire. Soit un classifieur est appris pour chaque classe. Soit un classifieur est appris pour chaque paire de classes et les classes retenues sont celles proposées par le plus de classifieurs.

3 Une approche hybride

L'approche que nous proposons vise à résoudre le problème d'indexation par termes-clés de manière globale en effectuant les deux catégories d'indexation, libre et contrôlée. L'objectif est de simuler l'indexation manuelle réalisée dans un cadre professionnel, celui des bibliothèques numériques.

La méthode que nous proposons étend la méthode TopicRank (Bougouin *et al.*, 2013). TopicRank est une méthode à base de graphe fonctionnant en cinq grandes étapes :

1. **Sélection des termes-clés candidats.** TopicRank suit les travaux précédents (Wan & Xiao, 2008; Hasan & Ng, 2010) en sélectionnant les plus longues séquences de noms et d'adjectifs en tant que termes-clés candidats.
2. **Groupement en sujets.** Tous les termes-clés candidats sont réduits à des sacs de mots racinisés et ceux partageant un quart de leurs mots racinisés sont groupés au sein du même sujet.
3. **Construction du graphe.** Le document est représenté par un graphe complet où les nœuds sont les sujets. Chaque sujet est connecté aux autres par une arête pondérée selon la force du lien entre les sujets. Plus les termes-clés candidats de deux sujets sont proches dans le document, plus la pondération de l'arête est élevée entre les deux sujets.
4. **Ordonnancement des sujets.** À la manière de TextRank (Mihalcea & Tarau, 2004), les sujets sont ordonnés par importance de sorte que plus un sujet est fortement connecté à un grand nombre de sujets, plus il gagne d'importance, et plus les sujets avec lesquels il est fortement connecté sont importants, plus l'importance qu'il gagne est forte.
5. **Extraction des termes-clés.** Un unique terme-clé est extrait pour chacun des k plus importants sujets. Bougouin *et al.* (2013) ont choisi de sélectionner dans chaque sujet le terme-clé candidat qui apparaît en premier dans le document.

Notre méthode modifie la construction du graphe, l'ordonnancement par importance et la sélection des termes-clés de TopicRank. La construction du graphe étend le graphe de sujet en l'unifiant à un graphe des termes-clés de référence du domaine. L'ordonnancement est désormais conjoint entre les sujets du document et les termes-clés du domaine. Enfin, la sélection des termes-clés ajoute la possibilité de puiser dans le graphe du domaine afin de réaliser une indexation contrôlée.

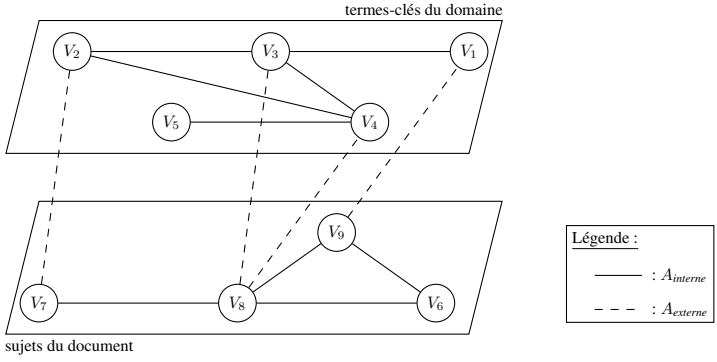


FIGURE 1 – Illustration du graphe unifié que nous proposons

3.1 Construction du graphe

Afin de réaliser simultanément indexation libre et contrôlée, nous unifions deux graphes : l'un représentant le document (graphe de sujets) et l'autre les termes-clés de référence de son domaine (graphe du domaine). Le premier graphe sert à l'indexation libre. Le second, construit à partir des termes-clés de référence de documents d'entraînement, sert à l'indexation contrôlée. À l'instar de Chaimongkol & Aizawa (2013) en extraction de termes techniques, nous faisons l'hypothèse que les termes-clés de référence des documents d'entraînement constituent la terminologie du domaine et nous les utilisons comme substituts au vocabulaire contrôlé usuel en indexation contrôlée. Contrairement aux termes-clés candidats sélectionnés dans le document, les termes-clés de référence ne sont pas redondants et ne nécessitent pas d'être groupés en sujets.

Soit le graphe unifié non orienté $G = (N, A = A_{interne} \cup A_{externe})$. N dénote indifféremment les sujets et les termes-clés du domaine. A regroupe les arêtes $A_{interne}$, qui connectent deux sujets ou deux termes-clés du domaine, et les arêtes $A_{externe}$, qui connectent un sujet à un terme-clé de référence (voir la figure 1). Le graphe de sujets et le graphe du domaine sont unifiés grâce aux arêtes $A_{externe}$. En considérant le domaine comme une carte conceptuelle, l'objectif des arêtes $A_{externe}$ est de connecter le document à son domaine par l'intermédiaire des concepts qu'ils partagent. Une arête $A_{externe}$ est créée entre un sujet et un terme-clé du domaine si ce dernier appartient au sujet, c'est-à-dire correspond à l'un de ses termes-clés candidats.

Pour permettre un ordonnancement conjoint des sujets et des termes-clés du domaine, le schéma de connexion entre deux sujets et entre deux termes-clés du domaine (arêtes $A_{interne}$) doit être homogène. En effet, si les conditions de connexion et si la pondération des arêtes ne sont pas équivalentes et du même ordre de grandeur, alors l'impact du domaine sur le document, et du document sur le domaine, sera marginal. Pour obtenir un graphe unifié homogène, nous connectons deux sujets ou deux termes-clés du domaine n_i et n_j lorsqu'ils apparaissent dans le même contexte et nous pondèrons leur arête par le nombre de fois que cela se produit (poids(n_i, n_j)). Lorsqu'il s'agit des sujets, le contexte est une phrase du document ; lorsqu'il s'agit des termes-clés du domaine, le contexte est l'ensemble des termes-clés de référence d'un document d'entraînement [1].

1. Les contextes étant utilisés pour la création du graphe, le graphe de sujets n'est plus complet comme celui de TopicRank.

3.2 Ordonnancement conjoint des sujets et des termes-clés du domaine

L'ordonnancement conjoint des sujets et des termes-clés du domaine établit leur ordre d'importance vis-à-vis du contenu du document et du domaine. Pour cela, un score d'importance est attribué simultanément aux sujets et aux termes-clés du domaine. Nous reprenons le principe de la recommandation de TopicRank et l'adaptons au problème d'ordonnancement conjoint. Les premières hypothèses de recommandation sont donc les mêmes que celles de TopicRank :

— un sujet est d'autant plus important s'il est fortement connecté à un grand nombre de sujets et si les sujets avec lesquels il est fortement connecté sont importants ;
— un terme-clé du domaine est d'autant plus important s'il est fortement connecté à un grand nombre de termes-clés du domaine et si les termes-clés du domaine avec lesquels il est connecté sont importants.

Ces hypothèses de recommandation, que nous qualifions d'internes, permettent d'établir l'importance des sujets les uns par rapport aux autres et l'importance des termes-clés du domaine les uns par rapport aux autres. Cependant, elles ne permettent pas de tirer profit des relations entre sujets et termes-clés du domaine. Par ailleurs, l'importance des termes-clés du domaine ne dépend pas du document. Nous ajoutons donc deux nouvelles hypothèses de recommandation, que nous qualifions d'externes :

— un sujet est d'autant plus important s'il est représenté par (connecté à) d'importants termes-clés du domaine ;
— un terme-clé du domaine est d'autant plus important vis-à-vis du contenu du document s'il véhicule (est connecté à) l'un de ses sujets importants.

Les sujets et termes-clés du domaine sont ainsi évalués d'après leur usage dans le document et leur importance dans le domaine. L'ordonnancement des uns joue un rôle sur celui des autres et permet ainsi d'effectuer conjointement une indexation libre et contrôlée.

L'équation 3 exprime le calcul de l'importance d'un sujet ou d'un terme-clé du domaine à partir de sa recommandation interne $R_{interne}$ et de sa recommandation externe $R_{externe}$:

$$S(n_i) = (1-\lambda)\, R_{externe}(n_i) + \lambda\, R_{interne}(n_i) \tag{3}$$

$$R_{interne}(n_i) = \sum_{n_j \in A_{interne}(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A_{interne}(n_j)} \text{poids}(n_j, n_k)} \tag{4}$$

$$R_{externe}(n_i) = \sum_{n_j \in A_{externe}(n_i)} \frac{S(n_j)}{|A_{externe}(n_j)|} \tag{5}$$

où $A_{interne}(n_i)$ représente l'ensemble de tous les nœuds connectés au nœud n_i par une arête $A_{interne}$, où $A_{externe}(n_i)$ représente l'ensemble de tous les nœuds connectés au nœud n_i par une arête $A_{externe}$ et où le facteur λ permet de définir la recommandation la plus influente entre $R_{interne}$ et $R_{externe}$. Par défaut, nous définissons $\lambda = 0,5$.

3.3 Sélection des termes-clés

Nous utilisons l'ordre d'importance établit avec le score S des sujets et des termes-clés du domaine pour déterminer les termes-clés du document. Les k nœuds du graphe unifié ayant obtenu les meilleurs scores sont retenus, qu'ils soient des sujets ou des termes-clés du domaine.

Lorsqu'un terme-clé du domaine doit être assigné (indexation contrôlée), une étape de vérification supplémentaire est effectuée pour s'assurer que son importance relève aussi bien du domaine que du document. En effet, il est possible que le graphe du domaine soit constitué de composantes connexes, soit de sous-graphes dont les nœuds ne sont connectés qu'entre eux. Dans ce cas, il se peut qu'un terme-clé du domaine d'un sous-graphe ne soit connecté, ni directement, ni indirectement (par l'intermédiaire d'un autre nœud), à un sujet du document. Son importance est donc déterminée uniquement à partir du domaine et il n'est donc pas pertinent de l'assigner au document.

Lorsqu'un nœud retenu représente un sujet, c'est la même stratégie que celle de TopicRank qui est appliquée. Pour un sujet donné, le terme-clé extrait est son terme-clé candidat qui apparaît en premier dans le document.

3.4 Exemple

La figure 2 donne un exemple d'indexation automatique par termes-clés (hybride) libre et contrôlée avec notre méthode, à partir d'une notice bibliographique d'un article d'archéologie. Dans cet exemple, nous observons une meilleure indexation par termes-clés qu'avec TopicRank. Tout d'abord, nous voyons que le graphe du domaine permet l'assignement du terme-clés générique « France » qui n'est pas présent dans le document. Ensuite, nous voyons que les relations de « diffusion », « analyse » et « répartition » dans le graphe du domaine permettent de les ordonner mieux qu'avec TopicRank (ils font maintenant partie des termes-clés correctement extraits).

4 Paramètres expérimentaux

4.1 Collections de données

Nous conduisons nos expériences sur quatre collections de notices bibliographiques en domaines de spécialité : linguistique, sciences de l'information, archéologie et chimie. Chaque collection est constituée d'environ 700 notices en français extraites depuis les bases de données de l'Inist. Les notices sont constituées d'un titre, d'un résumé d'article scientifique et de termes-clés annotés manuellement par des indexeurs professionnels. Le tableau 1 présente ces collections. Chacune d'elles est répartie en deux sous-ensembles : un ensemble d'apprentissage que nous utilisons pour représenter le domaine, et un ensemble de test pour l'évaluation.

La quantité de termes-clés contrôlés indiquée dans le tableau 1 montre l'importance de l'indexation contrôlée. En effet, plus de la moitié des termes-clés ne peuvent pas être obtenus par indexation libre.

4.2 Méthodes de référence

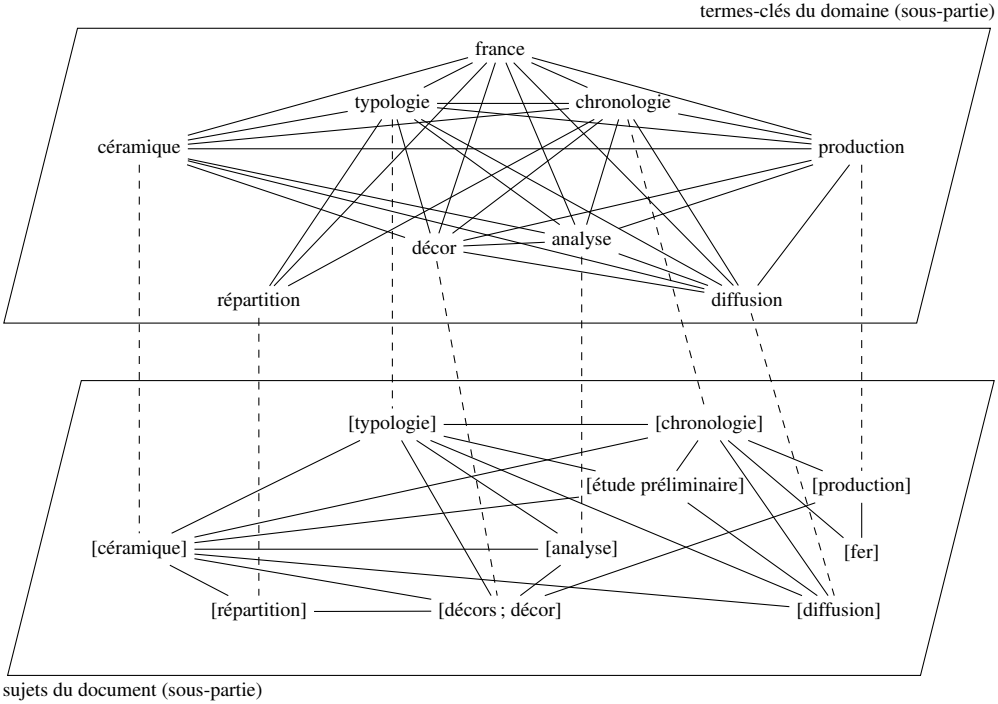
Dans nos expériences, nous comparons notre méthode, que nous appelons TopicRankSpe, à TF-IDF, TopicRank et KEA++. Pour cette dernière, nous utilisons les thésaurus décrivant les vocabulaires contrôlés de l'Inist en linguistique, sciences de l'information, archéologie et chimie.

Afin de mesurer l'efficacité de l'ordonnancement conjoint, nous comparons aussi TopicRankSpe à deux variantes. La première, $\text{TopicRankSpe}_{\text{libre}}$, ne réalise que l'indexation libre. La seconde,

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.



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

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

FIGURE 2 – Exemple d'extraction de termes-clés avec notre méthode sur le résumé d'une notice d'un artice d'archéologie. Les termes-clés soulignés sont les termes-clés correctement extraits.

Collection	Documents		Termes-clés		
	Quantité	Mots moy.	Quantité moy.	Contrôlés	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

TABLE 1 – Collections de données

$\text{TopicRankSpe}_{\text{contrôlé}}$, n'effectue que l'indexation contrôlée.

4.3 Mesures d'évaluation

Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f1-mesure (F). En accord avec l'évaluation menée dans les travaux précédents, les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont effectuées sur leurs racines préalablement calculées par le racineur de Porter (1980).

Nous représentons aussi les résultats sous la forme de courbes de rappel–précision. Celles-ci permettent d'observer si une méthode domine les autres pour les critères de rappel et de précision. Une méthode dominante est une méthode dont les valeurs de précision et de rappel, quelles qu'elles soient, sont toujours supérieures à celles des autres méthodes. Pour générer ces courbes, nous calculons la précision et le rappel lorsque le nombre de termes-clés proposés varie de un jusqu'au nombre total de termes-clés candidats.

5 Résultats

Nous réalisons ici une série d'expériences destinées à comparer TopicRankSpe à l'existant, puis à observer son comportement selon différentes configurations.

Le tableau 2 montre les performances de TopicRankSpe en domaines de spécialité (linguistique, sciences de l'information, archéologie, chimie) comparées à celles des méthodes de référence. De manière générale, les résultats montrent le bien fondé de TopicRankSpe : la variante $\text{TopicRankSpe}_{\text{contrôlé}}$ réalise les meilleures performances, suivie par TopicRankSpe et $\text{TopicRankSpe}_{\text{libre}}$. Les faibles performances de KEA++ sont surprenantes, d'autant plus que la seule autre méthode d'indexation contrôlée, $\text{TopicRankSpe}_{\text{contrôlé}}$, est celle qui réalise les meilleures performances. Contrairement à $\text{TopicRankSpe}_{\text{contrôlé}}$, KEA++ se limite aux entrées du thésaurus qui occurrent dans le document,

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
$\text{TopicRankSpe}_{\text{libre}}$	14,3	16,5	15,1	15,4	15,9	$15,2^{\ddagger}$	36,7	24,6	$28,8^{\dagger}$	15,8	12,1	13,1
$\text{TopicRankSpe}_{\text{contrôlé}}$	**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	$17,0^{\ddagger}$	38,3	25,7	$30,1^{\dagger}$	17,2	13,4	$14,4^{\ddagger}$

TABLE 2 – Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, $\text{TopicRankSpe}_{libre}$, $\text{TopicRankSpe}_{contrôlé}$ 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.

	Libre (%)	Contrôlé (%)
Linguistique	61,7	38,3
Sciences de l'info.	66,4	33,6
Archéologie	69,1	30,9
Chimie	68,4	31,6

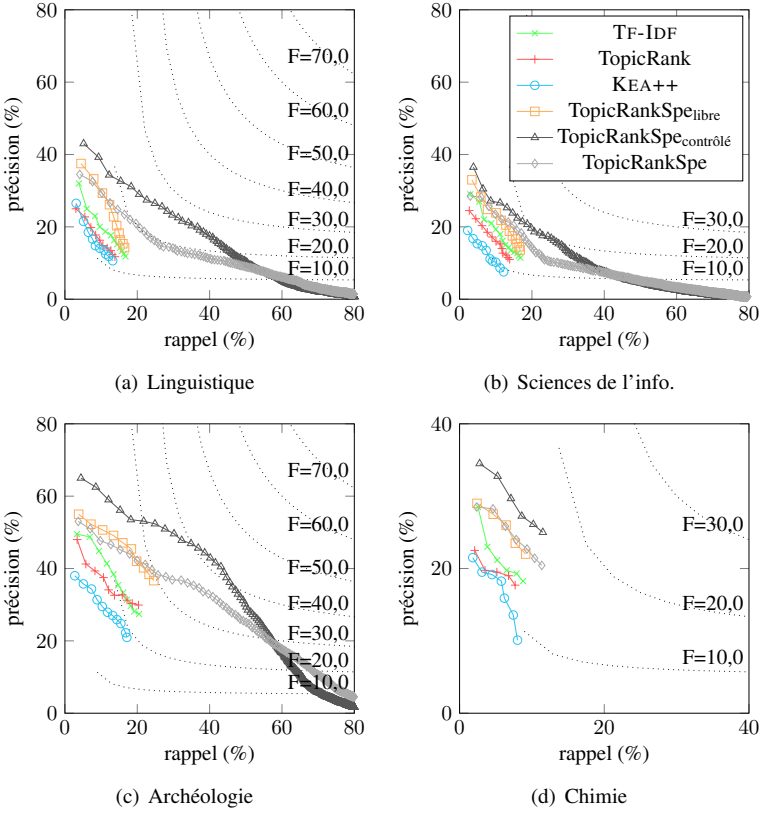
TABLE 3 – Taux moyens d'indexation libre et contrôlée réalisée par TopicRankSpe sur les données Termith

alors que la majorité des termes-clés des collections Termith n'apparaissent pas dans les documents. De plus les thésaurus de l'Inist ne sont pas aussi riches que ceux utilisés par Medelyan & Witten (2006) dans leurs expériences : moins de relations y sont définies entre les concepts. TopicRankSpe et ses variantes sont significativement meilleurs que les méthodes de référence. Comparées à celles de TopicRank, les performances de $\text{TopicRankSpe}_{\text{libre}}$ montrent que le domaine apporte des informations permettant d'ordonner plus précisément les sujets du document. Le fait que $\text{TopicRankSpe}_{\text{contrôlé}}$ obtienne les meilleures performances montre aussi que les termes-clés du domaine sont ordonnés efficacement d'après le contenu du document (ses sujets).

Les courbes de précision et de rappel de la figure 3 permettent de comparer le comportement respectif des méthodes de référence, de TopicRankSpe et de ses variantes [2]. Elles montrent que TopicRankSpe et ses variantes dominent les méthodes de référence selon les critères de précision et de rappel. Parmi elles, nous observons aussi que la variante $\text{TopicRankSpe}_{\text{contrôlé}}$ domine la variante $\text{TopicRankSpe}_{\text{libre}}$, mais que TopicRankSpe n'est, ni dominante, ni dominé par elles. Bien que l'amélioration significative de TopicRank par TopicRankSpe et ses variantes montrent l'apport de l'ordonnancement conjoint entre sujets du document et termes-clés du domaine, la réalisation simultanée de l'indexation libre et contrôlée reste difficile.

Afin d'observer la place que prend l'indexation contrôlée dans TopicRankSpe, nous nous intéressons maintenant aux taux de termes-clés proposés par indexation libre ou contrôlée par TopicRankSpe. Le table tableau 3 montre les taux d'indexation libre et contrôlée pour chaque collection de notices. Nous observons que l'indexation libre est légèrement prédominante face à l'indexation contrôlée. Les deux catégories d'indexation par termes-clés sont effectivement réalisées conjointement, mais

2. Notez que le rappel maximum affiché pour TopicRankSpe est supérieur à celui des autres méthodes, car TopicRankSpe propose des termes-clés absents des documents, contrairement aux autres méthodes.



(a) Linguistique (b) Sciences de l'info.

(c) Archéologie (d) Chimie

FIGURE 3 – Courbes de rappel-précision de TF-IDF, TopicRank KEA++, TopicRankSpe$_{libre}$, TopicRankSpe$_{contrôlé}$ et TopicRankSpe appliqués aux données Termith

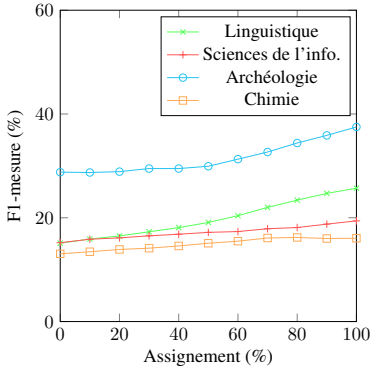


FIGURE 4 – Performance de TopicCoRank appliqué aux données Termith, lorsque le taux d'indexation contrôlée varie

l'ordonnancement donne plus d'importance aux sujets du document qu'aux termes-clés du domaine. En domaines de spécialité où l'indexation contrôlée est préférée, cela peut être résolu en travaillant sur un affinage des schémas de connexion des nœuds de chaque graphe.

Enfin, nous réalisons une dernière expérience pour déterminer si l'ordonnancement des termes-clés est efficace. Pour cela, nous faisons une expérience dans laquelle nous modifions l'étape de sélection des termes-clés de TopicRankSpe et forçons le taux d'indexation contrôlée. Un ordonnancement efficace des termes-clés du domaine doit induire une courbe de performance cumulative quand nous faisons croître le taux d'indexation contrôlée [3]. La figure 4 montre la performance de TopicRankSpe lorsque le taux d'indexation contrôlée varie de 0 % à 100 % avec un pas de 10 %. À chaque augmentation du taux d'indexation contrôlée, la performance de TopicRankSpe augmente. L'ordonnancement des termes-clés du domaine fait donc émerger efficacement les plus importants vis-à-vis du document.

6 Conclusion

Dans cet article, nous avons présenté une méthode à base de graphe permettant de réaliser simultanément une indexation par termes-clés libre et contrôlée. Notre méthode, qui étend TopicRank, unifie le graphe des sujets du document à un graphe représentant les termes de son domaine de spécialité, puis les ordonne conjointement. Comparée à l'état de l'art, notre méthode a montré de meilleures performances en termes de précision, de rappel et de f-mesure. En domaines de spécialité, la connaissance du domaine joue un rôle crucial pour améliorer l'indexation automatique par termes-clés.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).

3. Dans cette situation, cela signifie que la performance obtenue avec $\text{TopicRankSpe}_{\text{contrôlé}}$ est la performance maximale avec TopicRankSpe

Références

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). Topicrank : Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543–551, Nagoya, Japan : Asian Federation of Natural Language Processing.

BRIN S. & PAGE L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, **30**(1), 107–117.

CHAIMONGKOL P. & AIZAWA A. (2013). Utilizing LDA Clustering for Technical Term Extraction. In *Proceedings of the 19th Annual Meeting of the Association for Natural Language Processing (ANLP)*, p. 686–689, Nagoya, Japan : Association for Natural Language Processing.

D'AVANZO E. & MAGNINI B. (2005). A Keyphrase-Based Approach to Summarization : the LAKE System at DUC-2005. In *Proceedings of DUC 2005 Document Understanding Conference*.

HAN J., KIM T. & CHOI J. (2007). Web Document Clustering by Using Automatic Keyphrase Extraction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, p. 56–59, Washington, DC, USA : IEEE Computer Society.

HASAN K. S. & NG V. (2010). Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters (COLING)*, p. 365–373, Stroudsburg, PA, USA : Association for Computational Linguistics.

HASAN K. S. & NG V. (2014). Automatic Keyphrase Extraction : A Survey of the State of the Art. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland : Association for Computational Linguistics.

JONES S. & STAVELEY M. S. (1999). Phrasier : a System for Interactive Document Retrieval Using Keyphrases. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 160–167, New York, NY, USA : ACM.

MEDELYAN O. & WITTEN I. H. (2006). Thesaurus Based Automatic Keyphrase Indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, p. 296–297 : ACM.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing Order Into Texts. In DEKANG LIN & DEKAI WU, Eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

PARTALAS I., GAUSSIER É. & NGOMO A.-C. N. (2013). Results of the First BioASQ Workshop. In *BioASQ@ CLEF*, p. 1–8.

PORTER M. F. (1980). An Algorithm for Suffix Stripping. *Program : Electronic Library and Information Systems*, **14**(3), 130–137.

SALTON G., WONG A. & YANG C. (1975). A Vector Space Model for Automatic Indexing. *Communication ACM*, **18**(11), 613–620.

WAN X. & XIAO J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, p. 855–860 : AAAI Press.

WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C. & NEVILL MANNING C. G. (1999). KEA : Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, p. 254–255, New York, NY, USA : ACM.