Représentation vectorielle de mots pour l'indexation de notices bibliographiques

Morgane Marchand[1] Geoffroy Fouquier[1] Guillaume Pitel[1]
(1) eXenSa, 41 rue Périer, 92120 Montrouge
morgane.marchand@exensa.com, geoffroy.fouquier@exensa.com,
guillaume.pitel@exensa.com

RÉSUMÉ

Cet article présente la contribution d'eXenSa à l'édition 2016 au DÉfi Fouille de Textes (DEFT) dont la tâche consiste à indexer des documents scientifiques par des mots clefs, préalablement sélectionnés par des professionnels. Le système proposé est purement statistique et combine une approche graphique et une approche sémantique. La première approche cherche dans le titre et le résumé du document des mots graphiquement proches des mots clefs du thésaurus. La seconde approche attribue à un nouveau document des mots clefs associés aux documents du corpus d'entraînement qui lui sont sémantiquement proches. Les deux approches utilisent des représentations vectorielles apprisent en utilisant l'algorithme NCISC, un algorithme stochastique de factorisation de matrices. Notre approche s'est classée première sur deux des corpus de spécialité proposés et deuxième sur les deux autres.

ABSTRACT

Word vector embeddings for bibliographic records indexing

This article presents the eXenSa contribution to the 2016 DEFT Workshop. The proposed task consists in indexing bibliographic records with keywords chosen by professional indexers. We propose a statistical approach which combines graphical and semantic approches. The first approach defines a document keywords as thesaurus terms graphically similars to terms contained in the title or the abstract of this document. The second approach assigns to document the keywords associated with semantically similar documents in training corpora. Both approach use models generated using NCISC, a stochastic matrix factorisation algorithm. Our system obtains the best F-score on half of the four test corpuses and ranks second for the two others.

MOTS-CLÉS : Indexation, mots clefs, espaces sémantiques, représentation vectorielle de mots.

KEYWORDS: Indexation, keywords, semantic spaces, word vector embedding.

1 Introduction

L'édition 2016 de DEFT propose de travailler sur l'indexation de documents scientifiques par des mots clefs qui ont été proposés par des indexeurs professionnels. Les corpus proposés sont issus de quatre spécialités : linguistique, sciences de l'information, archéologie et chimie. Pour chaque document du corpus d'entrainement, nous disposons d'un titre, d'un résumé, ainsi que d'une liste de mots clefs attribués par un ingénieur documentaliste. Les mots clefs sont soit issus d'un thésaurus de spécialité, soit ajoutés par l'ingénieur documentaliste en fonction de leur pertinence.

L'édition 2012 présentait une tâche similaire, consistant à indexer des articles scientifiques par l'intermédiaire de mots clefs d'auteurs (Paroubek *et al.*, 2012). L'équipe ayant obtenu les meilleurs scores a choisi une approche mixte utilisant des espaces sémantiques (El Ghali *et al.*, 2012). L'approche que nous proposons pour le défi de cette année utilise également des espaces sémantiques. Néanmoins, nous avons pour notre part choisi d'aborder cette tâche de façon purement statistique, et, mis à part le passage en minuscule et la suppression des accents, nous n'utilisons aucun prétraitement. Cela nous permet d'avoir un protocole facilement applicable à des langues peu dotées.

Les représentations vectorielles que nous utilisons ont été générées par l'algorithme NCISC présenté dans la section suivante. Nous présenterons ensuite les deux approches que nous avons définies avant de discuter les résultats dans la section 4.

2 L'algorithme NCISC

NCISC [1] est un algorithme stochastique de factorisation de matrices de la famille de RandNLA (Drineas & Mahoney, 2016), performant sur de grandes dimensions. Il permet d'apprendre des représentations vectorielles sur des tâches non supervisées ou semi-supervisées. Ces vecteurs peuvent être utilisés directement dans des recherches de plus proches voisins ou bien servir d'étape de prétraitement pour d'autres tâches d'apprentissage.

A l'instar de la LSA (Landauer & Dumais, 1997), NCISC effectue une réduction de dimension en réalisant une factorisation d'une grande matrice creuse en trois matrices denses : une matrice contenant des représentations vectorielles pour la première dimension de la matrice, une matrice contenant les valeurs propres des dimensions latentes et une matrice contenant des représentations vectorielles pour la seconde dimension de la matrice. Cette réduction est effectuée de manière itérative. Des connaissances a priori peuvent être introduites au cours du processus sur chacune des dimensions sous forme de vecteurs pré-calculés qui sont concaténés aux matrices en cours de calcul. Cela permet par exemple d'infléchir le calcul afin que deux objets appartenant à une même catégorie (ou deux mots utilisés dans un même contexte) aient au final des représentations approchantes.

Pour ce défi, deux types de matrices ont été utilisés : une matrice mot-graphèmes pour construire des représentations vectorielles graphiques et une matrice document-mots pour construire des représentations vectorielles sémantiques.

3 Stratégies de sélection des mots clefs

3.1 Approche par modèle graphique

Nous avons émis une première hypothèse formulée ainsi : le titre et le résumé d'un article permettent de connaître la teneur de celui-ci. De cette hypothèse découle une règle de sélection des mots clefs à partir du thésaurus thématique correspondant : un mot clef est un terme du thésaurus dont tous les mots pleins qui le composent apparaissent dans le titre ou le résumé d'un article (recherche exacte). Les scores obtenus en recherchant les mots clefs de cette manière sont assez faibles, surtout pour

1. L'algorithme NCISC est développé par la société eXenSa dans son moteur eXenGine (http ://www.exensa.com/about-us/). Une démonstration d'eXenGine analysant les données de wikipédia anglais est disponible à l'adresse wikinsights.org.

le rappel. En effet, un mot ou un expression clef peut apparaître sous une forme différente dans le titre ou le résumé d'une part, et le thésaurus d'autre part. Or, étant donné que nous n'utilisons pas de prétraitement linguistique comme la lemmatisation, les mots au pluriel (ou accordés au féminin) par exemple sont donc considérés comme des mots différents d'une version au singulier (resp. accordée au masculin). Pour que les différentes versions d'un mot puissent être rapprochées, nous avons construit un modèle de représentation vectorielle graphique. Un mot est décomposé en graphèmes de 2 à 5 lettres ordonnées. Nous utilisons des marqueurs de début de mot mais pas de marqueur de fin, afin de ne pas donner de poids à la désinence d'un mot. Par exemple le mot « laches » est composé des graphèmes suivants :

```
[::s::l; la; ac; ch; he; es;
 ::s::la; lac; ach; che; hes;
 ::s::lac; lach; ache; ches;
 ::s::lach; lache; aches]
```

On compose ainsi une matrice mot-graphèmes dont on réduit la dimension à l'aide de NCISC afin d'obtenir des vecteurs de représentation graphique des mots. Ainsi, les vecteurs de « lexique » et « lexiques » ont une similarité très élevée entre eux, mais sont également proches du vecteur représentant le mot « lexical ».

Munis de ces vecteurs graphiques, nous pouvons définir une nouvelle règle de sélection des mots clefs ainsi : chaque terme du thésaurus dont tous les mots pleins qui le composent ont une similarité suffisante avec un terme du titre ou du résumé d'un article est sélectionné comme mot clef (recherche approchée).

	Précision	Rappel	F-mesure
Archéologie	42,12	44,09	41,12
Linguistique	20,36	24,62	21,29
Sciences de l'information	18,86	22,12	19,34
Chimie	23,85	15,36	17,15

TABLE 1 – Résultats obtenus par recherche approchée de mots clefs dans les titres et résumés en utilisant des représentations vectorielles graphiques et un seuil de similarité de sélection de $0,98$.

Le tableau 1 présente les scores obtenus avec une recherche approchée. Cette approche permet une forte augmentation du rappel ainsi qu'une faible baisse de la précision (en moyenne). L'amélioration en terme de F-mesure va de $9,39$ à $14,68$ par rapport à l'approche exacte.

3.2 Approche par modèle sémantique

L'approche précédente, par construction, ne permet d'attribuer que des mots clefs déjà présents dans le titre ou le résumé d'un document. A l'inverse, l'intérêt des espaces sémantiques est justement de prendre en compte des informations de similarité sémantique non explicites. Notre seconde approche utilise ce type de représentation : à partir d'un modèle sémantique issu d'un apprentissage, nous construisons une représentation vectorielle d'un document et nous l'utilisons pour trouver les mots clefs qui lui sont associés. L'approche suivie est la suivante : pour un document donné, nous déterminons quels sont les documents les plus similaires puis, connaissant leur mots clefs respectifs, nous en déduisons les mots clefs liés au document. Il serait possible de chercher directement les mots

clefs associés à un document, sans passer par la recherche de documents voisins, mais cette approche ne donne pas les meilleurs résultats. D'autres travaux ont déjà utilisé l'information contenue dans des documents proches pour l'extraction de mots clefs, comme (Wan & Xiao, 2008) qui réalisent une partition thématique avant de repérer les mots saillants à l'aide d'un algorithme de graphe.

Le modèle sémantique utilisé est une matrice document-mots représentant les documents des corpus en sacs de mots d'uni-, bi- et tri-grammes, sélectionnés selon un critère d'information mutuelle. La dimension de la matrice est réduite à l'aide de l'algorithme NCISC. Au cours de la réduction, nous introduisons deux a prioris selon la procédure présentée dans la section 2 :

— un a priori sur les documents, généré à partir d'une matrice document-mots clefs,
— un a priori sur les mots, généré à partir d'une matrice de cooccurrences issue de la base wikipédia en français.

3.2.1 Vecteur de réprésentation des documents tests

Après apprentissage, chaque document du corpus d'entraînement dispose d'une représentation vectorielle tout comme chaque mot apparaissant dans le corpus. Pour qu'il soit possible de comparer les documents du corpus de test aux documents du corpus d'apprentissage, chaque document du corpus de test doit être représenté par un vecteur compatible. La représentation d'un document est contruite à partir des vecteurs des mots qu'il contient, qui sont sommés membre à membre. Il faut ensuite ôter de ce vecteur les valeurs propres de la matrice document-mots réduite, calulées sur le corpus d'entraînement par NCISC, afin d'obtenir une représentation comparable aux vecteurs des documents d'entraînement. Pour estimer la similarité entre deux documents dans cet article, nous utilisons le complément à 1 de la distance cosinus entre leur représentation vectorielle.

Il est à noter que nous aurions pu apprendre les représentations des mots et des documents sur un corpus élargi constitué à la fois des documents d'entraînement et des documents de test. Entraînées sur plus de données, les représentations vectorielles des tests auraient sans doute été plus précises. Nous avons cependant choisi de rester dans un cas d'utilisation où, quand de nouveaux documents se présentent, il est possible de trouver leurs mots clefs sans avoir besoin de refaire l'apprentissage.

3.2.2 Attribution des mots clefs via les plus proches voisins

Après le calcul de la représentation de chaque document du corpus de test, l'attribution des mots clefs s'effectue de la manière suivante :

— identification des K plus proches voisins parmi les documents du corpus d'entrainement de la spécialité associée,
— suppression des documents voisins ayant une similarité inférieure à un seuil S avec le document test,
— remplacement des documents voisins par leurs mots clefs associés, pondérés par leur similarité respective avec le document test. Si un même mot clef est lié à plusieurs documents voisins du document test, alors les scores de similarité sont additionnés,
— réordonnancement des mots clefs candidats en fonction de leur popularité dans le corpus d'entraînement : leur score de similarité est multiplié par le logarithme du nombre d'apparitions du mot clef dans le corpus d'entraînement.

Les scores présentés dans le tableau 2 montrent que cette approche favorise la précision au détriment du rappel. Les différents paramètres (dimension de la représentation vectorielle, nombre de documents

voisins sélectionnés avec seuil et nombre final de mots clefs proposés) ont été optimisés pour maximiser la F-mesure. La dernière colonne du tableau indique la précision au rang 1. La performance

	Précision	Rappel	F-mesure	Précision au rang 1
Archéologie	63,57	13,35	20,71	76,28
Linguistique	30,26	22,38	24,35	46,98
Sciences de l'information	40,69	11,62	15,88	42,45
Chimie	31,14	7,94	11,07	39,57

TABLE 2 – Résultats obtenus par similarité sémantique entre documents voisins

en terme de rappel est par contre bien moindre qu'avec le modèle précédent. C'est pourquoi nous avons cherché à combiner les deux solutions précédentes.

3.3 Combinaison des approches précédentes

Chacune des approches précédentes fournit une liste de mots clefs candidats associés à un score de similarité. Cependant, ces scores étant issus de deux modèles différents, calculés sur deux matrices différentes, ils ne sont pas directement comparables. Les scores de similarité des mots clefs issus du modèle graphique, plus élevés, sont donc multipliés par le score de similarité du meilleur mot clef issu du modèle sémantique. Les résultats sont ensuite fusionnés. Si un mot clef est proposé à la fois par le modèle sémantique et le modèle graphique, alors leurs scores sont additionnés. Une fois la liste finale ordonnée, les N premiers parmi ceux dont le score final dépasse un certain seuil de sélection sont conservés. Cette simple statégie de fusion donnant déjà de bons résultats, il serait intéressant par la suite d'en tester des plus élaborées.

Nous avons choisi de développer notre système avec peu de paramètres spécifiques au domaine. Cela permet une certaine stabilité pour de nouveaux domaines. De plus, notre approche statistique bénéficie d'un plus grand corpus d'apprentissage. Les modèles sont donc entraînés sur les quatre corpus d'entraînement réunis et les différents seuils ont été choisis pour maximiser la F-mesure moyenne sur les quatre corpus de développement. La dimension des vecteurs de représentation une fois les matrices réduites a ainsi été empiriquement fixée à 130. Seul le nombre maximal N de mots clefs finalement proposés dépend du type de corpus.

	Précision	Rappel	F-mesure	N
Archéologie	43,48	52,71	45,59	44
Linguistique	23,28	32,73	26,30	22
Sciences de l'information	21,26	30,32	23,86	29
Chimie	24,92	21,73	21,46	47

TABLE 3 – Résultats obtenus en combinant les deux approches avec un seuil de sélection de 0,4.

La précision obtenue avec ce modèle mixte est moindre que celle obtenue avec le modèle sémantique mais reste au dessus de celle du modèle graphique. Le véritable apport de cette combinaison est sur le score de rappel. En effet, bien que le modèle sémantique ait un rappel faible, son ajout augmente le rappel déjà élevé du modèle graphique. Le gain observé va de $6,37$ à $8,62$ points selon les cas. Cela confirme que l'utilisation de représentations vectorielles entraînées sur de grands corpus permet de capter de l'information sémantique qui n'est pas explicite dans les documents.

4 Discussion des résultats

Lors de ce défi, sur cinq participants, notre méthode s'est classée première sur les corpus de chimie et d'archéologie, deuxième sur les sciences de l'information et sur la linguistique, validant notre approche statistique. Nos représentations vectorielles sémantiques permettent bien de récupérer des mots clefs non présents dans les documents. Cependant, ce type d'approche de ne permet pas d'attribuer à un document un mot clef qui n'apparaît pas dans le corpus d'entraînement. La proportion de mots clefs du corpus d'apprentissage effectivement présents dans le corpus de test va donc avoir un impact sur les résultats. Cette proportion dépasse en effet en moyenne les 80 % dans les corpus d'archéologie et de sciences de l'information qui obtiennent de hauts scores de précision avec l'approche sémantique.

Des trois approches présentées dans cet article, l'approche sémantique est celle à privilégier pour obtenir une forte précision. A l'inverse, les approches graphiques et mixtes donnent un rappel meilleur que la précision pour trois des quatre corpus. Seul le corpus de chimie fait exception. L'hypothèse de l'approche graphique est que les mots clefs apparaîssent dans le titre ou le résumé. Les représentations vectorielles graphiques permettent d'intégrer certaines variations, en particulier, les graphèmes n'ont pas de marqueur de fin afin de donner peu de poids aux désinences. Par contre, la présence d'un marqueur de début rend difficile de rapprocher des mots qui diffèrent par leur préposition. Or en chimie, beaucoup de mots clefs sont des noms de molécules ou composés chimiques dont les dénominations se font par agglomération. Un texte peut contenir le mot « hexadécylsulfonate » et être indexé par le terme plus générique « sulfonate » qui ne sera donc pas reconnu. Il serait possible d'essayer un modèle entraîné sur des graphèmes ne comportant pas ce marqueur de début. Un autre problème survient lorsqu'un mot clef comprend un mot générique du domaine qui est absent du résumé. Par exemple, un texte peut être indexé par « cellule tumorale » et contenir « tumeurs » mais pas « cellule ». Ces mots, très fréquents dans un domaine, pourraient être repérés à l'aide d'un tf-idf et ne pas être recherchés dans les notices pour la validation des mots-clefs les contenant.

5 Conclusion

Cet article présente la contribution d'eXenSa à DEFT 2016 sur la tâche d'indexation de notices bibliographiques par des mots clefs. Notre approche statistique combine deux parties, l'une graphique et l'autre sémantique. La première cherche dans la notice du document des mots graphiquement proches des mots clefs utilisés dans les corpus d'entraînement. La seconde attribue à un nouveau document les mots clefs associés aux documents du corpus d'entraînement qui lui sont sémantiquement les plus proches. Les deux approches utilisent des représentations vectorielles apprises en utilisant notre algorithme NCISC, un algorithme stochastique de factorisation de matrices. À des fins de généralisation, nous utilisons le même corpus d'entraînement et le même paramétrage pour tous les domaines de spécialité. Notre système final se classe premier sur deux des quatre corpus de test, deuxième sur les autres. De plus, une fois l'apprentissage effectué, l'attribution de mots clefs à un nouveau document ne prend que quelques millisecondes. Une approche statistique est donc bien apropriée pour cette tâche bien qu'elle soit ici pénalisée par la petite taille des corpus. Par construction, notre système n'attribue que des mots clefs déjà présents dans les corpus d'entraînement. La proportion de mots clefs du corpus d'apprentissage effectivement présents dans le corpus de test a donc un impact sur les résultats. Les conventions de mots clefs propres à chaque spécialité ont également une influence, notamment pour la partie graphique de notre méthode.

Références

DRINEAS P. & MAHONEY M. W. (2016). Randnla : randomized numerical linear algebra. *Communications of the ACM*, **59**(6), 80–90.

EL GHALI A., HROMADA D. & EL GHALI K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. volume 2012, p.77̃.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211.

PAROUBEK P., ZWEIGENBAUM P., FOREST D. & GROUIN C. (2012). Indexation libre et contrôlée d'articles scientifiques présentation et résultats du défi fouille de textes deft2012.

WAN X. & XIAO J. (2008). Collabrank : towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, p. 969–976 : Association for Computational Linguistics.