Étiquetage thématique automatisé de corpus par représentation sémantique

Lucie Martinet[**,***], Hussein T. Al-Natsheh[*,***,****], Fabien Rico[*,♯], Fabrice Muhlenbach[*,♯♯], Djamel A. Zighed[*,***]

[*]Université de Lyon, France
[**]CESI EXIA/LINEACT, 19 Avenue Guy de Collongue, F-69130 Écully, France
[***]Lyon 2, ERIC EA 3083, 5 Avenue Pierre Mendès France - F69676 Bron Cedex
[****]CNRS, ISH FRE 3768, 14 avenue Berthelot - 69363 Lyon Cedex 07
[♯]Lyon 1, ERIC EA 3083, 5 Avenue Pierre Mendès France, F69676 Bron Cedex
[♯♯]UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint Etienne

Résumé. Dans les corpus de textes scientifiques, certains articles issus de communautés de chercheurs différentes peuvent ne pas être décrits par les mêmes mots-clés alors qu'ils partagent la même thématique. Ce phénomène cause des problèmes dans la recherche d'information, ces articles étant mal indexés, et limite les échanges potentiellement fructueux entre disciplines scientifiques.

Notre modèle permet d'attribuer automatiquement une étiquette thématique aux articles au moyen d'un apprentissage des représentations sémantiques d'articles du corpus déjà étiquetés. Passant bien à l'échelle, cette méthode a pu être testée sur une bibliothèque numérique d'articles scientifiques comportant des millions de documents. Nous utilisons un réseau sémantique de synonymes pour extraire davantage d'articles sémantiquement similaires et nous les fusionnons avec ceux obtenus par un modèle de classement thématique. Cette méthode combinée présente de meilleurs taux de rappel que les versions utilisant soit le réseau sémantique seul, soit la seule représentation sémantique des textes.

1 Introduction

L'activité des chercheurs a été bouleversée par un accès toujours plus important aux bibliothèques numériques en ligne. La recherche d'information dans ces bibliothèques numériques se fait le plus souvent au moyen de mots-clés entrés dans des moteurs de recherche. Néanmoins, l'appariement entre les mots-clés entrés et ceux utilisés pour décrire les documents scientifiques pertinents présents dans ces bibliothèques numériques peut s'avérer limité si la terminologie employée n'est pas la même dans les deux cas. Tout chercheur appartient à une communauté avec laquelle il partage des connaissances et un vocabulaire communs. Cependant, lorsque celui-ci souhaite étendre l'exploration bibliographique au-delà de sa communauté d'appartenance afin de recueillir des éléments d'information qui le conduisent à de nouvelles connaissances, il convient de lever plusieurs verrous scientifiques et techniques induits par la grande taille des bibliothèques numériques, l'hétérogénéité des données et la complexité du