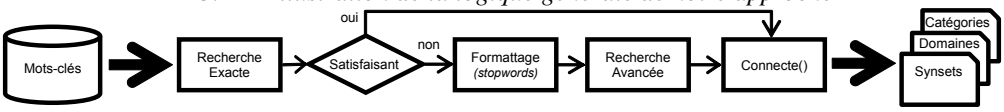
Catégorisation d'articles scientifiques basée sur les relations sémantiques des mots-clés

Bastien Latard*,** Jonathan Weber*
Germain Forestier*, Michel Hassenforder*

*MIPS, Université de Haute-Alsace, Mulhouse, France
**MDPI AG, Bâle, Suisse

Introduction. La recherche bibliographique est une étape cruciale pour tout chercheur. En effet, la connaissance des travaux existant peut faire gagner un temps précieux tant pour le choix de la méthode à adopter que pour être à jour des dernières avancées. Néanmoins, trouver des articles similaires reste une tâche compliquée et pénible autant pour les domaines étendus que réduits. Les chercheurs passent un temps considérable à chercher des travaux proches de leurs intérêts de recherche, disséminés dans 47'000 revues scientifiques appartenant à quelques 6000 éditeurs différents. Cette étape est cependant incontournable dans tout projet de recherche afin de confronter de nouvelles idées à des solutions existantes, ansi que pour l'acquisition de connaissance à propos d'un domaine spécifique. Dans cet article, une nouvelle méthode d'extraction de connexions entre les catégories des mots-clés d'articles scientifiques est proposée. Les limites de notre approche naïve héritée de la recherche exacte ont été soulignées dans Latard et al. (2017), et cet article fournit une amélioration qui s'attaque à ce problème. Notre recherche a pour but d'intégrer les relations sémantiques dans les moteurs de recherche scientifiques afin de les rendre plus intelligents. Effectivement, en fonction du nombre de résultats renvoyés, une requête plus raffinée / étendue pourrait alors être proposée à l'utilisateur.

FIG. 1 – *Illustration de la logique générale de notre approche*



Approche Proposée. Notre approche utilise BabelNet (Navigli et Ponzetto (2012)), une base de données fusionnant lexiques sémantiques (WordNet, VerbNet) et autres bases de données collaboratives (Wikipedia et autres données Wiki). Une requête pour un terme renvoie des "entrées de dictionnaire", des synonymes, des catégories ou des domaines. Cette base de connaissance est intégrée afin d'ajouter de l'information sémantique à partir de tous les mots-clés des articles de la base de données de littérature scientifique, Scilit [1]. Scilit contient à ce jour les métadonnées de plus de 97 millions d'articles. La Figure 1 illustre la logique principale de notre framework. La recherche exacte est l'approche naïve de notre framework qui prend des mots-clés sans préformatage et tente de faire une recherche exacte sur BabelNet. Ses limites sont rapidement atteintes lorsqu'un article comporte des mots-clés composés (plusieurs

1. `http://www.scilit.net` – *développée par MDPI (`http://www.mdpi.com`)*

mots). Ceci est problématique étant donné que 76% des mots-clés Scilit sont composés. Lorsqu'une recherche exacte ne renvoie aucun résultat pour un mot-clé composé, les mots vides (*stopwords*) sont supprimés et le mot-clé est divisé sur les espaces. Cette étape s'appelle la recherche avancée. Le mode utilisé (aussi appelé Multi) est une version étendue des approches

FIG. 2 – *Le mode Multi*

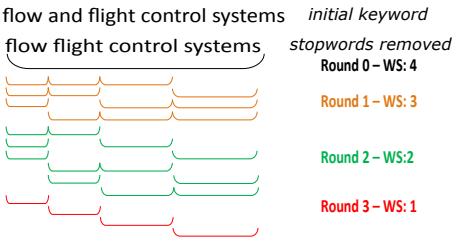


FIG. 3 – *Métriques pour* $\alpha = 4$

Recherche / Métrique	F1	F1($R0$)
Exacte	0.63	0.94
Multi	0.84	0.90

type n-grammes. Il démarre à partir de la plus grande fenêtre (sélection de mots) possible, mais tente toutes les combinaisons linéaires au lieu de seulement les adjacentes. La Figure 2 illustre la logique de propagation de ce mode sur un exemple réel et souligne son avantage principal par rapport aux approches n-grammes où *"flow control"* est plus significatif que *"flow flight"* pour le mot-clé *"flow and flight control systems"*. Cependant des résultats inattendus provenant de combinaisons indésirables, pour de long mots-clés, peuvent aussi être générés.

Analyse et Conclusion. Finalement, notre approche extrait avec succès *"Aerodynamics"* comme catégorie principale de *"lift coefficient ; normalized lift ; flapping flight"*, grâce aux recherches exactes et avancées qui retournent *"Aerodynamics"* respectivement pour *"lift coefficient"* et *"flapping flight"*. Les 43 synsets hérités de *"lift"* (*"normalized lift"*) sont conservés car il n'y a aucun chevauchement de catégories entre les deux mots, et les synsets n'ayant aucun rapport sont naturellement filtrés par la connexion des catégories des mots clés. La Figure 3 montre $F1$, un indicateur unique représentant le ratio précision/rappel. $F1(R0)$ est une variante évaluant précision et rappel seulement pour les catégories pour lesquelles nous pouvons estimer un degré de certitude minimum (2 mots-clés partagent la même catégorie). Les résultats détaillés sont disponibles depuis ce lien [2]. Le dossier compressé contient 595 articles de 7 journaux provenant de 2 éditeurs, ainsi que les résultats des différentes méthodes et modes testés (plus de détails dans le fichier readme.txt). En validant les entrées dans le dictionnaire à partir des catégories principales, le sens des mots-clés (et de l'ensemble de leurs synsets) est également vérifié, ce qui permet une exploitation plus poussée des données BabelNet.

Références

Latard, B., J. Weber, G. Forestier, et M. Hassenforder (2017). Towards a Semantic Search Engine for Scientific Articles. In *TPDL*. Springer.

Navigli, R. et S. P. Ponzetto (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.

2. `http://img.mdpi.org/data/latard_egc2018.zip`