Proposition de Stage de Master 2 / 5è année Ingénieur

Le TAL au service de la fouille d'articles scientifiques. Méthodes, objets d'études, sites d'études et impact des publications en écologie : une question de genre ?

Anne Loison, Laboratoire d'Ecologie Alpine, Université Savoie Mont-Blanc
Patrice Bellot, Laboratoire Informatique et Systèmes (LIS), Professeur, Université Aix-Marseille CNRS

Résumé.

Le stage vise à identifier, à partir d'une analyse automatisée d'un corpus d'articles scientifiques issus de revues d'écologie, si le genre des auteurs impacte les méthodes, modèles, espèces et type d'écosystèmes étudiées, et les sites d'études. Dans un deuxième temps, la relation entre genre des auteurs, performance individuelle des auteurs, contenu et l'impact des articles sera aussi étudiée.

Sur le plan informatique, le stage combine des problématiques du traitement automatique des langues, de la recherche d'information et de la fouille de données : extraction d'information (notamment reconnaissance d'entités nommées et identification de mots-clés), représentations de documents et partitionnement à partir d'approches neuronales (modèles de thèmes de type BERTopic), analyse de graphes et détection de communautés.

Contexte et objectif applicatif.

Les statistiques concernant l'équilibre des genres dans la recherche académique montrent que le pourcentage d'hommes et de femmes varie en fonction de la discipline concernée, et que dans les domaines dans lesquels les femmes sont traditionnellement minoritaires, les trajectoires de carrière des femmes sont en moyenne plus lentes que celles des hommes. La question de l'évaluation de la carrière et de la promotion repose en grande partie sur des critères de performance en termes de production scientifique. Se pose toutefois le défi d'évaluer l'originalité, la quantité, la qualité, et l'impact des recherches menées par une personne en particulier, que ce soit en termes d'impacts académique ou sociétal. Dans ce cadre, la disponibilité **d'outils bibliométriques** relativement faciles d'utilisation (logiciels dédiés, packages R et Python, outils fournis dans Google Scholar, Web of Knowledge, Altmetric) permet non seulement de calculer des critères usuels de performance (nombres d'articles, position, nombre de citations, « h-index ») et de visibilité en ligne (le nombre de consultations, de commentaires ou de partages), mais également **des métriques propres aux analyses de réseaux complexes**, reposant entre autres sur les listes de coauteurs, les adresses, les mots clés, les listes de référence. Des logiciels tels que Gargantext[1] et VosViewer[2] permettent d'explorer de tels réseaux de publications et d'auteurs au moyen de visualisations avancées.

Ceci ouvre donc la possibilité d'étudier plus finement où se jouent éventuellement des différences de performance et d'impact entre les genres. Par ailleurs, un des aspects peu étudiés des différences de genre est le contenu des articles publiés, qui permet cependant d'accéder à d'éventuelles différences de genre dans la production de connaissances, c'est-à-dire les méthodes utilisées, les organismes étudiés ou les lieux. Si ces caractéristiques diffèrent entre genre, et qu'elles influencent aussi l'impact des articles, alors, elles pourraient expliquer certaines des différences de performance bibliométriques et par suite, du déroulement de la carrière[3].

---

[1] https://iscpif.fr/projects/gargantext/

[2] https://www.vosviewer.com

[3] Depuis la Déclaration de San Francisco sur l'évaluation de la recherche (*San Francisco Declaration on Research Assessment*, DORA, 2012), le Manifeste de Leiden (2015) et la coalition européenne CoARA (*Coalition for Advancing Research Assessment*), les pratiques d'évaluation des activités de recherche évoluent. Si elles doivent dépasser le seul usage de quelques métriques, telles le h-index et le facteur d'impact des revues (JIF) et des articles (AIF), dont les limites sont bien identifiées, force est de constater que l'évolution des carrières en dépendent toujours.

Étapes de réalisation du stage pour le/la stagiaire :

1. Constitution de corpus et annotation automatique des noms d'auteurs : constituer un corpus d'articles de journaux en écologie à partir des APIs des services en lignes tels que HAL, Istex ou Google Scholar (SerpAPI) puis attribuer automatiquement le genre des auteurs à partir des noms et prénoms ;
2. extraction d'information et modèles de thèmes : identifier à partir du contenu des articles et des métadonnées, les mots-clés caractérisant les articles parmi lesquels figurent les organismes étudiés, les sites d'études, les méthodes, les écosystèmes, les pays dans lesquels les études ont pris place. Ces ensembles de mots clés pourront être comparés aux mots caractérisant les thèmes extraits automatiquement par une approche de type allocation latente de Dirichlet (LDA) ;
3. modèles de thèmes et partitionnements : un partitionnement des articles, à partir du texte intégral ou des mots-clés identifiés précédemment, sera effectué à partir de modèles de langues pré-entraînées et de représentations continues distribuées (approches BERTopic ou Top2Vec). Il s'agira alors de déterminer les liens potentiels entre genres et *clusters* ;
4. analyse de graphes et détection de communautés : analyser le graphe auteurs/mots-clés/articles/indicateurs, afin d'identifier si le genre influence la façon dont se produisent les connaissances, et influence l'impact des articles et la notoriété individuelle. Un premier point consistera en une analyse du réseau de collaborateurs en fonction du genre et de la « notoriété » des auteurs telle qu'exprimée par diverses métriques.

NB : le stage s'appuiera sur la méthodologie et les codes sources développés par Davide Rendina durant son stage de Master Recherche de Centrale-Supélec, intitulé « *Semantic Analysis of Web Archive Historical Data* » et co-dirigé par P. Bellot, S. Gebeil et M. Génois (AMU), 2023.

Problématiques « informatique » et « fouille de données textuelles ».

La première étape ci-dessus repose sur l'exploitations de services Web et d'APIs existantes. En ce qui concerne l'annotation automatique des noms d'auteurs, il s'agira d'évaluer, dans la lignée de Sebo (2021), la performance et les limites des services possibles (dont *gender*[4] et *genderizeR*[5] avec R, *Genderize*[6] avec Python) sur un échantillon du corpus, notamment en fonction des zones géographiques des auteurs.

L'extraction d'information (étape 2) consiste ensuite en l'identification automatique de certaines entités nommées présentes dans les contenus des articles. Ceux-ci étant généralement disponibles seulement en version PDF, il faudra tout d'abord extraire le contenu textuel avant de l'analyser avec une bibliothèque telle que spaCy et différents modèles de langue pré-entraînés (par ex. WikiNEuRal) pour l'identification des entités candidates. Parmi ces dernières, seules celles caractérisant l'article devront être automatiquement retenues : noms des méthodes et approches, noms des lieux et des êtres vivants et organismes étudiés… On s'appuiera pour cela d'une part sur les métadonnées disponibles et, d'autre part, sur des approches d'extraction automatique de mots-clés ou encore de résumé automatique [Firoozeh et al., 2020] [Hernández-Castañeda et al. 2022]. Ces listes de mots-clés pourront être comparées à celles obtenues par allocation latente de Dirichlet (LDA) sur le texte intégral avec scikit-learn.

Les modèles de thèmes (étape 3) seront mis en œuvre afin d'obtenir des représentations thématiques du corpus ou de sous-ensembles d'articles (par ex. selon le genre). Des outils logiciels tels que BERTopic [Grootendorst, 2022] seront utilisés avec différents modèles de langues généralistes ou spécialisés.

Enfin, des graphes de données peuvent être construits (étape 4). Les mots-clés, les thèmes et les entités sont associés aux auteurs et une analyse des co-occurrences (nombre ou information mutuelle) permet d'associer les thèmes entre eux tout comme le sont les co-auteurs. Des indicateurs bibliométriques extraits de services tels que Google Scholar peuvent enrichir les données. Un logiciel tel que Gargantext permettant la détection de communautés dans des graphes pourra être utilisé de manière interactive, tout comme certaines APIs Python mettant en œuvre les algorithmes de Louvain et ou de Girvan-Newman.

---

[4] https://cran.r-project.org/web/packages/gender/index.html
[5] https://github.com/kalimu/genderizeR
[6] https://pypi.org/project/Genderize/

Pour l'ensemble des étapes, le concours d'experts en écologie, permettra d'orienter l'étude, de raffiner les modèles numériques et, le cas échéant, d'aider à la constitution de données d'entraînement. Il sera bien sûr indispensable pour l'évaluation des modules intermédiaires et l'analyse des résultats finaux.

Localisation et encadrement.

Le/la stagiaire réalisera son stage au Où. Le/la stagiaire sera sous la responsabilité de Anne Loison, directrice de recherche au CNRS au LECA et de Patrice Bellot, professeur à l'université Aix-Marseille.

Compétences requises :

- Méthodes du traitement automatique des langues à base d'apprentissage machine pour l'extraction d'information et la classification automatique de textes (transformeurs, LDA, représentations vectorielles…)
- Intérêt pour les analyses bibliométriques et scientométriques
- Langage Python et bibliothèques spaCy, scikit-learn, Pandas et Keras ou PyTorch

Durée :

5 à 6 mois entre janvier et mi-juillet 2024
Poursuite en thèse envisagée

Divers :

Des déplacements entre l'Université Aix-Marseille et l'université Savoie Mont-Blanc sont à prévoir (frais de déplacement et d'hébergement pris en charge).

Quelques références :

Boekhout, H., van der Weijden, I., & Waltman, L. (2021). Gender differences in scientific careers: A large-scale bibliometric analysis. arXiv preprint arXiv:2106.12624.

Bradshaw, C. J., & Courchamp, F. (2018). Gender bias when assessing recommended ecology articles. Rethinking Ecology, 3, 1-12.

Campbell, S. E., & Simberloff, D. (2022). The Productivity Puzzle in Invasion Science: Declining but Persisting Gender Imbalances in Research Performance. BioScience, 72(12), 1220-1229.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2023). Gender gaps at the academies. Proceedings of the National Academy of Sciences, 120(4), e2212421120.

Ceci, S. J., Kahn, S., & Williams, W. M. (2023). Exploring gender bias in six key domains of academic science: An adversarial collaboration. Psychological Science in the Public Interest, 15291006231163179.

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), 259-291.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., & Millán-Hernández, C. E. (2022). Language-independent extractive automatic text summarization based on automatic keyword extraction. *Computer Speech & Language*, *71*, 101267.

Sebo P. (2021). Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association : JMLA*, *109*(3), 414–421. https://doi.org/10.5195/jmla.2021.1185