v

La Rochelle Université
Institut Littoral Urbain
Durable Intelligent

PROPOSITION DE SUJET POUR UN CONTRAT DOCTORAL

Laboratoire

L3i - LRUniv

Titre de la thèse

Évaluation ouverte, plurielle et flexible d'indices de production de la recherche

Direction de la thèse *directeur·trice·s (grade, HDR) et éventuels co-directeur·trice·s*

Antoine Doucet, PR

Co-encadrant

Esma Talhi, MCF (EIGSI)

Adéquation scientifique avec les priorités de l'établissement

Un objectif central de ce projet est de savoir qualifier et quantifier la performance d'une spécialisation thématique, avec pour cas d'usage celle du LUDI.

Scientifiquement, ce travail en traitement automatique des langues s'insère dans la continuité des activités en humanités numériques que l'établissement porte depuis une quinzaine d'années. L'analyse des contenus de bibliothèques numériques d'articles scientifiques rentre dans ce cadre et trouve un écho avec un récent projets Horizon 2020 du L3i (Embeddia), étendu dans plusieurs projets RNA ESR, notamment Termitrad. Fournir une estimation thématique de la production scientifique trouve un écho particulier dans le cadre de la stratégie de spécialisation de LRUniv et de son université européenne EU-CONEXUS. Il est en effet à l'heure actuelle très difficile de mesurer l'effectivité de cette spécialisation, et généralement de toutes les activités de recherche inter- et pluridisciplinaire.

Descriptif du sujet *(enjeux scientifiques, applicatifs, sociétaux…)*

L'enseignement supérieur a connu un développement important et ce, depuis quelques années. Le croisement entre les différentes disciplines et la diversification des parcours et des spécialités ont permis l'avènement de nouvelles institutions, universités, laboratoires de recherches, programmes pédagogiques, etc. Vincke [1] explique que ce développement se traduit par deux caractéristiques :

1. Une augmentation significative du nombre d'acteurs de l'enseignement supérieur et de la recherche : de nouvelles universités richement dotées apparaissent dans les pays émergents, des instituts privés se créent autour de nous, des universités prestigieuses installent des antennes aux quatre coins du monde, des programmes universitaires diplômants sont proposés sur Internet, …

2. Une mobilité croissante des étudiants et des enseignants-chercheurs, mobilité encouragée par les pouvoirs publics de tous les pays.

Devant cette croissance, étudiants, chercheurs et directions d'universités ont besoin de disposer d'informations sur les contenus proposés ainsi que les domaines de recherches des laboratoires avec lesquels ils sont amenés à collaborer.

Les classements mondiaux des universités (Shanghai, The Times Higher Education Supplement, US news, QS ranking, sont devenus des outils bien établis que les étudiants, les directions d'université et les décideurs politiques lisent et utilisent. Aussi discutables soient-ils, ils sont devenus un facteur clé dans le secteur de l'enseignement supérieur, suscitant un intérêt croissant et exerçant une influence extraordinaire sur un large éventail de parties prenantes [2].

Chaque classement prétend avoir une méthodologie unique capable de mesurer la qualité des universités, cependant, des études ont montré que ces classements comportent un biais de réputation [3,4,5]. La réputation ici pouvant être définie comme étant "une représentation collective qu'ont les parties prenantes (étudiants, alumni, évaluateurs, etc) d'une institution". [6] estime que la réputation est le résultat des performances d'une université. Ce biais de la réputation a une incidence sur le taux d'acceptation des articles de recherche par les éditeurs de revues. En effet, [7] explique que les rédacteurs en chef des revues les plus prestigieuses peuvent être enclins à donner plus de chances aux articles d'auteurs affiliés à des universités telles que *Oxford* ou *Harvard*, soit en réduisant le taux de rejet, soit en augmentant la probabilité d'acceptation même si les rapports des examinateurs ne concordent pas. D'autres critères font intervenir des nombres de citations d'auteurs ou d'articles. Or dans ce qui est expliqué précédemment, ce critère est fortement influencé par la réputation. De manière générale, les experts en bibliométrie admettent que des domaines tels que les sciences de l'ingénieur, les sciences de l'environnement, les sciences sociales, le droit et, plus globalement les sciences humaines (à l'exception de certaines parties de la psychologie et de l'économie) ne peuvent, pour le moment du moins, être traitées de façon satisfaisante par les techniques habituelles d'analyse bibliométrique [1]. Le même auteur indique que les disciplines des chercheurs les plus cités sont : Mathématiques, Physique, Chimie, Biologie – Biochimie, Informatique, Géosciences, Sciences de l'espace, Ingénierie, Science des matériaux, Agriculture, Environnement, Médecine clinique, Médecine vétérinaire, Pharmacologie, Biologie moléculaire et génétique, Microbiologie, Immunologie, Neurosciences. Ce problème de catégories génériques empêche par ailleurs l'identification d'universités et laboratoires interdisciplinaires ou travaillant sur de nouvelles spécialités comme le cas de l'université de La Rochelle qui s'est spécialisée dans la thématique du Littoral Urbain, Durable et Intelligent (LUDI). Le biais de réputation joue également bien entendu contre les établissements plus petits et jeunes, comme LRUniv et l'EIGSI.

Dans ce cadre, le travail doctoral aura pour objectif d'explorer des approches permettant de s'appuyer sur l'intelligence artificielle afin de proposer une nouvelle méthodologie de classification des activités des universités. Cette méthodologie devrait s'affranchir des biais exprimés plus haut en définissant des critères qui permettent d'évaluer de façon objective les universités.

D'autre part, afin d'identifier les équipes et universités travaillant sur des thématiques spécifiques, il sera nécessaire d'effectuer une analyse discursive de tous les documents décrivant ces activités, notamment des communications mais aussi des articles de recherche. En effet, ces derniers permettent d'identifier les domaines de recherche d'une université en analysant le contenu et l'affiliation des auteurs. Cette analyse vient améliorer la méthodologie en la dotant d'un modèle de connaissance, organisé sous la forme de taxonomies permettant d'organiser et visualiser les connaissances extraites : les domaines des universités, leur collaboration, thématique de

recherches, les universités qui traitent les mêmes domaines, etc. La construction d'une coloration thématique des publications scientifiques s'appuiera sur les avancées récentes des modèles à base de transformer tels que BERT, permettant une description sémantique pré-entraînée. Son paramétrage et sa mise en œuvre dans le contexte de ce travail doctoral sera un des verrous scientifiques importants de ce travail doctoral.

Une finalité applicative est de pouvoir prendre en compte les activités interdisciplinaires, mais aussi de développer une méthode fonctionnant sur la base de mots- ou phrases-clés données en entrée.

Bibliographie

1. Vincke, Philippe. "Les classements d'universités." *Pyramides. Revue du Centre d'études et de recherches en administration publique* 14 (2007): 71-94.
2. Hazelkornis, Ellen. "Global rankings and the geopolitics of higher education." (2017).
3. Safón, Vicente. "Inter-ranking reputational effects: an analysis of the Academic Ranking of World Universities (ARWU) and the Times Higher Education World University Rankings (THE) reputational relationship." Scientometrics 121.2 (2019): 897-915.
4. Shin, Jung Cheol. "Organizational effectiveness and university rankings." University rankings. Springer, Dordrecht, 2011. 19-34.
5. Toutkoushian, Robert K., and Karen Webber. "Measuring the research performance of postsecondary institutions." University rankings. Springer, Dordrecht, 2011. 123-144.
6. Keith, Bruce. "Organizational contexts and university performance outcomes: The limited role of purposive action in the management of institutional status." Research in Higher Education 42.5 (2001): 493-516.
7. Safón, Vicente, and Domingo Docampo. "Analyzing the impact of reputational bias on global university rankings based on objective research performance data: the case of the Shanghai Ranking (ARWU)." *Scientometrics* 125.3 (2020): 2199-2227.

Contexte partenarial *(cotutelle internationale, EU-CONEXUS, partenariat avec un autre laboratoire, une entreprise…)*

La thèse est une collaboration entre LRUniv et l'EIGSI, cette dernière finançant cette allocation doctorale à 50%.

Le travail se fera en collaboration avec les universités partenaires d'EU-CONEXUS afin d'établir un tour d'horizon de leurs pratiques et attentes. Cet état des lieux visera à nourrir le cas d'usage autour de la spécialisation LUDI.

Impacts *(scientifiques, technologiques, socio-économiques, environnementaux, sociétaux…)*

De ce travail doctoral découle la réponse à plusieurs besoins, pour lesquels il n'existe pas d'alternative connue permettant :

- une analyse **ouverte** des activités de recherche de différentes entités ;
- une analyse **ajustable** à toute spécialité, inter- ou trans-discipline ;
- une analyse par **mots-clés** ;
- la **recherche d'experts** pour les dimensions listées précédemment.

Les applications immédiates se situent dans le conseil, l'aide à la décision et/ou la recherche d'experts. Un transfert pourrait se faire en approchant des cabinets de conseil de l'ESR tels que SIRIS Academic. La recherche d'experts permet aussi de renforcer le lien entre les mondes industriels et académiques, en allant au-delà des classements globaux (aller des classements d'universités en informatique vers un référencement de laboratoires/chercheurs sur la base de requêtes par mots clés : "analyse de documents", "ambiance dans les bâtiments", etc.).

Du point de vue sociétal, une ambition est de permettre une alternative aux classements rigides et thématiques qui occupent l'espace médiatique.

Programme de travail du doctorant *(tâches confiées au doctorant)*

Le travail doctoral consistera en 2 volets liés, avec l'ambition de traiter les questions de recherche suivantes :

1. Comment construire une base de connaissance regroupant toutes les informations relatives aux universités et leurs équipes de recherche et la combiner à des représentations statiques telles que les plongements de mots contextuels ?
2. Comment prendre en compte de façon multimodale la terminologie fournie par les auteurs (mots clés), celle fournie par les éditeurs (taxonomie) et celle extraite automatiquement (topic modelling tel que LDA), et pondération vis-à-vis de facteurs bibliométriques classiques ?

Concrètement, le doctorant aura en charge les missions suivantes :

- Analyse bibliographique et synthèse ;
- Proposition de méthodologie(s) d'évaluation représentative(s) des tâches à aborder pour adresser les questions de recherche ;
- Définition des critères d'évaluation à prendre en compte pour classifier des universités, leurs poids et leur classification ;
- Extraction d'information à partir de la segmentation discursive des documents représentant les universités ainsi que leurs travaux, formations et équipes de recherche ;
- Création d'une taxonomie en ordonnant les informations extraites et en définissant les règles qui les lient ;
- Évaluation des résultats et validation de la méthodologie proposée ;
- Dissémination et publication des résultats dans des revues et/ou colloques.

Le plan de travail prévisionnel du doctorat est le suivant :

Semestre 1 :

- Étude approfondie de l'état de l'art sur les différentes problématiques du sujet de thèse ;
- Définition d'une méthodologie d'évaluation représentative des tâches à aborder pour aborder les questions de recherche ;
- Analyse des méthodes de classement existant et identification des biais de ces évaluations

Semestre 2 :

- Identification des critères d'évaluation, leur poids, impact, utilité, etc
- Prise en main des outils de la littérature et évaluation, notamment la partie Extraction des connaissances ;
- Proposition d'un modèle de connaissance d'une méthode pour identifier les mots clés (informations relatives aux universités) et identifier les règles qui les lient afin de créer une taxonomie.

Semestre 3

- Développement d'un démonstrateur pour valider la première partie de la méthodologie d'évaluation
- Prise en main des outils de création de taxonomie, comme les ontologies pour une représentation semi-automatique ;
- Validation du modèle avec les outils choisis

Semestre 4

- Intégration du modèle sémantique (de connaissances) dans la méthodologie globale
- Développement de la partie 2 du démonstrateur après intégration du modèle

Semestre 5 :

- Fusion des solutions proposées, afin de mettre en place une proposition globale composée d'une méthodologie et une implémentation permettant de s'appuyer sur l'Intelligence Artificielle afin de d'évaluer les universités et d'analyser, extraire et représenter la connaissance relative à leurs travaux de recherche, spécialités, domaines et parcours d'études.

Semestre 6 :

- Rédaction du manuscrit et préparation de la soutenance.

Calendrier de réalisation