Rapport Network Analysis

Julien Bastian and Colin Fourment

Contributing authors: julien.bastian@univ-lyon2.fr;
colin.fourment@univ-lyon2.fr;

1 Introduction, présentation du corpus

Dans ce travail, nous explorons la création d'un graphe à partir de publications académiques issues du portail Persée, en mettant l'accent sur la construction d'un réseau de coauteurs. Notre corpus de données est composé d'articles académiques provenant de différentes disciplines, et notre objectif est de comprendre les relations entre ces articles à travers les références citées.

Le corpus de données que nous utilisons est constitué de neuf ensembles de données, chacun contenant des informations sur des articles académiques, y compris des identifiants, des auteurs et des références. Pour cette démonstration, nous nous concentrons uniquement sur le neuvième ensemble (nommé *rscir_xxs*) de données, mais la méthode et le code sont adaptées à l'utilisation sur l'ensemble du corpus.

Nous faisons le choix d'utiliser les auteurs dans chaque article comme base pour la construction du graphe, car cela nous permet de capturer les relations entre les articles et les auteurs à travers la relation de coautorage. En construisant un graphe de coauteurs, nous pouvons explorer la structure du réseau académique et découvrir des informations importantes sur les collaborations et les domaines de recherche.

Le corpus complet contient 909780 articles et l'ensemble de donnée sur lequel nous produisons nos illustrations en contient 59529 soit 15% des articles. On y retrouve 16 domaines scientifique différents : *Archéologie, Arts, Droit, Etudes classiques, Etudes g. a. cultur., Etudes regionales, Geographie, Histoire, Litterature, Pluridisciplinaire par essence, Religion theologie, Science de l'éducation, Science de l'environnement, Science politique, Sciences de la Terre, Sociologie.* Chaque texte est rattaché à l'un d'entre eux uniquement. La répartition de ces domaines est Illustré en Figure 1, on y constate une large sur-représentation des articles du domaine *Religion theologie* et *Histoire.*

Dans les sections suivantes, nous détaillons les différentes étapes de notre approche, chacune offrant une contribution distincte à notre analyse des relations entre les articles académiques. Tout d'abord, dans la section sur la préparation du graphe, nous

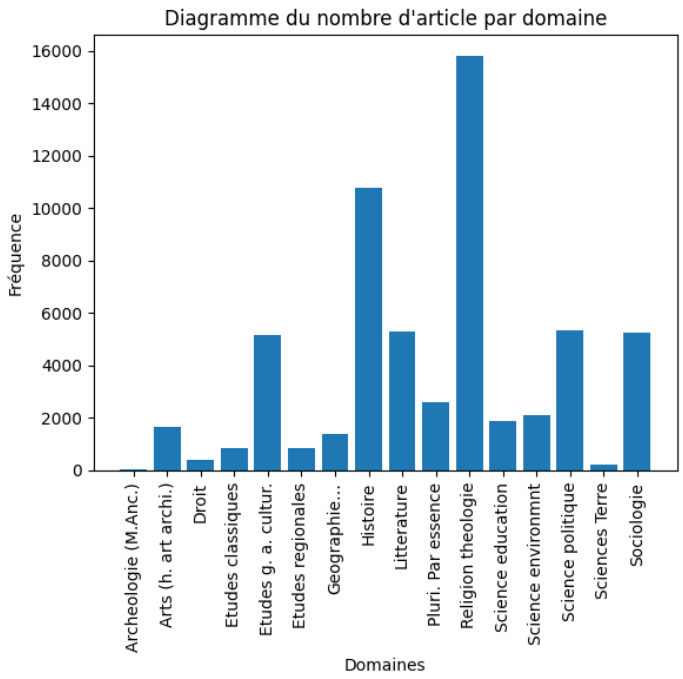

Fig. 1 Diagramme des classes dans le dataset considérée

décrivons en détail le processus de construction du graphe à partir des données fournies par le portail Persée. Cette étape est essentielle car elle établit la base de notre analyse en capturant les relations entre les articles à travers les coauteurs, ce qui nous permet de visualiser et d'explorer la structure du réseau académique.

Ensuite, nous abordons la création d'un moteur de recherche basé sur l'encodage de texte avec le modèle CamemBERT. Cette méthode nous permet de rechercher des textes similaires à partir d'une requête textuelle et du titre des articles, offrant ainsi un outil précieux pour explorer et naviguer dans le corpus de manière efficace et pertinente.

Dans la section suivante, nous explorons le clustering de Louvain comme méthode pour détecter les communautés dans notre graphe. Cette approche nous permet d'identifier les groupes d'articles étroitement liés au sein du réseau académique, ce qui peut révéler des tendances, des thématiques ou des domaines de recherche spécifiques. Nous proposons également dans notre code de comparer les résultats du clustering avec les domaines scientifiques des articles en uilisant le score de Rand ajustée.

Enfin, dans la dernière section, nous examinons la classification des articles scientifiques par propagation d'étiquettes (*Label propagation*). Cette méthode nous permet d'attribuer des étiquettes à des articles non étiquetés en exploitant la structure du graphe et les relations entre les articles. En utilisant une approche semi-supervisée, nous pouvons prédire avec précision les étiquettes des articles, même avec un petit nombre d'étiquettes disponibles, en utilisant les liens entre articles pour propager

les domaines des articles pour lesquels ils sont connus aux articles avec lesquels ils possédent une relation dans le graphe.

2 Préparation du graphe

La construction du graphe se déroule en plusieurs étapes importantes. Tout d'abord, nous chargeons les données à partir des fichiers de données pickle, en extrayant les informations sur les articles, y compris les identifiants, les auteurs et les références citées. Nous sélectionnons un sous-ensemble de données pour faciliter la manipulation du graphe, en choisissant 500 articles . Nous choisissons de stratifier les données par domaine pour conserver la structure du corpus et garantir une représentation équilibrée de chaque domaine dans le graphe. Le choix de produire un graphe de coauteurage est motivée par 2 choses. D'abord, nous avons décider de nous concentrer sur des informations qui sont disponibles uniformément à travers le corpus qui seront les auteurs, les citations et les titres. En effet, ce sont les seules informations disponibles pour tous les articles sans exception. Ensuite, parmi ces trois informations nous avons fait le choix de travailler sur les auteurs car elle permet de produire des relations porteuses de sens (il est assez intuitif de connecter deux articles s'ils partagent un auteurs commun) mais aussi car après plusieurs expérimentations nous avons réalisé que c'était la relation qui permettait de créer le plus de lien au sein du graphe. En effet, si l'adjacence des articles est fabriqué à partir de citations communes ou si l'un cite l'autre le graphe obtenu est extrêmement creux et donc pauvre a analyser.

Un choix important que nous faisons est d'ajouter les classes de domaine comme attributs aux nœuds du graphe, ce qui nous permet de colorer les nœuds en fonction de leur domaine pour une visualisation plus claire de la structure du graphe.

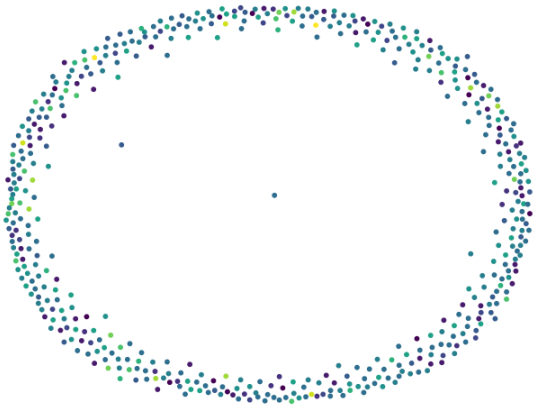

Fig. 2 Graphe obtenu, aucune relation n'est présente dans le dataset.

La création du graphe est assez coûteuse en temps (plusieurs heures sur nos ordinateurs portables) en raison de la forme particulière des données qui entraîne le besoin de naviguer un très grand nombre d'entrée du dataframe, nous n'excluons malgré tout pas

la possibilité d'optimiser grandement le code pour permettre une exécution efficace. Le graphe obtenu pour illustration ne contient malheureusement pas de connexion (Figure 2). Nous imputons cet échec à la malchance dans la sélection des données. En effet, d'autres expérimentations préliminaires nous renseignées ici ont prouvé que cette méthode permet la création d'un nombre relativement élevé de connexion. De plus, d'autres camarades de classe ont produit le graphe à partir des relations de coauteurage également et ont obtenu un graphe non vide. Le temps requis pour la fabrication des liens nous empêche de répéter cette procédure sur notre matériel et nous continuerons donc l'illustration des procédures proposées sur ce graphe.

3 Moteur de recherche

La création d'un moteur de recherche basé sur l'encodage de texte avec CamemBERT est une méthode efficace pour trouver des textes similaires dans un corpus donné. Cette méthode repose sur l'utilisation d'un modèle de langue pré-entraîné spécifiquement conçu pour la langue française, CamemBERT, qui capture les nuances et la structure sémantique des textes.

Méthode de création

Pour créer un moteur de recherche avec CamemBERT, plusieurs étapes sont nécessaires. Tout d'abord, les textes du corpus doivent être encodés en utilisant le modèle CamemBERT pour obtenir des embeddings vectoriels qui représentent le contenu sémantique des textes. Ces embeddings sont ensuite stockés pour une utilisation ultérieure afin d'éviter de les recalculer à chaque requête de recherche, ce qui améliore les performances du moteur de recherche.

Ensuite, la similarité entre le texte d'entrée et chaque texte du corpus est calculée en utilisant la mesure de similarité cosinus. Cette mesure compare les embeddings des textes dans un espace vectoriel et fournit une mesure de la similitude entre les textes. Les textes du corpus sont ensuite triés en fonction de leur similarité avec le texte d'entrée, et les n textes les plus similaires sont renvoyés en tant que résultats de la recherche.

Avantages de la méthode

La méthode de création d'un moteur de recherche avec CamemBERT offre plusieurs avantages. Tout d'abord, en utilisant un modèle spécifiquement conçu pour le français, nous obtenons des résultats précis et pertinents pour la recherche de similarité de texte en français. De plus, l'utilisation de la similarité cosinus comme mesure de similarité est une approche robuste et largement acceptée pour comparer des vecteurs dans un espace vectoriel, ce qui garantit des résultats précis et fiables pour la recherche de similarité de texte.

Possibilités d'amélioration

La limite de cette méthode est qu'elle ne prend pas en compte la structure du graphe. Plusieurs pistes d'améliorations sont alors imaginables. On pourrait par exemple calculer une mesure de centralité des noeuds et l'utiliser comme informations supplémentaires sur le corpus. Cela peut-être réalisé en utilisant la mesure de centralité comme une pondération sur la similarité entre texte d'entrée et titre de l'article.

Une autre amélioration (possiblement complémentaire de la précédente) consisterait à utiliser les résultats du clustering. Cela peut-être réalisé en prenant le noeud central de chaque cluster et de calculer la similarité de son titre avec la requête en entrée. A partir de là une forme de topologie de l'information serait obtenue, en identifiant les clusters d'articles potentiellement correspondant à la requête même si le titre de certains n'est pas proche du texte d'entrée.

4 Clustering

Le clustering de Louvain est une méthode populaire pour la détection de communautés dans les graphes. Cette méthode permet de regrouper les nœuds du graphe en communautés ou clusters, de manière à maximiser la modularité du graphe. La modularité mesure la qualité du partitionnement du graphe en communautés, en évaluant la densité des connexions à l'intérieur des communautés par rapport à celles entre les communautés.

Dans la première étape, chaque nœud du graphe est initialement assigné à sa propre communauté. Ensuite, l'algorithme de Louvain procède à une série d'itérations où il cherche à optimiser la modularité du partitionnement en fusionnant les communautés de manière itérative.

L'algorithme commence par sélectionner aléatoirement un nœud et l'affecte à sa meilleure communauté voisine, ce qui maximise la modularité. Cette étape est répétée pour chaque nœud du graphe jusqu'à ce qu'aucune amélioration de la modularité ne soit possible.

Phase d'optimisation de la modularité

Dans la deuxième étape, l'algorithme de Louvain cherche à optimiser la modularité globale en itérant sur les communautés obtenues lors de la phase précédente. À chaque itération, chaque communauté est considérée comme un seul nœud dans un nouveau graphe pondéré, où les poids des arêtes entre les communautés représentent le nombre de liens entre elles dans le graphe d'origine.

L'algorithme de Louvain applique ensuite la phase de construction de la hiérarchie des communautés sur ce nouveau graphe pondéré, ce qui entraîne la fusion de certaines communautés pour maximiser la modularité globale du graphe.

Avantages et applications

Le clustering de Louvain présente plusieurs avantages, notamment sa rapidité et son efficacité pour détecter des communautés dans de grands graphes. Ce qui en fait une méthode adapté pour les données de Persée. Malheureusement, comme nous l'avons

expliqué plus tôt le graphe d'illustration ne contient aucun lien et donc le clustering de Louvain ne produit aucun cluster (ou plutôt chaque noeud est son propre cluster) ce qui empêche toute analyse expérimentale dans ce rappport. Cependant, le code proposé est fonctionnelle est peut-être utilisé immédiatement avec un graphe complet. Nous avons également mis en place le calcul du score de Rand ajusté qui permet d'évaluer le degré d'adéquation entre clusters prédits et domaines initiaux. L'analyse de ce score sur le graphe complet sera particulièrement importante comme compréhension initiale de la relation entre clusters et domaines. En particulier, nous supposons que le nombre de clusters sera largement supérieurs au nombre de domaines et qu'au sein de chaque domaine existerons différents clusters représentant des directions de recherches spécifiques. On fait également l'hypothèse selon laquelle les clusters seront cloisonnés dans les classes et ne contiendra pas d'articles de plusieurs domaines. Cette hypothèse est particulièrement importante à expérimenter car dans le cas où elle ne serait pas vérifiée elle motiverait la remise en compte des domaines et peut-être la création de nouveaux.

5 Classification

L'implémentation de la classification des articles scientifiques par propagation d'étiquettes. Cette méthode de classification semi-supervisée est particulièrement pertinente dans le domaine de la recherche scientifique et avec ce graphe de co-auteurs.

La méthode de propagation d'étiquettes exploite la structure du graphe pour prédire les étiquettes des échantillons non étiquetés en se basant sur les étiquettes des échantillons voisins dans le graphe. Son intérêt réside dans sa capacité à capturer les relations entre les échantillons dans un graphe, ce qui est crucial dans le domaine de la recherche scientifique où les articles liés par des auteurs communs ont une connection significative.

Cette méthode est particulièrement adaptée à la classification dans des domaines tels que la recherche scientifique en raison de ses caractéristiques uniques. Tout d'abord, elle utilise la structure du graphe pour propager les étiquettes des échantillons étiquetés aux échantillons non étiquetés, ce qui permet de capturer les relations non linéaires entre les échantillons dans un graphe.

De plus, la méthode de propagation d'étiquettes est une technique de classification semi-supervisée, ce qui signifie qu'elle utilise à la fois des données étiquetées et non étiquetées pour effectuer la classification. Cela est particulièrement utile dans les cas où seules quelques étiquettes sont disponibles, ce qui est fréquent dans le domaine de la recherche scientifique où l'étiquetage manuel peut être coûteux et fastidieux.

L'implémentation de la méthode de propagation d'étiquettes commence par la construction d'un graphe où chaque nœud représente un article scientifique et les arêtes représentent les co-auteurs communs entre les articles. Certaines étiquettes sont attribuées à des articles spécifiques pour créer un ensemble de données partiellement étiqueté.

Ensuite, la méthode de Label Propagation est appliquée au graphe pour propager les étiquettes des échantillons étiquetés aux échantillons non étiquetés. Cette méthode est de beaucoup plus simple que les réseaux de neurones types Graphical Convolutional

Networks, mais elle nous semble tout à fait indiqué pour ce travail car les relations de coauteurage dans les articles scientifiques sont porteuse d'une information importante. En effet, deux articles partageant un auteur proviendront dans la très large majorité des cas du même domaine de recherche. De plus, si l'on prend en compte une pondération des liens entre articles par le nombre de co auteurs on pourra évaluer un niveau de confiance dans la propagation du label, un nombre de co auteurs élevé sera un fort indicateur de l'appartenance à leur domaine de recherche.

A nouveau à cause du fait que notre graphe ne contienne aucune arête l'analyse des résultats est vide de sens. Cependant, notre code produit un rapport de classification qui informe sur la qualité de la prédiction pour chaque classe séparément et sur l'ensemble du graphe. Mais aussi une matrice de confusion pour une évaluation rapide des résultats, la matrice de confusion de la propagation de label pour le graphe d'illustration est en Figure 3, du fait de l'absence de liens toutes les étiquettes sont prédites comme la classe majoritaire *Religion théologie*.

Nous croyons que cette méthode de classification a toute légitimité pour le corpus Persée car elle extrêmement flexible et rapide d'usage. Mais surtout car la propagation de label dans un réseau d'articles par coauteurage correspond à une prise en compte de la structure, justifié et explicable ce qui est un avantage bienvenu pour la critique des résultats.

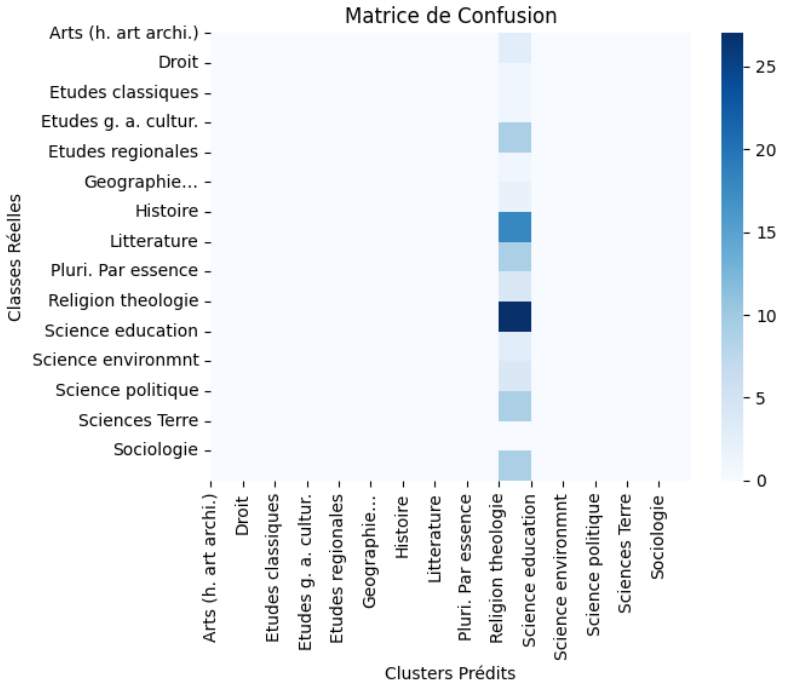

Fig. 3 Matrice de confusion

6 Conclusion

En conclusion, notre travail a exploré différentes méthodes pour analyser les relations entre les articles académiques à partir du corpus de données du portail Persée. Nous avons commencé par la construction d'un graphe de coauteurs pour représenter les liens entre les articles, en mettant en évidence les collaborations et les domaines de recherche. Malgré l'absence de liens dans le graphe d'illustration, nous avons présenté les étapes de sa création et souligné l'importance de cette approche pour comprendre la structure du réseau académique.

Ensuite, nous avons proposé un moteur de recherche basé sur l'encodage de texte avec CamemBERT, offrant une méthode efficace pour explorer et naviguer dans le corpus en recherchant des textes similaires. Bien que cette méthode ne prenne pas en compte la structure du graphe, nous avons discuté des possibilités d'amélioration en intégrant les résultats du clustering et en utilisant la mesure de centralité des nœuds.

Nous avons également exploré le clustering de Louvain comme méthode pour détecter les communautés dans le graphe, soulignant son efficacité pour identifier les regroupements d'articles étroitement liés. Bien que le clustering n'ait pas pu être réalisé sur le graphe d'illustration, nous avons expliqué sa méthode et discuté des pistes pour une analyse plus approfondie en particulier au regard des domaines fournis.

Enfin, nous avons présenté la classification des articles scientifiques par propagation d'étiquettes comme une méthode prometteuse pour prédire les domaines de recherche des articles non étiquetés. Malgré les limitations de notre graphe d'illustration, nous avons souligné la pertinence de cette approche pour le corpus Persée en raison de sa capacité à capturer les relations de coauteurage et à prendre en compte la structure du réseau.

En résumé, notre travail offre un aperçu complet des différentes approches pour analyser les relations entre les articles académiques, en mettant en évidence les avantages et les limitations de chaque méthode. Bien que notre étude ait été entravée par des contraintes techniques, nous croyons que les méthodes proposées ont le potentiel de fournir des perspectives pertinentes pour la compréhension du réseau académique.