Network Analysis for Information Retrieval

Jules Odje

M2 MIASHS, Université Lyon 2, Laboratoire ERIC

Résumé Ce projet présente une analyse complète d'un corpus de publications scientifiques, exploitant à la fois l'information textuelle et la structure relationnelle des documents. Nous avons développé cinq fonctionnalités principales : (1) acquisition et statistiques du corpus, (2) modélisation et analyse de graphe, (3) moteur de recherche hybride, (4) clustering automatique, et (5) classification supervisée. Notre corpus de 40 596 documents scientifiques révèle une structure de réseau fragmentée et une distribution thématique déséquilibrée. Nos expérimentations montrent qu'une approche combinant contenu textuel et structure de réseau améliore les performances du moteur de recherche, tandis que la classification supervisée atteint une exactitude de 30,79% avec une régression logistique basée sur le contenu textuel. Ce travail met en évidence les défis et opportunités de l'analyse de réseaux bibliographiques pour l'organisation et la recherche d'information.

1 Introduction

L'analyse de réseaux bibliographiques constitue un domaine de recherche en pleine expansion, à l'intersection de la recherche d'information, de l'analyse de graphes et de l'apprentissage automatique. L'exploitation simultanée du contenu textuel des documents et de leur structure relationnelle offre des perspectives prometteuses pour améliorer l'organisation, la recherche et la recommandation d'information scientifique.

Dans ce projet, nous explorons un corpus de publications scientifiques sous ces deux angles complémentaires. Notre objectif est de développer une solution complète d'analyse permettant de :

— Caractériser statistiquement le corpus et sa structure relationnelle
— Proposer un moteur de recherche exploitant à la fois contenu et structure
— Découvrir automatiquement des groupes thématiques cohérents
— Prédire la catégorie thématique des documents

Cette approche multi-facettes permet d'appréhender les différentes dimensions du corpus et d'évaluer l'apport respectif de l'information textuelle et structurelle dans chaque tâche.

Toutes les analyses et visualisations présentées dans cet article ont été implémentées dans un notebook Jupyter, disponible en complément de ce rapport. Pour visualiser les différents graphiques et tableaux mentionnés, nous invitons le lecteur à consulter directement les sorties générées par le notebook lors de l'exécution du code.

Le reste de cet article est organisé comme suit : la Section 2 présente l'acquisition et l'analyse statistique des données. La Section 3 détaille la modélisation en graphe et l'analyse de sa structure. La Section 4 décrit le moteur de recherche développé. La Section 5 présente l'approche de clustering et son évaluation. La Section 6 analyse les résultats de classification supervisée. Enfin, la Section 7 conclut et propose des perspectives.

2 Acquisition et prétraitement des données

2.1 Description du jeu de données

Le jeu de données analysé est constitué de 40596 articles scientifiques issus principalement du domaine de l'informatique. Chaque article est caractérisé par 9 attributs comprenant des informations textuelles (titre, résumé), structurelles (auteurs, références) et des métadonnées (année, citations, classe). Comme le montre le notebook (entrées [1-5]), le chargement et l'exploration préliminaire des données révèlent la richesse et la complexité de ce corpus.

2.2 Statistiques descriptives

Le corpus contient 40596 documents publiés entre 1960 et 2017, avec une concentration particulière sur la période 2000-2015. Ces documents sont répartis en 8 classes avec une distribution déséquilibrée : la classe 1 représente 24,9% du corpus (10099 documents), tandis que la classe 3 n'en représente que 5,6% (2291 documents), comme illustré dans le notebook (entrée [9]).

Une analyse des valeurs manquantes révèle que 55,6% des résumés (abstracts) sont absents, ce qui pourrait influencer les analyses textuelles ultérieures. De même, 32,9% des références sont manquantes, ce qui pourrait affecter la construction du graphe de citations. Ces analyses sont détaillées dans les entrées [6-8] du notebook.

2.3 Analyse temporelle

La distribution temporelle des publications montre une évolution significative du volume de production scientifique sur la période couverte. Après une croissance lente jusqu'aux années 1990, on observe une accélération remarquable à partir des années 2000, culminant autour de 2010-2012 avec environ 3800 publications annuelles, suivie d'une chute après 2015. Pour visualiser cette évolution, veuillez consulter la sortie graphique générée dans l'entrée [10] du notebook Jupyter associé.

2.4 Analyse des auteurs

Le corpus implique environ 38990 auteurs uniques. L'analyse des auteurs les plus productifs révèle une forte hétérogénéité, avec les auteurs les plus prolifiques

(comme Mario Piattini et John H. L. Hansen) contribuant à plus de 30 publications chacun. Cette distribution asymétrique, typique des réseaux scientifiques, suggère l'existence de quelques "hubs" centraux dans le réseau de collaboration. L'exploration détaillée des patterns d'autorat est visible dans les entrées [11-12] du notebook.

3 Analyse de la structure du corpus

3.1 Modélisation en graphes

Nous avons construit trois types de graphes à partir du corpus, comme détaillé dans les entrées [20-25] du notebook :

- Un graphe de co-autorat où les nœuds sont les auteurs (95726) et les arêtes représentent la co-écriture d'articles (148025).
- Un graphe bipartite reliant les articles (40596) aux auteurs (95726) avec 115757 liens.
- Un graphe basé sur le contenu similaire des documents, obtenu par une approche KNN, comportant 40596 nœuds et 258678 arêtes.

Ces modélisations offrent différentes perspectives sur les relations entre documents et auteurs, avec des densités très faibles (de l'ordre de 10^{-5}), caractéristiques des grands réseaux bibliographiques.

3.2 Propriétés topologiques

L'analyse des propriétés structurelles révèle plusieurs caractéristiques importantes :

Composantes connexes Le graphe de co-autorat comporte 24473 composantes connexes, la plus grande regroupant 8955 auteurs (9,35% du total), comme calculé dans l'entrée [26] du notebook. Cette fragmentation importante suggère que la communauté scientifique étudiée est composée de nombreux groupes isolés, avec un noyau principal relativement limité.

Distribution des degrés La distribution des degrés suit une loi de puissance, caractéristique des réseaux sans échelle. La majorité des auteurs collaborent avec peu de personnes, tandis qu'un petit nombre d'auteurs très connectés servent de hubs dans le réseau. Cette analyse et sa visualisation en échelle log-log sont disponibles dans les entrées [28-29] du notebook.

Centralité Les mesures de centralité (degré, PageRank, intermédiarité) ont permis d'identifier les acteurs clés du réseau, comme implémenté dans les entrées [30-31] du notebook. Les auteurs comme John H. L. Hansen et Bo Xu présentent des valeurs élevées, indiquant leur rôle central dans la diffusion de l'information et la connexion entre différentes communautés.

3.3 Analyse des communautés

L'application de l'algorithme de Louvain a révélé 23125 communautés dans le graphe de co-autorat (entrée [32]). Cette multitude de communautés corrobore l'observation d'un réseau très fragmenté. La visualisation d'un échantillon des 500 nœuds les plus influents (selon PageRank) montre néanmoins une structure claire avec 181 communautés distinctes, disponible dans l'entrée [32] du notebook. Cette organisation modulaire pourrait refléter les différentes sousdisciplines ou équipes de recherche dans le domaine.

3.4 Implications structurelles

La structure de réseau observée révèle plusieurs phénomènes intéressants :

— La forte fragmentation (24473 composantes) suggère un domaine de recherche composé de nombreuses niches spécialisées avec peu d'échanges interdisciplinaires.
— La présence d'auteurs à haute centralité constitue des points d'articulation potentiels entre différentes communautés.
— La distribution en loi de puissance confirme l'hypothèse du "rich-get-richer" dans les collaborations scientifiques, où les chercheurs établis attirent davantage de nouvelles collaborations.

Ces observations découlent des analyses structurelles réalisées dans les entrées [25-32] du notebook.

4 Implémentation du moteur de recherche

4.1 Conception du moteur de recherche

Notre moteur de recherche vise à exploiter à la fois le contenu textuel des documents et la structure du réseau pour améliorer la pertinence des résultats. Nous avons conçu une architecture modulaire intégrant plusieurs approches complémentaires : une recherche basique par TF-IDF, une approche exploitant la structure du graphe, une méthode tirant parti des clusters, et une stratégie combinée fusionnant ces signaux. L'implémentation complète est disponible dans les entrées [40-55] du notebook.

4.2 Méthodologie

Indexation et prétraitement Nous avons indexé le corpus à l'aide de la méthode TF-IDF (Term Frequency-Inverse Document Frequency), en limitant le vocabulaire aux 5000 termes les plus discriminants et en éliminant les mots vides en anglais (entrées [41-43]). Cette représentation vectorielle permet de capturer l'importance relative des termes dans chaque document.

Intégration de la structure du réseau Pour enrichir la recherche textuelle, nous avons implémenté une approche exploitant la structure du graphe de coautorat (entrées [46-48]). Notre algorithme calcule d'abord une liste de documents candidats par similarité textuelle, puis ajuste les scores en fonction des relations structurelles via un mécanisme de propagation de pertinence inspiré de PageRank. Un paramètre α contrôle l'équilibre entre signaux textuel et structurel.

Exploitation des clusters Nous avons également tiré parti du clustering préalablement réalisé pour améliorer les résultats de recherche (entrées [49- 50]). Cette approche repose sur l'hypothèse que des documents appartenant aux mêmes communautés thématiques que les documents les plus pertinents textuellement sont susceptibles d'être également pertinents.

Stratégie combinée Notre approche finale fusionne les signaux issus des trois méthodes précédentes (entrées [51-53]). Cette combinaison permet de bénéficier des avantages complémentaires de chaque approche tout en atténuant leurs limites respectives.

4.3 Évaluation et résultats

Pour évaluer notre moteur de recherche, nous avons défini un ensemble de requêtes représentatives du domaine étudié, notamment "machine learning algorithms", "network analysis social media" et "information retrieval" (entrées [55-58]). Les résultats montrent que la méthode par clusters améliore systématiquement les scores de pertinence de 12% en moyenne par rapport à l'approche TF-IDF de base. L'approche par graphe, bien que prometteuse conceptuellement, a rencontré des limitations techniques liées au calcul du PageRank sur notre graphe de grande taille. Le détail des scores par requête et par méthode est disponible dans les entrées [55-58] du notebook.

Une analyse qualitative des résultats montre que notre moteur est capable de retrouver des documents très pertinents, même lorsque les termes exacts de la requête ne sont pas présents dans le titre du document.

4.4 Interface utilisateur

Nous avons développé une interface interactive permettant à l'utilisateur de :
- Saisir librement sa requête
- Sélectionner la méthode de recherche (TF-IDF, Graphe, Cluster, Combinée)
- Définir le nombre de résultats à afficher

Cette interface, implémentée dans l'entrée [60] du notebook, facilite l'exploration interactive du corpus et permet de comparer empiriquement l'efficacité des différentes approches.

4.5 Limitations et perspectives

Notre implémentation présente certaines limitations :
— L'erreur dans le calcul du PageRank limite l'exploitation effective de la structure du graphe
— L'absence d'une vérité terrain ne permet pas une évaluation quantitative rigoureuse des performances
— Le modèle TF-IDF ne capture pas les relations sémantiques sophistiquées entre termes

Des améliorations futures pourraient inclure :
— L'intégration de modèles de plongements sémantiques plus avancés (word2vec, BERT)
— Une optimisation du calcul de PageRank pour les grands graphes
— Une évaluation plus formelle basée sur des métriques standards de recherche d'information (précision, rappel, nDCG)

5 Analyse par clustering

5.1 Approche de clustering

L'objectif de cette partie est d'identifier des groupes thématiques cohérents au sein du corpus, permettant une organisation plus structurée des publications. Nous avons opté pour une approche en trois étapes : (1) vectorisation et réduction de dimensionnalité, (2) détermination du nombre optimal de clusters, et (3) application de l'algorithme K-means pour obtenir la partition finale. Cette méthodologie est implémentée dans les entrées [65-75] du notebook.

Vectorisation et réduction de dimensionnalité Le clustering nécessite une représentation vectorielle des documents. Nous avons utilisé une approche TFIDF avec 5000 caractéristiques, suivie d'une décomposition en valeurs singulières tronquée (LSA) pour réduire la dimensionnalité à 100 composantes (entrées [65- 67]). Cette réduction capture 18,25% de la variance totale, offrant un compromis entre fidélité de représentation et efficacité computationnelle.

Détermination du nombre optimal de clusters La détermination du nombre optimal de clusters constitue un défi classique en apprentissage non supervisé. Nous avons utilisé le score de silhouette comme métrique d'évaluation, testant systématiquement des valeurs de k allant de 2 à 20. Cette analyse est détaillée dans les entrées [68-70] du notebook, avec un graphique montrant l'évolution du score en fonction de k. Contrairement à ce qu'on observe classiquement, le score de silhouette augmente de façon quasi monotone avec k. Cette tendance suggère que notre corpus pourrait bénéficier d'une granularité encore plus fine. Pour des raisons d'interprétabilité, nous avons retenu $k = 20$ comme valeur optimale (score de silhouette de 0,4793).

5.2 Analyse des clusters obtenus

L'application de K-means avec $k = 20$ a produit une partition du corpus présentant une distribution déséquilibrée. Le cluster le plus important englobe 56% des documents (22 888 publications), tandis que le plus petit n'en contient que 89. Cette analyse est présentée dans les entrées [71-73] du notebook, avec une visualisation graphique qui illustre clairement ce déséquilibre.

Pour caractériser chaque cluster, nous avons extrait les termes les plus discriminants selon leur poids TF-IDF (entrée [74]). Les résultats de cette analyse sont disponibles dans l'entrée [74] du notebook, présentant les mots-clés principaux pour chaque cluster.

5.3 Interprétation thématique

L'analyse des mots-clés permet d'identifier clairement les domaines de recherche représentés dans chaque cluster (entrées [74-75]) :

- **Cluster 0** regroupe les publications sur les algorithmes de clustering euxmêmes
- **Cluster 2** concerne la théorie des graphes et les algorithmes associés
- **Cluster 3** se concentre sur le traitement d'images et la segmentation
- **Clusters 5 et 8** regroupent des publications en français et allemand respectivement
- **Cluster 15** est dédié aux réseaux informatiques et sans fil
- **Cluster 16** concerne l'apprentissage et l'éducation

Cette organisation thématique offre une vue structurée du corpus, facilitant la navigation et la recherche ciblée d'information. Notons toutefois que le cluster 1, qui contient plus de la moitié des documents, présente des termes peu discriminants qui suggèrent une catégorie "générale" ou un artefact de la méthode de clustering.

5.4 Discussion

Le clustering révèle une structure thématique cohérente dans le corpus, avec des domaines clairement identifiables. Cependant, plusieurs observations méritent d'être soulignées :

- La forte asymétrie dans la distribution des documents, avec un cluster dominant, suggère soit une réelle prédominance d'une thématique générale, soit une limitation de l'approche K-means qui tend à créer des clusters de taille similaire.
- L'émergence de clusters linguistiques (français, allemand) indique que la langue influence fortement la représentation vectorielle des documents, parfois au détriment de la thématique.
- La croissance monotone du score de silhouette avec k suggère que le corpus pourrait bénéficier d'une subdivision plus fine, potentiellement avec des techniques de clustering hiérarchique.

Ces observations informeront notre approche pour la classification supervisée (Section 6), notamment en considérant l'appartenance aux clusters comme caractéristique potentielle.

6 Classification supervisée

6.1 Problématique de classification

L'objectif de cette dernière partie est de prédire automatiquement la catégorie thématique des publications à partir de leurs caractéristiques textuelles et/ou structurelles. Le corpus comporte 8 classes avec une distribution déséquilibrée, la classe majoritaire représentant 24,88% des documents et la classe minoritaire seulement 5,64%, ce qui constitue un défi supplémentaire pour les algorithmes de classification. Cette problématique est abordée dans les entrées [80-85] du notebook.

6.2 Approche méthodologique

Nous avons implémenté trois approches complémentaires pour évaluer l'apport respectif du contenu et de la structure des documents :

- **Classification basée sur le contenu textuel** utilisant les abstracts traités comme source d'information (entrées [86-90])
- **Classification basée sur la structure** exploitant les caractéristiques topologiques des nœuds dans le graphe (entrées [91-93])
- **Classification hybride** combinant les deux types de caractéristiques (entrées [94-96])

Pour chaque approche, nous avons testé plusieurs algorithmes de classification, notamment Random Forest, SVM et Régression Logistique, en utilisant une division 70%/30% pour les ensembles d'entraînement et de test.

6.3 Résultats et analyse comparative

Les performances des différentes approches de classification en termes d'exactitude et de F1-score moyen sont détaillées dans les entrées [87-96] du notebook. Contre toute attente, l'approche basée uniquement sur le contenu textuel surpasse l'approche hybride, avec une exactitude maximale de 30,79% obtenue par la régression logistique. Plusieurs facteurs peuvent expliquer ce résultat :

- Les caractéristiques structurelles extraites pourraient être insuffisamment discriminantes pour les catégories thématiques
- Le déséquilibre des classes peut affecter différemment les approches
- La combinaison linéaire de caractéristiques hétérogènes pourrait nécessiter une pondération plus sophistiquée

6.4 Performances par classe

Les performances détaillées par classe pour la meilleure approche (Régression Logistique sur le contenu textuel) ont été analysées dans l'entrée [90] du notebook. Les métriques de précision, rappel et F1-score montrent des variations importantes entre les classes, reflétant l'impact du déséquilibre des données sur les performances du modèle.

6.5 Analyse des caractéristiques déterminantes

Pour comprendre les facteurs influençant la classification, nous avons extrait les termes les plus discriminants dans l'entrée [97] du notebook. Les termes génériques comme "paper" et "based" figurent parmi les plus importants, mais on note également la présence de termes spécifiques à certains domaines comme "image", "learning" et "model". Cette analyse révèle que le vocabulaire technique constitue un signal fort pour distinguer les différentes catégories de publications scientifiques.

6.6 Discussion et limitations

Les performances obtenues, bien que modestes (30,79% d'exactitude pour la meilleure approche), représentent une amélioration significative par rapport à une classification aléatoire (12,5%) dans ce contexte multiclasse déséquilibré.

Plusieurs limitations méritent d'être soulignées :

- **Problèmes techniques** avec l'exploitation du graphe, limitant l'exploration complète de l'approche structurelle
- **Biais en faveur de la classe majoritaire**, avec une tendance à surclassifier dans la classe 1
- **Faible performance sur les classes minoritaires**, particulièrement les classes 3, 7 et 8
- **Absence d'approches spécifiques au déséquilibre de classes** comme le sous-échantillonnage ou la pondération des classes

7 Conclusion et perspectives

Ce projet a permis d'explorer un corpus de publications scientifiques sous différents angles complémentaires, en exploitant à la fois l'information textuelle et la structure relationnelle des documents.

7.1 Synthèse des résultats

L'analyse statistique initiale a révélé un corpus riche de 40 596 documents répartis en 8 catégories thématiques. La modélisation en graphe a mis en évidence une structure fortement fragmentée, avec 24 473 composantes connexes, suggérant un paysage scientifique composé de multiples communautés isolées.

Le moteur de recherche développé a démontré l'intérêt d'une approche hybride, intégrant à la fois le contenu textuel (TF-IDF) et des signaux issus du clustering pour améliorer la pertinence des résultats. Bien que l'exploitation directe du graphe ait rencontré des limitations techniques, l'approche par clustering a permis d'identifier 20 groupes thématiques cohérents, offrant une granularité plus fine que les 8 classes d'origine.

Finalement, la classification supervisée a confirmé la prédominance de l'information textuelle comme signal discriminant pour la catégorisation des publications, avec une exactitude maximale de 30,79% obtenue par la régression logistique. Contre toute attente, l'approche hybride n'a pas surpassé l'approche purement textuelle, suggérant que l'information structurelle, telle qu'elle a été exploitée, n'apporte pas de signal complémentaire significatif pour cette tâche spécifique.

7.2 Limites et perspectives

Ce travail présente plusieurs limitations qui constituent autant de pistes d'amélioration :

- **Représentation du texte** : L'approche TF-IDF, bien qu'efficace, ne capture pas les relations sémantiques subtiles entre termes. L'exploration de modèles de langue pré-entraînés comme BERT pourrait significativement améliorer la représentation textuelle.
- **Exploitation de la structure** : Les difficultés rencontrées avec le graphe de citations ont limité l'exploration complète de cette dimension. Des approches plus robustes d'extraction et d'exploitation des caractéristiques topologiques pourraient révéler un potentiel inexploité.
- **Déséquilibre des données** : Tant dans le clustering que dans la classification, le déséquilibre des classes a posé un défi majeur. Des techniques spécifiques comme SMOTE ou la pondération adaptative pourraient améliorer significativement les performances.
- **Intégration texte-structure** : L'approche hybride par simple juxtaposition de caractéristiques n'a pas donné les résultats escomptés. Des architectures d'apprentissage profond comme les Graph Neural Networks pourraient offrir un cadre plus naturel pour cette intégration.

Au-delà de ces améliorations techniques, ce projet ouvre des perspectives intéressantes pour la compréhension de l'évolution des domaines scientifiques, la détection de tendances émergentes ou encore l'identification de ponts interdisciplinaires, autant de dimensions qui mériteraient d'être explorées dans des travaux futurs.

7.3 Reproductibilité des analyses

Afin de garantir la transparence et la reproductibilité des résultats présentés, l'ensemble des analyses statistiques, graphiques et algorithmes d'apprentissage a été implémenté dans un notebook Jupyter. Ce notebook est structuré selon les cinq fonctionnalités principales de ce projet et contient l'intégralité du code

commenté permettant de reproduire les expérimentations décrites dans cet article. Les visualisations mentionnées dans ce rapport peuvent être générées en exécutant le code correspondant dans le notebook.

Annexe : Structure du notebook

Table 1. Organisation du notebook Jupyter

Entrée	Fonction	Description
1-12	Acquisition des données	Chargement et statistiques descriptives du corpus
20-32	Structure du corpus	Construction et analyse des graphes
40-60	Moteur de recherche	Implémentation des différentes méthodes de recherche
65-75	Clustering	Détermination du nombre optimal de clusters et analyse
80-97	Classification supervisée	Comparaison des approches de classification

Références

1. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of massive datasets (3rd ed.). Cambridge University Press.
2. Newman, M. (2018). Networks : An introduction (2nd ed.). Oxford University Press.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.
4. Hamilton, W. L. (2020). Graph representation learning. Morgan & Claypool Publishers.
5. Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern information retrieval : The concepts and technology behind search (2nd ed.). Addison-Wesley.
6. Langville, A. N., & Meyer, C. D. (2012). Google's PageRank and beyond : The science of search engine rankings. Princeton University Press.
7. Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3-5), 75-174.