Graphes de citation pour la recherche d'information

Hugo LEGUILLIER[1] Imran NAAR[1]
(1) M1 ATAL, Nantes Université - UFR Sciences et Techniques, 44100, France
hugo.leguillier@etu.univ-nantes.fr,
mohamed-imran.naar@etu.univ-nantes.fr

RÉSUMÉ
L'augmentation exponentielle du nombre d'articles scientifiques publiés sur les bibliothèques numériques rend de plus en plus difficile la recherche d'informations pertinentes pour les chercheurs. Les réseaux de citations, qui fournissent des informations complémentaires sur les textes scientifiques, ont été largement étudiés pour leur utilisation en TAL, tels que l'analyse de la structure d'un domaine de recherche, la découverte de nouvelles relations entre les articles et la génération de catégories. Dans ce travail, nous proposons une approche générant des graphes de citations et retrouvant les contextes de citations afin de potentiellement améliorer la génération de catégories d'articles scientifiques. Lorsque le corpus est soumis au système, celui-ci extrait plusieurs informations, notamment l'auteur, le titre, les citations et leurs contextes, afin de générer le réseau de citations,les contextes de citations et la catégorie de chaque article. Cette approche offre une nouvelle perspective pour la recherche d'informations dans le domaine de la classification de texte en utilisant les contextes de citations comme outil d'analyse.

MOTS-CLÉS : citations, graphe de citation, arXiv.org, data set.

1 Introduction

La recherche d'information est un domaine de recherche en pleine expansion, en particulier avec la prolifération d'articles scientifiques disponibles sur les bibliothèques numériques. La quantité d'articles publiés a augmenté de manière exponentielle ces dernières années, les relations entre articles comme les citations ont donc aussi augmenté. Il devient donc de plus en plus difficile pour les chercheurs de trouver les informations pertinentes pour leurs travaux de recherche. Les réseaux de citations ont été largement étudiés dans la littérature scientifique pour leur capacité à fournir des informations complémentaires sur les textes scientifiques. Selon (Garfield, 1972), un réseau de citation est un système complexe de liens qui relie les articles scientifiques à travers leurs références bibliographiques. Ces réseaux peuvent être utilisés pour plusieurs applications en TAL, telles que l'analyse de la structure d'un domaine de recherche donné, la découverte de nouvelles relations entre les articles et la génération de catégories. Selon (Small, 1973), les réseaux de citations peuvent également être utilisés pour la prédiction de la qualité future d'un article scientifique. D'autres travaux ont montré que l'analyse des réseaux de citations peut être utilisée pour la détection de communautés (Fortunato, 2010), et la classification de texte (Chaney & Blei, 2012).

Le problème de la classification de texte préoccupe de nombreux scientifiques avec différentes approches. Parmi celles-ci, l'apprentissage automatique (machine learning) est couramment utilisé, avec de nombreux algorithmes tels que les k plus proches voisins (kNN), Naïve Bayes, les machines à

vecteurs de support (SVM), les arbres de décision et les réseaux de neurones artificiels qui cherchent tous à résoudre le problème (Aggarwal & Zhai, 2012).

Dans ce travail, nous proposons une approche générant des graphes de citations et trouvant les contextes de citations utilisés par la suite pour la construction d'un dataset de travail pour une tâche de classification de catégories d'articles scientifiques. Lorsque le corpus (*.pdf) est soumis au système , celui-ci extrait plusieurs informations telles que l'auteur, le titre, les citations et leurs contextes afin de générer le réseau de citations et la catégorie de chaque article.

2 État de l'art

Dans le contexte du développement exponentiel de l'information numérique, les réseaux de citations et les techniques de fouille de textes jouent un rôle important dans la gestion de l'information et des connaissances, attirant l'attention des chercheurs. Les réseaux de citations sont des graphes qui représentent les relations entre les publications scientifiques. Chaque nœud dans le graphe représente une publication, et les liens entre les nœuds représentent les citations entre les publications. Les réseaux de citations peuvent être utilisés pour des tâches d'analyse de corpus scientifiques, telles que la classification automatique d'articles ou la recommandation d'articles pertinents. Les méthodes d'apprentissage automatique telles que les K plus proches voisins (KNN), les SVM (Support Vector Machines) et les Naïve Bayes classifier peuvent être appliquées sur les caractéristiques des nœuds et des arêtes du réseau de citations pour effectuer ces tâches.

Plusieurs travaux de recherches ont utilisé les réseaux de citations pour l'analyse de corpus scientifiques. Par exemple (Caragea *et al.*, 2014) ont proposé une méthode basée sur les réseaux de citations pour l'extraction de phrases clés décrivant et résumant un article en utilisant des caractéristiques de citation et de contenu. (Huynh *et al.*, 2012) ont proposé un système de recommandation qui suggère des articles aux utilisateurs sur la base du réseau de citations de l'article de départ.

Ces approches ont montré des résultats prometteurs dans l'analyse de corpus scientifiques, mais il reste encore des défis à relever pour améliorer la précision et l'efficacité de celle-ci. Par exemple, certains travaux ont proposé des approches hybrides combinant des méthodes basées sur les réseaux de citation et la fouille de texte pour améliorer la précision de la classification (Caragea *et al.*, 2015). En somme, les réseaux de citations offrent des perspectives intéressantes et perspicaces pour l'analyse de corpus scientifiques.

3 Expérimentations

Dans cette section, nous détaillons notre approche pour la génération de catégories grâce aux graphes de citations et aux contextes de citations. Par soucis de généralisations et pour pouvoir travailler avec le plus grand nombre d'articles possible, nous nous sommes focalisés exclusivement sur les articles arXiv. ArXiv est une plateforme en ligne gratuite et ouverte où des chercheurs peuvent publier des articles scientifiques dans des domaines tels que la physique, les mathématiques, l'informatique et d'autres sciences connexes. En 2020, elle hébergeait plus de 1,7 million d'articles scientifiques dans une grande variété de domaines de recherche. L'utilisation d'arXiv comme source pour notre étude de génération de graphes de citations, nous permet de travailler avec un grand nombre d'articles et de

données de qualité.

3.1 Génération de graphe de citation

Dans cette étude, nous proposons en premier une méthode pour générer des graphes de citations à partir d'un corpus de documents. Cette méthode se compose de quatre étapes.

La première étape consiste à fouiller les données des articles en utilisant un parser Python spécialement conçu à cet effet. Ce parser est capable d'extraire le texte des PDF des articles, d'identifier les auteurs et les années de chaque citation, ainsi que de récupérer les DOI des articles cités commençant par "arXiv" grâce à des expressions régulières. Cette étape est essentielle pour récupérer toutes les informations nécessaires à la création du graphe de citation.

La deuxième étape consiste à récupérer les informations des articles cités en utilisant l'API arXiv. Nous utilisons les DOI des articles cités pour obtenir leur titre, les auteurs et l'abstract, que nous stockons dans un dataset [1].

	DOI object	Title object	Author(s) object	Abstract object
	2203.15827 ..... 8.3% 2107.06955 ..... 8.3% 10 others 83.3%	LinkBERT: Pr... . 8.3% HTLM: Hype... . 8.3% 10 others 83.3%	Michihiro Ya... . 8.3% Armen Aghaj... . 8.3% 10 others 83.3%	Language m... . 8.3% We introduc... . 8.3% 10 others 83.3%
0	2203.15827	LinkBERT: Pretraining...	Michihiro Yasunaga, Jure...	Language model (LM) pretraining...
1	2107.06955	HTLM: Hyper-Text Pre-Training and...	Armen Aghajanyan, Dmytro Okhonko,...	We introduce HTLM, a hyper-te...
2	2108.07258	On the Opportunities an...	Rishi Bommasani, Drew A. Hudson,...	AI is undergoing a paradigm shift wi...
3	2101.00406	CDLM: Cross-Document...	Avi Caciularu, Arman Cohan, Iz...	We introduce a new pretraining...
4	1704.05179	SearchQA: A New Q&A Dataset...	Matthew Dunn, Levent Sagun, Mi...	We publicly release a new large-scale...
5	2007.15779	Domain-Specific Language Model...	Yu Gu, Robert Tinn, Hao Cheng,...	Pretraining large neural language...
6	2110.04541	The Inductive Bias of In-Context...	Yoav Levine, Noam Wies, Daniel...	Pretraining Neural Language Models...
7	1907.11692	RoBERTa: A Robustly Optimiz...	Yinhan Liu, Myle Ott, Naman Goyal...	Language model pretraining has le...
8	2012.14610	UniK-QA: Unified Representations ...	Barlas Oguz, Xilun Chen, Vladimir...	We study open-domain question...
9	1908.08962	Well-Read Students Learn...	Iulia Turc, Ming-Wei Chang, Kenton Le...	Recent developments in...

FIGURE 1 – Dataset obtenu à partir d'un article exemple

La troisième étape consiste à construire le graphe de citation en utilisant la bibliothèque NetworkX. Nous créons un nœud pour chaque article et une arête pour chaque citation. Le graphe est ensuite dessiné en utilisant la bibliothèque Matplotlib [2].

Enfin, la quatrième étape consiste à générer des graphes pour tout un corpus. Nous avons développé une fonction qui prend en entrée un seul article et génère son graphe de citation. Nous avons ensuite créé un script qui boucle à travers tous les fichiers PDF d'un corpus spécifié en entrée, applique la fonction à chaque article et stocke les graphes de citation dans une liste.

Cette méthode de génération de graphes de citation permet également d'extraire si on le souhaite le contexte de la citation, que nous avons défini dans notre cas comme étant la phrase avant et après la citation (Caragea *et al.*, 2014). En effet grâce au dataset obtenu à la deuxième étape de notre méthode

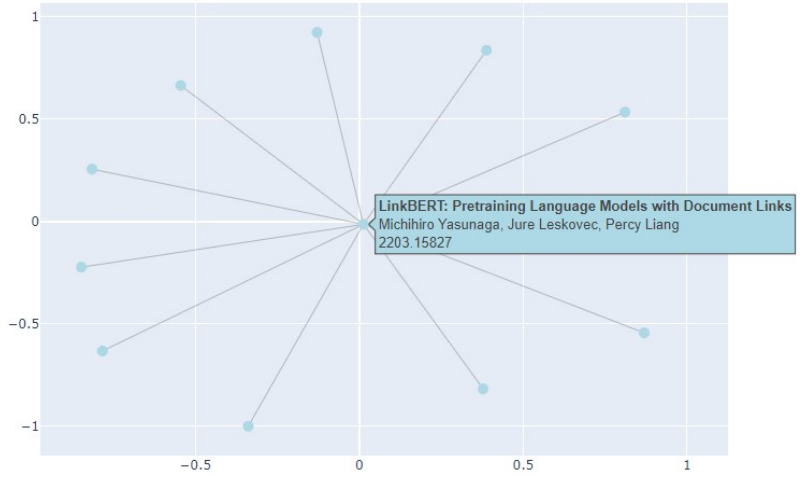

FIGURE 2 – Graphe de citation obtenu à partir d'un article

[1] nous avons créé une fonction qui extrait le nom du premier auteur de l'article à partir de la colonne "Author". Ensuite, la fonction cherche la position de ce nom dans le texte, une fois trouvé elle extrait le contexte à gauche et à droite du nom de l'auteur avec des bornes définies au préalables afin de correspondre à une phrase avant et après la citation.

Author object	context_left object	context_right object
Gu et al. 5.6% Devlin et al. 4.6% 87 others 89.8%	in both domains.LinkBERT consist... 5.6% . LinkBERT is especially effective ... 4.6% 87 others 89.8%	Gu et al., 2020) and sets new state... 5.6% Devlin et al., 2019; Brown et al., 202... 4.6% 87 others 89.8%
Devlin et al.	. LinkBERT is especially effective for multi-hop reasoning and few-shot QA (+5% absolute...	Devlin et al., 2019; Brown et al., 2020), have shown remarkable performance on many natura...
Brown et al.	ally effective for multi-hop reasoning and few-shot QA (+5% absolute improvement on...	Brown et al., 2020), have shown remarkable performance on many natural language...
Bommasani et al.	ur biomedical LinkBERT sets new states of the art on various BioNLP tasks (+7% on BioASQ...	Bommasani et al., 2021). By performing self-supervised learn- ing, such as masked languag...
Devlin et al.	ntroduction Pretrained language models (LMs), like BERT and GPTs (Devlin et al., 2019; Brow...	Devlin et al., 2019), LMs learn to encode various knowl- edge from text corpora and produce...
Petroni et al.	shown remarkable performance on many natural language processing (NLP) tasks, suc...	Petroni et al., 2019; Bosselut et al., 2019; Raffel et al., 2020). Equal senior authorship. 1Available at...
Bosselut et al.	0), have shown remarkable performance on many natural language processing (NLP) task...	Bosselut et al., 2019; Raffel et al., 2020). Equal senior authorship. 1Available at...
Raffel et al.	l language processing (NLP) tasks, such as text classification and question answering,...	Raffel et al., 2020). Equal senior authorship. 1Available at https://github.com/michiyasunaga/...
Liu et al.	us applications, including answering a question "What trees can you see at Tidal Basin?". We...	Liu et al., 2019; Joshi et al., 2020) and do not model links between documents. This can pose...
Joshi et al.	. "Tidal Basin has Japanese cherry trees"), which can be useful for various applications,...	Joshi et al., 2020) and do not model links between documents. This can pose limitations...
Margolis et al.	ontext (linked), besides the existing options of placing a single document (contiguous) o...	Margolis et al., 1999). In this work, we propose LinkBERT , an effective language model...

FIGURE 3 – Dataset des contextes de citations extrait depuis un article exemple

Cet outil peut donc servir à la création ou à l'augmentation du dataset que nous présentons dans la prochaine section.

3.2 Génération de catégorie

La prédiction de catégorie en Traitement Automatique du Langage (TAL) consiste à prédire la catégorie à laquelle appartient un article scientifique, en utilisant différentes caractéristiques telles que le titre, les auteurs, l'abstract, etc. Traditionnellement, ces caractéristiques sont utilisées pour entraîner un modèle de classification qui prédit la catégorie de l'article (Hasan & Ng, 2014), (Das Gollapalli & Caragea, 2014). Par exemple en utilisant l'algorithme de classification SVM on observe une précision moyenne de 91% pour de la génération de catégories d'articles scientifiques (Dien *et al.*, 2019), il reste donc encore quelques pourcent de précision à améliorer. Dans notre approche, nous avons donc également pris en compte le contexte de citation de chaque article en plus des caractéristiques traditionnelles (Caragea *et al.*, 2015). Le but étant d'obtenir un dataset de travail pour la génération de catégorie puis d'éventuellement entraîner un modèle sur celui-ci.

En ajoutant le contexte de citation, nous avons pu capturer des informations supplémentaires sur l'article, telles que le domaine de recherche dans lequel il est cité et la relation entre l'article citant et l'article cité. Nous avons donc créé notre dataset de travail basé sur l'intersection de 2 autres dataset existant unarXive (Saier & Färber, 2020) basé sur toutes les publications disponible sur arXiv.org avant 2021 (aprox. 1 millions d'articles 29.2 millions de citations avec leur contexte et autres métadonnées), et arxiv-abstract (Clement *et al.*, 2019) basé lui aussi sur un grand nombre d'articles d'arXiv.org (approx. 1,5 millions d'articles avec leur métadonnées dont la catégorie de l'article).

Notre dataset de travail contient un total de 144 636 articles, répartis dans 155 catégories différentes. Un exemple d'un article du dataset est montré [4]. Les caractéristiques de chaque article incluent le titre, les auteurs, l'abstract, le DOI, ainsi que le contexte de citation avant et après la citation. Nous avons également effectué une analyse descriptive de notre dataset, qui a révélé que la distribution des catégories était déséquilibrée, avec un nombre important d'articles dans certaines catégories et un nombre beaucoup plus faible dans d'autres [5].


```
{'id': '1412.8192',
'title': 'Generalized complex Monge-Ampere type equations on
closed Hermitian manifolds',
'citation_contexts': "['Later, the author MAINCIT generalized the
result to more general cases by piecewise continuity method. In
local coordinate charts, we may write FORMULA and FORMULA as
FORMULA and FORMULA For convenience, we denote FORMULA Therefore,
equation (REF )']",
'abstract': '  We study generalized complex [...]. Moreover, the
gradient estimate is improved.\n',
'categories': "['math.AP math.DG']"}
```


FIGURE 4 – Un exemple des métadonnées pour un article du dataset

Dans l'objectif de pouvoir entraîner un modèle de classification de catégorie (sur le dataset de travail), nous avons d'abord effectué une étape de tokenisation pour transformer les features des articles en une séquence de tokens. Nous avons utilisé le modèle BART (Lewis *et al.*, 2020) pré-entraîné sur la langue anglaise pour cette étape. Le modèle BART est un modèle d'encodeur-décodeur (seq2seq)

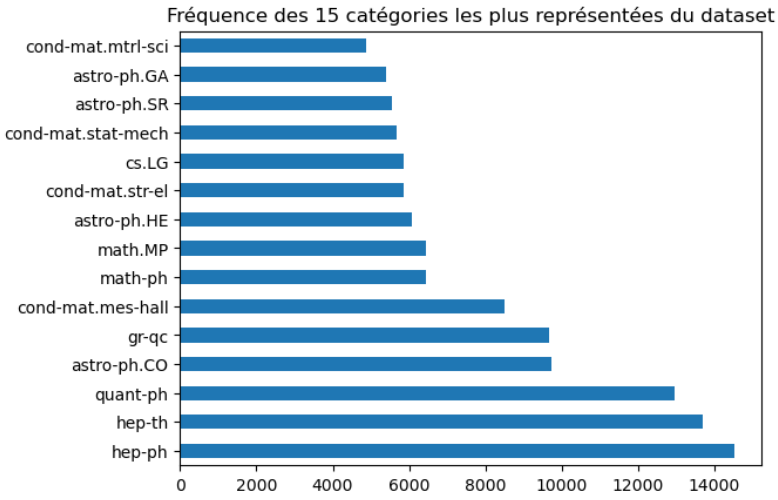

FIGURE 5 – Fréquence des 15 catégories les plus représentées du dataset

basé sur les transformers. BART est particulièrement efficace lorsqu'il est finement réglé pour la génération de texte (par exemple, la résumé, la traduction), mais fonctionne également bien pour les tâches de compréhension (par exemple, la classification de texte, la réponse aux questions). Chaque contexte de citation à été transformé en une séquence de tokens. Nous avons ensuite analysé le nombre de tokens par catégorie. Cette analyse a révélé des déséquilibres, avec certaines catégories ayant un nombre beaucoup plus élevé de tokens que d'autres [6].

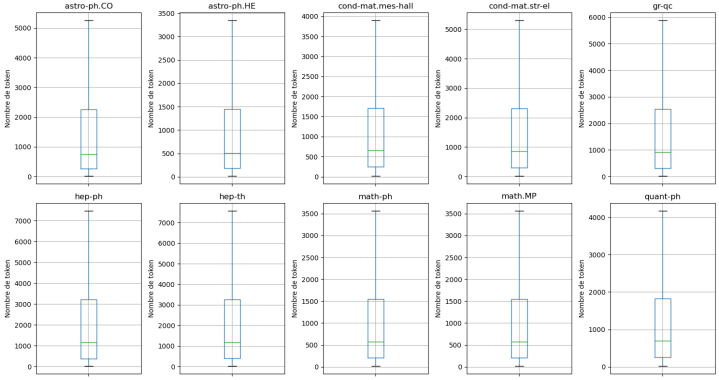

FIGURE 6 – Distribution du nombre de token par catégorie (10 premières catégories)

3.3 Problèmes rencontrés

Dans le cadre de notre étude, nous avons rencontré plusieurs problèmes. Tout d'abord en ce qui concerne la génération de graphes de citation, nous avons dû construire un parser Python capable d'extraire les informations nécessaires des articles. Cette tâche s'est révélée être plus complexe que prévue, en raison de la variété de formats de citations présents dans les articles. En particulier, la gestion des expressions régulières a été une étape très chronophage, car nous n'étions pas familiers avec leur utilisation pour ce type de traitement.

Concernant la création de notre dataset de travail, nous avons du manipuler des flux de données assez important en effet les datasets unarXive et arxiv-abstract ne pouvant être manipulé facilement sur nos machine sans des temps de traitement rébarbatif. Ainsi, nous avons du apprendre à utiliser la bibliothèque Datasets de Hugging Face [1] optimisée spécialement pour le traitement d'important jeu de données en TAL.

4 Limites et perspectives

Notre approche présente cependant certaines limites qui pourraient être améliorées à l'avenir. Tout d'abord, la fenêtre d'extraction des citations utilisée pour extraire les informations pertinentes des articles pourrait être améliorée. Actuellement, nous utilisons une approche simple qui récupère un certain nombre de mots, correspondant à une phrase, avant et après la citation. Cependant, il existe des approches plus sophistiquées pour identifier le texte pertinent pour une citation donnée, telles que celles proposées par (Abu-Jbara & Radev, 2012, 2011). Une étude sur l'influence de la longueur du contexte sur la qualité des mots-clés extraits pourrait également être menée.

Une autre piste d'amélioration possible concerne la visualisation du graphe de citation généré par notre approche. Actuellement, nous utilisons une représentation graphique simple qui montre les articles comme des nœuds et les citations comme des arêtes. Cependant, il existe des approches plus avancées pour la visualisation de graphes qui pourraient être explorées, telles que les représentations en 3D. Ces techniques pourraient permettre une visualisation plus claire et informative du graphe de citation, en mettant en évidence des relations structurelles entre les articles et les citations. Par ailleurs, on pourrait également envisager d'ajouter des fonctionnalités interactives pour permettre une exploration plus approfondie du graphe, telles que la possibilité de zoomer sur des zones spécifiques ou de rechercher des articles ou des citations spécifiques.

En outre, notre dataset de travail présente des déséquilibres en termes de nombre de tokens par catégorie d'articles. Bien que nous ayons pris en compte ces déséquilibres, il faudrait pour entraîner un modèle de classification utiliser des techniques de pondération, il serait aussi intéressant de prendre d'autres mesures pour gérer ce déséquilibre. Par exemple, on pourrait envisager de collecter davantage d'articles dans les catégories moins représentées pour équilibrer le nombre de tokens par catégorie.

Enfin il est important de noter que malgré les efforts pour la création d'un dataset de travail à partir de notre algorithme de génération de graphes de citations, nous n'avons pas pu entraîner de modèle de classification sur ce dataset en raison de contraintes de temps. Par conséquent, il serait essentiel de réaliser une évaluation rigoureuse d'un tel modèle si on venait à l'entraîner. En particulier, il faudrait s'assurer que le modèle est capable de généraliser à des données qui n'ont pas été vues

1. https://huggingface.co/docs/datasets/index

pendant l'entraînement, et vérifier que les performances sont cohérentes avec les attentes en termes de précision, de rappel. Une autre étape importante serait de comparer les performances de notre modèle avec celles d'autres approches existantes pour la prédiction de catégorie dans les articles scientifiques, en utilisant des métriques standardisées pour la comparaison. Cela permettrait de mieux évaluer les avantages et les limites de notre approche en termes de performance et de qualité de la prédiction de catégorie.

5 Conclusion

En conclusion, notre travail a permis de développer un outil de génération de graphes de citations pour la recherche d'informations scientifiques, ainsi qu'un dataset de travail pour une tâche de classification de catégories d'articles scientifiques. Nous avons également exploré l'utilisation du contexte de citation pour améliorer la performance de la prédiction de catégorie. En outre, nous avons discuté des limites de notre approche et proposé des pistes pour l'amélioration de l'extraction de citation, la gestion des déséquilibres de données et la visualisation du graphe de citation. Bien que nous n'ayons pas pu entraîner de modèle sur notre dataset de travail, nous avons souligné l'importance d'une évaluation rigoureuse de tout modèle potentiellement entraîné.

5.1 Code

L'ensemble de nos programmes et nos jeux de données sont disponibles sur notre dépôt Github à l'adresse : https://github.com/MHoubre/TER_citation_graph.

Remerciements

Nous tenons à remercier nos encadrants durant ce TER, Maël Houbre et Florian Boudin qui nous ont orienté et soutenu dans ce projet de recherche.

Références

ABU-JBARA A. & RADEV D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 500–509, Portland, Oregon, USA : Association for Computational Linguistics.

ABU-JBARA A. & RADEV D. (2012). Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 80–90, Montréal, Canada : Association for Computational Linguistics.

AGGARWAL C. C. & ZHAI C. (2012). A survey of text classification algorithms. In *Mining Text Data*. DOI : 10.1007/978-1-4614-3223-4_6.

CARAGEA C., BULGAROV F. & MIHALCEA R. (2015). Co-training for topic classification of scholarly data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2357–2366, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1283.

CARAGEA C., BULGAROV F. A., GODEA A. & DAS GOLLAPALLI S. (2014). Citation-enhanced keyphrase extraction from research papers : A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1435–1446, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1150.

CHANEY A. J.-B. & BLEI D. M. (2012). Visualizing topic models. *Proceedings of the International AAAI Conference on Web and Social Media*. DOI : 10.1609/icwsm.v6i1.14321.

CLEMENT C. B., BIERBAUM M., O'KEEFFE K. P. & ALEMI A. A. (2019). On the use of arxiv as a dataset. DOI : 10.48550/arXiv.1905.00075.

DAS GOLLAPALLI S. & CARAGEA C. (2014). Extracting keyphrases from research papers using citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**(1). DOI : 10.1609/aaai.v28i1.8946.

DIEN T. T., LOC B. H. & THAI-NGHE N. (2019). Article classification using natural language processing and machine learning. *2019 International Conference on Advanced Computing and Applications (ACOMP)*, p. 78–84.

FORTUNATO S. (2010). Community detection in graphs. *Physics Reports*, **486**(3-5), 75–174. DOI : 10.1016/j.physrep.2009.11.002.

GARFIELD E. (1972). Citation analysis as a tool in journal evaluation. *Science*, **178**(4060), 471–479. DOI : 10.1126/science.178.4060.471.

HASAN K. S. & NG V. (2014). Automatic keyphrase extraction : A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1262–1273, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-1119.

HUYNH T., HOANG K., DO L., TRAN H., LUONG H. P. & GAUCH S. (2012). Scientific publication recommendations based on collaborative citation networks. *2012 International Conference on Collaboration Technologies and Systems (CTS)*, p. 316–321. DOI : 10.1109/CTS.2012.6261069.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

SAIER T. & FÄRBER M. (2020). unarXive : A Large Scholarly Data Set with Publications' Full-Text, Annotated In-Text Citations, and Links to Metadata. *Scientometrics*, **125**(3), 3085–3108. DOI : 10.1007/s11192-020-03382-z.

SMALL H. G. (1973). Co-citation in the scientific literature : A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.*, **24**, 265–269.