EXPLORATION DE TECHNIQUES DE FOUILLE DE TEXTES POUR LA CRÉATION D'ONTOLOGIES AFIN D'ASSISTER LA SYNTHÈSE DES CONNAISSANCES SCIENTIFIQUES

Résumé
En raison de leur capacité à formaliser les connaissances d'un domaine en une spécification explicite interprétable par ordinateur (Gruber, 1995), le modèle de représentation des ontologies s'avère idéal pour assister la conduite de méthodes de synthèse des connaissances telles que les revues systématiques de la littérature (Sahlab et al., 2022). Considérant l'accroissement rapide et continu de la production savante, automatiser l'élaboration de ces produits de connaissance s'avère indispensable (Al-Aswadi et al., 2020). Cette présentation fait état des premiers jalons d'une recherche en cours visant à évaluer l'apport de méthodes d'apprentissage d'ontologies pour assister les revues de synthèses de la littérature.

1. Introduction

Les ontologies sont des systèmes d'organisation des connaissances permettant de modéliser les concepts d'un domaine et leurs relations (Gruber, 1995 ; Zacklad, 2010). Puisqu'elles formalisent les connaissances dans un langage interprétable par ordinateur, celles-ci sont utilisées dans des secteurs variés, notamment le développement de moteurs de recherche (M. N. Asim et al., 2019), de systèmes d'aide à la décision (Dissanayake et al., 2020) ou encore celui de systèmes de recommandation (Alsobhi et Amare, 2022).

Récemment, le modèle de description des ontologies a également été proposé pour assister la conduite de synthèses des connaissances telles que les revues systématiques de la littérature, puisqu'il offre les mécanismes de généralisation d'intérêt pour leur réalisation (Ali et Gravino, 2018 ; Sahlab et al., 2022). Parallèlement, considérant l'accroissement rapide de la production savante, automatiser la conduite de ces revues de synthèse s'avère indispensable. Combiner les avancées en fouille de textes aux mécanismes de représentation des ontologies pour assister la réalisation de synthèses des connaissances présente donc une perspective de recherche prometteuse répondant à un besoin d'actualité en sciences de l'information.

Apprentissage et population d'ontologies à partir de textes

La construction d'ontologies à partir de textes peut être désignée par le terme d'« apprentissage d'ontologies » (*ontology learning*). Cette démarche consiste à extraire un modèle conceptuel formel à partir de données textuelles non structurées. Plus spécifiquement, il s'agit d'appliquer diverses méthodes de traitement automatique du langage et d'apprentissage automatique afin d'extraire les principaux concepts d'un domaine et leurs relations (Cimiano, 2006 ; Cimiano et al., 2009). Ce processus peut être perçu comme une tâche de rétro-ingénierie visant à restituer une conceptualisation verbalisée dans un ensemble de textes en procédant à l'extraction des termes représentatifs du domaine, au regroupement de leurs synonymes ainsi qu'à l'extraction de relations hiérarchiques et associatives entre ceux-ci (Cimiano, 2006). En contrepartie, le processus de population d'ontologie (*ontology population*) consiste à alimenter une ontologie existante avec des instances associées aux classes conceptuelles du modèle, et ce à partir de nouvelles sources d'information (Cimiano, 2006). Concrètement, cela consiste à appliquer des outils d'extraction d'information à des documents textuels afin de détecter les réalisations des entités du modèle dans

le corps des textes (Lubani et al., 2019 ; Petasis et al., 2011). Pour ce faire, diverses méthodes proposent de paramétrer un système de reconnaissance d'entités nommées (*named-entity recognition [NER]*) afin de détecter les classes d'intérêt, notamment en entraînant des algorithmes de classification automatique à partir de corpus annotés (Lubani et al., 2019).

Les principales approches proposées pour l'apprentissage et la population d'ontologies à partir de textes reposent sur des mécanismes d'apprentissage hiérarchique (Al-Arfaj et Al-Salman, 2015), d'inférence à base de règles (Al-Aswadi et al., 2020) ou encore sur l'emploi d'algorithmes de regroupement (Mahmood et al., 2023) et de classification automatique (M. Asim et al., 2018). Cependant, plusieurs de ces approches dépendent d'une supervision humaine et ne sont que partiellement automatisées (Zulkipli et al., 2022). Récemment, l'emploi de grands modèles de langue comme BERT (Devlin et al., 2019) ou GPT-4 (OpenAI, 2022) a été suggéré pour automatiser la construction d'ontologies. Malgré les limites de ces modèles, cette avenue s'annonce prometteuse considérant leur capacité à encoder l'information textuelle grâce au volume colossal de leurs données d'apprentissage (Pan et al., 2023).

Revues systématiques des connaissances

La conduite d'une revue systématique mobilise une démarche rigoureuse visant à analyser, synthétiser et résumer de manière critique et exhaustive l'ensemble des travaux de recherche sur une question donnée (Hong et Pluye, 2018). Ce type de synthèse est hautement valorisé puisqu'il offre un accès rapide et fiable aux résultats de recherche de manière à guider leur utilisation dans la pratique professionnelle et à orienter les investissements de recherche futurs (Fortin et Gagnon, 2022). Cette démarche repose sur la sélection d'un ensemble de publications liées à une question de recherche, suivie de l'extraction de données sur la base d'une lecture attentive des publications, puis de la rédaction d'une synthèse critique sous la forme d'un rapport structuré (Hong et Pluye, 2018 ; van Dinter et al., 2021). Diverses méthodologies (e.g. PRISMA, PICO, GRADE, CERQual) ont été proposées pour standardiser la réalisation des revues systématiques (Fortin et Gagnon, 2022) ; néanmoins, leur réalisation s'avère de plus en plus laborieuse considérant l'accroissement massif des publications dans de nombreuses disciplines (van Dinter et al., 2021).

Un certain nombre de travaux ont été menés dans l'objectif d'automatiser certaines étapes de réalisation des revues systématiques (van Dinter et al., 2021). Si plusieurs études se sont intéressées à l'étape de sélection préliminaire (*screening*), peu se sont penchées sur l'extraction des métadonnées ainsi que des composantes méthodologiques à partir du texte intégral des publications (Jonnalagadda et al., 2015 ; van Dinter et al., 2021). Ces étapes sont donc encore à ce jour en majeure partie réalisées manuellement (Li et al., 2023).

2. But et objectifs de recherche

Le but de ce projet est de comparer différentes méthodes d'apprentissage d'ontologies pour assister la conduite de revues systématiques des connaissances. Il vise à remplir les objectifs spécifiques suivants : (1) caractériser les enjeux informationnels et techniques relatifs à la représentation des connaissances scientifiques par le biais d'ontologies ; (2) comparer diverses méthodes permettant de structurer et de populer des ontologies à partir de publications savantes ; (3) identifier les apports et les défis soulevés par ces méthodes pour assister l'étape d'extraction des données liées à la conduite de revues systématiques.

3. Méthodologie

La méthodologie repose sur un devis en fouille de textes orienté vers l'apprentissage et la population automatique d'ontologies. La fouille de textes (*text analytics*) est un domaine de la science des données caractérisé par l'application de techniques d'intelligence artificielle sur de grands corpus textuels afin d'en extraire de nouvelles connaissances (Feldman et Sanger, 2006 ; Jo, 2018 ; Weiss et al., 2005). Le devis se décline selon les étapes de réalisation suivantes.

Une étape préliminaire consiste à identifier des ontologies existantes pouvant être adaptées aux besoins de modélisation du présent projet (Kendall et McGuinness, 2019). Ces ontologies devront permettre de représenter l'ensemble des composantes méthodologiques d'un article de recherche (questions, hypothèses, outils de collecte et d'analyse, résultats, etc.). Quelques exemples incluent l'ontologie *Information Artifact Ontology* (IAO) (Ceusters, 2012), ou plus récemment l'ontologie *Metadata4Ing* (Arndt et al., 2023), permettant de représenter les hypothèses et les conclusions d'articles de recherche ainsi que les méthodes d'expérimentation employées.

Par la suite, la collecte des données est réalisée par la constitution de corpus d'articles scientifiques issus de différents domaines académiques. Ces domaines sont ciblés en fonction de leurs particularités respectives quant à la conduite de revues systématiques. Les corpus sont constitués d'articles rendus disponibles au sein de dépôts d'établissements de recherche (ex. uO Research, TSpace [ABRC, 2023]). Pour les besoins d'évaluation des ontologies générées, ces articles sont sélectionnés sur la base de questions de recherche ayant déjà fait l'objet de revues systématiques.

Une fois les données textuelles nettoyées et uniformisées, celles-ci sont converties sous forme de représentations numériques exploitables par les outils de fouille mobilisés dans les étapes subséquentes. Ces représentations sont utilisées pour paramétrer différents algorithmes d'apprentissage automatique dans le but d'extraire les instances conceptuelles contenues dans les textes et de les associer aux classes ontologiques des modèles préalablement définis.

Finalement, les ontologies ainsi alimentées sont évaluées selon deux approches : (1) au regard de mesures documentées spécifiquement pour la tâche d'apprentissage d'ontologies à partir de textes (ex. analyse de couverture, distance vectorielle entre concepts [Gao et Langlais, 2023]) ; (2) au moyen de mesures de rappel et de précision relatives à des unités de référence, c'est-à-dire des revues systématiques ayant déjà été publiées relativement aux questions de recherche sélectionnées (qui serviront de *Gold standard*) (Jonnalagadda et al., 2015 ; van Dinter et al., 2021).

4. Contributions attendues

Ce projet propose une contribution à l'avancement applicatif de la recherche en fouille de textes pour la représentation des connaissances. Sa réalisation vise une meilleure compréhension des exigences associées à l'enrichissement automatique d'ontologies à partir de textes ainsi que des possibilités qui en découlent pour la conduite de revues systématiques. Ses principales contributions consistent en l'évaluation de différentes méthodes relativement aux applications visées, de même qu'à la formulation de recommandations quant aux outils pour y répondre. Finalement, l'approche ontologique préconisée propose un moyen concret d'appliquer les principes de données FAIR au processus de synthèse des connaissances scientifiques, favorisant ainsi leur partage et leur réutilisation (Wilkinson et al., 2016).

5. Conclusion

Cette présentation fait état d'une recherche en cours visant à évaluer l'apport de méthodes d'apprentissage d'ontologies pour assister les revues de synthèse des connaissances. Dans le cadre de notre présentation, nous présenterons les enjeux du projet ainsi qu'un devis en fouille de textes visant à populer un modèle ontologique représentatif des principales composantes méthodologiques d'un article de recherche. Ce projet constitue un saut considérable d'un point de vue disciplinaire et technique : d'une part, parce qu'il s'insère dans le domaine applicatif de l'intelligence artificielle en sciences de l'information ; d'autre part, parce qu'il s'intéresse aux applications de la fouille de textes pour assister la représentation des connaissances dans le contexte déferlant des IA génératives et des grands modèles de langue.

1499 mots

References

ABRC. (2023). *Dépôts au Canada*. Association des bibliothèques de recherche du Canada. https://www.carl-abrc.ca/fr/faire-avancer-la-recherche/depots-institutionnels/depots-au-canada/

Al-Arfaj, A. et Al-Salman, A. (2015). Ontology Construction from Text: Challenges and Trends, 15-26.

Al-Aswadi, F. N., Chan, H. Y. et Gan, K. H. (2020). Automatic Ontology Construction from Text: a Review from Shallow to Deep Learning Trend. *Artificial Intelligence Review, 53*(6), 3901-3928. https://doi.org/10.1007/s10462-019-09782-9

Ali, A. et Gravino, C. (2018, 1 décembre). *An Ontology-Based Approach to Semi-Automate Systematic Literature Reviews* (p. 09-16). https://doi.org/10.1109/ICOSST.2018.8632205

Alsobhi, A. et Amare, N. (2022). Ontology-Based Relational Product Recommendation System. *Computational and Mathematical Methods in Medicine*, *2022*, 1591044. https://doi.org/10.1155/2022/1591044

Arndt, S., Farnbacher, B., Fuhrmans, M., Hachinger, S., Hickmann, J., Hoppe, N., Horsch, M. T., Iglezakis, D., Karmacharya, A., Lanza, G., Leimer, S., Munke, J., Terzijska, D., Theissen-Lipp, J., Wiljes, C. et Windeck, J. (2023). Metadata4Ing: An Ontology for Describing the Generation of Research Data within a Scientific Activity. https://doi.org/10.5281/ZENODO.5957103

Asim, M. N., Wasim, M., Ghani Khan, M. U., Mahmood, N. et Mahmood, W. (2019). The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval. *IEEE Access, 7*, 21662-21686. https://doi.org/10.1109/ACCESS.2019.2897849

Asim, M., Wasim, M., Khan, M., Mahmood, W. et Abbasi, H. (2018). A Survey of Ontology Learning Techniques and Applications. *Database-The Journal OF Biological Databases AND Curation*. https://doi.org/10.1093/database/bay101

Ceusters, W. (2012). *An Information Artifact Ontology Perspective on Data Collections and Associated Representational Artifacts.* (p. 68-72).

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer.

Cimiano, P., Mädche, A., Staab, S. et Völker, J. (2009). Ontology Learning. Dans S. Staab et R. Studer (dir.), *Handbook on Ontologies* (p. 245-267). Springer. https://doi.org/10.1007/978-3-540-92673-3_11

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019, 24 mai). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dissanayake, P. I., Colicchio, T. K. et Cimino, J. J. (2020). Using Clinical Reasoning Ontologies to Make Smarter Clinical Decision Support Systems: a Systematic Review and Data Synthesis. *Journal of the American Medical Informatics Association*, *27*(1), 159-174. https://doi.org/10.1093/jamia/ocz169

Feldman, R. et Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Fortin, M.-F. et Gagnon, J. (2022). *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives* (4e édition). Chenelière éducation.

Gao, T. et Langlais, P. (2023, 18 juillet). RaTE : a Reproducible Automatic Taxonomy Evaluation by Filling the Gap. arXiv. http://arxiv.org/abs/2307.09706

Gruber, T. R. (1995). Toward Principles for the Design of Ontologies used for Knowledge Sharing? *International Journal of Human-Computer Studies*, *43*(5-6), 907-928. https://doi.org/10.1006/ijhc.1995.1081

Guarino, N. (1997). Understanding, Building and Using Ontologies. *International Journal of Human-Computer Studies*, *46*(2), 293-310. https://doi.org/10.1006/ijhc.1996.0091

Guarino, N., Oberle, D. et Staab, S. (2009). What Is an Ontology? Dans S. Staab et R. Studer (dir.), *Handbook on Ontologies* (p. 1-17). Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Hong, Q. N. et Pluye, P. (2018). Systematic Reviews: A Brief Historical Overview. *Education for Information*, *34*(4), 261-276. https://doi.org/10.3233/EFI-180219

Ji, X., Ritter, A. et Yen, P.-Y. (2017). Using Ontology-Based Semantic Similarity to Facilitate the Article Screening Process for Systematic Reviews. *Journal of biomedical informatics*, *69*, 33-42. https://doi.org/10.1016/j.jbi.2017.03.007

Jo, T. (2018). *Text Mining: Concepts, Implementation, and Big Data Challenge*. Springer Science+Business Media.

Jonnalagadda, S. R., Goyal, P. et Huffman, M. D. (2015). Automating Data Extraction in Systematic Reviews: a Systematic Review. *Systematic Reviews*, *4*(1), 78. https://doi.org/10.1186/s13643-015-0066-7

Kendall, E. F. et McGuinness, D. L. (2019). *Ontology Engineering* (vol. 1-1 online resource). Springer. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5763942

Li, T., Higgins, J. P. T. et Deeks, J. J. (2023). Chapter 5: Collecting data. Dans J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page et V. A. Welch (dir.), *Cochrane Handbook for Systematic Reviews of Interventions* (6.4). Cochrane. www.training.cochrane.org/handbook

Lubani, M., Noah, S. A. M. et Mahmud, R. (2019). Ontology Population: Approaches and Design Aspects. *Journal of Information Science, 45* (4), 502-515. https://doi.org/10.1177/0165551518801819

Mahmood, K., Mokhtar, R., Raza, M. A., Noraziah, A. et Alkazemi, B. (2023). Ecological and Confined Domain Ontology Construction Scheme Using Concept Clustering for Knowledge Management. *Applied Sciences*, *13*(1), 32. https://doi.org/10.3390/app13010032

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. et Wu, X. (2023, 14 juin). Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv. http://arxiv.org/abs/2306.08302

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A. et Zavitsanos, E. (2011). Ontology Population and Enrichment: State of the Art. Dans G. Paliouras, C. D. Spyropoulos et G. Tsatsaronis (dir.), *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap* (p. 134-166). Springer. https://doi.org/10.1007/978-3-642-20795-2_6

Sahlab, N., Kahoul, H., Jazdi, N. et Weyrich, M. (2022). A Knowledge Graph-Based Method for Automating Systematic Literature Reviews. *Procedia Computer Science*, *207*, 2814-2822. https://doi.org/10.1016/j.procs.2022.09.339

van Dinter, R., Tekinerdogan, B. et Catal, C. (2021). Automation of Systematic Literature Reviews: A Systematic Literature Review. *Information and Software Technology*, *136*, 106589. https://doi.org/10.1016/j.infsof.2021.106589

Weiss, S. M., Indurkhya, N., Zhang, T. et Damerau, F. J. (2005). *Text Mining*. Springer. https://doi.org/10.1007/978-0-387-34555-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zacklad, M. (2010). Évaluation des systèmes d'organisation des connaissances. *Les Cahiers du numérique*, *6* (3), 133-166.

Zulkipli, Z., Maskat, R. et Ibrahim Teo, N. H. (2022). A Systematic Literature Review of Automatic Ontology Construction. *Indonesian Journal of Electrical Engineering and Computer Science*, *28*, 878. https://doi.org/10.11591/ijeecs.v28.i2.pp878-889