Forum
Jeunes Chercheuses
Jeunes Chercheurs d'INFORSID
2024

Actes de la 12ème édition

Introduction au Forum Jeunes Chercheuses Jeunes Chercheurs d'INFORSID 2024

Mario CORTES-CORNAX

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
150 Pl. du Torrent, 38400 Saint-Martin-d'Hères, France
Mario.Cortes-Cornax@univ-grenoble-alpes.fr

1. Introduction

Le Forum Jeunes Chercheuses Jeunes Chercheurs (JCJC) d'Inforsid 2024 s'est tenu à Nancy lors du congrès annuel de l'association Inforsid. Cet événement offre aux doctorant(e)s de première ou deuxième année de thèse l'opportunité de présenter leurs travaux à l'ensemble de la communauté française de recherche spécialisée en systèmes d'information. Il s'agit d'un moment important pour la communauté permet de découvrir via les doctorant(e)s, une vue générale des travaux en cours dans les différentes équipes de recherche liés à la communauté Inforsid. La douzième édition du FJCJC n'a pas dérogé à cette tradition, avec treize doctorant(e)s présentant leurs travaux en session plénière, représentant neuf équipes de recherche françaises. Chaque article a bénéficié d'une relecture approfondie de ma part ou de la part d'autres chercheurs du domaine pour garantir la qualité scientifique. Tous les articles ont été acceptés avec de légères modifications ou des révisions de niveau moyen. Tous ont ensuite été relus dans leur version finale, assurant ainsi la rigueur sur la mise en forme et la qualité attendues.

Lors de cette édition, l'essor de l'intelligence artificielle s'est faite fortement ressentir dans les présentations de différents travaux, notamment comme outil pour la gestion, le traitement et l'utilisation de données très variés (ex. publications scientifiques, scènes vidéos, séquences sémantiques temporelles, charge de travail des infirmières, dessins techniques aéronautiques, ...). La communauté se pose de plus la question de l'impact de l'IA dans les méthodes d'apprentissage et l'enseignement.

Les treize travaux présentés reflètent comment la communauté Inforsid est de plus en plus engagée dans la responsabilité des systèmes d'informations en terme sociétal et environnemental. Nous observons aussi une inter-disciplinarité croissante dans les thèmes proposés. Des questions critiques sont abordées telles que la gestion frugale des données, la potentielle fragilité des systèmes de sécurité, l'impact de la collecte massive de données mais aussi son exploitation dans des buts tels que la détection de l'objectification dans les scènes vidéos ou la détection de citations scientifiques erronées.

La présentation des travaux s'est déroulée sous le format « Dragons et Chevaliers », comme lors de l'édition précédente, avec quelques ajustements.

Chaque doctorant a endossé trois rôles : présentateur, dragon, et chevalier. Le présentateur disposait de 5 minutes pour exposer son travail, suivi de 5 minutes durant lesquelles le dragon apportait trois critiques constructives, tandis que chaque chevalier soulignait trois aspects positifs du projet. L'ordre de passage et les rôles ont été définis à l'avance pour assurer une bonne fluidité tout au long de la session. Au fond de la salle, nous avons affiché en format A3 le titre et le résumé de chaque présentation, accompagnés de post-its distribués aux autres chercheurs du public pour noter des questions spécifiques pour chaque travail. La session était organisée en trois blocs de 3 à 4 étudiants. À la fin de chaque bloc, une pause de 15 minutes permettait au public de se lever, de coller leurs post-its avec des questions sur les fiches A3 des étudiants concernés et de discuter avec eux si souhaité. Chaque post-it était signé, permettant aux doctorants de retrouver l'auteur de la question pour des échanges plus approfondis. Ce format a favorisé des échanges interactifs entre doctorants grâce au jeu de rôles, tout en créant une ambiance conviviale. Les échanges se sont également poursuivis après les présentations grâce aux questions laissées sur les post-its. Je remercie chaleureusement nos auteurs pour leurs contributions scientifiques et leur participation engagée aux sessions du Forum JCJC, ainsi que les chercheurs dans le public, qui ont stimulé les étudiants avec leurs questions.

Je tiens à exprimer aussi toute ma gratitude au bureau d'INFORSID pour la confiance qu'il m'a accordée en me confiant l'organisation de ce Forum. Ce fut un réel plaisir de prendre en charge le contenu scientifique, de préparer les sessions en présentiel et de bénéficier des pratiques d'organisation des années passées. Grâce à cet événement, nous avons pu offrir une expérience enrichissante et stimulante à nos jeunes chercheuses et chercheurs, et je suis heureux d'avoir pu contribuer à développement de la communauté INFORSID.

Mario CORTES CORNAX

Responsable du Forum JCJC d'INFORSID 2024

Similarité de séquences sémantiques

Hiba MERAKCHI

Laboratoire LIFAT, Université de Tours
Campus Universitaire de Blois, 3 place Jean Jaurès
41000 Blois, France

hiba.merakchi@univ-tours.fr

MOTS-CLÉS : *fouille de données, séquences sémantiques, trajectoires sémantiques, distance d'édition, logique floue, ontologie*

RÉSUMÉ: *Les séquences sémantiques offrent un potentiel pour comprendre et anticiper les comportements humains. Cette recherche explore deux nouvelles mesures de similarité, la Distance d'Édition Contextuelle (CED) et la Distance de Hamming Temporelle Floue (FTH). Nos travaux visent à améliorer ces méthodes, développer un langage de requêtes pour trouver des séquences similaires, améliorer l'explicabilité des résultats, et valider leur applicabilité dans divers contextes de données.*

ENCADREMENT. *Thomas DEVOGELE, Veronika PERALTA, Cyril DE RUNZ.*

1. Introduction

Les séquences sémantiques désignent des ensembles d'éléments chronologiquement ordonnés, chaque élément ayant une sémantique et potentiellement une durée. Ces éléments, définis et interprétés à l'aide d'une ontologie, peuvent représenter une multitude d'événements, actions et activités humaines. Les séquences sémantiques permettent donc de représenter divers processus et enchaînements d'activités humaines, par exemple, des trajectoires de mobilité, des parcours de vie, des dossiers patients et des flux d'activités diverses (comme des étapes dans les chaînes de productions, des exercices d'e-learning, des requêtes dans un système d'information, ou des chansons d'une playlist).

Par exemple, la séquence de la figure 1 représente la mobilité d'une personne pendant une partie de la journée. La personne est chez elle de 4h à 8h du matin. Elle

prend le bus de 8h à 8h30 pour se rendre au travail. À 16h30, elle finit le travail et marche pendant 15 minutes pour aller à la piscine, où elle nage jusqu'à 18h. Ensuite, elle marche pendant 30 minutes et fait du télétravail de 18h30 à 19h30.

L'analyse de ces séquences sémantiques, comme discutée par (Parent C., 2013), ouvre la voie à une multitude d'applications aussi variées que cruciales pour la société contemporaine. En effet, au-delà de leur simple description, ces séquences offrent un potentiel considérable pour répondre à des défis sociétaux, industriels et individuels. L'analyse des séquences sémantiques offre également la possibilité d'apprendre des modèles de comportement. Cette capacité d'apprentissage permet de créer des groupes homogènes, d'observer des caractéristiques communes et même de recommander des actions ou d'anticiper des intérêts potentiels.

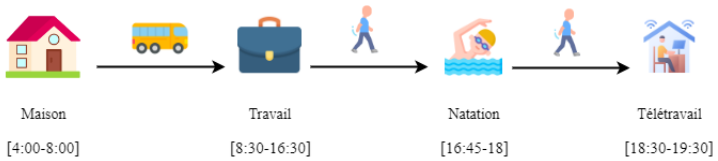


Figure 1: Exemple de séquence avec annotations sémantiques (Moreau C., 2021).

2. Motivations : Mesures de similarité entre séquences

La thèse de (Moreau C., 2021), portant sur l'exploration des séquences de mobilité sémantique, introduit deux nouvelles mesures de similarité de séquences : la Distance d'Édition Contextuelle (CED) et la Distance de Hamming Temporelle Floue (FTH). Ces mesures sont inspirées des méthodes existantes. Elles sont particulièrement adaptées à l'analyse des séquences grâce à l'utilisation d'ontologies et de logique floue, et sont au coeur d'un processus de clustering et d'une méthodologie d'analyse de séquences. En mettant en œuvre cette méthodologie sur des jeux de données synthétiques et réels issus de différents domaines de la mobilité, Clément Moreau a pu améliorer significativement la capacité à interpréter et découvrir des comportements.

CED étend la Distance d'Édition en tenant compte du contexte, qui désigne le contenu sémantique ou une partie de la séquence. Elle repose sur le produit de deux fonctions. La première mesure la similarité sémantique à l'aide d'une ontologie. La deuxième prend en compte l'écart de position.

Ainsi, elle répond à des propriétés mathématiques comme l'homogénéité sémantique, la temporalité des activités, le décalage temporel, la permutation d'activités et la redondance d'activités (Moreau C., 2021).

FTH, quant à elle, est une extension floue de la distance de Hamming, pour les séquences sémantiques-temporelles. Cette mesure améliore la capacité de comparer des séquences de durées égales en introduisant une fenêtre temporelle floue pour gérer

les distorsions temporelles telles que les décalages et les permutations. FTH garantit également d'autres propriétés telles que la temporalité et l'homogénéité sémantique.

3. Actions réalisées : Uniformisation et analyse de sensibilité

Dans le cadre de notre recherche en cours, visant à améliorer les méthodes CED et FTH, nous avons commencé par une étude comparative, en tenant compte de plusieurs axes : la dimension sémantique, contextuelle, floue et temporelle. Nous souhaitons déterminer quelles caractéristiques sont à privilégier lors du choix et du réglage des paramètres de la mesure de similarité.

En premier temps, nous avons travaillé sur l'unification des mesures FTH et CED en intégrant la logique floue dans CED. Plus concrètement, nous utilisons une fonction floue trapézoïdale comme fonction de contexte (plutôt que la fonction exponentielle utilisée par CED). Nous avons également commencé à tester l'impact du choix du support de la fonction floue qui caractérise et contrôle le degré de proximité temporelle entre les activités sur les deux mesures et ultimement sur la comparaison des séquences.

De plus, nous avons testé les performances de CED et FTH sur des séquences avec des activités de même durée et avons constaté que les deux méthodes donnent des résultats similaires, comme attendu. Cependant, FTH pourrait être plus intéressante dans le cas des séquences comportant des activités de durées différentes. Cela reste à explorer.

4. Actions futures : Recherche de sous-séquences, explicabilité et généricité

En nous inspirant des travaux sur les requêtes floues par l'exemple (Moreau A., 2018, Smits G., 2013), nous souhaitons définir un langage de requêtes favorisant la recherche de ***top-k*** séquences similaires à un motif (une sous-séquence). Par exemple, si nous cherchons à trouver toutes les personnes ayant travaillé environ 8 heures et qui ont marché pendant à peu près une heure pour s'y rendre et en revenir (figure 2). Nous désirons retourner des séquences où les horaires peuvent varier, les activités peuvent être segmentées, voire inversées ou remplacées par des activités proches. L'utilisation des distances classiques comme la distance d'édition ne permettrait pas de trouver la séquence de la figure 3. En effet, la marche à pied est remplacée par la trottinette et l'activité "Travail" a été segmentée par l'activité "Repas".

Figure 2: Exemple de sous-séquence recherchée.

La séquence illustrée en figure 3 pourrait être retournée par une telle requête. Elle comprend des déplacements doux (en trottinette, à pied) ainsi que des périodes de travail pas nécessairement de la même durée recherchée ni dans le même ordre.

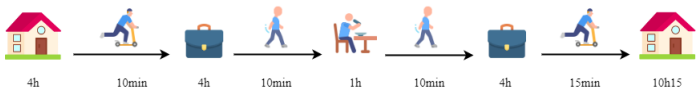


Figure 3: Exemple de séquences retournées.

L'objectif est donc de concevoir un langage permettant de définir des critères de similarité pour identifier des sous-séquences proches en termes de contexte temporel et sémantique, tout en préservant des propriétés de permutation, répétition et homogénéité sémantique. Cela faciliterait l'analyse et la compréhension des habitudes de mobilité d'une personne.

De plus, nous allons nous atteler à développer davantage la partie ***explicabilité*** de nos méthodes, en cherchant à identifier des motifs fréquents ou centraux représentatifs de sous-ensembles (par exemple, des *clusters*) de séquences. Ces avancées nous permettront d'offrir à l'utilisateur une compréhension plus approfondie des résultats et de faciliter leur interprétation dans divers contextes d'application.

Nous envisageons également d'étendre l'utilisation de CED et FTH à différentes sources de données pour valider leur généricité. Nous utiliserons des données d'activités provenant de maisons intelligentes, des Enquêtes Ménages Déplacements (EMD), des listes de lecture de chansons, ainsi que des fichiers de logs pour tester, confirmer et valider l'applicabilité de ces méthodes dans divers contextes.

Bibliographie

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., et al. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, *45*(4), 1–32. ACM New York, NY, USA.

Moreau, C. (2021). Fouille de séquences de mobilité sémantique: sur l'élaboration de mesures pour la comparaison, l'analyse et la découverte de comportements. Thèse de doctorat, Université de Tours.

Moreau, A., Pivert, O., Smits, G. (2018). Fuzzy query by example. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 688–695).

Smits, G., Pivert, O., Girault, T. (2013). Reqflex: fuzzy queries for everyone. *Proceedings of the VLDB Endowment*, *6*(12), 1206–1209. VLDB Endowment.

Détection automatique de citations erronées : jeu de données et méthodes

Qinyue LIU

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
150 Pl. du Torrent, 38400 Saint-Martin-d'Hères, France
qinyue.liu@univ-grenoble-alpes.fr

RESUME. Les citations jouent un rôle important dans la recherche scientifique. Cependant, de nombreuses citations erronées sont identifiées au sein des publications scientifiques. Ces citations erronées, également appelées miscitations, peuvent conduire à une mauvaise interprétation des recherches citées, à une distorsion du message que l'auteur original souhaitait transmettre, et, potentiellement, à des conséquences plus graves. L'objectif de notre recherche est de détecter automatiquement les citations erronées selon le contexte de citation, et de construire un jeu de données contenant différents types de citations. Pour l'instant, nous avons constitué un jeu de données équilibré comprenant à la fois des citations fiables et erronées, issues de publications scientifiques en libre accès. En plus de ce jeu de données, notre étude propose deux méthodes basées sur le Traitement automatique des langues (TAL) pour distinguer automatiquement les citations erronées : une utilisant la similarité cosinus et l'autre, un classifieur de paraphrases, avec des plongements BERT en entrée. Selon nos résultats expérimentaux préliminaires, la similarité cosinus offre les meilleures performances sur notre base de données. Avec nos méthodes et le jeu de données équilibrés, nous nous concentrons pour l'instant sur la faisabilité de la détection automatique des citations fiables et erronées.

MOTS CLES. Citations erronées, Traitement automatique des langues, Article scientifique
ENCADREMENT. Cyril Labbé, Amira Barhoumi

1. États de l'art sur l'étude des citations

Certaines recherches sur les citations se concentrent quantifier la fréquence des citations erronée. Dans une étude sur les citations parue dans OHNS, 50 références aléatoires ont été analysées, révélant des erreurs dans 17% des cas, dont 34% considérées majeures (Armstrong et al., 2018). Il y a également des chercheurs qui ont analysé le contexte des citations pour identifier les tendances dans les sciences biomédicales (Jebari et al., 2021). Une autre recherche a développé une méthode pour étudier les thèmes cachés dans les publications, en analysant les résumés et les

citations d'un article source (Liu and Chen, 2013). D'autres études utilisant des techniques de TAL se sont consacrées à diverses tâches analytiques, telles que l'analyse du sentiment des citations (Liu, 2017) et la classification de la polarité des citations (Bordignon, 2022; Te et al., 2022).

2. Problématique : détecter automatiquement les citations erronées

De nombreuses études antérieures ont utilisé des techniques de TAL pour l'analyse des citations. Cependant, peu de recherches se sont concentrées sur l'évaluation automatique de la fiabilité des citations. À cet égard, notre étude vise à distinguer automatiquement les citations fiables et les citations erronées. Pour ceci, nous constituons un jeu de données en collectant des exemples de différentes types de citations, et testons différentes méthodes de TAL pour classifier automatiquement les citations.

3. Travaux réalisés : constitution du jeu de données et premiers tests de méthodes

Actuellement, nous concentrons sur l'évaluation de la similarité entre le contexte de citation et l'abstract des articles cités. Cette approche suppose qu'une citation fiable présente une grande similarité avec l'abstract. Nous avons donc collecté des citations pour créer notre jeu de données, sur laquelle nous avons ensuite testé nos méthodes.

3.1. Définition de différentes catégories de citations

Dans notre jeu de données, un contexte d'une citation "fiable" reflète correctement le contenu de l'article cité. À l'opposé, une citation "erronée" n'a aucun rapport avec le contenu de l'article cité (Voir Tableau 1).

3.2. Construction du jeu de données

Nous avons identifié deux méthodes pour collecter les données. La première méthode consiste à partir d'un article scientifique, parcourir la liste de ses références (articles cités par cet article). Puis, collecter dans cet article le contexte de citation pour chaque référence, ainsi que l'abstract de chaque référence. De cette manière, nous sommes capables d'évaluer la fiabilité des citations au sein d'un article scientifique.

La deuxième méthode consiste à partir d'un article et de son abstract , examiner tous les autres articles qui ont référencé cet article original. Pour chaque article qui a référencé l'article cité, nous identifions le contexte de citation, et nous le comparons avec l'abstract du article cité. De cette manière, nous pouvons découvrir combien de fois un article a été correctement cité.

Dans notre travail, nous avons appliqué principalement la deuxième méthode pour collecter les données. Les contextes de citation ont été manuellement rassemblés et

annotés à partir de divers articles en libre accès qui citaient 6 articles[1] dans des domaines scientifiques différentes. Au total, 199 citations ont été collectées pour le jeu de données. Pour garantir l'équilibre, 100 de ces citations sont fiables, tandis que 99 sont erronées.

Tableau 1. Exemples d'une citation erronée et d'une citation fiable

Catégorie	Contexte de citation	Abstract d'article cité
Fiable	For instance, other approaches for topic modelling can be tested.	Semantic similarity detection is a fundamental task in natural language understanding. Adding topic information has been useful for previous feature-engineered semantic similarity models, as well as neural models for other tasks. (Peinelt et al., 2020)
Erronée	Eddy covariance devices or lysimeters can be used to determine ET0	Male moths compete to arrive first at a female releasing pheromone. A new study reveals that additional pheromone cues released only by younger females may prompt males to avoid them in favor of older but more fecund females. (Vickers, 2017)

3.3. Méthodes pour classifier les citations

3.3.1. Similarité Cosinus

La similarité cosinus est largement utilisée pour mesurer la similarité entre deux textes sous formes de vecteurs (plongements qui capturent le contenu sémantique). Nous avons utilisé le modèle BERT (Devlin et al., 2018) pour générer les plongements du contexte de citation et de l'abstract d'article cité et comparer leurs similarités.

3.3.2. Classifieur de paraphrase

Nous avons fine-tune un classifieur également basé sur BERT (Devlin et al., 2018) en utilisant le corpus MSRP[2] pour différencier les citations fiables et erronées. La sortie du classifieur est catégorisée soit comme 'paraphrase', soit comme 'non paraphrase'. Dans notre cas, une sortie 'paraphrase' signifie une citation fiable ; Une sortie 'non paraphrase' signifie une citation erronée.

4. Travaux Futurs

Dans cette étude, notre objectif principal est d'évaluer la fiabilité des citations dans les articles scientifiques. Nous avons construit un jeu de données comprenant 199 contextes de citation, proposé et évalué deux méthodes sur nos données. La méthode

[1] Les DOI de 6 articles pour construire notre base de données : 10.1177/1609406919841251, 10.1371/journal.pone.0090972, 10.1007/s10900-017-0360-5, 10.18653/v1/2020.acl-main.630, 10.48550/ARXIV.1706.03762, 10.1016/j.cub.2017.05.0641

[2] https://www.microsoft.com/en-us/download/details.aspx?id=52398

de Similarité Cosinus a donné de meilleurs résultats, atteignant une précision de 93% sur notre jeu de données. La méthode de classifieur a atteint une précision de 87.4%. Pour les travaux futurs, nous envisageons d'ajouter d'autres types de citations erronées et d'agrandir le jeu de données. Nous souhaiterions également tester nos méthodes sur des articles scientifiques où le nombre de citations fiables dépasse celui des citations erronées, plutôt que tester sur notre jeu de données équilibré. De plus, nous envisageons de mener une recherche statistique pour évaluer si la section abstract d'un article cité suffit à justifier le contexte de citation dans l'article citant. Cette analyse aiderait à déterminer s'il est nécessaire d'analyser l'ensemble du article cité ou si se concentrer uniquement sur la section abstract est suffisant.

Remerciements

Nous remercions le projet NanoBubbles, financé par Conseil Européen de la Recherche (ERC) sous forme de subvention Synergie, dans le cadre du programme Horizon 2020 de l'Union Européenne, accord de subvention n° 951393.

Bibliographies

Armstrong, M.F., Conduff, J.H., Fenton, J.E., Coelho, D.H., 2018. Reference Errors in Otolaryngology–Head and Neck Surgery Literature. Otolaryngol.--head neck surg. 159, 249–253. https://doi.org/10.1177/0194599818772521

Bordignon, F., 2022. Critical citations in knowledge construction and citation analysis: from paradox to definition. Scientometrics 127, 959–972. https://doi.org/10.1007/s11192-021-04226-0

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

Jebari, C., Herrera-Viedma, E., Cobo, M.J., 2021. The use of citation context to detect the evolution of research topics: a large-scale analysis. Scientometrics 126, 2971–2989. https://doi.org/10.1007/s11192-020-03858-y

Liu, H., 2017. Sentiment Analysis of Citations Using Word2vec. https://doi.org/10.48550/ARXIV.1704.00177

Liu, S., Chen, C., 2013. The differences between latent topics in abstracts and citation contexts of citing papers. J Am Soc Inf Sci Tec 64, 627–639. https://doi.org/10.1002/asi.22771

Payton, E., Khubchandani, J., Thompson, A., Price, J.H., 2017. Parents' Expectations of High Schools in Firearm Violence Prevention. J Community Health 42, 1118–1126. https://doi.org/10.1007/s10900-017-0360-5

Te, S., Barhoumi, A., Lentschat, M., Bordignon, F., Labbé, C., Portet, F., n.d. Citation Context Classification: Critical vs Non-critical.

Vickers, N.J., 2017. Animal Communication: When I'm Calling You, Will You Answer Too? Current Biology 27, R713–R715. https://doi.org/10.1016/j.cub.2017.05.064

Graphes de connaissances : Construction d'une plateforme versatile pour la gestion du risque cyber

Marin François

Université Paris-Dauphine, PSL, LAMSADE UMR CNRS 7243, DRM UMR CNRS 7088, Place du Maréchal de Lattre de Tassigny, 75775 Paris, France
marin.francois@dauphine.psl.eu

RÉSUMÉ. *Dans cet article, nous décrivons une plateforme Graph-ETL soutenant cinq applications pour la gestion du risque cyber. Cette plateforme permet le traitement de données socio-techniques à travers des graphes de connaissances, élargissant le spectre des données utilisables pour l'apprentissage statistique.*

MOTS-CLÉS: *Sécurité Informatique, Graph-ETL, Apprentissage Statistique*
ENCADREMENT: *Myriam MERAD (LAMSADE), Pierre-Emmanuel ARDUIN (DRM)*

1. Introduction, positionnement et problématique

La « *softwarisation* » des infrastructures d'information et communication (Chatras *et al*., 2016), sous-tendue par l'évolution des processus métier, s'accompagne d'une expansion de leur surface d'attaque. Dans un contexte de croissance des menaces cyber, le recours aux technologies d'aide à la décision dans l'incertain capables de détecter rapidement les menaces nouvelles est inévitable (Nanda *et al*., 2016). Les modèles d'apprentissage automatique (ML) pour la cybersécurité, reposant avant tout sur l'utilisation de données structurées (Martínez Torres *et al*., 2019), répondent partiellement à ce besoin : pour la protection contre les logiciels malveillants par exemple, les méthodes basées sur des architectures neuronales ont dépassé les performances des heuristiques (Aslan, Samet, 2020); pour la détection d'anomalies réseaux, les méthodes non-supervisées sont plus efficcaces q ue l es m odèles d 'inférence asymptomatique (Sharma *et al*., 2018). Cependant, si l'on se positionne d'un point de vue des systèmes d'information (SI) (Tisdale, 2015), c'est dire d'un problème de décision socio-technique dans l'incertain, la capture de l'information et de la connaissance « abstraite » par ces modèles manque pour supporter la décision de manière plus complète. Ces données socio-techniques résultent d'un aggregat de données techniques

(*ex*, journaux d'évènements, octets de signature d'un executable) et de données sociologiques (*ex*, âge, hierarchie, contexte géopolitique, *etc.*)

Les récentes avancées en modélisation des connaissances ont ouvert la porte à un nouveau champ d'application de ML (Liu *et al*., 2022) où, à partir d'une ontologie (graphes sémantiques et graphes de connaissances), il est possible d'intégrer les données socio-techniques non structurées dans le raisonnement d'apprentissage. Ainsi, nous nous positionnons dans la continuité des recherches existantes sur l'utilisation des modèles de ML sur graphes de connaissance pour le traitement des données socio-techniques dans le cadre de la sécurisation des SI. Nous proposons de répondre à la problématique suivante : quelles caractéristiques une plateforme d'apprentissage doit-elle présenter et quelles applications doit-elle supporter pour permettre à l'organisation de tirer profit des données socio-techniques pour l'amélioration de la sécurité du SI ? La construction d'une plateforme pour l'apprentissage automatique pose de nombreux défis (Paleyes *et al*., 2022), cependant la littérature académique se focalise essentiellement sur des problèmes spécifiques, plutôt en lien avec la construction des modèles d'apprentissage. L'objectif de cette recherche est donc double. D'une part la construction d'une architecture système pour l'extraction, transformation et traitement des données socio-techniques, d'autre part la construction d'applications pour l'aide à la décision exploitant ces données socio-techniques.

2. Actions Réalisées et Futures

Nous avons implémenté partiellement cette plateforme (Figure 1). Celle-ci a pour principale fonctionnalité l'aggrégation de plusieurs bases de données (BDD) relationnelles en un graphe de connaissances cyber (CSKG). Par ailleurs, nous avons analysé l'état de l'art en matière d'utilisation des graphes de connaissance pour l'aide à la décision cyber, et avons identifié plusieurs opportunités d'application (François *et al*., 2023), dont deux ont déjà été développées. La plateforme est ainsi composée : (I) Lac de données, (II) BDD Topologie Réseau, (III) BDD Configurations, (IV) BDD Vulnérabilités, (V) BDD IDS, (VI) BDD Organisation, (VII) Serveur de collecte, (VIII) Stockage CSKG, (IX) Graph de connaissance traité, (X) Serveur de pré-traitement, (XI) Stockage modèles, (XII) Visualisation, (XIII) Métriques de risque, (XIV) Supervision.

La première application developpée est basée sur un modèle de Markov Caché (HMM) permettant d'identifier le régime d'intensité de l'activité cyber. En analysant plusieurs sources d'alertes, le modèle infère le régime de perturbation dans lequel se trouve le réseau parmis trois seuil : stable, dégradé, ou critique. Les alertes sont liées à des entités du graphe : organisation(s), utilisateurs, *etc.* Sur cette base, nous sommes en mesure de fournir sept métriques de résilience du réseau. Ces métriques sont precieuses et difficiles à définir pour un réseau étendu sans ce modèle. La seconde permet d'étendre la logique du score d'exploitabilité technique (EPSS) à l'intégralité des composants du SI : données, utilisateurs, organisation(s) (Figure 1 - A). Notre modèle s'appuie sur un algorithme d'étiquettage et trois réseaux de neurones: (1) un

réseau d'apprentissage par représentation inductive (GraphSAGE), (2) un réseau profond pour la régression des scores de risque (DNN), et (3) auto-encodeur pour la classification des actifs par niveau de risque. Nous avons obtenu un score de précision de classification des éléments à risque de 97,6% et sommes en mesure de fournir un niveau de risque associé à chaque actif du SI, en prenant en compte le context local, le renseignement sur les menaces (CTI), et les connaissances « abstraites » liées à ces éléments. Ces éléments sont représentés dans la Figure 1: (1) Nœuds systèmes, (2) informations abstraites associées, (3) Induction.

Nous allons développer trois nouvelles applications sur cette plateforme. La première pour optimiser la sélection des mesures de protection au regard de la topologie du graphe et des informations sur les menaces (Figure 1 - B, (6) Optimisation des mesures de protection). Nous nous appuierons pour cela sur un algorithme de regroupement hiérarchique (Salva, Regainia, 2019). La seconde pour optimiser le choix du positionnement de leurres (« défense active ») (Figure 1 - C). Nous allons pour cela nous appuyer sur l'algorithme d'embranchement optimal (Edmonds *et al.*, 1967). La troisième vise à modifier dynamiquement le positionnement et les caractéristiques des leurres en fonction de signaux variés, via un système d'apprentissage par renforcement (Figure 1 - D). Ces éléments sont représentés en Figure 1: (4) Nœud attaquant, (5) Analyse des chemins d'attaque, (7) Leurre positionné, (8) Leurre ajusté.

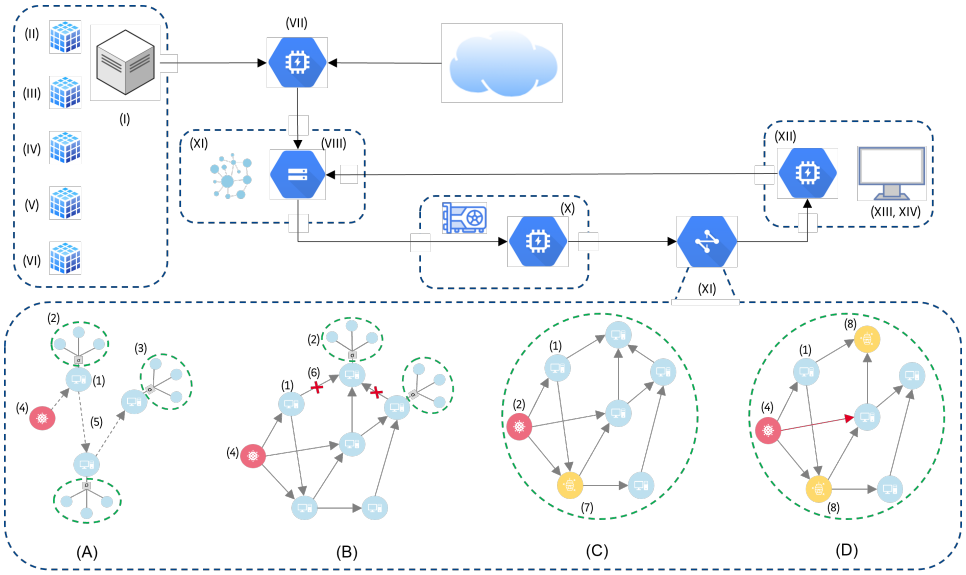


FIGURE 1. *Schéma de l'architecture de notre plateforme et modèles applicatifs*

3. Conclusions

Dans cet article, nous avons présenté une plateforme d'extraction, transformation, chargement des données en graphes pour la sécurité des systèmes d'information. Cette

architecture permet à une organisation de construire sur le long terme une capacité d'apprentissage sur données socio-techniques, qu'il serait difficile d'intégrer sans recourir au modèle de données proposé. Par ailleurs, cette architecture de traitement permet de « recycler » les données, réalisant ainsi des économies d'échelle : les applications existantes et celles proposées, bien qu'ayant des routines de traitement différentes, utilisent ce même modèle de données et sont complémentaires. Par exemple, nous pouvons appliquer la routine d'inférence de régime en tenant compte de la toplogie de l'infrastructure et l'espacement géographique. L'organisation se positionne ainsi en capacité d'améliorer sa prise de décision dans l'incertain, notamment pour l'identification des risques, la protection des actifs, la détection des menaces, et la réponse dynamique aux menaces.

Bibliographie

Aslan Ö. A., Samet R. (2020). A comprehensive review on malware detection approaches. *IEEE access*, vol. 8, p. 6249–6271.

Chatras B., Barthel D., Bertin E., Bertin P., Chemouil P., Guillemin F. *et al*. (2016). Softwarisation et webification, la révolution logicielle des réseaux. In *De nouvelles architectures de communication.*

Edmonds J. *et al*. (1967). Optimum branchings. *Journal of Research of the national Bureau of Standards B*, vol. 71, n° 4, p. 233–240.

François M., Arduin P.-E., Merad M. (2023). Classification of decision support systems for cybersecurity. In *15th mediterranean conference on information systems (mcis) and the 6th middle east & north africa conference on digital information systems (menacis).*

Liu K., Wang F., Ding Z., Liang S., Yu Z., Zhou Y. (2022). A review of knowledge graph application scenarios in cyber security. *arXiv preprint arXiv:2204.04769*.

Martínez Torres J., Iglesias Comesaña C., García-Nieto P. J. (2019). Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, vol. 10, p. 2823–2836.

Nanda S., Zafari F., DeCusatis C., Wedaa E., Yang B. (2016). Predicting network attack patterns in sdn using machine learning approach. In *2016 ieee conference on network function virtualization and software defined networks (nfv-sdn)*, p. 167–172.

Paleyes A., Urma R.-G., Lawrence N. D. (2022). Challenges in deploying machine learning: a survey of case studies. *ACM Computing Surveys*, vol. 55, n° 6, p. 1–29.

Salva S., Regainia L. (2019). A catalogue associating security patterns and attack steps to design secure applications. *Journal of Computer Security*, vol. 27, n° 1, p. 49–74.

Sharma R., Guleria A., Singla R. (2018). An overview of flow-based anomaly detection. *International Journal of Communication Networks and Distributed Systems*, vol. 21, n° 2, p. 220–240.

Tisdale S. M. (2015). Cybersecurity: Challenges from a systems, complexity, knowledge management and business intelligence perspective. *Issues in Information Systems*, vol. 16, n° 3.

Une « révolution éducative » ? Exploration de l'impact de l'intelligence artificielle sur l'apprentissage et les pratiques d'enseignement

Marine CLOUX

Laboratoire ERPI, Université de Lorraine
8, rue Bastien Lepage, 54000 Nancy, France
marine.cloux@univ-lorraine.fr

RESUME. L'introduction de l'intelligence artificielle dans les écoles et les universités est aujourd'hui facilitée grâce à la démocratisation de l'accès aux intelligences artificielles génératives. Cette « révolution », largement annoncée, impacte-t-elle autant l'acte d'apprendre qu'elle le laisse entendre ? L'objectif du travail de thèse présenté ici est d'évaluer l'impact d'outils d'intelligence artificielle sur l'acte d'apprendre et les pratiques d'enseignement. Dans cet article, nous nous attarderons particulièrement sur les questionnements, les verrous scientifiques identifiés et le plan d'action qui sera mis en œuvre pour trouver des réponses.

MOTS-CLES. Intelligence artificielle, apprendre, impact.

ENCADREMENT. Davy MONTICOLO (Professeur des Universités), Raphaël BARY (Maître de Conférences).

1. Contexte

L'intelligence artificielle (IA) n'est pas un sujet nouveau puisque son émergence remonte à 1956 lors d'une conférence à Dartmouth (Le Cun, 2019a). En 1969, Marvin Minsky définissait l'IA comme « *un domaine de recherche qui développe des technologies capables de faire des choses qui exigeraient de l'intelligence si elles étaient faites par des humains* » (Minsky, 1969). Une des approches de l'IA, le Deep Learning (DL), « *n'est devenue largement utilisée qu'au début des années 2010, alimentant ainsi la récente vague d'intérêt portée à l'IA* » (Le Cun, 2019b). En 2022, les avancées du DL ont conduit à rendre l'application ChatGPT plus facile d'accès pour tous. « *Depuis qu'il a fait l'objet d'une large publicité en janvier 2023, ChatGPT a attiré plus de 100 millions d'utilisateurs actifs* » (Bisi et al., 2023) et s'affirme ainsi comme l'une des applications d'IA connaissant une croissance des plus rapides. ChatGPT est une intelligence artificielle générative (IAG) ; un « *terme qui fait référence à des techniques informatiques capables de générer un contenu*

apparemment nouveau et significatif, tel que du texte, des images ou du son, à partir de données d'apprentissage » (Feuerriegel et al., 2023). L'émergence récente des IAG soulève de nombreuses interrogations concernant leur impact ; notamment dans le domaine de l'éducation. On parle ainsi d'« IAEd », c'est-à-dire de l'IA pour l'éducation, comme étant l'adoption des technologies de l'IA à des fins d'apprentissage (Chen et al., 2020).

2. Question de recherche

Si les IAG sont désormais accessibles à tous sans frais d'inscription, les étudiants et les enseignants sont également concernés. Il semble ainsi important de s'interroger sur l'impact de ces outils sur l'apprentissage :

- Avec l'IAG, enseigne-t-on de la même manière ?
- Avec l'IAG, apprend-on de la même manière ?
- Avec l'IAG, les connaissances sont-elles les mêmes ? Lorsqu'un enseignant prépare un cours, il décrit les connaissances à acquérir et les rend compréhensibles pour les apprenants. Les IAG fournissent-elles des connaissances avec les mêmes caractéristiques ?

Notre question de recherche se formule donc de manière plus générale de la façon suivante : quel est l'impact de l'intelligence artificielle générative sur l'acte d'apprendre ?

Holmes et al. (2019) proposent une taxonomie tripartite des IAEd : axées sur les apprenants, les enseignants ou les institutions. Ces catégories désignent des outils assistés par IA spécifiquement conçus pour aider respectivement les apprenants, les enseignants et les institutions. Il existe des IAEd génératives telles que Quillionz[1] ou encore Grammarly[2]. Cependant, il est essentiel de noter, comme l'ont souligné Holmes et Tuomi en 2022, que les apprenants n'utilisent pas forcément et uniquement des IAEd. Avec la récente facilitation de l'accès aux IAG au grand public, les apprenants bénéficient d'outils qui ne sont pas spécifiquement conçus pour les aider dans leur processus d'apprentissage. Des outils comme ChatGPT, Mistral AI et DALL-E apportent de nouvelles perspectives aux apprenants. Nos travaux de recherche s'intéressent ainsi particulièrement à l'impact de ces IAG sur les processus d'apprentissage.

3. Verrous scientifiques

La construction d'une séquence d'apprentissage implique notamment l'identification des compétences et des connaissances à acquérir par les élèves, ainsi que la sélection des activités qui serviront de support à l'enseignement. Chaque séance se construit alors à partir des évaluations formatives et en tenant compte des

1. https://www.quillionz.com/ (construction de quiz et d'évaluations)
2. https://www.grammarly.com/ (partenaire de rédaction)

progrès des élèves. On comprend donc bien que l'usage d'IAG dans l'acte d'apprendre sera spécifique à l'objectif et au contexte d'apprentissage, ainsi qu'au niveau des élèves. De plus, le boom récent de la disponibilité gratuite des IAG qui remonte à 2022 implique que leurs usages par les apprenants et les enseignants restent à identifier. Comment générer des savoirs sur l'impact de l'IAG sur l'acte d'apprendre en général alors que les usages sont spécifiques à un objectif, un contexte et un niveau d'apprentissage ; à fortiori dans un contexte où les usages de l'IAG par les apprenants et les enseignants ne sont pas encore identifiés, est le premier verrou scientifique.

Zheng et al. (2021) ont effectué une méta-analyse sur l'évaluation de l'impact de l'IA sur l'apprentissage. Ils ont notamment constaté que la plus grande proportion des études portait sur l'enseignement supérieur et que les sciences sociales, l'ingénierie ainsi que les sciences technologiques sont les matières les plus fréquemment étudiées. Pour poursuivre l'analyse, nos recherches bibliographiques ont démontré qu'aucune étude n'a déjà défini un modèle d'analyse de l'impact qui soit généralisable à plusieurs niveaux d'apprentissage et à plusieurs sujets d'enseignement. Nos recherches démontrent également que la majorité des études ne s'appuient que sur un nombre restreint de métriques, cognitives essentiellement, sans étudier l'ensemble des catégories d'indicateurs nécessaires pour évaluer l'impact sur l'acte d'apprendre. Lee et al. (2021) expliquent d'ailleurs que la plupart des études s'intéressent à évaluer les livrables de l'apprentissage alors que « *apprendre est un processus plutôt qu'un résultat* ». C'est ainsi que s'écrit le second verrou scientifique de notre recherche : il ne semble pas exister de modèle exhaustif d'évaluation de l'impact de l'IA sur l'acte d'apprendre.

Le manque de recul historique sur l'usage des IAG explique l'absence actuelle de recommandations sur la manière d'enseigner à des apprenants qui utilisent des IAG. Emettre des recommandations nécessiterait de récolter des données issues de nouveaux contextes d'apprentissage utilisant des IAG. Diverses pistes peuvent être explorées, telles que l'utilisation de données d'observation, de questionnaires ou de données issues de l'analyse de l'apprentissage ("learning analytics"). Comment collecter des données issues de nouvelles situations pédagogiques utilisant des IAG afin d'apprendre et recommander des bonnes pratiques est le troisième verrou scientifique.

4. Contributions et méthodologie

Pour répondre à la problématique de recherche et lever les verrous identifiés, la stratégie qui sera employée propose 3 contributions. La première portera sur la construction d'une ontologie pour représenter les concepts de l'acte d'apprendre avec des IAG. Pour cela, des interviews et des observations d'experts en éducation et d'apprenants seront réalisées. Notre seconde contribution sera de construire un modèle d'impact pour évaluer l'IAG en éducation. La consultation d'experts, en IA et en éducation, ainsi que la réalisation d'une revue systématique de la littérature permettront de comprendre comment a déjà été étudié l'impact de l'IA sur l'acte

d'apprendre. Nous chercherons ainsi à identifier quels impacts ont déjà été étudiés et comment ils ont été mesurés. Enfin, nous nous concentrerons sur la conception d'un système de recommandation pour promouvoir des bonnes pratiques. Nous réaliserons des expérimentations et appliquerons notre modèle d'évaluation ; ce qui permettra de collecter des données, de les entraîner et de définir des stratégies de recommandation. Ces préconisations seront partagées pour savoir, d'une part, comment concevoir des IAG qui favorisent l'acte d'apprendre, d'autre part, comment intégrer l'utilisation des IAG dans les situations pédagogiques.

Bibliographie

Bisi, T., Risser, A. H., Clavert, P., Migaud, H., & Dartus, J. (2023). What is the rate of text generated by artificial intelligence over a year of publication in Orthopedics & ; Traumatology : Surgery & ; Research ? Analysis of 425 articles before versus after the launch of ChatGPT in November 2022. *Orthopaedics & Traumatology : Surgery & Research*, 109(8), 103694. https://doi.org/10.1016/j.otsr.2023.103694

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. *Business & Information Systems Engineering*. https://doi.org/10.1007/s12599-023-00834-7

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in Education: Promises and implications for teaching & learning. *The Center for Curriculum Redesign*.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal Of Education*, 57(4), 542-570. https://doi.org/10.1111/ejed.12533

Le Cun, Y. (2019b). 1.1 Deep Learning Hardware : Past, Present, and Future. *2019 IEEE International Solid-State Circuits Conference*. https://doi.org/10.1109/isscc.2019.8662396

Lee, C., Tzeng, J., Huang, N., & Su, Y. (2021). Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors. *DOAJ (DOAJ : Directory Of Open Access Journals)*. https://doaj.org/article/10e4c1e72a0943ad90dde67649593651

Minsky, M. (Ed.). (1969). Semantic information processing. *The MIT Press*.

Quand la machine apprend, Yann Le Cun, Odile Jacob, 2019a, 400 pages, 22,90 euros. (2020). *Cerveau & Psycho*, N° 117(1), 93. https://doi.org/10.3917/cerpsy.117.0093

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2021). The effectiveness of artificial intelligence on learning achievement and learning perception : A meta-analysis. *Interactive Learning Environments*, 31(9), 5650-5664. https://doi.org/10.1080/10494820.2021.2015693

Correspondance Exigences-Normes via les Grands Modèles de Langage

Abdelkarim EL-HAJJAMI

Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne
abdelkarim.el-hajjami@univ-paris1.fr

RÉSUMÉ. L'innovation dans le génie logiciel a conduit à l'adoption croissante de réglementations pour garantir des développements conformes et éthiques. Face à la complexité des textes juridiques et à la fréquente mise à jour des lois, notre recherche explore l'amélioration de la correspondance entre les exigences spécifiées dans les cahiers des charges (SRS) et les normes réglementaires. Cette étude se penche particulièrement sur l'utilisation des Grands Modèles de Langage (LLMs) pour pallier les limitations des méthodes traditionnelles de traitement du langage naturel (NLP) et d'apprentissage machine (ML), qui peinent souvent à capturer les nuances légales complètes. Notre approche proposée vise à tester différentes configurations de LLMs pour déterminer leur efficacité potentielle dans l'amélioration de la correspondance entre les exigences des SRS et les normes réglementaires. Ce travail ambitionne de démontrer comment les LLMs pourraient offrir une méthode plus robuste pour assurer la conformité continue des logiciels aux réglementations en évolution, tout en éclairant les défis associés.

MOTS-CLÉS: Conformité Réglementaire, Ingénierie des Exigences, Grands Modèles de Langage.

ENCADREMENT: Camille Salinesi.

1. Introduction

L'innovation en génie logiciel a transformé notre existence, introduisant de nouveaux produits et services dans presque tous les aspects de notre quotidien. Avec l'intégration croissante des systèmes logiciels dans nos sociétés, les réglementations évoluent continuellement pour garantir que le cycle de vie du développement logiciel soit conforme aux lois et éthique. Assurer la conformité aux lois et réglementations applicables est une préoccupation majeure dans le domaine de l'ingénierie des exigences (RE). La communauté RE a étudié en profondeur le traitement automatisé des textes juridiques à l'aide du traitement du langage naturel (NLP) et de l'apprentissage machine (ML) pour faciliter la définition et l'analyse des exigences légales (Abualhaija et al, 2024). Malgré les progrès significatifs dans le traitement automatisé des textes juridiques, de nombreuses entreprises peinent encore à respecter la législation. Même les grandes entreprises, pourtant bien dotées en ressources juridiques et techniques, ne parviennent pas à se conformer aux lois,

s'exposant ainsi à de lourdes amendes. Un exemple récent et notable est celui de Meta Platforms Ireland Limited (Meta IE), qui a été sanctionnée par une amende de 1,2 milliard d'euros à la suite d'une enquête sur son service Facebook pour violation des exigences de transfert de données personnelles selon le Règlement Général sur la Protection des Données (RGPD) (EDPB, 2023). Un des principaux défis pour implémenter correctement les exigences légales dans les systèmes logiciels est la complexité des structures linguistiques présentes dans les textes juridiques. Cette difficulté souligne le rôle potentiel des Grands Modèles de Langage (LLMs).

2. État de l'Art

De nombreuses études ont exploré l'utilisation du NLP, de ML, et des LLMs pour assurer la conformité des documents réglementés. Torre et al. (2020) ont développé une solution combinant NLP et ML supervisé pour identifier les types d'informations pertinentes au RGPD dans les politiques de confidentialité (PP) et détecter les violations potentielles. Cejas et al. (2023) ont présenté une solution automatisée qui utilise les techniques de NLP pour comparer les représentations au niveau des phrases du texte de l'accord de traitement des données (DPA) avec des représentations prédéfinies des exigences du RGPD. Azeem et Abualhaija (2023) ont examiné dix approches basées sur ML traditionnel, apprentissage profond, et LLMs pour évaluer la complétude des DPA vis-à-vis du RGPD. D'autres études ont mis l'accent sur l'automatisation de la correspondance entre les normes réglementaires et les exigences spécifiques aux produits. Cleland-Huang et al. (2010) ont proposé des méthodes pour générer automatiquement des liens de correspondance, évaluant plusieurs approches dont un modèle probabiliste, un classificateur ML et une méthode de minage web. Guo et al. (2017) ont traité la disparité des termes dans la correspondance en introduisant des techniques d'augmentation de requête : classification, minage web, et méthode basée sur l'ontologie, chacune présentant des compromis distincts en termes de performance et de coûts de mise en œuvre.

3. Problématique

Nous estimons que la correspondance exigences-normes est essentielle dans la conformité réglementaire pour deux raisons principales. Premièrement, l'approche de correspondance établit un lien direct et explicite entre les exigences techniques spécifiées dans un cahier des charges (SRS) et les textes réglementaires. Contrairement aux méthodes qui se concentrent uniquement sur la conformité des documents réglementés tels que les PPs et les DPAs, la correspondance assure une couverture plus complète des dispositions légales, réduisant ainsi le risque de négliger des exigences non mentionnées dans les documents réglementés. Deuxièmement, les lois et réglementations changent fréquemment, et la correspondance joue un rôle important dans la gestion de l'impact de ces changements légaux sur les systèmes logiciels. En maintenant des liens clairs et à jour entre les exigences du système et les dispositions légales, il devient plus facile de comprendre comment les modifications de la loi affectent le système et de déterminer les ajustements nécessaires. Cette capacité à répondre aux changements

légaux est indispensable pour assurer la conformité continue du logiciel. Cependant, l'activité de correspondance fait face à d'importants défis en raison de la disparité terminologique entre les documents SRS et les textes juridiques, exacerbée par la complexité et souvent le langage ambigu utilisé dans les deux domaines. Les documents juridiques sont denses et remplis de terminologie spécialisée, susceptibles de multiples interprétations. De même, les documents SRS peuvent incorporer des descriptions vagues en raison des contributions variées des parties prenantes. Cette double ambiguïté pose un défi significatif pour garantir que les systèmes logiciels se conforment aux réglementations requises. Les approches précédentes en matière de correspondance ont principalement reposé sur la similarité sémantique au niveau des termes, ce qui peut ne pas capturer entièrement le contexte nécessaire pour une compréhension complète. Cette méthode peut échouer à saisir des nuances légales plus subtiles et des implications qui vont au-delà de la simple correspondance de certains mots-clés. Avec l'émergence des LLMs, de nouvelles opportunités se présentent pour pallier ces limitations. Compte tenu de ces considérations, notre recherche sera guidée par les questions suivantes :

– **QR1.** Comment les LLMs améliorent-ils la performance de la correspondance des exigences SRS aux normes réglementaires comparativement aux méthodes traditionnelles de NLP et ML?

– **QR2.** Quelle est la configuration de LLMs la plus efficace pour la correspondance exigences-normes?

– **QR3.** Quels sont les défis et les limitations potentiels de l'utilisation des LLMs dans la correspondance exigences-normes, et comment peuvent-ils être abordés pour améliorer leur efficacité et leur utilisabilité?

4. Approche Proposée

Cette section décrit notre approche structurée en quatre phases pour atteindre nos objectifs de recherche :

1. Phase de préparation : Nous collecterons et préparerons les jeux de données, débutant par le RGPD, et collaborerons avec des industries pour obtenir des documents SRS. Ces documents serviront à créer une matrice de correspondance pour l'entraînement des modèles et l'évaluation quantitative.

2. Phase de développement: Nous allons reproduire des modèles de référence en utilisant les techniques traditionnelles de NLP et ML décrites dans les études antérieures (Cleland-Huang et al., 2010 ; Guo et al., 2017). Parallèlement, diverses configurations de LLMs seront testées pour déterminer la configuration la plus efficace pour une correspondance précise des exigences SRS avec les dispositions légales du RGPD.

3. Phase d'évaluation: L'efficacité des modèles sera mesurée à l'aide de métriques quantitatives (précision, rappel, F-score) et complétée par une analyse qualitative réalisée par des experts pour évaluer la pertinence et l'exhaustivité des correspondances.

4. Phase d'optimisation : Nous affinerons les modèles basés sur les résultats de l'évaluation, en portant une attention particulière à l'usabilité et à l'intégration pratique des modèles dans les flux de travail de conformité.

En conclusion, ce projet de thèse évaluera l'efficacité des LLMs dans l'amélioration de la correspondance entre les exigences SRS et les dispositions légales du RGPD, et identifiera leurs avantages, leurs limites, ainsi que la configuration optimale pour améliorer les pratiques de conformité réglementaire et le succès des projets logiciels.

References

Abualhaija, S., Ceci, M., Briand, L. (2024). Legal Requirements Analysis. arXiv preprint arXiv:2311.13871.

Torre, D., Abualhaija, S., Sabetzadeh, M., Briand, L., Baetens, K., Goes, P., Forastier, S. (2020). An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR. In Proc. 2020 IEEE 28th International Requirements Engineering Conference (RE), Zurich, Switzerland, 136-146.

Cejas, O., Azeem, M., Abualhaija, S., Briand, L. (2023). NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR. IEEE Transactions on Software Engineering, 49(09), 4282-4303.

Azeem, M. I., Abualhaija, S. (2023). A Multi-solution Study on GDPR AI-enabled Completeness Checking of DPAs. arXiv preprint arXiv:2311.13881.

Cleland-Huang, J., Czauderna, A., Gibiec, M., Emenecker, J. (2010). A machine learning approach for tracing regulatory codes to product specific requirements. In Proc. 2010 ACM/IEEE 32nd International Conference on Software Engineering, Cape Town, South Africa, 155-164.

Guo, J., Gibiec, M., Cleland-Huang, J. (2017). Tackling the term-mismatch problem in automated trace retrieval. Empirical Software Engineering, 22, 1103–1142.

European Data Protection Board (EDPB). (2023). '1.2 billion euro fine for Facebook as a result of EDPB binding decision.' https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision.

Une approche pour garantir la conformité des informations entre deux dessins techniques mécaniques

Alexandre MONNIER WEIL

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
Kaizen Solutions, 38330, Montbonnot-Saint-Martin, France
alexandre.monnier@univ-grenoble-alpes.fr

RÉSUMÉ.
Pour moderniser leurs systèmes d'information, les entreprises opérant dans de nombreux secteurs, tels que l'hydraulique et l'aéronautique, se trouvent confrontées à la tâche de numériser leurs plans techniques mécaniques originaux, élaborés à la main, en plans réalisés via des logiciels de Conception Assistée par Ordinateur (CAO). Cependant, en raison du volume important de plans à numériser, c'est une tâche chronophage pour les ingénieurs. De plus, une vérification rigoureuse est nécessaire pour prévenir les coûts élevés associés à la non-conformité des plans de CAO par rapport aux plans originaux. Notre objectif est de simplifier, pour les ingénieurs, la vérification de la conformité des informations entre les plans CAO en comparant les dessins techniques du plan original, préalablement vectorisés, avec les dessins correspondants sur le plan CAO. Ainsi, nous présentons dans cet article la vision de notre approche qui repose sur la définition d'un formalisme permettant de décrire la sémantique des dessins techniques mécaniques. À terme, ce formalisme nous permettra de procéder à une comparaison significative des informations contenues dans les dessins.

MOTS-CLÉS : Système d'information, Dessins techniques, Comparaison, Similarité
ENCADRANTS: Nicolas HILI, Yves LEDRU

1. Contexte

Le passage des plans techniques mécaniques réalisés à la main dans les années 1960 aux plans réalisés sous logiciel de CAO a marqué une évolution majeure pour les entreprises de nombreux secteurs, telles que l'hydraulique, l'aéronautique et le génie civil, cherchant à renforcer leur système d'information (SI). En effet, les plans CAO améliorent considérablement l'accès et l'exploitation des données, ouvrant la voie à des applications telles que la gestion de projet et l'assurance qualité. Les entreprises, pour profiter de cette modernisation, entreprennent la numérisation des plans originaux de pièces mécaniques réalisés à la main en plans CAO. Cependant, en raison du volume important de plans à numériser, cette tâche est très complexe et chronophage pour les ingénieurs. De plus, si les informations du nouveau plan ne sont pas conformes au plan original, cela peut rendre l'assemblage de la pièce impossible.

Cette non-conformité peut provoquer d'importants retards de production et un gaspillage de matériaux, entraînant des coûts élevés. Il est donc nécessaire d'avoir une étape de vérification rigoureuse. Dans cet article, nous présentons la vision d'une approche partiellement implémentée pour comparer les dessins originaux préalablement vectorisés aux versions CAO, afin de faciliter la tâche des ingénieurs. L'objectif est de repérer les dissimilarités, c'est-à-dire les éléments divergents entraînant une non-conformité, afin que les ingénieurs puissent ajuster le plan CAO.

De nombreux travaux sur la numérisation automatique des dessins techniques ont été réalisés, la méthodologie se divisant généralement en trois phases : vectorisation, détection de formes et contextualisation (Moreno-García *et al.*, 2019). Le prototype Celesstin (Vaxiviere, Tombre, 1992) spécialisé dans la numérisation de dessin technique mécanique se distingue comme l'un des plus aboutis. Cependant, le manque de maturité de ces approches empêche leur déploiement dans les entreprises. C'est la raison pour laquelle nous proposons une approche alternative à la numérisation automatique, en considérant que la numérisation est effectuée manuellement par les ingénieurs, tandis que l'étape de vérification est automatisée autant que possible pour aider à identifier le maximum de dissimilarités. Nous partons de l'hypothèse que les plans originaux scannés sont nettoyés algorithmiquement de toutes informations parasites (tâches, pliures) puis vectorisés. Cette supposition nécessite une validation ultérieure basée sur des travaux hors du périmètre actuel de notre étude.

2. Présentation de l'approche

Notre approche se divise en deux volets : d'une part, le calcul d'un taux de similarité, et d'autre part, notre principal objectif, la production d'une description explicite des dissimilarités (cf. Fig 1). Les réseaux de neurones siamois sont efficaces pour calculer un degré de similarité entre deux entrées (Bromley *et al.*, 1993). Pour notre approche, nous prévoyons aussi d'y associer le mécanisme d'auto-attention (Vaswani *et al.*, 2017), efficace dans le traitement du langage, qui serait également adapté ici aux langages vectoriels. Les modèles d'apprentissage profond abstraient automatiquement les problèmes à partir d'exemples représentatifs. Toutefois, leur opacité compliquerait l'extraction de connaissances et l'explication des décisions, rendant difficile la compréhension de l'origine des dissimilarités dans les dessins. Pour ces raisons, nous adoptons également des méthodes d'Intelligence Artificielle (IA) symbolique comme la reconnaissance de plans (Castellanos-Paez *et al.*, 2022) et la distance d'édition de graphe (Riesen, 2015), qui utilisent des connaissances expertes pour définir explicitement des abstractions. Ainsi, nous établissons un formalisme pour décrire les dessins techniques.

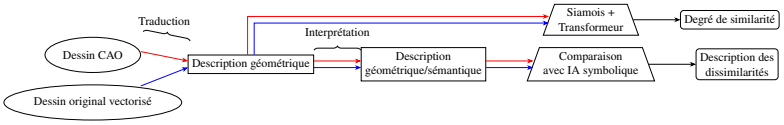
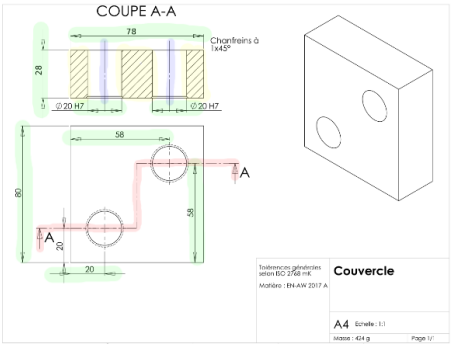


FIGURE 1. *Processus de comparaison des dessins techniques*

Comparer les dessins techniques uniquement d'un point de vue géométrique n'est pas suffisant. Deux dessins sont considérés similaires d'un point de vue sémantique, si toutes les informations présentes dans le dessin original sont également présentes dans sa version CAO, quand bien même, les deux dessins peuvent être différents d'un point de vue géométrique.



Dans cet exemple, le plan CAO présente deux dessins techniques offrant des perspectives différentes d'une section d'une pièce, avec sa représentation en 3D dans le coin supérieur droit. Il inclut des annotations : cotations (vert), lignes de centre (bleu), hachures (jaune) et une ligne de coupe (rouge), qui indique la section de la pièce pour en révéler l'intérieur.

Figure 2, reprise de Cdang, licence CC BY-SA 3.0, https://commons.wikimedia.org/wiki/File:Maquette_plan_cc_pct_couvercle.svg

FIGURE 2. *Exemple de plan CAO*

Il est donc important de pouvoir formaliser la notion de dissimilarité à un niveau conceptuel. Pour cela, nous présentons l'ébauche d'un méta-modèle dont le but est de servir de langage pivot afin d'uniformiser l'information des dessins. Il se divise en deux : une partie géométrique, détaillant la géométrie des dessins, et une partie sémantique (cf. Fig 3) qui a pour but d'organiser des ensembles d'éléments géométriques (lignes, courbes, arcs, etc.) en catégories sémantiques (lignes de contours, zones hachurées, cotations, etc.).

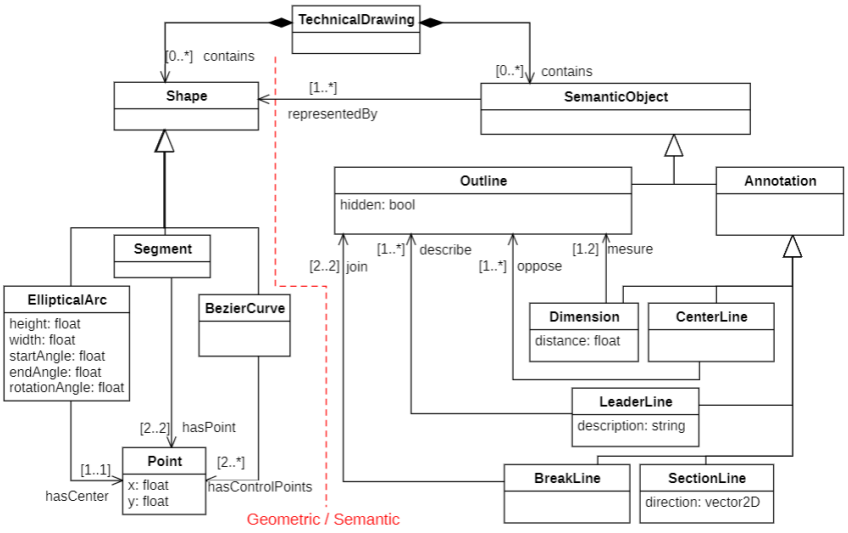


FIGURE 3. *méta-modèle sémantique-géométrique*

La partie géométrique du méta-modèle vise à standardiser la description géométrique, inspirée des graphes de contraintes géométriques (Seff *et al*., 2020), contrairement aux langages vectoriels qui ne permettent pas une représentation uniforme de la géométrie, car plusieurs directives peuvent être utilisées pour représenter une même forme. La partie sémantique attribue des labels à des groupes de formes, alignés sur les conventions des dessins techniques mécaniques. Le méta-modèle, analogue à une grammaire, permet de décrire un dessin. Le langage issu sert de langage pivot dont le rôle est d'abstraire la description géométrique en concepts de haut niveau. Par exemple, la classe *Dimension* spécifie les mesures entre deux contours définis par *Outline*. Des dimensions similaires avec différentes représentations géométriques sont ainsi unifiées, ce qui facilite la comparaison assistée par l'IA symbolique.

3. Conclusion et perspectives

Dans cet article, nous décrivons une méthode en trois phases : traduction, interprétation et comparaison, utilisant un méta-modèle sémantique pour focaliser la comparaison sur les informations essentielles des dessins. Nous visons à développer premièrement l'interprétation en combinant l'apprentissage profond et l'IA symbolique. Pour illustrer comment l'interpréteur pourrait fonctionner, considérons les hachures : si des lignes L sont parallèles $par(l1, l2)$ et proches $proche(l1, l2)$, alors L définit un groupe de hachures $HatchingGroup(L)$. Nous testerons également un réseau de neurones siamois sur le jeu de données SketchGraph (Seff *et al*., 2020) pour pouvoir obtenir un degré de similarité.

Remerciements

Ce travail est financé par Kaizen Solutions - Entreprise n°799348255.

Bibliographie

Bromley J., Bentz J. W., Bottou L., Guyon I., Lecun Y., Moore C. *et al*. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 07, n° 04, p. 669–688.

Castellanos-Paez S., Hili N., Albore A., Pérez-Sanagustín M. (2022). Board-ai: A goal-aware modeling interface for systems engineering, combining machine learning and plan recognition. *Frontiers in Physics*, vol. 10, p. 944086.

Moreno-García C. F., Elyan E., Jayne C. (2019). New trends on digitisation of complex engineering drawings. *Neural Computing and Applications*, vol. 31, n° 6, p. 1695–1712.

Riesen K. (2015). *Structural pattern recognition with graph edit distance*. Springer International Publishing.

Seff A., Ovadia Y., Zhou W., Adams R. P. (2020). *Sketchgraphs: A large-scale dataset for modeling relational geometry in computer-aided design.*

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N. *et al*. (2017). Attention is all you need. *CoRR*, vol. abs/1706.03762.

Vaxiviere P., Tombre K. (1992). Celesstin: Cad conversion of mechanical drawings. *Computer*, vol. 25, n° 7, p. 46-54.

Analyse multimodale de scène : vers une intégration des données contextuelles?

Ibrahim MOHAMED SEROUIS

Université Paul Sabatier, Laboratoire IRIT
118 Route de Narbonne
31062 Toulouse

ibrahim.mohamed-serouis@irit.fr

RÉSUMÉ : *Dans divers domaines comme la communication, le cinéma et les interactions humaines, l'avènement de l'apprentissage multimodal ouvre de nouvelles perspectives dans l'analyse des interactions, des scènes et des publicités. Cependant, peu d'études explorent les modalités autres que l'image et l'audio, négligeant des informations contextuelles cruciales. Cette étude propose un modèle de données pour l'analyse de scènes centrées sur l'humain et une méthodologie pour automatiser l'extraction de ces données à partir de vidéos, ainsi qu'une méthode pour intégrer les données contextuelles dans l'analyse multimodale des scènes.*

MOTS-CLÉS : *Analyse de scène, Apprentissage multimodal, Modélisation de données, Extraction de données multimédias.*

ENCADREMENT : *Florence SÈDES (PR), Lucile SASSATELLI (PR)*

1. Introduction et problématique

Dans les domaines de la communication, du cinéma, ou lors d'interactions humaines, les canaux de communication peuvent être à la fois visuels, textuels, auditifs et contextuels. L'émergence croissante de l'apprentissage multimodal a ouvert de nouvelles perspectives, notamment en ce qui concerne l'analyse d'interactions, de scènes ou de publicités, avec des résultats de plus en plus prometteurs (Schauerte *et al.*, 2011) (Xu *et al.*, 2016) (Gasparini *et al.*, 2018) (Kukleva *et al.*, 2020).

Cependant, très peu d'études à l'instar des travaux de (Gasparini *et al.*, 2018) et (Vicol *et al.*, 2018) exploitent des modalités autres qu'une image associée à sa

description et/ou un audio, négligeant ainsi les informations contextuelles telles que le contexte de l'interaction ou les relations entre les personnages à l'écran. Or, ne pas tenir compte du contexte pourrait entraîner une perte d'informations cruciales lors de l'analyse d'une interaction, le point de vue pouvant être influencé par exemple par le caractère formel ou non de la situation, ou encore de la relation entre les acteurs de l'interaction. De même, rares sont les études exploitant l'aspect relationnel entre les différentes informations en entrée, et encore plus celles mettant un accent sur l'extraction et/ou la représentation de ces données, à l'instar de (Al-Jarrah *et al.*, 2015), (Panta *et al.*, 2018), et plus récemment (Qodseya, 2020).

Cette étude a pour objectif d'ouvrir de nouvelles perspectives pour une compréhension plus approfondie des situations humaines par les systèmes automatisés. Pour ce faire, nous proposons, d'une part, un modèle de données permettant la représentation des données pour un problème d'analyse de scène centrée sur l'humain, ainsi qu'une méthodologie permettant d'automatiser l'extraction de ces données à partir d'une vidéo. D'autre part, nous proposerons une méthodologie interprétable permettant d'intégrer les données contextuelles à l'analyse multimodale de scène.

2. Actions réalisées et limites

2.1. *Modèle de données et extraction des données*

Dans le cadre de cette étude, une première itération du module AMDER (acronyme de **A**dvancing **M**ultimedia **D**ata **E**xtraction and **R**epresentation) a été développée. Comme illustré dans les figures 1a et 1b, ce module prend en entrée une vidéo en et génère en sortie un graphe de scène. Ce graphe comprend les interactions vidéo, les coordonnées de détection des personnages, ainsi que leurs attributs physiques et les émotions exprimées au cours de la scène. Ce module peut être utilisé comme un outil de pré-annotation, auquel on pourrait appliquer un algorithme d'analyse de relation entre les personnages, tel que celui mentionné dans la section 2.2, afin d'enrichir les données contextuelles. La sortie du module est une représentation qui s'inscrit dans une première tentative de modélisation des données relatives aux scènes. Cette modélisation inclut les personnages, leurs attributs (tels que le sexe, la race, les attributs particuliers), les coordonnées de détection, les émotions exprimées pendant la scène, l'ensemble des interactions réalisées dans une scène, ainsi que le discours tenu lors de l'interaction.

2.2. *Modèle d'apprentissage*

Les approches basées sur les Graph Neural Networks (GNNs) sont de plus en plus populaires pour les problèmes de classification sur des données multimodales, bien que certaines réticences aient été exprimées dans la littérature (Ektefaie *et al.*, 2023). Les GNNs permettent d'exploiter l'aspect relationnel entre les données et de traiter des données de tailles variables, ce qui en fit une piste intéressante à explorer pour notre

(a) Capture d'écran du clip

(b) Sortie du module : graphe de scène

Figure 1: Exemple d'exécution du module AMDER sur un extrait de *Chasing Amy*.

problème. Nous avons donc développé une méthodologie en trois étapes pour tirer parti des données contextuelles. La première étape consiste en l'encodage des données d'entrée, via un modèle d'*embedding*. La deuxième étape consiste en l'apprentissage de la représentation des noeuds du graphe. La dernière étape consiste en la classification des noeuds d'intérêt (scène, interaction), et la génération de statistiques sur la contribution des noeuds au résultat final. Cette méthodologie a été évaluée sur différentes tâches d'analyse de scène, telles que la détection d'objectification, la classification d'interactions et la détection de relations entre les personnages. Les résultats obtenus sont prometteurs, dépassant même ceux de (Kukleva *et al.*, 2020) sur la classification d'interactions et de relations entre les personnages, sur un même jeu de données. Pour résoudre le problème de boîte noire (Hussain, 2019) associé à ce type d'algorithme, nous avons également proposé une approche permettant d'obtenir l'influence de chaque élément de notre graphe. Néanmoins, cette dernière reste perfectible notamment concernant les choix d'architecture et d'opérations de traitement des données.

3. Conclusions et perspectives

Cette étude a pour objectif d'ouvrir de nouvelles perspectives pour une compréhension plus approfondie des situations humaines par les systèmes automatisés.

Dans un premier temps, nous présentons un module d'extraction de données vidéo, qui peut servir de pré-annotation pour la création de jeux de données contenant des scènes. Les premiers résultats étant globalement satisfaisants, nous envisageons d'intégrer des techniques de Human Parsing, telles que celle proposée par (Liang *et al.*, 2018), afin d'obtenir des détails plus fins sur les tenues vestimentaires des personnages à l'écran. Cela permettrait par exemple de détecter la nudité, ou une tenue inappropriée dans un contexte sérieux. Nous prévoyons également d'ajouter des techniques de réduction de bruit pour améliorer l'extraction du discours dans les vidéos, ainsi que l'utilisation de grands modèles de langage pour obtenir une description de la scène basée sur les éléments extraits.

Enfin, nous proposons une méthodologie interprétable de raisonnement sur les scènes, qui peut intégrer des données contextuelles (comme les relations entre les personnages, le lieu de la scène) en plus des entrées traditionnelles (images, transcription du discours). Bien que cette méthodologie soit plus performante que certaines méthodes de référence, telles que celle proposée par (Kukleva *et al.*, 2020), pour la classification d'interactions, elle pourrait être améliorée en intégrant des connaissances métiers, en particulier pour le problème de détection d'objectification. Nous envisageons donc d'introduire une approche neuro-symbolique qui bénéficierait des retours d'experts pour la partie symbolique, et des sorties de notre modèle comme connaissances préalables.

4. Bibliographie

Al-Jarrah O. Y., Yoo P. D., Muhaidat S., Karagiannidis G. K., Taha K., "Efficient machine learning for big data: A review", *Big Data Research*, vol. 2, n° 3, p. 87–93, 2015. Publisher: Elsevier.

Ektefaie Y., Dasoulas G., Noori A., Farhat M., Zitnik M., "Multimodal learning with graphs", *Nature machine intelligence*, vol. 5, n° 4, p. 340–350, avril, 2023.

Gasparini F., Erba I., Fersini E., Corchs S., "Multimodal Classification of Sexist Advertisements:", *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, SCITEPRESS - Science and Technology Publications, Porto, Portugal, p. 399–406, 2018.

Hussain J.*, "Deep Learning Black Box Problem"*, Master's thesis, Uppsala University, Department of Informatics and Media, 2019. Backup Publisher: Uppsala University, Department of Informatics and Media.

Kukleva A., Tapaswi M., Laptev I., "Learning Interactions and Relationships between Movie Characters", 2020. arXiv:2003.13158 [cs].

Liang X., Gong K., Shen X., Lin L., "Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. Publisher: IEEE.

Panta F. J., Roman-Jimenez G., S'edes F., "Modeling metadata of CCTV systems and Indoor Location Sensors for automatic filtering of relevant video content", *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, IEEE, p. 1–9, 2018.

Qodseya M., Managing heterogeneous cues in social contexts. A holistic approach for social interactions analysis, PhD thesis, Universite Toulouse 3 Paul Sabatier, 2020.

Schauerte B., Kühn B., Kroschel K., Stiefelhagen R., "Multimodal saliency-based attention for object-based scene analysis", *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 1173–1179, 2011.

Vicol P., Tapaswi M., Castrejon L., Fidler S., "MovieGraphs: Towards Understanding Human-Centric Situations from Videos", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xu P., Davoine F., Bordes J.-B., Zhao H., Denœux T., "Multimodal information fusion for urban scene understanding", *Machine Vision and Applications*, vol. 27, n° 3, p. 331–349, avril, 2016.

Nurses' workload prediction in hospitals – a Machine Learning-based approach

Mohamed GHARBI

Calystene,3C rue Irène Joliot Curie, 38320 Eybens, France.
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.*
Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000, Grenoble, France.*
** Institute of Engineering Univ. Grenoble Alpes*
mohamed.gharbi1@univ-grenoble-alpes.fr

ABSTRACT. In hospitals, nurses represent nearly half of the staff and they play an important role in healthcare delivery. However, there is an existing shortage that leads to increased workloads, which affects their performances and the quality of care they provide. To address this issue, workload prediction has been suggested as a solution. Despite its potential benefits, workload prediction presents challenges such as data selection, data size, and model selection. In this article, we will examine various methods and models utilized for predicting nurses' workload, as well as in other areas such as energy. Additionally, we will propose a methodological approach to overcome the limitations associated with workload prediction in our future works.

KEYWORDS. Workload, data processing, Machine Learning.

SUPERVISION. Christine Verdier (LIG), Maria Di Mascolo (G-SCOP).

1. Introduction

In contemporary healthcare, predicting nurses' workload has emerged as one of the top priorities for hospitals seeking to ensure optimal patient care. Despite the crucial role that nurses play in healthcare delivery, there is a significant shortage of the nursing-staff, which is affecting the hospitals ability to meet the evolving healthcare needs. To address this challenge, considerable efforts have been dedicated to understand and explore the factors influencing nurses' workload and to develop predictive models supporting nursing work management.

Although predicting workload for nurses' remains an ongoing attempt, progress has been made in several fields, including human resources, and energy. In Data Science, numerous Machine Learning models such as linear regression, sequence models like recurrent neural networks (RNNs), especially Long Short-Term Memory

(LSTM) and probabilistic models have been applied to tackle this issue. These models can handle different data types like sparse, heterogeneous, and big data.

In this paper, we describe and develop our global architecture in the field of nurses' workload prediction by outlining the possible paths for our data and models preparation and selection.

2. State of the art

Numerous studies have explored methods for calculating nurses' workload and identifying the factors that influence it within hospital settings. In our prior research (Gharbi *et al.*, 2024), we reviewed these methodologies and outlined the envisioned advancements in the area. In addition, we highlighted the several approaches that were used for the nurses' workload prediction and there limits in combining and using the influential factors such as patients characteristics, nurses experience, the complexity of cares and their variability status. Some studies have highlighted the important role of Machine Learning in predicting nurses' workload; for instance, addressing flexible staffing issues (Kortbeek *et al*., 2015).

Machine learning has already proven beneficial across other domains of healthcare industry, like in clinical trials ranging from preclinical drug discovery to pre-trial planning, facilitated by robust data management and analysis (Weissler *et al.*, 2021). Additionally, (Mirza *et al*., 2023) focused on source data verification (SDV) prediction using LSTM and time series analysis. Furthermore, (Mughees et *al.*, 2021) applied Bi-LSTM for day-ahead peak load forecasting within the power industry. While these Machine Learning techniques are adopted to manage extensive and diverse datasets, (Alsafadi and Wu, 2023) emphasized the utility of Generative Deep Models in tackling challenges associated with smaller datasets. They also discussed the application of random forest models for feature capture.

Nonetheless, when it comes to predicting nurses' workload – due to its complex nature influenced by the cited factors and the different data types– the application of machine learning in this area remains a field for improvement.

3. Problematic

In our research, we encounter various challenges and limitations related to understanding and processing the data, especially considering its size, variation, and complexity. Additionally, selecting of the appropriate models can be challenging based on the expected outcomes. **Data understanding** phase has the goal of identifying and validating the key factors influencing nurses' workload, drawing from their professional experience. This necessitates engaging a significant number of nursing staff to gather meaningful insights. As for **data processing**, it presents its own set of challenges. Our dataset is heterogeneous, consisting of diverse types of data such as numeric, qualitative and textual, in addition, several data sources (medical record, nurses' schedules, PRN etc.). The task is to effectively integrate and leverage this multifaceted data. A notable issue we face is the limited size and sequential nature

of our dataset. This causes some difficulties when considering the application of deep learning models, such as LSTM. While these models have demonstrated effectiveness in prior research, as shown by (Mirza *et al.*, 2023), they typically require extensive data, which in their study spanned from 2014 to 2020.

This leads us to several **pertinent questions:** Where can we access a more extensive dataset? Is a large dataset indispensable for our objective? Could an alternative approach involving the generation of sequential data, while fine-tuned using methods such as Hidden Markov models (Hanif *et al.*, 2017), be a viable solution? Additionally, we need to address how to manage missing data effectively while preserving as much valuable information as possible.

4. Contribution

Based on the literature and the perspectives mentioned in the article (Gharbi *et al*, 2024), data understanding, handling and the use of predictive models are considered as potential contributions. As a follow up, we will discuss these contributions in detail. We propose a structured approach that encompasses the various steps of our future contributions, drawing from a standard data-mining model derived from the CRISP-DM framework (Mirza *et al*., 2023; Wirth and Hipp, 2000). In this approach, we will outline our planned work for the data processing and data modelling stages.

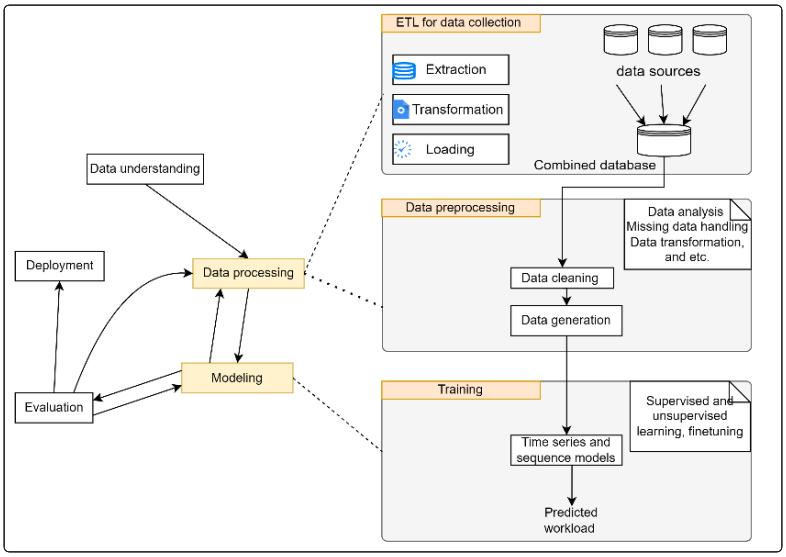


Figure 1. Nurses' workload prediction process: from data understanding to deployment

Figure 1, illustrates the process which begins with **data understanding**; this initial phase enables us to grasp our data, identify relevant variables such as patient and

nurses' characteristics, care types, and specific details about wards and hospitals. The second phase is **data processing** that is divided into two steps. First, we collect data from different sources, including the computerized patient record for patient related data and HR systems for staff-related data, ensuring anonymization for ethical and security reasons. These different sources are consolidated into a unified database following the ETL (extract, transform, and load) process (Hendayun *et al.*, 2021). Second, we analyse and clean our data by addressing missing values, examining correlations between variables, and handling unbalanced classes. Given the small size of our dataset, we aim to explore the potential of generating sequential data inspired by unsupervised learning especially the Generative Deep Models. Finally, techniques such as normalization, outlier detection, and statistics might be employed. For the **prediction model**, entry data might be processed and transformed using machine learning model such as time-series. Subsequently, a ML model that deals with sequential data, can be applied. We then predict workload by ward, expressed in hours across different time slots based on the ward organization (eg. morning, afternoon, and night). Following this, we evaluate our model and deploy it. The result is a dashboard that exploits the several key performance indicators.

As **future work**, we plan to refine our methodological approach on which we will have a process that generates meaningful data and train models that address our challenges. We plan to study and employ Generative Deep Models (GDMs), Long Short-Term Memory (LSTMs) networks, and Time-Series analysis. Additionally, we are interested in exploring the potential of incorporating spatial data into our research.

References

Alsafadi, F., Wu, X., (2023). *Deep generative modeling-based data augmentation with demonstration using the BFBT benchmark void fraction datasets*. Nucl. Eng. Des. 415, 112712.

Gharbi, M., Di Mascolo, M., Verdier. C. (2024). Charge de travail du personnel infirmier dans les hôpitaux -étude bibliographique. Congrès conjoint Alass-Giseh 2024. 3-6 juillet 2024. Liège. Belgique

Hendayun, M., Yulianto, E., Rusdi, J.F., Setiawan, A., Ilman, B., (2021). *Extract transform load process in banking reporting system*. MethodsX 8, 101260.

Mirza, B., Li, X., Lauwers, K., Reddy, B., Muller, A., Wozniak, C., Djali, S., (2023). *A clinical site workload prediction model with machine learning lifecycle.* Healthc. Anal. 3, 100159.

Mughees, N., Mohsin, S.A., Mughees, Abdullah, Mughees, Anam, (2021). *Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting.* Expert Syst. Appl. 175, 114844.

Weissler, E.H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D.F., Benoit, J., Hughes, M.C., Khan, F., Slater, P., Shameer, K., Roe, M., Hutchison, E., Kollins, S.H., Broedl, U., Meng, Z., Wong, J.L., Curtis, L., Huang, E., Ghassemi, M., (2021). *The role of machine learning in clinical research: transforming the future of evidence generation.* Trials 22, 537.

Détection d'anomalies lexicales par fouille d'articles scientifiques : exploration du voisinage d'expressions torturées

Wendeline SWART

Université Toulouse III – Paul Sabatier
Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505 CNRS)
118 route de Narbonne
31062 Toulouse cedex 9

wendeline.swart@irit.fr

RÉSUMÉ : *La méconduite scientifique, notamment le plagiat par paraphrasage, compromet l'intégrité de la recherche. Les articles frauduleux utilisent des expressions torturées pour échapper à la détection de plagiat. Nous explorons une approche automatisée pour détecter ces expressions dans les articles scientifiques. Notre méthode consiste à analyser les termes voisins des expressions torturées connues, basée sur l'observation récurrente de l'utilisation de synonymes. Nous avons développé un algorithme qui identifie les expressions suspectes en examinant les paragraphes contenant déjà des expressions torturées. L'évaluation de notre algorithme sur des documents frauduleux a montré des résultats prometteurs, avec des coefficients d'accord inter-annotateurs de 0,78 et 0,84 calculés sur deux jeux de données.*

MOTS-CLÉS : *Expressions torturées, fouille de textes, intégrité scientifique.*

ENCADREMENT : *Guillaume Cabanac.*

1. Contexte

Un article scientifique présente des résultats de recherche originaux dans un domaine spécifique, généralement soumis à un processus d'évaluation par les pairs avant publication. Malheureusement, à cause des pressions institutionnelles liées à la « course à la publication » ou par appât du gain, certains articles soumis et publiés sont le résultat de méconduites, c'est-à-dire des comportements contraires à l'éthique pouvant

entraîner des conséquences négatives. L'une de ces méconduites consiste à utiliser des paraphraseurs afin, pour un faussaire, d'éviter que des algorithmes anti-plagiats ne détectent ce vol. À cet effet, les faussaires s'approprient un texte, le paraphrasent et l'ajoutent à leur propre article tel quel, sans citer l'auteur originel (action de copier-coller). Ces publications présentent des expressions torturées, propres à l'utilisation de cette méthode de plagiat (Cabanac *et al.*, 2021). Une expression torturée est une expression qui ne fait pas sens en science ou dans le domaine de l'article dans lequel elle est employée. Une expression torturée résulte du remplacement de mots par des synonymes générés par un logiciel de paraphrasage. Chaque expression torturée correspond à une version établie dans un domaine donné, que les lecteurs auraient dû trouver dans l'article. En recourant à plusieurs synonymes pour un même mot, les paraphraseurs créent plusieurs variantes d'expressions torturées pour une même phrase attendue, comme illusté dans le tableau 1. Cette variante est due à l'utilisation d'un dictionnaire de synonymes par les paraphraseurs tel que SpinBot[1]. Par exemple *"counterfeit consciousness"* (conscience contrefaite) est employé au lieu d'*"Artificial Intelligence"* (intelligence artificielle).

TABLEAU 1. *Trois variantes torturées de l'expression établie 'squared error' (erreur quadratique)*

Tortured Phrases	**Expressions torturées**
squared blunder	boulette quadratique
square mistake	méprise quadratique
squared fault	faute quadratique

Les articles frauduleux ou suspects sont référencés sur la plateforme Problematic Paper Screener (PPS)[2], un site d'évaluation post-publication d'articles, dans la section '*Tortured Detector*'. Sur le PPS sont référencées les 5 000 expressions torturées connues à ce jour. Le PPS est un site web collaboratif et libre d'accès qui soutient une communauté de chercheurs, experts et toute autre personne souhaitant contribuer à la dépollution de la science en commentant des articles repérés par les détecteurs du PPS. Les commentaires sont principalement postés sur le site PubPeer[3] dédié au recueil de commentaires post-publication.

L'étude systématique des commentaires postés sur PubPeer par des dizaines de personnes permet d'identifier de nouvelles expressions torturées. En effet, les auteurs de ces commentaires peuvent suggérer de nouvelles expressions que ces personnes ont relevées lors de leur lecture, et qui n'étaient pas référencées dans le PPS[4]. Par la suite, nous vérifions manuellement chacune de ces propositions afin de déterminer s'il s'agit d'un faux positif, c'est-à-dire une expression légitime. Pour chaque com-

1. https://spinbot.com/
2. https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener/tortured
3. https://pubpeer.com/
4. cf. ex. https://pubpeer.com/publications/78F3363D36B17782173CFC0D3F8246

mentaire il est donc nécessaire d'annoter les publications afin de mettre en évidence les expressions torturées en les surlignant dans leur contexte, en jaune par exemple. Ces tâches manuelles sont chronophages et fastidieuses, notamment sur des corpus comme ceux des maisons d'édition renommées, telles que l'Institute of Electrical and Electronics Engineers (IEEE) pour laquelle nous avons signalé plus de 4 000 articles torturés (Swart *et al.*, 2023).

2. Motivation des travaux

Comme le processus d'ajout de nouvelles expressions torturées est long, il est intéressant de s'interroger sur la possibilité de l'automatiser. Nous avons déjà réussi à automatiser la tâche d'annotation des articles à partir des expressions torturées connues.[5] Cependant, maintenant, nous souhaitons automatiser la détection des expressions suspectées d'être torturées en analysant les passages susceptibles d'être les plus affectés dans les articles : les paragraphes qui contiennent déjà des phrases torturées.

3. Méthode

Nous avons conçu un algorithme capable de trouver des expressions candidates à partir des expressions connues, en analysant les termes voisins de celles-ci. Cette démarche est justifiée car nous avons observé l'emploi fréquent de synomymes pour chaque expression torturée. De plus, un même article contient fréquement plusieurs synonymes d'une même expression (voir le tableau 1, par exemple).
Pour un document donné, l'algorithme développé va rechercher l'ensemble des expressions torturées connues à ce jour : celles qui sont présentes dans le document analysé et déjà repertoriées dans le PPS comme expressions torturées. Une fois ces expressions extraites, on les découpe. Par exemple, *"square blunder"* deviendra d'un côté *"square"* et de l'autre *"blunder"*. Puis, on recherche sur ce même document les termes découpés afin d'examiner l'ensemble des termes voisins de ceux-ci. Ces termes voisins sont examinés dans des fenêtres de grandeurs différentes afin de déterminer la longueur de l'expression la plus optimale. Pour le moment nous avons exploré le contenu de deux documents reconnus comme étant frauduleux, avec une fenêtre des termes voisins égale à 5 mots et une autre fenêtre de termes égale à 3 mots.
Une fois la liste des expressions suspectes extraites, un autre algorithme les trie par rapport à leur taux de problèmes, celui-ci étant calculé en fonction de critères subjectivement définis. Pour ces critères nous avons essayé de rester au plus proche de ceux utilisés lors de la vérification humaine, que je réalise depuis 2 ans. Cette démarche est possible grâce à l'API de Dimensions [6], puisqu'actuellement les vérifications manuelles sont réalisées avec cette plateforme, qui est la principale source de données du PPS. Parmi ces critères, nous pouvons citer le nombre de publications qui contiennent

5. cf. ex. `https://pubpeer.com/publications/36448420457798DEC2236D7048CF3C`
6. `https://app.dimensions.ai/discover/publication`

l'expression, toutes les années confondues, et depuis 2015. Nous quantifions également les pays qui ont le plus de publications contenant l'expression suspecte afin de constater s'il s'agit de pays soumis à de fortes pressions institutionnelles tels que la Chine ou l'Inde (Cabanac, 2024).

4. Résultats

L'algorithme a permis de détecter pour le premier document 72 expressions dont 53 suspectes et pour le second 138 expressions dont 49 suspectes. Pour déterminer la pertinence de nos résultats, nous avons effectué une évaluation grâce au coefficient d'accord inter-annotateurs Kappa Cohen. Pour une recherche des termes voisins menés sur une fenêtre de 5 mots, le coefficient Kappa Cohen est de 0,78 tandis que pour une fenêtre de 3 mots nous il est de 0,84. Ces résultats sont donc prometteurs mais il est possible de les améliorer.

5. Actions futures

Afin d'améliorer les résultats de découverte d'expressions torturées, il faut réussir à déterminer la meilleure fenêtre possible de recherche des termes voisins. Actuellement, nos résultats sont très bruités, il faudrait donc parvenir à le réduire à un nombre minime. Pour filtrer nos résultats, nous nous appuyons sur le taux de faux positifs et le taux de faux négatifs, c'est-à-dire le nombre d'expressions légitimes ou torturées, dans les expressions récupérées. Cependant, nous aimerions automatiser cette tâche. Pour cela, nous allons mettre au point un algorithme d'apprentissage machine et utilisant les arbres de décisions. Le modèle apprendrait sur les évaluations déjà effectuées manuellement. Une amélioration envisagée est de rajouter dans l'algorithme la prise en compte du domaine de recherche de la publication. En effet, certains termes sont torturés dans un domaine mais pas dans tous. Par exemple *"profound learning"* fait sens en psychologie mais n'en a pas en informatique, où l'expression attendue est *"deep learning"*.

6. Bibliographie

Cabanac G., « Fake science : panorama des méconduites et contre-feux pour déjouer les pièges », 2024, Conférence donnée le 18/03/2024 à l'Université Paul Sabatier, `https://hal-lara.archives-ouvertes.fr/IRIT-IRIS/hal-04225515v7`.

Cabanac G., Labbé C., Magazinov A., « Tortured phrases : A dubious writing style emerging in science. Evidence of critical issues affecting established journals », 2021. preprint, `https://doi.org/10.48550/arXiv.2107.06751`.

Swart W., Cabanac G., « Year after year : Tortured conference series thriving in Computer Science », 2023, preprint, `https://doi.org/10.48550/arXiv.2401.02422`.

Découverte d'acronymes torturés dans des publications scientifiques

Alexandre CLAUSSE

Université Toulouse III – Paul Sabatier
Institut de recherche en informatique de Toulouse (IRIT UMR 5505 CNRS)
118 route de Narbonne
31062 Toulouse cedex 9

alexandre.clausse@univ-tlse3.fr

RÉSUMÉ : *Dans un contexte de course à la publication, du contenu plagié est régulièrement publié, amenant à une pollution croissante de la littérature scientifique. Une telle fraude peut être caractérisée par l'utilisation d'expressions torturées. Des solutions ont été développées afin de contribuer à la détection et au signalement de tels contenus. D'une part, elles reposent sur des méthodes et ressources de nature hétérogène, avec des biais qui leur sont propre. D'autre part, elles nécessitent la collaboration d'experts permettant l'alimentation d'une liste d'expressions connues. Ainsi, nous proposons une approche peu coûteuse en ressources et indépendante de tout domaine, reposant sur la détection d'acronymes torturés, étant visuellement facile à mettre en œuvre. Afin de détecter la présence de l'ensemble de ces expressions dans une publication donnée, nous mettons à disposition un jeu de données de publications torturées, un algorithme d'extraction et de classification d'acronymes, ainsi qu'une méthode permettant d'évaluer cette ligne de base. Les résultats obtenus sont biaisés par l'utilisation du jeu de données de développement, annoté par une seule personne, lors de l'évaluation de la solution proposée. Nos futures recherches seront focalisées sur l'élaboration de méthodes permettant la détection de formes particulières d'expressions telles que les hallucinations et les termes polysémiques, toujours dans une optique de faciliter la détection d'expressions torturées.*

MOTS-CLÉS : *Extraction d'information, détection d'anomalie, plagiat, intégrité scientifique.*

ENCADREMENT : *Guillaume Cabanac, Pascal Cuxac, Cyril Labbé.*

1. Contexte

Les politiques de recherche publique exercent une pression constante sur les chercheurs, en les incitant à publier le plus régulièrement possible dans des revues réputées, afin d'obtenir les meilleures performances dans les classements internationaux. Cette situation, communément appelée « publier ou périr », amène un petit nombre de personnes peu scrupuleuses à manquer de considération pour la rigueur nécessaire aux travaux de recherche scientifique, en ayant recours à la fabrication, à la falsification et au plagiat. Ces problèmes ont été décrits dans un ouvrage édité par Biagioli *et al.* (2020). Le plagiat peut être déguisé par l'utilisation d'expressions torturées, définies par Cabanac *et al.* (2021) comme étant le remplacement de termes scientifiques établis par des synonymes, les vidant ainsi de toute signification, principalement en utilisant des outils de paraphrase. Par exemple, le terme informatique « *machine learning* » peut être paraphrasé en « *computer mastering* ». De plus, un texte contenant des expressions torturées peut échapper à la vigilance d'un comité de relecture, et être publié tel quel, amenant à une pollution croissante de la littérature scientifique. Cela met aussi en doute leur probité quant à la tenue des expériences et la rédaction sincère des résultats. Pour contrer un tel problème, il est envisageable de permettre aux différents acteurs de l'édition savante de vérifier la conformité d'un contenu, notamment vis-à-vis du respect des valeurs et principes d'honnêteté et de rigueur attendu de la part des chercheurs. Une solution consiste à développer une brique logicielle permettant d'identifier en amont de la publication les articles contenant de telles expressions, par leur détection automatique, dans le but d'y apporter une attention accrue.

2. État de l'art

En 2021, le *Problematic Paper Screener* (PPS)[1] a été développé pour permettre le recensement de publications scientifiques frauduleuses ou suspectes, incluant les contenus plagiés par l'utilisation d'expressions torturées. Ce site facilite l'évaluation post-publication en identifiant des articles problématiques et en préparant des rapports d'évaluation que les utilisateurs peuvent publier sur PubPeer[2], une plateforme de réévaluation collaborative d'articles scientifiques. Ainsi, plus de 5 000 expressions torturées distinctes ont été recensées, et plus de 13 000 articles contenant de telles expressions ont été listés sur le PPS. Dans l'optique de détecter du contenu plagié, des travaux ont constitué des corpora de documents web, journalistiques et scientifiques, pour expérimenter plusieurs approches utilisant des modèles de langue et d'apprentissage machine. Gehrmann *et al.* (2019) ont proposé une méthode de détection de textes générés par des modèles de langue, axée sur des caractéristiques syntaxiques et distributionnelles. Les résultats obtenus présentent un biais lié à l'utilisation de données similaires. Wahle *et al.* (2022) ont cherché des façons d'identifier le plagiat par paraphrasage en constituant un jeu de données composé de paragraphes paraphrasés, et

1. https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener
2. https://www.pubpeer.com

en comparant leurs résultats de classification avec ceux de deux détecteurs de plagiat. Leurs résultats sont biaisés par l'utilisation d'un jeu d'entraînement sur les mêmes poids lors de l'apprentissage par transfert de leurs modèles d'apprentissage machine. Lay *et al.* (2022) ont complété l'étude précédente par l'apport d'un ensemble de cinq-grammes (suite de cinq mots), comprenant ou non des expressions torturées. Leurs résultats varient d'un modèle et d'une fonction d'agrégation à l'autre. De plus, les résultats de classification sont biaisés par l'utilisation d'un jeu de données déséquilibré, dont certaines sont dupliquées. Martel *et al.* (2023) ont également réutilisé ce corpus afin de réaligner une centaine de paires de phrases et compléter l'étude de Wahle *et al.* (2022), leurs résultats présentent des biais similaires, en plus de comporter des caractéristiques peu exploitables car extraites depuis des modèles de plongement lexical non contextuel, posant un problème de polysémie.

3. Problématique

Étant données les limites soulignées dans la section 2, il faudrait arriver à extraire d'autres caractéristiques permettant la détection d'expressions torturées. De plus, les méthodes actuelles de détection se font par lecture humaine et la collecte participative de nouvelles expressions torturées (notamment via le PPS). Il faudrait donc automatiser cette tâche et compléter ces méthodes afin de se prémunir d'autres formes de plagiat.

4. Contribution

Très récemment, nous avons exploré l'extraction d'acronymes torturés avec un algorithme de correspondance (notamment en comparant les initiales), développé en utilisant un corpus vérité-terrain de 75 articles scientifiques manuellement annotés (Clausse *et al.*, 2024). Les résultats obtenus sont biaisés car l'algorithme a été évalué sur l'ensemble de données de développement, et notre approche ne peut pas détecter les formes hallucinées (par exemple, le « *Bolster Vector Machine (BVM)* » où l'acronyme correspond à sa forme développée). Les prochaines étapes consistent donc à améliorer la qualité du corpus vérité-terrain par l'annotation de celui-ci (par au moins une autre personne) et le calcul d'un accord inter-annotateur.

5. Recherches futures

Dans la mesure où nous travaillons sur des données textuelles en anglais et dans un nombre restreint de domaines, il est nécessaire de disposer d'ensembles de données axés sur des articles scientifiques torturés provenant de plusieurs domaines, générés par différents outils de paraphrase et dans différentes langues. Des échantillons synthétiques pourraient être paraphrasés à l'aide d'auto-encodeurs ou générés à l'aide de modèles auto-régressifs (les expressions torturées recensées par le PPS pourraient

être utilisées), et pourraient être comparés à des échantillons existants. Les domaines pourraient être pris en compte à l'aide de la modélisation thématique. Dans le but de détecter les expressions torturées en utilisant des ressources informatiques limitées, il serait intéressant de prendre en compte plusieurs caractéristiques pour réaliser des algorithmes de correspondance, avec la difficulté de prendre en compte les différentes formes de ces expressions (par exemple les acronymes ou encore la spécificité d'un domaine). Il serait aussi intéressant de développer une méthode permettant d'identifier les sources de plagiat. Pour cela, l'utilisation de modèles de langues permettrait de détecter du contenu suspect, associés à des dictionnaires, pour retrouver le contenu original en fonction d'un seuil de score de similarité.

6. Bibliographie

Biagioli M., Lippman A. (éd.), *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, The MIT Press, 2020. DOI: https://doi.org/10.7551/mitpress/11087.001.0001.

Cabanac G., Labbé C., Magazinov A., "Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals", 2021. Preprint arXiv. DOI: https://doi.org/10.48550/arXiv.2107.06751.

Clausse A., Cabanac G., Cuxac P., Labbé C., "Extraction d'acronymes torturés dans la littérature scientifique", *in* P. Cuxac, C. Lopez (éd.), *Atelier TextMine de la conférence Extraction et Gestion des Connaissances (EGC) de 2024*, Dijon, France, p. 27–37, 2024. URL : https://hal.science/hal-04426448.

Gehrmann S., Strobelt H., Rush A., "GLTR: Statistical Detection and Visualization of Generated Text", *in* M. R. Costa-jussà, E. Alfonseca (éd.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, p. 111–116, 2019. DOI: https://doi.org/10.18653/v1/P19-3019.

Lay P., Lentschat M., Labbe C., "Investigating the detection of Tortured Phrases in Scientific Literature", *in* A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, L. L. Wang (éd.), *Proceedings of the Third Workshop on Scholarly Document Processing*, Association for Computational Linguistics, p. 32–36, 2022. URL: https://aclanthology.org/2022.sdp-1.4.

Martel E., Lentschat M., Labbe C., "Detection of Tortured Phrases in Scientific Literature", *in* T. Ghosal, F. Grezes, T. Allen, K. Lockhart, A. Accomazzi, S. Blanco-Cuaresma (éd.), *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, Association for Computational Linguistics, p. 43–48, 2023. DOI: https://doi.org/10.18653/v1/2023.wiesp-1.6.

Wahle J. P., Ruas T., Foltýnek T., Meuschke N., Gipp B., "Identifying Machine-Paraphrased Plagiarism", *in* M. Smits (éd.), *Information for a Better World: Shaping the Global Future*, Springer, p. 393–413, 2022. DOI: https://doi.org/10.1007/978-3-030-96957-8_34.

Bien Vivre et Bien Vieillir sur son territoire

Analyser le cadre de vie avec trajectoires sémantiques

Yunji ZHANG

LIUPPA, Université de Pau et des Pays de l'Adour (UPPA)
2 allée du Parc Montaury, 64600 ANGLET, France
yunji.zhang@univ-pau.fr

RESUME. Bien Vivre et Bien Vieillir sur son territoire représente un enjeu reconnu par les décideurs locaux. Les travaux existants se concentrent sur une analyse mono-thématique d'un composant du « Bien Vivre et Bien Vieillir » (ex : éducation pour bien-vivre). Il y a un manque d'analyses multi-vues qui donnent aux décideurs une image complète du cadre de vie. Pour combler cette lacune, nos travaux de recherche visent à proposer : (i) une cartographie avec des dimensions impliquées dans le Bien Vivre et le Bien Vieillir ; (ii) une intégration croisée des données et de leur modélisation pour favoriser leur accessibilité ; et (iii) une analyse multi points de vue basée sur le concept de trajectoires sémantiques afin de représenter les différentes dimensions relatives au « Bien Vivre et Bien Vieillir ».

MOTS-CLES : analyse de données, trajectoires sémantiques, bien vivre, bien vieillir

ENCADREMENT : Philippe ROOSE (LIUPPA), Franck RAVAT (IRIT), Sébastien LABORIE (LIUPPA)

1. Introduction

On estime qu'à l'horizon 2050, il y aura 30 000 Landais et Landaises dépendants alors qu'ils étaient 17 000 en 2015. Les décideurs locaux des Landes ont par conséquent besoin d'une analyse multi-vues du cadre de vie et des recommandations pour mieux vivre sur leur territoire. Cependant, la plupart des études se concentrent principalement sur une analyse mono-thématique du Bien Vivre et Bien Vieillir (ex : éducation pour bien-vivre (Collado-Ruano J. *et al*., 2019). Il y a un manque d'une vision complète du cadre de vie locale pour les décideurs.

Compte tenu de ce contexte, la chaire Industrielle « Bien Vivre et Bien Vieillir » est soutenue par la technopole Domolandes, le Conseil départemental des Landes, l'Université de Pau et des Pays de l'Adour (UPPA) et le Conseil Régional Nouvelle Aquitaine. Le sujet de ma thèse s'inscrit dans cette chaire et porte sur les « Trajectoires du Bien Vivre et Bien Vieillir sur son territoire » pour offrir une analyse multi-vues du cadre de vie grâce au concept de trajectoires sémantiques (Noël *et al*., 2017).

La problématique peut donc être résumée au travers des trois questions suivantes : « Quelles sont les dimensions analytiques impliquées dans Bien Vivre et Bien Vieillir ? », « Comment établir un modèle de données relatif au cadre de vie quotidienne ? », « Comment proposer un modèle d'analyse de données facilitant la prise de décision des acteurs locaux ? ».

Cet article vise tout d'abord à présenter le contexte ainsi que la problématique de notre recherche en vue de clarifier le rôle des trajectoires sémantiques dans la construction du cadre de vie. Dans un second temps, nous avons présenté l'avancement de ma recherche, qui peut être divisée en trois parties distinctes : (i) une cartographie avec des dimensions impliquées dans Bien Vivre et Bien Vieillir ; (ii) une intégration croisée des pour aboutir à une modélisation favorisant les analyses ; et (iii) une analyse multi-points de vue basée sur le concept de trajectoires sémantiques.

2. Contexte et problématique

Pour construire une analyse multi-points de vue, nous avons besoin de données partant sur des dimensions différentes. Ces données sont dispersées dans différentes sources de données sans lien direct. Comme les relations causales entre les données ne sont pas claires et qu'il manque des métadonnées permettant d'expliciter ces sources, nous devons procéder à une intégration des données afin d'être capable de construire une trajectoire de vie à travers des données spatio-temporelles et de l'interpréter selon des dimensions multiples. Nous reprenons le concept de trajectoires sémantiques comme outil d'analyse de données sur le bien vivre et bien vieillir (Figure 1).

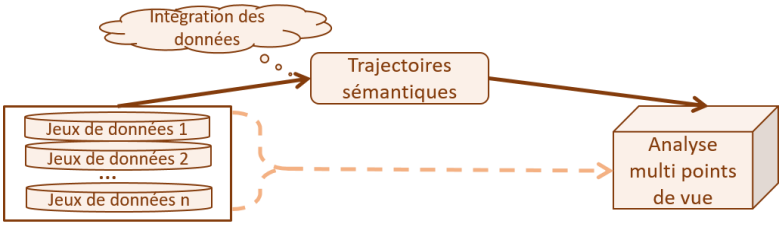


Figure 1. Concept d'analyse multi points de vue

3. Cartographie des dimensions du Bien Vivre et Bien Vieillir

À cette étape, nous avons identifié les dimensions analytiques du Bien Vivre et Bien Vieillir et les combinés avec des données ouvertes accessibles.

Notre étude sur des mesures de la qualité de vie (WHO, 1998 ; OECD, 2011) nous a permis de proposer un cadre analytique pour le « Bien Vivre » avec 9 dimensions (cf., Figure 2). Pour chacune, nous avons identifiées les sous-branches analytiques (nombre associé aux dimensions) (ex : sous-branches de la dimension

« Physique », Figure 3). La difficulté est que certaines dimensions sont associées à plusieurs données qu'il va falloir combiner alors que d'autres sont non-valuées.

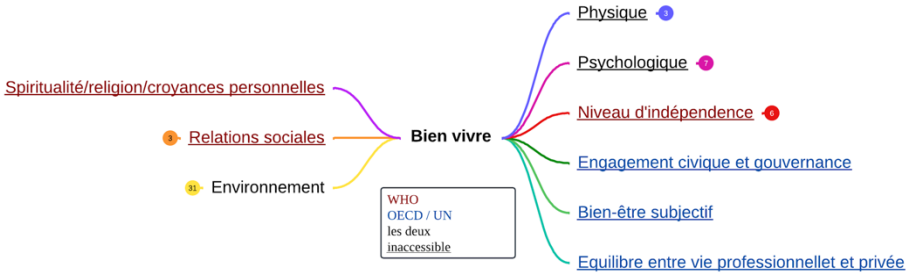


Figure 2. Cadre analytique de Bien Vivre[1]

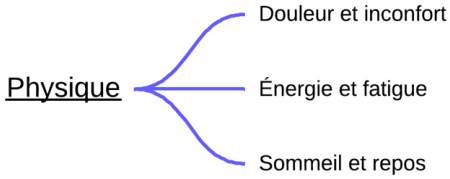


Figure 3. Sous-branches de Physique

Le cadre analytique du « Bien Vieillir » possède des dimensions communes mais intègre des dimensions spécifiques au vieillissement (données de l'OMS (WHO, 2020)) comme le montre la Figure 4.

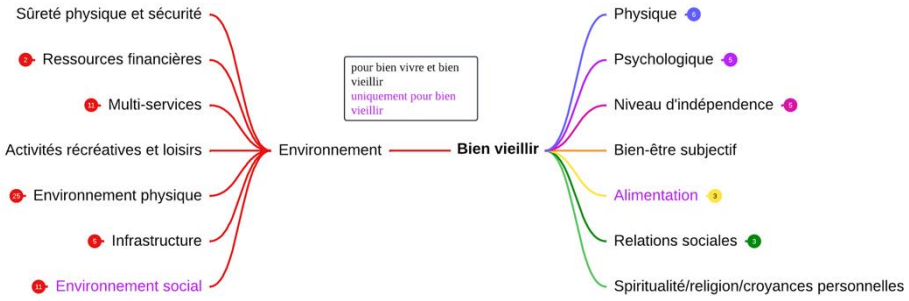


Figure 4. Cadre analytique de Bien Vieillir[2]

[1] Cadre complet de Bien Vivre : https://bit.ly/4dgnGdm
[2] Cadre complet de Bien Vieillir dans le même lien au-dessus

4. Futurs axes de recherche

Nos premiers travaux de cartographie ont répondu à la première question de notre problématique « Quelles sont les dimensions analytiques impliquées dans Bien Vivre et Bien Vieillir ? ». Pour résoudre entièrement notre problématique, nous avons identifié les trois obstacles technologiques suivants :

- Le premier obstacle porte sur le manque de données pour modéliser l'intégration. Nous proposons 3 scénarios pour obtenir des données : (i) trouver plus de collaborateurs pour obtenir plus de données réelles ; (ii) simuler les données locales selon des données d'autres régions ou pays ; et (iii) construire et entraîner des modèles avec des données d'autres régions, puis les importer pour représenter le territoire considéré.

- Le deuxième obstacle est l'intégration croisée de données reposant sur différents niveaux d'agrégation. Cette intégration multi-dimensions et multi-échelles nécessite le développement de nouveaux modèles de données permettant de représenter des croisements à différents niveaux avec des données manquantes voire erronées tout en tenant compte de l'évolution temporelle des données.

- Le dernier obstacle concerne la construction des modèles de trajectoires sémantiques avec des capacités prédictives. L'objectif est d'offrir des analyses originales descriptives et prédictives de données, centrées sur l'humain afin de faciliter la prise de décision. L'intérêt des trajectoires sémantiques est d'offrir une vision spatio-temporelle d'activités humaines selon différentes dimensions.

5. Conclusion

Nos premiers travaux ont consisté en une analyse approfondie de la littérature sociologique sur Bien Vivre et Bien Vieillir, qui nous a permis de spécifier un cadre analytique multi-dimensions. Notre contribution vise à définir un processus d'ingestion de données afin d'alimenter notre modèle multi-dimensions pour fournir des analyses descriptives et prédictives basées sur des trajectoires sémantiques.

Bibliographie

Collado-Ruano J., Morillo M. M., et González F. J. A. (2019). Education and Good-Living: Transdisciplinary Skills for Teachers´ Training. *Athenea Digital*. vol. 19, n°3, e2216.

Noël D., Villanova-Oliver M., Gensel J., Le Quéau P. (2017). Design Patterns for Modelling Life Trajectories in the Semantic Web. *Web and Wireless Geographical Information Systems*. vol. 10181, p 51-65

Organisation for Economic Co-operation and Development (2011). *How's Life?: Measuring Well-Being*. OECD.

World Health Organization (2020). *Age-friendly environments in Europe: Indicators, monitoring and assessments*.

World Health Organization (1998). *WHOQOL User Manual*.

La gestion frugale de données

Vlada STEGARESCU

IRIT, CNRS (UMR 5505), Université Toulouse Capitole, Akkodis
Toulouse, France

vlada.stegarescu@(irit.fr, akkodis.com)

RÉSUMÉ : *La frugalité est devenue un concept tendance, englobant des considérations économiques, sociétales et environnementales. En informatique, le terme « frugalité des données » est souvent utilisé à tort de manière interchangeable avec la durabilité, l'éco-responsabilité, ou encore la minimisation des données. Pour combler cette lacune en matière de recherche, nous proposons : (i) une définition complète de la frugalité de la gestion des données ; et (ii) plusieurs futurs axes de recherche visant à parvenir à une gestion frugale des données.*

MOTS-CLÉS: *Frugalité, Durabilité, Eco-responsabilité, Gestion de données*

ENCADREMENT : *Professeur Franck Ravat (directeur de thèse), Maître de conférence Jiefu Song (co-directeur de thèse), Benoit Baurens (encadrant scientifique industriel)*

1. Introduction

L'essor du Big Data et des systèmes scalables a incité les entreprises à privilégier une utilisation exhaustive des données, approche qui entraîne souvent une accumulation de données inutiles et une complexité croissante dans leur utilisation. Ce gaspillage de ressources est particulièrement préoccupant sur le plan écologique. Ainsi, la "Frugalité" émerge comme une solution potentielle pour prévenir ces problèmes.

La frugalité des données est souvent utilisée à tort pour parler de minimisation des données (Biega *et al.*, 2020) ou de réduction des données (Esubalew Aman Mezmir, 2020), ou comme synonyme de *responsabilité environnementale* et de *durabilité*. Cependant, ces concepts ne sont pas interchangeables. Alors, la problématique de recherche de ma thèse peut être résumée comme suit : "Comment définir une gestion frugale de données ?" et "Comment mesurer le niveau de frugalité du processus de gestion de données ?".

L'objectif de ce papier est de présenter les problématiques associées à la frugalité dans la gestion de données. Dans un second temps, nous présenterons l'avancement de nos recherche qui peut être décliné à deux niveaux différents : (i) une définition exhaustive de la frugalité pour la gestion des données, (ii) plusieurs futurs axes de recherche visant une gestion frugale de données.

2. Gestion frugale de données

La confusion entre *responsabilité environnementale*, *durabilité* et *frugalité* vient principalement de leurs piliers communs : *Économique*, *Sociétal* et *Environnemental*.

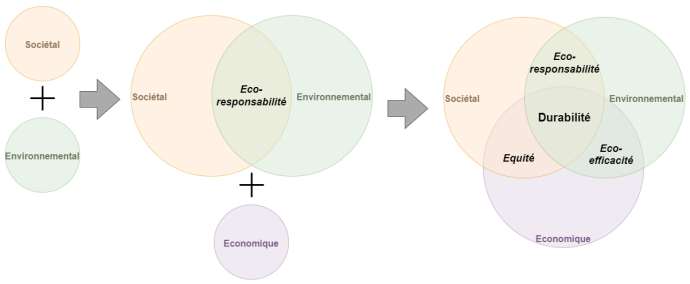


Figure 1. *Interdépendances entre l'éco-responsabilité et la durabilité*

Comme représenté dans la figure 1, la convergence entre les aspects sociétal et environnemental aboutit au concept d'éco-responsabilité. En d'autres termes, l'éco-responsabilité s'intéresse à la diminution de l'impact écologique conformément aux réglementations environnementales, et aux mesures sociétales adoptées dans ce sens, comme la prévention et la sensibilisation.

Cependant, la croissance économique reste un objectif d'actualité, notamment dans un contexte tenant compte de l'environnement. Alors, la durabilité émerge comme résultat de la convergence des trois piliers, comme représenté sur la figure 1. Cette conceptualisation de la durabilité est souvent retrouvé dans la littérature scientifique, mais elle n'a pas de point d'origine unique (Purvis *et al.*, 2019). La durabilité, qui fait référence à la capacité d'une solution ou d'un processus à conserver ses caractéristiques spécifiques au fil du temps, assure l'efficience de l'activité économique en respect avec les limites environnementales et sociétales.

Comme présenté sur la figure 2, nous définissons la frugalité en informatique comme une solution satisfaisant les principes de la durabilité tout en visant l'optimisation de ressources, la consolidation ou l'amélioration du niveau de performance et la réponse aux besoins métier.

Nos premiers travaux nous ont permis également de proposer une définition complète de la **frugalité du management de données**. La gestion frugale des données est une approche visant un résultat durable, et qui consiste à respecter les principes **RePRO**:

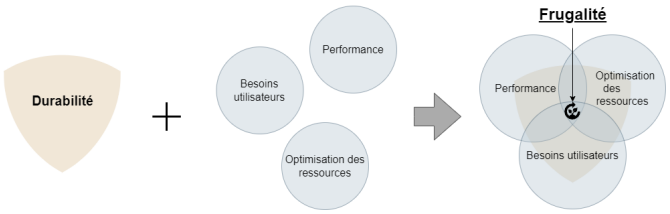


Figure 2. *Interdépendances entre la durabilité et la frugalité*

assurer au moins le même niveau de **P**erformance, tout en **O**ptimisant les **R**essources utilisées et vérifier la conformité de la solution avec les exigences exprimées par les utilisateurs (**RE**quirements) à chaque phase du **cycle de vie des données**.

La gestion de données peut être décrite comme étant frugale lorsqu'elle remplit trois conditions principales :

1) Les nouveaux systèmes ou les systèmes existants modernisés doivent prendre en compte les exigences du décideur (**RE**quirements) dans toutes les phases du cycle de vie des données. Cette propriété pourrait nous permettre d'éviter de collecter ou de traiter des données inutiles.

2) Les actions prises par les entreprises dans un objectif final de frugalité doivent assurer au moins le même niveau de **P**erformance qu'une approche classique de gestion de données. Par performance, nous entendons que lorsqu'une approche frugale est suivie par la stratégie de gestion des données choisie, la qualité, la cohérence, la sécurité, la disponibilité et la fiabilité des données ne doivent pas être affectées négativement.

3) Une gestion frugale des données ne concerne pas seulement l'**O**ptimisation des **R**essources nécessaires à l'analyse future, mais aussi les conséquences de cette nouvelle conception, comme l'empreinte carbone. La gestion frugale des données implique de prendre des mesures pro-environnementales à chaque étape du cycle de vie des données. Cependant, les mesures prises dans le cadre du développement durable ne doivent pas contredire les aspects législatifs de l'activité commerciale.

3. Axes de recherche

Nos premiers travaux sur la frugalité de la gestion de données nous ont permis de répondre à la première question de notre problématique au travers de la définition que nous avons proposée. Pour répondre en intégralité à notre problématique, nous avons identifié trois verrous technologiques :

– Le premier obstacle vise à **qualifier** les données et les traitements associés selon deux axes : (i) réduction des données en fonction du besoin métier et (ii) minimisation des coûts et de l'impact environnemental. Nous souhaiterons proposer un système de métriques nous permettant d'évaluer le niveau de frugalité du processus de gestion

de données mise en place. Notre objectif est de proposer un système d'évaluation multicritères ayant comme résultat un score global de frugalité. Les critères permettant ce calcul peuvent être formalisés par des éléments mesurant le niveau d'optimisation de ressources, le niveau de performance ou le niveau de cohérence entre la solution et le besoin utilisateur.

– Le deuxième obstacle vise à définir une politique de frugalité **mesurable**, conforme aux exigences réglementaires et environnementales en vigueur, tout en optimisant l'utilisation des ressources disponibles et les exigences métier. La frugalité telle que nous l'avons définie, remet en question les architecture existantes, comme les lacs de données qui reposent sur les principes de stockage de la totalité des données brutes et des transformations associées. Nous envisageons de mesurer la frugalité selon plusieurs critères : le type de données, les architectures choisies, les environnements de stockage (on *cloud* ou on *premise*) et le niveau de satisfaction des besoins exprimés par les utilisateurs. Notre objectif est d'introduire une nouvelle technique de conception frugale ("frugality by design") qui permettrait d'intégrer les principes de frugalité dès les premières phases du développement des nouvelles solutions.

– Le troisième obstacle consiste dans la mise en place d'un **pilotage** du management frugal de données, reposant sur un monitoring en temps réel des mesures de frugalité définies précédemment et sur des recommandations d'actions d'amélioration de la frugalité. L'objectif est de fournir une stratégie de pilotage générique et facilement configurable en fonction des besoins et des ressources.

4. Conclusion

Nos travaux de recherche se centrent sur la frugalité associée au management de données. Notre première contribution se matérialise à travers d'une définition de la frugalité pour la gestion de données.

Nous avons présenté plusieurs axes de recherche : (i) qualification des données et traitements associés, (ii) mesure de la frugalité et (iii) pilotage d'un management frugal de données. La partie la plus cruciale de nos futurs travaux concerne la définition de métriques tenant compte des trois piliers sociétal, économique et environnemental.

5. Bibliographie

Biega A. J., Potash P., Daumé H., Diaz F., Finck M., "Operationalizing the Legal Principle of Data Minimization for Personalization", *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, Association for Computing Machinery, New York, NY, USA, p. 399–408, juillet, 2020.

Esubalew Aman Mezmir, "Qualitative Data Analysis: An Overview of Data Reduction, Data Display and Interpretation", *Research on Humanities and Social Sciences*, novembre, 2020.

Purvis B., Mao Y., Robinson D., "Three pillars of sustainability: in search of conceptual origins", *Sustainability Science*, vol. 14, n° 3, p. 681–695, mai, 2019.

Utilisation des graphes de connaissances dans la génération augmentée de récupération

Marion SINAEVE

Laboratoire Informatique de Bourgogne
9 avenue A. Savary
21000 Dijon

Marion_Sinaeve@etu.u-bourgogne.fr

RÉSUMÉ. *Les grands modèles de langage (LLM) ont révolutionné le domaine du traitement automatique des langues, mais ils présentent certaines limites, notamment la présence d'hallucinations dans le contenu généré. Pour remédier à cela, la génération augmentée par récupération (RAG) a été développée. Elle vise à améliorer la pertinence des réponses des LLM en intégrant des mécanismes de recherche d'informations dans des bases de connaissances externes. Toutefois, la RAG présente elle aussi des défauts, et ses limitations peuvent conduire à une diminution de la confiance des utilisateurs dans la fiabilité des contenus générés. Face à ces défis, une alternative prometteuse émerge avec les graphes de connaissances. Ces derniers offrent une façon structurée et interconnectée de représenter l'information, ouvrant ainsi de nouvelles perspectives pour améliorer la pertinence et la fiabilité des contenus générés par les LLM. L'exploration des manières dont les graphes de connaissances peuvent compléter ou renforcer les approches basées sur les LLM et le RAG représente une direction intéressante pour les recherches futures dans le domaine de l'intelligence artificielle générative.*

MOTS-CLÉS. *RAG, LLM, graphe de connaissances*

ENCADREMENT. *Lylia ABROUK, Christophe CRUZ, Laurent GAUTIER*

1. Contexte

Les grands modèles de langage (LLM) marquent une avancée significative dans le domaine du traitement automatique du langage naturel (TALN). Malgré leur haute performance, ils comportent des lacunes notables nécessitant souvent une supervision et une évaluation humaines. Parmi ces limitations figure leur difficulté à comprendre

pleinement le contexte et les subtilités des situations, ce qui peut mener à des incohérences dans le contenu généré. De plus, ces modèles éprouvent des difficultés avec des sujets très spécialisés ou exigeants des informations précises. Ils sont aussi critiqués pour leur tendance à produire des hallucinations, c'est-à-dire à générer du contenu comportant des informations fausses tout en paraissant cohérent et fluide.

Pour pallier au phénomène d'hallucinations, la technique de la RAG (Retrievied Augmented Generation) à été introduite en 2020 par Lewis et al (Lewis *et al.*, n.d.). Dans une perspective plus large, la RAG vise à améliorer les résultats produits par les LLMs. Cette méthode consiste à intégrer des documents externes dans la base de connaissances des LLMs, de sorte que l'extraction d'information s'effectue sur ces données externes plutôt que sur les données d'entraînement du modèle. Par ailleurs, les données récupérées dans la RAG sont classées avant d'être réintroduites dans le LLM, permettant une meilleure adéquation des données générées avec la requête de l'utilisateur. Utiliser la RAG permet donc de produire un texte basé sur des informations vérifiables, évitant les hallucinations et offrant une meilleure compréhension contextuelle de la demande de l'utilisateur.

Le fonctionnement de la RAG se divise en plusieurs étapes, pouvant varier en fonction de la catégorie à laquelle elle appartient : naïve, avancée ou modulaire (Gao *et al.*, n.d.). La RAG naïve se compose d'une étape d'indexation, de récupération et de génération. Au cours de l'*indexation* des données, les documents que l'on souhaite intégrer dans la base de connaissances sont divisés en plusieurs morceaux (chunks) où chaque morceau représente une à plusieurs informations. Ces derniers sont par la suite placés dans un espace vectoriel, permettant ainsi à chaque chunk d'être converti en un vecteur représentatif de l'information qu'il comporte. Ces vecteurs sont ensuite stockés dans une base de données sous la forme clé (vecteur) / valeur (chunk) formant ainsi la base de connaissances qui sera exploitée par le LLM lors de la génération du texte. Lors de la *récupération*, l'entrée (demande de l'utilisateur) est placée dans l'espace vectoriel. Un calcul de similarité est ensuite réalisé entre ce vecteur et les clés stockées dans la base de connaissances. Les données les plus proches sont ainsi sélectionnées. Enfin, la *génération* se base sur les données retournées afin de générer du contenu.

S'inscrivant dans une démarche d'amélioration de la RAG naïve, la RAG avancée apporte une meilleure récupération des données. Cela se traduit par un travail sur le modèle de vectorisation des données. Plus ce modèle est adapté au contexte, plus les informations récupérées seront pertinentes et la compréhension du contexte accrue. Par ailleurs, la RAG avancée apporte deux nouvelles étapes de pré-récupération et de post-récupération. Le processus de *pré-récupération* constitue une optimisation de la méthode d'indexation des données au sein de la base de connaissances. La *post-récupération* quant à elle apporte des méthodes afin de maximiser les chances que le LLM performe en sortie. Cela implique une phase de *reclassement* des informations récupérées, visant à promouvoir les données les plus pertinentes pour répondre de manière adéquate à la requête de l'utilisateur. Enfin, la commande finale est reformulée avant d'être transmise au LLM, réduisant ainsi les risques de bruit.

Enfin, la RAG modulable apporte quant à elle beaucoup plus de flexibilité concernant l'architecture. Elle introduit le concept de modules qui peuvent venir se greffer ou se substituer à certaines étapes de la RAG. On trouve ainsi des modules de recherche, de mémoire, de fusion, de routage, de prédiction ou encore d'adaptation à la tâche.

L'utilisation de la RAG pour la génération de texte demeure sujette à plusieurs limitations décrites dans le papier "Seven Failure Points When Engineering a Retrieval Augmented Generation System" (Barnett *et al.*, n.d.)

– *Contenu manquant*: il se peut que la base de connaissances ne contienne pas les informations adéquates pour répondre à la question de l'utilisateur. En pratique, les sorties produites sont plausibles mais erronées.

– *Mauvaise récupération*: les données retournées lors de la récupération peuvent ne pas correspondre aux informations les plus pertinentes pour répondre à l'entrée initiale.

– *Manque de compréhension du contexte*: le contexte transmis au LLM peut s'avérer trop restreint ou au contraire trop large, induisant respectivement une absence de réponse ou des hallucinations.

– *Format incorrect*: la sortie n'est pas dans le format attendu.

– *Mauvais niveau de spécificité*: le contenu généré est trop général par rapport au contexte.

De nombreuses limitations constatées dans la RAG peuvent être liées à la manière dont l'information est représentée, tant dans les documents formant la base de connaissance que dans la question posée par l'utilisateur. C'est pourquoi une nouvelle approche est apportée afin de représenter les données : les graphes de connaissances. En effet, ces structures de données permettent de modéliser et d'organiser les informations sous forme de réseaux sémantiques composés de nœuds représentant les entités et reliés par des arêtes définissant leurs relations. Cette représentation offre une formalisation explicite et structurée des connaissances, facilitant leur gestion, leur raisonnement et leur exploitation.

Cet article a pour objectif de présenter des travaux alliant la RAG et les graphes de connaissances ainsi qu'un positionnement pour mes futurs travaux de recherche.

2. Orientation des travaux

De récents travaux ont été menés quant à l'utilisation des graphes de connaissances appliqués au RAG. On retrouve notamment GraphRAG (Potts, n.d.) qui a été introduit en février 2024 par Microsoft Research. Son principe est de créer une base de connaissance sous forme de graphe à l'aide d'un LLM. Cette base est par la suite interrogée par un module de récupération de données orienté graphe et est exploitée avec un module de génération compatible avec les graphes de connaissances. Bien que très prometteur, cette technologie présente des limitations, principalement axées sur le graphe. En effet, la construction du graphe de connaissances et la formulation des requêtes pour l'interroger peuvent s'avérer complexes. De même, la définition des bornes pour la

récupération des nœuds pertinents est difficile. Des bornes trop restrictives nuiraient à la compréhension du contexte, tandis que des bornes trop larges introduiraient du bruit indésirable.

G-Retriever est un module développé par (He *et al.*, n.d.) en février 2024 permettant d'interagir avec un graphe de connaissances (Graph QA). Il allie les atouts des réseaux de neurones sur graphes (GNN), des LLMs et de la RAG. Le principe est de projeter un graphe de connaissances dans un espace vectoriel à l'aide d'un modèle de langage pré-entraîné (indexation). Lors de la récupération, les nœuds les plus proches de la requête utilisateur sont extraits et servent à construire un sous-graphe de connaissances en se basant sur le problème d'optimisation d'arbre de Steiner. Ce sous-graphe reçoit ensuite un traitement visant à le transformer en un texte interprétable par le LLM. Ainsi, la génération de la réponse s'appuie sur ce texte. Contrairement à GraphRAG, G-Retriever peut bénéficier d'un entraînement afin d'optimiser ses performances. Néanmoins, il demeure plus sensible aux variations induites par les transformations de représentation texte vers graphe ou graphe vers texte.

De nombreux travaux mettant en œuvre l'utilisation de graphes de connaissances dans le cadre de la RAG ont été publiés au cours des dernières semaines. Cet article ne les répertorie pas tous mais présente simplement certains cas d'utilisation. Mes prochaines recherches se concentreront sur l'impact que peuvent avoir les graphes de connaissances à chaque étape de la RAG, ainsi que sur les modalités d'intégration de ces graphes dans les différents modules.

3. Bibliographie

Barnett S., Kurniawan S., Thudumu S., Brannelly Z., Abdelrazek M., "Seven Failure Points When Engineering a Retrieval Augmented Generation System", n.d.

Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., Dai Y., Sun J., Guo Q., Wang M., Wang H., "Retrieval-Augmented Generation for Large Language Models: A Survey", n.d.

He X., Tian Y., Sun Y., Chawla N. V., Laurent T., LeCun Y., Bresson X., Hooi B., "G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering", n.d.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S., Kiela D., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", n.d. version: 1.

Potts B., "GraphRAG: A new approach for discovery using complex information", n.d.