INFORSID

FORUM JCJC
2 juin 2022, Dijon

v

Forum Jeunes Chercheuses Jeunes Chercheurs

Actes de la 11ème édition

Dijon, Juin 2022

Table de Matières

Introduction au Forum Jeunes Chercheuses Jeunes Chercheurs d'INFORSID 2022

Elena Kornyshova

CEDRIC, Conservatoire National des Arts et Métiers
2 rue Conté, 75003 Paris, France
elena.kornyshova@cnam.fr

La douzième édition du Forum Jeunes Chercheuses Jeunes Chercheurs (JCJC) de la conférence INFORSID s'est déroulée en juin 2022 à Grenoble. Le Forum JCJC permet aux jeunes chercheuses et jeunes chercheurs en première ou deuxième année de doctorat de présenter leurs problématiques de recherche, d'établir des contacts avec des équipes travaillant sur des domaines identiques ou connexes, d'offrir un aperçu des axes de recherche actuels et ainsi d'élargir le champ de leurs connaissances.

L'objectif de ce Forum étant d'accompagner les doctorants dans leurs premiers pas de chercheurs, le processus de sélection a été élaboré sur des principes de bienveillance et de conseil. Douze articles ont été soumis, tous dans les thématiques d'INFORSID. Chaque article a reçu deux relectures et une méta-relecture. Deux articles ont été acceptés avec les modifications mineures, cinq articles ont été acceptés avec des modifications de niveau moyen et cinq articles ont été acceptés conditionnellement. Tous les articles ont reçu une relecture de leur version finale.

Tous les douze jeunes chercheuses et jeunes chercheurs ont présenté leurs travaux de thèse lors du Forum JCJC. Dans ses travaux de recherche, Chloé Vigneau propose une approche qui permet d'évaluer les activités de création des jeux vidéo à vocation scolaire en classe. Antoine Aubé développe une approche aidant la conception de migrations performantes des logiciels vers un nuage public en se fondant sur les exigences. Le travail d'Anouck Chan vise à élaborer une méthode de conception simultanée d'un produit et de son système de production dans le domaine aéronautique. Une méthode pour identifier les meilleurs partenaires dans un écosystème d'affaires numériques en s'appuyant sur les intentions des utilisateurs et des informations de contexte à l'aide d'algorithmes de machine learning est développée par Mustapha Kamal Benramdane. Aly Benjilany étudie l'alignement opérationnel entre le métier et les systèmes logiciels. Maël Lesavourey propose une approche facilitant la recherche d'information en biomédecine en élaborant une représentation sémantique multi-contextuelle de la terminologie. La thèse de Solweig Hennechart vise à établir une base de données intégrée portant sur la chimiodiversité des produits naturels couvrant tout le vivant en complétant les

données marquantes avec les modèles prédictifs. Le projet de recherche de Maxime Masson s'appuie sur les données générées par des utilisateurs sur les réseaux sociaux afin de proposer des services augmentés pour un tourisme intelligent. Ramona Elali propose une méthode de recommandations contextuelles en se basant sur la fouille des intentions et les ontologies de domaine. Dans sa thèse, Kaoutar Sadouki s'intéresse aux problèmes d'adaptation des composants de l'industrie 4.0 aux objectifs des utilisateurs internes et externes. L'utilisation des réseaux de neurones en graphes (GNN) pour une tâche de classification fait l'objet des travaux d'Ikram Boukharouba. Olivier de Casanove travaille sur la gestion des incidents de sécurité à l'échelle inter-systèmes en étudiant l'impact du facteur humain.

La présentation des travaux s'est déroulée sous le format « Dragons et Chevaliers ». Pour chaque thésard, il y avait trois rôles : présentateur, dragon et chevalier. Un présentateur a été tiré aléatoirement et présente son travail. Puis un dragon a prodigué une critique constructive de ce qui venait d'être présenté. Un chevalier a expliqué ensuite à tous pourquoi ce qui a été présenté était vraiment excellent. Nous avons itéré ce processus pour chacun des jeunes présentant ses travaux. Chaque présentation durait 4-5 minutes. Chaque dragon donnait trois critiques constructives et chaque chevalier présentait trois aspects positifs. Tous les doctorants ont joué tous les rôles. Ce format leur a permis non seulement de présenter leur travail de recherche et d'avoir des retours, mais également de développer leurs capacités analytiques sur les travaux des autres chercheurs. En plus, le format « Dragons et Chevaliers » a contribué à des discussions très riches et à une bonne ambiance lors du Forum. Je tiens à remercier nos auteurs pour leurs contributions scientifiques et leur participation très engagée dans les sessions du Forum JCJC.

Je souhaite également remercier les personnes suivantes qui ont consacré du temps pour fournir à nos jeunes chercheuses et jeunes chercheurs des commentaires très constructifs qui leur ont permis de progresser, à savoir les membres du comité du programme du Forum JCJC d'INFORSID 2022 :

– *Lylia Abrouk*, Université de Bourgogne, France,
– *Iness Ahriz*, Conservatoire National des Arts et Métiers, France,
– *Isabelle Astic*, Conservatoire National des Arts et Métiers, France,
– *Judith Barrios Albornoz*, University of Los Andes, Venezuela,
– *Samia Bouzefrane*, Conservatoire National des Arts et Métiers, France,
– *Guillaume Cabanac*, Université Toulouse 3, France,
– *Rebecca Deneckere*, Université Paris 1, France,
– *Agnès Front*, Université Grenoble, France,
– *Eric Gressier-Soudan*, Conservatoire National des Arts et Métiers, France,
– *Elisabeth Métais*, Conservatoire National des Arts et Métiers, France,
– *Thomas Polacsek*, Onera, France,
– *Irina Rychkova*, Université Paris 1, France,
– *Nathalie Vallès-Parlangeau*, Université Toulouse 1 Capitole, France.

J’aimerais tout particulièrement remercier le bureau d’INFORSID de m’avoir fait confiance pour organiser ce Forum, pour gérer le contenu scientifique et la préparation des sessions en présentiel, ainsi que pour le partage des pratiques d’organisation des années précédentes afin de fournir une meilleure expérience à nos jeunes chercheuses et jeunes chercheurs.

Elena Kornyshova

Responsable du Forum JCJC d’INFORSID 2022

Conception d'un système d'évaluation d'une activité de création de vidéo pour les enseignants de seconde

Chloé Vigneau

Chaire Science et Jeu vidéo, Laboratoire Leprince-Ringuet, Ecole Polytechnique, Route de Saclay, 91120 Palaiseau
Laboratoire CEDRIC, CNAM, 2 rue conté 75003 PARIS
chloe.vigneau@polytechnique.edu

MOTS-CLES : Jeu vidéo, Evaluation, Projet interdisciplinaire.

KEYWORDS : Video game, Assessment, Interdisciplinary project.

ENCADREMENT : Axel Buendia, Stéphanie Mader, Catherine Rolland.

1. Contexte

Notre sujet porte sur l'évaluation d'une activité de création de jeu vidéo en classe. D'après les retours d'expérience, cette modalité d'intégration du jeu vidéo a un impact bénéfique sur l'apprentissage et la motivation des élèves (Reynolds *et al*, 2011). Cependant, il n'est pas toujours facile pour les enseignants de prendre en main et d'évaluer cette activité qui mobilise des compétences et connaissances de différentes disciplines, c'est-à-dire « *de mesurer le degré d'acquisition des connaissances et des compétences ainsi que la progression de l'élève* [1]». La compétence est définie comme la « *capacité d'action efficace face à une famille de situations, qu'on arrive à maîtriser parce qu'on dispose à la fois des connaissances nécessaires et de la capacité de les mobiliser à bon escient* » (Perrenoud, 2011). Nous avons observé les outils informatiques utilisés au cours de l'activité de création de jeu vidéo et lors de son évaluation dans quatre classes de lycée en France : les moteurs de création de jeux et la plateforme d'évaluation Pronote. Parmi les difficultés relevées, nous avons choisi de nous focaliser sur la récolte et le traitement de données pertinentes dans le cadre de l'évaluation de cette activité.

[1] Loi n°2013-595 du 8 juillet 2013 d'orientation et de programmation pour la refondation de l'école de la République

2. Etat de l'art

L'atelier de création de jeu vidéo est une « *situation d'apprentissage informatisée* (SAI) », c'est-à-dire « *une situation d'apprentissage intégrant un ou plusieurs logiciels qui y jouent un rôle particulier du point de vue de l'apprentissage* » (Tchounikine, 2009). Le logiciel central de l'activité est le moteur de création de jeu (Wu et Wang, 2012) qui est la source d'information principale sur le travail effectué par les élèves à condition que l'enseignant puisse analyser des « *traces d'apprentissage* » (Boyer, 2019). Il s'agit donc de convertir des « *traces numériques* », « *enregistrement automatique d'éléments d'interaction entre un utilisateur et son environnement, dans le cadre d'une activité donnée* » (Laflaquière et Prié, 2007) en « *indicateurs* », « *variable au sens mathématique à laquelle est attribuée une série de caractéristiques* » (Dimitracopoulou et Bruillard, 2006). Pour permettre l'évaluation, il faut aussi pouvoir stocker et traiter ces données. Or, il n'existe pas d'EIAH (environnement informatique pour l'apprentissage humain) adapté à cette activité donc aucun « *logiciel spécifiquement conçu dans le but d'amener un apprenant à développer une activité favorable à l'atteinte des objectifs de la* […] *SAI considérée* » (Tchounikine, 2009). Enfin, en l'absence de connaissance des processus de production du jeu vidéo, les enseignants ne peuvent pas effectuer d'évaluations formatives dans le but d'identifier les forces et faiblesse de l'élève, de lui donner des recommandations en vue de l'aider à progresser (Andrade et Cizek, 2009). Ils ne parviennent pas toujours non plus à faire le lien entre l'activité et les compétences visées, ce qui rend aussi difficile l'évaluation sommative qui reconnaît l'atteinte des objectifs d'apprentissage par les étudiants (Legendre, 1993).

3. Problématique

Nous avons vu la limite des outils informatiques utilisés dans les processus de production et d'évaluation d'une activité de création de jeu vidéo. Nous souhaitons proposer une solution pour résoudre ces difficultés et répondre aux deux questions suivantes : (1) comment identifier des données pertinentes à collecter et les mettre en place ? et (2) comment traiter ces données pour en faire des informations signifiantes pour suivre et organiser la progression des élèves ainsi que pour réaliser l'évaluation de compétences (évaluations formatives et sommatives) ?

4. Actions réalisées

Nous avons réalisé un modèle de donnée théorique permettant de répondre aux deux questions identifiées. Pour commencer, nous nous appuyons sur des ouvrages de référence en game design et en production de jeu vidéo pour décrire le processus de réalisation d'un jeu vidéo (Schell, 2014). Un jeu se compose d'unités de conception ou briques fonctionnelles que les élèves ou l'enseignant souhaitent intégrer dans leur jeu (par ex : créer un personnage). Pour créer ces unités de conception, les élèves accomplissent des tâches réparties par rôle (ex :

programmation, graphisme). Ces unités de conception évoluent au fur et à mesure de l'avancement du projet donc de la réalisation des tâches : elles prennent différents états d'une phase de projet à l'autre. Chaque tâche mobilise une ou plusieurs compétences qui peuvent être associées à des disciplines scolaires et/ou à des connaissances issues des programmes scolaires. Nous pouvons vérifier la réalisation des tâches en comparant les changements d'état des unités de conception d'une phase à l'autre. Pour contrôler ces changements d'états nous devons implémenter des données associées à chaque tâche qui fonctionneront comme des traces d'apprentissage. Ces données ne peuvent être vérifiées que dans un environnement maîtrisé. Nous proposons donc de mettre en place des templates de jeux composés d'un certain nombre d'unités de conception qui sont toutes dans le même état. Le passage d'un état à l'autre, donc la bonne réalisation de la tâche, peut être validé en s'appuyant sur les données accessibles dans les templates. Nous nous appuyons sur deux hypothèses de travail : (1) la vérification des tâches d'une phase à l'autre du projet va permettre de suivre la progression des élèves donc de réaliser des évaluations formatives et (2) la complétion de toutes les tâches rattachées à une compétence valide la compétence et sert donc de base à l'évaluation sommative.

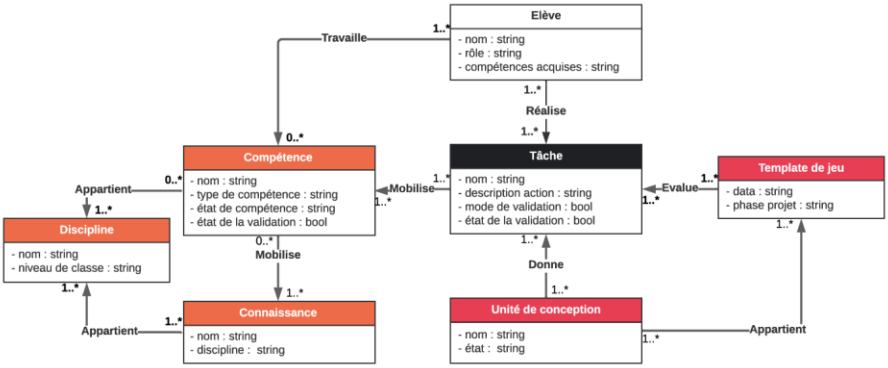

Figure 1. Schéma du système d'évaluation de l'activité de création de jeu vidéo

Nous avons classé les tâches en fonction de leur mode de validation : « automatique », où l'enseignant reçoit un rapport sur l'état de la tâche (validé ou non) en fonction des données récoltées, et « libre », où il visualise les données associées à chaque tâche et réalise lui-même l'évaluation.

Tableau 1. Instanciation du système d'évaluation

Compétences	**Tâches**	**Actions à réaliser**	**Data**	**Mode de validation**
Programmer	Intégrer la vie du personnage	Créer variable	Variable	Automatique (test variable)
Rédiger	Ecrire un dialogue	Remplir zone texte	Texte	Manuelle (évaluation texte)

5. Actions futures

Pour mettre en place ce système d'information et vérifier sa pertinence dans le cadre de l'évaluation de l'activité de création de jeu vidéo, nous avons besoin de créer une base de données et de développer des templates de jeux comportant des données associées à chacune des tâches. Nous souhaitons également approfondir le concept de validation d'une tâche. En effet, l'activité de création de jeu vidéo s'appuyant sur une méthode d'apprentissage par « essai-erreur » (Lafarge, 2007), il serait intéressant de prendre en compte ce concept dans le suivi de la progression des élèves soit en pondérant les résultats de l'évaluation du nombre d'essais effectués pour réaliser la tâche, soit en constituant une banque d'erreurs possibles également associées à des données pour être tracées.

Bibliographie

Andrade, H. L, et Cizek, G. J. (2009). *Handbook of formative assessment*. New York and London: Routledge

Boyer, A. (2019). Quelques réflexions sur l'exploration des traces d'apprentissage. *Distance et Médiation des Savoirs*, n° 27.

Dimitracopoulou, A. et Bruillard, E. (2006). Enrichir les interfaces de forums par la visualisation d'analyses automatiques des interactions et du contenu. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation. ATIEF*, 13

Lafarge, V. (2007). L'analyse d'erreurs comme outil dans la détermination des aides proposées par un EIAH. *Apprentissage des langues et systèmes d'information et de communication*, vol 10, n°1

Legendre, R. (1993). *Dictionnaire actuel de l'éducation*. Montréal : Éditions Guérin.

Laflaquière, J. et Prié, Y. (2007). Des traces modélisées, un nouveau support pédagogique ?, *Actes de la 4e conférence scientifique de Lornet*

Perrenoud, P. (2011). *Construire des compétences dès l'école. Pratiques et enjeux pédagogiques*, ESF, Pari**s.**

Reynolds R., Harel Caperton, I. (2011). Contrasts in student engagement, meaning-making, dislikes, and challenges in a discovery-based program of game design learning. *Education, Tech Research Dev.* n° 59, pp 267-289.

Schell, J. (2014). *The Art of Game Design : A Book of Lenses*. A K Peters/CRC Press.

Tchounikine, P. (2009). Précis de recherche en EIAH. En ligne

Wu, B. et Wang, A-I. (2012). A guideline for game development-based learning: a literature review. International Journal of Computer Technology.

Co-conception de logiciels et de leur environnement d'exécution infonuagique

Antoine Aubé

Office National d'Études et de Recherches Aérospatiales
2, avenue Édouard Belin, 31055 Toulouse, France
antoine.aube@onera.fr

MOTS-CLES : Informatique en nuage, Conception simultanée, Ingénierie des exigences.

KEYWORDS : Cloud Computing, Co-design, Requirements engineering.

ENCADREMENT : Thomas Polacsek, Clément Duffau.

1. Contexte

L'*informatique en nuage* (en anglais : « *Cloud Computing* ») est défini comme un modèle d'accès à des ressources informatiques exposés sous la forme de services accessibles via l'Internet au sein d'un catalogue appelé *nuage* (en anglais : « *cloud* ») qui assure élasticité et observabilité (Mell, Grance, 2011). Les nuages publics sont ouverts à toutes les organisations et facturent leurs services à l'usage, ce qui permet aux organisations qui hébergent leurs systèmes sur ces nuages de régler finement les coûts de leur infrastructure et leur épargne des investissements. Pour cette raison, nombre d'entre elles migrent leurs systèmes informatiques vers un nuage public.

Dans un système infonuagique, des logiciels (e.g. serveurs web, bases de données) sont hébergées sur un environnement infonuagique, constitué de ressources (e.g. machines virtuelles) déployées par la configuration de services. Les choix des services et de leur configuration sont faits pour répondre aux exigences du système et de la migration, et peuvent rendre des adaptations des logiciels nécessaires. À ce titre, certains logiciels peuvent être remplacés par des logiciels similaires fournis sous forme de services. Zhao et Zhou (2014) suggèrent que la migration d'un système en *SaaS*[2] requiert la migration au préalable de son architecture pour qu'elle soit orientée services. Enfin, de nombreux services ont des interfaces propriétaires et

[2]1. Service de logiciel à la demande (en anglais : « Software as a Service », *SaaS*), qui déploie une application prête à l'usage (e.g. Outlook, Google Workspace).

non-interopérables avec d'autres services ou logiciels similaires. C'est le cas de nombreux services de bases de données non-relationnelles infogérées, qui chacun a un modèle de fonctionnement singulier et impose des contraintes aux logiciels qui interagissent avec lui (e.g. usage d'une bibliothèque dédiée). Des effets de bord sur les exigences peuvent accompagner ces adaptations : remise en question d'exigences jusqu'ici satisfaites par le système, tâches de développement incompatibles avec les exigences de la migration.

Nos travaux de recherche visent à concevoir un cadre méthodologique pour la co-conception de logiciels et de leur environnement d'exécution infonuagique dans le cadre d'une migration.

2. État de l'art

Plusieurs travaux caractérisent les migrations vers un nuage. Il en découle des catégorisations de *stratégies de migration*, qui regroupent les migrations suivant divers critères (e.g. activités mises en œuvre, types de services utilisés). Zhao et Zhou (2014) présentent une revue de ces catégorisations et proposent des pistes de travaux à mener. Parmi ces dernières, ils constatent que les travaux sur les migrations se concentrent chacun sur une stratégie de migration en particulier, mais il manque une méthode holistique permettant à une organisation de décider de la stratégie de migration à employer.

Il existe de nombreux travaux sur la sélection de services et de leur configuration. Par exemple, Quinton (2014) propose une approche basée sur les lignes de produits logiciels pour choisir et configurer les services afin qu'ils correspondent aux besoins d'un logiciel fixé en amont. Ferry *et al.* (2013) présentent *CloudML* qui présentent un langage de modélisation des *IaaS*[3] agnostique de tout nuage, et un système permettant d'assister, entre autres, la configuration et la mise à l'échelle d'un environnement infonuagique distribué sur plusieurs nuages. *COCA-PT* est un outil qui automatise le choix de *IaaS* et leur configuration en tenant compte de leurs coûts ; il repose sur l'observation d'un système pour constituer un modèle de consommation des ressources (e.g. durée d'allocation d'une machine virtuelle) qu'il le transpose sur un environnement cible pour en estimer le coût (Belli *et al.*, 2016). *CoCoOn* est une ontologie des *IaaS* qui a été utilisée dans un système de recommandation des services (M. Zhang *et al.*, 2012), puis a été étendue pour capturer les aspects performance et tarification (Q. Zhang *et al.*, 2019). Ces approches reposent sur les contraintes imposées par les logiciels et sont à ce titre adaptés pour les migrations sans adaptation de ces logiciels qui ne permettent pas de tirer pleinement profit des avantages de nuages publics, comme l'élasticité des ressources.

[3]2. Service d'infrastructure à la demande (en anglais : « Infrastructure as a Service », *IaaS*), qui déploie des machines, disques et réseaux

3. Problématique

Si la littérature compte plusieurs approches répondant à des préoccupations très techniques, peu de travaux abordent les besoins des utilisateurs des systèmes informatiques. De fait, il manque de travaux centrés sur les exigences qui dirigent la migration vers un nuage public.

La satisfaction de ces exigences conditionne la réussite d'une migration. Cependant, des retours recueillis au sein de *Stack Labs*, société de conseil spécialisée dans les migrations infonuagiques, indiquent que ces exigences sont souvent insuffisamment élicitées et que certaines d'entre elles sont difficiles à prendre en compte (e.g. concernant le budget, la sécurité, les performances) dans les choix de conception d'un environnement. Ces difficultés sont en lien avec les caractéristiques-clés de l'informatique en nuage (e.g. l'élasticité) et des nuages publics (e.g. facturation à l'usage). Or, des choix malheureux peuvent avoir de grandes répercussions négatives (e.g. coûts importants, failles de sécurité) qu'il est coûteux de corriger. C'est pour ces raisons que nous souhaitons aborder les phases d'élicitation et d'analyse des exigences, et de conception des migrations infonuagiques.

L'objectif de nos recherches est de déterminer comment aider à concevoir la migration infonuagique la plus performante. Nous devons, pour cela, définir ce qui fait la performance d'une migration en comprenant et identifiant, à partir de retours industriels, ce qui est important dans le cadre d'une migration ***(i)***. L'aide apportée, pour être utilisable dans l'industrie, prendra la forme d'outils automatiques pour évaluer les qualités d'une conception préliminaire d'un système infonuagique ***(ii)***.

4. Travaux réalisés

Étude des exigences lors d'une migration vers un nuage public – Nous avons mené des entretiens auprès de spécialistes qui effectuent des migrations dans les entreprises afin d'identifier les préoccupations qui dirigent la sélection d'un environnement infonuagique (verrou ***(i)***). Une première analyse de ces entretiens indique la réduction des coûts comme principale raison de migrer, et des difficultés à tenir compte de l'élasticité dans l'estimation des coûts. Nous avons identifié un ensemble d'exigences de haut niveau qui dirigent le choix des services et de leur configuration.

Estimation des coûts d'un environnement infonuagique – Les entretiens ont mis en lumière des difficultés à estimer les coûts d'un environnement. Actuellement, ces estimations reposent soit sur un avis d'expert non chiffré, soit sur des tâches chronophages qui poussent à la négligence ; or, la mauvaise maîtrise des coûts qui découle de cette négligence sont la source d'insatisfactions suite à une migration.

Nous sommes ainsi en train d'aborder le verrou ***(ii)*** sous l'angle des coûts en développant un outil de simulation de l'utilisation et de la facturation d'un environnement infonuagique. La simulation repose, d'une part, sur la description d'un système infonuagique (environnement et logiciels hébergés) grâce à la modélisation préalable du catalogue des services et de leur tarification et, d'autre

part, sur la description du système et des interactions des utilisateurs avec ce système, sous forme de scénarios.

4. Travaux futurs

Étude des exigences lors d'une migration vers un nuage public – Sur la base des résultats des entretiens, nous avons préparé un sondage que nous sommes en train de diffuser afin de confronter les exigences de haut niveau identifiées à un grand nombre de retours de professionnels et les compléter : identification d'autres exigences, établissement de liens de priorités entre elles (verrou ***(i)***).

Estimation des coûts d'un environnement infonuagique – Nous travaillons actuellement à intégrer l'élasticité de l'environnement dans l'outil présenté ci-dessus afin de lever les difficultés liées au manque de compréhension des coûts liés à l'élasticité (verrou ***(ii)***). En effet, s'il est aisé de calculer les coûts d'un système composé de ressources non-élastiques (e.g. machine virtuelle), calculer ceux d'un système avec des ressources élastiques (e.g. groupe de machines virtuelles infogéré) demande de prendre en compte la charge générée par les utilisateurs à tout instant. Il faut pour cela comprendre le comportement des utilisateurs du système, les interactions entre les composants de ce même système et les règles qui automatisent la mise à l'échelle des ressources.

Bibliographie

Belli O., Loomis C., Abdennadher N. (2016). Towards a cost-optimized cloud application placement tool. *IEEE International Conference on Cloud Computing Technology and Science*, p. 43-50.

Ferry N., Rossini A., Chauvel F., Morin B., Solberg A. (2013). Towards model-driven provisioning, deployment, monitoring, and adaptation of multi-cloud systems. *2013 IEEE Sixth International Conference on Cloud Computing*, p. 887-894.

Lamsweerde A. van (2001). Goal-oriented requirements engineering: A guided tour. *5th IEEE International Symposium on requirements engineering*, p. 249.

Mell P., Grance T. (2011). *The NIST definition of Cloud Computing*. Rapport technique n°800-145.

Quinton C. (2014). *Cloud Environment Selection and Configuration: a software product lines-based approach*. Thèse de doctorat, Université Lille I.

Zhang M., Ranjan R., Nepal S., Menzel M., Haller A. (2012). A declarative recommender syste for cloud infrastructure services selection. *Economics of grids, clouds, systems, and services – 9th International Conference*, vol. 7714, p. 102-113.

Zhang Q., Haller A., Wang Q. (2019). CoCoOn: Cloud Computing Ontology for IaaS price and performance comparison. *18th International Semantic Web Conference*, part. II, vol. 11779, p. 325-341

Zhao J., Zhou J. (2014). Strategies and Methods for cloud migration. *Int. J. Autom. Comput.*, vol. 11, p. 143-152.

Aide à la co-conception de produits complexes

Anouck Chan

ONERA/DTIS, Université de Toulouse,
F-31055 Toulouse, France
anouck.chan@onera.fr

MOTS-CLÉS : Ingénierie des exigences, Modélisation, Optimisation, Formalisation.

KEYWORDS : Requirements engineering, Modelling, Optimisation, Formalisation.

ENCADREMENT : Thomas Polacsek, Stéphanie Roussel.

1. Contexte

La création d'un produit complexe, comme un avion ou une constellation de satellites, implique non seulement la conception du produit lui-même, mais aussi de son moyen de production (son système industriel, son usine, ou plus particulièrement sa ligne d'assemblage). Cette nécessité de double conception est motivée par deux facteurs : le premier concerne les choix de design du produit qui conditionnent le moyen de production. Par exemple, le choix du matériau du fuselage d'un avion peut impliquer l'usage et l'achat de machines spécifiques. Le second facteur est le fait que le produit et son système industriel possèdent des objectifs pouvant s'influencer entre eux de manière positive ou négative. Un choix de design côté produit pourrait empêcher l'usine d'atteindre ses objectifs en termes de vitesse production. De même, certaines exigences du produit et de son système industriel peuvent s'avérer incompatibles, s'opposant notamment à travers les méthodes choisies pour les accomplir. Dans de telles situations, des compromis doivent être effectués afin d'obtenir un ensemble d'exigences cohérent et satisfaire les objectifs des deux partis de manière optimale.

Dans le domaine aéronautique, le cycle de développement produit/moyen de production est traditionnellement *séquentiel*. C'est à dire que l'avion est tout d'abord conçu, puis sa ligne d'assemblage. Cette approche comporte plusieurs inconvénients présentés dans (Polacsek et al., 2017). En effet, les impacts des choix de design de l'avion sur son système industriel ne sont évalués qu'après la conception finale de l'avion. Il est ainsi possible que certains designs s'avèrent difficilement réalisables ou même que cela soit impossible. Dans de telles situations, il est alors nécessaire que les concepteurs de l'avion modifient leurs plans, puis les transmettent de nouveau aux concepteurs du système industriel, et ceci jusqu'à obtenir une solution

constructible et satisfaisante. Ce processus peut s'avérer long et coûteux, et s'achever sur une solution globale non optimale. Les auteurs proposent donc la méthode d'ingénierie simultanée, où le produit et son moyen de production sont conçus en même temps. Cette approche permettrait de détecter les situations de blocage plus précocement dans le processus et ainsi d'y remédier plus aisément.

2. Problématique

Nous nous intéressons à élaborer une aide à la conception simultanée de produits complexes. Pour cela plusieurs conditions sont nécessaires, dont notamment : (i) la capacité des concepteurs du produit à pouvoir estimer les conséquences de leurs choix sur le système industriel à différentes étapes de la conception, et en particulier dès les phases précoces, (ii) pouvoir identifier et utiliser les opportunités qu'offrent le système industriel à la conception du produit, (iii) définir des critères d'évaluation, notamment des objectifs du produit et de son système de production et enfin (iv) élaborer de méthodes de mesures des influences d'un choix de conception sur ces critères. Les objectifs des travaux de cette thèse est de proposer une méthode qui permette d'éliciter les objectifs des concepteurs produits et du système industriels, de caractériser et d'évaluer les influences que peuvent avoir leur satisfaction les unes sur les autres et enfin de proposer des outils et des méthodes d'aide à la co-conception.

3. Etat de l’art

En ingénierie des exigences plusieurs auteurs se sont intéressés à éliciter et définir des ensembles consistants et complets d'objectifs comme (van Lamsweerde, 2001; Kavakli, Loucopoulos, 2005). Les modèles orientés buts *(Goal-Oriented Requirements Engineering, GORE)* comme *IStar 2.0* (Dalpiaz et al., 2016), *Kaos* ou *Techne* (Jureta et al., 2010) permettent de représenter graphiquement les acteurs et leurs exigences notamment en organisant et liant les objectifs et leurs moyens de réalisation. Ces modèles assistent la construction et l'évaluation de solutions (Elahi, Yu, 2011), c'est à dire d'ensembles d'éléments permettant la satisfaction des objectifs des acteurs. Pour cela, une méthode est de propager les conséquences de la satisfaction d'éléments (Mylopoulos et al., 1992; Zhang, Wang, 2019). Cette propagation permet également de détecter des conflits entre les exigences. Ces derniers peuvent être résolus par négociation entre les objectifs (Elahi, Yu, 2011), notamment via des méthodes d'argumentation comme proposé dans (Murukannaiah et al., 2015; ElRakaiby et al.,2020).

4. Actions réalisées

Dans un premier temps, nous nous sommes intéressés à la représentation en modèles orientés buts de la conception d'un avion et de sa ligne d'assemblage. Nous nous avons étudié la conception traditionnelle séquentielle et celle simultanée où l'on recherche un optimal avion/ligne d'assemblage. Les objectifs de ces modèles sont d'éliciter les exigences des concepteurs d'avions et des concepteurs de la ligne

d'assemblage et de définir les relations de dépendances entre ces deux acteurs. Cela met en exergue un blocage dû à la dépendance cyclique des deux acteurs, c'est à dire, qu'aucun d'eux ne peut satisfaire ses exigences car il a besoin que l'autre agisse en premier pour cela.

Cette situation s'approche de celles de coopétition décrites dans (Pant, 2021), c'est à dire une situation où les acteurs sont en compétition par certains aspect (ici, dans la satisfaction de leurs exigences personnelles) et en coopération dans d'autres (ici, satisfaire l'objectif commun). Dans (Pant, Yu, 2019), une solution est proposée par l'auteur pour obtenir une situation gagnante-gagnante, où tous les acteurs sont satisfaits. Il s'agit d'ajouter un acteur intermédiaire. Nous avons ainsi introduit un acteur facilitateur de co-conception nommé *global designers* (Aquieta-Nuñez et al., 2021). Ce dernier est capable de négocier entre les exigences des deux autres acteurs et ses objectifs (buts) combinent ceux des acteurs initiaux. Cette introduction nous permet de briser le cercle de dépendances. Les compétences, les actions et la caractérisation de l'entité facilitatrice de co-conception comme acteur, ne sont néanmoins pas définies.

À l'aide d'outils de recherche opérationnelle, et notamment la modélisation sous contraintes, nous avons pu évaluer différents processus de construction d'avion, en fonction d'exigences de performance du système industriel. Cela nous a permis de définir des critères d'évaluation qui, selon les choix de design côté avion, pouvaient être satisfaits ou non (Chan et al., 2022).

Nous nous sommes intéressés ensuite aux situations de co-conception dans d'autres domaines que l'aéronautique, notamment en architecture. Cela nous a permis de généraliser les exigences du concepteur produit, de celui du système industriel et des différentes attentes/dépendances qu'il pourrait exister entre eux, et quelles caractéristiques l'entité facilitatrice de co-conception devrait posséder.

Au vu de ces résultats, nous nous sommes penchés sur l'approche présentée dans (Bryl et al., 2006) pour répartir des tâches/buts d'un système global entre ses différents acteurs notamment en permettant aux acteurs de déléguer (transmettre la responsabilité de la satisfaction) un but qu'ils ne savent pas satisfaire à un acteur plus approprié. Nous avons adapté la méthode pour obtenir à partir de buts globaux et haut (voir très haut) niveau, un ensemble de buts plus simples ainsi que l'ensemble d'acteurs capable de les satisfaire. L'application de cette méthode à notre problème de co-conception nous a notamment permis d'éliciter certains objectifs *clés*, c'est à dire qui pourraient nous aider à définir plus concrètement l'entité facilitatrice de co-conception.

5. Actions futures

Les travaux réalisés à ce jour nous ont permis de distinguer quelques pistes pour définir les caractéristiques de l'entité facilitatrice de co-conception d'un produit complexe et de son système de production. Une future action pourra ainsi consister à étudier les objectifs clés trouvés grâce à la méthode de délégation. Les relations entre les acteurs, et notamment celles de dépendances pourront également être

approfondies et enrichies, à la lumière des résultats obtenus avec cette même méthode. Nous pourrons aussi, au moyen d'outils de recherche opérationnelle, proposer une aide à la conception de systèmes industriels optimaux, et en particulier de lignes d'assemblage, en fonction des critères et des objectifs du produit et de la production.

Bibliographie

Aquieta-Nuñez A., Chan A., Donoso-Arcinieg A., Polacsek T., Roussel S. (2021, 11). A collaborative model for connecting product design and production line design: an aeronautical case study. *In POEM'21*. vol. 432. Springer.

Bryl V., Giorgini P., Mylopoulos J. (2006, 10). Designing cooperative is: Exploring and evaluating alternatives. In *OTM 2006*, vol. 4275, p. 533–550. Springer.

Chan A., Polacsek T., Roussel S. (2022, 02). Une approche basée sur l'ordonnancement pour évaluer la performance de la production d'avions à haut niveau. In *Roadef 2022*.

Dalpiaz F., Franch X., Horkoff J. (2016, 05*). iStar 2.0 language guide*. coRR.

Elahi G., Yu E. (2011, 01). Requirements trade-offs analysis in the absence of quantitative measures: A heuristic method. *In ACM - SAC 2011*, p. 651-658. ACM.

ElRakaiby Y., Borgida A., Ferrari A., Mylopoulos J. (2020, 11). A refinement calculus for requirements engineering based on argumentation theory. *In ER 2020*, vol. 12400, p. 3-18. Springer.

Jureta I. J., Borgida A., Ernst N. A., Mylopoulos J. (2010, 09). Techne: Towards a new generation of requirements modeling languages with goals, preferences, and inconsistency handling. *In RE 2010*, p. 115-124. IEEE Computer Society.

Kavakli E., Loucopoulos P. (2005, 01). Goal modeling in requirements engineering: Analysis and critique of current methods. *In Information Modeling Methods and Methodologies*, p.102–124. Idea Group.

Murukannaiah P. K., Kalia A. K., Telang P. R., Singh M. P. (2015, 08). Resolving goal conflicts via argumentation-based analysis of competing hypotheses. In *RE 2015*, p. 156–165. IEEE Computer Society.

Mylopoulos J., Chung L., Nixon B. (1992, 06). Representing and using nonfunctional requirements: A process-oriented approach. *IEEE -TSE*, vol. 18, no 6, p. 483-497.

Pant V. (2021). *Strategic coopetition - a conceptual modeling framework for analysis and design*. Thèse de doctorat non publiée, University of Toronto.

Pant V., Yu E. (2019, 07). A modeling approach for getting to win-win in industrial collaboration under strategic coopetition. *CSIMQ*, vol. 19, p. 19-41.

Polacsek T., Roussel S., Bouissiere F., Cuiller C., Dereux P., Kersuzan S. (2017, 10). Towards thinking manufacturing and design together: An aeronautical case study. *In ER 2017*, vol.10650, p. 340–353. Springer.

van Lamsweerde A. (2001, 08). Goal-oriented requirements engineering: A guided tour. *In RE2001*, p. 249-262. IEEE Computer Society.

Intent-based Contextual Orchestration inside Digital Business Ecosystems and Platforms

Mustapha Kamal Benramdane

CEDRIC, CNAM, 292 rue Saint Martin, 75003 Paris, France
mustapha-kamal.benramdane@lecnam.net

MOTS-CLES : Écosystèmes et Plateformes Digitaux, Orchestration des DBE, Configuration Contextuelle, Edge Cloud et Edge Computing, Machine Learning.

KEYWORDS : Digital Business Ecosystems, Digitally Enabled Collaborations, DBE entities' Orchestration, Contextual Orchestration, DBE configuration, Machine Learning.

ENCADREMENT : Samia Bouzefrane, Elena Kornyshova, Hubert Maupas.

1. Introduction

The development of technologies contributes to the growth of digital platforms which allow users to exchange content and data (De Reuver et al., 2018). However, this strong presence of users and organizations is likely to evolve into a business ecosystem. However, the complexity of exchanges and the abundance of data make it difficult to find partners or services that can meet the needs of a certain organization. Digital Business Ecosystems (DBE) is an economic community that produces goods and services for customers who are themselves members of the ecosystem (Moore, 1993).

One way to connect ties within a digital platform is through matchmaking. This process attempts to assess the interest profiles of market actors, with the aim of matching the supply chain agents with the least conflicting interests, thus expected to have better benefits during the arbitration phases and subsequent execution (Medjahed et al., 2003). The use of an intermediary actor that collects data about market actors, helps potential customers and suppliers find trading partners and improves the efficiency of the matchmaking process (Ouksel et al., 2004) as in digitized financial and commodity trading systems.

2. Problem

The notion of intention is essential for member organizations of a digital ecosystem because it allows them to define their needs and expectations and to meet

the requirements of internal and external users. The intent-oriented perspective is being considered in many areas (Deneckère & Kornyshova, 2011), including different organizational aspects, as it allows the considered artifacts to be connected to business needs and others. However, the methods explored for organization of DBEs do not include the notion of intention (Senyo et al., 2019). The characteristics of the context (characteristics of the sector, market segment, financial parameters, geographical proximity, market history, digital compatibility, etc.) as well as the rules of the trade must be taken into account in the orientation of the choices of the members of a digital ecosystem. This orientation will allow them to find the best partners with whom to associate. Our problem is *how to join business rules, context characteristics, and user intentions in guiding users through DBEs.* We intend to resolve this problem using Machine Learning algorithms that would take into account the users' needs and business characteristics in the matchmaking process.

3. State of the Art

To establish a state of the art on matchmaking in DBEs, we questioned the Scopus database with the following query.

```
TITLE-ABS-KEY(matchmaking) AND TITLE-ABS-KEY(algorithm)AND
(TITLE-ABS-KEY(ecosystem) OR TITLE-ABS-KEY(platform) OR
TITLE-ABS-KEY(BtoB)  OR TITLE-ABS-KEY(DBE) OR TITLE-ABS-
KEY(digital) OR TITLE-ABS-KEY(business))
```

We have chosen to limit our search since 2010. We obtained 62 publications which we consulted to keep only the closest ones to our research topic. We could not find in the literature a method to orchestrate or structure the contribution of the different entities within the DBEs. None of these contributions considered combining the different matchmaking techniques with machine learning in order to consolidate its performance.

We were also interested in the organizational aspect of DBEs. A DBE can be formed from a simple supply-chain, in which all actors participate in the development of a product or service. Among the main actors that come into play in organization there are mostly Competitors, Customers, Suppliers, Distributors, Partners, Manufacturers, and Influencers. New entrants can either adapt to the ecosystem or make changes to its structure. Each member of the ecosystem occupies a place in its own landscape of opportunity, and each landscape involves collaborators, competitors and complements (Lewin, 1999). The ecosystem leadership function is valued by the community because it allows members to move toward common visions, and find mutually reinforcing roles (Moore, 1993). The evolution of a DBE goes through three essential phases: Creation, monitoring and evaluation (D'Andrea et al., 2013), although they are not punctual in time and can be spread out over the life of the DBE.

However knowledge about the challenges associated with the development of DBE platforms and associated solutions is limited since these studies do not use theories (Senyo et al., 2019). It is possible that this is due to the novelty of the DBE

concept (Senyo et al., 2019). It is important to understand how to strategically control platforms for the benefit of all participants (Koch & Windsperger, 2017).

4. Achievements and Future Works

In the perspective of developing a generic approach to orchestrate entities within a DBE, we designed a recommendation system to guide organizations in their choice of partners. This system is based on two types of matchmaking algorithms, the first called "Static" and the second "Smart Recommendation" as depicted in Figure 1.

Static matchmaking uses static business rules, correspondence tables and graphs that represent business context to generate a list of recommended partners at the request of the user and according to the data he has inserted as input. These algorithms provide similar results for organizations with the same business characteristics (Benramdane et al., 2021).

The Smart Recommendation fills the gap in static matchmaking. It proceeds in two phases: a) for a pair (A, B), A as the organization in input, and B each organization from the list, the model returns a match probability which is considered as a score to each organization B appearing of the previous list; b) the list is reordered according to the new score of phase a) in descending order. We trained the Logistic Regression and Decision Tree algorithms on a database of 10 000 organizations profiles and obtained more than 90% accuracy rate. These models aim at classifying organizations by assigning them a score. The score is calculated based on the information related to the organizations and the needs of the user (Benramdane et al., 2021).

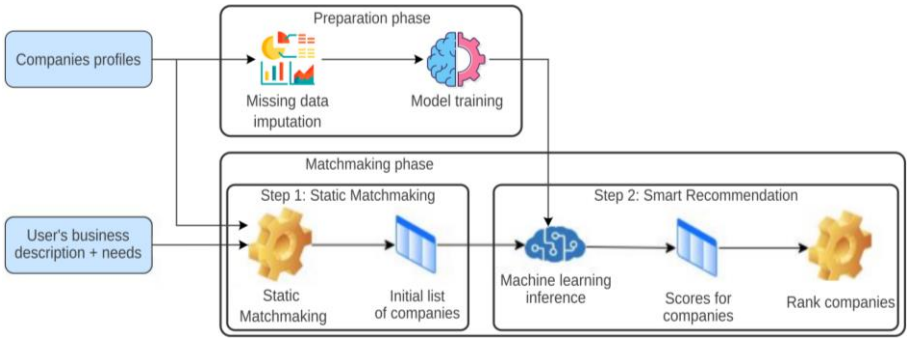

Figure 1: Overview of the Recommendation System for Matchmaking in DBEs.

Our **future work** consists in identifying all the actors involved in the elaboration, structuring and evolution of a DBE. The knowledge of their roles will provide useful information to simulate the evolution of a DBE, and subsequently develop a method to guide these actors in their choice of partners within the ecosystem. In an environment where several digital platforms come into play, we select to explore DBEs using Edge computing architecture. One of the sources of

data about intentions and context should be provided from Edge Cloud applications in addition to data gathered from other sources.

Considering the volume of data going through DBE platforms, we aim to explore several hybrid machine learning algorithms in order to include the teleological dimension of the different actors within the DBE in the matchmaking process, with the objective of developing an intent-based and contextual orchestration method inside DBE and platforms.

References

Benramdane, M. K., Maupas, H., Kornyshova, E., & Banerjee, S. (2021). Business Recommender System through Matchmaking with Supervised Machine Learning in Distributed Digital Platforms: Energy Complexity Analysis. *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*, 372–376.

D'Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2013). Digital eco-systems: The next generation of business services. *Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems*, 40–44.

De Reuver, M., Sørensen, C., & Basole, R. C. (2018). The digital platform: A research agenda. *Journal of Information Technology*, *33*(2), 124–135.

Deneckère, R., & Kornyshova, E. (2011). Processus téléologique et variabilité: Utilisation de la sensibilité au contexte. *Revue Des Sciences et Technologies de l'Information-Série ISI: Ingénierie Des Systèmes d'Information*, *16*(1), 61–88.

Koch, T., & Windsperger, J. (2017). Seeing through the network: Competitive advantage in the digital economy. *Journal of Organization Design*, *6*(1), 1–30.

Lewin, R. (1999). *Complexity: Life at the edge of chaos*. University of Chicago Press.

Medjahed, B., Benatallah, B., Bouguettaya, A., Ngu, A. H., & Elmagarmid, A. K. (2003). Business-to-business interactions: Issues and enabling technologies. *The VLDB Journal*, *12*(1), 59–85.

Moore, J. F. (1993). Predators and prey: A new ecology of competition. *Harvard Business Review*, *71*(3), 75–86.

Ouksel, A. M., Babad, Y. M., & Tesch, T. (2004). Matchmaking software agents in b2b markets. *37th Annual Hawaii International Conference on System Sciences*.

Senyo, P. K., Liu, K., & Effah, J. (2019). Digital business ecosystem: Literature review and a framework for future research. *International Journal of Information Management*, *47*, 52–64.

Cohérence des systèmes d'information :
Alignement opérationnel des applications sur le métier

Ali Benjilany

Nantes Université, CNRS, LS2N, F-44000 Nantes,
52, rue de la Houssinière, BP 92208 - 44322 Nantes Cedex 3, France
ali.benjilany@ls2n.fr

MOTS-CLES : Système d'information, Urbanisation, Architecture d'entreprise, Alignement Business-IT, Méthodes d'alignement

KEYWORDS : Information System, Urbanization, Enterprise Architecture, Business-IT Alignment, Alignment Methods.

ENCADREMENT : Pascal André, Hugo Brunelière, Dalila Tamzalit.

1. Contexte général

La cohérence des systèmes d'information est un problème ancien et épineux, qui exprime le fait que les actions stratégiques et opérationnelles sont en phase entre elles et avec l'organisation. Cette cohérence est notamment exprimée au travers de la notion d'*alignement* entre différents points de vue sur un système donné. Nous nous intéressons ici à l'alignement entre le système d'information et le système d'information informatisé, appelé *alignement Business-IT (BITA)*. L'article fondateur (Henderson *et al,* 1999) pose clairement les bases d'un modèle général d'alignement qui a ensuite pris une forme plus précise dans les cadres d'architecture d'entreprise[4] (Lankhorst, 2013, Mandel, 2006). Ces cadres d'architecture définissent des **couches d'abstraction**, dont le nombre varie d'un cadre à un autre, allant des niveaux très opérationnels relatifs aux infrastructures jusqu'aux niveaux stratégiques du système décisionnel. Maîtriser la complexité de l'exercice d'alignement revient souvent à rendre des couches adjacentes cohérentes entre elles plutôt que de tenter d'aligner directement les plus hauts niveaux stratégiques avec les plus bas niveaux opérationnels. La distance sémantique entre les concepts de ces couches rend l'exercice ardu. Nous nous focalisons dans nos travaux sur l'*alignement opérationnel* (Henderson *et al,* 1999) et plus particulièrement entre le métier et les systèmes logiciels qui le supportent au niveau fonctionnel. L'alignement opérationnel est

[4] En France, le terme « urbanisation » est aussi souvent utilisé, en lien avec la métaphore d'organisation des villes.

fondamental pour accompagner l'évolution des systèmes d'information car, lorsque les applications logicielles ne sont pas ou plus en phase avec l'organisation, des dysfonctionnements apparaissent engendrant dette technique et coûts imprévus. Malgré son aspect incontournable pour la cohérence globale d'une organisation, il a pourtant fait l'objet de relativement peu d'attention comparé à l'alignement stratégique. Nous le verrons dans l'état de l'art ci-après avant d'aborder la problématique de la thèse et les pistes suivies.

2. État de l'art

Les contributions sur l'alignement opérationnel couvrent des sujets divers et difficilement comparables car abordés selon différentes préoccupations (e.g., cartographie, architecture, évolution des SI). Dans (Aversano *et al.* 2012), les auteurs montrent l'importance de l'alignement dans la performance des systèmes d'information mais aussi la disparité des approches (90 publications sélectionnées) tant dans la terminologie que les modèles, leur alignement, évaluation et évolution. Ils montrent que le volet opérationnel (appelé *fonctionnel* ici) a été peu exploré, peu formalisé et finalement peu automatisé. Ils proposent dans (Aversano *et al.* 2016) une méthode en trois étapes (modélisation, évaluation, évolution). L'évaluation s'appuie sur un ensemble de métriques (*coverage/adequacy*) pour calculer un degré d'alignement. L'apport majeur se situe dans l'assistance à l'utilisateur : utilisation, si besoin, de la rétro-ingénierie pour générer des modèles UML à partir du code source, analyse sémantique pour proposer des liens de traçabilité permettant le calcul des métriques, historisation des versions pour observer l'évolution. Dans (Habba e*t al.* 2019) l'alignement est considéré entre (i) les besoins métiers, (ii) les processus métiers et (iii) le système logiciel. Les questions posées sont relatives à la définition de l'alignement, sa mesure et son application pratique. Les auteurs ont déterminé 63 références sur le sujet et mis en exergue la variété des approches. L'alignement des besoins métiers et processus métier (23 références mentionnées) se fait selon des règles de correspondance assez naturelles. L'alignement métier/applicatif est moins naturel et repose principalement sur des approches à base de règles pour lier (ou générer) les modèles applicatifs à partir des processus métiers. Les auteurs mettent globalement en évidence le manque de méthodologie, d'évaluation de l'alignement et d'outillage, ainsi que la grande variété des langages existants pour la modélisation (e.g., UML, BPMN, SoaML, i*, MAP) et par conséquent le peu d'homogénéité pour la formalisation de l'alignement. Si le choix de BPMN semble faire consensus pour les processus métier, il n'émerge en revanche pas de standard de référence pour la couche applicative. UML est le plus utilisé mais de manière hétérogène. La thèse (Pepin 2016) présente, à notre connaissance, une des rares approches traitant des différents éléments rentrant en compte dans une problématique d'alignement : des métamodèles décrivant les couches d'abstraction, la génération de modèles par rétro-ingénierie à partir du code source et l'évaluation des alignements réalisés. Une approche pragmatique est ainsi proposée pour un alignement opérationnel de la couche relative aux processus métiers (BPM) et de la couche applicative abstraite depuis le code. En pratique, l'alignement est réalisé via des techniques de tissage de facettes non intrusives sur les modèles concernés. Enfin, la validation de l'approche est assurée grâce à des cas réels. En revanche, la construction des modèles, lorsque non produits par les architectes, n'est pas automatisée.

3. Problématique

L'état de l'art a montré que les sujets de recherche à traiter pour mettre en œuvre un alignement opérationnel sont nombreux : choix des langages pour exprimer les modèles des différentes couches d'abstraction et des relations d'alignement, construction (semi) automatisée de ces modèles, évaluation des relations d'alignement (métriques, visualisation), suivi et maîtrise des évolutions du SI au travers de l'alignement, sont les principaux que nous avons identifiés à l'heure actuelle.

De nombreux défis liés à l'alignement du SI sur le métier se situent entre les deux couches métier et applicative (Pereira et Souza, 2005) : couplages forts, coûts de maintenance, taux d'échecs élevé des projets, lenteur d'adaptation aux changements (réglementaires ou stratégiques), etc. (Hinkelmann et al 2016, Yeow et al 2018). Dans le cadre de nos travaux, nous considérons uniquement l'alignement des couches métiers (structuration des processus opérationnels de l'organisation) et applicatives (abstraction des programmes sous forme d'architectures logicielles). Nous posons alors les questions de recherche suivantes : 1) Sur quel(s) modèle(s) de référence s'appuyer pour réaliser l'alignement ? 2) Existe-t-il une taxonomie des relations ? 3) Comment évaluer et/ou valider ces relations entre les entités métiers et applicatives en question ? 4) Comment assurer la maintenance d'un alignement acceptable vis-à-vis des changements continus dans le temps que subissent les organisations ? Comme premiers éléments de réponse, nous posons les principes suivants : se **baser sur des standards** pour améliorer l'interopérabilité, identifier et considérer uniquement les **concepts concernés par l'alignement**, favoriser une vision **fractale** de l'alignement (données/traitements, services/patrimoine monolithique, etc.) pour viser à la fois la pertinence et la cohérence des alignements, ainsi que fournir une assistance en termes de méthodologie et d'outils automatisés aux architectes. Dans notre perspective, un bon alignement est celui qui tend d'une part à éviter des couplages et dépendances inappropriés et souvent non voulues, et d'autre part à assurer l'existence des liens nécessaires au travers des différentes couches.

4. Actions

Ces premiers mois de thèse ont été consacrés principalement à : 1) l'étude du domaine (architecture d'entreprise, urbanisation, alignement, standards du domaine tels que SAM, Cigref, Togaf, Archimate) ; 2) l'étude des techniques d'ingénierie et rétro-ingénierie logicielle pour le développement futur de prototypes favorisant l'alignement, sa mesure et son évolution et 3) des expérimentations menées sur SmartEA (https://www.obeosmartea.com/fr/), un outil de cartographie d'entreprise. Nous allons compléter ce travail par une étude détaillée qui va approfondir et actualiser les *surveys* de l'état de l'art. Concernant notre proposition, la première étape est d'établir un (méta)modèle initial permettant l'interconnexion des deux couches du SI par des liens d'alignement pour pouvoir automatiquement analyser et détecter des problèmes d'alignement selon les différents points de vue considérés (alignement fractal). Ce (méta)modèle servira alors à construire une analyse quantitative (dans un premier temps) puis qualitative (par la suite) de l'alignement à base d'indicateurs

(métriques, agrégation de données, etc.) permettant de détecter des situations favorables ou défavorables sous la forme de patterns/anti-patterns d'alignement (Gouigoux et Tamzalit, 2021). La seconde étape concerne la construction effective des liens d'alignement en intégrant différentes techniques mentionnées dans l'état de l'art. Dans une troisième étape, il est prévu d'analyser l'évolution de l'alignement sur différentes versions d'un système d'information informatisé et de proposer des pistes d'amélioration aux architectes. En termes d'implémentation, nous visons l'extension du standard de facto Archimate- un langage de modélisation d'Architecture d'Entreprise-. Ce choix d'étendre Archimate en particulier est justifié par le fait qu'il est correctement outillé et extensible. Il nous permettra d'y définir des relations d'alignement entre concepts qui n'existent pas tel quel. Nous visons également la mise en œuvre d'outils basés sur l'ingénierie des modèles, notamment les solutions open source qui intègrent déjà le méta-modèle Archimate dans leur environnement de modélisation/développement.

5. Bibliographie sommaire

Aversano, L., Grasso, C., & Tortorella, M. (2012). A literature review of Business/IT Alignment Strategies. Procedia Technology, 5, 462-474.

Aversano, L., Grasso, C., & Tortorella, M. (2016). Managing the alignment between business processes and software systems. Information and Software Technology, 72, 171-188.

Gouigoux, J. P., & Tamzalit, D. (2021, September). Business-IT alignment anti-patterns: a thought from an empirical point of view. In 29THInternational Conference On Information Systems Development (ISD2021 Valencia, Spain).

Habba M. et al. (2019). Alignment between Business Requirement, Business Process, and Software System : A Systematic Literature Review. Journal of Engineering, 2019 :6918105, October 2019. Publisher : Hindawi.

Henderson, J. C., & Venkatraman, H. (1999). Strategic alignment: Leveraging information technology for transforming organizations. IBM Systems Journal, 38(2/3), 472.

Hinkelmann, K., Gerber, A., Karagiannis, D., Thoenssen, B., Van der Merwe, A., & Woitsch, R. (2016). A new paradigm for the continuous alignment of business and IT: Combining EA modelling and enterprise ontology. Comp. in Industry, 79, 77-86.

Lankhorst, M.(2013) Enterprise Architecture at Work - Modelling, Communication and Analysis (3. ed.). The Enterprise Engineering Series. Springer, 2013.

Mandel, R (2006). De la stratégie business aux systèmes d'information. Collection Management et Informatique, 292 pages, 2006 EAN13 9782746212978.

Pépin, J. (2016). Architecture d'entreprise: alignement des cartographies métiers et applicatives du système d'information (Doctoral dissertation, Nantes).

Pereira, C. M., & Sousa, P. (2005, March). Enterprise architecture: business and IT alignment. In Proceedings of the 2005 ACM symposium on Applied computing (pp. 1344-1345).

Yeow, A., Soh, C., & Hansen, R. (2018). Aligning with new digital strategy: A dynamic capabilities approach. The Journal of Strategic Information Systems, 27(1), 43-58.

Approche pour la recherche d'information multifacette en biomédecine

Maël Lesavourey

Institut de Recherche en Informatique de Toulouse (IRIT)
118 route de Narbonne
31062 Toulouse
mael.lesavourey@irit.fr

MOTS-CLES : Recherche d'information, Reconnaissance d'Entités Nommées, Indexation, Biomédecine, Toxicologie.

KEYWORDS : Information Retrieval, Named Entity Recognition, Indexing, Biomedicine, Toxicology.

ENCADREMENT : Gilles Hubert, Fabien Jourdan, Yoann Pitarch.

1. Contexte

Les menaces pour la santé humaine se font de plus en plus nombreuses. Les populations sont en effet touchées par différentes épidémies (Ebola, Zika, Covid-19…) mais aussi exposées à certaines substances chimiques dangereuses (Bisphénol-A et autres perturbateurs endocriniens). Les chercheurs en biomédecine ou toxicologie sont soumis à un double effort. D'une part, ils doivent faire preuve de réactivité lorsqu'il s'agit d'évaluer les risques d'un nouveau composé mis sur le marché ou qu'une population fait face à une épidémie. D'autre part, ils mettent en place des travaux anticipatifs pour faire avancer la connaissance sur le long terme, par exemple lors de l'étude de l'activité virale d'agents pathogènes. Ces efforts interviennent à des temporalités bien différentes et peuvent sembler contradictoires mais ils reposent sur un socle commun : l'analyse de l'existant en matière de connaissances. L'essor de la science ouverte combiné à la croissance du nombre de publications par année facilite l'accès au savoir produit par la communauté scientifique. Cependant le flot de connaissances étant toujours plus grand, il devient difficile de faire une navigation et une assimilation aisée de ces connaissances.

Une solution envisageable est de fournir des outils visant à automatiser et accélérer la recherche bibliographique nécessaire aux chercheurs en biomédecine et toxicologie. Pour se faire il est nécessaire de mettre en place des systèmes d'extraction et de recherche d'information répondant aux besoins spécifiques des chercheurs dans ces domaines. Deux possibilités découlent de ces observations. La première est l'application au domaine biomédical de méthodes connues et

performantes. La seconde est la construction de nouveaux critères d'intérêts propres à ces domaines et non-exploités dans les moteurs de recherche d'articles scientifiques. Ils permettraient notamment de fournir une information plus ciblée.

2. État de l'art

La reconnaissance d'entité nommée (NER) joue un rôle important dans les performances des systèmes d'extraction ou de recherche d'information. Dans le domaine de la biomédecine, de tels systèmes permettent d'identifier avec précision des concepts comme une maladie, une protéine, un biomarqueur… Des outils pour mener à bien ces tâches existent et peuvent être classés dans deux catégories. La première rassemble les méthodes basées sur des règles précises définies par des experts du domaine pour identifier différentes entités nommées dans un article. C'est le cas des systèmes basés sur le « term matching » (Aronson, 2001) ou le « pattern matching » (Eftimov *et al.,* 2017). Le problème de ces méthodes est qu'elles demandent l'intervention d'experts et qu'elles sont limitées au domaine pour lequel les règles ont été établies.

La seconde catégorie est celle des systèmes basés sur l'apprentissage automatique. Dans ce cas c'est le modèle qui va apprendre des motifs pour distinguer les entités nommées en utilisant un large corpus d'articles annotés. Ces méthodes ont souvent un pouvoir de généralisation plus puissant que celles décrites précédemment mais leurs performances dépendent énormément de la quantité de données annotées disponibles. Ce travail de labellisation est très coûteux et se restreint souvent à un seul type d'entité (gène, protéine…) par article (Wang *et al.,* 2019). Certains travaux tentent d'enrichir les jeux de données en utilisant des méthodes d'apprentissage semi-supervisé (Gao *et al.,* 2021) qui consistent à générer, à partir d'un corpus annoté, des pseudos labels sur des articles non annotés. Ceci permet d'obtenir des systèmes plus performants en limitant la quantité de données labellisées manuellement et donc en réduisant le coût humain.

Une deuxième tâche de traitement automatique du langage naturel (TALN) permettant d'améliorer la recherche d'articles biomédicaux est l'indexation de termes. La Librairie Nationale de Médecine des États-Unis (U.S. National Library of Medicine ou NLM) maintient la base de données bibliographique MEDLINE comprenant plus de 28 millions de références d'articles de journaux. Chaque référence est indexée à la main par des experts à l'aide d'une ontologie nommée Medical Subject Headings (MeSH). Cet effort permet de faciliter la recherche d'information en catégorisant les articles avec un vocabulaire reconnu et évolutif mais son coût est gigantesque. Plusieurs méthodes d'apprentissage automatique ont été développées pour permettre d'automatiser ce traitement qui correspond, d'un point de vue strictement informatique, à une classification multi-labels à grande échelle. La première, Medical Text Indexer (MTI), est le système proposé par la NLM en 2002 pour recommander des termes MeSH à indexer pour chaque article. Bien que des études montrent que les performances de MTI sont toujours pertinentes (Mork *et al.,* 2017), d'autres approches ont été développées au fil des années. On trouve notamment des méthodes basées sur le Learning To Rank (Peng *et al.,* 2016 ;

Dai *et al.*, 2020) s'appuyant aussi bien sur des représentations bag-of-words que sur des représentations basées sur l'apprentissage profond (Word2Vec, Doc2Vec). Plus récemment, plusieurs travaux (Jin *et al.*, 2018 ; You *et al.*, 2021) ont proposé une résolution de ce problème de classification en utilisant des mécanismes d'attention.

3. Problématique

Afin de mettre au point des systèmes performants de recherche et synthèse d'information, il est nécessaire de comprendre les documents du corpus traité. Le TALN est une branche de l'Intelligence Artificielle- qui a pour but de donner du sens aux textes notamment via les méthodes de NER ou d'indexation de texte. La question de recherche est, par conséquent, de définir une approche d'indexation qui tienne compte des différentes facettes liées aux besoins d'information spécifiques des chercheurs en toxicologie et biomédecine.

Une fois le problème posé, un des enjeux est de trouver des jeux de données annotées pour pouvoir évaluer les performances de différentes approches et se comparer aux approches de l'état de l'art (baselines). Le workshop BioCreative7 (https://biocreative.bioinformatics.udel.edu/) qui s'est tenu en Novembre 2021 a permis de promouvoir le développement d'approches destinées à répondre à plusieurs problématiques de NLP appliqué aux articles biomédicaux. Contrairement à la tâche de NER, il s'est avéré que l'indexation était très peu traitée et que les performances des systèmes des participants étaient faibles (seulement une équipe a dépassé la baseline fournie par les organisateurs). C'est à cette dernière problématique que s'intéresse plus particulièrement le challenge BioASQ (http://bioasq.org/) hébergé par CLEF (Conference and Labs of the Evaluation Forum) qui propose chaque année depuis 10 ans une tâche d'indexation de termes MeSH à grande échelle.

4. Actions futures

Pour répondre à la problématique de l'indexation de termes MeSH à grande échelle deux approches sont envisagées. La première consiste à utiliser un plongement (« embedding ») de documents (Doc2Vec) pour représenter les articles dans un espace de dimension fixée, l'hypothèse étant que des documents indexés avec les mêmes termes seront proches sémantiquement. Il sera possible dans cet espace de calculer le vecteur d'un terme MeSH M1 à partir des vecteurs documents indexés avec M1. Ainsi pour un document non indexé, son vecteur pourra être projeté (en utilisant un produit scalaire à définir) sur chacun des vecteurs MeSH pour obtenir un score évaluant la proximité du document avec un terme MeSH. Enfin, il faudra appliquer un seuillage pour garder les scores les plus hauts et sélectionner uniquement les MeSH pertinents pour chaque article. Une seconde approche, fondée sur la même hypothèse, consisterait à exploiter des systèmes basés sur des modèles de « transformers » pré-entraînés comme BioBERT et PubMedBERT, et utilisant les mécanismes d'attention.

L'indexation des articles suivant les termes d'une ressource ontologique n'est pas suffisante pour comprendre le contexte dans lequel ils sont ancrés. Par exemple, un composé chimique peut apparaître dans un document en tant que médicament ou élément toxique. Une substance peut avoir un effet différent suivant le sexe des sujets, leur espèce ou encore les tissus dans lesquels elle est présente.

La représentation sémantique multi-contextuelle des termes MeSH pourrait permettre de capter des informations supplémentaires inhérentes aux documents. Un même terme serait alors représenté de plusieurs façons suivant les contextes dans lesquels il apparaît.

Prendre en compte ces facettes non encore intégrées dans les moteurs de recherche de publications scientifiques faciliterait significativement le travail bibliographique des chercheurs en leur fournissant des informations plus ciblées sur les documents recherchés.

Bibliographie

Aronson A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, p. 17–21.

Dai S, You R, Lu Z, Huang X, Mamitsuka H, Zhu S. (2020). FullMeSH: improving large-scale MeSH indexing with full text. *Bioinformatics*, vol. 6, n° 5, p. 1533-1541, PMID: 31596475, PMCID: PMC7523651.

Eftimov T, Koroušić Seljak B, Korošec P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One*, vol.12, n° 6, PMID: 28644863, PMCID: PMC5482438.

Gao S, Kotevska O, Sorokine A, Christian JB. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PLoS ONE*, vol. 16, n° 2, e0246310.

Jin Q, Dhingra B, Cohen W and Lu X. (2018). AttentionMeSH: Simple, Effective and Interpretable Automatic MeSH Indexer. *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, p. 47–56, Brussels, Belgium. Association for Computational Linguistics.

Mork J, Aronson A and Demner-Fushman D. (2017). 12 years on – Is the NLM medical text indexer still useful and relevant? *J Biomed Semant*, vol. 8, n° 8.

Peng, Shengwen, Ronghui You, Hongning Wang, ChengXiang Zhai, Hiroshi Mamitsuka and Shanfeng Zhu. (2016). DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, vol. 32, p. 70-79.

Wang X, Zhang Y, Ren X, Zhang Y, Zitnik M, Shang J, Langlotz C, Han J. (2019). Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, vol. 35, n° 10, p. 1745–1752.

You R, Liu Y, Mamitsuka H, Zhu S. (2021). BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, vol. 37, n° 5, p. 684-692.

Intégration de données et inférence au service de l'étude de la chimiodiversité du vivant

Solweig Hennechart

Laboratoire de Recherche en Sciences Végétales (LRSV) UMR5546
Institut de Recherche en Informatique de Toulouse (IRIT) UMR5505
Université Toulouse 3 Paul Sabatier
118 route de Narbonne, 31062 Toulouse
solweig.hennechart@univ-tlse3.fr

MOTS-CLES : Chimiodiversité, Métabolomique, Prédiction de données,Iintégration de données, Fusion de données.

KEYWORDS: Chimiodiversity, Metabolomics, Data prediction, Data integration, Data fusion.

ENCADREMENT : Guillaume Cabanac, Guillaume Marti.

1. Contexte

La chimiodiversité du vivant correspond à l'ensemble des molécules d'origine naturelle et s'organise autour des taxons, des milieux trophiques et des écosystèmes en une dynamique et une cartographie encore peu documentée. À l'instar de sa grande sœur la biodiversité, il existe un écart entre la diversité chimique connue (~ 1 million) et celle estimée (plusieurs millions de composés), alors même que cette estimation est très complexe à réaliser (Kind *et al.,* 2009). Une connaissance plus fine et exhaustive de ce domaine a pourtant le potentiel de révéler des composés dans des groupes d'organismes à fort potentiel de découverte, notamment pour les industries pharmaceutique et agroalimentaire. Cette cartographie de la diversité chimique des espèces est d'autant plus prégnante dans un contexte d'érosion de la biodiversité mondiale.

L'étude expérimentale de cette chimiodiversité se fait à l'aide d'approche métabolomique. La métabolomique représente le dernier élément de la cascade des « omiques » initié par la génomique. Car si le génome correspond à ce que l'organisme à la capacité de faire, l'analyse du métabolome, correspondant à l'ensemble des composés chimiques d'un organisme. Cette analyse permet d'appréhender un système biologique dans son environnement en captant ce qui se produit réellement à un instant T. Les applications de la métabolomique sont larges et vont de la compréhension de mécanismes (biologie fondamentale) au diagnostic

clinique de patients (Ellero-Siratos *et al.*, 2019) en passant par la détection de substances dopantes, ou encore la détection de contrefaçon de grands millésimes de vins par exemple.

Le constat interne au laboratoire de biologie à l'origine du développement de ce projet, en parallèle avec la publication d'un outil de traitement de données métabolomiques (Fraisier-Vannier *et al.*, 2020), est que l'une des pistes d'amélioration des analyses métabolomiques repose sur une présélection des composés que l'on peut s'attendre à trouver dans un échantillon. Il n'existe pas à l'heure actuelle de base de données exhaustive permettant de faire cela. C'est pour cela que les laboratoires IRIT et LRSV ont initié la base de données biochimiques Pharmakon en 2017. Maintenant que l'intérêt pour la métabolomique est croissant et que la plus-value d'une telle base de données est confirmée, cette thèse pluridisciplinaire a pu voir le jour. Elle se propose de perfectionner la cartographie de la chimiodiversité en fédérant les informations de différentes sources de données biochimiques pour mettre au point une base de données unique, la plus exhaustive possible, tout en s'efforçant de compléter les informations sur l'origine biologique des différents composés pour lesquels cela n'est pas renseigné.

2. État de l'art

Pour d'autres champs de connaissance en biologie, il existe une base de données de référence, comme UniProt pour les protéines, mais ce n'est pas le cas concernant les composés d'origine naturelle. Depuis les années 2000, il a été recensé plus de 120 sources de données distinctes (Sorokina et Steinbeck, 2020), publiques ou commerciales, spécialisés en classe chimique comme lipidMaps pour les lipides ou en branche taxonomique comme KNApSAcK pour les métabolites des plantes. Toutes ces sources cataloguent différentes informations, se comptant en plusieurs dizaines de champs distincts dont les valeurs sont similaires mais rarement identiques entre les sources, exprimés sans formalisme ni ontologie commune.

La version 2020 de Pharmakon résulte d'une agrégation de onze sources de données de référence en produits naturels, publiques ou sous licence. Elle référencie plus 600 000 composés « uniques » selon des critères chimiques. L'information sur l'origine biologique (règne, famille, genre, espèce) n'est cependant connue que pour seulement un tiers d'entre eux.

La diversité de formats et de méthodes de récupération, spécifique à chaque source, rend difficile l'extraction d'un référentiel de données commun. Pour pouvoir mettre en lumière les différentes informations récupérables et leurs formats, il est nécessaire de les déduire par rétro-ingénierie, car les schémas de données ne sont pas toujours mis à la disposition des utilisateurs, avant de transformer les données pour les charger dans un référentiel de données intégré. Ce processus est appelé ETL (Extract, Transform and Load), et il permet de lister les informations d'intérêt avant de réaliser un chargement uniforme de plusieurs sources hétérogènes en suivant un schéma conçu en fonction des besoins (Kathiravelu *et al.*, 2019).

3. Problématique

La métabolomique se différencie des études biologiques plus courantes par son approche sans a priori, ne partant pas initialement d'une hypothèse pour analyser un échantillon. L'analyse consiste à capturer l'empreinte chimique d'un système biologique complexe avec un spectromètre de masse, l'un des instruments les plus utilisés en métabolomique. Cette approche requiert une étape de traitement pour l'annotation des centaines de signaux acquis par l'instrument qui est, soit excessivement long, soit peu efficace. En effet, il s'appuie sur la comparaison des signaux résultant de la fragmentation des molécules présentes dans l'échantillon au cours de l'analyse avec des molécules connues dont les fragmentations sont prédites par des calculs *in silico*.

C'est pourquoi il est intéressant de développer une base de données exhaustive permettant d'extraire et/ou de prioriser l'annotation des signaux à partir d'une liste de molécules que l'on s'attend à retrouver dans un extrait. Cela permettrait de réduire les limites techniques et biologiques. Mais comment concevoir une base de données issue de la fusion de plusieurs sources hétérogènes ? Comment compléter les informations nécessaires et pourtant manquantes dans les sources aujourd'hui disponibles ?

4. Actions réalisées

Il n'existe pas, à notre connaissance, d'étude portant sur la prédiction de la source biologique d'un composé. En s'appuyant sur des travaux mettant en lumière des tendances dans la répartition de différentes classes chimiques en fonction de la branche phylogénétique (plante, animal, bactérie, champignon) (Chassagne *et al.*, 2019), nous avons voulu confirmer qu'il est possible de prédire le règne taxonomique d'un composé chimique.

Pour cela, je suis partie d'un jeu de données réduit et équilibré issue de la version 2020 de Pharmakon, soit 9 736 composés, ayant une seule origine biologique connue et une structure chimique standardisée. J'ai calculé des descripteurs chimiques, données numériques permettant de caractériser les molécules grâce à des propriétés chimiques (nombre de cycles aromatiques par exemple) et ai conservé ceux ayant significativement un lien avec la biosource, soit 91 caractéristiques. J'ai d'abord testé différents modèles dont un classifieur *KNN* et un *Random Forest* permettant d'obtenir des prédictions avec une accuracy (nombre de composés correctement prédits sur le nombre total de composés) de 80 %. C'est une preuve de concept validant qu'il est possible de prédire une biosource pour une molécule dont la formule et la structure est connue et avec des modèles simples. Il faudra ensuite aller plus loin avec la prochaine version de Pharmakon et des modèles de classification plus adaptés.

5. Actions futures

Après avoir effectué différentes analyses sur la base de données Pharmakon 2020 de nouveaux besoins ont été identifiés. Des pistes d'explorations sont régulièrement

proposées au regard de la diversité d'information, qu'il s'agisse d'analyses scientométriques, évolutives, et/ou géographiques. Mais il semblerait surtout que les critères pour regrouper les molécules identiques ne soient pas assez rigoureux.

La première étape consiste à reformater la base de données relationnelles Pharmakon, pour répondre aux besoins nouvellement identifiés, en référentiel de données intégré. Pharmakon 2022 sera composé d'une dizaine de champs qui restent à déterminer et nécessiteront tous un traitement. Même les SMILES et les InChI, champs alphanumériques communs à toutes les sources correspondant à la composition et à la structure des molécules, ne sont pas homogènes, car il existe plusieurs façons de les obtenir.

Après avoir récupéré les données, il faudra mettre en évidence les molécules identiques soit parce qu'il existe un lien entre les composés dans les informations récupérées, soit en s'appuyant sur des critères chimiques prédéfinis. Cette étape devrait permettre de collecter plus d'un million de composés uniques. Ceci pose bien évidemment le problème de fusion de données dans le cas où une même molécule est présente dans plusieurs sources, avec l'ajout d'un indicateur de fiabilité de l'information si elle est confirmée en étant présente dans différentes sources ou encore référencée dans une ressource bibliographique. Nous pouvons imaginer trouver des informations contradictoires à cette étape, qu'il faudra identifier et résoudre.

Une fois la base de données relationnelles Pharmakon 2022 construite, la prochaine étape visera à développer un ou plusieurs modèles d'inférences pour prédire la biosource des composés. Une validation expérimentale permettra de confirmer la robustesse des prédictions ou d'apporter des modifications.

6. Bibliographie

Chassagne, F., Cabanac, G., Hubert, G., David, B., & Marti, G. (2019). The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products®. *Phytochemistry Reviews*, *18*(3), 601-622.

Ellero-Simatos, S., Claus, S., & Guillou, H. (2019). La métabolomique : Applications médicales. *Médecine des Maladies Métaboliques*, *13*(3), 263-267.

Fraisier-Vannier, O., Chervin, J., Cabanac, G., Puech, V., Fournier, S., Durand, V., Amiel, A., André, O., Benamar, O. A., Dumas, B., Tsugawa, H., & Marti, G. (2020). MS-CleanR: A Feature-Filtering Workflow for Untargeted LC–MS Based Metabolomics. *Analytical Chemistry*, *92*(14), 9971-9981.

Kathiravelu, P., Sharma, A., Galhardas, H., Van Roy, P., & Veiga, L. (2019). On-demand big data integration: A hybrid ETL approach for reproducible scientific research. *Distributed and Parallel Databases*, *37*(2), 273-295.

Kind, T., Scholz, M., & Fiehn, O. (2009). How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry. *PLoS ONE*, *4*(5), e5440.

Sorokina, M., & Steinbeck, C. (2020). Review on natural products databases: Where to find data in 2020. *Journal of Cheminformatics*, *12*(1), 20.

Services augmentés pour le tourisme intelligent et l'analyse des pratiques

Maxime Masson

Laboratoire LIUPPA, Collège STEE, Université de Pau et des Pays de l'Adour
Avenue de l'Université, 64000 Pau, France
maxime.masson@univ-pau.fr

MOTS-CLES : Réseaux Sociaux, Tourisme, Traitement Automatique du Langage (TAL), Analyses Multidimensionnelles, Proxémique, Extraction d'Informations.

KEYWORDS : Social Media, Tourism, Natural Language Processing (NLP), Multidimensional Analysis, Proxemics, Information Extraction.

ENCADREMENT : Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, Philippe Roose, Annig Le Parc Lacayrelle.

1. Introduction

Les dernières décennies ont vu une croissance significative des sources de données disponibles couvrant de nombreux sujets ainsi que l'essor du contenu généré par les utilisateurs. Le tourisme, en particulier, est un domaine dans lequel l'obtention et l'étude d'ensembles de données thématiques massifs et précis sont essentiels pour mieux comprendre certains phénomènes et tendances.

Ces jeux de données peuvent être utilisées par les professionnels du tourisme à différentes échelles. Tout d'abord, pour aider au processus de prise de décision des acteurs du tourisme en leur permettant de faire les choix les plus éclairés possibles. Cela se fait en explorant les données pour arriver à mieux comprendre la pratique et les besoins des visiteurs. Ce type d'analyse est aussi particulièrement utile pour les entreprises, notamment celles spécialisées dans le marketing touristique (comme les *organisations de gestion de destination*), pour lesquelles il est crucial de bien cerner les désirs et les attentes des touristes (Belias *et al*., 2021). D'autre part, les systèmes de recommandation touristique sont devenus un enjeu important à la fois pour les touristes et les acteurs locaux. Ils permettent, en examinant la pratique touristique et les lieux visités par un grand nombre de personnes, de suggérer les activités ou itinéraires touristiques les plus adaptés possible au profil du visiteur.

Différentes ressources peuvent être utilisées pour accéder à des données touristiques. Historiquement, il s'agit principalement de bases de données, avec : (1) les bases de données commerciales, telles que celles des agences de voyages en

ligne (OTA, *Online Travel Agencies*) et (2) les bases de données publiques, par exemple celles produites par les agences gouvernementales ou celles fonctionnant sur un principe de *crowdsourcing*. Ces dernières font partie de ce que l'on appelle le « contenu généré par les utilisateurs ». Cette catégorie de données s'est considérablement développée, via les réseaux sociaux généralistes (*Twitter*, *Facebook*) ou les sites d'avis (*TripAdvisor*). L'Internet des objets est également un nouveau vecteur de données clef pour le tourisme intelligent (ex : surveillance et prédiction des flux de visiteurs). Enfin, les systèmes de traçabilité des touristes sont de plus en plus utilisés pour suivre avec précision les itinéraires empruntés.

Notre projet s'inscrit dans la continuité de ce phénomène. Nous souhaitons concevoir et développer une plateforme visant à collecter, traiter, analyser puis valoriser des données relatives à la pratique du tourisme, aux flux de visiteurs et à la fréquentation des points d'intérêt dans la région du Pays Basque, un territoire hautement touristique qui s'étend entre la France et l'Espagne. Il convient de déterminer si la construction de trajectoires touristiques multidimensionnelles (spatiale, temporelle, thématique) individuelles dans un premier temps puis agrégées corrélée à la mise en place de services d'analyse et l'introduction de la notion de proxémique à ces dernières peut contribuer à des analyses pertinentes, à fois pour les touristes (*recommandation*) et les professionnels (*aide à la décision*). Nous évoquerons les motivations derrière notre travail puis les actions réalisées et futures.

2. Motivations

Afin d'illustrer les grandes phases de mon travail de thèse, nous prenons appui sur le concept de *WaterWheel* (Bucher *et al*., 2021), une roue montrant le cycle de vie des données dans le système d'information que nous allons mettre en place.

La 1ère phase (Fig. 1, *Collecte*) consiste en la recherche de ressources liées au tourisme et à la collecte des données associées. A cette fin, il faut définir précisément les jeux de données nécessaires en fonction des besoins (*tourisme, zone et temporalité ciblés*) ; mettre en place une méthodologie de collecte de données sur les réseaux sociaux générique, multilingue et indépendante du domaine. Cette méthodologie doit également s'assurer de l'évaluation du jeu de données résultant. Au cours de la 2ème phase (Fig. 1, *Traitement*), nous souhaitons mettre en place une batterie de services dédiés à l'enrichissement, l'édition, et l'agrégation des données touristiques précédemment collectées. Le but est de les transformer et les croiser pour augmenter leur valeur ajoutée (*extraction de concepts, classification, détection de sentiments, etc.*). Les données seront ensuite enrichies afin d'être en mesure de définir la pratique touristique de la manière la plus complète possible. Une fois les données nécessaires en main (*collecte*) et dans la forme souhaitée (*traitement*), nous allons concevoir des services d'analyse (Fig. 1, *Analyse*) basés sur les dimensions : spatiale (*ensemble des lieux visités*), temporelle (*période et durée du séjour*) et thématique (*sémantique de la pratique touristique* : *type d'activités, condition du séjour, profil touristique, coût*). L'une des originalités de notre travail est que nous souhaitons définir des trajectoires touristiques multidimensionnelles à partir des données collectées et y intégrer la notion de *proxémique* (Hall, 1966). La proxémique est la science qui étudie l'organisation de

l'espace et l'effet des distances dans les relations interpersonnelles. Cette dernière est généralement appliquée au monde physique (*mouvement dans l'espace*) mais l'un des défis de mon travail est de l'adapter au monde virtuel au service du tourisme intelligent. Nous cherchons à concevoir un modèle proxémique permettant d'offrir une perspective nouvelle d'analyse des pratiques touristiques et qui soit applicable aux réseaux sociaux, un lieu d'expression fréquent pour les touristes. De plus, nous souhaitons que la dimension thématique prenne le pas sur le côté strictement géographique. L'espace pourrait s'apparenter à l'espace thématique du tourisme sur le réseau social (*types de tourisme, d'activité, typologies de visiteur*).

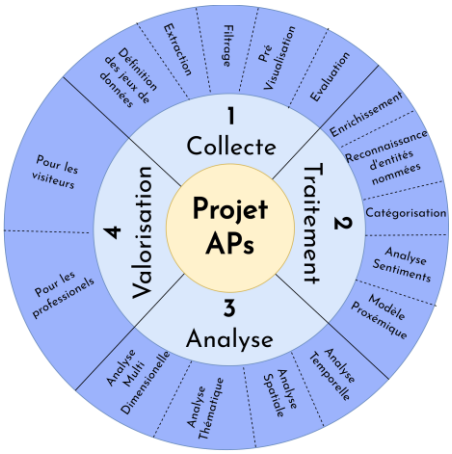

Figure 1 : Cycle de vie de l'information touristique dans le cadre de notre projet. Représentation inspirée de la WaterWheel (Bucher et al., 2021).

La 4ème phase (Fig. 1, *Valorisation*) s'intéresse à la valorisation des données précédemment collectées et traitées en visant deux cibles : les **professionnels** (i) et les **visiteurs** (ii). L'expérience des visiteurs pourrait être présentée dans des tableaux de bord multimodaux corrélant l'offre touristique et les pratiques afin d'orienter la prise de décision des aménageurs. Il serait ainsi possible de mettre en lumière les points chauds (*zones d'activité touristique intense*) passés, actuels et futurs, les couloirs de passage, ou les pratiques touristiques récurrentes. Nous souhaitons aussi construire un système de recommandation touristique calibré pour cette région si particulière qu'est le Pays Basque en prenant appui sur l'analyse des trajectoires touristiques multidimensionnelles précédemment citées afin d'être en mesure de suggérer les itinéraires les plus adaptés en fonction du profil du visiteur.

3. Actions réalisées

Rappelons que la 1ère phase de notre travail de recherche est la mise en place d'une méthodologie générique pour la collecte de données touristiques. Nous avons choisi les réseaux sociaux comme principale ressource pour plusieurs raisons : (i)

leur facilité d'accès, (ii) pas de mise en place de longues campagnes de collecte, (iii) pas de nécessité d'acheter de coûteux systèmes de suivi des touristes ou même des données commerciales à des sociétés spécialisées. De plus, ces données sont massives et très diverses, c'est-à-dire que de multiples aspects du tourisme sont couverts. L'extraction sélective à partir de cette masse de données est un défi complexe et a été l'objet du début de notre travail.

A notre connaissance, il n'existe pas d'outil ou de méthode entièrement générique et applicable à n'importe quel domaine pour extraire des données thématiques très ciblées à partir de n'importe quel réseau social, les méthodes actuelles étant fortement corrélées au domaine (Scholz et Jezni, 2020). Nous avons donc conçu une nouvelle méthode générique, indépendante du domaine et itérative destinée à faciliter ce processus de collecte et l'avons expérimentée sur le thème du tourisme avec le réseau social *Twitter*. Cette méthode s'architecture autour des 3 dimensions au cœur de nos futures analyses : spatial (*étendue spatiale des données*), temporelle (*période temporelle concernée*) et thématique (*sujet d'étude représenté sous la forme d'un vocabulaire plus ou moins complexe*). Afin de s'assurer de la qualité des données collectées, un processus d'évaluation et de prévisualisation des données a été mis en place. Les expérimentations menées nous permettent de construire des jeux de données thématiques exhaustifs avec un niveau de bruit minimal et tenant compte des contraintes inhérentes aux réseaux sociaux.

4. Actions futures

En suivant les phases décrites dans notre cycle de vie, nous allons démarrer la phase 2, avec l'enrichissement puis l'extraction de concepts fins et d'entités nommées de ces données brutes via des méthodes de Traitement Automatique du Langage. En parallèle, nous travaillerons à l'intégration du concept de proxémique dans nos travaux, en l'adaptant aux données issues du cyber espace, et avec un focus sur les interactions dans les réseaux sociaux ayant trait au tourisme tout en proposant des méthodes d'analyse et d'extraction novatrices. Le but final restant toujours de concevoir un système de recommandation pour les touristes qui soit adapté à la région et différents outils d'aide à la décision pour les professionnels.

Bibliographie

Belias A., Malik A., S. Rossidis, I. Mantas (2021). Use of Big Data in Tourism: Current Trends and Directions for Future Research. *Academic Journal of Interdisciplinary Studies, vol.* 10, n° 5, p. 357.

Bucher B., Hein C., Raines D., Gouet Brunet V. (2021) Towards Culture-Aware Smart and Sustainable Cities: Integrating Historical Sources in Spatial Information Infrastructures. *ISPRS International Journal of Geo-Information,* vol. 10, n° 9, 588.

Scholz J., Jezni J. (2020) Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria. *ISPRS International Journal of Geo-Information,* vol. 9, n° 11, 681.

Hall E. T. (1966). *The hidden dimension,* Anchor, vol. 609.

Recommandations contextuelles fondées sur la fouille d'intentions et les ontologies

Ramona Elali

Centre de recherche informatique, Université Paris 1 Panthéon – Sorbonne
90 rue Tolbiac, Paris, France
ramona.elali@etu.univ-paris1.fr

MOTS-CLES : Fouille d'intentions, Recommandation, Modèle de processus intentionnel, Ontologie, Evènements, Contexte.

KEYWORDS : Intention Mining, Recommendation, Intentional Process Model, Ontology, Events, Context.

ENCADREMENT : Camille Salinesi, Rebecca Déneckère, Elena Kornyshova.

1. Introduction

Aujourd'hui, le défi de valoriser les données pour obtenir plus d'informations, augmenter la productivité, produire de meilleures performances et réduire les coûts devient une nécessité (Van der Aalst, 2016). Ainsi, les systèmes de recommandation sont d'une importance significative afin d'augmenter les revenus dans différents types de domaines. Néanmoins, les systèmes de recommandation existants ont montré certaines limites car ils se basent seulement sur les préférences de l'utilisateur et non pas sur ses intentions réelles. En général, les utilisateurs agissent selon leur objectif (Cheng *et al.*,2017) et leurs intentions. Il est donc important d'étudier la partie intentionnelle du processus en se basant sur la fouille d'intentions et les modèles de processus intentionnels. En effet, les modèles intentionnels sont flexibles et permettent de détecter le raisonnement de l'utilisateur qui affecte ses activités journalières (Khodabandelou *et al.*, 2013 ; 2014). La fouille d'intentions est devenue un sujet de recherche important dans de nombreux domaines de l'informatique et elle est considérée comme une technique intéressante pour la recommandation car elle se base sur la partie intentionnelle de l'activité de l'utilisateur (Khodabandelou *et al.*, 2013). L'objectif principal de la fouille d'intentions est d'identifier les intentions de l'utilisateur à la volée en découvrant les modèles de processus intentionnels reflétant son comportement et ses stratégies. L'utilisateur utilise ses propres intentions et stratégies pour accomplir ses tâches sans tenir compte des processus prescrits pour réaliser ses activités. MAP (Rolland et *al.*, 1999) est un modèle de processus intentionnel qui peut être utilisé pour prendre en compte les intentions de l'utilisateur tout en fournissant une flexibilité sous-jacente aux processus pour adopter différentes

stratégies afin d'atteindre une intention spécifique. D'autre part, les intentions et les stratégies de l'utilisateur sont affectées par divers facteurs contextuels tels que le lieu, la date et l'heure, la météo, le profil, etc. Ainsi, il est important de prendre en considération de multiples sources de données qui composent le contexte pour fournir des recommandations de meilleure qualité. Les informations contextuelles peuvent être structurées dans une ontologie de domaine en décrivant la relation entre les entités de l'ontologie de domaine et les informations contextuelles. Les ontologies sont en pleine évolution au sein des systèmes d'information (Viinikkala, 2004). De nombreux chercheurs utilisent les ontologies pour classer les connaissances du domaine (Mansingh *et al.*, 2011) telles que les concepts, les entités, et les relations qui existent entre eux (Viinikkala, 2004). Par la suite, ces ontologies de domaine seront utilisées comme entrées dans l'algorithme de fouille d'intentions afin de construire un modèle de processus intentionnel tel que MAP. Les modèles intentionnels découverts seront utilisés pour fournir des recommandations utiles en proposant à l'utilisateur les connaissances nécessaires et en lui indiquant la séquence optimale d'actions pour atteindre son intention. Ainsi, au cours de ce programme doctoral, notre objectif principal est d'élaborer une nouvelle approche pour faire des recommandations en s'appuyant sur la fouille d'intentions et en se basant sur la combinaison de modèles de processus intentionnels, d'ontologies de domaine et d'informations contextuelles afin de guider l'utilisateur.

2. Questions de recherche et approche proposée

2.1. Questions de recherche

La méthode de recherche du travail proposé comprend les étapes suivantes : la définition du problème et des objectifs de recherche, l'étude de la littérature, l'élaboration et la mise en œuvre de la proposition, l'évaluation et la validation. L'objectif principal de ce projet de recherche est de guider les utilisateurs en leur proposant des recommandations utiles qui sont basées sur le modèle de processus intentionnel découvert en fonction de ses activités actuelles. Par conséquent, le principal problème de recherche défini est le suivant : *Peut-on fournir des recommandations de bonne qualité aux utilisateurs en adoptant une combinaison entre les modèles de processus intentionnels, d'ontologies de domaine et d'informations contextuelles* ?

Afin de résoudre ce problème principal et de construire le modèle de processus intentionnel et le système de recommandation, nous avons défini les questions de recherche suivantes qui doivent être traitées :

- QR1 : Comment peut-on combiner l'ontologie du domaine, les informations contextuelles, etc. avec les logs d'événements sous une forme qui peut être utilisée comme une entrée pour l'algorithme de fouille d'intentions et celui de la recommandation ?
- QR2 : Comment peut-on définir un algorithme de fouille d'intentions qui va utiliser ces différentes sources de logs avec l'ontologie de domaine afin de construire un modèle de processus intentionnel ?

- QR3 : Comment peut-on créer un algorithme de recommandation qui va avoir comme entrées ces différentes sources de logs avec le modèle de processus intentionnel ?
- QR4 : Comment l'utilisation des ontologies de domaine, des informations contextuelles et d'un modèle de processus intentionnel (comme MAP) va améliorer les recommandations ?

2.2. Approche proposée

Comme les modèles intentionnels sont proches du raisonnement humain, nous trouvons que ces modèles peuvent fournir des recommandations efficaces aux utilisateurs s'ils sont construits en utilisant l'environnement contextuel. En effet, dans des contextes différents, le même utilisateur choisira des stratégies ou des intentions différentes pour atteindre son objectif. Notre approche comprend deux phases principales (voir Figure 1) :

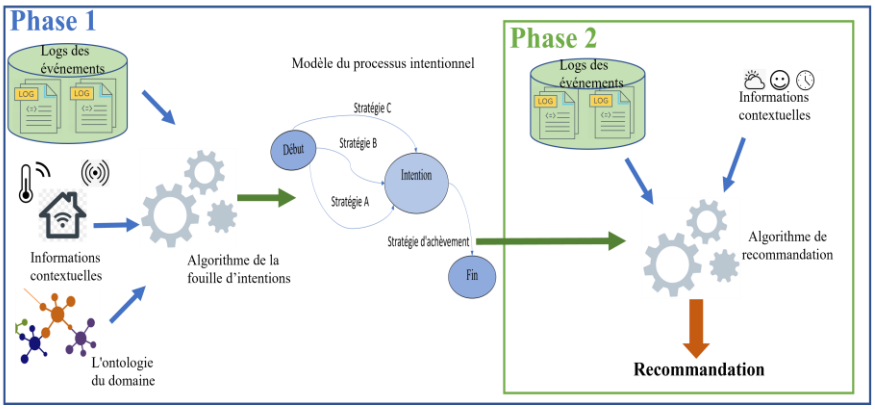

Figure 1. Approche proposée

- Dans la phase 1, différentes sources de logs (logs des événements, logs de capteurs, etc.) avec leurs informations contextuelles sont collectées. Ensuite, un algorithme de fouille d'intentions construit un modèle intentionnel. Cet algorithme découvrira les intentions et les stratégies du modèle intentionnel à partir des logs d'activités, avec l'aide de l'ontologie du domaine. Le modèle créé utilisera le formalisme MAP.

- Dans la phase 2, le modèle de processus intentionnel construit dans la phase 1 et les logs des activités actuelles de l'utilisateur avec ses informations contextuelles sont donnés comme paramètres d'entrée à l'algorithme de recommandation. Ensuite, après avoir traité les paramètres d'entrée, l'algorithme de recommandation fournira une recommandation appropriée pour l'utilisateur en fonction de son activité actuelle.

3. Conclusion et avancement

Nous sommes en train de réaliser une revue systématique de la littérature sur la fouille d'intentions basée sur (Déneckère *et al.*,2021) afin de fournir un état de l'art détaillé. En parallèle, nous travaillons sur un log d'activités relatives à une Smart Home afin de tester les hypothèses liées à la première question de recherche. Nous avons pu découvrir le lien entre les activités et les capteurs et construire les ontologies reliées au domaine étudié. Dans un avenir proche, nous prévoyons de commencer à travailler sur l'algorithme pour créer un modèle de processus intentionnel en utilisant les différents paramètres d'entrée de l'approche. Une vérification va être effectuée pour découvrir dans quelle mesure l'utilisation d'ontologies de domaine, d'informations contextuelles et de modèles de processus intentionnels est susceptible d'améliorer la qualité des recommandations. La méthode proposée sera validée par différents types d'évaluations pour vérifier la précision, la cohérence, la fiabilité et la confiance des résultats obtenus.

Bibliographie

Cheng, J., Lo, C., Leskovec, J., (2017). Predicting Intent Using Activity Logs: How Goal Specificity and Temporal Range Affect User Behavior. *Proceedings of the 26th International Conference on World Wide Web Companion*.

Déneckère, R., Kornyshova, E., Hug, C., (2021). A Framework for Comparative Analysis of Intention Mining Approaches. *International Conference on Research Challenges in Information Science (RCIS), May 2021*, Limassol, Cyprus. pp.20-37.

Khodabandelou, G., Hug, C., Deneckère, R., Salinesi, C., (2013). Process Mining Versus Intention Mining. *Lecture Notes in Business Information Processing*. 147. 466-480.

Khodabandelou, G., Hug, C., Deneckere, R., Salinesi, C., (2013). Supervised Intentional Process Models Discovery using Hidden Markov Models. *Seventh International Conference on Research Challenges in Information Science*, May 2013, Paris, France. pp.1-11.

Khodabandelou, G., Hug, C., Deneckere, R., Salinesi, C., (2014). Unsupervised discovery of intentional process models from event logs. *11th Working Conference on Mining Software Repositories*, May 2014, Hyderabad, India. pp.282-291.

Mansingh, G., Osei-Bryson, K. M., Reichgelt, H., (2011). Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences*, Volume 181, Issue 3, Pages 419-434.

Rolland, C., Prakash, N., Benjamen, A., (1999). A Multi-Model View of Process Modelling. *Requirements Engineering, Springer*, Verlag, pp.169 – 187.

Van der Aalst, W. M. P., (2016). Process Mining: Data Science in Action. *Springer*, Heidelberg. ISBN: 978-3-662-49850-7.

Viinikkala, M., (2004). Ontology in Information Systems.

Intent-Based Configuration of Information and Communication Components for Industry 4.0 Applications

Kaoutar Sadouki

Centre d'Etudes et de Recherche en Informatique et Communications (CEDRIC), Conservatoire National des Arts et Métiers (CNAM),
292 Rue Saint Martin, 75003 Paris, France
kaoutar.sadouki@lecnam.net

MOTS-CLES : Industrie 4.0, Technologies d'Information et de Communication, Approches Basées sur les Intentions.

KEYWORDS : Industry 4.0, Information and Communication Technologies, Intent-Based approaches.

ENCADREMENT : Elena Kornyshova, Eric Gressier Soudan.

1. Introduction

Smart factories with decentralized, flexible, self-organizing, and automated production environments are one of the main key features of Industry 4.0 (I4.0) (Büchi et al., 2020). I4.0 provides many opportunities to companies; however, it challenges them with the successful adoption and configuration of Information and Communication Technologies (ICT). To overcome these challenges, companies need to create and implement new business models based on distributed systems, to include the new ICT into their infrastructure.

Many studies have investigated the impact of I4.0-related technologies on business models (Grabowska *et al.*, 2020) or supply chains (Ghadge *et al.*, 2020). Yet, there is a lack of alignment of I4.0 ICT to the business strategy and business intentions. The existing literature still does not meet this urgent and pressing need.

The notion of intention or intent is very crucial for companies in this case, as it allows to understand the goals of internal and external users of a system. The perspective (intention-based) is gaining recognition in many fields like accounting (Leisenring *et al.*, 2012), intelligent mobility aid (Zhou *et al.*, 2010) and intent-based networking (IBN) (Han *et al.*, 2016) which is an emerging approach allowing the

configuration of the physical and virtual network infrastructure depending on business strategies requirements.

2. State of the Art

Numerous works are done in different scientific fields to deal with intentions (like intention mining, intention processing, and so on) or to use intentions as a mean for other purposes (intent-based networking, intention-oriented modelling for information system engineering, and so on). For instance, a taxonomy of intentions is proposed in the context of intention mining and contains six categories: feature request, opinion asking, problem discovery, solution proposal, information seeking, and information giving (Di Sorbo *et al.*, 2015). Intents can be written by humans or extracted using intention mining techniques from (user feedback, event logs, developer discussions ...), then they need to be transformed into device-consumable or lower-level forms (Cisco public, 2018). They can be thought of as high-level business or operational targets that a system should meet, so the basic goal of developing intent-based solutions is to meet system requirements without specifying how to accomplish them (Zeydan et Turk, 2020).

To prepare the literature review on intention-based approaches in ICT and their adoption by organizations, we need to understand the characterization of intents in the literature, the nature of intentions that will be beneficial for our configuration, the potential threats and the benefits gained, as well as the requirements to apply intent-based approaches in a given context. We chose the Systematic Mapping Study method (SMS) (Petersen *et al.*, 2008) to provide an overview of this research issue, present an unbiased assessment of the current literature, identify research gaps, and gather evidence for future research possibilities. We are following the steps of this method which contains: (a) Definition of the research question, (b) Searching for primary papers, (c) Analysis of papers, (d) Identification of keywords and classification of results, and (e) Data extraction and mapping process.

We have selected papers using the SCOPUS Search API. We searched for papers containing the term “Intent-based” or “Intention-based” only in the title and the number of obtained papers was more than 2100. For the time being, we are analyzing titles, the abstracts, and introductions to keep only the relevant papers. We have seen 140 papers until now, with 52 papers that were completely out of context, i.e. papers related to psychology.

Our main observations while surveying recent works related to intentions can be summarized as follows:

– Intent-based approaches: The principal motivation here is to introduce an intent-based perspective to gain new abilities that would not be in reach with the human workforce. The frequent approaches are linked to network management and orchestration like IBN, but intent-based approaches are increasingly becoming important in different areas, that we found techniques such as intent-based accounting, intent-based management, intent-based 3D illustration, etc.

– Technologies or solutions related to intents: different approaches to identify and to process intentions. Recent advances in artificial intelligence (AI) and natural language processing (NLP) have also led the knowledge to convert the user queries expressed into various representations of intentions that are adequately structured to be processed by an automated service.

3. Proposal Overview

Our proposal (illustrated at Figure 1) aims at connecting companies' objectives to digital technologies deployed to support it through a dynamic contextual configuration of new ICT components, using an intent-based approach.

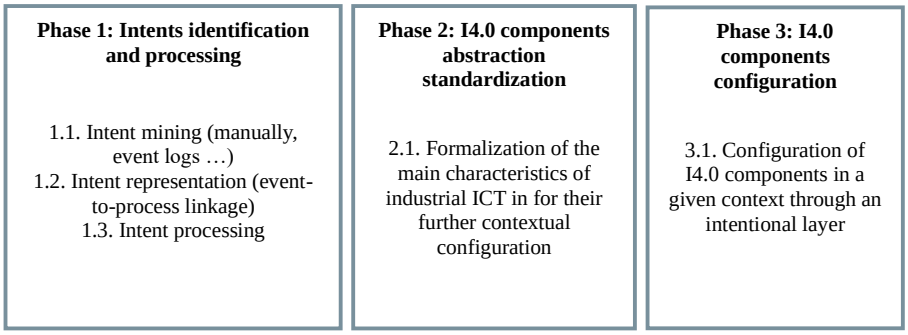

Figure 1. Overview of the Proposal.

Figure 1 presents the steps we will adopt to respond to our research goal: (i) elaboration a taxonomy of intentions adapted to the usage of ICT components; (ii) development of a framework for characterization and standardization of industrial ICT components, and their contextual selection and configuration; and (iii) elaboration of a framework allowing the configuration of ICT components depending on business and other goals through an intentional approach.

3. Discussion and Future Work

Organizations are facing some weaknesses in adopting efficient I4.0 solutions, and the main reasons can be related to the absence of formalized processes, lack of ICT knowledge as well as low-cost commercial systems (Dassisti et al., 2019). Thus, the necessity of a link between business intentions and I4.0 components would provide a context-aware adoption, by allowing the artifacts to be connected to business and other needs, which will provide the satisfaction of internal and external users. We intend to validate our approach by applying it to an industrial case related to 5G networks.

References

Brecher C., Müller A., Dassen Y., Storms S. (2021). Automation technology as a key component of the Industry 4.0 production development path. Int J Adv Manuf Technol 117, 2287–2295.

Büchi G., Cugno M., Castagnoli R. (2020). Smart factory performance and Industry 4.0. Technological Forecasting and Social Change. 150. 119790.

Cisco public (2018), Intent-Based Networking: Building the bridge between business and IT. https://www.cisco.com/site/us/en/products/networking/index.html.

Dassisti M., Giovannini A., Merla P., Chimienti M., Panetto H. (2019). An approach to support Industry 4.0 adoption in SMEs using a core-metamodel. Annual Reviews in Control, Volume 47.

Di Sorbo A., Panichella S., Visaggio C. A., Penta M. Di, Canfora G., Gall H. C. (2015). Development emails content analyzer: Intention mining in developer discussions (t). In Automated Software Engineering (ASE), 30th IEEE/ACM International Conference on. IEEE, pp. 12–23.

Ghadge, A., Er Kara, M., Moradlou, H. and Goswami, M. (2020), The impact of Industry 4.0 implementation on supply chains. In Journal of Manufacturing Technology Management, Vol. 31 No. 4, pp. 669-686.

Grabowska, S., Gajdzik, B., Saniuk, S. (2020). The Role and Impact of Industry 4.0 on Business Models. 10.1007/978-3-030-33369-0_3.

Han Y., Li J., Hoang D., Yoo J.-H., Won-Ki Hong J. (2016). An intent-based network virtualization platform for SDN. In the 12th International Conference on Network and Service Management (CNSM), 353-358.

Leisenring, James & Linsmeier, Thomas & Schipper, Katherine & Trott, Edward. (2012). Business-model (Intent)-based Accounting. Accounting and Business Research. 42. 329-344.

Petersen K., Feldt R., Mujtaba, S., Mattsson, M. (2008). Systematic mapping studies in software engineering. In the 12th Intl Conference on Evaluation and Assessment in Software Engineering, volume 17.

Zeydan E., Turk Y. (2020). Recent Advances in Intent-Based Networking: A Survey. In the 91st Vehicular Technology Conference (VTC2020-Spring), pp. 1-5. IEEE.

Zhou W., Xu L., Yang J. (2010). An intent-based control approach for an intelligent mobility aid. In the 2nd international asia conference on informatics in control, automation and robotics. Vol. 2. IEEE.

Application de GNN sur des graphes non-attribués : benchmark de performances.

Ikram Boukharouba

IRIT, Université Toulouse III Paul Sabatier, Toulouse/France
Berger Levrault, Labège/France

ikram.boukharouba@irit.fr

MOTS-CLES : Traces d'activité utilisateurs, Logs, Application industrielle, Graphe de navigation, Graph Neural Network (GNN), Graphe non-attribué.

KEYWORDS : Traces of user activity, Logs, Industrial application, Navigation graph, Graph Neural Network (GNN), Non-attributed graph.

ENCADREMENT : Florence Sèdes, Christophe Bortolaso, Florent Mouysset.

1. Introduction

Les traces utilisateurs peuvent être modélisées sous forme de graphes qui traduisent la navigation de l'utilisateur au sein d'un logiciel. Analyser ces graphes afin d'éliciter les comportements des utilisateurs nécessite l'utilisation de méthodes de machine learning traditionnelles. Or, les graphes de navigation générés à partir des ensembles de traces ne conviennent pas à l'application directe de ces méthodes : graphes incomplets, non labellisés,... Les Graph Neural Networks (GNNs) représentent des algorithmes de machine learning non traditionnels dédiés aux graphes qui permettent de pallier ces insuffisances en rendant possible la prédiction des liens et des nœuds manquants. L'avantage des GNNs réside dans leur capacité à prendre en considération la structure de graphe ainsi que les features de ses nœuds pour la tâche d'apprentissage**.** Néanmoins, cet avantage entrave aussi l'applicabilité des GNNs sur certains graphes « du monde réel », **lorsque peu de features sur les nœuds sont disponibles.** Plus précisément, les graphes de terrain manipulés peuvent s'avérer incomplets ; par exemple, en raison de problèmes de confidentialité : ces graphes sont qualifiés de non-attribués.

Dans cet article nous abordons le problème d'application de GNN sur les graphes non-attribués et montrons **l'importance du nombre de features** des nœuds pour le bon fonctionnement de GNN (section 1). Dans la section 2, un état de l'art sur les défis d'application de GNN sur des graphes non-attribués est présenté. Ensuite dans la section 3, nous présentons notre étude de cas sur la classification des nœuds sur notre graphe de terrain non-attribué ; ainsi qu'une comparaison entre les résultats sur

notre graphe de navigation et ceux obtenus sur des jeux de données standards du domaine dans la section 4. Enfin, dans la section 5. Nous concluons cet article et ouvrons par quelques perspectives.

2. État de l'art

L'extraction des informations pertinentes dans les graphes est devenue un problème important pour la communauté de l'exploration des données. Différents types de réseaux de neurones pour graphes (GNNs) ont prouvé leur efficacité pour la classification des nœuds (Wu *et al.*, 2019a) et la prédiction des liens (Wu *et al.*, 2019b). À notre connaissance, il existe très peu de travaux qui abordent l'importance de features des nœuds dans le fonctionnement des GNNs et qui proposent des solutions pour remédier au problème de manque de features dans les graphes. Duong *et al.* (Duong *et al.*, 2019) montrent la sensibilité des GNNs aux modifications des features des nœuds. Leur protocole expérimental consiste à intervertir les features des nœuds du graphes sans changer les étiquettes. Une revue sur les techniques d'initialisation de features artificielles pour l'application des GNNs sur les graphes non attribués est présentée aussi. Ces techniques peuvent être classées en 2 catégories : les techniques basées sur la centralité comme Degree (Rossi *et al.*, 2017) et Egonet (Henderson *et al.*, 2012). Et les techniques basées sur l'apprentissage : DeepWalk (Perozzi *et al.*, 2014), HOPE(*Ou et al.*, 2016)). Les résultats prouvent que les features artificielles donnent les mêmes résultats que les features originales voir mieux dans certains cas. Cependant, ce travail a été validé que sur des datasets standards.

Dans la continuité de Duong, Cui *et al* (Cui *et al.*, 2021) proposent une revue sur les techniques de générations des features artificielles des nœuds pour les graphes non-attribuées. Ils regroupent les features de nœuds en deux grandes familles : les features de nœuds positionnelles, basées sur la position dans le graphe. Ainsi que les features de nœuds structurels qui capturent les informations structurelles des nœuds. Néanmoins, leur travail n'est testé et validé que sur des datasets standards. Zhou *et al.* (Zhou *et al.*, 2021), proposent une approche pour la classification des nœuds dans les graphes multi-étiquetées. Leur contribution se base sur l'exploitation simultanée des informations de structure dans le graphe, ainsi que les features et les étiquettes de nœuds. Cette méthode prouve son efficacité aussi bien sur les graphes attribués que non attribués où les résultats étaient fortement affectés par l'absence de features.

3. Cas d'utilisation et expérimentations

Lorsqu'un utilisateur navigue d'une page à une autre, il crée un lien entre ces pages et ça génère un graphe de navigation. Dans notre cas d'étude, nous analysons les traces utilisateurs de logiciel de gestion Sedit. Ce que nous cherchons à faire est de trouver la distribution des nœuds par module de l'application ; un module regroupant des pages métier. Nous considérons 3 datasets standards avec des features de nœuds: Cora (Sen *et al.*, 2008), Citesser (Sen *et al.*, 2008), et Pubmed (Namata *et al.*, 2012). Également nous considérons notre ensemble de données

Sedit. Les statistiques des données sont présentées dans le Tableau 1 Afin de montrer l'importance des features des nœuds, notre protocole d'évaluation est le suivant : tester les GNNs avec l'algorithme GraphSAGE (Hamilton et al., 2017) sur les 4 datasets avec les mêmes paramètres. Les deux méthodes d'agrégation : Sum et Mean ont été utilisées (voir Tableau 1).

Tableau 1. Jeux de données et résultats de la classification de nœuds avec les GNNs.

Jeux de données	Nb nœuds	Nb liens	Nb features	Nb classes	Taux de classification	
					Mean	Sum
Cora	2.708	8.278	1.433	7	80,15%	70,32%
Citesser	3.312	4.614	3.703	6	73,81%	59,01%
PubMed	19.717	44.325	500	3	77,88%	75,15%
Sedit	**108**	**17.192**	**6**	**6**	**19,45%**	**16,25%**

4. Observations et discussion

Méthodes d'agrégation : la méthode Mean présente de meilleur taux que la méthode Sum sur les 4 datasets. En effet, la méthode d'agrégation Sum peut filtrer efficacement l'influence de la structure du voisinage, ce qui contribue peu à la performance de la classification des nœuds dans les datasets positionnels comme les datasets de citations. De manière similaire aux précédents datasets Sedit semble sensible à la position. En effet compte tenue de peu d'informations sur les transitions métiers sur les pages ces résultats sont prévisibles

Les performances : le Tableau1 montre des résultats similaires entre les datasets attribués. Les performances de classification des nœuds avec les GNNs sur les datasets de graphes attribués ont largement surpassés ceux sur le jeu de données non-attribué. Par exemple, sur l'ensemble de données Cora, le taux de précision des classifications le plus élevé, parmi les méthodes de centralité, est de 80.15%, soit 60% de plus que le taux le plus élevé de classification sur Sedit. Ce qui est prévisible. Étant donnée le peu de features sur Sedit. La comparaison entre les performances obtenues avec les GNNs sur les datasets des graphes attribués et notre jeu de données industriel Sedit prouvent l'importance des features pour le bon fonctionnement de GNN.

5. Conclusion et futurs travaux

Dans cet article, nous avons abordé l'application des GNNs sur des graphes non attribués. Notre état de l'art montre l'importance des features de nœuds dans le fonctionnement des GNNs et la dégradation de ces performances en leurs absence. Notre hypothèse est que les GNN peuvent classifier les écrans d'une application par

module métier. . Quelques travaux existants montrent que le faible nombre de features est un inconvénient pour les GNNs. Cet article reproduit des conditions d'expérimentation et conduit à confirmer ces résultats. Enfin, notre originalité est dans la vérification de ces limites avec un corpus de terrains orienté « traces d'activités utilisateurs ».

Dans la continuité de notre travail, nous envisageons de générer des features artificielles sur nos corpus de données afin de valider les techniques proposées dans l'état de l'art sur un jeu de données de terrain.

Bibliographie

Cui H., Lu Z., Li P., Yang C., « On positional and structural node features for graph neural networks on non-attributed graphs », arXiv preprint arXiv :2107.01495, 2021.

Duong C. T., Hoang T. D., Dang H. T. H., Nguyen Q. V. H., Aberer K., « On node features for graph neural networks », arXiv preprint arXiv :1911.08795, 2019.

Hamilton W., Ying Z., Leskovec J., « Inductive representation learning on large graphs », Advances in neural information processing systems, 2017.

Henderson K., Gallagher B., Eliassi-Rad T., Tong H., Basu S., Akoglu L., Koutra D., Faloutsos C., Li L., « Rolx: structural role extraction & mining in large graphs », Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 1231-1239, 2012.

Namata G., London B., Getoor L., Huang B., EDU U., « Query-driven active surveying for collective classification », 10th International Workshop on Mining and Learning with Graphs, vol. 8, p. 1, 2012.

Ou M., Cui P., Pei J., Zhang Z., Zhu W., « Asymmetric transitivity preserving graph embedding », Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, p. 1105-1114, 2016.

Perozzi B., Al-Rfou R., Skiena S., « Deepwalk: Online learning of social representations », Proceedings of the 20th ACM SIGKDD international conference on Knowledge disco very and data mining, p. 701-710, 2014.

Rossi R. A., Zhou R., Ahmed N. K., « Deep feature learning for graphs », arXiv preprint arXiv :1704.08829, 2017.

Sen P., Namata G., Bilgic M., Getoor L., Galligher B., Eliassi-Rad T., « Collective classification in network data », AI magazine, vol. 29, no 3, p. 93-93, 2008.

Wu J., He J., Xu J., « Net: Degree-specific graph neural networks for node and graph classification », Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 406-415, 2019a.

Wu Y., Lian D., Jin S., Chen E., « Graph Convolutional Networks on User Mobility Heterogeneous Graphs for Social Relationship Inference. », IJCAI, p. 3898-3904, 2019b.

Zhou C., Chen H., Zhang J., Li Q., Hu D., Sheng V. S., « Multi-label graph node classification with label attentive neighborhood convolution », Expert Systems with Applications, vol. 180, p. 115063, 2021.

Cybersécurité à l'échelle intersystème d'information avec prise en compte du facteur humain

Olivier de Casanove

IRIT, Université Toulouse III – Paul Sabatier
118 route Narbonne 31062 Toulouse CEDEX 9
olivier.decasanovenom@irit.fr

MOTS-CLES : Cybersécurité, Sécurité des Systèmes d'Information, Détection d'Evènements, Prévention.

KEYWORDS : Cybersecurity, Information System Security, Event Detection, Prevention.

ENCADREMENT : Florence Sèdes.

1. Introduction

L'ANSSI (Agence Nationale de la Sécurité de Systèmes d'Information) définit la cybersécurité ainsi (ANSSI, 2022) : "État recherché pour un système d'information lui permettant de résister à des évènements issus du cyberespace susceptibles de compromettre la disponibilité, l'intégrité ou la confidentialité des données stockées, traitées ou transmises [...]". Cette définition ne fait pas l'unanimité et chaque pays a sa propre vision de la cybersécurité, mais elle suffit pour en comprendre les enjeux principaux. Le système d'information (SI) est au cœur des politiques de cybersécurité, l'objectif est de le protéger. Les évènements (ou incidents) de sécurité auxquels fait référence la définition sont décrits par une chronologie qui peut être découpée en plusieurs étapes. Comme pour la définition de la cybersécurité, il existe de nombreuses versions de la chronologie d'un évènement de sécurité. La plus simple et explicite est celle du NIST (National Institute of Standards and Technology) qui reconnait quatre étapes (Scarfone *et al.*, 2008) : préparation, prévention ; détection, identification ; isolement, éradication, remédiation ; apprentissage. Les solutions de cybersécurité développées à chaque étape de cette chronologie sont déployées à l'échelle du SI. Cependant, avec l'étude de Code Red, un ver informatique, Moore *et al.* ont pour la première fois formalisé le fait que la cybersécurité d'un SI dépend aussi de la sécurité des autres SI (Moore *et al*., 2002). Ce changement de paradigme implique que les outils développés actuellement ne sont pas suffisants pour détecter les incidents de sécurité, puisqu'ils ne sont

utilisables qu'à une échelle locale. Dans cet article, nous proposons une approche non plus à l'échelle du SI, mais à une échelle intersystème d'information. Dans la section 2 nous expliquons comment nous comptons détecter les attaques informatiques à l'échelle intersystème d'information. Dans la section 3 nous proposons une approche permettant de tirer avantage de la détection pour améliorer la cybersécurité. Nous concluons dans la section 4.

2. Détection à l'échelle intersystème d'information

Pour détecter les attaques à l'échelle intersystème d'information, il faut un jeu de données à la même échelle. Nous avons choisi d'utiliser les réseaux sociaux pour cela, car la surveillance des réseaux sociaux pour détecter les attaques informatiques s'est déjà montrée efficace (Khandpur *et al.*, 2017, Ritter et al., 2015, Sabottke et al., 2015, Sceller *et al.*, 2017). Pour détecter les attaques, nous utiliserons donc des algorithmes de détection d'évènements : (Atefeh et Khreich, 2015) en fournissent une revue de la littérature.

Le principal inconvénient de cette approche est que n'importe quel utilisateur malveillant peut poster des messages sur les réseaux sociaux. Les données d'entrées sont donc facilement corruptibles. La littérature sur la détection de spams, sujet sur lequel nous avons déjà travaillé dans l'équipe (Washha et al., 2016), est abondante et détaillée dans plusieurs revues de la littérature récentes (Gheewala et Patel, 2018, Tingmin et al., 2018, Yurtseven *et al*., 2021). Cependant, dans notre cas, le problème s'avère plus complexe. Les algorithmes de détection sont des algorithmes d'apprentissage automatique et sont donc sensibles à l'apprentissage adverse (ou *adversarial learning* en anglais). L'apprentissage adverse est une technique utilisée par un attaquant pour exploiter à son avantage un algorithme d'apprentissage automatique qui ne lui appartient pas, le plus souvent dans le but d'altérer les données de sorties de l'algorithme (Xianmin *et al.*, 2019). Dans notre cas cela voudrait dire qu'un attaquant publie des messages sur les réseaux sociaux, spécialement pour que la détection d'évènements soit moins efficace. Dans notre contexte, détecter des évènements de cybersécurité, il n'est pas pensable de concevoir un système qui ne puisse pas être résilient à des acteurs malveillants. Il existe une littérature sur l'apprentissage adverse sur les réseaux sociaux, mais elle est centrée autour de la question de détection de spams (Imam et Vassilakis, 2018). Nous avons contribué à l'état de l'art grâce à une première typologie pour modéliser les enjeux de l'apprentissage adverse dans le contexte de la détection d'évènements et nous travaillons actuellement à développer des algorithmes permettant de s'en défendre.

La détection d'évènements de sécurité à l'échelle intersystème d'information a déjà un intérêt en soi, mais il est possible d'en tirer une plus-value supplémentaire grâce à la prévention.

3. Prévention, Éducation et Sensibilisation à la sécurité

Les *malwares* pouvant passer d'un ordinateur à l'autre, la sécurité d'un SI ne dépend plus de sa propre sécurité uniquement, mais de la sécurité de toutes les machines connectées à internet. Moore *et al.* formalisent alors pour la première fois l'aspect épidémique des attaques informatiques *(*Moore *et al.*, 2002). Les auteurs proposent trois moyens de lutte : la prévention, la remédiation et l'isolation. Les auteurs ont traité le sujet de l'isolation, nous choisissons de nous intéresser à la prévention dans ce travail. La prévention en sécurité informatique n'est pas un concept nouveau, mais est souvent vue comme un plus, quelque chose de non nécessaire. Une des raisons de la mauvaise utilisation de la prévention est la difficulté de pouvoir quantifier le nombre d'attaques évitées et ainsi prouver l'utilité de la prévention. Pour contrer ce phénomène, nous avons établi une revue de la littérature sur la prévention dans le milieu de la cybersécurité (de Casanove et Sèdes, 2021). Dans notre revue nous proposons d'adapter les campagnes de prévention au cycle PDCA (Plan Do Check Adjust) pour que la prévention puisse être utilisée comme un outil de cybersécurité à part entière qui s'intègre dans une politique globale, et non plus comme un agrément. L'objectif est de déclencher des campagnes de prévention au moment opportun pour lutter contre les attaques informatiques, moment qui est détecté au plus tôt grâce à notre proposition de système de détection intersystème d'information. Dans le meilleur des cas, la campagne de prévention arrive avant que le SI soit compromis et on évite une compromission. Dans le pire des cas, le système est déjà compromis et la campagne permet aux utilisateurs d'être plus vigilants et de noter plus tôt des comportements étranges ou inhabituels du SI.

4. Conclusion

Dans la première section, nous avons énoncé les limites du modèle actuel de la cybersécurité qui se concentre sur la détection et le traitement des attaques informatiques à l'échelle du SI. Dans la deuxième section, nous avons vu comment nous comptons implémenter un système de détection de cyberattaque à l'échelle intersystème qui puisse être résilient à l'apprentissage adverse. Dans la troisième section, nous valorisons la détection intersystème en utilisant la prévention comme moyen de lutte contre les cyberattaques.

Bibliographie

ANSSI. (2022). *Glossaire cybersécurité*, https://www.ssi.gouv.fr/administration/glossaire/c/

Atefeh F., Khreich W. (2015). *A Survey of Techniques for Event Detection in Twitter.Computational Intelligence*, vol. 31, no1, p. 132–164.

de Casanove O., Florence Sèdes S. (2022). *Extracting Guidelines for Security Education,Training and Awareness Programme from the Literature.* (working paper or preprint).

Gheewala S., Patel R.. (2018). Machine Learning Based Twitter Spam Account Detection: A Review. *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 79-84.

Imam N. H., Vassilakis V.G. (2019).. A Survey of Attacks Against Twitter Spam Detectors in an Adversarial Environment. *Robotics* 8, no. 3: 50.

Khandpur R. P., Ji T., Jan S., Wang G., Lu C.-T., Ramakrishnan N. (2017). Crowdsourcing cybersecurity: Cyber attack detection using social media. InProceedings of the *2017 ACM on Conference on Information and Knowledge Management*, p. 1049–1057.

Le Sceller Q., Karbab E. M. B., Debbabi M., Iqbal F. (2017). SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream. *In Proceedings of the 12th International Conference on Availability, Reliability and Security (ARES '17).* Association for Computing Machinery, New York, NY, USA, Article 23, 1–11.

Moore D., Voelker G. M., Savage S. (2002). *Quantitative network security analysis.* Cooperative Association for Internet Data Analysis (CAIDA), NSF-01-160, vol. 7.

Ritter A., Wright E., Casey W., Mitchell T. (2015). Weakly supervised extraction of computer security events from twitter. InProceedings of the 24th international conference on worldwide web, p. 896–905. Republic and Canton of Geneva, CHE, *International World Wide Web Conferences Steering Committee.*

Sabottke C., Suciu O., Dumitras T. (2015). Vulnerability disclosure in the age of so-cial media: Exploiting twitter for predicting real-world exploits. In *24th USENIX security symposium* (USENIX security 15), p. 1041–1056. Washington, D.C., USENIX Association.

Scarfone K. A., Grance T., Masone K. (2008). *Computer Security Incident Handling Guide.*

Tingmin Wu, Sheng Wen, Yang Xiang, Wanlei Zhou (2018). Twitter spam detection: Survey of new approaches and comparative study, *Computers & Security*, Volume 76, Pages 265-284.

Washha M., Qaroush A., Sedes F. (2016). Leveraging time for spammers detection on twitter. InProceedings of the *8th international conference on management of digital ecosystems*, p. 109–116. New York, NY, USA, Association for Computing Machinery.

Xianmin Wang, Jing Li, Xiaohui Kuang, Yu-an Tan, Jin Li. (2019). The security of machine learning in an adversarial setting: A survey, *Journal of Parallel and Distributed Computing*, Volume 130, Pages 12-23.

Yurtseven I., Bagriyanik S. and Ayvaz S. (2021). A Review of Spam Detection in Social Media, *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pp. 383-388.

Biographies des auteurs

Antoine Aubé est doctorant CIFRE en deuxième année dans le département Traitement de l'Information et Système (DTIS) de l'Office National d'Études et de Recherches Aérospatiales (ONERA) de Toulouse et au sein de Stack Labs, entreprise de conseil spécialisée dans l'informatique en nuage et les nuages publics. Diplômé ingénieur de l'école Polytech Nice, il débute sa thèse après trois ans passés dans l'industrie en tant qu'ingénieur "cloud". Ses recherches se concentrent sur la conception de systèmes infonuagiques.

Ali Benijlany est en première année de doctorat au Laboratoire des Sciences du Numérique de Nantes. Originaire du Maroc, il y a fait toutes ses classes : 2ans de classes préparatoires MPSI/MP au Lycée Salmane Al Farissi (Salé) puis le cursus d'ingénieur à l'École Nationale d'Informatique et d'Analyse des Systèmes à Rabat, avant de rejoindre la France en 2021 pour préparer sa thèse. Une thèse qui s'intitule : « Une approche orientée services pour l'alignement Business-IT » et qui vise à proposer des solutions pour réconcilier les processus métiers et les portefeuilles applicatifs des entreprises.

Mustapha Kamal Benramdane est ingénieur d'Etat et titulaire d'un master en informatique de l'Ecole nationale supérieure d'informatique (ESI) d'Alger (2019-2020). A la fin de ses études, il intègre la société de logiciels MUST SAS en tant que développeur web back-end junior. En 2021, il rejoint le CNAM en tant que doctorant sous la direction de Pr. Samia Bouzefrane et de Dr. Elena Kornyshova et travaille sur un sujet intitulé : Orchestration contextuelle basée sur l'intention au sein des écosystèmes et des plateformes d'affaires numériques.

Ikram Boukharouba est doctorante en première année dans le cadre d'une thèse CIFRE supervisée par la professeure Florence Sèdes à l'Institut de Recherche en Informatique de Toulouse (IRIT). Le titre de sa thèse est : "Analyses Comportementales d'Utilisateurs de Logiciels de Gestion". Auparavant, elle a travaillé sur la l'analyse de traces logicielles pour la détection de fraude et la génération de graphes de navigation ; et sur d'autre projets en sécurité informatique et cryptographie. Elle occupe le poste d'ingénieure R&D au sein du département de recherche et d'innovation à Berger-Levrault depuis 2020.

Anouck Chan est doctorante depuis janvier 2021 dans le département Traitement de l'Information et Systèmes de l'ONERA (Office National d'Études et de Recherches Aérospatiales) sous la direction de Thomas Polacsek et de Stéphanie Roussel. Sa thèse s'intituleModélisation d'exigences multidimensionnelles pour la conception simultanée de satellites et d'avions et combine les domaines de l'ingénierie des exigences et de la recherche opérationnelle.

Olivier de Casanove est en 2ème année de doctorat à l'Université Toulouse III – Paul Sabatier, financé par le ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Il est membre de l'Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505 CNRS). Son sujet de recherche touche à la sécurité des systèmes d'information et plus particulièrement la place de l'humain dans la conception de solutions de sécurité.

Ramona Elali est une doctorante en informatique en 2ème année à l'Université Paris 1 – Panthéon Sorbonne à Paris. Son champ de recherche est la fouille des intentions sous la direction du Prof. Camille Salinesi, Dr. Rébecca Déneckère et Dr. Elena Kornyshova. Elle a reçu un master de recherche en Système d'information et Intelligence de données et un master professionnel en Développement Web de l'Université Libanaise au Liban.

C'est en dernière année de licence de biologie que **Solweig Hennechart** a découvert la plus-value que représente l'informatique pour la science. Elle a donc une formation initialement orientée biologie qu'elle a pu compléter en intégrant un master de bio-informatique dans lequel ont été enseignées les bases de la programmation. Mais c'est en stage à l'INRAE sur la question de la réutilisation de données en métabolomique qu'elle a réellement confirmé son intérêt, et qu'elle a eu la chance de pouvoir intégrer ce projet de thèse pluridisciplinaire.

Maël Lesavourey a suivi un cursus classes préparatoires au lycée Louis Barthou de Pau puis s'est spécialisé en mathématiques appliquées grâce au parcours Science et Ingénierie des Données (SID) de l'Université Toulouse 3 Paul Sabatier. Il est doctorant en recherche d'information au laboratoire IRIT (Institut de Recherche en Informatique de Toulouse). Il travaille sous la direction de Gilles Hubert et Yoann Pitarch (IRIT) et Fabien Jourdan (Toxalim - INRAE). Sa thèse, cofinancée par la région Occitanie et l'Université Fédérale de Toulouse, a pour but d'améliorer la recherche bibliographique des chercheurs en toxicologie et biomédecine.

Maxime Masson est doctorant en 1ère année au laboratoire LIUPPA de l'Université de Pau et des Pays de l'Adour. Il a reçu son diplôme de Master « Technologie de l'Internet » de cette même université en 2021 et réalisé un stage de recherche de 6 mois dans le cadre du projet régional Nouvelle-Aquitaine DA3T (Dispositif d'Analyse des Traces numériques pour la valorisation des Territoires Touristiques). Il travaille désormais, dans le cadre de sa thèse intitulée « Services proxémiques augmentés pour le patrimoine culturel et les pratiques touristiques », sur l'extraction ciblée de données issues des réseaux sociaux, le traitement automatique du langage (TAL) et l'adaptation de la théorie de la proxémique.

Kaoutar Sadouki est une doctorante de première année sous la supervision d'Elena Kornyshova et Eric Gressier Soudan au Centre d'Etudes et de Recherche en Informatique et Communications (CEDRIC) du Conservatoire National des Arts et Métièrs (CNAM) de Paris. Au travers de sa thèse, elle s'intéresse aux défis liés à l'adaptation et à la configuration des technologies de l'information et de la communication à l'écosystème Industrie 4.0 à partir du concept d'intention et des approches associées comme le machine learning et l'intelligence artificielle.

Après une formation initiale en sciences de l'information et de la communication et un parcours professionnel en gestion de projets (jeux pédagogiques), **Chloé Vigneau** s'est orientée vers la recherche en informatique en lien avec l'éducation. Elle s'est inscrite depuis janvier 2021 à la Chaire Science et Jeu Vidéo de l'Ecole Polytechnique et au CNAM - ENJMIN sous la direction d'Axel Buendia et la coordination de Catherine Rolland et Stéphanie Mader. Ses recherches portent sur l'apprentissage en classe à travers une activité de création de jeu vidéo. Elle pilote également un groupe de travail numérique du Ministère de l'Education Nationale initié autour de ces recherches.

Résumés des articles

Conception d'un système d'évaluation d'une activité de création de vidéo pour les enseignants de seconde

Chloé Vigneau[1]

1. Chaire Science et Jeu vidéo, Laboratoire Leprince-Ringuet, Ecole Polytechnique, Route de Saclay, 91120 Palaiseau
Laboratoire CEDRIC, CNAM, 2 rue conté 75003 PARIS
chloe.vigneau@polytechnique.edu

RESUME : Cet article porte sur l'évaluation d'une activité de création de jeu vidéo en classe. L'objectif est de répondre à une double problématique de récolte et de traitement de données pertinentes pour évaluer les compétences acquises par les élèves lors de cette activité. Nous proposons une analyse du processus créatif, d'une part, et du socle pédagogique, d'autre part, en vue de créer un modèle de données permettant de produire et d'analyser des traces d'apprentissage. Nous proposons d'implémenter ce modèle dans des « templates de jeu », base d'un environnement informatique dédié à l'apprentissage à travers cette activité de création de jeu vidéo.

ABSTRACT. This article deals with the evaluation of a video game creation activity in class. The objective is to answer the twofold problem of collecting and processing relevant data to evaluate the skills acquired by students during this activity. We propose an analysis of the creative process, on the one hand, and of the pedagogical base, on the other hand, in order to create a data model allowing the production and analysis of learning analytics. We propose to implement this model in "game templates", the basis of a computer environment dedicated to learning through this video game creation activity.

MOTS-CLES : Jeu vidéo, Evaluation, Projet interdisciplinaire.

KEYWORDS : Video game, Assessment, Interdisciplinary project.

ENCADREMENT : Axel Buendia, Stéphanie Mader, Catherine Rolland.

Co-conception de logiciels et de leur environnement d'exécution infonuagique

Antoine Aubé

Office National d'Études et de Recherches Aérospatiales
2, avenue Édouard Belin, 31055 Toulouse, France
antoine.aube@onera.fr

RESUME : L'émergence de l'informatique en nuage et des nuages publics est une opportunité pour les organisations de réduire le budget de leur système d'information. Migrer vers un nuage inclue la sélection d'un environnement d'exécution pour les logiciels du système d'information via la configuration de services infonuagiques. Si de nombreuses approches assistent la sélection de l'environnement, elles l'abordent sous l'angle des contraintes techniques des logiciels existants (e.g. quantité de mémoire nécessaire à l'exécution), ce qui limite dans les faits le nombre de services infonuagiques utilisables : la plupart d'entre eux requièrent des adaptations des logiciels. Dans l'industrie, les migrations sont plutôt dirigées par les besoins des utilisateurs et de l'organisation, mais ils sont difficiles à prendre en compte à cause de spécificités de l'informatique en nuage. L'objectif de notre projet de recherche est d'aider la conception de migrations performantes vers un nuage public en satisfaisant les exigences des utilisateurs et des organisations.

ABSTRACT. The emergence of Cloud Computing and public clouds is an opportunity for organizations to reduce the budget for their information system. To migrate to a cloud involves the selection of an information system's software execution environment through the configuration of cloud services. If many approaches assist the selection of the environment, they address it from the perspective of the existing software's technical constraints (e.g. RAM amount needed for execution) which limits the number of usable cloud services: most of them require software adaptations. In the industry, migration is rather driven by the needs of both users and organization, but they are difficult to take into account due to specificities of Cloud Computing. The objective of our research project is to help the design of efficient migration to a public cloud by meeting the requirements of users and organizations.

MOTS-CLES : Informatique en nuage, Conception simultanée, Ingénierie des exigences.

KEYWORDS : Cloud Computing, Co-design, Requirements engineering.

ENCADREMENT : Thomas Polacsek, Clément Duffau.

Aide à la co-conception de produits complexes

Anouck Chan

ONERA/DTIS, Université de Toulouse,
F-31055 Toulouse, France
anouck.chan@onera.fr

RESUME : La création d'un produit complexe implique non seulement la conception du produit lui-même, mais aussi de son moyen de production. Dans le domaine aéronautique, le cycle de développement produit/moyen de production est traditionnellement séquentiel : l'avion est d'abord conçu, puis sa ligne d'assemblage. Cette méthode peut cependant mener à des designs d'avions qui s'avèrent difficilement ou non réalisables. Une méthode alternative permettrait d'éviter ces écueils : la conception simultanée. Ainsi, dans cette thèse, nous nous intéressons à élaborer une aide à la conception simultanée d'un produit et de son moyen de production, notamment dans le domaine aéronautique. Pour cela nous utilisons des modèles orientés buts pour éliciter les objectifs, les relations et les acteurs de cette double conception. Nous employons également des outils de recherche opérationnelle pour définir des critères d'évaluation du moyen de production en fonction de son produit.

ABSTRACT. When a complex product such as an aircraft is conceived, its means of production has to be too. This double conception is due to different factors like the impacts of aircraft design choice on the assembly line organisation. In the aeronautical field, the aircraft is usually designed first, then assembly line. This order may lead to messy situations where the aircraft is hardly manufacturable. This is why we are interested in the simultaneous design method, where the aircraft and its assembly line are designed together. We investigate support for the simultaneous design of complex products. To do this we use goal-oriented methods to elicit goals, relationships and actors of this double conception, and operational research tools help us to define some evaluation criteria of a good assembly line of a specific product.

MOTS-CLÉS : Ingénierie des exigences, Modélisation, Optimisation, Formalisation.

KEYWORDS : Requirements engineering, Modelling, Optimisation, Formalisation.

ENCADREMENT : Thomas Polacsek, Stéphanie Roussel.

Intent-based Contextual Orchestration inside Digital Business Ecosystems and Platforms

Mustapha Kamal Benramdane

CEDRIC, CNAM, 292 rue Saint Martin, 75003 Paris, France
mustapha-kamal.benramdane@lecnam.net

RESUME : Différents acteurs économiques, notamment des entreprises appartenant à différents secteurs de marchés peuvent se rencontrer sur les plateformes numériques afin d'échanger. Le développement de ce réseau peut aboutir à l'apparition d'un écosystème d'affaire numérique (DBE). Toutefois, la complexité de ces échanges, et l'intensité du flux d'information échangés peut rendre difficile pour une entreprise de trouver des partenaires répondant à ses besoins. Dans notre travail de recherche, nous proposons une méthode permettant d'identifier les meilleurs clients, fournisseurs et autre type de partenaires basée sur les intentions des utilisateurs et le contexte économique et commercial de ces entreprises. Ce système de recommandation est basé sur le matchmaking avec des algorithmes d'apprentissage automatique hybride supervisé. Le travail actuel s'insère dans la recherche d'une méthode plus générique d'orchestration de ces entreprises, dites entités, au sein d'un écosystème d'affaire numérique.

ABSTRACT. Different economic actors, notably companies belonging to different market segments can meet on digital platforms in order to exchange. The development of this network can lead to the appearance of a Digital Business Ecosystem (DBE). However, the complexity of these exchanges, and the intensity of the flow of information exchanged can make it difficult for a company to find partners that meet its needs. In our research, we propose a method to identify the best customers, suppliers and other kinds of partners based on the intentions of users and the economic and business context of these companies. This recommendation system is based on matchmaking with supervised hybrid machine learning algorithms. The current work is part of the search for a more generic method of orchestrating these companies, called entities, within a Digital Business Ecosystem.

MOTS-CLES : Écosystèmes et Plateformes Digitaux, Orchestration des DBE, Configuration Contextuelle, Edge Cloud et Edge Computing, Apprentissage Automatique.

KEYWORDS : Digital Business Ecosystems, Digitally Enabled Collaborations, DBE entities' Orchestration, Contextual Orchestration, DBE configuration, Machine Learning.

ENCADREMENT : Samia Bouzefrane, Elena Kornyshova, Hubert Maupas.

Cohérence des systèmes d'information

Alignement opérationnel des applications sur le métier

Ali Benjilany

1. Nantes Université, CNRS, LS2N, F-44000 Nantes,
52, rue de la houssinière, BP 92208 - 44322 Nantes Cedex 3, France
ali.benjilany@ls2n.fr

RESUME : *Une préoccupation majeure des DSI est de fournir des moyens adaptés pour mettre en œuvre le système d'information informatisé. Il faut en particulier rendre cohérents les visions des différents acteurs, on parle alors d'alignement Business-IT. L'architecture d'entreprise est une approche qui permet d'exprimer les différents points de vue. Elle doit s'accompagner d'indicateurs d'aliment pour exprimer cette cohérence. Si l'alignement stratégique a fait l'objet de nombreux travaux de recherche, l'alignement opérationnel, qui est la vraie préoccupation des entreprises, reste limitée. Pourtant il est fondamental pour prédire l'évolution des systèmes d'information. Nous proposons ici un état de l'art des approches existantes, une critique ce ces approches et des pistes de recherche en guise de solution.*

ABSTRACT. *A major concern of CIOs is to provide appropriate means to implement the computerized information system. In particular, it is necessary to make the visions of the different actors coherent, we then speak of Business-IT alignment. Enterprise architecture is an approach that allows different points of view to be expressed. It must be accompanied by indicators to express this consistency. While strategic alignment has been the subject of much research, operational alignment, which is the real concern of companies, remains limited. Yet it is fundamental to predict the evolution of information systems. We propose here a state of the art of the existing approaches, a criticism of these approaches and avenues of research as a solution.*

MOTS-CLES : *Système d'information, Urbanisation, Architecture d'entreprise, Alignement Business-IT, Méthodes d'alignement.*

KEYWORDS : *Information System, Urbanization, Enterprise Architecture, Business-IT Alignment, Alignment Methods.*

ENCADREMENT : *Pascal André, Hugo Brunelière, Dalila Tamzalit.*

Approche pour la recherche d'information multifacette en biomédecine

Maël Lesavourey

Institut de Recherche en Informatique de Toulouse (IRIT)
118 route de Narbonne
31062 Toulouse
mael.lesavourey@irit.fr

RESUME : Tout travail de recherche commence par une étude bibliographique de la connaissance existante. La science ouverte permet l'accès à un océan de contenu mais la navigation y devient toujours plus difficile. Afin d'améliorer la recherche bibliographique des chercheurs en biomédecine et toxicologie nous proposons d'étudier des approches de traitement automatique du langage naturel ayant un impact sur les performances des systèmes de recherche d'information. La tâche d'indexation d'articles suivant les termes d'une ressource ontologique est particulièrement intéressante. Un exemple est la base de données MEDLINE où chaque article est indexé avec les termes de l'ontologie Medical Subject Headings (MeSH). Cependant le sens de ces termes spécifiques peut varier suivant leur contexte d'utilisation, limitant la pertinence des résultats retournés par les moteurs de recherche de publications scientifiques. Nous envisageons de faire une représentation sémantique multi-contextuelle des termes MeSH pour considérer des facettes non encore exploitées par ces moteurs de recherche.

ABSTRACT. Any research work starts with a bibliographic study of existing knowledge. Open science allows access to vast amount of content but the drawback is that it is becoming increasingly difficult to navigate through it. In order to improve the bibliographic search of biomedical and toxicology researchers, we propose to study natural language processing approaches that have an impact on the performance of information retrieval systems. The task of indexing articles according to the terms of an ontological resource is particularly interesting. An example is the MEDLINE database where each article is indexed with the terms of the Medical Subject Headings (MeSH) ontology. However, the meaning of these specific terms can vary according to their context of use, limiting the relevance of the results returned by scientific publication search engines. We consider a multi-contextual semantic representation of MeSH terms to take into account facets not yet exploited by these search engines.

MOTS-CLES : Recherche d'information, Reconnaissance d'Entités Nommées, Indexation, Biomédecine, Toxicologie.

KEYWORDS : Information Retrieval, Named Entity Recognition, Indexing, Biomedicine, Toxicology.

ENCADREMENT : Gilles Hubert, Fabien Jourdan, Yoann Pitarch.

Intégration de données et inférence au service de l'étude de la chimiodiversité du vivant

Solweig Hennechart

Laboratoire de Recherche en Sciences Végétales (LRSV) UMR5546
Institut de Recherche en Informatique de Toulouse (IRIT) UMR5505
Université Toulouse 3 Paul Sabatier
118 route de Narbonne, 31062 Toulouse
solweig.hennechart@univ-tlse3.fr

RESUME : La chimiodiversité du vivant, ensemble des composés d'origine naturelle, est encore peu connue, limitant ainsi des connaissances pharmacologiques et agroalimentaires pouvant être déterminantes. Il n'existe pas à l'heure actuelle de base de données unique regroupant tous les composés connus et leurs métadonnées, notamment l'origine biologique du composé, ou encore les molécules proches structurellement. Une telle source de données pourrait pourtant permettre d'améliorer l'efficacité de l'analyse et le temps de traitement des études métabolomiques, qui ont pour but d'identifier les molécules présentes dans un échantillon, reflet le plus fidèle aujourd'hui d'un système biologique. L'objectif de ce projet est donc de regrouper toutes les informations connues dans une base exhaustive des produits naturels et compléter les données manquantes grâce à des modèles prédictifs s'appuyant sur les données intégrées et confirmés par validation expérimentale.

ABSTRACT. The chimiodiversity of living organisms, corresponding to all natural compounds, is not still well known, thus limiting pharmacological and agri-food knowledge that can be decisive. Currently, there is no single database grouping together all the known compounds and their metadata, in particular the biological origin of the compound, or the structurally similar molecules. Such a data source could however enhance the quality of analyses and reduce the processing time of metabolomic studies, which aim to identify the molecules present in a sample, which best reflect biological systems today. The objective of this project is therefore to merge all the known information in an exhaustive database of natural products and to complete the missing information thanks to inferences using predictive models based on the integrated data and confirmed by experimental validations.

MOTS-CLES : Chimiodiversité, Métabolomique, Prédiction de données, Intégration de données, Fusion de données.

KEYWORDS: Chimiodiversity, Metabolomics, Data prediction, Data integration, Data fusion.

ENCADREMENT : Guillaume Cabanac, Guillaume Marti.

Services augmentés pour le tourisme intelligent et l'analyse des pratiques

Maxime Masson

Laboratoire LIUPPA, Collège STEE, Université de Pau et des Pays de l'Adour
Avenue de l'Université, 64000 Pau, France
maxime.masson@univ-pau.fr

RESUME : Les sources de données touristiques se sont particulièrement développées ces dernières décennies. Les réseaux sociaux, et plus généralement le "contenu générés par les utilisateurs" (CGU) sont devenus des ressources cruciales pour l'étude approfondie des pratiques et comportements touristiques. Notre projet s'inscrit dans la continuité de ce phénomène et vise à exploiter cette source de données au service des acteurs touristiques. Nous décrivons le cycle de vie de ce dernier et faisons l'hypothèse que l'étude des trajectoires multidimensionnelles (spatiale, temporelle, thématique) de touristes issues des réseaux sociaux couplée à l'application de la théorie de la proxémique comme prisme d'analyse peut contribuer à des résultats pertinents à la fois pour les professionnels du tourisme (aide à la décision) et les visiteurs eux-mêmes (recommandation). A cet effet, nous mettons également en place une méthode générique et itérative dédiée à la construction de jeux de données précis à partir des réseaux sociaux et leur évaluation.

ABSTRACT. Tourism data sources have grown a lot in the last decades. Social media, and more generally "user-generated content" (UGC), became crucial for the in-depth study of tourism practices and behaviors. Our project is a continuation of this phenomenon and aims at exploiting this data source for tourism stakeholders. We present its life cycle and hypothesize that the study of multidimensional trajectories (spatial, temporal, thematic) of tourists from social media paired with the application of the proxemics theory as an analysis standpoint can contribute to relevant results for both tourism professionals (decision support) and visitors themselves (recommendation). To this end, we also designed a generic and iterative method dedicated to the construction of focused datasets from social media and their assessment.

MOTS-CLES : Réseaux Sociaux, Tourisme, Traitement Automatique du Langage (TAL), Analyses Multidimensionnelles, Proxémique, Extraction d'Informations.

KEYWORDS : Social Media, Tourism, Natural Language Processing (NLP), Multidimensional Analysis, Proxemics, Information Extraction.

ENCADREMENT : Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, Philippe Roose, Annig Le Parc Lacayrelle.

Recommandations contextuelles fondées sur la fouille d'intentions et les ontologies

Ramona Elali

Centre de recherche informatique, Université Paris 1 Panthéon – Sorbonne
90 rue Tolbiac, Paris, France
ramona.elali@etu.univ-paris1.fr

RESUME : Différentes techniques sont utilisées pour améliorer les processus métier. La fouille d'intentions permet de fournir de nouveaux services avec une meilleure qualité car elle se concentre sur la partie intentionnelle d'un processus, plus proche du raisonnement de l'utilisateur. Les principaux objectifs de la fouille d'intentions sont d'analyser le comportement des utilisateurs d'une manière fiable et de fournir des recommandations de bonne qualité. Généralement, l'activité du l'utilisateur est affectée par plusieurs facteurs contextuels. Ces informations contextuelles peuvent être obtenues à partir de nombreuses sources de données externes. Cette thèse explore comment il est possible de combiner différents types de sources riches en contexte afin d'obtenir une meilleure qualité des recommandations. Bien que les recherches existantes se concentrent principalement sur des données d'activité standards, quelques-unes utilisent les ontologies pour combiner les diverses sources de données. Nous proposons une nouvelle approche qui (a) combine plusieurs types de sources dans des ontologies de domaine, (b) utilise ces ontologies pour construire le modèle de processus intentionnel, et (c) utilise un modèle intentionnel pour faire des recommandations contextuelles.

ABSTRACT. Different techniques are used by companies to enhance their business processes. Intention Mining can provide us with a new and higher quality of services because it focuses on the intentional part of a process that is closer to the user's thinking. Analyzing users' behavior reliably and providing quality recommendations are principal objectives of IM. Many contextual factors influence the user activity. This contextual information could be obtained from many external sources that differ from the standard activity logs. This PhD project is exploring how the combination of different types of sources with a context-rich intentional process mining can lead to a higher quality of recommendations. Although existing research mainly focuses on single activity log datasets; only a few consider ontologies to combine various sources. Therefore, we propose a novel approach that (a) combines several types of sources into domain ontologies, (b) uses those ontologies to build the intentional process model, and (c) uses an intentional model for contextual recommendations.

MOTS-CLES : Fouille d'intentions, Recommandation, Modèle de processus intentionnel, Ontologie, Evènements, Contexte.

KEYWORDS : Intention Mining, Recommendation, Intentional Process Model, Ontology, Events, Context.

ENCADREMENT : Camille Salinesi, Rebecca Déneckère, Elena Kornyshova.

Intent-Based Configuration of Information and Communication Components for Industry 4.0 Applications

Kaoutar Sadouki

Centre d'Etudes et de Recherche en Informatique et Communications (CEDRIC),
Conservatoire National des Arts et Métiers (CNAM),
292 Rue Saint Martin, 75003 Paris, France
kaoutar.sadouki@lecnam.net

RESUME : Les applications de l'industrie 4.0 (I4.0) évoluent rapidement, donnant aux organisations la capacité de faire face au défi d'une concurrence mondiale qui s'accroît rapidement dans diverses conditions économiques. Cependant, il n'est pas toujours simple de tirer parti de la transformation numérique. L'objectif de ce projet de recherche est de répondre aux difficultés de l'I4.0 et de fournir une nouvelle configuration conceptuelle pour les composants de l'I4.0 sans négliger sa rentabilité économique pour l'entreprise à travers une couche intentionnelle. L'Internet des intentions gagne déjà du terrain dans de nombreux domaines. La notion d'intentions est cruciale pour l'adoption des nouvelles technologies de l'information et de la communication (TIC), car elle permet un niveau d'intelligence plus profond tout en atteignant les objectifs des utilisateurs externes et internes des TIC.

ABSTRACT. Industry 4.0 (I4.0) applications are evolving fast, giving organizations the ability to deal with the challenge of rapidly increasing global competition under various economic conditions. However, it is not always simple to take advantage of digital transformation. The aim of this research project is to respond to I4.0 difficulties and to provide a new conceptual configuration for I4.0 components without disregarding its economic profitability for the enterprise through an intentional layer. The Internet of intents is already gaining traction in many fields. The notion of intent is crucial for new information and communication technologies (ICT) adoption, as it allows a deeper level of intelligence while achieving the goals of external and internal users of ICT.

MOTS-CLES : Industrie 4.0, Technologie d'Information et de Communication, Approches Basées sur les Intentions.

KEYWORDS : Industry 4.0, Information and Communication Technologies, Intent-Based Approaches.

ENCADREMENT : Elena Kornyshova, Eric Gressier Soudan.

Application de GNN sur des graphes non-attribués : Benchmark de performances.

Ikram Boukharouba[1, 2]

[1] *IRIT, Université Toulouse III Paul Sabatier, Toulouse/France*
[2] *Berger Levrault, Labège/France*
ikram.boukharouba@irit.fr

RESUME : Les Graph Neural Networks (GNNs) sont des méthodes d'apprentissage profond construisant un modèle lié à un graphe. Cet apprentissage rend possible la classification et la prédiction de nœuds ou d'arêtes manquantes. Cependant, ces approches produisent des résultats significatifs lorsque les features sur les nœuds sont suffisamment nombreuses, ce qui n'est pas le cas pour les problèmes de terrain. En effet, du fait de l'origine des données et de leur génération non maîtrisée, les graphes résultants contiennent peu de features. En outre certaines contraintes légales, peuvent limiter l'utilisation des données. Dans cet article, nous montrons que des graphes issus de traces d'activité utilisateur (logs) comportent également de telles limites. Notre démonstration commence par une synthèse de l'état de l'art justifiant notre utilisation des GNNs pour la tâche de classification. Par la suite, nous détaillons les expérimentations conduites sur notre corpus de données de terrain (logs issus du produit Sedit produit par Berger-Levrault). Les résultats démontrent les limites effectives des GNNs. Enfin, certaines approches pouvant proposer des solutions à ce problème sont avancées.

ABSTRACT. Graph Neural network (GNN) is a deep learning method that builds a model linked to a graph. This learning allows the classification and prediction of nodes and/or missing edges. However, these approaches produce noteworthy results only when the features on the nodes are enough numerous. In real-world, especially industrial problems, these features may not always be as numerous which is not the case for "real-world problems". Indeed, due to the origin of the data and their uncontrolled generation process, the resulting graphs may contain few features or even none. In addition, certain legal constraints may limit the use of data. In this article, we show that graphs derived from traces of user activity (logs) also have such limits. Our demonstration begins with a summary of the state of the art justifying the use of GNNs for the classification task in our use case. Thereafter, we detail the experiments conducted on our corpus of field data (logs from the Sedit product produced by Berger-Levrault). The results demonstrate the effective limitations of GNNs. Finally, some approaches that can offer solutions to this problem are put forward.

MOTS-CLES : Traces d'activité utilisateurs, Logs, Application industrielle, Graphe de navigation, Graph Neural Network (GNN), Graphe non-attribué.

KEYWORDS : Traces of user activity, Logs, Industrial application, Navigation graph, Graph Neural Network (GNN), Non-attributed graph.

ENCADREMENT : Florence Sèdes, Christophe Bortolaso, Florent Mouysset.

Cybersécurité à l'échelle intersystème d'information avec prise en compte du facteur humain

Olivier de Casanove

IRIT, Université Toulouse III – Paul Sabatier
118 route Narbonne 31062 Toulouse CEDEX 9
olivier.decasanovenom@irit.fr

RESUME : Le NIST (National Institute of Standards and Technology) reconnait quatre étapes dans la gestion des incidents de sécurité : prévention, détection, remédiation et apprentissage. Actuellement, les technologies de détection utilisées ne fonctionnent qu'à l'échelle d'un seul système d'information. Se cantonner à une telle détection c'est négliger l'interconnexion de la sécurité des systèmes d'information entre eux. Nous proposons dans ce papier une solution de détection de cyberattaques, complémentaire à l'état de l'art, en travaillant à l'échelle intersystème d'information. Nous verrons également en quoi une telle détection permet d'améliorer les effets de la prévention pour la sécurité informatique.

ABSTRACT. The NIST (National Institute of Standards and Technology) recognise four steps in security incident management: prevention, detection, recovering and assessment. Detection technologies currently used only function at the information system scale only. This type of detection neglects the interconnection of the information systems security. We propose in this paper a new solution to detect security incidents at inter-information system scale. This approach is complementary to what already exists in the literature. We will also see how to value such a detection.

MOTS-CLES : Cybersécurité, Sécurité des Systèmes d'Information, Détection d'Evènements, Prévention.

KEYWORDS : Cybersecurity, Information System Security, Event Detection, Prevention.

ENCADREMENT : Florence Sèdes.