Actes de la 6e conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle

Amal El Fallah-Seghrouchni, Stephan Brunessaux

AfIA

Association française
pour l'Intelligence Artificielle

APIA

Conférence Nationale
sur les
Applications Pratiques de l'Intelligence Artificielle

PFIA 2020

Table des matières

Éditorial

L'Intelligence Artificielle poursuit son essor sans précédent. Les recherches menées ces dernières années ont abouti à des résultats spectaculaires dans certains domaines et des résultats très prometteurs dans d'autres.

Aujourd'hui, l'IA se trouve au cœur de nombreuses applications très performantes qui révolutionnent notre vie quotidienne.

Plus que jamais, l'objectif de cette 5ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA 2020) était de faire connaitre les applications concrètes de l'IA qui couronnent de succès l'opérationnalisation de l'IA et des travaux de recherche.

Ces applications peuvent mettre en œuvre une ou plusieurs facettes de l'IA comme :

— Agents autonomes et systèmes multi-agents : simulation, planification, décision individuelle ou collective ;
— Applications de l'Intelligence Artificielle, méthodologie, évaluation ;
— Apprentissage numérique et symbolique ;
— Environnements informatiques d'apprentissage humain ;
— Évolution artificielle, systèmes situés, systèmes adaptatifs ;
— Fouille de données, bases de données avancées, web sémantique ;
— Ingénierie et partage des connaissances ;
— Intelligence collective, intelligence sociale, réseaux sociaux ;
— Interfaces Intelligentes, interaction homme-machine, intelligence ambiante ;
— Langages de programmation pour l'IA, programmation logique, programmation orientée agent ;
— Logique formelle et outils pour l'Intelligence Artificielle, sémantique ;
— Passage à l'échelle, organisation de systèmes, émergence ;
— Plates-formes et environnements de développement en IA ;
— Raisonnement à base de modèles, raisonnement à partir de cas ;
— Raisonnement probabiliste et incertain, logique floue ;
— Raisonnement spatial et temporel, environnements physiques ;
— Recherche opérationnelle, programmation par contraintes, ordonnancement ;
— Représentation des connaissances, extraction et gestion des connaissances ;
— Réseaux de neurones artificiels, approches neuromimétiques ;
— Robotique, vision par ordinateur, capteurs intelligents, systèmes physiques ;
— Sciences cognitives et Intelligence Artificielle, cognition, informatique affective ;
— Systèmes à base de règles, aide à la décision ;
— Traitement automatique du langage, terminologie, langage naturel contrôlé, explication ;
— Traitement du signal et de l'image, traitement de la parole ;
— Web intelligence, internet du futur, protection de la vie privée.

Quelles sont les applications qui s'appuient sur de l'IA ou qui nécessitent de l'IA ? Du système de surveillance militaire au système d'aide au diagnostic médical, du climatiseur à l'assistant personnel, du système d'aide à la conduite à l'analyse de données massives, etc., les applications sont nombreuses. Qu'elles soient industrielles, sociétales, économiques, politiques, environnementales, artistiques ou autres, cette conférence est l'occasion de présenter des applications concrètes et des travaux dont l'objet d'étude n'est pas uniquement des cas de laboratoire. Les contributions peuvent illustrer des domaines très divers automobile, robotique, militaire, logistique, télécommunication, finance, domotique, agronomie, réseaux sociaux, risque, grandes masses de données, santé, aide à la personne, jeux vidéo, réalité virtuelle/mixte, musées, EIAH, serious games, récit interactif, et bien d'autres encore.

L'objectif est également de comprendre comment ces applications concrètes font remonter des verrous scientifiques que l'IA doit résoudre pour démocratiser encore plus son utilisation. Par exemple, comment la prise en compte des réalités concrètes vues dans leur globalité amènent la prise en compte par l'IA de problèmes complexes décuplés qu'il n'est pas possible de rendre compte dans des approches réductionnistes de problèmes de laboratoire ? L'IA est-elle suffisamment expressive et intelligible pour être utilisée ? Est-elle fiable et robuste ? Est-elle capable de passer à l'échelle ? Quels sont les problèmes éthiques liés à son utilisation ? Comment garantir l'interprétabilité ou l'explicabilité de l'IA ?

Il ne s'agit pas d'opposer recherche fondamentale/appliquée mais laboratoire/terrain et comment les applications concrètes apportent des problématiques fondamentales ou encore comment des recherches fondamentales apportent des solutions à des problèmes complexes difficiles à résoudre sans IA.

APIA 2020 a favorisé l'échange entre chercheurs académiques et industriels pour qu'ils puissent partager leurs expériences, débattre des différents verrous qu'ils rencontrent, présenter les méthodes qu'ils mettent en œuvre pour enrichir le potentiel applicatif des modèles et outils de l'IA et les besoins naissants, en mettant en valeur l'IA de ces applications.

Cette édition 2020 a été organisée en virtuel du fait de la pandémie de COVID-19.

Amal ELFALLAH-SEGHROUCHNI, Stéphan BRUNESSAUX

Comité de programme

Président

— Amal ELFALLAH-SEGHROUCHNI, Sorbonne University
— Stéphan BRUNESSAUX, Airbus

Membres

— Ghislain ATEMEZING, Mondeca
— Jérôme AZÉ, LIRMM-UM-CNRS
— Assia Belbachir, IPSA
— Alain BERGER, Ardans
— Stéphan BRUNESSAUX, Airbus
— Amal ELFALLAH-SEGHROUCHNI, Sorbonne University
— Christophe GUETTIER – Safran
— Céline HUDELOT, Ecole Centrale Paris
— Christophe LABREUCHE, Thales
— Arnaud LALLOUET, Huawei
— Christine LARGOUËT, Irisa /Agrocampus Ouest
— Vincent LEMAIRE, Orange Labs
— Dominique LENNE, Heudiasyc – Université de Technologie de Compiègne
— Sylvain MAHE, EDF R&D
— Philippe MATHIEU, University of Lille 1
— Nada MATTA, University of Technology of Troyes
— Juliette MATTIOLI, Thales
— Philippe MORIGNOT, Aspertise
— Selmin NURCAN, Université Paris 1 Panthéon – Sorbonne
— Brigitte TROUSSE, Université Côte d'Azur, INRIA Sophia Antipolis – Méditerranée

Prévisions geographiques du nombre d'interventions des pompiers respectant la confidentialité différentielle locale

H. H. Arcolezi[1], J-F. Couchot[1], S. Cerna[1], C. Guyeux[1], G. Royer[2], B. Al Bouna[3], X. Xiao[4]

[1] Femto-ST Institute, Univ. Bourgogne Franche-Comté, UBFC, CNRS, Belfort France
[2] SDIS 25 - Service Départemental d'Incendie et de Secours du Doubs, France
[3] Lab., Antonine University, Hadath-Baabda, Lebanon
[4] School of Computing, National University of Singapore, Singapore

Résumé

Les travaux présentés ici visent à prédire le nombre d'interventions des pompiers par Communauté d'Agglomération tout en respectant la vie privée des utilisateurs. Une approche basée sur la confidentialité différentielle locale a été développée pour anonymiser les données de localisation, puis une approche d'apprentissage supervisé a été mise en place pour faire les prédictions. Les expériences réalisées montrent que la méthode complète d'anonymisation-prédiction est très précise : l'introduction de bruit n'affecte pas la qualité des prédictions, et ces dernières reflètent avec une bonne précision ce qui s'est passé dans la réalité.

Mots-clés

confidentialité différentielle locale, RAPPOR, lieu d'intervention des pompiers, prévision multi-cibles, XGBoost.

Abstract

The work presented here aims to predict the number of firefighters' interventions by region while respecting the privacy of users. An approach based on local differential privacy was used to anonymize the location data, then a supervised learning approach was used to make the predictions. The experiments carried out show that the complete anonymization-prediction method is very precise : the introduction of noise does not affect the quality of the predictions, and the predictions reflect with good accuracy what happened in reality.

Keywords

local differential privacy, RAPPOR, firemen intervention location, multi-target forecasting, XGBoost.

1 Introduction

Le transport médical d'urgence comprend les différents services utiles au transport des personnes blessées depuis leur domicile ou depuis le lieu de l'incident jusqu'à la structure médicale la plus apte à prendre soin du patient. La manière dont ce transport médical d'urgence est organisé dépend du pays considéré, de son histoire et des choix politiques qui ont été faits dans le passé. En France, par exemple, les Sapeurs Pompiers ne sont pas seulement responsables de l'extinction des incendies, mais ils doivent aussi prendre en charge une partie des transports médicaux d'urgence, et cette charge représente plus de 80% de leur activité.

Cette structuration a bien fonctionné dans le passé. Cependant, depuis un certain temps, en France comme dans divers autres pays, nous sommes confrontés à une crise majeure du transport médical d'urgence, pour diverses raisons (par exemple, le vieillissement de la population, la crise financière). Ces éléments et d'autres encore conduisent donc à une crise du transport sanitaire d'urgence dans diverses régions du monde.

Une des solutions envisagées pour soulager la pression sur ces transporteurs est d'optimiser l'utilisation de leurs ressources, afin de renforcer les équipes pendant les périodes de pointe, tout en les réduisant pendant les périodes creuses. Mais la situation de crise est telle qu'il est maintenant nécessaire d'aller beaucoup plus loin dans ces optimisations, ce qui nécessite une vision relativement claire des besoins à court, moyen et long terme.

Ainsi certains auteurs ont récemment cherché à exploiter les techniques d'intelligence artificielle [3, 2, 1, 9] afin de prévoir la demande future en matière de transport médical d'urgence. Cependant, pour être supervisé, l'apprentissage automatique nécessite d'avoir accès au flux d'intervention des opérateurs dont nous essayons de prédire la charge (ambulanciers privés, pompiers, etc.). Ces derniers n'ont généralement ni les ressources humaines et matérielles ni la compétence pour déployer des solutions basées sur l'intelligence artificielle et sont donc obligés de transmettre ces données à un tiers de confiance ayant cette compétence.

S'agissant de données sensibles liées à des accidents, au sauvetage de personnes, à des décès éventuels,..., comment ces données doivent-elles être publiées après avoir été anonymisées ? En France par exemple, deux bases de données récentes de tels flux ont été publiées sur `data.gouv.fr`. La première concerne les interventions 2007-2017 du Service Départemental d'Incendies et de Secours de Saône-et-Loire (SDIS 71) [1], contenant le nombre d'in-

1. `https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers-od71/`

terventions par type et par commune, tandis que la seconde concerne les mêmes types de données [2] du SDIS 91 (département de l'Essonne) pour la période 2010-2018. Dans chaque cas, l'anonymisation a été effectuée par agrégation : mensuelle pour la première série de données, et hebdomadaire pour la seconde.

Cependant, la manière dont ces données ont été diffusées pose deux problèmes : l'anonymisation obtenue est à la fois trop forte et trop faible. Trop forte, tout d'abord, parce que la réalisation d'une agrégation par mois entraîne une perte complète de l'utilité de celle-ci : elle résume les interventions à 12 points par an, pour lesquels seule une simple régression linéaire reste possible. Ensuite, trop faible, car cette agrégation par mois, ou par semaine, a été faite de manière aveugle et généralisée. Par exemple, dans le cas de l'agrégation mensuelle, il existe plus de 600 situations où il n'y a eu qu'une seule intervention dans une commune au cours d'un mois donné : à ce niveau, le simple 2-anonymat [13] n'est plus satisfait, et la fuite d'informations est évidente. Ces fuites d'informations sont également nombreuses dans le cas de données hebdomadaires, et l'anonymat a échoué pour les deux séries de données.

L'objectif de cet article est donc de montrer qu'il est possible de traiter de tels flux de manière à ce que l'anonymat soit garanti d'une part et que des prévisions correctes puissent être faites par apprentissage automatique sur ces données d'autre part. Ceci est vrai même si les données considérées ont des densités spatiales très variables. Plus spécifiquement, dans cet article, notre objectif est d'appliquer une version de la Confidentialité Différentielle Locale (CDL), afin de transformer les données réelles pour qu'elles soient à la fois correctement anonymisées, et utiles pour l'apprentissage automatique. Ces données traitées sont ensuite exploitées à des fins d'apprentissage et de prédiction : une tâche de prédiction du nombre d'interventions par communauté d'agglomération (CA), avec des données brutes et anonymes, est alors proposée. Les approches envisagées comprennent l'utilisation d'une mémoire longue à court terme pour le nombre total d'interventions [1] ; un perceptron multicouche pour le nombre total d'interventions à nouveau [9] ; et enfin l'utilisation de XGboost sur une tranche de temps de 3h, un modèle pour deux CA importantes, et des modèles par motif [2].

Le reste de cet article est organisé comme suit. La section suivante présente les données qui seront traitées dans cet article. La section 3 est consacrée au contexte théorique lié à la confidentialité différentielle et à sa version locale. Son application est expliquée à la section 4 et évaluée expérimentalement à la section 5. La capacité à effectuer des prédictions précises d'apprentissage machine sur la base de ces données est enfin détaillée dans la section 6. Cet article se termine par une section de conclusion, dans laquelle la contribution est résumée et des perspectives sont énoncées.

2. https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers/

2 Présentation des données

La base de données à notre disposition a été fournie par le service d'incendie et de secours, SDIS 25, du département du Doubs (France). Ce fichier contient des informations sur 382046 interventions auxquelles les pompiers ont participé de 2006 à 2018 au sein de ce département. Chaque intervention est enregistrée dans un fichier sous la forme d'une ligne et les principaux attributs de ce fichier sont indiqués dans le tableau 1 avec des informations artificielles et décrits comme suit :

ID	SDate	Caserne	Ville	Lieu
8	2008/08/08 08 :08	Besançon Est	Besançon	(47.2380, 6.0243)

TABLE 1 – Principaux attributs des données relatives aux opérations des pompiers

- *ID* est l'identifiant d'intervention, qui est utilisé dans les dossiers supplémentaires ;
- *SDate* est la date de début de l'intervention ;
- *Caserne* est le nom de la caserne de pompiers qui a participé à l'intervention ;
- *Ville* est le nom de la municipalité où l'opération a eu lieu ;
- *Lieu* donne le lieu précis (latitude, longitude) de l'intervention.

En analysant les données, nous avons constaté une forte augmentation du nombre d'interventions au fil des ans. Autrement dit, en 10 ans, le nombre d'interventions a doublé de 17333 en 2006 à 34436 en 2016 et a continué d'augmenter jusqu'à 40510 en 2018.

3 Contexte théorique sur la confidentialité différentielle locale

Soit $\mathcal{A}$ un algorithme utilisé pour publier des informations issues d'une base de données privée. La confidentialité différentielle (CD) [6] est une contrainte sur $\mathcal{A}$ qui limite la divulgation d'informations privées des enregistrements se trouvant dans la base de données. Intuitivement, $\mathcal{A}$ est différentiellement privé si un observateur qui en voit un extrait ne peut pas dire si les informations d'un individu particulier ont été utilisées dans le calcul.
Soit ϵ un nombre réel positif qui correspond intuitivement au niveau de fuite. Plus la valeur de cette variable est élevée, plus la fuite d'informations est importante. Soit im $(\mathcal{A})$ représente l'image de $\mathcal{A}$, *i.e.*, l'ensemble de tous les résultats possibles par $\mathcal{A}$. On dit que l'algorithme $\mathcal{A}$ fournit ϵ -confidentialité différentielle si, pour tous les ensembles de données D_1 et D_2 qui diffèrent sur les données d'une personne, et pour tous les sous-ensembles R de im $(\mathcal{A})$, nous avons

$$\Pr[\mathcal{A}(D_1) \in R] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in R], \qquad (1)$$

avec $\Pr[\mathcal{A}(D_2) \in R]$ la probabilité qu'un ensemble de données D_2 puisse être anonymisé en un élément de R selon $\mathcal{A}$ et ϵ le montant de la fuite. Cette équation donne une limite supérieure de la probabilité qu'un ensemble de

données D_1 puisse être anonymisé en un élément de R, ce qui constitue donc une fuite d'informations. Le lecteur intéressé pourra lire [5, 7].
Toutefois, cette approche exige que l'ensemble des données soit complet et stocké de manière sure. L'anonymisation ne se fait pas avant. La Confidentialité Différentielle Locale (CDL) [8] est une solution à ce problème. Dans cette approche, les données sont aseptisées par l'utilisateur de manière probabiliste avant d'être envoyées au collecteur. Un exemple simple consiste à demander à une personne de répondre à la question "Habitez-vous à Belfort ?", selon la procédure suivante :
Lancez une pièce de monnaie.

— Si c'est "pile", lancez à nouveau la pièce (en ignorant le résultat) et répondez honnêtement à la question.
— Si c'est "face", alors relancez la pièce et répondez "Oui" si face et "Non" sinon.

Soit t_y la proportion de réponses "Oui" vraies et c_y la proportion de réponses "Oui" observées. L'équation suivante donne une relation estimée entre ces deux variables

$$\frac{1}{2}t_y + \frac{1}{4} \approx c_y.$$

Plus le nombre d'expériences est élevé, plus la proportion de réponses "Oui" aléatoires sera proche de 1/4 et plus le nombre de fois où la vérité est dite sera proche, plus l'estimation sera précise. Dans ce cas, t_y peut être estimé par

$$t_y \approx 2.c_y - \frac{1}{2}.$$

L'algorithme $\mathcal{A}$ est censé fournir ϵ-CDL si, pour toutes les paires de données privées possibles de l'utilisateur v_1 et v_2 et tous les sous-ensembles R de im$\mathcal{A}$:

$$\Pr[\mathcal{A}(v_1) \in R] \leq e^{\epsilon} \times \Pr[\mathcal{A}(v_2) \in R]. \quad (2)$$

4 Collecte de données sur la localisation des interventions des pompiers dans le respect de la vie privée

La première question que l'on peut se poser est de savoir si une intervention est un attribut sensible. La réponse est oui, car les pompiers n'auraient pas été appelés si la situation n'avait pas été suffisamment grave. Prenons par exemple le scénario où une personne, qui habite dans un petit village, a contracté une maladie très particulière. Si l'on sait que, pendant cette période, une intervention a eu lieu dans cette ville, alors qu'elle est normalement rare, il y a une forte probabilité que les pompiers soient intervenus pour cette personne.
Par conséquent, le but de cette tâche est de mettre en œuvre un mécanisme de préservation de la vie privée pour la localisation de l'intervention des pompiers en utilisant le concept de confidentialité différentielle locale décrit précédemment. Ensuite, pour évaluer le compromis vie privée/utilité, le défi consiste à estimer approximativement le nombre d'interventions des pompiers sur les lieux analysés en utilisant les données anonymes.
Les deux sous-sections suivantes présentent l'approche de préservation de la vie privée basée sur la CDL appliquée à la collecte de données de localisations des interventions des pompiers et la méthode statistique pour estimer le nombre d'interventions par localisation.

4.1 Approche de la collecte de données basée sur la CDL

La figure 1 illustre un aperçu de l'approche et est résumée dans ce qui suit.

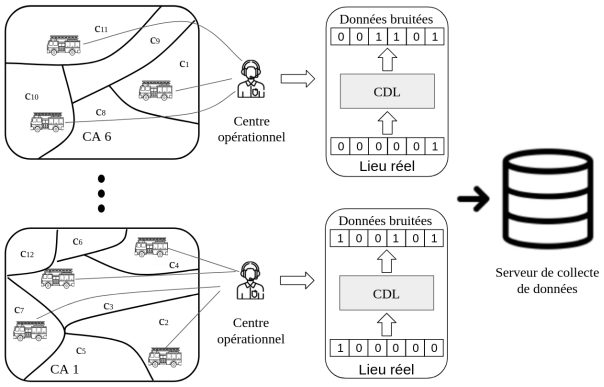


FIGURE 1 – Un aperçu de l'approche appliquée pour collecter les données de localisation des interventions des pompiers en préservant la vie privée.

Dans l'approche proposée, la première étape pour garantir la confidentialité du lieu de chaque intervention consiste à regrouper les villes qui ont fait l'objet de chaque intervention au niveau d'une communauté d'agglomération pour obtenir des événements suffisamment représentatifs en nombre. Par exemple, on peut remarquer dans la figure 1 qu'un ensemble de petites villes $C = \{c_1,c_2,...,c_{12},...,c_m\}$ sont regroupées en $n = 6$ CA.
Dans ce contexte, en utilisant les données dont nous disposons, 608 villes où des interventions ont eu lieu dans le département du Doubs sont agrégées en $n = 17$ CA en utilisant l'ensemble de données publiques disponibles [3]. Les 17 CA sont : (1) CA du Grand Besançon, (2) CA Pays de Montbéliard Agglomération, (3) CC Altitude 800, (4) CC de Montbenoit, (5) CC des Deux Vallées Vertes, (6) CC des Lacs et Montagnes du Haut-Doubs, (7) CC des Portes du Haut-Doubs, (8) CC du Doubs Baumois, (9) CC du Grand Pontarlier, (10) CC du Pays d'Héricourt, (11) CC du Pays de Maîche, (12) CC du Pays de Sancey-Belleherbe, (13) CC du Plateau de Frasne et du Val Rasne et du Val de Drugeon (CFD), (14) CC du Plateau de Russey, (15) CC du Val de Morteau, (16) CC du Val Marnaysien, (17) CC Loue-Lison.
Deuxièmement, pour améliorer le niveau de confidentialité de chaque intervention, le mécanisme "Basic One-time

3. https://www.collectivites-locales.gouv.fr/liste-et-composition-2018/

RAPPOR" introduit par [8] est appliqué. Cet algorithme est une simplification du mécanisme RAPPOR, qui utilise des filtres de Bloom et des fonctions de hachage pour cartographier les rapports envoyés par les utilisateurs et il comporte deux niveaux de réponses aléatoires, à savoir les réponses permanentes et les réponses instantanées.
Cependant, dans le "Basic One-time RAPPOR", il n'est appliqué qu'une seule étape de réponse aléatoire en utilisant une cartographie déterministe des 17 CA en tableau de bits. La motivation pour utiliser cet algorithme simple est la suivante : les communautés d'agglomération sont connues a priori en permettant la cartographie déterministe plutôt qu'en utilisant des fonctions de hachage et des filtres de Bloom ;
Une application technique de cet algorithme dans notre étude de cas est décrite ci-dessous :

1. **Vrai signal de localisation.** Soit $R = \{r_1,r_2,...,r_n\}$ un ensemble de n CA, où chaque indice représente un identifiant de CA unique. Ainsi, un tableau de n bits, B (qui indique le lieu d'intervention actuel) est défini comme

$$B_k = \begin{cases} 1 & \text{si } k = i \text{ et} \\ 0 & \text{sinon,} \end{cases} \tag{3}$$

   où B_k représente la valeur du $k^{\text{ème}}$ bit dans B, $k \in \{1,\ldots,n\}$. Autrement dit, le bit correspondant à l'identifiant des CA est mis à 1, tandis que les autres sont mis à 0.

2. **Réponse aléatoire permanente.** Ensuite, chaque bit dans B (de l'étape précédente) est perturbé en appliquant le concept de réponse aléatoire comme suit :

$$U_k = \begin{cases} 1 & \text{avec probabilité } \frac{1}{2}f, \\ 0 & \text{avec probabilité } \frac{1}{2}f \text{ et} \\ B_k & \text{avec probabilité } 1-f, \end{cases} \tag{4}$$

   où f est une valeur de probabilité entre 0 et 1, qui contrôle le niveau de garantie de confidentialité différentielle de ϵ.

3. **Rapport final.** La réponse aléatoire permanente B est transmise au serveur du collecteur de données.

Il a été démontré [8] que le niveau différentiel local de confidentialité ϵ est au pire ϵ_∞ défini comme

$$\epsilon_\infty = 2\ln\left(\frac{1-\frac{1}{2}f}{\frac{1}{2}f}\right). \tag{5}$$

4.2 Estimation du nombre d'interventions par lieu

En considérant un moment précis que l'on souhaite analyser, l'objectif est d'estimer approximativement le nombre d'interventions par lieu associé à la $i^{\text{ème}}$ CA, r_i. Soit donc *set*(U) un ensemble de réponses permanentes randomisées et *set*(B) l'ensemble correspondant de tableaux de bits de localisation d'origine. En outre, supposons que $|$*set*$(U)|$ et $|$*set*$(B)|$ indiquent le nombre d'éléments dans chaque ensemble respectif. Naturellement, $|$*set*$(U)| = |$*set*$(B)|$.
Par conséquent, le nombre estimé d'interventions *NBint*$_{est}$ par emplacement de CA r_i pour $i \in [1,n]$ est acquis par une approche basée sur les statistiques (SB) comme suit [8] :

$$\textit{NBint}_{est}(r_i) = \frac{1}{1-f} \cdot \left(N_i - \frac{f \cdot N_{total}}{2}\right) \tag{6}$$

où N_{total} est le nombre de réponses permanentes randomisées $|$*set*$(U)|$ et N_i est le nombre total de réponses permanentes randomisées dont le $i^{\text{ème}}$ bit est fixé à 1. N_i est calculé en utilisant les données de *set*(U). Il convient de noter que l'équation (6) peut estimer des nombres négatifs, d'où l'utilisation de la fonction $max(0,$*NBint*$_{est})$. Dans la littérature, d'autres techniques comme l'algorithme de maximisation des attentes [4] sont également utilisées pour estimer les densités.
Pour évaluer le résultat SB, l'estimation de la densité d'un emplacement de la $i^{\text{ème}}$ CA associée à r_i est calculée comme suit [11] :

$$\textit{Density}_{est}(r_i) = \frac{\textit{NBint}_{est}(r_i)}{\sum_{y=1}^{n} \textit{NBint}_{est}(r_y)} \tag{7}$$

où n est le numéro de la CA, et, par conséquent, la métrique du taux d'erreur (*ER*) est définie comme

$$\textit{ER} = \frac{1}{n}\sum_{i=1}^{n} |\textit{Density}_{actual}(r_i) - \textit{Density}_{est}(r_i)| \tag{8}$$

où *Density*$_{actual}(r_i)$ et *Density*$_{est}(r_i)$ correspondent respectivement à la densité réelle et à la densité estimée de la CA associée à la $i^{\text{ème}}$ localisation. Plutôt que de calculer la racine carrée de l'erreur quadratique moyenne sur le nombre estimé et réel d'interventions, le taux d'erreur est calculé sur la valeur de densité motivée par des valeurs normalisées comprises entre 0 et 1.

5 Expériences d'anonymisation

Pour évaluer l'approche consistant à anonymiser le lieu d'intervention des pompiers, plusieurs simulations sont réalisées avec différentes valeurs de f, ce qui détermine le niveau de ϵ_∞-différence de vie privée. Ainsi, en utilisant l'approche statistique (Equation (6)), l'objectif est d'estimer le nombre d'interventions par CA en considérant différents scénarios de temps. Ces expériences permettront d'évaluer la relation entre *ER* et la taille des données (période d'analyse) en fonction de f afin de trouver le meilleur compromis vie privée-utilité pour différentes applications.
Les scénarios de temps sont décrits ci-dessous. Chaque scénario permet aux pompiers de disposer d'une base de données anonyme où des entreprises tierces ou le département des ressources humaines lui-même pourraient acquérir des statistiques de grande utilité. Dans les expériences, f variera dans $[0{,}1; 0{,}2; \ldots; 0{,}8; 0{,}9]$, ce qui garantit ϵ_∞-une différence de confidentialité dans $[5{,}89; 4{,}39; \ldots; 0{,}81; 0{,}4]$ respectivement.

La première analyse co ncerne des données sur un an (13 points de données), ce qui permet au début d'une année aux pompiers de mieux répartir leur budget autour de leurs centres en fonction du nombre d'interventions par CA. Ensuite, un scénario d'un mois (156 points de données) est envisagé. Les pompiers peuvent ainsi disposer de statistiques pour réorganiser les budgets et le personnel chaque mois. Enfin, un scénario d'un jour (4748 points de données) est pris en considération de sorte que des tâches d'apprentissage machine puissent être appliquées dans cette quantité de données.

5.1 Resultats

Le but de cette tâche est d'anonymiser les données de localisation des interventions des pompiers avec différents niveaux de confidentialité (mesuré à l'aide de ϵ_∞) et trois scénarios de temps et d'évaluer les relations entre *ER* et la taille des données.

Pour mieux illustrer les résultats, la figure 2 montre la relation entre *ER* et ϵ_∞ pour chaque année (2006-2018), avec un zoom pour les 8 derniers mois de 2018, et avec un zoom pour les 8 derniers jours de décembre de 2018, respectivement. De plus, la figure 3 illustre les statistiques acquises sur le nombre d'interventions pour l'année 2013, le premier mois de 2017, et un jour précis du mois de janvier 2016 avec deux valeurs pour ϵ_∞ égal à 4,394 et 2,773 engendrées respectivement à partir de f égal à 0,2 et 0,4 en suivant l'équation (5).

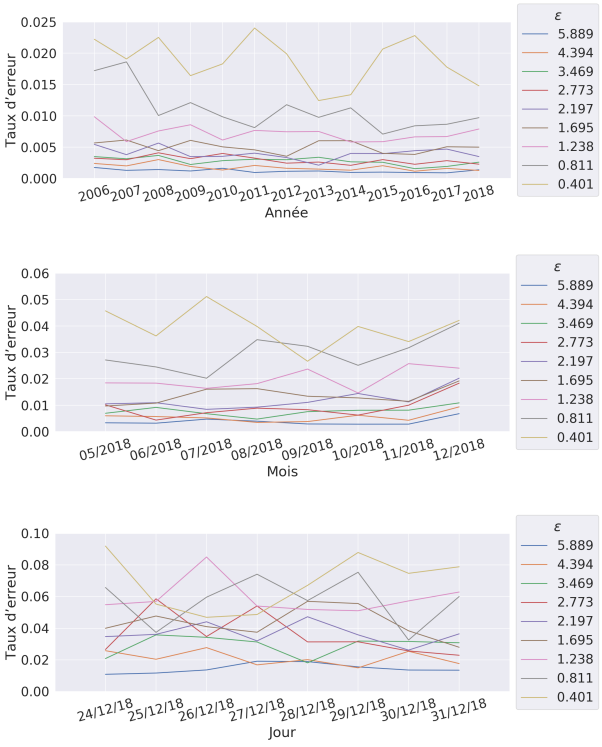


FIGURE 2 – Relation entre le taux d'erreur *ER* et la durée de l'étude variant ϵ_∞.

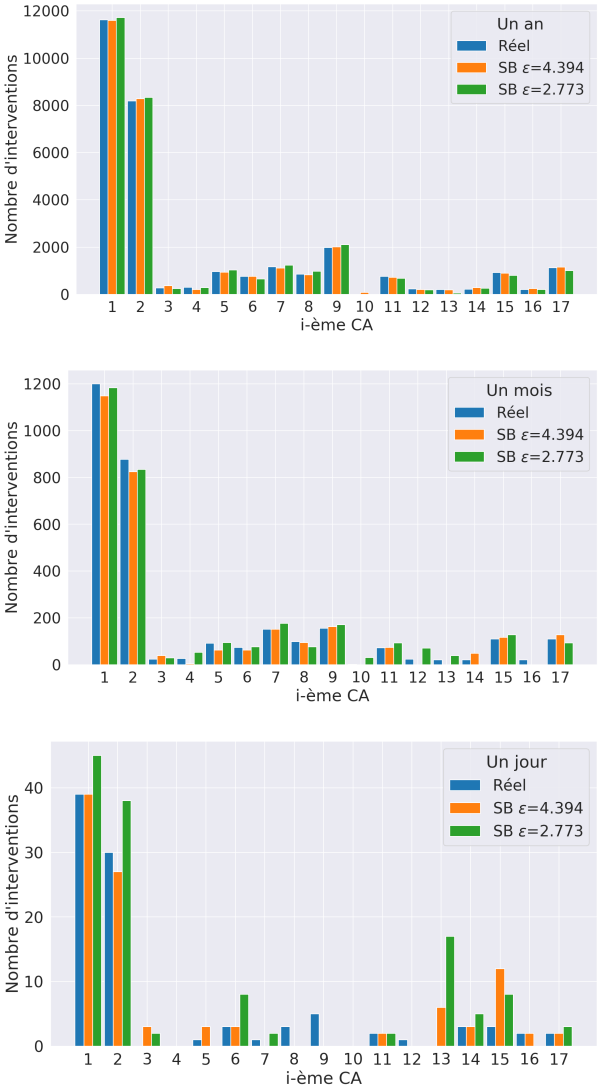


FIGURE 3 – Analyse entre le nombre réel et le nombre estimé d'interventions par CA.

5.2 Discussions

Comme on peut le remarquer dans les figures 2 et 3, le mécanisme basé sur la CDL peut être bien appliqué à la collecte de la localisation des interventions des pompiers dans le but de déduire le nombre d'interventions par CA. Comme la CDL garantit la confidentialité des personnes en perturbant les données avant de les envoyer au collecteur de données, dans ce cas, le pompier chargé de signaler les interventions appliquera la perturbation à la localisation des interventions avant de l'envoyer au serveur de données (comme l'illustre la figure 1).

Il est à noter que plus le nombre de points de données diminue, plus le *ER* augmente. Comme le souligne également la littérature, la CDL nécessite une grande quantité de données pour garantir un bon équilibre du bruit, car le bruit est ajouté à chaque point de données dans cette approche. Alors que pour une analyse sur un an, le nombre

d'interventions est d'au moins 17333 pour l'année 2006, la moyenne par jour n'est que de 47 pour la même année. Pour cette raison, l'utilité des données diminue si l'on veut des scénarios plus précis, comme les cas d'un mois et d'un jour présentés dans cet article.

De plus, la relation entre les *ER* et les garanties de confidentialité est naturelle : plus ϵ_∞ est petit, plus la garantie de confidentialité différentielle est renforcée, plus le bruit ajouté aux données est important et plus l'utilité de celles-ci est faible. Prenons par exemple la figure 3, particulièrement celle détaillant les prédictions journalières. On constate dans celle-ci que les prévisions issues d'une anonymisation avec ϵ_∞ petit (*i.e.* en vert sur la courbe) sont plus éloignées de la réalité que celles obtenues à la suite d'une anonymisation avec un ϵ_∞ plus grand.

Toutefois, comme nous l'avons déjà dit, la taille des données a une grande influence à ce stade.

Supposons que nous ayons une densité suivant une distribution normale $\textit{Density}_{actual}(r_i) = 1/n$. Un $ER = 0{,}01$ représente en moyenne 10% d'erreur par CA, r_i. La figure 2 montre la relation entre *ER* et la taille des données : pour le scénario d'un an (resp. un mois, une jour) avec le niveau maximum de garanties de confidentialité, le taux d'erreur *ER* oscille entre $0{,}001$ et $0{,}025$ (resp. entre $0{,}004$ et $0{,}05$, entre $0{,}01$ et $0{,}09$).

La Figure 3 montre les statistiques obtenues sur le nombre d'interventions pour chaque période : des petites erreurs sont obtenues pour le cas d'un an (ER normalement inférieur à $0{,}01$) ; des erreurs raisonnable pour le scénario d'un mois (la moitié de ER est inférieure à $0{,}01$), et des erreurs considérables pour le scénario d'un jour (tous $ER \geq 0{,}01$). Par conséquent, comme le souligne également la littérature, le choix de ϵ dépend de plusieurs facteurs (taille des données, domaine d'application) et il faut trouver un équilibre approprié entre confidentialité des données et utilité de celles-ci. Par exemple, dans ce cas, où les 608 villes ont été agrégées en $n = 17$ CA, certains jours avec peu de données (moins de ~ 100 par ex.), la confidentialité pourrait être légèrement diminuée afin d'en préserver l'utilité. Dans la littérature, les valeurs d'ϵ se situent classiquement dans la fourchette $[0{,}01; 10]$ [10].

6 Prévision des interventions des pompiers par CA

Le but de cette tâche est d'implémenter un algorithme d'apprentissage machine efficace, à savoir le renforcement du gradient extrême (XGBoost), pour prévoir le nombre d'interventions par jour des 17 CA du Doubs-France. L'objectif principal est d'utiliser des fichiers anonymes pour construire des modèles afin d'évaluer l'utilité des données avec différents niveaux de confidentialité ϵ_∞ différents par rapport à l'original.

6.1 Préparation des données

Trois sources initiales ont été prises en compte :

— Une liste de lieux géométriques pour chaque ville appartenant au département du Doubs, obtenue auprès du SDIS 25 ;
— Une liste de villes regroupées en 17 CA pour le département du Doubs ;
— Une liste des interventions de 2006 à 2018 du SDIS 25. Elle a été organisée en un ensemble de données, où chaque ligne, représentant une journée, comprend le nombre d'interventions sur une zone géographique donnée.

De la première source, on a extrait les polygones qui décrivent chaque ville. Ensuite, ils ont été regroupés par CA en tenant compte de la deuxième source. La troisième source comporte 5 versions : les données réelles et les données anonymes (4 au total avec des f variant en $[0{,}2; 0{,}4; 0{,}6; 0{,}8]$, qui garantissent ϵ_∞ en $[4{,}394; 2{,}773; 1{,}695; 0{,}811]$). Pour les deux types d'ensembles de données, on a ajouté des informations temporelles telles que l'année, le mois, le jour, le jour de la semaine, le jour de l'année, des valeurs (1 pour "Oui", 0 pour "Non") pour indiquer les années bissextiles, le premier ou le dernier jour du mois, et le premier ou le dernier jour de l'année comme attributs.

Dans la plupart des cas, les données anonymisées décrivent un nombre d'interventions plus élevé que le nombre réel. Afin de préserver l'intégrité des données, un filtre est appliqué à chaque ensemble anonymisé. Par exemple, un ensemble de données anonymisées spécifique est pris ; pour chaque ville qu'il contient, un ratio est obtenu. Ce ratio r est le résultat de la division de la moyenne du nombre d'incidents survenus l'année précédente (2017) par le jeu de données réel et le jeu de données anonymes, selon la ville. Ainsi, le nouveau nombre d'interventions anonymisées dans chaque point de données d'une ville est le résultat de la division du nombre d'interventions anonymisées par leur ratio calculé respectif.

Les données sont considérées comme séquentielles dans chaque ensemble de données. La cible est un vecteur, où chaque position et chaque valeur représentent respectivement la CA et le nombre de ses interventions, pour l'heure suivante $(t + 1)$ d'un échantillon actuel (t). Un échantillon actuel est composé du nombre actuel d'interventions dans chaque CA et des variables temporelles à ce moment. Comme la base de données fournie par le SDIS25 contient des informations sur les interventions suivies de 2006 à 2018, les modèles sont formés en utilisant les années 2006-2017 et testés en 2018.

6.2 Modélisation

Afin de faire une prévision multiple du nombre d'interventions par CA, une régression multicible est utilisée pour résoudre cette tâche. Ainsi, le "MultiOutputRegressor" de la bibliothèque scikit-learn [12] est appliqué. À cet égard, un régresseur par cible (CA) est ajusté en utilisant le régresseur XGBoost avec le paramètre $objective = `count : poisson'$ et le reste par défaut.

Six modèles ont été construits. Deux modèles formés avec les données réelles : un modèle de base qui décrit le nombre moyen d'interventions par CA pour chaque jour de la semaine ; et un second modèle construit avec XGBoost qui

prédit le nombre d'interventions par CA pour une journée entière. En outre, quatre modèles ont été construits avec des données anonymes en tenant compte de différents niveaux de garanties de confidentialité en utilisant également XGBoost.

L'hypothèse retenue est la suivante : les pompiers communiquent les données anonymes et les informations sur le ratio r de l'année précédente à des tiers afin de construire des modèles appropriés. L'efficacité des modèles est ensuite évaluée en comparant les résultats avec les données réelles de 2018, non apprises.

6.3 Résultats

Les modèles sont évalués à l'aide des mesures de l'erreur quadratique moyenne (EQM) et de l'erreur absolue moyenne (EAM). De plus, comme il s'agit d'un scénario à sorties multiples, les scores de chaque cible sont moyennés avec une moyenne uniformément pondérée sur les sorties.

Le tableau 2 présente les résultats des mesures pour une prédiction de base, pour les modèles formés avec les données originales et pour les modèles formés avec des données anonymes. Pour les ensembles de données anonymisées, les résultats sont présentés dans les deux cas où le ratio r est utilisé pour normaliser le nombre d'interventions par CA en fonction de l'année 2017 et dans celui où il ne l'est pas. Dans la figure 4, les meilleurs résultats de prédiction sont illustrés pour chaque CA en comparant le nombre initial d'interventions avec les modèles formés avec les données brutes et anonymes ($f = 0{,}60; \epsilon_\infty = 1{,}69$) pour une seule journée.

Modèle	Ratio normalisé		Ratio non normalisé	
	EAM	EQM	EAM	EQM
Modèle de base (moyenne)	-	-	2,5556	3,3237
Original	-	-	1,8552	2,5821
$f = 0{,}20$ $\epsilon_\infty = 4{,}39$	1,8666	2,5963	2,1748	2,8822
$f = 0{,}40$ $\epsilon_\infty = 2{,}77$	1,9271	2,7194	2,7436	3,6736
$f = 0{,}60$ $\epsilon_\infty = 1{,}69$	**1,9151**	**2,6848**	4,2475	4,9567
$f = 0{,}80$ $\epsilon_\infty = 0{,}81$	1,9403	2,7002	7,8542	8,4985

TABLE 2 – Erreurs de prédiction du nombre d'interventions journalières par CA en 2018 en utilisant des données originales et anonymes normalisées et non normalisées.

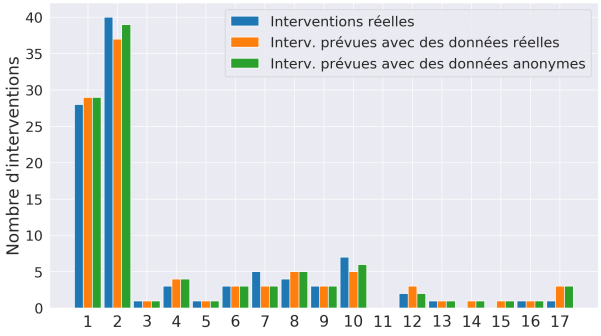


FIGURE 4 – Comparaison entre les nombres réels et prévus d'interventions par CA pour une seule journée.

6.4 Discussion

Étant donné une mise en œuvre d'un mécanisme de confidentialité différentielle locale pour la collecte de données sur la localisation des interventions, cette recherche a mis en œuvre un algorithme d'apprentissage automatique pour prédire le nombre d'interventions par CA. En comparaison avec la littérature, ce travail introduit un modèle de prévision d'interventions par période pour plusieurs CA plutôt que pour l'ensemble du département, ce qui est une tâche plus difficile. De plus, il est remarquable de constater l'amélioration du score avec les modèles formés pour une tâche aussi complexe plutôt que de développer un simple modèle de prévision comme la moyenne supposée dans cet article.

Comme on peut le constater dans le tableau 2 et dans le figure 4, les modèles formés avec des données anonymisées et normalisées peuvent également garantir une bonne utilité des données à des fins de prédiction. Il convient de noter l'utilisation d'un ratio r pour normaliser le nombre d'interventions par CA et par jour, où dans ce cas la performance de prédiction n'a pas trop diminué par rapport au modèle formé avec les données brutes. En revanche, pour les ensembles de données non normalisés, les résultats diminuent très rapidement à mesure que la garantie de confidentialité est appliquée, et après $f = 0{,}4$, les mesures de le EAM et de le EQM sont toutes deux pires que le modèle de base (la moyenne).

Les nombres en gras dans le tableau 2 représentent le meilleur résultat pour l'ensemble de données anonymisées maximisant le compromis vie privée/utilité garantissant un $\epsilon_\infty = 1{,}69$. De meilleurs résultats ont été obtenus (par exemple avec $f = 0{,}05$ et $f = 0{,}15$), cependant, les deux ensembles ont un niveau de garantie de confidentialité inférieur : $\epsilon_\infty = 7{,}33$ et $\epsilon_\infty = 5{,}02$ respectivement.

Par conséquent, dans la figure 4, il est montré pour un jour donné du mois de mars 2018 la comparaison du nombre réel et prévu d'interventions par CA en utilisant les données brutes et anonymisées ($f = 0{,}60$,$\epsilon_\infty = 1{,}69$). Avec un tel résultat, la prédiction du nombre d'interventions par CA pour le lendemain, les pompiers peuvent se préparer efficacement pour des scénarios à court, moyen et long terme. En particulier, sachant que certaines CA sont plus susceptibles de connaître des incidents, les pompiers peuvent mieux répartir les ressources humaines et matérielles ainsi que planifier la construction de nouvelles casernes. Cette approche de prédiction peut aider l'expérience humaine à prendre des décisions réelles.

Pour l'instant, un bon modèle serait celui qui peut prévoir le nombre d'interventions quotidiennes pour chaque région avec une marge d'erreur de ± 2. À l'avenir, il serait nécessaire de développer différents modèles de prévision pour chaque CA et d'établir différentes marges d'erreur en tenant compte du nombre moyen d'interventions quotidiennes. Par exemple, les deux premiers CA de la Figure 4, qui ont normalement un nombre élevé d'interventions, pourraient maintenir ± 2. Et les autres, qui ont un faible nombre d'interventions par jour, avec ± 1.

7 Conclusion

La confidentialité différentielle locale est une approche efficace utilisée pour protéger la vie privée d'une personne lors du processus de collecte de données. Plutôt que de faire confiance à un conservateur de données pour stocker les données brutes et les rendre anonymes en réponse aux requêtes (comme l'approche générale de confidentialité différentielle), la CDL permet aux utilisateurs de rendre leurs propres données anonymes avant de les envoyer au serveur du collecteur de données.

Dans cet article, l'application d'un mécanisme de CDL pour la collecte de données à des fins de préservation de la vie privée sur les lieux d'intervention des pompiers est présentée. Comme le montrent les résultats, le mécanisme "Basic One-Time RAPPOR" peut acquérir de manière adéquate des statistiques avec un bon niveau de garanties de confidentialité.

En outre, il est possible de prévoir le nombre d'interventions par CA avec des données anonymes aussi précisément qu'avec les données brutes. Ces prédictions peuvent être utilisées pour développer des outils prédictifs. Par exemple, les prévisions à court terme permettent d'optimiser les effectifs pour la semaine à venir. A moyen terme, elles permettent de redéployer de façon saisonnière les ressources matérielles et humaines dans les casernes existantes. Enfin, à plus long terme, elles peuvent aider à choisir l'emplacement géographique des futures casernes.

Nous prévoyons ainsi d'étudier dans un futur proche des approches permettant d'augmenter le volume données de manière fictive mais réaliste. L'influence de l'étape d'agrégation sur la précision des prédictions n'est pas évaluée dans cet article, qui se concentre davantage sur la partie d'anonymisation par confidentialité différentielle locale. Cette partie sera aussi adressée en travaux futurs. Enfin, la généralisation à d'autres territoires est une perspective pratique que nous travaillerons.

Remerciements

Ce travail a été soutenu par le projet CADRAN de la Région Bourgogne Franche-Comté, par l'école doctorale EIPHI-BFC (contrat "ANR-17-EURE-0002"), par le projet Interreg RESponSE et par la brigade de pompiers SDIS25.

Références

[1] S. Cerna, C. Guyeux, H. Hwang Arcolezi, A. D. P. Lotufo, R. Couturier, and G. Royer. Long short-term memory for predicting firemen interventions. In *6th International Conference on Control, Decision and Information Technologies (CoDIT 2019)*, Paris, France, apr 2019.

[2] Jean-François Couchot, Christophe Guyeux, and Guillaume Royer. Anonymously forecasting the number and nature of firefighting operations. In *Proceedings of the 23rd International Database Applications & Engineering Symposium on - IDEAS19*. ACM Press, 2019.

[3] T. T. Dang, Y. Cheng, J. Mann, K. Hawick, and Q. Li. Fire risk prediction using multi-source data : A case study in humberside area. In *2019 25th International Conference on Automation and Computing (ICAC)*, pages 1–6, Sep. 2019.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[5] Cynthia Dwork. Differential privacy : A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confidentiality*, 7(3) :17–51, 2016.

[7] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4) :211–407, 2014.

[8] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor : Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

[9] Christophe Guyeux, Jean-Marc Nicod, Christophe Varnier, Zeina Al Masry, Nourredine Zerhouny, Nabil Omri, and Guillaume Royer. Firemen prediction by using neural networks : A real case study. In *Advances in Intelligent Systems and Computing*, pages 541–552. Springer International Publishing, August 2019.

[10] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. Differential privacy : An economic method for choosing epsilon. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium*, CSF '14, pages 398–410, Washington, DC, USA, 2014. IEEE Computer Society.

[11] Jong Wook Kim, Dae-Ho Kim, and Beakcheol Jang. Application of local differential privacy to collection of indoor positioning data. *IEEE Access*, 6 :4276–4286, 2018.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[13] Latanya Sweeney. k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :557–570, October 2002.

Inférence Grammaticale pour la Prévision de la Consommation Énergétique

G. Guerard[1], H. Pousseur[2], M. Rivoire[2]

[1] Pole Universitaire Léonard de Vinci, Research Center, DVRC
[2] Research Student, DVRC
92 916, Paris La Défense, France

prenom.nom@devinci.fr

Résumé

La domotique est programmable, contrôlable, et la plupart du temps a une consommation connue. Néanmoins, la prévision d'une consommation des ménages représente un problème difficile généralement résolu à l'aide de l'apprentissage en profondeur ou de modèles mathématiques coûteux en temps et en mémoire. L'approche proposée est basée sur l'inférence grammaticale pour prédire la consommation d'un appareil à partir du passé. L'automate de prédiction consomme peu de ressources et peut être facilement intégré dans des appareils intelligents, des compteurs intelligents ou des maisons intelligentes.

Mots-clés

Prédiction, Inférence grammaticale, ALERGIA

Abstract

Home automation is programmable, schedulable, and most of the time has a known consumption. Nevertheless, forecasting a household consumption represents a challenging problem typically resolved using deep learning or huge mathematical models. The proposed approach is based on grammatical inference to predict a device's consumption from the past. The prediction machine consumes little resources and can be easily embedded into smart devices, smart meters or smart houses.

Keywords

Forecast, Grammar Induction, ALERGIA

1 Introduction

Au cours de la dernière décennie, les applications de prévision énergétique ont été développées non seulement du côté réseau des réseaux électriques, mais aussi du côté client afin d'équilibrer la charge et de la demande. Dans ce contexte, la prévision à court terme de la consommation d'énergie électrique est une condition nécessaire à la gestion de l'énergie et à la planification de tous les bâtiments, des ménages et des résidences à petite échelle aux énormes complexes immobiliers à grande échelle. Elle permet de surveiller la consommation d'énergie et de trouver les creux et les pics de demande, de réduire les pertes, de minimiser les risques, de garantir la fiabilité pour un fonctionnement ininterrompu. La prévision des charges joue un rôle actif dans la prise de décisions viables en ce qui concerne la planification de la maintenance et les investissements futurs, y compris les technologies énergétiques renouvelables et non renouvelables.

Plus récemment, l'omniprésence de l'Internet des objets rend les systèmes énergétiques distribués plus intelligents en optimisant l'efficacité énergétique pour réduire les pertes et crée une nouvelle aire appelée Internet de l'énergie (IoE) qui est équipée de systèmes de prévision intelligents. Ces derniers utilisent des prévisions météorologiques et d'autres explications pour prévoir la consommation d'énergie future. Selon l'un des derniers rapports de l'Agence internationale de l'énergie [1], les bâtiments représentent la plus grande partie de la consommation finale d'énergie avec une part de 36% du marché mondial, ce qui augmente l'importance des prévisions énergétiques des bâtiments afin de rétablir l'équilibre entre l'offre et la demande pour un avenir plus économe en énergie.

Il n'existe actuellement pas de norme définissant les types de prévisions. Hong et Fan ont regroupé les catégories de prévisions en très court terme, court terme, moyen terme et long terme avec des horizons de coupure de 1 jour, 2 semaines et 3 ans [9]. Principalement, les prévisions à court terme se réfèrent à des prévisions à l'heure, à la journée ou à la semaine à venir et il est considéré que ce concept peut également être appliqué à la prévision de la consommation d'énergie électrique des bâtiments [23].

Les algorithmes d'apprentissage automatique de la littérature sont couramment utilisés pour prévoir la consommation d'énergie électrique des bâtiments à court terme. La littérature contient une variété d'étude résumant les méthodologies de prévision de la consommation d'énergie des bâtiments sous différents angles.

Zhao et Magoules ont examiné les prévisions de consommation d'énergie des bâtiments en classant les méthodologies telles que les méthodes d'ingénierie, les méthodes statistiques et les méthodes d'IA [22]. Ahmad et al. ont ré-

1. 2019 Global Status Report for Buildings and Construction : Towards a zero-emission, efficient and resilient buildings and construction sector.

sumé les applications des réseaux de neurones artificiels (ANN) et des machines à vecteurs de support (SVM) [1]. Raza et Khosravi ont mené une étude sur les techniques de prévision de la demande de charge basées sur l'IA, non seulement pour les bâtiments, mais aussi pour les réseaux intelligents, en expliquant toutes les phases de la prévision de charge à court terme [16]. Daut et al. ont examiné la prédiction de la consommation d'énergie électrique des bâtiments en divisant les méthodologies en méthodes conventionnelles, IA et hybrides [5]. Wang et Srinivasan ont comparé des modèles simples et d'ensemble pour la prévision de la consommation d'énergie des bâtiments basée sur l'IA [20].
Wei et al. ont présenté des approches d'apprentissage machine basées sur l'étude des données pour la prévision et la classification de la consommation d'énergie des bâtiments [21]. De la même manière, Amasyali et El-Gohary ont examiné les études de prévision de la consommation d'énergie des bâtiments basées sur les données en se concentrant particulièrement sur les domaines de prédiction, les propriétés des données et les méthodes de prétraitement, les algorithmes d'apprentissage automatique et les mesures de performance [2]. Enfin, Runge et Zmeureanu ont présenté une étude de la prévision de la consommation d'énergie dans les bâtiments utilisant des ANN en mettant en évidence les applications, les données, les modèles de prévision et les mesures de performances [17].
L'apprentissage machine et l'apprentissage profond sont efficaces et proposent des modèles avec des dizaines de colonnes en entrée (en fonction des capteurs / données externes et internes aux bâtiments) pour prédire la consommation énergétique. Cependant, un détail n'est jamais abordé dans ces études : la consommation énergétique d'un tel dispositif par rapport aux gains potentiels.
Notre étude propose une méthode de prédiction peu consommatrice en temps et en mémoire, donc énergétiquement sobre. La méthode se base sur l'inférence grammaticale, c'est-à-dire un apprentissage par construction d'un automate stochastique à partir d'une série temporelle représentant la consommation énergétique d'un appareil.
Le papier est construit comme suit : la deuxième section présente la méthodologie ; cette dernière est décomposée dans les sections 3, 4 et 5 par les traitements sur le jeu de données, les opérations sur ces dernières et la méthode d'inférence grammaticale. La section 6 présente les résultats et le papier se conclut par la section 7.

2 Méthodologie

2.1 Jeu de données

Chaque appareil produit une courbe de consommation. Celui-ci est enregistré par un compteur intelligent sous la forme d'une série temporelle où chaque tuple fournit les informations suivantes : ID_périphérique, consommation (en Wh), date. La courbe de consommation est considérée sans erreur car le compteur intelligent transmet une valeur consolidée. Le système d'éclairage, l'ordinateur et les petits appareils électroménagers ne sont pas pris en compte.

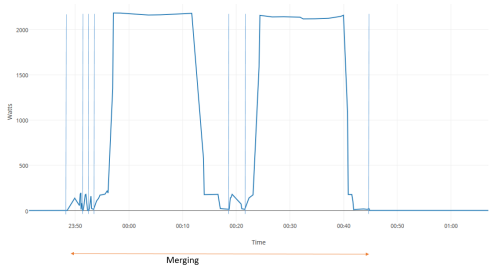


FIGURE 1 – Cycle de consommation d'une machine à laver.

Afin d'apprendre de la consommation d'un appareil, nous devons définir des séquences de consommation. Cette dernière commence lors du passage d'une valeur de consommation nulle à une valeur de consommation non nulle sur deux étapes consécutives de la série temporelle. Dès qu'il y a une valeur nulle de consommation à l'étape suivante, la séquence s'arrête.
Certains appareils peuvent avoir des cycles de consommation programmés avec de courtes périodes à consommation nulle. La méthode suivante fusionne deux séquences de consommation si elle est considérée sur un même cycle de consommation :

$$\Delta B \leqslant \Delta T_i \text{ and } \Delta B \leqslant \Delta T_j$$

avec les variables suivantes : ΔB le temps de césure ; ΔT_i le temps de consommation de la séquence i ; ΔT_j pour la séquence j.
La figure 1 présente un exemple de fusion de plusieurs séquences (séparées par des lignes verticales) pour constituer un cycle de consommation unique.

2.2 Méthodes

L'approche proposée dans ce papier est composée de trois étapes distinctes :

- TRAITEMENT DES DONNÉES : tout d'abord, les séquences sont définies comme présentées précédemment. Le bruit est éliminé par une méthode de regroupement.
- PRÉRÉGLER LES DONNÉES : ensuite, les séquences sont analysées en motifs récurrents.
- INFÉRENCE GRAMMATICALE : un automate stochastique est construit par inférence grammaticale des séquences
- PRÉDICTION ET SIMULATION : à partir d'un début de séquence, il est possible de prédire les possibilités futures, mais aussi d'obliger l'automate stochastique à adopter un comportement donné.

3 Traitements des données

La consommation des appareils est considérée en Wh. Étant donné le très grand nombre de symboles que cela produit, il est pertinent de nettoyer les données : un premier algorithme pour discrétiser les données, nommé *Discr. Algo.* ; un deuxième algorithme partitionne l'alphabet, nommé *Clust. Algo.*

3.1 Réduction des données

Discr. Algo. a pour objectif, à partir d'un échantillon de toutes les données de consommation appelé *data*, à fournir un ensemble de données discrétisé à un timer constant qui commence à la valeur de 5 minutes. *Discr. Algo.* calcule entre chaque valeur la pente. En effet, les appareils peuvent avoir des cycles de consommation similaires, mais avec des variations de la puissance requise (par exemple les types distinctifs de cycle de lavage d'une machine à laver).
Une fois que toutes les pentes sont calculées dans l'ensemble de données suivant *données discrétisées*, l'intégrale de la courbe *données* est comparée à l'intégrale de la courbe *données discrétisées*. Le rapport entre ces deux courbes fournit une estimation de la perte d'informations. Ce ratio ne doit pas dépasser une valeur α.
Tant que ce rapport est strictement inférieur à α, il est possible de considérer les données avec un timer plus grand. Au contraire, si le rapport est supérieur à α, il faut raccourcir le timer. L'algorithme s'arrête lorsque le rapport est proche de α.

3.2 Partitionnement des données

Discr. Algo. génère un grand alphabet alors que certaines valeurs de pente sont proches. Pour réduire la taille de l'alphabet et surtout pour réduire le bruit entre les valeurs, des valeurs similaires sont assimilées à une unique. Ce processus est appelé partitionnement.
Le partitionnement représente le processus de détermination de groupes typiques, appelés partitions, dans un ensemble de données. L'objectif est de trouver les partitions les plus homogènes et les plus distincts possible des autres partitions. Plus formellement, le regroupement devrait maximiser la variance inter-partition tout en minimisant la variance intra-partition.
Une classification des méthodes de partitionnement pour diverses données statiques est proposée dans [10]. Dans le contexte de notre ensemble de données, le partitionnement basé sur les centroïdes convient.
Le problème contient une dimension (les valeurs de consommation), donc le centre de la partition est une moyenne arithmétique des valeurs. L'algorithme k-moyennes convient exactement à l'ensemble de données et a une petite complexité de temps et de mémoire. Cet algorithme est linéaire en fonction du k, du nombre de points et du nombre d'itérations.
Comme dans l'algorithme précédent, il ne doit pas perdre trop d'informations. Le rapport entre la différence des courbes intégrales *data clustered* et *data* ne doit pas dépasser une valeur α'.
Tant que ce ratio est strictement inférieur à α', il est possible de diminuer la valeur k des k-moyennes. Au contraire, si le rapport est supérieur à α', on augmente la valeur de k. Notez que la k-moyenne peut générer des résultats divers pour la même valeur de k.

3.3 Extraction des données isolées

Une fois que les k-moyennes atteignent un partitionnement approprié, une mesure spécifique est calculée : la silhouette. Elle fait référence à une méthode d'interprétation et de validation de la cohérence dans le regroupement des données et fournit une représentation graphique de la façon dont chaque point a été classé.
Soit $a(i)$ la distance moyenne entre un point i et tous les autres points de données de la même partition ; soit $b(i)$ la plus petite distance moyenne du point i à tout autre groupe dont i ne fait pas partie de l'ensemble. La silhouette du point i est définie comme suit :

$$s(i) = \begin{cases} 1 - a(i)/b(i), \text{if } a(i) < b(i) \\ 0, \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, \text{if } a(i) > b(i) \end{cases}$$

la silhouette donne un résultat compris entre -1 et 1. Un $s(i)$ proche de 1 signifie que les données sont correctement regroupées.
Pour chaque partition, la silhouette est calculée pour chaque point. Si la silhouette de certains points est inférieure à un seuil $0 < silh < 1$, alors ces points sont considérés comme $noise$. Si la silhouette de certains points est supérieure au seuil $silh$, alors ces points sont considérés comme $accessibles$. Si la silhouette fournit une valeur négative, le point est concédé à la partition correspondante. Une nouvelle partition est construite pour chaque point $noise$.
La méthode proposée n'est pas optimale mais linéaire sur le nombre de clusters et le nombre d'itérations, et log-linéaire sur le nombre de points.

4 Opérations sur le jeu de données

Un appareil peut avoir des motifs de consommation qui se produisent dans plusieurs ou une même séquence. Pour limiter la longueur des séquences et améliorer l'efficacité de la prédiction, ces motifs sont regroupés comme un unique symbole, c'est-à-dire qu'une fois le motif détecté ou suggéré, la prédiction le produira dans son entièreté.
Cette section est composée de deux algorithmes : *Disco. Motif. Algo.* est conçu pour découvrir un motif fréquent à l'intérieur de l'ensemble des séquences ; *Ext. Motif. Algo* étend un motif à l'intérieur des séquences.

4.1 Découverte des motifs

L'exploration de modèle séquentiel est une méthode d'exploration de données importantes qui peut extraire des séquences similaires tout en conservant leur ordre. Cependant, il est essentiel d'identifier les intervalles des éléments des modèles séquentiels extraits par exploration de modèles séquentiels. *Disco. Motif. Algo.* est basé sur l'algorithme de Hirate et Yamana [8].
Ces intervalles peuvent être un intervalle entre deux symboles ou un intervalle de temps. Dans notre jeu de données, les deux approches sont identiques. De plus, le motif doit être continu et contigu. L'exploration de modèle séquentiel est donc définie sur des données contiguës sans intervalle de temps.
En plus du motif, l'algorithme Hirate et Yamara fournit aussi son support. Le support fournit la fréquence de cette

TABLE 1 – Jeu de données après *Clust. Algo.*

ID	Séquence
1	aaabacc
2	bacaaa
3	ccbacaaa

TABLE 2 – Algorithme de Hirate and Yamana.

Motif	Support (>50%)
bac	100%
aaa	100%
ba	100%
...	...
cc	66%
ca	66%

sous-séquence parmi l'ensemble de données mais ne donne pas si le motif est récurrent dans la même séquence. Ce processus est illustré dans les tableaux 1 et 2.

Après avoir trouvé les motifs, seuls ceux dont la longueur et le support sont les plus importants sont pris en compte sans aucun conflit entre eux. Le tableau 3 montre les séquences de l'exemple avant et après traitement. Les motifs sont entre parenthèses.

4.2 Extension des motifs

L'objectif de *Ext. Motif. Algo.* est de découvrir des versions étendues des motifs dans les séquences. Par exemple, les séquences $aaaeaaa$ et $aaafaaa$ ne diffèrent que par un seul symbole. Si les symboles e et f ont des valeurs proches, les deux séquences sont considérées comme égales avec un nouveau symbole $g = \frac{e+f}{2}$.

Á partir de deux mêmes motifs, l'extension se réalise sur le suffixe et / ou le préfixe. Si la pente du motif généré diffère de α'' pour cent de celui d'origine, alors le processus d'agglomération s'arrête pour le côté concerné.

5 Inférence grammaticale et prédiction

Suivant les étapes précédentes, les cycles de consommation sont désormais composés de symboles et de motifs. L'objectif de cette section est de construire un automate de prédiction à partir de ces séquences. À partir d'un arbre de préfixe de fréquence, une inférence grammaticale construit l'automate stochastique. À ce stade, une marche aléatoire détermine les schémas de consommation future de l'appareil concerné.

TABLE 3 – Jeu de données après *Disco. Motif. Algo.*

ID	Séquence
1	(aaa)(bac)c
2	(bac)(aaa)
3	cc(bac)(aaa)

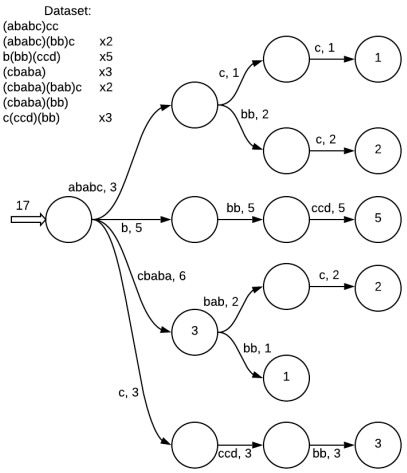


FIGURE 2 – Exemple d'arbre de préfixe.

5.1 Arbre de préfixe

Un arbre de préfixe est un arbre dont les transitions représentent les symboles d'une séquence et qui possède les deux propriétés suivantes : pour chaque état, la somme des fréquences (des symboles) entrantes est égale à la somme des fréquences sortantes ; le nombre de séquences démarrant à la racine est égal à la somme des arrêts sur les noeuds.

Pour chaque appareil, chaque séquence est entrée dans l'arbre de préfixe comme suit :

1. Pour commencer à analyser une séquence, on se positionne à la racine de l'arbre de préfixe et on incrémente la fréquence d'entrée de 1.
2. Pour chaque symbole d'un mot (ordre des préfixes).
   (a) Si un arc contenant ce symbole existe déjà, incrémentez sa valeur de fréquence de un et continuez d'interpréter le mot sur le nœud enfant correspondant.
   (b) Sinon, créez un nœud enfant correspondant au symbole avec un arc de valeur de fréquence 1.
   (c) S'il s'agit du symbole terminal du mot, alors le nœud enfant correspondant devient terminal avec une valeur de terminaison incrémentée de 1.
3. Renvoyer l'arborescence des préfixes de fréquences.

Un exemple est montré sur la figure 2 où chaque mot est suivi de sa fréquence. Les fréquences deviennent des fréquences relatives (probabilité) en divisant chaque fréquence sortante par la somme de toutes les fréquences sortantes.

5.2 Inférence grammaticale

Comme nous avons un arbre de préfixe probabiliste, il est possible de faire une inférence grammaticale sur celui-ci. L'inférence grammaticale a été développée de manière significative par Colin de la Higuera dont les travaux sont présentés dans ces livres [6, 7]. L'inférence grammaticale représente le processus d'apprentissage d'une grammaire formelle à partir d'un ensemble d'observations (dans notre cas, l'arbre de préfixe probabiliste).

Il existe de nombreuses méthodes d'inférence grammaticale, souvent en compétition dans le concours PAutomaC [19]. Parmi les algorithmes, ALERGIA est un algorithme non déterministe permettant la réduction d'un automate déterministe probabiliste par un calcul d'équivalence d'automate probabiliste [4, 18].

ALERGIA a été appliqué à l'extraction d'informations à partir de textes ou de documents structurés et à la modélisation du langage vocal. Lorsque la probabilité d'apparition d'une chaîne suit une distribution bien approximée, ALERGIA a la capacité de fusionner des noeuds lorsque les automates résultants sont compatibles avec la fréquence observée de chaînes.

Soit n_i le nombre de séquences arrivant au noeud i ; f représente la fréquence : $f_i(a)$ est le nombre de séquences arrivant au noeud i et ayant comme prochain symbole a ; $f_i(.)$ est le nombre de séquence se terminant en i. ALERGIA a besoin du calcul préliminaire suivant : $p_i(a) = \frac{f_i(a)}{n_i}$ pour chaque transition sortante.

Deux noeuds i et j sont dit compatibles si pour toutes les transitions (dont les terminaisons) sortantes et récursivement pour l'ensemble de leurs enfants :

$$|p_i(a) - p_j(a)| < \sqrt{0.5 \ln \frac{2}{\epsilon}} \left(\frac{1}{\sqrt{n_i}} + \frac{1}{\sqrt{n_j}} \right)$$

où ϵ est appelé plage d'acceptation.

L'algorithme ALERGIA vérifie toutes les paires de noeuds, si aucune paire n'est compatible, l'algorithme s'arrête. La fusion et la plage d'acceptation sont présentées dans les deux sous-sections suivantes.

5.3 Fusion

Si deux noeuds i et j sont compatibles, ALERGIA traite la fusion et le repli de ceux-ci et de leurs enfants. Ce processus est connu sous le nom d'algorithme RPNI [12].

La fusion de deux noeuds signifie les réduire à un unique, dont la position est la plus petite profondeur des deux fusionnés. Ce noeud est considéré comme terminal si le noeud fusionné était terminal.

Quant aux transitions sortantes, elles sont elles-mêmes fusionnées si elles sont étiquetées avec le même symbole, et dans un tel cas, les deux noeuds pointés sont fusionnés récursivement.

Étant donné que i et j sont fusionnés, dans un premier temps, la transition entre j et ses parents est rompus. Ces transitions se situent actuellement entre i et les parents de j (pour chaque transition).

Ensuite, l'algorithme RPNI absorbe récursivement les enfants de j comme suit :

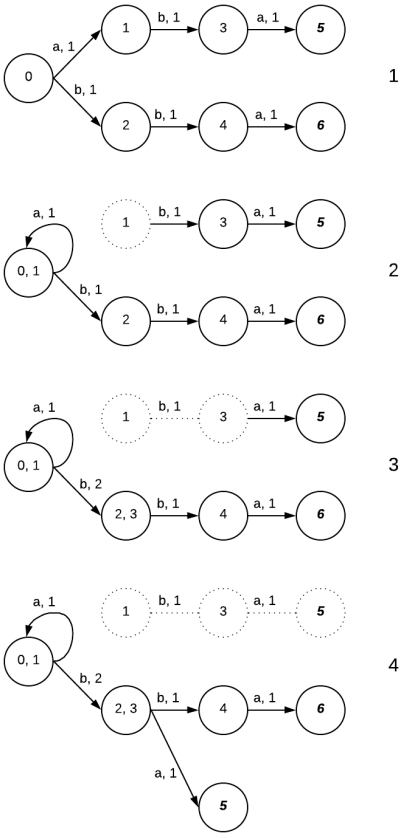


FIGURE 3 – Fusion entre les sommets 0 et 1.

1. Si la transition entre j et un successeur existe entre i et un successeur ; alors la fréquence de ce chemin est incrémentée. Si le successeur de j est un noeud terminal, alors l'état correspondant de i devient ou reste un noeud terminal.
2. Si la transition entre j et un successeur n'existe pas entre i et un successeur, et récursivement pour toutes les transitions entre j et le successeur choisi ; alors le chemin de j au successeur est attaché à i.

Dans la figure 3, nous fusionnons les noeuds 0 et 1. La transition montre le symbole et la fréquence, le noeud terminal est en italique et en gras. À la première étape, la transition entre 1 et ses parents est rompue puis ajoutée entre 0 et le parent de 1, dans ce cas entre 0 et 0. Cela crée une boucle étiquetée a sur le noeud 0 comme indiqué dans la deuxième image.

On considère maintenant le chemin entre 1 et 3, la transition b existe déjà après le noeud 0, mais pas le chemin ba. Ainsi, l'état 3 fusionne avec le noeud 2 (troisième image) mais le noeud 5 reste seul. Ce dernier se réfère à la deuxième règle, le noeud est apposé à l'état 2 comme le montre la quatrième image.

5.4 Influence de la plage d'acceptation

Les scientifiques spécialistes dans l'inférence grammaticale posent la plage d'acceptation à $\epsilon = \frac{1}{N^r}$ avec N le

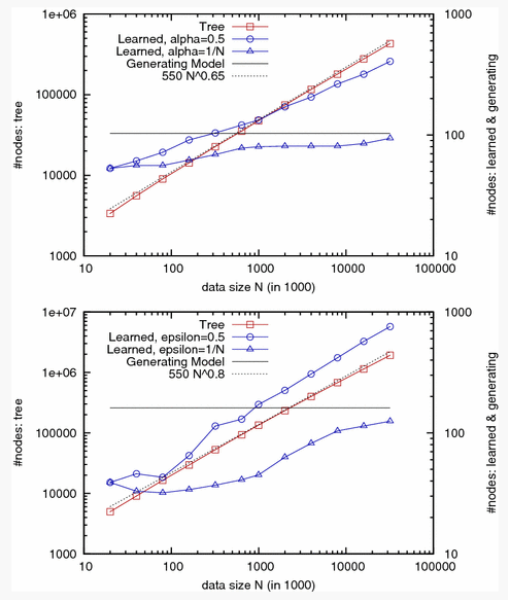


FIGURE 4 – Taille de l'automate finale (en haut $r = 3$, en bas $r = 5$) [14].

nombre de symboles différents dans l'ensemble des séquences, et le facteur de puissance à $r > 2$. Mao et al. [14] ont effectué de nombreuses expériences sur la valeur de r. La figure 4 provient de leur papier, les deux graphes présents indiquent le nombre de sommets de l'automate stochastique en fonction du nombre de sommets initiaux et de la valeur de r.
Afin de comprendre la similitude entre l'automate réduit et l'automate de départ, Mao et al. ont implanté la divergence de Kullback-Leiber [11]. La divergence diminue à r constant et en augmentant le nombre de sommets de départ ; la divergence augmente quand r augmente (donc si ϵ diminue).
L'automate généré par la présente méthode doit générer les mots de départ mais doit être suffisamment flexible afin de reconnaître des comportements non rencontrés lors de la phase d'apprentissage. Le compromis est trouvé avec $r = 3$.

5.5 Prédiction

Une fois l'automate réduit par l'algorithme ALERGIA, il est possible de l'exploiter pour fournir des prédictions sur la consommation future de l'appareil. Les prédictions sont obtenues en effectuant des marches aléatoires dans l'automate probabiliste sur un certain nombre w d'étapes. Cela est réalisé en considérant l'automate probabiliste comme une chaîne de Markov, le passage à travers un arc produira le symbole associé comme une prédiction. La prédiction est un suffixe de la consommation actuelle.
A noter que la prédiction par marche aléatoire peut se faire en générant des individus. L'autre possibilité, adapté à des chaînes de petite taille, consiste à multiplier le vecteur de population (donc le vecteur de distribution après lecture du préfixe) par la matrice représentant la chaîne de Markov associé à l'automate stochastique.

TABLE 4 – Prédiction des suffixes à partir du mot b.

Longueur	Suffixes et probabilités d'apparition
1	{b, 0.86} ; {ϵ, 0.14}
2	{ba, 0.11} ; {bb, 0.43} ; {bϵ, 0.32}, {ϵ, 0.14}
3	{bab, 0.07} ; {baa, 0.04} ; {bbϵ, 0.43} ; {bϵ, 0.32} ; {ϵ, 0.14}

Étant donné que la fréquence d'un chemin avec des individus et la probabilité d'un suffixe convergent lorsque le nombre d'individus est vaste, ils sont un moyen plus pratique de générer des prévisions. En effet, un arbre de possibilités de longueur croissante utilise plus de mémoire que certaines marches aléatoires. Étant donné que la machine de prédiction peut être intégrée, les marches aléatoires sont préférées.
Expliquons le processus avec un exemple. La figure 5 présente à gauche la machine de prédiction avec les symboles et les fréquences ; et à droite la chaîne de Markov correspondante. Le tableau 4 montre une prédiction de l'augmentation de la longueur du mot b avec la probabilité correspondante du suffixe (où ϵ signifie la fin du mot).

6 Expériences

6.1 Résultats

Les tests ont été effectués avec un ensemble de données de plus de 700.000 lignes référençant la quantité de kW consommée avec un timer d'une seconde, soit plus d'une semaine de données. Le dispositif étudié est un réfrigérateur industriel fréquemment utilisé et pouvant contenir des produits différents d'une journée à l'autre. Sa consommation dépend des éléments à l'intérieur, de l'heure d'ouverture de la porte et de sa variable environnementale. La figure 6 présente un échantillon de l'ensemble de données.
La prédiction a été générée plusieurs fois et toujours après un départ de consommation de l'appareil. Toutes les prédictions fournissent une prévision précise. En prenant un intervalle de 95% des prédictions faites à partir d'un instant donné (nous écartons les prédictions trop éloignées de la masse), au moins 92% (97% en moyenne) de toutes les prévisions sont similaires à la consommation réelle avec un écart-type d'au plus 5% (3% en moyenne). Cet écart-type s'explique par les différentes approximations faites dans la méthode proposée ; l'écart type dépend de la valeur des seuils.

6.2 Discussions

La méthode présentée dans ce papier obtient de bons résultats sur des appareils ayant une latence dans sa consommation. C'est-à-dire que suite à un stimulus, la consommation de l'appareil est perturbé après et pendant un certain temps qui est bien plus grande que la durée du stimulus. En effet, l'automate de prédiction est une méthode assez simple

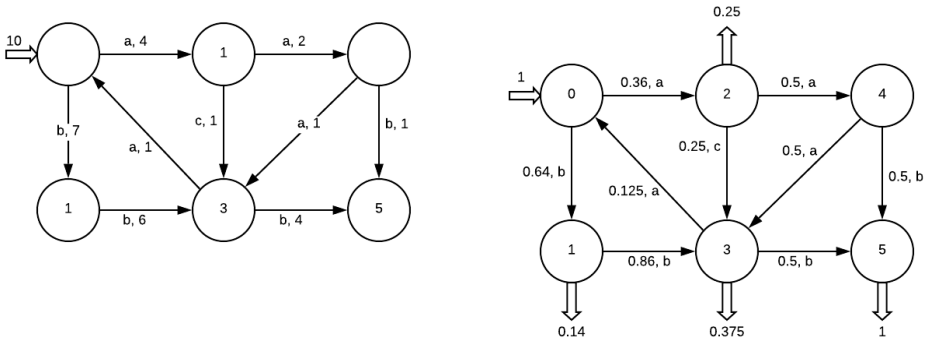


FIGURE 5 – De l'automate à fréquence à un automate stochastique.

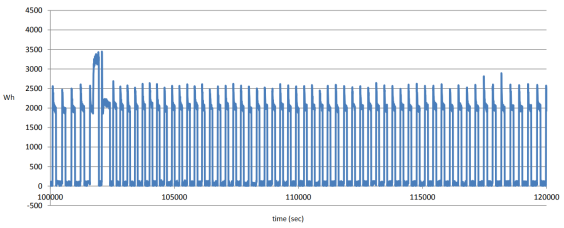


FIGURE 6 – Données utilisées.

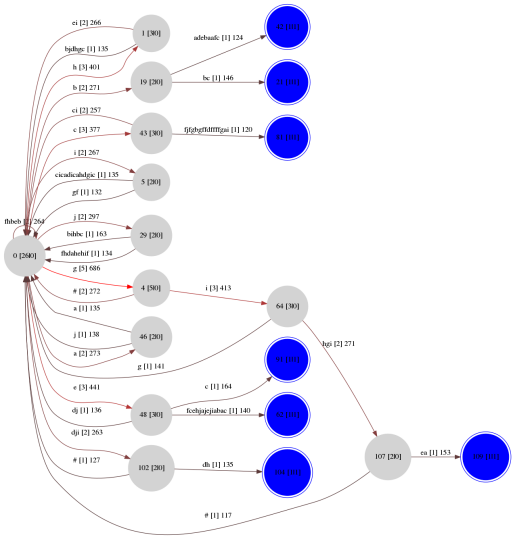


FIGURE 7 – Automate stochastique obtenu par la méthode présentée.

représentant des cycles de consommations. Nous avons vu précédemment que le processus de la méthode repère les schémas récurrents puis réduit l'automate par inférence. Il est donc logique que l'automate soit capable de reconnaître des schémas à des types de stimuli mais ait plus de difficulté à prédire la consommation d'un appareil plus erratique.
Si l'automate présente des résultats non satisfaisants, il est possible de créer un nouvel automate stochastique avec des nouvelles valeurs.

7 Conclusion

La prédiction de la consommation est un problème fondamental du réseau électrique intelligent. La plupart des méthodes de la littérature font appel à des méthodes coûteuses, consommatrices mais ayant des résultats similaires à la réalité. Cependant, notre approche soulève et répond à une problématique sous-jacente à la prédiction : peut-on prédire la consommation future avec peu de mémoire et une puissance de calcul minime ? si possible de manière embarquée dans l'appareil ou chez le consommateur.
La méthode proposée est adaptée à des appareils ayant des schémas distincts et répondant à des stimuli avec une grande amplitude dans le temps comparé à la durée du stimulus. Le principal défaut de la méthode est la mise à jour de son automate stochastique en repartant de zéro. Les travaux futurs ont pour objectif de changer l'automate stochastique en chaîne de Markov à état caché. En effet ces dernières sont capables d'apprendre les nouveaux cycles de consommation au fil de l'eau si la prédiction s'avère trop médiocre.
De plus, la première partie de l'algorithme sur la discrétisation des données doit être retravaillée et comparée à des méthodes comme [3], SAX [13] et Persist [15]

Références

[1] AS Ahmad, MY Hassan, MP Abdullah, HA Rahman, F Hussin, H Abdullah, and R Saidur. A review on applications of ann and svm for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33 :102–109, 2014.

[2] Kadir Amasyali and Nora M El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81 :1192–1205, 2018.

[3] Marc Boullé. Modl : A bayes optimal discretization method for continuous attributes. *Machine learning*, 65(1) :131–165, 2006.

[4] Rafael C Carrasco and José Oncina. Learning stochastic regular grammars by means of a state merging method. In *International Colloquium on Grammatical Inference*, pages 139–152. Springer, 1994.

[5] Mohammad Azhar Mat Daut, Mohammad Yusri Hassan, Hayati Abdullah, Hasimah Abdul Rahman, Md Pauzi Abdullah, and Faridah Hussin. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods : A review. *Renewable and Sustainable Energy Reviews*, 70 :1108–1118, 2017.

[6] Colin De la Higuera. *Grammatical inference : learning automata and grammars*. Cambridge University Press, 2010.

[7] Rémi Eyraud, Colin De La Higuera, Makoto Kanazawa, and Ryo Yoshinaka. Introduction to the grammatical inference special issue of fundamenta informaticae, 2016.

[8] Yu Hirate and Hayato Yamana. Generalized sequential pattern mining with item intervals. *JCP*, 1(3) :51–60, 2006.

[9] Tao Hong and Shu Fan. Probabilistic electric load forecasting : A tutorial review. *International Journal of Forecasting*, 32(3) :914–938, 2016.

[10] Han Jaiwei and Micheline Kamber. Data mining : concepts and techniques. *ed : Morgan Kaufmann San Francisco*, 2006.

[11] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[12] Kevin J Lang. Random dfa's can be approximately learned from sparse uniform examples. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 45–52, 1992.

[13] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax : a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2) :107–144, 2007.

[14] Hua Mao, Yingke Chen, Manfred Jaeger, Thomas D Nielsen, Kim G Larsen, and Brian Nielsen. Learning deterministic probabilistic automata from a model checking perspective. *Machine Learning*, 105(2) :255–299, 2016.

[15] Fabian Mörchen and Alfred Ultsch. Finding persisting states for knowledge discovery in time series. In *From Data and Information Analysis to Knowledge Engineering*, pages 278–285. Springer, 2006.

[16] Muhammad Qamar Raza and Abbas Khosravi. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, 50 :1352–1372, 2015.

[17] Jason Runge and Radu Zmeureanu. Forecasting energy use in buildings using artificial neural networks : a review. *Energies*, 12(17) :3254, 2019.

[18] Franck Thollard, Pierre Dupont, Colin de la Higuera, et al. Probabilistic dfa inference using kullback-leibler divergence and minimality. In *ICML*, pages 975–982, 2000.

[19] Sicco Verwer, Rémi Eyraud, and Colin De La Higuera. Pautomac : a probabilistic automata and hidden markov models learning competition. *Machine learning*, 96(1-2) :129–154, 2014.

[20] Zeyu Wang and Ravi S Srinivasan. A review of artificial intelligence based building energy use prediction : Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75 :796–808, 2017.

[21] Yixuan Wei, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, and Xiaoyun Zhao. A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82 :1027–1047, 2018.

[22] Hai-xiang Zhao and Frédéric Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6) :3586–3592, 2012.

[23] Kasım Zor, Oğuzhan Timur, and Ahmet Teke. A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. In *2017 6th International Youth Conference on Energy (IYCE)*, pages 1–7. IEEE, 2017.

Classification d'instabilité d'équipements réseaux FTTH à l'aide de techniques de Machine Learning

Susie BRUNESSAUX, Jean-Luc COURANT

Orange Labs Networks, 2 avenue Pierre Marzin, 22300 Lannion

{ susie.brunessaux@orange.com, jeanluc.courant@orange.com }

Résumé

Pour un opérateur télécom, la qualité de service rendue au client est essentielle, en particulier sur les offres d'accès fixe sur support fibre qui sont amenées à se généraliser. Le fonctionnement des équipements sur ce support est considéré comme binaire : en fonctionnement ou en panne. Malgré tout, nous constatons que des équipements supportant le service présentent des instabilités. Il convient de déterminer l'impact pour le client et d'apporter une solution. Pour cet objectif, les techniques de Machine Learning appliquées à des données archivées ont permis d'une part de mettre en lumière les clients souffrant d'instabilité au niveau de leurs services et de détecter d'autre part la cause racine afin de les réparer dans les plus brefs délais.

Mots-clés

Classification, fibre optique, FTTH, Machine Learning, Big Data, agrégats, données déséquilibrées.

Abstract

For a telecom operator, the quality of service delivered to the customer is essential, in particular on fixed access offers on fiber support which should become widespread. The operation of the equipment on this support is considered binary: in operation or out of order. Despite everything, we note that equipment supporting the service can have instabilities. It is necessary to determine the impact for the customer and provide a solution. For this objective, Machine Learning techniques applied to archived data have made it possible, on the one hand, to highlight customers suffering from instability in their services and, on the other hand, to detect the root cause in order to repair them as soon as possible.

Keywords

Classification, optical fiber, FTTH, Machine Learning, Big Data, aggregated data, unbalanced data.

1 Introduction

Avec l'accélération des déploiements de services d'accès à internet par fibre optique, l'amélioration des techniques de diagnostic FTTH (Fiber To The Home) du réseau d'accès est un sujet primordial pour rétablir au plus vite le client après un dysfonctionnement. Pour cela, il est important d'apporter aux techniciens d'intervention une préconisation pertinente afin de gagner du temps dans la durée de rétablissement du service. D'autant plus, si la panne se manifeste de façon aléatoire, auquel cas, le technicien pourra difficilement mettre en œuvre ses compétences pour analyser le problème.

Cet article décrit la démarche adoptée pour trouver la cause racine du dysfonctionnement puis proposer une solution pour ces cas de clients instables, c'est-à-dire dont les équipements réseau FTTH génèrent des alarmes de manière intermittente avec un risque d'impact sur l'accès aux services. Des investigations à l'aide de techniques de Machine Learning ont permis de classifier ces cas selon l'impact sur la qualité de services. Cette classification était difficilement réalisable via des techniques classiques étant donné la volumétrie importante d'indicateurs relatifs à l'état du service. Enfin, en corrélant les cas issus de la classification avec l'efficacité des actions de résolution tentées par les techniciens ou les téléconseillers, une préconisation a pu être proposée. L'objectif est d'identifier les variables influentes et leurs seuils, puis de rédiger une règle métier à intégrer dans le système expert utilisé actuellement pour le diagnostic à distance.

2 Contexte FTTH

2.1. FTTH

Cette étude se consacre à la recherche d'incidents sur les accès internet de clients raccordés sur un réseau d'accès fibre optique. Il s'agit de la partie « terminale » d'un réseau télécom entre le « central », où se trouve le dernier équipement actif, et le logement. Ces réseaux d'accès, historiquement en cuivre, ayant permis des accès aux offres internet, voix et TV par ADSL et VDSL, sont progressivement remplacés par des réseaux d'accès en fibre optique ; FTTH, car la fibre est alors installée jusque dans le logement. L'intérêt majeur de ces accès en fibre est de rendre la performance du système indépendante de la distance au central et d'apporter du très haut débit aux clients. Le déploiement de cette technologie a une forte dynamique dans le monde.

Afin de diagnostiquer les clients, différents équipements sont interrogés pour constituer le puits de données

nécessaire. La figure 1 montre la chaîne technique mise en œuvre pour les accès FTTH. On notera en particulier l'OLT (OLT au central pour Optical Line Termination localisé au « central téléphonique ») qui génère le signal fibre et l'ONT (Optical Network Termination) directement raccordé à la fibre d'une part et à la passerelle résidentielle (box) d'autre part.

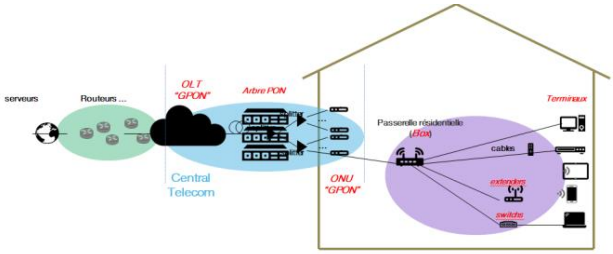


Figure 1 : Schéma d'un réseau d'accès FTTH, des équipements opérateur et des équipements client

2.2. Instabilité d'équipements

En FTTH, il existe peu de perturbations contrairement à une ligne cuivre qui peut subir celles liées à l'environnement (perturbations électromagnétiques par exemple). On a souvent tendance à penser qu'il n'existe pas d'instabilité pour un client FTTH : soit il fonctionne très bien, soit il ne fonctionne pas. Cependant, ce postulat ne se vérifie pas toujours sur le terrain et on peut observer des instabilités sur quelques équipements de la chaîne technique délivrant le service internet au client. Une analyse des données des équipements a mis en évidence que certains accès fibre subissaient des anomalies sous la forme d'instabilités. Une instabilité est un changement d'état répété plusieurs fois dans la même journée : ce changement d'état peut se manifester pour un client FTTH par des déconnexions des sessions internet intempestives tout au long de la journée. Ces instabilités au niveau des équipements peuvent se traduire par des impacts perçus par le client sur l'accès aux services (TV, internet, téléphonie).

3 Données

Pour traiter des instabilités, il est nécessaire d'utiliser l'historique de l'état de la ligne du client les jours précédant le dysfonctionnement. De plus, afin de qualifier le ressenti du client, les signalisations (appels au service client, tchat avec un conseiller Orange, etc.), les remontées terrains des techniciens en interventions ainsi que les résultats des tests effectués avec le système expert de diagnostic distant ont été pris en compte.

3.1. Type de données

Le jeu de données créé pour ce sujet est composé de trois parties : la qualité du réseau d'accès FTTH, la qualité de l'expérience TV et internet, et le ressenti du client.

La première partie est constituée des données suivantes : les puissances optiques, les erreurs de transmissions, des informations sur les équipements Orange (OLT, ONT, Box, Décodeur TV) tels que les constructeurs, les modèles, les types, les versions logicielles déployées, etc., ainsi que tous les logs d'alarmes OLT et ONT terminées [7].

Pour refléter l'expérience du client, nous utilisons des données telles que : les informations sur la qualité du service TV (erreurs de transmissions, notes de qualité de l'expérience, etc.), et des informations de qualité de connexions internet (durée de la session, nombre d'erreurs, nombre de déconnexions).

Enfin, les données telles que les signalisations clients (appels effectués au service client, tchat avec un conseiller Orange, etc.), les remontées terrains des techniciens en interventions et également les résultats des tests effectués avec notre système expert permettent d'obtenir le ressenti du client face à l'état de sa ligne. Ces informations sont récupérées et agrégées à la journée.

3.2. Agrégats créés

Du fait de cette forte volumétrie et cette grande variété de données, il a été décidé de créer des agrégats afin de faciliter l'analyse, mais aussi de permettre d'avoir rapidement une vue d'ensemble sur la qualité du réseau et de l'expérience client sur une journée pour chaque ligne FTTH Orange.

Les agrégats ayant apporté le plus de visibilité sont ceux créés à partir des logs d'alarmes, c'est-à-dire des rapports de défauts transmis par les équipements.

En effet, un log d'alarme est composé d'une date et heure de début, d'une date et heure de fin et d'une information indiquant le type d'alarme. Ainsi, si le client fait l'objet de 300 alarmes dans la journée, 300 logs auront donc été générés.

Pour rendre l'analyse de ces alarmes plus simple, plusieurs agrégats pour chaque client, chaque jour et chaque type d'alarme (déconnexion, dégradation ou coupure d'alimentation) ont été créés. Ces agrégats permettent ainsi d'avoir immédiatement la signature des alarmes auxquels nous faisons face : alarmes courtes mais répétées, alarmes longues mais peu nombreuses, alarmes ponctuelles, alarmes régulières, etc.

4 Démarche

L'utilisation de techniques de Machine Learning semble incontournable a priori pour ce sujet compte tenu de la volumétrie et de l'enjeu d'identification des cas instables.

Le sujet peut être vu comme une classification binaire supervisée avec pour cible l'instabilité : le client est instable ou stable sur une journée.

La complexité de ce sujet repose principalement sur le manque de fiabilité de la cible choisie. En effet, la présence d'instabilité au niveau des équipements ne traduit pas forcément un impact service perçu par le client. Soit parce qu'il n'utilisait pas le service au moment où des défauts

sont apparus, soit parce que des mécanismes réseau ou services ont masqué le défaut en assurant une continuité de service. En effet, seul le client est en mesure de nous indiquer si ses services ont été impactés.

De plus, les données sont déséquilibrées (majorité de clients stables contre une minorité de clients instables). Il est indispensable d'en tenir compte dans le choix des modèles de Machine Learning.

4.1. Première approche

Une première approche a été réalisée dans ce sens. La cible était construite à partir des signalisations clients : le client était considéré insatisfait s'il avait effectué au moins un appel en lien avec un problème technique au service client, un tchat avec un conseiller ou un test réalisé depuis l'une des applications expertes de support client et que l'OLT raccordé à sa ligne remontait des alarmes.

Avec une cible construite de cette façon, le modèle de Machine Learning apprend que certaines lignes instables et avérées comme telles par un expert métier, sont stables – e.g. un client constatant un impact sur ces services mais ne se manifestant pas ou un client absent ne constatant donc pas le défaut. Cela fausse complètement les résultats. Le problème d'une cible dépendant trop de l'humain ayant été mis en évidence, nous avons procédé à une autre approche.

4.2. Deuxième approche

Cette nouvelle approche consiste à choisir une cible créée à partir d'un indicateur technique révélateur de la satisfaction du client et mesurant un impact de déconnexion sur un de ses services.

Après une analyse des différents indicateurs dont nous disposons, nous avons pu mettre en évidence, grâce à des techniques de Machine Learning, le fait que la durée de la session internet et le nombre de déconnexions de celle-ci sont des indicateurs traduisant l'insatisfaction du client.

Le jeu de données reste similaire à celui utilisé lors de la première approche.

La nouvelle cible est donc fiable et indépendante de la présence du client. L'idée étant de détecter le dysfonctionnement même en son absence. Cela nous permet également de bénéficier d'un jeu de données plus volumineux.

Après analyse du modèle de Machine Learning obtenu, il apparait que les résultats ont peu de sens d'un point de vue diagnostic. En effet, l'identification d'un seuil franc pour chaque variable influente n'était pas possible avec cette approche. Cela est dû au comportement et à l'impact différent des alarmes.

4.3. Troisième approche

Nous décidons de mener une nouvelle approche en séparant les types d'alarmes par gravité : déconnexion, dégradation de la ligne et rupture d'alimentation d'équipements client. Nous créons ainsi 3 jeux de données.

Nous faisons ensuite apprendre les mêmes algorithmes de Machine Learning qu'utilisés pour les approches précédentes sur chacun des jeux. Ainsi, cette démarche a permis de mettre en évidence des comportements différents suivants les types de gravités et de faire apparaitre pour chaque cas les variables influentes. Cette approche a permis de révéler des causes racines de nature différentes pour chaque type de gravité validant cette troisième approche.

La suite de cet article est dédiée au traitement des instabilités liées à des alarmes de type déconnexions car elles nous sont apparues plus impactantes d'un point de vue métier.

5 Machine Learning

Plusieurs approches ont été réalisées avec des modèles différents de Machine Learning.

Le jeu de données est constitué d'un mois de données composé d'agrégats calculés par jour et par client. Grâce à la volumétrie importante, nous avons pu réaliser un équilibre en sélectionnant aléatoirement un taux de clients stables légèrement supérieur au taux de clients instables (52% de clients stables contre 48% instables). Nous vous présentons ci-après les résultats obtenus sur un jeu de tests correspondant à 15 jours de données (hors période d'apprentissage).

5.1. Khiops

La première étude a consisté en l'utilisation de l'outil Khiops. Il s'agit d'un outil de Machine Learning développé par Orange et basé sur des réseaux bayésiens [1], [2], [3]. Lors du paramétrage de Khiops, il est possible de choisir différents modèles à tester. Ici, deux modèles sont choisis :

- Selective Naive Bayes
- MAP (Maximum A Posteriori) Naive Bayes

Le Selective Naives Bayes obtient une AUC (Area Under the Curve), qui correspond à l'aire sous la courbe de ROC (Receiver Operating Characteristic), de 0,8904.

L'accuracy, proportion de prédictions correctes (Stable et Instable), est de 0,8237. La précision et le rappel sont à calculer car non fournis par Khiops. Ainsi, on obtient :

$$Précision = \frac{TP}{TP + FP} \qquad Rappel = \frac{TP}{TP + FN}$$

La précision, proportion de prédictions positives qui était effectivement positive dans le test, est ici de 0,7823.

Le rappel, proportion d'actuels positifs trouvés par le modèle, est de 0,8437.

Enfin, la matrice de confusion montre bien que les deux classes sont prédites correctement : False représentant les cas « Stable » et True les cas « Instable ». Les cas observés sont précédés de « % » et ceux prédits de « $ ».

Cible	%False	%True
$False	80,72	19,28
$True	15,63	84,37

Pour le MAP Naive Bayes, les résultats sont similaires :

AUC	Accuracy	Précision	Rappel
0,8835	0,8216	0,7725	0,8471

De même que pour le Selective Naive Bayes, la matrice de confusion montre que les deux classes sont bien prédites.

Cible	%False	%True
$False	80,12	19,88
$True	15,28	84,72

Le Selective Naives Bayes obtient des résultats très légèrement meilleurs que celui du MAP. Les performances des modèles peuvent se différencier d'un point de vue métier. En effet, les variables influentes identifiées par le Selective Naive Bayes ont plus de sens que celles sélectionnées par le MAP Naive Bayes.

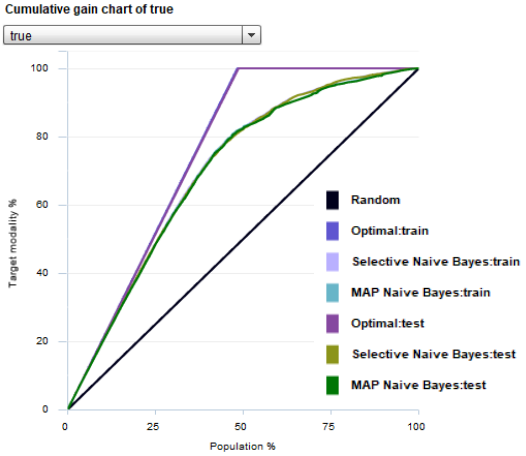


Figure 2 : Evaluation des performances Khiops

5.2. Forêts aléatoires (Random Forest)

Les Random Forest sont un des algorithmes de classification supervisée les plus puissants [4].

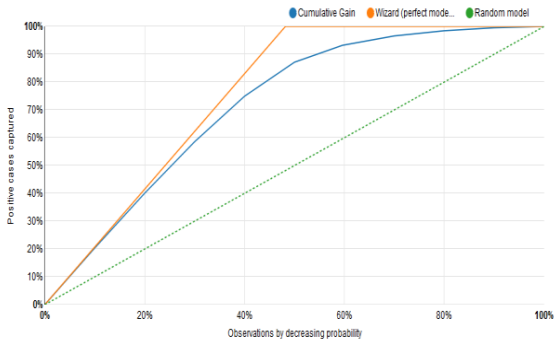


Figure 3 : Evaluation des performances Random Forest

Pour ce modèle, les métriques sont les suivantes :

AUC	Accuracy	Précision	Rappel
0,9257	0,8593	0,8466	0,8649

Le modèle ici a donc fourni des performances tout à fait intéressantes et les résultats obtenus sont également cohérents d'un point de vue métier.

Cible	%False	%True
$False	85,42	14,58
$True	13,51	86,49

5.3. Régression logistique

La régression logistique est un modèle de classification linéaire qui permet de générer une probabilité comprise entre 0 et 1 [6].

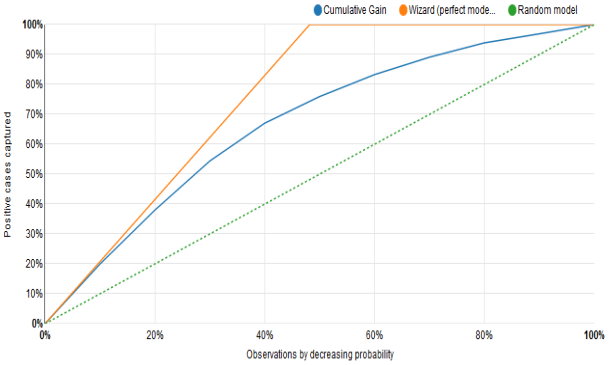


Figure 4 : Evaluation des performances de la régression logistique

Les métriques pour le modèle sont les suivantes :

AUC	Accuracy	Précision	Rappel
0,8286	0,7422	0,7061	0,7966

Le modèle présenté ici est donc moins bon que le modèle précédent. Les résultats restent néanmoins tout à fait corrects autant d'un point de vue Machine Learning que d'un point de vue métier.

Cible	%False	%True
$False	69,16	30,84
$True	20,34	79,66

5.4. Arbres de décision

Les arbres de décision sont une technique qui peut être utilisée à la fois dans le cadre d'une classification mais aussi pour une régression. Ce type de technique, basée sur du « IF », « THEN », « ELSE » est très pratique pour mettre en évidence le lien entre des variables et donc particulièrement adapté à notre problématique [5].

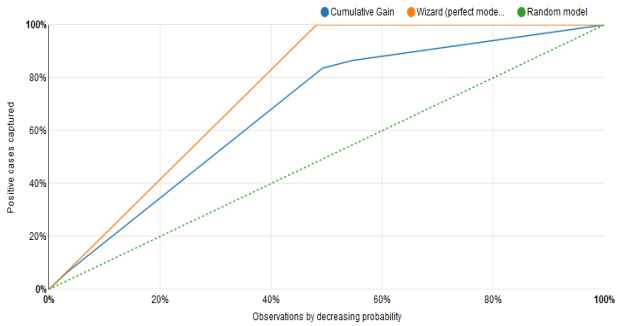


Figure 5 : Evaluation des performances du modèle d'arbres de décision

Les métriques pour le modèle d'arbres de décision sont les suivantes :

AUC	Accuracy	Précision	Rappel
0,8751	0,8313	0,8178	0,8362

Les résultats fournis par ce modèle sont meilleurs que ceux fournis par la régression logistique étudiée précédemment, mais moins bons que ceux obtenus par le Random Forest.

Cible	%False	%True
$False	82,67	17,33
$True	16,38	83,62

5.5. Choix de l'algorithme

Au vue des performances, le choix des modèles pour la poursuite de l'étude s'est porté sur le Random Forest et le Selective Naives Bayes de Khiops. Les variables influentes choisies par ces deux modèles sont semblables. Elles ont un sens métier ce qui n'était pas autant le cas pour les trois autres algorithmes de Machine Learning testés. En effet, les autres modèles mettaient en évidence des variables qui ne permettaient pas de conclure d'un point de vue métier.

6 Application métier : règle de diagnostic

Les résultats de Machine Learning nous ont permis d'établir une règle de diagnostic, à partir des alarmes de déconnexion, capable de détecter les cas d'instabilités ayant un impact sur les services client. Les algorithmes de Machine Learning ont mis en évidence des seuils francs pour les variables influentes. Ces seuils ont ensuite été traduits sous la forme suivante :

$$\text{SI variable}_{\text{influente}} > seuil_{déterminé}$$
$$ET\ autres_{paramètres}\ ALORS\ situation_{instable}$$

Nous avons validé ces résultats en analysant les échanges avec les clients. Une fois la règle validée, nous avons pu analyser les comptes rendus d'appels SAV et ceux des techniciens d'interventions afin de déterminer la cause racine de ce problème d'instabilité. Nous avons ainsi pu préconiser un processus de réparation. Cette règle est actuellement en cours de déploiement dans le système de diagnostic opéré par Orange et sera opérationnelle d'ici la fin de l'été 2020.

7 Conclusion

L'étude présentée a montré l'intérêt du Machine Learning dans le cadre de l'amélioration du diagnostic et la réparation de réseaux d'accès de type FTTH, de par nature très stables. L'utilisation de données historiques et agrégées a permis de mettre en évidence des problématiques non traitées jusqu'à présent puisqu'indétectables avec une vue statique.

Cette approche a produit une analyse sur un volume très important de données de natures différentes, de classifier les clients selon un critère technique représentatif des perturbations subies au niveau des services et de créer une règle de diagnostic. Ainsi, nous mettons en visibilité des problèmes d'instabilités chez les clients sans qu'ils ne contactent le SAV. Cette combinaison, règle et préconisation, est en cours de déploiement dans le système de diagnostic à destination des intervenants et des clients d'Orange afin de donner les conclusions précises et fiables lorsque qu'un SAV s'avère nécessaire. La règle sera opérationnelle d'ici la fin de l'été 2020. A la suite de cela, nous pourrons mesurer l'impact sur les appels au SAV et sur le taux d'interventions et ainsi mesurer le bénéfice obtenu.

Remerciements

Les auteurs tiennent à remercier Joachim Flocon-Cholet et Frédéric Neddam pour leur soutien au cours de cette étude.

Références

[1] M. Boullé, *MODL: A Bayes optimal discretization method for continuous attributes*, Springer, vol. 65, no. 1, pp. 131-165, 2006

[2] M. Boullé, *Compression-based averaging of selective naive Bayes classifiers*, Journal of Machine Learning Research, vol. 8, no. Jul, pp. 1659-1685, 2007

[3] M. Boullé, *Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données*, EGC, pp.229-230, 2008

[4] L. Breiman, *Random Forests*, Springer, vol. 45, no. 1, pp. 5-32, October 2001.

[5] M. Chen, A. X. Zheng, J. Lloyd, M. I. Jordan, E. Brewer, *Failure diagnosis using decision trees*, Proceedings of International Conference on Autonomic Computing, pp.36-46, 2004

[6] J. M. Hilbe, *Logistic Regression*, www.encyclopediaofmath.org

[7] *G.984.3 : Réseaux optiques passifs gigabitaires : Transmission convergence layer specification*, Union Internationale des Telecommunications : série G, 2014

MLP4NIDS : Application pratique d'un réseau de neurones pour la détection d'intrusions réseau dans les voitures connectées

A. Rosay[1 2] F. Carlier[1] P. Leroux[1]

[1] Centre de Recherche en Éducation de Nantes (CREN), Le Mans Université, Le Mans, France
[2] STMicroelectronics, Le Mans, France

arnaud.rosay@st.com, {arnaud.rosay, florent.carlier, pascal.leroux}@univ-lemans.fr

Résumé

La connexion Internet devient omniprésente dans les systèmes embarqués, entraînant des victimes potentielles d'intrusion. Bien qu'ayant gagné en popularité ces dernières années, les algorithmes d'apprentissage approfondi ont tendance à produire de moins bons résultats que l'apprentissage automatique traditionnel. Dans cet article, nous proposons une méthodologie basée sur le réseau de neurones à action anticipée pour la détection des intrusions. De meilleures performances que les techniques traditionnelles d'apprentissage machine peuvent être obtenues lorsque toutes les étapes de la méthodologie sont appliquées. Les performances sont évaluées sur CICIDS2017, montrant une précision supérieure à 99% et un taux de faux positifs inférieur à 0,5%. Après analyse des études précédentes, leurs résultats sont comparés aux performances de l'approche proposée. Enfin, le réseau de neurones a été porté sur un processeur automobile pour caractériser des aspects de performance et valider les résultats.

Mots Clef

Apprentissage automatique, perceptron multi-couche, détection d'intrusion réseau, corpus CICIDS2017.

Abstract

The Internet connection is becoming ubiquitous in embedded systems, making them potential victims of intrusion. While in the age of deep learning, these algorithms tend to produce worse results than traditional machine learning. In this paper, we propose a methodology based on feed-forward neural network for intrusion detection. Better performances than traditional machine learning techniques can be achieved when all steps of the methodology are applied. Performance is evaluated on CICIDS2017, showing accuracy better than 99% and a false positive rate lower than 0.5%. After analysis of previous studies, their results are compared to the performance of the proposed approach. Finally, the neural network trained on a PC has been implemented on an automotive processor to characterize performance aspects.

Keywords

Machine Learning, Multi-Layer Perceptron, Network Intrusion Detection, CICIDS2017 data set.

1 Introduction

Ces dernières années, l'Internet des Objets (IoT) s'est développé dans tous les domaines. Un véhicule est un type spécifique d'IoT qui intègre un modem cellulaire pour les appels d'urgence et devient ainsi une voiture connectée. Depuis le 1er avril 2018, il est obligatoire en Europe pour toutes les voitures particulières et les véhicules utilitaires légers [1]. La présence d'un modem permet l'émergence de nouveaux services nécessitant le transfert de données par cette connexion cellulaire. Dans un tel contexte, les voitures deviennent des victimes potentielles de pirates informatiques capables de lancer des attaques à distance sur l'ensemble d'un parc de véhicules. En réponse à ces risques, une première approche consiste à utiliser des outils de systèmes de détection des intrusions sur le réseau (IDS) tels que Snort [2] pour analyser le trafic réseau et extraire des signatures qui peuvent être comparées aux signatures d'attaques connues. Ce type d'approche présente deux limites : d'une part, la signature d'une attaque doit déjà être connue et, d'autre part, elle génère un taux élevé de faux positifs [3]. Une autre approche consiste à utiliser des techniques d'apprentissage automatique (ML). Ces techniques comprennent l'apprentissage supervisé pour classer les flux de réseau dans différentes catégories et l'apprentissage non supervisé permettant la détection d'anomalies.

L'apprentissage supervisé n'est possible que si l'on dispose d'un corpus numérique, en l'occurrence l'enregistrement de trames de réseau contenant du trafic normal et des attaques. Dans le secteur automobile, il n'existe pas de tels corpus publics. En effet, les constructeurs automobiles conservent soigneusement leurs informations et ne souhaitent pas prendre de risques en divulguant des données qui ne seraient pas ou peu anonymées. En se concentrant sur les attaques via une connexion cellulaire, l'intrusion sur le réseau automobile peut être considérée comme similaire à des attaques menées sur un réseau infor-

matique conventionnel. Il est donc possible d'utiliser des corpus communs de détection d'intrusion réseau comme KDD-Cup99 [4], NSL-KDD [5] ou CICIDS2017 [6] qui contiennent différents types d'attaques contre des ordinateurs ou des serveurs. Les méthodes d'apprentissage supervisé peuvent être divisées en deux groupes : les algorithmes d'apprentissage automatique traditionnels tels que les arbres de décision (DT) ou les forêts aléatoires (RF) et les algorithmes basés sur les réseaux neuronaux tels que le perceptron multi-couche (MLP) ou le réseau neuronal convolutif (CNN). Dans le cadre des recherches sur la détection d'intrusion réseau, de nombreux articles couvrent les approches traditionnelles, tandis qu'un nombre plus restreint d'articles se concentrent sur les techniques de réseaux neuronaux.

Nos principales contributions se composent de trois parties. Tout d'abord, nous proposons une approche basée sur le perceptron multi-couche exercé sur un corpus numérique récent CICIDS2017 contenant des attaques et des trafics représentatifs de l'utilisation actuelle des réseaux informatiques. Toutes les étapes sont détaillées, de l'analyse du corpus à la mise au point du réseau de neurones. Nous analysons ensuite les travaux antérieurs sur ce même corpus et comparons leurs performances avec nos résultats expérimentaux. Enfin, nous complétons le travail sur le réseau neuronal par une implémentation sur un microprocesseur (MPU) de STMicroelectronics. Ce MPU est généralement utilisé dans les unités de contrôle électronique automobile pour des cas d'usage télématique tel que décrit dans [7]

Ce document présente dans la section 2 les travaux antérieurs. La méthodologie que nous proposons est décrite dans la section 3. La section 4 présente les résultats de notre approche en utilisant plusieurs métriques et les compare aux travaux précédents. Le temps d'exécution sur un processeur automobile est fourni dans la section 4.3. Enfin, la section 5 conclut ce document et identifie des idées pour des travaux futurs.

2 Travaux antérieurs

Le problème de détection des intrusions réseau peut être décrit mathématiquement [8] et a été abordé dans plusieurs études avec diverses méthodes sur de multiples corpus numériques. Dhanabal et Shantharajah [9] ont analysé le contenu du corpus NSL-KDD et ont étudié plusieurs classificateurs issus des techniques traditionnelles d'apprentissage automatique. Ils ont obtenu une assez bonne détection des intrusions avec les algorithmes SVM [10] et arbre de décision J48, une implémentation Java du C4.5 [11]. Tang et al. [12] ont utilisé une approche d'apprentissage profond sur le même jeu de données. Ils ont proposé un petit réseau de neurones avec un nombre très limité de caractéristiques en entrée et ont obtenu des performances inférieures.

Sharafaldin et al. [6] ont accompagné la publication de leur corpus CICIDS2017 de performances de détection d'intrusion. Un écart significatif est observé sur la précision, le rappel et la mesure f_1- entre les algorithmes traditionnels de ML et les MLP. K-NN et l'analyse discriminante quadratique (QDA) ont surpassé le MLP de près de 20%. La sélection des caractéristiques, le prétraitement des données et les détails sur les classificateurs ne sont pas décrits et les résultats ne peuvent donc pas être reproduits. Dans leur article, Jiang et al. [13] se sont concentrés uniquement sur 4 classes correspondant à des attaques par déni de service parmi les 14 classes du corpus CICIDS2017. Ils ont proposé de nouvelles caractéristiques en entrée d'un MLP et ont comparé le résultat avec l'utilisation de toutes les caractéristiques fournies dans CICIDS2017.

Un modèle à deux niveaux a été proposé par Ullah et Mahmoud [14]. Le premier niveau classifie le trafic comme normal ou comme attaque grâce à un arbre de décision et le second niveau identifie le type d'attaque avec une forêt aléatoire après augmentation des données basée sur les techniques de sur-échantillonnage synthétique des minorités [15] et l'édition des plus proches voisins [16]. La méthode a été testée sur deux corpus : CICIDS2017 et UNSW-NB15. Ustebay et al. [17] ont proposé une approche en deux étapes pour le corpus CICIDS2017. L'élimination récursive des caractéristiques (RFE) basée sur la méthode RF est utilisée pour identifier les caractéristiques les plus utiles qui sont ensuite injectées dans un réseau neuronal.

Tous ces travaux obtiennent soit des performances inférieures avec le réseau neuronal, soit, lorsqu'ils obtiennent de bons résultats, ne prennent pas en compte tous les types d'attaques. Nous proposons une méthodologie détaillée pour obtenir des performances élevées avec un réseau neuronal sur toutes les attaques disponibles dans l'ensemble des données.

3 Méthodologie et solution à perceptron multi-couche

Après la sélection d'un corpus numérique, nous proposons une approche consistant à former un réseau neuronal perceptron multi-couche afin de quantifier les avantages et les limites potentielles de l'utilisation de l'apprentissage profond dans l'IDS.

3.1 Choix du corpus numérique

KDD-Cup99 est un corpus ancien publié en 1999. Un premier examen de celui-ci a été effectué par McHugh [18]. Tavallaee et al. [5] en ont fourni une analyse détaillée et en ont dérivé un nouveau corpus appelé NSL-KDD pour remédier à certaines des lacunes décrites par McHugh. NSL-KDD est largement utilisé pour les IDS depuis 2009. Bien que les attaques aient évolué dans le temps, KDD-Cup99 et NSL-KDD restent un sujet d'étude [19] [20] [21]. Pourtant les deux souffrent de l'absence de certains protocoles importants comme HTTPS qui représente aujourd'hui plus de 70% du trafic réseau.

En 2015, Moustafa et Slay ont proposé un nouveau corpus appelé UNSW-NB15 [22] qui a été généré par simulation et représente 31 heures de trafic avec un total de 2 540 044 instances. Il contient des attaques modernes qui sont re-

groupées en neuf familles d'attaques différentes. UNSW-NB15 comprend 49 caractéristiques.

L'institut canadien de cyber-sécurité de l'Université du Nouveau-Brunswick a publié le corpus CICIDS2017 [23] selon le cadre défini dans [6] dans le but de résoudre les lacunes des jeux de données existants. Les données brutes sous forme de fichiers PCAP sont fournies avec un ensemble de 84 caractéristiques dans des fichiers CSV. Le trafic réseau a été enregistré sur 5 jours, ce qui a donné un total de 2 830 743 cas. Comme les corpus mentionnés précédemment, chaque instance correspond à un flux et contient des informations relatives à un groupe de paquets entre une adresse/un port source donnés et une adresse/un port destination donnés. En raison de leurs lacunes connues, KDDcup99 et NSL-KDD n'ont pas été considérés comme un choix pertinent. Nous observons que CICIDS2017 est plus récent, comprend plus de caractéristiques et contient plus d'exemples que UNSW-NB15. La date de création et la quantité de données sont des éléments clés. Tout d'abord, les données les plus anciennes ne sont plus représentatives du trafic réseau et des attaques actuelles. Deuxièmement, un plus grand nombre de données permet un meilleur apprentissage. Pour ces raisons, le corpus numérique CICIDS2017 est sélectionné pour l'évaluation de notre solution de détection d'intrusion.

3.2 Préparation des données

Nettoyage du corpus. Une série d'opérations a été menée pour détecter la présence de lignes vides, de caractéristiques redondantes et de valeurs non numériques dans des champs numériques. Ce contrôle systématique a permis de supprimer plus de 280 000 cas où toutes les caractéristiques étaient vides. La longueur de l'entête des messages émis apparaissant deux fois, une instance a été supprimée. La plupart des caractéristiques sont numériques mais certains de ces champs peuvent contenir des chaînes de caractères comme "NaN" ou "Infinity". Ces cas apparaissent dans 6 types d'attaques : BENIGN, FTP-PATATOR, DoS Hulk, Bot, PortScan et DDoS. Comme la quantité de ces cas est négligeable dans chaque type de trafic, ces instances ont simplement été supprimées.

Création des jeux de données d'entraînement, de validation et de test. CICIDS2017 est un corpus très déséquilibré à double titre. D'une part, le trafic normal (BENIGN) représente 80% du trafic total et d'autre part, certaines attaques (Heartbleed, Infiltration, WebAttack SQL injection) sont très largement sous-représentées. Sachant qu'un corpus déséquilibré est un défi pour l'apprentissage supervisé, nous avons préparé des jeux d'entraînement, de validation croisée et de test en prenant de manière aléatoire respectivement 50%, 25% et 25% de chaque type d'attaque tout en garantissant que chaque instance n'est utilisée qu'une seule fois. Ensuite, chaque jeu de données a été complété par des instances de trafic normal également choisies aléatoirement sans remise. Ce procédé permet d'équilibrer la quantité de trafic normal et d'attaques mais reste déséquilibré en termes de type d'attaque. Seuls les jeux d'entraînement et de validation croisée servent à l'apprentissage des poids du réseau et au réglage des hyperparamètres. Les mesures de performances ont été effectuées avec les données du jeu de test qui n'ont jamais vues par le réseau durant l'apprentissage.

Sélection des caractéristiques. La sélection des caractéristiques parmi les 84 proposées est l'une des étapes fondamentales de l'apprentissage automatique influençant grandement les performances du modèle. Des caractéristiques non pertinentes peuvent avoir un impact négatif et entraîner une diminution de la précision des prédictions. L'analyse du corpus numérique a révélé 8 caractéristiques qui ne sont pas informatives. Leur valeur est constante quel que soit le type de trafic. Par conséquent, les caractéristiques "Bwd PSH Flags", "Bwd URG Flags", "Fwd Avg Bytes/Bulk", "Fwd Avg Packets/Bulk", "Fwd Avg Bulk Rate", "Bwd Avg Bytes/Bulk", "Bwd Avg Packets/Bulk", "Bwd Avg Bulk Rate" ont été supprimées. Afin que le modèle n'apprenne pas quand les attaques se produisent, la caractéristique "Timestamp" n'est pas considérée comme informative pour notre étude. Chaque instance est caractérisée par son port et son adresse IP d'origine et de destination, son protocole et un identifiant de flux contenant les mêmes informations. Comme la caractéristique "FlowID" est redondante avec d'autres champs, elle a été retirée du jeu de données. Nous avons également décidé de retirer l'adresse IP source/destination et le port source parce que ces caractéristiques ne sont pas pertinentes dans un système générique de détection des intrusions. Il en résulte un corpus numérique contenant 70 caractéristiques.

Standardisation. Le pré-traitement des données est essentiel pour préparer l'ensemble des données en vue d'un apprentissage efficace. En particulier, un réseau neuronal apprend mieux lorsque toutes les caractéristiques sont mises à l'échelle dans la même plage de valeurs. Ceci est utile lorsque les entrées sont à des échelles très différentes comme c'est le cas avec le corpus CICIDS2017. Dans notre implémentation, les données sont standardisées car c'est le pré-traitement qui donne le meilleur résultat sur notre corpus. Pour chaque caractéristique $\mathcal{F}_j$, chaque valeur x_i est transformée selon l'équation (1) où $\mu^{(\mathcal{F}_j)}$ et $\sigma^{(\mathcal{F}_j)}$ sont respectivement la moyenne et l'écart-type de la caractéristique $\mathcal{F}_j$.

$$x_i^{(\mathcal{F}_j)} = \frac{x_i^{(\mathcal{F}_j)} - \mu^{(\mathcal{F}_j)}}{\sigma^{(\mathcal{F}_j)}} \tag{1}$$

3.3 Création du modèle

Description du modèle. Le perceptron multi-couche est un réseau de neurones à propagation avant pouvant être utilisé comme classificateur [24]. Le modèle proposé prend en entrée les valeurs standardisées des caractéristiques sélectionnées. Il contient deux couches cachées de 256 nœuds. Le choix du nombre de couches et de nœuds par couche résulte d'une analyse de performance et d'un compromis

minimisant le total des erreurs dues au biais et à la variance. En effet, un modèle trop simple n'arrive pas à capturer la relation entre les données d'entrée et la sortie (erreurs importantes à la fois sur les données d'entraînement et de validation croisée) alors qu'un modèle trop complexe engendre un sur-ajustement (erreurs faibles sur les données d'entraînement et importantes sur celles de validation croisée). Bien qu'un modèle à une seule couche cachée soit un approximateur universel [25], il est généralement plus efficace d'augmenter le nombre de couches plutôt que d'augmenter le nombre de nœuds d'une couche unique. Compte tenu du temps d'entraînement nécessaire au bon dimensionnement du réseau et à l'analyse du compromis biais-variance, l'espace de recherche a été contraint à des nombres de nœuds valeurs multiples de 16. Ce choix est lié à la plateforme de déploiement visée pour laquelle chaque ligne de cache du processeur en charge des inférences peut contenir 16 valeurs flottantes de 32 bits. Le classificateur possède 15 sorties pour les 14 types d'attaque et le trafic bénin. La technique du *dropout* est utilisée sur les couches cachées pour éviter le sur-ajustement aux données d'apprentissage en éteignant des nœuds aléatoirement avec une probabilité $drop\_rate$.

La sortie du réseau de neurones $\boldsymbol{h}^{(3)}$ est calculée en chaînant les sorties des différentes couches comme décrit dans les équations (2), (3) et (4) dans lesquelles $\boldsymbol{x}$, $\boldsymbol{W}^{(i)}$ et $\boldsymbol{b}^{(i)}$ sont respectivement le vecteur d'entrée contenant les caractéristiques sélectionnées, la matrice de poids et le vecteur de biais pour la couche i.

Compte tenu de la topologie du réseau, on a $\boldsymbol{x} \in \mathbb{R}^{70}$, $\boldsymbol{W}^{(1)} \in \mathbb{R}^{70\times256}$, $\boldsymbol{b}^{(1)} \in \mathbb{R}^{256}$, $\boldsymbol{W}^{(2)} \in \mathbb{R}^{256\times256}$, $\boldsymbol{b}^{(2)} \in \mathbb{R}^{256}$, $\boldsymbol{W}^{(3)} \in \mathbb{R}^{256\times15}$ et $\boldsymbol{b}^{(3)} \in \mathbb{R}^{15}$.

$$\boldsymbol{h}^{(1)} = g_1\left(\boldsymbol{W}^{(1)T}.\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) \tag{2}$$

$$\boldsymbol{h}^{(2)} = g_1\left(\boldsymbol{W}^{(2)T}.\boldsymbol{h}^{(1)} + \boldsymbol{b}^{(2)}\right) \tag{3}$$

$$\boldsymbol{h}^{(3)} = g_2\left(\boldsymbol{W}^{(3)T}.\boldsymbol{h}^{(2)} + \boldsymbol{b}^{(3)}\right)' \tag{4}$$

Les fonctions d'activation g_1 et g_2 apportent de la non-linéarité dans le réseau. Comme le MLP ne fournit pas une normalisation intrinsèque de ses sorties, la fonction d'activation SELU g_1 dont la formule est donnée en (5) est utilisée pour les couches cachées avec les valeurs $\lambda = 1.0507$ et $\alpha = 1.6733$ comme définies dans [26]. Les valeurs de λ et α permettent d'obtenir à la sortie de chaque couche une distribution centrée en 0 avec un écart-type de 1. Quant à la couche de sortie, elle utilise la fonction softmax g_2 définie en (6) afin que chaque sortie puisse être interprétée comme la probabilité de prédire une classe donnée. Le label prédit $\hat{y}$ est obtenu par $\hat{y} = argmax\ \boldsymbol{h}^{(3)}$.

$$g_1(z) = \lambda.\begin{cases}\alpha.(e^z - 1) & \text{for } z < 0\\ z & \text{for } z \geq 0\end{cases} \tag{5}$$

$$g_2\left(z\right)_j = \frac{e^{z_j}}{\sum_{k=1}^{15} e^{z_k}} \text{ for } j \in [1;15] \tag{6}$$

Entraînement. Le modèle a été implémenté en python et en utilisant l'infrastructure logicielle TensorFlow pour la bibliothèque d'apprentissage profond. Les données d'entraînement ont été divisées en mini-lots de 32 instances. La taille du mini-lot est choisie en fonction de la taille du cache de données de la machine servant à entraîner le réseau. 32 instances correspondent à la plus grande puissance de 2 permettant de contenir le mini-lot, les poids et les sorties de la première couche du réseau dans le cache de niveau 1 avec suffisamment de marge (40%) pour limiter les risques d'éviction de nos données pendant les traitements. Le MLP apprend à classifier en ajustant les poids entre les nœuds du réseau afin de réduire une fonction de coût $\mathcal{L}(w)$. Celle-ci est basée sur l'entropie croisée définie par l'équation (7) dans laquelle y est le label correct et $\hat{y}$ la classe prédite. Plusieurs algorithmes permettent l'optimisation de la fonction de coût. [27] définit un algorithme appelé Adam et le compare à plusieurs alternatives. Celui-ci étant généralement plus performant, $\mathcal{L}(w)$ est optimisée avec Adam dont les trois paramètres α, β_1 et β_2 permettent une configuration fine de l'algorithme.

$$\mathcal{L}(w) = -\left[y.\log\left(\hat{y}\right) + \left(1-y\right).\log\left(1-\hat{y}\right)\right] \tag{7}$$

Le niveau de sur-ajustement du réseau est contrôlé en évaluant les performances de prédiction sur le jeu de validation croisée. Le paramètre de *dropout* $drop\_rate$ est affiné pour garder la différence de performance entre les jeux d'entraînement et de validation croisée la plus faible possible. Le réglage des hyper-paramètres est la tâche délicate qui consiste à trouver le bon ensemble de valeurs pour obtenir les meilleures performances. Une stratégie de recherche aléatoire [28] a été appliquée afin de régler $drop\_rate$, α et β_1 tout en gardant la valeur par défaut $\beta_2 = 0.999$.

3.4 Métriques d'évaluation

Plusieurs informations clefs peuvent être extraites en comparant les prédictions aux labels fournis dans le corpus. Le nombre de vrai positif (TP) est le nombre d'attaques correctement prédites comme étant des attaques. Le nombre de vrai négatif est le nombre d'instances normales classées comme trafic normal alors que les faux positifs (FP) et faux négatifs (FN) sont respectivement le nombre d'instances normales classées comme attaques et le nombre d'attaques prédites comme trafic normal. Différentes métriques peuvent être dérivées de l'information contenue dans la matrice de confusion.

Comme les travaux publiés sur le corpus CICIDS2017 utilisent un sous-ensemble de métriques, nous proposons de baser notre évaluation sur un ensemble plus complet de métriques permettant une comparaison avec les études existantes et futures sur ce corpus.

Le taux de vrai positif (TNR), la justesse (Acc), la précision (Prec) et le rappel (Rec) sont définis dans [29]. Le score f_1 est la moyenne harmonique de la précision et du rappel. Le taux de faux positif (FPR) dans l'équation (8)

est le pourcentage de trafic bénin classé comme attaques.

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

Le coefficient de corrélation de Matthews (MCC) est une métrique intéressante prenant en compte tous les éléments de la matrice de confusion de manière correcte dans le cas des corpus déséquilibrés contrairement à la justesse qui peut donner des valeurs élevées même lorsque la totalité de la classe minoritaire est mal prédite. Puisque les corpus de détection d'intrusion sont intrinsèquement déséquilibrés, c'est une métrique judicieuse pour notre application. MCC prend une valeur comprise entre -1 et $+1$. Une prédiction parfaite correspond à $MCC = 1$. A contrario, $MCC = -1$ dénote un désaccord total entre les prédictions et les classes réelles. Une prédiction aléatoire correspond à $MCC = 0$.

4 Résultats expérimentaux

Cette section couvre l'évaluation de performance du modèle et une comparaison aux résultats antérieurs. De plus, la sous-section 4.3 explique comment le modèle a été déployé sur une plate-forme STMicroelectronics (ST) pour l'automobile et fournit une évaluation de performance dans un tel système embarqué.

4.1 Évaluation des performances

Après entraînement d'un modèle de référence, le jeu de test a permis la mesure des performances de notre classificateur sur 15 classes. La matrice de confusion est simplifiée dans la TABLE 1 en regroupant les classes prédites en 3 colonnes : bénin, attaque correcte, autre attaque. Dans le cas d'une classification multi-classes, un type d'attaque peut être prédit comme étant un autre type d'attaque. Dans le cas de la détection d'intrusion, ce type d'erreur n'est pas un problème car il s'agit toujours d'une attaque. Par conséquent, les prédictions pour tous les types d'attaque peuvent être fusionnées en une seule classe pour ainsi obtenir un classificateur binaire et calculer les métriques de la TABLE 2.

La TABLE 1 montre que notre MLP classe correctement pratiquement tous les types de trafic à l'exception des classes "Bot", "Infiltration" et les différents "WebAttack". Ceci est probablement dû au nombre d'instances extrêmement bas pour ces classes. Le MLP n'est pas en mesure d'apprendre des seuls 18 exemples d'infiltration inclus dans le jeu d'entraînement. Cela confirme le fait bien connu que la quantité de donnée est un point clef pour le succès de l'apprentissage automatique.

Les résultats pourraient probablement être améliorés en gardant les adresses IP et les ports. Mais dans ce cas, on peut imaginer que le réseau de neurones apprendrait à reconnaître l'adresse/port de la machine conduisant l'attaque.

TABLE 1 – Matrice de confusion simplifiée.

		Prédictions		
		BENIGN	***Correct attack***	***Other attack***
Actual classes	BENIGN	138,460	-	675
	Bot	184	296	9
	DDoS	38	31,964	4
	DoS GoldenEye	43	2,528	2
	DoS Hulk	106	57,425	0
	DoS Slowhttptest	16	1,342	16
	DoS Slowloris	7	1,431	11
	FTP-PATATOR	6	1,968	9
	Heartbleed	0	2	0
	Infiltration	9	0	0
	PortScan	2	39,675	24
	SSH-PATATOR	16	1,450	8
	WebAttack BruteForce	258	102	16
	WebAttack SQL Injection	2	0	3
	WebAttack XSS	131	4	32

TABLE 2 – Résultats de performance.

Métrique	**Données d'entraînement**	**Données de test**
TNR (%)	99.52	99.51
FPR (%)	0.48	0.49
Rappel (%)	99.40	99.41
Précision (%)	99.52	99.51
Justesse (%)	99.46	99.46
Score f_1 (%)	99.46	99.46
MCC	0.9892	0.9893

4.2 Comparaison avec les résultats antérieurs

La TABLE 3 compare les résultats avec des papiers de référence sur ce corpus numérique. Comme différentes métriques sont utilisées, certaines cellules de la table ne peuvent pas être remplies. Cette comparaison couvre non seulement les solutions à base de réseau de neurones mais aussi celles utilisant les algorithmes traditionnels d'apprentissage automatique.

Notre modèle surpasse tous les résultats reportés par Sharafaldin et al, à la fois pour les réseaux de neurones et les techniques traditionnelles d'apprentissage automatique. Jiang et al. obtiennent des performances similaires à notre solution. Toutefois, il est à noter que leur étude se limite aux attaques de type déni de service (slowloris, slowhttptest, hulk, GoldenEye), soit 4 classes parmi les 14 types d'attaque. Comme le montre la TABLE 1, ces attaques ne sont pas les plus difficiles à détecter. On peut imaginer que les performances de leur solution diminueraient si tous les types d'attaque étaient pris en compte.

L'approche en deux étapes utilisant des arbres de décision proposée par Ullah and Mahmoud [14] donne des "métriques moyennes de 100%". Une analyse plus détaillée révèle que 4 types d'attaques ne sont pas parfaitement dé-

tectées par les arbres de décision. Ces classes sont exactement celles pour lesquelles notre classificateur rencontre des difficultés. Le manque de chiffres significatifs dans [14] ne permet pas une analyse fine mais la comparaison globale montre que notre solution basée sur un MLP atteint le même niveau de performance que les algorithmes traditionnels d'apprentissage automatique.
Ustebay et al. [17] utilisent également un MLP mais n'atteignent pas de performances très élevées. Néanmoins, leurs résultats ne peuvent pas être directement comparés avec [13] et [14] car ces derniers utilisent les adresses IP conduisant les attaques. Dans un scénario réel, les adresses des pirates ne sont pas connues et ne peuvent pas être utilisées pour détecter des intrusions. Lorsque l'adresse IP est supprimée, l'analyse RFE montre que le port d'origine de l'attaque est la caractéristique la plus importante. Ceci prouve que l'usage des adresses et des ports source et destination améliorent la détection d'intrusion même si ce n'est pas réaliste pour une application réelle.

TABLE 3 – Comparaison de performances.

Papier	Algorithme	Acc (%)	Prec (%)	Rec (%)	FPR (%)
our work	MLP	99.46	99.51	99.41	0.49
Sharafaldin et al. [6]	MLP	-	77	83	-
	QDA	-	97	88	-
	K-NN	-	96	96	-
Jiang et al. [13]	MLP	99.23	99.87	99.60	0.77
Ullah and Mahmoud [14]	DT+RF	-	100	100	-
Ustebay et al. [17]	MLP	91	-	-	-

4.3 Déploiement du MLP sur un processeur pour le secteur automobile

La plupart des études sur les IDS se limite à une analyse des performances sur un PC ou un serveur. Pour aller plus loin, notre solution a été déployée sur un microprocesseur dédié à l'industrie automobile. Dans un système embarqué contraint en termes de capacité de calculs et de mémoire, le coût de calculs des réseaux neuronaux peut conduire à préférer l'usage de techniques traditionnelles. Avec le développement du "Edge AI", de plus en plus de circuits électroniques embarquent des accélérateurs matériels pour les calculs d'inférence des réseaux neuronaux. L'utilisation d'un tel accélérateur permettrait de libérer du temps d'occupation du processeur pour adresser d'autres tâches.
Le système sur puce (SoC) Telemaco3P de ST est une solution permettant une connexion entre un véhicule et le nuage. Son architecture multi-cœurs asymétrique fournit des processeurs performants pour les applications ainsi qu'un sous-système indépendant en tant que contrôleur CAN. Sa qualification pour le milieu automobile en fait un bon candidat pour l'implémentation d'une large gamme d'applications télématiques sécurisées supportant une connectivité sans fil à haut débit pouvant nécessiter un système de détection d'intrusion. Un exemple typique d'usage de cette plate-forme est décrit en FIGURE 1. Telemaco3P contient deux cœurs ARM Cortex-A7 à 600MHz avec les extensions NEON et FPU. Ce SoC ne possèdent pas d'accélérateur matériel dédié aux réseaux neuronaux mais permet d'analyser les performances d'un système embarqué existant correspondant au cas d'utilisation visé.

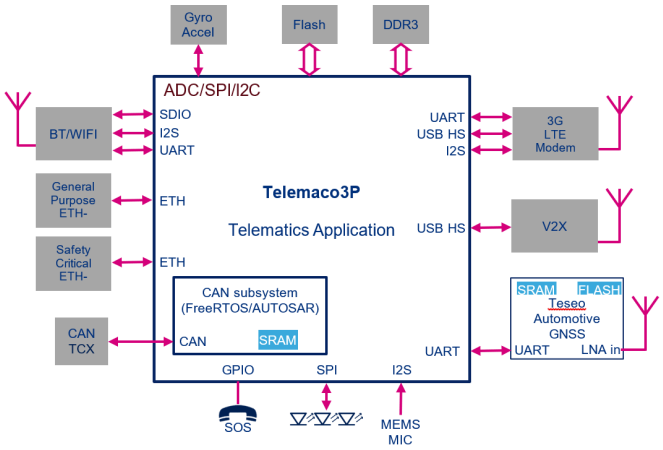
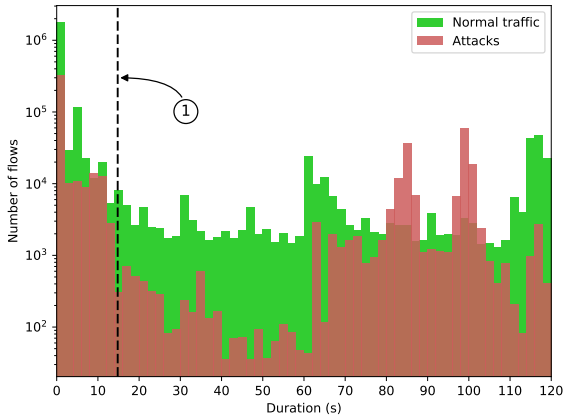


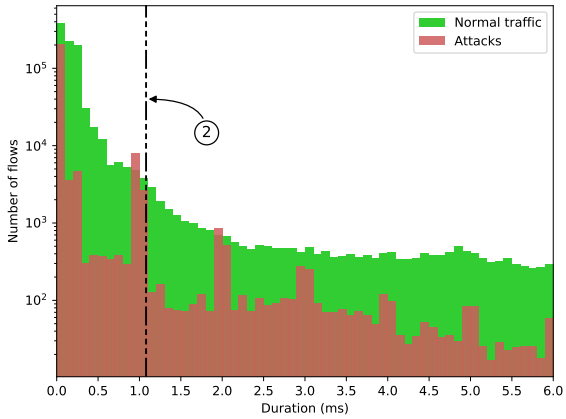
FIGURE 1 – Exemple d'usage de Telemaco3P.

Notre modèle de référence a été entraîné avec TensorFlow-2.0 (TF) [30]. Pour le déploiement, nous avons utilisé TF-Lite, un ensemble logiciel pour l'apprentissage profond dédié à l'inférence sur des dispositifs embarqués. Bien que TF-Lite supporte aujourd'hui un sous-ensemble limité d'opérations de TensorFlow, toutes celles nécessaires à notre modèle sont supportées. Comme le format des nombres flottants de TF-lite est sur 32 bits alors que le modèle expérimental évalué au paragraphe 4.1 utilise une représentation en double précision (64 bits), le modèle déployé a été ré-entraîné en simple précision. La TABLE 4 montre la différence de performance entre le modèle de référence et le modèle déployé pour l'inférence du jeu de test. Les différences sont mineures et dues à la représentation réduite sur 32 bits des nombres flottants.

L'inférence est exécutée dans un processus unique sur un seul cœur Cortex-A7. Le temps moyen de calcul pour une seule inférence est 1.08ms. La FIGURE 2 donne la distribution des durées des flux réseau pour l'ensemble du corpus. On remarque que la majorité des flux sont de très courte durée. Comme indiquée par le marqueur 1 sur la FIGURE 2a, la durée moyenne des flux est de 14.8s alors que la médiane est de 31.3ms. Le marqueur 2 de la FIGURE 2b montre le temps nécessaire à la classification d'une inférence. La durée des flux est plus longue que le temps de calcul dans 60% des cas. Une analyse plus approfondie de la distribution temporelle des flux est nécessaire. Nous formulons l'hypothèse que l'implémentation logicielle courante sur cette plate-forme ne permet pas la classification des flux en temps réel et nécessite de futures investigations. Les améliorations possibles consisteraient à utiliser une quantification sur des entiers 8 bits afin de réduire drastiquement le temps de calcul et de regrouper les



(a) Vue d'ensemble, intervalles de 2s



(b) Flux de moins de 6ms, intervalles de 0.1ms

FIGURE 2 – Distribution de la durée des flux.

flux à classifier afin de tirer parti de la vectorisation. Il faudra également garder des ressources de calculs pour faire le pré-traitement des données avant de les injecter dans le réseau de neurones. Si ces mesures devaient se révéler insuffisantes pour un traitement en temps réel, il faudrait envisager l'utilisation d'un accélérateur matériel.

TABLE 4 – Performance du modèle déployé et du modèle de référence.

Métrique	Modèle déployé	Modèle de référence
Rappel (%)	99.44	99.41
Précision (%)	99.34	99.51
Justesse (%)	99.40	99.46

5 Conclusion

L'approche suivie couvre toutes les étapes de l'analyse du corpus numérique à l'apprentissage et au déploiement dans un système embarqué d'un réseau de neurones à propagation avant. L'exécution de l'ensemble des étapes permet d'atteindre des performances élevées et démontre la nécessité de nettoyer les données, de sélectionner les caractéristiques avant la phase d'apprentissage utilisant des hyperparamètres optimisés. Notre méthode obtient de meilleurs résultats que les implémentations antérieures sur le même corpus.
Le déploiement sur un processeur dédié au domaine automobile démontre que de la détection d'intrusion réseau est possible avec un réseau de neurones malgré des limitations sur les capacités à exécuter cette détection en temps réel. Néanmoins, avec des techniques d'optimisation et la présence croissante d'accélérateur matériel dans les composants électroniques des systèmes embarqués, l'usage de l'apprentissage profond pourrait plus facilement être acceptable dans un futur proche.
La détection d'intrusion sur le corpus CICIDS2017 pourrait être amélioré. Les classes qui sont mal détectées sont sous-représentées et les techniques d'augmentation des données pourraient étendre nos travaux. D'autre part, l'apprentissage supervisé n'est en mesure de détecter que les types d'attaques présentes dans le corpus servant à l'apprentissage. Afin de détecter de nouveaux types d'attaque, il serait nécessaire de passer à des méthodes d'apprentissage non supervisé.

Références

[1] E. Parliament, "Regulation (EU) 2015/758 of the European Parliament and of the Council of 29 April 2015 concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC," *Official Journal of the European Union*, May 2015.

[2] M. Roesch, "Snort - lightweight intrusion detection for networks," in *Proceedings of the 13th USENIX Conference on System Administration*, LISA '99, (USA), p. 229–238, USENIX Association, 1999.

[3] A. Garg and P. Maheshwari, "Performance analysis of Snort-based Intrusion Detection System," in *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 01, pp. 1–5, Jan. 2016.

[4] W. Lee, S. J. Stolfo, and K. W. Mok, "Mining in a data-flow environment : Experience in network intrusion detection," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, (New York, NY, USA), pp. 114–124, ACM, 1999.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, July 2009.

[6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Pro-*

ceedings of the 4th International Conference on Information Systems Security and Privacy - Volume 1 : ICISSP,, pp. 108–116, SciTePress, Jan. 2018.

[7] H. Dakroub, A. Shaout, and A. Awajan, "Connected car architecture and virtualization," *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, vol. 9, 04 2016.

[8] P. Patel, C. Langin, F. Yu, and S. Rahimi, "Network Intrusion Detection Types and Computation," in *International Journal of Computer Science and Information Security*, vol. 10, pp. 14–21, 2012.

[9] L. Dhanabal and D. S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," in *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, pp. 446–452, 2015.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, p. 273–297, Sept. 1995.

[11] J. Quinlan, *C4.5 : Programs for Machine Learning*. Ebrary online, Elsevier Science, 2014.

[12] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for Network Intrusion Detection in Software Defined Networking," in *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 258–263, Oct. 2016.

[13] J. Jiang, Q. Yu, M. Yu, G. Li, J. Chen, K. Liu, C. Liu, and W. Huang, "ALDD : A Hybrid Traffic-User Behavior Detection Method for Application Layer DDoS," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 1565–1569, Aug. 2018.

[14] I. Ullah and Q. H. Mahmoud, "A Two-Level Hybrid Model for Anomalous Activity Detection in IoT Networks," in *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pp. 1–6, Jan. 2019.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic minority oversampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[16] R. Alejo, J. Sotoca, R. Valdovinos, and P. Toribio, "Edited nearest neighbor rule for improving neural networks classifications," in *Advances in Neural Networks - ISNN 2010*, pp. 303–310, Springer Berlin Heidelberg, 2010.

[17] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pp. 71–76, Dec. 2018.

[18] J. McHugh, "Testing Intrusion Detection Systems : A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations As Performed by Lincoln Laboratory," *ACM Trans. Inf. Syst. Secur.*, vol. 3, pp. 262–294, Nov. 2000.

[19] B. Riyaz and S. Ganapathy, "An Intelligent Fuzzy Rule based Feature Selection for Effective Intrusion Detection," in *2018 International Conference on Recent Trends in Advance Computing (ICRTAC)*, pp. 206–211, Sept. 2018.

[20] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, pp. 41–50, Feb. 2018.

[21] E. Zyad, A. Taha, and B. Mohammed, "Improve R2L Attack Detection Using Trimmed PCA," in *2019 International Conference on Advanced Communication Technologies and Networking (CommNet)*, pp. 1–5, Apr. 2019. ISSN :.

[22] N. Moustafa and J. Slay, "UNSW-NB15 : A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, Nov. 2015. ISSN :.

[23] "Intrusion detection evaluation dataset (cicids2017)." http ://205.174.165.80/CICDataset/CIC-IDS-2017/. (Accessed May 11, 2020).

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[25] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251 – 257, 1991.

[26] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems, 2017, pp. 971-980*, 2017.

[27] D. P. Kingma and J. Ba, "Adam : A method for stochastic optimization," *2015, 3rd International Conference for Learning Representations*, 2014.

[28] J. Bergstra and Y. Bengio, "Random search for hyperparameter optimization," *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb. 2012.

[29] P. Maniriho and T. Ahmad, "Analyzing the Performance of Machine Learning Algorithms in Anomaly Network Intrusion Detection Systems," in *2018 4th International Conference on Science and Technology (ICST)*, pp. 1–6, Aug. 2018.

[30] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

Apprentissage de variété pour l'identification de lauréats potentiels de financements pour l'innovation

V. Grollemund[1,2] G. Le Chat[2] J-F. Pradat-Peyre[1,3] F. Delbot[1,3]

[1] LIP6, Sorbonne Université
[2] FRS Consulting
[3] Université Paris Nanterre

vincent.grollemund@lip6.fr

Résumé

finElink est un moteur de recommandation qui oriente les entreprises innovantes dans leur stratégie de financement. La recommandation se base en partie sur l'analyse des données financières d'entreprises précédemment lauréates. Les données financières d'une population représentative d'entreprises sont réduites et projetées dans un espace 2D grâce à l'algorithme d'apprentissage de variété Uniform Manifold Approximation and Projection. La projection des entreprises précédemment lauréates fait ressortir une distribution non uniforme avec une zone où se concentrent les lauréats. Cette zone est identifiée avec une méthode de partitionnement par densité. Les entreprises candidates appartenant à cette zone sont étiquetées comme potentiellement lauréates et le système de recommandation oriente ces entreprises vers les dispositifs existants les plus compétitifs.

Mots Clef

UMAP, DBSCAN, réduction de dimension, apprentissage de variété, partitionnement.

Abstract

finElink is a recommendation system that provides guidance to innovative companies with regard to their financing strategy. Analysis of financial data from former funding recipients partially feeds the recommendation system. Financial company data from a representative French population sample are reduced and projected onto a 2D space with Uniform Manifold Approximation and Projection, a manifold learning algorithm. Former funding recipients' data are projected onto the 2D space. Their distribution is non-uniform, with data concentrating in one region of the projection space. This region is identified using a density-based clustering method. Applicant companies which are projected within this region are labelled potential funding recipients and will be suggested the most competitive funding mechanisms.

Keywords

UMAP, DBSCAN, dimension reduction, manifold learning, clustering.

1 Introduction

Les entreprises innovantes ont un besoin d'orientation dans leur stratégie de financement de par la diversité et la quantité d'information non structurée des dispositifs compétitifs existants. finElink[3] est un système de recommandation qui répond à ce besoin. Conçu au sein de FRS Consulting, il s'est d'abord appuyé seulement sur l'expertise métier des collaborateurs de l'entreprise. Afin d'améliorer la pertinence des recommandations, l'analyse de données financières d'entreprises françaises ayant été lauréates par le passé est étudiée. L'apprentissage automatique des caractéristiques financières de ces dernières peut aider à isoler de nouvelles entreprises candidates à fort potentiel.

Cependant, les données sur les entreprises précédemment lauréates ne peuvent être utilisées seules pour évaluer la pertinence d'une nouvelle entreprise, car ces données ont de fortes contraintes de rareté des données, de biais et de données manquantes. Ces données d'entreprises lauréates sont obtenues en croisant les informations disponibles sur les sites des dispositifs de financement et du registre des entreprises. Ainsi, les données disponibles ne représentent que les lauréats de dispositifs qui communiquent sur l'attribution de financement et les données financières d'entreprises jeunes sont rarement disponibles dans le registre des entreprises. Pour contourner ces problèmes, un apprentissage non supervisé est effectué sur un autre jeu de données d'entreprises représentatif de l'ensemble des entreprises françaises. Ces autres données sont échantillonnées par un logiciel propriétaire, sont représentatives de l'ensemble des entreprises françaises et sont disponibles en grande quantité.

Ces données d'entreprises représentatives sont réduites et projetées dans un espace bidimensionnel avec l'algorithme d'apprentissage de variété, Uniform Manifold Approximation and Projection (UMAP)[8]. Les données d'entreprises

lauréates sont ensuite projetées dans ce nouvel espace. La distribution de ces entreprises lauréates dans l'espace réduit montre une concentration de celle-ci dans une zone de l'espace. Cette zone est identifiée grâce à une méthode de partitionnement par densité : Density-based Spectral Clustering on Applications with Noise (DBSCAN)[2]. La chaîne de traitement des données est synthétisée dans la figure 1.

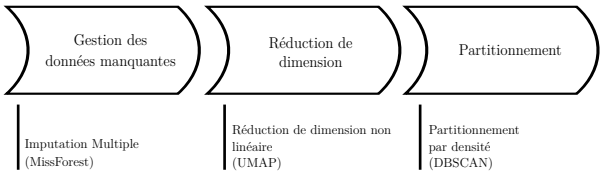


FIGURE 1 – Flux de traitement des données.

Cette étude présente une méthodologie pour exploiter une population cible d'entreprises lauréates pour isoler une sous-population d'entreprises potentiellement lauréates au sein d'une population représentative d'entreprises. L'objectif est l'identification d'une partition de données d'entreprises dont le voisinage est dense en entreprises précédemment lauréates au travers d'un apprentissage non supervisé. Cette approche se base sur la proximité : l'hypothèse sous-jacente est qu'une proximité définie par des attributs financiers induit une plus forte chance d'être lauréat. Cette approche combine un algorithme d'apprentissage de variété avec une méthode de partitionnement par densité. Ces travaux enrichiront le système de recommandation existant finElink, constitué de plusieurs analyses concomitantes de données saisies sur l'entreprise et son projet de recherche. Les données et la méthodologie de cette étude sont introduites plus précisément dans la section 2. Les résultats quant à la réduction et le partitionnement des données sont abordés dans la section 3. Enfin, la conclusion est présentée dans la section 4.

1.1 Travaux antérieurs

La réduction de dimensions permet de représenter des données de grande dimension dans un espace de basse dimension tout essayant de préserver au maximum la structure des données. Les algorithmes de réduction de dimensions linéaire, comme l'Analyse en Composantes Principales (ACP), préservent la structure globale des données sans pour autant réussir à décrire la géométrie réelle des données non linéaires[6]. Dans notre cas applicatif, les entreprises lauréates et les entreprises représentatives avaient des distributions spatiales similaires dans l'espace réduit. L'ACP n'a pas réussi à isoler la population cible de la population représentative. Les algorithmes de réduction de dimensions non linéaire, aussi nommés d'apprentissage de variété, peuvent décrire un plus large éventail d'interactions entre les différentes variables[13]. Ils sont généralement divisés en deux catégories selon la manière dont la structure des données est préservée : soit à un niveau local, soit à un niveau global. Les algorithmes de réduction de dimensions non linéaire préservant la structure globale des données tels que l'ACP à noyaux (kernel PCA)[9] ou l'Isometric Feature Mapping (ISOMAP)[16] s'efforcent de préserver la géométrie des données d'entrée à toutes les échelles : le voisinage et l'éloignement entre points sont préservés entre les espaces d'entrée et de sortie. Les algorithmes de réduction de dimension non linéaire préservant la structure locale des données tels que le Locally Linear Embedding (LLE)[10] ou le t-distributed Stochastic Neighbor Embedding (t-SNE)[7] se concentrent principalement sur la préservation de la géométrie locale au niveau de voisinages de points de la variété[13]. L'algorithme UMAP[8], développé récemment, entre dans cette dernière catégorie. L'approche locale présente deux avantages par rapport à l'approche globale : d'abord, une complexité de calcul plus faible, car les calculs se basent sur une manipulation de matrices creuses, ensuite, une capacité accrue pour représenter une gamme plus large de variétés, en particulier lorsque la géométrie est euclidienne à l'échelle locale sans l'être à l'échelle globale. t-SNE a fait ses preuves sur des données réelles en équilibrant la structure locale et globale des données, ce qui lui donne un avantage concurrentiel. Cela n'a pas été le cas pour le LLE et ses autres homologues non linéaires[7]. UMAP concurrence t-SNE en contournant certaines de ses limites[12, 17] comme :

- le passage à l'échelle difficile en temps de calcul pour les implémentations des librairies fréquemment utilisées en python [1] ;
- la non-convexité de la fonction de coût engendrant potentiellement des résultats qui dépendent de l'initialisation ;
- la non-préservation de la densité et des distances de l'espace de départ dans l'espace d'arrivée (mais préservation du voisinage).

En raison de l'impossibilité d'utiliser la distance et la densité dans l'espace de sortie et de pouvoir projeter de nouvelles données directement dans l'espace réduit, l'algorithme UMAP a été préféré à t-SNE. Bien que récent, UMAP a déjà été testé avec succès dans un contexte médical pour des applications clinique[5], génétique[4] et épidémiologique[1]. Appliquer UMAP dans le contexte de financement de l'innovation est nouveau tant sur l'utilisation de l'algorithme que le contexte applicatif en raison de la faible disponibilité des données financières.

2 Méthode

2.1 Données

La première population cible d'entreprises lauréates est constituée de 3 350 échantillons. La seconde population représentative d'entreprises françaises a été obtenue grâce au logiciel propriétaire Amadeus et inclut 152 899 échan-

1. Depuis décembre 2019, la librairie scikit-learn propose une implémentation de t-SNE parallélisable.

tillons. Amadeus agrège des informations financières sur des millions d'entreprises européennes et permet d'obtenir une population d'entreprises sans biais de sélection. Les attributs sélectionnés pour cette étude sont le chiffre d'affaires, les fonds propres, l'effectif et les bénéfices sur une période de trois années sans que les données aient été traitées comme des séries temporelles. Le choix des attributs est lié au système de recommandation finElink : les informations demandées à l'utilisateur sur son entreprise doivent être facilement accessibles pour faciliter la saisie. De plus le chiffre d'affaires, les bénéfices et les fonds propres décrivent tous trois des réalites financières différentes : le chiffre d'affaires traduit le poids de l'entreprise sur le marché, les bénéfices renseignent sur la performance et rentabilité économiques, et les fonds propres informe quant à la capacité de l'entreprise à entreprendre des projets (et puiser dans ses réserves). Les attributs d'âge, de secteur d'activité et d'emplacement géographique ont été exclus de la projection, car étant discrétisés (respectivement en année, code NACE et région), ils induisaient une séparation lors de la projection par UMAP qui ne permettait pas d'isoler correctement les lauréats. De fait, si ces attributs étaient inclus, la projection reproduisait les catégories initialement présentes dans les données sans séparer la population d'entreprises lauréates de la population représentative.

2.2 Imputation de données manquantes

Les données manquantes ont été imputées en utilisant une méthode d'imputation multiple MissForest [14] qui se base sur un modèle de forêt aléatoire. MissForest a une bonne tolérance à un taux élevé d'absence de données et peut gérer des cas où les données ne sont pas manquantes au hasard (Non Missing At Random - NMAR)[15]. Les méthodes d'imputations multiples préservent mieux les distributions de données que les méthodes d'imputation simple. Multiple Imputation by Chained Equations (MICE)[11] est une autre méthode d'imputation multiple de données qui se base sur de la régression. Les deux fonctionnent sur des données mixtes (catégorielles et/ou continues). L'avantage principal de MissForest sur MICE est que MissForest est non-paramétrique et peut ainsi gérer les non-linéarités et interactions entre variables. Le taux initial de données manquantes pour la population cible d'entreprises lauréates était de 58% tandis que celui de la population représentative d'entreprises était de 34%. De par les taux observés, l'imputation de données n'est pas pertinente sur l'intégralité des données. 15 attributs sont utilisés pour l'imputation de données. Sur ces 15 attributs, les attributs de région, de secteur d'activité et d'âge étaient faiblement manquants car ordinairement disponibles dans le registre des sociétés. Les données d'entreprises avec au plus 7 attributs financier manquants, sur un total de 12 attributs financiers sont gardées pour que l'imputation de données manquantes ne modifie pas fortement la distribution initiale des données. Les 12 attributs financiers sont traités de manière équivalente pour la sélection des échantillons. Le nombre de données fut réduit à 1 413 et 114 628 pour respectivement la population cible et la population représentative.

2.3 Réduction de dimension

La capacité de séparation des populations d'entreprises représentatives et lauréates est présentée dans la figure 2 pour l'ACP, t-SNE et UMAP. L'ACP n'arrive pas à produire une représentation réduite des données qui permet de séparer la population représentative de la population cible comme présenté dans figure 2a. t-SNE obtient une distribution plus informative de la population représentative (figure 2b) sans toutefois pouvoir directement projeter de nouvelles données dans l'espace réduit : il est nécessaire de passer par un modèle de régression - ici un modèle de fôret aléatoire à 10 arbres - pour pouvoir projeter la nouvelle population dans cet espace réduit. De plus, il n'est pas possible d'exploiter complètement l'espace réduit produit par t-SNE car les notions de densité et de distance de l'espace de départ ne sont pas préservées dans l'espace d'arrivée : effectuer un partitionnement dans cet espace se révélerait potentiellement trompeur. Enfin la projection résultante d'UMAP est présentée dans la figure 2c, celle-ci obtient des distributions de populations différentes pouvant être exploitées car l'espace d'arrivée conserve les propriétés de distance, voisinage et densité.

UMAP réduit et projette la population représentative d'entreprises dans un espace bidimensionnel. L'algorithme a une approche par voisinage. Il se déroule en deux étapes : la première consiste en la modélisation des données d'entrées en utilisant des outils de topologie. Un graphe de voisinage des données est construit par recouvrement [2] de l'ensemble des données avec l'aide de simplexes [3]. Cela permet d'obtenir une structure de la variété des données par une approche de voisinage. La seconde est une étape de compression : la représentation réduite des données est obtenue par optimisation par descente de gradient selon l'entropie croisée [4]. L'objectif est de trouver la représentation de faible dimension la plus proche du graphe de voisinage obtenu lors de la première étape de modélisation. L'hypothèse initiale sur les données pour pouvoir appliquer UMAP est qu'il faut que les données soient distribuées de manière uniforme sur une variété Riemmanienne. En faisant l'hypothèse qu'il existe une métrique Riemannienne pour la variété qui ne découle pas de l'espace ambiant alors il est possible de trouver une métrique pour laquelle les données soient distribuées de manière uniforme sur celle-ci : cela implique de créer une distance spécifique à chaque échan-

2. Le recouvrement d'un espace E est une famille d'ensembles dont l'union contient E.

3. En géométrie, un simplexe est défini comme un ensemble de points dont aucun ne peut être obtenu comme barycentre des autres. Sa réalisation géométrique correspond à l'enveloppe convexe de ces points. Plus simplement, un n-simplexe peut être vu comme la généralisation du triangle à une dimension n.

4. En apprentissage, l'entropie croisée est fréquemment utilisée comme fonction de coût pour comparer deux distributions : une à optimiser p, l'autre, fixe, que l'on souhaite approximer, q

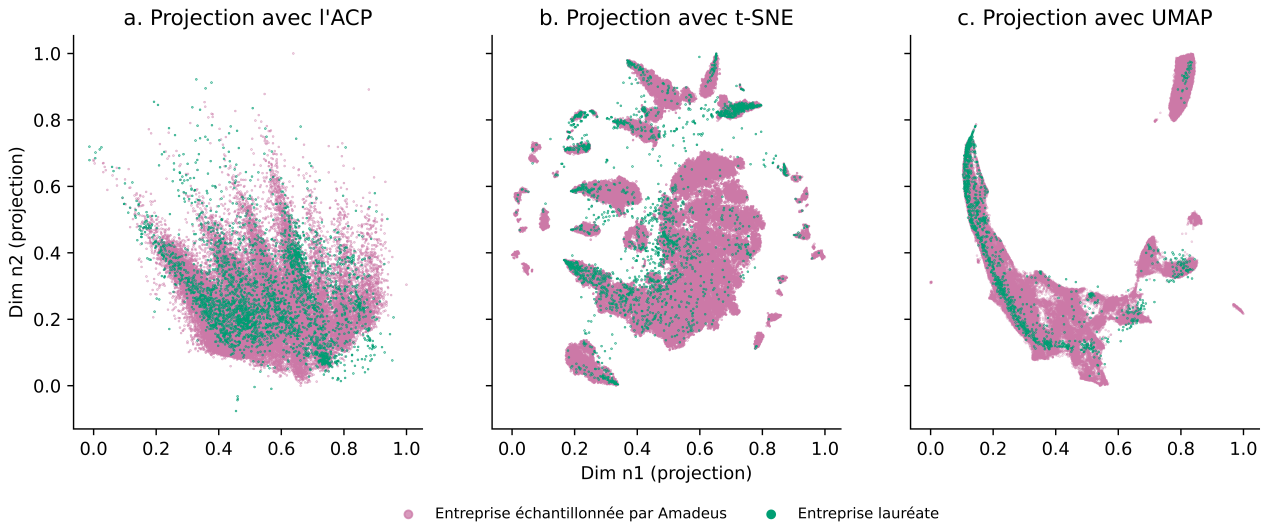


FIGURE 2 – Projection des données d'entreprises représentatives avec l'ACP (a), t-SNE (b) et UMAP (c). Les données d'entreprises lauréates sont ensuite projetées dans cet espace réduit. Les axes résultant de la réduction sont sans dimensions pour les trois algorithmes.

tillon et permet de valider l'hypothèse de distribution uniforme des données sur la variété dans les cas réels [8].

Le paramétrage de l'algorithme est principalement lié pour la phase de modélisation à l'importance portée sur la structure locale ou globale des données d'entrées selon l'étendu du voisinage considéré pour chaque échantillon. Pour la phase de compression, les facteurs essentiels sont le nombre de dimensions choisies pour l'espace de sortie et la distance minimale entre deux points dans l'espace réduit, soit le degré de tolérance d'obtenir une représentation compacte des données. Dans notre étude, la dimensionalité de sortie a été fixée à 2, l'ajout de dimensions supérieures n'apportant pas d'informations supplémentaires dans notre cas. Une vision globale de la structure des données a été privilégiée, avec un nombre de voisins explorés aussi élevé que possible étant donné le temps de calcul (6 500) car la structure locale des données n'a pas permis d'isoler les entreprises lauréates. Nous avons toléré le chevauchement des points dans l'espace réduit en choisissant 0 comme distance minimum.

2.4 Partitionnement

L'espace de projection est découpé en grille et les différences de densité sont examinées, en ce qui concerne le ratio de population de lauréats sur la population totale. Les projections de la population cible sont données en entrée de la méthode de partitionnement DBSCAN pour isoler la zone avec une forte densité d'entreprises lauréates du reste de la population cible. Avec DBSCAN, une partition est identifiée en évaluant la densité de voisinage de chaque échantillon du jeu de données, c'est-à-dire en évaluant le nombre de voisins dans un rayon de ϵ de cet échantillon. Si le nombre de voisins est supérieur au seuil défini par l'utilisateur, cet échantillon est considéré comme un point central de la partition (« cluster core point »). Si l'échantillon n'a pas assez de voisins dans un rayon de ϵ tout en ayant au moins un point central comme voisin, alors ce point est attribué à la partition de ce point central et il est défini comme point frontière (« border point »). Dans le reste des cas, ce point est qualifié d'aberrant (« noise point ») et n'est pas associé à une partition. Les données réduites de la population d'entreprises lauréates sont utilisées comme entrée pour DBSCAN. L'objectif est d'isoler la zone de projection spatiale avec une forte densité d'échantillons de la population cible. Les échantillons restants ont été étiquetés comme bruit. Le choix des paramètres de DBSCAN a été le suivant : la distance ϵ a été fixée à la valeur du premier centile de la distribution des distances entre les échantillons de la population cible. Le nombre minimum de points dans un rayon ϵ nécessaire pour la constitution d'une partition a été fixé à 20.

3 Résultats

3.1 Analyse des attributs d'entrée

L'analyse de la projection des attributs d'entrée, pour la population représentative, indique l'importance relative des attributs dans la construction de la projection d'UMAP. Les résultats sont présentés pour l'année N-1 seulement, mais les tendances de chaque variable sont les mêmes pour les trois années étudiées. Les variables de chiffre d'affaires et d'effectif, présentées respectivement dans la figure 3a et 3c, semblent avoir un impact global sur la distribution des entreprises dans l'espace réduit. La variable de bénéfice n'a pas le même effet d'ensemble sur la projection, comme l'indique la figure 3b. Cette tendance est similaire

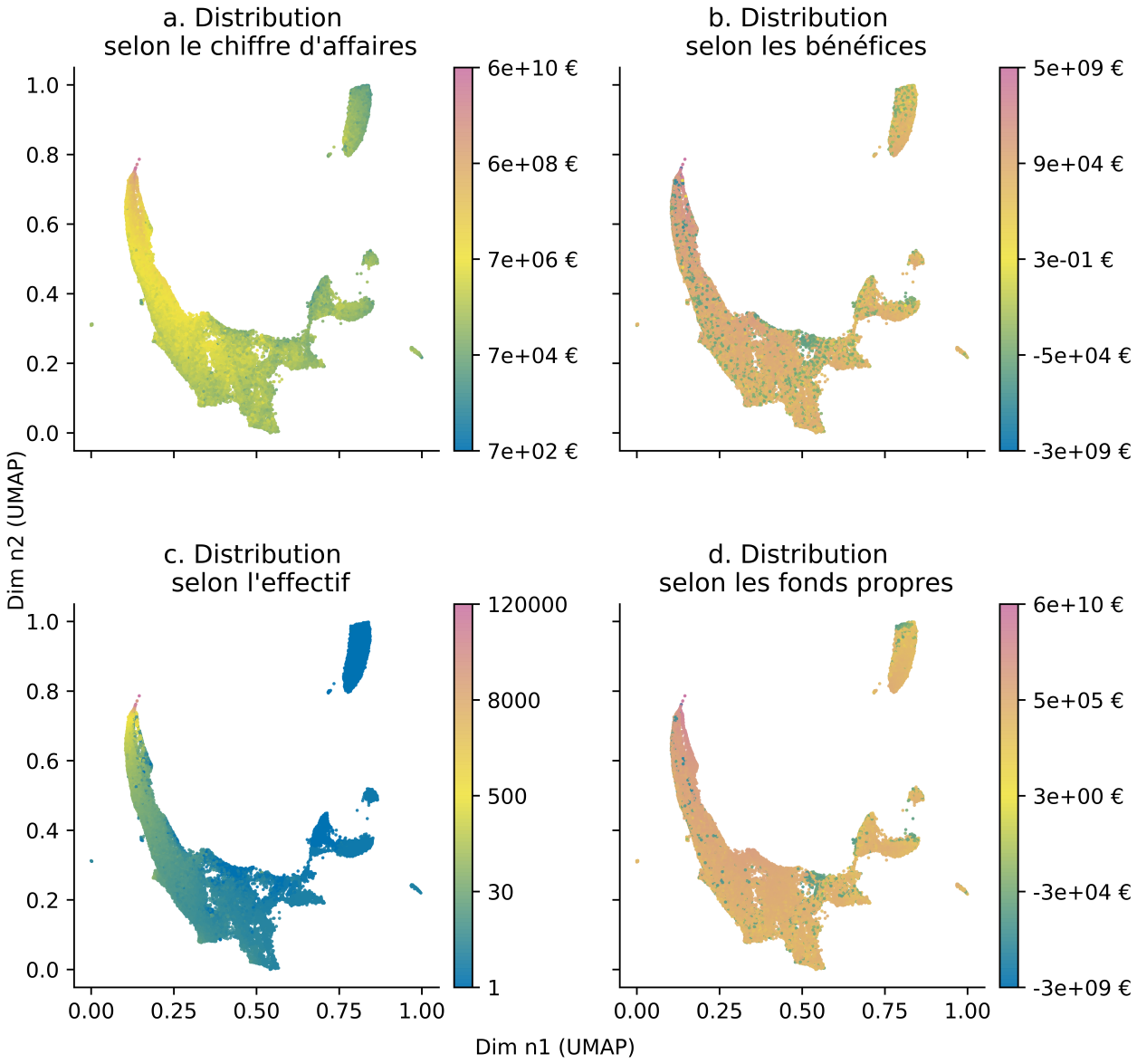


FIGURE 3 – Coloration de la projection des entreprises représentatives selon les variables d'entrée de chiffre d'affaires (a), de bénéfices (b), d'effectif (c) et de fonds propres (d) pour l'année N-1. Les axes résultant de la réduction sont sans dimensions.

pour la variable de fonds propres représentée dans la figure 3d. Bien que cette coloration soit informative quant au fonctionnement de l'algorithme, aucune conclusion ne peut être faite sur le rôle précis de chacune des variables dans la projection apprise par UMAP.

3.2 Analyse des données de lauréats

La projection des entreprises lauréates dans l'espace réduit montre que la distribution des lauréats est différente de la distribution globale des entreprises non lauréates comme présentée dans la figure 4a. La majorité des entreprises lauréates semblent se concentrer dans l'espace gauche de la projection selon une courbe qui part de l'extrémité supérieure gauche de la forme principale et qui suit une ligne pour se terminer dans la partie inférieure centrale, formant une « dorsale ». Le découpage de l'espace de projection en grille dans la figure 4b permet de s'intéresser à la densité de lauréats dans chacune des cellules confirmant la tendance observée précédemment. Les données d'entreprises lauréates se concentrent principalement dans cette « dorsale » avec 74% de la population d'entreprises lauréates appartenant à cette zone (soit 1 041 sur les 1 413 échantillons). Cette concentration spatiale des entreprises lauréates dans l'espace de l'ensemble des entreprises justifie l'utilisation d'une approche par proximité pour identifier les entreprises ayant certains des attributs nécessaires pour être de po-

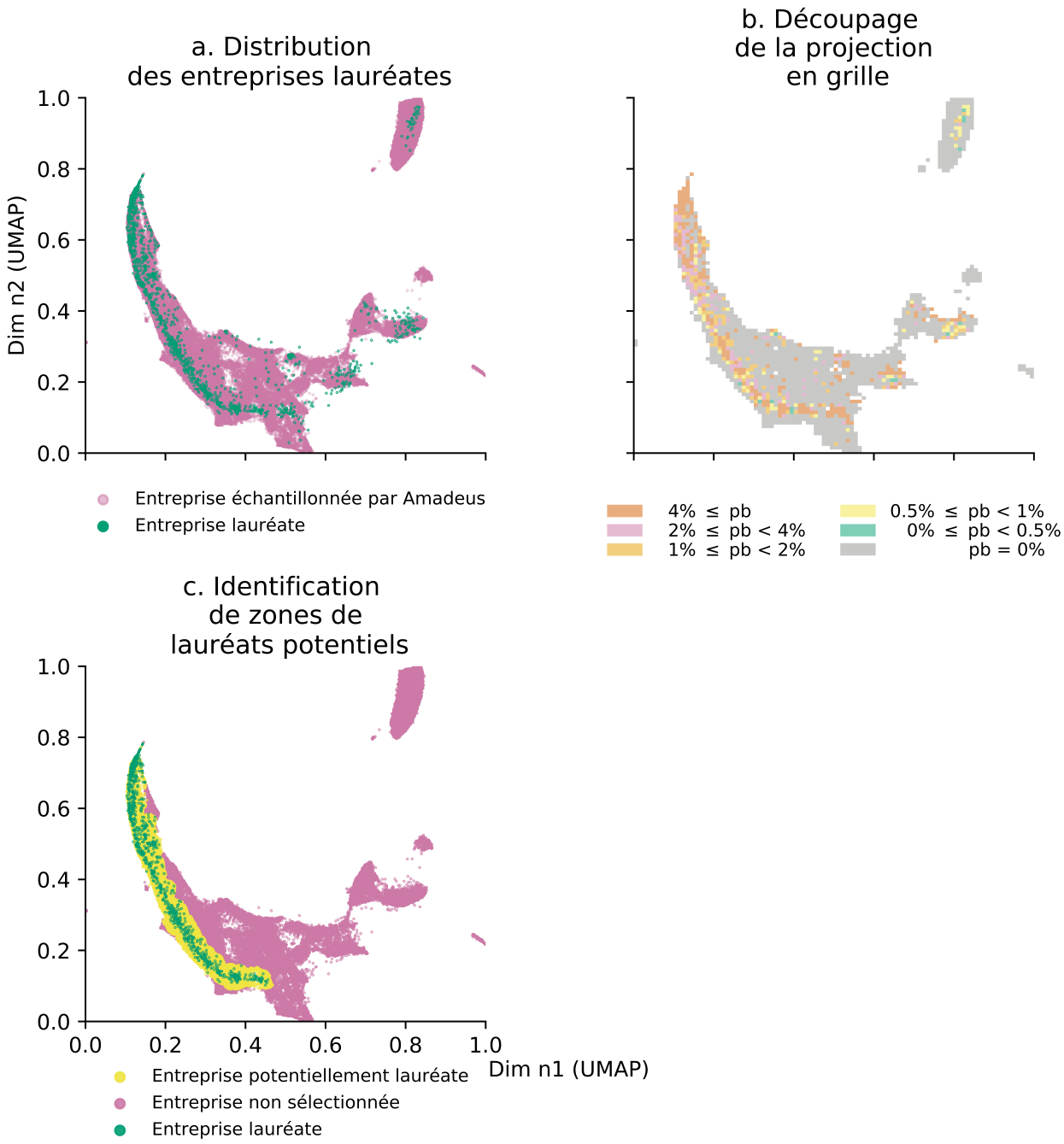


FIGURE 4 – La projection des entreprises lauréates dans l'espace réduit montre une distribution non uniforme dans *a*, qu'un découpage de l'espace de projection confirme dans *b*. La partition d'entreprises potentiellement lauréates est isolée avec DBSCAN dans *c*. Les axes résultant de la réduction sont sans dimensions.

tentielles lauréates. Pour identifier les entreprises les plus proches de cette « dorsale » d'entreprises lauréates, l'algorithme de DBSCAN est appliqué pour identifier la partition principale des entreprises lauréates (et considérer les autres entreprises comme du bruit). Les entreprises non lauréates proches de cette partition sont identifiées selon leur proximité avec les points centraux de la partition et sont considérées comme potentiellement lauréates comme présenté dans la figure 4c. L'appartenance ou non-appartenance à la partition des potentiels lauréats influera sur le choix du système de recommandation de diriger l'utilisateur vers des dispositifs compétitifs.

4 Conclusion

Nous avons présenté notre méthodologie qui permet d'isoler un sous-ensemble d'entreprises partageant une certaine proximité avec notre population cible d'entreprises lauréates. Cette proximité est estimée sur la base d'un nombre restreint d'attributs financiers que sont le chiffre d'affaires, les bénéfices, les fonds propres et l'effectif, sur trois années fiscales. Notre approche se décompose en deux étapes : une première apprend une représentation réduite de nos données représentatives grâce à l'algorithme d'apprentissage de variété UMAP. Les données de notre population cible d'entreprises lauréates sont ensuite projetées dans cet espace réduit. La seconde étape consiste à identifier la zone avec la plus forte concentration d'échantillons d'entreprises lauréates avec la méthode de partitionnement par densité DBSCAN. Les entreprises de la population représentative ou nouvellement projetées suffisamment proches de cette partition sont isolées et traitées comme entreprises avec un plus grand potentiel de succès. Cette différenciation est intégrée au système de recommandation finElink pour affiner les résultats proposés et orienter ces entreprises vers les dispositifs les plus compétitifs. La recommandation de finElink pourrait se raffiner avec un découpage plus fin de l'espace de lauréats pour identifier des sous-groupes de lauréats issus de dispositifs avec une typologie similaire. La méthodologie proposée peut s'appliquer dans d'autres contextes applicatifs lorsque l'objectif est d'isoler une population minoritaire au sein d'une population représentative.

Références

[1] R. Agius, C. Brieghel et al., Machine learning can identify newly diagnosed patients with CLL at high risk of infection, *Nature Communications*, Vol. 11, 2020.

[2] M. Ester, H. Kriegel et al., A density-based algorithm for discovering clusters in large spatial databases with noise, *Kdd*, Vol. 96, pp. 226-231, 1996.

[3] finElink,*https ://www.finelink.eu*, 2018.

[4] A. Fornito, A. Arnatkevičiūtė et al., Bridging the Gap between Connectome and Transcriptome, *Trends in Cognitive Sciences*, Vol. 23, pp. 34-50, 2019.

[5] V. Grollemund, J.-F. Pradat-Peyre et al., Manifold learning for ALS prognosis : development and validation of a prognosis model, *Scientific Reports*, manuscript soumis à la publication.

[6] J. Lee, M. Verleysen et al., *Nonlinear dimensionality reduction* Spriner Science, 2007.

[7] L. van der Maaten et G. Hinton, Visualizing data using t-SNE, *Journal of machine learning research*, Vol. 9, pp. 2579-2605, 2008.

[8] L. McIness, J. Healy et al., UMAP : Uniform Manifold Approximation and Projection for dimension reduction, *arXiv* , preprint arXiv :1802.03426, 2018.

[9] S. Mika, B. Schölkopf et al., Kernel PCA and De-Noising in Feature Space, *11th International Conference on Neural Information Processing Systems*, 1998.

[10] S. Roweis, Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, 2000.

[11] J. Schafer, *Analysis of incomplete multivariate data*, Chapman and Hall/CRC, 1997.

[12] E. Schubert et M. Gertz, Intrinsic t-SNE for visualization and outlier detection, *International Conference on Similarity Search and Applications*, pp. 188-203, 2017.

[13] V. Silva et J. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, *Advances in neural information processing system*, 2003.

[14] D. Stekhoven et P. Bühlmann, MissForest - nonparametric missing value imputation for mixed-type data, *Bioinformatics*, Vol. 28, pp. 112-118, 2012.

[15] F. Tang et H. Ishwaran, Random Forest Missing Data Algorithms, *arXiv*, 2017.

[16] J. Tenenbaum, A Global Geometric Framework for Nonlinear Dimensionality, *Science*, 2000.

[17] M. Wattenberg, F. Viégas et al., How to Use t-SNE Effectively, *Distill*, 2016.

Bandit et Semi-Bandit avec Retour Partiel : Une Stratégie d'Optimisation du Retour Utilisateur

A. LETARD[1, 2], T. AMGHAR[2], O. CAMP[3], N. GUTOWSKI[3]

[1] Dpt R&D, KARA TECHNOLOGY
[2] Université d'Angers, LERIA
[3] Groupe ESEO, ERIS

alexandre.letard@kara.technology

Résumé

Aujourd'hui, dans de nombreux secteurs d'activités, les entreprises renforcent leur numérisation et proposent de nouveaux services à leurs usagers. Ces dernières années, nombre de ces services ont reposé sur des techniques d'apprentissage automatique. Pour les algorithmes de bandits-manchots combinatoires, particulièrement employés pour la recommandation, le retour utilisateur joue un rôle crucial dans l'apprentissage en ligne. Cependant, les stratégies de prise en compte de ce retour reposent essentiellement sur l'observation d'un vecteur de récompenses complet. Celui-ci reste difficile à acquérir lorsque l'utilisateur doit être directement et trop fréquemment sollicité. Dans cet article, nous proposons une nouvelle approche permettant de pallier cette problématique et maintenant une précision globale proche de celles des méthodes classiques.

Mots-clés

Apprentissage par Renforcement, Bandits-Manchots Combinatoires, Retour Utilisateur, Systèmes de Recommandation, Vitesse d'Apprentissage

Abstract

Nowadays, in most fields of activities, companies are strengthening their digitization process and offer new services to their users. In recent years, many of these services have relied on machine learning techniques. Concerning combinatorial multi-armed bandit algorithms, which are particularly employed for recommendation, user feedbacks play a crucial role for online learning. However, strategies for considering those feedbacks are essentially based on the observation of a full rewards vector which can be hard to acquire when users must be directly and too frequently solicited. Herein, we propose a novel approach which overcomes these limitations, while providing a level of global accuracy similar to that obtained by classical competitive methods.

Keywords

Reinforcement Learning, Combinatorial Multi-Armed Bandits, User Feedbacks, Recommendations System, Learning Speed

1 Introduction

De nos jours, les systèmes de recommandations basés sur des méthodes d'apprentissage automatique sont devenus courants dans de nombreux domaines [12]. Parmi les techniques employées, celles reposant sur les bandits-manchots obtiennent de bons résultats en matière de précision globale [12, 6]. C'est le cas, plus particulièrement des bandits-manchots combinatoires [17]. Dans un cadre industriel, certains secteurs d'activités, comme la navigation de plaisance [7], amorcent une transition numérique afin de proposer des services similaires à leurs usagers. Dans le domaine de l'habitat mobile, auquel appartient la navigation de plaisance [7], on observe une concentration des problématiques inhérentes à un lieu de vie et à un véhicule. Il existe ainsi de multiples modes d'utilisation d'un véhicule habitable qui peut être une résidence, principale ou secondaire, ou un moyen de transport, ou encore un outil de dépassement de soi. Ces modes d'utilisation sont propres à chaque usager et à chaque contexte d'usage. Au cours des dernières années, des travaux ont été entrepris pour favoriser la transition numérique du nautisme [14, 22]. Ces travaux portent essentiellement sur l'automatisation des manoeuvres de navigation et ne traitent pas des autres aspects de l'habitat mobile. Or, afin d'apporter des recommandations pertinentes aux navigateurs et ainsi améliorer leurs expériences maritimes, il est nécessaire de prendre en considération l'ensemble de ces aspects. À terme, notre objectif est la mise en oeuvre d'un bateau intelligent, EVA [1], dont le comportement sera guidé par les besoins des utilisateurs, en regard de leur mode d'utilisation du bateau. Nous viserons à réduire les risques de pénuries d'énergie en mer et les impacts environnementaux en optimisant l'utilisation du bateau. Nous mettrons en oeuvre des techniques de bandits-manchots combinatoires afin d'identifier les fonctionnalités "juste nécessaires" parmi celles disponibles à bord, pour satisfaire les usagers.

Afin d'effectuer des recommandations personnalisées, les techniques de bandits-manchots combinatoires considèrent le retour utilisateur exprimé à chaque recommandation [10]. Dans la littérature, les stratégies de prise en compte

1. Entité de Voyages Automatisée

du retour utilisateur les plus fréquemment exploitées reposent sur un vecteur de récompenses complet pour l'ensemble des recommandations effectuées [20, 10]. Ce vecteur peut se révéler difficile à acquérir, notamment lorsque les récompenses dépendent d'un retour explicite de l'utilisateur, p.ex sous la forme d'un score ou d'une évaluation. À ce titre, dans cet article nous expérimentons et évaluons une autre méthode permettant de favoriser l'application des techniques de bandits-manchots combinatoires au sein de systèmes interactifs, en vue de leur intégration dans des secteurs d'activités nouvellement connectés.

Nous proposons une approche appliquant à la fois des considérations de type bandit et semi-bandit sur un sous-ensemble d'éléments recommandés de taille variable. Nous nommons cette méthode "Partial Bandit with Semi-Bandit" (*P-BSB*). Nous proposons trois stratégies pour déterminer le sous-ensemble à observer : *Reinforce - RE*, *Optimal Exploration - OE* et *Randomized - RD*.

Au travers de nos expérimentations, nous appliquons une approche combinatoire à plusieurs algorithmes de bandits manchots à tirages simples. Nous observons que cette stratégie permet l'emploi des techniques de bandits-manchots combinatoires avec un retour utilisateur restreint. Aussi, nous constatons que les résultats de précision globale et d'itération de convergence obtenus pour un horizon supérieur à 10 000 itérations sont proches de ceux obtenus avec des méthodes classiques de bandit et semi-bandit.

En résumé, nos contributions visent à permettre : 1) l'application d'une méthode proposant un tirage multiple à des algorithmes de bandits manchots de l'état de l'art couramment utilisés en tirage simple, notamment *UCB2* [4], qui, à notre connaissance, n'a pas été employé dans un cadre combinatoire ; 2) l'évaluation et la comparaison de ces algorithmes en termes de précision globale et d'itération de convergence sur plusieurs jeux de données issus d'applications réelles ; 3) la formalisation et la proposition d'une nouvelle méthode de considération du retour utilisateur – "Partial Bandit with Semi-Bandit" – pour les algorithmes de bandits manchots que nous déclinons selon trois variantes, afin de réduire les contraintes liées à l'acquisition du retour utilisateur.

Cet article est organisé comme suit : la section 2 expose les notions fondamentales inhérentes aux techniques employées, quelques travaux connexes de la littérature, et nos motivations. La section 3 formalise notre problématique et l'approche que nous proposons pour y répondre. La section 4 analyse les résultats d'expérimentations effectuées avec des jeux de données issus d'applications réelles. Enfin, dans la section 5, nous concluons et exposons les futurs travaux envisagés.

2 Travaux Connexes et Motivations

Cet article traite des bandits-manchots combinatoires et notamment des stratégies de prise en compte du retour utilisateur dans le cadre de leur apprentissage. Ainsi, dans cette section, avant d'aborder les spécificités des bandits-manchots combinatoires [9] (*Combinatorial Multi-Armed Bandit* : *COM-MAB* et *Combinatorial Contextual Multi-Armed Bandit* : *COM-CMAB*) et des approches couramment employées pour exploiter le retour de l'utilisateur, nous rappellerons le problème du bandit-manchot [19] et le problème des bandits-manchots contextuels [15]. Enfin, nous évoquerons les travaux connexes de la littérature et présenterons nos motivations à envisager la combinaison de deux prises en compte partielles du retour utilisateur au sein d'applications réelles de bandits manchots combinatoires.

2.1 Les bandits manchots

Il existe une vaste littérature sur le problème de bandits-manchots (*MAB*) largement étudiés depuis leur première formulation par Robbins en 1952 [19]. Il en résulte aujourd'hui de nombreuses approches [12] : stochastiques [3], non stochastiques [5] ou bayésiennes [1]. Un problème de bandits-manchots est composé d'un ensemble $\mathcal{A} = \{a_1, ...a_m\}$ de *m* bras indépendants, où chaque bras $a \in \mathcal{A}$ correspond à un élément à recommander. Dans le cadre des systèmes de recommandation, à chaque itération $t \in [1,\ T]$, *T* étant l'horizon, un agent sélectionne, suivant une politique π, un bras $a_t \in \mathcal{A}$, correspondant à un élément à recommander et le propose à l'utilisateur. Une partie du vecteur de récompenses Y_t [2] associé à $\mathcal{A}$ est dévoilée à l'agent qui perçoit alors une unique récompense $r_{t,a}$ pour l'élément recommandé. Dans cet article, nous nous intéresserons plus particulièrement au problème de *MAB* de type Bernoulli où $r_{t,a} \in \{0, 1\}$ avec $r_{t,a} = 0$ si l'utilisateur ne valide pas la recommandation qui lui a été faite et $r_{t,a} = 1$ s'il la valide [12]. Dans le cadre d'une approche stochastique, où les récompenses sont considérées comme étant des variables indépendantes et identiquement distribuées entre les bras, l'objectif d'un algorithme de *MAB* est de minimiser le regret $\rho_T = T\mu^* - \sum_{t=1}^{T} r_{t,a}$, où μ^* désigne l'espérance de récompense associée au bras optimal, sans connaissance a priori de la distribution des probabilités de récompenses $\mu_a \in [0, 1]$ associées à chacun des bras a de $\mathcal{A}$. Chercher à minimiser le regret observé revient à maximiser la précision globale $Acc(T) = \frac{\sum_{t=1}^{T} r_{t,a}}{T}$, qui est également couramment exploitée comme critère d'évaluation au sein de la littérature [12].

Dans le cadre particulier des bandits-manchots contextuels (*CMAB*), le contexte de l'utilisateur doit être pris en considération. Il est traduit sous la forme d'un vecteur $x \in \mathcal{X}$, $x \subseteq \mathbb{R}^d$ encodant les *d* caractéristiques de l'utilisateur et de l'environnement dans lequel il évolue, p.ex., le profil (âge, sexe, métier), les préférences, l'environnement (localisation, quartier), ou encore l'activité qu'il réalise. Dans cette variante contextuelle, on considère qu'il existe une dépendance entre l'espérance de récompense d'un bras et le contexte observé. Dans les cas d'une dépendance linéaire, l'espérance s'exprime en fonction du contexte comme suit : $\mathbb{E}[r_{t,a}|x_t] = \hat{\theta}_{t,a}^{\top} x_t$, où $\hat{\theta}_{t,a}$ est un vecteur de coefficients, associé au bras *a*, initialement nul et estimé à l'itération *t*.

2. Y_t est supposé existant mais ne peut être, en réalité, observé qu'en partie ($r_{t,a}$ en cas de tirage simple, R_t dans un cadre combinatoire).

2.2 Les bandits-manchots combinatoires

Les problèmes de bandits manchots combinatoires correspondent à une généralisation des problèmes de *MAB* et *CMAB* où l'utilisateur se voit proposer un super-bras, S_t, constitué de *k* éléments, tels que $S_t = \cup_{i=1}^{k} a_i$, avec $a_i = argmax^{\pi}_{a \in \mathcal{A} \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t]$. À notre connaissance,la principale approche combinatoire est le tirage multiple. Cette méthode constitue S_t dynamiquement en répétant l'action de recommandation d'une ou plusieurs instances à tirages simples pour sélectionner k bras [9, 17, 10] ensuite agrégés en S_t. Ainsi, l'apprentissage s'effectue toujours au travers des bras individuels $a \in \mathcal{A}$.

Ainsi, la valeur de récompense associée à la recommandation S_t, utilisée dans l'apprentissage de l'agent, est exprimée par $S_t^\top R_t = \sum_{i=1}^{k} S_{t,i}\, R_{t,i}$, où R_t est le vecteur de récompense observé, de dimension *k*. Nous nommons ϕ la stratégie de prise en compte du retour utilisateur déterminant : 1) la construction de R_t à partir de Y_t et S_t ; 2) la politique d'apprentissage de l'agent.

Cette évolution des techniques *MAB*/*CMAB* a été employée dans de nombreux secteurs d'activités tels que les systèmes de recommandation, la finance ou le domaine médical [6]. Ainsi, l'algorithme 1 [9, 17, 10] est utilisé dans les expérimentations présentées ici. Dans le cadre de cet article, une valeur de récompense globale, $r_t \in \{0, 1\}$, où $r_t = 1$ si au moins un des éléments de S_t est validé par l'utilisateur, 0 sinon, est employée pour le calcul de la précision globale de l'algorithme tel que défini dans la sous-section 2.1.

2.3 Prise en compte du retour utilisateur

Différentes stratégies de prise en compte du retour utilisateur ont été développées pour les bandits-manchots combinatoires. Ces variantes peuvent être majoritairement regroupées en deux approches principales : **bandit** [13] et **semi-bandit** [20, 10].

Dans l'approche **bandit**, l'agent observe uniquement une récompense cumulée pour le "super-bras" S_t, sans connaître la valeur de retour propre à chacun des *k* bras le constituant : $R_{t,\,\phi_B} = S_t^\top R_t$. L'approche **semi-bandit** permet l'observation de la récompense spécifique de chaque bras $a_{t,i}$ constituant S_t : $R_{t,\,\phi_{SB}} = \cup_i\, S_{t,i}\, R_{t,i}$. Dans les deux cas, l'ensemble du vecteur de récompense R_t, de dimension *k*, reste nécessaire à l'apprentissage.

Dans de récents travaux, il est remarqué que l'approche de type **semi-bandit** est prépondérante [20]. Il en existe également de nombreuses déclinaisons permettant son exploitation dans certains cadres applicatifs, p.ex., le modèle en cascade où le retour utilisateur est exprimé par un clic sur une recommandation et où la position de l'élément cliqué est exploitée pour déterminer implicitement les autres valeurs de retours [16].

2.4 Retours partiels

Une stratégie *partielle* de prise en compte des retours utilisateur correspond à une approche où le vecteur de récompenses observées R_t est seulement défini sur une partie $P_t \subseteq S_t$ de l'ensemble des éléments recommandés à l'utilisateur. Formellement, lors d'une considération non-partielle, la dimension du vecteur de récompenses observées s'exprime par $|R_t| = |S_t|$, tandis que pour les approches partielles elle s'exprime $|R_t| = l$ avec $l < |S_t|$. Parmi les exemples de la littérature [11, 18], l est considéré comme une constante. Ainsi, Grant et al. [11] emploie une approche de semi-bandit partiel reposant sur un filtrage par application d'une loi binomiale. Luedtke et al. [18] exploite aussi une approche de type semi-bandit partiel où un sous-ensemble de S_t, est sélectionné uniformément parmi tous les sous-ensembles de S_t de cardinalité *l*.

L'approche **P-BSB** se différencie par la détermination d'un sous-ensemble de taille variable. L'objectif applicatif de cette nuance est de permettre, à chaque itération, l'exploitation du nombre maximal de retours que l'usager est prêt à prodiguer sans l'excéder. Une autre distinction porte sur la stratégie employée pour construire R_t. Avec **P-BSB**, l'identification des retours à solliciter auprès de l'utilisateur est soit aléatoire (variante **RD**), soit basée sur l'apprentissage réalisé jusqu'à l'itération $t-1$ (variantes **RE** et **OE**). Enfin, notre méthode diffère par l'observation d'une récompense double (R_{t,ϕ_B} et $R_{t,\phi_{SB}}$) lorsqu'un bras a effectivement apporté satisfaction à l'utilisateur.

Algorithme 1 : Bandit à tirages multiples

Entrées : π : Une instance d'une politique de bandit à tirages simples et ses paramètres particuliers.
$\mathcal{A}$: L'ensemble des bras disponibles.
k : Le nombre d'éléments à recommander à chaque itération.
Y_t, Le vecteur de récompenses réelles.
T : L'horizon.
$x \in \mathcal{X}$: Le contexte utilisateur.
$\phi(S_t, Y_t)$: La stratégie de considération du retour utilisateur.

Initialisation : Initialiser l'instance conformément aux besoins de π

pour $t = 1$ *à* T **faire**
  Considérer $x_t \in X$: un utilisateur u et son contexte
  $S_t \leftarrow \emptyset$
  pour $i = 1$ *à* k **faire**
    Sélectionner l'élément $a_i \in \mathcal{A} \setminus \{S_t\}$ selon x_t et π
    $S_t \leftarrow S_t \cup a_i$
  fin
  Recommander S_t à l'utilisateur u
  Recevoir la récompense globale r_t de la recommandation $S_t, r_t \in \{0, 1\}$
  Déterminer R_t à partir de Y_t et S_t selon ϕ
  Mettre à jour la politique π avec R_t selon π et ϕ
fin

2.5 Motivations

Les algorithmes de *MAB* ou de *CMAB* visent à maximiser leur précision globale [12]. À cette fin, la prise en compte du retour utilisateur joue un rôle majeur. Cependant, l'acquisition d'un vecteur de récompense R_t complet, nécessaire dans les considérations bandit et semi-bandit, peut s'avérer difficile voire impossible en situation réelle.

De nombreux secteurs d'activités, comme la navigation de plaisance, amorcent leur transition numérique et ne disposent donc aujourd'hui d'aucun jeu de données permettant l'entraînement hors ligne d'un agent. Dans ce cadre applicatif, les usagers risquent de se détourner d'une application s'ils ne sont pas satisfaits des recommandations proposées. Il est alors crucial pour le système de recommandation d'acquérir rapidement en ligne une connaissance suffisante. Ainsi, nous soutenons que l'itération de convergence de la précision globale, indicatrice de la vitesse d'apprentissage de l'agent, doit être prise en considération comme un critère d'évaluation à part entière des algorithmes étudiés.

Par ailleurs, des approches telles que celles en cascade [16] restent délicates à employer si le retour utilisateur doit être explicitement sollicité, p.ex., sous la forme d'un score pour déterminer les points de valeur parmi les étapes et activités - qui seraient les bras disponibles d'un algorithme de bandit manchot combinatoire - d'un voyage défini et recommandé par l'agent. Le nombre potentiellement important de retours demandés auprès de l'usager pour satisfaire une telle configuration pourrait en effet le détourner de la solution.

Ainsi, l'approche *Partial Bandit with Semi-Bandit* (*P-BSB*), proposée dans cet article, repose sur l'identification d'un sous-ensemble de R_t, $\mathcal{P}_t$, de cardinalité ρ variable, correspondant au nombre de sollicitations auxquelles l'usager accepte de répondre. **P-BSB** emploie ensuite une approche de type bandit sur S_t et une approche de type semi-bandit sur P_t. Cette double attribution de récompense à certains bras vise à accroître la vitesse d'apprentissage en avantageant les bras pour lesquels une satisfaction de l'utilisateur est effectivement observée. Au travers de cette approche, l'objectif est de faciliter l'utilisation des techniques de bandits-manchots combinatoires sur un plus large spectre d'applications du monde réel en : 1) réduisant les sollicitations auprès des utilisateurs ; 2) conservant des performances similaires à celles observées avec les méthodes classiques.

3 Définition du problème et méthodes

Dans cette section, nous formulons notre problème et décrivons notre nouvelle méthode. Celle-ci porte sur la prise en compte du retour utilisateur et repose sur la combinaison des stratégies de "bandit" et "semi-bandit" couramment employées dans la littérature. Nous les appliquons avec un nombre restreint de retours utilisateur observés.

3.1 Énoncé du problème

Soit $\mathcal{X} \subseteq \{0,1\}^d$ l'ensemble des vecteurs de contexte de dimension d caractérisant un utilisateur et son environnement p.ex., $x \in \mathcal{X}$ est un vecteur binaire encodant les caractéristiques des utilisateurs demandant une recommandation p.ex., des activités à réaliser au cours d'un voyage, à l'instant *t* d'un horizon fini *T* déterminé à l'avance. Dans le cas non contextuel c.-à-d., en l'absence de contexte ou sans prise en compte du contexte par l'algorithme, alors $\forall t \in T, x_t = \vec{0}$, le vecteur x_t est alors limité à un simple identifiant.

Soit $\mathcal{A} = \{a_1, ..., a_m\}$ l'ensemble des éléments pouvant être recommandé par un algorithme de *MAB* ou *CMAB* donné de politique π et $\mu = \{\mu_1, ..., \mu_m\}$ la distribution des espérances de récompenses associées à chaque bras *a* de $\mathcal{A}$ selon π. Soit S_t le sous-ensemble constitué d'éléments de $\mathcal{A}$, de dimension $k < m$ à l'itération $t \in [1, T]$. À chaque pas de temps t on recommande un super-bras S_t, déterminé selon π et μ_t, à un utilisateur u_t se présentant avec son vecteur de contexte x_t. Enfin, soient $r_t \in \{0,1\}$ la récompense globale associée à S_t utilisée pour le calcul de la précision globale, $Acc^{\pi}(T)$ de l'agent, Y_t un vecteur associant une récompense réelle à chacun des bras $\{a_1, ..., a_m\}$ de $\mathcal{A}$ et $R_t \subseteq Y_t$ le vecteur de récompense exprimé par l'usager et effectivement observé par l'agent. Dans les cas des stratégies de type bandit et semi-bandit il est supposé que $R_t = Y_t$ pour les *k* bras de S_t. Or, dans plusieurs applications du monde réel, lorsque les récompenses observées ne peuvent être obtenues que par une sollicitation explicite de l'usager pour l'ensemble des bras constituant S_t, cette nécessité devient alors très difficile à satisfaire. Ce constat peut se révéler critique pour nombre d'applications du monde réel. Ainsi, une nouvelle approche visant à réduire les sollicitations des utilisateurs tout en maintenant un apprentissage efficace semble nécessaire. La sous-section suivante présente une nouvelle méthode exploitant une combinaison des stratégies de type bandit et semi-bandit sur un sous-ensemble restreint de S_t.

3.2 Partial Bandit with Semi-Bandit : P-BSB

Soit un cadre applicatif où le retour utilisateur doit être explicitement demandé à l'usager et où $|R_t| = \rho \leq k$ est donc une variable aléatoire représentative de la capacité de l'utilisateur à effectuer un retour à l'agent, pouvant dépendre p.ex., de sa disponibilité, de son intérêt à répondre, de son humeur. Dans cet article, cette capacité est désignée sous le terme de "patience" de l'utilisateur. L'approche *P-BSB* vise à construire R_t et à déterminer son utilisation pour l'apprentissage de l'agent. Cette méthode correspond à une application des lignes 10 et 11 de l'algorithme 1. La première étape de l'approche *P-BSB* consiste à identifier un sous-ensemble $\mathcal{P}_t \subseteq \mathcal{S}_t$ de cardinalité ρ pour lequel des récompenses pourront effectivement être observées par l'agent, tel que $P_t = \cup_{i=1}^{\rho} a_i$. Pour ce faire, *P-BSB* est décliné selon trois variantes pour déterminer les bras a_i considérés :

— **Reinforce - RE** : sélectionne les ρ bras de S_t

ayant l'espérance de récompense $\mathbb{E}[R_{t,a}|x_t]$ la plus haute, c.-à-d. :

$$a_i = argmax^{\pi}_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t] \quad (1)$$

— **Optimal-Exploration - OE** : sélectionne les ρ bras de S_t ayant été le moins fréquemment observés à l'itération *t*, c.-à-d. :

$$a_i = argmin_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} obs_{a,t} \quad (2)$$

où $obs_{a,t}$ est le nombre de fois qu'une récompense a été observée pour le bras a à l'itération t.

— **Randomized - RD** : sélectionne aléatoirement ρ bras distincts de S_t, c.-à-d. :

$$a_i = random(S_t) \quad (3)$$

où $random(S_t)$ correspond à la sélection aléatoire d'un bras dans S_t n'ayant pas été précédemment choisi.

La seconde étape de *P-BSB* est commune à toutes ces variantes et consiste à acquérir le vecteur R_t de récompenses des ρ bras considérés à partir de Y_t, ou autrement dit, de l'utilisateur :

$$R_t = \cup_{i \in P_t} Y_{t,i}$$

Enfin, la troisième étape consiste à appliquer une stratégie de type bandit sur l'ensemble des *k* bras de S_t et une stratégie de type semi-bandit sur les ρ bras de P_t à partir de l'échantillon R_t observé [3] :

$$\forall a \in S_t, r_{t,a} = r_{t-1,a} + R_{t,B}$$

et si $a \in P_t$ alors $r_{t,a} = r_{t,a} + R_{t,SB_a}$

Où $r_{t,a}$ désigne l'ensemble des récompenses perçues pour le bras *a* à l'itération *t*, et avec :

$$R_{t,B} = \mathcal{P}_t^{\top} \; R_t$$

et $\forall i \in P_t$:

$$R_{t,SB} = \cup_i \; P_{t,i} \; R_{t,i}$$

Cet article détaille la variante **RE** dans l'Algorithme 2. Les autres variantes suivent la même trame globale et ne diffèrent que par leur stratégie de constitution de $\mathcal{P}_t$. Ainsi, pour employer ces variantes, il convient de remplacer la ligne 2 de l'Algorithme 2 par les éléments de l'Équation 2 pour **OE** et de l'Équation 3 pour **RD**.

L'objectif de **RE** est de favoriser une exploitabilité applicative rapide en favorisant une action optimiste de l'agent. **OE**, quant à lui, suit un objectif exploratoire et cherche donc à renforcer la connaissance de l'agent sur la distribution des espérances de récompenses $\mu_1, ..., \mu_k$ des bras recommandés. Enfin, **RD** applique une stratégie de sélection aléatoire permettant un comportement plus proche d'une prise en compte semi-bandit classique.

3. Lorsque $a \in \mathcal{P}_t$, l'agent observe donc deux récompenses pour le bras a : $R_{t,B}$ et R_{t,SB_a}. Si $\rho = 0$, l'agent n'observe alors aucune récompense à l'itération *t*.

4 Expérimentations et résultats

Nous présentons dans cette section l'évaluation empirique hors ligne de notre méthode. Cette phase d'expérimentation est préliminaire à l'intégration en ligne de notre méthode dans notre système de recommandation en environnement marin.

Ainsi dans cette section, nous commençons par présenter les jeux de données et les algorithmes employés pour évaluer notre approche. Nous exposons ensuite le protocole expérimental. Enfin nous présentons et analysons les résultats obtenus.

Algorithme 2 : *P-BSB* - RE

Entrées : S_t, Le super-bras recommandé à l'utilisateur.
Y_t, Le vecteur de récompenses réelles.
ρ_t, Le nombre de bras de S_t dont les récompenses peuvent être observées.

1 **tant que** $|\mathcal{P}_t| < \rho_t$ **faire**
2 Constituer $\mathcal{P}_t$ tel que $\mathcal{P}_t = \cup_i \; argmax^{\pi}_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t]$ (selon l'Équation 1)
3 **fin**
4 **pour** $i \in P_t$ **faire**
5 Construire R_t tel que $R_t = \cup_i \; Y_{t,i}$
6 Appliquer la stratégie semi-bandit à R_t : $R_{t,SB} = \cup_i \; P_{t,i} \; R_{t,i}$
7 **fin**
8 Appliquer la stratégie bandit à R_t : $R_{t,B} = \mathcal{P}_t^{\top} \; R_t$
9 **pour** $a \in S_t$ **faire**
10 Mettre à jour la politique π avec $r_{t,a} = r_{t-1,a} + R_{t,B}$
11 **si** $a \in \mathcal{P}_t$ **alors**
12 Mettre à jour la politique π avec $r_{t,a} = r_{t,a} + R_{t,SB_a}$
13 **fin**
14 **fin**

4.1 Jeux de données

Nous évaluons notre approche sur cinq jeux de données issus d'applications réelles (cf. Tableau 1) :

— **Covertype** [4] ainsi que **Poker Hand** [5] offrent un nombre important de contextes utilisateur et permettent ainsi de passer à l'échelle ;

4. https ://archive.ics.uci.edu/ml/datasets/covertype
5. https ://archive.ics.uci.edu/ml/datasets/Poker+Hand
6. https ://www.kaggle.com/vikashrajluhaniwal/jester-17m-jokes-ratings-dataset
7. https ://www.kaggle.com/assopavic/recommendation-system-for-angers-smart-city
8. https ://grouplens.org/datasets/movielens/100k/

Jeu de données	Instances	Bras	Variables contextuelles
Covertype [4]	581 012	7	54
Poker [5]	1 025 010	10	10
RS-ASM [7]	2 152	18	50
Jester [6]	59 132	150	0
MovieLens [8]	942	1682	23

Tableau 1 – Jeux de données employés dans nos expérimentations

— **RS-ASM** [7] est un jeu de données pour la recommandation de services dans la ville intelligente [12] ;
— **Jester** [6] est un jeu de données pour la recommandation de blague où aucune information de contexte n'est disponible ;
— **MoviesLens** [8] est un jeu de données spécifique pour la recommandation de films.

Jester et **MoviesLens** représentent les cas où le nombre de bras disponibles est important et où le retour utilisateur est exprimé sous la forme d'un score allant de 0 à 5. Pour cette expérience, nous définissons un seuil $s = 4$ où $R_{t,a} = 1$ si le score est supérieur ou égal à s, 0 sinon.

4.2 Algorithmes

La méthode évaluée porte sur la prise en compte du retour utilisateur. Ainsi, elle fonctionne indépendamment de la politique π suivie par l'agent c.-à-d., indépendamment de l'algorithme de *COM-MAB* / *COM-CMAB* employé. Nous mettons en oeuvre l'algorithme 1 avec plusieurs algorithmes à tirages simples connus de la littérature avant de les évaluer en termes de précision globale et d'itération de convergence.
Ainsi, dans cet article et à la lumière des précédentes évaluations effectuées dans la littérature sur les banditscombinatoires [9, 17, 8], nous considérons les algorithmes suivants :

— **MAB** : *ε-greedy* [21] avec $\varepsilon = 0.0009$, *Thompson Sampling (TS)* [1], *UCB* [3] et *UCB2* [4] ;
— **CMAB** : *LinUCB* [15] et *LinTS* [2].

4.3 Protocole expérimental

Au cours de nos expériences et pour chaque algorithme, afin de simuler un flux de données d'utilisateurs se présentant dans un contexte donné pour recevoir une recommandation (voir ligne 2 de l'algorithme 1), une sélection aléatoire est réalisée séquentiellement parmi les contextes disponibles dans l'ensemble du jeu de données jusqu'à un horizon fixe $T = 10000$. L'itération de convergence t_c considérée dans cet article correspond à la première itération *t* à partir de laquelle la précision globale demeure équivalente à la précision globale finale $Acc(T)$ (voir calcul à la Sous-Section 2.1), à $\delta = 1\%$ près :

$\forall t \geq t_c$:

$$Acc(T) - \delta \leq Acc(t) \leq Acc(T) + \delta, \text{ avec } \delta = 0.01$$

Algorithme	**Stratégie**	$Acc(T)$	t_c
ε-greedy	Bandit	$0,833_{\pm 0,002}$	$1461_{\pm 1746}$
	Semi-Bandit	$0,859_{\pm 0,004}$	$840_{\pm 912}$
	P-BSB-RE	$0,840_{\pm 0,005}$	$2476_{\pm 1359}$
	P-BSB-OE	$0,836_{\pm 0,004}$	$411_{\pm 265}$
	P-BSB-RD	$0,838_{\pm 0,002}$	$1288_{\pm 1113}$
TS	Bandit	$0,825_{\pm 0,002}$	$928_{\pm 1472}$
	Semi-Bandit	$0,857_{\pm 0,003}$	$1938_{\pm 1511}$
	P-BSB-RE	$0,845_{\pm 0,004}$	$1206_{\pm 1211}$
	P-BSB-OE	$0,837_{\pm 0,003}$	$545_{\pm 541}$
	P-BSB-RD	$0,839_{\pm 0,007}$	$1079_{\pm 1396}$
UCB	Bandit	$0,832_{\pm 0,005}$	$1171_{\pm 1304}$
	Semi-Bandit	$0,842_{\pm 0,002}$	$4163_{\pm 1512}$
	P-BSB-RE	$0,830_{\pm 0,004}$	$1530_{\pm 1566}$
	P-BSB-OE	$0,823_{\pm 0,004}$	$774_{\pm 905}$
	P-BSB-RD	$0,826_{\pm 0,002}$	$1580_{\pm 1542}$
UCB2	Bandit	$0,796_{\pm 0,002}$	$948_{\pm 1041}$
	Semi-Bandit	$0,796_{\pm 0,002}$	$886_{\pm 950}$
	P-BSB-RE	$0,790_{\pm 0,002}$	$1554_{\pm 1611}$
	P-BSB-OE	$0,801_{\pm 0,003}$	$889_{\pm 1213}$
	P-BSB-RD	$0,792_{\pm 0,001}$	$1734_{\pm 2235}$

Tableau 2 – Résultats pour une application non-contextuelle avec $k = 10$ sur le jeu de données **Jester**.

Chacun des algorithmes *COM-MAB / COM-CMAB* est employé sur les cinq jeux de données pour réaliser des recommandations de 3 éléments ($k = 3$). Ces expériences sont réalisées en employant successivement les stratégies de considération du retour utilisateur **bandit**, **semi-bandit**, **RE**, **OE** et **RD**, afin de permettre une comparaison des approches.
Lorsque l'une des variantes de **P-BSB** est appliquée, la "patience" ρ de l'usager est simulée par une variable aléatoire comprise entre 0 et *k*, générée à chaque itération.
Le même procédé est employé pour effectuer des recommandations à 10 éléments ($k = 10$), en faisant varier ρ entre 0 et 4, sur les jeux de données **Jester** et **MoviesLens**, disposant d'un nombre de bras important, afin d'expérimenter des situations où $\rho \ll k \ll m$.
Ainsi, pour chacun des différents cas et pour chaque approche, 10 expériences de 10 000 itérations sont simulées.
Les tableaux 2 et 3 présentent les moyennes et écarts-type obtenus pour les critères de précision globale et d'itération de convergence dans les expérimentations où $k = 10$ avec $0 \leq \rho \leq 4$. Ce cas est particulièrement intéressant dans la mesure où le nombre de bras pour lesquels l'agent ne pourra pas observer de récompense à l'itération *t* est plus important, l'expérience est donc plus représentative des résultats pouvant être attendus pour l'application visée à terme.
À la sous-section suivante nous nous focaliserons sur l'interprétation et l'analyse de ces résultats et indiquerons si les tendances observées sont vérifiées au travers de nos autres expérimentations.

4.4 Analyse des résultats

Afin d'observer l'impact d'une approche dans l'apprentissage d'un agent dans un cadre applicatif spécifique indépendamment de l'algorithme *COM-MAB / COM-CMAB*

Algorithme	Stratégie	$Acc(T)$	t_c
LinTS	Bandit	$0,996 \pm 0,001$	330 ± 253
	Semi-Bandit	$0,995 \pm 0,001$	732 ± 472
	P-BSB-RE	$0,989 \pm 0,001$	627 ± 505
	P-BSB-OE	$0,986 \pm 0,001$	391 ± 417
	P-BSB-RD	$0,989 \pm 0,001$	491 ± 330
LinUCB	Bandit	$0,994 \pm 0,001$	944 ± 513
	Semi-Bandit	$0,992 \pm 0,001$	1582 ± 596
	P-BSB-RE	$0,982 \pm 0,001$	1034 ± 541
	P-BSB-OE	$0,979 \pm 0,001$	644 ± 399
	P-BSB-RD	$0,982 \pm 0,002$	1028 ± 662

Tableau 3 – Résultats pour une application contextuelle avec $k = 10$ sur le jeu de données **MoviesLens**.

choisi, nous considérons la moyenne des résultats obtenus par les algorithmes employés. Ainsi, les approches sont comparées à partir des résultats obtenus avec l'équation suivante :

$$\forall \pi \in \Pi : M_\phi = \sum_{\pi=1}^{|\Pi|} \frac{Acc_\phi^\pi(T)}{|\Pi|} \quad (4)$$

Ce procédé est également exploité pour les comparaisons d'itération de convergence : il suffit de remplacer les valeurs $Acc(T)$ par les valeurs correspondantes des colonnes t_c des tableaux de résultats 2 et 3.

4.4.1 Observations spécifiques - Jester et MoviesLens

Le tableau 4 présente les résultats de précision globale - $Acc(T)$ - et d'itération de convergence - t_c - observés en moyenne (selon l'équation 4) pour chacune des approches considérées sur les jeux de données Jester et MoviesLens :

	Jester		**MovieLens**	
Stratégie	$Acc(T)$	t_c	$Acc(T)$	t_c
Bandit	$0,822$	1127	$0,995$	637
Semi-Bandit	$0,839$	1957	$0,994$	1157
P-BSB-RE	$0,826$	1692	$0,986$	831
P-BSB-OE	$0,824$	655	$0,983$	517
P-BSB-RD	$0,824$	1420	$0,986$	759

Tableau 4 – Résultats observés en moyenne pour les approches considérées sur **Jester** et **MoviesLens**.

4.4.2 Tests statistiques

Nous réalisons en premier lieu des tests de *Kruskal-Wallis (KW)* afin de metre en évidence les inégalités entre les résultats obtenus par chacun des algorithmes, c.-à-.d., nous testons l'hypothèse nulle H_0 : « Il n'y a pas de différence significative entre les résultats des différentes approches (médianes) ». Si le test de *KW* indique qu'il existe des différences entre les résultats, il sera alors nécessaire de réaliser des tests de *Rang signés de Wilcoxon (RW)* deux à deux sur la précision globale et l'itération de convergence, c.-à-.d., nous testons l'hypothèse nulle H_0 : « Il n'y a pas de différence significative entre les résultats entre chaque paire d'approches appliquées à chaque algorithme ». Par la suite, nous indiquerons donc : si l'hypothèse nulle est rejetée ou non, et la valeur de *p* correspondante pour chaque comparaison que nous observerons.

4.4.3 Analyse des résultats

Itération de convergence : Même si nous observons un léger avantage à employer l'approche **P-BSB-OE**, cela reste en revanche statistiquement non significatif (Tests *KW* : $p > 0.05$) pour les cas contextuels comme non contextuels.

Précision globale : Les tests de *KW* nous indiquent qu'il existe une différence significative entre les mesures de précision globale obtenues par l'application de chacune des 5 approches considérées au travers d'un même algorithme, et cela pour chacun des algorithmes appliqués dans les cas contextuels comme non contextuels ($p < 0.01$). Dans le cas non-contextuel, les trois déclinaisons de **P-BSB** obtiennent une précision globale significativement supérieure à l'approche **bandit** (Tests *RW* : $p < 0.01$) et l'approche **Semi-Bandit** obtient une précision globale significativement supérieure aux autres approches (Tests *RW* : $p < 0.01$). Les approches **P-BSB-RD** et **P-BSB-OE** obtiennent des résultats équivalents et les tests de *RW* indiquent qu'ils ne présentent pas de différences significatives (Tests *RW* : $p > 0.05$). Dans le cas contextuel, l'approche **bandit** obtient une précision globale significativement supérieure aux autres approches (Tests *RW* : $p < 0.01$). L'approche **P-BSB-RE** obtient une précision globale non significativement supérieure à l'approche **P-BSB-RD** (Tests *RW* : $p > 0.05$). L'approche **P-BSB-RE** quant à elle obtient des résultats significativement supérieurs à l'approche **P-BSB-OE** (Tests *RW* : $p < 0.01$).

Observations : Ces résultats sont obtenus alors que dans les meilleurs cas, c-à-d lors des itérations ou ρ prend sa valeur maximale : 4, **P-BSB** n'emploie que 40% des retours utilisateurs considérés par les stratégies **bandit** et **semi-bandit** et que dans les pires cas, c-à-d lors des itérations ou ρ prend sa valeur minimale : 0, aucun retour utilisateur n'est exploitable pour l'apprentissage de l'agent. Nos résultats sur les jeux de données **RS-ASM**, **Poker Hand** et **Covertype** avec les algorithmes *COM-MAB* confirment les tendances observées lorsque la part de récompenses non observée est moins importante ($k = 3$ et $0 \leq \rho \leq 3$) et nous permettent de confirmer l'adéquation de notre approche dans ce type d'application.

Conclusion : À la vue des résultats expérimentaux, l'objectif applicatif visé par notre approche - acquérir une précision globale proche de celles obtenues avec les approches classiques malgré un nombre de retours utilisateur restreint, voire inexistant à certaines itérations - est atteint par les variantes de **P-BSB** proposées.

5 Conclusions et Perspectives

Notre objectif final est d'intégrer un système de recommandations guidé par les besoins utilisateurs en environnement marin où un vecteur complet de récompenses R_t serait difficile à observer.

Ainsi, dans cet article, nous avons proposé et appliqué une approche combinatoire à plusieurs algorithmes de bandits-

manchots à tirages simples issus de la littérature. Nous les avons évalués en termes de précision globale et d'itération de convergence sur plusieurs jeux de données du monde réel. Les résultats que nous avons obtenus sont en faveur d'une utilisation de l'approche combinatoire pour les systèmes de recommandation à choix multiples.
La principale contribution de cet article porte sur la mise au point et l'expérimentation d'une nouvelle méthode de prise en compte du retour utilisateur : **P-BSB**. Cette approche propose trois variantes : 1) **RE** qui observe les récompenses associées aux ρ bras de S_t de plus haute espérance de récompense ; 2) **OE** qui consulte les récompenses des bras de S_t ayant été le moins observés à l'itération t ; 3) **RD** qui emploie une sélection aléatoire de ρ bras parmi S_t. Dans les cadres contextuels comme non contextuels, l'approche partielle combinant les stratégies bandit et semi-bandit offre des performances proches des approches classiques, malgré un nombre restreint de retours utilisateur.
L'acquisition et la valorisation du retour utilisateur constitue un défi majeur dans le domaine de l'apprentissage automatisé et les résultats obtenus par *P-BSB* encouragent des perspectives d'une mise en application réelle pour un apprentissage en ligne. À ce titre, l'une des perspectives imminentes que nous envisageons est d'étudier une approche complémentaire où la stratégie de prise en compte du retour utilisateur serait déterminée dynamiquement par l'agent à chaque itération.

Remerciements

Ces travaux ont été menés par l'entreprise KARA TECHNOLOGY en collaboration avec les laboratoires du LERIA et ESEO-TECH et avec le soutien de l'Association Nationale de la Recherche et de la Technologie (ANRT).

Références

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, pages 397–422, 2002.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3) :235–256, 2002.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SICOMP*, 32(1) :48–77, 2002.

[6] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *ARXIV*, abs/1904.10040, 2019.

[7] Fondation Bénéteau. Les attentes des futurs plaisanciers. *Rapport FIN*, 2014.

[8] Lixing Chen, Jie Xu, and Zhuo Lu. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In *NIPS*, pages 3247–3256. Curran Associates, Inc., 2018.

[9] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit : General framework and applications. In *ICML*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[10] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *NIPS*, pages 2116–2124. Curran Associates, Inc., 2015.

[11] James A. Grant, David S. Leslie, Kevin Glazebrook, and Roberto Szechtman. Combinatorial multi-armed bandits with filtered feedback. *ARXIV*, 2017.

[12] Nicolas Gutowski. *Context-aware recommendation systems for cultural events recommendation in Smart Cities*. Theses, Université d'Angers, November 2019.

[13] Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Improved regret bounds for bandit combinatorial optimization. In *NIPS*, pages 12027–12036. Curran Associates, Inc., 2019.

[14] Luc Jaulin, Fabrice Bars, Benoit Clement, Yvon Gallou, Olivier Menage, Olivier Reynet, Jan Sliwka, and Benoit Zerr. Suivi de route pour un robot voilier. *CIFA*, 07 2012.

[15] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670. ACM, 2010.

[16] Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 1245–1253, New York, New York, USA, 20–22 Jun 2016. PMLR.

[17] Jonathan Louedec. *Bandit strategies for recommender systems*. Theses, Université Paul Sabatier - Toulouse III, November 2016.

[18] Alexander Luedtke, Emilie Kaufmann, and Antoine Chambaz. Asymptotically optimal algorithms for multiple play bandits with partial feedback. *ARXIV*, 06 2016.

[19] H. Robbins. Some aspects of the sequential design of experiments. *Bull. of the AMS*, pages 527–535, 1952.

[20] Karthik Abinav Sankararaman. Semi-bandit feedback : A survey of results. , 2016.

[21] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning : An introduction*, volume 1. MIT press Cambridge, 1998.

[22] M. van Aartrijk, C. Tagliola, and P. Adriaans. Ai on the ocean : the robosail project. *ECAI*, pages 653–657, 2002.

Suivi thérapeutique intelligent par recommandation à base d'ontologie et de règles

X. Goblet[1], C. Rey[2]

[1] Jeolis Solutions – Chamalières, France
[2] LIMOS – Université Clermont Auvergne, France

xavier.goblet@lojelis.com, christophe.rey@uca.fr

Résumé

Le domaine de l'Education Thérapeutique du Patient vise à habituer un patient à adopter les bons comportements qui lui permettront au quotidien de renforcer l'action d'un traitement médical. Dans ce contexte, nous présentons une application proposant automatiquement des défis ludiques à un patient souffrant d'obésité. Le processus de recommandation est basé sur l'usage d'une ontologie des patients, des défis et de règles pédagogiques pour le choix des meilleurs défis. Nous utilisons les standards du web sémantique que sont OWL2, SWRL, et le raisonneur Pellet, orchestrés par un programme en Python grâce au module Owlready2. Nous discutons des résultats obtenus ainsi que de l'intérêt d'utiliser des langages déclaratifs tels que OWL2 et SWRL du point de vue du génie logiciel.

Mots-clés

Education thérapeutique du patient, recommandation, OWL2, SWRL, Owlready2, langage déclaratif

Abstract

The field of patient education aims at accustoming a patient to adopt the right behaviors which will allow him daily to reinforce the action of a medical treatment. In this context, we present an application automatically offering fun challenges to an obese patient. The recommendation process is based on the use of an ontology that models the patients, the challenges and on educational rules for choosing the best challenges. We use the semantic web standards OWL2, SWRL, and the Pellet reasoner, orchestrated by a program in Python thanks to the Owlready2 module. We discuss the obtained practical results as well as the interest to use declarative languages such as OWL2 and SWRL in software engineering.

Keywords

patient education, recommendation, OWL2, SWRL, Owlready2, declarative language

1 Introduction

De plus en plus de pathologies chroniques sont prises en charge par des entretiens de suivi avec les professionnels de santé (PS). Ce suivi est efficace si les entretiens sont le moins espacés dans le temps, le plus souvent en présentiel (face-à-face) et si le patient est pleinement acteur de sa thérapie. L'objectif d'un entretien est, après un diagnostic, la mise en place de bonnes pratiques et leur maintien. C'est en cela que l'on peut parler d'éducation thérapeutique du patient (ETP) [6]. Les principaux freins à une ETP efficace sont : le manque de temps des praticiens et patients, les contraintes économiques du monde de la santé, les contraintes écologiques (limiter les déplacements physiques par exemple), une accessibilité grandissante d'informations sur internet (pas toujours fiables), un contenu peu adapté à l'individu, une faible adhésion si la motivation du patient n'est pas présente [6, 18]. Ces freins sont par exemple vérifiés dans le contexte d'une expérience d'ETP sur le bassin clermontois pour prévenir l'obésité infantile dans une famille à risques (au moins un enfant en surpoids) [17]. Pendant six mois, une famille incluse dans ce protocole est suivie sur les axes Nutrition, Activités Physiques Adaptées (APA), et Parentalité/Sociabilité par trois spécialistes qui se déplacent chacun dans la famille pour cinq ateliers de suivi. Les six mois suivant cette première période, la famille n'est plus suivie ; elle rentre alors en phase d'autonomie. Au bout d'un an, le protocole se finit par des entretiens de bilan avec chaque PS et le(s) médecin(s). Il a ainsi été constaté, en l'absence d'un suivi régulier, une baisse de motivation des participants. De plus, la phase d'autonomie est souvent préjudiciable aux progrès observés les six premiers mois. Dans ce contexte, dans le même esprit que [7], un projet de digitalisation de cette ETP est en cours d'élaboration. L'objectif à partir d'une application mobile est d'augmenter singulièrement le suivi des patients, de l'individualiser en utilisant des outils basés sur de l'IA et de maintenir la motivation via une ludification du protocole au moyen de défis personnels en complément des entretiens de suivi. Ce concept de défis (cf. tableau 1) est le pendant des tâches à accomplir par le patient dans le cadre des thérapies cognitives et comportementales (TCC). « Chaque patient progresse à son propre rythme » est aussi le principe fondateur des psychologues cliniciens des TCC. De plus, mis à part la toute première fois, c'est toujours le patient qui choisit son prochain défi. Nous avons conçu « ORALOOS » un Outil de Recommandation d'Activités Ludiques basé sur une Ontologie Opération-

Nutrition (nb=499)	- Buvez seulement de l'eau à vos repas. - La télévision est éteinte lors de vos repas.
Activités Physiques Adaptées (nb=207)	- Descendez un arrêt de bus/métro/tram avant destination et faites-le reste à pieds. - Marchez d'un bon pas sur les trajets courts.
Parentalité /Sociabilité (nb=144)	- Essaie de parler sans pouce ni doudou dans la bouche. - Prenez un moment pour faire une activité avec votre conjoint(e).

TABLE 1 – Des exemples de défis, définis par une psychologue de Jeolis avec les praticiens de Proxob [17].

nelle de Suivi, pour permettre un suivi adaptatif du patient. Nous l'avons défini en utilisant les langages standard du Web sémantique OWL2 et SWRL [9, 10], et le raisonneur Pellet, le tout orchestré par du code Python (via le framework Owlready2 [11]). L'ontologie est dite opérationnelle puisque son contenu va déterminer les traitements effectués par l'application. Les deux contributions de cet article sont : (i) la description de la conception d'ORALOOS, de son ontologie et de son usage dans la gestion et la recommandation des défis personnels pour une meilleure ETP, et (ii) les conséquences, en termes de génie logiciel, de la nature déclarative d'OWL2 et de SWRL, dans le contexte d'une application en langage impératif Python.

2 Travaux antérieurs

Nous nous sommes inspirés des travaux autour de l'utilisation de la logique du premier ordre dans les AEHS (Adaptive Educational Hypermedia Systems) [14, 8]. Chaque système adaptatif, en tant que système hypermédia éducatif, fait des hypothèses sur les documents et leurs relations dans un espace documentaire. Il propose un modèle d'utilisateur pour spécifier diverses caractéristiques d'utilisateurs individuels ou de groupes d'utilisateurs. Pendant l'exécution, il collecte des observations sur les interactions de l'utilisateur. Sur la base de l'organisation de l'espace documentaire sous-jacent, ainsi que des informations du modèle utilisateur et de l'observation du système, une fonctionnalité adaptative est fournie. Un AEHS se conceptualise sous la forme de quatre composants ou espaces : DOCS pour DOCument Space, UM pour User Model, OBS pour OBServations et AC pour Adaptation Component. Cette architecture correspond à nos besoins dans le domaine de l'ETP si l'on remplace les utilisateurs par les patients et si l'on considère que la fonctionnalité adaptative consiste à recommander les meilleurs défis aux patients. Pour la modélisation du domaine, nous nous sommes tournés vers les langages de la logique du premier ordre. En effet, des sous-langages de cette logique permettent la représentation de connaissances dynamiques sous forme de règles et la représentation de connaissances statiques sous forme d'énoncés et de faits. Ces deux types de connaissances sont cruciaux pour un système hypermédia éducatif adaptatif qui doit permettre l'expression et la réutilisation des règles d'adaptation (connaissances dynamiques) dans différents contextes en prenant en compte des métadonnées pour l'adaptation (connaissances statiques) [2]. Dans le domaine du web sémantique, OWL2 et SWRL figurent parmi les langages du premier ordre les plus utilisés et permettent le découplage entre représentation de connaissances dynamiques (en SWRL) et statiques (en OWL2). Le profil EL d'OWL2 [12] assure de plus de bonnes performances dans les raisonnements associés (subsomption, satisfaisabilité, traitement de requêtes), ces derniers étant traitables dans la plupart des cas. SWRL est quant à lui un langage de règles basé sur les clauses définies (sous-ensemble de la logique du premier ordre) étendu par des prédicats qui sont des concepts OWL2. En dehors des langages du web sémantique, nous avons aussi regardé le langage IDP (premier ordre augmenté par la récursivité) [4] qui apparaît très intéressant sur le plan théorique étant donné la possibilité d'exécuter plusieurs types d'inférences différentes. Cependant sur le plan pratique, il semble moins intéressant que les langages du web sémantique étant donné une plus grande difficulté d'apprentissage a priori, l'absence de recommandation W3C associée, ainsi que d'outil de modélisation du type Protégé [15]. De plus on s'interroge sur la possibilité de passage à l'échelle d'un tel langage. C'est pourquoi nous avons choisi OWL2 avec un profil EL et SWRL. Plus précisément, nous utilisons le profil EL d'OWL2 augmenté avec des propriétés d'objet fonctionnelles ainsi qu'avec le constructeur d'union de classes et la possibilité de définir des types intervalles de valeurs ayant une valeur minimale. Nous gagnons donc en expressivité au détriment de la complexité des raisonnements qui perd son caractère traitable mais n'en devient pas prohibitive pour autant en pratique (voir la section 5). Dans la suite, nous supposons le lecteur familier avec OWL2 et SWRL qu'on ne peut redéfinir ici par soucis de synthèse.

3 Approche

En section 3.1, nous expliquons la modélisation déclarative sous la forme des quatre espaces/modèles d'un AEHS en utilisant OWL2 et SWRL. En section 3.2, nous présentons le module python permettant d'accéder à l'ontologie et de raisonner avec elle.

3.1 Modélisation déclarative de l'ontologie

L'ontologie OWL2 et les règles SWRL ont été définies en utilisant l'éditeur Protégé. Comme évoqué précédemment, OWL2 est utilisé pour la modélisation statique du domaine de l'ETP, c'est-à-dire pour décrire les notions qui structurent le domaine (comme par exemple les catégories de patients ou les caractéristiques d'un défi), et SWRL est utilisé pour décrire les aspects dynamiques du domaine, c'est-à-dire les règles métier (notamment celles qui définissent le processus de recommandation de défis aux patients). Dans ce qui suit, nous présentons certaines parties de l'ontologie OWL2 sous forme de diagrammes de classes UML. En effet, la connaissance pouvant être décrite avec le profil EL de OWL2 utilisé correspond aux notions de classe et de relation en UML. Plus précisément, une classe OWL2 est représentée par une classe UML, une propriété OWL2

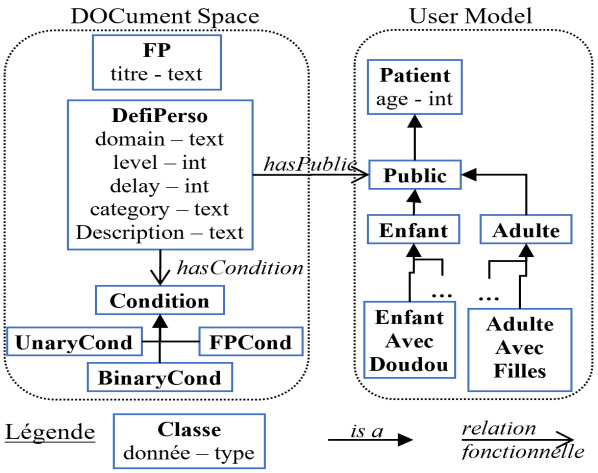


FIGURE 1 – Les espaces Document et Utilisateur

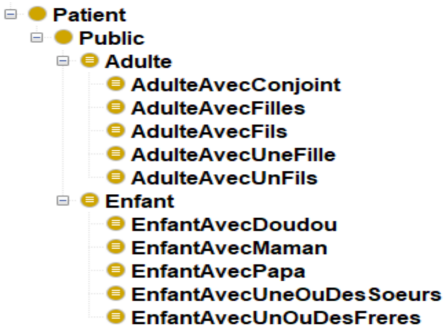


FIGURE 2 – Hiérarchie des différents types de patients

par une relation (ou association) UML [1] , et une relation de subsomption entre classes par une relation d'héritage. Cela nous permet de regrouper en une seule vue graphique, les classes et les propriétés OWL2. Nous voyons maintenant la modélisation des quatre parties du cadre des AEHS.

DOCument Space (cf. Fig.1) : c'est l'espace documentaire de l'ontologie qui contient les défis personnels, ainsi que des fiches pédagogiques (FP). Un défi personnel est défini obligatoirement par un domaine d'application (Nutrition, APA, ou Parentalité), le public concerné par la relation fonctionnelle [2] hasPublic, un niveau de difficulté (niveau croissant avec un entier), et une échéance (entier croissant : 1 → Jour, 2 → WE, 3 → Semaine, 4 → Mois). Les autres données ou relations sont optionnelles : une condition d'exécution du défi par la relation hasCondition (avoir lu une fiche pédagogique ou une caractéristique spécifique du public ciblé), une catégorie du défi dans le domaine et une description textuelle du défi. Dans un domaine, les défis sont définis indépendamment et il n'y pas non plus de dépendances entre les domaines.

User Model (cf. Fig.1) : c'est l'espace des utilisateurs qui porte sur la classe Patient défini par un âge qui permet de classifier le public auquel il appartient. Les classes Adulte et Enfant peuvent être spécialisées chacune en sous-

1. Dans la suite, on parlera de manière interchangeable de propriété (OWL2) ou de relation (UML).

2. Une relation est fonctionnelle lorsqu'il n'y a qu'une seule valeur possible entre les deux classes (1 → 1).

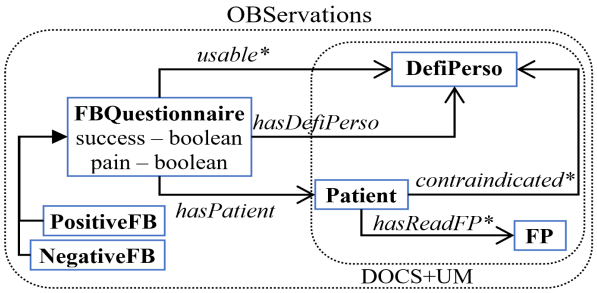


FIGURE 3 – Espace des observations

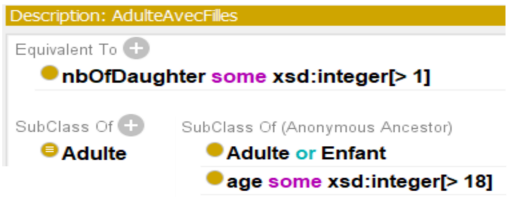


FIGURE 4 – La classe AdulteAvecFilles sous Protégé

classes qui portent une caractéristique optionnelle du patient (exemples : nombre de frères ou de sœurs, avoir un conjoint. . .). La figure 2 montre la hiérarchie des différents types de patients et la figure 4 montre la modélisation de la classe AdulteAvecFilles sous Protégé. Un individu peut appartenir à plusieurs classes : un enfant de 12 ans, vivant uniquement avec sa mère et possédant un doudou est : [Enfant, EnfantAvecDoudou, EnfantAvecMaman]. A contrario un adulte sans conjoint, ni fils, ni fille appartient uniquement à la classe Adulte.

OBServations (cf. Fig.3) : c'est l'espace des observations qui contient toutes les interactions des utilisateurs avec le système. Le premier type d'interaction est le questionnaire d'auto-évaluation (classe FBQuestionnaire, FB pour "feedback"). A échéance d'un défi, un questionnaire d'auto-évaluation est envoyé au patient sur son application mobile. En répondant à une série de questions, il évalue son succès ou non dans la réalisation du défi (booléen success), ainsi que sa difficulté ou non d'exécution (booléen pain). Chaque feed-back est contextualisé par le défi courant (relation fonctionnelle hasDefiPerso) et le patient ayant réalisé le défi (relation fonctionnelle hasPatient). La relation non fonctionnelle [3] usable filtre en amont tous les défis applicables de même domaine et même public que le défi courant à l'instant du feed-back. Ainsi tous les défis qui ne sont plus applicables après un feed-back ne sont pas dans usable. Un défi est applicable dans deux cas. Premièrement, il ne doit pas être contre-indiqué pour le patient. Cela correspond à la relation non fonctionnelle contraindicated. Par exemple un défi réalisé avec succès n'est plus proposé, ou bien un professionnel de santé peut explicitement interdire certains défis pour certains patients. Dans les deux cas le défi est dans contraindicated. Deuxièmement, si ce défi porte une condition, elle doit être vérifiée à cet instant, sinon le défi n'est pas applicable. C'est le cas de la relation

3. Une relation est non fonctionnelle lorsqu'il y a plusieurs valeurs possibles entre les deux classes (1 → *). Ce type de relation exprime aussi une liste dynamique d'informations.

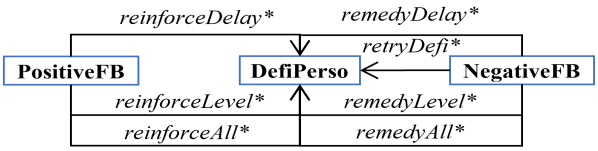


FIGURE 5 – Les relations selon la classe de feed-back

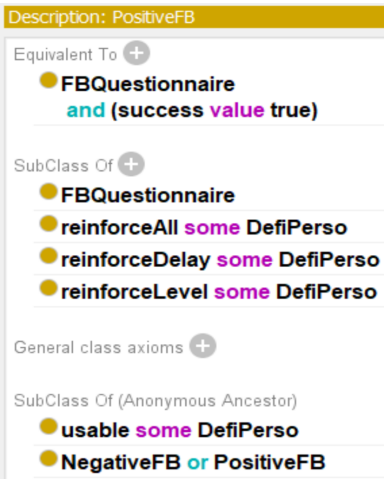


FIGURE 6 – Spécification Protégé de PositiveFB

non fonctionnelle hasReadFP qui concerne les fiches pédagogiques lues/vues par le patient et qui conditionne l'exécution de certains défis. En résumé, la relation usable est une liste dynamique des défis applicables. Comme évoqué à la section 5, mettre à jour cette liste nécessite du code Python. A contrario, une fois usable calculée, seules les règles SWRL (cf. Fig.7) sont nécessaires pour que le système établisse dynamiquement ses recommandations de défis à venir.

Adaptation Component (cf. Fig.5) : c'est l'espace des règles d'adaptation qui implémente les principes de progression au rythme de chaque patient et qui détermine les défis à recommander. Nous avons vu que l'espace des défis est important et que, sans guide, le patient ne progressera pas. Nous avons donc, en relation avec les experts psychologues et pédagogues, défini des règles métier en s'inspirant de la pédagogie behavioriste [5]. Le premier cas est celui où l'on renforce le succès en proposant des défis un peu plus difficiles. Cela se traduit dans la classe DefiPerso par une incrémentation du niveau, de l'échéance ou bien des deux. Cette incrémentation est exprimée par les règles R1, R2 et R3 de la figure 7. Le second cas est celui où l'on remédie à l'échec en proposant des défis un peu moins difficiles. Cela se traduit dans la classe DefiPerso par une décrémentation du niveau, de l'échéance ou bien des deux. Cette décrémentation est exprimée par les règles R5, R6 et R7 de la figure 7. On peut vérifier que les classes et les propriétés non fonctionnelles OWL2 présentes dans les figures 5 et 6 sont bien utilisées en prédicats dans les règles précédemment citées. Pour un même type de feed-back, chaque règle associée à une relation non fonctionnelle peut être activée en même temps (cependant, elles ne contiennent pas les mêmes défis). La règle R4 de la figure 7 implémente le fait qu'un défi évalué avec succès n'est plus proposé au patient en augmentant sa relation non fonctionnelle contraindicated. De même, les règles R8 et R9 de la figure 7 stipulent qu'un défi en échec peut soit être proposé de nouveau (retryDefi) au patient si une cause extérieure sans lien avec sa pathologie en a empêché la réussite (par ex. une météo défavorable), soit n'être plus proposé au patient s'il a échoué et éprouvé une difficulté.

R1 PositiveFB(?fb) ^ hasDefiPerso(?fb, ?df) ^
hasLevel(?df, ?l) ^ hasDelay(?df, ?delay) ^
swrlb:add(?nl, ?l, 1) ^ usable(?fb, ?ndf) ^
hasLevel(?ndf, ?nl) ^ hasDelay(?ndf, ?delay)
-> reinforceLevel(?fb, ?ndf)
R2 PositiveFB(?fb) ^ hasDefiPerso(?fb, ?df) ^
hasDelay(?df, ?delay) ^ hasLevel(?df, ?l) ^
swrlb:add(?ndelay, ?delay, 1) ^ usable(?fb, ?ndf) ^
hasLevel(?ndf, ?l) ^ hasDelay(?ndf, ?ndelay)
-> reinforceDelay(?fb, ?ndf)
R3 hasLevel(?ndf, ?nl) ^ usable(?fb, ?ndf) ^
hasDefiPerso(?fb, ?df) ^ swrlb:add(?nl, ?l, 1) ^
hasDelay(?df, ?delay) ^ hasDelay(?ndf, ?ndelay) ^
swrlb:add(?ndelay, ?delay, 1) ^ PositiveFB(?fb) ^
hasLevel(?df, ?l)
-> reinforceAll(?fb, ?ndf)
R4 PositiveFB(?fb) ^ hasDefiPerso(?fb, ?df) ^
hasPatient(?fb, ?pat)
-> contraindicated(?pat, ?df)
R5 hasLevel(?ndf, ?nl) ^ usable(?fb, ?ndf) ^
hasDefiPerso(?fb, ?df) ^ swrlb:greaterThan(?l, 1) ^
hasDelay(?ndf, ?delay) ^ hasDelay(?df, ?delay) ^
swrlb:subtract(?nl, ?l, 1) ^ NegativeFB(?fb) ^
hasLevel(?df, ?l)
-> remedyLevel(?fb, ?ndf)
R6 hasLevel(?ndf, ?l) ^ usable(?fb, ?ndf) ^
hasDefiPerso(?fb, ?df) ^
swrlb:greaterThan(?delay, 1) ^
swrlb:subtract(?ndelay, ?delay, 1) ^
hasDelay(?df, ?delay) ^
hasDelay(?ndf, ?ndelay) ^ NegativeFB(?fb) ^
hasLevel(?df, ?l)
-> remedyDelay(?fb, ?ndf)
R7 hasLevel(?ndf, ?nl) ^ usable(?fb, ?ndf) ^
hasDefiPerso(?fb, ?df) ^ swrlb:greaterThan(?l, 1) ^
swrlb:greaterThan(?delay, 1) ^
swrlb:subtract(?ndelay, ?delay, 1) ^
hasDelay(?df, ?delay) ^ swrlb:subtract(?nl, ?l, 1) ^
hasDelay(?ndf, ?ndelay) ^ NegativeFB(?fb) ^
hasLevel(?df, ?l)
-> remedyAll(?fb, ?ndf)
R8 NegativeFB(?fb) ^ pain(?fb, false) ^
hasDefiPerso(?fb, ?df)
-> retryDefi(?fb, ?df)
R9 NegativeFB(?fb) ^ pain(?fb, true) ^
hasDefiPerso(?fb, ?df) ^ hasPatient(?fb, ?pat)
-> contraindicated(?pat, ?df)

FIGURE 7 – Les règles SWRL de recommandation de défis. Les paramètres ?x sont des variables, les prédicats sont des classes ou des relations OWL2, et swrlb préfixe des fonctions prédéfinies SWRL

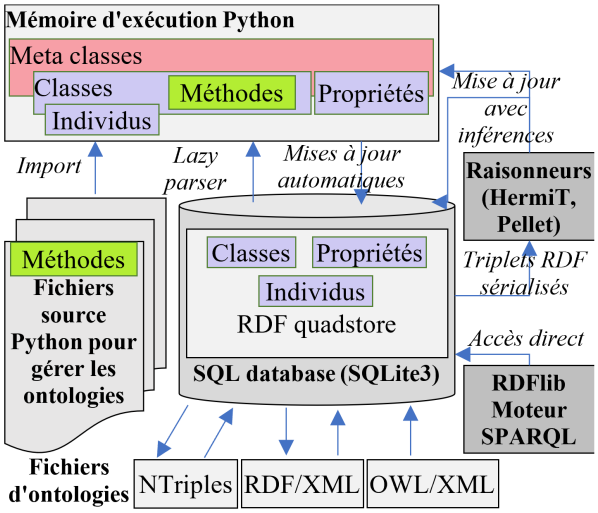


FIGURE 8 – Architecture d'Owlready2 extrait de [11]

```
with onto:
 class FBQuestionnaire(Thing):
  def preFilter(self):
   dom = self.hasDefiPerso.domain
   pub = self.hasDefiPerso.hasPublic
   contraindicates = self.hasPatient.contraindicated
   usables = onto.search(type=onto.DefiPerso,
                         domain=dom,
                         hasPublic=pub)
   # le defi courant doit etre enleve
   current = self.hasDefiPerso
   usables.remove(current)
   if len(contraindicates) != 0:
    for df in usables:
     if df in contraindicates:
      usables.remove(df)
   for df in usables:
    cond = df.hasCondition
    if (cond is None)
       or
       (cond.checkApplicability(self.hasPatient)):
     self.usable.append(df)
   return True
```

FIGURE 9 – Classe FBQuestionnaire

3.2 Programmation impérative d'ontologie

Nous avons programmé notre application en langage Python, avec l'aide d'Owlready2 un module Python (version 3 du langage) facilitant la programmation d'ontologies (accès aux données et connaissances, exécutions d'inférences, lecture et écriture avec les formats standards du Web sémantique). Ce module inclut nativement une base de triplets (appelée quadstore) et les raisonneurs HermiT et Pellet[4]. En plus des raisonnements fournis par ces deux raisonneurs, l'utilisateur d'Owlready2 peut en développer d'autres en Python. L'architecture d'Owlready2 est présentée à la figure 8. Avec Owlready2, toutes les classes OWL2 correspondent à des classes Python. Nous illustrons ce principe avec la classe OWL2 FBQuestionnaire contenant la méthode preFilter() qui met à jour la relation usable. La figure 9 montre le code python de cette classe.

Par ailleurs les classes FPCond et UnaryCond implémentent la vérification dynamique d'une condition d'applicabilité d'un défi (cf. Fig.10). On ne peut pas implémenter cette vérification grâce à des règles SWRL car il faudrait pouvoir exprimer des négations dans le corps des règles, voire même écrire des règles hors de la logique du premier ordre (cf. section 5).

4. Voir les sites web `http://www.hermit-reasoner.com/` et `https://github.com/stardog-union/pellet`.

```
class FPCond(Thing):
 def checkApplicability(self, patient):
  fp = self.hasFP
  readings = list(patient.hasReadFP)
  if fp in readings:
   return True
  return False

class UnaryCond(Thing):
 def checkApplicability(self, patient):
  classe = self.type
  subclasses = str(patient.is_a)
  if classe in subclasses:
   return True
  return False
```

FIGURE 10 – Les classes FPCond et UnaryCond

```
with onto:
 pat = onto.Patient(patient.ident,
                    age=patient.age,
                    **patient.kargs)
 try:
  sync_reasoner_pellet([onto],
          fer_property_values = True,
          infer_data_property_values = True)
 except OwlReadyInconsistentOntologyError:
  destroy_entity(pat)
  raise HTTPException(status_code=500,
            detail="Inconsistent ontology!"+
            "Check if patient age " +
            "is aligned with optional" +
            "characteristics...")
 return patient
```

FIGURE 11 – Création et classement d'un patient

La figure 11 illustre le raisonnement permettant de classer le patient en adulte ou enfant ; si l'individu possède des caractéristiques optionnelles, le raisonnement affine la classification dans l'une des deux branches de la hiérarchie des patients (cf. Fig.2), tout en gérant les inconsistantes éventuelles (comme par exemple le fait qu'un enfant n'a pas de conjoint).

Le second raisonnement est la recherche des réponses à une requête et est implémenté en utilisant le raisonneur Pellet couplé à notre fonction ad-hoc preFilter() (cf. Fig.12) : la première étape consiste à créer un individu de classe FBQuestionnaire avec les différents paramètres d'entrées. La seconde étape consiste à appeler la méthode preFilter() qui permet de contextualiser dynamiquement le feed-back en calculant les défis applicables du domaine pour le patient (c'est-à-dire en mettant à jour la relation usable). La troisième étape consiste à lancer le raisonneur Pellet pour trouver le type de feed-back (de la classe FB) positif ou négatif et appliquer en conséquence les règles SWRL.

4 Intégration en application de suivi

Dans le cadre d'une architecture orientée services, ORALOOS est un composant micro-service qui interagit avec d'autres applications via une API Web REST comme illustré sur la figure 13. Il utilise pleinement la base de données nativement intégrée d'Owlready2 pour persister la base ontologique.

```
with onto:
 fdbk = onto.FBQuestionnaire(fb.fbID,
                             success=fb.success,
                             pain=fb.pain,
                             hasDefiPerso=defi,
                             hasPatient= pat)
 fdbk.preFilter()
 try:
  sync_reasoner_pellet([onto],
            infer_property_values = True,
            infer_data_property_values = True)
  except OwlReadyInconsistentOntologyError:
   raise HTTPException(status_code=500,
                       detail="Inconsistent ontology!")
  if isinstance(fdbk, onto.PositiveFB):
   response = PosFBAPI(ID = fb.fbID,
              patientID = fb.patientID,
                   defiID = fb.defiID,
    reinforceAll = format_defis(fdbk.reinforceAll),
    reinforceLevel = format_defis(fdbk.reinforceLevel),
    reinforceDelay = format_defis(fdbk.reinforceDelay)
                    )
  else:
   response = NegFBAPI(ID = fb.fbID,
              patientID = fb.patientID,
                   defiID = fb.defiID,
    remedyAll = format_defis(fdbk.remedyAll),
    remedyLevel = format_defis(fdbk.remedyLevel),
    remedyDelay = format_defis(fdbk.remedyDelay),
    retry = format_defis(fdbk.retryDefi)
                    )
return response
```

FIGURE 12 – Mise à jour des défis à recommander

4.1 API Web

Les principales interactions entre le back office et Oraloos permettent de peupler la base d'individus : createPatient, createChallenge, createFeedBack. Ces requêtes sont initiées par le serveur BO. Seules createPatient et createFeedBack déclenchent le raisonneur Pellet comme présenté en section 3.2. La réponse principale d'ORALOOS, contient les recommandations calculées par les règles dans les différentes listes, selon le type de feed-back. C'est le serveur BO qui les notifie alors au Patient.

4.2 Premiers résultats

Le tableau 2 montrent quelques métriques issues de Protégé. La figure (a) du tableau 2 concerne l'ontologie avant insertion des défis. Après peuplement de la base avec les 850 défis, et après inférence, nous constatons normalement une augmentation du nombre d'individus, mais aussi une augmentation importante du nombre d'axiomes (figure (b) du tableau 2). Evidemment, il y a une augmentation d'individus et d'axiomes lors des créations des patients et des feedbacks. Nos premiers tests montrent un temps de raisonnement moyen de l'ordre de 3-4 secondes lors de la création d'un patient et un temps de 6-7 secondes lors de la création d'un feed-back. Ces temps peuvent sembler importants mais nous rappelons que notre besoin de recommandations n'est pas en temps réel ; l'échéance minimale d'un défi est la journée. Ces résultats sont à confirmer avec une utilisation par des patients en conditions réelles lorsque l'application sera industrialisée.

5 Apports et limites de l'ontologie

Le principal avantage lié à l'usage d'une ontologie et d'un raisonneur dans notre application d'ETP est l'ajout d'une dimension déclarative forte au développement et à la maintenance de cette application. Cela permet ainsi d'éviter

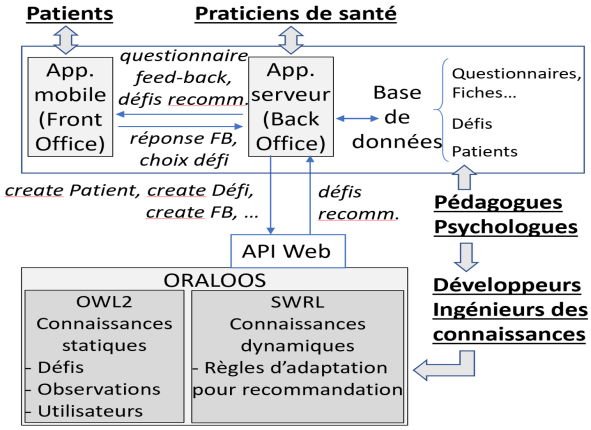


FIGURE 13 – Architecture de l'application de suivi à distance pour une ETP interactive, ludique et adaptative

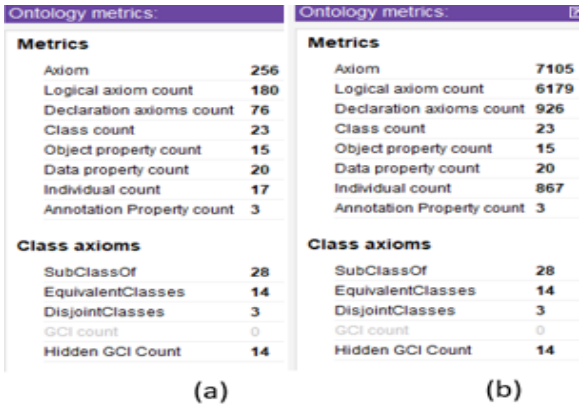


TABLE 2 – Quelques métriques Protégé

l'écriture d'un code Python complexe puisque toute la partie nécessitant la classification des connaissances et données ainsi que le requêtage avec les règles sont entièrement pris en charge par le raisonneur. Par ailleurs, la connaissance stockée dans l'ontologie est facilement extensible et maintenable puisqu'il suffit d'éditer le fichier avec un éditeur dédié comme Protégé.

Nous mettons ici en avant le fait que le niveau de déclarativité permis par le couple (ontologie avec règles, raisonneur) va bien au-delà de celui permis par un couple (fichiers texte de paramètres, bibliothèque de fonctions en python). En effet, grâce aux règles notamment, le développeur (cf. Fig.13) a la possibilité, par exemple, d'ajouter une nouvelle règle métier à l'application et ce déclarativement, c'est-à-dire en décrivant seulement les états initial et final de l'application de cette règle. Avec un langage impératif tel que Python, cela correspondrait à l'ajout d'une nouvelle fonction au sein d'une bibliothèque où, en plus des états initial et final associés à la fonction, nous devrions aussi décrire la gestion de la mémoire et les structures de contrôle nécessaires au calcul fait par la fonction. En résumé, nous pouvons étendre ou modifier très simplement les règles métier sans manipulations compliquées dans le code de l'application. Une autre conséquence intéressante de cette approche déclarative en OWL2 est la possibilité d'ajouter de nouveaux concepts du domaine uti-

lisables dans de nouvelles règles. C'est donc bien plus puissant qu'un fichier texte contenant des paramètres qu'on ne peut que modifier sans pouvoir en ajouter d'autres non prévus par le développeur d'origine. Ainsi une telle application nécessite-t-elle moins de développeurs, de temps de développement et de maintenance. Cependant, l'ontologie OWL2 avec règles SWRL présente quelques limites. Comme évoqué en section 3.2, certains traitements, notamment de réification , sortent du cadre de la logique du premier ordre. Par exemple, la méthode checkApplicability() de la classe FPCond de la figure 10 pourrait être traduite par la règle suivante :
UnaryCond(?cond) ^ type(?cond, ?classe) ^?classe(?pat) -> applicable(?cond, ?pat)
où la variable ?classe est tantôt utilisée en objet du triplet type(?cond, ?classe) de propriété type, tantôt utilisée en propriété dans le triplet ?classe(?pat). Une autre limite importante citée précédemment est celle de l'impossibilité d'exprimer la négation dans le corps d'une règle SWRL. Par exemple, à la figure 9, le calcul d'une différence ensembliste avec l'instruction :

```
if df in contraindicates:
        usables.remove(df)
```

pourrait être exprimée avec la règle non SWRL (avec ?x0 un FBQuestionnaire et ?x un DefiPerso) :
NOT contraindicates(?x0, ?x) -> usables(?x0, ?x)
Enfin, l'impossibilité bien connue d'exprimer certains cas de jointures en OWL2 impose de décrire certaines connaissances statiques sous forme de règles SWRL. Ceci peut être une gêne dans la mesure où cela brise le principe de modélisation réservant les règles SWRL à la modélisation des règles métier, c'est-à-dire à la partie dynamique des connaissances (cf. section 3.1).

6 Perspectives

Un axe de nos futurs travaux est de conceptualiser les notions de profil et de parcours Patient. Nous allons évidemment nous intéresser au choix du défi par le patient suite aux recommandations faites par ORALOOS. Comme le feed-back mémorise de quelle relation non fonctionnelle est issue le défi choisi par le patient, nous allons définir avec les psychologues et pédagogues des profils pour le patient, par exemple : performant, persévérant, prudent, etc. Ensuite, en analysant plus finement le parcours de chaque patient et son profil, à partir de modèles machine learning ou de métarègles, nous pourrons moduler les règles de recommandation behavioristes.
Par ailleurs, nous voulons améliorer la modélisation statique OWL2 de l'ontologie : (i) introduire une hiérarchie dans les défis et l'espace documentaire, (ii) étudier le concept d'échéance (delay/deadline) d'un défi qui peut être interprété différemment selon les domaines et/ou professionnels de santé, (iii) retravailler la modélisation des conditions et (iv) conceptualiser la notion d'espace pour améliorer l'organisation de l'ontologie. Nous discutons maintenant des moyens d'atteindre un plus grand niveau de déclarativité en proposant des solutions aux problèmes évoqués en section 5.

Le problème du besoin de réification pourrait être résolu en utilisant la sémantique basée sur RDF de OWL2 qui permet des cas de réification. Cette sémantique n'est cependant pas la même que celle utilisée par les règles SWRL ou par les raisonneurs HermiT et Pellet inclus dans Owlready2 (qui est la sémantique directe d'OWL2 basée sur la théorie des modèles). Utiliser la sémantique basée sur RDF apparaît donc très compliqué car cela implique de changer une bonne partie du code de l'application et de gérer la cohabitation de plusieurs sémantiques différentes. Ainsi, soit nous changeons, dans la mesure du possible, la modélisation de l'ontologie pour éviter les cas de réification, soit nous gardons le code Python permettant ce genre de traitement.
La question de l'ajout de la négation dans les règles SWRL est plus ouverte. La première solution serait d'utiliser dans les règles des noms de prédicats définis dans la partie OWL2 de l'ontologie par des descriptions de classes utilisant la négation (c'est possible avec OWL2). De fait, nous augmentons alors la complexité de raisonnement relative à la partie OWL2 (en s'éloignant un peu plus de la traitabilité du profil EL). De plus, les raisonneurs HermiT et Pellet ne peuvent pas raisonner avec des négations dans des descriptions de propriétés. Ainsi cette solution ne couvrirait pas les cas de négation de prédicats binaires dans les règles.
La deuxième solution pour ajouter la négation dans les règles est d'adopter une stratégie de raisonnement différente de celle des raisonneurs HermiT et Pellet. En effet, il est possible de traduire les connaissances d'une ontologie OWL2 en profil EL (sans les extensions évoquées en section 2). en règles existentielles (ou datalog$^{\pm}$), qui sont des règles SWRL auxquelles nous ajoutons des variables existentielles en tête [13]. Il existe depuis quelques années des travaux permettant d'étendre à la négation les algorithmes des deux grandes familles permettant de raisonner avec ces règles existentielles (les algorithmes de saturation et les algorithmes de réécriture de requêtes) [1]. Cependant, ces travaux restent pour le moment essentiellement théoriques. En dehors de ces algorithmes, une autre possibilité est de skolémiser les variables existentielles des règles, c'est-à-dire de les remplacer par des termes fonctionnels uniques pour chaque variable. Nous obtenons alors des programmes datalog avec des fonctions, passant ainsi dans le monde de la programmation logique où de nombreux travaux décrivent l'ajout de la négation aux règles. Citons par exemple : (i) la SLDNF-resolution avec stratification et complétion du programme [16], (ii) la sémantique bien fondée [16], et même l'answer set programming (ASP) avec la sémantique des modèles stables [3]. Le choix de l'approche dépend alors de la sémantique de la négation souhaitée. Globalement cette solution impose donc de changer de raisonneur (et donc sans doute de framework et de langage de programmation) et de corriger la partie OWL2 de l'ontologie pour rester dans le profil EL strict.
Le dernier problème évoqué en section 5 est celui de la perte de séparation nette entre la modélisation statique en OWL2 et la modélisation dynamique en SWRL, dans le cas où des connaissances statiques ne peuvent être exprimées que par des règles. Une solution peut être d'abandon-

ner cette distinction pour s'orienter vers une modélisation initiale de toute la connaissance sous forme de règles existentielles. Mais alors en plus de perdre le principe de séparation statique-dynamique permettant une modélisation claire de la connaissance, nous abandonnons aussi tous les standards et outils du web sémantiques dont la diffusion est de plus en plus large, notamment en lien avec les bases de données de graphes et plus généralement avec le domaine des linked data.

7 Conclusion

Nous montrons comment créer une application d'ETP basée sur une ontologie avec recommandation automatique et adaptative de défis. Les temps d'exécution sont encourageants et permettent une mise en production de l'application à court terme. D'un point de vue génie logiciel, la présence d'une ontologie et d'un raisonneur amènent une dimension déclarative au développement, se traduisant par un code impératif plus petit, des besoins en développeurs moins grands et des temps de développement plus courts. Aller vers une plus grande déclarativité imposerait d'étendre l'expressivité des langages OWL2 et SWRL (en ajoutant par exemple la négation aux règles). Cela pourrait avoir comme conséquence une remise en question de l'application en profondeur (avec un changement possible de langage de programmation). De même, une mise à jour de l'ontologie vers le profil EL sans extension pourrait être nécessaire, ce qui, paradoxalement, constitue une réduction de l'expressivité. Ainsi la question importante devient-elle celle du meilleur compromis entre expressivité et déclarativité. Dans cette optique, le langage des règles existentielles est intéressant puisqu'il peut généraliser SWRL et le profil EL d'OWL2. Cependant l'ajout de la négation reste encore un problème ouvert et principalement théorique.

Références

[1] M. Alviano, M. Morak, and A. Pieris. Stable model semantics for tuple-generating dependencies revisited. In *36th ACM Symposium on Principles of Database Systems (PODS'17)*, page 377–388, 2017.

[2] S. Angeletou, M. Rigou, and S. Sirmakessis. A logic-based approach to learner assessment. In *the 1st Int. Conf. on Educational Technologies, Tenerife, Spain*, pages 200–205, December 2005.

[3] J.-F. Baget, L. Garcia, F. Garreau, C. Lefevre, S. Rocher, and I. Stéphan. Bringing existential variables in answer set programming and bringing non-monotony in existential rules : two sides of the same coin. *Annals of mathematics and artificial intelligence*, 82(1-3) :3–41, 2018.

[4] B. De Cat, B. Bogaerts, M. Bruynooghe, and M. Denecker. Predicate Logic as a Modelling Language : The IDP System. *CoRR*, 2014.

[5] A. Giordan. Education thérapeutique du patient : les grands modèles pédagogiques qui les sous-tendent. *Médecin des maladies métaboliques*, 4(3), 2010.

[6] S. Hamy-Shoshany. Freins et dynamiques à la mise en place de programmes d'éducation thérapeutique du patient en soins primaires. *Thèse de Docteur en Médecine, Université Claude Bernard - Lyon 1*, 2015.

[7] B. Hansel, P. Giral, L. Gambotti, A. Lafourcade, G. Peres, C. Filipecki, D. Kadouch, A. Hartemann, J.-M. Oppert, E. Bruckert, M. Marre, A. Bruneel, E. Duchene, and R. Roussel. A fully automated web-based program improves lifestyle habits and hba1c in patients with type 2 diabetes and abdominal obesity : Randomized trial of patient e-coaching nutritional support (the anode study). *J Med Internet Res*, 19(11), 2017.

[8] N. Henze and W. Nejdl. Logically characterizing adaptive educational hypermedia systems. In *In International Workshop on Adaptive Hypermedia and Adaptive Web-based Systems (AH 2003)*.

[9] P. Hitzler, M. Krötzsch, B. Parsia, P. Patel-Schneider, and S. Rudolph. Owl 2 web ontology language primer (second edition), 2012. https ://www.w3.org/TR/owl2-overview/.

[10] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosofand, and M. Dean. SWRL : A semantic web rule language combining OWL and RuleML, 2004. http ://www.w3.org/Submission/SWRL/.

[11] J.-B. Lamy. Owlready : Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, 80 :11 – 28, July 2017.

[12] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz. Owl 2 web ontology language : Profiles (second edition), 2012.

[13] M.-L. Mugnier and M. Thomazo. An introduction to ontology-based query answering with existential rules. In *Reasoning Web Summer School*, 2014.

[14] C. Mulwa, S. Lawless, M. Sharp, I. Arnedillo-Sanchez, and V. Wade. Adaptive educational hypermedia systems in technology enhanced learning : A literature review. SIGITE '10, page 73–84, 2010.

[15] M. A. Musen. The protégé project : A look back and a look forward. *AI Matters*, 1(4) :4–12, June 2015.

[16] U. Nilsson and J. Maluszynski. *Logic, programming and PROLOG (2ed)*. 2001.

[17] R. Rigondet, A. Rigal, C. Desblès, Q. Lesaichot, J. Masurier, C. Cardenoux, D. Thivel, B. Pereira, C. Lambert, Y. Boirie, and M. Miolanne. Accompagnement familial à domicile et de proximité de l'obésité infanto-juvénile proxob : étude pilote de faisabilité en recherche interventionnelle en santé. *Nutrition Clinique et Métabolisme*, 33(1) :83 – 84, 2019.

[18] K. Sandid. Usage des nouvelles technologies en éducation thérapeutique du patient., 2019. https ://www.slideshare.net/KarimSandid/.

Définition d'une méthodologie d'indexation de documents textuels par étiquetage de séquences : application aux offres d'emploi

H.Ramdani[1 2 3] A.Brun[2] E.Bonjour [1] D.Monticolo [1]

[1] Equipe de Recherche sur les Processus Innovatifs, Université de Lorraine, ERPI
[2] Laboratoire lorrain de recherche en informatique et ses applications, Universite de Lorraine, LORIA
[3] Xtramile, 11 Rempart Saint-Thiebault, 57000 Metz

h.ramdani@myxtramile.com

Résumé

Automatiser la tâche de mise en correspondance d'une offre d'emploi et d'un CV est un sujet d'intérêt dans un nombre grandissant de travaux. Elle repose sur une identification automatique du profil souhaité dans l'offre et du profil candidat dans le CV. Dans cet article, nous nous intéressons à l'identification du profil souhaité dans l'offre et considérons cette identification comme un problème d'indexation de document textuel semi-structuré , rédigé en langage naturel et dont le vocabulaire est évolutif. Les approches d'indexation de documents présentées dans la littérature prennent généralement en compte une de ces caractéristiques, mais pas les trois à la fois. Dans cet article, nous proposons une méthodologie pour l'indexation automatique de documents, reposant sur l'étiquetage de séquences et qui prend en compte l'ensemble de ces points. Cette méthodologie repose sur la création manuelle d'un corpus étiqueté, étape de la plus haute importance pour obtenir un étiquetage automatique de qualité. Cette méthodologie est validée au travers d'algorithmes d'apprentissage supervisé sur un corpus réel d'offres d'emploi.

Mots-clés

Indexation de textes, étiquetage de séquences, e-recrutement, indexation d'offres d'emploi, apprentissage supervisé, méthodologie.

Abstract

A number of research studies have focused on automate the matching of a job offers to CVs. This relies on the identification of the profile sought by the recruiter in the offer and its matching with the candidate profile in the CV. In this paper we suggest a methodology for indexing semi-structured textual documents written in natural language. Document indexing approaches presented in the literature generally take into account one of these characteristics, but not all three at once. In this paper, we propose a methodology for document indexing, based on sequence labeling, that allows all of these points to be taken into account. This methodology is based on the creation of a learning corpus. This methodology is validated through supervised learning algorithms on a real corpus of job offers.

Keywords

Automatic text indexing, sequence labeling, e-recruitment, job offer indexing, supervised learning, methodology.

1 Introduction

Le domaine du recrutement, à l'instar de nombreux autres domaines, évolue rapidement vers une digitalisation de ses activités. Du côté du recruteur, recruter se déroule classiquement en trois étapes : 1) diffuser l'offre d'emploi, 2) réceptionner les CV et les mettre en correspondance avec l'offre, 3) sélectionner les meilleurs candidats. A l'instar du recrutement, le e-recrutement adopte ces trois mêmes étapes, mais a la caractéristique de se dérouler majoritairement sur le Web. En effet, les offres et les CV sont désormais sous forme numérique et diffusés en ligne [8]. Dans ce contexte, il est naturel de chercher à automatiser les différentes étapes du recrutement, en particulier la mise en correspondance des CV et des offres d'emploi, étape la plus longue. Pour permettre cette mise en correspondance, il est nécessaire d'identifier dans un premier temps le profil dont il est question que ce soit dans l'offre d'emploi ou dans chaque CV. Le profil souhaité et le profil candidat peuvent être exprimés au travers de la formation, des compétences professionnelles, de qualités personnelles, etc. [4].

Nous proposons de voir la tâche d'identification du profil comme un problème d'indexation de document, c'est-à-dire identifier les informations pertinentes dans le document afin de faciliter l'accès à son contenu [7]. D'un point de vue applicatif, l'objectif est donc d'identifier les informations présentes dans une offre d'emploi, qui permettront d'identifier au mieux le profil souhaité.

Une offre est un document qui a un certain nombre de caractéristiques. Une offre est certes rédigée librement en langage naturel, mais une offre se présente sous la forme d'un texte semi-structuré dans sa rédaction et le vocabulaire utilisé évolue. En effet, une offre est un enchaînement de sections représentant chacune un type d'information. Néan-

moins, l'ordre des sections peut varier d'une offre d'emploi à une autre. De plus, avec l'apparition de nouveaux métiers, de nouvelles compétences, etc. le vocabulaire utilisé tend à évoluer. Par ailleurs, certaines informations partagent leur vocabulaire, c'est le cas des compétences techniques et des missions, ce qui peut provoquer une ambiguïté lors de l'indexation d'une offre. Les approches classiques d'indexation ne peuvent donc pas être utilisées. Plusieurs travaux de la littérature traitent de la mise en correspondance d'une offre d'emploi et de CV reponsant sur une première étape d'indexation. Les méthodes d'indexation exploitées sont cependant limitées puisqu'elles ne prennent pas en compte la structure du texte et ne considèrent pas l'évolution possible du vocabulaire et leur ambiguïté. Pourtant, la prise en compte de ces deux caractéristiques pourrait améliorer la qualité de l'indexation de l'offre. L'hypothèse sur laquelle repose le travail mené dans cet article est que l'étiquetage de séquences est une approche qui pourrait non seulement prendre en compte la structure du texte, mais également d'être robuste à l'évolution du vocabulaire. En effet, les tâches d'étiquetage de séquences sont depuis longtemps d'un intérêt particulier pour le TAL : étiquetage en parties du discours, annotation sémantique, extraction d'informations, etc. [3]. Cependant, elle n'a jamais été utilisée sur les offres d'emploi. La tâche d'étiquetage de séquences considère un texte comme un enchaînement de mots ayant un sens dans un contexte : les séquences, auxquelles il cherche à attribuer une étiquette. Dans notre travail, nous considérons que les étiquettes et les séquences associées forment l'indexation du document. Partant d'un corpus étiqueté en séquences (un corpus d'apprentissage), les travaux de la littérature en étiquetage de séquences analysent et comparent différents algorithmes d'apprentissage supervisé afin d'améliorer la qualité de l'indexation automatique. Le corpus d'apprentissage, et en particulier l'étiquetage utilisé, est pourtant important et influence la qualité de l'étiquetage automatique, mais il n'est jamais considéré. Par conséquent, le travail que nous proposons ici se concentre sur la création d'un corpus d'apprentissage étiqueté de qualité et à étudier son impact sur la qualité de l'étiquetage automatique afférent. Il vise donc à proposer une méthodologie de création manuelle de corpus étiqueté, dans l'objectif d'un étiquetage automatique de séquences. Cette méthodologie est validée expérimentalement au travers d'algorithmes classiques d'apprentissage supervisé, sur un corpus réel d'offres d'emploi. Nous introduisons de plus dans la suite de cet article de nouvelles étiquettes pour l'identification d'un profil à partir d'une offre d'emploi.

La section 1 présente un état de l'art de l'indexation de documents. La section 2 définit et propose la méthodologie d'indexation de documents semi-structurés. La section 3 présente l'expérimentation menée sur un corpus réel d'offres d'emploi. La section 4 est consacrée à la discussion et aux perspectives de recherche.

2 Indexation de documents

Cette première section est dédiée à la présentation des principales approches de la littérature adoptées pour l'indexation de documents textuels.

2.1 L'approche par règles

L'indexation par règles vise à associer une catégorie à chaque information contenue dans un texte. Les paires catégories/séquences représentent l'indexation des documents. Trois techniques sont classiquement utilisées :
Les expressions régulières, qui offrent la possibilité d'identifier dans le document des chaînes de caractères, construites avec des caractères ou des méta-caractères [2]. C'est par exemple le cas des dates, codes postaux, expériences etc. L'utilisation d'expressions régulières permet ensuite d'associer une chaîne identifiée à une catégorie ou étiquette. Les expressions régulières peuvent être utilisées sur un texte rédigé en langage naturel, quelle que soit sa structure.
Les dictionnaires de mots, qui sont une énumération de mots associés à une étiquette [16]. Citons par exemple dans les offres d'emploi les mots CDI, CDD, alternance, etc. qui sont associés à l'étiquette "type de contrat". L'utilisation d'un dictionnaire permet donc d'associer une étiquette à une séquence (ou à un mot) identifiée dans le document. Ces paires catégories/séquences représentent à nouveau l'indexation du document. Comme précédemment, les dictionnaires de mots ont l'avantage de pouvoir être utilisés sur des documents textuels rédigés en langage naturel, quelle que soit sa structure.
Contrairement aux deux approches précédentes, *les règles d'indexation* sont une technique qui se situe au niveau du document et non au niveau des mots et de leurs enchaînements. Ces règles représentent la structure des documents au travers d'informations représentant l'enchaînement d'étiquettes au sein des documents [13]. Un exemple de règle peut être : une offre d'emploi contient toujours la description de l'entreprise en premier paragraphe et est suivie de la description des missions. Les règles d'indexation permettent ainsi d'associer une séquence à une catégorie définie dans les règles. Les règles d'indexation peuvent donc être utilisées sur des documents textuels rédigés en langage naturel et contrairement aux approches précédentes, sur des structurés. Cette technique est particulièrement adéquate lorsque l'on souhaite indexer des documents dont on a une connaissance relative à l'enchaînement des séquences et à la structure du texte.
Ces trois techniques peuvent être utilisées séparément ou conjointement lorsque la nature de l'information connue porte sur le texte ou sur les informations souhaitées dans l'indexation, que les informations à extraire peuvent être énumérées (dans un dictionnaire ou au travers d'une expression régulière) et que le document à indexer est structuré (et représentable par des règles). Cette approche, utilisée dans plusieurs domaines tels que le recrutement [2]

ou la santé [1], est cependant limitée puisqu'elle ne prend en compte l'évolutivité du domaine et des termes ou encore l'ambiguïté associée à la langue et ne peut pas être appliquée sur des documents dont la structure est incertaine (semi-structurés) (dans le cas de l'utilisation des règles d'indexation).

2.2 L'approche par ontologie

Une ontologie peut être décrite comme un vocabulaire organisé, comme pour un dictionnaire, mais qui donne en plus une représentation formelle des connaissances d'un domaine. Le plus souvent, il s'agit d'un ensemble de concepts et instances (valeurs), organisés hiérarchiquement et structurés par des relations. L'indexation de documents textuels par ontologie consiste à déterminer le(s) concept(s) au(x)quel(s) le texte à indexer est associé. Les travaux de la littérature utilisent soit des ontologies existantes [20], soit une ontologie qu'ils construisent [6]. Construire sa propre ontologie a l'avantage de créer un vocabulaire et des relations propres au domaine à indexer. Néanmoins, ce choix est très coûteux en temps puisque cette création par un expert du domaine est manuelle [15]. Utiliser une ontologie existante offre un gain de temps. Néanmoins, il se peut que cette ontologie ne contienne pas le niveau de détails souhaité ou ne corresponde pas parfaitement au domaine d'intérêt. Dans les deux cas, à chaque nouveau concept apparaissant, il est nécessaire de l'ajouter dans l'ontologie. Le coût associé à cet ajout est également conséquent. Par conséquent, l'approche par ontologie, bien que précise, ne permet pas de gérer simplement et automatiquement l'évolutivité du vocabulaire et des termes. Par ailleurs, elle ne permet pas non plus de prendre en compte la structure des documents. Cependant, elle permet de gérer le langage naturel.

2.3 L'approche par étiquetage de séquences

Principe. L'approche d'indexation par étiquetage de séquences est à l'origine de nombreux travaux, en particulier dans le domaine du TAL [10, 5]. Cette approche peut être appliquée sur tout type de document textuel : rédigé en langage naturel, structuré, non structuré et semi-structuré. L'indexation par étiquetage de séquences considère un texte comme un enchaînement de séquences. L'objectif est d'assigner une étiquette par séquence [17]. Dans l'exemple "nous recherchons un développeur back-end pour notre application", la séquence "développeur back-end" est étiquetée par "poste". Comme précédemment, les paires étiquette/séquence représentent l'indexation du document. La mise en œuvre de cette approche nécessite :

La disposition d'un corpus, dit corpus d'apprentissage, composé de documents dont les séquences sont étiquetées. Cet étiquetage se fait généralement manuellement par des experts du domaine [5]. Cette première étape d'étiquetage manuel est nécessaire car la notion de séquence n'est pas suffisamment formalisable pour envisager un étiquetage automatique des séquences dans un premier temps.

La disposition d'outils de traitement de documents textuels pour offrir une représentation enrichie. Nous pouvons citer [10] l'analyse morphosyntaxique, qui comprend la segmentation du texte en mots, la lemmatisation des mots [9], l'étiquetage morphosyntaxique [19], le calcul de la position du mot dans une phrase, de la position de la phrase dans le texte, etc. Cette représentation enrichie permettra d'améliorer la qualité du point suivant.

Un modèle d'apprentissage qui permet d'identifier les séquences d'un document et de les étiqueter. Ce modèle repose à la fois sur le corpus d'apprentissage étiqueté manuellement et sur des outils d'enrichissement du corpus.

La littérature sur l'étiquetage de séquences a montré que cette approche a l'avantage de s'adapter à l'évolutivité de la langue [11] En effet, elle utilise un modèle d'apprentissage sur une base étiquette/séquence (qui ne nécessite pas l'ajout de nouveau vocabulaire). Les algorithmes d'apprentissage permettent par la suite une analyse des positions et relations entre les séquences et les étiquettes associées.

Algorithmes d'apprentissage. L'étiquetage de séquences est souvent abordé dans la littérature comme un problème d'apprentissage supervisé. De nombreux travaux s'attaquent à ce problème en utilisant les machines à vecteurs support (SVM) [10] ou les réseaux de neurones récurrents (RNN) [5]. Les SVM sont utilisées pour la simplicité de leur utilisation, néanmoins elles n'utilisent aucune information sur l'enchaînement des séquences et donc sur la structure du document. Les réseaux de neurones récurrents sont couramment utilisés pour l'étiquetage de séquences, néanmoins, en raison du défi lié à la capture de dépendances entre les séquences à long terme, certains travaux ont privilégié le Long Short-Term Memory (LSTM) qui est un réseau de neurones récurrents à mémoire court et long terme qui répond au défi lié à la dépendance entre les séquences. Les réseaux de neurones récurrents ont l'avantage de pouvoir aussi traiter les séquences dans les deux sens (bi-LSTM) (enchaînement de gauche à droite et de droite à gauche) [14]. Par ailleurs, plus récemment, beaucoup de travaux se sont penchés sur les algorithmes CRF (Conditional Random Field) [12]. Ils capturent les dépendances entre les étiquettes prédites à différentes étapes de l'apprentissage en analysant les probabilités de transition d'une étape à une autre. Ils s'adaptent ainsi à des données semi-structurées. De ce fait, plusieurs travaux ont utilisé les LSTM couplés au CRF, et bi-LSTM couplés au CRF [14] pour leur adaptativité à l'étiquetage de séquences. Les algorithmes CRF et les réseaux de neurones récurrents sont ainsi deux algorithmes qui ont permis une modélisation plus précise et efficace des étiquettes [5] comparés aux autres méthodes d'apprentissage, tel qu'un réseau de neurones récurrent simple ou SVM sur des documents textuels non structurés rédigés en langage naturel. La question qui se pose naturellement est de savoir s'ils conservent leur avantage par rapport aux autres algorithmes dans le cas de documents semi-structurés.

2.4 Indexation des offres d'emploi

Rappelons que l'indexation d'une offre d'emploi a pour objectif d'identifier le profil souhaité dans l'offre, en attribuant des étiquettes aux informations de profil : type de contrat, compétences requises, formation requise, etc. Certaines des approches d'indexation mentionnées ci-dessus ont été étudiées dans le secteur du recrutement sur des offres d'emploi. C'est le cas de l'approche par règles utilisée avec des expressions régulières pour extraire les années d'expérience et avec un dictionnaire pour extraire les compétences [2]. Il a été montré que l'approche par règles permet de bien indexer les années d'expériences et les compétences contenues dans le dictionnaire. Néanmoins, pour les séquences non contenues, elle ne permet pas une indexation correcte.

La construction d'une ontologie destinée à la modélisation du vocabulaire des offres d'emploi a été proposée par [21]. Néanmoins, ce travail se limite à un secteur particulier du recrutement : informatique et télécommunications. Au vu de cet article, il est difficile d'en déduire la généralisation de l'approche ontologique au domaine général du recrutement. Nous constatons cependant que cette démarche n'est pas facilement généralisable à l'ensemble des métiers et secteurs en raison de la complexité de construction d'une ontologie complète du domaine et qu'elle ne s'adapte pas à l'évolution du vocabulaire, ce qui est pourtant fréquent dans le domaine du recrutement. Par ailleurs, [18] ont proposé d'utiliser de ontologies existantes : ROME [1] pour pallier le problème de la complexité de construction. Cela offre un gain de temps significatif. Cependant, ces travaux se concentrent uniquement sur les compétences relatives à l'expérience professionnelle, délaissant ainsi d'autres caractéristiques des offres tels que le savoir-être, le salaire, etc. Ces derniers éléments sont pourtant importants pour le recruteur. Nous pouvons dire que les travaux réalisés sur l'indexation des offres d'emploi à l'aide des approches par ontologie et par règles semblent pertinents puisqu'ils permettent bien d'indexer les offres. Néanmoins, ces travaux ne montrent pas l'adaptativité à l'évolutivité des métiers, des compétences, etc. et ne tirent pas profit de l'information, même partielle, sur la structure du texte.

Dès lors, il nous semble pertinent d'utiliser l'approche par étiquetage de séquences puisqu'elle s'adapte à l'évolutivité du domaine. Notons que cette approche a déjà été utilisée dans d'autres domaines sur des textes structurés, mais n'a jamais été utilisée ni sur des textes semi-structurés ni sur des offres d'emploi. Rappelons que l'indexation par étiquetage de séquences repose sur un corpus d'apprentissage étiqueté manuellement. Nous proposons donc dans la suite de l'article une méthodologie pour l'indexation de documents textuels semi-structurés, appliquée aux offres d'emploi, reposant sur l'étiquetage de séquences. Cette méthodologie porte en grande partie sur la partie étiquetage manuel. Cette indexation prend en compte l'évolutivité de la langue et du domaine d'application en utilisant la structure du document.

1. Répertoire opérationnel des métiers et des emplois

3 Proposition d'une méthodologie

L'approche par étiquetage de séquences proposée nécessite la création d'un corpus étiqueté, qui jouera le rôle de corpus d'apprentissage, et qui sera exploité pour l'apprentissage automatique des paires étiquettes/séquences. Cette méthodologie propose donc un schéma de création manuelle de corpus étiqueté en séquences, visant à limiter la variabilité et l'incertitude de l'étiquetage, notamment pour le cas de documents rédigés en langage naturel qui contiennent du bruit ou des incertitudes. Cette méthodologie est composée de 5 grandes phases :

1. Préparation. C'est cette première étape qui constitue la contribution principale de ce travail, elle constitue la partie amont de l'étiquetage. Elle vise à comprendre le domaine d'application, à analyser la structure des données et à identifier les étiquettes pertinentes :

   Analyse du domaine et structuration des documents textuels : cette étape vise à créer un modèle de la structure des documents. Elle analyse le domaine des documents pour en déduire des informations de structure. Etant donné le caractère semi-structuré des offres, le modèle défini n'est pas généralisable à toutes les offres puisque le contenu et l'ordre des séquences peut différer entre les offres d'emploi. De ce fait, la structure incertaine sur quelques offres d'emploi peut créer des ambiguïtés.

   Recueil des étiquettes à extraire : cette étape vise à obtenir un ensemble d'étiquettes jugées utiles en vue d'une indexation complète des documents. Une réflexion est menée sur la modélisation du domaine et sur les étiquettes à extraire. Cette étape identifie les informations importantes à extraire. Plusieurs itérations sont nécessaires pour débuter l'annotation manuelle et permettre une indexation qui répond au besoin d'identification d'informations.

   Création d'un guide d'étiquetage des documents : cette étape vise à proposer une unique façon d'étiqueter aux différents annotateurs humains. L'objectif est d'éviter d'éventuels biais liés à la subjectivité des annotateurs et les ambiguïtés liées à la semi-structure des documents, d'autant plus lorsque l'étiquetage d'un corpus est réparti entre plusieurs personnes. Le résultat de cette étape est un guide d'annotation qui contient la description de chaque étiquette, des exemples, des règles d'annotation communes pour chaque étiquette, etc, et qui vise à laisser aussi peu de latitude que possible aux annotateurs. Cette étape se fait en trois temps. Tout d'abord la création d'un premier guide d'annotations sur la base de la structure définie dans l'étape 1. Ensuite distribuer les mêmes documents aux annotateurs, pour un premier étiquetage de séquences. Enfin, calculer le taux d'accord inter-annotateurs, et modifier puis valider le guide d'annotations en conséquence pour répondre aux ambiguïtés.

2. Etiquetage manuel d'un corpus de documents : cette étape vise à faire étiqueter manuellement les séquences des documents en utilisant le guide créé dans l'étape précédente. C'est ce corpus de données qui servira ensuite pour l'apprentissage automatique d'un modèle d'étiquetage. Certains logiciels existent pour simplifier cette étape en proposant des interfaces d'étiquetage pour les annotateurs en incluant éventuellement une phase de pré-étiquetage automatique. Nous pouvons par exemple citer Dataturks [2] ou encore Gate [3].
3. Enrichissement du corpus de données : Cette étape vise à enrichir le corpus de données avec la phase de traitements de données (comme présenté dans l'état de l'art) pour représenter la richesse des informations lexicales, l'information nécessaire à la désambiguïsation des étiquettes morphosyntaxiques. Cette étape vise aussi à déterminer les entrées et sorties du modèle et à définir les méthodes d'évaluation et validation des algorithmes choisis.
4. Apprentissage du modèle d'indexation : cette étape vise à apprendre automatiquement le modèle d'indexation. Pour cela, elle exploite le corpus de données étiqueté et enrichi. Il est impossible de savoir *a priori* quel(s) algorithme(s) seront les plus adaptés à un domaine applicatif donné. Plusieurs expérimentations seront donc nécessaires pour le(s) identifier.
5. Evaluation de la qualité de l'étiquetage : cette étape consiste à évaluer le/les modèle(s) d'étiquetage automatique, en fonction de mesures prédéfinies.

4 Expérimentations

Nous cherchons maintenant à évaluer la méthodologie d'indexation de documents textuels semi-structurés que nous avons proposée. Nous exploitons un corpus réel d'offres d'emploi que nous avons constitué, composé de 3 335 offres d'emploi (1 094 562 mots) extraites de plusieurs sites d'offres d'emploi français (indeed [4], leboncoin [5] etc.) entre les années 2017 à 2019. Les offres d'emploi sont réparties en 25 secteurs d'activités (ressources humaines, juridique, enseignement, etc.) de façon homogène et contiennent en moyenne 328 mots. Il y a 21 790 mots distincts sur l'ensemble des offres.
Nous détaillons ici chacune des étapes de la méthodologie proposée.

4.1 Préparation

Analyse du domaine et structuration des documents. Nous avons choisi d'acquérir des connaissances du domaine du recrutement avec l'aide d'experts des ressources humaines. Nous avons pour cela réalisé des interviews, étudié et analysé la structure de plusieurs centaines d'offres d'emploi. Cette étape nous a permis de définir une structure-type des offres : une première section présente généralement l'entreprise au travers de ses activités, sa taille, ses valeurs etc. Une deuxième section est généralement dédiée aux missions ou aux compétences techniques requises. La dernière section relate le type de contrat, le salaire, les horaires, etc. Cela valide l'hypothèse qu'une structure, même approximative, est présente dans les offres d'emploi.

Recueil des étiquettes à extraire. Dans le cadre des offres d'emploi, les étiquettes classiquement utilisées dans la littérature sont le métier, le diplôme, le type de contrat, la ville, etc. Nous avons choisi d'exploiter ces étiquettes. Suite aux discussions avec les experts RH, nous avons choisi d'intégrer une étiquette supplémentaire, relative aux missions. Par exemple, une mission est "répondre aux appels téléphoniques". Cette étiquette n'est pas souvent intégrée dans la littérature, pourtant elle identifie des compétences transverses. Dans le cadre de nos échanges nous avons également identifié que le savoir-être (le travail en équipe, l'autonomie etc.) est une information importante exploitée par les recruteurs. En effet, les recruteurs sont aujourd'hui à la recherche de profils qui possèdent, en plus de leurs connaissances techniques, des compétences nécessaires pour s'intégrer aux équipes et aux valeurs de l'entreprise. Cette information est une information implicitement présente dans les offres et nous pensons qu'elle peut être déduite du texte d'une offre d'emploi.

Création d'un guide d'étiquetage des documents. Dans cette étape, nous créons le guide d'annotation pour que l'étiquetage des séquences des offres d'emploi soit objectif et uniforme pour tous les annotateurs. Une des difficultés de la mise en place de ce guide d'annotations est de pouvoir différencier les étiquettes et les séquences associées pour éviter des ambiguïtés liées à un vocabulaire commun entre certaines étiquettes. Par exemple "Développeur informatique en Python" est le poste mais il contient aussi un langage de programmation qui est considéré comme une compétence. La règle d'annotation choisie dans ce guide est de ne pas annoter une compétence lorsqu'elle est contenue dans une séquence qui est spécifique à un autre label (dans ce cas une expérience, une mission ou encore le poste). Par exemple, les missions et les compétences qui ont un vocabulaire commun mais ne représentent pas le même type d'informations. Cette règle a été choisie pour éviter une ambiguïté du modèle associée au vocabulaire commun entre les étiquettes, et appuyer sur la distinction celles-ci. Le guide d'annotations proposé comporte 25 règles qui peuvent aller de la prise en compte des ponctuations, à la distinction du vocabulaire entre différentes étiquettes.
Dans nos expérimentations, trois annotateurs ont été choisis qui se répartissent l'ensemble des offres d'emploi. La Figure 1 présente un exemple d'offre d'emploi étiquetée manuellement par un annotateur. Dans cet exemple, la séquence "Opérateur Régleur CNC" est étiquetée "poste", c'est-à-dire le métier.

2. Outil d'annotation https ://dataturks.com/
3. Outil d'annotation https ://gate.ac.uk/
4. Métamoteur de recherche d'emploi
5. Site web d'annonces commerciales

FIGURE 1 – Exemple d'offre d'emploi annotée manuellement en utilisant le logiciel Dataturks.

Étape d'étiquetage manuel. Dans un premier temps, il faut choisir l'outil d'annotation qui facilitera le travail de l'annotateur. Dans notre cas, l'outil choisi est Dataturks. Pour vérifier la qualité de l'annotation, nous avons calculer le taux d'accord inter-annotateurs qui est de 0,92.

Enrichissement du corpus. Un pré-traitement des données a été effectué en utilisant une analyse morphosyntaxique. Le corpus comporte 317 693 séquences étiquetées qui représentent 66 % des séquences. 34 % des séquences ne sont donc pas importantes pour l'indexation. Cette phase de création de corpus a nécessité au total 300h de travail d'étiquetage. La Figure 2 présente la répartition des différentes étiquettes dans le corpus. Nous pouvons constater que "compétences techniques" et "missions" représentent les étiquettes les plus fréquentes dans les offres. De plus, "savoir-être" représente également une part importante des étiquettes. Cela confirme qu'il s'agit d'une information importante pour le recruteur. Enfin, nous constatons que le salaire est l'étiquette la moins utilisée, ce qui tend à signifier que ce n'est pas une information obligatoire dans les offres. Ces valeurs constitueront une base pour l'interprétation des résultats du modèle d'indexation automatique.

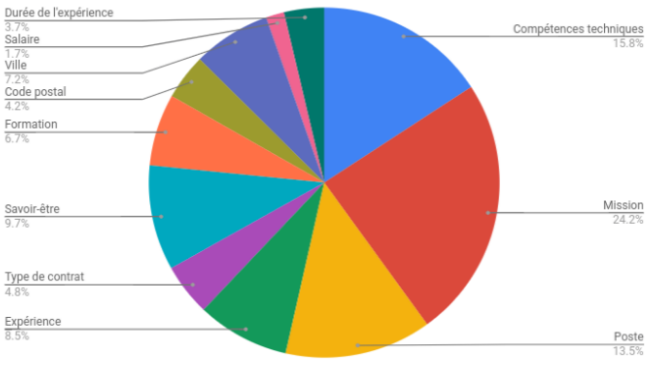


FIGURE 2 – Distribution des paires étiquettes dans les offres d'emploi

4.2 Apprentissage du modèle d'indexation

Choix des algorithmes d'apprentissage. Nous nous sommes intéressés aux modèles CRF, LSTM CRF et bi-LSTM CRF pour l'étiquetage de séquences puisqu'ils se distinguent dans la littérature. Les entrées de ces modèles d'apprentissage sont les séquences, l'étiquetage morpho-syntaxique, la lemmatisation du mot, la position de la séquence dans le document et des mots qui la composent. La sortie de l'étiquetage automatique sont les étiquettes.
La première expérimentation, utilise le modèle LSTM avec les mêmes paramètres que ceux utilisés dans [5], sauf pour certains d'entre eux que nous avons optimisés : les plongements des étiquettes (représentations distributionnelles des étiquettes) ont une taille de 150, les couches cachées du réseau de neurones récurrents ont une taille de 512, le dropout sur tous les plongements [5] qui représente une méthode de régularisation qui vise à empêcher le surapprentissage pendant l'entraînement est de 0,2. Pour le modèle LSTM CRF nous choisissons les mêmes paramètres que le LSTM, auxquels nous ajoutons une couche de sortie. Pour le modèle CRF nous avons utilisé les paramètres standards. Nous nous sommes également intéressés à l'algorithme d'apprentissage SVM, très utilisé dans la littérature et qui a montré de très bonnes performances, mais qui ne prend pas en compte l'enchaînement des séquences et donc la structure du texte. La classification par SVM a été choisie pour valider l'impact de la structure du texte, l'évolutivité des termes, la minimisation de l'ambiguïté associée au vocabulaire commun entre les étiquettes et le choix de l'approche par étiquetage de séquences.

Protocole d'expérimentation. La technique d'évaluation choisie est la validation croisée 10-fold. L'échantillon de validation et de test sont choisis aléatoirement. Les mesures d'évaluation que nous avons choisies sont les mesures classiques de précision (P), rappel (R) et F1.

4.3 Evaluation de la qualité de l'étiquetage

Cette étape vise à évaluer la méthodologie que nous avons proposée.

Performances des algorithmes de la littérature. La Table 1 présente les performances des modèles CRF, LSTM CRF et bi-LSTM CRF. Nous observons dans un premier temps que la valeur moyenne de F1 est très élevée : supérieure à 0,89 pour les trois modèles. Nous pouvons en déduire que la méthodologie d'indexation que nous proposons est valide. Nous constatons dans un second temps que les performances associées à chaque étiquette pour tous les modèles est très élevée (au minimum 0,74). De plus, les étiquettes "salaire", "code postal", "formation" et "durée de l'expérience" sont très bien modélisées par les différents modèles avec un minimum de 0,90, alors que la quantité de données d'apprentissage sur ces étiquettes est relativement faible. Nous notons que le modèle bi-LSTM CRF est celui qui a les meilleures performances en moyenne. De plus, pour l'étiquette "savoir-être" que nous avons introduite et

Métriques Etiquettes	CRF P.	CRF R.	CRF F1	LSTM CRF P.	LSTM CRF R.	LSTM CRF F1	bi-LSTM CRF P.	bi-LSTM CRF R.	bi-LSTM CRF F1	SVM P.	SVM R.	SVM F1
Ville	0,94	0,90	**0,92**	0,88	0,84	0,86	0,95	0,85	0,90	0,67	0,50	0,57
Type contrat	0,92	0,85	0,89	0,90	0,91	**0,91**	0,96	0,87	**0,91**	0,90	0,46	0,63
Formation	0,92	0,93	0,91	0,94	0,92	**0,95**	0,94	0,94	0,93	0,89	0,64	0,74
Experience	0,77	0,81	**0,91**	0,78	0,84	0,87	0,83	0,85	0,86	0,84	0,70	0,76
Durée exp.	0,97	0,96	**0,97**	0,96	0,96	0,96	0,99	0,99	**0,97**	0,60	0,36	0,45
Comp. tech.	0,84	0,75	0,80	0,77	0,76	0,76	0,76	0,86	**0,81**	0,69	0,90	0,78
Missions	0,86	0,90	0,88	0,91	0,87	0,89	0,93	0,91	**0,91**	0,82	0,64	0,72
Poste	0,96	0,93	**0,95**	0,95	0,92	0,93	0,99	0,88	0,93	0,86	0,38	0,52
Code postal	0,99	0,99	**0,99**	1	0,95	0,97	1	0,97	**0,99**	0,83	0,69	0,75
Salaire	0,93	0,91	**0,92**	0,93	0,90	0,91	0,91	0,93	**0,92**	0,88	0,80	0,83
Savoir être	0,91	0,87	0,89	0,88	0,79	0,89	0,92	0,90	**0,91**	0,77	0,60	0,68
Moyenne	0,91	0,87	0,89	0,91	0,87	0,89	0,92	0,90	**0,91**	0,81	0,58	0,67

TABLE 1 – Performance des trois modèles d'apprentissage sur le corpus d'offres d'emploi

Métriques Etiquettes	Dictionnaire P.	Dictionnaire R.	Dictionnaire F1	bi-LSTM CRF P.	bi-LSTM CRF R.	bi-LSTM CRF F1
Comp. tech.	0,57	0,26	0,36	0,76	0,86	**0,81**
Missions	0,62	0,47	0,54	0,93	0,91	**0,91**

TABLE 2 – Comparaison des scores pour les compétences techniques et les missions entre une méthode basée sur un dictionnaire et le bi-LSTM CRF

qui exploite des données implicites, les performances, notamment pour le bi-LSTM CRF sont particulièrement encourageantes elle est la sixième meilleure étiquette avec un score de précision et de rappel respectivement de 0,92 et 0,90. Nous constatons aussi que pour les étiquettes "ville", et "poste", le modèle bi-LSTM sous performe sur le Rappel par rapport aux deux modèles CRF et LSTM CRF puisque celui-ci nécessite plus de données de par sa particularité de bi-direction. Notons par ailleurs que lors de l'étiquetage automatique pour les modèles LSTM CRF et CRF, les étiquettes "expérience" et "compétences techniques" sont confondues dans certaines offres d'emploi (conclusions tirées d'exemples d'étiquetage). Citons la séquence : "vous avez une expérience en langage Python". Le langage Python peut être une expérience mais aussi une compétence. Ces deux modèles ne permettent pas de distinguer ces deux étiquettes contrairement au bi-LSTM CRF qui lui a un bon rappel mais une précision plus faible. La bidirectionnalité du bi-LSTM améliore la prise en compte de la structure d'un texte qui favorise la performance de ce modèle.

Prise en compte de la structure des offres. La table 2, vise à comparer bi-LSTM CRF, qui a permis d'avoir la meilleure précision, à une classification par SVM qui ne prend pas en compte la structure des documents. Nous notons une différence significative entre ces deux algorithmes. En effet, le SVM a une précision moyenne de 0,81 qui est significativement plus faible que celle de bi-LSTM CRF, mais également que celle des autres algorithmes. Ces performances montrent que les offres d'emploi sont bien au moins partiellement structurées et que les modèles d'apprentissage prenant en compte la structure d'un texte sont plus performants.

Evolution du vocabulaire. Pour évaluer la capacité de la méthodologie à prendre en compte l'évolution du vocabulaire, nous avons testé le modèle bi-LSTM sur une offre d'emploi rédigée il y a 19 ans. Cette offre contient 4 séquences non présentes dans le corpus d'apprentissage. Sur cette offre, F1=0,82 et les séquences inconnues par le système ont été correctement étiquetées. Par exemple, la séquence "fort esprit d'entreprise" a été correctement étiquetée "savoir-être".

Ambiguïté du vocabulaire. Nous avons choisi de comparer un dictionnaire complet de missions et compétences pour vérifier que notre méthodologie et le choix des algorithmes offrent une minimisation de l'ambiguïté liée à l'utilisation d'un vocabulaire commun entre les étiquettes. Les scores présentés Table 3 montrent une différence significative entre ces deux approches et montre la validation de notre approche.

5 Discussion et perspectives

Le travail présenté dans cet article chercher à identifier automatiquement le profil dans des offres d'emploi dans l'objectif de permettre, à terme, leur mise en correspondance avec des CV. D'un point de vue scientifique, ce travail vise à proposer une méthodologie d'indexation de documents textuels écrits en langage naturel, semi-structurés et dont

le vocabulaire est amené à changer. La méthodologie proposée repose à la fois sur la définition d'un protocole d'indexation de textes semi-structurés, dans le but de limiter la variabilité entre les annotateurs, et est basée sur l'apprentissage d'un modèle d'indexation automatique. C'est l'approche par étiquetage de séquences qui a été retenue tout au long de ce travail. Les expérimentations menées sur un corpus d'offres d'emploi montrent que la méthodologie proposée offre des performances très élevées : en moyenne plus de 0,91 de précision et plus de 0,87 pour la mesure F1. Outre les performances très élevées, cette méthodologie gère un grand nombre de secteurs d'activités. Nous avons aussi pu constater que cette méthodologie est coûteuse en temps. Néanmoins, elle permet tout d'abord de prendre en compte l'évolutivité des termes grâce à un apprentissage automatique continuel des nouveaux termes. De plus, elle permet de limiter les ambiguïtés. Dans un second temps, nous allons chercher à améliorer l'étiquetage de séquences du "savoir-être" et des "compétences" techniques en utilisant un système hybride reposant sur l'approche à base de règles, pour créer un dictionnaire de motspour représenter cette étiquette. Enfin, nous allons ajouter des étiquettes dans l'annotation pour inclure le poids de chaque étiquette pour la comparaison des offres d'emploi et des CVs.

Remerciements

Ce travail a été réalisé avec le soutien de l'Association Nationale Recherche Technologie (ANRT) (convention CIFRE N 2081/0190) ainsi que l'entreprise Xtramile.

Références

[1] D. D. Bui and Q. Zeng-Treitler. Learning regular expressions for clinical text classification. *J. of the American Medical Informatics Association*, 21(5) :850–857, 02 2014.

[2] A. Casagrande, F. Gotti, and G. Lapalme. Cerebra, un système de recommandation de candidats pour l'e-recrutement. In *AISR 2017*, 2017.

[3] V. Claveau and A. Ncibi. Découverte de connaissances dans les séquences par crf non-supervisés. 2013.

[4] G. De Larquier and G. Rieucau. Job ads : A public but targeted information. A French Labour Force Surveys analysis (2003-2012). *Revue Economique*, 2017.

[5] Marco Dinarelli and Isabelle Tellier. New Recurrent Neural Network Variants for Sequence Labeling. *Lecture Notes in Computer Science*, April 2016.

[6] K. Drame. *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical*. PhD thesis, Université de Bordeaux, 2014.

[7] K. Fort. *Annotated resources, a key issue in content analysis :towards a methodology for manual corpus annotation*. Theses, Université Paris-Nord, 2012.

[8] C. Georgy. *Visibilité numérique et recrutement. Une sociologie de l'évaluation des compétences sur Internet*. PhD thesis, Sociologie. Université Paris-Saclay, 2017.

[9] M. Kestemont and J. De Gussem. Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. working paper or preprint, 2016.

[10] E. Knyazeva, G. Wisniewski, and F. Yvon. Apprentissage par imitation pour l'étiquetage de séquences : vers une formalisation des méthodes d'étiquetage "easy-first". In *Conf. TALN*, page (12), Caen, France, 2015.

[11] J. Krantz, M. Dulin, and P. De Palma. Language-agnostic syllabification with neural sequence labeling. 2019.

[12] T. Lavergne, O. Cappé, and F. Yvon. Practical very large scale CRFs. In *Proc. of the 48th ACL*, pages 504–513, Uppsala, Sweden, 2010.

[13] F. Lévy, A. Nazarenko, and A. Guissé. *Annotation, indexation et parcours de documents numériques Consulter, explorer et maintenir les textes réglementaires*, volume 13. 2010.

[14] X. Ma and E. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Technical report, Language Technologies Institute Carnegie Mellon University Pittsburgh, PA 15213, USA, 2016.

[15] W. Njomgue Sado and D. Fontaine. Un traitement sémantique par ontologie pour l'indexation de documents dans un référentiel métier. In *IC - 16èmes Journées francophones d'Ingénierie des Connaissances*, pages 181–192, Nice, 2005.

[16] C. Pelissier. Les connaissances liées à l'apprentissage de la lecture dans un dictionnaire électronique. *CORELA*, 2007.

[17] C. Raymond and J. Fayolle. Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Conf. TALN'10*, 2010.

[18] B. Sidahmed, N. Mellouli, M. Lamolle, and P. Paroubek. Smart4Job : A Big Data Framework for Intelligent Job Offers Broadcasting Using Time Series Forecasting and Semantic Classification. *Big Data Research*, 7 :16–30, mar 2017.

[19] S. Sidhom. *Morpho-syntactic Analysis Platform for Automatic indexing and Information retrieval :from text-writing to knowledge management*. Theses, Université Claude Bernard - Lyon I, March 2002.

[20] D. Vallet, M. Fernández, and P. Castells. An Ontology-Based Information Retrieval Model. In *The Semantic Web :Research and Applications*, pages 455–470, Berlin, Heidelberg, 2005.

[21] L. Yahiaoui, Z. Boufaida, and Y. Prié. Automatisation du e-recrutement dans le cadre du Web sémantique. 2006.

Réflexion sur le choix d'un classifieur sémantique destiné à aider le cogniticien dans l'élaboration d'une base de connaissance et la garantie de sa consistance dans le temps

Alain Berger[1], François Vexler[1], Corentin Mary[1], Jean-Pierre Cotton[1]

[1]Ardans SAS,
6 rue Jean Pierre Timbaud, « Le Campus » Bâtiment B1,
78180 Montigny-le-Bretonneux, France
{aberger, fvexler, cmary, jpcotton }@ardans.fr,
www.ardans.fr

25 mai 2020

Résumé

L'ingénierie de la connaissance nécessite un effort substantiel lors de l'élaboration de la base de connaissance par les choix engendrés lors de sa structuration et dans la production du contenu qui est validé par l'expert humain. Pour les industriels, l'implantation d'un dispositif n'a de sens que dans sa vie et son maintien en conditions opérationnelles ; ce qui impose de maintenir une cohérence et une consistance dans le dispositif formalisé au travers de cette base de connaissance. Comment faciliter l'alimentation en continu de la base de connaissance par ses contributeurs au fil du temps ? Ce point essentiel de l'intérêt et de la pérennité de ces dispositifs au sein des organisations est celui qui a motivé ces travaux.

Les bases étant construites à l'aide de l'outil Ardans Knowledge Maker® (AKM), sont donc fortement structurées en termes de modèles et d'arborescences de classification (ou vues). La méthode habituelle s'appuie sur l'expertise du contributeur. Ce mode de fonctionnement peut toutefois être mis en défaut notamment lors de contributions occasionnelles ou encore lorsque l'on souhaite intégrer des contenus de manière automatisée.

Comment traiter ce problème de classification automatique ? A partir d'un état de l'art, Ardans exprime comment il en est arrivé à utiliser la sémantique vectorielle pour mettre en œuvre un processus de classification guidé par les connaissances. Cette approche constitue une aide appréciable au maintien en cohérence.

Après un rappel des principales techniques du domaine, cet article discute de l'expérimentation produite dans un récent « Proof Of Concept » pour la classification automatique de nouvelles entrées de retour d'expérience (REx). Ces travaux sont issus de l'industrie et de la vingtaine d'années d'opérations d'ingénierie des connaissances réalisée par Ardans en France et en Europe. Ils préfigurent l'usage prochain du module compagnon d'AKM appelé « Semantic Analysis ».

Mots-clés

Cogniticien, classifieur sémantique, base de connaissance, Management des connaissances augmenté, actualisation automatique, maintien cohérence et qualité, sémantique vectorielle, Retour d'expérience, POC Industriel.

Abstract

Knowledge engineering requires a substantial effort in the development of the knowledge base by the choice generated during its structuring and the production of the content which is validated by the human expert. For industrialists, the implantation of such a process has no meaning except in its life and its maintain in operational conditions ; which requires maintaining a consistency in the system formalized through this knowledge base. How to facilitate the continuous supply of the knowledge base by its contributors over time ? This essential point of interest of sustainability of these systems within organizations is the one that motivated these works. The bases being built using the Ardans Knowledge Maker® (AKM) tool, are basically strongly structured in terms of models and trees classification (or views), the usual method is based on the expertise of the contributor. This operating mode can however be faulted, in particular during occasional contributions or when one wishes integrate content automatically. How to deal with this automatic classification problem ? Starting with a state of art,Ardans express how he came to use semantics vector to implement a classification process guided by knowledge. This approach is a significant aid in maintaining coherence. After a reminder of the main techniques in the field, this article discusses The experimentation produced in a recent « Proof Of Concept » for the automatic classification of new experience feedback inputs (REx). These works come from industry and from the twenty years of knowledge engineering operations carried out by Ardans en France and in Europe. They foreshadow the upcoming use of the AKM companion module called « Semantic Analysis ».

Keywords

Cogniticien, semantic classifier, knowledge base, Increased knowledge management, automatic updating, maintaining coherence and quality, vector semantics, Feedback, Industrial POC.

1 Introduction

L'ingénierie des connaissances propose aux organisations de s'appuyer sur une méthode et des outils afin de modéliser et structurer leurs métiers et les connaissances de leurs experts. Depuis sa création, Ardans conforte sa méthode Ardans *MAKE*®[1] au fil de ses opérations industrielles et implante dans sa plate-forme Ardans Knowledge Maker® (Ci-après AKM)[2] les fonctionnalités qui satisfont aux exigences des utilisateurs. Ainsi, dans le cadre de la validation qualitative des bases de connaissances mises à disposition en tant que livrable, l'intérêt de l'usage des représentations graphiques a été démontré[31].

Les présents travaux s'attachent plus à la stratégie de conservation de ce niveau de qualité compte-tenu des évolutions inéluctables qu'une base de connaissance est amenée à connaître : évolution des connaissances elles-mêmes suite au Retour d'expérience (REx) ou aux efforts de R&D, mais aussi évolution de l'organisation et de ses acteurs. Les principes de similarité syntaxique ou sémantique mis en œuvre s'adaptent bien à la structuration native d'AKM.

Dans le format épistolaire de ce document, nous vous invitons dans une premier temps à apprécier les contextes opérationnels concrets à la fois fonctionnel pour l'industriel et technique pour les ingénieurs de la connaissance dans lequel se positionne cette action de recherche appliquée. Puis nous apprécierons les solutions connues et envisageables relativement à la question de l'analyse et de la comparaison de contenu afin d'élaborer notre choix. Enfin nous recalerons dans le contexte de la technologie AKM les contributions des solutions évaluées afin de produire le résultat attendu.

Nous conclurons par rapport à ce retour d'expérience sur un premier jalon que nous avons analysé comme très encourageant pour les industriels désireux de valoriser au mieux des fonds de connaissances hétérogènes dans un environnement fédérateur.

2 Contextes fonctionnel industriel et technologique opérationnel initiaux

2.1 Origine fonctionnelle de la démarche

Le propre des grandes structures industrielles est la culture de l'écrit. Avec les contraintes imposées ou exprimées par les normes (ISO9001, ISO30401, etc.), par la législation ou la règlementation (la sûreté, le code du travail, etc.), par l'économie (les projets, les contrats, le marché, l'utilisateur final, etc.) et bien entendu par les règles de la physique (mécanique, hydraulique, chimie, métallurgie, optique, etc.), les organismes se sont dotés d'outils pour conserver leurs bonnes règles métiers, leurs meilleures pratiques, leur retour d'expérience, etc.

Naturellement des silos de connaissances se sont élevés dans différents services d'un même département, d'une même direction, et chacun ayant sa finalité singulière avec ses réponses dédiées. Avec le temps, les espaces de chacun de ces silos ont des intersections ou des zones de recouvrement qui s'étendent.

Puis, un jour, suite à un événement opérationnel, un manager pose la question : « pourquoi en est-on arrivé à ce stade ? », « Comment un nouvel arrivant s'y retrouve ? », « Comment est-on sûr, avec cette dispersion, de ne pas perdre la maîtrise de ce que l'on a appris ? ».

Pour résumer l'articulation du système d'information pour soutenir l'activité métier est réalisée autour de serveurs de fichiers, de système de Gestion électronique de document, de système de gestion collaborative, d'intranet métier, de système de gestion de données technique. L'organisation technique intègre les grandes phases du cycle de vie du produit (conception, production, exploitation, formation) pour simplifier de manière presque caricaturale à la complexité intrinsèque de l'activité s'ajoute celle de la sensibilité de l'information.

Fort de ce constat une réflexion sur la gestion du patrimoine de connaissance est lancée avec une cible claire de simplification à des fins d'efficience opérationnelle tout en garantissant les usages multiples actuels et sans perdre le patrimoine aujourd'hui épars.

2.2 Structuration des bases et outil utilisé

Le méta-modèle de gestion des connaissances implanté dans AKM met en œuvre deux principes simples. Les contenus sont stockés dans des fiches ou éléments de connaissances typés par des modèles définis en fonction du besoin. L'exemple présenté dans cet article se réfère à une base de connaissance utilisant quatre modèles (Fondamental, Procédé, Fiche Technique et Fiche Rex).

Dans l'exemple de modélisation avec AKM présenté en FIGURE 1, les nœuds rouges représentent les items des vues et les nœuds bleus les fiches ou éléments de connaissances.

- La relation entre deux vues « V-V » est unidirectionnelle afin de représenter la hiérarchie dans la classification. Le nœud « Racine » est l'origine de l'arborescence.
- La relation entre une fiche et une vue « FI-V » correspondra selon la modélisation à une méta-donnée par exemple ou à une propriété partagée.
- La relation entre deux fiches « FI-FI » est orientée uni ou bidirectionnelle afin de traduire le degré de proximité, de justification, d'explicitation entre les deux concepts.

Notre sujet serait ici de considérer une nouvelle fiche F8 par son contenu et de savoir à quel(s) item(s) de vue l'accrocher et à quelle(s) fiche(s) proposer de la lier.

Chacune des fiches produites peut être liée à une ou plusieurs fiches via un lien bidirectionnel enregistré en base permettant la navigation de proche en proche. Ce premier

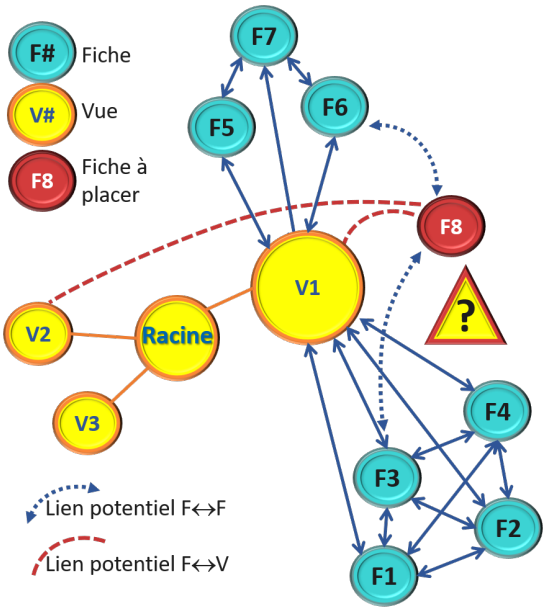


FIGURE 1 – Exemple de structuration AKM - les réseaux « FI-FI » et « FI-V »

réseau de liens est appelé le réseau « FI-FI ». Par ailleurs, des structures arborescentes sont déclarées et jouent le rôle de support de classification pour les éléments de connaissance que l'on accroche en posant un lien. Ces structures arborescentes, que l'on appelle « Vues » dans le vocabulaire AKM, représentent les concepts métier. Elles sont susceptibles de jouer le rôle de taxonomies ou de structure de décomposition [1]. Un élément de connaissance pouvant être lié à plusieurs items d'arbre via l'accrochage « multi-vues », jouera *de facto* un rôle de transversalité. Ce second réseau de liens est appelé le réseau « FI-V ». La FIG. 1 montre un graphe de cette organisation. On retrouve ici les principes utilisés dans la gestion des documentations techniques modulaires (Data-Module et Applicabilité) mais avec un progiciel qui garantit une évolutivité des structures, aussi bien des vues que des modèles. Bien entendu, les structures décrites précédemment sont autant de points d'entrée et d'accès aux contenus que de liens de navigation.

Une telle structure issue d'un processus d'explicitation au sens « recueil d'expertise » peut être vue comme un véritable « réseau sémantique », réseau sur lequel nous nous appuyons pour parfaire notre objectif d'analyse. En ce sens, ce dernier point est comparable aux approches de similarité basées sur la connaissance ou approches topologiques [23, 10].

1. Lors de la conception d'une base de connaissance, on distingue bien les deux cas. Dans le cas d'une taxonomie on a une relation hiérarchique de type « Is-a » ou « SubClassOf » (ontologies) alors que dans les structures de décomposition la relation est de type « PartOf ». La distinction doit être notée car les raisonnements que l'on peut mener sur ces structures sont différents.

2.3 Retranscription synthétique de la problématique

La question posée pour l'équipe du Lab d'Ardans est comment consolider une base de connaissance distribuée car fondée sur des contenus dispersés, la fédérer autour d'un noyau élaboré sur AKM à partir d'une modélisation structurée sur une expertise humaine, la faire grandir en maintenant la consistance des éléments de connaissance entre eux (évitant les doublons) et en appuyant l'utilisateur en consultation ou en contribution dans le respect de ses prérogatives.
A ce point, il convient de rappeler l'état de l'art sur les approches d'analyse et de comparaison de contenu pour envisager de renforcer le moteur AKM par un module efficient pour remplir cette fonction.

3 Approches d'analyse et de comparaison de contenus connues

Une connaissance minimale concernant les techniques de mesure de similarité entre contenus est présentée dans [23] et[10]. On distingue les techniques de similarité syntaxique et celles de similarité sémantique. Notre approche du module compagnon utilise clairement ces deux techniques.

1. La similarité syntaxique : la plus simple à mettre en œuvre, est à la base de l'enrichissement du réseau « FI-FI ». Il s'agit donc ici de repérer des contenus existants s'approchant d'une nouvelle entrée et de les proposer [2] comme étant des voisins.
2. La similarité sémantique : qui nécessite de plus l'exploitation des structures de classification représentant les différents axes est à la base de l'enrichissement du réseau « FI-V ». Il s'agit donc ici de repérer les items d'arborescences ayant un lien de signification avec une nouvelle entrée.

3.1 Techniques de similarité syntaxique

La similarité syntaxique entre deux contenus est classiquement une fonction de proximité sur les termes communs, sans faire intervenir de ressources sémantiques. Il s'agit donc d'une mesure de similarité fondée sur la comparaison des chaînes de caractères qui composent ces contenus. Pour cela, il est habituel de passer par une représentation vectorielle de chaque contenu afin de le positionner dans l'espace vectoriel commun à l'ensemble des contenus. Le Vector Space Model (VSM) fait partie des premiers modèles de représentation des contenus sous forme d'un « sac de mots » [27]. C'est à partir de ce positionnement que l'on pourra décider d'une valeur de proximité en appliquant une métrique dans cet espace vectoriel. On aura pris soin au préalable de constituer le vecteur représentatif de chaque contenu par une technique d'extraction des

2. Le terme de proposer est ici générique. Soit il s'agit d'une vraie proposition faite au contributeur au moment de sa production, soit il s'agit d'une intégration automatisée basée sur un paramétrage qui peut ensuite être ajustée.

termes avec lemmatisation [3](ou racinisation) et par un calcul de poids pour chaque composante du vecteur [4].
Il existe pour ce dernier point différentes méthodes : booléenne, tf (Term Frequency) ou encore tf-idf (Term Frequency x Inverse Document Frequency). Ce sont ces dernières techniques qui sont utilisées dans notre développement avec leurs différentes déclinaisons en fonction du sujet traité. La mesure de similarité s'effectue ensuite par application d'une métrique qui quantifie le rapport entre les contenus. On distingue plusieurs métriques que l'on utilise en fonction des besoins. Dans notre développement nous avons mis en œuvre trois métriques très courantes :

1. La similarité cosinus : fréquemment utilisée et qui consiste à calculer le produit scalaire des deux vecteurs représentatifs des contenus à comparer. Ce calcul utilise directement le concept d'espace vectoriel et d'opérations sur les vecteurs.
2. Le coefficient de Jaccard : qui n'utilise pas le concept de vecteur mais celui d'ensemble et qui mesure une similarité de deux ensembles par le rapport des cardinalités entre leur intersection et leur union. Chaque contenu n'est donc pas vu ici comme un vecteur mais comme un ensemble de termes.
3. La distance euclidienne : qui calcule la similarité comme la distance entre les extrémités des deux vecteurs qui sont donc ramenés à deux points dans l'espace euclidien.

Les autres métriques sont à découvrir dans l'article[23] qui en donne les fondements et propose des exemples de calculs pour illustrer le propos. Nous les citons ci-après pour les plus connues : coefficient de corrélation de Pearson, distance de Levenshtein, indice de Dice.
Enfin, il est intéressant de savoir que chaque métrique possède des performances qu'il convient d'adapter. On retiendra par exemple que plus le texte est de petite taille, meilleur est le résultat de la distance euclidienne par rapport au calcul cosinus ou à l'indice de Jaccard [23] ce qui explique le choix de proposer cette métrique étant donné le caractère modulaire [5] de nos bases de connaissance [28].

3.2 Techniques de similarité sémantique

La similarité sémantique entre deux contenus [6] est classiquement une fonction de proximité de leur signification. Ces techniques peuvent faire intervenir des ressources sémantiques présentées sous différents formats (réseaux sémantiques, taxonomies ou ontologies, dictionnaires, structuration XML [34]). Ces ressources sont alors utilisées pour exploiter la distance sémantique entre deux (ou plusieurs) mots (ou concepts) afin d'établir des ponts entre contenus et qui ne seraient pas identifiés par une technique syntaxique.
On en distingue trois approches essentielles : les approches vectorielles, les approches topologiques basées sur la connaissance et les approches statistiques ou basée sur un corpus.

Les approches vectorielles

Les approches vectorielles sont, dans leurs principes de calcul, équivalentes à celles que l'on trouve en similarité syntaxique. On introduit toutefois dans la construction des vecteurs la notion de contexte qui caractérise chaque mot par l'ensemble des *k* mots les plus associés.
Ainsi, chaque mot possède une représentation vectorielle dans un espace vectoriel de dimension *k* sur lequel on sait pratiquer les métriques vues précédemment. Il est à noter que certaines variantes font la distinction entre le contexte gauche et le contexte droit ce qui aboutit à un vecteur représentatif de dimension $2 \times k$.
Par extension, à partir de ces vecteurs de mot (ou de terme), on applique des opérations linéaires [7] pour composer des vecteurs de phrases ou de groupes [5].
On détermine dans ces conditions une similarité entre deux mots : le mot vectorisé et le mot base d'une des dimensions de l'espace. Ce principe de similarité sémantique se fonde sur l'hypothèse de [15]qui considère que deux mots sont sémantiquement plus proches qu'ils sont plus souvent utilisés dans un même contexte. Dans ce sens, les mesures comme la Probabilité Conditionnelle (SCP ou Semantics of Conditional Probability), le coefficient de Dice ou de Jaccard ou encore l'Information Mutuelle peuvent alors être qualifiées de « mesures associatives » [6]. On trouvera dans la littérature plusieurs articles sur ce sujet avec des propositions de variantes.
L'autre approche vectorielle connue est celle de la double classification où les corpus de contenus sont représentés comme un tableau de vecteurs. L'utilisation d'algorithmes de clustering permet ensuite d'identifier des blocs au sein de cette matrice, blocs qui établissent le lien entre un groupe de documents et un groupe de mots. Il est à noter que cette mesure de co-similarité développée par [3] a été étendue dans le cadre de la classification de données multi-relationnelles [13].
On trouve dans la littérature un nombre important d'article traitant de ce sujet et qui représentent pour notre objectif une source importante de réflexion [21].

Les approches topologiques basées sur la connaissance

Les approches topologiques ou à base de connaissances sont conçues pour déterminer la similarité entre mots et

3. Regroupement des mots d'une même famille sous un même lemme.
4. Il s'agit donc de déterminer les coordonnées du vecteur de contenu dans l'espace vectoriel à N dimensions, N étant le nombre de termes retenus représentatif de l'ensemble des contenus.
5. C'est effectivement un point important dans la conception d'une base de connaissance et qui est la traduction d'une approche holistique. On aura idéalement la même contrainte de taille vis-à-vis des pièces jointes qui sont des documents et qui peuvent être analysées à la demande.
6. On sera très vigilent sur ces notions de similarité sémantiques qui peuvent s'appliquer entre mots ou entre contenus. Certaines techniques utilisent les deux notions de similarité comme Chi-Sim (similarité croisée ou co-similarité) [13].
7. On distingue les opérations vectorielles suivantes : somme normée, combinaison normée, multiplication par un scalaire et produit terme à terme : chacune de ces opérations peut être utilisée pour paramétrer la détermination de la représentation vectorielle et influer sur le résultat.

s'appuient sur des ressources sémantiques de type taxonomie telles que WordNet [11] et [14].
On distingue deux techniques principales : la similarité à base de distance taxonomique (Edge-based) et la similarité à base de contenu informationnel (Node-based).

1. Dans la première technique « Edge-based » [24], la similarité se traduit par le calcul du nombre d'arcs entre les termes. Il existe différents modes de calcul qui permettent d'adapter les résultats et dont nous citons ici les plus connus : la mesure de [25] est considérée comme étant la méthode d'origine et compte simplement le nombre d'arc séparant les deux termes ; la mesure de [33]qui ont utilisé leur approche dans le cadre du projet KBMT (Knowledge Based Machine Translation) en introduisant la profondeur du plus bas ancêtre commun aux deux termes (LCS pour Lowest Common Subsumer) et enfin la mesure de [17] qui ont adapté la mesure de Rada en considérant que les arcs les plus bas dans la hiérarchie reflètent des similarités plus fortes et des distances plus faibles. Il est à noter que l'on trouve régulièrement des articles proposant de nouvelles variantes de calcul [29].
2. La seconde technique « Node-based » encore appelée « information content-based » [26], n'utilise pas les distances vues précédemment mais directement la mesure de l'information commune. La similarité entre deux concepts c1 et c2 est calculée par l'intermédiaire des contenus appartenant à l'ensemble des concepts qui subsument c1 et c2. La mesure proposée par [26] retourne simplement l'IC [8] du LCS de c1 et c2. D'autres propositions de ce principe existent parmi lesquelles [19] et [16].

Les approches statistiques ou basée sur un corpus

Les approches statistiques ou à base de corpus fondent la similarité entre les mots issus de l'ensemble des documents sur leurs liens de cooccurrences. Il s'agit d'identifier les composants sémantiques sous-jacents des mots. Le principe repose sur l'idée que des mots d'un corpus de textes apparaissant dans les mêmes contextes linguistiques ont des significations similaires.

— Le modèle LSA (Latent Semantic Analysis) est lui-même issu d'un des principes de la linguistique distributionnelle [15] où le sens d'un mot peut être défini statistiquement à partir de l'ensemble de ses contextes [18]. [7] propose une technique d'indexation LSI (Latent Semantic Indexing) pour regrouper les termes sémantiquement liés par cooccurrence dans un même concept afin de pouvoir pratiquer la Recherche d'Information (RI) par concept.
Cette technique utilise une représentation matricielle dont la dimension est ensuite réduite par une décomposition en valeurs singulières [9] et par filtrage. LSA réalise de cette manière une sorte d'analyse factorielle avec, sur les données textuelles, un « *mécanisme inductif qui rapproche sémantiquement les termes cooccurrents. Ainsi, aucune connaissance du domaine, formalisé par un humain, n'est nécessaire au préalable* » [9].
— Les autres méthodes statistiques (ESA, PMI-IR et NPMI) ne sont pas décrites dans cet article et on trouvera en [12], en [30] et en [4] les explications nécessaires.

4 Processus général

Le cadre de ce travail est l'indexation automatisée de fiches de connaissances de spécialité, complétées ou non par des pièces jointes documentaires, pour laquelle on dispose de ressources sémantiques sous la forme d'une base AKM validée. Il s'agit donc dans un premier temps d'un dispositif supervisé.
L'originalité de la démarche consiste à exploiter des recueils d'expertise et il s'agit donc là d'une manière supplémentaire de valoriser ce travail, par ailleurs utilisé de manière classique au sein des entreprises (Capitalisation & Transmission des connaissances). On trouvera dans [20], dans [10] et dans [8] les éléments de réflexion qui avalisent notre démarche.

4.1 Positionnement du contexte du POC dans le management de REx

Les fonctionnalités développées ont été mises en œuvre industriellement dans le cadre d'un « Proof Of Concept » (ci-après POC) dont l'objectif était de proposer un moyen d'intégrer des faits techniques d'exploitation (à l'origine du Retour d'Expérience ou Rex) dans un référentiel commun aux différentes entités de l'entreprise et notamment les métiers [10]. L'intérêt réside ici en deux volets organisationnels :

1. Un volet qui s'intéresse à la mobilisation du ou des métiers impliqués dans le fait technique : c'est un principe de mobilisation synchrone des ressources.
2. Un volet qui s'intéresse au ciblage du sujet à traiter et au repérage des faits techniques similaires, ce dernier point entrant dans le cadre du raisonnement par analogie.

On pourra se référer à divers articles et notamment ceux de la Foncsi (Fondation pour une culture de la sécurité industrielle) [22] pour comprendre que la gestion du REx est d'une manière générale un processus complexe à organiser. Il l'est d'autant plus que l'objet sur lequel il porte est lui-même complexe et que l'environnement d'exploitation est compliqué, avec notamment différents niveaux d'acteurs.
Chaque entité dispose alors de bases de données aux modèles plus ou moins compatibles et qui en bout de chaîne sont plus ou moins bien exploitées.
Il est donc intéressant dans ces conditions, d'exploiter au

8. IC (Informational Content) au sens de Shannon : IC(c) = - log P(c) où P est la probabilité de rencontrer le concept c.
9. SVD : Singular Value Decomposition.
10. Notre expérimentation revêt un intérêt d'autant plus important que l'entreprise est en charge de réaliser des produits complexes mettant en jeux plusieurs métiers (mécanique, électronique, contrôle commande, physique et mesures, etc.).

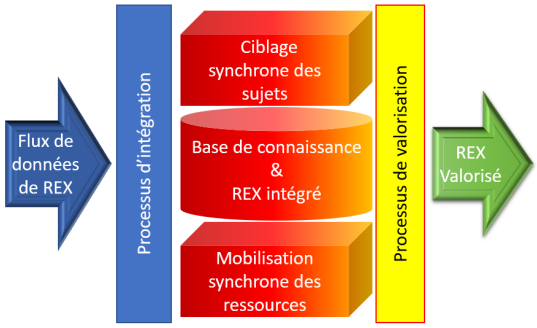


FIGURE 2 – Schéma du dispositif du référentiel de REx

mieux la partie textuelle du fait technique avec des taxonomies prenant en compte la transversalité des sujets.
Le schéma FIG. 2 illustre le dispositif avec la fonction d'intégration qui met en œuvre ces techniques de classification automatique. Dans son principe, les techniques utilisées dans le POC sont relativement élémentaires et doivent être vues comme une première version de notre applicatif. L'une est directement issue de la similarité syntaxique et l'autre introduit *a minima* une similarité sémantique. Dans ce dernier cas, on considère que nous nous rapprochons plutôt d'une technique à base de corpus.

4.2 Recherche de similarité de fiches

Cette première fonctionnalité permet d'établir d'éventuels liens de fiche à fiche (réseau « FI-FI ») et réalise une véritable intégration d'une nouvelle entrée dans la base. Compte-tenu de l'organisation par modèles de cette dernière, on distingue la possibilité d'une part de relier un nouveau fait technique à un processus ou à une fiche technique (typiquement le REx de procédure et le REx de matériel) et d'autre part à un fait technique (ou REx) similaire. Le schéma FIG. 3 rend compte de cette structuration.

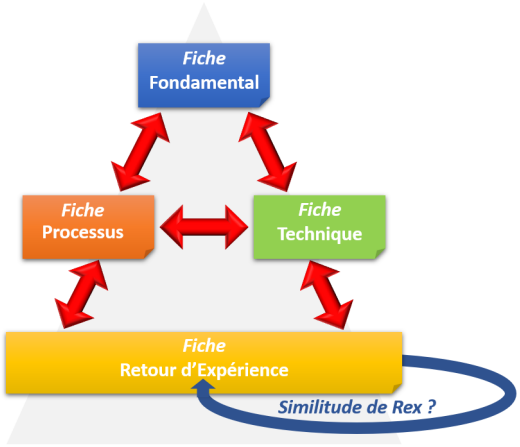


FIGURE 3 – Structuration des modèles et intégration d'une nouvelle entrée REx

Le calcul de similarité est directement issu du Vector Space Model (VSM) et s'effectue selon une des trois métriques déjà décrites : cosinus, indice de Jaccard ou distance euclidienne. Dans la partie actuellement implantée, la comparaison des fiches se fait deux à deux ce qui se justifie par le volume en général modéré des bases de connaissance [11].
La détermination du poids se fait en utilisant une technique de normalisation entre fiches ce qui permet d'obtenir des vecteurs représentatifs indépendamment de la taille du contenu. Cette technique est celle du tf-adapté (Term Frequency-adjusted). Un paramétrage de poids peut être réalisé par l'intermédiaire de coefficients d'importance sur les différentes parties du contenu (titre, rubrique ou pièce jointe qui sont les éléments de connaissances manipulés dans nos bases).

4.3 Recherche d'items de classification

La fonction de classification sémantique est une opération originale qui consiste à déterminer un « ADN sémantique » à chaque item de la taxonomie et qui facilita la réalisation du rapprochement entre une nouvelle entrée et un ou plusieurs items. On convient que ce principe s'approche d'une factorisation des contenus des fiches déjà rattachées à l'item qui devient alors un concept. L'approche de type « sac de mots » alias « bag of words » avec compteur brut consiste à simplement mettre en commun toutes les racines de mot de l'ensemble des fiches attachées à l'item de vue, et de traiter ce résultat comme une fiche. On aura au préalable effectué une opération de mise en exergue des termes les plus significatifs pour un item par un complément de calcul idf (Inverse Document Frequency).
Le calcul de similarité proprement dit se produira entre la nouvelle entrée et l'« ADN sémantique » de la même manière que la similarité entre deux fiches avec normalisation de type tf-adapté (Term Frequency-adjusted).

4.4 Premiers résultats et suite des travaux

4.4.1 A la rédaction de l'article soit fin 2019

Les premiers essais industriels du POC [12][32] ont été réalisés dans des délais tendus et la stratégie a été de valider le concept à partir de fiches fictives extraites de REx existants et volontairement dénaturés, c'est à dire moins précises ou utilisant des expressions synonymes. Nous avons pris soin notamment de supprimer certains termes fortement utilisés dans les métiers concernés. Nous avons constaté que ces fiches REx ont quand même été rattachées à ces métiers et aux différentes arborescences qui nous intéressaient (Système Principal, Système de Soutien, Paramètres de fonctionnement, Exigences,etc.). Concernant les valeurs de seuils pour les fonctions de similarité implantées, nous avons opté pour une valeur par défaut afin de réduire l'espace des possibles, en laissant à l'utilisateur toute capacité à agir sur les différents curseurs.
De manière générale, nous avons constaté pour la fonction de classification que plus un item de vue était fourni en

11. En général, les bases de connaissance « d'expertise » rassemblent quelques centaines de fiches tout au plus.
12. Cette application expérimentale dans le cadre du REx est un cas particulier. Plus généralement, cette fonctionnalité est destinée à maintenir le niveau de qualité d'une base.

fiches, meilleure en était la réponse. En effet, certains items n'ayant que deux ou trois fiches rattachées [13] semblaient trop « binaires », *ite est* manquant de diversité.
Les résultats ont été suffisamment encourageants et probants pour envisager un développement plus avant vers un applicatif de type « module compagnon ».
Par ailleurs, les travaux vont changer d'échelle car nous allons disposer d'un patrimoine de connaissance d'origine experte plus important et donc d'arborescences de classification plus riche en item et avec une sémantique induite par les réseaux de liens vers les fiches validées par les experts plus riche. Nous allons aussi pouvoir consolider et renforcer le mécanisme de rapprochement entre ce noyau d'expertise et le patrimoine « *legacy* » disséminé avec la comparaison entre différentes stratégies pour enrichir le classifieur sémantique.

4.4.2 Après l'acceptation de l'article soit mi mai 2020

La base de connaissance pour le POC était constituée de 110 éléments de connaissances pour les 4 modèles de représentation de connaissance majeurs évoqués supra. Son évolution actuelle qui regroupe 989 éléments de connaissances pour 5 des 6 modèles de représentation de connaissance majeurs actuels confirme ces dires bien que l'essentiel des fiches contenues dans la base ne sont pas en état « validé expert ». Le dispositif implanté aide effectivement le cogniticien quant à la maîtrise du contenu de cette base d'expertise, et l'usage fonctionnel ouvre de nouvelles perspectives quant à l'usage de la base de connaissance au quotidien pour les acteurs experts du métier ; en effet, à chaque ajout d'un nouvel élément de connaissance le cogniticien ou l'utilisateur métier accède à la mesure entre cet élément et l'ensemble de la base existante. Il agit alors en parfaite connaissance de l'existant et sur les éventuelles nuances par rapport aux Retours d'expérience antérieurs appréciés comme « proches » : un atout majeur pour gérer la consistance de la base de connaissance dans le temps.

5 Conclusion et perspectives

Nous avons présenté un développement adapté à une expérimentation de classification et d'intégration automatique de faits techniques (REx) au sein d'une base de connaissance. Les techniques utilisées sont jusqu'alors relativement élémentaires au regard des techniques existantes. Cependant, elles répondent déjà correctement à la problématique qui nous a été posée considérant les ressources mobilisées.
L'état de l'art que nous avons réalisé nous offre plusieurs voies d'amélioration. Ces pistes de progrès sont sereinement accessibles étant donné le matériel dont nous disposons en matière de base de connaissances et notamment la structuration avec laquelle nous construisons ces bases nous permet d'envisager toute solution qui nous semblerait pertinente parmi celle exposées dans cet article. De plus le code qui a été réalisé dans cet objectif a été conçu de manière très modulaire afin d'obtenir une adaptabilité maximale pour une efficience optimale. L'aide au maintien de la consistance de la base dans le temps reste aujourd'hui un objectif ambitieux mais il semble plus que jamais accessible.

13. On rappelle que le POC s'est effectué sur une base de connaissance restreinte ce qui explique l'existence de ces cas de figure.

Références

[1] A. Berger, JP. Cotton, and P. Mariot. Accompagner au début du 21e siècle les organisations dans la mise en place d'une gestion des connaissances : retour d'expérience. *Revue des Nouvelles Technologies de l'Information - EGC 2009*, E15 :475–479, 2009.

[2] Vincent Besson and Alain Berger. To initiate a corporate memory with a knowledge compendium: ten years of learning from experience with the Ardans method. In Extraction Gestion des Connaissances 2015, Luxembourg, Luxembourg, volume RNTI Extraction Gestion des Connaissances 2015, E-28, pages 401–412, January 2015.

[3] Gilles Bisson and Fawad Hussain. Chi-sim : A new similarity measure for the co-clustering task. In *Seventh International Conference on Machine Learning and Applications, ICMLA 2008, San Diego, California, USA, 11-13 December 2008*, pages 211–217, 2008.

[4] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning : Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen, 2009.

[5] Chauche.J. Application des vecteurs sémantiques à la fouille de texte. In *Revue des Nouvelles Technologies de l'Information*, volume Défi Fouille de Textes : reconnaissance automatique des auteurs de discours - Campagne DEFT'05 (TALN'05), RNTI-E-10, pages 131–150, 2007.

[6] G. Cleuziou and G. Dias. Apprentissage de mesures de similarité sémantiques : étude d'une variante de la mesure infosimba. In *First joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society*, volume Hal-00466038, pages 223–236, 01 2008.

[7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. Indexing by latent semantic analysis. jasis. In *Journal of the American Society for Information Science*, volume 41 (6), pages 391–407, 1990.

[8] O. Desfriches Doria. *Faceted Classification for Knowledge Management in specific trades, Method for the development of FolkFaceted Classifications*. Theses, Conservatoire national des arts et metiers - CNAM, November 2013.

[9] P. Dessus. Vérification sémantique de liens hypertextes avec lsa. In *5ème Conférence internationale Hypertextes, hypermédias et internet (H2PTM'99)*, pages 119–129, Paris, 1999.

[10] Bassetto.S Elbadiry.A and Ouali.M. Etude comparative des méthodes d analyse de similarité des défaillances de système aéronautiques. In *11ème Congrès International de Génie Industriel, CIGI2015, 26-28 octobre*, Québec, 2015.

[11] C. Fellbaum. *WordNet : An Electronic Lexical Database*. Language, Speech and Communication. Mit Press, 1998.

[12] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Manuela M. Veloso, editor, *IJCAI*, pages 1606–1611, 2007.

[13] Clément Grimal and Gilles Bisson. Un cadre général pour les mesures de co-similarité. In *SFC'11 - Rencontres de la Société Francophone de Classification*, pages 141–144, 10 2011.

[14] B. Habert and L. Monceaux. Wordnet : la mère (le père) de tous les réseaux de mots ? In *Séminaire du groupe LIR (Langue, Information, Représentation) du LIMSI du 29/05*, 05 2001.

[15] Harris.Z and al. *The form of Information in Science : Analysis of an immunology sublanguage*. Boston Studies in the Philosophy and History of Science. Springer Netherlands, Berlin, 1989.

[16] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33, 1997.

[17] C. Leacock and M. Chodrow. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet : An Electronic Lexical Database*, pages 265–283. MIT Press, 1998.

[18] Lemaire.B and Dessus.P. Modèles cognitifs issus de l'analyse de la sémantique latente. In *Cahiers Romans de sciences cognitives*, volume 1(1), pages 55–74, 2003.

[19] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann and San Francisco and CA, 1998.

[20] H. P. Luhn. A Business Intelligence System. *IBM Journal of Research and Development*, 2(4) :314–319, 1958.

[21] M. Marchand, G. Fouquier, E. Marchand, and G. Pitel. Représentation vectorielle de documents pour l'indexation de notices bibliographiques. In *Contribution eXenSa à l'édition 2016 du Défi Fouille de Textes (DEFT)*. Recherche d information, document et web sémantique, 2017.

[22] Y. Mortureux. Comparaison de textes : quelques approches... *Techniques de l'Ingénieur, Réf : SE1040 v1*, 04 2004.

[23] E. Negre. Comparaison de textes : quelques approches... *Cahier du Lamsade*, 338, 04 2013.

[24] Aly Ngom. Etude des mesures de similarité sémantique basées sur les arcs. In *CORIA, Paris, France, 2015.*, pages 535–544, 03 2015.

[25] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1) :17–30, 1989.

[26] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95 : Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[27] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620, 1975.

[28] Prajol Shrestha. Corpus-based methods for short text similarity. In *Rencontre des étudiants Chercheurs en Informatique pour le TAL, Jun 2011, Montpellier, France. 297. hal-00609909*, 06 2011.

[29] T. Slimani, B. Ben Yaghlane, and K. Mellouli. Une extension de mesure de similarité entre les concepts d'une ontologie. In *4th International Conference : Sciences of Electronic, Technologies of Information and Telecommunications (SETIT), March 25-29, Tunisie*, pages 211–217, 03 2007.

[30] Peter D. Turney. Mining the web for synonyms : Pmi-ir versus lsa on toefl. In *EMCL '01 : Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag.

[31] F. Vexler, A. Berger, JP. Cotton, and A. Belloni. Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans. In *AIDE EGC 2013, 2ème édition Atelier aIde à la Décision à tous les Étages, Toulouse, France*, volume Actes Atelier AIDE EGC'2013, pages 59–72, Janvier 2013.

[32] François Vexler, Corentin Mary, Alain Berger, and Jean-Pierre Cotton. Management des connaissances augmenté : usage d'un classifieur sémantique pour l'aide à l'élaboration et au maintien en cohérence d'une base de connaissance. In Extraction Gestion des Connaissances 2020, Bruxelles, Belgique, volume Extraction et Gestion des Connaissances, RNTI-E-36, pages 393–400, January 2020.

[33] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[34] Haïfa Zargayouna and Sylvie Salotti. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents xml. In *15èmes Journées francophones d'Ingénierie des Connaissances, Lyon, France.*, volume hal-00380573, pages 249–260, 05 2004.

A Hybrid Bi-LSTM-CRF Model for Sequence Labeling Applied to the Sourcing Domain

Hasnaa Daoud[1], Molka Tounsi Dhouib[1,2], Jerôme Rancati[1],
Catherine Faron[2], Andrea G. B. Tettamanzi[2]

[1] Silex France
[2] Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France

{hasnaa.daoud, molka.tounsi, jerôme.rancati}@silex-france.com,
{dhouib,faron,tettaman}@i3s.unice.fr

Résumé

Dans un certain nombre de domaines, les entreprises sont souvent confrontées à la tâche de traiter au quotidien des quantités importantes de demandes textuelles. L'extraction automatique des informations clés à partir des demandes clients, peut aider à accélérer le processus de traitement. Silex France est aujourd'hui confrontée à ces enjeux dans le cadre du traitement des demandes de sourcings.
Dans cet article, nous partageons nos résultats d'étiquetage de séquences en nous basant sur une méthode hybride BiLSTM-CRF, dans un contexte industriel. Le travail est intégré dans la plate-forme B2B Silex pour la recommandation des prestataires de services. Les expériences faites sur les données de la plateforme B2B Silex montrent qu'avec un bon choix de features à extraire et des hyperparamètres, la combinaison du modèle Bi-LSTM-CRF permet de réussir l'extraction d'infomation à partir des demandes textuelles, même dans un contexte de petites données (small data). En effet, le contenu textuel traité est sous forme de phrases complètes générées par des utilisateurs, et est ainsi exposé à des erreurs de frappe. Pour gérer ce type de données, nous combinons plusieurs types de features extraites décrivant le contenu textuel tels que : (i) la sémantique, (ii) la syntaxe, (iii) les caractères des mots, (iv) la position des mots.

Mots-clés

Traitement du Langage Naturel, Extraction d'Information, Etiquetage de Séquences, Réseaux de neurones artificiels

Abstract

In a number of areas, companies are often faced with the task of dealing with large amounts of textual costumers' requests. Automating information extraction like key phrases from customers' requests can help to accelerate the processing process. Silex France is currently facing this challenge in the context of processing sourcing requests.
In this article, we share our sequence labeling results based on a hybrid method Bi-LSTM-CRF, in an industrial context. This work was integrated in the B2B Silex platform for service providers recommendation. Experiments with the B2B Silex platform data show that, with a good choice of features to extract and optimal choice of hyperparameters, the combination of the Bi-LSTM and CRF helps to achieve good results even in a context of small data. Indeed, the textual content processed is in the form of complete sentences generated by users, and thus is subject to typing errors. To handle this type of data we combine several types of extracted features describing the textual content such as : (i) semantics, (ii) syntax, (iii) word characters, (iv) position of words.

Keywords

Natural Language Processing, Information Extraction, Sequence Labeling, Artificial Neural Networks

1 Introduction

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that helps computers understand, interpret and manipulate human language. Information Extraction (IE) is a crucial task in the field of NLP and linguistics that is widely used for tasks such as Question Answering, Machine Translation, Entities Extraction, Event Extraction...
In this paper, we report on an IE task we conducted in the context of *Silex* company which develops a SaaS e-sourcing platform for the identification of the best providers that meet expressed needs. It takes into consideration the requested service, costs, deadlines, innovation, and quality [4]. There are two main user communities within *Silex* platform : (i) service providers, and (ii) service buyers. Silex aims to automatically analyze the textual descriptions of service requests and providers in order to better evaluate opportunities in a faster way with more targeted sourcings.
We aim to extract key phrases from costumers' requests. In the context of sourcing requests, a key phrases usually indicates a product, a service, an occupation (job title) or a skill. These are the main entities in Sourcing domain. These types of information can be considered as *contextualized* named entities. In fact, it turns out that in some requests, a costumer may talk about his own services, but

we aim to extract only the services he needs.

We propose an IE approach based on a Bi-LSTM-CRF architecture, able to analyze textual descriptions (service providers and service requests) and extract the relevant parts of the text that summarize a provider's offer/ a customer need (such as services, products, occupations, skills). In this paper, we focus on information extraction from service requests.

Processing of service requests is more challenging than that of service offers : (i) text requests are generally short (50 words on average in our case) ; (ii) the content of these descriptions is user-generated, and then is subject to typing errors ; (iii) finally, a user may describe his own products or services to contextualize his request, which would create confusion. This raises the issue of distinguishing between the real user's need and the general context of the sourcing request description. Therefore the task is not the extraction of *any* expression describing a service or a product in a sourcing request.

Our main research question is : How to extract relevant information that summarizes a costumer's need ? We focus on the following sub-questions :

- Which is the best approach to extract information from short texts ?
- Which types of embeddings must we use to extract relevant information in our case ?
- How to deal with the limited number of data ?

Our approach is based on the Bi-LSTM-CRF framework [11] ; a Bidirectional Long-term and Short-term Mermory (Bi-LSTM) encodes an input sequence words and a Conditional Random Fields model (CRF) labels every sequence word. In [11], the authors use two types of embeddings : *word embeddings* and *character based embeddings* (computed using a Bi-LTSM). The final embeddings, which constitutes inputs of the main Bi-LSTM-CRF are obtained by concatenating *word embeddings* and *character based embeddings*.

Our contribution lies in the addition of two other types of embeddings computed with two Bi-LSTMs : (i) Syntactic embeddings and (ii) Position embeddings. Also, to answer *Silex* use case, since we have only 858 annotated service requests, we adapted the architecture hyper-parameters to our context of small data.

This paper is organized as follows : Section 2 presents the related works. Section 3 describes our information extraction approach. Section 4 describes our data and our implementation. Section 5 reports and discusses the results of our experiments. Section 6 concludes with an outline of future work.

2 Related Works

There are three common techniques in the literature for the NER task [17, 12] : (i) knowledge-based unsupervised systems, (ii) feature-engineered supervised systems and (iii) feature-inferring neural network systems.

Our work is mostly related to feature-inferring neural network systems. [3] proposes semi-supervised method where a deep neural network model learns internal representations on the basis of vast amounts of mostly unlabeled training data. This model presents two main limitations : (i) it doesn't take into account long-term dependencies between words because it is based on a simple feed-forward neural network and limits the use of context to a fixed-size window ; (ii) by using only word embedding, the model is unable to exploit other features such as letter's cases or complex aspects of user-generated content. These types of considerations could be useful, especially for rare words.

To overcome some of the limitations of [3]'s model, deep learning algorithms such as Recurrent Neural Networks (RNN) are successfully applied for sequence labeling task. Authors of [18] present a detailed comparison between the Convolution Neural Networks (CNN), and RNN in the particular context of NER. The specificity of RNN is that it allows a neural network to exhibit temporal dynamic behavior to process input sequences with no limitation on the input size. However, RNN are not adapted when input sequences are getting too long ; in a RNN, gradients are back-propagated through all time steps as well. This means that the longer our sequence size is, the more gradients we will be taking the product of. This leads to the vanishing gradient problem. Long-term and Short-term Memory Networks (LSTMs) are a particular type of RNN that are designed to avoid this problem through the use of LSTM cells, which makes it easy to learn about long term dependencies [5].

[11] proposes a more powerful neural network model that incorporates a bi-directional LSTM (Bi-LSTM) and CRF. This model is based on robust learning with the dropout, which allows good recognition results of NER. The bi-directional LSTM model takes into account the whole context enables to effectively train a model with the flexible use of long range context [6].

In addition to the architecture, the way we present input data to a neural network matters. For example, we can see a textual sentence in different ways : (i) sequence of words, (ii) sequence of letters, (iii) sequence of chunks... Enriching the input sequences with accessible additional data can also help to have better results. Character-based representations are very important in our use case. In fact, our data are user-generated. Then, it is important to capture morphological and orthographic patterns. [2] presents a hybrid bidirectional LSTM and bidirectional CNN neural network architecture that helps to exploit explicit character level features such as prefixes and suffixes, which could be useful especially with rare words for which word embeddings are poorly (or are not) trained. [14] introduces the neural character embedding in the NER task for English and achieves the state-of-the-art. [1] explored ways to improve point-of-sale labeling using different types of auxiliary losses, and different representations of words. They built their model based on Bi-LSTM layers and showed that introducing word representations through their characters gives better results in terms of model speed and performance. [9] proposes a simple yet effective dependency-guided LSTM-CRF model that takes the complete depen-

dency trees and captures syntactic properties for the Named Entities Recognition task. Furthermore, [11] incorporated character-level structure into word representation. Each input vector consists of two parts : (i) pre-trained word-level representation [13] and (ii) task-related character-level representation. The authors of [11] adopted a bidirectional LSTM to capture information in both forward and backward directions and concatenate the outputs of these two LSTMs.
Most related works cited in this paper use only word embedding, or combine them with character based representations [9, 1], while we enrich input sequences with Part Of Speech tagging and word position information. The way we represent sequences improves the results in our use case.

3 Methodology

Bi-LSTM-CRF model is introduced by Huang and al [8]. It has been compared to LSTM, Bi-LSTM and LSTM-CRF models. Best results in the paper are achieved with Bi-LTSM-CRF model in different sequence tagging tasks (Part-of-Speech Tagging, Chunking, and NER tasks). In the first part of this section, we present the state of the art Bi-LSTM-CRF model and how we use it in our architecture. In the second part, we detail the features we introduce to enrich word embeddings.

3.1 Architecture

3.1.1 Bi-LSTM :

Bidirectional LSTM or Bi-LSTM model [7] is composed of two LSTMs each one processes the sequence in a different direction.
The main characteristics of LSTM are (i) the ability to operate on sequential data and (ii) the ability to capture long term dependencies thanks to a memory-cell.
LSTM takes as input a sequence of vectors $X = (x_1, x_2, ..., x_n)$ and returns another sequence of vectors, named hidden states vectors $H = (h_1, h_2, ..., h_n)$ as output.
Below, we detail mathematical equations in LSTM network we used in this work :

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + b_{xi} + W_{hi}h_{(t-1)} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \\
C_t &= \tanh(W_{iC}x_t + b_{iC} + W_{hC}h_{(t-1)} + b_{hC}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \\
c_t &= f_t \,\mathbf{x}\, c_{(t-1)} + i_t \mathbf{x} C_t \,^{1} \\
h_t &= o_t \,\mathbf{x}\, \tanh(c_t)
\end{aligned}
$$

Where x_t is the input at time t, h_t is the hidden state at time t, c_t is the cell state at time t, h_t is the hidden state of the layer at time t-1 or the initial hidden state at time 0, and i_t, f_t, C_t , o_t are the input, forget, cell, and output gates, respectively. σ is the sigmoid function.
Figure 1 presents the LSTM architecture, where pink

1. **x** is Hadamard product

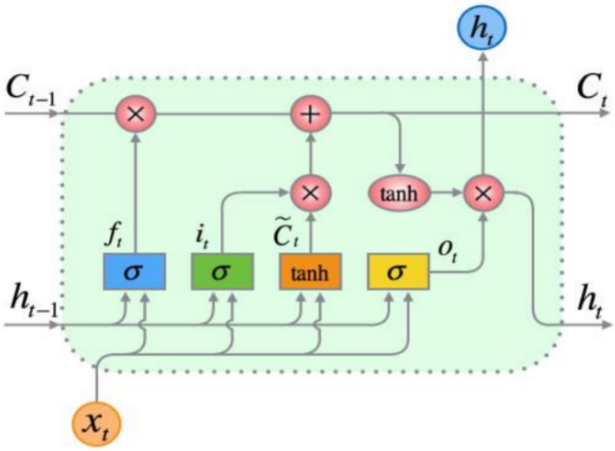


FIGURE 1 – LSTM Architecture.

circles represent arithmetic operators and rectangles represent LSTM gates.
The drawback of the LSTM model is that it processes input from left to right, which involves that it can only encode dependencies on previous tokens. That is why a second LSTM is used to process input in the reverse direction (i.e., from right to left). This new layer makes it possible to detect dependencies on the right context of a token.
In our model, we use Bi-LTSMs to :

— Extract character based embeddings, syntactic representations, and position representations. This part will be detailed in section 3.2.
— Combine all of these representations to extract complete dependency information. The idea is to concatenate the above representations and pre-trained word embeddings to have complete representations of words, then to pass the obtained representations by a Bi-LSTM.

3.1.2 Linear Conditional Random Field

The way sequences are encoded is not the only important part in a sequence labeling problem. The chosen classifier also plays a crucial role. One simple approach is to classify each word independently. The problem with this approach is that it assumes that given the input, all of the entity labels are independent.
In order to relax this assumption, *Generative Models* named HMMs (Hidden Markov Models) make two assumptions : (i) Each state depends only on its immediate predecessor. (ii) Each observation variable depends only on the current state. This last statement makes it impossible to add additional knowledge in the model (e.g contextual information). Using conditional Random Fields (CRF) [10] is a solution to overcome this issue. CRFs are undirected graphical models and are partially similar to HMMs. Nonetheless, they are not *Generative* but *Discriminative Models* trained to maximize the conditional probability of observation and state sequences [15]. The primary advantage of CRFs over HMMs is their conditional nature, resulting in the relaxation of the independence assumptions required by

HMMs in order to ensure tractable inference.
In our Bi-LSTM-CRF model, after extracting a sequence hidden vectors $(h_1, h_2, h_3, ..., h_n)$ with Bi-LSTMs (with n the sequence length), we compute scores associated with each label j at position t in the sequence with a linear layer. We consider P the resulted matrix of size $n \times k$ (with k the number of labels) :

$$P_t = W_{hp} h_t + b_{hp} \forall t \in [[1, n]] \quad (1)$$

with :

— $h_t = (h_{t_l}, h_{t_r})$ hidden vector at position t in the sequence (concatenation of two hidden vectors ; right context hidden vector and left context hidden vector)
— W_{hp} weight matrix of dimension (k, m) (with m the hidden vectors size)
— b_{hp} bias vector of dimension k

As explained earlier, CRF model does not rely only on matrix P since we assume that predicted variables in a sequence depend on each other.
Inference in linear CRF models is based on maximizing the following conditional probability :

$$P(y|H) = \frac{\exp(score(H, y))}{\sum_{y' \in Y} \exp(score(H, y'))} \quad (2)$$

with :

— H the matrix of hidden vectors : $H = [h_1, h_2, h_3, ..., h_n]$
— $P(y|H)$ the conditional probability of a sequence of tags y .
— Y the set of possible tag sequences.

$score$ function is defined as follows :

$$score(H, y) = \sum_{t=1}^{n} A_{y_t, y_{t+1}} + \sum_{t=1}^{n} P_{t, y_t} \quad (3)$$

With A the transition scores matrix ($A_{i,j}$ transition score from state i to state j)

3.2 Features

Our main contribution in this paper is the addition of new kinds of features extracted using Bi-LSTMs. Our model uses four different kinds of embeddings for each sequence word. Three types of these embeddings are trained with Bi-LSTMs, they are concatenated with pretrained *FastText* word embeddings, and used as the input sequence to our main Bi-LSTM-CRF model.
Figure 2 schematizes how the different types of embeddings are trained.

3.2.1 Word embedding

Word embeddings are vector representations trained on a large corpus of texts such as encyclopedias, news texts, or literature. There are many language modeling and feature learning techniques that help to map words to vectors that allow to represent the contextual proximity of different words. It is possible to train from scratch these vector representations on the particular task of keyword sequence tagging using our data. However, it turns out that the size of our data will not allow us to cover a large vocabulary. Therefore, we chose to use pre-trained word vectors for French, learned using fastText [2] on a Wikipedia dump. The French model we used contains 1.152.449 words mapped to 300-dimensional vectors [13].

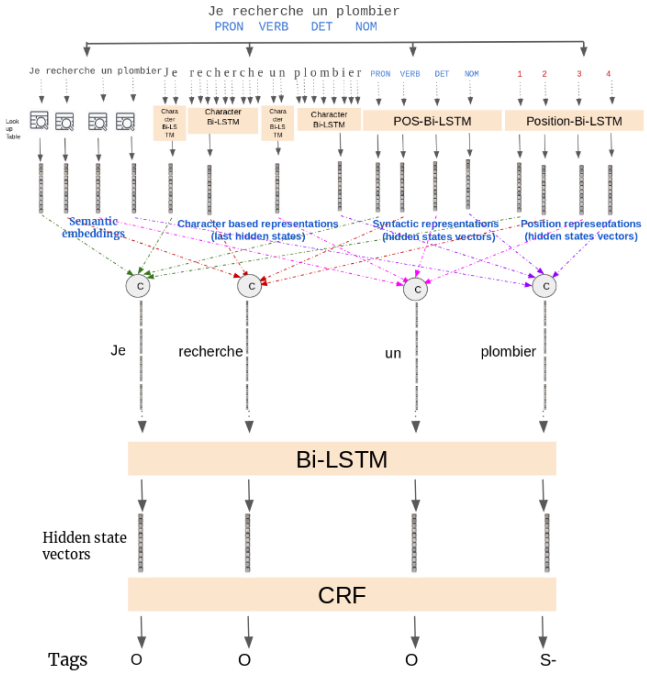


FIGURE 2 – Model architecture (Embeddings Extraction + Main BiLSTM-CRF)

3.2.2 Syntactic Word Representations

Manual analysis of our data shows that syntax plays an important role to locate the user's need in a sequence. Hence the need to extract features representing syntax and dependency between words. This type of information (i.e., syntactic structure and dependencies) is important to complete the semantic information. For example, many sourcings requests contain verbs like (*rechercher, souhaiter, chercher ...*), in these cases recognizing the sentence object would help the model to recognize the customer's need. We then trained part-of-speech (POS) embeddings, using a Bi-LSTM Model.

3.2.3 Character-level Word Representations

Fasttext pretrained Word embedding allows to extract information about the meaning of words, but the covered vocabulary remains limited to the vocabulary of the training corpus. Therefore, rare words or words with spelling errors cannot be represented. Hence the importance of Character-level Word Representations since any word is made up of a number of characters and characters set is finite. For every word, we use a BiLSTM that takes as input the sequence of word's characters, and returns the last hidden states vector. We consider this vector as a character level based represen-

2. https ://fasttext.cc/docs/en/pretrained-vectors.html

tation.

3.2.4 Position Representation

This kind of information is important in our use case : based on manual analysis of our data, it turns out that the main subject is mentioned at the beginning of the text. We are convinced that this is generally valid for written text requests. Thus, it would be interesting if the model takes into account the position of words. This way, the model will be able to understand that it is highly likely that the words at the beginning of the text are relevant information. We use a Bi-LSTM model to extract this type of embedding.
All these types of representations are concatenated and used as input of the main Bi-LSTM-CRF bloc as shown in Figure 2.

4 Experiments

4.1 Dataset

Our dataset is composed of 883 descriptions of service requests distributed as follows : (i) 594 service request descriptions in the train set, (ii) 198 service request descriptions in the development set, and (iii) 90 service request descriptions in the test set.
Data are annotated by Sourcing experts at *Silex* according to the BIOES format, which stands for Beginning, Inside, Outside, End, and Single to indicate the position of a token in a tagged segment. Descriptions are annotated as follow :

- **B-** : to mark the beginning of an entity
- **I-** : to mark the inside of an entity
- **E-** : to mark the end of an entity
- **S-** : to mark a single entity
- **O** : to mark a token outside all of entities

Table 1 shows an example annotation using the BIOES format.

Word	POS	Label
Je	PRON	O
recherche	VERB	O
un	DET	O
plombier	NOUN	S-
La	DET	O
société	NOUN	O
BNB	NOUN	O
souhaite	VERB	O
créer	VERB	O
des	DET	O
supports	NOUN	B-
de	ADP	I-
communication	NOUN	E-
.	.	.
.	.	.

TABLE 1 – Example of input training data

Artificial neural networks generally require a large corpus of training data in order to learn efficiently the task. To overcome this limitation of the training set size (594 service requests), we augmented the training set. We generated new training examples by applying some transformations of the available ones. These transformations must preserve the entities' labels.
In the field of NLP, it is difficult to increase text due to the complexity of natural language. Some methods in the literature use shuffling, which involves changing the order of sentences in the text or random deletion of certain words from the text. However, this augmentation techniques may change the whole context of the text. In natural language processing, it is not a good idea to augment data using signal transformation techniques (images [16], speech recognition...), because the order of characters is important to keep a strict syntactic and semantic meaning. Other augmentation techniques make more sense in the context of our work, such as injecting punctuation noise or modifying certain characters in words. Otherwise, it is difficult to add more semantics, the best way to do so is to use human sentence rephrasing, but it is an expensive operation. Therefore, the most natural choice in data augmentation for us is to replace words with their synonyms based on a dictionary of synonyms of the most frequent words in the field of sourcing.
We therefore chose to augment our training dataset by :

- introducing punctuation noise
- changing characters of some words (which simulates spelling errors in user-generated data)
- replacing some words with their synonyms

This way, we multiplied our training set size by three.

4.2 Implementation

To implement our model, we used the Pytorch library, which supports GPU computing. For all Bi-LSTMs in our model, we used the *torch.nn.LSTM* class, which allows to create LSTMs and to set parameters such as the number of layers, bidirectionality, input feature size, hidden state vector size. We implemented the CRF based on equations of section 3.1.2. We use the *Viterbi* algorithm for finding the most likely sequence of hidden states.

4.3 Hyper-parameters

Table 2 shows the hyper-parameters we use in Bi-LSTM layers used for character-based, position and syntactic features extraction.

Features	**Embedding dimensions**	**Hidden vectors dimensions**
Syntactic features	25	30
Character-level based features	25	50
Position features	25	30

TABLE 2 – Hyper-parameters for the features extraction layers

For the final Bi-LSTM bloc (see Figure 2) used to combine all the features of Table 2 with Fasttext word embeddings, we chose a hidden layer of size 400.
In our experiments, we started by initializing the semantic word embedding layer with FastText pre-trained embeddings, and we updated them during training. We found that

this causes overfitting because of a high number of parameters. We then chose to **freeze the semantic word embedding layer**, which helped us to improve the results.
In addition to increasing the data, to avoid over-fitting, we chose a high dropout value of 0.5.
In the literature, the number of Bi-LSTMs hidden layers is usually equal to the same number of embedding units. We compared the results with different units' numbers in the main Bi-LSTM and we were able to get better results when the size of the hidden layers is equal to 400.
In this paper, we conducted four experiments to compare the performance of four kinds of models :

- I : Bi-LSTM model with only word embedding and logistic regression classification model
- II : Bi-LSTM model with only word embedding and CRF classification model
- III : Bi-LSTM-CRF model with word embedding, character-based representations, and Bi-LSTM position based representation.
- IV : Bi-LSTM-CRF model with word embedding, character-based representations, Bi-LSTM position-based representation and syntactic word representations.

5 Result and Discussion

The problem we deal with is different from ordinary NER tasks, since we detect text segments summarizing a user's need. Let us note that even expert annotators find it sometimes hard to decide on the segment to annotate. Thus, we chose to evaluate the four models in two different ways : (i) Precision and Recall, (ii) Cosine similarity.

5.1 Precision and Recall

We started with an evaluation based on precision, recall, and the F1 score as the basis for choosing the best model. In our evaluation, we do not consider the complete annotated entity, but rather words of the entity separately. For example, in the sentence *"I am looking for a **plumber** able to **repair a faucet Sprinkle**"*, we do not fully penalize the algorithm if it does not detect the complete **repair a faucet Sprinkle** segment. But we count words that it could detect in that segment. Indeed, if we suppose that the model detects only **repair a faucet**, this may be enough to understand the need's subject. We also ignore conjunctions, determinants and punctuation in the evaluation. Table 3 presents the results obtained in terms of precision, recall and F_1 score.

	Recall		Precision		F1	
Model	Dev	Test	Dev	Test	Dev	Test
I	58.88	61.31	77.02	76.61	66.74	68.11
II	64.03	66.61	75.21	76.66	69.17	71.29
III	63.06	66.61	75.62	**80.11**	68.77	**72.74**
IV	67.57	**70.20**	76.03	73.77	**71.55**	71.94

TABLE 3 – Precision, Recall and F1-score without data augmentation

Figure 3 and Figure 4 respectively show the evolution of the F1 score and the evolution of the loss function across epochs.

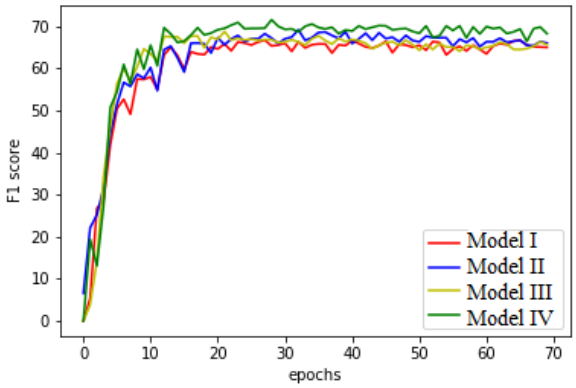


FIGURE 3 – Evolution of score $F1_{Dev}$

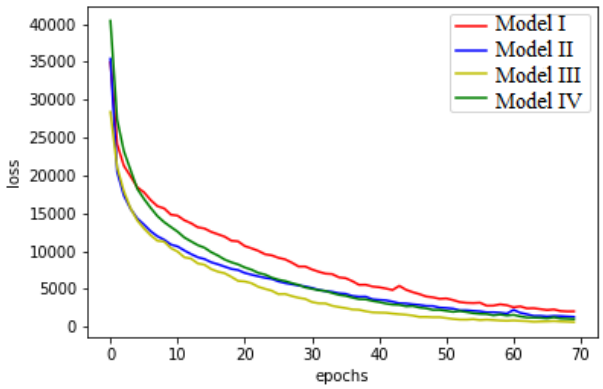


FIGURE 4 – Evolution of the function (*loss*)

One can see that model IV which uses all types of features is the best model across all epochs. Model I, which takes as word representations only word embedding with logistic regression for tagging, has the lowest score. Figure 4 shows that from epoch 30, the loss function continues to decrease without improving the F1 score. From this epoch, the model starts to overfit on training data.
Using syntactic information with POS tagging significantly improves Dev and Test recall, and balances well precision and recall (see Table 3). We also note that with this model, we were able to get a similar F1 scores in dev and test data (difference of 0.39%). The switch from CPU to GPU computing allowed us to gain in terms of performance (more than 1 difference for the $F1\ score$, which is due to the difference between implementations of the libraries used on CPU and GPU) and in terms of training time (from 350 seconds per epoch to 75 seconds per epoch for model IV) .
With data augmentation (Table 4), we have mainly improved the $F1_{Dev}$ scores of the different models. The recall

	Recall		Precision		F1	
Model	Dev	Test	Dev	Test	Dev	Test
I	60.94	61.93	77.31	73.79	68.15	67.35
II	65.64	67.71	75.50	70.23	70.22	68.94
III	63.64	**68.49**	75.28	72.68	70.13	70.52
IV	67.05	68.02	75.34	**77.58**	**70.96**	**72.49**

TABLE 4 – Results with data augmentation

associated with test dataset was relatively improved for the first three models, but the precision decreased relatively, except for model IV.

We can deduce that the best model in terms of Test and Dev scores is model IV that uses a Bi-LSTM-CRF applied on the concatenation of semantic embedding, POS embedding, Character based embedding, and Position embedding.

We point out that the difference between dev and test scores is due to the small amount of annotated data.

The disadvantage of this evaluation method in the particular case of our work is that it gives equal importance to all extracted words. However, it turns out that some words are more important than others, especially words for which meaning appears several times in a service request extracted segments. In the sentence *"I am looking for a **plumber** able to **repair a faucet Sprinkle**"*, "plumber" and "faucet" are two words that generally appear in the same context and could be considered as the most important extracted words. "Sprinkle" ; the faucet brand, is the least important word. However, with precision and recall evaluation, all words are given the same weight, and the model will be penalized in the same way if it does not detect the word **Sprinkle** or the word **plumber**.

We mentioned earlier that the manual annotation was not obvious to the experts. Here is a second possible scenario : in the previous example, an annotator decides to annotate only "plumber" as a unique key phrase, and the model annotates only "repair a faucet", even if there is an important semantic similarity between the two segments, the model will be penalized in terms of Precision and Recall scores. Hence, to deal with this issue, we propose a meaning-based evaluation.

5.2 Evaluation using cosine similarity distribution

To address all of the above problems, we tried to present the results otherwise, which corresponds to the purpose of this work.

Indeed, the goal is to be able to reduce a service request description into a set of keywords. So we thought of calculating the cosine similarity between embedding of expert annotated segments and embedding of segments detected by the algorithm for each service request, and then plot the histogram of the results. A service request extracted segments' embedding is computed by averaging on FastText pretrained word embeddings of their words. We ignore stop words and punctuation. In Figure 5, we present four histograms each associated to a different model.

Note that each type of features added to the model helps to move the distribution a little bit to the right. We can clearly see the difference between the distribution of model I and the distribution of model IV : the distribution of model IV is tightest around 1, with a highest peak.

6 Conclusion

In this paper, we proposed a method that relies on Bi-LSTM-CRF for sequence labeling to summarize sourcing requests. We combined several types of features to represent every word in a sequence. The key of success of our method is an original combination of input features. In addition to word embedding, we extract three other kinds of embeddings : (i) character-level based embeddings, (ii) syntax based embeddings and (iii) position embeddings. These additional embeddings are extracted using Bi-LSTMs and are concatenated with word embedding. We have shown that syntax and position of words help to improve the quality of the information extraction in our use case. We also shared hyper-parameters that give us the best training and choices we made to overcome overfitting problems. Moreover, we have shown that Bi-LSTM-CRF architectures for information extraction can provide value even in a small data context.

This work was integrated into Silex sourcing platform to recommend similar service requests, which considerably reduces the processing time in 60% of cases. This recommendation is based on semantic similarity between requests based on their extracted segments.

As future work we aim to experiment new extraction approaches based on transformers like BERT or Camembert for French texts. We also aim to evaluate the generality of our approach designed for the sourcing domain by experimenting it in a general benchmark.

Références

[1] Daniil Anastasyev, Ilya Gusev, and Eugene Indenbom. Improving part-of-speech tagging via multitask learning and character-level word representations. *arXiv preprint arXiv :1807.00818*, 2018.

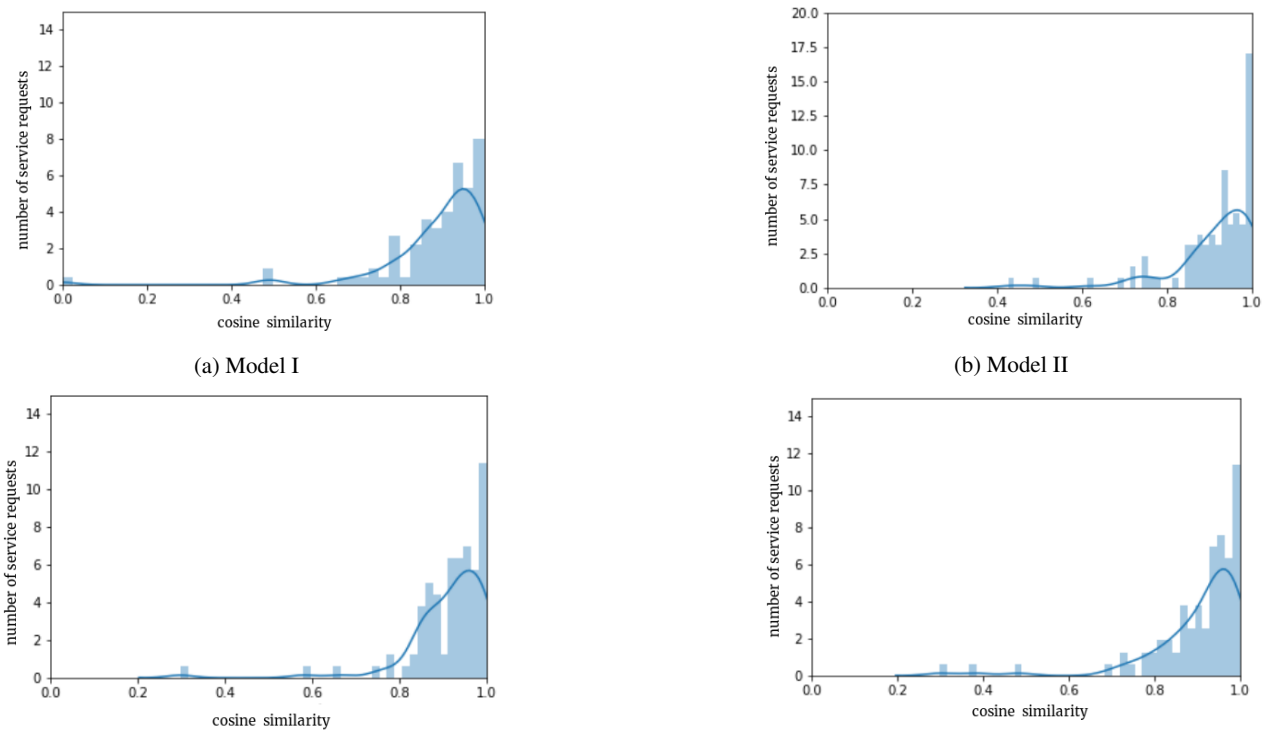
[2] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4 :357–370, 2016.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug) :2493–2537, 2011.

[4] Rémi ESCHENLAUER. Le sourcing, June 2013.

[5] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget : Continual prediction with lstm. 1999.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference*



(a) Model I

(b) Model II

(c) Model III

(d) Model IV

FIGURE 5 – Distribution of cosine similarity between segments automatically detected and segments annotated by human experts

on acoustics, speech and signal processing, pages 6645–6649. IEEE, 2013.

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6) :602–610, 2005.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv :1508.01991*, 2015.

[9] Zhanming Jie and Wei Lu. Dependency-guided lstm-crf for named entity recognition. *arXiv preprint arXiv :1909.10148*, 2019.

[10] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. 2001.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*, 2016.

[12] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *arXiv preprint arXiv :1812.09449*, 2018.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] Cicero Nogueira dos Santos and Victor Guimaraes. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv :1505.05008*, 2015.

[15] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4) :267–373, 2012.

[16] Luke Taylor and Geoff Nitschke. Improving deep learning using generic data augmentation. *arXiv preprint arXiv :1708.06020*, 2017.

[17] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv :1910.11470*, 2019.

[18] Zenan Zhai, Dat Quoc Nguyen, and Karin Verspoor. Comparing cnn and lstm character-level embeddings in bilstm-crf models for chemical and disease named entity recognition. *arXiv preprint arXiv :1808.08450*, 2018.