Appel à projets d'études stratégiques et prospectives 2018

Présentation résumée du projet (mai 2018)

Titre

Vers l'Intelligence artificielle territoriale : apprentissage massif pour la visualisation et la prédiction de décisions administratives

Thème 2 (IA) L'intelligence artificielle appliquée à l'administration territoriale et aux affaires intérieures et sécurité

Durée de l'étude : 9 mois (mai à décembre 2018)

Principaux objectifs : réaliser la preuve de concept d'un outil d'aide à la décision publique, à la fois **visuel, prédictif et généralisable** :

- Visuel : cartographie géo-référencée et temporelle de documents administratifs ;
- Prédictif : identification par apprentissage automatique de corrélations fortes entre contenus des documents : dates, lieux, thèmes, éléments techniques, et décision finale ;
- Généralisable : idéalement indépendant d'un territoire, d'un thème ou d'un format de données particulier, et dont les données sont réutilisables dans les systèmes d'information des administrations ou des services extérieurs, pour croisement avec d'autres données.

Un tel outil doit permettre, par exemple, à partir d'un corpus d'études d'impact, d'identifier rapidement les zones d'un territoire peu ou trop sollicitées sur les 5 dernières années par un type d'installation, et d'estimer la décision la plus probable pour une nouvelle installation similaire. Cet outil permettrait ainsi d'identifier les territoires présentant les plus grandes chances d'accueillir favorablement de nouveaux projets. Tout nouveau document devrait pouvoir être pris en compte au fil de l'eau. Les techniques envisagées sont l'extraction semi-automatique de connaissances, la visualisation de données et l'apprentissage automatique.

Nom du directeur ou responsable scientifique : David GROSS AMBLARD

Laboratoire ou entité responsable de la recherche (nom développé, acronyme) : Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)

Adresse (postale et mail) :

IRISA, 263 avenue du Général Leclerc, CS 74205, 35042 RENNES CEDEX

david.gross-amblard@irisa.fr

Thème de recherche : 2 (IA)

Contexte et état des connaissances

Ces cinq dernières années ont vu l'émergence d'applications de l'intelligence artificielle (IA) qui semblaient jusqu'alors inatteignables : reconnaissance automatique de visages, analyse automatique de vidéos, moyens de transports autonomes, pour n'en nommer que quelques unes. En particulier, de nombreux auteurs [MA:2018] voient dans l'IA une opportunité d'aide à la décision publique [HGL+:2018] par son caractère **prédictif** (ou au moins indicatif), du point de vue de la santé des citoyens, de la qualité des transports, de la sécurité. Couplés à des outils de **visualisation de données,** il est possible de constituer des tableaux de bords pertinents à l'échelle d'un territoire [KN:2018].

Ces nouvelles potentialités sont rendues possibles notamment grâce à l'amélioration récente des techniques d'apprentissage automatique (*machine learning*). Mais ces techniques sont **particulièrement gourmandes en données d'exemple**, permettant leur entraînement (techniques dites supervisées). Ceci est particulièrement vrai pour la célèbre méthode de l'apprentissage profond (*deep learning*). À ce titre, les données administratives constituent un corpus de choix, corpus devenu progressivement accessible par le mouvement d'ouverture des données (*OpenData).* Ainsi, la directive 2007/2/CE INSPIRE (Infrastructure d'information géographique dans la Communauté européenne) a promu dans chaque état membre des plateformes de référencement et de partage d'informations géographiques structurées. La directive s'est traduite en Bretagne par un partenariat d'acteurs publics : GéoBretagne[1], qui oeuvre pour le partage des données, les mutualisations d'acquisition de données, l'aide à leur réutilisation, et la fabrication de services pour tous publics dans tous domaines. Grâce à cet outil, un responsable, un citoyen, peut aujourd'hui faire **l'exploration visuelle** d'un territoire depuis son bureau.

Cependant, si ces données très structurées et d'une très grande qualité sont directement utilisables pour des techniques d'intelligence artificielle, elles restent cantonnées aux données géographiques ou issues de la statistique publique. Or les administrations ont produit et produisent toujours nombre de documents qui ne sont **pas structurés** (donc difficilement exploitables par une machine, comme le format PDF), mais dont l'importance est cruciale dans la vie du citoyen : études d'impact, enquêtes publiques, arrêtés, évaluation environnementale, délibérations des collectivités, …

Le principal verrou au déploiement de techniques d'intelligence artificielle pour l'aide à la décision publique est donc l'accès à de vastes corpus de données peu structurées, desquels il faut extraire l'information pertinente pour en faciliter l'apprentissage automatique.

[1] http://geobretagne.fr

Objectifs

L'objectif de notre projet est de déterminer dans quelle mesure les documents non structurées produits et publiés par l'administration peuvent être analysés, pour en extraire les descripteurs nécessaires à l'**entraînement d'algorithmes d'apprentissage**, en particulier d'apprentissage profond. Ces descripteurs incluent le temps et le lieu concernés par le document, les décisions publiques associées, les critères retenus (par exemple, mesures et conclusions d'une étude d'impact).

Ces descripteurs, et l'apprentissage de corrélations entre ces descripteurs, permettra de produire de nouveaux jeux de données **superposables** aux données géographiques ou aux données statistiques existantes. Cette analyse et cet apprentissage doivent permettre par exemple de répondre aux questions suivantes :

- **Exploration visuelle** :
    - Ponctuelle : quelles sont les informations associées à tel point sur le territoire ?
    - Thématique, géographique : quelles sont les décisions prises sur tel périmètre géographique, tel domaine ?
    - Temporelle : quelle est **l'évolution** passée de tel domaine sur telle période de temps ?
    - Agrégative : quelles sont les préoccupations de tel territoire (en lien par exemple aux enquêtes publiques) ?
- **Prédiction** :
    - Thématique, géographique : quelle est la décision la plus probable pour un thème donné, sur un territoire **non encore analysé jusqu'à présent** ?
    - **Tendancielle** : quelle est l'évolution future la plus probable de tel domaine ?
    - **Veille** : y-a-t-il un changement significatif dans les informations observées ?

Ce projet est l'opportunité de rassembler différents acteurs complémentaires en Bretagne :

- Le laboratoire IRISA[2]: les équipes DRUID (gestion de données hétérogènes, crowdsourcing, machine learning), SHAMAN (extraction de connaissances interprétables), INTUIDOC (extraction d'information dans des documents peu structurés), SEMLIS (extraction de connaissances, apprentissage symbolique, traitement du langage naturel) ;
- La DREAL Bretagne (compétences en géomatique et statistique publique) ;
- Le partenariat GéoBretagne (regroupe les données structurées de 130 acteurs publics, dispose d'une fabrique de services et d'outils de visualisation) ;
- Le SGAR Bretagne et la préfecture d'Ille-et-Vilaine (production de décisions administratives, connaissance des enjeux portés par les décideurs publics, connaissance des mécanismes de prise de décision) ;
- Un.e ingénieur.e temps plein sera recruté sur la durée du projet.

[2] http://www.irisa.fr

Hypothèses

Dans son expression générale, ce projet est très ambitieux. Il est nécessaire pour aboutir à une preuve de concept en un temps raisonnable, de limiter sa portée. Nous nous focaliserons d'abord sur des jeux de données certes peu structurés, mais dont nous avons une représentation structurée par ailleurs (vérité terrain par annotation manuelle) : décisions de l'autorité environnementale, enquêtes publiques, données contenues dans les applications d'instruction de l'état et des collectivités. De même, nous procéderons par difficulté croissante pour l'extraction d'information, en commençant par le lieu et le temps, puis des thèmes généraux. Pour choisir les cibles adéquates, nous procéderons à plusieurs entrevues avec les spécialistes métier (services du SGAR Bretagne, DREAL Bretagne, Préfecture d'Ille-et-Vilaine, ...). L'extraction de données sera réalisée dans le respect des règles d'utilisation d'éventuelles données nominatives.

Méthodes : sources y compris méthodes d'analyse des résultats

Une chaîne de traitement sera mise en place pour 1) la récupération de documents cibles multi-sources[3], 2) l'extraction de descripteurs, 3) l'apprentissage automatique de ces descripteurs, 4) l'enrichissement de bases de données en vue de leur visualisation. Pour l'extraction des descripteurs de documents, nous utiliserons les techniques classiques ou développées à l'IRISA (*entity resolution*, TALN avec apprentissage automatique [BF:2016] [CFF+:2016], *crowdsourcing* pour le "dernier kilomètre" de l'apprentissage *[MGM:2016]*). À moyen terme, il sera possible d'exploiter des techniques d'analyse d'images de documents [PLC:2014] [CLC:2017] pour extraire automatiquement la structure et le contenu de fichiers PDF non annotés. Nous évaluerons l'entraînement de plusieurs méthode d'apprentissage automatique (*deep learning* ou plus classiques) et comparerons les connaissances acquises. L'évaluation se fera par comparaison à nos données de vérité terrain et par validation croisée. Les résultats obtenus par apprentissage seront consolidables par les experts métier [SPL:2017] [SYP:2017]. Pour évaluer l'utilité des données produites, GéoBretagne est en lien avec plusieurs projets désireux d'associer information structurée et documents non structurés, sous la forme de services.

Résultats attendus

Le livrable du projet comportera un outil preuve de concept montrant l'**extraction** de descripteurs de documents administratifs non-structurés, le **positionnement** dans le temps et l'espace des documents selon une interface cartographique, et l'obtention de **prévisions /suggestion** de décisions sur un certain nombre de thématiques. L'approche est conçue pour être **généralisable** à plusieurs types de documents, plusieurs régions. L'infrastructure proposée sera orientée service, pour permettre à d'autres acteurs de composer des services à partir du nôtre (selon un modèle *Freemium*). Enfin, nous proposerons un certain nombre de **recommandations** afin de systématiser l'extraction de descripteurs pertinents, et donc de faciliter à l'avenir le déploiement de techniques d'IA pour l'aide à la décision publique.

[3] Par exemple https://www.toutsurlenvironnement.fr/ et http://communes.bretagne-environnement.org/

Bibliographie indicative

Références externes

Apprentissage machine pour la ville intelligente et l'e-gouvernement

[MA:2018] Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges. Mehdi Mohammadi, Ala Al-Fuqaha. In IEEE Communications Magazine (Volume: 56, Issue: 2, Feb. 2018).

[HGL+:2018] Emerging Trends, Issues, and Challenges in Big Data and Its Implementation toward Future Smart Cities: Part 2. Guangjie Han ; Mohsen Guizani ; Jaime Lloret ; Sammy Chan ; Liangtian Wan ; Wael Guibene. In IEEE Communications Magazine (Volume: 56, Issue: 2, Feb. 2018).

[Mehr:2017] Artificial Intelligence for Citizen Services and Government. Hila Mehr. Harvard Kennedy School, ASH center for democratic governance and innovation, August 2017.

Visualisation de données pour la ville intelligence et l'e-gouvernement

[KN:2018] Big data dashboards as smart decision support tools for i-cities – An experiment on stockholm. Karima Kourtit, Peter Nijkamp. In Land Use Policy 71 (2018) 24–35.

[SS:2017] Big Data Analytics Towards a Framework for a Smart City. Srivastava D.K., Singh A. In: Satapathy S., Joshi A. (eds) Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1. ICTIS 2017. Smart Innovation, Systems and Technologies, vol 83. Springer, Cham.

Références des membres du projet

Extraction de données à partir de sources non-structurées

[CLC:2017] Eyes Wide Open: an interactive learning method for the design of rule-based systems. Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon. *International Journal on Document Analysis and Recognition*, Springer Verlag, 2017, 20 (2), pp.91-103.

[PLC:2014] Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases. Baptiste Poirriez, Aurélie Lemaitre, Bertrand Coüasnon. International Conference on Frontiers in Handwriting Recognition (ICFHR), Sept 2014.

Data Mining

[DGG+:2017] Yann Dauxais, Thomas Guyet, David Gross-Amblard, André Happe: Discriminant Chronicles Mining - Application to Care Pathways Analytics. AIME 2017: 234-244

Crowdsourcing, données participatives

[MGM:2016] Panagiotis Mavridis, David Gross-Amblard, Zoltán Miklós:
Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. WWW 2016: 843-853.

Exploration interactive de données

[SYP:2017] Interactive data exploration on top of linguistic summaries. Smits, G., Yager, R. R., & Pivert, O. (2017, July). In *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on* (pp. 1-8). IEEE.

[SPL:2017] Vocabulary elicitation for informative descriptions of classes. Smits, G., Pivert, O., & Lesot, M. J. (2017, June). In *Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), 2017 Joint 17th World Congress of International* (pp. 1-8). IEEE.

Traitement automatique de la langue naturelle

[BF:2016] Categorial Dependency Grammars with Iterated Sequences. Denis Béchet, Annie Foret. LACL2016: 34-51.

[BCC+:2012] Nicolas Béchet, Peggy Cellier, Thierry Charnois, Bruno Crémilleux: Discovering Linguistic Patterns Using Sequence Mining. CICLing (1) 2012: 154-165.

[CFF+:2016] Exploration des Données du Défi EGC 2016 à l'aide d'un Système d'Information Logique. Peggy Cellier, Sébastien Ferré, Annie Foret, Olivier Ridoux. EGC 2016: 443-448.

Equipe d'étude ou de recherche

Directeur d'étude ou responsable scientifique

Nom	GROSS AMBLARD
Prénom	David
Date de naissance	30/07/1973
Titre	PU
Organisme d'appartenance	Université Rennes 1 / Laboratoire IRISA
Adresse	263 avenue du Général Leclerc, CS 74205, 35042 RENNES CEDEX
Courriel	david.gross-amblard@irisa.fr
Téléphone	06 43 76 19 11

Laboratoire ou entité de rattachement du directeur d'étude ou responsable scientifique du projet

Nom du directeur d'étude, de laboratoire ou de l'entité de recherche	IRISA, dirigé par M. JEZEQUEL
Prénom	Jean-Marc
Titre (CR, DR, MCU, MCA, PU PA) ou références	PU
Organisme d'appartenance	Université Rennes 1 / Laboratoire IRISA
Nom développé du cabinet d'étude ou du laboratoire, acronyme et numéro	Institut de recherche en Informatique et Systèmes aléatoires (IRISA, UMR 6074)
Adresse	263 avenue du Général Leclerc, CS 74205, 35042 RENNES CEDEX
Mail	Jean-Marc.Jezequel@irisa.fr
Téléphone / fax	+33 (0)2 99 84 71 00

Membres de l'équipe d'étude (un tableau par personne)

DREAL Bretagne

Nom	PHUNG
Prénom	Fabrice
Références	DSI, Chef de projet GéoBretagne http://cms.geobretagne.fr/
Organisme d'appartenance	DREAL Bretagne

IRISA

Nom	GROSS AMBLARD
Prénom	David
Références	PR, http://www-druid.irisa.fr/David.Gross Amblard
Organisme d'appartenance	Université Rennes 1, IRISA, équipe DRUID

Nom	MIKLOS
Prénom	Zoltan
Références	MCF, http://people.irisa.fr/Zoltan.Miklos/
Organisme d'appartenance	Université Rennes 1, IRISA, équipe DRUID

Nom	JEANTET
Prénom	Ian
Références	Doctorant, https://www.linkedin.com/in/ian-jeantet-64a010122/
Organisme d'appartenance	Université Rennes 1, IRISA, équipe DRUID

Nom	COÜASNON
Prénom	Bertrand
Références	MCF HDR
Organisme d'appartenance	INSA Rennes, IRISA, équipe INTUIDOC

Nom	LEMAITRE
Prénom	Aurélie
Références	MCF
Organisme d'appartenance	Université Rennes 2, IRISA, équipe Intuidoc

Nom	SMITS
Prénom	Grégory
Références	MCF
Organisme d'appartenance	Université Rennes 1, équipe SHAMAN

Nom	FORET
Prénom	Annie
Références	MCF HDR, http://www.irisa.fr/prive/foret/
Organisme d'appartenance	Université Rennes 1, IRISA, équipe SemLIS

Nom	CELLIER
Prénom	Peggy
Références	MCF
Organisme d'appartenance	INSA Rennes, IRISA, équipe SemLIS

SGAR Bretagne

Nom	CHANTRELLE
Prénom	Fanny
Références	Chargée de mission Développement des usages du numérique
Organisme d'appartenance	Préfecture de région Bretagne / SGAR

Calendrier indicatif des travaux jusqu'à la livraison

Mois	**Objectifs (O), Communication (C)**	**Soutiens (hors participants au projet)**	**Recrutement**
Mai	O1 : Identification des corpus structurés et plein texte, entrevues avec acteurs métiers administration, pour scénarios de classification et de prédiction C1 : site web, C2 : organisation atelier	Stage M1 Miage (5 étudiants temps plein, à confirmer)	
Juin	O2 : Spécification d'une preuve de concept		Ingénieur de développement temps plein, sur 7 mois
Juillet	O3 : Développement algorithme prédictif sur corpus structuré et de méthodes d'exploration de données		
Août	O4 : Développement méthode d'étiquetage sur corpus plein texte Validation croisée sur O3, essai d'intégration visualisation DREAL		
Sept.	O5 : Evaluation qualité O4, intégration avec O3, validation croisée O3,O4	- Projet M1 Info (10 étudiants, 1j/semaine, sur 1 an, à confirmer) - Proposition d'un sujet de Master recherche (à confirmer)	
Oct.	O6 : Intégration visualisation DREAL		
Nov.	O7 : Retour vers acteurs métiers administration		
Déc.	O8 : Rapport de conclusion, recommandations techniques, proposition de généralisation de la solution, livraison C3 : atelier, C4 : début rédaction pour publication scientifique		

Prévision de budget :

La trame ci-dessous est indicative sur les ressources qui seront mobilisées par l'entité partenaire ou le laboratoire

Le budget ci-dessous inclus, en ressources propre, la participation de 8 enseignants-chercheurs de l'IRISA, 4 agents de la DREAL Bretagne et 1 agent du SGAR Bretagne sur la durée du projet. Le besoin en financement couvre le recrutement d'un.e ingénieur.e temps plein sur la durée du projet pour les besoins en développement et expérimentation (1 candidat déjà identifié), une unité de calcul supplémentaire pour l'environnement de calcul massif de l'IRISA, l'organisation d'un atelier avec invitations et les frais de présentation des publications scientifiques issues du projet.

	Total des coûts	**Nombre en équivalent journées de travail**
Personnel (A)	57 000	305
Fonctionnement (B)	5000	
Équipement (C)	5000	
Prestations (D)	0	
Coût complet du projet TTC (Total= A+B+C+D)	67 000	

La demande de co-financement de l'étude peut se situer entre 15 000 et 50 000 € et l'entité est invitée à préciser ses autres ressources ou financeurs.

Demande de co-financement du MI	49 000
Autres co-financements	Temps agents

Le projet bénéficiera de l'appui de la Fondation Rennes 1[4] "Progresser, Innover, Entreprendre" qui est la fondation universitaire de l'Université de Rennes 1. Elle œuvre au quotidien à l'ouverture de l'Université de Rennes 1 vers le monde socio-économique. Grâce à de nombreuses actions, la Fondation Rennes 1 vise à favoriser l'innovation et le développement socio-économique. Une de ses actions prioritaires est la promotion et la valorisation des travaux de recherche ce qui est en adéquation avec notre projet. Fort de ses nombreux partenariats, la Fondation Rennes 1 a, depuis 2010, su se positionner comme facilitateur entre l'Université de Rennes 1 et le monde socio-économique.

[4] https://fondation.univ-rennes1.fr/