Rapport final - Projet IAT 2018

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

Nom du directeur ou responsable scientifique : David GROSS AMBLARD (Irisa - Univ Rennes)

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

Objectif

L'objectif du projet IAT est de déterminer dans quelle mesure les documents non structurés produits et publiés par l'administration peuvent être analysés, pour en extraire les descripteurs nécessaires à l'**entraînement d'algorithmes d'apprentissage**. 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 a été l'opportunité de rassembler différents acteurs complémentaires en Bretagne :

- Le laboratoire IRISA[1]: 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) ;

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

Précision des besoins

Le lancement du projet IAT a été effectué lors d'une réunion plénière des participants, le 16 novembre 2018, en présence de Fanny Chanterelle (SGAR Bretagne), David Gross Amblard, Zoltan Miklos (DRUID/IRISA), Ségolène Poisson (Master 1 ISTIC), Bertrand Coüasnon, Aurélie Lemaitre (IntuiDoc/IRISA), Peggy Cellier, Annie Forêt (SemLis/IRISA) (étant excusé M. Fabrice Phung - DREAL Bretagne).

Bertrand Couäsnon (Intuidoc) a présenté le système DEMOS-PI, langage de description de documents utilisant la reconnaissance d'écriture manuscrite (*OCR*). Si les documents sont réguliers, alors il est relativement aisé de programmer des règles DEMOS-PI pour extraire l'information utile, même à partir de PDF image. Pour que cela soit effectif, il est important de veiller à la qualité de la numérisation. Les documents exemples utilisés lors du montage du projet présentent par exemple des inconvénients (comme un taux de compression trop élévé).

S'en est suivi une longue phase de discussion entre les participants, en se basant sur les documents de l'étude cas par cas environnementale. Cette discussion a permis d'identifier les questions importantes suivantes :

- Est-ce que les champs (du formulaire) sont peu nombreux et fixes, ou est-ce la notion de champs peut être généralisée (s'il y a beaucoup de champs ou qu'ils changent de nature d'un document à l'autre).
- Quelle est la meilleure version des documents numérisés accessible ? Il est suggéré de refaire une numérisation de haute qualité si besoin.
- Quelle est l'ancienneté du corpus disponible : sur combien d'années peut-on remonter ? A-t-on une vérité terrain ?
- Comment un expert obtient sa décision sur un document ? En particulier, peut-on décider à partir des champs CERFA ou bien est-ce qu'une lecture des annexes (très complexes) est obligatoire dans la majorité des cas ?
- Quel est le niveau d'accès des documents qui vont être utilisés, et quels sont les délais d'obtention ?

Faisabilité de l'extraction de descripteurs

Extraction automatique

Le financement de ce projet a été l'occasion de recruter une ingénieure spécialisée sur les outils en usage dans l'équipe Intuidoc de l'Irisa. Ainsi, Mme Tarride a travaillé sur les documents administratifs intitulés *Demande d'examen au cas par cas préalable à la réalisation d'une étude d'impact* (CERFA N°14752). Après discussion avec la DREAL, les champs les plus pertinents ont été ciblé, permettant une décision rapide :

- Les coordonnées GPS : la zone d'implantation envisagée est déterminante pour prendre une décision rapide. Ce champ est une case à remplir avec la longitude et la latitude de localisation du projet.
- L'adresse : ce champ donne un complément d'information quant à la localisation GPS. Elle permet de corriger si besoin les coordonnées GPS détectées. C'est un champ libre à remplir avec l'adresse de la zone d'implantation.
- L'auto-évaluation : ce champ contient la réponse à la question : "*Au regard du formulaire rempli, estimez-vous qu'il est nécessaire que votre projet fasse l'objet d'une évaluation environnementale ou qu'il devrait en être dispensé ? Expliquez pourquoi.*". La façon dont cette section est rédigée impacte l'avis donné par des agents administratifs.
- Les cases à cocher : ces champs sont une série de questions concernant la sensibilité environnementale de la zone d'implantation (zone protégée, montagne, zone humide...). Les réponses à ces questions peuvent être *oui* ou *non* et permettent de cibler rapidement les projets qui doivent impérativement se soumettre à une étude d'impact.

Après sélection d'un échantillon représentatif du jeu de données, il a été conduit différentes mesures sur la qualité de l'extraction effectuée (voir figure suivante). La distance choisie est la distance de Levenshtein. Il en ressort naturellement que les meilleurs résultats de reconnaissances sont obtenus pour les documents ayant été numérisés avec la plus forte résolution. La faisabilité de l'extraction des différents champs est clairement établie. Le champ le plus difficile a été la coordonnées GPS (présence de caractères spéciaux, superpositions de caractères). Quelques difficultés peuvent apparaître sur la reconnaissance des cases à cocher, lorsque leur contour est effacé.

Coordonnées géographiques	Basse qualité	Haute qualité
Levenshtein (toute la chaîne)	7.7	6.8
Levenshtein (nombres)	1.3	0.8

Adresse	Basse qualité	Haute qualité
Levenshtein	2.7	0.75
Word Error Rate	13.6	6.6

Cases à cocher	Basse qualité	Haute qualité
Pourcentage de cases reconnues par document	69% (29/42)	98% (41/42)

Evaluation de la reconnaissance des documents selon le niveau de numérisation

Extraction par *crowdsourcing*

Afin d'augmenter la quantité d'information disponible, et de posséder une vérité-terrain d'envergure, nous avons soumis le corpus à une plateforme de crowdsourcing, permettant d'obtenir une annotation "à la main" du corpus, après passage dans un OCR de l'état de l'art pour les cases non traitées par l'approche précédente. Ainsi, la totalité des champs du corpus a pu être obtenue. Il est à noter que le coût unitaire de cette méthode manuelle est bien supérieur au coût de la méthode automatique. Elle ne nécessite cependant pas ou peu d'étude préalable. Ce travail a permis d'obtenir 150 champs sur 1387 documents.

Entraînement d'algorithmes d'apprentissage à partir des descripteurs

A partir des descripteurs obtenus, nous avons entraîné plusieurs modèles d'apprentissage, en particulier le modèle des arbres de décision, permettant une bonne compréhension des prédictions effectuées. Lors d'une étude préliminaire, la prédiction de la décision de la demande CERFA à partir du seul descripteur de la commune visée avait été effectuée. Un score d'environ 50% avait été obtenu, ce qui est considéré comme médiocre. Fort des nouvelles données, en particulier grâce à la prise en compte de 42 « cases à cocher » extraites, la même méthode d'apprentissage obtient des scores situés entre 63% et 65% de prédictions conformes à la réalité-terrain. Le meilleur résultat est obtenu avec la méthode des *random forest,* aux alentours des 70%.
Il est à noter que ce résultat est modeste (un bon résultat étant plus proche des 90 %). Cependant, il est obtenu de façon agnostique : aucune sémantique particulière n'a été injectée dans les données (par exemple, les coordonnées GPS ne sont pas interprétées comme des quantités liées, ou les liens entre communes ne sont pas explicitées). Même si ce n'était pas le but de cet étude, une amélioration substantielle de la prédiction pourrait être obtenue en:

- augmentant considérablement le volume de données exemples,
- injectant de la sémantique,
- exploitant les mots clés dans les formulaires texte,
- Équilibrant le jeu de données, qui possède une population de projets acceptés plus grande que de projets refusés (une hypothèse est que les projets soumis sont déjà de grande qualité),
- utilisant une mesure de comparaison pour l'apprentissage qui tient compte de ce déséquilibre (F-mesure par exemple).

Intégration dans Géobretagne

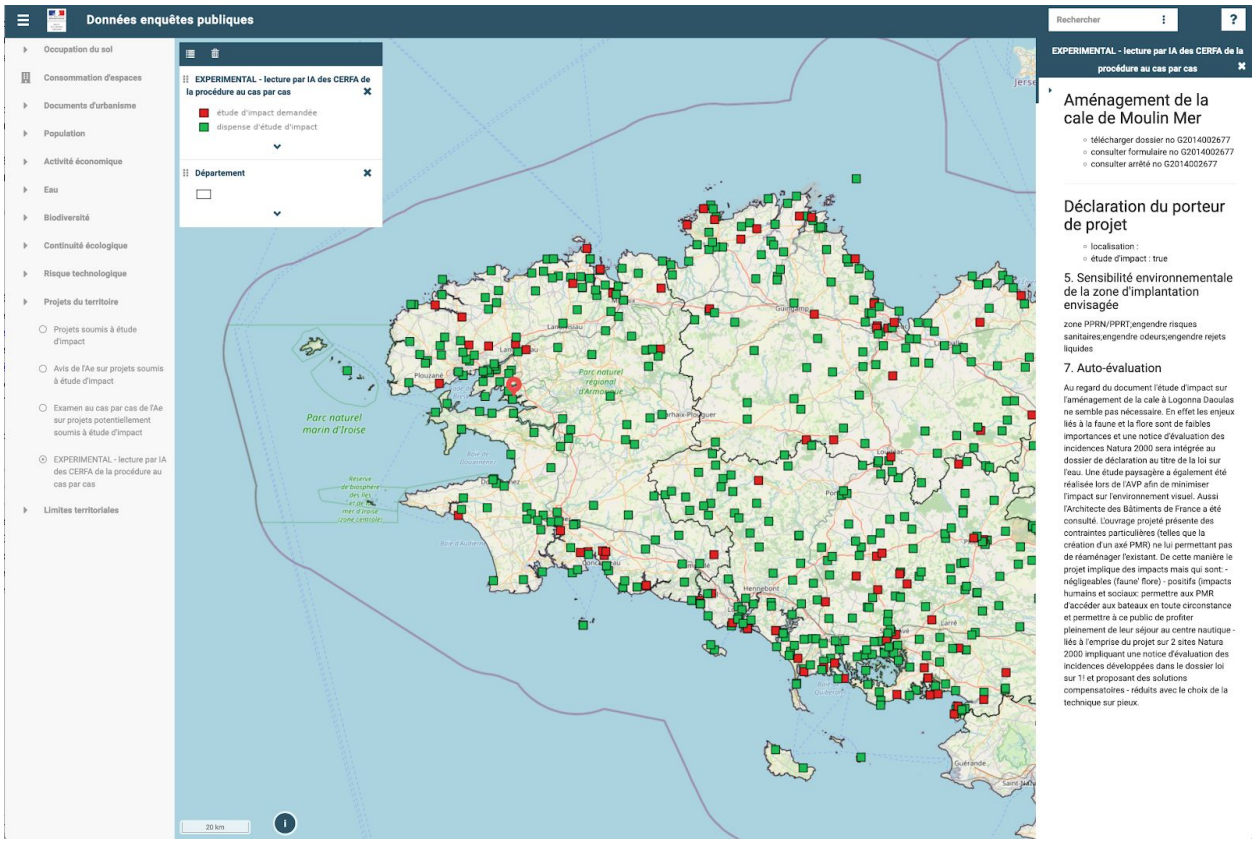
L'objectif de l'intégration des données dans GéoBretagne est de vérifier que les données sortant des algorithmes peuvent être intégrées et utilisées par les agents publics dans l'exercice de leurs missions.

Les données de sortie (CERFA et prédictions) ont été traitées par script puis publiées sur la plateforme geobretagne.fr. Elles deviennent alors ***découvrables*** pour qui en a besoin. GéoBretagne étant relié à la plateforme nationale data.gouv.fr, les données de sortie se sont trouvées automatiquement visibles et découvrables sur le catalogue national de données publiques.

Les données ont ensuite été exposées dans l'application "commissaire enquêteurs" de la DREAL Bretagne. Cette application a pour objectif de fournir aux commissaires enquêteurs les

informations de contexte dont ils pourraient avoir besoin lors d'une enquête publique. L'application est consultable sur http://geobretagne.fr/app/enquetepublique.

Sur cette application, dans le menu à gauche, développer le thème "projets du territoire" et cocher la case "EXPERIMENTAL" : les données de sortie apparaissent sur la carte interrogeable. En cliquant sur un point les données lues par IA depuis les CERFA deviennent consultables.



Capture d'écran des données de numérisation et de prédiction dans l'outils GeoBretagne

Cette expérimentation prouve que, grâce aux normes d'interopérabilité que la France promeut pour les données publiques localisées (directive européenne INSPIRE), des informations provenant de sources différentes et les résultats d'algorithmes de traitement automatique peuvent être facilement superposées et exposées à une cible utilisateur.

Pour ces utilisateurs, le bénéfice est d'accéder à de nouvelles données, potentiellement prédictives ou analytiques, sans changer leurs habitudes. Un premier effet est la réduction du temps passé à rechercher les informations auprès de sources différentes et hétérogènes. Un deuxième effet est d'augmenter la capacité d'analyse croisée des enjeux territoriaux grâce à la

superposition aisée de ces enjeux, présentés directement à l'utilisateur sur une unique interface qui parle son langage.

Ce démonstrateur ouvre d'autres possibilités. Les utilisateurs procèdent par superposition d'enjeux territoriaux pour analyser des projets. Ces enjeux (zonages, points d'intérêt, indicateurs socio-économiques...) sont déjà décrits sous forme de données interopérables. Les modèles prédictifs pourraient donc prendre en compte les principaux enjeux de façon à mieux conseiller l'utilisateur : en lui proposant des dossiers et décisions similaires, ou en l'avertissant sur la présence d'enjeux. Ce potentiel peut s'exercer en contrôle de légalité, en évaluation, mais pourrait aussi être mis à disposition des porteurs de projet en amont, avec pour effet d'améliorer la qualité des dossiers reçus par l'administration.

Conclusion

Le projet AIT a permis de montrer qu'à partir de documents CERFA standardisés et numérisés, il est possible d'extraire des champs d'analyse pertinents pour nourrir des algorithmes d'apprentissage machine. Cette extraction peut être automatique, et peut se faire par simple adaptation d'outils puissants qui sont déjà à disposition. La qualité de la numérisation influe beaucoup sur le résultat. Pour les champs qui échapperaient à cette méthode, une approche par *crowdsourcing* permet de compléter le jeu de données. Les informations ainsi produites sont compatibles avec les méthodes d'apprentissage machine standards. En utilisant les données brutes, et sans injecter de sémantique au modèle, des résultats d'apprentissage raisonnables sont obtenus, mais peuvent être améliorés. Enfin, nous avons montré que grâce aux standards de données, il est possible de présenter ses résultats au sein de plateformes publiques comme GéoBretagne.

Recommandations

- Privilégier les formulaires PDF (PDF natif)
- Privilégier la numérisation en haute résolution
- Augmenter la taille du corpus d'apprentissage, en équilibrant les avis positifs et négatifs

Données complementaires

https://www-druid.irisa.fr/iat/

Annexe: détails techniques pour l'extraction automatique

Données utilisées

Les données sont disponibles à cette adresse sur le site www.geobretagne.fr. Les avis dits "cas par cas" de l'évaluation environnementale peuvent être téléchargés ici https://geobretagne.fr/pub/dreal_b/ae/casparcas/. Les avis correspondant sont disponibles ici https://geobretagne.fr/pub/dreal_b/ae/avis/. Le formulaire CERFA a évolué au fil du temps : il en existe 3 versions (CERFA N°14752*01, 14752*02 et 14752*03).

Le travail s'est focalisé aux documents dont les caractères sont imprimés. On laisse aussi de côté les formulaires remplis en PDF natif car leurs champs peuvent être extraits directement.
Au total, 34 documents compressés ont été traités ainsi que 5 documents en haute résolution.

Travail réalisé

DEMOS-PI est un langage de description de document développé par IntuiDoc. Puisque les documents CERFA sont réguliers, on peut créer des règles logiques pour extraire les zones d'intérêt. Par exemple, les cases à cocher peuvent être définies comme deux composantes connexes de même taille et côte à côte. DMOS peut aussi être utilisé avec un OCR (Reconnaissance Optique de Caractères), ABBYY, qui permet d'obtenir une transcription du document ainsi que la localisation de chaque mot. Par exemple, le champ GPS est toujours situé à droite de l'intitulé "Coordonnées géographiques".

Les règles doivent être suffisamment précises car elles doivent réussir uniquement sur le champ voulu, mais elles doivent aussi ne pas être trop strictes pour réussir sur les 3 versions du CERFA et sur les documents de mauvaise qualité (sur lesquels l'OCR fonctionne moins bien et où des mots clés peuvent ne pas être reconnus).

La qualité de la numérisation a une influence considérable sur la performance de l'OCR.

Chaîne de traitement

A partir des documents PDF on convertit chaque page en fichier JPG grâce à la commande

```
convert -colorspace RGB -density 200 pdf_name -quality 100 jpg_name
```

Ensuite, on utilise l'OCR (ABBYY v9 ou v11) qui est installé sur une machine virtuelle. En sortie, on va obtenir un fichier XML contenant les informations sur le contenu de chaque image (mot reconnu, taille de la police, localisation du mot...). Ce fichier XML est ensuite transformé en calque utilisable par DMOS. On va donc pouvoir utiliser les positions des mots reconnus pour reconnaître des champs. Dans DMOS il faut charger le calque obtenu puis écrire des règles pour détecter chaque type de champ. En cas de succès d'une règle, le contenu du du champ est écrit dans un fichier csv.

Les coordonnées GPS

On commence par chercher un mot clé spécifique à la page des coordonnées géographiques (ex : "4.6" ou "Localisation") afin de ne pas chercher le champ sur les autres page. On cherche ensuite un mot clé du champ (ex : "coordonnées" ou "géographiques") et on délimite la zone de recherche en fonction du mot clé trouvé. Puisque l'OCR ABBYY ne donne pas de bons résultats sur ce champ, on va transformer la zone en imagette et la passer dans un OCR plus performant (tesseract4 -LSTM).

/!\ Si ABBYY intègre du preprocessing, ce n'est pas le cas de tesseract. Il faut donc débruiter l'imagette, la rogner, puis ajouter des marges de 10 px (conformément aux recommandations du wiki de tesseract). Il faut aussi passer une option particulière pour que tesseract ne considère pas l'imagette comme un document entier ("-psm 7" pour signifier que l'imagette est une ligne de texte ou "-digits" pour lui dire qu'on cherche des nombres).

L'adresse

On commence par chercher un mot clé spécifique à la page de l'adresse (ex : "4.6" ou "Localisation"). On cherche ensuite un mot clé du champ (ex : "Adresse" ou "commune(s)") et on délimite la zone de recherche en fonction du mot clé trouvé. On récupère dans le bon ordre tous les mots trouvés par ABBYY dans cette zone.

L'auto-évaluation

On commence par chercher un mot clé spécifique à la page de l'auto-évaluation (ex : "7." ou "Auto-évaluation"). On cherche ensuite un mot clé du champ (ex : "Expliquez" ou "pourquoi") et on délimite la zone de recherche en fonction du mot clé trouvé. On va ensuite récupérer toutes les lignes de texte tant qu'elles ne sont pas trop éloignées les unes des autres. On définit la fin du champ comme étant une ligne de texte suivie d'un espace suffisamment grand. Cela permet de s'adapter aux différences de mise en page sur les 3 versions du CERFA (parfois le champ est grand, parfois il est petit).

Les cases à cocher

On commence par chercher un mot clé spécifique aux pages des cases à cocher. On cherche ensuite le premier couple de cases (vu par DMOS comme des composantes connexes) de même taille. On compte le nombre de pixels noirs dans chaque case, et on en déduit la réponse (la case cochée aura plus de pixels noirs: si elle est à gauche c'est oui, si elle est à droite c'est non). On va ensuite chercher le prochain couple de composantes connexes juste en dessous. Une fois trouvé, on va pouvoir chercher l'intitulé de la question à gauche, et s'arrêter au niveau des cases du dessous pour être sûr de récupérer seulement la question voulue, et pas celle du dessous, comme sur cet exemple :

Post-processing

Une fois qu'on a traiter tous les documents, on obtient une sortie .csv par champ. On utilise Python 3.6.8 pour post-traiter les résultats obtenus.

- Les coordonnées GPS

On a obtenu plusieurs prédictions de champs GPS grâce à :

- DMOS + ABBYY9
- DMOS + ABBYY11
- DMOS + ABBYY9 + tesseract4
- DMOS + ABBYY11 + tesseract4

Il faut maintenant lire chaque fichier de résultat et séparer les champs Lat et Long. Il faut aussi enlever les caractères spéciaux qui ont mal été détectés. On va ensuite calculer des distances à la vérité de chaque résultat et ne garder que le meilleur.

- L'adresse

On enlève juste les caractères spéciaux parasites qui ont éventuellement été détectés ("@", "/", "^"...). On fait aussi en sorte d'obtenir un code postal en 5 chiffres en supprimant les espaces

entre les chiffres (ex : “35000 Rennes” au lieu de “35 000Rennes”). A noter que ce champ est souvent bien détecté.

- L’auto-évaluation

On enlève juste les caractères spéciaux parasites qui ont éventuellement été détectés. Ce champ est souvent correctement détecté.

- Les cases à cocher

On va comparer l’intitulé de chaque question à la liste des vraies questions (Word Error Rate). Lorsque l’intitulé est suffisamment proche d’une vraie question, on change l’intitulé par la vraie question. De cette manière, on peut normaliser la sortie. Lorsque que le champ n’a pas été détecté par DMOS, la réponse est “N/A”.

Toutes ces informations sont regroupées en un fichier JSON.

Note pour utiliser ABBYY

- demander la VM à IntuiDoc (chaque conversion est payante).
- ABBYY9 : Générer le XML :
    - Sur la VM dans le dossier contenant les images à traiter

    make -f ..LP/Finereader/Makefile.example -k (sinon échec)

- Conversion en calque
    - sur machine physique dans le dossier contenant les fichiers XML (java n’est pas installé sur la VM)

    make -f ../LP/Finereader/Makefile.conversion

ABBY11 : Générer le XML

- sur la VM make -f LP/Finereader/Makefile.fr11 (tout marche bien sur cette version) Conversion en calque :

- sur la VM make -f LP/Finereader/Makefile.conversion

DMOS : ne pas oublier la commande envEPFUnicode avant de compiler sinon les accents ne sont pas détectés.