Journées de méthodologie statistique de l'Insee 2022

EXTRACTION AUTOMATIQUE DE DONNÉES ISSUES D'IMAGES SCANNÉES : UNE ILLUSTRATION PAR LES COMPTES SOCIAUX D'ENTREPRISES

Laura GAIMARD (*), Adem KHAMALLAH (**)

(*) Insee, Direction des statistiques d'entreprises
(**) Insee, PSAR Analyse Territoriale

laura.gaimard@insee.fr
adem.khamallah@insee.fr

Mots-clés : Deep learning, machine learning, classification, open data

Domaine concerné : **Analyse des données et *data science***

Résumé

Dans le cadre du règlement européen European Business Statistics et pour répondre à de nombreux besoins nationaux, la direction des statistiques d'entreprises (DSE) de l'Insee produit et diffuse différents agrégats économiques structurels sur les entreprises, notamment celles profilées (EP). Ces agrégats sont élaborés à partir de deux sources de données existantes au niveau des unités légales (UL): les données d'enquêtes, avec les réponses aux enquêtes annuelles (EAP (industrie) et ESA (autres secteurs)) et les données administratives, avec les liasses fiscales transmises par la DGFIP.
Ces données, collectées au niveau des UL sont incomplètes pour la détermination des caractéristiques des EP. D'autres sources et méthodes sont alors utilisées pour constituer les données consolidées sur les entreprises profilées mais le processus actuel est perfectible.

L'analyse d'une autre source de données, les comptes sociaux des entreprises, est nécessaire pour améliorer la qualité des données. Ils comportent toute la documentation comptable et financière de la société qui les publie, sous un format d'images scannées ou de document structuré (format pdf), avec différents tableaux et textes de format différent selon chaque compte social. Actuellement, ils sont analysés manuellement par un nombre important de gestionnaires.
Or, depuis 2020, les comptes sociaux sont mis à disposition en open data par l'institut national de la propriété industriel (INPI), et sont récupérables massivement par le biais d'une interface de programmation applicative (API), ce qui permettrait d'automatiser la collecte de ces informations. L'objectif de cet exposé est de présenter l'expérimentation de l'automatisation de la collecte des données contenues dans un compte social, par l'exemple du tableau des filiales et participations.

Le processus d'extraction des données comporte deux étapes : dans un premier temps, la page comportant le tableau des filiales et participation est prédite parmi l'ensemble des

pages du compte social, dans un second temps, l'emplacement du tableau est prédit au sein de cette page.

Pour la première étape, la base de travail est composée de plusieurs centaines de pages annotées manuellement de la présence ou non du tableau d'intérêt.
Des pré-traitements sont réalisés sur l'ensemble de ces pages, avant la construction du modèle. Ils consistent à récupérer le texte contenu sur l'ensemble des pages par des techniques de reconnaissances optique de caractères (OCR), puis à réduire le nombre de mots à analyser afin d'optimiser les temps de calculs.
Un modèle de classification supervisé est entraîné (Random Forest) sur une partie de la base de travail, avant d'être évalué sur un échantillon test comportant la partie restante de la base. Ce modèle est ensuite appliqué à l'ensemble des pages pré-traitées d'un compte social afin d'obtenir la page recherchée, page ayant la plus grande probabilité de présence du tableau. Celui-ci n'étant pas présent dans tous les comptes sociaux, un seuil minimal de probabilité de présence est calculé pour éviter les cas de présence à tort du tableau.
Ainsi, ce modèle est perfectible, notamment en enrichissant la base de travail de pages annotées pour améliorer l'apprentissage et la validation du modèle, ainsi qu'en analysant manuellement les erreurs de prédictions du modèle.

Pour la seconde étape, le modèle Tablenet (Paliwa et al., 2019), basé sur les réseaux de neurones convolutif, est utilisé. Il prédit l'emplacement du tableau et de ses colonnes sur une image, à partir d'une base d'images comportant des tableaux avec leurs coordonnées dans une image. Cette base d'image provient du jeu de données Marmot, ainsi qu'une annotation manuelle des coordonnées des tableaux étudiés. La qualité de la prédiction est mesurée par la comparaison entre l'intersection de la surface du tableau réel et prédit avec l'union des surfaces du tableau réel et prédit (indice de Jaccard). Ensuite, ces emplacements prédits sont appliqués à l'image pour en extraire l'information des lignes et colonnes par reconnaissance optique de caractères. La qualité de l'extraction est déterminée en comparant le nombre de lignes et colonnes du tableau réel avec celui de l'extraction finale.

Une fois les informations extraites du tableau, une analyse devra être menée pour associer les données avec les variables correspondantes. Ce travail permettra la comparaison avec d'autres productions réalisées par l'Insee : dans le cas du tableau des filiales et participations, des liens capitalistiques obtenus seront comparés aux données du répertoire Lifi pour une identification plus simple des UL.

D'autres instituts nationaux de statistiques, comme StatCan, ont eu des réflexions semblables sur l'extraction automatique de données issues de documents comptables, afin d'obtenir les données sur le chiffre d'affaires et le total du bilan de la société. Leur processus est similaire à celui présenté ici, avec deux étapes : recherche de la page puis du chiffre d'intérêt. Toutefois, leurs documents n'étant pas des images scannées, mais des documents structurés, la seconde partie du processus diffère (techniques de parcours de graphes contre analyse d'images).

Ce processus, une fois mis en production, permettra de récupérer des informations nécessaires à l'amélioration de la qualité des statistiques structurelles d'entreprises (liens financiers, restructurations d'UL, ventilation du chiffre d'affaires par branches agrégées, etc).

Bibliographie

[1] Anurag Bejju, Saeid Malladavoudi, Monica Pickard, Automation of Information Extraction from Financial Statements using Graph-Based Techniques. Travaux pour StatCan, 2019.
[2] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, Lovekesh Vig, TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images. TCS Research, 2020.
[3] Leo Breiman, Random Forests. University of California, Berkeley, 2001.
[4] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, University of Oxford, 2014.