ANALYSE DES OFFRES D'EMPLOI EN LIGNE : COMMENT CODER LE MÉTIER ?

Paul ANDREY, Maxime BERGEAT

Dares, Département Métiers et Qualifications

paul.andrey@ensae-paristech.fr
maxime.bergeat@travail.gouv.fr

Mots-clés : analyse textuelle, appariement, données massives, emplois à pourvoir, *machine learning*, *scraping*

Résumé

Cette communication vise à présenter les travaux expérimentaux effectués par la Dares sur l'analyse de descriptifs des offres postées sur des sites d'offres d'emploi. Dans le cadre du projet européen *ESSNet Big Data – Webscraping job vacancies*, des offres d'emploi publiées en ligne sont collectées de façon automatique avec l'objectif d'améliorer la précision des statistiques sur les emplois à pourvoir grâce à la mobilisation de ces nouvelles données.

Dans cette communication, après un rappel sur le contexte du projet et sur les enjeux concernant la mesure des emplois à pourvoir, les éléments techniques concernant la collecte et la structuration de ces sources de données sur les offres d'emploi encore inexploitées sont présentés : récupération des données par *scraping*, nettoyage de l'information textuelle et structuration de l'information contenue dans les offres d'emploi. Une étude de la codification du métier selon la nomenclature « métiers » des Familles Professionnelles à partir de données récoltées sur 4 sites d'offres d'emploi est ensuite présentée.

Deux approches sont comparées. L'une mobilise le descriptif complet de l'offre et d'éventuelles variables annexes pour prédire le métier à l'aide d'algorithmes de *machine learning*, en se fondant sur des données d'entraînement où le métier est déjà codé. On utilise notamment les méthodes de régression logistique, de forêt aléatoire, et de réseau de neurones, en testant différents paramètres dans les modèles d'apprentissage. L'autre se fonde sur une analyse de proximité lexicale et confronte le libellé de l'offre d'emploi à un référentiel d'appellations métiers constitué par Pôle emploi. Lorsqu'il est possible d'évaluer la performance des résultats, nous atteignons jusqu'à 80% d'exactitude de la codification pour le niveau le plus agrégé de la nomenclature des Familles Professionnelles, qui comporte 22 modalités, et jusqu'à 60% pour le niveau le moins agrégé de cette nomenclature, qui comporte 225 modalités. Plusieurs pistes d'amélioration pour améliorer les performances sont également discutées.

Dans une dernière section, d'autres questions méthodologiques pour utiliser les offres d'emploi en ligne sont évoquées. On pose en particulier en conclusion la question de la validation des méthodes mises en place pour structurer l'information à partir de sources *Big Data* : quelles données de référence utiliser pour entraîner et/ou valider les algorithmes utilisés pour la structuration des données ?