

Sujet de Master 2 Recherche ou École d'ingénieurs 3A

Extraction automatique de documents cibles en utilisant des techniques d'apprentissage artificiel

(Mars - Août 2012)

Lieu : UMR MIA (Mathématiques et Informatique Appliquées) à AgroParisTech
16, rue Claude Bernard, 75005 Paris

Direction : Antoine CORNUÉJOLS (AgroParisTech) (antoine.cornuejols@agroparistech.fr)
Christine MARTIN (AgroParisTech) (christine.martin@agroparistech.fr)

1. Sujet

Le stage se place dans le contexte du projet ExtraEx (*Extraction de paramètres cruciaux pour une aide à la synthèse et à la prise de décision à partir de grandes bases de données bibliographiques. Application à l'estimation des émissions de gaz à effets de serre*) qui associe quatre équipes de AgroParisTech et de l'INRA.

Son objectif est de fournir des outils et des méthodes d'estimation de chiffres à partir de bases de données bibliographiques et d'y associer une mesure d'incertitude. Très spécifiquement, dans le cadre de ce projet, l'objectif est de ré-examiner l'estimation de deux nombres. L'un est la quantité de N₂O produite par hectare de culture en moyenne dans le monde. L'autre est la quantité de méthane produite par les animaux d'élevage. La valeur actuellement admise pour ces deux nombres, et utilisée par le GIEC dans ses prévisions sur le réchauffement climatique, résulte d'analyses datant d'une dizaine d'années à partir de corpus de quelques centaines d'articles. En outre, l'incertitude sur la valeur de ces deux nombres est importante, et, plus encore, il manque un intervalle de confiance fiable autour de ces valeurs. L'impact de ces incertitudes sur les prévisions du GIEC est potentiellement significatif. Dans ce contexte, le projet vise à reprendre et étendre ces travaux en recourant à des méthodes d'apprentissage artificiel, de méta-analyse et de statistiques avancées développées à cet effet et en incluant des données récemment publiées sur ces questions environnementales.

Le stage va concerner l'extraction automatique d'articles pertinents au sein d'une base bibliographique. En utilisant des **techniques d'apprentissage et de fouille de données** appliquées à une base d'articles étiquetés, on cherchera à identifier des mots-clés et une ou des formules logiques associées afin de classer correctement les articles.

Les développements informatiques seront réalisés préférentiellement en Python.

2. Environnement du stage

Le stage se déroulera dans le laboratoire d'informatique de l'UMR MIA (Mathématiques et Informatique Appliquées) d'AgroParisTech au 16 rue Claude Bernard, 75005 Paris. Il sera rémunéré sur la base légale (environ 436 euros/mois).