Stage Master 2

Intégration d'informations sémantiques pour identifier les variables essentielles à partir de données textuelles hétérogènes : application à la Malherbologie

Durée : 5 à 6 mois à partir de février 2020

Encadrement : Sandrine Auzoux (AIDA) et Mathieu Roche (TETIS)

Contact : sandrine.auzoux@cirad.fr et mathieu.roche@cirad

Description :

Les adventices (mauvaises herbes) sont une contrainte majeure de la production agricole tropicale, induisant des pertes de récoltes de 30 à 80%. Le calage des pratiques de désherbage dans les itinéraires techniques nécessite une bonne connaissance de leur comportement. Le développement de l'agroécologie en région tropicale nous amène à considérer les dimensions négatives et positives des adventices.

Le travail proposé dans le cadre de ce stage au Cirad (TETIS/AIDA) consiste à proposer et mettre en œuvre une **méthode automatique d'identification de variables essentielles** pour la gestion des adventices qui implique la mise en place de nouvelles pratiques agricoles et la mobilisation de la biodiversité. Nous définissons les variables essentielles comme une combinaison d'éléments caractéristiques, par exemple *le climat*, *le milieu*, *la localisation* et le *nom vernaculaire*.

Le but du stage est d'identifier, par des méthodes de fouille de textes, les variables essentielles de manière automatique à partir de données textuelles. Dans l'extrait ci-dessous, les variables essentielles du *pissenlit* à extraire sont par exemple une combinaison de *climat tempéré, milieux humides et échelle mondiale.*

> Faire connaissance avec le Pissenlit et ses bienfaits.
>
> L'herbacée, Taraxucum officinale, connue sous le nom de pissenlit, elle est une plante originaire d'Europe de l'Ouest. Le pissenlit pousse à l'état sauvage dans les climats tempérés et milieux humides de toutes les régions du monde, pouvant vivre jusqu'à 2 000 mètres d'altitude.

Dans le processus de fouille de textes à mettre en place, deux verrous scientifiques seront particulièrement étudiés :

- Adapter les méthodes de fouille de textes aux différents types de données mobilisées (scientifique *vs.* grand public).
- Intégrer des ressources sémantiques et scientifiques (par exemple, thésaurus) au processus proposé.

Dans ce cadre, le processus reposera sur 3 grandes étapes qui seront mises en place et évaluées avec des experts du domaine :

1) Acquisition de données textuelles en anglais par des méthodes semi-automatiques (web crawling / web scraping). Deux types de documents seront étudiés : (1) des documents « grand public » issus du web (blogs, sites touristiques, presse) et (2) des publications scientifiques (articles scientifiques).

2) Extraction de variables essentielles dans ces données par des méthodes adaptées au domaine de la Malherbologie. Ces méthodes s'appuieront sur l'intégration de connaissances sémantiques notamment spatiales (par exemple, Geonames, OpenStreetMap, etc.) et thématiques (par exemple, Agrovoc, dictionnaire des plantes, etc.)

3) Evaluation de ces informations dans un cadre pluridisciplinaire et mise en lien avec des bases de données de référence.

Lieu et gratification :

Ce stage basé au Cirad à Montpellier (https://www.cirad.fr/) bénéficie d'une gratification mensuelle de 580 euros.

Profil :

Master 2 ou École d'Ingénieur en Informatique / Science des Données