

Normalisation automatique de variables issues de bases de données en agroécologie

Oussama Mechhour^{*,**,***}, Sandrine Auzoux^{*,**}, Mathieu Roche^{*,***}

*CIRAD, UPR AIDA & UMR TETIS, F-34398 Montpellier, France

**AIDA, Univ Montpellier, CIRAD, La Réunion, France

***TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
oussama.mechhour@cirad.fr

La mesure de similarité entre variables textuelles est un défi reconnu. Un exemple illustratif de l'importance de cette mesure est le "products matching", qui recommande automatiquement des produits similaires aux préférences des clients de manière efficace en termes de temps et de coût. L'objectif de ce travail est de se concentrer sur les aspects méthodologiques avec une démarche finalement assez générique sans détailler les concepts d'agroécologie. Nous nous concentrons sur la mesure de similarité des variables utilisées pour décrire la culture de la canne à sucre en association culturale.

L'objectif principal de ce travail est double :

1. Résoudre la problématique de l'hétérogénéité des variables agronomiques utilisées par les chercheurs, appelée variables sources (chaque chercheur ayant sa propre méthode de nomination et de description de ces variables), grâce à l'aide d'experts pour normaliser les bases de données et lier ces variables aux variables candidates. Le tableau 1 montre des exemples des noms¹ et descriptions des variables sources et candidates (84 variables sources et candidates au total). Pour cela, des mesures lexicales, contextuelles et des combinaison ont été appliquées entre ces variables :
 - a) Dans un premier temps, notre démarche consiste à évaluer la similarité entre chaque variable source et l'ensemble de 84 variables candidates (Auzoux et al., 2023), en les classant de la plus similaire à la moins similaire. Pour cela nous nous appuyons uniquement sur les noms et les descriptions des variables (nous utilisons donc une mesure de similarité approchée).
 - b) Par la suite, nous avons enrichi notre méthode en incorporant 15 articles en anglais représentatifs du domaine de l'agroécologie, constitués manuellement avec l'aide de 3 experts. Ces articles permettent d'apporter des informations contextuelles (sacs de mots). Ces ajouts ont amélioré des résultats récents (Ngaba, 2022), qui utilisaient 122 documents sur le thème de la canne à sucre, rédigés en anglais. Les tableaux 2 et 3 présentent les résultats précédents et actuels fondés sur les mesures de précision au rang i ($p@i$). Ceci permet d'évaluer, dans quelle mesure, la variable candidate qui correspond réellement à la variable source se situe parmi les i premières variables candidates proposées.

1. Le nom de chaque variable source et candidate se compose de deux parties : son nom avant le dernier '_' et son unité de mesure après le dernier '_'.