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 noms [1] 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 '_'.

2. Le développement d'une interface web [2] vise à offrir aux chercheurs un outil pratique, leur permettant d'appliquer les mesures de similarité entre les variables sources et candidates, ainsi que d'autres fonctionnalités.

Plusieurs mesures ont été proposées : une mesure lexicale fondée sur la distance de Levenshtein et des mesures dites contextuelles (c'est-à-dire, TF-IDF, BERT-base) qui comparent respectivement les variables à travers leurs noms et leurs descriptions. Enfin, une méthode de combinaison linéaire intègre les différentes mesures de similarité. Les résultats ont montré que BERT-base (Devlin et al., 2018) avec 2 sous-couches cachées et TF-IDF (Salton et Buckley, 1988) (avec l'enrichissement contextuel de 15 articles scientifiques en anglais) ont respectivement donné de meilleurs résultats (cf. tableaux 2 et 3).

Variable	**Nom**	**Description**
Source	Yield_CAS_t.ha-1	Cane yield (in fresh machinable stem)
Candidate	stem_juice_crop_yield_l.ha-1	stem juice yield

TAB. 1 – *Exemples du nom et de la description d'une variable source et candidate.*

Précision	Levenshtein (noms des variables)	TF-IDF + cosinus (descriptions des variables)	Combinaison
p@1	15.48%	33.33%	44.05%
p@3	19.05%	42.86%	55.95%
p@5	23.81%	51.19%	64.29%
p@10	42.86%	60.71%	73.81%

TAB. 2 – *Résultats précédents en termes de précision (avec articles).*

Précision	BERT-base + cosinus (noms des variables)	TF-IDF + cosinus (descriptions des variables)	Combinaison
p@1	11.90%	29.76%	52.38%
p@3	28.57%	42.86%	66.67%
p@5	36.90%	51.19%	71.43%
p@10	53.57%	64.29%	80.95%

TAB. 3 – *Résultats actuels en termes de précision (avec articles).*

Références

Auzoux, S., B. Ngaba, M. Christina, B. Heuclin, et M. Roche (2023). Experimental variables in sugarcane intercropping in reunion island for data matching. *Data in Brief 46*, 108869.

Devlin, J., M. W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pp. 4171–4186.

Ngaba, B. (2022). Rapport de stage : Couplage d'un modèle de culture avec une plateforme de capitalisation des données issues d'agroécosystèmes à la réunion.

Salton, G. et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management 24*(5), 513–523.

2. https://drive.google.com/file/d/14V9ElV1878GZKbnzTMk1DUBhJv0kLIiC/view?usp=sharing