

Exploration par Text Mining d'Articles en Pathologie Végétale

Cas des Maladies Foliaires de l'Hévéa

 Ayman BENAZZOUZ






Avril - Août
2025



UMR TETIS - CIRAD






Contexte économique

-  Hévéa : principale source de caoutchouc naturel
-  Culture stratégique en régions tropicales
-  Vulnérable aux maladies foliaires fongiques



Problématique de recherche

-  Recherches fragmentées géographiquement
-  Approches cloisonnées disciplinairement
-  Masse d'informations difficile à traiter





Principaux pathogènes

- Pestalotiopsis
- Colletotrichum
- Corynespora
- Phytophthora



Solution proposée

-  Text mining et analyse bibliométrique
-  Intelligence artificielle et TALN



Projet BibEx4MaFhé : Identifier les lacunes disciplinaires et opportunités de collaboration

Objectifs et Résultats Attendus

🎯 Objectifs Principaux

1 Extraction des entités d'intérêt

- Espèces végétales hôtes
- Informations spatiales

2 Identification des disciplines

Cartographier les disciplines scientifiques via approches supervisées et non-supervisées

3 Identification des lacunes disciplinaires

Détecter les approches sous-représentées et opportunités de collaboration

🏆 Résultats Attendus

📄 Système d'extraction automatique

Outil d'extraction d'entités à partir de résumés scientifiques

📖 Cartographie disciplinaire

Vision intégrée des disciplines mobilisées

💡 Recommandations stratégiques

Axes de recherche et partenariats futurs

♻️ Ressources réutilisables

Corpus annotés et lexiques spécialisés

23,614

Articles analysés

2003-2024

Période couverte

Méthodologie

🗄️ Constitution du Corpus

- 📄 23 614 articles scientifiques
- 📅 Période : 2003-2024
- 🌐 Source : **Web of Science**
- ⚙️ Prétraitement linguistique

🔍 Extraction d'Entités

🌿 Espèces hôtes

Noms communs et scientifiques

📍 Informations spatiales

Pays, régions, villes

📍 Techniques d'extraction

- NER (Named Entity Recognition)
- Dictionnaires spécialisés
- Ressources taxonomiques

🧠 Identification des Disciplines

👁️ Approches supervisées

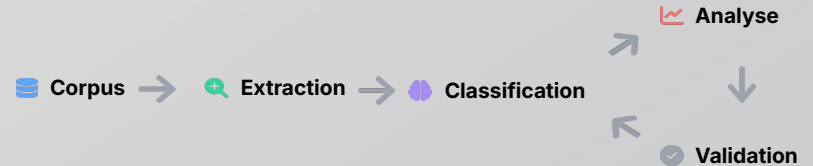
- Définition avec experts domaine
- LLMs pour génération de données
- Classification automatique

🔍 Approches non-supervisées

- Topic Modeling (BERTopic)
- LLMs avec prompts ciblés

📈 Analyse des Lacunes

- 📊 Mesure d'occurrence
- 🔗 Analyse de cooccurrence
- 🔄 Comparaison inter-espèces



Résultats Intermédiaire et Conclusions

☰ Résultats Obtenus

🏗️ Système d'Extraction Développé

- Extraction automatique des espèces végétales
- Reconnaissance des informations spatiales

📄 Cartographie des pays

- Les pays extraits des métadonnées : affiliations des auteurs
- Les pays extraits des résumés : zones d'études
- Cooccurrence : affiliations des auteurs vs zones d'études

🔍 Analyse des mots clés

- Identifier les thématiques dominantes
- Comparaisons inter-espèces réalisées

💡 Conclusions et Impact

🎓 Contributions Scientifiques

- Méthodologie TALN adaptée à la pathologie végétale
- Vision intégrée des recherches sur l'hévéa

➔ Recommandations Futures

- Collaborations interdisciplinaires**
Renforcer les liens entre disciplines identifiées
- Transfert méthodologique**
Adapter les approches d'autres pathosystèmes

🚀 Perspectives d'Extension

- Application à d'autres pathosystèmes
- Développement d'outils prédictifs

🌐 Impact Attendu pour la Recherche en Pathologie Végétale

👥 Collaboration

Renforcement des liens interdisciplinaires

💡 Innovation

Nouvelles approches méthodologiques

🛡️ Résilience

Stratégies durables contre les pathogènes