



جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B



جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediène

Faculté d'Informatique

Département d'Intelligence Artificielle et Sciences
des Données

Mémoire de Master

Spécialité: Master Systèmes Informatiques Intelligents

Thème

Deep learning pour la recherche des experts scientifiques

Sujet Proposé par :

M. BELLAZZOUGUI Djamal

M. CHAA Messaoud

Présenté par :

HOUATI Chakib Mouloud

MEZIANI Serine

Soutenu le : 20 /09/2022

Devant le jury composé de:

M. HAMMAL Youcef

M. KHENNAK Ilyes

Président

Membre

N° Projet : SII_E_012 / 2022

Résumé

La recherche d'experts scientifiques (*scientific expert finding*) est une tâche récurrente dans le milieu académique. En effet, s'exprime souvent le besoin de trouver des encadreurs, membres de jury pour des projets de fin d'études ou de thèses, évaluateurs de propositions de projets de recherches, référés pour des articles soumis à des revues scientifiques ou des membres de comités de programmes de conférences. Cette recherche est rendue très difficile par l'augmentation du nombre de chercheurs et des travaux scientifiques ainsi que par la grande diversification et spécialisation des domaines de recherches scientifiques.

L'objectif de ce travail sera, de concevoir un système de recherche d'experts par des requêtes exprimées en langage naturel. Ce système sera basé sur les données des articles publiés par les chercheurs (titre, résumé, mots-clés). Le système proposé intégrera les techniques les plus récentes de traitement du langage naturel (NLP) basées sur le deep-learning, en particulier la technique d'apprentissage automatique à base de *transformer BERT* (*SciBERT* et *RoBERTa*), développée par *Google*, et publiée en 2018. Nous présenterons trois contributions principales : une nouvelle approche de l'indexation, l'expansion des requêtes à l'aide de la définition, et une méthode de distribution des scores. En plus de cela, la construction d'un nouveau corpus de test basé sur l'ACM.

”Une plateforme nationale de recherche d'experts académiques a été créée à l'issue de ce projet. Ce portail intègre les méthodes que nous avons proposées, et les valide sur le cas de la production scientifique algérienne.

Mots clés :

Scientific Expert Finding ; Deep Learning ; Word Embeddings ; Bibliographic Databases ; Voting Models ; Query Expansion ; BERT ; SciBERT ; RoBERTa ; FAISS.

Abstract

Scientific expert finding is a recurrent task in the academic world. Indeed, it is often necessary to look for supervisors, jury members for end-of-study or thesis projects, evaluators for research project proposals, referees for articles submitted to scientific journals or members of conference program committees. This research is made very difficult by the increase in the number of researchers and scientific work as well as the great diversification and specialization of the scientific research fields.

The objective of this work will be to design an expert finding system that enhanced through the expansion of queries expressed in natural language. This system will be based on the data of the articles published by the researchers (title, abstract, keywords). The proposed system will integrate the most recent techniques in natural language processing (NLP) based on deep-learning, in particular the transformer-based machine learning technique *BERT* (*SciBERT* and *RoBERTa*), developed by *Google*, and published in 2018. We will present three main contributions : a new indexation, expansion of queries by using the definition, and a score distribution method. In addition to this, the construction of a new test corpus based on ACM.

A national platform for expert research was created at the end of this project. This portal integrates the methods that we proposed, and validates them on the case of the Algerian scientific production.

Key words :

Scientific Expert Finding ; Deep Learning ; Word Embeddings ; Bibliographic Databases ; Voting Models ; Query Expansion ; BERT ; SciBERT ; RoBERTa ; FAISS.

Remerciements

*It is with genuine gratitude and warm regard that I dedicate this thesis to my loving parents, **Abdelaziz** and **Salima**, who have been always by my side throughout my life and especially in my educational journey. To my beloved siblings, **Tahar** and little **Nonat**.*

*I would also like to acknowledge my appreciation and gratefulness to my supervisors **Mr. BELLAZZOUGUI Djamal** and **Mr. CHAA Messaoud** for their valuable advises and orientations that allowed us to carry out our work successfully.*

*I sincerely thank the members of the jury, **Mr. HAMMAL Youcef** and **Mr. KHENNAK Ilyes** for the honor they give us by accepting to judge this work.*

*Finally, a special thanks to my aunt **Souhila**, my partner **Chakib**, and to all the people who have reviewed this thesis.*

Remerciements

*J'adresse mes sincères remerciements à mes deux encadreurs, **Mr. BELLAZZOUGUI Djamel** et **Mr. CHAA Messaoud** pour toutes leurs remarques précieuses qui ont guidé notre travail tout au long de notre stage au CERIST.*

*Je remercie également le juré composé de **Mr. HAMMAL Youcef** et **Mr. KHENNAK Ilyes** pour nous avoir honoré de juger notre projet.*

*Je tiens à exprimer toute ma reconnaissance à mes deux chers parents, ma petite soeur **Amel**, et mes grand-parents pour leur soutien inconditionnel durant tout mon parcours académique.*

*J'exprime toute ma gratitude à ceux qui ont revu notre mémoire et corrigé les fautes, incluant **Mme. Souhila**, **Mlle Ikram** ainsi que mes amis **Nesrine** et **Djalil**.*

*Je remercie spécialement **Dr. BOUCENNA Fateh**, qui fut le responsable de notre découverte du sujet de notre PFE. Ainsi que mon oncle **Mouloud** qui a grandement contribué au bon déroulement du projet*

*Je voudrais Pour finir, exprimer mes vifs remerciements envers mes amis **Hocine**, **Ines**, **Chakib**, **Zo**, **Yacine**, **Lint**, sans oublier mon cousin **Raouf**, et ma binôme **Serine** qui m'ont apporté leur soutien moral et intellectuel.*

À ma très chère regrettée grand-mère

Table des matières

Introduction générale	10
1 Organisme d'accueil	12
1.1 Introduction	12
1.2 Historique du CERIST	12
1.3 Missions du CERIST	13
1.4 Produits/Services phares du CERIST	13
1.4.1 PNST	13
1.4.2 ARN	13
1.4.3 ASJP	13
1.4.4 SNDL	14
1.4.5 SYNGEB	14
1.4.6 CERIST WebTV	14
1.5 Les divisions de recherche du CERIST	14
1.5.1 La Division de recherche et développement en Théories et Ingénierie des Systèmes Informatiques (DTISI)	14
1.5.2 Division Sécurité Informatique	15
1.5.3 Division Systèmes d'Information et Systèmes Multimédia	15
1.5.4 Division Réseaux et systèmes distribués	15
1.5.5 Division Recherche et Développement en Sciences de l'Information et Humanités Numériques (DRDHN)	15
1.6 Organigramme et structure du CERIST	16
1.7 Conclusion	17
2 État de l'art	18
2.1 Introduction	18
2.2 Recherche d'information (RI)	18
2.2.1 Les modèles de RI	19
2.2.2 Expansion de la requête	23
2.2.3 Métriques d'évaluation	24
2.3 Réseaux de neurones (RN) dans la RI	25
2.3.1 Définitions	25
2.3.2 Représentation des données	29
2.4 La recherche d'expert	33
2.4.1 La source de l'expertise	33
2.4.2 Les modèles utilisés	34
2.4.3 Les bases de données utilisées	36
2.5 Conclusion	36

3	Approches proposées pour la recherche d'experts académiques	37
3.1	Introduction	37
3.2	Représentation des documents et calcul de similarité	37
3.2.1	Approche de base	37
3.2.2	Indexation par phrases	38
3.2.3	Domaine de l'auteur pour la distribution du score	41
3.2.4	Techniques proposées pour l'expansion de la requête	41
3.3	Conclusion	44
4	Réalisation et évaluation des approches proposées	45
4.1	Introduction	45
4.2	Environnement de travail	45
4.2.1	Environnement matériel	45
4.2.2	Environnement logiciel	46
4.3	Optimisations	48
4.3.1	Multiprocessing	48
4.3.2	CUDA (Compute Unified Device Architecture)	48
4.4	Datasets utilisées	49
4.4.1	Dataset Arxiv + MAG	49
4.4.2	Dataset ACM	50
4.5	Méthodes d'évaluation	50
4.5.1	Évaluation par des requêtes thématiques exactes	50
4.5.2	Évaluation avec approximation de la requête	50
4.6	Évaluation des approches proposées	51
4.6.1	Plongement avec <i>RoBERTa</i>	51
4.6.2	Plongement avec <i>SciBERT</i>	54
4.7	Création d'un nouveau corpus d'évaluation	56
4.7.1	Présentation de la base de données de l'ACM	56
4.7.2	Collecte et nettoyage des données	57
4.7.3	Échantillonnage pondéré pour la sélection des requêtes	58
4.7.4	Évaluation des approches proposées avec le nouveau corpus	59
4.8	Discussion générale	61
4.9	Conclusion	63
5	Un portail pour la recherche d'experts académiques algériens	64
5.1	Introduction	64
5.2	Outils et langage de développement	64
5.3	Présentation du site web	65
5.3.1	Page d'accueil	65
5.3.2	Page de recherche avancée	66
5.3.3	Page des résultats de la recherche	67
5.3.4	Page des informations de l'expert	67
5.4	Exemple de recherche dans notre site web	68
5.5	Conclusion	71
	Conclusion générale	72
	Annexe A Complément de la réalisation et l'évaluation des approches proposées	73
A.1	Collecte et nettoyage des données	73
A.2	Les requêtes qui ont été utilisées dans l'évaluation des deux bases de données	74
A.2.1	Requêtes utilisées pour la base de données Arxiv+MAG	74

A.2.2	Requêtes utilisées pour la base de données ACM	74
A.3	L'architecture des bases de données utilisées	75
A.3.1	Base de données Arxiv + MAG	75
A.3.2	Base de données ACM (International)	76
A.3.3	Base de données OpenAlex + ACM (Algérienne)	77